ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΚΙΝΔΥΝΩΝ

Ο ΡΟΛΟΣ ΤΩΝ BIG DATA ΣΤΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΔΙΑΧΕΙΡΙΣΗΣ ΑΣΦΑΛΙΣΤΙΚΩΝ ΑΠΑΙΤΗΣΕΩΝ

Αθανάσιος Ζέρβας

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 Διαχείριση Κινδύνων

Πειραιάς

Σεπτέμβριος 2024

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

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Ξένος Παναγιώτης
- Καθηγητής Χατζηκωνσταντινίδης Ευστάθιος
- Αναπληρωτής Καθηγητής Τζαβελάς Γεώργιος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και

Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN **ACTUARIAL SCIENCE AND RISK MANAGEMENT**

THE ROLE OF BIG DATA IN OPTIMIZING THE INSURANCE CLAIMS MANAGEMENT PROCESS

Athanasios Zervas

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

March 2024

Περίληψη

Η παρούσα εργασία έχει ως στόχο τη διερεύνηση του ρόλου των Big Data στη βελτιστοποίηση της διαδικασίας διαχείρισης ασφαλιστικών απαιτήσεων και χωρίζεται σε έξι κεφάλαια. Στο 1ο κεφάλαιο γίνεται μια αναφορά στις πλατφόρμες μεγάλων δεδομένων. Στο 2ο κεφάλαιο γίνεται μια ανάλυση στα μεγάλα δεδομένα και ειδικότερα στα 5Vs που τα χαρακτηρίζουν, στην αρχιτεκτονική τους, καθώς και στις προκλήσεις και τις ευκαιρίες που εμφανίζονται για τον κόσμο της ασφάλισης. Στο 3ο κεφάλαιο πραγματοποιείται μια ιστορική αναδρομή από τις συμβατικές μεθόδους ανάλυσης δεδομένων, έως την ανάλυση μεγάλων δεδομένων. Στο 4ο κεφάλαιο αναλύεται η χρήση και η σημασία των μεγάλων δεδομένων στον κόσμο της ασφάλισης. Στο 5ο κεφάλαιο αναλύονται στατιστικές μέθοδοι μάθησης οι οποίοι χρησιμοποιούνται στο πλαίσιο των μεγάλων δεδομένων για την βελτιστοποίηση της διαδικασίας διαχείρισης ασφαλιστικών απαιτήσεων. Στο 6ο Κεφάλαιο παρατίθενται έρευνες, οι οποίες έχουν ως σκοπό να τεκμηριώσουν επιστημονικά την αξιοποίηση της θεωρίας στην πράξη. Τέλος, παρατίθεται ο επίλογος όπου πραγματοποιείται η συλλογή των συμπερασμάτων σχετικά με τον ρόλο που διαδραματίζουν τα μεγάλα δεδομένα στον κόσμο της ασφάλισης.

Λέξεις Κλειδιά: Μεγάλα δεδομένα, Ασφάλιση, διαχείριση απαιτήσεων, στατιστικές μέθοδοι μάθησης.

Abstract

The present thesis aims to investigate the role of Big Data in the optimization of the insurance claims management process and is divided into five chapters. In the 1st chapter, a reference is made to big data platforms. In the 2nd chapter there is an analysis of big data and in particular the 5Vs that characterize it, their architecture as well as the challenges and opportunities that appear for the world of insurance. In chapter 3, a historical review is carried out from conventional data analysis methods to big data analysis. Chapter 4 analyzes the use and importance of big data in the world of insurance. In the 5th chapter, statistical learning methods are analyzed which are used in the context of big data to optimize the insurance claims management process. Chapter 6 presents research that scientifically substantiates the use of theory in practice. Finally, the conclusion is presented where the conclusions regarding the role of big data in the world of insurance are collected.

Keywords: Big data, Insurance, claims management, statistical learning methods.

Περιεχόμενα

Εισαγωγή

Ο όγκος των δεδομένων που δημιουργήθηκαν τα τελευταία χρόνια ήταν άνευ προηγουμένου. Αυτό δεν οφείλεται μόνο στην επικράτηση των διαδικτυακών κοινωνικών δικτύων και στις πανταχού παρούσες συσκευές που είναι συνδεδεμένες στο Διαδίκτυο, αλλά και ως αποτέλεσμα της προόδου της τεχνολογίας σε άλλα πεδία, όπως για παράδειγμα, η αλληλουχία ολόκληρου του γονιδιώματος. Ως εκ τούτου, είναι δίκαιο να πούμε ότι ζούμε στην εποχή των μεγάλων δεδομένων (Big Data). Τα Big Data αναφέρονται σε μεγάλα σύνολα δεδομένων ή ροές δεδομένων που έχουν ξεπεράσει τις δυνατότητές μας για αποθήκευση και επεξεργασία και δεν μπορούν να αναλυθούν με παραδοσιακά μέσα. Πιο συγκεκριμένα, οι προκλήσεις προκύπτουν κυρίως για έναν ή περισσότερους από τους ακόλουθους λόγους (Mayer-Schönberger and Cukier 2013):

- Όγκος: όταν αντιμετωπίζουμε τεράστια δεδομένα σε μέγεθος, π.χ. δεδομένα από ανίχνευση στον ιστό ή δεδομένα αλληλουχίας γονιδιώματος, τα παραδοσιακά συστήματα αποθήκευσης και επεξεργασίας υπολείπονται. Επομένως, πρέπει να δημιουργήσουμε νέα συστήματα, τεχνικές και αλγόριθμους που αποθηκεύουν, ανακτούν και επεξεργάζονται αποτελεσματικά τεράστιους όγκους δεδομένων.
- Ταχύτητα: τα μεγάλα δεδομένα δεν αφορούν μόνο το μέγεθος. Ο υψηλός ρυθμός παραγωγής δεδομένων είναι επίσης σημαντικός. Για παράδειγμα, τα δεδομένα που παράγονται στο Twitter ή στα δίκτυα επικοινωνίας έρχονται με τη μορφή συνεχών ροών δεδομένων με πολύ υψηλό ρυθμό. Πολλά συστήματα απαιτούν την ανάλυση αυτού του είδους δεδομένων σε πραγματικό χρόνο.

- Ποικιλία: μερικές φορές, τα δεδομένα προέρχονται από πολλές πηγές και σε διάφορες μορφές, για παράδειγμα, ως συνδυασμός δομημένων, ημιδομημένων και μη δομημένων δεδομένων. Ως εκ τούτου, είναι σημαντικό να υπάρχουν συστήματα που χειρίζονται διαφορετικά μοντέλα δεδομένων χωρίς συμβιβασμούς στην απόδοση.

Παρουσία αυτών των προκλήσεων, οι παραδοσιακές πλατφόρμες αποτυγχάνουν να δείξουν την αναμενόμενη απόδοση και ως εκ τούτου, είναι ζωτικής σημασίας να εμφανιστούν νέα συστήματα για την αποθήκευση και την επεξεργασία δεδομένων μεγάλης κλίμακας (Warren and Marz, 2015). Σε αυτή την εργασία, εξερευνούμε μερικές από τις νέες τάσεις της τεχνολογίας για το χειρισμό μεγάλων δεδομένων και ειδικότερα στο πλαίσιο της βελτιστοποίησης της διαδικασίας διαχείρισης ασφαλιστικών απαιτήσεων.

Κεφάλαιο 1ο : Πλατφόρμες μεγάλων δεδομένων: προκλήσεις και απαιτήσεις

Μια πλατφόρμα μεγάλων δεδομένων θα πρέπει να παρέχει μέσα για την αποτελεσματική αποθήκευση, ανάκτηση και επεξεργασία τεράστιου όγκου δεδομένων. Μία από τις κύριες προκλήσεις που πρέπει να αντιμετωπίσει μια πλατφόρμα μεγάλων δεδομένων είναι η επεκτασιμότητα. Πιο συγκεκριμένα, η πλατφόρμα θα πρέπει να διαθέτει όσους πόρους απαιτούνται για τη διαχείριση μεγάλων δεδομένων. Υπάρχουν δύο πιθανές λύσεις για να γίνει ένα σύστημα επεκτάσιμο:

1. να κλιμακωθεί κάθετα, με την προσθήκη περισσότερων πόρων σε ένα μόνο μηχάνημα ή
2. να κλιμακωθεί οριζόντια, με την προσθήκη περισσότερων μηχανών σε ένα δίκτυο και να χρησιμοποιήσουν όλους τους συλλογικούς πόρους τους (Fan and Bifet, 2013).

Η αγορά μιας εξαιρετικά ισχυρής μηχανής για κλιμάκωση είναι πιθανώς λιγότερο δύσκολη, αλλά πολύ δαπανηρή. Το πιο σημαντικό, υπάρχει η δυνατότητα κάθετης κλιμάκωσης ενός συστήματος μόνο σε έναν ορισμένο βαθμό, δηλαδή, υπάρχει ένα όριο στο πόσους πόρους μπορεί κάποιος να προσθέσει σε ένα μόνο μηχάνημα και αυτό το όριο είναι πολύ μικρότερο από αυτό που απαιτούν οι περισσότερες εφαρμογές επεξεργασίας μεγάλων δεδομένων. Αντίθετα, η εκμετάλλευση των συλλογικών πόρων ενός δικτύου μηχανών εμπορευμάτων είναι μια οικονομικά και τεχνικά ελκυστική λύση και ως εκ τούτου, η οριζόντια κλιμάκωση είναι η προσέγγιση που ακολουθούν σχεδόν όλες οι υπάρχουσες πλατφόρμες (Kitchin, 2014).

Ωστόσο, λόγω της διανομής δεδομένων και υπολογισμών σε ένα δίκτυο, προκύπτουν νέες προκλήσεις και απαιτήσεις (Chen, Mao and Liu, 2014; Chen and Zhang, 2014):

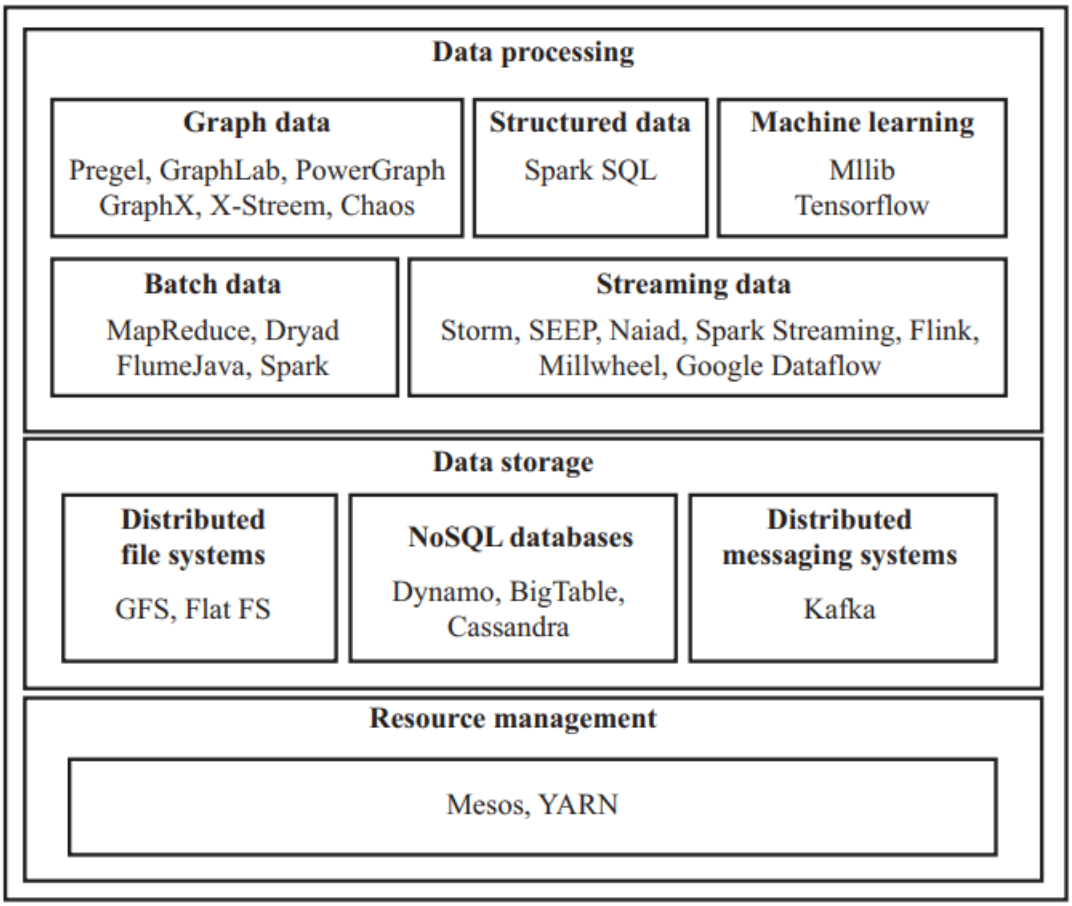
- Ανοχή σφαλμάτων: ένα ή περισσότερα μηχανήματα ενδέχεται να παρουσιάσουν βλάβη κατά την εκτέλεση μιας εργασίας. Ας υποθέσουμε ότι ένα μηχάνημα μπορεί να παραμείνει σε λειτουργία για 1.000 ημέρες. Εάν υπάρχουν 1.000 μηχανές σε ένα δίκτυο, αναμένουμε να παρατηρούμε ένα μηχάνημα που έχει αποτύχει την ημέρα κατά μέσο όρο. Όταν υπάρχουν εκατομμύρια μηχανήματα σε ένα δίκτυο, όπως στους ιστότοπους της Google, ενδέχεται να έχουμε 1.000 βλάβες μηχανημάτων την ημέρα. Ως εκ τούτου, είναι ζωτικής σημασίας η πλατφόρμα να είναι ανθεκτική στις αστοχίες.
- Διαφάνεια: ενώ οι πόροι μιας πλατφόρμας διανέμονται, είναι ευρέως αποδεκτό ότι οι χρήστες θα πρέπει να έχουν την ψευδαίσθηση ότι εργάζονται με ένα μόνο μηχάνημα. Πιο συγκεκριμένα, οι λεπτομέρειες της διαχείρισης πόρων, συμπεριλαμβανομένης της κατανομής πόρων και της εξισορρόπησης φορτίου, θα πρέπει να είναι κρυφές από έναν απλό χρήστη της πλατφόρμας. Αυτή είναι μια από τις απαιτήσεις κάθε πλατφόρμας επεξεργασίας μεγάλων δεδομένων.
- Μοντέλο παράλληλου προγραμματισμού: Τα παραδοσιακά μοντέλα προγραμματισμού υποθέτουν ότι ο κώδικας, τα δεδομένα και όλοι οι απαιτούμενοι πόροι για την εκτέλεση του κώδικα (π.χ. CPU και μνήμη), είναι διαθέσιμα τοπικά. Αυτή η υπόθεση δεν ισχύει πλέον σε οριζόντια κλιμακούμενες πλατφόρμες. Στο νέο μοντέλο, τα δεδομένα και/ή οι λειτουργίες θα πρέπει να παραλληλίζονται, έτσι ώστε διαφορετικά μέρη των δεδομένων να μπορούν να υποβάλλονται σε παράλληλη επεξεργασία. Επιπλέον, δεδομένου ότι η μεταφορά μεγάλων ποσοτήτων δεδομένων μέσω του δικτύου είναι δαπανηρή, είναι συχνά ο κώδικας που αποστέλλεται εκεί όπου αποθηκεύονται τα δεδομένα. Αυτή η αλλαγή παραδείγματος απαιτεί την ανάπτυξη πολλών νέων παράλληλων και κατανεμημένων αλγορίθμων.

- Μοντέλο επικοινωνίας κοινόχρηστο: οι διεργασίες μπορούν να επικοινωνούν μέσω ενός δικτύου με τρεις διαφορετικούς τρόπους: μέσω αποθήκευσης, μνήμης ή δικτύου. Αυτά τα μοντέλα είναι γνωστά ως κοινόχρηστη αποθήκευση, κοινόχρηστη μνήμη και κοινόχρηστο δίκτυο αντίστοιχα.

Επί του παρόντος, υπάρχουν αρκετές πλατφόρμες μεγάλων δεδομένων που παρέχουν τις παραπάνω δυνατότητες. Η ποικιλομορφία αυτών των πλατφορμών μπορεί να κάνει δύσκολη την επιλογή της καλύτερης για την εκτέλεση μιας εργασίας. Ορισμένες πλατφόρμες έχουν σχεδιαστεί για ένα συγκεκριμένο τύπο επεξεργασίας, για παράδειγμα, το GraphLab για την επεξεργασία γραφημάτων και το Storm για την επεξεργασία ροής, ενώ ορισμένες άλλες είναι πιο γενικές και χειρίζονται ένα ευρύτερο φάσμα τύπων επεξεργασίας. Παράδειγμα τέτοιων πλατφορμών περιλαμβάνει το MapReduce, το Spark και το Flink (Almeida and Bernardino, 2015).

Ενώ η συνολική αρχιτεκτονική αυτών των πλατφορμών μοιράζεται πολλά κοινά χαρακτηριστικά, οι ίδιες οι πλατφόρμες μπορούν να ενσωματωθούν σε μια στοίβα, που απεικονίζεται στην Εικόνα 1, η οποία αποτελείται από τα ακόλουθα επίπεδα (Taheri, 2018):

Εικόνα 1: Πλατφόρμες Big Data



Πηγή: Taheri (2018)

- Διαχείριση πόρων (Resource management): αυτό το επίπεδο περιέχει πλατφόρμες που χρησιμοποιούνται για τη διαχείριση πόρων ενός συμπλέγματος και την κοινή χρήση τους μεταξύ των πλατφορμών στα ανώτερα επίπεδα.
- Αποθήκευση δεδομένων (Data storage): οι πλατφόρμες σε αυτό το επίπεδο χρησιμοποιούνται για την αποθήκευση και ανάκτηση μαζικών δεδομένων. Περιλαμβάνουν κατανεμημένα συστήματα αρχείων που διατηρούν δεδομένα σε κατανεμημένους δίσκους, σύστημα ανταλλαγής μηνυμάτων για το χειρισμό

δεδομένων σε πραγματικό χρόνο και βάσεις δεδομένων για τη διατήρηση δομημένων δεδομένων σε κλίμακα.

- Επεξεργασία δεδομένων (Data processing): αυτό το επίπεδο περιέχει τις πλατφόρμες για παράλληλη επεξεργασία δεδομένων σε μεγάλο αριθμό υπολογιστών βασικών προϊόντων. Αυτές οι πλατφόρμες κατηγοριοποιούνται σε μερικές υποομάδες, με βάση την εφαρμογή-στόχο και το μοντέλο εισόδου, για παράδειγμα, για δεδομένα παρτίδας, δεδομένα ροής, δεδομένα γραφημάτων, δομημένα δεδομένα ή για ανάλυση υψηλότερου επιπέδου, π.χ. αλγόριθμους μηχανικής μάθησης.

Κεφάλαιο 2ο : Μεγάλα δεδομένα

Είναι δύσκολο να δοθεί ένας όρος ως γενικός, ευρέως χρησιμοποιούμενος για τα μεγάλα δεδομένα. Σύμφωνα με τη Wikipedia:

«Ο όρος Μεγάλα δεδομένα ή Μεγα-δεδομένα (αγγλικά: Big data) χρησιμοποιείται για να περιγράψει σύνολα δεδομένων τόσο μεγάλα ή σύνθετα που ξεφεύγουν από τις δυνατότητες καταγραφής, αποθήκευσης και ανάλυσης των παραδοσιακών τεχνικών επεξεργασίας δεδομένων.»

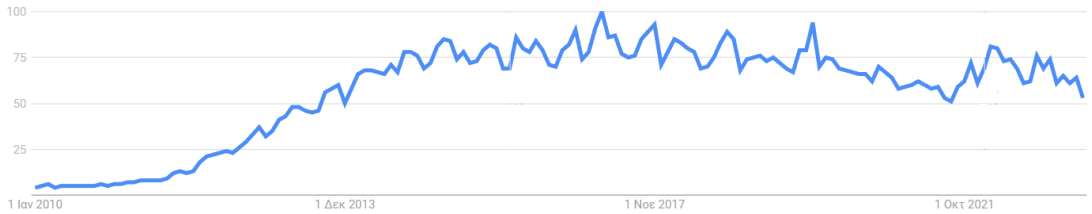
Αυτός ο ορισμός του φαινομένου των μεγάλων δεδομένων παρουσιάζει μια ενδιαφέρουσα άποψη. Επικεντρώνεται στην απώλεια της ικανότητας των κλασικών εργαλείων να επεξεργάζονται τόσο μεγάλους όγκους δεδομένων. Αυτή η άποψη προτάθηκε σε μια έκθεση από την εταιρεία συμβούλων McKinsey and Company που περιγράφει τα μεγάλα δεδομένα ως δεδομένα των οποίων η κλίμακα, η κατανομή, η ποικιλομορφία και η παροδικότητα απαιτούν νέες αρχιτεκτονικές και τεχνικές ανάλυσης που μπορούν να ξεκλειδώσουν νέες πηγές προστιθέμενης αξίας (Manyika et al., 2011).

Φυσικά, αυτή η άποψη επικρατεί έως και σήμερα και ένας καθολικός ορισμός πρέπει να χρησιμοποιεί πιο γενικά χαρακτηριστικά που θα αντέξουν στο χρόνο. Ωστόσο, όπως πολλές νέες έννοιες, υπάρχουν τόσοι ορισμοί όσοι και οι συγγραφείς για το θέμα. Παραπέμπουμε τον αναγνώστη στο άρθρο των Ward and Barker (2013) για μια ενδιαφέρουσα συζήτηση σχετικά με αυτό το θέμα.

Προκειμένου να διαπιστωθεί ο χρόνος γένεσης των μεγάλων δεδομένων, ο τεχνολογικός γίγαντας Google, ως ένας από τους μεγαλύτερους προμηθευτές τους μπορεί να φανεί πολύ χρήσιμος. Με τη βοήθεια του εργαλείου Google Trends, εντοπίστηκε η αύξηση του αριθμού των αναζητήσεων για τον όρο «μεγάλα δεδομένα»

στη διάσημη μηχανή αναζήτησης. Η Εικόνα 2 δείχνει μια σχεδόν εκθετική αύξηση του ενδιαφέροντος των ατόμων που χρησιμοποιούν τη μηχανή αναζήτησης από το 2010 και μετά, ένα σημάδι της νεότητας του όρου και ίσως κάποιο βαθμό έκπληξης από έναν ξαφνικά ανεξέλεγκτο όγκο δεδομένων, όπως ο ορισμός της Wikipedia, που εξακολουθεί να είναι επίκαιρος το 2023, προτείνει.

Εικόνα 2: Διαχρονική εξέλιξη δημοτικότητας του όρου Big Data από το 2010 έως σήμερα



Πηγή: Google Trends (06/04/2023)

Σημείωση: Οι αριθμοί αναπαριστούν το ενδιαφέρον αναζήτησης σε σχέση με το υψηλότερο σημείο του γραφήματος για τη δεδομένη περιοχή και χρονική περίοδο. Η τιμή 100 αντιστοιχεί στην υψηλότερη δημοτικότητα για τον όρο. Η τιμή 50 σημαίνει ότι ο όρος έχει τη μισή δημοτικότητα.

Ωστόσο, η βιβλιογραφία χρησιμοποιεί ευρέως τον όρο Big Data από το 1998, για να συσχετίσουν μια μελλοντική ανάπτυξη ποσοτήτων δεδομένων και βάσεων δεδομένων προς μεγαλύτερες και ευρύτερες κλίμακες (Diebold, 2012; Fan and Bifet, 2013). Το άρθρο αναφοράς, το οποίο αναφέρεται ευρέως από την επιστημονική κοινότητα, χρονολογείται από το 2001 και αποδίδεται στον Doug Laney από την εταιρεία συμβούλων Gartner (Laney, 2001). Περιέργως, το έγγραφο δεν αναφέρει ποτέ τον όρο μεγάλα δεδομένα, αν και διαθέτει τον χαρακτηρισμό αναφοράς τριών Vs: όγκος, ταχύτητα και ποικιλία. Ο «όγκος» περιγράφει το μέγεθος των δεδομένων, ο όρος «ταχύτητα» αποτυπώνει την ταχύτητα με την οποία παράγονται, κοινοποιούνται και πρέπει να υποστούν επεξεργασία, ενώ ο όρος «ποικιλία» αναφέρεται στην ετερογενή

φύση αυτών των νέων ροών δεδομένων. Τα περισσότερα άρθρα συμφωνούν στα βασικά τρία Vs (βλ. Chen, Mao and Liu, 2014; Fan and Bifet, 2013; Fan, Han and Liu, H., 2014), στα οποία προστίθεται το τέταρτο V της αξιοπιστίας (που αποδίδεται στην IBM, καθώς και το πέμπτο V, της αξίας (Corlosquet-Habart and Janssen, 2018).

Ο όρος «αξιοπιστία» εστιάζει στην αξιοπιστία των διαφόρων δεδομένων. Πράγματι, τα δεδομένα μπορεί να είναι λανθασμένα, ελλιπή ή πολύ παλιά για την προβλεπόμενη ανάλυση. Το πέμπτο V μεταφέρει το γεγονός ότι τα δεδομένα πρέπει πάνω από όλα να δημιουργούν αξία για τις εμπλεκόμενες εταιρείες ή την κοινωνία γενικότερα. Από αυτή την άποψη, ακριβώς όπως ορισμένοι συγγραφείς μας υπενθυμίζουν ότι οι μικροί όγκοι μπορούν να δημιουργήσουν αξία, δεν πρέπει να ξεχνάμε ότι οι εταιρείες, υιοθετώντας πρακτικές κατάλληλες για μεγάλα δεδομένα, πρέπει κυρίως να αποθηκεύει, να επεξεργάζεται και να δημιουργεί έξυπνα δεδομένα. Ίσως θα έπρεπε να μιλάμε για έξυπνα δεδομένα και όχι για μεγάλα δεδομένα; (Gu and Zhang, 2014)

2.1. Χαρακτηρίζοντας τα μεγάλα δεδομένα με τα πέντε Vs

Στην αρχική αξιολόγηση του φαινομένου των μεγάλων δεδομένων, θα πρέπει να σημειωθεί ότι το πλαίσιο 3 Vs όγκου, ταχύτητας και ποικιλίας, που διαδόθηκε από την ερευνητική εταιρεία Gartner, είναι πλέον στάνταρ (Laney, 2001). Θα ξεκινήσουμε λοιπόν με αυτά τα κλασικά 3 Vs πριν εξετάσουμε και άλλα V, τα οποία θα αποδειχθούν χρήσιμα για την ανάπτυξη αυτής της αρχικής περιγραφής.

2.1.1. Ποικιλία (Variety)

Η ποικιλία των Μεγάλων Δεδομένων αναφέρεται στους διαφορετικούς τύπους δεδομένων που είναι διαθέσιμα σήμερα. Τα δεδομένα προέρχονται από παντού, για παράδειγμα (Mayer-Schönberger and Cukier, 2013):

- κείμενα, φωτογραφίες και βίντεο (Διαδίκτυο κ.λπ.)
- χωροχρονικές πληροφορίες (κινητές συσκευές, έξυπνοι αισθητήρες κ.λπ.)
- μεταδεδομένα σε τηλεφωνικά μηνύματα και κλήσεις (κινητές συσκευές κ.λπ.)
- ιατρικές πληροφορίες (βάσεις δεδομένων ασθενών, έξυπνα αντικείμενα κ.λπ.)
- αστρονομικά και γεωγραφικά δεδομένα (δορυφόροι, επίγεια παρατηρητήρια κ.λπ.)·
- δεδομένα πελάτη (βάσεις δεδομένων πελατών, αισθητήρες και δικτυωμένα αντικείμενα κ.λπ.).

Τα ελάχιστα παραδείγματα που αναφέρονται παραπάνω απεικονίζουν την ετερογένεια των πηγών και των δεδομένων - «κλασικά» δεδομένα όπως αυτά που παρατηρήθηκαν πριν από την εποχή των μεγάλων δεδομένων, προφανώς, καθώς και σήματα βίντεο, σήματα ήχου, μεταδεδομένα κ.λπ. (Mayer-Schönberger and Cukier, 2013)

Αυτή η ποικιλομορφία περιεχομένου έχει επιφέρει μια αρχική αλλαγή παραδείγματος από δομημένα σε μη δομημένα δεδομένα. Στο παρελθόν, πολλά δεδομένα θα μπορούσαν να θεωρηθούν δομημένα με την έννοια ότι θα μπορούσαν να αποθηκευτούν σε σχεσιακές βάσεις δεδομένων (Zikopoulos and Eaton, 2011).

Αυτός ήταν ο τρόπος με τον οποίο αποθηκεύονταν τα δεδομένα πελατών ή εμπορικών στοιχείων. Σήμερα, μεγάλο ποσοστό δεδομένων δεν είναι δομημένο (φωτογραφίες, ακολουθίες βίντεο, ενημερώσεις λογαριασμού, καταστάσεις κοινωνικών δικτύων, συνομιλίες, δεδομένα αισθητήρων, εγγραφές κ.λπ.) (Sakr and Gaber, 2014).

2.1.2. Όγκος (Volume)

Εάν κάποιος ζητήσει από μια σειρά διαφορετικών ανθρώπων να ορίσουν τα Μεγάλα Δεδομένα, οι περισσότεροι από αυτούς θα αναφέρουν την έννοια του μεγέθους, του όγκου ή της ποσότητας. Αρκεί κάποιος να κλείσει τα μάτια του και να φανταστεί τον αριθμό των μηνυμάτων, των φωτογραφιών και των βίντεο που ανταλλάσσονται ανά δευτερόλεπτο παγκοσμίως (Sakr and Gaber, 2014). Παράλληλα με το αυξανόμενο ενδιαφέρον για την έννοια των μεγάλων δεδομένων στη μηχανή αναζήτησης Google, η χρήση του Διαδικτύου έχει επίσης εκτοξευθεί μέσα σε λίγα μόλις χρόνια (Hariri, Fredericks and Bowers, 2019), όπως μαρτυρεί ο ετήσιος αριθμός αναζητήσεων στο Google (Πίνακας 1).

Πίνακας 1: Ετήσια στατιστικά Google

Έτος	Ετήσιος αριθμός αναζητήσεων	Μέσος όρος αναζητήσεων ανά ημέρα
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000	9,800

Πηγή: Taheri (2018)

Η έκρηξη στη χρήση του Διαδικτύου, και ιδιαίτερα στο κινητό Διαδίκτυο, όπως έγινε δυνατή από smartphone και πρότυπα υψηλής ταχύτητας, οδήγησε σε μια ασταμάτητη αύξηση των όγκων δεδομένων, προς μονάδες που σίγουρα οι περισσότεροι ανακάλυψαν πρόσφατα: gigabyte, terabyte, petabytes, exabyte και ακόμη και zettabyte (ένα zettabyte είναι 1021 byte!) (Taheri, 2018).

Σύμφωνα με μια ετήσια έκθεση για το Διαδίκτυο των Πραγμάτων (GSM, 2015), μέχρι το τέλος του 2015, υπήρχαν 7,2 εκατομμύρια συνδέσεις κινητής τηλεφωνίας, με τις προβλέψεις μόνο για smartphone να ξεπερνούν τα 7 εκατομμύρια το 2019. Αυτός ο εκτεταμένος όγκος δεδομένων είναι αυτό που οδήγησε στο φαινόμενο των μεγάλων δεδομένων. Καθώς τα σημερινά καταστήματα δεδομένων δεν μπορούν να απορροφήσουν τέτοια αύξηση του όγκου δεδομένων, οι εταιρείες, οι μηχανικοί και οι ερευνητές αναγκάστηκαν να δημιουργήσουν νέες λύσεις, ιδίως προσφέροντας κατανεμημένη αποθήκευση και επεξεργασία αυτών των μαζών δεδομένων (Dhamodharavadhani, Gowri and Rathipriya, 2018).

Τα μέρη που αποθηκεύουν αυτά τα δεδομένα, τα διάσημα κέντρα δεδομένων, εγείρουν επίσης σημαντικά ερωτήματα όσον αφορά την κατανάλωση ενέργειας. Μια έκθεση υπογραμμίζει το γεγονός ότι τα κέντρα δεδομένων που χειρίζονται αμερικανικά δεδομένα κατανάλωσαν 91 δισεκατομμύρια kWh ηλεκτρικής ενέργειας το 2013, που ισοδυναμεί με την ετήσια παραγωγή 34 μεγάλων σταθμών ηλεκτροπαραγωγής με καύση άνθρακα (Delforge, 2015). Αυτός ο αριθμός είναι πιθανό να φτάσει τα 140 δισεκατομμύρια το 2020, που ισοδυναμεί με την ετήσια παραγωγή 50 σταθμών ηλεκτροπαραγωγής, κοστίζοντας στον αμερικανικό πληθυσμό 13 δισεκατομμύρια δολάρια ετησίως σε λογαριασμούς ηλεκτρικής ενέργειας. Αν προσθέσουμε σε αυτό την εκπομπή 100 εκατομμυρίων μετρικών τόνων CO2 ετησίως, είναι εύκολο να καταλάβουμε γιατί οι μεγάλοι οργανισμοί άρχισαν πολύ γρήγορα να παίρνουν σοβαρά

αυτό το πρόβλημα, όπως αποδεικνύεται από τη συχνή εγκατάσταση κέντρων δεδομένων σε ψυχρές περιοχές σε όλο τον κόσμο, με έξυπνα συστήματα ανακύκλωσης φυσικής ενέργειας (Corlosquet-Habart and Janssen, 2018).

2.1.3. Ταχύτητα (Velocity)

Το τελευταίο από τα τρία ιστορικά Vs, το V για την ταχύτητα, αντιπροσωπεύει αυτό που πιθανώς πιο φυσικά θα ονομαζόταν ταχύτητα. Καλύπτει επίσης πολλαπλά στοιχεία και είναι εγγενές στο φαινόμενο των μεγάλων δεδομένων. Αυτό είναι ξεκάθαρο από τα παραπάνω σχήματα σχετικά με την ανάπτυξη της έννοιας και του όγκου των δεδομένων, όπως μια ταινία σε γρήγορη μετάβαση. Η ταχύτητα μπορεί να αναφέρεται στην ταχύτητα με την οποία παράγονται τα δεδομένα, στην ταχύτητα με την οποία μεταδίδονται και επεξεργάζονται, καθώς και στην ταχύτητα με την οποία μπορούν να αλλάξουν μορφή, να δώσουν αξία και, φυσικά, να εξαφανιστούν. Σήμερα, πρέπει να αντιμετωπίσουμε μεγάλα κύματα μαζών δεδομένων που πρέπει να υποβληθούν σε επεξεργασία σε πραγματικό χρόνο. Αυτά τα ηλεκτρονικά επεξεργασμένα δεδομένα επιτρέπουν στους λήπτες αποφάσεων να κάνουν στρατηγικές επιλογές που δεν θα γνώριζαν καν στο παρελθόν (Hariri, Fredericks and Bowers, 2019).

2.1.4. Προς τα πέντε Vs: εγκυρότητα (veracity) και αξία (value)

Ένας εμπλουτισμένος ορισμός των μεγάλων δεδομένων διαμορφώθηκε γρήγορα με την εμφάνιση ενός τέταρτου στοιχείου, του V της εγκυρότητας (veracity), που αποδόθηκε στην IBM (Corlosquet-Habart and Janssen, 2018). Η λέξη εγκυρότητα μας επαναφέρει στην ποιότητα των δεδομένων, μια ζωτική ιδιότητα για όλες τις διαδικασίες αναζήτησης δεδομένων. Και πάλι, αυτή η έννοια καλύπτει διάφορες πτυχές, όπως η

ανακρίβεια, η ατέλεια, η ασυνέπεια και η αβεβαιότητα. Σύμφωνα με την IBM, η κακή ποιότητα δεδομένων κοστίζει κατά μέσο όρο 3,1 τρισεκατομμύρια δολάρια ετησίως. Η εταιρεία προσθέτει ότι το 27% των ερωτηθέντων δεν είναι σίγουροι για τις πληροφορίες που εισάγουν και ότι ένας στους τρεις υπεύθυνους λήψης αποφάσεων έχει αμφιβολίες σχετικά με τα δεδομένα στα οποία βασίζει την απόφασή του (Sakr and Gaber, 2014). Πράγματι, η ποικιλία των ροών δεδομένων, οι οποίες συχνά δεν είναι δομημένες, περιπλέκει τη διαδικασία πιστοποίησης δεδομένων. Αυτό φέρνει στο νου, για παράδειγμα, την ποιότητα των δεδομένων στο κοινωνικό δίκτυο Twitter, του οποίου η επιβεβλημένη μορφή 140 χαρακτήρων δεν προσφέρεται για ακριβή πεζογραφία που μπορεί εύκολα να αναγνωριστεί με εργαλεία αυτόματης επεξεργασίας φυσικής γλώσσας (Corlosquet-Habart and Janssen, 2018).

Η πιστοποίηση δεδομένων αποτελεί προϋπόθεση για τη δημιουργία αξίας, η οποία αποτελεί το πέμπτο V που έχει καθιερωθεί στις σύγχρονες πρακτικές. Η ικανότητα αποθήκευσης, κατανόησης και ανάλυσης αυτών των νέων κυμάτων μεγάλου όγκου, υψηλής ταχύτητας, ποικίλων δεδομένων και η διασφάλιση αξιοπιστίας ενσωματώνοντάς τα σε ένα οικοσύστημα επιχειρηματικής ευφυΐας, αναμφίβολα θα επιτρέψει σε όλες τις εταιρείες να δημιουργήσουν νέες ενότητες συμβουλών αποφάσεων (για παράδειγμα, προγνωστική ανάλυση) με υψηλή προστιθέμενη αξία. Ένα εντυπωσιακό παράδειγμα αφορά τις αμερικανικές πωλήσεις αθλημάτων και εισιτηρίων που επί του παρόντος βασίζονται σε δυναμικές μεθόδους τιμολόγησης που ενισχύονται από ιστορικά δεδομένα και δεδομένα σε πραγματικό χρόνο (Ghasemaghaei and Calic, 2019).

Όπως πολλές άλλες αμερικανικές αθλητικές ομάδες, η ομάδα μπέιζμπολ San Francisco Giants έχει προσαρμόσει έτσι το σύστημα έκδοσης εισιτηρίων αγώνων για να κάνει χρήση μεγάλων δεδομένων, κάνοντας χρήση των υπηρεσιών της εταιρείας QCUE για

να δημιουργήσουν αλγοριθμικές τεχνικές συναλλαγών εμπνευσμένες από αεροπορικές εταιρείες. Οι τιμές των εισιτηρίων ενημερώνονται σε πραγματικό χρόνο σε συνάρτηση με την προσφορά και τη ζήτηση. Ειδικότερα, τα ιστορικά δεδομένα για την ποιότητα των αγώνων και των συμμετοχών χρησιμοποιούνται για την προσαρμογή των τιμών των εισιτηρίων για τη βελτιστοποίηση της κατοχής θέσεων/γηπέδων και των κερδών της εταιρείας. Στον ιστότοπό τους, οι QCUE αναφέρουν πιθανή αύξηση κερδών έως και 46% σε σύγκριση με το προηγούμενο σύστημα (Corlosquet-Habart and Janssen, 2018).

Σε παγκόσμιο επίπεδο, τα μεγάλα δεδομένα αντιπροσωπεύουν μια επικερδή επιχείρηση. Το Ινστιτούτο McKinsey έχει προτείνει ότι ακόμη και η απλή χρήση των δεδομένων τοποθεσίας πελατών θα μπορούσε να αποφέρει ένα δυνητικό ετήσιο πλεόνασμα καταναλωτή 600 δισεκατομμυρίων δολαρίων (Manyika et al., 2011). Η συμβουλευτική ομάδα Wikibon εκτιμά ότι η αγορά μεγάλων δεδομένων, που περιλαμβάνει υλικό, λογισμικό και συναφείς υπηρεσίες, θα αυξηθεί από 19,6 δισεκατομμύρια δολάρια το 2013 σε 84 δισεκατομμύρια δολάρια το 2026 (Kelly, 2015).

2.1.5. Άλλα πιθανά Vs

Ανατρέχοντας στον τεράστιο αριθμό άρθρων αφιερωμένων στο θέμα, ο αναγνώστης σύντομα συνειδητοποιεί ότι κάθε συγγραφέας μπαίνει στον πειρασμό να προσθέσει το δικό του προσωπικό V, καθένας από τους οποίους συνεισφέρει στις διάφορες πτυχές των μεγάλων δεδομένων. Έτσι, οι όροι μεταβλητότητα (variability) και εγκυρότητα (validity), που σχετίζονται άμεσα με τις προηγούμενες έννοιες της ποικιλίας και της εγκυρότητας, μπορούν επίσης να προστεθούν στη λίστα. Η λέξη μεταβλητότητα

εστιάζει στην ευέλικτη φύση των δεδομένων, η οποία μπορεί να αλλάξει με την πάροδο του χρόνου, ενώ η εγκυρότητα είναι μια πιο ρητή αναφορά σε μια διαδικασία πιστοποίησης κλασικών δεδομένων (Corlosquet-Habart and Janssen, 2018; Hariri, Fredericks and Bowers, 2019; Jane and Ganesh, 2019).

Τέλος, αξίζει τον κόπο να αναφέρουμε ένα τελευταίο V, την οπτικοποίηση (visualization). Το V της ορατότητας (visibility) μερικές φορές συνδέεται με αυτό. Τα μεγάλα δεδομένα, με όλα τα χαρακτηριστικά τους όπως περιγράφηκαν μέχρι τώρα, απαιτούν νέες μορφές οπτικοποίησης για να κάνουν τα δεδομένα κατανοητά και εμφανή για τους υπεύθυνους λήψης αποφάσεων. Αυτό μπορεί να κυμαίνεται από απλά εργαλεία αναφοράς που προσφέρουν μια γενική άποψη των κύριων χαρακτηριστικών δεδομένων έως πιο προηγμένες μεθόδους που συνδυάζουν οπτικοποίηση και ανάλυση δεδομένων. Για παράδειγμα, οι τεχνικές οπτικοποίησης με γραφήματα που καταδεικνύουν τις περίπλοκες σχέσεις μεταξύ των συντελεστών στα κοινωνικά δίκτυα, των πελατών, των κοινοτήτων ή των φυσικών ομάδων που σχηματίζονται είναι πλέον συνηθισμένες (Naeem et al., 2022).

2.2. Αρχιτεκτονική

Η εποχή των μεγάλων δεδομένων πείθει τις επιχειρήσεις όλων των μεγεθών να εφαρμόσουν διαδικασίες για να βοηθήσουν στη λήψη αποφάσεων με βάση την ανάλυση δεδομένων. Η πρόβλεψη του τι θα ικανοποιήσει έναν πελάτη, η βελτιστοποίηση των διαδικασιών και γενικότερα, η παραγωγή αξίας από δεδομένα, έχουν πλέον γίνει ουσιαστικά για κάθε επιχείρηση που θέλει να παραμείνει ανταγωνιστική. Αν και αυτές ήταν πάντα κεντρικές προκλήσεις για τους ασφαλιστές, δεν επηρεάζονται λιγότερο από το πιο περίπλοκο περιβάλλον της οικονομίας

δεδομένων. Οι αυξανόμενοι όγκοι δεδομένων, διαφόρων διαφορετικών φύσεων, με μεταβλητή διάρκεια ζωής και διαφορετικής ποιότητας, που θέλουμε να διερευνήσουμε σε πραγματικό χρόνο, επηρεάζουν τα εργαλεία που χρησιμοποιούνται, τα οποία συνεχίζουν να εξελίσσονται (Sangeetha and Sudha Sadasivam, 2019).

Το επιστημονικό και τεχνικό περιβάλλον γίνεται πλουσιότερο και πιο περίπλοκο μέρα με τη μέρα. Νέοι αλγόριθμοι σχεδιάζονται για την αντιμετώπιση προβλημάτων και δημιουργούνται νέα εργαλεία για τη δοκιμή και την εφαρμογή τους. Στο πλαίσιο αυτό, το κύριο καθήκον των εταιρειών είναι να ενσωματώσουν αυτές τις καινοτομίες μαζί με τα υπάρχοντα εργαλεία προκειμένου να ενσωματώσουν νέες διαδικασίες ανάλυσης προγνωστικών δεδομένων με τις υπάρχουσες επιχειρηματικές διαδικασίες. Αυτό απαιτεί χρόνο και τεχνογνωσία, για να καθοριστεί το έργο, να εκτελεστεί και, στη συνέχεια, να συντηρηθεί και να ενημερωθεί (Balas et al., 2019).

2.2.1. Μετανάστευση προς μια στρατηγική προσανατολισμένη στα δεδομένα

Υπάρχουν ακόμη πολύ λίγες εταιρείες που μπορούν να καυχηθούν ότι έχουν μεταναστεύσει προς μια στρατηγική προσανατολισμένη στα δεδομένα. Οι Martinez, Viles and Olaizola (2021) προτείνουν τέσσερις αναγνωρίσιμες φάσεις υιοθέτησης μεγάλων δεδομένων:

1. πειραματισμός με την πλατφόρμα μεγάλων δεδομένων.
2. υλοποίηση: ανάπτυξη περιπτώσεων πρώτης χρήσης.
3. επέκταση: ανάπτυξη σε περιπτώσεις πολλαπλών χρήσεων.
4. βελτιστοποίηση: ενοποίηση με το επιχειρηματικό σύστημα πληροφορικής.

Η φάση πειραματισμού αφορά στο όταν διερευνάται η δυνατότητα χρήσης μιας υποδομής μεγάλων δεδομένων. Ο στόχος σε αυτή τη φάση είναι η εγκατάσταση και η

διαμόρφωση. Ο κύριος στόχος είναι να δούμε πόσο συμβατή είναι η τεχνολογία με την υπάρχουσα αρχιτεκτονική. Τέτοιοι πειραματισμοί δεν χρειάζεται να κοστίζουν πολύ, επειδή το μόνο που απαιτείται είναι μερικοί διακομιστές κατώτερου εύρους εξοπλισμένοι με λογισμικό ανοιχτού κώδικα όπως το Hadoop/Spark. Αυτή η πειραματική φάση έχει πολύ συχνά ως αποτέλεσμα τη χρήση ενός επιπέδου αποθήκευσης δεδομένων με προϋπάρχοντα δεδομένα, στο οποίο προστίθεται ένα νέο επίπεδο διαχείρισης δεδομένων, όπως ερωτήματα βάσης δεδομένων (Corlosquet-Habart and Janssen, 2018).

Μόλις η τεχνική πλατφόρμα κατακτηθεί, κατά τη δεύτερη φάση υλοποίησης, η επιχείρηση αντιμετωπίζει μια περίπτωση χρήσης που καταδεικνύει την αξία των μεγάλων δεδομένων. Αυτό συνίσταται στην ανάπτυξη μιας αλυσίδας επεξεργασίας δεδομένων για προϋπάρχοντα δεδομένα και στη συνέχεια στην ανάπτυξη αυτής της απόδειξης της ιδέας σε ένα πλαίσιο παραγωγής. Οι συνήθεις περιπτώσεις χρήσης σε αυτό το στάδιο περιλαμβάνουν τον εντοπισμό απάτης, την ανάλυση αρχείων καταγραφής για βελτιωμένη κατανόηση των μοτίβων χρήσης, την πρόβλεψη της παρέκκλισης ή, πιο κοντά στην εμπειρία του χρήστη, την εισαγωγή συστημάτων συστάσεων. Οι βιβλιοθήκες ανάλυσης δεδομένων, όπως το MLib for Spark, έχουν μεγάλες λίστες εγγενών (και βελτιστοποιημένων) αλγορίθμων για την αντιμετώπιση αυτών των τύπων προβλημάτων. Ο στόχος εδώ είναι να καταδειχθεί η προστιθέμενη αξία και ο οικονομικός αντίκτυπος της δημιουργίας μιας αρχιτεκτονικής μεγάλων δεδομένων (Corlosquet-Habart and Janssen, 2018).

Η τρίτη φάση είναι φυσικά η γενίκευση των περιπτώσεων χρήσης σε διαφορετικά επίπεδα της αλυσίδας αξίας της επιχείρησης. Οι ομάδες που είναι υπεύθυνες για τα μεγάλα δεδομένα θα έχουν μέχρι τώρα παραδείγματα πρώιμων επιτυχιών, για να πείσουν τους διαφορετικούς ενδιαφερόμενους στην επιχείρηση και το κόστος

ανάπτυξης μιας νέας περίπτωσης χρήσης θα μειωθεί, καθώς η υποδομή υπάρχει ήδη. Αυτό είναι όπου οι επιχειρηματικές εφαρμογές βλέπουν το φως της δημοσιότητας, κάθε υπηρεσία αξιοποιεί την τεχνολογία για να βελτιστοποιήσει την υπάρχουσα ανάλυση, να την επεκτείνει, να προτείνει νέα ανάλυση ή απλώς να αποκτήσει καλύτερη κατανόηση του τομέα της (Corlosquet-Habart and Janssen, 2018). Μια χρηματοοικονομική υπηρεσία θα επιδιώξει να βελτιώσει τη διαχείριση κινδύνων ή τον εντοπισμό απάτης, μια υπηρεσία υγείας θα ξεκινήσει στοχευμένα προγράμματα πρόληψης, θα στοχεύει στη μείωση της επανεισδοχής ή θα αναλύσει τις εσωτερικές διαδικασίες για τη βελτίωση του συντονισμού τους (Martinez, Viles and Olaizola, 2021).

Τέλος, η τελευταία φάση συνίσταται στην πραγματική ενσωμάτωση της ανάλυσης δεδομένων και των γνώσεών της στη συνολική στρατηγική της επιχείρησης. Η βελτίωση των επιχειρηματικών διαδικασιών ή/και των οικονομικών οφελών μετατρέπεται σε ανταγωνιστικά πλεονεκτήματα. Τα αποτελέσματα από την προγνωστική ανάλυση συμβάλλουν στη λήψη αποφάσεων. Σε αυτό το στάδιο, οι υπεύθυνοι λήψης αποφάσεων συμβουλεύονται κάποιον υπεύθυνο για τα δεδομένα (ο τίτλος θέσης Chief Data Officer αρχίζει να εμφανίζεται) και μια ειδική ομάδα δεδομένων διατηρεί την υποδομή και εργάζεται για την επίλυση νέων, ad-hoc προβλημάτων ειδικά για την επιχείρηση (Corlosquet-Habart and Janssen, 2018). Ο αναλυτής δεδομένων, ειδικός στα στατιστικά, βοηθά στη δημιουργία πινάκων εργαλείων που εμφανίζουν τα δεδομένα και στη βέλτιστη χρήση των αλυσίδων επεξεργασίας δεδομένων, ενώ ο επιστήμονας δεδομένων, με εξειδίκευση στα μαθηματικά, στα στατιστικά και στους υπολογιστές, παράγει νέες αλυσίδες επεξεργασίας δεδομένων και ξεκλειδώνει νέες ευκαιρίες, φροντίζοντας παράλληλα να

διατηρείται οπτικοποίηση της απόδοσης της εταιρείας σε πραγματικό χρόνο (Martinez, Viles and Olaizola, 2021).

2.2.2. Είναι απαραίτητη η μετάβαση προς μια αρχιτεκτονική μεγάλων δεδομένων;

Οι εταιρείες αναπόφευκτα εξετάζουν το ενδεχόμενο μετάβασης ή όχι προς μια αρχιτεκτονική μεγάλων δεδομένων. Χρειάζεται αντικατάσταση του υπάρχοντος συστήματος επιχειρηματικής ευφυΐας (BI); Για απλοποίηση, αυτός ο τύπος συστήματος αποτελείται από δύο κύρια μέρη (Balas et al., 2019):

- Τη διαδικασία ETL (εξαγωγή, μετατροπή και φόρτωση δεδομένων), η οποία συνίσταται στην εξαγωγή από τις πηγές επιχειρησιακών δεδομένων της εταιρείας όλων των (ετερογενών) δεδομένων που θα μπορούσαν να βοηθήσουν στην απάντηση στις ερωτήσεις των υπευθύνων λήψης αποφάσεων. Στη συνέχεια, τα δεδομένα υποβάλλονται σε επεξεργασία (καθαρίζονται, κανονικοποιούνται, συγκεντρώνονται κ.λπ.) και ενσωματώνονται έτσι ώστε να μπορούν να φορτωθούν στην αποθήκη δεδομένων ακολουθώντας προκαθορισμένα πρωτόκολλα.
- Την αποθήκη δεδομένων, που επιτρέπει την ενοποίηση όλων των δεδομένων μιας εταιρείας και ως εκ τούτου προσφέρει μια οριζόντια και ολοκληρωμένη επισκόπηση όλων των πτυχών της επιχειρηματικής δραστηριότητας της εταιρείας. Μπορεί να αποτελείται από πολλά υποσύνολα που ονομάζονται datamarts και το καθένα χαρακτηρίζει μια καθορισμένη επιχειρηματική διαδικασία. Αυτά τα δεδομένα είναι δομημένα με τη μορφή πολυδιάστατων λογικών σχημάτων που επιτρέπουν την προετοιμασία της πρόσβασης σε

προκαθορισμένους δείκτες, για την εκπλήρωση μιας απαίτησης αναφοράς για παράδειγμα, ενώ παράλληλα επιτρέπει την ανάλυσή τους σε διάφορες διαστάσεις (για παράδειγμα, αναλύοντας τον δείκτη «εσόδων» «ανά περιοχή» , «κατά περίοδο» ή «κατά κατάστημα»).

Αυτή η μοντελοποίηση μπορεί να χρησιμοποιηθεί για τη δημιουργία πολυδιάστατων κύβων (ή υπερκύβων) σε διακομιστές OLAP, επιτρέποντας σημαντική διαδραστικότητα κατά την αναζήτηση. Τα γραφικά εργαλεία BI για ανάλυση και αναφορά, όπως το Excel, το Table ή το Business Object, χρησιμοποιούνται συχνά για τη δημιουργία πινάκων εργαλείων και αναφορών σε συνεννόηση με την αποθήκη (Naeem et al., 2022).

Η άφιξη των μεγάλων δεδομένων συνοδεύτηκε από την εμφάνιση νέων αναλυτικών διεργασιών (ή φόρτου εργασίας) που θα δυσκολευόταν να ολοκληρώσει η κλασική τεχνολογία ETL ή αποθήκευσης (Sakr and Gaber, 2014):

- διερευνητική ανάλυση ακατέργαστων, μη μοντελοποιημένων και μη δομημένων δεδομένων·
- επεξεργασία σε πραγματικό χρόνο, σε αντίθεση με τις διαδικασίες ETL που εκτελούνται σε παρτίδες.
- ταχεία επεξεργασία κατά παρτίδες για μεγάλους όγκους δεδομένων.
- ευελιξία και ταχεία αρχειοθέτηση δεδομένων, με δυνατότητα ταχείας επανάληψης της επεξεργασίας που απαιτείται για την ενημέρωση των δεδομένων της αποθήκης·
- περίπλοκη ανάλυση, όπως η παράλληλη εφαρμογή πολλών εκατομμυρίων μοντέλων βαθμολόγησης σε εκατομμύρια τραπεζικούς λογαριασμούς για τον εντοπισμό απάτης, για παράδειγμα.

Τα καλά νέα είναι ότι είναι δυνατό να ενωθούν οι δύο κόσμοι και να χρησιμοποιηθεί το Hadoop ως μια αποτελεσματική και επεκτάσιμη λύση ETL για δεδομένα που απαιτούν συγκεκριμένους φόρτους εργασίας. Μόλις εξαχθούν και φορτωθούν τα δεδομένα στο Hadoop, μπορούν να υποβληθούν σε σύνθετους μετασχηματισμούς σε παρτίδες προγραμματίζοντας εργασίες MapReduce ή Spark ή χρησιμοποιώντας γλώσσες υψηλού επιπέδου όπως HiveQL ή Pig. Είναι δυνατή η ανάλυση της σύνταξης μη δομημένων ή ημιδομημένων δεδομένων και η διεξαγωγή υπολογισμών, ενώσεων και συναθροίσεων προκειμένου να ενσωματωθούν δεδομένα από διαφορετικές πηγές ή να δομηθούν έτσι ώστε να μπορούν να εισαχθούν σε αποθήκες δεδομένων μετά κλασικές επιχειρηματικές ροές εργασίας (Mehmood and Anees, 2022).

Το Hadoop μπορεί επίσης να χρησιμοποιηθεί για τη δημιουργία μιας ευέλικτης και επεκτάσιμης αποθήκης δεδομένων και για τη διασύνδεσή της με κλασικά εργαλεία BI, για παράδειγμα για αναφορές. Ωστόσο, η πλειονότητα των εκδοτών λύσεων αποθήκευσης δεδομένων, όπως η Oracle ή η Teradata, προτιμούν να ενσωματώνουν το Hadoop μόνο σε επίπεδο ETL, κάτι που επιτρέπει στις λύσεις τους να επαυξηθούν αντί να αντικατασταθούν. Αντίθετα, οι υποστηρικτές των λύσεων ανοιχτού κώδικα υπερασπίζονται τη διαχείριση φόρτου εργασίας στην οποία το κατανεμημένο περιβάλλον Hadoop παίζει το ρόλο ενός κόμβου δεδομένων, μέσω του οποίου διέρχονται όλα τα δεδομένα στο οικοσύστημα της εταιρείας, πριν τροφοδοτηθούν σε πολλαπλές αναλυτικές πλατφόρμες (Raj and D'Souza, 2019).

Η ανάλυση όλων αυτών των προσεγγίσεων είναι περίπλοκη. Μερικοί συγγραφείς έχουν δημιουργήσει πλέγματα συγκρίνοντας τις απαιτήσεις διαφορετικών τεχνικών επιλογών, όπως οι ιδιότητες των αλγορίθμων ανάλυσης δεδομένων (Landset et al., 2015), καθώς και οι πιθανές επιπτώσεις τους, για παράδειγμα, όσον αφορά τις δεξιότητες και τους ανθρώπινους πόρους (Chalmers, Bothorel and Clemente, 2013).

2.3. Προκλήσεις και ευκαιρίες για τον κόσμο της ασφάλισης

Τα δεδομένα βρίσκονται στην καρδιά της ασφάλισης. Είναι η πρώτη ύλη για τη βαθμολόγηση μοντέλων, που επιτρέπει την τμηματοποίηση των κατόχων premium, για να τους γνωρίσουν καλύτερα και να τους προσφέρουν προϊόντα κατά παραγγελία, να εκτιμήσουν καλύτερα τον τρέχοντα και μελλοντικό κίνδυνο και να λάβουν αποφάσεις. Τα μεγάλα δεδομένα και η ψηφιακή μετάβαση αλλάζουν ριζικά τον ασφαλιστικό τομέα. Όπως για όλους τους οικονομικούς παράγοντες, οι ασφαλιστές θα αντιμετωπίσουν φυσικά αλλαγές οργάνωσης, κουλτούρας και ανταγωνισμού. Θα επεξηγήσουμε αυτή την εξέλιξη με δύο παραδείγματα στα οποία τα μεγάλα δεδομένα παίζουν κεντρικό ρόλο: το πρώτο απεικονίζει τον αντίκτυπο της ανάπτυξης της οικονομίας διαμοιρασμού και το δεύτερο τον αντίκτυπο της αλλαγής συμπεριφορών στην τμηματοποίηση (Corlosquet-Habart and Janssen, 2018).

Η ασφάλιση είναι ήδη μέρος της οικονομίας του διαμοιρασμού. Νέοι παράγοντες, όχι απαραίτητα από τον κόσμο της ασφάλισης, δημιουργούν κοινότητες ατόμων με συγκεκριμένες ασφαλιστικές ανάγκες προκειμένου να διαπραγματευτούν εξαιρετικά εξατομικευμένα συμβόλαια για αυτούς από ασφαλιστές και να μειώσουν το κόστος καθώς το κάνουν. Εάν οι πλατφόρμες της κοινότητας επιτρέπουν στα άτομα να διατυπώνουν τις ανάγκες τους, τα μεγάλα δεδομένα επιτρέπουν σε αυτούς τους νέους φορείς να είναι προορατικοί στην εύρεση μικρών ομάδων πελατών, των οποίων η απογοήτευση συσσωρεύεται στο διαδίκτυο (Billot, Bothorel and Lenca, 2018).

Πράγματι, το μόνο που απαιτείται είναι να αναλυθούν τα ερωτήματα των μηχανών αναζήτησης, τα ιστολόγια και τα κοινωνικά δίκτυα για να προσδιοριστούν συγκεκριμένες ασφαλιστικές ανάγκες. Αυτοί οι νέοι φορείς αλλάζουν έτσι τη σχέση μεταξύ του ασφαλισμένου και των ασφαλιστών τους, αλλά διευκολύνουν επίσης την καινοτομία, καθώς οι (πολύ) εξατομικευμένες λύσεις είναι είτε προσαρμογές

υφιστάμενων συμβολαίων είτε εντελώς νέες συμβάσεις. Αν και αυτός ο τύπος αγοράς εξακολουθεί να είναι οριακός, φαίνεται πιθανό μια τέτοια αγορά εξειδικευμένων θέσεων να μπορεί να αναπτυχθεί. Αυτό ισχύει ιδιαίτερα για τις συνεργατικές πρακτικές κοινής χρήσης αγαθών ή υπηρεσιών (διαμοιρασμός αυτοκινήτου, ενοικίαση οχημάτων/διαμερισμάτων μεταξύ ατόμων κ.λπ.) που συνεχίζουν να αναπτύσσονται. Αυτά αλλάζουν τον τρόπο αξιολόγησης των κινδύνων και πάλι συγκεκριμένες, ή ακόμα και κατά παραγγελία, εγγυήσεις πρέπει να προσφέρονται. Ουσιαστικά, αυτές οι πρακτικές αλλάζουν το παράδειγμα από «ένα αγαθό για έναν ιδιοκτήτη» σε «πλήθος χρηστών για ένα αγαθό». Αυτή η μετατόπιση από την ιδιοκτησία προς τη χρήση, επιφέρει νέους τύπους κινδύνων και αποτελεί πρόκληση για τους ασφαλιστές (Corlosquet-Habart and Janssen, 2018)

Τα μεγάλα δεδομένα παρέχουν επίσης εύκολη πρόσβαση σε ορισμένες από τις πληροφορίες που είναι απαραίτητες για την τιμολόγηση και σταδιακά θα μειώσουν τη χρήση κλασικών έντυπων ερωτηματολογίων. Ως εκ τούτου, επιτρέπει ταχύτερη λήψη αποφάσεων. Ακόμη καλύτερα, παρέχοντας πρόσβαση σε προηγουμένως απρόσιτες πληροφορίες, θα επιτρέψει τη μείωση της υφιστάμενης ασυμμετρίας πληροφοριών μεταξύ του ασφαλισμένου, ο οποίος γνωρίζει σχεδόν όλες τις πληροφορίες που τον αφορούν, και του ασφαλιστή που έχει μόνο μερικές πληροφορίες. Ως εκ τούτου, τα μεγάλα δεδομένα επιτρέπουν μεγαλύτερη γνώση των ασφαλισμένων και των κινδύνων που συνδέονται με αυτούς, ακριβέστερη αξιολόγηση της συμπεριφοράς και ως εκ τούτου βελτιστοποιημένη επιλογή του ποιος θα ασφαλιστεί και δικαιότερες τιμές ασφαλίστρων. Όσοι είναι ασφαλισμένοι μπορούν, ιδιαίτερα αν είναι προς το συμφέρον τους, να δώσουν πρόσβαση σε πολύ ιδιωτικά δεδομένα σχετικά με τον τρόπο ζωής τους. Η αποδοχή μιας τέτοιας προσέγγισης, για τους καταναλωτές και τις ρυθμιστικές αρχές, είναι προφανώς κρίσιμη (Belhadi, Abdellah and Nezai, 2023).

Το σύνθημα «πλήρωσε όσο ζεις, οδηγείς κ.λπ.» είναι ήδη εδώ, ειδικά στην ασφάλιση αυτοκινήτου. Για παράδειγμα, η συνδεδεμένη οδήγηση επιτρέπει την ακριβή ανάλυση του στυλ οδήγησης (ταχύτητα, επιτάχυνση, φρενάρισμα, στρίψιμο κ.λπ.), ανάλογα με τον δρόμο και τις καιρικές συνθήκες. Αυτή η τάση αναπτύσσεται και στην ασφάλιση υγείας με συνδεδεμένα αντικείμενα, επιτρέποντας τη μέτρηση της φυσικής κατάστασης (καρδιακός ρυθμός, ύπνος κ.λπ.) και της δραστηριότητας (αριθμός βημάτων που έγιναν, συμμετοχή σε αθλήματα κ.λπ.) του ασφαλισμένου. Η ποιότητα του καθημερινού τους περιβάλλοντος μπορεί να αξιολογηθεί χρησιμοποιώντας εξωτερικά και ανοιχτά δεδομένα. Ωστόσο, η «υπερ-εξατομικευμένη» τιμολόγηση premium θα μπορούσε να αμφισβητήσει το τρέχον μοντέλο τμηματοποίησης και αμοιβαίας απόδοσης του κινδύνου, την υποκείμενη αρχή του τρόπου καθορισμού των τιμών και το ερώτημα πώς θα δομηθούν τα χαρτοφυλάκια κινδύνου. Η εισβολή των ασφαλιστών στην καρδιά της ιδιωτικής ζωής των ατόμων, θέτει προφανώς το πρόβλημα της προστασίας των δεδομένων. Υπάρχουν επίσης ερωτήματα σχετικά με το πώς θα αναπτυχθούν νέες πρακτικές και πώς θα μπορούσαν να επηρεάσουν την κοινωνία (Barry and Charpentier, 2020).

Μέσα από αυτά τα δύο παραδείγματα, δείξαμε μερικές από τις ευκαιρίες που προσφέρουν τα μεγάλα δεδομένα (νέες αγορές, καινοτομία και μείωση της ασυμμετρίας πληροφοριών). Η βελτίωση της αποτελεσματικότητας των διαφημιστικών εκστρατειών, η στόχευση και η μείωση της απάτης, αποτελούν περαιτέρω παραδείγματα. Εμφανίζονται νέες προκλήσεις (η είσοδος διαμεσολαβητών, τα θεμελιώδη στοιχεία της υπό αμφισβήτηση ασφάλισης, ασφάλεια δεδομένων, αναλογιστικές προκλήσεις), ενώ εγείρονται και ζητήματα ηθικής, ασφάλειας και νομικής φύσεως. Οι ρυθμιστικές αρχές ενδέχεται να περιορίσουν τη χρήση προσωπικών δεδομένων ή δεδομένων που οδηγούν σε τμηματοποίηση, που θεωρείται

ότι εισάγει διακρίσεις. Θα μπορούσαν να αναπτυχθούν αγορές για δόλια προφίλ και οι κάτοχοι premium ειδοποιήσεων θα δημιουργήσουν διαφορετικά προφίλ για ιδιωτική και δημόσια χρήση, οδηγώντας έτσι σε αμφισβήτηση το όφελος από τη μείωση της ασυμμετρίας πληροφοριών. Τέλος, εάν τα μεγάλα δεδομένα αντιπροσωπεύουν μια κερδοφόρα επένδυση, κινδυνεύει να αποσταθεροποιήσει ολόκληρη την ασφαλιστική αγορά. Από τη μία πλευρά, οι εταιρείες που δεν διαθέτουν τα μέσα πρόσβασης σε μεγάλα δεδομένα και τις απαραίτητες τεχνολογίες και δεξιότητες εργατικού δυναμικού θα δουν την ανταγωνιστικότητά τους να ξετυλίγεται. Ως εκ τούτου, κινδυνεύουν να εξαφανιστούν ή να εξαγοραστούν (Cassel and Bindman, 2019).

Από την άλλη πλευρά, οι ενδιάμεσες πλατφόρμες, ιδίως η GAFA (Google, Apple, Facebook, Amazon), που ελέγχουν ολόκληρη την αλυσίδα αξίας δεδομένων (συλλογή, τεχνολογία αποθήκευσης και υπολογισμούς, σχετική τεχνογνωσία), θα μπορούσαν να επιδιώξουν να λάβουν σημαντικό ποσοστό κερδών, ή θα μπορούσαν ακόμη και να μπουν στον πειρασμό να γίνουν οι ίδιοι ασφαλιστές. Η εξαγορά αποδυναμωμένων εταιρειών θα μπορούσε έτσι να τους επιτρέψει να εισέλθουν στην ασφαλιστική αγορά. Μια νέα μορφή ασυμμετρίας, ελέγχου των δεδομένων, είναι πιθανώς ήδη σε ισχύ (Charpentier, 2020).

Κεφάλαιο 3ο : Από τις συμβατικές μεθόδους ανάλυσης δεδομένων έως την ανάλυση μεγάλων δεδομένων

3.1. Από την ανάλυση δεδομένων στην εξόρυξή τους: εξερεύνηση και πρόβλεψη

Η ανάλυση δεδομένων εδώ, σημαίνει κυρίως περιγραφικές και διερευνητικές μεθόδους γνωστές και ως μη εποπτευόμενες. Στόχος είναι να περιγραφεί και να δομηθεί ένα σύνολο δεδομένων που μπορούν να αναπαρασταθούν με τη μορφή ενός ορθογώνιου πίνακα που διασταυρώνει n στατιστικές μονάδες και p μεταβλητές. Γενικά θεωρούμε n παρατηρήσεις ως σημεία στον διανυσματικό χώρο p διαστάσεων, ο οποίος εάν παρέχεται με απόσταση, είναι ένας Ευκλείδειος χώρος. Οι αριθμητικές μεταβλητές είναι διανύσματα ενός n διαστάσεων χώρου (Larose and Larose, 2014).

Οι μέθοδοι ανάλυσης δεδομένων είναι ουσιαστικά μέθοδοι μείωσης της διάστασης που χωρίζονται σε δύο κατηγορίες: – αφενός μέθοδοι παραγόντων (ανάλυση κύριας συνιστώσας για αριθμητικές μεταβλητές, αναλύσεις αντιστοιχίας για μεταβλητές κατηγορίας) που οδηγούν σε νέες αριθμητικές μεταβλητές, συνδυασμούς των αρχικών μεταβλητών, επιτρέποντας αναπαραστάσεις σε χώρους χαμηλών διαστάσεων. Μαθηματικά, αυτές είναι παραλλαγές της αποσύνθεσης μοναδικών τιμών του πίνακα δεδομένων – αφετέρου, οι μη εποπτευόμενες μέθοδοι ταξινόμησης ή ομαδοποίησης που χωρίζουν τις παρατηρήσεις ή τις μεταβλητές σε ομοιογενείς ομάδες (Saporta, 2008). Οι κύριοι αλγόριθμοι είναι είτε ιεραρχικοί (βήμα προς βήμα κατασκευή των κλάσεων με διαδοχική ομαδοποίηση μονάδων), είτε άμεσες αναζητήσεις διαμερισμάτων με k-means (Zhang, Zhang and Yang, 2003).

Ωστόσο, η ανάλυση δεδομένων είναι επίσης μια στάση που συνίσταται στο «να αφήνουμε τα δεδομένα να μιλήσουν» βάζοντας τίποτα, ή τουλάχιστον πολύ λίγο a

priori, στον μηχανισμό παραγωγής. Ας θυμηθούμε εδώ την αρχή που δηλώνει ο (Smyth, 2000): «Το μοντέλο πρέπει να ακολουθεί τα δεδομένα και όχι το αντίθετο». Η ανάλυση δεδομένων αναπτύχθηκε στις δεκαετίες του 1960 και του 1970 ως αντίδραση στις καταχρήσεις της επισημοποίησης, (βλέπε Anscombe, 1967, σχετικά με τον John Tukey): «Αυτός (Tukey) φαίνεται να ταυτίζει τις στατιστικές με το γκροτέσκο φαινόμενο που είναι γενικά γνωστό ως μαθηματική στατιστική και θεωρεί απαραίτητο να αντικατασταθούν οι στατιστικές με την ανάλυση δεδομένων».

Η εξόρυξη δεδομένων, ένα κίνημα που ξεκίνησε τη δεκαετία του 1990 στη διασταύρωση στατιστικών και τεχνολογιών πληροφοριών (βάσεις δεδομένων, τεχνητή νοημοσύνη, μηχανική μάθηση κ.λπ.), στοχεύει επίσης στην ανακάλυψη δομών σε μεγάλα σύνολα δεδομένων και προωθεί νέα εργαλεία, όπως κανόνες συσχέτισης. Η μεταφορά της εξόρυξης δεδομένων σημαίνει ότι υπάρχουν θησαυροί ή ψήγματα κρυμμένοι κάτω από βουνά δεδομένων που μπορούν να ανακαλυφθούν με εξειδικευμένα εργαλεία. Η εξόρυξη δεδομένων είναι ένα βήμα στη διαδικασία ανακάλυψης γνώσης, η οποία περιλαμβάνει την εφαρμογή αλγορίθμων ανάλυσης δεδομένων (Friedman, 1998). Ο Hand (1999) το όρισε ως εξής: «Θα ορίσω την εξόρυξη δεδομένων ως την ανακάλυψη ενδιαφέρουσων, απροσδόκητων ή πολύτιμων δομών σε μεγάλα σύνολα δεδομένων».

Η εξόρυξη δεδομένων αναλύει δεδομένα που συλλέγονται για άλλους σκοπούς: είναι συχνά μια δευτερεύουσα ανάλυση βάσεων δεδομένων, σχεδιασμένη για τη διαχείριση μεμονωμένων δεδομένων και όπου δεν υπάρχει ανησυχία για την αποτελεσματική συλλογή αυτών (έρευνες, πειραματικά σχέδια) (Azzalini and Scarpa, 2012).

Η εξόρυξη δεδομένων επιδιώκει επίσης να βρει μοντέλα πρόβλεψης μιας απόκρισης που δηλώνεται με Y, αλλά από μια πολύ διαφορετική προοπτική από αυτή της

συμβατικής μοντελοποίησης. Ένα μοντέλο δεν είναι τίποτα άλλο παρά ένας αλγόριθμος και όχι μια αναπαράσταση του μηχανισμού που παρήγαγε τα δεδομένα. Στη συνέχεια, προχωράμε διερευνώντας ένα σύνολο γραμμικών ή μη γραμμικών αλγορίθμων, σαφών ή μη, προκειμένου να επιλέξουμε τον καλύτερο, που είναι αυτός που παρέχει τις πιο ακριβείς προβλέψεις, χωρίς να πέφτει στην παγίδα της υπερπροσαρμογής. Διακρίνουμε μεθόδους παλινδρόμησης, όπου το Y είναι ποσοτικές, εποπτευόμενες μέθοδοι ταξινόμησης (ονομάζονται επίσης μέθοδοι διάκρισης) όπου το Y είναι κατηγορηματικό, πιο συχνά με δύο τρόπους. Η μαζική επεξεργασία δεδομένων έχει απλώς ενισχύσει τις τάσεις που υπάρχουν ήδη στην εξόρυξη δεδομένων (Haoxiang and Smys, 2021).

3.2. Απαρχαιωμένες προσεγγίσεις

Τα συμπεράσματα στατιστικά αναπτύχθηκαν σε ένα πλαίσιο σπάνιων δεδομένων, τόσο που ένα δείγμα άνω των 30 μονάδων θεωρήθηκε μεγάλο! Ο όγκος των δεδομένων αλλάζει ριζικά την πρακτική της στατιστικής. Να μερικά παραδείγματα (Wolf et al., 2013):

- οποιαδήποτε απόκλιση από μια θεωρητική τιμή γίνεται «σημαντική». Έτσι, ένας συντελεστής συσχέτισης 0,01 που υπολογίζεται μεταξύ δύο μεταβλητών σε ένα εκατομμύριο παρατηρήσεις (και ακόμη λιγότερο, όπως εύκολα θα επαληθεύσει ο αναγνώστης) θα δηλωθεί σημαντικά διαφορετικός από το μηδέν. Είναι χρήσιμο αποτέλεσμα;
- τα διαστήματα εμπιστοσύνης των παραμέτρων ενός μοντέλου γίνονται μηδενικά στο πλάτος, αφού το τελευταίο είναι γενικά σε 1/n. Αυτό σημαίνει ότι το μοντέλο θα είναι γνωστό με βεβαιότητα;

Γενικά, δεν υπάρχει πλέον ένα παραγωγικό μοντέλο που να ισχύει για μεγάλο όγκο δεδομένων, όχι περισσότερο από τους κανόνες επιλογής του μοντέλου με κυρώσεις πιθανοτήτων που αποτελούν αντικείμενο τόσων πολλών δημοσιεύσεων (Saporta, 2008).

Πρέπει να σημειωθεί ότι τα κριτήρια του τύπου:

$$AIC = -2\ln(L) + 2k$$

$$BIC = -2\ln(L) + \ln(n)\,k$$

για να γίνει επιλογή μεταξύ απλών μοντέλων όπου k είναι ο αριθμός των παραμέτρων και L η πιθανότητα, γίνεται αναποτελεσματικό όταν γίνεται σύγκριση αλγόριθμων πρόβλεψης, όπου ούτε η πιθανότητα, ούτε ο αριθμός των παραμέτρων είναι γνωστά, όπως στα δέντρα αποφάσεων και σε πιο σύνθετες μεθόδους που συζητούνται στο επόμενο κεφάλαιο. Θα πρέπει να σημειωθεί ότι είναι παράλογο, όπως φαίνεται συχνά, η ταυτόχρονη χρήση AIC και BIC, καθώς προέρχονται από δύο ασύμβατες θεωρίες: πληροφορίες Kullback–Leibler για την πρώτη και Bayesian επιλογή μοντέλων a priori ισοπιθανές για τη δεύτερη (Yang, 2005).

Ο μεγάλος όγκος δεδομένων θα μπορούσε να είναι ένα επιχείρημα υπέρ των ασυμπτωτικών ιδιοτήτων του BIC, αν ήταν υπολογίσιμο, αφού έχει αποδειχθεί ότι η πιθανότητα επιλογής του αληθινού μοντέλου τείνει στο 1 όταν ο αριθμός των παρατηρήσεων τείνει στο άπειρο. Το αληθινό μοντέλο, ωστόσο, πρέπει να είναι μέρος της οικογένειας που μελετάται και είναι ιδιαίτερα απαραίτητο να υπάρχει αυτό το «αληθινό» μοντέλο, που είναι μυθοπλασία: ένα μοντέλο (με τη γενετική έννοια) είναι απλώς μια απλοποιημένη αναπαράσταση της πραγματικότητας. Πριν από τριάντα χρόνια, πολύ πριν μιλήσουμε για μεγάλα δεδομένα, ο Box (1979) δήλωσε «Όλα τα μοντέλα είναι λάθος, μερικά είναι χρήσιμα».

Οι καταχρήσεις των λεγόμενων συμβατικών στατιστικών είχαν καταγγελθεί σθεναρά από τον John Nelder, τον συν-εφευρέτη των γενικευμένων γραμμικών μοντέλων, σε αυτό το κείμενο του 1985 που συζητούσε το άρθρο του Chatfield (1985): «Η στατιστική συνδέεται στενά με την επιστήμη και την τεχνολογία και λίγοι μαθηματικοί έχουν εμπειρία ή κατανοούν τις μεθόδους των δύο. Αυτό πιστεύω ότι είναι που κρύβεται πίσω από την τρομακτική έμφαση στα τεστ σημασίας στα στατιστικά μαθήματα όλων των ειδών. Έχει δημιουργηθεί μια μαθηματική συσκευή με τις έννοιες της δύναμης, ομοιόμορφα πιο ισχυρά τεστ, ομοιόμορφα πιο ισχυρά αμερόληπτα τεστ, κλπ. κ.λπ. και αυτό διδάσκεται στους ανθρώπους, οι οποίοι, αν δεν έχουν άλλη ιδέα, θα θυμούνται ότι η στατιστική είναι για σημαντικές διαφορές […]. Ο μηχανισμός πάνω στον οποίο έχει κατασκευαστεί το μάθημα της στατιστικής τους είναι συχνά χειρότερο από άσχετο, είναι παραπλανητικό σχετικά με το τι είναι σημαντικό για την εξέταση δεδομένων και την εξαγωγή συμπερασμάτων».

3.3. Κατανόηση ή πρόβλεψη;

Η χρήση αλγορίθμων εκμάθησης οδηγεί σε μεθόδους γνωστές ως «μαύρα κουτιά» που εμπειρικά δείχνουν ότι δεν είναι απαραίτητο να κατανοήσουμε για να προβλέψουμε. Αυτό το γεγονός, το οποίο είναι ανησυχητικό για τους επιστήμονες, υποστηρίζεται ρητά από θεωρητικούς της μάθησης, όπως ο Vapnik (2006) που γράφει «Καλύτερα μοντέλα αποκτώνται μερικές φορές αποφεύγοντας σκόπιμα την αναπαραγωγή των αληθινών μηχανισμών».

Ο Breiman (2001) το επιβεβαίωσε στο διάσημο άρθρο του με τίτλο «Statistical Modeling: The Two Cultures»: «Η σύγχρονη στατιστική σκέψη κάνει μια σαφή διάκριση μεταξύ του στατιστικού μοντέλου και του κόσμου. Οι πραγματικοί

μηχανισμοί που διέπουν τα δεδομένα θεωρούνται άγνωστοι. Τα στατιστικά μοντέλα δεν χρειάζεται να αναπαράγουν αυτούς τους μηχανισμούς για να μιμηθούν τα παρατηρήσιμα δεδομένα». Έτσι ο Breiman αντιπαραβάλλει δύο πολιτισμούς μοντελοποίησης προκειμένου να εξαχθούν συμπεράσματα από δεδομένα: η μία υποθέτει ότι τα δεδομένα παράγονται από ένα δεδομένο στοχαστικό μοντέλο και η άλλη θεωρεί τον μηχανισμό δημιουργίας ως άγνωστο και χρησιμοποιεί αλγόριθμους.

Στην πρώτη περίπτωση, δίνεται προσοχή στην προσαρμογή του μοντέλου στα δεδομένα (goodness of fit) και στη δεύτερη, εστίαση στην ακρίβεια της πρόβλεψης. Ο Donoho (2015) πιο πρόσφατα ανέλαβε αυτή τη συζήτηση μιλώντας για κουλτούρα γενετικής μοντελοποίησης και κουλτούρα προγνωστικής μοντελοποίησης. Η διάκριση μεταξύ μοντέλων κατανόησης και μοντέλων πρόβλεψης ήταν επίσης σαφής στα άρθρα των Saporta (2008) και Shmueli (2010).

Κεφάλαιο 4ο : Χρήση μεγάλων δεδομένων στην ασφάλιση

4.1. Ασφάλειες, ένας κλάδος ιδιαίτερα κατάλληλος για την ανάπτυξη μεγάλων δεδομένων

4.1.1. Ένας κλάδος που αναπτύχθηκε μέσω της χρήσης δεδομένων

4.1.1.1. Μακροχρόνια βιομηχανία ηλεκτρονικών υπολογιστών

Τα τελευταία 60 χρόνια αποτέλεσαν την αφορμή για μια ραγδαία και ολοκληρωμένη ανάπτυξη της πληροφορικής και των χρήσεών της σε όλους τους τομείς. Η περίοδος 1960-1970 χαρακτηρίστηκε από την υιοθέτηση της τεχνολογίας πληροφοριών για σκοπούς διαχείρισης back-office, προκειμένου να καταγραφούν, να επεξεργάζονται και να αποκαθίστανται μεγάλοι όγκοι πληροφοριών που είχαν προηγουμένως επεξεργαστεί σε έντυπη μορφή. Η τεχνολογία της πληροφορίας εφαρμόστηκε από πολύ περιορισμένο αριθμό χρηστών, αλλά επέτρεψε ήδη σημαντική μείωση του χρόνου επεξεργασίας και σημαντική εξοικονόμηση πόρων (Campbell-Kelly et al., 2023).

Η ανάπτυξη μηχανών, γλωσσών προγραμματισμού και προγραμμάτων κατά την περίοδο 1970-1980 επέτρεψε την ανάπτυξη προσαρμοσμένων εφαρμογών διαχείρισης και την υιοθέτηση από μεγαλύτερο αριθμό χρηστών. Ως εκ τούτου, η χρήση της πληροφορικής επεκτάθηκε σταδιακά σε όλους τους τομείς διαχείρισης: λογιστικός και διαχειριστικός έλεγχος, εκτέλεση πληρωμών και διαχείριση μετρητών, επιχειρησιακός και οικονομικός σχεδιασμός, διαχείριση μισθοδοσίας και γενικότερα θέματα που σχετίζονται με το ανθρώπινο δυναμικό, παρακολούθηση και μέτρηση της επιχειρηματικής δραστηριότητας. Αυτός ο εκδημοκρατισμός της χρήσης της τεχνολογίας των πληροφοριών εντός των εταιρειών οδήγησε σταδιακά στην ανάπτυξη εσωτερικών υπηρεσιών πληροφορικής και στη διαφοροποίηση των προφίλ σε αυτές τις υπηρεσίες (Campbell-Kelly et al., 2023).

Η περίοδος 1980–1990 χαρακτηρίστηκε από τη διαφοροποίηση των γλωσσών προγραμματισμού, τη σταδιακή σμίκρυνση και τυποποίηση του εξοπλισμού, την ανάπτυξη της διαλειτουργικότητας μεταξύ των συστημάτων και τη σταδιακή εφαρμογή των master plans εντός των εταιρειών με σκοπό τον εξορθολογισμό των επενδύσεων και την υιοθέτηση ενός στρατηγικού οράματος για την ανάπτυξη των πόρων πληροφορικής. Ταυτόχρονα, ο αριθμός των χρηστών εντός των εταιρειών συνέχισε να αυξάνεται και η τεχνολογία πληροφορικής χρησιμοποιήθηκε για όλες σχεδόν τις τρέχουσες λειτουργίες, ανεξάρτητα από το μέγεθος της συγκεκριμένης εταιρείας, αν και σε αυτό το στάδιο δεν ήταν επαρκώς ανεπτυγμένο όσον αφορά τα άτομα, ώστε να αποτελέσει ένα κανάλι διανομής (Yin, Stecke and Li, 2018).

Κατά την περίοδο 1990–2000, η περαιτέρω ανάπτυξη λύσεων αφιερωμένων στις διάφορες χρήσεις της τεχνολογίας της πληροφορίας και η βελτίωση της εργονομίας συνοδεύτηκαν από πτώση του κόστους του εξοπλισμού. Αυτές οι εξελίξεις ενθάρρυναν σταδιακά τα άτομα να εξοπλιστούν. Για τις εταιρείες, η ανάπτυξη λύσεων αφιερωμένων στην αποθήκευση και επεξεργασία δεδομένων συνέβαλε στην υλοποίηση αποθηκών δεδομένων και στην ανάπτυξη μεθόδων εξόρυξης δεδομένων (Cardona, Kretschmer and Strobel, 2013).

Η ταχεία ανάπτυξη του Διαδικτύου, παρά την έκρηξη της φούσκας του το 2000, από το δεύτερο μισό της δεκαετίας συνέβαλε σημαντικά στις προσπάθειες επέκτασης της τεχνολογίας της πληροφορίας στις εμπορικές σχέσεις μεταξύ εταιρειών αλλά και μεταξύ εταιρειών και ιδιωτών. Επιπλέον, η εμφάνιση διαδικτυακών εφαρμογών την περίοδο 2000–2010 και η εξάπλωση των δεδομένων, οδήγησαν σε πιο κατακερματισμένους τρόπους αστικοποίησης των πληροφοριακών συστημάτων. Ως απάντηση, η ανάπτυξη μεθοδολογιών και ικανοτήτων σχετικά με τους κατανεμημένους

υπολογιστές οδήγησε σε αυξανόμενες απαιτήσεις ροής και στην ανάπτυξη ενός συστήματος προσανατολισμένου στη λύση (Cardona, Kretschmer and Strobel, 2013).

Από το 2010, η ανάπτυξη της χρήσης του Διαδικτύου μέσω smartphone και η ανάπτυξη κοινωνικών δικτύων συνέβαλαν επίσης στην αύξηση της ανάγκης για διαθέσιμους πόρους και στην ποικιλομορφία των χρήσεων της τεχνολογίας πληροφοριών και των συναφών υπηρεσιών. Εκτός από τα δεδομένα που παράγονται άμεσα από τους χρήστες, τα κυβερνητικά κίνητρα για ανοιχτά δεδομένα συνέβαλαν στον πολλαπλασιασμό και τη διαφοροποίηση των διαθέσιμων δεδομένων, καθώς και στην ανάπτυξη των ενδιαφερομένων μέσω της αποτίμησής τους (Campbell-Kelly et al., 2023).

Ο ασφαλιστικός κλάδος ήταν από τους πρώτους που έκανε εκτεταμένη χρήση της τεχνολογίας της πληροφορίας, ένα ιδανικό εργαλείο για την εκτέλεση πολυάριθμων, διαδοχικών και τμηματικών εργασιών. Αυτή η υιοθέτηση έχει αλλάξει βαθιά τη σχέση μεταξύ των ασφαλιστικών εταιρειών και των πελατών τους και η βελτιστοποίηση της διαχείρισης έδωσε σταδιακά τη θέση της σε έναν πραγματικό εμπορικό αγώνα, καθιστώντας σταδιακά την ασφάλιση καταναλωτικό προϊόν, στο οποίο η θέση των ενδιαφερομένων έχει γίνει πρωταρχικής σημασίας, λόγω ενός ολοένα και πιο ανταγωνιστικού περιβάλλοντος (Nicholson, 2019).

Τα επόμενα χρόνια κινούνται προς την έκρηξη του Διαδικτύου των Πραγμάτων (IoT) και της τεχνητής νοημοσύνης. Οι πιο προηγμένες ασφαλιστικές εταιρείες επενδύουν ήδη σε μεγάλο βαθμό σε αυτές τις τεχνολογίες. Ως εκ τούτου, δεν υπάρχει αμφιβολία ότι ορισμένοι από αυτούς θα αποκτήσουν ανταγωνιστικά πλεονεκτήματα βελτιώνοντας την ποιότητα των σχέσεων με τους πελάτες τους, αυξάνοντας τα περιθώρια κέρδους τους και προσφέροντας στους υποψήφιους πελάτες τους και στους ήδη υπάρχοντες

πελάτες τους εγγυήσεις και υπηρεσίες που είναι πάντα πιο προσαρμοσμένες στις ανάγκες τους (Spender et al., 2019).

Η ανάπτυξη της προσβασιμότητας στο Διαδίκτυο σε ορισμένες αναπτυσσόμενες χώρες θα συμβάλει επίσης στην αύξηση του όγκου των δεδομένων που παράγονται ετησίως και μπορούν να εκμεταλλευτούν σε διάφορους τομείς. Καθώς τα σημερινά τους συστήματα πληροφορικής είναι ουσιαστικά ανύπαρκτα, θα μπορούν να υιοθετούν απευθείας τεχνολογίες αιχμής, ειδικά επειδή το κόστος που σχετίζεται με αυτές τις τεχνολογίες έχει μειωθεί σημαντικά και ενόψει του νομικού και δημογραφικού τους περιβάλλοντος, παρατηρείται η εμφάνιση ταχέως αναπτυσσόμενων εταιρειών (Nicholson, 2019).

4.1.1.2. Ένας κλάδος που τα δεδομένα του είναι πρώτη ύλη

Από την εμφάνιση της τεχνολογίας της πληροφορίας στον κλάδο, τα δεδομένα αντιπροσωπεύουν ένα θεμελιώδες πλεονέκτημα για τις ασφαλιστικές εταιρείες. Τα δεδομένα βρίσκονται στο επίκεντρο της σχέσης μεταξύ των αντισυμβαλλομένων και των ασφαλιστών και βοηθούν στη δημιουργία των συνθηκών για την αμοιβαία δέσμευσή τους. Οι ασφαλιστές ασκούν τις δραστηριότητές τους και διαχειρίζονται τους κινδύνους στους οποίους υπόκεινται προβλέποντας ιδίως την πιθανότητα εμφάνισης μελλοντικών απαιτήσεων (Kaswan et al., 2022).

Η γνώση του κινδύνου και της συμπεριφοράς των πελατών, είναι de facto βασικό συστατικό για τη διασφάλιση της βιωσιμότητας των ασφαλιστικών εργασιών, καθώς και για την ανταγωνιστική θέση των ασφαλιστών και, κατ' επέκταση, τη διατήρηση και ανάπτυξη των μεριδίων αγοράς τους. Όσον αφορά την πρόληψη και τη διαχείριση των ζημιών, η πρόσβαση σε αυξανόμενο όγκο πληροφοριών και η δυνατότητα

επεξεργασίας αυτών των δεδομένων σε πραγματικό χρόνο ή σχεδόν σε πραγματικό χρόνο, συμβάλλουν στην καλύτερη υποστήριξη του αντισυμβαλλομένου και στη μείωση του κόστους των αποζημιώσεων (Thouvenin et al., 2019).

Οι αλγόριθμοι μηχανικής μάθησης, λόγω της ικανότητάς τους να επεξεργάζονται μεγάλους όγκους δεδομένων και με αυτοματοποιημένο τρόπο μετά την εκβιομηχάνιση, αποτελούν λογικά ουσιαστικά εργαλεία, όσον αφορά την αποτίμηση δεδομένων. Παράλληλα, επιτρέπουν τη βέλτιστη χρήση των δεδομένων που διατίθενται μέσω ανοιχτών δεδομένων και πέρα από τα δεδομένα που είχαν συλλέξει προηγουμένως οι ασφαλιστές, τα δεδομένα που συγκεντρώθηκαν από τους τελευταίους μέσω της αυξανόμενης ψηφιοποίησης των διαδικασιών τους (Kaur, Sharma and Mittal, 2018).

4.1.1.3. Ένας κλάδος του οποίου το ρυθμιστικό πλαίσιο αποτελεί κίνητρο στην περιοχή

Το άρθρο 82 της Οδηγίας 2009/138/ΕΚ (Solvency II) εισάγει απαιτήσεις ποιότητας δεδομένων για τον υπολογισμό των τεχνικών αποθεματικών. Σύμφωνα με το άρθρο αυτό, οι ασφαλιστικές και αντασφαλιστικές εταιρείες υποχρεούνται να εφαρμόζουν εσωτερικές διαδικασίες και διαδικασίες για να διασφαλίζουν την καταλληλότητα, την πληρότητα και την ακρίβεια των δεδομένων που χρησιμοποιούνται σε αυτό το πλαίσιο. Τα άρθρα 19 έως 21 του κεφαλαίου ΙΙΙ, τμήμα 2 του κατ' εξουσιοδότηση κανονισμού προσδιορίζουν αυτές τις απαιτήσεις (Berthelé, 2018).

Οι απαιτήσεις ποιότητας δεδομένων σχετικά με το Solvency II αποτελούν δυνητικά ισχυρό περιορισμό στη χρήση μεγάλων δεδομένων για τον υπολογισμό των τεχνικών προβλέψεων, ιδίως ενόψει των αυστηρών απαιτήσεων, όσον αφορά τη διαδρομή ελέγχου και των συμπληρωματικών απαιτήσεων που αφορούν τη χρήση εξωτερικών δεδομένων επίσης που ορίζεται στο άρθρο 19 του κατ' εξουσιοδότηση κανονισμού που

αναφέρεται ανωτέρω. Αυτός ο περιορισμός είναι ένας από τους παράγοντες που προκαλούν τη χρήση μεγάλων δεδομένων σε αυτό το περιορισμένο στάδιο για αναλογιστικούς σκοπούς (Boobier, 2016).

Πίνακας Ι: Απαιτήσεις Solvency II σχετικά με την ποιότητα των δεδομένων που χρησιμοποιούνται για τον υπολογισμό των τεχνικών αποθεματικών

Κριτήρια	Αναμενόμενα όσον αφορά τα δεδομένα που χρησιμοποιούνται για τον υπολογισμό των τεχνικών αποθεματικών (TPs)
Πληρότητα	Τα δεδομένα περιλαμβάνουν επαρκείς ιστορικές πληροφορίες για την αξιολόγηση των χαρακτηριστικών των υποκείμενων κινδύνων και τον προσδιορισμό των τάσεων αυτών των κινδύνων. Είναι διαθέσιμα για καθεμία από τις σχετικές ομοιογενείς ομάδες κινδύνου.
Ακρίβεια	Τα δεδομένα είναι απαλλαγμένα από ουσιώδη σφάλματα. Είναι συνεπή με την πάροδο του χρόνου εάν χρησιμοποιούνται για την ίδια εκτίμηση. Καταγράφονται με κανονικό και συνεκτικό τρόπο.
Καταλληλότητα	Τα δεδομένα είναι συνεπή με τη χρήση τους. Ο όγκος και η φύση τους είναι τέτοια ώστε να διασφαλίζουν ότι οι εκτιμήσεις που γίνονται για τον υπολογισμό των τεχνικών αποθεματικών είναι απαλλαγμένες από οποιοδήποτε σημαντικό σφάλμα εκτίμησης (που είναι πιθανό να επηρεάσει τη λήψη αποφάσεων ή την κρίση των χρηστών του αποτελέσματος υπολογισμού, συμπεριλαμβανομένων των εποπτικών αρχών).

	Είναι συνεπείς με τις παραδοχές στις οποίες βασίζονται οι αναλογιστικές και στατιστικές τεχνικές που εφαρμόζονται σε αυτές για τον υπολογισμό των τεχνικών αποθεματικών. Αντικατοπτρίζουν επαρκώς τους κινδύνους στους οποίους εκτίθεται η ασφαλιστική ή αντασφαλιστική εταιρεία όσον αφορά τις ασφαλιστικές ή αντασφαλιστικές της δεσμεύσεις. Συλλέγονται, επεξεργάζονται και εφαρμόζονται με διαφανή και δομημένο τρόπο βάσει τεκμηριωμένης διαδικασίας. Οι ασφαλιστικές ή αντασφαλιστικές εταιρείες διασφαλίζουν ότι τα δεδομένα τους χρησιμοποιούνται διαχρονικά με συνέπεια για τον υπολογισμό των τεχνικών αποθεματικών.

Πηγή: Boobier (2016)

Η θέσπιση ενιαίων προτύπων σε ευρωπαϊκό επίπεδο (το Solvency II αποτελεί μόνο ένα μέρος αυτής της τάσης προς κανονιστική τυποποίηση, που αφορά ολόκληρο το χρηματοπιστωτικό σύστημα) και οι ανάγκες ασφαλιστικής κάλυψης ορισμένων τμημάτων του ευρωπαϊκού πληθυσμού θα πρέπει να τείνουν προς την τυποποίηση των προϊόντων και εμφάνιση πανευρωπαϊκών προϊόντων. Η συμπεριφορική ανάλυση που επιτρέπει η επιστήμη των δεδομένων, μας δίνει τη δυνατότητα σε αυτό το πλαίσιο να αντικειμενοποιήσουμε το σχηματισμό νέων ομοιογενών ομάδων κινδύνου μέσω μιας προσέγγισης που βασίζεται στα δεδομένα (Kemp, 2014).

4.1.2. Σχέση μεταξύ δεδομένων και ασφαλίσιμων περιουσιακών στοιχείων

4.1.2.1. Ασφαλίσιμο

Ο κίνδυνος, που είναι η πρώτη ύλη της ασφάλισης, είναι η αντιμετώπιση κινδύνων με ανθρώπινα υλικά και κατ' επέκταση οικονομικά διακυβεύματα.

Σύμφωνα με τον Berthelé (2018), για να είναι ασφαλίσιμος ένας κίνδυνος, πρέπει να είναι τυχαίος, μελλοντικός, νόμιμος, ανεξάρτητος από τη βούληση του αντισυμβαλλομένου και αρκετά κοινός, ώστε να υπόκειται σε υπολογισμό της πιθανότητας εμφάνισής του χωρίς να είναι σχεδόν βέβαιο.

Η ψηφιοποίηση της σχέσης μεταξύ ασφαλιστών και αντισυμβαλλομένων αυξάνει σημαντικά τη συχνότητα συλλογής δεδομένων. Οι δυνατότητες αντιστοίχισης των δεδομένων που συλλέγονται με εξωτερικά δεδομένα, αυξάνουν την ποικιλομορφία των διαθέσιμων πληροφοριών και οι μέθοδοι μηχανικής μάθησης καθιστούν δυνατό τον εντοπισμό των πιο μεροληπτικών μεταβλητών προκειμένου να προβλέψει μια συμπεριφορά ή την εμφάνιση ενός δεδομένου κινδύνου. Ουσιαστικά, το όραμα των κινδύνων τείνει να βελτιωθεί και ενδέχεται να προκύψουν διακυμάνσεις τιμών: διακυμάνσεις υπέρ του έτους που σχετίζονται με αλλαγές στις καταναλωτικές συνήθειες των αντισυμβαλλομένων κατά τη διάρκεια του έτους (για παράδειγμα, εποχικότητα χρήσης αυτοκινήτου ή κατάληψη δευτερεύουσας κατοικίας). Εξωτερικά, θα μπορούσε να εξεταστεί η τιμολόγηση και κατ' επέκταση, η ασφάλιση της πράξης, βάσει δεδομένων από συνδεδεμένα αντικείμενα - εκλεπτυσμένη κατάτμηση των εγγυήσεων στο πλαίσιο μιας σύμβασης και δυνατότητα του λήπτη της ασφάλισης, να αναλαμβάνει μόνο ένα μέρος των προσφερόμενων εγγυήσεων (για παράδειγμα, τα συμβόλαια ασφάλισης κατοικίας συχνά περιλαμβάνουν κατ' αποκοπή εγγυήσεις που μπορεί να είναι ακατάλληλες για νέους πελάτες) (Zheng and Guo, 2020).

Πέρα από τη διαφοροποίηση των υφιστάμενων προϊόντων και εγγυήσεων, θα μπορούσε να προβλεφθεί η ανάπτυξη εγγυήσεων για κινδύνους που μέχρι τώρα θεωρούνταν μη ασφαλίσιμοι λόγω της διαθεσιμότητας νέων πηγών δεδομένων. Τέλος, διευκολύνεται και η ασφάλιση αναδυόμενων κινδύνων (για παράδειγμα, κυβερνοκινδύνων) (Berthelé, 2018).

4.1.2.2. Έννοια της ομαδοποίησης

Η ομαδοποίηση είναι βασική ασφαλιστική αρχή. Περιλαμβάνει τον επιμερισμό του κόστους των απαιτήσεων που προκύπτουν από την εμφάνιση ενός κινδύνου μεταξύ των μελών μιας ομάδας που δυνητικά υπόκειται σε αυτόν και έχουν στην πραγματικότητα συνάψει ασφάλιση για την προστασία τους. Η ασφάλιση οργανώνει με δικαιοσύνη την οικονομική αλληλεγγύη μεταξύ ατόμων που αποτελούν μια ομοιογενή ομάδα κινδύνου, όσον αφορά τον εξεταζόμενο κίνδυνο: την αμοιβαιότητα. Οι τιμές προσαρμόζονται ανάλογα με την εξέλιξη του κινδύνου και η αμοιβαιότητα τιμωρείται σε περίπτωση απάτης από τα μέλη της. Η καλύτερη γνώση του κινδύνου που σχετίζεται με τις διάφορες ασφαλισμένες ομοιογενείς ομάδες κινδύνων, επιτρέπει τη βελτίωση της τιμής των προτεινόμενων εγγυήσεων, καθιστώντας την να αντιστοιχεί ακριβώς στον κίνδυνο στον οποίο υπόκειται ο ασφαλιστής λόγω των δεσμεύσεων που έχει αναλάβει (Borch, Sandmo and Aase, 2014).

4.1.2.3. Το όραμα των εξατομικευμένων τιμών

Η εμφάνιση των μεγάλων δεδομένων στην ασφάλιση δημιούργησε μια πίστη στην ικανότητα των ασφαλιστών να έχουν πολύ ακριβή γνώση του κινδύνου, έτσι ώστε η τιμή των εγγυήσεων να μπορεί να εξατομικεύεται. Αν και η ιδέα της εξατομίκευσης

των τιμών μπορεί να φαίνεται ελκυστική εκ πρώτης όψεως, αρκετά σημεία, εν τούτοις έρχονται σε αντίθεση με αυτήν (Picard, 2018).

Πρώτον, η εξατομίκευση συνεπάγεται πλήρη γνώση των παραγόντων κινδύνου που συνιστούν τον κίνδυνο. Ωστόσο, η απόκτηση ενός τέτοιου επιπέδου γνώσης είναι αδύνατη (επιρροή του περιβάλλοντος, αστάθεια στη συμπεριφορά των αντισυμβαλλομένων κ.λπ.) και μια συγκέντρωση, πιθανώς πιο λεπτή, αλλά ελάχιστη σε μια ομοιογενή ομάδα κινδύνου, θα είναι πάντα τεχνικά απαραίτητη για τη διασφάλιση της βιωσιμότητα των προϊόντων και των προτεινόμενων εγγυήσεων. Επιπλέον, δεδομένου ότι αυτή η εξατομίκευση εξαλείφει τη συγκέντρωση μεταξύ των αντισυμβαλλομένων που αποτελούν μια ομοιογενή ομάδα κινδύνου, θα παρέμενε μόνο η προσωρινή ομαδοποίηση των αντισυμβαλλόμενων ως προς τον αναλαμβανόμενο κίνδυνο και θα είχε δύο κύριες συνέπειες (McFall, 2019):

– ορισμένοι ασφαλισμένοι μπορεί να έχουν την τάση να οργανώνουν ιδιωτικά τη δική τους προσωρινή συγκέντρωση περιουσιακών στοιχείων, χωρίς να καταφεύγουν σε ασφάλειες, αλλά δίνοντας προτεραιότητα στα χρηματοοικονομικά προϊόντα·

– άλλοι, για τους οποίους η τιμή θα είχε γίνει απαγορευτική, θα έτειναν να μην αναλάβουν ασφάλιση, η οποία πέρα από τις δυνητικά δραματικές συνέπειες για αυτούς σε περίπτωση αξιώσεων, δημιουργεί πραγματικό κοινωνικό πρόβλημα, λόγω της προκύπτουσας αδυναμίας των ασφαλισμένων να πληρώσουν πιθανές ζημίες που προκλήθηκαν σε τρίτους.

Τέλος, η υπερβολική τμηματοποίηση (και κατ' επέκταση η εξατομίκευση τιμών), αν και οδηγεί σε αύξηση της κατά κεφαλήν κερδοφορίας των αντισυμβαλλομένων που διατηρούνται στο χαρτοφυλάκιο, έχει ως αποτέλεσμα μια ανταγωνιστική αγορά για τη δημιουργία εξειδικευμένης αγοράς και τη μείωση του ασφαλισμένου πληθυσμού σε

ένα δεδομένο ασφαλιστή (οι ασφαλισμένοι που τιμωρούνται περισσότερο από τις αλλαγές των τιμών, θα επιλέγουν έναν ασφαλιστή με λιγότερο ακριβή κατάτμηση) που οδηγεί σε μείωση του κύκλου εργασιών και της συνολικής κερδοφορίας σε ολόκληρο το ασφαλισμένο χαρτοφυλάκιο (Berthelé, 2018).

4.1.2.4. Προσαρμογή της ασφάλισης στις χρήσεις χάρη στα διαθέσιμα δεδομένα

Οι αλλαγές στον τρόπο ζωής και στην κατανάλωση, συμβάλλουν σημαντικά στην αύξηση του όγκου των διαθέσιμων δεδομένων. Η ψηφιοποίηση των δραστηριοτήτων, επιτρέπεται μόνο με την υιοθέτηση από τους ασφαλισμένους ενός τρόπου ζωής, όπου όλες οι υπηρεσίες πρέπει να είναι μόνιμα διαθέσιμες, ανεξαρτήτως τοποθεσίας. Αυτή η μετάβαση σε μια οικονομία της αμεσότητας συνοδεύεται από την ανάπτυξη της συνεργασίας και τη συνεχή αξιολόγηση των αγαθών που καταναλώνονται. Η ασφάλιση δεν αποτελεί εξαίρεση στην τάση και γίνεται όλο και περισσότερο καταναλωτικό αγαθό, για το οποίο οι αποτιμήσεις που παρέχονται από ασφαλισμένους πελάτες θα επιτρέψουν μεγαλύτερη προσαρμογή των εγγυήσεων και των προσφερόμενων υπηρεσιών (van Valkengoed and Steg, 2019).

Παρόμοια με την τάση που παρατηρείται σε άλλους τομείς, οι ασφαλισμένοι πελάτες τείνουν να αναζητούν όλο και περισσότερο εξατομικευμένες υπηρεσίες και εγγυήσεις που είναι απόλυτα προσαρμοσμένες στις ανάγκες τους. Στο πλαίσιο αυτό, φαίνονται ολοένα και περισσότερο διατεθειμένοι να διαθέσουν τα δεδομένα τους, εάν αυτό μπορεί να επιτρέψει μεγαλύτερη προσωποποίηση των προτάσεων των ασφαλιστικών εταιρειών (Porrini, 2017).

Η αύξηση του όγκου των δεδομένων προϊόντων είναι εκθετική, κάθε χρόνο ο όγκος των δεδομένων που παράγονται είναι υπερδιπλάσιος του όγκου που παρήχθη το

προηγούμενο έτος, με αποτέλεσμα το 90% των υπαρχόντων δεδομένων δημιουργήθηκαν τα τελευταία 2 χρόνια (Berthelé, 2018).

4.1.3. Πολλαπλασιασμός πηγών δεδομένων δυνητικού ενδιαφέροντος

4.1.3.1. Δεδομένα άμεσα διαθέσιμα στον ασφαλιστή

Οι τρέχουσες εξελίξεις, δεν θέτουν υπό αμφισβήτηση την απόκτηση ορισμένων δεδομένων που παραδοσιακά συλλέγονται από τους ασφαλιστές (Boobier, 2016):

- δεδομένα που σχετίζονται με τους αντισυμβαλλομένους, αν και ενδέχεται να υπάρχουν περιορισμοί τιμολόγησης σε αυτό τον τομέα (για παράδειγμα, μετά την οδηγία για το φύλο στην ασφάλιση αυτοκινήτου, το φύλο δεν μπορεί πλέον να εισάγει διακρίσεις όσον αφορά την τιμολόγηση, αν και μπορεί να συνεχίσει να ισχύει ως έχει σχετικά με τα αποθεματικά):
- δεδομένα που σχετίζονται με ασφαλισμένα πρόσωπα ή περιουσιακά στοιχεία·
- δεδομένα που σχετίζονται με συμβάσεις και εγγυήσεις που έχουν εγγραφεί·
- δεδομένα που σχετίζονται με τα καταβληθέντα ασφάλιστρα και τις προηγούμενες αποζημιώσεις του λήπτη της ασφάλισης και γενικότερα με τις διαφορετικές χρηματοοικονομικές ροές που υπάρχουν μεταξύ του ασφαλιστή και του ασφαλισμένου.

Ωστόσο, η ψηφιοποίηση των δραστηριοτήτων δημιουργεί δύο ανταγωνιστικές τάσεις (Boobier, 2016):

- η απλούστευση των διαδικασιών αναδοχής τείνει να περιορίσει τα δεδομένα που ζητούνται από τους αντισυμβαλλόμενους κατά την αναδοχή. Ωστόσο, η δυνατότητα αντιστοίχισης εξωτερικών δεδομένων θα αντισταθμίσει την απώλεια των πληροφοριών που δημιουργούνται.

- ο πολλαπλασιασμός των καναλιών επικοινωνίας (ιδιαίτερα, ιστοσελίδες και εφαρμογές για κινητά) και η προσαρμογή των ασφαλιστικών προσφορών στην οικονομία χρήσης πολλαπλασιάζουν τις ευκαιρίες για επαφές μεταξύ ασφαλισμένων και ασφαλιστών και συμβάλλουν στην αύξηση των εκμεταλλεύσιμων δεδομένων από τους ασφαλιστές, τόσο ως προς τη συχνότητα απόκτησης και ποικιλομορφίας δεδομένων (αρχεία καταγραφής σύνδεσης στον ιστότοπο του ασφαλιστή συγκεκριμένα ή τηλεφωνικά κέντρα που μπορούν να αποτιμηθούν μέσω της χρήσης ομιλίας σε κείμενο και σημασιολογικής ανάλυσης).

Αυτές οι δύο τάσεις δεν αντισταθμίζουν η μία την άλλη. Ο όγκος των δεδομένων που διαθέτουν οι αντισυμβαλλόμενοι και η έκθεσή τους σε ασφαλισμένους κινδύνους αυξάνεται σημαντικά, έτσι ώστε οι ασφαλιστές να χρησιμοποιούν τεχνολογικά και οικονομικά μέσα για να τα αξιοποιήσουν (Boobier, 2016).

4.1.3.2. Δεδομένα συνδεδεμένων αντικειμένων

Αν και τα δεδομένα που προέρχονται από συνδεδεμένα αντικείμενα θεωρούνται σημαντική πηγή πληροφοριών σχετικά με τη συμπεριφορά του ασφαλισμένου και επειδή επιτρέπουν μια πολύ λεπτομερή ανάλυση με την πάροδο του χρόνου του τρόπου χρήσης του συνδεδεμένου αντικειμένου, η χρήση τους για ασφαλιστικούς σκοπούς παραμένει οριακή. Δεδομένα που σχετίζονται με το ποσοτικοποιημένο άτομο, δεν μπορούν σε αυτό το στάδιο να θεωρηθούν ότι μπορούν να χρησιμοποιηθούν για ασφαλιστικούς σκοπούς, λόγω των υφιστάμενων κανονιστικών περιορισμών και της περιορισμένης επιθυμίας του ασφαλισμένου να τα διαθέσει στους ασφαλιστές. Δεδομένα που σχετίζονται με συνδεδεμένα αυτοκίνητα, αν και συλλέγονται ήδη,

εξακολουθούν να υφίστανται ανεπαρκή εκμετάλλευση. Ωστόσο, η ισχυρή ανάπτυξη αυτών των αντικειμένων και η γενίκευσή τους σε όλους τους τύπους αντικειμένων που σχεδιάζονται, θα πρέπει να τα καταστήσει επαρκώς ενσωματωμένα στους τρόπους ζωής και στις καταναλωτικές συνήθειες, ώστε να προβλεφθούν ρυθμιστικές αλλαγές και ένας πιο φυσικός τρόπος διάθεσης των δεδομένων (Berthelé, 2018).

Το θέμα που θα διευθετήσουν οι ασφαλιστές θα αφορά μάλλον την κατεύθυνση των επενδύσεών τους σε αυτόν τον τομέα, καθώς δεν θα είναι όλα τα αναπτυγμένα αντικείμενα του ίδιου επιπέδου επιτυχίας. Η τάση σύνδεσης ορισμένων αντικειμένων μπορεί μερικές φορές να οδηγήσει σε κάτι χωρίς νόημα και ο ασφαλισμένος θα μπορούσε να υιοθετήσει αυτά τα αντικείμενα για να κάνει πρωτογενή χρήση των ίδιων των αντικειμένων και όχι για δευτερεύουσα χρήση τους για ασφαλιστικούς σκοπούς (Boobier, 2016).

4.1.3.3. Δεδομένα κοινωνικών δικτύων

Δεδομένης της μαζικής υιοθέτησής τους από τον πληθυσμό και του σημαντικού όγκου πληροφοριών που μεταφέρουν τα προσωπικά και επαγγελματικά κοινωνικά δίκτυα, με την πρώτη εντύπωση φάνηκαν να είναι μια σημαντική πηγή δεδομένων για τους ασφαλιστές και τη στιγμή της συνειδητοποίησης της πιθανής συμβολής των μεγάλων δεδομένων στην αγορά, ορισμένοι ασφαλιστές σκόπευαν να τα αποτιμήσουν. Επί του παρόντος, ο ενθουσιασμός για αυτή την πιθανή αποθήκη δεδομένων, έχει σε μεγάλο βαθμό υποχωρήσει. Αν και τα διαθέσιμα δεδομένα είναι πράγματι πολυάριθμα και δυνητικά ποικίλα, είναι πέρα από τα προβλήματα που σχετίζονται με τη χρήση προσωπικών δεδομένων και σε μεγάλο βαθμό μεροληπτικά, πολλά από αυτά αντικατοπτρίζουν αυτό που ο χρήστης επιθυμεί να επικοινωνήσει στο δίκτυό του. Στην

πραγματικότητα είναι πιο αντιπροσωπευτικά της εμφάνισης, παρά της πραγματικότητας και είναι δύσκολο να χρησιμοποιηθούν για τη μελέτη της συμπεριφοράς ή την ανάλυση του κινδύνου (Ghani et al., 2019).

4.1.3.4. Διατιθέμενα εξωτερικά δεδομένα

Πέρα από τα άμεσα διαθέσιμα δεδομένα μέσω των σχέσεων μεταξύ των ασφαλιστών και των ασφαλισμένων τους και των δεδομένων από συνδεδεμένα αντικείμενα, οι εξωτερικές πηγές δεδομένων μπορεί να έχουν ουσιαστικό ενδιαφέρον τόσο για τα ενδιαφερόμενα μέρη του τραπεζικού, όσο και του ασφαλιστικού τομέα. Σε αυτό το πλαίσιο τα δεδομένα που διατίθενται από ορισμένους δημόσιους φορείς, όπως - για την περίπτωση της Ελλάδας- η ΕΛΣΤΑΤ ή ορισμένοι ιστοχώροι κλιματολογίας/μετεωρολογίας ειδικότερα-, ενδέχεται να έχουν μεγάλο ενδιαφέρον για την κατανόηση της τυπολογίας του ασφαλισμένου ή της έκθεσης στους κινδύνους της ασφαλισμένης περιουσίας (Berthelé, 2018).

Τα δεδομένα που διατίθενται υπό την επιφύλαξη της πληρωμής από ορισμένους ιδιωτικούς οργανισμούς όπως το Bloomberg ή οργανισμούς υπεύθυνους για τη συγκέντρωση δεδομένων, θα μπορούσαν επίσης να αποτιμηθούν, ιδίως για την κατανόηση του μακροοικονομικού περιβάλλοντος και των αλλαγών συμπεριφοράς που επέρχονται με την πάροδο του χρόνου, όσον αφορά τους αντισυμβαλλομένους. Παρόλα αυτά, ακόμα κι αν ο όγκος των διαθέσιμων δεδομένων και οι δυνατότητες που προκαλούνται στη βελτίωση των τιμών αυξηθούν σημαντικά, η εξατομίκευση των τιμών δεν είναι ούτε πλήρως εφικτή, ούτε επιθυμητή για τους ασφαλιστές, τους ασφαλισμένους και την κοινωνία γενικότερα (Boobier, 2016).

Κεφάλαιο 5ο : Στατιστικές Μέθοδοι Μάθησης

5.1. Εισαγωγή

Ο στόχος αυτού του κεφαλαίου είναι να παρουσιάσει τις στατιστικές μεθόδους μάθησης που χρησιμοποιούνται πιο συχνά στην αναλογιστική επιστήμη. Αυτές είναι συμπληρωματικές μέθοδοι με τα πιο συμβατικά στατιστικά μοντέλα, όπως η γραμμική και η λογιστική παλινδρόμηση, που εφαρμόζονται εδώ και πολύ καιρό στην αναλογιστική επιστήμη. Με τη μαζική εισροή ψηφιακών δεδομένων, καθίσταται βολικό να εφαρμόζονται πιο εξελιγμένες μέθοδοι επεξεργασίας δεδομένων και πρόβλεψης. Παραδείγματα πιθανών εφαρμογών περιλαμβάνουν (James et al., 2013):

- εκτίμηση των ποσών και της συχνότητας των ατομικών απαιτήσεων περιουσίας και της ασφάλισης ατυχημάτων·
- εκτίμηση των ατομικών ιατρικών εξόδων στην ασφάλιση υγείας.
- εντοπισμός απάτης κατά τη δήλωση περιουσίας και ασφάλισης ζημιών·
- εντοπισμός δανειοληπτών που διατρέχουν κίνδυνο αθέτησης υποχρεώσεων·
- ανάπτυξη ασφαλιστικών ζωνών λαμβάνοντας υπόψη γεωγραφικές πληροφορίες, όπως ανοιχτά δεδομένα
- λαμβάνοντας υπόψη δεδομένα που συλλέγονται σε πραγματικό χρόνο για οχήματα με σκοπό τον καθορισμό ενός πιο στοχευμένου ποσοστού ασφάλισης αυτοκινήτου (τύπου pay-how-you-drive).

Πριν από την ακριβέστερη παρουσίαση των διαφορετικών μεθόδων μάθησης, θα πρέπει να γίνει μια γενική διάκριση μεταξύ εποπτευόμενων και μη εποπτευόμενων μεθόδων.

5.1.1. Επίβλεψη μάθησης

Σε αυτήν την περίπτωση, η παρουσία μιας μεταβλητής Y που πρέπει να εξηγηθεί είναι θεμελιώδης: έχοντας από κοινού παρατηρήσει τη μεταβλητή Y και τις επεξηγηματικές μεταβλητές σε ένα δείγμα ατόμων, ο στόχος είναι να κατασκευαστεί ένα μοντέλο που θα επιτρέπει την πρόβλεψη της μεταβλητής εξόδου Y όταν εισάγονται νέες ανεξάρτητες μεταβλητές. Το μαθηματικό πλαίσιο της εποπτευόμενης στατιστικής μάθησης είναι αυτό. Ορίζουμε ένα δείγμα μάθησης:$L_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, δηλαδή, μια ακολουθία ανεξάρτητων και πανομοιότυπα κατανεμημένων τυχαίων διανυσμάτων (i.i.d.), που έχουν τον ίδιο νόμο με ένα τυχαία διάνυσμα (X,Y). Ο νόμος του ζεύγους (X,Y) είναι άγνωστος και ο στόχος της εποπτευόμενης στατιστικής μάθησης είναι να τον μάθει παρατηρώντας τη βάση μάθησης L_n. Οι τυχαίες μεταβλητές ή τα διανύσματα X και Y δεν παίζουν τον ίδιο ρόλο: στην πράξη, το X αναφέρεται στη μεταβλητή εισόδου (συνήθως ένα διάνυσμα που αποτελείται από επεξηγηματικές μεταβλητές) και το Y αναφέρεται στη μεταβλητή εξόδου (ονομάζεται επίσης μεταβλητή που πρέπει να εξηγηθεί). Η μεταβλητή Y μπορεί επίσης να είναι πολυδιάστατη. Ο σκοπός της εποπτευόμενης μάθησης είναι να μάθουμε τη σύνδεση μεταξύ των μεταβλητών X και Y: σημειώστε ότι η γνώση των οριακών νόμων των X και Y δεν είναι αρκετή για να γνωρίζουμε το νόμο του ζεύγους (X,Y). Είναι μέσω μιας στατιστικής προσέγγισης που βασίζεται στην παρατήρηση του L_n που θα λάβουμε πληροφορίες, καθιστώντας δυνατή την πρόβλεψη μιας τιμής $\hat{y}$ που σχετίζεται με μια νέα παρατήρηση της μεταβλητής εισόδου x, για την οποία η μεταβλητή y που πρέπει να εξηγηθεί είναι άγνωστη. Ανάλογα με τη φύση της μεταβλητής εξόδου y, μπορούμε να διακρίνουμε δύο εποπτευόμενα πλαίσια στατιστικής μάθησης: παλινδρόμηση και ταξινόμηση (James et al., 2013).

Στην περίπτωση της παλινδρόμησης, η μεταβλητή εξόδου Y είναι μια ποσοτική μεταβλητή, με πραγματικές τιμές. Το τυπικό πλαίσιο είναι τότε:

$$Y = f(X) + \varepsilon,$$

όπου f είναι η άγνωστη συνάρτηση, που ονομάζεται συνάρτηση παλινδρόμησης που επιδιώκουμε να εκτιμήσουμε. Η μεταβλητή ε είναι μια πραγματική τυχαία μεταβλητή μηδενικής προσδοκίας: αντιπροσωπεύει το σφάλμα και καθιστά δυνατό να ληφθεί υπόψη το γεγονός ότι η μεταβλητή Y εξηγείται μόνο εν μέρει από τη μεταβλητή X. Στη συνέχεια, η μαθησιακή βάση αποτελείται από ζευγάρια:$(X_i, Y_i) = f(X_i) + \varepsilon_i$, όπου οι ε_i είναι ανεξάρτητες τυχαίες μεταβλητές και του ίδιου νόμου (Berk, 2008).

Υποθέτουμε επίσης ότι η μεταβλητή σφάλματος ε είναι υπό όρους κεντραρισμένη στο X. Τότε έχουμε: $E[Y|X] = f(X)$, το οποίο ορίζει τη συνάρτηση f με μοναδικό τρόπο. Χωρίς κανέναν πρόσθετο περιορισμό στη συνάρτηση, μπορούμε να εξετάσουμε ένα μη παραμετρικό μοντέλο παλινδρόμησης, το οποίο επομένως συνίσταται στην κατασκευή ενός προγνωστικού $\hat{f}$ από τη βάση εκμάθησης L_n, καθιστώντας δυνατή την καλύτερη πρόβλεψη της μεταβλητής εξόδου y που σχετίζεται με μια είσοδο x. Η ποιότητα του $\hat{f}$ μπορεί να μετρηθεί, για παράδειγμα, με το γενικευμένο μέσο τετράγωνο σφάλμα που ορίζεται από:$E[\left(\hat{f}(x) - Y\right)^2]$. Δεδομένου ότι ο νόμος του ζεύγους (X, Y) είναι άγνωστος, στην πράξη, η απόδοση του προγνωστικού δείκτη μετριέται μάλλον με ένα εμπειρικό μέσο τετραγωνικό σφάλμα:

$$R_m\left(\hat{f}, \tilde{L}_m\right) = \frac{1}{m}\sum_{i=1}^{m}(\tilde{Y}_i - \hat{f}(\tilde{X}_i))^2,$$

που ορίζεται σε ένα δείγμα δοκιμής: $\tilde{L}_m = \{\left(\tilde{X}_1, \tilde{Y}_1\right), \ldots, \left(\tilde{X}_m, \tilde{Y}_m\right)$ μεγέθους m, ανεξάρτητο από το L_n, όπου τα $\left(\tilde{X}_i, \tilde{Y}_i\right)$ είναι του ίδιου νόμου με το ζεύγος (X, Y).

Για να εφαρμοστεί πρακτικά μια εποπτευόμενη μέθοδος μάθησης, είναι απαραίτητο να περιοριστεί η κατηγορία των συναρτήσεων παλινδρόμησης που εξετάζονται και οι τεχνικές για τον προσδιορισμό ενός προγνωστικού παράγοντα θα εξαρτηθούν στη συνέχεια από αυτήν την επιλογή. Μια συμβατική επιλογή που δεν θα συζητήσουμε εδώ είναι η γραμμική παλινδρόμηση, όπου η f(X) υποτίθεται ότι είναι ένας γραμμικός συνδυασμός επεξηγηματικών μεταβλητών. Στην περίπτωση αυτή, η ελαχιστοποίηση του εμπειρικού μέσου τετραγώνου σφάλματος που υπολογίζεται στο σύνολο εκμάθησης οδηγεί στην επίλυση ενός συστήματος γραμμικών εξισώσεων, το οποίο παρέχει μια ρητή λύση. Για πιο εξελιγμένα μοντέλα, κυρίως μη γραμμικά, η επιλογή και η αναζήτηση του «καλύτερου» μοντέλου δεν είναι πλέον αρκετά απλή, όπως θα δούμε στις επόμενες ενότητες (Berk, 2008).

Στην περίπτωση της ταξινόμησης, η μεταβλητή εξόδου Y είναι διακριτή και αντιπροσωπεύει την κλάση στην οποία ανήκει η παρατήρηση. Αν το K υποδηλώνει τον αριθμό των κλάσεων, μπορούμε να θεωρήσουμε ότι το Y έχει τιμές στο {1,…, K}.Ο στόχος είναι να κατασκευάσουμε ένα μοντέλο που να συσχετίζει σε κάθε διάνυσμα εισόδου X μια κλάση στο {1,…, K}. Μια κοινή προσέγγιση είναι να προχωρήσουμε σε δύο βήματα: το μοντέλο υπολογίζει πραγματικά τις ποσότητες: $P[Y = k|\ X = x]$, για k ∈ {1,…, K}, από την οποία επιλέγεται η κλάση που σχετίζεται με την παρατήρηση x (James et al., 2013).

Στις επόμενες ενότητες θα παρουσιαστούν οι ακόλουθες εποπτευόμενες μεθόδους μάθησης: δέντρα αποφάσεων, πολυεπίπεδα νευρωνικά δίκτυα, SVM και μέθοδοι συνάθροισης μοντέλων (random forests, bagging, stacking, boosting).

5.1.2. Εκμάθηση χωρίς επίβλεψη

Σε αυτή την περίπτωση, δεν υπάρχει καμία μεταβλητή προς εξήγηση, η οποία επομένως αφορά μάλλον ένα πρόβλημα ομαδοποίησης. Ο στόχος είναι η κατασκευή ομοιογενών κλάσεων που ομαδοποιούν τα πιο παρόμοια άτομα (σε σχέση με τις μεταβλητές που τα περιγράφουν) και οι κλάσεις πρέπει να είναι όσο το δυνατόν ανόμοιες. Στόχος είναι να οργανωθούν οι πληροφορίες που περιέχονται στα δεδομένα, ώστε να γίνουν πιο ορατές και καλύτερα εκμεταλλεύσιμες. Μεταξύ των συμβατικών μεθόδων, έχουμε την αύξουσα ιεραρχική ταξινόμηση και αλγόριθμους με δυναμική ανακατανομή (k-means) (Hiran et al., 2021). Σε επόμενη ενότητα, θα παρουσιαστεί η μέθοδος αυτοοργάνωσης του χάρτη Kohonen, η οποία έχει το πλεονέκτημα ότι επιτρέπει τη γραφική αναπαράσταση των τάξεων σε μικρό χώρο.

5.2. Δέντρα απόφασης

Η αρχή των δέντρων απόφασης είναι να διαιρείται ο χώρος των τιμών των επεξηγηματικών μεταβλητών σε ορθογώνια, στα οποία η μεταβλητή που πρέπει να εξηγηθεί είναι σταθερή. Αυτή η πολύ απλή ιδέα, η οποία αναπτύχθηκε από τους Breiman et al. (1984) με το ακρωνύμιο CART (δέντρο ταξινόμησης και παλινδρόμησης), καθιστά δυνατή την απόκτηση ενός εύχρηστου και απλούστερου στην ερμηνεία μοντέλου, το οποίο αποτελεί έτσι ένα γνήσιο εργαλείο για την υποστήριξη της λήψης αποφάσεων. Επιπλέον, αυτή η μέθοδος, η οποία συνίσταται στην αναδρομική κατάτμηση του χώρου εισόδου με δυαδικό τρόπο, μπορεί να εφαρμοστεί τόσο στην παλινδρόμηση όσο και στην ταξινόμηση (Steinberg and Colla, 2009).

Ας περιγράψουμε τη μέθοδο στην περίπτωση του παρακάτω μοντέλου (Trendowicz et al., 2014): (X, Y) είναι ένα τυχαίο διάνυσμα με τιμές στο $R^p \times G$, όπου $G = \mathbb{R}$ στην περίπτωση παλινδρόμησης και $G = \{1, \ldots, K\}$ στην περίπτωση της ταξινόμησης. Υποθέτουμε ότι γνωρίζουμε ένα σύνολο εκμάθησης: $L_n = \{(x_i, y_i) \in R^p \times G, i = 1, \ldots, n\}$, όπου (x_i, y_i) είναι ανεξάρτητες πραγματοποιήσεις τυχαίων μεταβλητών με τον ίδιο νόμο με τα (X, Y). Σε κάθε στάδιο της κατάτμησης, μέρος του χώρου εισόδου χωρίζεται σε δύο υποτμήματα και ένα δυαδικό δέντρο συνδέεται φυσικά με τον κατασκευασμένο διαχωρισμό:

- η ρίζα n_1 του δέντρου συσχετίζεται με ολόκληρο τον χώρο εισόδου και επομένως περιέχει όλες τις παρατηρήσεις του L_n
- το πρώτο βήμα του CART περιλαμβάνει τη διαίρεση αυτού του χώρου στα δύο, επιλέγοντας ένα διαχωρισμό της μορφής: $\{X^j \leq s\} \cup \{X^j > s\}$, όπου j ∈{1, …, p}, $X = (X^1, \ldots, X^p)$ και s ∈ $\mathbb{R}$. Έτσι, ο διαχωρισμός σημαίνει ότι όλες οι παρατηρήσεις που έχουν τιμή της j-στης μεταβλητής μικρότερη από s αντιστοιχίζονται στο αριστερό υποδέντρο και οι άλλες στο δεξιό υποδέντρο. Για να γίνει αυτό, η μέθοδος επιλέγει τον καλύτερο δυνατό διαχωρισμό (j, s), ελαχιστοποιώντας μια συνάρτηση κόστους C(j, s).

Αν $x_i = (x_i^1, \ldots, x_i^p)$,

$$n_{1,-}(j, s) = \{1 \in \{1, \ldots, n\}: x_i^j \leq s\}$$

και

$$n_{1,+}(j, s) = \{1 \in \{1, \ldots, n\}: x_i^j > s\}$$

Στην περίπτωση της παλινδρόμησης, η συνάρτηση προς ελαχιστοποίηση είναι

$$C(s, j) = \sum_{i \in n_{1,-}(j,s)} (y_i - \overline{y_-})^2 + \sum_{i \in n_{1,+}(j,s)} (y_i - \overline{y_+})^2$$

όπου

$$\overline{y_-} = \frac{1}{card(n_{1,-}(j,s)} \sum_{i \in n_{1,-}(j,s)} y_i$$

και

$$\overline{y_+} = \frac{1}{card(n_{1,+}(j,s)} \sum_{i \in n_{1,+}(j,s)} y_i$$

Επομένως, επιδιώκουμε να ελαχιστοποιήσουμε τη διακύμανση των δύο υποδέντρων που προκύπτουν. Στην περίπτωση ταξινόμησης, η συνάρτηση που πρέπει να ελαχιστοποιηθεί είναι:

$$C(j,s) = \sum_{k=1}^{K} \hat{p}^k_{n_{1,-(j,s)}} \left(1 - \hat{p}^k_{n_{1,-(j,s)}}\right) + \sum_{k=1}^{K} \hat{p}^k_{n_{1,+(j,s)}} \left(1 - \hat{p}^k_{n_{1,+(j,s)}}\right) \qquad [3.8]$$

όπου $\hat{p}^k_{n_{1,-(j,s)}}$ και $\hat{p}^k_{n_{1,+(j,s)}}$ είναι η αναλογία των παρατηρήσεων της τάξης k στο σύνολο $n_{1,-}(j,s)$ και $n_{1,+}(j,s)$, αντίστοιχα. Σε αυτήν την περίπτωση, προσπαθούμε να ελαχιστοποιήσουμε τον δείκτη Gini κάθε συνόλου και έτσι να λάβουμε τα πιο ομοιογενή υποδέντρα (ένα σύνολο είναι απολύτως ομοιογενές εάν όλες οι παρατηρήσεις είναι στην ίδια κατηγορία).

Μόλις διαμεριστεί η ρίζα του δέντρου, η διαδικασία επαναλαμβάνεται σε καθένα από τα δύο υποδέντρα που λαμβάνονται, αναζητώντας ξανά τη βέλτιστη διαίρεση για την επιλεγμένη συνάρτηση κόστους και ούτω καθεξής μέχρι να επιτευχθεί το κριτήριο διακοπής. Ένα κλασικό κριτήριο διακοπής περιλαμβάνει τη μη διάσπαση ενός κόμβου του δέντρου που περιέχει λιγότερο από έναν σταθερό αριθμό παρατηρήσεων. Οι τερματικοί κόμβοι, που δεν χωρίζονται πλέον, ονομάζονται φύλλα του δέντρου. Θα πρέπει να σημειωθεί ότι ένας καθαρός κόμβος, δηλαδή ένας κόμβος που περιέχει μόνο παρατηρήσεις με την ίδια τιμή εξόδου, δεν χωρίζεται. Το δέντρο T_{max} που λαμβάνεται με αυτή τη διαδικασία ονομάζεται μέγιστο δέντρο, και για μια παρατήρηση που ανήκει

σε ένα φύλλο, αυτό το μοντέλο προβλέπει ως τιμή εξόδου παλινδρόμησης τη μέση τιμή: $\overline{y_n} = \frac{1}{card(n)} \sum_{i \in n} y_i$ (και η πλειοψηφική τάξη των παρατηρήσεων του L_n που είναι παρούσα στο φύλλο, στην ταξινόμηση).

Το δεύτερο βήμα του αλγορίθμου CART αποτελείται από ένα βήμα κλαδέματος, το οποίο θα επιλέξει το δέντρο που έχει κλαδευτεί καλύτερα από το μέγιστο δέντρο T_{max} (με την έννοια του γενικευμένου σφάλματος). Αυτό το βήμα είναι απαραίτητο επειδή το T_{max}, λόγω της πολύ λεπτής κατασκευής του, είναι χαμηλής προκατάληψης, αλλά μπορεί να έχει πολύ μεγάλη διακύμανση: εξαρτάται πολύ από τις παρατηρήσεις που χρησιμοποιούνται, αλλά υπόκειται επίσης σε υπερμάθηση. Είναι επομένως απαραίτητο να επιλεγεί ένα μοντέλο που μπορεί να είναι κάπως λιγότερο ακριβές αλλά ικανό να παρέχει προβλέψεις ισοδύναμης ποιότητας για νέες παρατηρήσεις.

Σε αντίθεση με το μέγιστο δέντρο, το δέντρο που αποτελείται μόνο από τη ρίζα του έχει μηδενική διακύμανση, αλλά σημαντική προκατάληψη. Ο στόχος είναι επομένως να βρεθεί ένα ενδιάμεσο δέντρο μεταξύ αυτών των δύο άκρων. Για να το κάνουμε αυτό, ξεκινάμε κατασκευάζοντας μια ακολουθία $(T_j)_{1 \leq j \leq J}$ υποδέντρων που κλαδεύονται το ένα από το άλλο, από το T_{max}, που αντιστοιχεί σε μια οικογένεια ένθετων κατατμήσεων. Ας περιγράψουμε τη διαδικασία στην περίπτωση της παλινδρόμησης: Λαμβάνουμε αυτήν την ακολουθία $(T_j)_{1 \leq j \leq J}$ ελαχιστοποιώντας ένα κριτήριο απόκλισης που ορίζεται για κάθε κλαδεμένο υποδέντρο T του T_{max} και για όλα τα $a \geq 0$, όπου:

$$Crit_a(T) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_{T,i}\right)^2 + a|T|$$

όπου |T|, είναι ο αριθμός των φύλλων του T και $\hat{y}_{T,i}$ είναι η τιμή που προβλέπεται από το μοντέλο που συσχετίζεται με το T για την είσοδο x_i. Αυξάνοντας σταδιακά το a, η

ελαχιστοποίηση του $Crit_a$ παρέχει μια ακολουθία $(T_j)_{1\leq j\leq J}$ με ολοένα και λιγότερα φύλλα. Στη συνέχεια απομένει να επιλεγεί ένας υποψήφιος με την ακολουθία που λαμβάνεται με αυτόν τον τρόπο. Οι Breiman et al.(1984) πρότειναν ουσιαστικά δύο μεθόδους: χρήση δείγματος δοκιμής ή διαδικασία διασταυρούμενης επικύρωσης. Η πρώτη μέθοδος προϋποθέτει ότι υπάρχει μια διαθέσιμη βάση δοκιμής: $\tilde{L}_{n_t} = \{(\tilde{x}_i, \tilde{y}_i) \in \mathbb{R}^p \times G, i = 1, \dots, n_t$ ανεξάρτητες από το L_n, όπου $(\tilde{x}_i, \tilde{y}_i)$ αποτελούν ανεξάρτητες πραγματοποιήσεις τυχαίων μεταβλητών με τον ίδιο νόμο με τα (X, Y). Στη συνέχεια επιλέγουμε το υποδέντρο του $(T_j)_{1\leq j\leq J}$ με δείκτη:

$$j^* = argmin_{1\leq j\leq J}\{\frac{1}{n_t}\sum_{i=1}^{n_t}\left(\tilde{y}_i - \hat{y}_{T_j,i}\right)^2\}$$

όπου το $\hat{y}_{T_j,i}$ είναι η μεταβλητή εξόδου που σχετίζεται με το $\tilde{x}_i$.

Όπως αναφέρθηκε προηγουμένως, ένα από τα πλεονεκτήματα των δέντρων απόφασης είναι η αναγνωσιμότητα και η ευκολία ερμηνείας των αποτελεσμάτων, χάρη στο δέντρο γράφημα που παρέχει μια φυσική κατανόηση του μοντέλου. Ένα άλλο πλεονέκτημα είναι ότι αυτή η μέθοδος μπορεί να αντιμετωπίσει ανεπεξέργαστα δεδομένα. Επιπλέον, σε αντίθεση με τα γενικευμένα γραμμικά μοντέλα, είναι δυνατό να ληφθούν υπόψη γραμμικά σχετιζόμενες επεξηγηματικές μεταβλητές και η σύνδεση μεταξύ της μεταβλητής εξόδου και των άλλων μεταβλητών μπορεί να είναι μη γραμμική (Quan and Valdez, 2018).

Ωστόσο, αυτή η μέθοδος έχει ορισμένα μειονεκτήματα: για παράδειγμα, τα μοντέλα που λαμβάνονται δεν είναι ιδιαίτερα αξιόπιστα επειδή εξαρτώνται σε μεγάλο βαθμό από το μαθησιακό δείγμα. Μικρές αλλαγές στη βάση μάθησης μπορούν να αλλάξουν σημαντικά το μοντέλο που προκύπτει. Αυτός είναι ένας από τους λόγους για τους οποίους τα δέντρα απόφασης χρησιμοποιούνται πλέον ως μπλοκ πιο εξελιγμένων

μεθόδων χρησιμοποιώντας έναν μεγάλο αριθμό μοντέλων παράλληλα: μέθοδοι συνάθροισης μοντέλων (boosting, bagging, random forests) που θα παρουσιαστούν σε επόμενη ενότητα. Σημειώνουμε επίσης ότι τα δέντρα απόφασης δεν δίνουν μια συνολική τμηματοποίηση για κάθε επεξηγηματική μεταβλητή: είναι δυνατό μια μεταβλητή να εμφανίζεται πολλές φορές σε διαφορετικές διαδρομές (Rokach and Maimon, 2005).

Η μέθοδος CART δέχεται παραλλαγές (Breiman et al., 1984) και υπάρχουν επίσης άλλες μέθοδοι για την κατασκευή δέντρων αποφάσεων, όπως ο αλγόριθμος C4.5 που εισήχθη από τον Quinlan (1996) και χρησιμοποιείται ευρέως στην κοινότητα της πληροφορικής. Είναι επίσης δυνατή η κατασκευή προγνωστικών βασισμένων σε πιο κανονική αναδρομική κατάτμηση από τα δέντρα απόφασης, τα οποία ορίζουν σταθερές συναρτήσεις τμηματικά (για παράδειγμα, ο αλγόριθμος MARS του Friedman (1991)).

Τα δέντρα αποφάσεων έχουν εφαρμογές σε πολλά πεδία. Μπορούν να χρησιμοποιηθούν στην αναλογιστική επιστήμη, για παράδειγμα, στην τιμολόγηση εκτός ζωής, όπως διαπιστώνεται από το άρθρο των Quan and Valdez (2018).

5.3. Νευρωνικά δίκτυα

Αυτή η ενότητα είναι μια εισαγωγή στα «νευρωνικά δίκτυα» ή «συνδεσιωτικές μέθοδοι», που μπορούν να οριστούν ως το σύνολο των αριθμητικών μεθόδων επίλυσης προβλημάτων που χρησιμοποιούν μοντέλα που προέρχονται από τη νευροβιολογία. Οι πρώτοι σημαντικοί προβληματισμοί έγιναν γύρω στο 1940 και μπορούν να αναφερθούν ορισμένα ονόματα όπως ο Turing, ο Pitts, ο Wiener, ο von Neumann και ο McCulloch. Ο στόχος ήταν τότε να χρησιμοποιήσουμε τη νέα γνώση που έφερε η βιολογία και οι γνωστικές επιστήμες στον εγκέφαλο για να σχεδιαστούν συστήματα υπολογιστών με

ορισμένες από τις ιδιότητές τους, όπως προσαρμοστική μάθηση με διαδοχικές τοπικές τροποποιήσεις και μετεγκατάσταση της αποθήκευσης πληροφοριών, με αποτέλεσμα, για παράδειγμα, την ευρωστία σε περιπτώσεις μερικής καταστροφής. Από αυτά τα αρχικά έργα προέκυψαν δύο σχολές σκέψης. Η πρώτη (Turing, von Neumann, κ.λπ.) υιοθέτησε μια συμβολική προσέγγιση και βρισκόταν στην αρχή των εννοιών που εξακολουθούν να χρησιμοποιούνται στους υπολογιστές μας (μνήμη, επεξεργαστές) και στην «παραδοσιακή» τεχνητή νοημοσύνη. Η δεύτερη (Minsky, Pitts, McCulloch, κ.λπ.) ακολούθησε οικειοθελώς μια «συνδεσιοκρατική» προσέγγιση πιο κοντά στη βιολογική περιγραφή (νευρομιμητισμός). Αυτή η προσέγγιση είδε μια αναβίωση όταν οι δυσκολίες της παραδοσιακής τεχνητής νοημοσύνης ήρθαν στο φως τη δεκαετία του 1980. Τα αποτελέσματα ήταν μερικές φορές θεαματικά και προκαλούν σήμερα πολύ προηγμένες πρακτικές εφαρμογές (αναγνώριση μορφών και γραφής, σύνθεση ομιλίας, πρόβλεψη χρονοσειρών, ιατρική διαγνωστική βοήθεια, κ.λπ.), αλλά η μαθηματική ανάλυση αυτών των μη γραμμικών μοντέλων παραμένει πολύ περίπλοκη (Müller, Reinhardt and Strickland, 1995).

Τα νευρωνικά δίκτυα αποτελούν ένα τεράστιο πεδίο έρευνας από θεωρητική σκοπιά και από εφαρμογές (βλ., για παράδειγμα, το ολοκληρωμένο βιβλίο των Du and Swamy (2006). Θα επικεντρωθούμε στα πολυεπίπεδα δίκτυα και τον γνωστό αλγόριθμο οπισθοδιαβίβασης κλίσης, που προσφέρουν πολύ ενδιαφέρουσες εφαρμογές στην εποπτευόμενη μάθηση και αποτελούν συμπληρωματικές μεθόδους σε απλούστερα και πιο συμβατικά στατιστικά μοντέλα (γραμμική και λογιστική παλινδρόμηση). Πρόσφατα, υπήρξε ανανεωμένο ενδιαφέρον για αυτές τις μεθόδους, ακολουθούμενο από την ανάπτυξη ακόμη πιο αποτελεσματικών μεθόδων εκμάθησης: η βαθιά μάθηση, ειδικότερα, έχει ήδη φέρει επανάσταση στον κόσμο της τεχνητής νοημοσύνης (Aggarwal, 2018).

Θα πρέπει να σημειωθεί ότι ο αλγόριθμος Kohonen και οι μέθοδοι SVM, που θα παρουσιαστούν στις επόμενες ενότητες, αντλούν επίσης την προέλευσή τους από τη νευρωνική μοντελοποίηση (Huang, 2009). Η ανάπτυξη του SVM, για παράδειγμα, επηρεάστηκε αρχικά από την εργασία του Rosenblatt για το Perceptron το 1962 (Aggarwal, 2018).

5.3.1. Από πραγματικό σε επίσημο νευρώνα

Μία από τις πρώτες νευρομιμητικές κατασκευές της βασικής μονάδας νευρωνικού υπολογισμού αποδίδεται στους McCulloch and Pitts (1943). Ο πραγματικός νευρώνας, πολύ σχηματικά, λειτουργεί ως εξής:

- σταθμισμένη άθροιση των νευρικών ερεθισμάτων (αναστολείς ή διεγέρτες) από τους νευρώνες με τους οποίους συνδέεται, μέσω δενδριτών και συνάψεων.
- εκπομπή στον άξονα μιας εισροής εάν το άθροισμα εισόδου υπερβαίνει ένα όριο ενεργοποίησης.

Ο επίσημος νευρώνας των McCulloch and Pitts αναπαράγει αυτές τις δύο ιδιότητες. Για μια είσοδο $x \in \mathbb{R}^p$, η μονάδα συναπτικών βαρών: $w \in \mathbb{R}^p$ και το κατώφλι (threshold) ενεργοποίησης: $\theta \in \mathbb{R}$ υπολογίζει:

- Το σταθμισμένο άθροισμα $w \cdot x = \sum_{i=1}^{p} w_i x_i$
- Την έξοδο $P_w(x) = \Phi(w \cdot x - \theta) = \mathbb{I}_{\{w \cdot x \geq \theta\}}$

όπου η συνάρτηση $\mathbb{I}_{\{A\}}$ είναι 1 αν το A είναι πραγματικός και 0 αν όχι. Η συνάρτηση Φ ονομάζεται συνάρτηση ενεργοποίησης ή συνάρτηση απόκρισης. Απλοποιούμε τους συμβολισμούς θέτοντας το $w_{l+1} = \theta$ και λαμβάνοντας υπόψη μια πρόσθετη είσοδο

$x_{p+1} = -1$ έτσι ώστε να μπορούμε να γράψουμε $P_w(x) = \mathbb{I}_{\{w \cdot x \geq \theta\}}$, όπου $w \cdot \tilde{x} = \sum_{i=1}^{p+1} w_i x_i$ και $\tilde{x} = (x_1, \ldots, x_{p+1})$.

Σημειώστε ότι ένας τέτοιος νευρώνας εκτελεί μια συνάρτηση του $\mathbb{R}^p$ στο $\{0,1\}$. Δεν υπήρχε λόγος να περιοριστεί κανείς στη συνάρτηση απόκρισης του Heaviside του $\Phi(u) = \mathbb{I}_{\{u \geq 0\}}$ και άλλες λειτουργίες την έχουν αντικαταστήσει κατάλληλα σε πολλές εφαρμογές:

- $\Phi(u) = u$ (γραμμικός νευρώνας ή μέθοδος Widrow–Hoff).

$\Phi_T(u) = \frac{1}{1+exp\left(-\frac{u}{T}\right)}$ που ονομάζεται θερμοκρασία T σιγμοειδής, που συγκλίνει απλώς στο $\Phi_0(u) = \mathbb{I}_{\{u \geq 0\}}$ όταν $T \to 0$. Ένα από τα πλεονεκτήματα αυτής της συνάρτησης ενεργοποίησης, σε σχέση με τη συνάρτηση απόκρισης του Heaviside, είναι ότι είναι συνεχής και διαφοροποιήσιμη παντού, κάτι που θα είναι απαραίτητο για την πλειοψηφία των αλγορίθμων εκμάθησης (Dubois, 1998).

- $\Phi_{T,a,b}(u) = a + (b-a)\Phi_T(u)$ σιγμοειδής με τιμές στο $[a, b]$ και συγκεκριμένα $\Phi_{T,-1,1}(u) = tanh\left(\frac{u}{2T}\right)$

Αυτά τα πρώιμα έργα του McCulloch and Pitts γέννησαν τη σχολή των συνδετικών. Ωστόσο, μόλις το 1960 παρατηρήθηκε η πρώτη εφαρμογή: το Perceptron του Rosenblatt. Εισηγμένο το 1962 από τον ψυχολόγο Rosenblatt, το Perceptron υποδηλώνει ένα σύνολο συνδεδεμένων επίσημων νευρώνων. Ο στόχος του Rosenblatt ήταν να μοντελοποιήσει την οπτική αναγνώριση εικόνων, εμπνευσμένη από τη βιολογική δομή της όρασης: ο αμφιβληστροειδής, η προβολή και οι συνειρμικές περιοχές μοντελοποιούνται από ένα φορέα εισόδου και διαδοχικά στρώματα νευρώνων που συνδέονται μεταξύ τους (Kussul et al., 2001). Θα δούμε παρακάτω ότι τα τεχνητά πολυστρωματικά νευρωνικά δίκτυα αναπαράγουν αυτή τη δομή. Πριν μελετήσουμε τα

νευρωνικά δίκτυα, ας δούμε ποιες λειτουργίες μπορεί πραγματικά να υπολογίσει το απλό Perceptron, δηλαδή ο απλός νευρώνας των McCulloch και Pitts. Ακολουθούμε την ιστορική προσέγγιση της συνδετικιστικής σχολής.

5.3.2 Απλό Perceptron ως γραμμικός διαχωριστής

Επικεντρωνόμαστε στην συνάρτηση: $P_w(x) = \mathbb{I}_{\{w \cdot x \geq \theta\}}$, από το $\mathbb{R}^p$ στο $\{0,1\}$που ορίζεται από τα συναπτικά βάρη: $w \in \mathbb{R}^{p+1}$. Εισάγουμε την παρακάτω έννοια (Aggarwal, 2018):

Δύο σύνολα $A, B \in \mathbb{R}^l$ λέγονται γραμμικά διαχωρίσιμα αν υπάρχει $w \in \mathbb{R}^{p+1}$ τέτοιο ώστε:

- *Κάθε x στο A ικανοποιεί τη σχέση: $\sum_{i=1}^{p} w_i x_i \geq w_{p+1}$, και*
- *Κάθε x στο B ικανοποιεί τη σχέση: $\sum_{i=1}^{p} w_i x_i < w_{p+1}$.*

Παρόμοια, μια συνάρτηση $f: D \subset \mathbb{R}^p \to \{0,1\}$ λέγεται γραμμικά διαχωρίσιμη αν τα σύνολα $A = \{x \in D: f(x) = 1\}$ και $B = \{x \in D: f(x) = 0\}$ είναι γραμμικά διαχωρίσιμα.

Εξ ορισμού, μια δοσμένη συνάρτηση $f: D \subset \mathbb{R}^p \to \{0,1\}$ είναι γραμμικώς διαχωρίσιμη αν και μόνο αν υπάρχει μια P_w συνάρτηση που ορίζεται όπως παραπάνω και τέτοια ώστε $f(x) = P_w(x)$ για κάθε $x \in D$.

Ακόμα και να περιοριστούμε στην περίπτωση των συναρτήσεων Boolean $f: \{0,1\}^p \to \{0,1\}$, δεν γνωρίζουμε έναν γενικό τύπο που να δίνει τον αριθμό των γραμμικά διαχωρίσιμων συναρτήσεων. Το πιο διάσημο παράδειγμα είναι η περίπτωση $p = 2$: οι λογικές συναρτήσεις *and* και *or* είναι γραμμικά διαχωρίσιμες, ενώ η συνάρτηση xor δεν είναι. Πράγματι, τα σύνολα $A = xor^{-1}(1) = \{(0,1), (1,0)\}$ και

$B = xor^{-1}(0) = \{(0,0), (1,1)\}$ δεν διαχωρίζονται με ευθεία γραμμή στο $\mathbb{R}^2$. Σημειώστε ότι αυτό το αρνητικό αποτέλεσμα (μη υπολογίσιμο με απλό Perceptron) παρατηρήθηκε το 1969 από τους Minsky and Papert. Αν και είχαν επίσης αποδείξει εποικοδομητικά ότι ένα απλό Perceptron μπορεί να υπολογίσει οποιαδήποτε γραμμικά διαχωρίσιμη συνάρτηση, αυτό το αποτέλεσμα έβαλε τέλος στη μόδα του συνδεσινισμού. Εάν για τα μικρά p είναι εύκολο να προσδιοριστεί το P_w με πραγματοποίηση μιας δεδομένης γραμμικά διαχωρίσιμης συνάρτησης Boole, δεν είναι το ίδιο σε μια πιο γενική περίπτωση. Ο Rosenblatt δημιούργησε έναν αλγόριθμο που καθορίζει τους συντελεστές w μέσω μιας μαθησιακής διαδικασίας (Yadav et al., 2015).

Έστω A και B δύο πεπερασμένα σύνολα του $\mathbb{R}^p$, αυστηρά διαχωρίσιμα, δηλαδή υπάρχει $w^* \in \mathbb{R}^{p+1}$, τέτοιο ώστε $w^* \cdot \tilde{x} > 0$ για κάθε $x \in A$ και $w^* \cdot \tilde{x} < 0$ για κάθε $x \in B$, όπου $\tilde{x} = (x_1, \dots, x_{p+1})$. Είναι εύκολο να δειχτεί ότι η συνθήκη να είναι τα A και B αυστηρά γραμμικά διαχωρίσιμα είναι ισοδύναμη με το να είναι τα A και B γραμμικά διαχωρίσιμα. Αυτό που ψάχνουμε είναι ένα w^* διάνυσμα που δημιουργεί αυτόν τον γραμμικό διαχωρισμό. Ο SPL (Simple Learning Perceptron) αλγόριθμος δίνει μια λύση (Popescu et al., 2009).

Υποθέτουμε το σύνολο $A \cup B = \{x(1), \dots, x(M)\}$.

Αλγόριθμος SPL:

1) Αρχικοποίηση του w: Επιλέγουμε τυχαία $w \ \ στο \ \ \mathbb{R}^{p+1}$.
2) $t \coloneqq 0$; test:=0
3) $Για$ i:=1 έως M

$$If\ x(i) \in A\ \ και\ w \cdot \tilde{x}(i) \leq 0,\ \ τότε\ w(t+1) \coloneqq w(t) + \tilde{x}(i);\quad test \coloneqq 1;$$

$$Αλλιώς$$

If $x(i) \in B$ $και$ $w \cdot \tilde{x}(i) \geq 0$, $τότε$ $w(t+1) \coloneqq w(t) - \tilde{x}(i)$; $test \coloneqq 1$;

$Αλλιώς$

$w(t+1) \coloneqq w(t)$;

End If $t \coloneqq t+1$

End For.

4) $If\ test = 1\ go\ to$ 3.

   $End\ If$.

5) End.

A priori, αυτός ο αλγόριθμος μπορεί να μην σταματήσει ποτέ. Ωστόσο, οι Minsky and Papert έδειξαν το 1969 ότι αυτό δεν συμβαίνει και ότι ο αλγόριθμος SPL συγκλίνει σε πεπερασμένο χρόνο t, δίνοντας ένα διάνυσμα $w(t) \in \mathbb{R}^{p+1}$, τέτοιο ώστε $w(t) \cdot \tilde{x} > 0\ για\ κάθε\ x \in A$ και $w(t) \cdot \tilde{x} < 0\ για\ κάθε\ x \in B$. Ο αλγόριθμος σταματά σε έναν πεπερασμένο αριθμό βημάτων. Ωστόσο, στην πράξη, η σύγκλιση μπορεί μερικές φορές να είναι αργή: οι Minsky και Papert έδειξαν ότι στις πιο δυσμενείς περιπτώσεις, ο αριθμός των βημάτων μάθησης αυξάνεται εκθετικά με τον αριθμό των στοιχείων που πρέπει να ταξινομηθούν. Υπάρχουν πιθανές επεκτάσεις: για παράδειγμα, διαχωρισμός με μια σφαίρα και όχι με ένα υπερεπίπεδο, ή ταξινόμηση όχι πλέον σε δύο τάξεις από ένα υπερεπίπεδο αλλά σε πολλές κατηγορίες, ή ακόμη και προσδιορισμός ενός υπερεπίπεδου που χωρίζει στην καλύτερη περίπτωση ένα σύνολο μη γραμμικά διαχωρίσιμων σημείων. Ωστόσο, ο Amaldi έδειξε το 1991 ότι η εύρεση του μεγαλύτερου γραμμικά διαχωρίσιμου υποσυνόλου σε ένα σύνολο που δεν είναι διαχωρίσιμο είναι ένα πλήρες πρόβλημα NP (Non-Polyonimal), δηλαδή δεν υπάρχει αλγόριθμος που να εκτελεί αυτήν την εργασία σε πολυωνυμικό χρόνο (Vermet, 2018).

5.3.3. Multilayer Perceptron ως εργαλείο προσέγγισης συναρτήσεων

Ας εισαγάγουμε τον μαθηματικό φορμαλισμό των δομημένων στρωμάτων νευρωνικών δικτύων, τα οποία είναι μοντέλα που χρησιμοποιούνται πολύ στην πράξη σε προβλήματα προσέγγισης νευρωνικών δικτύων (Popescu et al., 2009).

Έστω $p, m, q \in \mathbb{N}^*$, $\varphi_c, \varphi_\varsigma$ συναρτήσεις από το $\mathbb{R}$ στο $\mathbb{R}$, $W^1 = \left(W_{ij}^1, i = 1, \dots, p+1, j = 1, \dots, m\right)$ και $\left(W_{jk}^2, j = 1, \dots, m+1, k = 1, \dots, q\right)$ πίνακες με πραγματικούς.

Η συνάρτηση:

$$P\left(W^1, W^2, \varphi_c, \varphi_\varsigma\right): \mathbb{R}^p \to \mathbb{R}^q$$

$$x \mapsto s = \left(s_k = \varphi_\varsigma\left(\sum_{j=1}^{m} W_{jk}^2 \, \varphi_c\left(\sum_{i=1}^{p+1} W_{ij}^1 \, x_i\right) - W_{m+1,k}^2\right), k = 1, \dots, q\right),$$

ονομάζεται Perceptron με κρυφά στρώματα συναπτικά βάρη $\left(W_{ij}^1\right)$ $και$ $\left(W_{jk}^2\right)$ των συναρτήσεων ενεργοποίησης φ_c $και$ φ_ς με τη σύμβαση

$x_{p+1} = -1$.

Οι λειτουργίες ενεργοποίησης είναι οι ίδιες για όλες τις μονάδες του ίδιου επιπέδου, δηλαδή, φ_c για αυτές του κρυφού στρώματος και φ_ς για εκείνες της εξόδου. Στην πράξη, αυτές οι συναρτήσεις είναι γενικά η συνάρτηση βήματος Heaviside, γραμμικές συναρτήσεις ή σιγμοειδείς συναρτήσεις. Ωστόσο, μπορούμε να φανταστούμε και άλλες επιλογές όπως τριγωνομετρικές συναρτήσεις (Haykin, 1998).

Για να απλοποιήσουμε την παρούσα ενότητα, έχουμε περιοριστεί στον ορισμό ενός δικτύου με ένα κρυφό επίπεδο. Αυτή η κατασκευή γενικεύεται εύκολα στην περίπτωση δικτύων με πολλά κρυφά επίπεδα. Αυτή η ιδέα βρίσκεται στην αρχή των μεθόδων βαθιάς μάθησης (LeCun, Hinton, Bengio) (Tappert, 2019).

Ήδη από το 1969, οι Minsky and Papert έδειξαν ότι ένα πολυστρωματικό δίκτυο θα μπορούσε να ξεπεράσει τους περιορισμούς του απλού Perceptron. Ωστόσο, χρησιμοποιήθηκαν στην πραγματικότητα οι μέθοδοι συνδεσιμότητας μόνο σε λίγα έργα, συμπεριλαμβανομένου του περίφημου αλγόριθμου εκμάθησης της «πίσω διάδοσης» (backpropagation) (Wythoff, 1993). Αυτός ο αλγόριθμος είναι στην πραγματικότητα μια γενίκευση ενός αλγορίθμου που προτάθηκε από τους Widrow-Hoff (1960) και η μέθοδος κλίσης που χρησιμοποιήθηκε είναι μια σχετικά κλασική μέθοδος για την επίλυση προβλημάτων μεταβλητής στον βέλτιστο έλεγχο (Rumelhart, Hinton and Williams, 1986). Αυτός ο αλγόριθμος, ο οποίος μετατρέπει το Perceptron από την κατάσταση ταξινομητή (προσεγγιστής των συναρτήσεων $\mathbb{R}^1$ στο {0,1}) σε αυτόν του καθολικού προσεγγιστή, θα δικαιολογηθεί μαθηματικά από, δείχνοντας, για παράδειγμα, ότι οποιαδήποτε συνάρτηση του $\mathbb{R}^1$ στο $\mathbb{R}^n$ που έχει έναν πεπερασμένο αριθμό ασυνεχειών μπορεί να προσεγγιστεί με αυθαίρετη ακρίβεια από ορισμένα δίκτυα με ένα κρυφό στρώμα (Hegazy, Fazio and Moselhi, 1994). Ας θέσουμε ένα από αυτά τα θεμελιώδη θεωρήματα στην περίπτωση συναρτήσεων με τιμές στο $\mathbb{R}$ για μια συγκεκριμένη κατηγορία δικτύων.

ΘΕΩΡΗΜΑ (Hornik, 1993): *Έστω* $I \subset \mathbb{R}$ *ένα μη κενό διάστημα. Έστω* U *στο* $\mathbb{R}^p$ *και* $\varphi: \mathbb{R} \longrightarrow \mathbb{R}$ *οριοθετημένο Borel, μη πολυωνυμικό στο I.*

Έστω $P(\varphi, I, U)$ *το σύνολο των Perceptrons με ένα κρυφό στρώμα από το* $\mathbb{R}^p$ *έως* $\mathbb{R}$ *με τη μορφή:*

$$P(\theta, W^1, W^2)(x) = \sum_{j=1}^{m} W_j^2 \varphi\left(\sum_{i=1}^{p} W_{ij}^1 x_i - \theta_j\right)$$

όπου

$m \in \mathbb{N}^*$, $W^2 \in \mathbb{R}^m$, $W^1 \in U^m$ *και* $\theta \in I^m$. *Έστω K είναι ένα συμπαγές σύνολο στο* $\mathbb{R}^p$.

Επομένως:

i) Για κάθε πεπερασμένο μέτρο μ στο K, το $P(\varphi, I, U)$ *είναι πυκνό σε όλο το* $L^r(\mu), r \in [1, +\infty)$ *: για κάθε συνάρτηση* $f \in L^r(\mu)$ *υπάρχει μια ακολουθία* $(f_k) \in P(\varphi, I, U)$ *τέτοια ώστε για κάθε ε>0, υπάρχει* $N \in \mathbb{N}$, *έτσι ώστε για κάθε* $k \geq N$,

$$\|f - f_k\|_r = \left(\int_K |f(x) - f_k(x)|^r d\mu(x)\right)^{1/r} < \varepsilon$$

ii) Από τη στιγμή που η φ είναι dx-σχεδόν σίγουρα συνεχής, το $P(\varphi, I, U)$ *είναι πυκνό στο σύνολο C(K) των συνεχών συναρτήσεων στο K που παρέχεται με την νόρμα απείρου* $\|g\|_\infty = sup\{|g(x)|, x \in K\}$.

Είναι ένα θεώρημα ύπαρξης, που αποδεικνύεται με μεθόδους συνέλιξης. Αυτό το θεωρητικό αποτέλεσμα είναι σημαντικό, αλλά δεν παρέχει πληροφορίες σχετικά με το μέγεθος του κρυφού στρώματος (αριθμός m μονάδων) που πρέπει να επιλεγεί για να προσεγγίσει μια δεδομένη συνάρτηση με την επιθυμητή ακρίβεια (Kofidis et al., 2006).

5.4. Μηχανές Διανυσμάτων Στήριξης (Support vector machines - SVM)

Οι μηχανές διανυσμάτων στήριξης (SVM) αποτελούν μια κατηγορία εποπτευόμενων αλγορίθμων μάθησης που αρχικά ορίστηκαν για τη διάκριση μεταξύ δύο κλάσεων. Στη συνέχεια γενικεύτηκαν για προβλήματα πολλαπλών τάξεων και για την πρόβλεψη ποσοτικών μεταβλητών. Στην περίπτωση διάκρισης μεταξύ δύο κλάσεων, βασίζονται στην αναζήτηση ενός βέλτιστου υπερεπίπεδου περιθωρίου, διαχωρίζοντας όσο το δυνατόν καλύτερα τα δύο υποσύνολα παρατηρήσεων. Εδώ, βρίσκουμε την αρχή του

απλού Perceptron, με την ιδέα να αναζητήσουμε μια βέλτιστη λύση, την καλύτερη για γενίκευση (Somvanshi et al., 2016).

Η προσέγγιση αυτής της μεθόδου προέρχεται από την εργασία του Vapnik στη θεωρία της στατιστικής μάθησης από το 1995 σχετικά με τους δεσμούς μεταξύ της πολυπλοκότητας ενός μοντέλου και των δυνατοτήτων γενίκευσής του (Vapnik, Guyon and Hastie, 1995).

5.4.1. Γραμμικός διαχωριστής

5.4.1.1. Υπερεπίπεδο διαχωριστή μέγιστου περιθωρίου

Όταν οι παρατηρήσεις είναι γραμμικά διαχωρίσιμες, έχουμε δει ότι το απλό Perceptron του Rosenblatt και ο σχετικός αλγόριθμος μάθησης καθιστούν δυνατό τον προσδιορισμό της εξίσωσης ενός διαχωριστικού υπερεπίπεδου (Bhavsar and Panchal, 2012). Θα επεξεργαστούμε αυτήν την έννοια αναζητώντας μια βέλτιστη λύση (Stitson et al., 1996) και θα δούμε ότι αυτή η έννοια μπορεί να επεκταθεί στη μη γραμμική περίπτωση.

Ας εξετάσουμε ένα μαθησιακό σύνολο $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}, i = 1, \dots, n\}$, όπου το y_i υποδηλώνει σε ποια από τις δύο κλάσεις ανήκει το άτομο που χαρακτηρίζεται από τις μεταβλητές $x_i = (x_{i,1}, \dots, x_{i,p})$.

Παρέχουμε στο $\mathbb{R}^p$ ένα βαθμωτό γινόμενο, που υποδηλώνεται με «.». Ένα υπερεπίπεδο H του $\mathbb{R}^p$ ορίζεται από την εξίσωσή του: $w.x + b = 0$, όπου w είναι ένα διάνυσμα ορθογώνιο στο H.

Το σύνολο των στοιχείων του L_n λέμε ότι είναι γραμμικά διαχωρίσιμο εάν υπάρχει ένα διάνυσμα $w \in \mathbb{R}^p$, με ||w||=1 και μια σταθερά $b \in \mathbb{R}$, έτσι ώστε οι παρακάτω ανισότητες να επαληθεύονται για i = 1, … , n:

$$w.x_i + b \geq 1, \ \ \alpha\nu \ \ y_i = 1, \quad \kappa\alpha\iota \ \ w.x_i + b \leq -1, \ \ \alpha\nu \ \ y_i = -1$$

Αυτό είναι ισοδύναμο με $y_i(w.x_i + b) \geq 1$, για i=1, …, n. Σε αυτή την περίπτωση, το πρόσημο του $f(x) = w.x + b$ δηλώνει σε ποια πλευρά του υπερεπίπεδου βρίσκεται το σημείο x και επίσης την κλάση που σχετίζεται με το x.

Η επιλογή της σταθεράς 1 είναι αυθαίρετη αφού ένα υπερεπίπεδο ορίζεται μέσα σε μια πολλαπλασιαστική σταθερά. Υπάρχει άπειρος αριθμός λύσεων: ψάχνουμε αυτή του μέγιστου περιθωρίου, δηλαδή το υπερεπίπεδο, το οποίο απέχει όσο το δυνατόν περισσότερο από όλα τα παραδείγματα. Περιθώριο ονομάζουμε την απόσταση μεταξύ των δύο αντίστοιχων εξισώσεων υπερεπίπεδων $w.x + b = -1$ και $w.x + b = 1$.

Συγχέουμε ένα διάνυσμα w και το σημείο συντεταγμένων w χωρίς τον κίνδυνο σύγχυσης. Η με πρόσημο απόσταση από ένα σημείο x έως ένα υπερεπίπεδο Η της εξίσωσης $w.x + b = 0$ είναι είναι $d(x, H) = (x - w_1).w_0$, όπου w_1 είναι το μοναδικό σημείο του Η τέτοιο ώστε το διάνυσμα w_1 να είναι συγγραμμικό με το διάνυσμα w και $w_0 = w/\|w\|$. Εφόσον το σημείο w_1 ανήκει στο Η, έχουμε $w_1.w_0 = -b/\|w\|$, όπου:

$$d(x, H) = \frac{w.x+b}{\|w\|} = f(x)/\|w\|$$

Εφόσον η συνάρτηση f είναι ίση με -1 και +1 στα υπερεπίπεδα που σχηματίζουν τις ακμές, συμπεραίνουμε ότι το περιθώριο που σχετίζεται με το υπερεπίπεδο Η είναι ίσο με $2/\|w\|$ Για να λάβουμε το μέγιστο περιθώριο, πρέπει επομένως να ελαχιστοποιήσουμε τη συνάρτηση $\Phi(w, b) = \|w\|$ (ή με ισοδύναμο τρόπο $\|w\|^2$) υπό τους περιορισμούς $y_i(w.x_i + b) \geq 1 \ \ \gamma\iota\alpha \ \ i = 1, \dots, n$.

Για αυτό, χρησιμοποιούμε την τυπική μέθοδο πολλαπλασιαστή Lagrange. Έτσι, το λαγκρατζιανό:

$$L(w, b, a) = \frac{1}{2} w.w - \sum_{i=1}^{n} \alpha_i (y_i (w.x_i + b) - 1)$$

Όπου α είναι το διάνυσμα των πολλαπλασιαστών Lagrange $\alpha_i \geq 0, i = 1, \dots, n$. Η λύση στο πρόβλημά μας καθορίζεται από το σημείο σέλας (σαγματικό σημείο) (w^*, b^*, a^*) του λαγκρατζιανού στον $n + p + 1$ διαστατικό χώρο που αντιστοιχεί στις μεταβλητές w, α και b, λαμβάνοντας το ελάχιστο σε σχέση με τις p συντεταγμένες του w και b και το μέγιστο σε σχέση με το $\alpha_i, i = 1, \dots, n.$. Στο σημείο πραγματοποίησης του ελάχιστου (στο w και b), έχουμε:

$$\left\{\frac{\partial L(w,b,a)}{\partial w}\right\}_{w=w^*} = w^* - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

$$\left\{\frac{\partial L(w,b,a)}{\partial b}\right\}_{b=b^*} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

Αντικαθιστώντας αυτές τις συνθήκες στο λαγκρατζιανό, λαμβάνουμε:

$$L(w^*, b^*, a) = \sum_{i=1}^{n} \alpha_\iota - \frac{1}{2} w^*.w^* = \alpha.u - \frac{1}{2} \alpha D \alpha^T$$

Όπου $u = (1,1, \dots, 1) \in \mathbb{R}^n$, α^Tείναι το ανάστροφο του α, (δηλαδή, το διάνυσμα στήλη) και D είναι ο συμμετρικός πίνακας: $D = (D_{ij} = y_i y_j x_i.x_j)_{i,j=1,\dots,n}$. Έτσι, για να βρείτε το σημείο σέλας, είναι απαραίτητο να προσδιοριστεί το μέγιστο του α στην:

$$\tilde{L}(a) = L(w^*, b^*, a) = \alpha.u - \frac{1}{2} \alpha D \alpha^T$$

υπό τους περιορισμούς $\alpha.y = 0$ και $\alpha \geq 0$, όπου $y = (y_1, \dots, y_n)$. Σύμφωνα με το θεώρημα βελτιστοποίησης Kuhn–Tucker, στο σημείο της σέλας (w^*, b^*, a), κάθε πολλαπλασιαστής Lagrange α_i^* και ο αντίστοιχος περιορισμός συνδέονται με την εξίσωση:

$\alpha_i^*[y_i(x_i.w^* + b^*) - 1] = 0,\ i = 1, \dots, n.$

Με άλλα λόγια, $\alpha_i^* \neq 0$ μόνο για τα παραδείγματα του L_n έτσι ώστε η ανισότητα $y_i(w.x_i + b) \geq 1$ είναι ισότητα. Αυτά τα σημεία, x_i, για τα οποία $y_i(w.x_i + b) = 1$, ονομάζονται «διανύσματα στήριξης», και η παραπάνω συνθήκη $\frac{\partial L}{\partial w} = 0$ δείχνει ότι η λύση w^*γράφεται ως γραμμικός συνδυασμός των διανυσμάτων στήριξης:

$w^* = \sum_{i:\alpha_i^* \neq 0} \alpha_i^* y_i x_i$

Η επίλυση αυτού του προβλήματος βελτιστοποίησης μέσου τετραγώνου παρέχει στη συνέχεια την εξίσωση του βέλτιστου υπερεπίπεδου $w^*.x + b^* = 0$ με $b^* = -\frac{1}{2}[w^*.x^+ + w^*.x^-]$, όπου x^+ και x^- είναι διανύσματα στήριξης κάθε κλάσης.

Στη συνέχεια, είναι δυνατό να χρησιμοποιηθεί το διαχωριστικό υπερεπίπεδο που λαμβάνεται για να γίνει η πρόβλεψη: για μια νέα παρατήρηση x που παρουσιάζεται στο μοντέλο, το πρόσημο της συνάρτησης $f(x) = w^*.x + b^*$ δίνει την κλάση για να της αποδωθεί.

5.4.1.2. Η μη γραμμικά διαχωρίσιμη περίπτωση

Όταν οι παρατηρήσεις δεν μπορούν να διαχωριστούν από ένα υπερεπίπεδο, είναι απαραίτητο να χαλαρώσουν οι περιορισμοί εισάγοντας ξ_i όρους σφάλματος που ελέγχουν την υπέρβασή τους (Stitson et al., 1996):

$y_i(w.x_i + b) \geq 1 - \xi_i, i = 1, \dots, n$

Έτσι, το μοντέλο εκχωρεί μια ψευδή απόκριση σε ένα διάνυσμα x_i αν το αντίστοιχο ξ_i είναι μεγαλύτερο από 1. Στη συνέχεια, επανερχόμαστε στο πρόβλημα της ελαχιστοποίησης της συνάρτησης:

$\Psi(w, b, \xi) = \frac{1}{2} w.w - \delta \sum_{i=1}^{n} \xi_i$

υπό τους περιορισμούς $y_i(w.x_i + b) \geq 1 - \xi_i$ και $\xi_i \geq 0$ για $i = 1, \ldots, n$, όπου δ είναι μια αυστηρά θετική παράμετρος. Η μέθοδος που χρησιμοποιεί τους πολλαπλασιαστές Lagrange εξακολουθεί να λειτουργεί σε αυτήν την περίπτωση (Cortes and Vapnik, 1995).

Θα πρέπει να σημειωθεί ότι υπάρχουν πολλοί αλγόριθμοι για την επίλυση των προβλημάτων βελτιστοποίησης του μέσου τετραγώνου, σε περίπτωση διαχωρισμού ή όχι. Ορισμένοι, προτείνοντας μια αποσύνθεση του συνόλου μάθησης, προσαρμόζονται πιο συγκεκριμένα ώστε να λαμβάνουν υπόψη έναν σημαντικό αριθμό περιορισμών όταν ο αριθμός των παρατηρήσεων n είναι μεγάλος (Stitson et al., 1996).

5.4.2. Μη γραμμικός διαχωριστής

Όταν οι παρατηρήσεις δεν είναι γραμμικά διαχωρίσιμες, μερικές φορές είναι δυνατό να βρεθεί ένα μη γραμμικό όριο. Ένας έξυπνος τρόπος για να προσεγγίσουμε αυτό το πρόβλημα είναι να παρατηρήσουμε ότι είναι συχνά δυνατό να γίνουν αυτές οι παρατηρήσεις γραμμικά διαχωρίσιμες εξετάζοντας τις εικόνες τους με μια μη γραμμική εφαρμογή K σε ένα χώρο G με διάσταση μεγαλύτερη από αυτή του αρχικού χώρου $F=\mathbb{R}^p$ (Corlosquet-Habart and Janssen, 2018).

Επιπλέον, δεδομένου ότι η μέθοδος που χρησιμοποιείται στη γραμμικά διαχωρίσιμη περίπτωση θα περιλαμβάνει μόνο κλιμακωτά προϊόντα y.y' για τα διανύσματα y και y' με τη μορφή y=K(x) και y'=K(x'), δεν είναι απαραίτητο να εξηγήσουμε τη συνάρτηση K, αν γνωρίζουμε μια συμμετρική συνάρτηση $k: F \times F \to \mathbb{R}$, που ονομάζεται kernel, τέτοια ώστε:

$k(x, x') = K(x).K(x')$

Η συνθήκη του Mercer διασφαλίζει ότι μια συμμετρική συνάρτηση k(.,.) είναι kernel αν για όλα τα δυνατά x_i, η μήτρα γενικών όρων k(xi, yi) είναι ένας θετικός ορισμένος πίνακας, δηλαδή ορίζει έναν πίνακα βαθμωτού γινομένου. Σε αυτήν την περίπτωση, δείχνουμε ότι υπάρχει ένας χώρος G και μια συνάρτηση K : F⟶ G τέτοια ώστε: $k(x, x') = K(x).K(x')$. Παραδείγματα πυρήνων (kernels) που χρησιμοποιούνται συνήθως περιλαμβάνουν:

- Πυρήνας πολυωνυμικός d βαθμού: $k(x, x') = (cx.x')^d$, με $c \in \mathbb{R}$
- υπερβολικός εφαπτομενικός πυρήνας: $k(x, x') = tanh(c_1 x.x' + c_2)$, με $c_1, c_2 \in \mathbb{R}$
- Gaussian (ή ακτινωτός) πυρήνας: $k(x, x') = exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, $\sigma^2 > 0$.

5.5. Μέθοδοι συνάθροισης μοντέλων

Στις προηγούμενες ενότητες, παρουσιάσαμε διαφορετικές μεθόδους εποπτευόμενης μάθησης, οι οποίες προσφέρουν μια πληθώρα πιθανών μοντέλων, αλλά συχνά είναι δύσκολο να γνωρίζουμε ποιο να διατηρήσουμε. Αντί να επιλέγουμε μόνο ένα από αυτά τα μοντέλα, η γενική αρχή των μεθόδων συνάθροισης (ή μεθόδων συνόλου) είναι η κατασκευή μιας συλλογής προγνωστικών και στη συνέχεια η συγκέντρωση όλων των προβλέψεών τους. Στην παλινδρόμηση, η τελική πρόβλεψη μπορεί να είναι ένας μέσος όρος των προβλέψεων που δίνονται από τα διαφορετικά μοντέλα, ενώ στο πλαίσιο της ταξινόμησης, μπορούμε να προχωρήσουμε με πλειοψηφία μεταξύ των κατηγοριών που παρέχονται από τους διάφορους προγνωστικούς παράγοντες. Αυτή η τεχνική είναι ενδιαφέρουσα μόνο εάν τα διαφορετικά βασικά μοντέλα που επιλέχθηκαν δίνουν σημαντικά διαφορετικά αποτελέσματα, όπως συμβαίνει για τις μη γραμμικές μεθόδους

όπως τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα για παράδειγμα, που έχουν λιγότερη προκατάληψη από τις γραμμικές μεθόδους σε σύνθετα προβλήματα, αλλά μεγαλύτερη διακύμανση. Η συνάθροιση διαφορετικών μοντέλων καθιστά στη συνέχεια δυνατή τη μείωση της διακύμανσης και συνεπώς της αστάθειας των βασικών μοντέλων. Τα βασικά μοντέλα μπορεί να είναι διαφορετικών φύσεων, αλλά οι μέθοδοι που χρησιμοποιούν δέντρα απόφασης είναι οι πιο ανεπτυγμένες, συμπεριλαμβανομένων ειδικότερα των «τυχαίων δασών» (random forests) (Corlosquet-Habart and Janssen, 2018).

Σε αυτή την υποενότητα θα παρουσιάσουμε τις συνήθεις μεθόδους συνάθροισης, εστιάζοντας σε δύο κατηγορίες:

- παράλληλες μέθοδοι που συγκεντρώνουν βασικά μοντέλα κατασκευασμένα ανεξάρτητα το ένα από το άλλο, όπως bagging και τυχαία δάση (random forests)
- προσαρμοστικές μέθοδοι όπου κάθε βασικό μοντέλο που περιλαμβάνεται στο μοντέλο εξαρτάται από το προηγούμενο και κατασκευάζεται σύμφωνα με την απόδοση του τελευταίου στο εκπαιδευτικό σύνολο. Αυτές οι μέθοδοι είναι γνωστές ως boosting.

Αυτές οι διαφορετικές μέθοδοι χρησιμοποιούνται στην αναλογιστική επιστήμη, ιδιαίτερα στην ασφάλιση περιουσίας και ατυχημάτων, για την πρόβλεψη μεμονωμένων ποσών αξιώσεων, τον εντοπισμό απάτης ή την ανάπτυξη ζωνών (Corlosquet-Habart and Janssen, 2018).

5.5.1. Bagging

Η μέθοδος bagging εισήχθη από τον Breiman (1996a). Η λέξη bagging είναι μια ένωση των λέξεων bootstrap και aggregating. Δεδομένου ενός δείγματος μάθησης L_n και μιας

βασικής μεθόδου, η αρχή του bagging είναι ο σχεδιασμός ανεξάρτητων δειγμάτων εκκίνησης $L_{n,1}, \dots, L_{n,B}$ στο L_n για τη βαθμονόμηση της βασικής μεθόδου σε καθένα από αυτά για την κατασκευή B προγνωστικών $\hat{\varphi}_1(.), \dots, \hat{\varphi}_B(.)$. Ο τελικός εκτιμητής είναι τότε:

- $\hat{\varphi}(.) = \frac{1}{B}\sum_{\beta=1}^{B} \hat{\varphi}_b(.)$, στην παλινδρόμηση
- $\hat{\varphi}(.) = argmax_j card\{b: \hat{\varphi}_b(.) = j\}$, στην ταξινόμηση.

Στην πρώτη περίπτωση, είναι απλώς η μέση τιμή των διαφορετικών εκτιμήσεων και στη δεύτερη περίπτωση, η πλειοψηφία για τις *B* προβλέψεις.

Δύο τεχνικές χρησιμοποιούνται γενικά για τη δημιουργία δειγμάτων bootstrap:

- τα δείγματα $(L_{n,b})$ λαμβάνονται με σχεδίαση n παρατηρήσεων με αντικατάσταση στο L_n κάθε παρατήρησης που σχεδιάζεται με πιθανότητα 1/n.
- τα δείγματα $(L_{n,b})$ λαμβάνονται με τη σχεδίαση $\mathtt{l}$ παρατηρήσεων (με ή χωρίς αντικατάσταση) στο L_n , με $\mathtt{l}$<n

Το bagging καθιστά δυνατό να γίνουν πιο αποτελεσματικές οι αδύναμες βασικές μέθοδοι. Αυτό αποδείχθηκε μαθηματικά σε ένα συγκεκριμένο παράδειγμα από τους Biau and Devroye [2010]: λαμβάνοντας τη μέθοδο $\mathtt{l}$-πλησιέστερου γείτονα ως βασική μέθοδο (η οποία δεν είναι καθολικά συνεπής) και δείγματα bootstrap με μέγεθος $\mathtt{l}_n$ όπως $\lim_{n\to\infty} l_n = +\infty$ και $\frac{\lim_{n\to\infty} l_n}{n} = 0$, απέδειξαν ότι ο ασυμπτωτικός εκτιμητής $\tilde{\varphi}(.) = \lim_{B\to\infty} \frac{1}{B}\hat{\varphi}_b(.)$ είναι καθολικά συνεπής.

Είναι δυνατή η χρήση μιας μεθόδου out-of-bag για την εκτίμηση του σφάλματος στην πρόβλεψη του μοντέλου που λήφθηκε σε ένα δείγμα δοκιμής: αυτό καθιστά δυνατό ιδίως τον έλεγχο του τρόπου με τον οποίο αυτό το σφάλμα εξελίσσεται ως συνάρτηση

του B, προκειμένου να επιλέγεται ένας επαρκής αριθμός βασικών μοντέλων (Breiman, 1996b). Στην περίπτωση του bagging, δεν είναι απαραίτητο να υπάρχει επιπλέον δείγμα ανεξάρτητο από το L_n. Στην πραγματικότητα, δεν περιέχουν όλα τα βασικά μοντέλα όλες τις παρατηρήσεις. Για να λάβουμε μια ανεξάρτητη από τη μάθηση πρόβλεψη για μια παρατήρηση (x_i, y_i) του L_n, περιοριζόμαστε επομένως στο σύνολο M των βασικών μοντέλων που δεν χρησιμοποιούν αυτή την παρατήρηση για βαθμονόμηση και η πρόβλεψη που σχετίζεται με το x_i είναι τότε:

$$\hat{y}_i = \frac{1}{card(M)} \sum_{b \in M} \hat{\varphi}_b(x_i)$$

Σε ένα πλαίσιο ταξινόμησης, προχωράμε με πλειοψηφία. Στη συνέχεια, το σφάλμα out-of-bag ορίζεται από:

- $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, στην παλινδρόμηση
- $\frac{1}{n} \sum_{i=1}^{n} I_{\hat{y}_i \neq y_i}$ στην ταξινόμηση

Για απλοποίηση, υποθέτοντας ότι όλοι οι βασικοί προγνωστικοί παράγοντες έχουν ως θεωρητική διακύμανση σ^2 και ότι δύο βασικοί προγνωστικοί παράγοντες που κατασκευάστηκαν με τη μέθοδο bootstrap έχουν ως συντελεστή συσχέτισης ρ, τότε είναι εύκολο να δούμε ότι ο προγνωστικός παράγοντας που λαμβάνεται από το bagging έχει ως διακύμανση $\rho\sigma^2 + \frac{(1-\rho)\sigma^2}{B}$. Έτσι, φαίνεται ότι η διακύμανση μειώνεται με τον αριθμό των βασικών μοντέλων που χρησιμοποιούνται και επίσης ότι η συσχέτιση μεταξύ των βασικών μοντέλων ποσοτικοποιεί το κέρδος της διαδικασίας συνάθροισης: όσο πιο αποσυσχετισμένα είναι τα βασικά μοντέλα (μικρό |ρ|), τόσο μικρότερη θα είναι η διακύμανση του τελικού μοντέλου. Αυτό αποτελεί αιτιολόγηση για τη λήψη βασικών μοντέλων ευαίσθητων στο μαθησιακό δείγμα, όπως δέντρα ή νευρωνικά δίκτυα:

διαφορετικά, τα βασικά μοντέλα θα συσχετίζονται ισχυρά και η άθροισή τους δεν θα οδηγήσει σε καμία βελτίωση.

5.5.2. Τυχαία δάση

Στην περίπτωση των μοντέλων CART, ο Breiman (2001) πρότεινε μια βελτίωση στη συσκευασία με την προσθήκη τυχαιοποίησης. Ο στόχος είναι να γίνουν τα βασικά μοντέλα λιγότερο συσχετισμένα προσθέτοντας τυχαιότητα στην επιλογή των μεταβλητών που χρησιμοποιούνται στα διάφορα βασικά μοντέλα. Είδαμε την τιμή στην προηγούμενη ενότητα: η χρήση λιγότερο συσχετισμένων βασικών μοντέλων βοηθά στη μείωση της διακύμανσης του προγνωστικού δείκτη που προκύπτει από τη συγκέντρωση. Για το σκοπό αυτό, ο Breiman (2001) πρότεινε να επιλεγούν τυχαία, σε κάθε στάδιο της κατασκευής ενός δέντρου, οι μεταβλητές μεταξύ των επεξηγηματικών μεταβλητών και να προσδιοριστεί η αντίστοιχη διαίρεση λαμβάνοντας υπόψη μόνο αυτές τις μεταβλητές. Σημειώστε ότι $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times G, i = 1, \dots, n\}$, το σύνολο εκμάθησης, όπου G = $\mathbb{R}$ στην περίπτωση παλινδρόμησης και G={1, … , K} σε αυτό της ταξινόμησης. Ο αλγόριθμος είναι ο εξής:

Αλγόριθμος τυχαίου δάσους:

1) Διάλεξε $B. m \in \mathbb{N}^*$.
2) Για b=1 έως B:

- σχεδιάστε ένα δείγμα Bootstrap $L_{n,b}$ στο L_n.
- με το δείγμα $L_{n,b}$ κατασκευάστε ένα δέντρο CART, το οποίο ορίζει έναν προγνωστικό παράγοντα $\hat{\varphi}_b(.)$. Κάθε διάσπαση του δέντρου προσδιορίζεται περιορίζοντας σε ένα σύνολο m μεταβλητών που έχουν επιλεγεί τυχαία μεταξύ των

p επεξηγηματικών μεταβλητών, με τις επιλογές να είναι ανεξάρτητες για τους διαφορετικούς διαχωρισμούς.

End For

3) Ορίστε $\hat{\varphi}(.) = \frac{1}{B}\sum_{b=1}^{B} \hat{\varphi}_b(.)$ (ή $\hat{\varphi}(.) = argmax_j card\{b: \hat{\varphi}_b(.) = j\}$).

Η αξία αυτής της προσέγγισης έχει αποδειχθεί όταν ο αριθμός των p επεξηγηματικών μεταβλητών είναι πολύ σημαντικός (Caruana, Karampatziakis and Yessenalina, 2008). Ο αριθμός των τυχαίων μεταβλητών m μπορεί να είναι μια ευαίσθητη παράμετρος. Οι κλασικές επιλογές είναι $m = \sqrt{p}$ στην ταξινόμηση και $m = p/3$ στην παλινδρόμηση. Αυτή είναι η προεπιλεγμένη επιλογή στη βιβλιοθήκη randomforest στην R. Όπως και με το bagging, η αξιολόγηση σφαλμάτων out-of-bag χρησιμοποιείται για τον έλεγχο της τιμής του B, και πιθανώς για τη βελτιστοποίηση της τιμής του m.

Από τη δημοσίευσή της το 2001, αυτή η μέθοδος έχει μελετηθεί και χρησιμοποιηθεί εκτενώς, λόγω των πολύ καλών προγνωστικών της ιδιοτήτων (βλ., για παράδειγμα, το συγκριτικό άρθρο των Fernandez-Delgado et al. (2014) και το άρθρο των Biau and Scornet (2016)). Από θεωρητικής σκοπιάς, τα αποτελέσματα σύγκλισης, τα οποία είναι δύσκολο να ληφθούν, επιδείχθηκαν το 2015 από τους Scornet, Biau and Vert (2015). Θα πρέπει να σημειωθεί επίσης το πιο πρόσφατο άρθρο των Biau, Scornet and Welbl (2016), το οποίο δημιουργεί μια ενδιαφέρουσα σύνδεση μεταξύ τυχαίων δασών και νευρωνικών δικτύων.

5.5.3. Boosting

Η μέθοδος boosting εισήχθη το 1996 από τους Freund and Schapire (1996) και έχει ένα κοινό χαρακτηριστικό με το bagging στο ότι χρησιμοποιεί βασικά μοντέλα για να συναγάγει έναν τελικό προγνωστικό παράγοντα με τη συσσωμάτωση. Ωστόσο, η

ουσιαστική διαφορά είναι ότι το boosting κατασκευάζει τα βασικά μοντέλα επαναληπτικά και προσαρμοστικά, προκειμένου να διορθωθούν σταδιακά τα λάθη που έγιναν. Η αρχική ιδέα ήταν να βελτιωθεί η απόδοση των μοντέλων που είχαν απόδοση ελάχιστα καλύτερη από έναν τυχαίο ταξινομητή για την πρόβλεψη μιας δυαδικής μεταβλητής. Αυτό είχε ως αποτέλεσμα τον αλγόριθμο AdaBoost (Freund and Schapire, 1996), ο οποίος είναι ο πιο δημοφιλής αλγόριθμος boosting (Corlosquet-Habart and Janssen, 2018). Έστω $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \{-1,1\}, i = 1, \dots, n\}$ σετ εκμάθησης. Ο αλγόριθμος λοιπόν έχει ως εξής:

Ο αλγόριθμος AdaBoost:

1) Διάλεξε $M \in \mathbb{N}^*$
2) Αρχικοποιήστε τα βάρη: $w = (w_i = \frac{1}{n}, i = 1, \dots, n)$
3) Για m=1 έως M:

- προσαρμόστε ένα βασικό μοντέλο $\hat{\varphi}_m(.)$ στα σταθμισμένα L_n με τα βάρη $w_1, \dots, w_n$
- υπολογίστε το εμφανές ποσοστό σφάλματος στο L_n για το μοντέλο $\hat{\varphi}_m(.)$:

$$\varepsilon_m = \frac{\sum_{i=1}^{n} w_i I_{\{y_i \neq \hat{\varphi}_m(x_i)\}}}{\sum_{i=1}^{n} w_i}$$

- υπολογίστε: $\alpha_m = \log\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$
- τροποποιήστε τα βάρη: $w_i = w_i \exp\left(\alpha_m I_{\{y_i \neq \hat{\varphi}_m(x_i)\}}\right), i = 1, \dots, n.$

End For

4) Ορίστε: $\hat{\varphi}(.) = sign(\sum_{m=1}^{M} a_m \hat{\varphi}_m(.))$.

Όλες οι παρατηρήσεις σταθμίζονται εξίσου για το πρώτο μοντέλο. Τότε, σε κάθε επανάληψη, το βάρος μιας παρατήρησης παραμένει αμετάβλητο εάν είναι καλά

ταξινομημένη και αυξάνεται διαφορετικά. Ο τελικός προγνωστικός παράγοντας είναι μια σταθμισμένη συνάθροιση των βασικών μοντέλων, με τη στάθμιση να είναι συνάρτηση των ιδιοτήτων προσαρμογής κάθε μοντέλου. Τα βασικά μοντέλα μπορεί να είναι αδύναμα, αλλά θα πρέπει να διασφαλιστεί στην πράξη ότι είναι καλύτερα από έναν τυχαίο προγνωστικό παράγοντα (μία στις δύο πιθανότητες να γίνει λάθος): πρέπει να έχουμε $\varepsilon_m > 0{,}5$, διαφορετικά το α_m γίνεται αρνητικό.

Το βασικό βήμα βαθμονόμησης του μοντέλου με τα βάρη $(w_1, \dots, w_n)$ προσφέρει πολλές δυνατότητες. Για παράδειγμα, τα βάρη w_i μπορούν να χρησιμοποιηθούν για τη στάθμιση του εμπειρικού σφάλματος που πρέπει να ελαχιστοποιηθεί:

$$\frac{1}{n}\sum_{i=1}^{n} w_i I_{\{y_i \neq \hat{\varphi}_m(x_i)\}}$$

Το βασικό μοντέλο μπορεί επίσης να προσαρμοστεί σε ένα υποδείγμα μεγέθους n του L_n, σχεδιάζοντας τυχαία παρατηρήσεις (με αντικατάσταση) σύμφωνα με τα βάρη $(w_1, \dots, w_n)$.

Ο αλγόριθμος AdaBoost έχει χρησιμοποιηθεί ευρέως με τα δέντρα CART ως βασικά μοντέλα. Έχει επίσης μελετηθεί εκτενώς από θεωρητική σκοπιά (βλ., για παράδειγμα, το άρθρο των Bartlett and Traskin (2006), το οποίο δείχνει ιδιότητες καθολικής συνέπειας).

Πολλαπλές παραλλαγές έχουν επίσης προταθεί για την προσαρμογή του στην περίπτωση μεταβλητών πολλαπλών κλάσεων ή ποσοτικών μεταβλητών εξόδου (βλ., για παράδειγμα, Schapire (2003)). Το 1997, ο Drucker πρότεινε μια έκδοση ενίσχυσης για παλινδρόμηση.

Έστω $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}$ σύνολο μάθησης. Ο αλγόριθμος του Drucker (1997) έχει ως εξής:

Αλγόριθμος boosting για παλινδρόμηση:

1) Διάλεξε $M \in \mathbb{N}^*$
2) Όρισε: $p_1 = (p_{1,i} = {}^1\!/_n, i = 1, \dots, n)$.
3) Για m=1 έως M:

- Σχεδιάστε ένα δείγμα $\bar{L}_n$ μεγέθους n με αντικατάσταση στο L_n που ακολουθεί το νόμο p_m
- Προσαρμόστε ένα βασικό μοντέλο $\hat{\varphi}_m(.)$ στο $\bar{L}_n$
- Υπολογίστε στο L_n για το μοντέλο $\hat{\varphi}_m(.)$
- $l_m(i) = Q(y_i, \hat{\varphi}_m(x_i))$ για i=1, ... , n, όπου Q είναι μια συνάρτηση απώλειας
- $\mathcal{L}_m = sup_{i=1,\dots,M}, l_m(i)$
- $d_m(i) = \frac{l_m(i)}{\mathcal{L}_m}$
- $\varepsilon_m = \sum_{j=1}^{n} p_{m,j} l_m(j)$
- $\beta_m = \frac{\varepsilon_m}{\varepsilon_m - \mathcal{L}_m}$
- $Αν\ \varepsilon_m < 0.5\mathcal{L}_m$, τότε όρισε $w_{m+1,i} = {\beta_m}^{1-d_m(i)} p_{m,i}$
- $Αλλιώς\ όρισε\ \ w_{m+1} = p_1$
- τροποποίησε τις πιθανότητες: $p_{m+1,i} = {}^{w_{m+1,i}}\!/_{\sum_{j=1}^{n} w_{m+1,j}}, i = 1, \dots, n$

End For

4) Υπολογίστε την $\hat{\varphi}(.)$, μέσος όρος ή διάμεσος των σταθμισμένων $\hat{\varphi}_m(.)$ προβλέψεων με τους συντελεστές $\left(\log\left({}^1\!/_{\beta_m}\right), m = 1, \dots, M\right)$.

Σε αυτόν τον αλγόριθμο, η συνάρτηση απώλειας Q μπορεί να είναι τετραγωνική, εκθετική ή η απόλυτη τιμή, η συνήθης επιλογή στην παλινδρόμηση είναι η τετραγωνική

συνάρτηση. Στο τελευταίο βήμα, η χρήση της διάμεσης τιμής εξαλείφει την επιρροή των υπερβολικά άτυπων προγνωστικών παραγόντων (Gey and Poggi, 2006).

Ο Breiman (1999) πρότεινε να εξεταστεί το boosting ως παγκόσμιος αλγόριθμος βελτιστοποίησης. Σύμφωνα με αυτή την ιδέα, οι Hastie et al. (2009) έδειξαν ότι η δυαδική περίπτωση μπορεί να θεωρηθεί ως μέθοδος προσέγγισης με ένα σταδιακά κατασκευασμένο προσθετικό μοντέλο, με τον αλγόριθμο να ελαχιστοποιεί μια συνάρτηση εκθετικής απώλειας που ορίζεται στη βάση εκμάθησης. Αυτή η αρχή αναπτύχθηκε από τον Friedman (2001) με το ακρωνύμιο MART (multiple additive regression trees) και στη συνέχεια με το ακρωνύμιο GBM (gradient boosting machine). Όσον αφορά τον αλγόριθμο AdaBoost, είναι ένα ζήτημα επαναληπτικής κατασκευής μιας σειράς μοντέλων έτσι ώστε τα διαδοχικά μοντέλα να βελτιώνονται και η τελική συγκέντρωση να παρέχει μια καλή λύση. Η συνεισφορά του GBM είναι ότι οι διαδοχικές τροποποιήσεις καθορίζονται με την εφαρμογή μιας μεθόδου βελτιστοποίησης κλίσης σε μια συνάρτηση απώλειας Q προς ελαχιστοποίηση (Friedman, 2001).

Αλγόριθμος GBM – Boosting με συναρτησιακή αρνητική κλίση:

1) Επιλέξτε $\lambda \in \mathbb{N}^*,\ \lambda \in [0,1]$.
2) Ορίστε $\hat{\varphi}_0(.) = argmin_c \frac{1}{n}\sum_{i=1}^{n} Q(y_i, c)$
3) Για m=1 έως M:

- υπολογίστε το αντίθετο της κλίσης $-\frac{\partial}{\partial z} Q(y, z)$ στα σημεία $z_i = \hat{\varphi}_{m-1}(x_i)$. Έστω οι τιμές αυτές $(u_1, \dots, u_n)$
- Προσαρμόστε ένα βασικό μοντέλο $\hat{\Psi}_m(.)$ στο $\{(x_1, u_1), \dots, (x_n, u_n)\}$
- Συνάγετε ένα νέο μοντέλο $\hat{\varphi}_m(.) = \hat{\varphi}_{m-1}(.) + \lambda\hat{\Psi}_m(.)$

End For

4) Η έξοδος είναι το μοντέλο $\hat{\varphi}_M(.)$

Η επιλογή της παραμέτρου λ είναι σχετικά ασήμαντη. Συχνά συνιστάται να το λαμβάνετε στην τάξη του 0,1. Αυτή η επιλογή σχετίζεται με αυτή του βέλτιστου αριθμού επαναλήψεων: μια μικρή τιμή λ θα απαιτήσει μεγάλο αριθμό επαναλήψεων και το αντίστροφο.

5.5.4. Stacking

Το stacking (ονομάζεται επίσης blending) είναι μια προσέγγιση που περιλαμβάνει τη χρήση ως εισόδου σε έναν αλγόριθμο στατιστικής εκμάθησης των τιμών εξόδου που δίνονται σε ένα πρώτο βήμα από διαφορετικά μοντέλα. Επομένως, πρόκειται για ένα μεταμοντέλο, το οποίο συγκεντρώνει βασικά μοντέλα με πιο περίπλοκο τρόπο από ό,τι η μέθοδος του bagging (Corlosquet-Habart and Janssen, 2018). Η αρχική ιδέα αναπτύχθηκε από τον Wolpert (1992), και αυτή η μέθοδος στη συνέχεια υιοθετήθηκε σε πολλές εφαρμογές. Μια ιδιαίτερη περίπτωση είναι όπου το μεταμοντέλο είναι ένα γραμμικό μοντέλο, ενώ τα βασικά μοντέλα είναι μη γραμμικά, όπως τα νευρωνικά δίκτυα ή τα δέντρα (Breiman, 1996a). Στην ταξινόμηση, ένα μοντέλο λογιστικής παλινδρόμησης μπορεί να χρησιμοποιηθεί ως μεταμοντέλο για να συνδυάσει τις προβλέψεις βασικών μοντέλων διαφορετικών τύπων (λογιστική παλινδρόμηση, δέντρο αποφάσεων, νευρωνικό δίκτυο κ.λπ.). Αυτή η μέθοδος είναι επομένως σχετικά παλιά, αλλά έχει αποδειχθεί πολύ αποτελεσματική στην επίλυση σύνθετων προβλημάτων (Corlosquet-Habart and Janssen, 2018).

5.6. Αλγόριθμος Kohonen ταξινόμησης χωρίς επίβλεψη

Η ανάλυση των πολυδιάστατων δεδομένων επικεντρώνεται συνήθως στη μελέτη των n ατόμων που καθορίζονται από μεταβλητές p (ποσοτικές ή ποιοτικές): ένα άτομο είναι ένα στοιχείο ενός χώρου p διαστάσεων. Οι παραγοντικές μέθοδοι (ACP και παράγωγα), που είναι στην πραγματικότητα μέθοδοι προβολής της γραμμικής άλγεβρας, επιτρέπουν γραφικές «αναπαραστάσεις» των δεδομένων στις διαστάσεις 1, 2 ή 3. Ωστόσο, είναι δύσκολο να κατασκευαστούν τάξεις εγγύτητας από προβολές εάν τα άτομα δεν αναπαρίστανται χωρίς απώλεια πληροφοριών σε ένα διαστατικό χώρο ≤ 3. Δύο άτομα των οποίων οι προβολές είναι κοντινές δεν είναι πάντα κοντά στον αρχικό p-διαστατικό χώρο (Corlosquet-Habart and Janssen, 2018).

Από την άλλη πλευρά, μπορούν να διακριθούν οι ακόλουθες δύο οικογένειες μεθόδων «ταξινόμησης» (Corlosquet-Habart and Janssen, 2018):

- ιεραρχική ταξινόμηση: επαναληπτική διαδικασία συγκέντρωσης των κλάσεων πλησιέστερου γείτονα. Αρχικά, κάθε άτομο είναι μια τάξη. Θεωρούμε τις n τάξεις και ομαδοποιούμε τις δύο πλησιέστερες τάξεις (για επιλεγμένη απόσταση) για να σχηματίσουμε μια νέα τάξη (δύο ατόμων). Και επαναλαμβάνουμε για τις n-1 κλάσεις που λήφθηκαν μέχρι να ληφθεί μια μεμονωμένη κλάση. Αυτό δημιουργεί μια αναπαράσταση από ένα δέντρο. Στη συνέχεια, μπορούμε να επιλέξουμε τον αριθμό των τάξεων που φαίνεται καλύτερα προσαρμοσμένο επιλέγοντας ένα επίπεδο ομαδοποίησης.
- μη ιεραρχική ταξινόμηση: ο αριθμός των κλάσεων καθορίζεται εκ των προτέρων και μια κλάση αποδίδεται σε κάθε άτομο με έναν αλγόριθμο που συγκλίνει προς μια κατανομή που ελαχιστοποιεί την ενδοταξική αδράνεια (δηλαδή τα αθροίσματα αποστάσεων κάθε ατόμου μιας κλάσης προς το κέντρο της τάξης). Ένα κλασικό παράδειγμα είναι ο αλγόριθμος του κινούμενου μέσου όρου.

Για αυτές τις μεθόδους ταξινόμησης, δύο σημεία της ίδιας κλάσης βρίσκονται κοντά στον αρχικό χώρο, αλλά πώς μπορούν οι κλάσεις να αναπαρασταθούν συνολικά διατηρώντας την αρχική τοπολογία των δεδομένων; Δεν υπάρχει η έννοια της γειτονιάς των τάξεων.

Ο αλγόριθμος Kohonen επιχειρεί να ταιριάξει την αναπαράσταση και την ταξινόμηση. Το 1982, ο Teuvo Kohonen πρότεινε έναν αλγόριθμο του οποίου η κύρια λειτουργία είναι να αντιστοιχίζει τα στοιχεία του χώρου εισόδων με μονάδες τεταγμένων σε έναν χάρτη δίνοντας μια γραφική αναπαράσταση (διαστάσεων 1, 2 ή 3), όπου κάθε μονάδα περιβάλλεται από τους γείτονές του (για προκαθορισμένη απόσταση). Το αποτέλεσμα είναι μια συνάρτηση του χώρου εισόδου στο σύνολο των μονάδων, έτσι ώστε οι εικόνες δύο γειτονικών στοιχείων με την έννοια μιας ορισμένης απόστασης στον χώρο εισόδου να είναι η ίδια μονάδα ή γειτονικές μονάδες στον χάρτη (Kohonen, 1991).

Ακόμα κι αν ο αλγόριθμος που προτείνει ο Kohonen μπορεί να θεωρηθεί μόνο ως αλγόριθμος ταξινόμησης χωρίς επίβλεψη, αρχικά σχετίζεται με τη νευρωνική μοντελοποίηση. Επινοήθηκε το 1982 ως μοντελοποίηση του αυτόματου σχηματισμού χαρτών στις αισθητήριες περιοχές του φλοιού, αν και έκτοτε η μαθηματική μελέτη του αλγορίθμου και οι εφαρμογές του ιδιαίτερα στην ανάλυση δεδομένων τον απομάκρυναν από το αρχικό βιολογικό του πλαίσιο. Για παράδειγμα, ας εξετάσουμε τις νευρικές συνδέσεις των κυττάρων του αμφιβληστροειδούς προς τον εγκεφαλικό φλοιό που επεξεργάζεται οπτικές πληροφορίες. Αν φανταστούμε τον αμφιβληστροειδή ως ένα πλέγμα δύο διαστάσεων, ένα σημαντικό γεγονός είναι ότι η τοπολογία του αμφιβληστροειδούς διατηρείται από το σύνολο των συνδέσεων με την έννοια ότι δύο στενά κύτταρα στον αμφιβληστροειδή συνδέονται με δύο στενά κύτταρα στον φλοιό. Δεν υπάρχει λόγος να συμβεί αυτό αυθόρμητα κατά τη γέννηση, και μπορούμε να υποθέσουμε ότι αυτό το φαινόμενο οφείλεται σε διαδικασία αυτοοργάνωσης και

επιλογής που διέπεται από τις αυθόρμητες δραστηριότητες των εγκεφαλικών νευρώνων. Ακριβώς αυτήν την ιδέα θέλησε να προσαρμόσει ο T. Kohonen όταν καθόρισε τον αλγόριθμο αυτοοργάνωσής του, προσθέτοντας μια σημαντική έννοια σε προηγούμενα συνδετικά μοντέλα που γνώριζε πολύ καλά. Από αλγοριθμική άποψη, η διαδικασία αυτοοργάνωσης πραγματοποιείται με τοπική ενημέρωση των συνδέσεων σύμφωνα με συμπληρωματικούς κανόνες ανταγωνισμού και συνεργασίας, σε κάθε παρουσίαση ενός πρωτοτύπου (Corlosquet-Habart and Janssen, 2018).

5.6.1. Σημειώσεις και ορισμός του μοντέλου

Σύμφωνα με τους Cottrell and Rousset (1997), ο χώρος δεδομένων είναι ένα κυρτό οριοθετημένο υποσύνολο $\chi \subset \mathbb{R}^p$, που παρέχεται με την Ευκλείδεια απόσταση. Θεωρούμε ένα δείγμα (x(1), …, x(t)) από διαδοχικές παρατηρήσεις στο χ (στατιστική προσέγγιση) ή πραγματοποιήσεις μιας ακολουθίας ανεξάρτητων τυχαίων μεταβλητών του ίδιου νόμου πιθανοτήτων μ με τιμές στο χ (πιθανολογική προσέγγιση).

Το δίκτυο αποτελείται από n μονάδες (ή νευρώνες) διατεταγμένες σύμφωνα με μια καθορισμένη τοπολογία:

- με d=1, 2 ή 3 διαστάσεις: μια γραμμή, ένα τετράγωνο ή ένας κύβος
- σύμφωνα με ένα πλέγμα του οποίου η γειτονική δομή καθορίζεται από μια συνάρτηση γειτνίασης. Οι μονάδες αντιπροσωπεύονται από ένα υποσύνολο I του $\mathbb{Z}^d$ και η συνάρτηση γειτονιάς είναι μια συνάρτηση Λ που ορίζεται στο $I \times I$, η οποία είναι:
    - συμμετρική (δηλαδή $\Lambda(i,j) = \Lambda(j,i)$
    - εξαρτώμενη μόνο από μια απόσταση D στο I
    - φθίνουσα ως προς την απόσταση: $\Lambda(i,j) \to 0 \quad \text{όταν} \quad D(i,j) \to \infty$.

Συχνά γίνεται αποδεικτό ότι $\Lambda(i,i) = 1$.

Όταν $\Lambda(i,j) = 1$, λέμε ότι τα i και j είναι στενά συνδεδεμένα και όταν $\Lambda(i,j) = 0$, τα i και j είναι πλήρως αποσυνδεδμένα και δεν έχουν αλληλεπιδράσεις.

Κάθε μονάδα i παρέχεται με ένα διάνυσμα κατάστασης $W_i(t) \in \mathbb{R}^p$ που δείχνει στον χώρο δεδομένων και μπορεί να τροποποιηθεί.. Η κατάσταση του δικτύου τη στιγμή t δίνεται από τα $(W_i(t), i \in I)$.

Ο στόχος είναι να βρεθούν διανύσματα που έχουν ιδιότητες:

- ποσοτικοποίηση: ο αριθμός των διανυσμάτων W_i σε μια περιοχή A του χ είναι περίπου ανάλογος με το μ(A).
- οργάνωση: δύο κοντινές μονάδες i και j (δηλαδή, $\Lambda(i,j) \approx 1$) έχουν κοντινά διανύσματα W_i και W_j.

Μόλις καθοριστούν τέτοια διανύσματα, χρησιμοποιούνται για να ορίσουν την κλάση οποιουδήποτε στοιχείου $x \in \chi$: αποδίδουμε στο x την κλάση i^* έτσι ώστε:

$$i^*(x, W) = argmin\{\|x - W_i\|, i \in I\}$$

Όπου $\|.\|$ είναι νόρμα στο $\mathbb{R}^p$. Ας μελετήσουμε τώρα τον αλγόριθμο Kohonen, ο οποίος καθορίζει τα διανύσματα W_i με μια μέθοδο εκμάθησης.

5.6.2. Αλγόριθμος Kohonen

Είναι ένας επαναληπτικός αλγόριθμος μάθησης που τροποποιεί τα διανύσματα κατάστασης W_i ανάλογα με τα δεδομένα που του παρουσιάζονται (Cottrell and Rousset, 1997):

- τα διανύσματα $W_i(t = 0)$ αρχικοποιούνται τυχαία.

- εάν τη στιγμή $t \in \mathbb{N}$, η κατάσταση του δικτύου δίνεται από την $W(t) = \{W_i(t), i \in I\}$, τότε η κατάσταση στη στιγμή t+1 θα καθοριστεί ως εξής: επιλέγεται τυχαία ένα διάνυσμα $x(t+1)$ του χώρου δεδομένων χ .

Στη συνέχεια, η «φάση του διαγωνισμού» ορίζει τη νικήτρια μονάδα:

$$i^*\big(x(t+1), W(t)\big) = argmin\{\|x(t+1) - W_i(t)\|, i \in I\}.$$

Στην περίπτωση που πολλές μονάδες ελαχιστοποιούν αυτή την απόσταση, συμφωνείται ένας κανόνας, για παράδειγμα, ο πρώτος δείκτης i για τη λεξικογραφική σειρά στο $\mathbb{Z}^d$.

Η «φάση συνεργασίας» τροποποιεί τα διανύσματα W_i:

$$\forall j \in I, W_j(t+1) = W_j(t) - \varepsilon_t \Lambda_t(i^*, j)\Big(W_j(t) - x(t+1)\Big),$$

όπου Λ_t είναι μια συνάρτηση γειτνίασης, η οποία μπορεί να εξαρτάται από το χρόνο και το $\varepsilon_t \in [0,1]$ είναι μια παράμετρος προσαρμογής.

Ο αλγόριθμος συνεχίζεται για όσο διάστημα το t είναι μικρότερο από μια τιμή κατωφλίου M που έχει καθοριστεί εκ των προτέρων ή μπορεί να επιβληθεί μια εναλλακτική ή πρόσθετη συνθήκη διακοπής, η οποία διακόπτει τον αλγόριθμο εάν δεν υπάρχει αξιοσημείωτη βελτίωση. Οι βασικές παράμετροι του αλγορίθμου είναι η διάσταση p του χώρου δεδομένων, ο νόμος πιθανότητας μ (που χαρακτηρίζει τον τρόπο με τον οποίο κατανέμονται οι παρατηρήσεις στο χώρο δεδομένων), η τοπολογία δικτύου, η συνάρτηση γειτνίασης Λ_t, σταθερή ή εξαρτώμενη από το χρόνο και η παράμετρος προσαρμογής ε_t, με τιμές στο [0,1], οι οποίες μπορεί να είναι σταθερές ή φθίνουσες στο t.

Η συνάρτηση γειτνίασης παίζει σημαντικό ρόλο εδώ: μόνο οι μονάδες που βρίσκονται κοντά στη νικήτρια μονάδα i^* έχουν το διάνυσμά τους W_i σημαντικά τροποποιημένο: εξ ου και η έννοια της συνεργασίας. Ακολουθούν μερικά κοινά παραδείγματα συνάρτησης γειτνίασης:

– $\Lambda_t(i,j) = \begin{cases} 1, & \alpha\nu\ D(i,j) \le k \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$ (όπου το k είναι μια σταθερή τιμή, για παράδειγμα k=1 ή k=2)

– $\Lambda_t(i,j) = g(D(i,j))$, όπου το g είναι μια καμπανοειδής συνάρτηση

– $\Lambda_t(i,j) = g(D(i,j)/\lambda(t))$, όπου $\lambda(t) \to 0$, όταν $t \to \infty$.

Ένα κοινό παράδειγμα είναι: $\Lambda_t(i,j) = \exp\left(-\frac{D(i,j)^2}{2\sigma(t)^2}\right)$, όπου:

- $\sigma(t) = \sigma_\iota(\sigma_f/\sigma_i)^{(t/M)}$, με $\sigma_i > \sigma_f > 0$ (για παράδειγμα $\sigma_i = 5$ και $f = 0.2$).

Η παράμετρος προσαρμογής ε_t παίζει σημαντικό ρόλο στη σύγκλιση του αλγορίθμου. Για παράδειγμα, μπορούμε να επιλέξουμε ε_t έτσι ώστε $\sum_{t\ge0}\varepsilon_t = +\infty$ και ώστε $\sum_{t\ge0}(\varepsilon_t)^2 < +\infty$ καθώς $\varepsilon_t = 1/t$ (συνθήκες Robbins–Monro που προέρχονται από τους στοχαστικούς αλγόριθμους).

Ένα άλλο σύνηθες παράδειγμα είναι: $\varepsilon_t = \varepsilon_i\left(\frac{\varepsilon_f}{\varepsilon_i}\right)^{(t/M)}$, με $\varepsilon_i > \varepsilon_f > 0$ (για παράδειγμα, $\varepsilon_i = 0.1\ \kappa\alpha\iota\ \varepsilon_f = 0.005$).

Η ανάλυση της σύγκλισης αυτού του αλγορίθμου θέτει δύσκολα μαθηματικά προβλήματα. Αν και η συντριπτική πλειονότητα των εφαρμογών αλγορίθμου Kohonen χρησιμοποιεί χώρους δεδομένων τουλάχιστον 2 διαστάσεων και εμφανίζει το φαινόμενο της αυτοοργάνωσης, η μεγάλη πλειοψηφία των μαθηματικών αποτελεσμάτων που το περιγράφουν ικανοποιητικά είναι έγκυρα μόνο στη διάσταση 1

(τοπολογία συμβολοσειράς και χώρος δεδομένων διάστασης 1)! Για μια επισκόπηση γνωστών μαθηματικών αποτελεσμάτων, μπορεί να γίνει αναφορά στο άρθρο σύνθεσης των Cottrell, Fort and Pagès (1998). Αυτό το άρθρο περιέχει επίσης μια πολύ εκτενή λίστα αναφορών για το θέμα.

5.6.3. Εφαρμογές

Αυτός ο αλγόριθμος και οι πολυάριθμες παραλλαγές του έχουν ένα ευρύ φάσμα εφαρμογών. Το πλεονέκτημα αυτής της μεθόδου ταξινόμησης είναι ότι καθιστά δυνατή την γραφική αναπαράσταση των παρατηρήσεων με σεβασμό της τοπολογίας του δικτύου. Οι παρατηρήσεις που ανήκουν σε αυτή την κατηγορία συσχετίζονται επομένως με κάθε μονάδα. Όταν τα δεδομένα και οι κλάσεις είναι πολυάριθμες, είναι δυνατό να πραγματοποιηθεί μια δεύτερη ταξινόμηση στα διανύσματα που αντιπροσωπεύουν τις κλάσεις, προκειμένου να ληφθεί ένα πιο χονδροειδές επίπεδο ταξινόμησης, το οποίο θα είναι ευκολότερο να ερμηνευτεί. Πολλές παραλλαγές αυτού του αλγορίθμου έχουν προταθεί. Συγκεκριμένα, είναι δυνατή η εφαρμογή της στην περίπτωση ποιοτικών μεταβλητών, συσχετίζοντας την με ανάλυση παραγοντικής αντιστοιχίας (Corlosquet-Habart and Janssen, 2018).

Κεφάλαιο 6ο : Αποτελέσματα εφαρμογών μεθόδων σε πραγματικά δεδομένα

Οι Farbmacher, Löw and Spindler (2022), στηριζόμενοι σε μοντέλα βαθιάς μάθησης που βασίζονται σε τεχνητά νευρωνικά δίκτυα ακολούθησαν αυτή την προσέγγιση για ένα σημαντικό βήμα στη διαδικασία διαχείρισης αποζημιώσεων: τον έλεγχο και τη διαχείριση του κινδύνου απάτης και λάθους. Σύμφωνα με τους ίδιους, αυτές οι μέθοδοι μπορούν να χειριστούν άμεσα ασυνήθιστες δομές δεδομένων και να εκτελέσουν μηχανική χαρακτηριστικών, ως μέρος της διαδικασίας εκμάθησης. Επιπλέον, οι ίδιοι συγγραφείς διερεύνησαν το βαθμό στον οποίο τέτοια μοντέλα παρέχουν ουσιαστικές εξηγήσεις για τις προβλέψεις τους. Στην εμπειρική τους εφαρμογή, εκπαίδευσαν τυποποιημένα και μη τυποποιημένα μοντέλα μηχανικής εκμάθησης σε πραγματικά δεδομένα που περιέχουν προηγούμενες αξιώσεις και τις σχετικές ετικέτες τους, οι οποίες ταξινομούν κάθε αξίωση, είτε ως σωστή, είτε ως ύποπτη. Το τελευταίο μπορεί να αναφέρεται σε οποιοδήποτε είδος λάθους, από τυπογραφικά λάθη και ακούσια ανακωδικοποίηση, έως απάτη. Η προτεινόμενη αρχιτεκτονική τους μπορεί να χρησιμοποιηθεί για τον μετριασμό των ασυμμετριών πληροφοριών και την επιτάχυνση των επακόλουθων ελεγκτικών αποφάσεων. Χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων από ιδιωτική ασφαλιστική εταιρεία στη Γερμανία, οι συγγραφείς ανέπτυξαν ένα βαθύ νευρωνικό δίκτυο που μπορεί να χρησιμοποιηθεί για την αυτόματη ανίχνευση δόλιων αξιώσεων. Το εν λόγω νευρωνικό δίκτυο, όχι μόνο ταξινομεί τους ισχυρισμούς, αλλά παρέχει επίσης δυνητικά σημαντικές εξηγήσεις. Το μοντέλο τους μπορεί να χειριστεί δεδομένα με εισόδους διαφορετικού μήκους και μεταβλητές με μεγάλο αριθμό κατηγοριών. Καθώς η ψηφιοποίηση πολλών πτυχών της ζωής μας συνεχίζεται, πολύ μεγάλοι όγκοι μη δομημένων δεδομένων με χαρακτηριστικά παρόμοια με τα δικά μας θα είναι διαθέσιμα και ως εκ τούτου είναι σημαντικό για τους εφαρμοσμένους

ερευνητές να έχουν στη διάθεσή τους τέτοιες μεθόδους. Τα εμπειρικά αποτελέσματα της συγκεκριμένης έρευνας, τεκμηριώνουν την απόδοση της βαθιάς μάθησης ως εργαλείου που βασίζεται σε δεδομένα στην προγνωστική ανάλυση στη διαχείριση αξιώσεων. Επιπλέον, αποδεικνύουν ότι το μοντέλο βαθιάς εκμάθησής τους όχι μόνο προσθέτει μια άλλη «κόκκινη» σημαία (red flag), αλλά παρέχει επίσης πρόσθετες πληροφορίες για τους ίδιους τους μηχανισμούς απάτης.

Οι Chun et al. (2019) παρατήρησαν ότι κατά τη συλλογή δεδομένων αποζημίωσης για ασφαλιστική απάτη, τα δεδομένα μεμονωμένων δειγμάτων ενδέχεται να αποκλίνουν από την πραγματική κατάσταση, λόγω ποικίλων ανθρώπινων και τυχαίων παραγόντων που θα προκαλέσουν παραμόρφωση του μοντέλου. Οι συγγραφείς αναγνώρισαν ότι το νευρωνικό δίκτυο BP έχει καλή στιβαρότητα και ανοχή σφαλμάτων και πως η απόκλιση μεμονωμένων δειγμάτων δεν μπορεί να επηρεάσει τη συνολική απόδοση του δικτύου και μπορεί να διορθωθεί μαθαίνοντας από το ίδιο το δίκτυο, χαρακτηριστικά τα οποία αποτελούν πλεονέκτημα της χρήσης του νευρωνικού δικτύου BP ως μοντέλου αναγνώρισης. Από την άλλη πλευρά, οι συγγραφείς αναγνωρίζουν ότι το παραδοσιακό νευρωνικό δίκτυο BP έχει τυχαία αρχικά βάρη, γεγονός που οδηγεί σε χαμηλή απόδοση εκμάθησης, αργή ταχύτητα σύγκλισης και ευκολία σχηματισμού τοπικών ελάχιστων χωρίς συνολική βελτιστοποίηση. Γι’ αυτό το λόγο, οι συγγραφείς πρότειναν έναν βελτιωμένο προσαρμοστικό γενετικό αλγόριθμο σε συνδυασμό με έναν αλγόριθμο νευρωνικών δικτύων BP για τον εντοπισμό απάτης ασφάλισης οχημάτων. Το μοντέλο τους συνδυάζει ένα γενετικό αλγόριθμο με ένα νευρωνικό δίκτυο BP, με βάση το γεγονός ότι τα νευρωνικά δίκτυα BP έχουν ισχυρή ικανότητα πρόβλεψης και ότι ο γενετικός αλγόριθμος έχει καλή ικανότητα αναζήτησης για βελτιστοποίηση. Στην εργασία των Chun et al. (2019), ο υπάρχων δείκτης δεδομένων απάτης ασφάλισης οχημάτων – με δεδομένα απάτης μιας συγκεκριμένης ασφαλιστικής εταιρείας -

ταξινομείται και ποσοτικοποιείται και στη συνέχεια επιλέγεται ο κύριος δείκτης απάτης ασφάλισης οχημάτων με ανάλυση κυρίων συνιστωσών (Principal component analysis), το οποίο χρησιμοποιείται ως είσοδος του νευρωνικού δικτύου BP. Ο βελτιωμένος προσαρμοστικός γενετικός αλγόριθμος των συγγραφέων (NAGA-BP), λαμβάνει υπόψη τον βαθμό καταλληλότητας του πληθυσμού και προσαρμόζει την πιθανότητα διασταύρωσης και την πιθανότητα μετάλλαξης του γενετικού αλγόριθμου. Στην τελική εμπειρική ανάλυση, ο βελτιωμένος γενετικός αλγόριθμος συγκρίθηκε με τους αλγόριθμους IAGA και GA όσον αφορά την ταχύτητα και την ακρίβεια σύγκλισης και το νευρωνικό δίκτυο BP βελτιστοποιήθηκε για την πρόβλεψη δεδομένων ασφαλιστικής απάτης. Τα αποτελέσματα έδειξαν ότι τα δεδομένα πρόβλεψης της απάτης στην ασφάλιση οχημάτων που ελήφθησαν από το βελτιωμένο μοντέλο NAGA-BP, είναι πιο κοντά στα αρχικά δεδομένα.

Στο άρθρο τους, οι Rukhsar et al. (2022) εκτελούν μια συγκριτική ανάλυση σε διάφορους αλγόριθμους ταξινόμησης, δηλαδή Support vector machine (SVM), Τυχαίο Δάσος (RF), Δέντρο Απόφασης (DT), Adaboost, K-Πλησιότερος Γείτονας (KNN), Γραμμική Παλινδρόμηση (LR), Naive Bayes (NB) και Multi-Layer Perceptron (MLP) για τον εντοπισμό της ασφαλιστικής απάτης, με δεδομένα από αξιόπιστο αποθετήριο ασφαλιστικής εταιρείας. Η αποτελεσματικότητα των αλγορίθμων παρατηρείται με βάση τις μετρήσεις απόδοσης: Precision, Recall και F1-Score. Τα συγκριτικά αποτελέσματα των αλγορίθμων ταξινόμησης καταλήγουν στο συμπέρασμα ότι το Δέντρο Απόφασης (DT) δίνει την υψηλότερη ακρίβεια 79% σε σύγκριση με τις άλλες τεχνικές. Επιπλέον, το Adaboost δίνει ακρίβεια 78% που είναι πιο κοντά στο Δέντρο Απόφασης DT. Ακολουθούν το KNN (77%), το RF (76%) και τέλος τα LR, SVM, NB και MLP (73%).

Ως μέρος της ανάπτυξης ενός αποτελεσματικού πλαισίου για τον εντοπισμό απάτης, οι Shamitha and Ilango (2020) με βάση ένα σύνολο δεδομένων αξιώσεων από την ασφαλιστική εταιρεία CMS Medicare, κατέληξαν στο συμπέρασμα ότι η εφαρμογή του Multi-Layer Perceptron, ενός νευρωνικού δικτύου τροφοδοσίας με βελτιστοποίηση γενετικού αλγορίθμου είχε βοηθήσει στη βελτίωση των αποτελεσμάτων και στην απόκτηση μεγαλύτερης ακρίβειας. Το PCA (Principal Component Analysis) εφαρμόστηκε επίσης για την επιλογή των πιο σημαντικών μεταβλητών. Η χρήση του PCA και άλλων κατάλληλων τεχνικών προεπεξεργασίας, βοήθησε επίσης στη μείωση του χρόνου εκπαίδευσης, επιτυγχάνοντας έτσι την αποτελεσματικότητα όσον αφορά την ακρίβεια και την ταχύτητα.

Οι Sowah et al. (2019), αξιοποίησαν ένα σύνολο δεδομένων αξιώσεων του Εθνικού Σχεδίου Ασφάλισης Υγείας που ελήφθη από νοσοκομεία της Γκάνα για τον εντοπισμό απάτης ασφάλισης υγείας και άλλων ανωμαλιών, ώστε να αξιολογήσουν το support vector machine στο πλαίσιο του συστήματος υποστήριξης αποφάσεων για τον εντοπισμό απάτης σε αξιώσεις ασφάλισης υγείας. Χρησιμοποίησαν μηχανές γενετικής διανυσματικής υποστήριξης (Genetic Support Vector Machines - GSVMs), ένα νέο εργαλείο υβριδοποιημένης εξόρυξης δεδομένων και στατιστικής μηχανικής μάθησης, που παρέχουν ένα σύνολο εξελιγμένων αλγορίθμων για τον αυτόματο εντοπισμό δόλιων αξιώσεων σε αυτές τις βάσεις δεδομένων ασφάλισης υγείας. Αξιολογήθηκαν τρεις ταξινομητές GSVM και συγκρίθηκαν τα αποτελέσματά τους. Τα πειραματικά αποτελέσματα δείχνουν σημαντική μείωση στον υπολογιστικό χρόνο κατά την επεξεργασία των αξιώσεων, ενώ αυξάνεται η ακρίβεια ταξινόμησης μέσω των διαφόρων ταξινομητών SVM (γραμμικός (80,67%), πολυωνυμικός (81,22%) και πυρήνας συνάρτησης ακτινικής βάσης (radial basis function - RBF) (87,91%).

Οι Muranda, Ali and Shongwe (2020) επιχείρησαν την ανίχνευση δόλιων αξιώσεων ασφάλισης αυτοκινήτων, με χρήση support vector machines με μέθοδο προσαρμοστικής συνθετικής δειγματοληψίας. Τα πειραματικά δεδομένα που χρησιμοποιήθηκαν σε αυτή την έρευνα, προέρχονται από ένα σύνολο δεδομένων ασφάλισης αυτοκινήτων πραγματικής ζωής από μια δημόσια διαθέσιμη πηγή δεδομένων. Οι συγγραφείς πρότειναν μια μέθοδο που χρησιμοποιεί την προσαρμοστική συνθετική μέθοδο δειγματοληψίας (ADASYN) για την άρση ανισορροπιών στο σύνολο δεδομένων. Στη συνέχεια χρησιμοποίησαν support vector machines (SVM) για να ταξινομήσουν τις υποθέσεις αξίωσης. Τα αποτελέσματα του αλγορίθμου τα συνέκριναν με τα μη ισορροπημένα σύνολα δεδομένων και άλλες υπάρχουσες μεθόδους. Τα αποτελέσματά τους έδειξαν αύξηση της συνολικής ακρίβειας του support vector machine από περίπου 93% σε υψηλό σχεδόν 98%. Ο αλγόριθμος έχει επίσης βελτιωμένη ακρίβεια, ανάκληση και βαθμολογία F1 (μέτρο που συνδυάζει τις παραμέτρους ανάκλησης και ακρίβειας με τέτοιο τρόπο που «τιμωρεί» τη μέτρηση με χαμηλή βαθμολογία και ευνοεί τους ταξινομητές με ίση ακρίβεια και ανάκληση) λόγω της εξισορρόπησης του αριθμού των δειγμάτων από κάθε κατηγορία. Σημείωσαν επίσης, ότι αυτό το μοντέλο μπορεί να βελτιωθεί περαιτέρω χρησιμοποιώντας άλλες ισορροπημένες τεχνικές, που βασίζονται σε περισσότερη γνώση των προτύπων δεδομένων, αντί να χρησιμοποιούν στατιστικές τεχνικές.

Οι Gong, Zhang and Du (2020) διεξήγαγαν έρευνα σχετικά με τη μέθοδο ανίχνευσης απάτης ολοκληρωμένης μάθησης, με βάση τη σύντηξη ταξινομητή συνδυασμού (tree hybrid bagging - THBagging), χρησιμοποιώντας ως μελέτη περίπτωσης το σύνολο δεδομένων ιατρικού διακανονισμού από το σχέδιο δράσης «Internet + Human Society» 2020 που εκδόθηκε από το Υπουργείο Ανθρώπινου Δυναμικού και Κοινωνικής Ασφάλισης της Κίνας. Το σύνολο δεδομένων περιλάμβανε δεδομένα

απευαισθητοποίησης ιατρικού διακανονισμού ιατρικής ασφάλισης και λεπτομέρειες κόστους 20.000 ασφαλισμένου προσωπικού σε 456 ιατρικά ιδρύματα, από τον Ιούλιο του 2016 έως τον Δεκέμβριο του 2016 στην επαρχία Hebei, στο Πεκίνο και στην Tianjin, στην Κίνα. Περιλαμβάνει κυρίως τα στοιχεία ιατρικών εξόδων και τα στοιχεία εξόδων του ασφαλισμένου προσωπικού, καθώς και πληροφορίες για το εάν υπάρχουν παράνομες συμπεριφορές απάτης από τα ταμεία ιατρικής ασφάλισης. Ανάμεσά τους, υπάρχουν 19.000 κανονικοί άνθρωποι (θετικά δείγματα) και 1.000 απατεώνες (αρνητικά δείγματα) και περιλαμβάνονται συνολικά 74 χαρακτηριστικά. Σε αυτή την εργασία, αναλύθηκε και συζητήθηκε το πρόβλημα έξυπνης αναγνώρισης της απάτης βασικής ιατρικής ασφάλισης. Σύμφωνα με το σενάριο του προβλήματος, διενεργείται λεπτομερής και σε βάθος ανάλυση και εξαγωγή χαρακτηριστικών μέσω ανάλυσης δεδομένων και εξόρυξης και δύο γύροι εξαγωγής χαρακτηριστικών πραγματοποιούνται με βάση την παραδοσιακή λειτουργία εξαγωγής χαρακτηριστικών. Με στόχο το πρόβλημα της μη ισορροπημένης κατανομής κατηγοριών στο σενάριο αναγνώρισης απάτης βασικής ιατρικής ασφάλισης, προτάθηκε ο αλγόριθμος THBagging. Αυτός ο αλγόριθμος χρησιμοποιήθηκε για την επίλυση προβλημάτων ανεπαρκούς χρήσης δείγματος, εύκολης υπερπροσαρμογής και χαμηλού ποσοστού αναγνώρισης στο πρόβλημα κατανομής κατηγορίας. Τα αποτελέσματα που λαμβάνονται με τη χρήση του μοντέλου σύντηξης, είναι καλύτερα από τις μεθόδους μηχανικής μάθησης και νευρωνικών δικτύων. Η ακρίβεια και η ανάκληση της προτεινόμενης μεθόδου που βασίζεται στην ολοκληρωμένη μάθηση είναι πάνω από 70% και 45% αντίστοιχα και τουλάχιστον υψηλότερα από τον βασικό αλγόριθμο 2,93% και 2% στα $F1$ και $macro$-$F1$. Η $Macro$-$F1$ αντιμετωπίζει όλες τις κατηγορίες εξίσου και δεν επηρεάζεται εύκολα από κοινές κατηγορίες. Ειδικά στην περίπτωση μη ισορροπημένων κατηγοριών δειγμάτων, το αποτέλεσμα του $Macro$-$F1$ ($Macro$-$F1$ είναι μια μέτρηση που αξιολογεί

τον μέσο όρο $F1$ όλων των διαφορετικών ετικετών τάξης) είναι καλύτερο. Δεν δείχνει μόνο την υπεροχή της μεθόδου διαίρεσης δεδομένων, αλλά δείχνει επίσης ότι το μοντέλο THBagging είναι πιο εύρωστο σε ανομοιόμορφα δείγματα. Το προτεινόμενο THBagging ανήκει στο μοντέλο σύντηξης και η τιμή $F1$ και η τιμή $macro\text{-}F1$ είναι υψηλότερες από όλους τους συνδυασμούς μοντέλων βασικής ταξινόμησης που χρησιμοποιούνται, υποδεικνύοντας ότι η έννοια της σύντηξης μοντέλου είναι επιτυχής. Μέσω του ερευνητικού περιεχομένου αυτού του άρθρου, αποδεικνύεται στα πειραματικά δεδομένα ότι ο αλγόριθμος THBagging είναι καλύτερος από τους παραδοσιακούς αλγόριθμους.

Οι Dhieb et al. (2020) ανέπτυξαν ένα ασφαλές και αυτοματοποιημένο ασφαλιστικό πλαίσιο συστήματος, που μειώνει την ανθρώπινη αλληλεπίδραση, προστατεύει τις ασφαλιστικές δραστηριότητες, ειδοποιεί και ενημερώνει για επικίνδυνους πελάτες, εντοπίζει δόλιες αξιώσεις και μειώνει τη χρηματική ζημία για τον ασφαλιστικό τομέα. Μετά την παρουσίαση του πλαισίου που βασίζεται σε blockchain για να επιτραπούν ασφαλείς συναλλαγές και κοινή χρήση δεδομένων μεταξύ διαφορετικών αλληλεπιδρώντων πρακτόρων εντός του ασφαλιστικού δικτύου, προτείνουν τη χρήση αλγορίθμου μηχανικής εκμάθησης ακραίας κλίσης (XGBoost) για τις προαναφερθείσες ασφαλιστικές υπηρεσίες και συγκρίνουν τις επιδόσεις του με αυτές άλλων αλγορίθμων τελευταίας τεχνολογίας. Τα αποτελέσματα αποκαλύπτουν ότι όταν εφαρμόζεται σε ένα σύνολο δεδομένων ασφάλισης αυτοκινήτου, το XGboost επιτυγχάνει υψηλά κέρδη απόδοσης σε σύγκριση με άλλους υπάρχοντες αλγόριθμους εκμάθησης. Για παράδειγμα, φτάνει το 7% υψηλότερη ακρίβεια σε σύγκριση με τα μοντέλα δέντρων αποφάσεων, όταν εντοπίζονται δόλιες αξιώσεις.

Οι Harjai, Khatri and Singh (2019) επιχείρησαν την ανίχνευση δόλιων αξιώσεων ασφάλισης με χρήση τυχαίων δασών και τεχνικής υπερδειγματοληψίας συνθετικών

μειονοτήτων. Σε αυτή τη μελέτη, οι συγγραφείς έδειξαν μια νέα προσέγγιση για την κατασκευή ενός ανιχνευτή απάτης αυτοασφάλισης με βάση τη μηχανική μάθηση, ο οποίος θα προβλέπει δόλιες αξιώσεις ασφάλισης από το σύνολο δεδομένων με πάνω από 15.420 αρχεία αξιώσεων αυτοκινήτου. Το προτεινόμενο μοντέλο έχει κατασκευαστεί χρησιμοποιώντας τεχνική συνθετικής μειοψηφίας υπερδειγματοληψίας (SMOTE), η οποία αφαιρεί την ανισορροπία κλάσης του συνόλου δεδομένων. Χρησιμοποίησαν τη μέθοδο ταξινόμησης τυχαίων δασών, για την ταξινόμηση των εγγραφών αξιώσεων. Τα αποτελέσματα της προσέγγισής τους συγκρίθηκαν με άλλα υπάρχοντα μοντέλα με βάση διάφορες μετρήσεις απόδοσης. Τα αποτελέσματα των μετρήσεών τους παρατίθενται στον παρακάτω πίνακα:

Performance Metrics (in %)	**Support Vector Machine (SVM)**	**Decision Tree**	**Multi-layer Perception (MLP)**	**Proposed Model**
Accuracy	58.41	57.39	74.98	94.33
Sensitivity (or Recall value)	90.53	86.94	47.83	99.9
Specificity	36.86	38.14	18.75	45.1

Πηγή: Harjai, Khatri & Singh (2019)

Οι Severino and Peng (2021) χρησιμοποίησαν αλγόριθμους μηχανικής μάθησης για την πρόβλεψη απάτης στην ασφάλιση περιουσίας με εμπειρικά στοιχεία μικροδεδομένα πραγματικού κόσμου. Στο άρθρο τους αξιολόγησαν μοντέλα

πρόβλεψης που βασίζονται σε μηχανική μάθηση για τον εντοπισμό απατών σε αξιώσεις ασφαλιστηρίων συμβολαίων περιουσίας, συγκρίνοντας τα προγνωστικά αποτελέσματα εννέα προγνωστικών μοντέλων χρησιμοποιώντας δεδομένα από μια μεγάλη βραζιλιάνικη ασφαλιστική εταιρεία. Τα αποτελέσματα έδειξαν ότι το μοντέλο τυχαίου δάσους πέτυχε σημαντικά καλύτερη απόδοση από την τυπική λογιστική παλινδρόμηση και άλλες μεθόδους μηχανικής μάθησης, όπως αποδεικνύεται από τις μετρήσεις της accuracy, του precision, της βαθμολογίας F1, του Cohen's Kappa και του MCC, ενώ το μοντέλο βαθιού νευρωνικού δικτύου, ξεπέρασε τα άλλα μοντέλα για τη μέτρηση ανάκλησης (recall). Επιπλέον, με βάση τις τεκμηριωμένες περιπτώσεις απάτης, παρέθεσαν ένα μακροοικονομικό προφίλ των απατεώνων και κατέταξαν τη σχετική σημασία των επεξηγηματικών μεταβλητών σύμφωνα με μια προσέγγιση που βασίζεται στη μετάθεση, επισημαίνοντας τα χαρακτηριστικά που συνέβαλαν περισσότερο στη συνολική προγνωστική ισχύ των μοντέλων και για την πρόβλεψη εξέχουσας ψευδώς θετικών και ψευδώς αρνητικών παρατηρήσεων. Ο πίνακας των αποτελεσμάτων τους δίνεται κάτωθι:

Πίνακας 2: Μέση τιμή και απόκλιση των μέτρων απόδοσης για 1000 γύρους με εκτός δείγματος προβλέψεις

Model	Accuracy	Precision	Recall	F1 Score	Kappa	MCC
Logistic Regression	80.67%	80.56%	78.99%	79.67%	61.26%	61.41%
	(1.60%)	(2.94%)	(3.79%)	(1.81%)	(3.21%)	(3.18%)
Penalized Logistic Regression	81.40%	81.27%	79.95%	80.48%	61.34%	62.93%
	(1.72%)	(3.36%)	(3.99%)	(1.88%)	(3.19%)	(3.36%)
Naive Bayes	71.16%	73.18%	73.02%	72.39%	47.66%	49.51%
	(5.66%)	(8.64%)	(5.53%)	(3.72%)	(10.28%)	(8.78%)
KNN	75.74%	77.74%	69.77%	73.39%	51.25%	51.66%
	(2.51%)	(3.76%)	(4.67%)	(2.88%)	(5.01%)	(4.98%)
Polynomial Kernel SVM	81.34%	79.22%	82.98%	80.93%	62.92%	63.00%
	(0.75%)	(1.15%)	(1.06%)	(0.84%)	(1.48%)	(1.48%)
Gaussian Kernel SVM	79.56%	79.07%	78.41%	78.53%	58.81%	59.00%
	(1.71%)	(3.08%)	(4.17%)	(1.98%)	(3.36%)	(3.31%)
Deep Neural Network	81.88%	78.41%	86.28%	82.06%	63.84%	64.32%
	(1.58%)	(3.11%)	(3.06%)	(1.32%)	(3.08%)	(2.80%)
Random Forest	84.56%	84.72%	82.77%	83.61%	69.05%	69.24%
	(1.43%)	(2.60%)	(3.72%)	(1.65%)	(2.97%)	(2.88%)
GBM	83.21%	83.55%	81.73%	82.44%	66.20%	66.39%
	(1.61%)	(2.97%)	(3.96%)	(1.82%)	(3.08%)	(3.02%)

Πηγή: Severino and Peng (2021)

Επίλογος

Οι ασφαλιστικές εταιρείες διαδραματίζουν θεμελιώδη οικονομικό ρόλο. Επιτρέπουν στις εταιρείες να αναλαμβάνουν κινδύνους, με την πάροδο του χρόνου εξομαλύνουν το κόστος των ατυχημάτων που μπορούν να διαταράξουν τις δραστηριότητές τους και αντιμετωπίζουν σοβαρές απαιτήσεις που δεν θα μπορούσαν να αντέξουν χωρίς την ασφάλεια που παρέχουν οι ασφαλιστές. Ομοίως, για τα άτομα, προσφέρουν μια συγκέντρωση κινδύνων της καθημερινής ζωής (αυτοκίνητο, πυρκαγιά, ασθένειες κ.λπ.), που επιβαρύνοντας το κόστος των ατυχημάτων σε μεγάλο αριθμό ανθρώπων στο χώρο και στο χρόνο, το καθιστά αποδεκτό για όλους (Trichet, 2005).

Δεδομένης της σημασίας του οικονομικού ρόλου που διαδραματίζουν τα ασφαλιστικά ιδρύματα όσον αφορά την κοινωνία, τέτοιοι φορείς υπόκεινται σε αυστηρή ρύθμιση και παρακολούθηση. Είναι έντονα κεφαλαιοποιημένα και ακολουθούν τον προληπτικό κανονισμό European Solvency II. Υπόκεινται σε πολλούς νόμους και κανονισμούς, πολλοί από τους οποίους προορίζονται να εξασφαλίσουν στους πελάτες ότι ο ασφαλιστής θα εκπληρώσει τις υποχρεώσεις τους. Επιπλέον, οι κανονισμοί αποσκοπούν στην προστασία των ασφαλισμένων, έτσι ώστε οι ασφαλιστές να μην κάνουν κατάχρηση των πληροφοριών των πελατών, τις οποίες αποκτούν λόγω του επαγγέλματός τους. Οι ασφαλιστές θεωρείται ότι βρίσκονται σε μια κατάσταση ασυμμετρίας πληροφοριών που τους δίνει μια θέση ισχύος σε σχέση με τους πελάτες και επομένως περιορίζονται στη χρήση κριτηρίων τιμολόγησης. Οι κανονισμοί ορίζουν επίσης ότι η χρήση των προσωπικών δεδομένων πρέπει να πραγματοποιείται με σεβασμό της ιδιωτικής ζωής, χωρίς διακρίσεις, με φειδωλό και ισότιμο τρόπο και ότι ο ασφαλιστής θα λάβει όλα τα απαραίτητα μέτρα ασφαλείας για να προστατευτεί από κλοπή ή διαρροή δεδομένων, με ποινή αυστηρών κυρώσεων (Linder and Ronkainen, 2004).

Ο ρόλος των αναλογιστών είναι να κάνουν συνετή χρήση δεδομένων σχετικά με την αξιολόγηση και τον έλεγχο κινδύνου, συμπεριλαμβανομένου του υπολογισμού των προβλέψεων για ζημίες. Η ποιότητα των δεδομένων πρέπει να έχει προηγουμένως επαληθευτεί. Οι εγγυήσεις που εκδίδουν οι ασφαλιστικές εταιρείες καλύπτουν όλους τους οικονομικούς τομείς: κάλυψη περιουσίας ιδιωτών ή εταιρειών, κάλυψη συνεπειών φυσικών γεγονότων, ασφάλιση αστικής ευθύνης, ασφάλιση υγείας, συνταξιοδότησης, θανάτου και πρόνοιας καθώς και ασφάλιση ζωής και αποταμίευσης κ.λπ. (Corlosquet-Habart and Janssen, 2018),

Αν και είναι ιδιωτικές, οι ασφαλιστικές εταιρείες διαδραματίζουν θεμελιώδη ρόλο στην κοινωνική χρησιμότητα, ιδίως για συμβάσεις υγείας και πρόνοιας, εκτός από τα υποχρεωτικά συστήματα. Επομένως, υπόκεινται σε κίνητρα ή περιορισμούς από τις δημόσιες αρχές, μέσω φορολογικών ή ρυθμιστικών μέτρων. Στην ασφάλιση υγείας για παράδειγμα, ο φορολογικός συντελεστής διαφοροποιείται ανάλογα με το εάν τα συμβόλαια υγείας χαρακτηρίζονται ως υπεύθυνα ή μη, ενώ λαμβάνεται υπόψη το γεγονός ότι οι δημόσιες αρχές είναι αυτές που ορίζουν τους όρους εντολής και το καλάθι περίθαλψης που οι συμβάσεις πρέπει να συμμορφώνονται με προκειμένου να επικυρωθούν ως «υπεύθυνες». Προφανώς οι πελάτες επιλέγουν ως επί το πλείστων συμβόλαια με τις πιο ανταγωνιστικές τιμές, δηλαδή υπεύθυνα συμβόλαια, αφού φορολογούνται λιγότερο. Σε άλλες περιπτώσεις, από νόμους ή κανονισμούς, για λόγους μη διάκρισης ή για λόγους αλληλεγγύης, η ελευθερία τιμολόγησης του ασφαλιστή περιορίζεται από την απαγόρευση της χρήσης ορισμένων κριτηρίων, όπως το φύλο, αν και είναι αρκετά μεροληπτική όσον αφορά προσδόκιμο ζωής, ή όσον αφορά το ιατρικό ιστορικό, μέσω του δικαιώματος στη λήθη που εμποδίζει να ληφθεί υπόψη (Corlosquet-Habart and Janssen, 2018).

Η ασφαλιστική επιχείρηση, η οποία αποτελείται από μια υπόσχεση πληρωμής ενός ποσού, συνήθως άγνωστου, εάν συμβεί ένα τυχαίο γεγονός στο μέλλον, βασίζεται σε στατιστικά στοιχεία. Αυτές οι πληροφορίες συλλέγονται προσεκτικά και αποθηκεύονται σε δομημένες βάσεις δεδομένων. Η συγκέντρωση των κινδύνων διαφορετικών πελατών επιτρέπει στον ασφαλιστή να βασιστεί στο νόμο των μεγάλων αριθμών για να μοντελοποιήσει τον δικό του κίνδυνο με μαθηματικούς νόμους πιθανοτήτων, χρησιμοποιώντας μερικά κριτήρια που συνοψίζουν τις περισσότερες πληροφορίες. Μετρά επίσης τον κίνδυνο λάθους αυτής της μοντελοποίησης, η οποία χρησιμοποιείται από τη ρυθμιστική αρχή για να δηλώσει την απαίτηση κινητοποίησης κεφαλαίων σε σχέση με τις υποσχέσεις κάλυψης που δίνουν οι ασφαλιστές. Επομένως, τα δεδομένα είναι η πρώτη ύλη για τους ασφαλιστές (Boobier, 2016).

Είναι δουλειά των αναλογιστών να τα επιλέγουν και να τα χρησιμοποιούν για την εκτίμηση και τον έλεγχο των κινδύνων και τον υπολογισμό των προβλέψεων που θα εγγραφούν στους λογαριασμούς, ενώ προοπτικά και στοχαστικά προβάλλουν σε αρκετά χρόνια στο μέλλον τις διάφορες τεχνικές και εμπορικές παραμέτρους, για να διασφαλίσουν τη βιωσιμότητα της εταιρείας και δεσμεύσεις. Μέχρι στιγμής, με περιορισμένο αριθμό δεδομένων, οι ασφαλιστές έχουν αξιολογήσει αποτελεσματικά τους κινδύνους τους και σε εύθετο χρόνο έχουν αντιμετωπίσει τους υποσχεμένους κανονισμούς σε περίπτωση καταστροφών. Είναι προφανές ότι η έκρηξη των δεδομένων της ψηφιακής βιομηχανικής επανάστασης που γνωρίζουμε, που ονομάζονται μεγάλα δεδομένα, έρχεται να κλονίσει αυτή την καθιερωμένη τάξη πραγμάτων (Porrini, 2017).

Είναι κοινότυπη η παρατήρηση του φαινομένου της ψηφιακής έκρηξης και των μεγάλων δεδομένων που χαρακτηρίζονται από τα 5Vs: όγκος, ποικιλία, ταχύτητα, ακρίβεια και αξία. Η πιο ορατή εκδήλωση των μεγάλων δεδομένων είναι η έκρηξη του

«όγκου» των δεδομένων, που αυξάνεται εκθετικά λόγω της ισχύος των υπολογιστών και της γενίκευσης του Ιστού, των smartphone, των κοινωνικών δικτύων, των συνδεδεμένων αντικειμένων, της προσβασιμότητας στο cloud κ.λπ. (Corlosquet-Habart and Janssen, 2018).

Ωστόσο, τα μεγάλα δεδομένα τροποποιούν επίσης σε βάθος το πλαίσιο από την «ποικιλία» των δεδομένων που ανταλλάσσονται: αριθμοί φυσικά, καθώς και κείμενο, εικόνα, ήχος και βίντεο. Ένα άλλο χαρακτηριστικό είναι η «ταχύτητα», η ταχύτητα με την οποία οι πληροφορίες φθάνουν άμεσα και ταυτόχρονα σε όλες τις γωνιές του πλανήτη, προερχόμενες από κάθε είδους εκδότες, άτομα που στέλνουν φωτογραφίες από τις διακοπές τους ή συλλέγουν πληροφορίες αξιώσεων μέσω δορυφόρου. Στην ψηφιακή εποχή όλα είναι στιγμιαία και οι τεχνολογίες που διατίθενται σε ιδιώτες γίνονται αρκετά παρόμοιες με αυτές των εταιρειών (Picard, 2018).

Η «αλήθεια» παραμένει ένα σημείο εστίασης επειδή τα δεδομένα συλλέγονται χωρίς γνώση του πλαισίου τους και αυτό πρέπει να λαμβάνεται υπόψη στις χρήσεις που θα γίνουν στη συνέχεια. Όλα αυτά τα δεδομένα, που μοιράζονται με νέους αλγόριθμους και τεχνητή νοημοσύνη αποτελούν τη νέα πηγή ενέργειας του 21ου αιώνα, δημιουργώντας «αξία». Όλοι οι παράγοντες της οικονομίας καταβάλλουν μεγάλες προσπάθειες για την εξαγωγή αξίας από όλα αυτά τα δεδομένα, παίρνοντας το μερίδιό τους στον εμπλουτισμό και επιτυγχάνοντας τον ψηφιακό τους μετασχηματισμό, ιδίως για να εκμεταλλευτούν αυτά τα νέα δεδομένα για σκοπούς μάρκετινγκ (Zheng and Guo, 2020).

Λόγω του αρχικού τους οικονομικού μοντέλου, του αντίστροφου κύκλου, οι ασφαλιστές ανησυχούν ιδιαίτερα από αυτή την ψηφιακή βιομηχανική επανάσταση, καθώς τα δεδομένα είναι η πρώτη ύλη τους. Ενόψει ενός ακριβού τυχαίου γεγονότος

(μια καταστροφή), ένα άτομο ή μια εταιρεία έχει την επιλογή μεταξύ του να αναλάβει την πρόκληση, να αναλάβει τον κίνδυνο και να αναλάβει το κόστος ή να αποφασίσει να μεταφέρει αυτόν τον κίνδυνο σε ένα αξιόπιστο τρίτο μέρος, για μια αρχική καταβληθείσα τιμή. Οι προμηθευτές «υποσχέσεων», οι οποίοι έχουν αξιολογηθεί προσεκτικά με βάση αποδεδειγμένες μεθόδους σχετικά με τη χρήση στατιστικών, είναι επαγγελματίες δεδομένων που μετατρέπουν μεμονωμένους κινδύνους που είναι δύσκολο να αναληφθούν σε ελεγχόμενο συνολικό κίνδυνο. Από αυτή την άποψη, η παροχή νέων δεδομένων μπορεί μόνο να βοηθήσει τους ασφαλιστές να κάνουν καλύτερα τη δουλειά τους (Molloy and Ronnie, 2020).

Δεν είναι όμως πολύ διαφορετικής φύσης ο αντίκτυπος της ψηφιακής τεχνολογίας; Δεν βρίσκεται σε διαδικασία τροποποίησης όχι μόνο του τρόπου άσκησης της ασφαλιστικής δραστηριότητας, αλλά και της ίδιας της ασφαλιστικής; Οι ασφαλιστικές δραστηριότητες υποστηρίζουν οικονομικές και χρηματοπιστωτικές δραστηριότητες και επομένως εξαρτώνται σε μεγάλο βαθμό από αυτές τις δραστηριότητες. Είναι στην υπηρεσία των ατόμων για να τα βοηθήσουν να αντιμετωπίσουν τους κινδύνους του τρόπου ζωής τους. Εάν η οικονομία αλλάξει και οι άνθρωποι ζουν διαφορετικά, οι ασφαλιστικές ανάγκες θα είναι επίσης διαφορετικές. Οι ασφαλιστικές εταιρείες πρέπει επομένως να κατανοήσουν και να προσαρμοστούν στις μεταβαλλόμενες ανάγκες και ακόμη και να τις προβλέψουν, γι’ αυτό και ο ρόλος των μεγάλων δεδομένων στον ασφαλιστικό τομέα είναι ιδιαιτέρως σημαντικός (Corlosquet-Habart and Janssen, 2018).

Βιβλιογραφία

Aggarwal, C.C., 2018. Neural networks and deep learning. *Springer*, *10*(978), p.3.

Almeida, P. and Bernardino, J., 2015. A comprehensive overview of open source big data platforms and frameworks. *Int. J. Big Data*, *2*, pp.1-19.

Anscombe, F.J., 1967. Topics in the investigation of linear relations fitted by the method of least squares. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(1), pp.1-29.

Azzalini, A. and Scarpa, B., 2012. *Data analysis and data mining: An introduction*. OUP USA.

Balas, V.E., Solanki, V.K., Kumar, R. and Khari, M. eds., 2019. *Internet of things and big data analytics for smart generation* (Vol. 154, p. 309). Heidelberg: Springer.

Barry, L. and Charpentier, A., 2020. Personalization as a promise: Can Big Data change the practice of insurance?. *Big Data & Society*, *7*(1), p.2053951720935143.

Bartlett, P. and Traskin, M., 2006. Adaboost is consistent. *Advances in Neural Information Processing Systems*, *19*.

Beard, R., 2013. *Risk theory: the stochastic basis of insurance* (Vol. 20). Springer Science & Business Media.

Belhadi, A., Abdellah, N. and Nezai, A., 2023. The Effect of Big Data on the Development of the Insurance Industry.

Berk, R.A., 2008. *Statistical learning from a regression perspective* (Vol. 14). New York: Springer.

Berthelé, E., 2018. Using big data in insurance. *Big data for insurance companies*, *1*, pp.131-161.

Berthelé, E., 2018. Using big data in insurance. *Big data for insurance companies*, *1*, pp.131-161.

Bhavsar, H. and Panchal, M.H., 2012. A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *1*(10), pp.185-189.

Biau, G. and Devroye, L., 2010. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, *101*(10), pp.2499-2518.

Biau, G. and Scornet, E., 2016. A random forest guided tour. *Test*, *25*, pp.197-227.

Biau, G., Scornet, E. and Welbl, J., 2016. Neural random forests. *arXiv preprint arXiv:1604.07143*.

Billot, R., Bothorel, C. and Lenca, P., 2018. Introduction to Big Data and Its Applications in Insurance. *Big Data for Insurance Companies*, *1*, pp.1-25.

Boobier, T., 2016. *Analytics for insurance: The real business of Big Data*. John Wiley & Sons.

Boobier, T., 2016. *Analytics for insurance: The real business of Big Data*. John Wiley & Sons.

Borch, K.H., Sandmo, A. and Aase, K.K., 2014. *Economics of insurance*. Elsevier.

Box, G.E., 1979. All models are wrong, but some are useful. *Robustness in Statistics*, *202*(1979), p.549.

Boyd, D. and Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), pp.662-679.

Breiman, L., 1996a. Bagging predictors. *Machine learning*, *24*, pp.123-140.

Breiman, L., 1996b. Out-of-bag estimation.

Breiman, L., 1999. *Using adaptive bagging to debias regressions* (p. 16). Technical Report 547, Statistics Dept. UCB.

Breiman, L., 2001. Random forests. *Machine learning*, *45*, pp.5-32.

Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), pp.199-231.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A., 1984. *Classification and regression trees*. CRC press.

Campbell-Kelly, M., Aspray, W.F., Yost, J.R., Tinn, H. and Díaz, G.C., 2023. *Computer: A history of the information machine*. Taylor & Francis.

Cardona, M., Kretschmer, T. and Strobel, T., 2013. ICT and productivity: conclusions from the empirical literature. *Information Economics and policy*, *25*(3), pp.109-125.

Caruana, R., Karampatziakis, N. and Yessenalina, A., 2008, July. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (pp. 96-103).

Cassel, C. and Bindman, A., 2019. Risk, benefit, and fairness in a big data world. *Jama*, *322*(2), pp.105-106.

Chalmers, S., Bothorel, C. and Clemente, R.P., 2013. Big data-state of the art, pp.28.

Charpentier, A., 2020, February. Big Data, GAFA et assurance. In *Annales des Mines-Réalités industrielles* (No. 1, pp. 53-57). Cairn/Softwin.

Chatfield, C., 1985. The initial examination of data. *Journal of the Royal Statistical Society: Series A (General)*, *148*(3), pp.214-231.

Chen, C.P. and Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, pp.314-347.

Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey. *Mobile networks and applications*, *19*, pp.171-209.

Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey. *Mobile networks and applications*, *19*, pp.171-209.

Chun, Y., Meixuan, L., Liu, W., & Qi, M. (2019). Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network. Theoretical Computer Science.

Corlosquet-Habart, M. and Janssen, J. eds., 2018. *Big data for insurance companies*. John Wiley & Sons.

Corlosquet-Habart, M. and Janssen, J. eds., 2018. *Big data for insurance companies*. John Wiley & Sons.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*, pp.273-297.

Cottrell, M. and Rousset, P., 1997. The Kohonen algorithm: a powerful tool for analysing and representing multidimensional quantitative and qualitative data. In *Biological and Artificial Computation: From Neuroscience to Technology: International Work-Conference on Artificial and Natural Neural Networks, IWANN'97*

Lanzarote, Canary Islands, Spain, June 4–6, 1997 Proceedings 4 (pp. 861-871). Springer Berlin Heidelberg.

Cottrell, M., Fort, J.C. and Pagès, G., 1998. Theoretical aspects of the SOM algorithm. *Neurocomputing*, *21*(1-3), pp.119-138.

Delforge, P., 2015. America's data centers consuming and wasting growing amounts of energy, Natural Resource Defence Council. Available at: https://www.nrdc.org/resources/americas-data-centers-consuming-and-wasting-growing-amounts-energy

Dhamodharavadhani, S., Gowri, R. and Rathipriya, R., 2018. Unlock different V's of big data for analytics. *International Journal of Computer Sciences and Engineering*, *6*(4), pp.183-190.

Dhieb, N., Ghazzai, H., Besbes, H. and Massoud, Y., 2020. A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. IEEE Access, 8, pp.58546-58558.

Diebold, F.X., 2012. *On the Origin (s) and Development of the Term â€œBig Data* (No. 12-037). Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Donoho, D., 2015. 50 years of Data Science, Tukey Centennial workshop. Available at: https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Du, K.L. and Swamy, M.N., 2006. *Neural networks in a softcomputing framework* (Vol. 1). London: Springer.

Dubois, D.M., 1998. A review of neural networks with direct learning based on linear or non-linear threshold logics. *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, pp.283-303.

Fan, J., Han, F. and Liu, H., 2014. Challenges of big data analysis. *National science review*, *1*(2), pp.293-314.

Fan, W. and Bifet, A., 2013. Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter*, *14*(2), pp.1-5.

Fan, W. and Bifet, A., 2013. Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter*, *14*(2), pp.1-5.

Farbmacher, H., Löw, L. and Spindler, M., 2022. An explainable attention network for fraud detection in claims management. Journal of Econometrics, 228(2), pp.244-258.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research*, *15*(1), pp.3133-3181.

Freund, Y. and Schapire, R.E., 1996, July. Experiments with a new boosting algorithm. In *icml* (Vol. 96, pp. 148-156).

Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), pp.119-139.

Friedman, J.H., 1991. Multivariate adaptive regression splines. *The annals of statistics*, *19*(1), pp.1-67.

Friedman, J.H., 1998. Data Mining and Statistics: What's the connection?. *Computing science and statistics*, *29*(1), pp.3-9.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.

Gey, S. and Poggi, J.M., 2006. Boosting and instability for regression trees. *Computational statistics & data analysis*, *50*(2), pp.533-550.

Ghani, N.A., Hamid, S., Hashem, I.A.T. and Ahmed, E., 2019. Social media big data analytics: A survey. *Computers in Human behavior*, *101*, pp.417-428.

Ghasemaghaei, M. and Calic, G., 2019. Does big data enhance firm innovation competency? The mediating role of data-driven insights. *Journal of Business Research*, *104*, pp.69-84.

Gong, J., Zhang, H. and Du, W., 2020. Research on integrated learning fraud detection method based on combination classifier fusion (THBagging): A case study on the foundational medical insurance dataset. Electronics, 9(6), p.894.

GSMA, 2015. Unlocking the Value of IoT Through Big Data, Report, GSM Association. Available at: https://www.gsma.com/iot/wp-content/uploads/2015/12/cl_iot_bigdata_11_15-004.pdf

Gu, J. and Zhang, L., 2014. Some comments on big data and data science. *Annals of data science*, *1*, pp.283-291.

Hand, D.J., 1999. Why data mining is more than statistics writ large. *Bulletin of the International Statistical Institute*, *1*, pp.433-436.

Haoxiang, W. and Smys, S., 2021. Big data analysis and perturbation using data mining algorithm. *Journal of Soft Computing Paradigm (JSCP)*, *3*(01), pp.19-28.

Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), pp.1-16.

Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), pp.1-16.

Harjai, S., Khatri, S.K. and Singh, G., 2019, November. Detecting fraudulent insurance claims using random forests and synthetic minority oversampling technique. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 123-128). IEEE.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Haykin, S., 1998. *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hegazy, T., Fazio, P. and Moselhi, O., 1994. Developing practical neural network applications using back-propagation. *Computer-Aided Civil and Infrastructure Engineering*, *9*(2), pp.145-159.

Hiran, K.K., Jain, R.K., Lakhwani, K. and Doshi, R., 2021. *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications.

Hornik, K., 1993. Some new results on neural network approximation. *Neural networks*, *6*(8), pp.1069-1072.

Huang, Y., 2009. Advances in artificial neural networks–methodological development and application. *Algorithms*, *2*(3), pp.973-1007.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jane, J.B. and Ganesh, E.N., 2019. A Review On Big Data With Machine Learning And Fuzzy Logic For Better Decision Making. *International Journal of Scientific & Technology Research*, *8*(10), pp.1121-1125.

Kaswan, K.S., Dhatterwal, J.S., Sharma, H. and Sood, K., 2022. Big data in insurance innovation. *Big Data: A game changer for insurance industry*, pp.117-136.

Kaur, P., Sharma, M. and Mittal, M., 2018. Big data and machine learning based secure healthcare framework. *Procedia computer science*, *132*, pp.1049-1059.

Kelly, J., 2015. Big Data Vendor Revenue and Market Forecast, 2011–2026, Report, WIKIBON. Available at: https://wikibon.com/executive-summary-big-data-vendor-revenue-and-market-forecast-2011-2026/

Kemp, R., 2014. Legal aspects of managing Big Data. *Computer Law & Security Review*, *30*(5), pp.482-491.

Kitchin, R., 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kofidis, N., Margaris, A., Roumeliotis, M. and Adamopoulos, M., 2006. Investigation of the determinism of complex dynamical systems using simple back propagation neural networks. *International Journal of Computer Mathematics*, *83*(4), pp.419-427.

Kohonen, T., 1982. Analysis of a simple self-organizing process. *Biological cybernetics*, *44*(2), pp.135-140.

Kohonen, T., 1991. Self-organizing maps: Ophmization approaches. In *Artificial neural networks* (pp. 981-990). North-Holland.

Kussul, E., Baidyk, T., Kasatkina, L. and Lukovich, V., 2001, July. Rosenblatt perceptrons for handwritten digit recognition. In *IJCNN'01. International Joint*

Conference on Neural Networks. Proceedings (Cat. No. 01CH37222) (Vol. 2, pp. 1516-1520). IEEE.

Landset, S., Khoshgoftaar, T.M., Richter, A.N. and Hasanin, T., 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, *2*(1), pp.1-36.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), p.1.

Larose, D.T. and Larose, C.D., 2014. *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.

Larose, D.T. and Larose, C.D., 2014. *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.

Linder, U. and Ronkainen, V., 2004. Solvency II–towards a new insurance supervisory system in the EU. *Scandinavian Actuarial Journal*, *2004*(6), pp.462-474.

Loh, W.Y., 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), pp.14-23.

Manyika J., Chui M., Brown B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A., 2011. Big data: The next frontier for innovation, competition, and productivity, Report, The McKinsey Global Institute. Available at: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_exec_summary.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A., 2011. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Martinez, I., Viles, E. and Olaizola, I.G., 2021. Data science methodologies: Current challenges and future approaches. *Big Data Research*, *24*, p.100183.

Mayer-Schönberger, V. and Cukier, K., 2013. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, pp.115-133.

McFall, L., 2019. Personalizing solidarity? The role of self-tracking in health insurance pricing. *Economy and society*, *48*(1), pp.52-76.

Mehmood, E. and Anees, T., 2022. Distributed real-time ETL architecture for unstructured big data. *Knowledge and Information Systems*, *64*(12), pp.3419-3445.

Minsky, M. and Papert, S., 1969. An introduction to computational geometry. *Cambridge tiass., HIT*, *479*(480), p.104.

Molloy, L. and Ronnie, L., 2020. Sustaining the life insurance industry in the Fourth Industrial Revolution. *South African Actuarial Journal*, *20*(1), pp.81-107.

Müller, B., Reinhardt, J. and Strickland, M.T., 1995. *Neural networks: an introduction*. Springer Science & Business Media.

Muranda, C., Ali, A. and Shongwe, T., 2020, October. Detecting fraudulent motor insurance claims using support vector machines with adaptive synthetic sampling method. In 2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS) (pp. 1-5). IEEE.

Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S.A., Montesano, N., Tariq, M.I., De-la-Hoz-Franco, E. and De-La-Hoz-Valdiris, E., 2022. Trends and future perspective challenges in big data. In *Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania* (pp. 309-325). Springer Singapore.

Nicholson, J.E., 2019. Challenges for the Insurance Industry in the Future. *Journal of Insurance Regulation*, *38*(6).

Picard, F., 2018. Current vision and market prospective. *Big Data for Insurance Companies, 1*, pp.83-129.

Popescu, M.C., Balas, V.E., Perescu-Popescu, L. and Mastorakis, N., 2009. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*(7), pp.579-588.

Porrini, D., 2017. Regulating Big Data effects in the European insurance market. *Insurance markets and companies*, (8), pp.6-15.

Quan, Z. and Valdez, E.A., 2018. Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), pp.377-407.

Quan, Z. and Valdez, E.A., 2018. Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), pp.377-407.

Quinlan, J.R., 1996, August. Bagging, boosting, and C4. 5. In *Aaai/Iaai, vol. 1* (pp. 725-730).

Raj, A. and D'Souza, R., 2019. A Review on Hadoop Eco System for Big Data. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(1), pp.343-348.

Rokach, L. and Maimon, O., 2005. Decision trees. *Data mining and knowledge discovery handbook*, pp.165-192.

Rukhsar, L., Bangyal, W.H., Nisar, K. and Nisar, S., 2022. Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology, 41(1), pp.33-40.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. *nature*, *323*(6088), pp.533-536.

Sakr, S. and Gaber, M. eds., 2014. *Large scale and big data: Processing and management*. Crc Press.

Sangeetha, S. and Sudha Sadasivam, G., 2019. Privacy of big data: a review. *Handbook of big data and iot security*, pp.5-23.

Saporta, G., 2008. Models for understanding versus models for prediction. In *COMPSTAT 2008: Proceedings in computational statistics* (pp. 315-322). Physica-Verlag HD.

Schapire, R.E., 2003. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, pp.149-171.

Scornet, E., Biau, G. and Vert, J.P., 2015. Consistency of random forests.

Severino, M.K. and Peng, Y., 2021. Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 5, p.100074.

Shamitha, S.K. and Ilango, V., 2020, July. A time-efficient model for detecting fraudulent health insurance claims using artificial neural networks. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-6). IEEE.

Shmueli, G., 2010. To explain or to predict? *Statistical Science*, *25*, pp. 289–310.

Smyth, P., 2000. Data mining: data analysis on a grand scale?. *Statistical methods in medical research*, *9*(4), pp.309-327.

Sowah, R.A., Kuuboore, M., Ofoli, A., Kwofie, S., Asiedu, L., Koumadi, K.M. and Apeadu, K.O., 2019. Decision support system (DSS) for fraud detection in health insurance claims using genetic support vector machines (GSVMs). Journal of Engineering, 2019.

Somvanshi, M., Chavan, P., Tambade, S. and Shinde, S.V., 2016, August. A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing communication control and automation (ICCUBEA)* (pp. 1-7). IEEE.

Spender, A., Bullen, C., Altmann-Richer, L., Cripps, J., Duffy, R., Falkous, C., Farrell, M., Horn, T., Wigzell, J. and Yeap, W., 2019. Wearables and the internet of things: Considerations for the life and health insurance industry. *British Actuarial Journal*, *24*, p.e22.

Steinberg, D. and Colla, P., 2009. CART: classification and regression trees. *The top ten algorithms in data mining*, *9*, p.179.

Stitson, M.O., Weston, J.A.E., Gammerman, A., Vovk, V. and Vapnik, V., 1996. Theory of support vector machines. *University of London*, *117*(827), pp.188-191.

Taheri, J., 2018. Big Data and Software Defined Networks. The Institution of Engineering and Technology. IET Computing Series 15.

Tappert, C.C., 2019, December. Who is the father of deep learning?. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 343-348). IEEE.

Thouvenin, F., Suter, F., George, D. and Weber, R.H., 2019. Big data in the insurance industry. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, *10*, p.209.

Trendowicz, A., Jeffery, R., Trendowicz, A. and Jeffery, R., 2014. Classification and regression trees. *Software Project Effort Estimation: Foundations and Best Practice Guidelines for Success*, pp.295-304.

Trichet, J.C., 2005. Financial stability and the insurance sector. *The Geneva Papers on risk and insurance-issues and practice*, *30*, pp.65-71.

van Valkengoed, A.M. and Steg, L., 2019. Meta-analyses of factors motivating climate change adaptation behaviour. *Nature climate change*, *9*(2), pp.158-163.

Vapnik, V., 2006. *Estimation of dependences based on empirical data*. Springer Science & Business Media.

Vapnik, V., Guyon, I. and Hastie, T., 1995. Support vector machines. *Mach. Learn*, *20*(3), pp.273-297.

Vermet, F., 2018. Statistical learning methods. *Big Data for Insurance Companies*, *1*, pp.43-82.

Ward, J.S. and Barker, A., 2013. Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.

Warren, J. and Marz, N., 2015. *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster.

Wolf, E.J., Harrington, K.M., Clark, S.L. and Miller, M.W., 2013. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and psychological measurement*, *73*(6), pp.913-934.

Wolpert, D.H., 1992. Stacked generalization. *Neural networks*, *5*(2), pp.241-259.

Wythoff, B.J., 1993. Backpropagation neural networks: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *18*(2), pp.115-155.

Yadav, N., Yadav, A., Kumar, M., Yadav, N., Yadav, A. and Kumar, M., 2015. History of neural networks. *An introduction to neural network methods for differential equations*, pp.13-15.

Yan, C., Li, M., Liu, W. and Qi, M., 2020. Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network. Theoretical Computer Science, 817, pp.12-23.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, *92*(4), pp.937-950.

Yin, Y., Stecke, K.E. and Li, D., 2018. The evolution of production systems from Industry 2.0 through Industry 4.0. *International Journal of Production Research*, *56*(1-2), pp.848-861.

Zhang, S., Zhang, C. and Yang, Q., 2003. Data preparation for data mining. *Applied artificial intelligence*, *17*(5-6), pp.375-381.

Zheng, L. and Guo, L., 2020, April. Application of big data technology in insurance innovation. In *International conference on education, economics and information management (ICEEIM 2019)* (pp. 285-294). Atlantis Press.

Zikopoulos, P. and Eaton, C., 2011. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.