

Αύγουστος 2023



**ΣΧΟΛΗ : ΟΙΚΟΝΟΜΙΚΩΝ, ΔΙΟΙΚΗΣΗΣ ΚΑΙ
ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΤΜΗΜΑ : ΠΛΗΡΟΦΟΡΙΚΗΣ

**Ο ΡΟΛΟΣ ΤΩΝ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ (BIG
DATA) ΚΑΙ ΤΗΣ ΠΡΟΓΝΩΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ
(PREDICTIVE ANALYTICS) ΣΤΗ ΒΙΟΜΗΧΑΝΙΑ
ΑΛΟΥΜΙΝΙΟΥ**

ΚΥΡΛΗΣ ΙΑΣΩΝ

ΑΥΓΟΥΣΤΟΣ/2023



**ΣΧΟΛΗ : ΟΙΚΟΝΟΜΙΚΩΝ, ΔΙΟΙΚΗΣΗΣ ΚΑΙ
ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΤΜΗΜΑ: ΠΛΗΡΟΦΟΡΙΚΗΣ

**Ο ΡΟΛΟΣ ΤΩΝ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ (BIG
DATA) ΚΑΙ ΤΗΣ ΠΡΟΓΝΩΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ
(PREDICTIVE ANALYTICS) ΣΤΗ ΒΙΟΜΗΧΑΝΙΑ
ΑΛΟΥΜΙΝΙΟΥ**

**Διατριβή η οποία υποβλήθηκε προς απόκτηση εξ
αποστάσεως μεταπτυχιακού τίτλου σπουδών στα
Πληροφοριακά Συστήματα και τη Ψηφιακή Καινοτομία στο
Πανεπιστήμιο Νεάπολις**

ΚΥΡΛΗΣ ΙΑΣΩΝ

ΑΥΓΟΥΣΤΟΣ/2023

Πνευματικά δικαιώματα

Copyright © Όνομα επίθετο φοιτητή, έτος κατάθεσης διατριβής

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της διατριβής από το Πανεπιστημίου Νεάπολις δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Πανεπιστημίου.

Ή ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Ο Κυρλής Ιάσων γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Ο ρόλος των μεγάλων δεδομένων (big data) και της προγνωστικής ανάλυσης (Predictive analytics) στη βιομηχανία αλουμινίου », αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Ο Δηλών

Κυρλής Ιάσων

A handwritten signature in black ink, appearing to read 'Kyrlis Iason', enclosed within a thin black rectangular border.

Περιεχόμενα

Περίληψη	1
Abstract.....	3
Κεφάλαιο 1 – Εισαγωγή	4
1.1 Χρήση των Μεγάλων Δεδομένων σε μεγάλες επιχειρήσεις.....	4
1.2 Τα Μεγάλα Δεδομένα στον κλάδο των βιομηχανιών αλουμινίου.....	6
1.3 Σκοπός και επιμέρους στόχοι.....	7
1.4 Βασικά Ερευνητικά Ερωτήματα.....	8
1.5 Αναμενόμενα Αποτελέσματα	9
Κεφάλαιο 2 - Βιβλιογραφική Ανασκόπηση.....	10
2.1.Ορισμός.....	10
2.2 Ιστορική αναδρομή	18
2.3 Αναλυτική (Analytics).....	20
2.4. Βιβλιογραφική Ανασκόπηση στις βιομηχανίες	22
2.4.1. Βιομηχανία παραγωγής.....	22
2.4.2. Βιομηχανία αλουμινίου	23
Κεφάλαιο 3 - Μεθοδολογία Έρευνας	26
3.1. Εισαγωγή	26
3.1.1 Συλλογή δεδομένων.....	26
3.1.2. Μεταβλητές.....	26
3.1.2. Προεπεξεργασία δεδομένων	28
3.2. Στατιστικές και ML προσεγγίσεις	29
3.3. Δενδρολογικές Μέθοδοι: Decision Trees and Forests.....	29
3.4. Διαχωρισμός και Μετασχηματισμός Δεδομένων (Splitting and Scalling).....	30
3.5. Η Αντιστάθμιση Μεροληψίας-Διακύμανσης (Bias-Variance Tradeoff).....	31
3.6. Μέθοδοι Ενίσχυσης και Ομαδοποίησης (Boosting and Ensemble Methods)	31
Κεφάλαιο 4 – Ανάλυση και Αποτελέσματα	35

4.1. Εισαγωγή	35
4.2. Ανάλυση Δεδομένων	35
Κεφάλαιο 5 - Συζήτηση αποτελεσμάτων και συμπεράσματα	46
5.1 Ερμηνεία αποτελεσμάτων.....	46
5.2. Συμπεράσματα	47
Βιβλιογραφία	50

Κατάλογος Γραφικών Παραστάσεων/Εικόνων/Διαγραμμάτων

Εικόνα 1. Τα 5V's των "Big Data"	11
Εικόνα 2. Όγκος δεδομένων (Volume).....	12
Εικόνα 3. Ετήσιο Μέγεθος της Παγκόσμιας Σφαίρας Δεδομένων.....	13
Εικόνα 4. Ταχύτητα δεδομένων (Velocity)	15
Εικόνα 5. Αξία (Value).....	17
Διάγραμμα 4.1. Ιστογράμματα και εκτιμήσεις (KDEs) των μεταβλητών-στόχων πρόβλεψης στη βάση δεδομένων του Αλουμινίου	36
Διάγραμμα 4.2.Συσχετίσεις μεταξύ των μεταβλητών του συνόλου δεδομένων.	37
Διάγραμμα 4.3. Scatter plot γραμμικής παλινδρόμησης OLS.....	38
Διάγραμμα 4.4. Doughnut plot για το μοντέλο γραμμικής παλινδρόμησης.....	39
Διάγραμμα 4.5. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από τον αλγόριθμο δέντρου αποφάσεων RandomForest	40
Διάγραμμα 4.6. Doughnut plot για το μοντέλο RandomForest.....	41
Διάγραμμα 4.7. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από την μέθοδο XGB	42
Διάγραμμα 4.8. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από την μέθοδο ExtraTrees.	43
Διάγραμμα 4.9. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από την μέθοδο LightGBM.	44
Διάγραμμα 4.10.Scatterplots και των 6 μοντέλων πρόβλεψης για το σύνολο δεδομένων του αλουμινίου.	45

Κατάλογος Πινάκων

Πίνακας 4.1. Περιγραφικά μέτρα των 5 μεταβλητών.....	35
--	----

Ονοματεπώνυμο Φοιτητή/Φοιτήτριας:

Τίτλος Μεταπτυχιακής Διατριβής:

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου στο Πανεπιστήμιο Νεάπολις και εγκρίθηκε στις [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή:

Πρώτος επιβλέπων (Πανεπιστήμιο Νεάπολις Πάφος).....[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Ευχαριστίες

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου κ. Ζαγόρη Κωνσταντίνο για την κατανόηση του κατά την εκπόνηση της παρούσας μελέτης και την στήριξή του. Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους ανθρώπους που ήταν δίπλα μου για την ηθική υποστήριξη που μου προσέφεραν. Ιδιαίτερως, θα ήθελα να ευχαριστήσω τον καθηγητή Χαράλαμπο Αγκυρόπουλο για τις ακαδημαϊκές προτάσεις του, τις υποδείξεις και συμβουλές, καθώς και τις πολύτιμες γνώσεις που μου προσέφερε στην στατιστική ανάλυση των δεδομένων, τις οποίες μοιράστηκε απλόχερα μαζί μου.

Αφιέρωση

Στην μνήμη του πατέρα μου

Περίληψη

Στο σημερινό ταχέως εξελισσόμενο επιχειρηματικό τοπίο, οι οργανισμοί σε όλους τους κλάδους αξιοποιούν τη δύναμη των δεδομένων για να προωθήσουν τη λήψη τεκμηριωμένων αποφάσεων και να αποκτήσουν ανταγωνιστικό πλεονέκτημα. Η βιομηχανία αλουμινίου, με τις περίπλοκες λειτουργίες και τη δυναμική αλυσίδα εφοδιασμού της, δεν αποτελεί εξαίρεση. Υιοθετώντας τα μεγάλα δεδομένα και την προγνωστική ανάλυση, ο κλάδος αυτός έχει τη δυνατότητα να φέρει επανάσταση στις διαδικασίες του, να αυξήσει τη λειτουργική αποδοτικότητα και να ξεκλειδώσει νέες ευκαιρίες.

Ένα από τα βασικά οφέλη των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη βιομηχανία είναι η βελτιστοποίηση των διαδικασιών. Αναλύοντας τεράστιες ποσότητες δεδομένων, οι εταιρείες μπορούν να εντοπίσουν μοτίβα και τάσεις που αποκαλύπτουν ανεπάρκειες στις δραστηριότητές τους. Αυτή η διορατικότητα τους επιτρέπει να εξορθολογίσουν τις ροές εργασίας παραγωγής, να μειώσουν τη σπατάλη και να βελτιώσουν τη συνολική παραγωγικότητα. Για παράδειγμα, αναλύοντας ιστορικά δεδομένα και πληροφορίες αισθητήρων σε πραγματικό χρόνο, οι οργανισμοί μπορούν να βελτιστοποιήσουν τις παραμέτρους των κλιβάνων, τις συνθέσεις κραμάτων και τις τεχνικές χύτευσης, με αποτέλεσμα να παράγουν προϊόντα υψηλότερης ποιότητας με μειωμένη κατανάλωση ενέργειας.

Η βιομηχανία αλουμινίου βασίζεται σε μια παγκόσμια και διασυνδεδεμένη αλυσίδα εφοδιασμού και τα μεγάλα δεδομένα σε συνδυασμό με την προγνωστική ανάλυση μπορούν να ενισχύσουν τη διαχείρισή της. Η ορατότητα σε πραγματικό χρόνο για τα επίπεδα αποθεμάτων, τα πρότυπα ζήτησης και τις αλυσίδες μεταφοράς και αποθήκευσης επιτρέπει στους οργανισμούς να βελτιστοποιούν τα επίπεδα αποθεμάτων, να προβλέπουν τις διακυμάνσεις της ζήτησης και να λαμβάνουν αποφάσεις βάσει δεδομένων σχετικά με την προμήθεια και τη διανομή. Κάτι τέτοιο έχει ως αποτέλεσμα τη βελτίωση της επιχειρησιακής απόδοσης, τη μείωση του κόστους και τη βελτίωση της ικανοποίησης των πελατών.

Η βιωσιμότητα αποτελεί αυξανόμενη ανησυχία στη βιομηχανία αλουμινίου. Τα μεγάλα δεδομένα και η προγνωστική ανάλυση μπορούν να συμβάλουν καθοριστικά στις πρωτοβουλίες και προοπτικές βιωσιμότητας κάθε είδους επιχείρησης. Με την ανάλυση δεδομένων αναφορικά για παράδειγμα με την κατανάλωση ενέργειας και τις παραμέτρους παραγωγής, οι επιχειρήσεις μπορούν να εντοπίσουν περιοχές αναποτελεσματικότητας και να εφαρμόσουν στοχευμένες βελτιώσεις, οδηγώντας σε εξοικονόμηση της ενέργειας και

μείωση των αποβλήτων. Η προγνωστική ανάλυση μπορεί επίσης να προβλέψει τη ζήτηση ενέργειας και να επιτρέψει τη δυναμική εξισορρόπηση φορτίου, βελτιστοποιώντας την κατανάλωση σε όλη τη διαδικασία παραγωγής.

Λέξεις κλειδιά: Μεγάλα Δεδομένα, προγνωστική ανάλυση, βιομηχανία αλουμινίου, παραγωγή, διαχείριση αποθεμάτων, βιωσιμότητα.

Abstract

In today's rapidly evolving business landscape, organisations across all industries are harnessing the power of data to drive informed decision making and gain competitive advantage. The aluminum industry, with its complex operations and dynamic supply chain, is no exception. By adopting big data and predictive analytics, this industry has the potential to revolutionise its processes, increase operational efficiency and unlock new opportunities.

One of the key benefits of big data and predictive analytics in the industry is process optimization. By analyzing vast amounts of data, companies can identify patterns and trends that reveal inefficiencies in their operations. This insight allows them to streamline production workflows, reduce waste and improve overall productivity.

The aluminium industry relies on a global and interconnected supply chain and big data combined with predictive analytics can enhance its management. Real-time visibility into inventory levels, demand patterns and transportation logistics allows organizations to optimize inventory levels, anticipate demand fluctuations and make data-driven decisions on supply and distribution. This results in improved operational efficiency, reduced costs and improved customer satisfaction.

Sustainability is a growing concern in the aluminum industry, and big data and predictive analytics can contribute to sustainability initiatives. By analyzing data, indicatively on energy consumption and production parameters, organizations can identify areas of inefficiency and implement targeted improvements, leading to energy optimization and waste reduction. Predictive analytics can also predict demand, energy demand and enable dynamic load balancing, optimising energy use throughout the production process.

Keywords: Big Data, predictive analytics, aluminum industry, production, demand, logistics, sustainability.

Κεφάλαιο 1 – Εισαγωγή

1.1 Χρήση των Μεγάλων Δεδομένων σε μεγάλες επιχειρήσεις

Η εξόρυξη δεδομένων ήταν μια πρόκληση για τους ερευνητές τις τελευταίες δεκαετίες. Οι αναλυτικές τεχνικές των βάσεων δεδομένων βοηθούν στην εξαγωγή πολύτιμων πηγών πληροφοριών για επιστήμονες και επαγγελματίες διαφορετικών ειδικοτήτων ως μέσο ανίχνευσης προτύπων συμπεριφοράς, εκμετάλλευσης ευρημάτων και βελτίωσης της κρίσης και της λήψης αποφάσεων.

Τα μεγάλα δεδομένα έχουν διαδραματίσει καθοριστικό ρόλο σε πολλές βιομηχανίες με αποτέλεσμα να προκαλέσουν επανάσταση στις επιχειρήσεις, δεδομένου ότι εξάγουν πολύτιμες πληροφορίες από τεράστιους όγκους δεδομένων. Η ανάλυση μεγάλων δεδομένων στη διαχείριση της εφοδιαστικής αλυσίδας, επιτρέπει στις εταιρείες να βελτιστοποιούν τις δραστηριότητές τους και να βελτιώνουν την απόδοσή τους (Chen et al., 2019). Μελετώντας και αναλύοντας ένα τεράστιο όγκο δεδομένων που αφορούν για παράδειγμα τα επίπεδα αποθεμάτων, τη ζήτηση των πελατών, τις διαδρομές μεταφοράς και την απόδοση των προμηθευτών, οι εταιρείες είναι σε θέση να λαμβάνουν αποφάσεις για τη βελτίωση της αποδοτικότητας και τη μείωση του κόστους.

Ο τομέας της μεταποίησης αξιοποιεί την ανάλυση μεγάλων δεδομένων για τη βελτιστοποίηση λειτουργιών μεγάλης κλίμακας. Η ανάλυση σε πραγματικό χρόνο των δεδομένων που παράγονται από αισθητήρες, μηχανές και γραμμές παραγωγής βοηθά στον εντοπισμό μοτίβων, στην ανίχνευση ανωμαλιών και στη βελτιστοποίηση των διαδικασιών παραγωγής (Zhang et al., 2019). Αυτό επιτρέπει στους κατασκευαστές να παρακολουθούν τον έλεγχο της ποιότητας, να προβλέπουν τις βλάβες του εξοπλισμού, να μειώνουν τον χρόνο διακοπής λειτουργίας και να βελτιώνουν την παραγωγικότητα.

Η ανάλυση μεγάλων δεδομένων είναι καθοριστική για τη διαχείριση και τη βελτιστοποίηση της παραγωγής, της διανομής και της κατανάλωσης ενέργειας. Η ανάλυση δεδομένων από έξυπνους μετρητές, αισθητήρες και καιρικές συνθήκες βοηθά στην πρόβλεψη της ζήτησης ενέργειας, στη βελτιστοποίηση των λειτουργιών του δικτύου, στον εντοπισμό διακοπών ρεύματος και στον εντοπισμό περιοχών για εξοικονόμηση ενέργειας (Rahmani et al., 2021). Η αξιοποίηση των μεγάλων δεδομένων βοηθά στην πρόβλεψη φορτίου, στην ανίχνευση βλαβών και στη διαχείριση της ενέργειας στον τομέα της ενέργειας.

Αυτά είναι μερικά μόνο παραδείγματα που παρουσιάζουν την αξιοποίηση των μεγάλων δεδομένων σε επιχειρήσεις μεγάλης κλίμακας. Άλλοι τομείς, όπως οι μεταφορές και τα logistics, οι έξυπνες πόλεις και άλλοι, επωφελούνται επίσης από την ανάλυση μεγάλων δεδομένων (Khan et al., 2020).

Για τη σύγχρονη βιομηχανία, τα δεδομένα που παράγονται από μηχανές και συσκευές, οι λύσεις που βασίζονται στο σύννεφο, η διαχείριση των επιχειρήσεων, έχουν φτάσει σε συνολικό όγκο άνω των 1000 Exabytes ετησίως και αναμένεται να 20πλασιαστούν τα επόμενα δέκα χρόνια. Οι McKinsey & Company αναφέρει ότι "η μεταποίηση αποθηκεύει περισσότερα δεδομένα κοντά στα 2 Exabyte νέων δεδομένων που αποθηκεύονται το 2010" (Bailly M. & Manyika J., 2013). Για παράδειγμα, σε μια εταιρεία καταναλωτικών συσκευασμένων προϊόντων που παράγει ένα προϊόν προσωπικής φροντίδας, ένα μόνο μηχάνημα παράγει 5000 δείγματα δεδομένων κάθε 33 ms, με αποτέλεσμα τέσσερα τρισεκατομμύρια (δηλ. 4 Tera) δείγματα ετησίως (Oussous et al., 2018).

Από την άποψη της βιομηχανίας, τα μεγάλα δεδομένα πρόκειται να διαδραματίσουν σημαντικό ρόλο στην τέταρτη βιομηχανική επανάσταση (Kagermann H. & Wahlster W., 2013). Η πρώτη βιομηχανική επανάσταση (από τα τέλη του 18ου έως τις αρχές του 20ού αιώνα) εξαρτιόταν από το νερό και ατμοηλεκτρική ενέργεια, η δεύτερη (από τον αρχές του 20ού αιώνα έως τις αρχές της δεκαετίας του εβδομήντα) εξαρτήθηκε από τη μαζική παραγωγή, βασισμένη στον καταμερισμό της εργασίας και την ηλεκτρική ενέργεια, και η τρίτη (από τις αρχές του δεκαετία του '70 έως σήμερα) εξαρτήθηκε από την ηλεκτρονική και την πληροφορική για περαιτέρω αυτοματοποίηση της παραγωγής. . Το καύσιμο της τέταρτης βιομηχανικής επανάστασης, που ονομάστηκε "Industrie 4.0" από την γερμανική κυβέρνηση, θα είναι τα μεγάλα δεδομένα που θα διατίθενται μέσω των κυβερνοφυσικών συστημάτων (CPS).

Στόχος είναι η υλοποίηση έξυπνων εργοστασίων, στα οποία οι μηχανές και οι πόροι επικοινωνούν όπως σε ένα κοινωνικό δίκτυο. Ένα τέτοιο έξυπνο εργοστάσιο θα παράγει ευφυή προϊόντα (έξυπνα προϊόντα) που θα γνωρίζουν πώς έχουν παραχθεί και θα συλλέγουν, θα μεταδίδουν δεδομένα κατά τη χρήση τους. Αυτές οι τεράστιες ποσότητες δεδομένων (big data) θα συλλέγονται και θα αναλύονται σε πραγματικό χρόνο. Έτσι, θα παράγονται νέες γνώσεις που θα χρησιμοποιούνται για να προχωρήσουν ένα επίπεδο πιο πάνω από το έξυπνα εργοστάσια σε έξυπνες διαδικασίες, και, τέλος, στο επίπεδο στο οποίο θα μπορούν να παρέχονται ευφυείς υπηρεσίες στον πελάτη μέσω του διαδικτύου.

1.2 Τα Μεγάλα Δεδομένα στον κλάδο των βιομηχανιών αλουμινίου

Η βιομηχανία αλουμινίου έχει σημαντικό αντίκτυπο σε διάφορους οικονομικούς τομείς σε όλο τον κόσμο, συμπεριλαμβανομένων των μεταφορών, των κατασκευών, της συσκευασίας και των ηλεκτρικών εξαρτημάτων. (Ratnasari et al., 2020). Η σύνθεση του αλουμινίου βασίζεται στο οξείδιο του αλουμινίου, το οποίο είναι ελαφρύ και ανθεκτικό στη διάβρωση, γεγονός που του επιτρέπει να χρησιμοποιείται σε διάφορες καταστάσεις. Η ανάπτυξη και η βιωσιμότητα του κλάδου εξαρτάται κατά κύριο λόγο από παράγοντες όπως η αποδοτικότητα της παραγωγής, η μείωση του κόστους, ο ποιοτικός έλεγχος και η βιωσιμότητα.

Πρόσφατα, η χρήση μεγάλων δεδομένων και προγνωστικών αναλύσεων σε βιομηχανικά περιβάλλοντα έχει αυξηθεί σε δημοτικότητα, ειδικά για τη βιομηχανία αλουμινίου. Τα μεγάλα δεδομένα χαρακτηρίζονται από τον μεγάλο όγκο και από πολυπλοκότητα, τα οποία είναι δύσκολο να χειριστούν με τις παραδοσιακές μεθόδους επεξεργασίας δεδομένων. (Nguyen et al., 2020). Αντίθετα, η προγνωστική ανάλυση περιλαμβάνει την εξαγωγή πληροφοριών για τη λήψη τεκμηριωμένων προβλέψεων και αποφάσεων βελτιστοποίησης. (Ahmadi et al., 2020).

Τα δυνητικά οφέλη από τη χρήση των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη βιομηχανία αλουμινίου είναι σημαντικά. Με την ανάλυση μεγάλου όγκου δεδομένων που παράγονται κατά τη διάρκεια της παραγωγικής διαδικασίας, οι εταιρείες μπορούν να αποκτήσουν πολύτιμες γνώσεις σχετικά με τους παράγοντες που επηρεάζουν την αποδοτικότητα, την ποιότητα και το κόστος (Karoor et al., 2021). Οι τεχνικές προγνωστικής ανάλυσης μπορούν να βοηθήσουν στην πρόβλεψη των προτύπων ζήτησης, στη βελτιστοποίηση της διαχείρισης αποθεμάτων και στη βελτίωση του σχεδιασμού συντήρησης. Αυτές οι εξελίξεις μπορούν να οδηγήσουν σε αυξημένη αποδοτικότητα της παραγωγής, μειωμένο χρόνο διακοπής λειτουργίας, καλύτερη κατανομή των πόρων και βελτιωμένη ικανοποίηση των πελατών (Das et al., 2021).

Ωστόσο, υπάρχουν αρκετές προκλήσεις που σχετίζονται με την αξιοποίηση των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη βιομηχανία αλουμινίου. Μια σημαντική πρόκληση είναι η διαχείριση και η ανάλυση τεράστιου όγκου δεδομένων που παράγονται από διάφορες πηγές, συμπεριλαμβανομένων των συστημάτων παραγωγής, των αισθητήρων, των αλυσίδων εφοδιασμού και των αλληλεπιδράσεων με τους πελάτες (Luo et al., 2020). Η διασφάλιση της ποιότητας των δεδομένων, της ασφάλειας και της ιδιωτικής ζωής αποτελούν επίσης σημαντικές ανησυχίες (Al-Turjman et al., 2021). Επιπλέον, η εφαρμογή των μεγάλων δεδομένων και της προγνωστικής ανάλυσης απαιτεί ισχυρή τεχνολογική υποδομή,

εξειδικευμένους αναλυτές δεδομένων και κατάλληλα πλαίσια διακυβέρνησης δεδομένων (Liang et al., 2021).

Ως εκ τούτου, είναι σημαντικό να εμβαθύνουμε στον ρόλο των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη βιομηχανία αλουμινίου. Η παρούσα έρευνα έχει ως στόχο να αναλύσει τα πιθανά οφέλη, τις προκλήσεις και τις ευκαιρίες που συνδέονται με την υιοθέτηση αυτών των τεχνολογιών στη βιομηχανία. Με την κατανόηση των επιπτώσεων και την ανάπτυξη στρατηγικών για την αντιμετώπιση των εμποδίων, η βιομηχανία αλουμινίου μπορεί να αξιοποιήσει τη δύναμη των μεγάλων δεδομένων και της προγνωστικής ανάλυσης για να ενισχύσει τη λειτουργική αποτελεσματικότητα, την ανταγωνιστικότητα και τη βιωσιμότητα.

1.3 Σκοπός και επιμέρους στόχοι

Ο κύριος σκοπός της παρούσας διατριβής είναι να διερευνήσει το ρόλο των μεγάλων δεδομένων, καθώς και της χρήσης τους στοχεύοντας σε μεγαλύτερη προγνωστική ακρίβεια στον κλάδο της βιομηχανίας αλουμινίου. Επιπλέον, στόχος της παρούσας εργασίας είναι να εξετάσει την τρέχουσα κατάσταση υιοθέτησης των μεγάλων δεδομένων, να διερευνήσει τα πιθανά οφέλη, και να παράσχει βέλτιστες πρακτικές και συστάσεις για την επιτυχή εφαρμογή των νέων τεχνολογιών στον κλάδο της βιομηχανίας αλουμινίου μέσω συγκεκριμένης μελέτης περίπτωσης.

Επιπρόσθετα, η διατριβή επιδιώκει να αξιολογήσει τα πλεονεκτήματα που μπορούν να αποκτηθούν από την εφαρμογή των νέων τεχνολογιών όσον αφορά την αποδοτικότητα της παραγωγής, τη βελτιστοποίηση του κόστους και τη βιωσιμότητα, καθώς και τον αντίκτυπό τους στις διαδικασίες λήψης αποφάσεων και τη συνολική επιχειρησιακή απόδοση.

Πιο συγκεκριμένα, στην παρούσα εργασία παρέχονται πολύτιμες συστάσεις για τους επαγγελματίες του κλάδου και τους υπεύθυνους χάραξης πολιτικής σχετικά με την υιοθέτηση των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη βιομηχανία αλουμινίου. Η αποτελεσματική εφαρμογή των νέων μεθόδων μέσω της χρήσης μεγάλων δεδομένων είναι βέβαιο πως θα ενισχύσει την ανταγωνιστικότητα, τη βιωσιμότητα και τη συνολική απόδοση, οδηγώντας σε βελτιωμένη επιχειρησιακή αποδοτικότητα και διαδικασίες λήψης αποφάσεων.

Ειδικότερα, η επίτευξη μιας αποτελεσματικής λειτουργίας των διαδικασιών, με την επίτευξη των επιθυμητών επιπέδων απόδοσης, ιδίως όσον αφορά την ποιότητα αποτελεί μια από τις

σημαντικότερες προκλήσεις τη δεδομένη στιγμή. Οι μηχανικοί διεργασιών, αντί να εργάζονται με ένα φυσικό μοντέλο του συστήματος (το οποίο μπορεί να είναι πολύ δύσκολο, αν όχι αδύνατο), προτιμούν ένα κατά προσέγγιση μοντέλο και χρησιμοποιούν προηγμένες τεχνικές για την παρακολούθηση, τον έλεγχο και τη βελτιστοποίηση της απόδοσης της διαδικασίας, με βάση μόνο τον τεράστιο όγκο μετρήσεων. Η αποτελεσματική καταγραφή και ανάλυση των μεγάλων δεδομένων δίνει τη δυνατότητα βελτίωσης της παραγωγικότητας με άμεσο επακόλουθο την υψηλή ανταγωνιστικότητα.

Από την άποψη της παραγωγής, η διαχείριση της εφοδιαστικής αλυσίδας μπορεί να βελτιωθεί μέσω των μεγάλων δεδομένων (Reichert, 2014). Επιπλέον, με τη σωστή ερμηνεία των μεγάλων δεδομένων, μπορούν να επιτευχθούν πιο αποτελεσματικά συστήματα διαχείρισης κινδύνων για να βοηθήσουν τη διοίκηση της εταιρείας να λαμβάνει καλύτερα τεκμηριωμένες αποφάσεις και να βελτιώσει την εταιρική διακυβέρνηση (H. Chen et al., 2020).

Με βάση αυτές τις παρατηρήσεις, είναι προφανές ότι η βασική έρευνα με επίκεντρο τα μεγάλα δεδομένα είναι επιτακτική και κρίσιμη για την επιχειρηματική κοινότητα και ιδιαίτερα για το μέλλον της βιομηχανίας.

1.4 Βασικά Ερευνητικά Ερωτήματα

Προκειμένου να διερευνήσουμε την ενσωμάτωση των μεγάλων δεδομένων και της προγνωστικής ανάλυσης με τη διαχείριση της αλυσίδας εφοδιασμού στη βιομηχανία αλουμινίου ανακύπτουν τα εξής ερευνητικά ερωτήματα:

- Πώς μπορούν αυτές οι τεχνολογίες να βελτιστοποιήσουν την παραγωγή σε συνάρτηση με τη διαχείριση αποθεμάτων και της ζήτησης;
- Ποια είναι τα πιθανά οφέλη της ενσωμάτωσης των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στις λειτουργίες της εφοδιαστικής αλυσίδας;

Εν συνεχεία για να εξετάσουμε τη δυνητική χρήση των μεγάλων δεδομένων και της προγνωστικής ανάλυσης για τη βελτίωση της βιωσιμότητας και της εκτίμησης των περιβαλλοντικών επιπτώσεων στη βιομηχανία αλουμινίου προκύπτουν τα εξής:

- Πώς μπορούν αυτές οι τεχνολογίες να βοηθήσουν στην παρακολούθηση και ανάλυση δεικτών περιβαλλοντικών επιδόσεων;

Τέλος, για να διερευνήσουμε το ρόλο των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στην ενίσχυση των απαιτήσεων των πελατών στη βιομηχανία αλουμινίου θα απαντήσουμε στα ακόλουθα:

- Πώς μπορούν να αξιολογηθούν οι δυνατότητες αξιοποίησης των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη διαχείριση κινδύνων και τη λήψη αποφάσεων στη βιομηχανία αλουμινίου.

1.5 Αναμενόμενα Αποτελέσματα

Με τη χρήση των μεγάλων δεδομένων και τις προηγμένες μεθόδους μηχανικής μάθησης αναμένεται να αναδειχθούν οι τρόποι με τους οποίους οι βιομηχανίες είναι δυνατόν να αυξήσουν την αποδοτικότητα, να μειώσουν τα κόστη, καθώς και να παράγουν πιο συνεπείς και ακριβείς προβλέψεις. Κάτι τέτοιο μπορεί να οδηγήσει σε βελτιωμένη λήψη αποφάσεων και μεγαλύτερη ικανοποίηση των πελατών. Ταυτόχρονα, με την ανάλυση δεδομένων και τη χρήση προβλεπτικών αλγορίθμων, θα μπορεί να βελτιωθεί η διαχείριση αποθεμάτων, αποφεύγοντας υπερβολικό ή ελλιπές απόθεμα. Αναλύοντας τις τάσεις της αγοράς θα μπορούν οι επιχειρήσεις να προβλέψουν τις μελλοντικές ανάγκες επιτρέποντας να προσαρμόσουν την παραγωγή τους αντίστοιχα.

Η ενσωμάτωση των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στη βιομηχανία αλουμινίου μπορεί να παράσχει λεπτομερή εικόνα της εφοδιαστικής αλυσίδας, βοηθώντας την να παρακολουθεί και να κατανοεί καλύτερα τις διεργασίες της. Η ορθή χρήση των μεγάλων δεδομένων θα μπορεί να βοηθήσει στην εξοικονόμηση πόρων μέσω της αυτοματοποίησης και ελαχιστοποίηση των κινδύνων που ελλοχεύουν, για παράδειγμα των δυσλειτουργιών ή ατελειών των προϊόντων, ή ακόμα και των εργατικών ατυχημάτων. Χρησιμοποιώντας ιστορικά δεδομένα και προγνωστικά μοντέλα, οι εταιρείες θα μπορούν να μειώσουν την κατανάλωση πόρων μειώνοντας έτσι το περιβαλλοντικό αποτύπωμα.

Τελικά, όσον αφορά τη διαχείριση κινδύνων η ανάλυση των δεδομένων από τη διαδικασία παραγωγής μπορεί να αποκαλύψει πολύτιμες πληροφορίες για πιθανούς κινδύνους και να βοηθήσει την εταιρεία να αντιδράσει αναλόγως, για παράδειγμα να έχει αποτελεσματική παραγωγή με βάση τη ζήτηση μειώνοντας τους κινδύνους σπατάλης.

Κεφάλαιο 2 - Βιβλιογραφική Ανασκόπηση

2.1.Ορισμός

Παρά το γεγονός ότι τα μεγάλα δεδομένα είναι ένας δημοφιλής όρος, ο ορισμός του εξακολουθεί να υπόκειται σε σημαντικό βαθμό εννοιολογικής ασάφειας. Ο ανωτέρω όρος χρησιμοποιείται για να περιγράψει μια ποικιλία ιδεών: από την τεχνολογική ικανότητα αποθήκευσης, συλλογής και επεξεργασίας πληροφοριών, έως την πολιτιστική επιρροή που είναι ευρέως διαδεδομένη στις επιχειρήσεις και την κοινωνία, και οι δύο είναι υπερφορτωμένες πληροφορίες. Η έλλειψη ενός επίσημου ορισμού οδήγησε στην εξέλιξη της ερευνητικής προόδου σε πολλαπλές, αντικρουόμενες κατευθύνσεις.

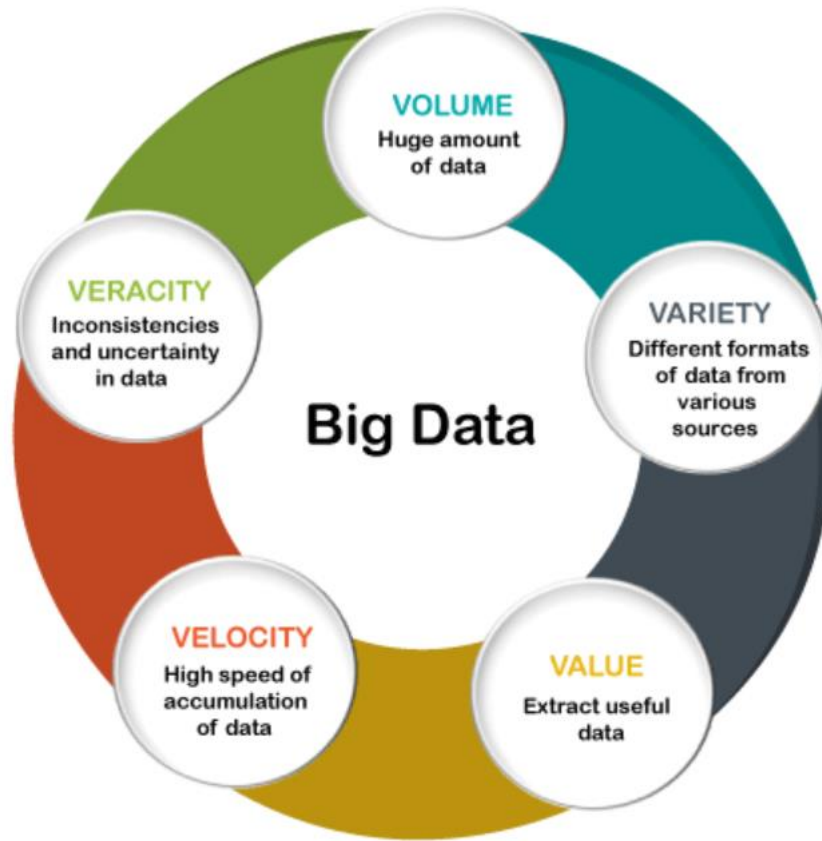
Σήμερα, ο όγκος των δεδομένων αυξάνεται δραματικά από διάφορες πηγές, όπως κάρτες, εφαρμογές για κινητά, μέσα κοινωνικής δικτύωσης και το διαδίκτυο. Όλα τα προαναφερθέντα πράγματα, παράγουν νέα δεδομένα κάθε λεπτό και αθροίζονται σε ένα εκπληκτικό σύνολο. Η IBM υπολογίζει ότι παράγονται 2,5 δισεκατομμύρια byte δεδομένων κάθε μέρα παγκοσμίως.

Για την κατανόηση του όρου (Μεγάλα Δεδομένα), τα δεδομένα ορίζονται ως ο πληθυντικός της λατινικής λέξης «dantum», που σημαίνει «πράγμα που δόθηκε». Τα μεγάλα δεδομένα αναφέρονται σε εξαιρετικά μεγάλα και πολύπλοκα σύνολα δομημένων, ημιδομημένων και αδόμητων δεδομένων που δεν μπορούν να διαχειριστούν ή να επεξεργαστούν εύκολα με τις παραδοσιακές τεχνικές επεξεργασίας δεδομένων. Περιλαμβάνουν δεδομένα μεγάλου όγκου, υψηλής ταχύτητας και ποικιλίας που απαιτούν προηγμένα εργαλεία και τεχνολογίες ανάλυσης για την εξαγωγή πολύτιμων πληροφοριών, μοτίβων και συσχετίσεων.

Ο ορισμός των μεγάλων δεδομένων εξελίσσεται με ρυθμό ανάλογο της ανάπτυξης της πληροφορικής. Για την ακρίβεια, ο Laney, αναλυτής της Gartner ¹(2001) εισήγαγε τον ορισμό ο οποίος αναφέρεται ως 3V: Όγκος (Volume), ταχύτητα (Velocity), ποικιλία Variety. Αργότερα, η Gartner (2015) διεύρυνε τον ορισμό σε 4V, συμπεριλαμβανομένης της αυθεντικότητας (Veracity). Ο ορισμός που εισήγαγε η IBM (2015) συμπεριλαμβάνει ένα πέμπτο V, την αξία (Value). Τα πρόσθετα Vs (4V, 5V και 7V) βασίζονται στην αρχική έννοια των 3V και παρέχουν μια πιο ολοκληρωμένη κατανόηση των προκλήσεων και των ευκαιριών που παρουσιάζουν τα μεγάλα δεδομένα. Αυτές οι έννοιες βοηθούν τους οργανισμούς να τους καθοδηγήσουν στην ανάπτυξη στρατηγικών για τη διαχείριση

¹ Gartner: Εταιρεία τεχνολογίας έρευνας και συμβουλευτικής τεχνολογίας

δεδομένων, την ανάλυση και τη λήψη αποφάσεων στην εποχή της τεράστιας αύξησης και πολυπλοκότητας των δεδομένων.



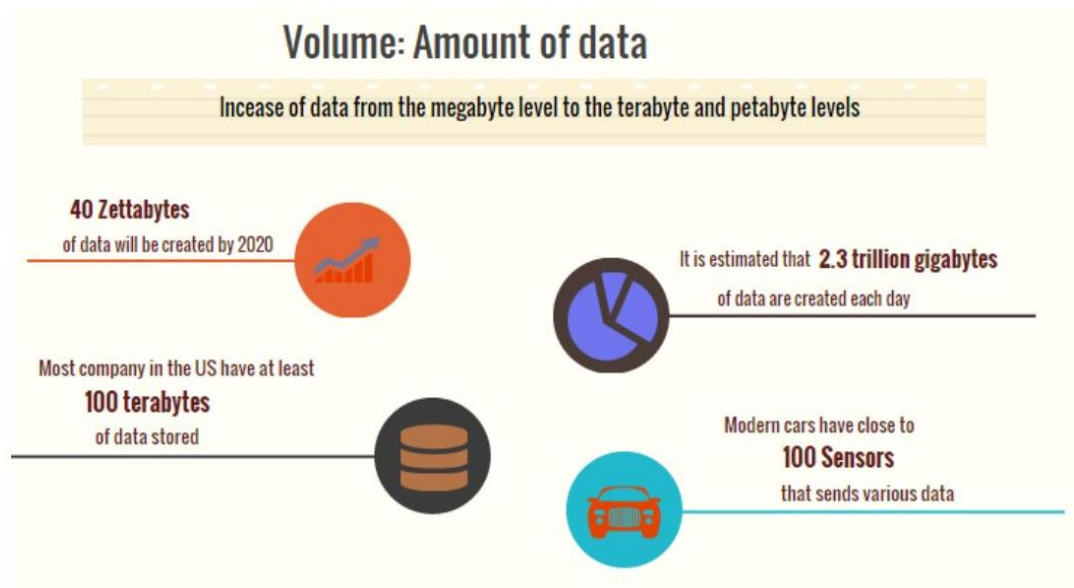
Εικόνα 1. Τα 5V's των "Big Data"

Πηγή: <https://www.javatpoint.com/big-data-characteristics>

Ακολουθεί μια σύντομη περιγραφή των χαρακτηριστικών των μεγάλων δεδομένων που με αυτά ασχολήθηκαν για αρκετά χρόνια πολλοί ερευνητές και εξακολουθούν να δίνουν βάση σε αυτά εμπλουτίζοντάς τα.

Ο όγκος (Volume) αντιπροσωπεύει το τεράστιο μέγεθος διαφόρων συνόλων δεδομένων (Philip Chen & Zhang, 2014). Συνήθως μετριέται σε terabytes, petabytes ή ακόμη και exabytes. Η διαχείριση και η επεξεργασία τέτοιων τεράστιων όγκων δεδομένων απαιτούν εξειδικευμένες τεχνολογίες και κατανομημένα υπολογιστικά συστήματα. Η έννοια του όγκου στο πλαίσιο των μεγάλων δεδομένων αναφέρεται στην απόλυτη κλίμακα ή το μέγεθος των δεδομένων που παράγονται, συλλέγονται και υποβάλλονται σε επεξεργασία. Τα σύνολα μεγάλων δεδομένων είναι συχνά τόσο μεγάλα που υπερβαίνουν τις δυνατότητες αποθήκευσης και επεξεργασίας των παραδοσιακών συστημάτων βάσεων δεδομένων. Αυτό το χαρακτηριστικό είναι ένας από τους καθοριστικούς παράγοντες των μεγάλων δεδομένων

και σχετίζεται στενά με την πρόοδο της τεχνολογίας που επέτρεψε τη δημιουργία και τη συσσώρευση τεράστιων όγκων δεδομένων.



Εικόνα 2. Όγκος δεδομένων (Volume)

Πηγή: : <https://blogs.sap.com/2019/06/24/what-is-big-data-and-why-do-we-need-hadoop-for-big-data/>

Μερικά παραδείγματα πηγών που συμβάλλουν σε μεγάλους όγκους δεδομένων είναι:

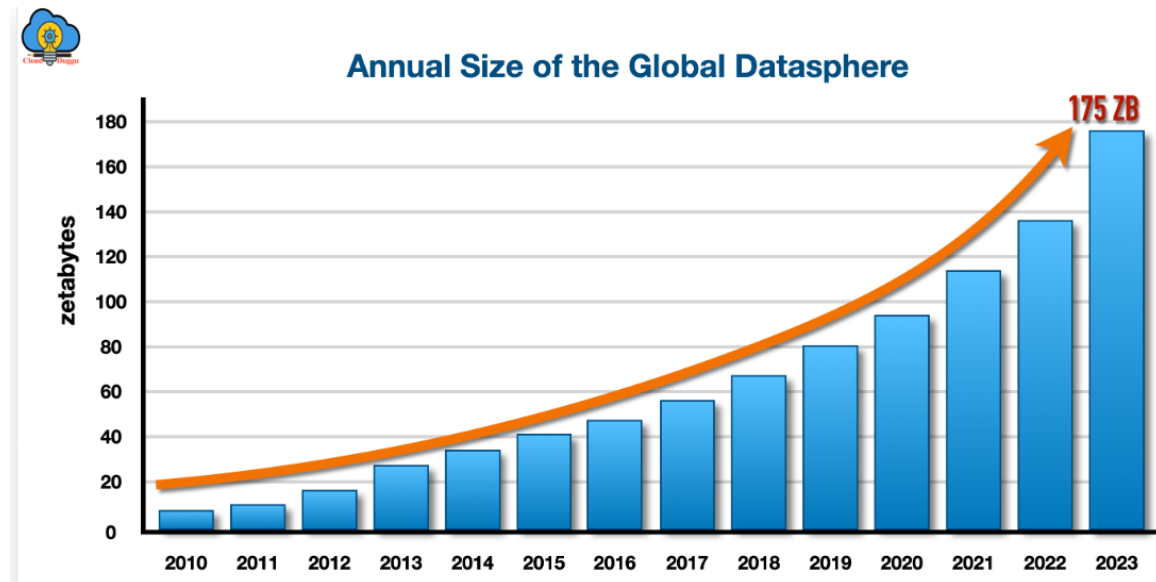
Social Media: Ο μεγάλος όγκος δεδομένων που δημιουργούνται από χρήστες σε πλατφόρμες όπως το Facebook, το Twitter και το Instagram έχει αυξήσει σημαντικά τον όγκο των Big Data. Αυτό περιλαμβάνει αναρτήσεις, σχόλια, likes, shares, εικόνες και βίντεο.

Συσκευές Internet of Things (IoT): Ο πολλαπλασιασμός συσκευών IoT, όπως αισθητήρες, έξυπνες συσκευές και φορητές συσκευές, παράγει μεγάλες ποσότητες δεδομένων σε πραγματικό χρόνο, με αποτέλεσμα την ταχεία ανάπτυξη δεδομένων.

Επιστημονική έρευνα: Πεδία όπως η αστρονομία και η κλιματική επιστήμη δημιουργούν τεράστια σύνολα δεδομένων μέσω πειραμάτων, προσομοιώσεων και παρατηρήσεων.

Ηλεκτρονικό εμπόριο και ηλεκτρονικές συναλλαγές: Οι ηλεκτρονικές αγορές και οι οικονομικές συναλλαγές παράγουν τεράστιες ποσότητες δεδομένων, συμπεριλαμβανομένου του ιστορικού αγορών, της συμπεριφοράς των πελατών και των στοιχείων πληρωμής.

Δεδομένα μηχανών και αισθητήρων: Ο βιομηχανικός εξοπλισμός, τα μηχανήματα και τα οχήματα παράγουν συνεχείς ροές δεδομένων από διάφορους αισθητήρες, συμβάλλοντας σε μεγάλους όγκους δεδομένων.



Εικόνα 3. Ετήσιο Μέγεθος της Παγκόσμιας Σφαίρας Δεδομένων

Πηγή: <https://www.cloudduggu.com/hadoop/big-data/>

Η ταχύτητα (Velocity) αφορά την ταχύτητα κίνησης των δεδομένων (Philip Chen & Zhang, 2014). Η ταχύτητα των μεγάλων δεδομένων είναι ο ρυθμός με τον οποίο παράγονται, συλλέγονται και επεξεργάζονται. Επικεντρώνεται στην ευαισθησία του χρόνου που χρειάζεται για να επεξεργαστούν και να αναλυθούν των δεδομένων, αναφέρει ότι τα δεδομένα πρέπει να χειρίζονται σε πραγματικό χρόνο ή σχεδόν σε πραγματικό χρόνο προκειμένου να έχουμε άμεσα αποτελέσματα και να λαμβάνουμε τεκμηριωμένες αποφάσεις. Οι αυξήσεις στις ταχύτητες μετάδοσης δεδομένων αποδίδονται στις τεχνολογικές εξελίξεις και στον αυξανόμενο αριθμό πηγών δεδομένων που παράγουν ροές δεδομένων. Επιπλέον, περιγράφει τον ρυθμό με τον οποίο παράγονται οι πληροφορίες και την ταχύτητα με την οποία αξιολογούνται και ενεργούνται (Gandomi & Haider, 2015).

Παρακάτω δίνονται ορισμένα παραδείγματα που εξηγούν την έννοια της ταχύτητας των δεδομένων:

Ροές δεδομένων σε πραγματικό χρόνο: Με την άνοδο του διαδικτύου, των συσκευών IoT και των πλατφορμών κοινωνικής δικτύωσης, τα δεδομένα παράγονται και μεταδίδονται

συνεχώς σε πραγματικό χρόνο. Παραδείγματα αποτελούν τα tweets στο Twitter, οι ενδείξεις αισθητήρων από βιομηχανικό εξοπλισμό και οι αλληλεπιδράσεις χρηστών στο διαδίκτυο σε ιστότοπους.

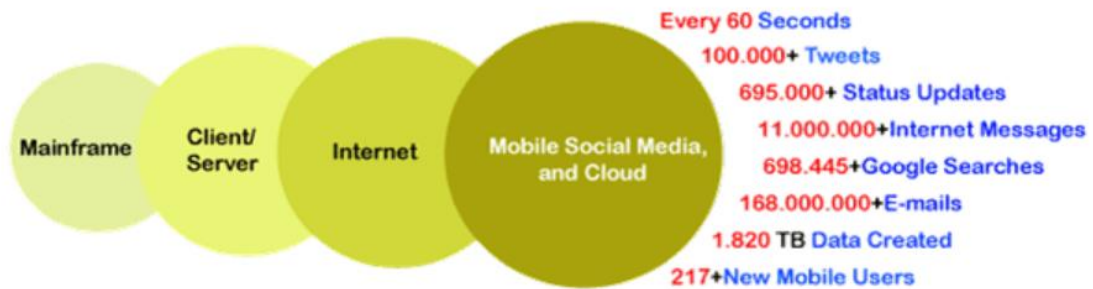
Σημασία στη λήψη αποφάσεων: Η ταχύτητα των δεδομένων είναι ζωτικής σημασίας σε σενάρια όπου απαιτούνται άμεσες γνώσεις για την ταχεία ανταπόκριση σε γεγονότα ή μεταβαλλόμενες συνθήκες. Για παράδειγμα, στις χρηματοοικονομικές συναλλαγές, τα δεδομένα σε πραγματικό χρόνο είναι απαραίτητα για τη λήψη αποφάσεων σε κλάσματα του δευτερολέπτου.

Αγωγοί δεδομένων: Για τον χειρισμό δεδομένων υψηλής ταχύτητας, οι οργανισμοί συχνά δημιουργούν αγωγούς δεδομένων που μεταφέρουν και επεξεργάζονται αποτελεσματικά τα δεδομένα από την πηγή τους στα συστήματα αποθήκευσης και ανάλυσης. Οι εν λόγω αγωγοί διασφαλίζουν ότι τα δεδομένα είναι διαθέσιμα για επεξεργασία το συντομότερο δυνατό.

Προκλήσεις στην επεξεργασία: Οι ροές δεδομένων υψηλής ταχύτητας μπορούν να υπερφορτώσουν τα παραδοσιακά συστήματα επεξεργασίας δεδομένων, οδηγώντας σε συμφόρηση και καθυστερήσεις στην ανάλυση. Για τον χειρισμό δεδομένων με υψηλή ταχύτητα χρησιμοποιούνται εξειδικευμένες τεχνολογίες, όπως πλαίσια επεξεργασίας ροής και εργαλεία ανάλυσης σε πραγματικό χρόνο.

Ανάλυση σε πραγματικό χρόνο: Η ανάλυση δεδομένων σε πραγματικό χρόνο περιλαμβάνει την επεξεργασία και την εξαγωγή συμπερασμάτων από τα δεδομένα καθώς αυτά παράγονται, επιτρέποντας άμεσες αντιδράσεις. Αυτό μπορεί να αποτελέσει πρόκληση λόγω του όγκου και της πολυπλοκότητας των ροών δεδομένων.

Εφαρμογές: Τα δεδομένα υψηλής ταχύτητας είναι απαραίτητα σε διάφορους τομείς, όπως η χρηματοοικονομική (συναλλαγές υψηλής συχνότητας), η υγειονομική περίθαλψη (παρακολούθηση δεδομένων ασθενών σε πραγματικό χρόνο), η εφοδιαστική (παρακολούθηση αποστολών σε πραγματικό χρόνο) και η διαδικτυακή διαφήμιση (υποβολή προσφορών σε πραγματικό χρόνο για διαφημιστικές εμφανίσεις).



Εικόνα 4. Ταχύτητα δεδομένων (Velocity)

Πηγή: www.javatpoint.com/big-data-characteristics

Ο χειρισμός δεδομένων υψηλής ταχύτητας απαιτεί ισχυρή υποδομή, αποδοτικούς αλγορίθμους επεξεργασίας δεδομένων και τη χρήση εξειδικευμένων εργαλείων και τεχνολογιών που έχουν σχεδιαστεί για την ανάλυση δεδομένων σε πραγματικό χρόνο. Επιπλέον, οι οργανισμοί πρέπει να διασφαλίζουν την ποιότητα και την ακρίβεια των δεδομένων, αντιμετωπίζοντας ταχέως μεταβαλλόμενες ροές δεδομένων.

Η ποικιλία (Variety), στο πλαίσιο των μεγάλων δεδομένων, αναφέρεται στο ποικίλο φάσμα τύπων και μορφών δεδομένων που συναντούν οι οργανισμοί και πρέπει να διαχειριστούν και να αναλύσουν. Σε αντίθεση με τις παραδοσιακές πηγές δεδομένων που είναι συνήθως δομημένες και ταιριάζουν με τάξη σε σχεσιακές βάσεις δεδομένων, τα μεγάλα δεδομένα περιλαμβάνουν ένα ευρύ φάσμα τύπων δεδομένων, συμπεριλαμβανομένων δομημένων, ημιδομημένων και μη δομημένων δεδομένων (Hashem et al., 2015).

Δομημένα δεδομένα: Πρόκειται για δεδομένα που είναι οργανωμένα σε ένα προκαθορισμένο και σταθερό σχήμα, με σαφή μορφή για κάθε εγγραφή δεδομένων (Assunção et al., 2015). Τα δομημένα δεδομένα βρίσκονται συνήθως σε σχεσιακές βάσεις δεδομένων και μπορούν εύκολα να ερωτηθούν και να αναλυθούν με τη χρήση SQL (Structured Query Language). Για παράδειγμα περιλαμβάνουν πίνακες με πληροφορίες πελατών, συναλλαγές πωλήσεων και οικονομικές εγγραφές.

Ημιδομημένα δεδομένα: Αυτός ο τύπος δεδομένων δεν έχει σταθερό σχήμα, αλλά περιέχει κάποια οργανωτική δομή, συνήθως με τη μορφή ετικετών ή χαρακτηριστικών. Τα ημιδομημένα δεδομένα είναι πιο ευέλικτα από τα δομημένα δεδομένα και επιτρέπουν ποικίλες μορφές εντός του ίδιου συνόλου δεδομένων. Ορισμένα ημιδομημένα δεδομένα,

περιλαμβάνουν δεδομένα XML (eXtensible Markup Language) και JSON (JavaScript Object Notation) που χρησιμοποιούνται σε εφαρμογές ιστού και API.

Μη δομημένα δεδομένα: Τα μη-δομημένα ή αλλιώς αδόμητα δεδομένα, δεν έχουν προκαθορισμένη δομή και δεν χωρούν εύκολα σε παραδοσιακές βάσεις δεδομένων.

Η αυθεντικότητα (Veracity), αυτό το πρόσθετο V επικεντρώνεται στην ακρίβεια, την αξιοπιστία και την αξιοπιστία των δεδομένων. Με την αύξηση του όγκου και της ποικιλίας των δεδομένων, η διασφάλιση της ποιότητας των δεδομένων καθίσταται κρίσιμη. Τα χαμηλής ποιότητας ή ανακριβή δεδομένα μπορούν να οδηγήσουν σε εσφαλμένα συμπεράσματα και αποφάσεις. Οι διαδικασίες επικύρωσης και εκκαθάρισης δεδομένων είναι απαραίτητες για τη διατήρηση της ειλικρίνειας των δεδομένων.

Ορισμένα παραδείγματα που αφορούν την αυθεντικότητα ώστε να γίνει αντιληπτή η έννοιά της παρατίθενται ακολούθως:

Ποιότητα δεδομένων: Η ποιότητα δεδομένων αναφέρεται στο βαθμό στον οποίο τα δεδομένα είναι ακριβή, αξιόπιστα, συνεπή και σχετικά με τον επιδιωκόμενο σκοπό τους. Τα δεδομένα υψηλής ποιότητας είναι απαλλαγμένα από σφάλματα, ασυνέπειες και επικαλύψεις, διασφαλίζοντας ότι τα αποτελέσματα της ανάλυσης είναι αξιόπιστα.

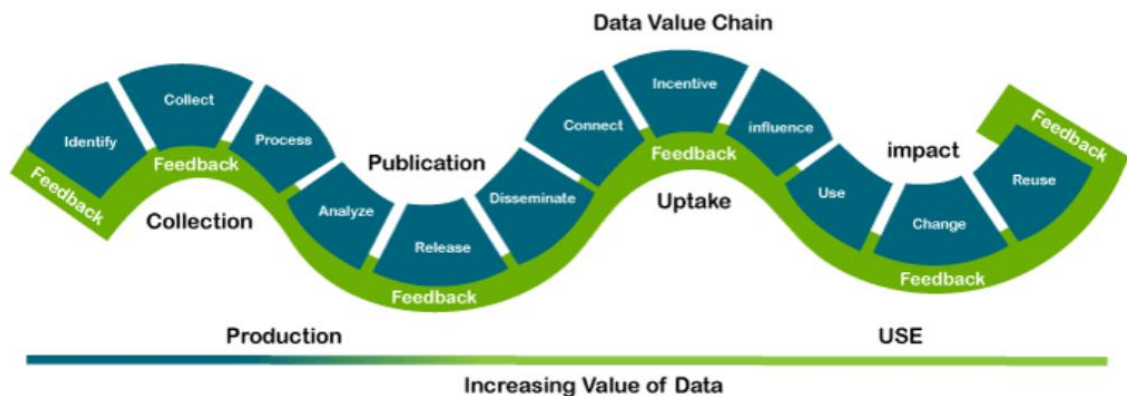
Πηγές αβεβαιότητας δεδομένων: Τα μεγάλα δεδομένα προέρχονται συχνά από διαφορετικές πηγές και κάθε πηγή μπορεί να έχει τα δικά της προβλήματα ποιότητας δεδομένων. Ορισμένες κοινές πηγές αβεβαιότητας δεδομένων περιλαμβάνουν σφάλματα καταχώρησης δεδομένων, ελλείπουσες τιμές, ξεπερασμένες πληροφορίες και ασυνεπή μορφοποίηση.

Επικύρωση δεδομένων: Η επικύρωση δεδομένων είναι η διαδικασία ελέγχου των δεδομένων ως προς την ακρίβεια και τη συνέπεια. Η επικύρωση περιλαμβάνει την επαλήθευση ότι τα δεδομένα πληρούν συγκεκριμένα κριτήρια, τηρούν προκαθορισμένους κανόνες και βρίσκονται εντός αποδεκτών ορίων.

Καθαρισμός δεδομένων: Η εκκαθάριση δεδομένων, περιλαμβάνει τον εντοπισμό και τη διόρθωση σφαλμάτων και ασυνεπειών στα δεδομένα. Περιέχει διαδικασίες όπως η αφαίρεση διπλών εγγραφών, η συμπλήρωση ελλειπόν τιμών και η διόρθωση λανθασμένων καταχωρίσεων.

Η αξία (Value) υπογραμμίζει τη σημασία της άντλησης ουσιαστικών συμπερασμάτων και αξίας από τα μεγάλα δεδομένα. Ο απώτερος στόχος της εργασίας με μεγάλα δεδομένα είναι η εξαγωγή αξιοποιήσιμων πληροφοριών που μπορούν να οδηγήσουν σε βελτιώσεις στις επιχειρηματικές λειτουργίες, στις εμπειρίες των πελατών, στις επιστημονικές ανακαλύψεις

και σε άλλες πτυχές της λήψης αποφάσεων. Η εξαγωγή αξίας από τα μεγάλα δεδομένα περιλαμβάνει προηγμένες τεχνικές ανάλυσης, εξόρυξης δεδομένων και μηχανικής μάθησης.



Εικόνα 5. Αξία (Value)

Πηγή: www.javatpoint.com/big-data-characteristics

Λήψη αποφάσεων με βάση τα δεδομένα: Οι γνώσεις που προκύπτουν από την ανάλυση μεγάλων δεδομένων μπορούν να δώσουν στους οργανισμούς τη δυνατότητα να λαμβάνουν πιο τεκμηριωμένες και καθοδηγούμενες από τα δεδομένα αποφάσεις. Οι αποφάσεις αυτές βασίζονται σε στοιχεία, μοτίβα και τάσεις που ανακαλύπτονται στα δεδομένα, αντί να βασίζονται αποκλειστικά στη διαίσθηση ή σε εμπειρίες του παρελθόντος.

Εξατομίκευση και εμπειρία πελατών: Η ανάλυση μεγάλων δεδομένων επιτρέπει στις επιχειρήσεις να κατανοούν τις προτιμήσεις, τη συμπεριφορά και τις ανάγκες των πελατών σε λεπτομερές επίπεδο. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για την εξατομίκευση των προϊόντων, των υπηρεσιών και των στρατηγικών μάρκετινγκ, με αποτέλεσμα τη βελτίωση της εμπειρίας των πελατών και την αύξηση της αφοσίωσής τους.

Βελτιωμένη επιχειρησιακή αποδοτικότητα: Η ανάλυση μεγάλων δεδομένων μπορεί να οδηγήσει στον εντοπισμό ανεπαρκειών και σημείων συμφόρησης στις επιχειρηματικές διαδικασίες. Βελτιστοποιώντας τις λειτουργίες με βάση τις γνώσεις που βασίζονται σε δεδομένα, οι οργανισμοί μπορούν να βελτιώσουν την παραγωγικότητα, να μειώσουν το κόστος και να ενισχύσουν τη συνολική αποδοτικότητα.

Προγνωστική ανάλυση: Η ανάλυση μεγάλων δεδομένων περιλαμβάνει συχνά την εφαρμογή προγνωστικών αναλύσεων, οι οποίες χρησιμοποιούν ιστορικά δεδομένα για να κάνουν προβλέψεις σχετικά με μελλοντικά γεγονότα ή τάσεις. Τα μοντέλα πρόβλεψης μπορούν να

χρησιμοποιηθούν για διάφορους σκοπούς, όπως η πρόβλεψη της ζήτησης, η προγνωστική συντήρηση και η αξιολόγηση κινδύνων.

Επιστημονικές ανακαλύψεις: Σε ερευνητικούς και επιστημονικούς τομείς, η ανάλυση μεγάλων δεδομένων μπορεί να διευκολύνει σημαντικές ανακαλύψεις και ιδέες.

Ανίχνευση απάτης και ασφάλεια: Η ανάλυση μεγάλων δεδομένων διαδραματίζει κρίσιμο ρόλο στην ανίχνευση απάτης και στην ασφάλεια στον κυβερνοχώρο. Αναλύοντας τεράστιες ποσότητες δεδομένων σε πραγματικό χρόνο, οι οργανισμοί μπορούν να εντοπίζουν ασυνήθιστα μοτίβα και ανωμαλίες που μπορεί να υποδεικνύουν δόλιες δραστηριότητες ή απειλές για την ασφάλεια.

2.2 Ιστορική αναδρομή

Η έννοια των μεγάλων δεδομένων έχει τις ρίζες της στις πρώτες ημέρες της πληροφορικής και στην εμφάνιση των ψηφιακών τεχνολογιών. Αν και η ιστορία του ξεκινά το 1600 η ευρέως χρήση του όρου πραγματικά ξεκίνησε το 1940 και 1950, με την ανάπτυξη στους ηλεκτρονικούς υπολογιστές. Ο «πατέρας της θεωρίας της πληροφορίας» ο Αμερικανός Μαθηματικός Κλοντ Σάνον (Claude Shannon) θεωρείται πως συνέβαλε σημαντικά στο να αποκτήσουμε όλες αυτές τις γνώσεις για τα μεγάλα δεδομένα.

Ωστόσο, απέκτησε σημαντική προβολή και προσοχή στο ξεκίνημα της δεκαετίας 2000 με την εκθετική αύξηση των ψηφιακών δεδομένων και τις εξελίξεις στις τεχνολογίες αποθήκευσης, επεξεργασίας και ανάλυσης δεδομένων. Τη δεκαετία του 1960 και του 1970, η έμφαση δόθηκε κυρίως στην αποθήκευση και τη διαχείριση δεδομένων. Αναπτύχθηκαν σχεσιακές βάσεις δεδομένων και συστήματα διαχείρισης δεδομένων για την αποτελεσματική οργάνωση και ανάκτηση δομημένων δεδομένων.

Στα τέλη της δεκαετίας του 1990 παρατηρήθηκε μια έξαρση στη δημιουργία ψηφιακών δεδομένων λόγω της ανόδου του διαδικτύου, του ηλεκτρονικού εμπορίου και των ψηφιακών τεχνολογιών. Ο πολλαπλασιασμός των δικτυακών τόπων, των πλατφόρμων κοινωνικής δικτύωσης και των ηλεκτρονικών συναλλαγών οδήγησε σε μαζική αύξηση του όγκου των δεδομένων.

Το 2001, ο αναλυτής της Gartner Doug Laney εισήγαγε την έννοια των "μεγάλων δεδομένων" προσδιορίζοντας τα τρία V: όγκος, ταχύτητα και ποικιλία. Ο όρος αυτός απέκτησε ευρεία προσοχή και έγινε σημείο εστίασης για τους ερευνητές και τους εμπειρογνώμονες του κλάδου.

Αν και ο όρος μεγάλα δεδομένα είναι σχετικά νέος συγκριτικά με την πράξη συλλογής και ανάλυσης δεδομένων που υπάρχει εδώ και χρόνια, τα δεδομένα δεν είναι νέα εφεύρεση. Ο άνθρωπος συλλέγει και αποθηκεύει πληροφορίες από τα πρώτα λογιστικά αρχεία στη Μεσοποταμία μέχρι τα πρώτα υπολογιστικά φύλλα και βάσεις δεδομένων. Οι επιχειρήσεις, επίσης, έχουν αντιληφθεί εδώ και αρκετό καιρό ότι μπορούν να αποκομίσουν σημαντικά οφέλη από την εφαρμογή των αναλυτικών στοιχείων. Στο παρελθόν, οι εταιρείες χρησιμοποιούσαν στατιστικούς και αναλυτές για να διευρύνουν σύνολα δεδομένων, αλλά η έκρηξη του όγκου των δεδομένων πια, απαιτεί διαφορετική προσέγγιση. Ο τεράστιος διαθέσιμος όγκος, προσφέρει νέες ευκαιρίες στους οργανισμούς με σκοπό να δημιουργήσουν ανταγωνιστικό πλεονέκτημα.

Οι εξελίξεις στις τεχνολογίες αποθήκευσης, επεξεργασίας και ανάλυσης δεδομένων έπαιξαν καθοριστικό ρόλο στην εξέλιξη των μεγάλων δεδομένων. Τεχνολογίες όπως το Hadoop, οι βάσεις δεδομένων NoSQL, η υπολογιστική νέφους και τα πλαίσια καταμεμημένου υπολογισμού εμφανίστηκαν για να αντιμετωπίσουν τις προκλήσεις που θέτουν τα μεγάλα δεδομένα.

Οι βιομηχανίες άρχισαν να αναγνωρίζουν τη δυνητική αξία των μεγάλων δεδομένων στη λήψη αποφάσεων, στη γνώση των πελατών και στις επιχειρηματικές λειτουργίες. Εταιρείες όπως η Google και η Yahoo ανέπτυξαν κλιμακούμενα συστήματα για την επεξεργασία και ανάλυση δεδομένων μεγάλης κλίμακας, θέτοντας τις βάσεις για την υιοθέτηση των μεγάλων δεδομένων.

Οι αρχές της δεκαετίας του 2010 σηματοδότησαν την επικράτηση των μεγάλων δεδομένων. Πολλοί οργανισμοί σε διάφορους τομείς αγκάλιασαν την ανάλυση μεγάλων δεδομένων για να αποκτήσουν ανταγωνιστικά πλεονεκτήματα, να βελτιώσουν τη λειτουργική αποδοτικότητα και να προσφέρουν εξατομικευμένες εμπειρίες.

Οι τεχνικές ανάλυσης μεγάλων δεδομένων, όπως η μηχανική μάθηση, η εξόρυξη δεδομένων και η προγνωστική ανάλυση, εξελίχθηκαν για να χειρίζονται μεγάλα και ποικίλα σύνολα δεδομένων. Η οπτικοποίηση δεδομένων και τα ταμπλό απέκτησαν επίσης σημασία για την ερμηνεία και την επικοινωνία σύνθετων πληροφοριών μεγάλων δεδομένων.

Τα μεγάλα δεδομένα βρήκαν εφαρμογές σε διάφορους τομείς, όπως η χρηματοδότηση, η υγειονομική περίθαλψη, το λιανικό εμπόριο, η μεταποίηση, οι μεταφορές, οι βιομηχανίες και άλλοι. Έδωσαν τη δυνατότητα στις εταιρείες να λαμβάνουν αποφάσεις βάσει δεδομένων, να βελτιώνουν τις εμπειρίες των πελατών και να αποκαλύπτουν νέες επιχειρηματικές ευκαιρίες.

Τα μεγάλα δεδομένα συνεχίζουν να εξελίσσονται με ταχείς ρυθμούς, λόγω των εξελίξεων στην τεχνητή νοημοσύνη, στο Διαδίκτυο των πραγμάτων (IoT) και τον υπολογισμό των άκρων. Το απόρρητο των δεδομένων, η ασφάλεια και οι δεοντολογικές ανησυχίες έγιναν επίσης κρίσιμα ζητήματα στο τοπίο των μεγάλων δεδομένων.

Η ιστορία των μεγάλων δεδομένων καταδεικνύει τον μετασχηματιστικό αντίκτυπό τους σε διάφορες πτυχές της κοινωνίας και αναδεικνύει τη συνεχιζόμενη ανάπτυξη και σημασία τους στον σημερινό κόσμο των δεδομένων.

2.3 Αναλυτική (Analytics)

Η Αναλυτική θεωρείται η διαδικασία χρήσης λογικής και μαθηματικών για την ερμηνεία δεδομένων προκειμένου να ληφθούν καλύτερες αποφάσεις (Herschel & Davis, 2015). Η ανάλυση μεγάλων δεδομένων χρησιμοποιεί στατιστικές και ποσοτικές μεθόδους ανάλυσης για μεγάλα δεδομένα, αυτές οι μέθοδοι χρησιμοποιούνται για την αξιολόγηση και τη διάδοση πληροφοριών από μεγάλα δεδομένα. Υπάρχουν πολλές μεγάλες μέθοδοι δεδομένων. Η ανάλυση μεγάλων δεδομένων προκύπτει από την εξόρυξη δεδομένων, είναι γνωστή ως ανακάλυψη γνώσης και εντοπίζει νέες ιδέες που απαρτίζονται από αναδυόμενες τάσεις στα δεδομένα. Οι πληροφορίες εξόρυξης δεδομένων μπορεί να έχουν σημαντικό οικονομικό αντίκτυπο, κάτι το οποίο είναι απαραίτητο για τις εταιρείες προκειμένου να αποκτήσουν ανταγωνιστικό πλεονέκτημα.

Η έννοια της αναλυτικής ωριμότητας εκφράζεται με διαφορετικούς τρόπους στη βιβλιογραφία. Οι Lustig et al. (2010) διέκρινε την ανάλυση δομημένων δεδομένων σε τρεις κατηγορίες ανάλυσης: η πρώτη ήταν περιγραφική, περιέγραφε δηλαδή, αυτό που είχε ήδη συμβεί. Η δεύτερη κατηγορία ήταν η προγνωστική, η οποία προέβλεπε τι θα συνέβαινε και η τρίτη κατηγορία, η αναλυτική, όπου υπαγόρευε τι έπρεπε να γίνει (Lustig et al., 2010). Επιπλέον, οι Herschel et al. (2015) ορίζουν τα αναλυτικά διαγνωστικά μελετώντας δεδομένα για τον προσδιορισμό της αιτίας των περιστατικών (Herschel & Davis, 2015). Ωστόσο, οι πρωταρχικές κατηγορίες ανάλυσης θεωρούνται γενικά η περιγραφική ανάλυση και οι προβλέψεις (Sanders, 2014). Σε μια πιο συνοπτική ταξινόμηση, η περιγραφική ανάλυση θεωρείται η βασική μορφή ανάλυσης, ενώ η προγνωστική ανάλυση θεωρείται η προηγμένη μορφή ανάλυσης (Lustig et al., 2010; Sanders, 2014). Τα παραδοσιακά συστήματα πληροφοριών αντιμετωπίζουν τις συναλλαγές ως διακριτές οντότητες και αυτοματοποιούν εργασίες όπως η διαχείριση λογαριασμού ή εγγραφές παραγγελιών, αλλά δεν είναι δυνατή η εξαγωγή δεδομένων σε διαφορετικά επίπεδα συγκέντρωσης. Ως αποτέλεσμα, απαιτούνται

αναλυτικές μέθοδοι για την αξιολόγηση των στοιχείων της επιχείρησης, της αλυσίδας εφοδιασμού και της αγοράς (Sahay & Ranjan, 2008). Η παρακολούθηση της φάσης υλοποίησης μεγάλων βαθμών ωριμότητας της ανάλυσης δεδομένων διασφαλίζει την ποιότητα των δεδομένων, τη βελτιωμένη απόδοση των KPIs (Key Performance Indicators) και την πρόσβαση σε σημαντικές πληροφορίες για τους υπεύθυνους λήψης αποφάσεων (Sanders, 2016). Είναι σημαντικό να σημειωθεί, ωστόσο, ότι δεν απαιτούν όλες οι εταιρείες τον υψηλότερο βαθμό ανάλυσης για να επιτύχουν σημαντικά κέρδη και δεν χρειάζεται να αφιερώνουν πόρους σε τεχνολογία αιχμής για να έχουν επιτυχία (Sanders, 2016).

Τα στοιχεία της ανάλυσης χρησιμοποιούνται για την περιγραφή, τη συγκέντρωση, τη συσχέτιση και τη σύνοψη συνόλων δεδομένων. Για παράδειγμα, ο αριθμός των επιστροφών προϊόντων, οι κλήσεις που ελήφθησαν σε ένα τηλεφωνικό κέντρο ή τα άτομα που αναζήτησαν έναν συγκεκριμένο όρο στη μηχανή αναζήτησης Google. Τα στοιχεία αυτά χρησιμοποιούν ιστορικά δεδομένα για να προσδιορίσουν τις τάσεις και να δημιουργήσουν αναφορές σχετικά με την προηγούμενη συμπεριφορά μιας επιχείρησης (Assunção et al., 2015). Οι περιγραφικές αναλύσεις χρησιμοποιούν απλές στατιστικές και υπολογισμούς (Sanders, 2014). Οι διαγνωστικές αναλύσεις είναι μια μορφή προηγμένης ανάλυσης που περιλαμβάνει τη μελέτη δεδομένων για την κατανόηση της αιτίας του κάτι που συμβαίνει χρησιμοποιώντας μεθόδους όπως η εξερεύνηση δεδομένων, η εξόρυξη δεδομένων και η συσχέτιση (Shi-Nash & Hardoon, 2017).

Η προγνωστική ανάλυση είναι η μελέτη του τρόπου με τον οποίο αναλύονται δεδομένα του παρελθόντος, του παρόντος και του μέλλοντος για την προσπάθεια πρόβλεψης του μέλλοντος (Assunção et al., 2015). Μπορεί να οδηγήσει σε ριζοσπαστικές νέες ιδέες και ευκαιρίες μέσω αυτοματοποιημένων διαδικασιών και άμεσης ανάλυσης δεδομένων (Sanders, 2016). Αυτή είναι μια από τις πιο σημαντικές πτυχές της ανάλυσης μεγάλων δεδομένων (Sanders, 2014). Η προγνωστική ανάλυση χρησιμοποιεί μεθόδους μηχανικής μάθησης και άλλους αλγόριθμους εξόρυξης δεδομένων που είναι υπολογιστικής φύσεως. Αυτές οι στατιστικές μέθοδοι περιλαμβάνουν, για παράδειγμα, ανάλυση χρονοσειρών που εξάγει σημαντικά μοτίβα, μοντέλα λογιστικής και γραμμικής παλινδρόμησης που προβλέπουν την ποιότητα και τις εξαρτήσεις των μεταβλητών (Manyika et al., 2011).

Η προδιαγραφική ανάλυση διευκολύνει τη λήψη αποφάσεων ορίζοντας ενέργειες και αξιολογώντας τον αντίκτυπό τους στις επιχειρηματικές ανάγκες, τους στόχους και τους περιορισμούς (Assunção et al., 2015). Η προδιαγραφική ανάλυση χρησιμοποιεί προβλέψεις που βασίζονται σε δεδομένα και προτείνει ενέργειες που είναι ωφέλιμες ή χρήσιμες για την

επιχείρηση. Επιπλέον, περιλαμβάνει την ανάλυση της μεταβλητότητας στα αναμενόμενα αποτελέσματα, χρησιμοποιώντας ανάλυση σεναρίων και θεωρία παιγνίων (Varela & Tjahjono, 2014). Για παράδειγμα, περιλαμβάνει τη λήψη αποφάσεων πολλαπλών κριτηρίων, τη βελτιστοποίηση και την προσομοίωση (Wang et al., 2016).

Η οπτικοποίηση μεγάλων δεδομένων περιλαμβάνει τεχνικές για τη δημιουργία διαγραμμάτων ή κινούμενων εικόνων με στόχο την επικοινωνία, την κατανόηση και τη βελτίωση των αποτελεσμάτων της ανάλυσης μεγάλων δεδομένων (Manyika et al., 2011). Για εργαλεία που εξετάζουν την ποιότητα και την παρουσίαση δεδομένων λόγω του συνεχώς αυξανόμενου όγκου δεδομένων που υποβάλλονται σε επεξεργασία στα αναλυτικά στοιχεία, είναι σημαντικό να ενεργοποιηθεί η πλοήγηση. Για παράδειγμα, όταν οι εσωτερικές πληροφορίες (όπως πληροφορίες τιμολογίων και συμβολαίων) συνδυάζονται με εξωτερικές πληροφορίες (όπως πληροφορίες αγοράς και ζήτησης), οι διαχειριστές μιας επιχείρησης μπορούν να αποκτήσουν πληροφορίες. Οι διαδραστικοί πίνακες ελέγχου επιτρέπουν την εύκολη παρουσίαση των αναλυτικών στοιχείων σε μορφή οπτικής παρουσίασης (Sanders, 2014).

2.4. Βιβλιογραφική Ανασκόπηση στις βιομηχανίες

2.4.1. Βιομηχανία παραγωγής

Οι Julian Krumeich, Dirk Werth και Peter Loos πρότειναν μια ιδέα για έναν κανονιστικό έλεγχο των επιχειρηματικών διαδικασιών με τη χρήση προβλέψεων διαδικασιών βάσει συμβάντων. Εν προκειμένω, διερεύνησαν νέες δυνατότητες μέσω της εφαρμογής της προγνωστικής ανάλυσης σε μεγάλα δεδομένα, ενώ εστίασαν στον προγραμματισμό και τον έλεγχο της παραγωγής στο πλαίσιο της βιομηχανίας παραγωγής διεργασιών. Αυτός ο τύπος βιομηχανίας αποτελεί κατάλληλο πεδίο εφαρμογής για τη συλλαμβανόμενη έννοια, δεδομένου ότι διαθέτει αρκετά χαρακτηριστικά που αντιτίθενται σε συμβατικές βιομηχανίες. Αυτές οι ιδιαιτερότητες περιλαμβάνουν αποκλίνουσες και κυκλικές ροές υλικών, μεγάλη ποικιλομορφία στις ιδιότητες των τελικών προϊόντων, καθώς και μη γραμμικές διαδικασίες παραγωγής που δεν είναι πλήρως ελεγχόμενες. Με βάση μια μελέτη περίπτωσης, μιας γερμανικής εταιρείας παραγωγής χάλυβα - ένα τυπικό παράδειγμα της βιομηχανίας διεργασιών - περιέγραψαν ποια δεδομένα καθίστανται διαθέσιμα κατά τη χρήση σύγχρονης τεχνολογίας αισθητήρων και παρείχαν έτσι την απαιτούμενη βάση για την υλοποίηση της προτεινόμενης έννοιας. Ωστόσο, η εξέταση του μεγέθους των δεδομένων αποκαλύπτει ότι απαιτούνται ειδικές μέθοδοι ανάλυσης μεγάλων δεδομένων για την

αξιοποίηση του πλήρους δυναμικού αυτών των δεδομένων. Κατά συνέπεια, εξάγουν επτά απαιτήσεις που πρέπει να αντιμετωπιστούν για την επιτυχή εφαρμογή της έννοιας.

- Παροχή κλιμακούμενων μέσων για την επέκταση δικτύων αισθητήρων σε όλες τις διαδικασίες παραγωγής και αποθήκευση των μαζικών δεδομένων σε περιγραφικά μοντέλα διεργασιών και περιβάλλοντος.
- Παροχή μέσων για την ανίχνευση και το φιλτράρισμα σύνθετων συμβάντων μέσα σε τεράστιες ροές δεδομένων αισθητήρων.
- Παροχή δυνατοτήτων για δεδομένα σε πραγματικό χρόνο αναλύσεων για τη συσχέτιση και την ανάλυση συλλογών δεδομένων και ροές που μπορούν να ταξινομηθούν ως "μεγάλες" από την άποψη των υψηλών όγκου, μεγάλης ποικιλίας και υψηλής ταχύτητας.
- Παραγωγή και συνεχής προσαρμογή μοντέλων πρόβλεψης (βάσει γεγονότων).
- Δημιουργία ειδοποιήσεων ως απάντηση σε προβλεπόμενες αποκλίσεις από τους προγραμματισμένους στόχους της διαδικασίας με βάση υπολογισμένες προβλέψεις.
- Παραγωγή συστάσεων και αυτόματων αποφάσεων για ενέργειες μετριασμού.
- Ενεργοποίηση εξελιγμένης προληπτικής διαδικασίας, προσαρμογές βάσει υπολογισμένων συστάσεων και αποφάσεων.

Επιπλέον, προτείνουν μια γενική αρχιτεκτονική των προδιαγραφικών επιχειρηματικών συστημάτων. Η αρχιτεκτονική αυτή περιλαμβάνει πέντε δομικά στοιχεία ενός συστήματος που είναι σε θέση να ανιχνεύει σύνθετα μοτίβα συμβάντων σε ένα περιβάλλον πολλαπλών αισθητήρων, να τα συσχετίζει με ιστορικά δεδομένα και να υπολογίζει προβλέψεις οι οποίες χρησιμοποιούνται τελικά για να προτείνουν την καλύτερη πορεία δράσης κατά την εκτέλεση της διαδικασίας, προκειμένου να ελαχιστοποιηθούν ή να μεγιστοποιηθούν ορισμένοι βασικοί δείκτες απόδοσης (Krumeich et al., 2016).

2.4.2. Βιομηχανία αλουμινίου

Οι κυψέλες αναγωγής αλουμινίου αντιμετωπίζουν συχνά υποβαθμισμένες επιδόσεις που μπορεί να οδηγήσουν σε δραματική αύξηση της κατανάλωσης ενέργειας και να αυξήσουν τις εκπομπές αερίων φθοράνθρακα. Το τεχνητό νευρωνικό δίκτυο και η μηχανική μάθηση μπορούν να διευκολύνουν την κατανόηση και τον εντοπισμό των μοτίβων θορύβου, όπως το κύλισμα των μεταλλικών επιφανειών και ο ήχος βραχυκυκλώματος, τα οποία με τη σειρά τους ειδοποιούν προληπτικά τους χειριστές να λάβουν τα κατάλληλα μέτρα. (Vikas Pandey, n.d.)

Μια άλλη σημαντική χρήση είναι η εκτίμηση της συγκέντρωσης αλουμινίου και της περισσειας $\text{Alf3} \%$ στις κυψέλες. Η διαδικασία περιλαμβάνει την εκτέλεση ανάλυσης διακύμανσης και την εφαρμογή αναλυτικών στοιχείων, τα οποία με τη σειρά τους βοηθούν στην ακριβή εκτίμηση των παραμέτρων και στη λεπτομερή διάγνωση (το νευρωνικό δίκτυο έχει δύο βήματα για αυτή την περίπτωση χρήσης - ανάλυση διακύμανσης και αναλυτικά στοιχεία. Πρώτα εκτελείται ανάλυση διακύμανσης και στη συνέχεια χρησιμοποιείται η αναλυτική για τον προσδιορισμό και την πρόβλεψη των πιθανών αποτελεσμάτων) (Vikas Pandey, n.d.).

Ένα μαθηματικό μοντέλο της διαδικασίας χρησιμοποιείται για τη μοντελοποίηση ενός συστήματος με βάση τη σχέση μεταξύ εισόδου και εξόδου. Η διαφορά μεταξύ των πραγματικών παραμέτρων και των παραμέτρων που μοντελοποιούνται υπολογίζεται και ονομάζεται κατάλοιπο. Αυτά τα κατάλοιπα αξιολογούνται περαιτέρω για την εκτίμηση των βέλτιστων παραμέτρων της διεργασίας (Vikas Pandey, n.d.).

Η προδιαγραφική ανάλυση εφαρμόζεται με τη χρήση αλγορίθμων βελτιστοποίησης και προσομοίωσης για την εκτίμηση των ακριβών παραμέτρων της διεργασίας. Μπορεί να χρησιμοποιηθεί για την ποσοτικοποίηση της επίδρασης στις κυψέλες αναγωγής με τη μεταβολή της συγκέντρωσης αλουμινίου και της περισσειας $\text{Alf3} \%$. Μπορεί να βοηθήσει στη σύσταση των πιθανών αποτελεσμάτων πριν από την πραγματική λήψη αποφάσεων. Δεν προβλέπει μόνο τι θα συμβεί στη θερμοκρασία της κυψέλης, αλλά παρέχει επίσης πληροφορίες σχετικά με το γιατί θα συμβεί και προδιαγράφει τη βέλτιστη θερμοκρασία της κυψέλης και την περισσεια $\text{Alf3} \%$ (Vikas Pandey, n.d.).

Παραδοσιακά, πολλές κρίσιμες λειτουργίες στα χυτήρια αλουμινίου λειτουργούν χειροκίνητα. Αυτές οι λειτουργίες μπορούν να αυτοματοποιηθούν με τη χρήση ρομποτικής και την αυτοματοποίηση της διαδικασίας. Ορισμένοι τομείς στους οποίους μπορεί να εφαρμοστεί η ψηφιοποίηση είναι:

Η αυτόματη έκχυση του καυτού λιωμένου μετάλλου στους κλιβάνους στο χυτήριο μπορεί να βελτιώσει τον έλεγχο της διαδικασίας για χύτευση υψηλότερης ποιότητας και χαμηλότερη κατανάλωση ενέργειας μέσω καλύτερου ελέγχου της θερμοκρασίας και της ταχύτητας έκχυσης.

Τα τηλεχειριζόμενα ρομπότ μπορούν να εκτελούν κρίσιμες μετρήσεις της διαδικασίας στα δοχεία αναγωγής και στους κλιβάνους, καταργώντας την τοποθέτηση ανθρώπων σε επικίνδυνες καταστάσεις.

Τα ψηφιακά εργαλεία έχουν μεταμορφώσει τον τρόπο με τον οποίο ένας πελάτης αγοράζει και μοιράζεται τις εμπειρίες του στη βιομηχανία αλουμινίου. Ψηφιακά κανάλια όπως

ιστότοποι, ιστολόγια και κοινωνικές πλατφόρμες μπορούν να χρησιμοποιηθούν για τη διαδικασία αγοράς και για τη συλλογή ανατροφοδότησης από τους πελάτες. Η ανάλυση των ανατροφοδοτήσεων των πελατών με τη χρήση ψηφιακών εργαλείων επιτρέπει την κατανόηση των προσδοκιών των πελατών, βοηθά στην απόκτηση και διατήρηση πελατών, καθώς και στην πρόβλεψη και υποβολή προτάσεων για τις μελλοντικές ανάγκες του πελάτη - γεγονός που ενισχύσει την ικανοποίησή του .

Το 60% των μεγάλων παραγωγών αλουμινίου θα επενδύσει σε ψηφιακές τεχνολογίες για μπορέσει να διατηρήσει και να βελτιώσει την ποιότητα και το κόστος. Η χρήση ψηφιακών τεχνικών μπορεί να βοηθήσει στον προληπτικό εντοπισμό των επιπτώσεων της ανόδου και στον εντοπισμό των μοτίβων θορύβου, ώστε να ληφθούν προληπτικά διορθωτικά μέτρα. Η ψηφιακή ανάλυση εικόνας μπορεί να διευκολύνει τον εντοπισμό της κατάλληλης σύνθεσης της κρούστας ανόδου που μπορεί να είναι αποτελεσματική στη διατήρηση της σωστής θερμικής ισορροπίας. Σίγουρα θα είναι απαραίτητο για τους παραγωγούς αλουμινίου να επανεκτιμήσουν τη διαδικασία παραγωγής τους και να σχεδιάσουν πώς θα ανοίξουν το δρόμο τους σε μελλοντικές ψηφιακές δεσμεύσεις στη διαδικασία παραγωγής και να αξιοποιήσουν τους ψηφιακούς παράγοντες για συνεχείς επιχειρηματικούς μετασχηματισμούς (Vikas Pandey, n.d.)

Κεφάλαιο 3 - Μεθοδολογία Έρευνας

3.1. Εισαγωγή

Η μεθοδολογία της παρούσας εργασίας επιδιώκει να διευκρινίσει το ρόλο των μεγάλων δεδομένων στην ενίσχυση των διαδικασιών λήψης αποφάσεων στη βιομηχανία αλουμινίου, εστιάζοντας ιδιαίτερα στη βελτιστοποίηση της παραγωγής. Για το σκοπό αυτό έχει επιλεγεί η μελέτη περίπτωσης της χρήσης της μηχανής TANDEM. Η προσέγγιση περιστρέφεται κυρίως γύρω από τη χρήση προηγμένων τεχνικών μηχανικής μάθησης (ML) για την πρόβλεψη της κατάλληλης παραγωγής της μηχανής. Στην ενότητα αυτή περιγράφονται οι υιοθετούμενες μέθοδοι, η διαδικασία συλλογής δεδομένων και η αιτιολόγηση της επιλογής των μοντέλων.

3.1.1 Συλλογή δεδομένων

Το σύνολο των δεδομένων που χρησιμοποιήθηκε αφορά την παραγωγή αλουμινίου από το έτος 2017 έως το 2022 της εταιρείας ELVAL, η οποία αποτελεί τον βασικό πυλώνα του τομέα έλασης αλουμινίου της ομίλου ElvalHalcor και είναι μια από τις κορυφαίες βιομηχανίες πλατέων προϊόντων έλασης αλουμινίου σε παγκόσμιο επίπεδο.

3.1.2. Μεταβλητές

Η παρούσα μελέτη βασίστηκε στην αρχή λειτουργίας ενός τετραπλού θερμού ελάστρου (Tandem) το οποίο παράγει φύλλα αλουμινίου που προκύπτουν από συγκεκριμένες παραγωγικές μεθόδους (θερμή έλαση). Μέχρι πριν λίγο καιρό, η παραγωγική διαδικασία παρόμοιων με το Tandem μηχανών πραγματοποιούνταν αυτοματοποιημένα ως προς την λειτουργικότητα τους, ενώ αντίθετα η παραγωγική διαδικασία επιλέγονταν χειροκίνητα, κυρίως βάσει εμπειρίας και ιστορικότητας των χειριστών ή/και στελεχών των μηχανημάτων των αντίστοιχων βιομηχανιών. Στη σημερινή εποχή, η ραγδαία εξέλιξη της τεχνολογίας εφαρμόζοντας μεθόδους μηχανικής μάθησης πάνω σε μεγάλα δεδομένα. Στο πλαίσιο της παρούσας εργασίας και στη μελέτη περίπτωσης που λεπτομερώς εξετάζεται, έχει επιλεγεί η προσομοίωση των κατάλληλων δεδομένων προκειμένου να εφαρμοστούν οι αντίστοιχες μέθοδοι μηχανικής μάθησης. Είναι προφανές, πως η συλλογή πραγματικών δεδομένων για τη συγκεκριμένη μηχανή Tandem, εμπίπτει στο επιχειρηματικό απόρρητο και ως εκ τούτου είναι αδύνατο να αποκτηθεί. Παρόλα αυτά, η διαδικασία της προσομοίωσης χρησιμοποίησε

πραγματικές ιστορικές τιμές των μεταβλητών που τελικά επιλέχθηκαν προκειμένου τα αποτελέσματα που ανακύπτουν να είναι ρεαλιστικά και να συνάδουν με την αρχή λειτουργίας του μηχανήματος, τουλάχιστον σε ότι έχει να κάνει με την κατεύθυνση (relationship direction) και τη σημαντικότητα (feature importance) των συντελεστών των μεταβλητών.

Το σύνολο δεδομένων που χρησιμοποιήθηκαν προέρχεται από διάφορες διαστάσεις των λειτουργιών της βιομηχανίας αλουμινίου και πιο συγκεκριμένα της χρήσης του μηχανήματος Tandem.

Ειδικότερα, η παρούσα μελέτη περίπτωσης χρησιμοποίησε ως διακριτές μεταβλητές τις εξής:

- Demand (Ζήτηση): Αντιπροσωπεύει την ζήτηση του τελικού προϊόντος σε τόνους ανά ημέρα. Επιλέχθηκε, καθώς θεωρήσαμε πως θα έχει καθοριστικό ρόλο στην διαδικασία της παραγωγής.
- Energy (Κατανάλωση ενέργειας): Αφορά την ενέργεια που καταναλώνεται (GJ/tn ανά προϊόντος), για παράδειγμα το πετρέλαιο κίνησης, το φυσικό αέριο, το υγραέριο, κατά τη διάρκεια της παραγωγικής διαδικασίας. Η κατανάλωση ενέργειας είναι μια σημαντική μεταβλητή που έχει τη δυνατότητα να επηρεάσει την παραγωγική διαδικασία.
- Logistics (Αποθήκευση): Αναφέρεται στην διαχείριση προϊόντων κατά την αποθήκευσή και τη διακίνησή τους (tn/ ημέρα). Μία ιδιαιτέρως σημαντική μεταβλητή για κάθε εταιρεία παραγωγής.
- SR (Severity Rate, Δείκτης σοβαρότητας συμβάντων): Ο δείκτης σοβαρότητας αναφέρεται σε κάθε τραυματισμό ή επαγγελματική ασθένεια, η οποία έχει ως αποτέλεσμα ο εργαζόμενος να μην μπορεί να εργαστεί στη θέση εργασίας του, καθώς και σε οποιαδήποτε άλλη ουσιαστική θέση κατά τη διάρκεια αποθεραπείας του. Χρησιμοποιείται κυρίως σε μεγάλες επιχειρήσεις που έχουν αριθμό περιστατικών απώλειας ημερών εργασίας και υπολογίζονται ως εξής:

$$SR = \frac{d}{e}$$

όπου, d ο αριθμός των ημερών εργασίας που χάθηκαν και e ο αριθμός των συμβάντων.

- Η εξαρτημένη μεταβλητή που επιλέχθηκε ως μέτρο ανάδειξης της σπουδαιότητας των μεγάλων δεδομένων είναι η παραγωγή (production) αλουμινίου από τη μηχανή Tandem σε τόνους ανά ημέρα.

3.1.2. Προεπεξεργασία δεδομένων

Πριν από την εισαγωγή των δεδομένων στα μοντέλα ML, πραγματοποιήθηκαν βήματα προεπεξεργασίας:

- I. Χειρισμός ελλিপών δεδομένων (Missing values): Οποιοσδήποτε ελλείπουσες τιμές αντιμετωπίστηκαν είτε με υπολογισμό, χρησιμοποιώντας μεθόδους όπως η αντικατάσταση του μέσου όρου ή της διάμεσης, είτε με διαγραφή των σχετικών γραμμών, ανάλογα με το ποσοστό των ελλিপών δεδομένων.
- II. Κανονικοποίηση δεδομένων (z-transformation): Το σύνολο των συνεχών μεταβλητών, κανονικοποιήθηκαν ώστε να διασφαλιστεί ότι έχουν παρόμοια κλίμακα, βοηθώντας έτσι τη σύγκλιση του μοντέλου και την ακρίβεια πρόβλεψης.
- III. Μηχανική χαρακτηριστικών (Feature Engineering): Νέες μεταβλητές προέκυψαν από τα υπάρχοντα δεδομένα για να προσφέρουν πιο διαφοροποιημένες γνώσεις, όπως ο ρυθμός μεταβολής της ζήτησης ή τα πρότυπα κατανάλωσης ενέργειας.

Οι βασικές μέθοδοι μηχανικής μάθησης έρχονται όχι μόνο να συμπληρώσουν τις συμβατικές μεθόδους αλλά και να τις αντικαταστήσουν με εντυπωσιακά αποτελέσματα. Η γραμμική παλινδρόμηση αν και ευρέως κατανοητή και γνωστή δεν δίνει πάντα όλες τις απαραίτητες πληροφορίες και τις πιο ακριβείς προβλέψεις. Κάποιες μέθοδοι μηχανικής μάθησης χρησιμοποιούν δέντρα αποφάσεων (decision trees) για τη διαίρεση των δεδομένων σε κατάλληλα υποσύνολα μεταβλητών, τα οποία με τη σειρά τους προκύπτουν από τα δέντρα αποφάσεων και αξιοποιούν τη γνώση πολλών επιμέρους προβλέψεων. Εν συνεχεία, η μέθοδος boosting συνδυάζει τις αδύναμες μικρότερες «εκπαιδευμένες» προβλέψεις σε κάποιον πιο ισχυρό παράγοντα.

Οι μέθοδοι που βασίζονται σε trees and forests (δέντρα και δάση) βαθμολογούν (score) τη σημασία των χαρακτηριστικών δεδομένων και έτσι αντισταθμίζουν την αδιαφάνεια των μαύρων κουτιών (black boxes). Τέλος, οι μέθοδοι μηχανικής μάθησης σε μεγάλα διαστρωματικά δεδομένα αναδεικνύουν την υπεροχή των big data έναντι των αλγορίθμων και την σοφή διατήρηση όλων των ακραίων τιμών. Όλα αυτά έχουν ως αποτέλεσμα τη

βελτίωση της προγνωστικής ανάλυσης και ακρίβειας των υποδειγμάτων που ξεπερνούν τις απλές μεθόδους της γραμμικής παλινδρόμησης.

3.2. Στατιστικές και ML προσεγγίσεις

Η γραμμική παλινδρόμηση είναι μακράν η πιο δημοφιλής μέθοδος για την αξιολόγηση των δεδομένων πίνακα. Η κυρίαρχη στατιστική προσέγγιση (στατιστική κουλτούρα) που οδηγεί σε αυτήν την προσέγγιση υποθέτει ότι τα δεδομένα προέρχονται από έναν συγκεκριμένο τύπο στοχαστικού μοντέλου (Breiman, 2001). Η μηχανική μάθηση αντιπροσωπεύει έναν αλγόριθμο που μπορεί να ανταγωνιστεί τις στατιστικές μεθόδους (Breiman, 2001). Κάνοντας υποθέσεις σχετικά με τη διανομή και τη δημιουργία δεδομένων ανοίγει η πόρτα σε αλγόριθμους που υπερβαίνουν τις γενικευμένες γραμμικές μεθόδους (Breiman, 2001). Αυτή η αλγοριθμική προσέγγιση επιδιώκει να επιτύχει μεγαλύτερη ακρίβεια και βαθύτερη κατανόηση των δεδομένων σε οποιαδήποτε κλίμακα.

Το θεώρημα του no-free-lunch υποστηρίζει ότι είναι αδύνατο να γνωρίζουμε εκ των προτέρων ποιο μοντέλο μηχανικής μάθησης ταιριάζει καλύτερα σε ένα συγκεκριμένο σύνολο δεδομένων (Wolpert, 1996). Κατά συνέπεια, η πιο πρακτική προσέγγιση έγκειται στην εφαρμογή όσο το δυνατόν περισσότερων μεθόδων. Αν και η επιστημονική κοινότητα άργησε να αποδεχθεί τη μηχανική μάθηση, τα τελευταία χρόνια οι ερευνητές και οι αναλυτές έχουν την ευκαιρία να επιλέξουν ελεύθερα τις κατάλληλες μεθόδους ανεξάρτητα σε ποια πλευρά μεταξύ των συμβατικών στατιστικών και αλγοριθμικών μεθόδων ανήκουν (Athey & Imbens, 2019).

Επειδή οι εναλλακτικές λύσεις μηχανικής μάθησης αντί της γραμμικής παλινδρόμησης εφαρμόζονται τόσο εύκολα, η συνδυαστική επιλογή στατιστικών και αλγοριθμικών μεθόδων γίνεται ακόμη πιο πειστική και απαραίτητη.

3.3. Δενδρολογικές Μέθοδοι: Decision Trees and Forests

Ο αλγόριθμος ταξινόμησης και παλινδρόμησης (CART – Classification And Regression Tree) υποστηρίζει έναν εντυπωσιακό καταγισμό μεθόδων (Breiman et al., 1984; Loh, 2008). Τα δέντρα απόφασης και τα στοχαστικά τους σύνολα (ή τα δάση) είναι γνωστά με το όνομα, δενδρολογική μηχανική μάθηση (dendrological machine learning). Οι προβλέψεις τους δεν είναι αυστηρά γραμμικές γι' αυτόν τον λόγο η δενδρολογική μηχανική μάθηση συχνά ξεπερνά τη γραμμική παλινδρόμηση. Όλοι οι δενδρολογικοί αλγόριθμοι είναι ισχυροί ακόμη και με την παρουσία ακραίων τιμών. Επιπλέον, αυτοί οι αλγόριθμοι εμφανίζονται αρκετά ανθεκτικοί ακόμη και για μη προκαθορισμένα μοντέλα. Η ενσωμάτωση ασθενώς

στατιστικά σημαντικών ή ακόμα και μη σημαντικών μεταβλητών σπάνια αποδυναμώνει ένα δέντρο απόφασης ή ένα σύνολο δασών (J. M. Chen, 2021).

3.4. Διαχωρισμός και Μετασχηματισμός Δεδομένων (Splitting and Scalling)

Η εποπτευόμενη (supervised) μηχανική μάθηση απαιτεί τον διαχωρισμό των δεδομένων σε τυχαία υποσύνολα για εκπαίδευση (training) και δοκιμή (testing). Αυτή η πρακτική, σπάνια, διασφαλίζει ότι η μηχανική μάθηση δεν απομνημονεύει απλώς τιμές που σχετίζονται με δεδομένα που πρόκειται να συμμετάσχουν στη διαδικασία πρόβλεψης (Muller & Guido, 2017). Η διατήρηση του 1/e για δοκιμή των δεδομένων διασφαλίζει τη γενίκευση οποιασδήποτε εποπτευόμενης μεθόδου μάθησης σε δεδομένα που δεν εμφανίζονται κατά τη διάρκεια της εκπαίδευσης (Muller & Guido, 2017).

Πολλοί αλγόριθμοι μηχανικής μάθησης έχουν μεγαλύτερη ακρίβεια σε μετασχηματισμένα (scaled) δεδομένα (Muller & Guido, 2017). Ο τυπικός μετασχηματισμός διασφαλίζει ότι η μηχανική μάθηση αξιολογεί όλες τις μεταβλητές και αναφέρει τα αποτελέσματα με όρους Gaussian z-scores. Ουσιαστικά, η κλιμάκωση των δεδομένων δοκιμής (test) πρέπει να προχωρήσει σύμφωνα με την κατανομή των τιμών στα δεδομένα εκπαίδευσης (train) προκειμένου να αποτραπεί η διαρροή δεδομένων (Muller & Guido, 2017).

Τα δέντρα απόφασης και τα δάση ποσοτικοποιούν τη συμβολή κάθε προγνωστικής μεταβλητής. Οι μέθοδοι που βασίζονται σε δέντρα απόφασης από τις βιβλιοθήκες SciKit-Learn παρουσιάζουν τη σημαντικότητα των μεταβλητών ως ένα διάνυσμα τιμών που αναφέρει τη συμβολή κάθε ανεξάρτητης μεταβλητής στις προβλέψεις του μοντέλου (Géron, 2019). Το διάνυσμα με τη στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών αντιπροσωπεύει “έναν σταθμισμένο μέσο όρο, όπου το βάρος κάθε κόμβου”, σε ένα δέντρο απόφασης ή σε όλα τα δέντρα του δάσους, “είναι ίσο με τον αριθμό των δειγμάτων εκπαίδευσης που σχετίζονται με αυτό”, (Géron, 2019). Όπως κάθε άλλο διάνυσμα πιθανοτήτων, το άθροισμά τους είναι πάντα 1.

Οι σπουδαιότητες των ανεξάρτητων μεταβλητών (feature importances) μοιάζουν περισσότερο με τους τυποποιημένους συντελεστές παλινδρόμησης (ή με άλλα λόγια τους συντελεστές βήτα) (Newman & Browner, 1991), η χρήση των οποίων στην στατιστική συμπερασματολογία είναι από μόνη της αμφιλεγόμενη (Greenland et al., 1986). Η σημασία των χαρακτηριστικών διαφέρει κατά πολύ σημαντικό τρόπο. Ενώ οι συντελεστές βήτα μπορεί να είναι θετικοί, αρνητικοί ή μηδενικοί, η σημασία των χαρακτηριστικών είναι πάντα μη αρνητική. Κατά συνέπεια, δεν μεταφέρουν πληροφορίες σχετικά με τη θετική ή αρνητική

συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών και της μεταβλητής στόχου (Agiropoulos Charalampos L., 2022).

3.5. Η Αντιστάθμιση Μεροληψίας-Διακύμανσης (Bias-Variance Tradeoff)

Η αντιστάθμιση μεροληψίας-διακύμανσης χρησιμεύει ως προοίμιο για μια εξερεύνηση τρόπων για τη βελτίωση της ακρίβειας του αλγορίθμου CART. Η ένταση μεταξύ της μεροληψίας και της διακύμανσης προκύπτει από μια εγγενή ιδιότητα της εποπτευόμενης μηχανικής μάθησης (Agiropoulos Charalampos L., 2022). Η μεγαλύτερη ανακρίβεια ή μεροληψία στις εκτιμήσεις των παραμέτρων του μοντέλου μπορεί να μειώσει τη διακύμανση μεταξύ των εκτιμήσεων παραμέτρων στα δείγματα (Kohavi & Wolpert, 1996). Η αντιστάθμιση της μεροληψίας είναι το κλειδί για την εφαρμογή των μεθόδων μηχανικής μάθησης πέρα από τα δεδομένα στα οποία έχουν εκπαιδευτεί αυτοί οι αλγόριθμοι (Geman et al., 1992). Δεδομένου ότι «είναι αδύνατο να επιτευχθεί ταυτόχρονα η χαμηλότερη δυνατή διακύμανση και μεροληψία», η πρόκληση είναι να δημιουργηθεί ένα μοντέλο με (λογικά) χαμηλή διακύμανση και χαμηλή μεροληψία, ως η προσέγγιση που μπορεί να γενικευτεί και να εφαρμοστεί επαρκώς σε εξωτερικά σύνολα δεδομένων (Dankers et al., 2018).

Η μεροληψία αναφέρεται στη συνολική ακρίβεια μιας μεθόδου, ιδιαίτερα στο υποσύνολο που αφορά την εκπαίδευση του μοντέλου. Ωστόσο, η υπερβάλλουσα μεροληψία οδηγεί σε ένα μοντέλο που δεν μπορεί να προσαρμοστεί (under-fit) καλά στα δεδομένα του. Όπως συμβαίνει συχνά με τις πολυωνμικές παραλλαγές γενικευμένων γραμμικών μεθόδων, τα μοντέλα υψηλής ακρίβειας δεν παρέχουν αξιόπιστα αποτελέσματα εκτός εάν γενικεύονται καλά σε νέα δεδομένα. Απ' την άλλη μεριά, τα μοντέλα υψηλής διακύμανσης τείνουν να υπερπροσαρμόζονται (over-fit) στο υποσύνολο που αφορά την εκπαίδευση του μοντέλου. Επομένως, η διακύμανση επηρεάζει τη γενίκευση και τη συνέπεια των αποτελεσμάτων σε νέα δεδομένα. Με τη βέλτιστη πολυπλοκότητα, ένα μοντέλο επιτυγχάνει την ιδανική ισορροπία μεταξύ υποπροσαρμογής και υπερπροσαρμογής δεδομένων (Agiropoulos Charalampos L., 2022).

3.6. Μέθοδοι Ενίσχυσης και Ομαδοποίησης (Boosting and Ensemble Methods)

Bagging and Pasting

Ο απλούστερος τρόπος διαφοροποίησης των αποτελεσμάτων ενός δέντρου αποφάσεων είναι η δειγματοληψία των υποσυνόλων εκπαίδευσης (training instances), με ή χωρίς

επανατοποθέτηση. Ο όρος bagging, αποτελεί τη συντομογραφία του όρου bootstrap aggregation (Breiman, 1996) για δείγματα με επανατοποθέτηση. Επειδή το $1/e$ οποιουδήποτε συνόλου δεδομένων θα διαφύγει από τη δειγματοληψία, ακόμη και αν ληφθεί άπειρος αριθμός δειγμάτων (Geron, 2019), τα υποσύνολα εκπαίδευσης (out-of-bag subsets) που δεν επιλέχθηκαν στο bagging παρέχει πρόσθετη επικύρωση της γενίκευσης ενός δέντρου αποφάσεων σε δεδομένα που δεν είχαν παρατηρηθεί προηγουμένως.

Random Forests και Extra Trees

Οι μέθοδοι ενίσχυσης και ομαδοποίησης λαμβάνουν υπόψη πολλά δέντρα αποφάσεων. Αυτή η ευρεία και ποικιλόμορφη κατηγορία αλγορίθμων περιλαμβάνει τις εξής μεθόδους: random forests, extremely randomized trees (extra trees), adaptive boosting (AdaBoost), gradient boosting, light gradient boosting machine (LightGBM) and extreme gradient boosting (XGBoost). Όλες αυτές οι μέθοδοι χρησιμοποιούν την ίδια σύνταξη στο περιβάλλον προγραμματισμού εφαρμογών του SciKit-Learn. Ο κώδικας για μία μέθοδο ισχύει, με μικρές τροποποιήσεις και για όλους τους άλλους.

Η random forests αποτελεί την απλούστερη μέθοδο των παραπάνω παραλλαγών (Ho, 1995). Έτσι, με αυτή τη μέθοδο αντί να αναζητείται το καλύτερο χαρακτηριστικό κατά τον διαχωρισμό ενός κόμβου, η random forests αναζητεί το καλύτερο χαρακτηριστικό μέσα σε ένα τυχαίο υποσύνολο. Για την εφαρμογή της απαιτείται ο συντονισμός μόνο δύο υπερπαραμέτρων: ο μέγιστος αριθμός χαρακτηριστικών σε ένα τυχαίο δέντρο, μαζί με το μέγιστο βάθος (maximum depth) κάθε δέντρου. Η τυχαιοποίηση των ορίων για κάθε χαρακτηριστικό, σε συνδυασμό με την αναζήτηση του βέλτιστου ορίου, αποδίδει έναν ακόμη πιο στοχαστικό αλγόριθμο που ονομάζεται random trees ή extra trees (Geurts et al., 2006).

Adaptive και Gradient Boosting

Οι μέθοδοι Boosting αντιπροσωπεύουν μια ειδική κατηγορία ομαδοποίησης που συνδυάζει τα αδύναμα (weak) με τα δυνατά (strong) σύνολα εκπαίδευσης (Drucker & Cortes, 1996). Κάθε βήμα στη διαδοχική κατάρτιση των προβλέψεων επιδιώκει να διορθώσει λάθη που έγιναν από τον προκάτοχό του (Geron, 2019). Ο αλγόριθμος AdaBoost βασίζεται σε δέντρα αποφάσεων που περικόπτονται (decision stumps) μετά από ένα διαχωρισμό (Freud & Schapire, 1997). Μετά από κάθε περίπτωση εκπαίδευσης, το AdaBoost ενημερώνει τα βάρη (weights) για κάθε προγνωστικό παράγοντα (Freud & Schapire, 1997). Η διαδοχική μάθηση καθιστά δύσκολη την εφαρμογή του AdaBoost μέσω παράλληλων υπολογιστών (parallel

computing) και την κλιμάκωσή του σε μεγαλύτερα σύνολα δεδομένων (Geron, 2019). Ο αλγόριθμος ενίσχυσης gradient (gradient boosting algorithm) προσθέτει επίσης προγνωστικούς παράγοντες διαδοχικά σε ένα σύνολο. Αντί να προσαρμόζονται τα βάρη για κάθε περίπτωση, όπως κάνει το AdaBoost, η ενίσχυση κλίσης προσαρμόζει κάθε νέο προγνωστικό παράγοντα στα υπολειπόμενα σφάλματα (Breiman, n.d.-b, n.d.-a; Friedman, 2002). Οι υπερπαράμετροι στην ενίσχυση της κλίσης (gradient boosting) ελέγχουν τον ρυθμό εκμάθησης του συνόλου, καθώς και το βάθος και την ανάπτυξη των δέντρων αποφάσεων εντός του συνόλου (Geron, 2019).

Η μηχανική μάθησης προσφέρει πολλές παραλλαγές gradient boosting. Το XGBoost ξεπερνά τα όρια στην ταχύτητα και την επεκτασιμότητα που έχουν ταλαιπωρήσει άλλους boosting algorithms (T. Chen & Guestrin, 2016). Η εκπαίδευση σε τυχαία υποδείγματα αποφέρει στοχαστική σταδιακή ενίσχυση, η οποία ανταλλάσσει υψηλότερη μεροληψία με χαμηλότερη διακύμανση και ταχύτερη εκπαίδευση (Friedman, 2002). Το Light Gradient Boosting Machine. διαφέρει σημαντικά από το μοντέλο XGBoost, χρησιμοποιώντας μια προσέγγιση ανάπτυξης κατά φύλλα με περιορισμούς βάθους. Αξιοποιεί τεχνικές που βασίζονται σε ιστογράμματα για να επιταχύνει την εκπαίδευση, να ελαχιστοποιήσει την κατανάλωση μνήμης και να το πράξει. Είναι δυνατή η ελαχιστοποίηση της χρήσης μνήμης στο 1/8 της αρχικής με τη χρήση της προσέγγισης ιστογράμματος, η οποία δεν χρειάζεται πρόσθετη αποθήκευση για τα προ-ταξινομημένα αποτελέσματα και μπορεί επίσης να διατηρήσει την τιμή μετά τη διακριτοποίηση των χαρακτηριστικών.

Ανάμειξη (Blending)

Το blending είναι η τεχνική ensemble, η οποία χρησιμοποιεί πολλαπλούς αλγορίθμους μηχανικής μάθησης για εκπαίδευση στο ίδιο σύνολο δεδομένων. Εδώ, σε αντίθεση με άλλες τεχνικές ensemble, χρησιμοποιούνται δύο στρώματα αλγορίθμων μηχανικής μάθησης, το ένα είναι τα βασικά μοντέλα και το δεύτερο στρώμα αλγορίθμων είναι γνωστό ως μετα-μοντέλα.

Τα μοντέλα βάσης εκπαιδεύονται στο σύνολο δεδομένων εκπαίδευσης που παρέχεται πρώτα, αφού τα μοντέλα βάσης εκπαιδευτούν στο σύνολο δεδομένων, το σύνολο δοκιμής τροφοδοτείται στα μοντέλα βάσης για πρόβλεψη και οι προβλεπόμενες τιμές από τα μοντέλα βάσης χρησιμοποιούνται για την εκπαίδευση του μετα-μοντέλου. Στη συνέχεια, τα μετα-μοντέλα εκπαιδεύονται στα δεδομένα πρόβλεψης, μετά την εκπαίδευση το μετα-μοντέλο αναθέτει τα βάρη στα μοντέλα βάσης και η έξοδος από τα μετα-μοντέλα θεωρείται η τελική

έξοδος. Το σύνολο δοκιμών που λαμβάνεται στο πρώτο βήμα μπορεί να χρησιμοποιηθεί για την αξιολόγηση του μοντέλου ανάμειξης (Ganaie et al., 2022).

Κεφάλαιο 4 – Ανάλυση και Αποτελέσματα

4.1. Εισαγωγή

Στοχεύοντας στην καλύτερη κατανόηση των μεθόδων που διεξοδικά αναφέρονται στην Ενότητα 3 κρίνεται σκόπιμη η αναλυτική παρουσίαση της μελέτης περίπτωσης της βιομηχανίας αλουμινίου Elval. Το συγκεκριμένο παράδειγμα, έρχεται να αναδείξει τη χρησιμότητα και τη σημαντικότητα των μεγάλων δεδομένων και των σύγχρονων τεχνικών μηχανικής μάθησης, οι οποίες δύναται να εφαρμοστούν για την προγνωστική ανάλυση. Όπως έχει αναφερθεί προηγούμενα, τα μοντέλα Machine Learning (ML) επιτυγχάνουν υψηλή ακρίβεια πρόβλεψης με σχεδόν κάθε τύπο ποσοτικών δεδομένων ανεξαρτήτως θέματος και ταξινομούν αποτελεσματικά διαφορών τύπων δεδομένα.

Στην παρούσα ενότητα θα πραγματοποιηθεί η αναλυτική παρουσίαση της περίπτωσης του μηχανήματος Tandem της βιομηχανίας αλουμινίου και μέσω αυτής θα αναδειχθεί η σπουδαιότητα των μεγάλων δεδομένων και της προγνωστικής ανάλυσης στην παραγωγή και έλαση αλουμινίου.

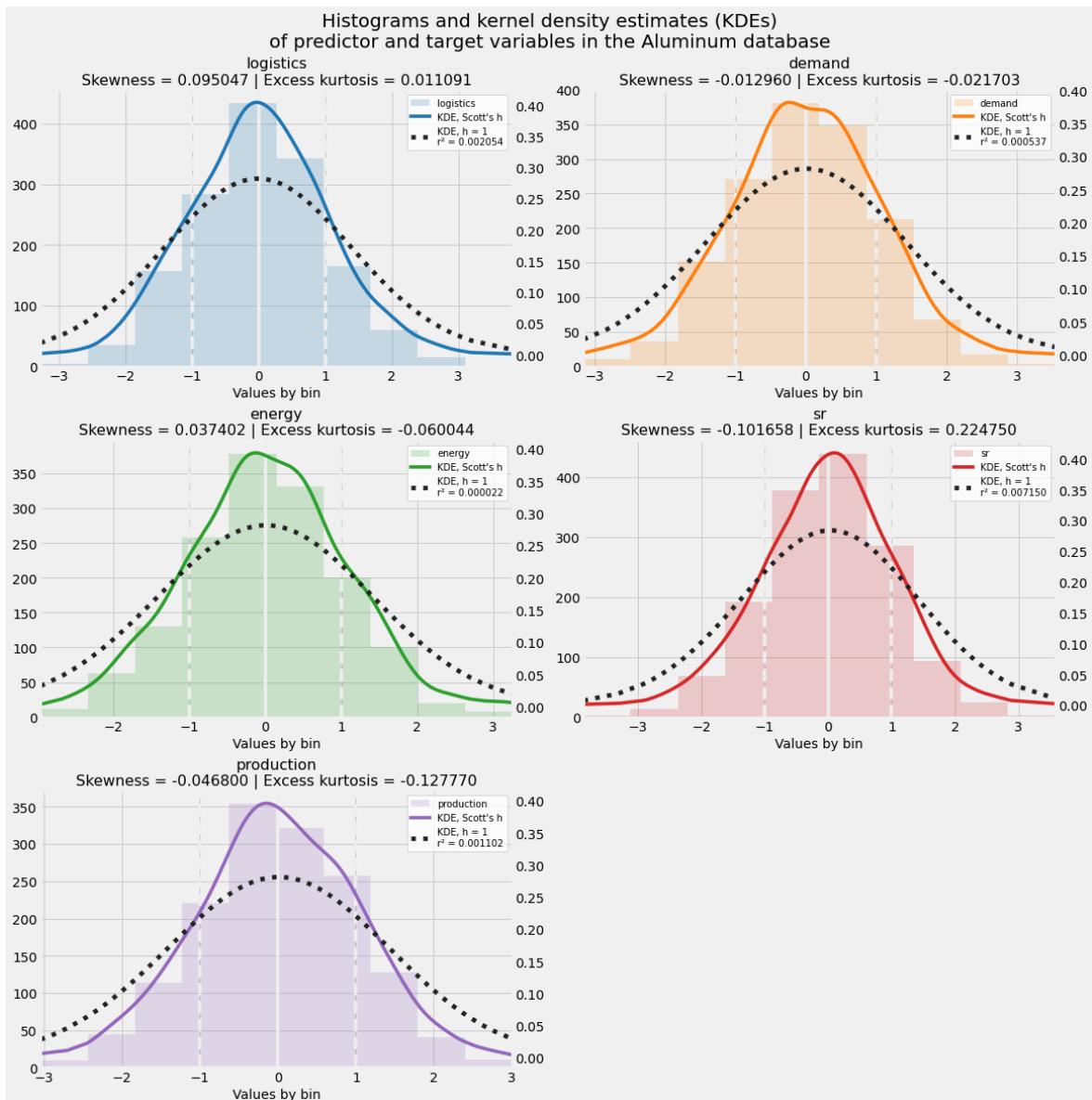
4.2. Ανάλυση Δεδομένων

Ο πίνακας 4.1. συνοψίζει τα περιγραφικά μέτρα των προαναφερθέντων μεταβλητών, όπως η μέση τιμή κάθε μεταβλητής, η τυπική απόκλιση, η διάμεσος, καθώς και τα ποσοστημόρια Q1 και Q3.

Πίνακας 4.1. Περιγραφικά μέτρα των 5 μεταβλητών

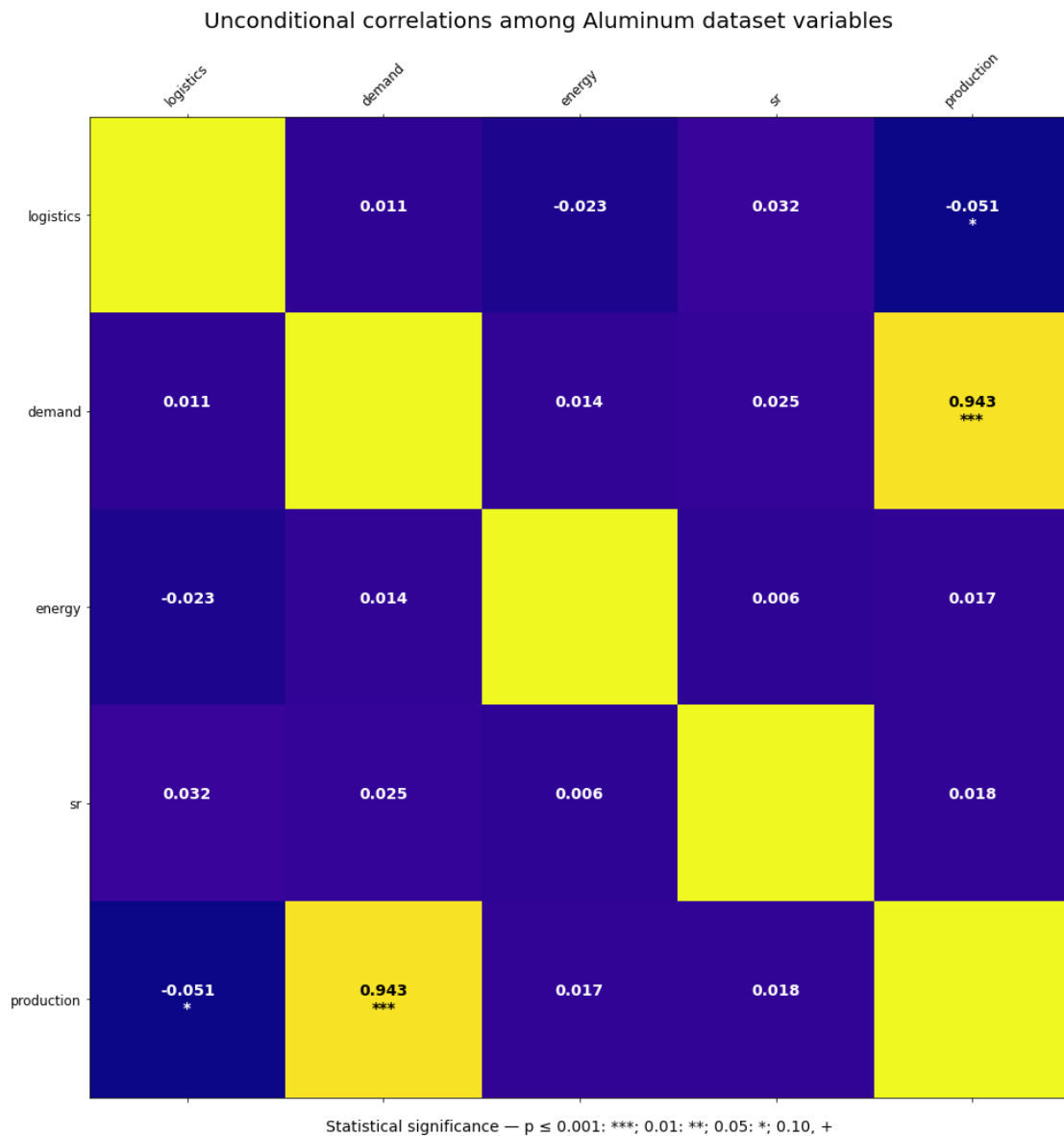
	Logistics	Demand	Energy	SR	Production
COUNT	1500	1500	1500	1500	1500
MEAN	800.0516174	902.6906726	4.747709482	16.97746326	109.1277459
STD	20.25009356	194.8384699	0.5024685747	1.993075021	61.51052741
MIN	733.7206469	288.5901375	3.261181733	9.295352013	-78.09922737
25%	786.3413724	772.6042539	4.41056957	15.67799624	68.78057107
50%	800.0679225	903.4334231	4.747664861	17.0525907	108.704514
75%	813.4554434	1033.506862	5.075532364	18.29045069	152.9128889
MAX	877.0589161	1590.889911	6.380051841	24.0994603	293.7767397

Με αυτές τις τιμές, τα δεδομένα θεωρούνται επαρκή ώστε να εφαρμοστούν κατάλληλες μέθοδοι μηχανικής μάθησης που βασίζονται σε σύνολα δέντρων αποφάσεων (decision – tree ensembles) ή μοντέλα ενίσχυσης (boosting models). Ένα σύνολο δεδομένων με 1500 παρατηρήσεις (περίπου 5 έτη δεδομένων) θεωρείται κατάλληλο για random forests, extra trees ή gradient boosting. Αυτές οι εποπτευόμενες μέθοδοι μηχανικής εκμάθησης, είναι εξαιρετικά ευάλωτες στην υπερπροσαρμογή, συνήθως απαιτούν διαχωρισμό των δεδομένων για σκοπούς εκπαίδευσης και δοκιμών. Ένα σύνολο αυτού του μεγέθους είναι ικανό να μετρήσει την προγνωστική ακρίβεια και να διατηρήσει αιτιώδη συμπεράσματα με μεγαλύτερη αξιοπιστία από τις παραδοσιακές γραμμικές μεθόδους.

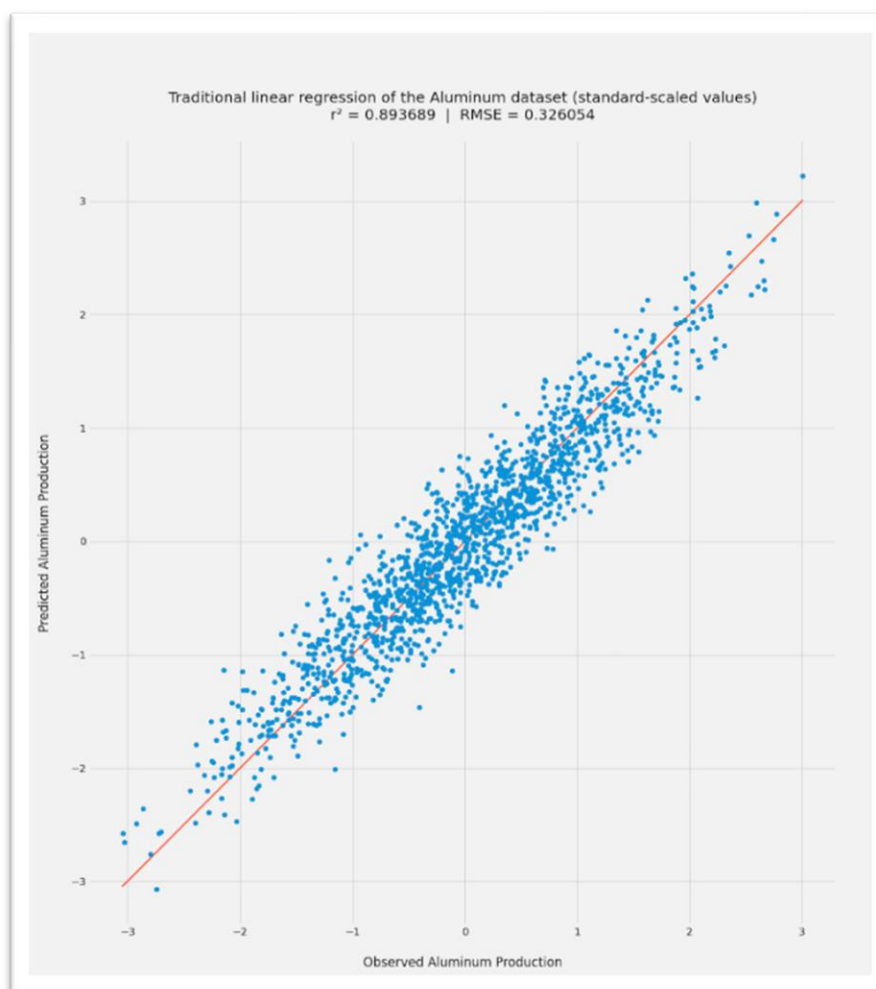


Διάγραμμα 4.1. Ιστογράμματα και εκτιμήσεις (KDEs) των μεταβλητών-στόχων πρόβλεψης στη βάση δεδομένων του Αλουμινίου

Όπως φαίνεται από το Διάγραμμα 4.2 που ακολουθεί, η μεταβλητή με τη μεγαλύτερη επιρροή, όπως υποδεικνύεται από τον κατά Pearson συντελεστή γραμμικής συσχέτισης και τη στατιστική σημαντικότητα 1%, είναι η ζήτηση (Demand), κάτι το οποίο είναι αναμενόμενο. Επίσης, η μεταβλητή logistics φαίνεται πως συσχετίζεται αρνητικά με την παραγωγή σε επίπεδο σημαντικότητας 10%.



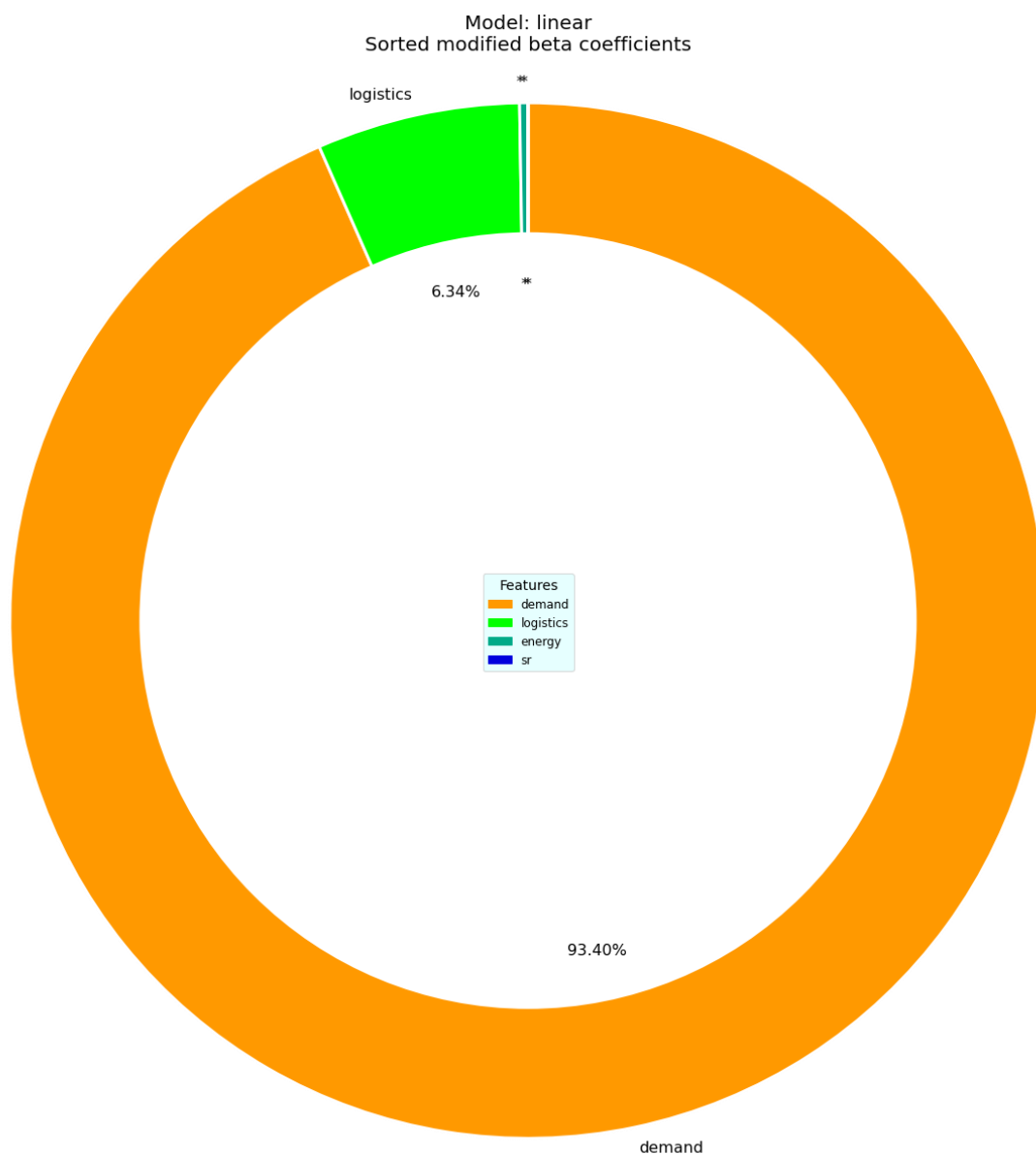
Διάγραμμα 4.2.Συσχετίσεις μεταξύ των μεταβλητών του συνόλου δεδομένων.



Διάγραμμα 4.3. Scatter plot γραμμικής παλινδρόμησης OLS

Όπως δείχνει το διάγραμμα 4.3, η παλινδρόμηση OLS επιτρέπει σε ένα γραμμικό μοντέλο να εκφραστεί και να ερμηνευτεί σε κλειστή μορφή. Η ακρίβεια, όπως μετράται με την μέθοδο για τις προβλέψεις στο σύνολο δεδομένων που αφορούν την παραγωγή του αλουμινίου, είναι ιδιαίτερα αξιοσημείωτη για την παραδοσιακή παλινδρόμηση ($R^2 = 0,893$) με $RMSE$ 0,326.

Στο διάγραμμα 4.2 παρουσιάζεται το μοντέλο KDE για ολόκληρο το σύνολο δεδομένων. Το συγκεκριμένο είναι πιο δύσκολο να το υπερασπιστούμε από ένα μοντέλο που βασίζεται σε διαχωρισμό train/test. Ο λόγος είναι ότι η εκτίμηση της πυκνότητας βασίζεται σε όλες τις "παρατηρούμενες" τιμές της μεταβλητής-στόχου. Αυτό την καθιστά μη χρήσιμη ως πρόβλεψη δεδομένων εντός του υπάρχοντος συνόλου, αφού δεν υπάρχει τίποτα που να κρατείται έξω.

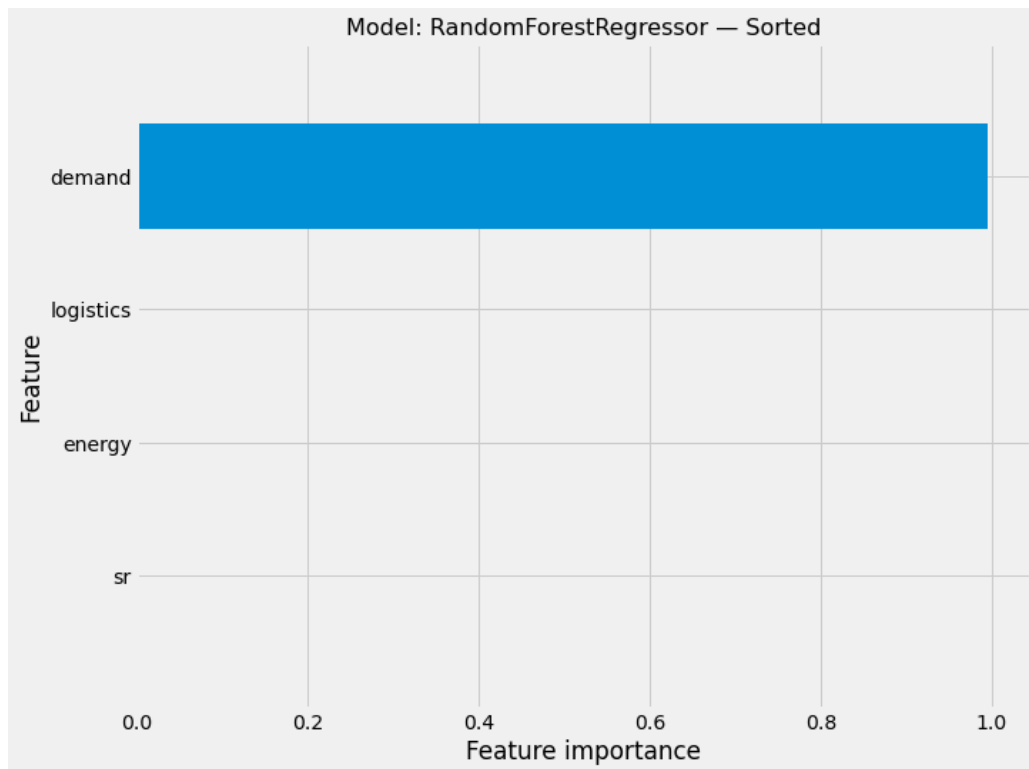


Διάγραμμα 4.4. Doughnut plot για το μοντέλο γραμμικής παλινδρόμησης

Παρατηρώντας το Διάγραμμα 4.4 βλέπουμε πως ο συντελεστής της μεταβλητής demand συνεισφέρει κατά 93,4% για την ερμηνεία της εκτίμησης της παραγωγής του αλουμινίου και κατά 6,34% ο συντελεστής της μεταβλητής logistics.

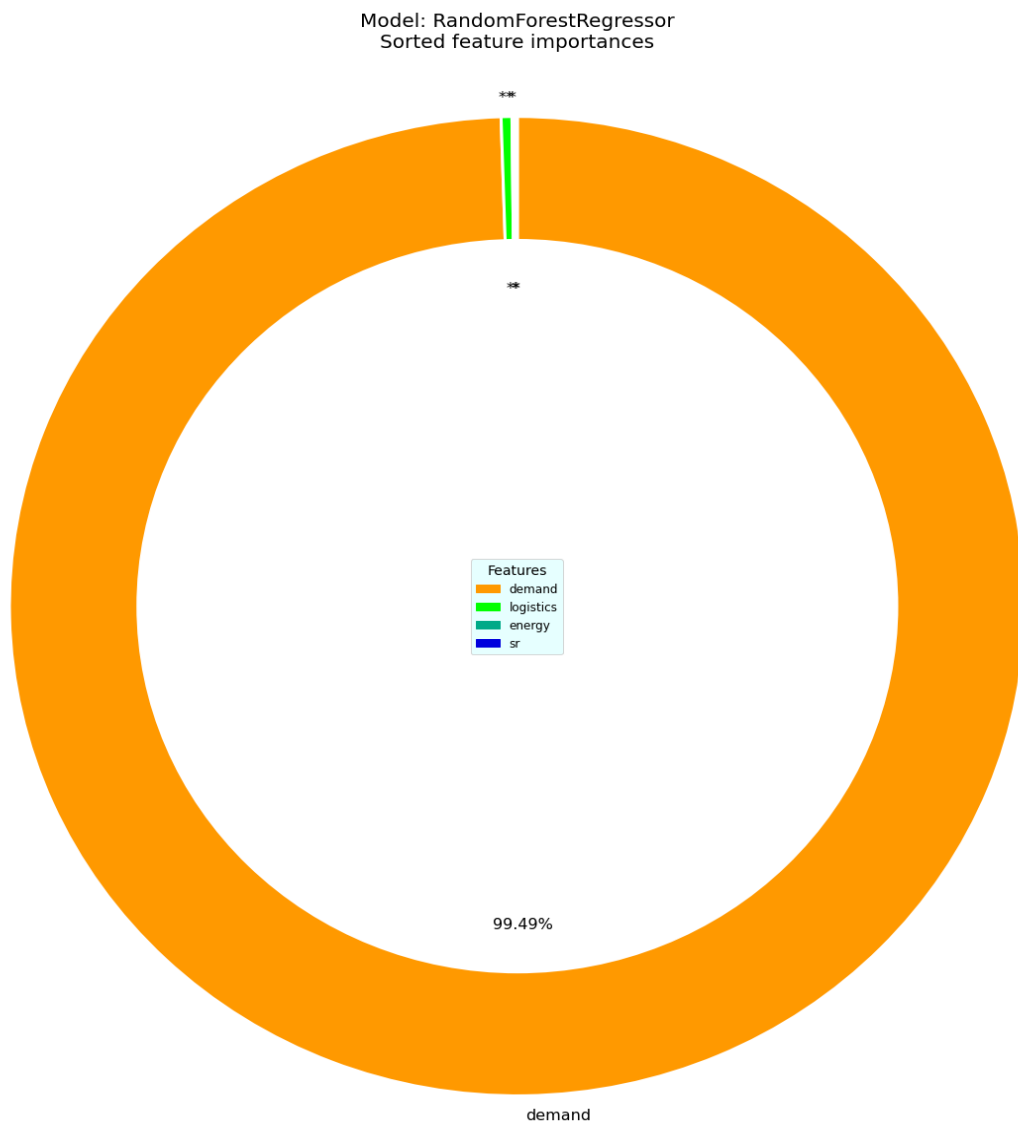
Ο διαχωρισμός των δεδομένων σύμφωνα με τις τιμές για κάθε ανεξάρτητη μεταβλητή δημιουργεί ένα δέντρο αποφάσεων που προβλέπει την μέση παραγωγή του αλουμινίου. Το μοντέλο μηχανικής μάθησης βελτιώνεται σε σχέση με την παραδοσιακή OLS μέθοδο.

Χρησιμοποιώντας τη μέθοδο RandomForest η τιμή του για το υποσύνολο της εκπαίδευσης είναι 0,905 και 0,876 για το σύνολο δοκιμής.



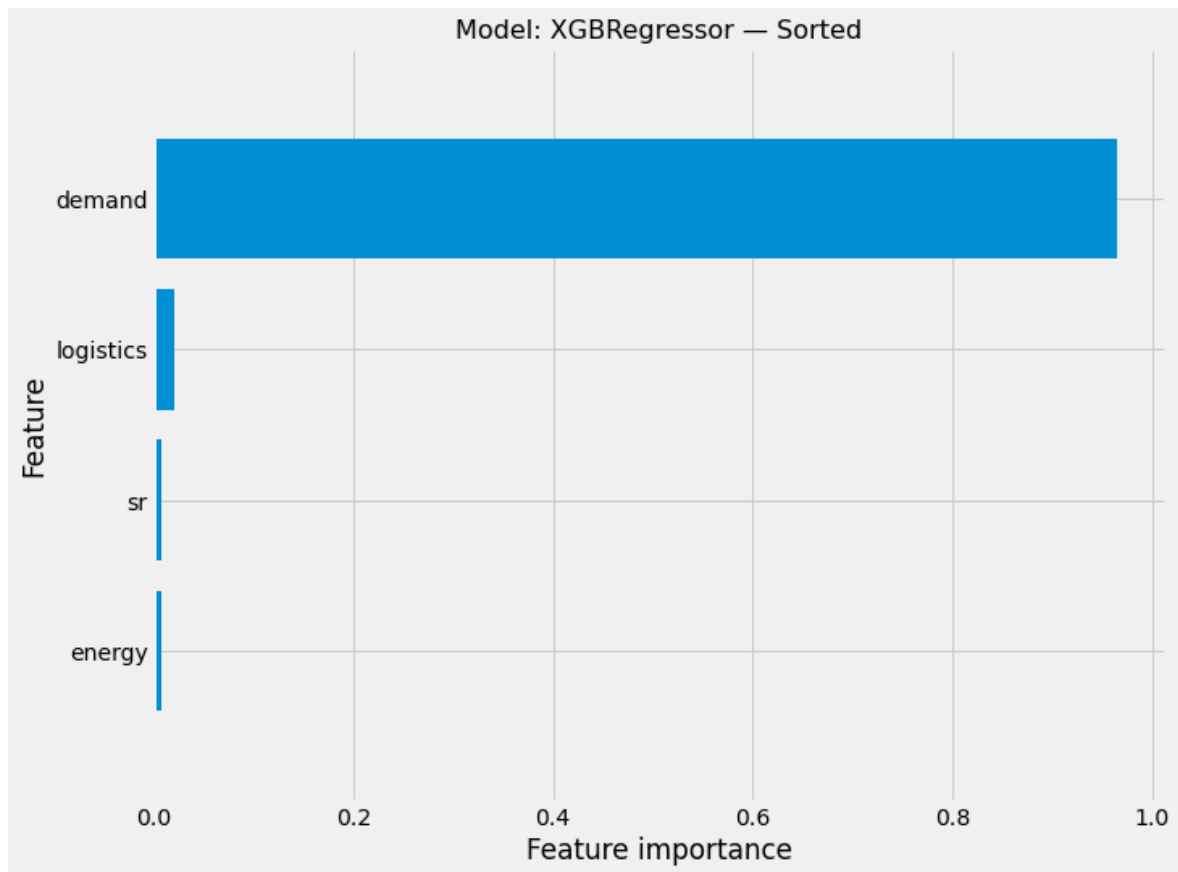
Διάγραμμα 4.5. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από τον αλγόριθμο δέντρου αποφάσεων RandomForest

Το Διάγραμμα 4.5 παρουσιάζει τη σημαντικότητα των ανεξάρτητων μεταβλητών για τη μέθοδο αυτή. Η ζήτηση αντιπροσωπεύει το 99,49% της προγνωστικής ισχύος ενός βασικού δέντρου αποφάσεων στο σύνολο δεδομένων.



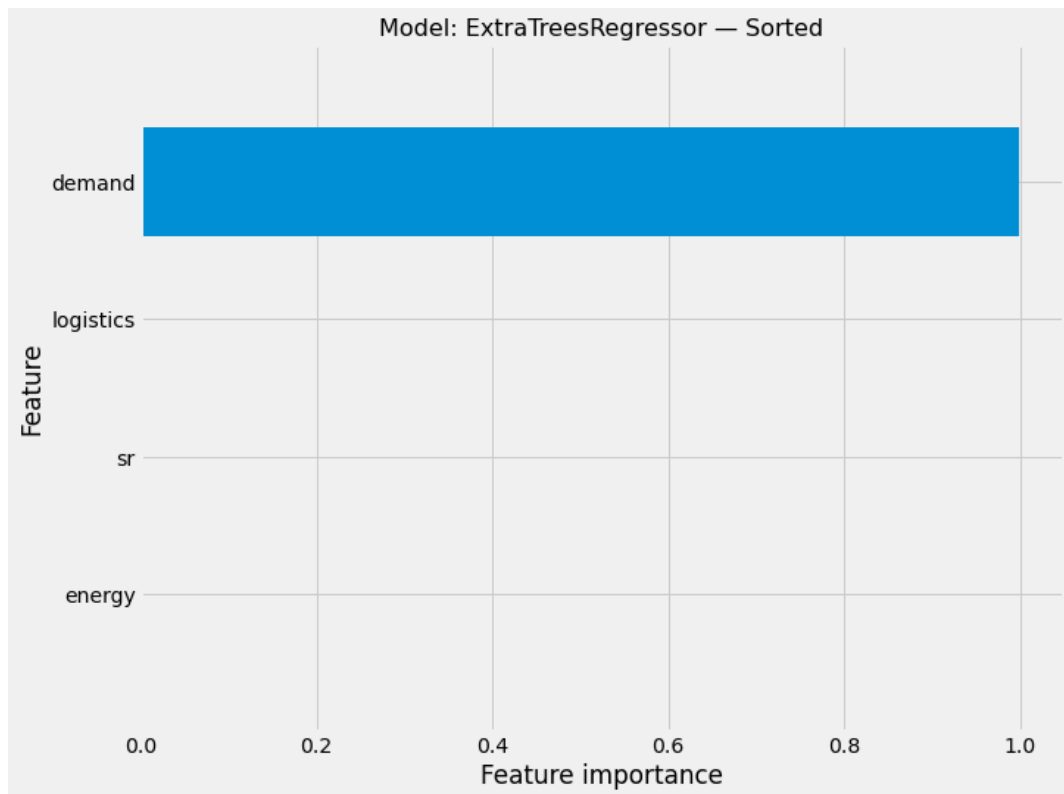
Διάγραμμα 4.6. Doughnut plot για το μοντέλο RandomForest

Χρησιμοποιώντας τη μέθοδο XGBRegressor η τιμή του για το υποσύνολο της εκπαίδευσης είναι 0,907 και 0,879 για το σύνολο δοκιμής. Το Διάγραμμα 4.7 παρουσιάζει τη σημαντικότητα των ανεξάρτητων μεταβλητών για τη μέθοδο αυτή. Η ζήτηση αντιπροσωπεύει το 96,47%, η μεταβλητή logistics το 2,06% και το υπόλοιπο 1,47% οι μεταβλητές SR και energy, της προγνωστικής ισχύος ενός βασικού δέντρου αποφάσεων στο σύνολο δεδομένων.



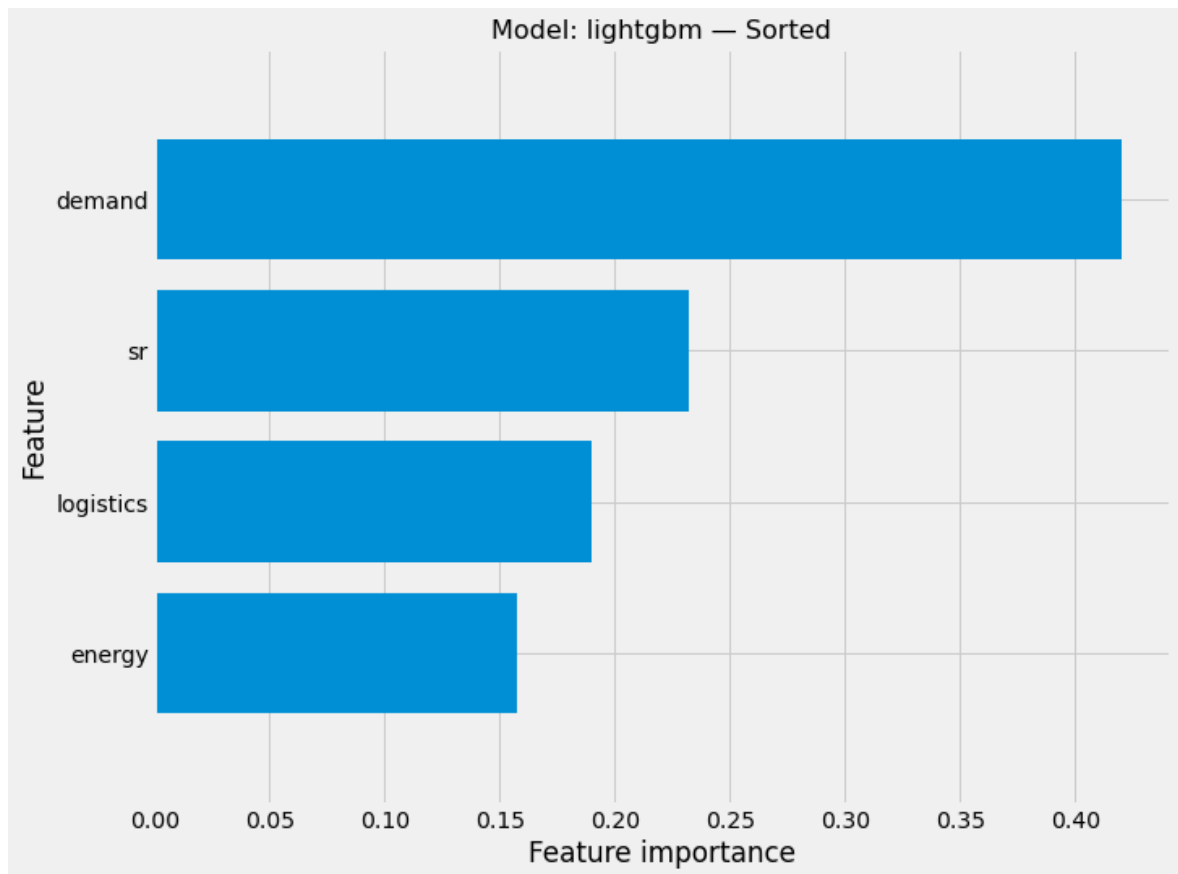
Διάγραμμα 4.7. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από την μέθοδο XGB

Η μέθοδος ExtraTreesRegressor δίνει τιμή για το υποσύνολο της εκπαίδευσης 0,869 και για το σύνολο της δοκιμής 0,845. Το Διάγραμμα 4.8 που ακολουθεί μας δείχνει ότι για αυτή τη μέθοδο η ζήτηση και πάλι πρωταγωνιστεί με ποσοστό 99,87% της προβλεπτικής ικανότητας της παραγωγής.



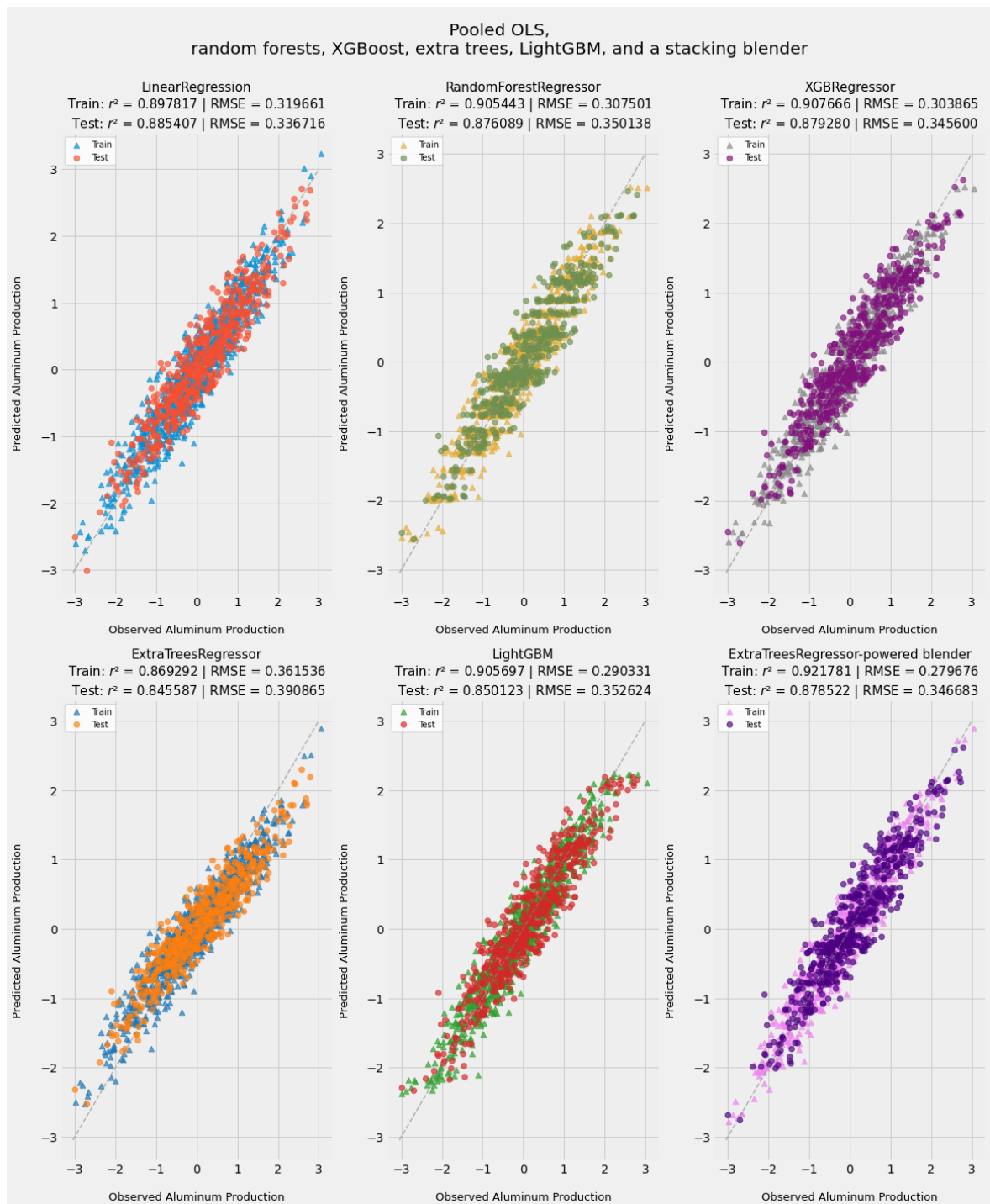
Διάγραμμα 4.8. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από την μέθοδο ExtraTrees.

Η μέθοδος LightGBM δίνει τιμές στο σύνολο της εκπαίδευσης 0,905 και στο σύνολο της δοκιμής 0,85, η ζήτηση μετέχει κατά 42%, η μεταβλητή SR κατά 23%, η μεταβλητή logistics 19% και τέλος η μεταβλητή energy 15,75% στη πρόβλεψη της παραγωγής.



Διάγραμμα 4.9. Σημαντικότητα ανεξάρτητων μεταβλητών που δημιουργούνται από την μέθοδο LightGBM.

Τέλος, στο Διάγραμμα 4.10, τα αποτελέσματα των δοκιμών παρέχουν περισσότερη πληροφορία από αυτά της εκπαίδευσης. Οι βαθμολογίες και το μέσο τετραγωνικό σφάλμα (RMSE) παρατηρούμε πως είναι καλύτερες στο απλό γραμμικό μοντέλο παλινδρόμησης.



Διάγραμμα 4.10. Scatterplots και των 6 μοντέλων πρόβλεψης για το σύνολο δεδομένων του αλουμινίου.

Τελικά, η διαφορά μεταξύ των περιπτώσεων όπου οι γενικευμένες γραμμικές μέθοδοι κυριαρχούν στη μηχανική μάθηση ή το αντίστροφο δεν είναι μαθηματική. Μάλλον, η προτίμηση για μια κατηγορία μεθόδων έναντι της άλλης είναι πρακτική. Σε γενικές γραμμές, το OLS και άλλες γραμμικές μέθοδοι προσφέρουν ανώτερη ερμηνεία, ενώ η μηχανική μάθηση υπόσχεται και συνήθως προσφέρει ανώτερη ακρίβεια.

Κεφάλαιο 5 - Συζήτηση αποτελεσμάτων και συμπεράσματα

5.1 Ερμηνεία αποτελεσμάτων

Στο κεφάλαιο 4 εφαρμόστηκαν διάφορες μέθοδοι ανάλυσης δεδομένων λαμβάνοντας υπόψη τις προσομοιωμένες τιμές των μεταβλητών, όπως αυτές δημιουργήθηκαν βάσει ιστορικών τιμών από τη λειτουργία του μηχανήματος διέλασης αλουμινίου Tandem. Αφού τα κανονικοποιημένα δεδομένα χωριστούν σε κατάλληλα σύνολα εκπαίδευσης, επικύρωσης και δοκιμών, επιχειρήθηκε η εφαρμογή των ανωτέρω μεθόδων μηχανικής μάθησης. Η σημασία των χαρακτηριστικών (feature importances) που δημιουργούνται από δένδρολογικές μεθόδους, όπως αυτές παρουσιάστηκαν στο προηγούμενο κεφάλαιο δεν είναι δυνατό να μεταφέρουν πληροφορίες που σχετίζονται με το πρόσημο των συντελεστών και τις συσχετίσεις σε γραμμικά μοντέλα. Η επιλογή μεταξύ συμβατικών γραμμικών μεθόδων και εναλλακτικών μεθόδων μηχανικής μάθησης, εξαρτάται από αυτή την ισορροπία μεταξύ ακρίβειας και ερμηνείας. Σε εφαρμογές ή περιστάσεις που δίνεται έμφαση στην προγνωστική ακρίβεια, η μηχανική μάθηση μπορεί να κυριαρχεί από τη συμβατική παλινδρόμηση. Σύνολα δεδομένων που είναι εγγενώς δύσκολο να ερμηνευτούν σε γραμμικά μοντέλα μπορεί να ωφεληθούν από την άμεση προσφυγή στη μηχανική μάθηση. Αντίθετα, όπου το πρόσημο και η κλίμακα των συντελεστών παλινδρόμησης έχουν μεγαλύτερη σημασία από την προγνωστική ακρίβεια, η μηχανική μάθηση μπορεί να παίξει έναν βοηθητικό ρόλο. Στο παράδειγμα

Πιο αναλυτικά, χρησιμοποιώντας το μοντέλο της γραμμικής παλινδρόμησης αναλύοντας τα αποτελέσματα που προέκυψαν γίνεται αντιληπτό πως η μεταβλητή η οποία συσχετίζεται θετικά και μάλιστα υψηλά με την παραγωγή του αλουμινίου είναι η ζήτηση. Αξίζει να σημειωθεί, πως ζήτηση (demand) εμφανίζεται ως η πιο σημαντική σε όλα τα μοντέλα παλινδρόμησης που χρησιμοποιήθηκαν.

Επιπλέον, σημαντική θεωρείται αλλά όχι με τον ίδιο βαθμό και η σχέση της μεταβλητής logistics με αρνητικό πρόσημο, όπως αυτό αναμενόταν. Αξίζει να σημειωθεί, πως η αποτελεσματική οργάνωση και διοίκηση της ροής προϊόντων και η συνετή διαχείριση αποθεμάτων σε μία εταιρεία είναι άρρηκτα συνδεδεμένη με την ποσότητα που θα πρέπει να παράγει μια εταιρεία, ώστε να ικανοποιήσει πλήρως τους πελάτες της, ενώ ταυτόχρονα να μπορεί να εξυπηρετήσει τις εσωτερικές διαδικασίες αποθήκευσης.

Οι υπόλοιπες μεταβλητές φαίνεται πως δεν συσχετίζονται ισχυρά με την παραγωγή. Ωστόσο, μέθοδοι όπως η Light GBM μπορούν να εξισορροπήσουν τις τιμές και να δώσουν

μεγαλύτερη σημασία σε μεταβλητές, οι οποίες φαίνεται να επηρεάζουν λιγότερο την παραγωγική διαδικασία. Το γεγονός αυτό επιβεβαιώνεται εν μέρη και από άλλες μεθόδους, οι οποίες έδειξαν ότι η κατανάλωση ενέργειας και ο δείκτης σοβαρότητας συμβάντων συμμετέχουν, έστω και με μικρό ποσοστό στην πρόβλεψη της παραγωγής αλουμινίου και ως εκ τούτου δεν μπορεί να θεωρηθούν αμελητέες.

Τέλος, από το Διάγραμμα 4.10 γίνεται αντιληπτό πως για το σύνολο των υποδειγμάτων που χρησιμοποιήθηκαν, τα δεδομένα εμφανίζουν υψηλή προσαρμοστικότητα ενώ ταυτόχρονα η προγνωστική αξία των μοντέλων κρίνεται ιδιαίτερος ικανοποιητική. Ως εκ τούτου, το τελευταίο υπόδειγμα ανάμειξης που εφαρμόστηκε, δύναται να χρησιμοποιηθεί ως το πλέον κατάλληλο προκειμένου να ληφθούν οι καθημερινές προβλέψεις αναφορικά με τη βέλτιστη παραγωγική διαδικασία του μηχανήματος διέλασης Tandem.

5.2. Συμπεράσματα

Στην παρούσα ενότητα, επιχειρείται μια σύντομη περιγραφή των ευρημάτων που προέκυψαν από την ανασκόπηση βιβλιογραφίας, καθώς και την αξιοποίηση των μεθόδων ανάλυσης δεδομένων στη βιομηχανία και πιο συγκεκριμένα στην παραγωγή διέλασης αλουμινίου. Όπως κατέστη φανερό, η ανάλυση δεδομένων μπορεί να αποδειχθεί ένα εξαιρετικά χρήσιμο εργαλείο για την κατανόηση, τον έλεγχο και τελικά τη βελτιστοποίηση της παραγωγικής διαδικασίας. Συγκεκριμένα, έγινε αντιληπτό πως με τη βοήθεια σύγχρονων και εξειδικευμένων μεθόδων ανάλυσης δεδομένων, μια βιομηχανική μονάδα μπορεί να αξιοποιήσει δεδομένα που παράγονται από τις εκτεταμένες ιστορικές βάσεις δεδομένων που διαθέτει, καθώς και τον μεγάλο όγκο δεδομένων που παράγει και μάλιστα σε πραγματικό χρόνο, σε καθημερινή βάση. Οι πληροφορίες αυτές στη συνέχεια μετατρέπονται σε έναν εξαιρετικά πολύτιμο πόρο, ο οποίος με τις κατάλληλες μεθόδους ανάλυσης, αποκαλύπτει τις ιδιαίτερες πτυχές της παραγωγικής διαδικασίας που είναι σημαντικές για τη βέλτιστη λειτουργία της επιχείρησης. Ταυτόχρονα, η χρήση των μεγάλων δεδομένων, δύναται να προσφέρει ιδιαίτερα σημαντική προγνωστική δυνατότητα, η οποία με την κατάλληλη αξιοποίηση της αποφέρει σημαντικά πλεονεκτήματα τόσο ως προς τη στρατηγική της επιχείρησης, μέσω της αντιστάθμισης κινδύνων, όσο και ως προς την αναβάθμιση του εσωτερικού επιχειρησιακού περιβάλλοντος και τελικά τη βελτίωση της παραγωγικότητάς της.

Η μελέτη περίπτωσης που παρουσιάστηκε αφορά τη μηχανή Tandem, η οποία είναι καθοριστική για την παραγωγή αλουμινίου υψηλής ποιότητας, ωστόσο απαιτεί σημαντική επενδυτική δαπάνη, τεχνολογική εξειδίκευση και συνεχή συντήρηση. Το γεγονός αυτό,

κάνει τη χρήση της προγνωστικής ανάλυσης και των μεγάλων δεδομένων ακόμη πιο σημαντική, ώστε να μεγιστοποιηθεί η αποτελεσματικότητα και η απόδοση αυτών των σημαντικών συστημάτων. Εκτός από την βασική λειτουργία της στην παραγωγή λεπτών φύλλων αλουμινίου, η μηχανή Tandem έχει κι άλλα ουσιαστικά χαρακτηριστικά και λειτουργίες όπως:

- **Ακρίβεια:** Επιτρέπει στους κατασκευαστές να ελέγχουν με ακρίβεια το πάχος και τη διάσταση των φύλλων αλουμινίου, διασφαλίζοντας έτσι ποιότητα και συνέπεια.
- **Αποδοτικότητα:** Παρόλο που απαιτεί σημαντική επενδυτική δαπάνη, είναι αποδοτική από πλευράς ενέργειας, επιτρέποντας την επεξεργασία μεγάλων όγκων αλουμινίου με σχετικά χαμηλή κατανάλωση ενέργειας.
- **Πολυεπίπεδη Επεξεργασία:** Η μηχανή Tandem επιτρέπει την επεξεργασία των φύλλων αλουμινίου σε διάφορα στάδια, κάθε ένα από τα οποία μπορεί να προσαρμοστεί για να επιτύχει συγκεκριμένα χαρακτηριστικά στο τελικό προϊόν.
- **Προσαρμοστικότητα:** Με την χρήση εξελιγμένων συστημάτων ελέγχου, η μηχανή Tandem μπορεί να προσαρμοστεί για να επεξεργαστεί διάφορα είδη και ποιότητες αλουμινίου, καθιστώντας την εξαιρετικά ευέλικτη για πληθώρα εφαρμογών.

Στην παρούσα μελέτη, αναπτύχθηκε ένα υπόδειγμα διαχείρισης μεγάλων δεδομένων μέσω του οποίου βελτιώνεται η λειτουργία του μηχανήματος Tandem. Πιο αναλυτικά, η προγνωστική δυνατότητα από τη χρήση των μεγάλων δεδομένων για τη λειτουργία του Tandem εμπεριέχει τα εξής:

- **Προβλεπτική Συντήρηση:** Με τη συλλογή και ανάλυση δεδομένων από το μηχανήμα Tandem, μπορούμε να προβλέψουμε πότε μπορεί να χρειαστεί συντήρηση ή επισκευή.
- **Βελτιστοποίηση της Απόδοσης:** Τα Big Data μπορούν να αναλυθούν για να κατανοηθούν οι τάσεις και τα μοτίβα που συνδέονται με την απόδοση του μηχανήματος. Αυτό μπορεί να οδηγήσει σε βελτιώσεις στις διαδικασίες και στην ρύθμιση του μηχανήματος για να επιτευχθεί η μέγιστη απόδοση.
- **Ελαχιστοποίηση των Αποβλήτων:** Η ανάλυση των Big Data μπορεί να βοηθήσει να εντοπιστούν οι πηγές αποβλήτων ή υπερβολικής κατανάλωσης υλικού. Αυτό μπορεί να οδηγήσει σε βελτιώσεις στη διαδικασία που ελαχιστοποιούν τα απόβλητα και τονίζουν την αποδοτικότητα.

Με βάση τα ανωτέρω, γίνεται αντιληπτό πως τα μεγάλα δεδομένα και η κατάλληλη ανάλυσή τους, μπορεί να διαδραματίσει κρίσιμο ρόλο στην ενίσχυση της λειτουργίας του κλάδου της βιομηχανίας αλουμινίου, αλλά και γενικότερα, εξασφαλίζοντας ποιότητα, αποδοτικότητα και βιωσιμότητα. Οι αποφάσεις πλέον βασίζονται σε στοιχεία, μοτίβα και τάσεις που ανακαλύπτονται από τα δεδομένα, σε αντίθεση με την μέχρι πρότινος αντιμετώπιση, η οποία χρησιμοποιούσε αποκλειστικά τη διαίσθηση ή/και την εμπειρία του παρελθόντος.

Βιβλιογραφία

- Aggiropoulos Charalampos L. (2022). *Quantitative Methods in Economic Theory through Innovative Machine Learning Techniques*. Research Center of the University of Piraeus.
- Ahmadi, H., Kaur, H., & Ammar, H. H. (2020). *Predictive Analytics Approaches in Big Data Environment: A Systematic Literature Review*. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC).
- Al-Turjman, F. M., Jeong, Y. S., Chong, I., & Jang, B. S. (2021). Security and Privacy Challenges in Big Data Analytics for the Internet of Things. *IEEE Internet of Things Journal*, 8(2), 1217–1224.
- Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79–80, 3–15.
- Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know about. *Annual Review of Economics*, 11, 685–725.
<https://doi.org/10.1146/ANNUREV-ECONOMICS-080217-053433>
- Bailly M., & Manyika J. (2013). *Is manufacturing ‘cool’ again?* McKinsey Global Inst.
- Breiman, L. (n.d.-a). *Arcing classifiers*.(1998a) *Annals of Statistics*, 26, 801–824.
- Breiman, L. (n.d.-b). *Arcing the edge*. *Annals of Probability*,(1998b). 26, 1683–1702.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
<https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–215. <https://doi.org/10.1214/SS/1009213726>
- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall/CRC.
- Chen, C., Zhang, C., Liu, J., & Liu, Z. (2019). *Big data analytics in supply chain management: A comprehensive literature review and a proposed framework*.
- Chen, H., Chiang, R. H., & Storey, V. C. (2020). *Business intelligence and analytics: From big data to big impact*. *MIS Quarterly*, 44(4). (pp. 1473–1496).
- Chen, J. M. (2021). An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, 27(1). <https://doi.org/10.1007/S11294-021-09815-6>

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Dankers, F. J. W. M., Traverso, A., Wee, L., & van Kuijk, S. M. J. (2018). Prediction modeling methodology. *Fundamentals of Clinical Data Science*, 101–120. https://doi.org/10.1007/978-3-319-99713-1_8
- Das, S., Bhattacharya, D., Ghosh, A., & Aditya, R. (2021). *Predictive Analytics and Decision Making in the Era of Big Data: State of the Art Review and Future Directions*. *Computers & Industrial Engineering*, 155, 107172.
- Drucker, H., & Cortes, C. (1996). *Boosting decision trees*. *Advances in Neural Information Processing Systems*. (Vol. 8).
- Freud, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and its application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4), 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
- Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115. <https://doi.org/10.1016/j.engappai.2022.105151>
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- Geman, S., Bienenstock, É., & Doursa, D. (1992). *Neural networks and the bias/variance dilemma*. *Neural Computation*, 4.
- Géron, A. (2019). *Hands-on machine learning with SciKit-learn, Keras & TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol: O'Reilly.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/S10994-006-6226-1>
- Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). *The fallacy of employing standardized regression coefficients and correlations as measures of effect*. *American Journal of Epidemiology*, 123.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115. <https://doi.org/10.1016/j.is.2014.07.006>

- Herschel, G., & Davis, M. (2015). *Understanding the Spectrum of Analytics Capabilities*. Gartner.
- Ho, T. K. (1995). *Random decision forests*. *Proceedings of 3rd International Conference on Document Analysis and Recognition*.
- Kagermann H., & Wahlster W. (2013). *Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0*. Working Group, Acatech National Academy of Science and Engineering, Germany,.
- Kapoor, K., Dwivedi, Y. K., Piercy, N., & Williams, M. D. (2021). *Big Data Analytics in Supply Chain Management: A Comprehensive Review*. *Decision Support Systems*, 145, 113482.
- Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Alam, M., Shiraz, M., & Gani, A. (2020). *Big IoT data analytics: architecture, opportunities, and open research challenges*. *Future Generation Computer Systems*. 375–390.
- Kohavi, R., & Wolpert, D. H. (1996). *Bias plus variance decomposition for zero-one loss functions*. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*.
- Krumeich, J., Werth, D., & Loos, P. (2016). *Prescriptive Control of Business Processes*. *Business & Information Systems Engineering*, 58(4), 261–280.
<https://doi.org/10.1007/S12599-015-0412-2>
- Liang, L., Zhang, J., Lu, C., & Hu, Z. (2021). *Big Data-Driven Decision-Making Model and Algorithms for Industrial Engineering*. *Applied Sciences*, 11(2), 575.
- Loh, W.-Y. (2008). *Classification and regression tree methods*. In F. Ruggeri, R. S. Kennet, & F. W. Faltin (Eds.), *Encyclopedia of statistics in quality and reliability*. Hoboken: Wiley.
- Luo, X., Zhang, J., Ren, X., Chen, T., & Zhou, X. (2020). *Big Data Analytics for Smart Manufacturing: Literature Review, Framework, and Future Research Agenda*. *Journal of Industrial Information Integration*, 19, 100133.
- Lustig, I., Dietrich, B., Johnson, C., & Dziekan, C. (2010). *The Analytics Journey*. Institute for Operations Research and the Management Sciences. INFORMS. <http://analytics-magazine.org/the-analytics-journey/>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. The McKinsey Global Institute (MGI).

- Muller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. 376.
- Newman, T. B., & Browner, W. S. (1991). *In defense of standardized regression coefficients*. *Epidemiology*, 2.
- Nguyen, V., Huang, W., Liu, C., & Yu, J. (2020). *Big Data Analytics for Smart Manufacturing: Case Studies in Semiconductor Manufacturing*. *Computers & Industrial Engineering*, 139, 106160.
- Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4), 431–448. <https://doi.org/10.1016/j.jksuci.2017.06.001>
- Philip Chen, C. L., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314–347. <https://doi.org/10.1016/j.ins.2014.01.015>
- Rahmani, A. M., Gia, T. N., Negash, B., Anzanpour, A., & Azimi, I. (2021). *Big data analytics for smart energy management in internet of things ecosystems: A review*. *Sustainable Cities and Society*, 73. (pp. 103–125).
- Ratnasari, A. D., Syahputra, M. F., & Musdalifah, R. (2020). Improving Aluminium Waste Recovery Efficiency through Material Flow Analysis: A Case Study of the Aluminium Industry in Indonesia. In *Conservation and Recycling* (p. 152).
- Reichert, P. (2014). *Comarch EDI Platform Case Study: The Advanced Electronic Data Interchange Hub as a Supply-Chain Performance Booster*. 143–155. https://doi.org/10.1007/978-3-319-07287-6_10
- Sahay, B., & Ranjan, J. (2008). Real time business intelligence in supply chain analytics. *Information Management & Computer Security*, 16(1), 28–48.
- Sanders, N. R. (2016). *How to Use Big Data to Drive Your Supply Chain*. California Management Review, Vol. 58, Issue. 3.
- Sanders, N. R. (2014). Big Data Driven Supply Chain Management: A Framework for Implementing Analytics and Turning Information into Intelligence. In *New Jersey: Pearson Education*.
- Shi-Nash, A., & Hardoon, D. R. (2017). Data analytics and predictive analytics in the era of big data. *Internet of Things and Data Analytics Handbook*, 329–345. <https://doi.org/10.1002/9781119173601.CH19>

- Varela, I. R., & Tjahjono, B. (2014). Big data analytics in supply chain management: trends and related research. *6th International Conference on Operations and Supply Chain Management*.
- Vikas Pandey. (n.d.). (<https://www.wipro.com/natural-resources/digital-transformation-of-aluminum-smelters/>).
- Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, 176, 98–110.
- Wolpert, D. H. (1996). The Lack of a Priori Distinctions between Learning Algorithms. *Neural Computation*, 8(7), 1341–1390. <https://doi.org/10.1162/NECO.1996.8.7.1341>
- Zhang, Q., Wang, J., Wang, S., Huang, J., Zhang, W., & Wang, H. (2019). *Big data analytics for smart manufacturing: Case studies and applications*. IEEE Access.