



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

**Πρόβλεψη Ψηφιακής Ευαλωτότητας Επιχειρήσεων σε
Οικονομική Κρίση με Χρήση Μεθόδων Μηχανικής Μάθησης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
Του
Νικολάου Αριστείδη

Επιβλέπων καθηγητής: Δρ Ευριπίδης Λουκής, Καθηγητής

Μέλη εξεταστικής επιτροπής:

Δρ Χαράλαμπος Αλεξόπουλος, Επίκουρος Καθηγητής

Δρ Βασιλική Διαμαντοπούλου, Επίκουρη Καθηγήτρια

Σάμος, Οκτώβριος, 2024

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Η παρούσα διπλωματική εργασία με τίτλο «Πρόβλεψη Ψηφιακής Ευαλωτότητας Επιχειρήσεων σε Οικονομική Κρίση με Χρήση Μεθόδων Μηχανικής Μάθησης» εκπονήθηκε στο πλαίσιο του προπτυχιακού προγράμματος σπουδών του Τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου.

Με το πέρας της παρούσας διπλωματικής εργασίας μου νιώθω την ανάγκη να εκφράσω τις ευχαριστίες μου προς τα άτομα που συνέβαλαν στην ολοκλήρωση της, παρέχοντάς μου είτε γνώση, είτε ηθική υποστήριξη. Αρχικά, ευχαριστώ θερμά τον επιβλέποντα Δρ. Ευριπίδη Λουκή, Καθηγητή του Τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου, για την εμπιστοσύνη και τον χρόνο που μου επέδειξε στην εκπόνηση της διπλωματικής εργασίας. Επιπλέον, θέλω να τον ευχαριστήσω θερμά για τις υποδείξεις και τη στήριξή του, καθώς και για την πολύτιμη γνώση που απέκτησα κατά την πορεία των φοιτητικών μου χρόνων. Ευχαριστίες θα ήθελα επίσης να απευθύνω σε όλους τους καθηγητές του τμήματος, για το γνωστικό υπόβαθρο, τον επαγγελματισμό και την έμπειρη καθοδήγηση που μου πρόσφεραν όλα αυτά τα χρόνια. Τέλος, θέλω να εκφράσω ιδιαιτέρως τις ευγνωμοσύνες μου προς την οικογένειά μου για τη συνεχή συμπαράστασή τους, τον αφιερωμένο χρόνο και την αγάπη τους, καθώς και για τις πολύτιμες συμβουλές τους και όλα όσα μου έχουν προσφέρει κατά τη διάρκεια της ζωής μου και των σπουδών μου.

© 2024

Του

ΝΙΚΟΛΑΟΥ ΑΡΙΣΤΕΙΔΗ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πανεπιστήμιο Αιγαίου, Τμήμα Μηχανικών Π.Ε.Σ.

Πρόβλεψη Ψηφιακής Ευαλωτότητας Επιχειρήσεων σε Οικονομική Κρίση με Χρήση Μεθόδων Μηχανικής Μάθησης

Νικολάου Αριστείδης

Περίληψη

Στόχος της Διπλωματικής αυτής Εργασίας είναι η χρήση ιστορικών δεδομένων από την περίοδο της οικονομικής κρίσης που είχαμε στην χώρα μας την προηγούμενη δεκαετία, σε συνδυασμό με μεθόδους τεχνητής νοημοσύνης από τον χώρο της μηχανικής μάθησης, για την δημιουργία ενός μοντέλου πρόβλεψης της ψηφιακής ευαλωτότητας (= της ευαλωτότητας όσον αφορά τις επενδύσεις για ψηφιακές τεχνολογίες) επιχειρήσεων. Αρχικά γίνεται παρουσίαση του θεωρητικού υποβάθρου, εστιάζοντας στην τεχνητή νοημοσύνη, την μηχανική μάθηση και την ανάλυση μεγάλων δεδομένων. Για τα τρία αυτά πεδία πραγματοποιείται μια επισκόπηση των βασικών αρχών και διαδικασιών τους, του πεδίου εφαρμογής, των κύριων κατηγοριών τους, αλλά και των προκλήσεων που αντιμετωπίζουν. Στην συνέχεια, πραγματοποιείται παρουσίαση της μεθοδολογίας που ακολουθήθηκε, παρουσιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε, το εργαλείο RapidMiner και οι διαδικασίες που ακολουθήθηκαν σε αυτό, οι μετρικές αξιολόγησης καθώς η παραμετροποίηση των μοντέλων. Τέλος, παρουσιάζονται τα αποτελέσματα όσον αφορά τα μοντέλα πρόβλεψης τα οποία δημιουργήθηκαν με βάση τα παραπάνω και την αξιοπιστία των προβλέψεών τους.

Λέξεις κλειδιά:

Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, RapidMiner, Οικονομική Κρίση

Abstract

The aim of this Thesis is to use historical data from the economic crisis period that our country experienced over the past decade, combined with artificial intelligence methods from the field of machine learning, to create a predictive model of digital vulnerability (i.e., vulnerability concerning investments in digital technologies) of firms. Initially, the theoretical background is presented, focusing on artificial intelligence, machine learning, and big data analysis. For these three fields, an overview of their basic principles and processes, areas of application, main categories, and the challenges they face is provided. Next, the methodology followed is presented, along with the dataset used, the RapidMiner tool, and the processes conducted within it, the evaluation metrics, and the parameterization of the models. Finally, the results concerning the predictive models created based on the above and the reliability of their predictions are presented.

Keywords:

Artificial Intelligence, Machine Learning, RapidMiner, Economic Crisis

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ	13
2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ	15
2.1. Τεχνητή Νοημοσύνη	15
2.1.1. Ορισμός Τεχνητής Νοημοσύνης	15
2.1.2. Τρόπος λειτουργίας	15
2.1.3. Σημασίας Τεχνητής Νοημοσύνης	16
2.1.4. Πλεονεκτήματα και μειονεκτήματα της τεχνητής νοημοσύνης	18
2.1.5. Δυνατή και Αδύνατη AI	22
2.1.6. Οι 4 τύποι τεχνητής νοημοσύνης.....	24
2.1.7. Παραδείγματα τεχνολογιών τεχνητής νοημοσύνης	25
2.1.8. Εφαρμογές της τεχνητής νοημοσύνης	28
2.1.9. Ηθική χρήση της τεχνητής νοημοσύνης	32
2.2. Μηχανική Μάθηση	34
2.2.1. Ορισμός μηχανικής μάθησης.....	34
2.2.2. Αλγόριθμος μηχανικής μάθησης	35
2.2.3. Τύποι μηχανικής μάθησης.....	36
2.2.4. Σημαντικά μοντέλα μηχανικής μάθησης.....	39
2.2.4.1. Δέντρα Απόφασης	39
2.2.4.2. Naïve Bayes	42
2.2.4.3. SVM.....	44
2.2.4.4. KNN	46
2.2.4.5. Νευρωνικά Δίκτυα	48
2.2.4.6. Random Forest	50
2.2.4.7. Γραμμική Παλινδρόμηση	52
2.2.5. Εφαρμογές Μηχανικής Μάθησης	54
2.3. Ανάλυση Μεγάλων Δεδομένων	56
2.3.1. Εισαγωγή στην Ανάλυση Μεγάλων Δεδομένων	56
2.3.2. Η Σημασία της Διαχείρισης και Ανάλυσης Μεγάλων Συνόλων Δεδομένων	56
2.3.3. Προκλήσεις στην Επεξεργασία Μεγάλων Δεδομένων	57

2.3.4. Εργαλεία και Τεχνολογίες για την Ανάλυση Μεγάλων Δεδομένων	58
2.3.5. Συλλογή και Προετοιμασία Δεδομένων	59
2.3.6. Αποθήκευση και Διαχείριση Δεδομένων	60
2.3.7. Επεξεργασία και Ανάλυση Δεδομένων	60
2.3.8. Οπτικοποίηση και Αναφορά Δεδομένων	61
2.3.9. Ασφάλεια και Ιδιωτικότητα Δεδομένων	62
3. ΜΕΘΟΔΟΛΟΓΙΑ ΠΕΙΡΑΜΑΤΙΚΟΥ ΜΕΡΟΥΣ	63
3.1. Βασική Φιλοσοφία	63
3.2. Σύνολο Δεδομένων Πειραματικής Διαδικασίας	64
3.3. Εργαλείο RapidMiner	67
3.3.1. Δυνατότητες.....	67
3.3.2. Οφέλη	69
3.3.3. Panels.....	70
3.4.Απαντήσεις ερώτησης προς πρόβλεψη	74
3.5.Μετρικές Αξιολόγησης	76
3.6. Εισαγωγή Δεδομένων	77
3.7. Εφαρμογή DecisionTree	78
3.7.1. Πρώτη παραλλαγή	78
3.7.2. Δεύτερη παραλλαγή	82
3.7.3. Τρίτη παραλλαγή	83
3.8. Εφαρμογή Random Forest	84
3.8.1. Πρώτη παραλλαγή	84
3.8.2. Δεύτερη και τρίτη παραλλαγή	85
4. ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΔΙΑΔΙΚΑΣΙΑΣ	87
4.1. Αποτελέσματα 1 ^{ης} παραλλαγής DecisionTree	87
4.2. Αποτελέσματα 2 ^{ης} παραλλαγής DecisionTree	88
4.3. Αποτελέσματα 3 ^{ης} παραλλαγής DecisionTree	89
4.4. Αποτελέσματα 1 ^{ης} παραλλαγής Random Forest.....	90
4.5. Αποτελέσματα 2 ^{ης} παραλλαγής Random Forest.....	92
4.6. Αποτελέσματα 3 ^{ης} παραλλαγής Random Forest.....	93
4.7. Σύγκριση Μέσων Αποτελεσμάτων.....	94

4.8. Αποτελέσματα εκτέλεσης με διαφορετική αναλογία training – testset.....	95
4.9. Εφαρμογή automodel	97
4.10. Αποτελέσματα automodel	102
4.11. Σχολιασμός αποτελεσμάτων	103
4.12. Νέος στόχος πρόβλεψης.....	105
4.13. Πρόβλεψη σημαντικότερων μεταβλητών με χρήση Automodel	109
Συμπεράσματα.....	114
ΒΙΒΛΙΟΓΡΑΦΙΑ – ΠΗΓΕΣ	115

Λίστα Εικόνων

Εικόνα 1: Δέντρο Απόφασης.....	41
Εικόνα 2: Μοντέλο SVM	45
Εικόνα 3: Νευρωνικό Δίκτυο	49
Εικόνα 4: Random Forest	51
Εικόνα 5: Γραμμική Παλινδρόμηση	53
Εικόνα 6: Ερωτηματολόγιο - Απόσπασμα 1	64
Εικόνα 7: Ερωτηματολόγιο - Απόσπασμα 2	64
Εικόνα 8: Ερωτηματολόγιο - Απόσπασμα 3	65
Εικόνα 9: Ερωτηματολόγιο - Απόσπασμα 4	65
Εικόνα 10: Ερωτηματολόγιο - Απόσπασμα 5	66
Εικόνα 11: Απόσπασμα Συνόλου Δεδομένων	66
Εικόνα 12: Περιβάλλον διεπαφής RapidMiner	68
Εικόνα 13: RepositoryPanel	71
Εικόνα 14: ProcessPanel	71
Εικόνα 15: OperatorsPanel	72
Εικόνα 16: ParametersPanel	73
Εικόνα 17: Επιλογή ImportData	77
Εικόνα 18: Επιτυχής Φόρτωση Δεδομένων	77
Εικόνα 19: Παράμετροι Retrieve.....	79
Εικόνα 20: Λίστα αποκλεισμένων Attributes.....	79
Εικόνα 21: NumericaltoPolynomial	80
Εικόνα 22: Παράμετροι SetRole	81
Εικόνα 23: Παράμετροι SplitData	81
Εικόνα 24: Παράμετροι DecisionTree	82
Εικόνα 25: Παράμετροι Performance	82
Εικόνα 26: : DecisionTreeProcess – 2 ^η παραλλαγή.....	83
Εικόνα 27: DecisionTree - αντιστοιχίσεις τιμών 2ης παραλλαγής	83
Εικόνα 28: DecisionTree - αντιστοιχίσεις τιμών 3ης παραλλαγής	84
Εικόνα 29: Random Forest Process – 1 ^η παραλλαγή	85
Εικόνα 30: Παράμετροι Random Forest.....	85
Εικόνα 31: Random Forest Process – 2η και 3η παραλλαγή	86
Εικόνα 32: Αποτελέσματα Accuracy - DecisionTree – 1 ^η παραλλαγή	87
Εικόνα 33: Αποτελέσματα Precision - DecisionTree 1 ^η παραλλαγή.....	87
Εικόνα 34: Confusion Matrix - Decision Tree– 1 ^η παραλλαγή	88
Εικόνα 35: Αποτελέσματα Accuracy - DecisionTree – 2η παραλλαγή.....	88
Εικόνα 36: Αποτελέσματα Recall - Decision Tree – 2η παραλλαγή	88
Εικόνα 37: Αποτελέσματα Precision - DecisionTree – 2η παραλλαγή.....	89
Εικόνα 38: Confusion Matrix - Decision Tree - 2 ^η παραλλαγή	89
Εικόνα 39: Αποτελέσματα Accuracy - DecisionTree - 3η παραλλαγή	89

Εικόνα 40: Αποτελέσματα Recall - Decision Tree - 3η παραλλαγή	90
Εικόνα 41: Αποτελέσματα Precision - DecisionTree - 3η παραλλαγή	90
Εικόνα 42: Confusion Matrix - Decision Tree - 3η παραλλαγή	90
Εικόνα 43: Αποτελέσματα Accuracy-Random Forest – 1 ^η παραλλαγή	90
Εικόνα 44: Αποτελέσματα Precision - DecisionTree 1 ^η παραλλαγή	91
Εικόνα 45: Confusion Matrix - Random Forest– 1 ^η παραλλαγή	91
Εικόνα 46: Αποτελέσματα Accuracy - Random Forest– 2η παραλλαγή	92
Εικόνα 47: Αποτελέσματα Recall - Random Forest– 2η παραλλαγή	92
Εικόνα 48: Αποτελέσματα Precision -Random Forest– 2η παραλλαγή	92
Εικόνα 49: Αποτελέσματα Accuracy - Random Forest- 3η παραλλαγή	93
Εικόνα 50: Αποτελέσματα Recall - Random Forest- 3η παραλλαγή	93
Εικόνα 51: Αποτελέσματα Precision - Random Forest- 3η παραλλαγή	93
Εικόνα 52: Αποτελέσματα DecisionTree - Παραλλαγή 1 - (60% - 40%)	95
Εικόνα 53: Αποτελέσματα DecisionTree - Παραλλαγή 2 - (60% - 40%)	96
Εικόνα 54: Αποτελέσματα DecisionTree - Παραλλαγή 3 - (60% - 40%)	96
Εικόνα 55: Αποτελέσματα Random Forest - Παραλλαγή 1 - (60% - 40%)	96
Εικόνα 56: : Αποτελέσματα Random Forest - Παραλλαγή 2 - (60% - 40%)	96
Εικόνα 57: : Αποτελέσματα Random Forest - Παραλλαγή 3 - (60% - 40%)	97
Εικόνα 58: AutoModel - Φόρτωση δεδομένων	99
Εικόνα 59: Καθορισμός στόχου και στήλης πρόβλεψης	99
Εικόνα 60: Κατανομή εγγραφών - Παραλλαγή 1	100
Εικόνα 61: Κατανομή εγγραφών - Παραλλαγή 2	100
Εικόνα 62: Κατανομή εγγραφών - Παραλλαγή 3	101
Εικόνα 63: Επιλογή στηλών	101
Εικόνα 64: Επιλογή μοντέλων	102
Εικόνα 65: Αποτελέσματα automodel - Παραλλαγή 1	102
Εικόνα 66: Αποτελέσματα automodel - Παραλλαγή 2	103
Εικόνα 67: Αποτελέσματα automodel - Παραλλαγή 3	103
Εικόνα 68: DecisionTree - νέος στόχος πρόβλεψης	105
Εικόνα 69: Random Forest - νέος στόχος πρόβλεψης	106
Εικόνα 70: AutoModel - επιλογή στήλης πρόβλεψης - νέος στόχος πρόβλεψης	106
Εικόνα 71: Κατανομή εγγραφών κλάσεων - νέος στόχος πρόβλεψης	107
Εικόνα 72: Αποτελέσματα DecisionTree - νέος στόχος πρόβλεψης	107
Εικόνα 73: Αποτελέσματα Random Forest - νέος στόχος πρόβλεψης	107
Εικόνα 74: Αποτελέσματα AutoModel - νέος στόχος πρόβλεψης	108
Εικόνα 75: Αποτελέσματα AutoModel για Decision Tree, Deep Learning, Random Forest και Naïve Bayes	110
Εικόνα 76: Αποτελέσματα για Decision Fast Large Margin, Gradient Boosting, Generalized Linear Regression	111
Εικόνα 77: Αποτελέσματα για Logistic Regression	112
Εικόνα 78: Αποτελέσματα για SVM	113

Λίστα Πινάκων

Πίνακας 1:Κλάσεις 1 ^{ης} παραλλαγής.....	74
Πίνακας 2: Κλάσεις 2ης παραλλαγής	75
Πίνακας 3: Κλάσεις 3ης παραλλαγής	75
Πίνακας 4: Συγκριτικές μέσες τιμές μετρικών	94
Πίνακας 5: Συγκριτικά αποτελέσματα (60% - 40%).....	97
Πίνακας 6: AutoModel - Μοντέλο με τα καλύτερα αποτελέσματα.....	103
Πίνακας 7: Συγκριτικά αποτελέσματα Decision Tree	104
Πίνακας 8: Συγκριτικά αποτελέσματα Random Forest	104
Πίνακας 9: Συγκριτικά αποτελέσματα DecisionTree – νέος στόχος πρόβλεψης	108
Πίνακας 10: Συγκριτικά αποτελέσματα Random Forest– νέος στόχος πρόβλεψης	108

1. ΕΙΣΑΓΩΓΗ

Στη σημερινή εποχή της πληροφορίας, η σύγκλιση της μηχανικής μάθησης και της ανάλυσης μεγάλων δεδομένων έχει φέρει επανάσταση στον τρόπο με τον οποίο οι οργανισμοί αντλούν αξία από τα δεδομένα τους. Η μηχανική μάθηση, ένα υποπεδίο της τεχνητής νοημοσύνης, δίνει τη δυνατότητα στα συστήματα να μαθαίνουν από τεράστιες ποσότητες δεδομένων, αναγνωρίζοντας πρότυπα και κάνοντας προβλέψεις χωρίς την ανάγκη ρητού προγραμματισμού. Όταν συνδυάζεται με την ανάλυση μεγάλων δεδομένων, η οποία ασχολείται με την επεξεργασία και ερμηνεία μεγάλων, σύνθετων συνόλων δεδομένων, οι αλγόριθμοι της μηχανικής μάθησης μπορούν να αποκαλύψουν πληροφορίες που προάγουν την καινοτομία, βελτιστοποιούν τις λειτουργίες και βελτιώνουν τις διαδικασίες λήψης αποφάσεων σε διάφορους τομείς.

Η ανάλυση μεγάλων δεδομένων παρέχει την πρώτη ύλη, τεράστιους όγκους δεδομένων, ποικίλων και μεταβαλλόμενων με γρήγορο ρυθμό που τροφοδοτεί τα μοντέλα της μηχανικής μάθησης. Μέσω προηγμένων τεχνικών όπως η εξόρυξη δεδομένων, η στατιστική ανάλυση και η προβλεπτική μοντελοποίηση, οι οργανισμοί μπορούν να επεξεργαστούν αυτά τα δεδομένα για να αποκαλύψουν τάσεις και συσχετισμούς που καθοδηγούν στρατηγικές αποφάσεις. Ωστόσο, η διαχείριση και η κατανόηση τόσο μεγάλου και ποικίλου όγκου δεδομένων απαιτεί ισχυρά εργαλεία, στα οποία το RapidMiner παίζει καθοριστικό ρόλο. Το RapidMiner απλοποιεί ολόκληρη τη ροή εργασίας της επιστήμης των δεδομένων, από την προετοιμασία και την οπτικοποίηση των δεδομένων μέχρι την κατασκευή και την υλοποίηση μοντέλων, καθιστώντας την προσβάσιμη σε χρήστες με διάφορα επίπεδα εξειδίκευσης.

Το RapidMiner λειτουργεί ως κρίσιμος σύνδεσμος μεταξύ της ανάλυσης μεγάλων δεδομένων και της μηχανικής μάθησης, προσφέροντας ένα διαισθητικό περιβάλλον όπου οι χρήστες μπορούν να αξιοποιήσουν τη δύναμη και των δύο πεδίων. Η εκτενής βιβλιοθήκη προκατασκευασμένων αλγορίθμων του, σε συνδυασμό με την ικανότητά του να

διαχειρίζεται μεγάλα σύνολα δεδομένων, επιτρέπει σε επιχειρήσεις και ερευνητές να κατασκευάζουν, να δοκιμάζουν και να υλοποιούν μοντέλα μηχανικής μάθησης γρήγορα.

Στα πλαίσια της συγκεκριμένης εργασίας, στόχος είναι να μελετηθεί πως μπορεί να χρησιμοποιηθεί το εργαλείο RapidMiner στην εφαρμογή μοντέλων μηχανικής μάθησης και την επεξεργασία μεγάλου όγκου δεδομένων. Η διάρθρωση της εργασίας έχει ως εξής:

Στο **δεύτερο κεφάλαιο** γίνεται παρουσίαση του θεωρητικού υποβάθρου, εστιάζοντας στην τεχνητή νοημοσύνη, μηχανική μάθηση και ανάλυση μεγάλων δεδομένων. Όσο αφορά και τα τρία πεδία πραγματοποιείται μια επισκόπηση των βασικών αρχών και διαδικασιών τους, του πεδίου εφαρμογής, των κύριων κατηγοριών τους, αλλά και των προκλήσεων που αντιμετωπίζουν.

Στο **τρίτο κεφάλαιο** πραγματοποιείται παρουσίαση της πειραματικής μεθοδολογίας που ακολουθήθηκε και παρουσιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε, το εργαλείο RapidMiner και οι διαδικασίες που ακολουθήθηκαν σε αυτό, οι μετρικές αξιολόγησης καθώς η παραμετροποίηση των μοντέλων.

Τέλος, στο **τέταρτο κεφάλαιο** παρουσιάζονται τα αποτελέσματα της πειραματικής διαδικασίας.

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1. Τεχνητή Νοημοσύνη

2.1.1. Ορισμός Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη [1] [2] είναι η προσομοίωση των ανθρώπινων διαδικασιών νοημοσύνης από μηχανές, ειδικά από συστήματα υπολογιστών. Παραδείγματα εφαρμογών της τεχνητής νοημοσύνης περιλαμβάνουν τα συστήματα ειδικών γνώσεων, την επεξεργασία φυσικής γλώσσας (NLP), την αναγνώριση ομιλίας και την μηχανική όραση.

Καθώς η δημοσιότητα γύρω από την τεχνητή νοημοσύνη έχει επιταχυνθεί, οι εκάστοτε προμηθευτές έχουν σπεύσει να προωθήσουν το πώς τα προϊόντα και οι υπηρεσίες τους ενσωματώνουν την τεχνητή νοημοσύνη. Συχνά, αυτό που αναφέρουν ως "AI" είναι μια καλά καθιερωμένη τεχνολογία όπως η μηχανική μάθηση.

Η τεχνητή νοημοσύνη απαιτεί εξειδικευμένο υλικό και λογισμικό για τη συγγραφή και την εκπαίδευση αλγορίθμων μηχανικής μάθησης. Δεν χρησιμοποιείται αποκλειστικά μία γλώσσα προγραμματισμού στην τεχνητή νοημοσύνη, αλλά οι Python, R, Java, C++ και Julia είναι όλες δημοφιλείς γλώσσες μεταξύ των προγραμματιστών AI.

2.1.2. Τρόπος λειτουργίας

Γενικά, τα συστήματα τεχνητής νοημοσύνης λειτουργούν [3] [4] καταναλώνοντας μεγάλες ποσότητες επισημασμένων δεδομένων εκπαίδευσης, αναλύοντας αυτά τα δεδομένα για συσχετίσεις και μοτίβα και χρησιμοποιώντας αυτά τα μοτίβα για να κάνουν προβλέψεις για μελλοντικές καταστάσεις.

Για παράδειγμα, ένα chatbot τεχνητής νοημοσύνης που τροφοδοτείται με παραδείγματα κειμένου μπορεί να μάθει να παράγει ρεαλιστικούς διαλόγους με ανθρώπους, και ένα εργαλείο αναγνώρισης εικόνων μπορεί να μάθει να αναγνωρίζει και να περιγράφει αντικείμενα σε εικόνες ανασκοπώντας εκατομμύρια παραδείγματα. Οι τεχνικές γεννητικής τεχνητής νοημοσύνης, οι οποίες έχουν εξελιχθεί ραγδαία τα τελευταία χρόνια, μπορούν να δημιουργούν ρεαλιστικά κείμενα, εικόνες, μουσική και άλλα μέσα.

Ο προγραμματισμός συστημάτων τεχνητής νοημοσύνης επικεντρώνεται σε γνωστικές δεξιότητες όπως οι εξής:

Μάθηση: Αυτή η πτυχή του προγραμματισμού AI περιλαμβάνει την απόκτηση δεδομένων και τη δημιουργία κανόνων, γνωστών ως αλγόριθμοι, για τη μετατροπή τους σε χρήσιμη πληροφορία. Αυτοί οι αλγόριθμοι παρέχουν στις συσκευές υπολογιστών βήμα προς βήμα οδηγίες για την ολοκλήρωση συγκεκριμένων εργασιών.

Συλλογιστική: Αυτή η πτυχή περιλαμβάνει την επιλογή του κατάλληλου αλγορίθμου για την επίτευξη ενός επιθυμητού αποτελέσματος.

Αυτοδιόρθωση: Αυτή η πτυχή περιλαμβάνει τους αλγόριθμους να μαθαίνουν και να προσαρμόζονται συνεχώς ώστε να παρέχουν τα πιο ακριβή αποτελέσματα.

Δημιουργικότητα: Αυτή η πτυχή χρησιμοποιεί νευρωνικά δίκτυα, συστήματα κανόνων, στατιστικές μεθόδους και άλλες τεχνικές τεχνητής νοημοσύνης για τη δημιουργία νέων εικόνων, κειμένων, μουσικής, ιδεών και άλλων.

2.1.3. Σημασίας Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη είναι σημαντική [5] [6] για το δυναμικό της να αλλάξει τον τρόπο που ζούμε, εργαζόμαστε και διασκεδάζουμε. Έχει χρησιμοποιηθεί αποτελεσματικά στις επιχειρήσεις για την αυτοματοποίηση εργασιών που παραδοσιακά εκτελούνταν από

ανθρώπους, συμπεριλαμβανομένων της εξυπηρέτησης πελατών, της παραγωγής leads, της ανίχνευσης απάτης και του ελέγχου ποιότητας.

Σε ορισμένους τομείς, η τεχνητή νοημοσύνη μπορεί να εκτελεί εργασίες πιο αποτελεσματικά και με μεγαλύτερη ακρίβεια από τους ανθρώπους. Είναι ιδιαίτερα χρήσιμη για επαναλαμβανόμενες, λεπτομερείς εργασίες, όπως η ανάλυση μεγάλου αριθμού νομικών εγγράφων για να διασφαλιστεί ότι τα σχετικά πεδία είναι σωστά συμπληρωμένα. Η ικανότητα της τεχνητής νοημοσύνης να επεξεργάζεται τεράστια σύνολα δεδομένων δίνει στις επιχειρήσεις πληροφορίες για τις λειτουργίες τους που μπορεί να μην είχαν παρατηρήσει διαφορετικά. Η ραγδαία ανάπτυξη των εργαλείων γεννητικής τεχνητής νοημοσύνης γίνεται επίσης σημαντική σε τομείς που κυμαίνονται από την εκπαίδευση έως το μάρκετινγκ και τον σχεδιασμό προϊόντων.

Οι εξελίξεις στις τεχνικές τεχνητής νοημοσύνης όχι μόνο έχουν συμβάλει σε μια έκρηξη στην αποδοτικότητα, αλλά έχουν ανοίξει το δρόμο για εντελώς νέες επιχειρηματικές ευκαιρίες για ορισμένες μεγάλες επιχειρήσεις. Για παράδειγμα, πριν από το σημερινό κύμα AI, θα ήταν δύσκολο να φανταστεί κανείς τη χρήση λογισμικού υπολογιστών για τη σύνδεση επιβατών με ταξί κατόπιν ζήτησης, όμως η Uber έχει γίνει μια εταιρεία Fortune 500 κάνοντας ακριβώς αυτό.

Η τεχνητή νοημοσύνη έχει αποτελέσει κεντρική επιλογή για πολλές από τις μεγαλύτερες και πιο επιτυχημένες εταιρείες σήμερα, συμπεριλαμβανομένων των Alphabet, Apple, Microsoft και Meta, οι οποίες χρησιμοποιούν την τεχνητή νοημοσύνη για να βελτιώσουν τις λειτουργίες τους και να ξεπεράσουν τους ανταγωνιστές τους. Στη θυγατρική της Alphabet, Google, για παράδειγμα, η τεχνητή νοημοσύνη είναι κεντρική για την ομώνυμη μηχανή αναζήτησης της, και η εταιρεία αυτοκινήτων χωρίς οδηγό Waymo ξεκίνησε ως τμήμα της Alphabet. Επιπλέον, το ερευνητικό εργαστήριο GoogleBrain επινόησε επίσης την αρχιτεκτονική του transformer, που υποστηρίζει τα πρόσφατα επιτεύγματα στην επεξεργασία φυσικής γλώσσας, όπως το ChatGPT της OpenAI.

2.1.4. Πλεονεκτήματα και μειονεκτήματα της τεχνητής νοημοσύνης

Η τεχνητή νοημοσύνη, ιδιαίτερα τα μοντέλα βαθιάς μάθησης όπως τα τεχνητά νευρωνικά δίκτυα, μπορεί να επεξεργαστεί μεγάλες ποσότητες δεδομένων πολύ πιο γρήγορα και να κάνει προβλέψεις πιο ακριβείς από ό,τι οι άνθρωποι. Ενώ ο τεράστιος όγκος δεδομένων που δημιουργείται καθημερινά θα ήταν αδύνατο να επεξεργαστεί ένας ανθρώπινος ερευνητής, οι εφαρμογές τεχνητής νοημοσύνης χρησιμοποιώντας μηχανική μάθηση μπορούν να λάβουν αυτά τα δεδομένα και να τα μετατρέψουν γρήγορα σε χρήσιμες πληροφορίες.

Ένα βασικό μειονέκτημα της τεχνητής νοημοσύνης είναι ότι είναι ακριβή στη διαδικασία των μεγάλων όγκων δεδομένων που απαιτεί. Καθώς οι τεχνικές τεχνητής νοημοσύνης ενσωματώνονται σε περισσότερα προϊόντα και υπηρεσίες, οι οργανισμοί πρέπει επίσης να προσέχουν τη δυνατότητα της τεχνητής νοημοσύνης να δημιουργεί μεροληπτικά συστήματα, είτε εσκεμμένα είτε ακούσια.

Τα ακόλουθα είναι μερικά πλεονεκτήματα [7] [8] της τεχνητής νοημοσύνης:

- **Ακρίβεια σε λεπτομερείς εργασίες:** Η ΑΙ είναι κατάλληλη για εργασίες που περιλαμβάνουν την αναγνώριση λεπτών προτύπων και σχέσεων σε δεδομένα που μπορεί να παραβλεφθούν από ανθρώπους. Για παράδειγμα, στην ογκολογία, τα συστήματα τεχνητής νοημοσύνης έχουν επιδείξει υψηλή ακρίβεια στην ανίχνευση καρκίνων σε πρώιμο στάδιο, όπως ο καρκίνος του μαστού και το μελάνωμα, επισημαίνοντας περιοχές ανησυχίας για περαιτέρω αξιολόγηση από επαγγελματίες υγείας.
- **Αποδοτικότητα σε εργασίες που απαιτούν μεγάλη ποσότητα δεδομένων:** Τα συστήματα ΑΙ και τα εργαλεία αυτοματοποίησης μειώνουν δραστικά το χρόνο που απαιτείται για την επεξεργασία δεδομένων. Αυτό είναι ιδιαίτερα χρήσιμο σε τομείς όπως οι χρηματοοικονομικές υπηρεσίες, η ασφάλιση και η υγειονομική περίθαλψη,

όπου εμπλέκεται μεγάλος όγκος επαναλαμβανόμενης καταχώρησης δεδομένων και ανάλυσης, καθώς και η λήψη αποφάσεων που βασίζεται σε δεδομένα.

- **Εξοικονόμηση χρόνου και αύξηση παραγωγικότητας:** Η τεχνητή νοημοσύνη και η ρομποτική μπορούν να αυτοματοποιήσουν τις λειτουργίες και να βελτιώσουν την ασφάλεια και την αποδοτικότητα. Για παράδειγμα, στη βιομηχανία, τα ρομπότ με τεχνητή νοημοσύνη χρησιμοποιούνται όλο και περισσότερο για να εκτελούν επικίνδυνες ή επαναλαμβανόμενες εργασίες ως μέρος της αυτοματοποίησης αποθηκών, μειώνοντας τον κίνδυνο για τους εργαζόμενους και αυξάνοντας την συνολική παραγωγικότητα.
- **Συνεπής αποτελέσματα:** Τα σημερινά εργαλεία ανάλυσης χρησιμοποιούν AI και μηχανική μάθηση για να επεξεργάζονται μεγάλα ποσά δεδομένων με ομοιόμορφο τρόπο, διατηρώντας παράλληλα τη δυνατότητα προσαρμογής σε νέες πληροφορίες μέσω συνεχούς μάθησης. Για παράδειγμα, οι εφαρμογές AI έχουν παράσχει συνεπή και αξιόπιστα αποτελέσματα στην αναθεώρηση νομικών εγγράφων και τη μετάφραση γλώσσας.
- **Εξατομίκευση και προσαρμογή:** Τα συστήματα AI μπορούν να βελτιώσουν την εμπειρία χρήστη εξατομικεύοντας τις αλληλεπιδράσεις και την παράδοση περιεχομένου σε ψηφιακές πλατφόρμες. Στις πλατφόρμες ηλεκτρονικού εμπορίου, για παράδειγμα, τα μοντέλα AI αναλύουν τη συμπεριφορά των χρηστών για να προτείνουν προϊόντα που ταιριάζουν στις προτιμήσεις του ατόμου, αυξάνοντας την ικανοποίηση και την αφοσίωση των πελατών.
- **Διαθεσιμότητα 24/7:** Τα προγράμματα τεχνητής νοημοσύνης δεν χρειάζονται ύπνο ή διαλείμματα. Για παράδειγμα, οι εικονικοί βοηθοί που τροφοδοτούνται από AI μπορούν να παρέχουν αδιάλειπτη εξυπηρέτηση πελατών 24/7, ακόμη και υπό μεγάλους όγκους αλληλεπιδράσεων, βελτιώνοντας τους χρόνους απόκρισης και μειώνοντας το κόστος.
- **Κλιμάκωση:** Τα συστήματα τεχνητής νοημοσύνης μπορούν να κλιμακωθούν για να χειριστούν αυξανόμενες ποσότητες εργασίας και δεδομένων. Αυτό καθιστά την

Αικατάλληλη για σενάρια όπου οι όγκοι δεδομένων και οι εργασίες μπορούν να αυξηθούν εκθετικά, όπως στην αναζήτηση στο διαδίκτυο και στην επιχειρηματική ανάλυση.

- **Επιτάχυνση έρευνας και ανάπτυξης:** Η τεχνητή νοημοσύνη μπορεί να επιταχύνει τον ρυθμό της έρευνας και ανάπτυξης σε τομείς όπως η φαρμακολογία και η επιστήμη των υλικών. Με την ταχεία προσομοίωση και ανάλυση πολλών πιθανών σεναρίων, τα μοντέλα τεχνητής νοημοσύνης μπορούν να βοηθήσουν τους ερευνητές να ανακαλύψουν νέα φάρμακα, υλικά ή ενώσεις πιο γρήγορα από τις παραδοσιακές μεθόδους.
- **Βιωσιμότητα και διατήρηση:** Η τεχνητή νοημοσύνη και η μηχανική μάθηση χρησιμοποιούνται ολοένα και περισσότερο για την παρακολούθηση περιβαλλοντικών αλλαγών, την πρόβλεψη μελλοντικών καιρικών φαινομένων και τη διαχείριση προσπαθειών διατήρησης. Τα μοντέλα μηχανικής μάθησης μπορούν να επεξεργάζονται δορυφορικές εικόνες και δεδομένα από αισθητήρες για να παρακολουθούν τον κίνδυνο πυρκαγιάς, τα επίπεδα ρύπανσης και τους πληθυσμούς απειλούμενων ειδών.
- **Βελτιστοποίηση διαδικασιών:** Η τεχνητή νοημοσύνη χρησιμοποιείται για την απλοποίηση και αυτοματοποίηση σύνθετων διαδικασιών σε διάφορες βιομηχανίες. Για παράδειγμα, τα μοντέλα AI μπορούν να εντοπίσουν αναποτελεσματικότητες και να προβλέψουν σημεία συμφόρησης στις ροές εργασίας παραγωγής, ενώ στον τομέα της ενέργειας μπορούν να προβλέψουν τη ζήτηση ηλεκτρικής ενέργειας και να καταναείμουν την προσφορά σε πραγματικό χρόνο.

Συνεχίζοντας, τα ακόλουθα είναι μερικά μειονεκτήματα [7] [8] της τεχνητής νοημοσύνης:

- **Υψηλό κόστος:** Η ανάπτυξη της τεχνητής νοημοσύνης μπορεί να είναι πολύ ακριβή. Η κατασκευή ενός μοντέλου AI απαιτεί σημαντική αρχική επένδυση σε υποδομές, υπολογιστικούς πόρους και λογισμικό για την εκπαίδευση του μοντέλου και την αποθήκευση των δεδομένων εκπαίδευσης. Μετά την αρχική εκπαίδευση,

υπάρχουν περαιτέρω συνεχή κόστη που σχετίζονται με την εφαρμογή του μοντέλου και την επανακατάρτιση του. Ως αποτέλεσμα, τα κόστη μπορούν να αυξηθούν γρήγορα, ιδιαίτερα για προηγμένα, σύνθετα συστήματα όπως οι εφαρμογές γεννητικής ΑΙ.

- **Τεχνική πολυπλοκότητα:** Η ανάπτυξη, η λειτουργία και η αντιμετώπιση προβλημάτων των συστημάτων ΑΙ -- ιδιαίτερα σε πραγματικά περιβάλλοντα παραγωγής -- απαιτεί μεγάλο βαθμό τεχνικής γνώσης. Σε πολλές περιπτώσεις, αυτή η γνώση διαφέρει από αυτήν που απαιτείται για την κατασκευή μη-τεχνητών νοημοσύνης λογισμικών.
- **Έλλειψη ταλέντου:** Επιπροσθέτως του προβλήματος της τεχνικής πολυπλοκότητας, υπάρχει σημαντική έλλειψη επαγγελματιών εκπαιδευμένων στην ΑΙ και στη μηχανική μάθηση σε σύγκριση με τη συνεχώς αυξανόμενη ζήτηση για τέτοιες δεξιότητες. Αυτό το κενό μεταξύ προσφοράς και ζήτησης ταλέντων ΑΙ σημαίνει ότι, παρόλο που το ενδιαφέρον για τις εφαρμογές ΑΙ αυξάνεται, πολλές οργανώσεις δεν μπορούν να βρουν αρκετά εξειδικευμένους εργαζόμενους για να στελεχώσουν τις πρωτοβουλίες τους.
- **Αλγοριθμική μεροληψία:** Οι αλγόριθμοι τεχνητής νοημοσύνης και μηχανικής μάθησης αντανακλούν τις μεροληψίες που υπάρχουν στα δεδομένα εκπαίδευσής τους. Σε ορισμένες περιπτώσεις, τα συστήματα τεχνητής νοημοσύνης μπορεί ακόμη και να ενισχύσουν τις λεπτές μεροληψίες στα δεδομένα εκπαίδευσης, ενσωματώνοντάς τις σε ενισχυτικά και ψευδοαντικειμενικά πρότυπα.
- **Δυσκολία γενίκευσης:** Τα μοντέλα ΑΙ συχνά υπερέχουν στις συγκεκριμένες εργασίες για τις οποίες εκπαιδεύτηκαν, αλλά δυσκολεύονται όταν καλούνται να αντιμετωπίσουν νέες καταστάσεις. Αυτή η έλλειψη ευελιξίας μπορεί να περιορίσει τη χρησιμότητα της ΑΙ, καθώς νέες εργασίες μπορεί να απαιτούν την ανάπτυξη ενός εντελώς νέου μοντέλου.
- **Απώλεια θέσεων εργασίας:** Η τεχνητή νοημοσύνη μπορεί να οδηγήσει σε απώλεια θέσεων εργασίας εάν οι οργανισμοί αντικαταστήσουν ανθρώπινους εργαζόμενους

με μηχανές -- μια αυξανόμενη ανησυχία καθώς οι δυνατότητες των μοντέλων AI γίνονται πιο εξελιγμένες και οι εταιρείες αναζητούν όλο και περισσότερο την αυτοματοποίηση ροών εργασίας χρησιμοποιώντας AI.

- **Ευπάθειες ασφάλειας:** Τα συστήματα τεχνητής νοημοσύνης είναι ευάλωτα σε ευρύ φάσμα κυβερνοαπειλών, όπως η δηλητηρίαση δεδομένων και η αντίπαλη μηχανική μάθηση. Οι κακόβολοι χρήστες μπορούν να εξάγουν ευαίσθητα δεδομένα εκπαίδευσης από ένα μοντέλο AI ή να εξαπατήσουν τα συστήματα AI ώστε να παράγουν λανθασμένα και επιβλαβή έξοδα.
- **Περιβαλλοντικός αντίκτυπος:** Τα κέντρα δεδομένων και οι υποδομές δικτύου που υποστηρίζουν τη λειτουργία των μοντέλων AI καταναλώνουν μεγάλες ποσότητες ενέργειας και νερού. Κατά συνέπεια, η εκπαίδευση και η λειτουργία των μοντέλων AI έχει σημαντικό αντίκτυπο στο κλίμα.
- **Νομικά ζητήματα:** Η τεχνητή νοημοσύνη θέτει σύνθετα ερωτήματα γύρω από την ιδιωτικότητα και την νομική ευθύνη, ιδιαίτερα σε ένα εξελισσόμενο νομικό πλαίσιο που διαφέρει ανάλογα με την περιοχή. Για παράδειγμα, η χρήση AI για την ανάλυση και λήψη αποφάσεων βάσει προσωπικών δεδομένων έχει σοβαρές επιπτώσεις στην ιδιωτικότητα, και παραμένει ασαφές πώς τα δικαστήρια θα αντιμετωπίσουν τη συγγραφή υλικού που παράγεται από μοντέλα AI που έχουν εκπαιδευτεί σε έργα με πνευματικά δικαιώματα.

2.1.5. Δυνατή και Αδύνατη AI

Η τεχνητή νοημοσύνη (AI) μπορεί γενικά να κατηγοριοποιηθεί σε δύο τύπους [9] [10]: Στενή (ή Αδύνατη) AI και Γενική (ή Δυνατή) AI.

Στενή AI (Αδύνατη AI):

Ορισμός: Αυτή η μορφή τεχνητής νοημοσύνης αναφέρεται σε μοντέλα που έχουν εκπαιδευτεί για να εκτελούν συγκεκριμένες εργασίες. Η Στενή τεχνητή νοημοσύνη λειτουργεί εντός του πλαισίου των εργασιών για τις οποίες έχει προγραμματιστεί, χωρίς τη δυνατότητα να γενικεύει ευρέως ή να μαθαίνει πέρα από τον αρχικό προγραμματισμό της.

Παραδείγματα: Παραδείγματα Στενής AI περιλαμβάνουν τους εικονικούς βοηθούς, όπως το Apple Siri και το Amazon Alexa, καθώς και τις μηχανές σύστασης, όπως αυτές που βρίσκονται σε πλατφόρμες streaming, όπως το Spotify και το Netflix.

Γενική AI (Δυνατή AI):

Ορισμός: Αυτός ο τύπος, που δεν υπάρχει προς το παρόν, αναφέρεται συχνά ως Τεχνητή Γενική Νοημοσύνη (AGI). Αν δημιουργηθεί, η AGI θα ήταν ικανή να εκτελεί οποιαδήποτε διανοητική εργασία που μπορεί να εκτελέσει ένας άνθρωπος. Για να το πετύχει αυτό, η AGI θα πρέπει να έχει τη δυνατότητα να εφαρμόζει λογική σε ένα ευρύ φάσμα τομέων για να κατανοεί σύνθετα προβλήματα που δεν έχουν προγραμματιστεί ειδικά για να επιλύσει. Αυτό, με τη σειρά του, θα απαιτούσε κάτι που στην τεχνητή νοημοσύνη είναι γνωστό ως ασαφής λογική (fuzzy logic): μια προσέγγιση που επιτρέπει γκρίζες περιοχές και διαβαθμίσεις αβεβαιότητας, αντί για δυαδικά, ασπρόμαυρα αποτελέσματα.

Σημαντικότητα: Το ερώτημα αν μπορεί να δημιουργηθεί η AGI - και οι συνέπειες αυτού - παραμένει έντονα συζητήσιμο μεταξύ των ειδικών στην τεχνητή νοημοσύνη. Ακόμη και οι πιο προηγμένες τεχνολογίες AI σήμερα, όπως το ChatGPT και άλλα εξαιρετικά ικανά μεγάλα γλωσσικά μοντέλα (LLMs), δεν επιδεικνύουν γνωστικές ικανότητες ισοδύναμες με αυτές των ανθρώπων και δεν μπορούν να γενικεύσουν σε ποικίλες καταστάσεις. Για παράδειγμα, το ChatGPT έχει σχεδιαστεί για την παραγωγή φυσικής γλώσσας και δεν είναι ικανό να ξεπεράσει τον αρχικό προγραμματισμό του για να εκτελέσει εργασίες όπως η σύνθετη μαθηματική λογική.

2.1.6. Οι 4 τύποι τεχνητής νοημοσύνης

Η τεχνητή νοημοσύνη μπορεί να κατηγοριοποιηθεί σε τέσσερις τύπους, ξεκινώντας από τα ευρέως χρησιμοποιούμενα έξυπνα συστήματα που εκτελούν συγκεκριμένες εργασίες και προχωρώντας προς τα αισθαντικά συστήματα, τα οποία δεν υπάρχουν ακόμη.

Οι κατηγορίες είναι οι εξής [11] [12]:

- **Τύπος 1 - Αντιδραστικές μηχανές:** Αυτά τα συστήματα AI δεν έχουν μνήμη και είναι συγκεκριμένα για συγκεκριμένες εργασίες. Ένα παράδειγμα είναι το DeepBlue, το πρόγραμμα σκακιού της IBM που νίκησε τον Ρώσο σκακιστή Garry Kasparov τη δεκαετία του 1990. Το DeepBlue μπορούσε να αναγνωρίσει τα κομμάτια σε μια σκακιέρα και να πραγματοποιεί προβλέψεις, αλλά επειδή δεν είχε μνήμη, δεν μπορούσε να χρησιμοποιήσει προηγούμενες εμπειρίες για να ενημερώσει τις μελλοντικές του κινήσεις.
- **Τύπος 2 - Περιορισμένη μνήμη:** Αυτά τα συστήματα AI διαθέτουν μνήμη, ώστε να μπορούν να χρησιμοποιούν προηγούμενες εμπειρίες για να ενημερώσουν μελλοντικές αποφάσεις. Ορισμένες από τις λειτουργίες λήψης αποφάσεων στα αυτοοδηγούμενα αυτοκίνητα έχουν σχεδιαστεί με αυτόν τον τρόπο.
- **Τύπος 3 - Θεωρία του νου:** Η θεωρία του νου είναι ένας όρος της ψυχολογίας. Όταν εφαρμόζεται στην AI, αναφέρεται σε ένα σύστημα ικανό να κατανοεί τα συναισθήματα. Αυτός ο τύπος AI μπορεί να αντλήσει ανθρώπινες προθέσεις και να προβλέψει συμπεριφορές, μια απαραίτητη δεξιότητα για τα συστήματα AI ώστε να γίνουν αναπόσπαστα μέλη ομάδων που παραδοσιακά αποτελούνται από ανθρώπους.
- **Τύπος 4 - Αυτογνωσία:** Σε αυτή την κατηγορία, τα συστήματα AI έχουν μια αίσθηση του εαυτού τους, που τους δίνει συνείδηση. Οι μηχανές με αυτογνωσία κατανοούν την τρέχουσα κατάστασή τους. Αυτός ο τύπος AI δεν υπάρχει ακόμα.

2.1.7. Παραδείγματα τεχνολογιών τεχνητής νοημοσύνης

Οι τεχνολογίες AI μπορούν να βελτιώσουν τις λειτουργίες των υπαρχόντων εργαλείων και να αυτοματοποιήσουν διάφορες εργασίες και διαδικασίες, επηρεάζοντας πολλές πτυχές της καθημερινής ζωής. Ακολουθούν μερικά εξέχοντα παραδείγματα [13] [14].

Αυτοματοποίηση: Η AI ενισχύει τις τεχνολογίες αυτοματοποίησης επεκτείνοντας το εύρος, την πολυπλοκότητα και τον αριθμό των εργασιών που μπορούν να αυτοματοποιηθούν. Ένα παράδειγμα είναι η ρομποτική αυτοματοποίηση διαδικασιών (RPA), η οποία αυτοματοποιεί επαναλαμβανόμενες, βασισμένες σε κανόνες εργασίες επεξεργασίας δεδομένων που παραδοσιακά εκτελούνταν από ανθρώπους. Επειδή η AI βοηθά τα ρομπότ RPA να προσαρμόζονται σε νέα δεδομένα και να ανταποκρίνονται δυναμικά σε αλλαγές διαδικασίας, η ενσωμάτωση δυνατοτήτων AI και μηχανικής μάθησης επιτρέπει στην RPA να διαχειρίζεται πιο σύνθετες ροές εργασίας.

Μηχανική μάθηση: Η μηχανική μάθηση είναι η επιστήμη της διδασκαλίας υπολογιστών να μαθαίνουν από δεδομένα και να λαμβάνουν αποφάσεις χωρίς να είναι ρητά προγραμματισμένοι για να το κάνουν. Η βαθιά μάθηση, ένας υποτομέας της μηχανικής μάθησης, χρησιμοποιεί εξελιγμένα νευρωνικά δίκτυα για να πραγματοποιήσει αυτό που ουσιαστικά είναι μια προηγμένη μορφή προβλεπτικής ανάλυσης.

Οι αλγόριθμοι μηχανικής μάθησης μπορούν γενικά να κατηγοριοποιηθούν σε τρεις κατηγορίες: επιβλεπόμενη μάθηση, μη επιβλεπόμενη μάθηση και ενισχυτική μάθηση.

- Η επιβλεπόμενη μάθηση εκπαιδεύει μοντέλα σε σύνολα δεδομένων με ετικέτες, επιτρέποντάς τους να αναγνωρίζουν με ακρίβεια πρότυπα, να προβλέπουν αποτελέσματα ή να ταξινομούν νέα δεδομένα.
- Η μη επιβλεπόμενη μάθηση εκπαιδεύει μοντέλα να ταξινομούν δεδομένα χωρίς ετικέτες για να βρουν υποκείμενες σχέσεις ή ομάδες.

- Η ενισχυτική μάθηση ακολουθεί μια διαφορετική προσέγγιση, όπου τα μοντέλα μαθαίνουν να λαμβάνουν αποφάσεις δρώντας ως πράκτορες και λαμβάνοντας ανατροφοδότηση για τις ενέργειές τους.

Υπάρχει επίσης η ημι-επιβλεπόμενη μάθηση, η οποία συνδυάζει πτυχές των επιβλεπόμενων και μη επιβλεπόμενων προσεγγίσεων. Αυτή η τεχνική χρησιμοποιεί μια μικρή ποσότητα δεδομένων με ετικέτες και μια μεγαλύτερη ποσότητα δεδομένων χωρίς ετικέτες, βελτιώνοντας έτσι την ακρίβεια της μάθησης ενώ μειώνει την ανάγκη για δεδομένα με ετικέτες, τα οποία μπορεί να είναι χρονοβόρα και εργασιοβόρα.

Υπολογιστική όραση: Η υπολογιστική όραση είναι ένας τομέας της τεχνητής νοημοσύνης που επικεντρώνεται στη διδασκαλία των μηχανών πώς να ερμηνεύουν τον οπτικό κόσμο. Αναλύοντας οπτικές πληροφορίες, όπως εικόνες και βίντεο από κάμερες, χρησιμοποιώντας μοντέλα βαθιάς μάθησης, τα συστήματα υπολογιστικής όρασης μπορούν να μάθουν να αναγνωρίζουν και να ταξινομούν αντικείμενα και να λαμβάνουν αποφάσεις βάσει αυτών των αναλύσεων.

Ο κύριος στόχος της υπολογιστικής όρασης είναι να αναπαράγει ή να βελτιώσει το ανθρώπινο οπτικό σύστημα χρησιμοποιώντας αλγόριθμους AI. Η υπολογιστική όραση χρησιμοποιείται σε ένα ευρύ φάσμα εφαρμογών, από την αναγνώριση υπογραφών έως την ανάλυση ιατρικών εικόνων και τα αυτόνομα οχήματα. Ο όρος "μηχανική όραση" αναφέρεται ειδικά στη χρήση της υπολογιστικής όρασης για την ανάλυση δεδομένων από κάμερες και βίντεο σε βιομηχανικά πλαίσια αυτοματοποίησης, όπως οι διαδικασίες παραγωγής στη βιομηχανία.

Επεξεργασία φυσικής γλώσσας (NLP): Η επεξεργασία φυσικής γλώσσας αναφέρεται στην επεξεργασία της ανθρώπινης γλώσσας από προγράμματα υπολογιστών. Οι αλγόριθμοι NLP μπορούν να ερμηνεύουν και να αλληλεπιδρούν με την ανθρώπινη γλώσσα, εκτελώντας εργασίες όπως η μετάφραση, η αναγνώριση ομιλίας και η ανάλυση συναισθημάτων. Ένα από τα παλαιότερα και πιο γνωστά παραδείγματα NLP είναι η ανίχνευση ανεπιθύμητης αλληλογραφίας, η οποία εξετάζει τη γραμμή θέματος και το κείμενο ενός email και

αποφασίζει αν είναι ανεπιθύμητο. Πιο προχωρημένες εφαρμογές του NLP περιλαμβάνουν μεγάλα γλωσσικά μοντέλα (LLMs) όπως το ChatGPT και το Claude της Anthropic.

Ρομποτική: Η ρομποτική είναι ένας τομέας της μηχανικής που επικεντρώνεται στον σχεδιασμό, την κατασκευή και τη λειτουργία ρομπότ: αυτοματοποιημένων μηχανών που αναπαράγουν και αντικαθιστούν ανθρώπινες ενέργειες, ιδιαίτερα εκείνες που είναι δύσκολες, επικίνδυνες ή κουραστικές για τους ανθρώπους. Παραδείγματα εφαρμογών ρομποτικής περιλαμβάνουν τη βιομηχανία, όπου τα ρομπότ εκτελούν επαναλαμβανόμενες ή επικίνδυνες εργασίες στη γραμμή παραγωγής, και τις εξερευνητικές αποστολές σε μακρινές, δύσκολα προσβάσιμες περιοχές όπως το διάστημα και τα βαθιά νερά.

Η ενσωμάτωση της AI και της μηχανικής μάθησης επεκτείνει σημαντικά τις δυνατότητες των ρομπότ, επιτρέποντάς τους να λαμβάνουν πιο ενημερωμένες αυτόνομες αποφάσεις και να προσαρμόζονται σε νέες καταστάσεις και δεδομένα. Για παράδειγμα, τα ρομπότ με δυνατότητες μηχανικής όρασης μπορούν να μάθουν να ταξινομούν αντικείμενα σε μια γραμμή παραγωγής ανάλογα με το σχήμα και το χρώμα τους.

Αυτόνομα οχήματα: Τα αυτόνομα οχήματα, γνωστά και ως αυτοκινούμενα αυτοκίνητα, μπορούν να αντιλαμβάνονται και να πλοηγούνται στο περιβάλλον τους με ελάχιστη ή καθόλου ανθρώπινη παρέμβαση. Αυτά τα οχήματα βασίζονται σε έναν συνδυασμό τεχνολογιών, συμπεριλαμβανομένων των ραντάρ, του GPS και μιας σειράς αλγορίθμων AI και μηχανικής μάθησης, όπως η αναγνώριση εικόνων.

Αυτοί οι αλγόριθμοι μαθαίνουν από δεδομένα πραγματικής οδήγησης, κυκλοφορίας και χαρτών για να λαμβάνουν ενημερωμένες αποφάσεις σχετικά με το πότε πρέπει να φρενάρουν, να στρίψουν και να επιταχύνουν. Αν και η τεχνολογία έχει προχωρήσει σημαντικά τα τελευταία χρόνια, ο απόλυτος στόχος ενός αυτόνομου οχήματος που μπορεί να αντικαταστήσει πλήρως έναν ανθρώπινο οδηγό δεν έχει ακόμη επιτευχθεί.

Γενετική τεχνητή νοημοσύνη: Ο όρος γενετική τεχνητή νοημοσύνη αναφέρεται σε συστήματα μηχανικής μάθησης που μπορούν να δημιουργήσουν νέα δεδομένα από

κειμενικές εντολές - πιο συχνά κείμενα και εικόνες, αλλά και ήχο, βίντεο, λογισμικό και ακόμη και γενετικές ακολουθίες και δομές πρωτεϊνών. Μέσω της εκπαίδευσης σε τεράστια σύνολα δεδομένων, αυτοί οι αλγόριθμοι μαθαίνουν σταδιακά τα μοτίβα των τύπων μέσω των οποίων θα τους ζητηθεί να δημιουργήσουν, επιτρέποντάς τους αργότερα να δημιουργήσουν νέο περιεχόμενο που μοιάζει με τα δεδομένα εκπαίδευσης.

Η γενετική τεχνητή νοημοσύνη γνώρισε ραγδαία αύξηση της δημοτικότητας μετά την εισαγωγή ευρέως διαθέσιμων εργαλείων δημιουργίας κειμένου και εικόνας το 2022, όπως το ChatGPT, το Dall-E και το Midjourney, και εφαρμόζεται ολοένα και περισσότερο σε επιχειρηματικά πλαίσια. Παρόλο που οι δυνατότητες πολλών εργαλείων γεννητικής τεχνητής νοημοσύνης είναι εντυπωσιακές, εγείρουν επίσης ανησυχίες σχετικά με θέματα όπως τα πνευματικά δικαιώματα, η δίκαιη χρήση και η ασφάλεια, τα οποία παραμένουν ανοικτά προς συζήτηση στον τομέα της τεχνολογίας.

2.1.8. Εφαρμογές της τεχνητής νοημοσύνης

Η τεχνητή νοημοσύνη έχει εισχωρήσει σε ποικίλους τομείς της βιομηχανίας και της έρευνας. Ακολουθούν ορισμένα από τα πιο αξιοσημείωτα παραδείγματα [15] [16].

Τεχνητή νοημοσύνη στην υγειονομική περίθαλψη: Η AI εφαρμόζεται σε μια σειρά από εργασίες στον τομέα της υγειονομικής περίθαλψης, με κύριους στόχους τη βελτίωση των αποτελεσμάτων των ασθενών και τη μείωση των συστημικών δαπανών. Μια σημαντική εφαρμογή είναι η χρήση μοντέλων μηχανικής μάθησης που έχουν εκπαιδευτεί σε μεγάλα σύνολα ιατρικών δεδομένων για να βοηθήσουν τους επαγγελματίες υγείας να κάνουν καλύτερες και ταχύτερες διαγνώσεις. Για παράδειγμα, λογισμικό με δυνατότητες AI μπορεί να αναλύσει αξονικές τομογραφίες (CTscans) και να ειδοποιήσει τους νευρολόγους για ύποπτα περιστατικά εγκεφαλικών.

Στην πλευρά των ασθενών, οι διαδικτυακοί εικονικοί βοηθοί υγείας και τα chatbots μπορούν να παρέχουν γενικές ιατρικές πληροφορίες, να προγραμματίζουν ραντεβού, να εξηγούν τις διαδικασίες χρέωσης και να ολοκληρώνουν άλλες διοικητικές εργασίες. Τα μοντέλα πρόβλεψης AI μπορούν επίσης να χρησιμοποιηθούν για την καταπολέμηση της εξάπλωσης πανδημιών όπως η COVID-19.

Τεχνητή νοημοσύνη στις επιχειρήσεις: Η τεχνητή νοημοσύνη ενσωματώνεται όλο και περισσότερο σε διάφορες επιχειρηματικές λειτουργίες και βιομηχανίες, με στόχο τη βελτίωση της αποδοτικότητας, της εμπειρίας των πελατών, του στρατηγικού σχεδιασμού και της λήψης αποφάσεων. Για παράδειγμα, τα μοντέλα μηχανικής μάθησης υποστηρίζουν πολλές από τις σημερινές πλατφόρμες ανάλυσης δεδομένων και διαχείρισης σχέσεων πελατών (CRM), βοηθώντας τις εταιρείες να κατανοήσουν πώς να εξυπηρετούν καλύτερα τους πελάτες τους μέσω της εξατομίκευσης των προσφορών και της παροχής πιο στοχευμένου μάρκετινγκ.

Οι εικονικοί βοηθοί και τα chatbots αναπτύσσονται επίσης σε εταιρικούς ιστότοπους και σε κινητές εφαρμογές για την παροχή 24ωρης εξυπηρέτησης πελατών και για να απαντούν σε συνήθεις ερωτήσεις. Επιπλέον, όλο και περισσότερες εταιρείες εξερευνούν τις δυνατότητες των εργαλείων γεννητικής AI, όπως το ChatGPT, για την αυτοματοποίηση εργασιών όπως η σύνταξη και η σύνοψη εγγράφων, ο σχεδιασμός προϊόντων και η παραγωγή ιδεών, καθώς και ο προγραμματισμός υπολογιστών.

Τεχνητή νοημοσύνη στην εκπαίδευση: Η AI έχει πολλές πιθανές εφαρμογές στην εκπαιδευτική τεχνολογία. Μπορεί να αυτοματοποιήσει πτυχές των διαδικασιών βαθμολόγησης, δίνοντας στους εκπαιδευτικούς περισσότερο χρόνο για άλλες εργασίες. Τα εργαλεία AI μπορούν επίσης να αξιολογήσουν την απόδοση των μαθητών και να προσαρμοστούν στις ατομικές τους ανάγκες, διευκολύνοντας πιο εξατομικευμένες εμπειρίες μάθησης που επιτρέπουν στους μαθητές να προχωρούν με τον δικό τους ρυθμό. Οι ψηφιακοί δάσκαλοι AI θα μπορούσαν επίσης να παρέχουν επιπλέον υποστήριξη στους μαθητές, εξασφαλίζοντας ότι παραμένουν στο σωστό δρόμο. Η τεχνολογία αυτή θα

μπορούσε επίσης να αλλάξει τον τρόπο και το μέρος που οι μαθητές μαθαίνουν, ίσως αλλάζοντας τον παραδοσιακό ρόλο των εκπαιδευτικών.

Καθώς οι δυνατότητες των μεγάλων γλωσσικών μοντέλων (LLMs) όπως το ChatGPT και το Google Gemini αυξάνονται, τέτοια εργαλεία θα μπορούσαν να βοηθήσουν τους εκπαιδευτικούς να δημιουργήσουν εκπαιδευτικό υλικό και να εμπλέξουν τους μαθητές με νέους τρόπους. Ωστόσο, η έλευση αυτών των εργαλείων αναγκάζει επίσης τους εκπαιδευτικούς να επανεξετάσουν τις πρακτικές εργασίας στο σπίτι και των εξετάσεων και να αναθεωρήσουν τις πολιτικές λογοκλοπής.

Τεχνητή νοημοσύνη στα χρηματοοικονομικά και στις τράπεζες: Οι τράπεζες και άλλοι χρηματοοικονομικοί οργανισμοί χρησιμοποιούν την AI για να βελτιώσουν τη λήψη αποφάσεων για εργασίες όπως η χορήγηση δανείων, ο καθορισμός πιστωτικών ορίων και ο εντοπισμός επενδυτικών ευκαιριών. Επιπλέον, η αλγοριθμική διαπραγμάτευση, που υποστηρίζεται από προηγμένη τεχνητή νοημοσύνη και μηχανική μάθηση, έχει μεταμορφώσει τις χρηματοοικονομικές αγορές, εκτελώντας συναλλαγές με ταχύτητες και αποδοτικότητα που υπερβαίνουν κατά πολύ αυτό που θα μπορούσαν να κάνουν οι άνθρωποι έμποροι χειροκίνητα.

Η τεχνητή νοημοσύνη και η μηχανική μάθηση έχουν επίσης εισχωρήσει στο χώρο των καταναλωτικών χρηματοοικονομικών. Για παράδειγμα, οι τράπεζες χρησιμοποιούν chatbots AI για να ενημερώνουν τους πελάτες για υπηρεσίες και προσφορές και για να χειρίζονται συναλλαγές και ερωτήσεις που δεν απαιτούν ανθρώπινη παρέμβαση. Παρομοίως, η Intuit προσφέρει χαρακτηριστικά γεννητικής AI στο προϊόν ηλεκτρονικής κατάθεσης TurboTax, τα οποία παρέχουν στους χρήστες εξατομικευμένες συμβουλές με βάση δεδομένα όπως το φορολογικό προφίλ του χρήστη και τον φορολογικό κώδικα της περιοχής τους.

Τεχνητή νοημοσύνη στο νομικό τομέα: Η AI αλλάζει τον νομικό τομέα αυτοματοποιώντας εργασίες που απαιτούν έντονη εργασία, όπως η αναθεώρηση εγγράφων και η απάντηση σε αιτήματα ανεύρεσης δεδομένων (discovery), τα οποία μπορεί να είναι κουραστικά και

χρονοβόρα για δικηγόρους και παρα-νομικούς. Σήμερα, τα δικηγορικά γραφεία χρησιμοποιούν την AI και τη μηχανική μάθηση για μια ποικιλία εργασιών, συμπεριλαμβανομένης της ανάλυσης δεδομένων και νομολογίας, της ταξινόμησης και της εξαγωγής πληροφοριών από έγγραφα μέσω υπολογιστικής όρασης, και της ερμηνείας και απάντησης σε αιτήματα ανεύρεσης δεδομένων μέσω NLP.

Εκτός από τη βελτίωση της αποδοτικότητας και της παραγωγικότητας, αυτή η ενσωμάτωση της AI επιτρέπει στους ανθρώπους νομικούς επαγγελματίες να αφιερώνουν περισσότερο χρόνο στους πελάτες τους και να επικεντρώνονται σε πιο δημιουργικές και στρατηγικές εργασίες που η AI είναι λιγότερο κατάλληλη να χειριστεί. Με την άνοδο της γενετικής AI στον νομικό τομέα, τα δικηγορικά γραφεία εξετάζουν επίσης τη χρήση των LLMs για τη σύνταξη κοινών εγγράφων, όπως οι τυποποιημένες συμβάσεις.

Τεχνητή νοημοσύνη στην ανάπτυξη λογισμικού και το IT: Η AI χρησιμοποιείται για την αυτοματοποίηση πολλών διαδικασιών στην ανάπτυξη λογισμικού, το DevOps και το IT. Για παράδειγμα, τα εργαλεία AI Ops επιτρέπουν την προβλεπτική συντήρηση των περιβαλλόντων IT αναλύοντας τα δεδομένα συστήματος για να προβλέψουν πιθανά προβλήματα πριν εμφανιστούν, και τα εργαλεία παρακολούθησης με AI μπορούν να βοηθήσουν στον εντοπισμό πιθανών ανωμαλιών σε πραγματικό χρόνο βάσει ιστορικών δεδομένων συστήματος. Τα εργαλεία γεννητικής AI όπως το GitHubCopilot και το Tabnine χρησιμοποιούνται επίσης όλο και περισσότερο για την παραγωγή κώδικα εφαρμογών βάσει εντολών φυσικής γλώσσας. Ενώ αυτά τα εργαλεία έχουν δείξει αρχικά υποσχέσεις και ενδιαφέρον μεταξύ των προγραμματιστών, είναι απίθανο να αντικαταστήσουν πλήρως τους μηχανικούς λογισμικού. Αντίθετα, χρησιμεύουν ως χρήσιμα εργαλεία παραγωγικότητας, αυτοματοποιώντας επαναλαμβανόμενες εργασίες και τη συγγραφή τυποποιημένου κώδικα.

Τεχνητή νοημοσύνη στην ασφάλεια: Η τεχνητή νοημοσύνη και η μηχανική μάθηση είναι συχνά λέξεις-κλειδιά στο μάρκετινγκ των προμηθευτών ασφάλειας, επομένως οι αγοραστές θα πρέπει να είναι προσεκτικοί. Ωστόσο, η τεχνητή νοημοσύνη είναι πράγματι

μια χρήσιμη τεχνολογία σε πολλούς τομείς της κυβερνοασφάλειας, όπως η ανίχνευση ανωμαλιών, η μείωση των ψευδώς θετικών αποτελεσμάτων και η διεξαγωγή ανάλυσης απειλών συμπεριφοράς. Για παράδειγμα, οι οργανισμοί χρησιμοποιούν μηχανική μάθηση σε λογισμικό διαχείρισης πληροφοριών και συμβάντων ασφαλείας (SIEM) για να ανιχνεύσουν ύποπτες δραστηριότητες και πιθανές απειλές. Αναλύοντας τεράστιες ποσότητες δεδομένων και αναγνωρίζοντας πρότυπα που μοιάζουν με γνωστό κακόβουλο κώδικα, τα εργαλεία AI μπορούν να ειδοποιούν τις ομάδες ασφαλείας για νέες και αναδυόμενες επιθέσεις, συχνά πολύ νωρίτερα από ό,τι θα μπορούσαν να το κάνουν οι άνθρωποι υπάλληλοι και οι προηγούμενες τεχνολογίες.

Τεχνητή νοημοσύνη στις μεταφορές: Εκτός από τον θεμελιώδη ρόλο της AI στη λειτουργία αυτόνομων οχημάτων, οι τεχνολογίες AI χρησιμοποιούνται στις αυτοκινητικές μεταφορές για τη διαχείριση της κυκλοφορίας, τη μείωση της συμφόρησης και τη βελτίωση της οδικής ασφάλειας. Στις αερομεταφορές, η AI μπορεί να προβλέπει καθυστερήσεις πτήσεων αναλύοντας δεδομένα όπως οι καιρικές συνθήκες και οι συνθήκες εναέριας κυκλοφορίας. Στις υπερπόντιες μεταφορές, η AI μπορεί να βελτιώσει την ασφάλεια και την αποδοτικότητα βελτιστοποιώντας τις διαδρομές και παρακολουθώντας αυτόματα τις συνθήκες των πλοίων. Τέλος, στις εφοδιαστικές αλυσίδες, η AI αντικαθιστά τις παραδοσιακές μεθόδους πρόβλεψης ζήτησης και βελτιώνει την ακρίβεια των προβλέψεων για πιθανά προβλήματα και σημεία συμφόρησης. Η πανδημία COVID-19 ανέδειξε τη σημασία αυτών των δυνατοτήτων, καθώς πολλές εταιρείες αιφνιδιάστηκαν από τις επιπτώσεις μιας παγκόσμιας πανδημίας στην προσφορά και τη ζήτηση αγαθών.

2.1.9. Ηθική χρήση της τεχνητής νοημοσύνης

Ενώ τα εργαλεία AI προσφέρουν μια σειρά από νέες λειτουργίες για τις επιχειρήσεις, η χρήση τους εγείρει σημαντικά ηθικά ερωτήματα [17] [18]. Είτε για καλό είτε για κακό, τα συστήματα AI ενισχύουν ό,τι έχουν ήδη μάθει, γεγονός που σημαίνει ότι αυτοί οι

αλγόριθμοι εξαρτώνται σε μεγάλο βαθμό από τα δεδομένα με τα οποία έχουν εκπαιδευτεί. Επειδή ένας άνθρωπος επιλέγει τα δεδομένα εκπαίδευσης, η πιθανότητα μεροληψίας είναι έμφυτη και πρέπει να παρακολουθείται στενά.

Η γενετική τεχνητή νοημοσύνη προσθέτει ένα επιπλέον επίπεδο ηθικής πολυπλοκότητας. Αυτά τα εργαλεία μπορούν να παράγουν εξαιρετικά ρεαλιστικά και πειστικά κείμενα, εικόνες και ήχους - μια χρήσιμη ικανότητα για πολλές νόμιμες εφαρμογές, αλλά και ένα δυνητικό μέσο παραπληροφόρησης και επιβλαβούς περιεχομένου όπως τα deepfakes.

Κατά συνέπεια, οποιοσδήποτε επιθυμεί να χρησιμοποιήσει τη μηχανική μάθηση σε συστήματα παραγωγής πραγματικού κόσμου πρέπει να ενσωματώσει την ηθική στις διαδικασίες εκπαίδευσης AI και να προσπαθήσει να αποφύγει την ανεπιθύμητη μεροληψία. Αυτό είναι ιδιαίτερα σημαντικό για αλγόριθμους AI που στερούνται διαφάνειας, όπως τα σύνθετα νευρωνικά δίκτυα που χρησιμοποιούνται στη βαθιά μάθηση.

Η εξηγησιμότητα, δηλαδή η ικανότητα κατανόησης του πώς ένα σύστημα τεχνητής νοημοσύνης λαμβάνει αποφάσεις, είναι ένας αναπτυσσόμενος τομέας ενδιαφέροντος στην έρευνα AI. Η έλλειψη εξηγησιμότητας παρουσιάζει ένα δυνητικό εμπόδιο στη χρήση της AI σε βιομηχανίες με αυστηρές κανονιστικές απαιτήσεις συμμόρφωσης. Για παράδειγμα, οι νόμοι περί δίκαιου δανεισμού απαιτούν από τα χρηματοπιστωτικά ιδρύματα των ΗΠΑ να εξηγούν τις αποφάσεις τους σχετικά με την έκδοση πιστώσεων σε αιτούντες δάνεια και πιστωτικές κάρτες. Ωστόσο, όταν τα προγράμματα AI λαμβάνουν τέτοιες αποφάσεις, οι λεπτές συσχετίσεις μεταξύ χιλιάδων μεταβλητών μπορεί να δημιουργήσουν το πρόβλημα του "μαύρου κουτιού", όπου η διαδικασία λήψης αποφάσεων του συστήματος είναι αδιαφανής.

Συνοψίζοντας, οι ηθικές προκλήσεις της AI περιλαμβάνουν τα εξής:

- Μεροληψία λόγω ακατάλληλα εκπαιδευμένων αλγορίθμων και ανθρώπινων προκαταλήψεων ή παραλείψεων.

- Κακή χρήση της γεννητικής AI για την παραγωγή deepfakes, phishing scams
- και άλλου επιβλαβούς περιεχομένου.
- Νομικές ανησυχίες, όπως η δυσφήμιση και τα ζητήματα πνευματικών δικαιωμάτων που σχετίζονται με την AI.
- Απώλεια θέσεων εργασίας λόγω της αυξανόμενης χρήσης της AI για την αυτοματοποίηση εργασιακών καθηκόντων.
- Ανησυχίες για την ιδιωτικότητα των δεδομένων, ιδιαίτερα σε τομείς όπως οι τράπεζες, η υγειονομική περίθαλψη και η νομική, που ασχολούνται με ευαίσθητα προσωπικά δεδομένα.

2.2. Μηχανική Μάθηση

2.2.1. Ορισμός μηχανικής μάθησης

Η μηχανική μάθηση [19] [20] είναι ένας υποτομέας της τεχνητής νοημοσύνης που εστιάζει στην ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και προηγούμενες εμπειρίες. Ο όρος "μηχανική μάθηση" εισήχθη για πρώτη φορά από τον Arthur Samuel το 1959 και ο ίδιος τον όρισε ως έναν "τομέα μελέτης που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να είναι ρητά προγραμματισμένοι". Οι αλγόριθμοι μηχανικής μάθησης δημιουργούν ένα μαθηματικό μοντέλο με τη βοήθεια δείγματος ιστορικών δεδομένων, γνωστών ως δεδομένα εκπαίδευσης, το οποίο βοηθά στην πρόβλεψη ή τη λήψη αποφάσεων χωρίς να είναι ρητά προγραμματισμένοι. Η απόδοση των μοντέλων μηχανικής μάθησης βελτιώνεται καθώς παρέχονται περισσότερες πληροφορίες, επιτρέποντάς τους να κάνουν πιο ακριβείς προβλέψεις.

Η μηχανική μάθηση βρίσκει εφαρμογή σε διάφορους τομείς. Για παράδειγμα, χρησιμοποιείται σε αυτόνομα οχήματα για την ανάλυση δεδομένων σε πραγματικό χρόνο

και τη λήψη αποφάσεων σχετικά με τις ενέργειες οδήγησης. Επίσης, χρησιμοποιείται σε συστήματα ανίχνευσης κυβερνοαπάτης για την αναγνώριση και την αποτροπή δόλιων δραστηριοτήτων αναλύοντας πρότυπα και ανωμαλίες στα δεδομένα. Η τεχνολογία αναγνώρισης προσώπου βασίζεται επίσης σε αλγόριθμους μηχανικής μάθησης για την αναγνώριση και την αυθεντικοποίηση ατόμων.

Επιπλέον, εταιρείες όπως η Netflix και η Amazon αξιοποιούν μοντέλα μηχανικής μάθησης για να αναλύσουν τις προτιμήσεις των χρηστών και να παρέχουν εξατομικευμένες προτάσεις για ταινίες και τηλεοπτικές εκπομπές. Η ικανότητα της μηχανικής μάθησης να επεξεργάζεται και να κατανοεί μεγάλες ποσότητες δεδομένων την καθιστά ένα ισχυρό εργαλείο για την εξαγωγή πολύτιμων πληροφοριών και τη λήψη ενημερωμένων αποφάσεων σε ένα ευρύ φάσμα βιομηχανιών και εφαρμογών.

2.2.2. Αλγόριθμος μηχανικής μάθησης

Ένας αλγόριθμος μηχανικής μάθησης [21] [22] είναι ένα σύνολο μαθηματικών κανόνων και διαδικασιών που επιτρέπουν σε ένα σύστημα τεχνητής νοημοσύνης να εκτελεί συγκεκριμένες εργασίες, όπως η πρόβλεψη αποτελεσμάτων ή η λήψη αποφάσεων, μαθαίνοντας από δεδομένα. Είναι ο μηχανισμός μέσω του οποίου το σύστημα αναλύει και επεξεργάζεται δεδομένα εισόδου, εντοπίζει πρότυπα ή σχέσεις στα δεδομένα και παράγει το επιθυμητό αποτέλεσμα ή την πρόβλεψη.

Ο αλγόριθμος είναι σχεδιασμένος να εξάγει πολύτιμες πληροφορίες από τα δεδομένα εισόδου και να τις χρησιμοποιεί για να λαμβάνει ενημερωμένες αποφάσεις ή προβλέψεις. Περνάει από μια φάση εκπαίδευσης όπου μαθαίνει από ένα σύνολο δεδομένων με ετικέτες (labeled/unlabeled dataset), όπου κάθε είσοδος συνδέεται με ένα γνωστό ή άγνωστο αποτέλεσμα. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος προσαρμόζει τις εσωτερικές του παραμέτρους και λειτουργίες για να ελαχιστοποιήσει τα λάθη και να βελτιώσει την απόδοσή του.

Μόλις εκπαιδευτεί, ο αλγόριθμος μπορεί να εφαρμοστεί σε νέα, μ - γνωστά δεδομένα για να κάνει προβλέψεις ή να λάβει αποφάσεις βασισμένος στα πρότυπα που έχει μάθει. Η ικανότητα του αλγορίθμου να γενικεύει και να προβλέπει με ακρίβεια τα αποτελέσματα για νέα δεδομένα είναι ένα βασικό μέτρο της αποτελεσματικότητάς του.

Διαφορετικές τεχνικές/αλγόριθμοι μηχανικής μάθησης έχουν σχεδιαστεί για διαφορετικές εργασίες και τύπους δεδομένων. Μερικοί αλγόριθμοι είναι πιο κατάλληλοι για προβλήματα ταξινόμησης, όπου ο στόχος είναι να εκχωρηθούν είσοδοι σε προκαθορισμένες κατηγορίες ή κλάσεις. Άλλοι είναι πιο αποτελεσματικοί για προβλήματα παλινδρόμησης, όπου στόχος είναι η πρόβλεψη μιας συνεχούς αριθμητικής τιμής. Επιπλέον, υπάρχουν αλγόριθμοι που έχουν σχεδιαστεί ειδικά για ομαδοποίηση, ανίχνευση ανωμαλιών, συστήματα συστάσεων και άλλες εργασίες.

Οι τεχνικές μηχανικής μάθησης αποτελούν τον βασικό πυρήνα ενός συστήματος τεχνητής νοημοσύνης που του επιτρέπει να μαθαίνει από δεδομένα, να εντοπίζει πρότυπα και να λαμβάνει προβλέψεις ή αποφάσεις βασισμένες σε αυτή τη μάθηση. Είναι το θεμελιώδες εργαλείο που επιτρέπει στις μηχανές να εκτελούν ευφυείς εργασίες και να προσαρμόζονται σε νέες καταστάσεις με βάση τη γνώση που αποκτήθηκε από τη διαδικασία εκπαίδευσης.

2.2.3. Τύποι μηχανικής μάθησης

Πριν προχωρήσουμε στις τεχνικές μηχανικής μάθησης, ας συζητήσουμε τους τύπους ή τις προσεγγίσεις της μηχανικής μάθησης. Η μηχανική μάθηση μπορεί να κατηγοριοποιηθεί ευρέως σε 4 κατηγορίες [23] [24]:

Εποπτευόμενη μάθηση (Supervised Learning): Η εποπτευόμενη μάθηση περιλαμβάνει την εκπαίδευση ενός αλγορίθμου χρησιμοποιώντας δεδομένα με ετικέτες, όπου κάθε σημείο δεδομένων είναι ετικετοποιημένο με το αντίστοιχο αποτέλεσμα. Σε αυτή την προσέγγιση, οι μηχανές εκπαιδεύονται χρησιμοποιώντας δεδομένα με ετικέτες για να κάνουν ακριβείς προβλέψεις. Για να το κατανοήσουμε καλύτερα, ας εξετάσουμε την εκπαίδευση ενός

μοντέλου μηχανικής μάθησης για την ταξινόμηση εικόνων γάτων και σκύλων. Το μοντέλο μαθαίνει από ένα σύνολο δεδομένων εικόνων ζώων με ετικέτες, χρησιμοποιώντας οπτικά χαρακτηριστικά όπως το σχήμα, το μέγεθος και το χρώμα για να αναγνωρίσει τα διακριτικά χαρακτηριστικά των γάτων και των σκύλων. Μέσω αυτής της διαδικασίας, η μηχανή μπορεί να αναγνωρίσει και να ταξινομήσει με ακρίβεια νέες εικόνες γάτων και σκύλων.

Η εποπτευόμενη μάθηση χωρίζεται σε δύο κατηγορίες:

- **Ταξινόμηση (Classification):** Η ταξινόμηση είναι μια τεχνική που χρησιμοποιείται όταν η μεταβλητή εξόδου αποτελείται από κατηγορίες. Περιλαμβάνει την ανάθεση σημείων δεδομένων σε προκαθορισμένες κατηγορίες ή κλάσεις. Για παράδειγμα, εξετάστε το πρόβλημα της εποπτευόμενης μάθησης για την αναγνώριση των μηνυμάτων ηλεκτρονικού ταχυδρομείου ως ανεπιθύμητα (spam) ή όχι. Τα δεδομένα εκπαίδευσης αποτελούνται από μια συλλογή μηνυμάτων ηλεκτρονικού ταχυδρομείου, καθένα από τα οποία έχει χαρακτηριστεί ως ανεπιθύμητο ή όχι. Ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται στα δεδομένα με ετικέτες για να αναγνωρίζει τα ανεπιθύμητα και τα μη ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου, εξετάζοντας τα χαρακτηριστικά και τα πρότυπα στα μηνύματα.
- **Παλινδρόμηση (Regression):** Η παλινδρόμηση χρησιμοποιείται όταν η μεταβλητή εξόδου είναι μια πραγματική ή συνεχής τιμή. Στοχεύει να καθορίσει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών, όπου οι αλλαγές σε μία μεταβλητή αντιστοιχούν σε αναλογικές αλλαγές στην άλλη. Για παράδειγμα, πρόβλεψη της τιμής ενός σπιτιού με βάση παράγοντες όπως το μέγεθός του, η τοποθεσία, ο αριθμός των δωματίων κ.λπ. Σε αυτή την περίπτωση, η τιμή του σπιτιού εξαρτάται από διάφορα χαρακτηριστικά ή προβλεπτικούς παράγοντες, και η ανάλυση παλινδρόμησης βοηθά στην εκτίμηση της σχέσης και στη λήψη ακριβών προβλέψεων.

Μη εποπτευόμενη μάθηση (Unsupervised Learning): Η μη εποπτευόμενη μάθηση είναι μια προσέγγιση μηχανικής μάθησης που περιλαμβάνει τη χρήση δεδομένων χωρίς ετικέτες

για την ανακάλυψη προτύπων και τάσεων χωρίς σαφείς οδηγίες ή προκαθορισμένα αποτελέσματα. Στο πλαίσιο των εικόνων ζώων, η μηχανή εξετάζει αυτόνομα τα δεδομένα εισόδου, τα κατηγοριοποιεί βάσει οπτικών χαρακτηριστικών, όπως το σχήμα, το μέγεθος και το χρώμα, και εντοπίζει κοινά σημεία και διαφορές μεταξύ των εικόνων. Μέσω της επαναληπτικής ανάλυσης, δημιουργεί σημαντικές γνώσεις και δομές στα δεδομένα, αποκαλύπτοντας κρυμμένα πρότυπα και πιθανές συσχετίσεις. Η μη εποπτευόμενη μάθηση είναι πολύτιμη για την εξερεύνηση σύνθετων συνόλων δεδομένων, επιτρέποντας ανακαλύψεις και λήψη αποφάσεων χωρίς την ανάγκη για δεδομένα με ετικέτες.

Ημι-εποπτευόμενη μάθηση (Semi-supervised Learning): Η ημι-εποπτευόμενη μάθηση είναι μια προσέγγιση εκπαίδευσης που γεφυρώνει το χάσμα μεταξύ εποπτευόμενης και μη εποπτευόμενης μάθησης. Συνδυάζει πολλά δεδομένα χωρίς ετικέτες με ένα μικρό σύνολο δεδομένων με ετικέτες για να δημιουργήσει μια προβλεπτική λειτουργία. Αυτή η μέθοδος χρησιμοποιεί τεχνικές μη εποπτευόμενης μάθησης για να ομαδοποιήσει παρόμοια δεδομένα, επιτρέποντας την κατηγοριοποίηση των δεδομένων χωρίς ετικέτες σε ετικετοποιημένα. Η ημι-εποπτευόμενη μάθηση είναι ιδιαίτερα ευεργετική όταν υπάρχει διαθέσιμη μεγάλη ποσότητα δεδομένων χωρίς ετικέτες, αλλά η επισήμανση όλων αυτών θα ήταν δαπανηρή ή δύσκολη. Για παράδειγμα, σε ένα σενάριο με 10.000 φωτογραφίες ζώων, όπου μόνο οι 1.000 είναι ετικετοποιημένες, ένα μοντέλο Συγκριτικού Νευρωνικού Δικτύου (CNN) μπορεί να εκπαιδευτεί χρησιμοποιώντας τις ετικετοποιημένες φωτογραφίες. Το εκπαιδευμένο μοντέλο μπορεί στη συνέχεια να χαρακτηρίσει τις υπόλοιπες 9.000 φωτογραφίες χωρίς ετικέτες, επεκτείνοντας το σύνολο δεδομένων με ετικέτες. Η επανεκπαίδευση του μοντέλου στο διευρυμένο σύνολο δεδομένων με ετικέτες επιτρέπει την αύξηση της ακρίβειας στην κατηγοριοποίηση σκύλων και γατών.

Ενισχυτική μάθηση (Reinforcement Learning): Η ενισχυτική μάθηση (Reinforcement Learning - RL) είναι μια μορφή μηχανικής μάθησης στην οποία ένας αυτόνομος πράκτορας μαθαίνει να λαμβάνει αποφάσεις και να αναλαμβάνει ενέργειες μέσα σε ένα περιβάλλον για να μεγιστοποιήσει ένα σήμα ανταμοιβής. Αλληλεπιδρώντας με το περιβάλλον, ο

πράκτορας λαμβάνει ανατροφοδότηση μέσω ανταμοιβών ή ποινών με βάση το αποτέλεσμα των ενεργειών του. Ο στόχος είναι να προσδιοριστεί η βέλτιστη πολιτική, ένα σύνολο κανόνων που καθοδηγεί τις επιλογές ενεργειών του πράκτορα για τη μεγιστοποίηση των μακροπρόθεσμων ανταμοιβών. Μέσω της διαδικασίας δοκιμής και σφάλματος, ο πράκτορας μαθαίνει από τις εμπειρίες του, παρατηρεί τις προκύπτουσες καταστάσεις και τις ανταμοιβές και ενημερώνει ανάλογα τη στρατηγική λήψης αποφάσεων. Η ενισχυτική μάθηση περιλαμβάνει τον ορισμό καταστάσεων, ενεργειών και μιας λειτουργίας ανταμοιβής για την εκπαίδευση του πράκτορα να λαμβάνει τις καλύτερες αποφάσεις με την πάροδο του χρόνου.

2.2.4. Σημαντικά μοντέλα μηχανικής μάθησης

2.2.4.1. Δέντρα Απόφασης

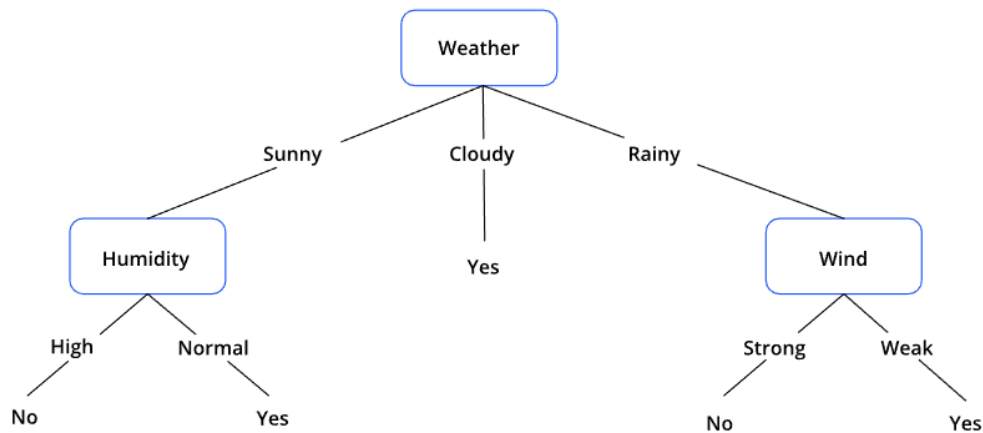
Ένα δέντρο απόφασης [25] είναι ένα ιεραρχικό μοντέλο που βοηθά στη λήψη αποφάσεων λαμβάνοντας υπόψη διάφορα αποτελέσματα και παράγοντες. Η δομή του δέντρου αποτελείται από μια ρίζα, κλάδους και φύλλα. Είναι σαν ένα διάγραμμα ροής που ακολουθεί διασπάσεις βάσει χαρακτηριστικών για να κάνει προβλέψεις. Ξεκινά με έναν κόμβο ρίζας και καταλήγει σε αποφάσεις που λαμβάνονται από τους κόμβους φύλλων. Τα δέντρα απόφασης είναι χρήσιμα σε διάφορους τομείς και μπορούν να χειριστούν προβλήματα ταξινόμησης και παλινδρόμησης.

Τα δέντρα απόφασης μπορούν να θεωρηθούν ως μια σειρά από δηλώσεις if-else. Το δέντρο ξεκινά από τον κόμβο ρίζας και ελέγχει εάν μια συγκεκριμένη συνθήκη είναι αληθής. Ανάλογα με το αποτέλεσμα, μετακινείται στον επόμενο κόμβο που συνδέεται με αυτή την

απόφαση. Αυτή η διαδικασία συνεχίζεται μέχρι να φτάσει σε έναν κόμβο φύλλου, ο οποίος αντιπροσωπεύει την τελική απόφαση ή το αποτέλεσμα. Για παράδειγμα, το δέντρο απόφασης ξεκινά ζητώντας πληροφορίες για τον καιρό. Τα επόμενα χαρακτηριστικά, όπως η υγρασία και ο άνεμος, καθορίζονται για να διαπιστωθεί εάν ο καιρός είναι ηλιόλουστος, νεφελώδης ή βροχερός. Στη συνέχεια, καθορίζεται αν ο άνεμος είναι δυνατός ή ασθενής· αν βρέχει και ο άνεμος είναι ασθενής, το άτομο μπορεί να παίξει σε εξωτερικούς χώρους. Το δέντρο συνεχίζει να διασπάται ανάλογα με τις συνθήκες μέχρι να ληφθεί μια απόφαση, όπως το να παίξει κάποιος σε εξωτερικό χώρο. Όταν η πρόσθετη διάσπαση δεν παρέχει νέες πληροφορίες ή δεν βοηθά στη διαδικασία λήψης αποφάσεων, το δέντρο απόφασης σταματά να διασπάται. Ορισμένες κοινές μετρικές που χρησιμοποιούνται στα δέντρα απόφασης είναι:

- **Εντροπία:** Είναι μια μετρική που χρησιμοποιείται στα δέντρα απόφασης για τη μέτρηση της αβεβαιότητας ή της αταξίας σε ένα σύνολο δεδομένων. Ποσοτικοποιεί την ακαθαρσία ενός κόμβου ή υποσυνόλου δεδομένων.
- **Κέρδος Πληροφορίας:** Είναι μια άλλη μετρική που χρησιμοποιείται στα δέντρα απόφασης. Μετρά τη μείωση της αβεβαιότητας ή της εντροπίας όταν ένα σύνολο δεδομένων διασπάται βάσει ενός συγκεκριμένου χαρακτηριστικού. Βοηθά στον προσδιορισμό του ποιο χαρακτηριστικό πρέπει να επιλεγεί ως κόμβος ρίζας ή απόφασης. Το χαρακτηριστικό με το μεγαλύτερο κέρδος πληροφορίας επιλέγεται για τη διάσπαση των δεδομένων.
- **Υπερπαράμετροι:** Οι υπερπαράμετροι μπορούν να χρησιμοποιηθούν για τον καθορισμό του πότε θα σταματήσει η διάσπαση. Αυτές οι παράμετροι ελέγχουν την πολυπλοκότητα και το μέγεθος του δέντρου απόφασης. Παραδείγματα περιλαμβάνουν το μέγιστο βάθος του δέντρου, τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για μια διάσπαση και τον ελάχιστο αριθμό δειγμάτων που απαιτούνται σε έναν κόμβο φύλλου.
- **Κλάδεμα:** Είναι μια τεχνική που χρησιμοποιείται για την αποφυγή υπερπροσαρμογής στα δέντρα απόφασης. Περιλαμβάνει την αφαίρεση ή το κόψιμο

κόμβων ή κλάδων που δεν είναι σημαντικοί. Το κλάδεμα μπορεί να γίνει κατά τη διάρκεια της κατασκευής του δέντρου (προ-κλάδεμα) ή μετά την κατασκευή του δέντρου (μετα-κλάδεμα), ανάλογα με τη σημασία των κόμβων. Βοηθά στη βελτίωση της απόδοσης και της ικανότητας γενίκευσης του δέντρου απόφασης.



Εικόνα 1: Δέντρο Απόφασης

Τα δέντρα απόφασης μπορούν να κατηγοριοποιηθούν σε διάφορους τύπους:

- **Δέντρα Ταξινόμησης (Classification Trees):** Αυτά τα δέντρα ταξινομούν τα δεδομένα εισόδου σε πολλές κατηγορίες βάσει των χαρακτηριστικών ή των ιδιοτήτων τους. Χρησιμοποιούνται για προβλήματα όπου το αποτέλεσμα ανήκει σε προκαθορισμένες κλάσεις, όπως η αναγνώριση αν ένα email είναι ανεπιθύμητο (spam) ή όχι.
- **Δέντρα Παλινδρόμησης (Regression Trees):** Αυτά τα δέντρα προβλέπουν συνεχείς αριθμητικές τιμές αναλύοντας τα δεδομένα εισόδου. Χρησιμοποιούνται για προβλήματα όπου το αποτέλεσμα είναι μια συνεχής τιμή, όπως η πρόβλεψη της τιμής ενός ακινήτου με βάση τα χαρακτηριστικά του.

- **Δυαδικά Δέντρα Απόφασης (Binary Decision Trees):** Αυτά τα δέντρα έχουν κόμβους με δύο δυνατά αποτελέσματα. Κάθε κόμβος χωρίζει τα δεδομένα σε δύο κλάδους, όπως ναι ή όχι, αληθές ή ψευδές.
- **Πολυδιάστατα Δέντρα Απόφασης (Multiway Decision Trees):** Αυτά τα δέντρα έχουν κόμβους με πολλαπλά εναλλακτικά αποτελέσματα. Κάθε κόμβος μπορεί να χωρίσει τα δεδομένα σε περισσότερους από δύο κλάδους, προσφέροντας πολλαπλές επιλογές ή κατηγορίες για το επόμενο βήμα.

2.2.4.2. Naïve Bayes

Ο αλγόριθμος Naive Bayes [26] είναι ένας απλός και αποτελεσματικός τρόπος για να ταξινομήσουμε αντικείμενα σε διαφορετικές κατηγορίες. Βασίζεται σε μια μαθηματική έννοια που ονομάζεται Θεώρημα του Bayes και υποθέτει ότι τα χαρακτηριστικά που χρησιμοποιούμε για να ταξινομήσουμε κάτι είναι ανεξάρτητα μεταξύ τους.

Ας υποθέσουμε ότι θέλουμε να ταξινομήσουμε emails ως spam ή μη spam. Ο Naive Bayes θα εξετάσει διαφορετικά χαρακτηριστικά ενός email, όπως τις λέξεις που χρησιμοποιούνται, τη διεύθυνση του αποστολέα και την παρουσία συγκεκριμένων φράσεων. Με βάση αυτά τα χαρακτηριστικά, υπολογίζει την πιθανότητα ένα email να είναι spam.

Το «naive» μέρος προέρχεται από την υπόθεση ότι αυτά τα χαρακτηριστικά είναι άσχετα μεταξύ τους, παρόλο που αυτό μπορεί να μην ισχύει πάντα. Για παράδειγμα, η λέξη «δωρεάν» σε ένα email μπορεί να αυξήσει τις πιθανότητες να είναι spam, αλλά ο NaiveBayes αντιμετωπίζει αυτό το χαρακτηριστικό ανεξάρτητα από άλλα χαρακτηριστικά όπως η διεύθυνση του αποστολέα.

Ας κατανοήσουμε τη λειτουργία του με ένα παράδειγμα: θέλουμε να προβλέψουμε εάν οι παίκτες θα παίξουν ποδόσφαιρο με βάση τις καιρικές συνθήκες. Ξεκινάμε με το να μετράμε πόσο συχνά εμφανίζεται κάθε καιρική συνθήκη και αποτέλεσμα παιχνιδιού στα δεδομένα εκπαίδευσης. Στη συνέχεια, υπολογίζουμε τις πιθανότητες χρησιμοποιώντας αυτές τις μετρήσεις. Για παράδειγμα, η πιθανότητα να παίξουν είναι 0,64 και η πιθανότητα για

«Νεφελώδη» καιρό είναι 0,29. Στη συνέχεια, χρησιμοποιούμε το Θεώρημα του Bayes για να κάνουμε προβλέψεις. Πολλαπλασιάζουμε την πιθανότητα να παίξουν (0,64) με την πιθανότητα της συγκεκριμένης καιρικής συνθήκης (π.χ., 0,29 για «Νεφελώδη») και τη διαιρούμε με τη συνολική πιθανότητα αυτής της καιρικής συνθήκης. Κάνουμε αυτό για κάθε καιρική συνθήκη και συγκρίνουμε τις πιθανότητες για να προβλέψουμε αν οι παίκτες θα παίξουν ποδόσφαιρο σε μια δεδομένη καιρική συνθήκη.

Ο Naive Bayes λειτουργεί καλά στην πράξη και χρησιμοποιείται ευρέως σε εφαρμογές όπως η ταξινόμηση κειμένων, η φιλτράρισμα ανεπιθύμητων μηνυμάτων και η ανάλυση συναισθημάτων. Είναι εύκολος στην υλοποίηση και αποδίδει αποτελεσματικά, ειδικά με μεγάλες ποσότητες δεδομένων. Υπάρχουν διαφορετικοί τύποι ταξινομητών NaiveBayes, όπως:

- **Ταξινομητής Multinomial Naive Bayes:** Αυτός ο ταξινομητής χρησιμοποιείται συνήθως για την ταξινόμηση εγγράφων. Είναι ένας πιθανολογικός αλγόριθμος που χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας (NLP) και προβλέπει την ετικέτα ενός κειμένου με βάση την υψηλότερη πιθανότητα που υπολογίζεται χρησιμοποιώντας το Θεώρημα του Bayes και την υπόθεση της συνθηματικής ανεξαρτησίας. Απλοποιεί την ταξινόμηση κειμένων αξιοποιώντας τη στατιστική ανάλυση των εμφανίσεων των λέξεων.
- **Ταξινομητής Bernoulli Naive Bayes:** Σε αυτόν τον ταξινομητή, τα χαρακτηριστικά αντιπροσωπεύονται ως ανεξάρτητες δυαδικές μεταβλητές (αληθές/ψευδές ή 0/1). Χρησιμοποιείται συχνά στην ταξινόμηση εγγράφων, όπου λαμβάνεται υπόψη η παρουσία ή η απουσία λέξεων σε ένα έγγραφο αντί για τις συχνότητές τους. Είναι χρήσιμος όταν θέλουμε να επικεντρωθούμε στη δυαδική εμφάνιση όρων.
- **Ταξινομητής Gaussian Naive Bayes:** Αυτός ο ταξινομητής υποθέτει συνεχή τιμή χαρακτηριστικών και ακολουθεί μια κανονική κατανομή (μια καμπύλη σε σχήμα καμπάνας). Είναι κατάλληλος για δεδομένα με συνεχή μεταβλητές. Υπολογίζει τον μέσο όρο και την τυπική απόκλιση κάθε χαρακτηριστικού και χρησιμοποιεί αυτές

τις πληροφορίες για να κάνει προβλέψεις βάσει της συνάρτησης πυκνότητας πιθανότητας της κανονικής κατανομής.

2.2.4.3. SVM

Το Support Vector Machine (SVM) [27] είναι ένας επιβλεπόμενος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης. Ο στόχος του είναι να βρει την καλύτερη γραμμή ή όριο που διαχωρίζει δύο κλάσεις δεδομένων. Σε αντίθεση με την λογιστική παλινδρόμηση, η οποία χρησιμοποιεί πιθανότητες, το SVM χρησιμοποιεί στατιστικές προσεγγίσεις.

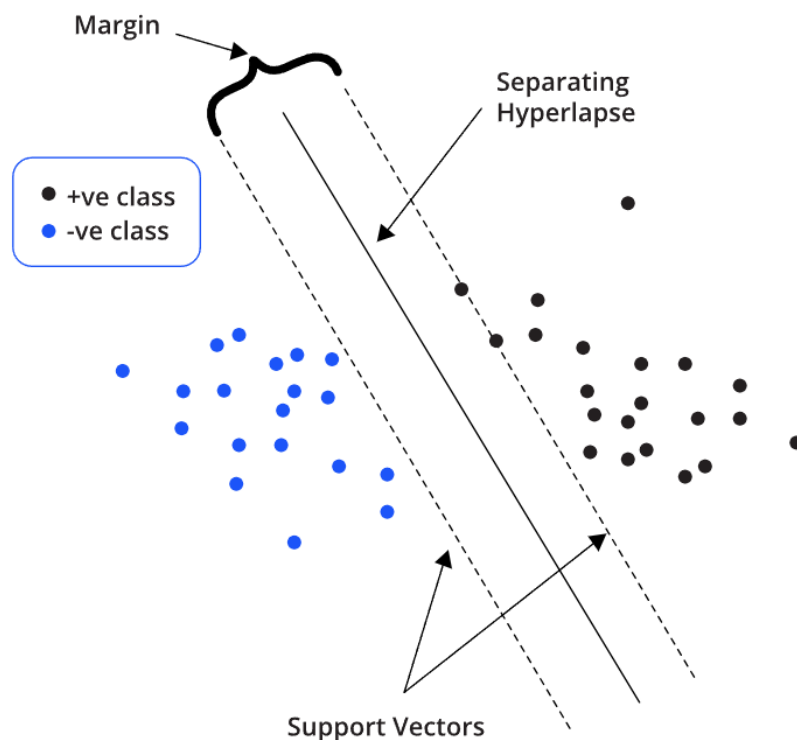
Τα SVMs απαιτούν τα δεδομένα εισόδου να εκπροσωπούνται ως διανύσματα χαρακτηριστικών (feature vectors). Αυτά τα διανύσματα χαρακτηριστικών μπορούν να προκύψουν από τα αρχικά δεδομένα εισόδου μέσω της εξαγωγής σχετικών χαρακτηριστικών ή με τη χρήση τεχνικών όπως η μείωση διαστάσεων.

Σε ένα πρόβλημα δυαδικής ταξινόμησης, το SVM στοχεύει να βρει ένα υπερεπίπεδο (hyper plane) που διαχωρίζει με τον καλύτερο τρόπο τις δύο κλάσεις. Το υπερεπίπεδο επιλέγεται ώστε να μεγιστοποιείται το περιθώριο (margin), δηλαδή η απόσταση μεταξύ του υπερεπιπέδου και των πλησιέστερων σημείων δεδομένων από κάθε κλάση. Αυτό το περιθώριο επιτρέπει στο SVM να γενικεύει καλύτερα και να μειώνει τον κίνδυνο υπερπροσαρμογής.

Αν τα δεδομένα μπορούν να διαχωριστούν τέλεια από ένα υπερεπίπεδο, τότε το πρόβλημα ονομάζεται γραμμικά διαχωρίσιμο (linear lyseparable). Σε αυτή την περίπτωση, το SVM καθορίζει το υπερεπίπεδο που επιτυγχάνει το μέγιστο περιθώριο μεταξύ των κλάσεων. Τα σημεία δεδομένων που βρίσκονται πιο κοντά στο υπερεπίπεδο ονομάζονται υποστηρικτικά διανύσματα (support vectors).

Σε πολλές περιπτώσεις πραγματικού κόσμου, τα δεδομένα μπορεί να μην είναι γραμμικά διαχωρίσιμα. Για να αντιμετωπιστούν τέτοιες περιπτώσεις, το SVM χρησιμοποιεί μια τεχνική που ονομάζεται kernel trick. Το kernel trick μετασχηματίζει τον αρχικό χώρο χαρακτηριστικών σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης όπου τα δεδομένα γίνονται γραμμικά διαχωρίσιμα. Αυτός ο μετασχηματισμός γίνεται έμμεσα, χωρίς να υπολογίζονται οι συντεταγμένες του νέου χώρου χαρακτηριστικών.

Οι συναρτήσεις πυρήνα (kernel functions) χρησιμοποιούνται για τον υπολογισμό της ομοιότητας μεταξύ ζευγών σημείων δεδομένων στον χώρο χαρακτηριστικών υψηλότερης διάστασης. Η επιλογή της συνάρτησης πυρήνα εξαρτάται από το πρόβλημα που εξετάζεται και τα χαρακτηριστικά των δεδομένων. Δημοφιλείς συναρτήσεις πυρήνα περιλαμβάνουν τη γραμμική, την πολυωνυμική, τη Radial Basis Function (RBF) και τη sigmoid.



Εικόνα 2: Μοντέλο SVM

Η εκπαίδευση του SVM περιλαμβάνει την εύρεση του βέλτιστου υπερεπιπέδου μέσω της επίλυσης ενός προβλήματος βελτιστοποίησης. Ο στόχος είναι να μεγιστοποιηθεί το περιθώριο ενώ παράλληλα ελαχιστοποιούνται τα λάθη ταξινόμησης. Αυτό επιτυγχάνεται μέσω της διατύπωσης του προβλήματος ως ένα κυρτό πρόβλημα βελτιστοποίησης, το οποίο μπορεί να επιλυθεί με τη χρήση αποδοτικών αλγορίθμων όπως ο προγραμματισμός τετραγωνικής βελτιστοποίησης (quadratic programming).

Αφού εκπαιδευτεί το μοντέλο SVM, μπορεί να χρησιμοποιηθεί για ταξινόμηση αξιολογώντας τη θέση νέων σημείων δεδομένων σε σχέση με το υπερεπίπεδο που έχει μάθει. Η θέση του σημείου δεδομένων καθορίζει την προβλεπόμενη ετικέτα κλάσης του.

Τα SVMs έχουν αρκετά πλεονεκτήματα, όπως η ικανότητα διαχείρισης δεδομένων υψηλής διάστασης, η αποτελεσματικότητα στην αντιμετώπιση μη γραμμικά διαχωρίσιμων δεδομένων μέσω συναρτήσεων πυρήνα και οι καλές δυνατότητες γενίκευσης. Ωστόσο, τα SVMs μπορεί να είναι ευαίσθητα στην επιλογή των υπερπαραμέτρων και μπορεί να απαιτούν προσεκτική ρύθμιση.

2.2.4.4. KNN

Ο αλγόριθμος K-nearest Neighbor (KNN) [28] είναι μια δημοφιλής τεχνική μηχανικής μάθησης για ταξινομητικά προβλήματα. Συγκρίνει ένα νέο σημείο δεδομένων με υπάρχοντα σημεία δεδομένων για να το ταξινομήσει σε μια συγκεκριμένη κατηγορία. Για παράδειγμα, αν έχουμε μια εικόνα ενός ζώου και θέλουμε να προσδιορίσουμε αν είναι γάτα ή σκύλος, μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο KNN.

Ξεκινάμε επιλέγοντας μια τιμή για το K, που αντιπροσωπεύει τους πλησιέστερους γείτονες που θα λάβουμε υπόψη. Στη συνέχεια, υπολογίζουμε την απόσταση μεταξύ του νέου και των υπάρχοντων σημείων δεδομένων, χρησιμοποιώντας συνήθως ένα μέτρο όπως η Ευκλείδεια απόσταση. Επιλέγουμε τα K σημεία δεδομένων με τις μικρότερες αποστάσεις ως τους πλησιέστερους γείτονες. Μεταξύ αυτών των γειτόνων, μετράμε τον αριθμό των

σημείων δεδομένων που ανήκουν σε κάθε κατηγορία (π.χ., γάτα ή σκύλος). Τέλος, αναθέτουμε το νέο σημείο δεδομένων στην κατηγορία με την υψηλότερη καταμέτρηση μεταξύ των πλησιέστερων γειτόνων.

Ο KNN βασίζεται στην ιδέα ότι τα σημεία δεδομένων της ίδιας κατηγορίας είναι παρόμοια μεταξύ τους. Βρίσκοντας τους πλησιέστερους γείτονες βάσει της απόστασης, μπορούμε να προβλέψουμε την κατηγορία ενός νέου σημείου δεδομένων. Αξίζει να σημειωθεί ότι η επιλογή της σωστής τιμής για το K απαιτεί κάποιους πειραματισμούς και προσεκτική εξέταση του συνόλου δεδομένων. Πολύ μικρή τιμή για το K μπορεί να κάνει τον αλγόριθμο ευαίσθητο σε θόρυβο, ενώ μια μεγάλη τιμή K μπορεί να δυσκολέψει την ακριβή ταξινόμηση. Ο KNN είναι ένας απλός και ευέλικτος αλγόριθμος που χρησιμοποιείται σε διάφορες εφαρμογές όπως η αναγνώριση εικόνας, τα συστήματα συστάσεων και η αναγνώριση προτύπων. Οι μετρικές απόστασης που χρησιμοποιούνται στον KNN είναι:

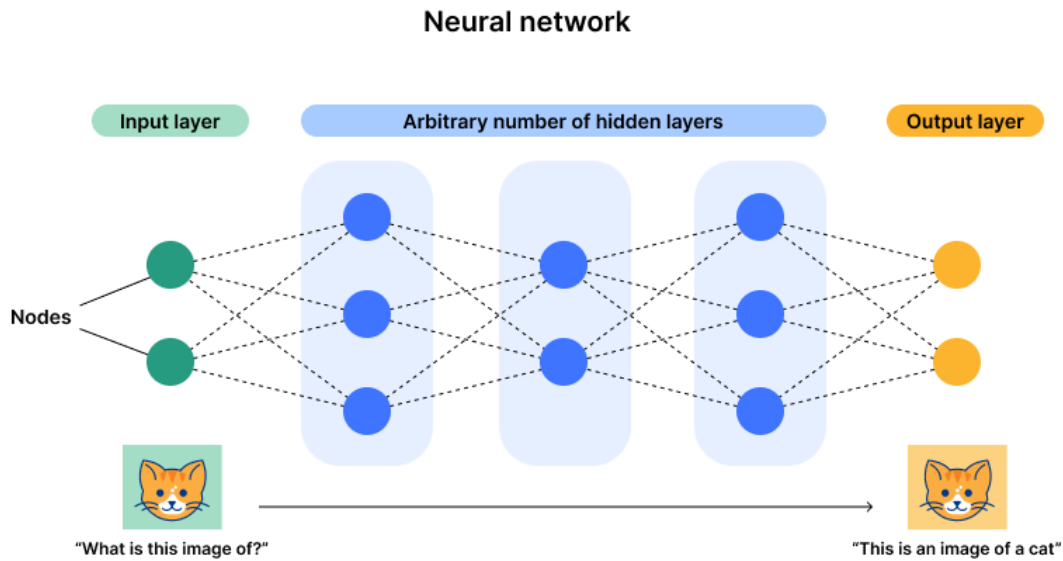
- **Ευκλείδεια απόσταση (Euclidean distance):** Η Ευκλείδεια απόσταση μετρά την ευθεία απόσταση μεταξύ δύο σημείων σε ένα επίπεδο ή υπερεπίπεδο, αντιπροσωπεύοντας τη συνολική μετατόπιση ή το μήκος της συνδετικής γραμμής. Ποσοτικοποιεί τη χωρική σχέση και ομοιότητα μεταξύ αντικειμένων με βάση τις συντεταγμένες τους σε έναν χώρο χαρακτηριστικών.
- **Απόσταση Manhattan (Manhattan distance):** Η απόσταση Manhattan, γνωστή και ως απόσταση πόλης ή "cityblock distance", είναι ένας τρόπος μέτρησης της απόστασης μεταξύ δύο σημείων σε μια δομή τύπου πλέγματος. Είναι σαν να κινείστε μέσα σε μια πόλη με δρόμους που σχηματίζουν πλέγμα. Αντί να κινείστε διαγώνια, μπορείτε να μετακινηθείτε μόνο οριζόντια ή κάθετα.
- **Απόσταση Minkowski (Minkowski distance):** Η απόσταση Minkowski είναι μια γενική μετρική που περιλαμβάνει την Ευκλείδεια απόσταση (για $p=2$) και την απόσταση Manhattan (για $p=1$) ως ειδικές περιπτώσεις. Επιτρέπει τον υπολογισμό της απόστασης μεταξύ δύο σημείων σε έναν n -διάστατο χώρο, λαμβάνοντας υπόψη

τις απόλυτες διαφορές υψωμένες σε κάποια δύναμη p και στη συνέχεια λαμβάνοντας τη ρίζα του αθροίσματος αυτών των διαφορών.

2.2.4.5. Νευρωνικά Δίκτυα

Ένα νευρωνικό δίκτυο [29] είναι μια τεχνική τεχνητής νοημοσύνης που επιτρέπει στους υπολογιστές να αναλύουν δεδομένα. Εμπνέεται από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Στον ανθρώπινο εγκέφαλο, διασυνδεδεμένοι νευρώνες στέλνουν σήματα για την επεξεργασία πληροφοριών. Παρομοίως, τα τεχνητά νευρωνικά δίκτυα αποτελούνται από λογισμικά μονάδες που ονομάζονται κόμβοι, οι οποίοι συνεργάζονται για την επίλυση προβλημάτων. Αυτά τα δίκτυα χρησιμοποιούν υπολογιστικά συστήματα για να πραγματοποιούν μαθηματικούς υπολογισμούς και να δημιουργούν προσαρμοστικά συστήματα που μαθαίνουν από τις αποτυχίες.

Ένα βασικό νευρωνικό δίκτυο αποτελείται από τρία στρώματα: το στρώμα εισόδου, το κρυφό στρώμα και το στρώμα εξόδου. Το στρώμα εισόδου λαμβάνει δεδομένα εξωτερικά και οι κόμβοι εισόδου επεξεργάζονται και αναλύουν τα δεδομένα προτού τα περάσουν στο κρυφό στρώμα. Τα κρυφά στρώματα λαμβάνουν τα δεδομένα από το προηγούμενο ή από άλλα κρυφά στρώματα και τα επεξεργάζονται περαιτέρω. Τα νευρωνικά δίκτυα μπορεί να έχουν πολλαπλά κρυφά στρώματα. Κάθε κρυφό στρώμα βελτιώνει την έξοδο του προηγούμενου στρώματος και την περνά στο επόμενο στρώμα. Τέλος, το στρώμα εξόδου παρουσιάζει τα επεξεργασμένα αποτελέσματα του νευρωνικού δικτύου.



Εικόνα 3: Νευρωνικό Δίκτυο

Υπάρχουν διάφοροι τύποι νευρωνικών δικτύων που χρησιμοποιούνται για διάφορους σκοπούς:

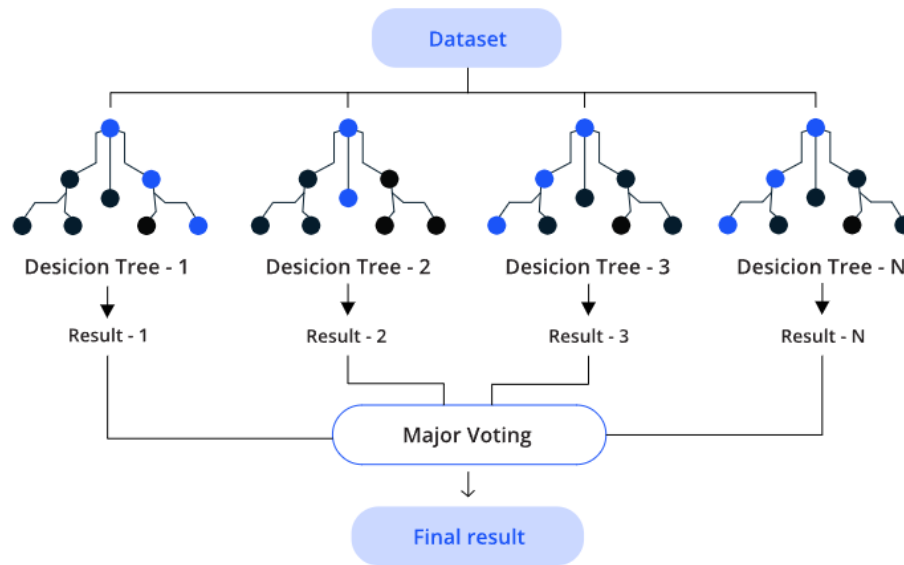
- **Νευρωνικά δίκτυα προώθησης (Feedforward neural networks):** Αυτά είναι επίσης γνωστά ως πολυστρωματικοί perceptrons (MLPs). Αυτά τα δίκτυα έχουν ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Τα MLPs χρησιμοποιούν σιγμοειδείς νευρώνες για να διαχειρίζονται αποτελεσματικά μη γραμμικά προβλήματα. Χρησιμοποιούνται ευρέως στην υπολογιστική όραση, στην επεξεργασία φυσικής γλώσσας και σε άλλες εφαρμογές.
- **Συνελκτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs):** Αυτά είναι ένας άλλος τύπος νευρωνικών δικτύων που χρησιμοποιείται συνήθως στην αναγνώριση εικόνας, στην αναγνώριση προτύπων και σε εργασίες υπολογιστικής όρασης. Αυτά τα δίκτυα αξιοποιούν έννοιες από την γραμμική άλγεβρα, όπως ο πολλαπλασιασμός πινάκων, για να αναγνωρίζουν πρότυπα μέσα σε εικόνες.

- **Αναδραστικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs):** Αυτά χαρακτηρίζονται από τους βρόχους ανάδρασης, που τους επιτρέπουν να επεξεργάζονται σειριακά δεδομένα ή δεδομένα χρονικής ακολουθίας. Είναι ιδιαίτερα χρήσιμα για την πρόβλεψη βάσει ιστορικών δεδομένων, όπως η πρόβλεψη χρηματιστηριακών τιμών ή οι προβλέψεις πωλήσεων. Τα RNNs έχουν εφαρμογές σε διάφορους τομείς όπου οι χρονικές εξαρτήσεις είναι κρίσιμες.

2.2.4.6. Random Forest

Ο αλγόριθμος Random Forest [30] είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης που ανήκει στη κατηγορία της επιβλεπόμενης μάθησης. Είναι ευέλικτος και μπορεί να εφαρμοστεί τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Ο αλγόριθμος βασίζεται στη μάθηση μέσω συνόλων (ensemble learning), συνδυάζοντας πολλαπλούς ταξινομητές για να επιλύσει σύνθετα προβλήματα και να βελτιώσει την απόδοση του μοντέλου.

Παίρνει το όνομά του επειδή από το γεγονός ότι αποτελείται από μια συλλογή δέντρων απόφασης, καθένα από τα οποία κατασκευάζεται με διαφορετικά υποσύνολα του αρχικού συνόλου δεδομένων. Ο αλγόριθμος κάνει προβλέψεις συλλέγοντας τις προβλέψεις από τα επιμέρους δέντρα και λαμβάνοντας την πλειοψηφική ψήφο. Αυτή η προσέγγιση συνόλου βελτιώνει την ακρίβεια του τελικού αποτελέσματος σε σύγκριση με την εμπιστοσύνη σε ένα μόνο δέντρο απόφασης.



Εικόνα 4:Random Forest

Ο αλγόριθμος έχει τρεις σημαντικές υπερπαραμέτρους που πρέπει να ρυθμιστούν πριν από την εκπαίδευση: το μέγεθος του κόμβου, τον αριθμό των δέντρων και τον αριθμό των χαρακτηριστικών που δειγματίζονται. Αυτές οι υπερπαραμέτροι καθορίζουν τη συμπεριφορά και την απόδοση του ταξινομητή τυχαίου δάσους για τα καθήκοντα παλινδρόμησης ή ταξινόμησης.

Ο αλγόριθμος αποτελείται από δέντρα απόφασης, όπου κάθε δέντρο κατασκευάζεται χρησιμοποιώντας ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης, το οποίο ονομάζεται bootstrap sample. Επιπλέον, ένα μέρος των δεδομένων, γνωστό ως Out-Of-Bag (OOB) sample, διατηρείται εκτός και δεν χρησιμοποιείται κατά την κατασκευή κάθε μεμονωμένου δέντρου. Αυτό το δείγμα OOB χρησιμεύει ως ένα σύνολο δοκιμής για την αξιολόγηση της απόδοσης του τυχαίου δάσους.

Η τεχνική του feature bagging εφαρμόζεται για να εισαγάγει περισσότερη τυχαιότητα και να μειώσει τη συσχέτιση μεταξύ των δέντρων. Αυτό περιλαμβάνει την τυχαία επιλογή ενός υποσυνόλου χαρακτηριστικών από το αρχικό σύνολο δεδομένων για κάθε δέντρο. Η

διαδικασία πρόβλεψης διαφέρει ανάλογα με το εκάστοτε πρόβλημα. Για προβλήματα παλινδρόμησης, οι προβλέψεις από τα μεμονωμένα δέντρα απόφασης λαμβάνονται ως μέσος όρος για την τελική πρόβλεψη. Σε καθήκοντα ταξινόμησης, η προβλεπόμενη κατηγορία καθορίζεται με πλειοψηφική ψηφοφορία, όπου επιλέγεται η πιο συχνή κατηγορική μεταβλητή μεταξύ των δέντρων.

Τέλος, το δείγμα ΟΟΒ χρησιμοποιείται για διασταυρωμένη επικύρωση (cross-validation) για την αξιολόγηση της ακρίβειας και της απόδοσης του μοντέλου τυχαίου δάσους πριν γίνει η τελική πρόβλεψη.

Χρησιμοποιώντας πολλά δέντρα, ο αλγόριθμος μειώνει τον κίνδυνο υπερπροσαρμογής (overfitting) και επιτυγχάνει μεγαλύτερη ακρίβεια στις προβλέψεις. Είναι ένα ισχυρό εργαλείο για την αντιμετώπιση διαφόρων εργασιών μηχανικής μάθησης και είναι ιδιαίτερα αποτελεσματικό όταν πρόκειται για σύνθετα σύνολα δεδομένων.

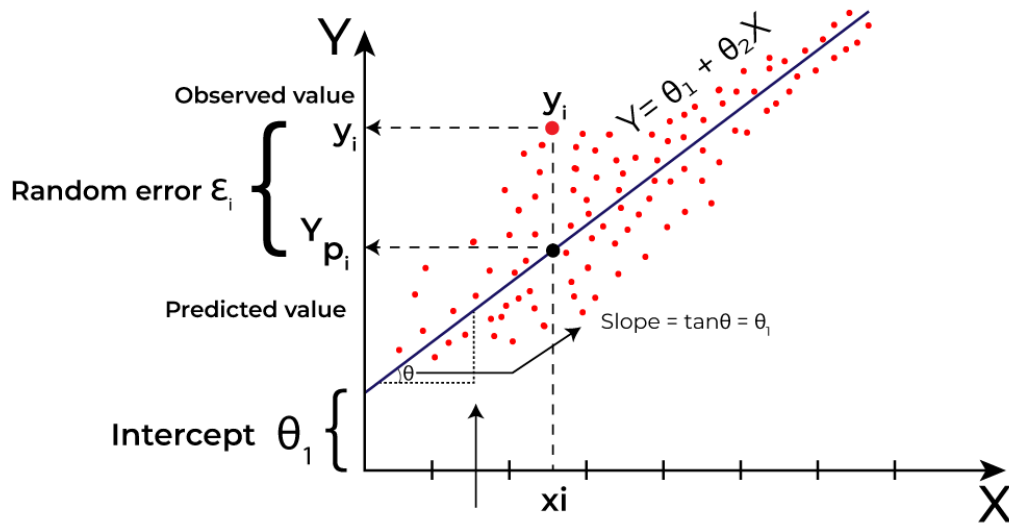
2.2.4.7. Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση [31] είναι ένας επιβλεπόμενος αλγόριθμος μηχανικής μάθησης που μοντελοποιεί τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής (Y) και μιας ή περισσότερων ανεξάρτητων μεταβλητών (X). Στην περίπτωση της μονομεταβλητής γραμμικής παλινδρόμησης, υπάρχει μόνο μία ανεξάρτητη μεταβλητή, ενώ στη πολυμεταβλητή γραμμική παλινδρόμηση υπάρχουν πολλές ανεξάρτητες μεταβλητές.

Ο στόχος της γραμμικής παλινδρόμησης είναι να βρει την καλύτερα προσαρμοσμένη γραμμική εξίσωση που μπορεί να προβλέψει την τιμή της εξαρτημένης μεταβλητής βάσει των ανεξάρτητων μεταβλητών. Η εξίσωση αυτή είναι μια ευθεία γραμμή σε ένα διδιάστατο χώρο, αλλά μπορεί να αντιπροσωπεύει σχέσεις υψηλότερης διάστασης στη πολυμεταβλητή παλινδρόμηση.

Ας υποθέσουμε ότι θέλουμε να προβλέψουμε τον μισθό ενός ατόμου (Y) βάσει της εργασιακής του εμπειρίας (X). Το μοντέλο γραμμικής παλινδρόμησης θα προσπαθήσει να

βρει τη γραμμή που αντιπροσωπεύει καλύτερα τη σχέση μεταξύ της εργασιακής εμπειρίας και του μισθού. Αυτή η γραμμή ονομάζεται γραμμή παλινδρόμησης ή καλύτερα προσαρμοσμένη γραμμή. Η γραμμή παλινδρόμησης καθορίζεται με την εκτίμηση της κλίσης και της τομής της γραμμής. Η κλίση αντιπροσωπεύει πόσο αλλάζει η εξαρτημένη μεταβλητή (μισθός) για μια μονάδα αλλαγής στην ανεξάρτητη μεταβλητή (εργασιακή εμπειρία). Η τομή αντιπροσωπεύει την προβλεπόμενη τιμή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή είναι μηδενική (δηλαδή το σημείο εκκίνησης της γραμμής στον άξονα y).



Εικόνα 5: Γραμμική Παλινδρόμηση

Αφού το μοντέλο έχει μάθει την γραμμική εξίσωση, μπορεί να χρησιμοποιηθεί για την πρόβλεψη νέων περιπτώσεων όπου η τιμή της ανεξάρτητης μεταβλητής (εργασιακή εμπειρία) είναι γνωστή, αλλά η τιμή της εξαρτημένης μεταβλητής (μισθός) είναι άγνωστη. Εισάγοντας την γνωστή τιμή της X στην εξίσωση, το μοντέλο μπορεί να εκτιμήσει την αντίστοιχη τιμή της Y .

Η γραμμική παλινδρόμηση χρησιμοποιείται ευρέως σε διάφορους τομείς για την κατανόηση και πρόβλεψη της συμπεριφοράς των μεταβλητών. Για παράδειγμα, στα οικονομικά, μπορεί να χρησιμοποιηθεί για την ανάλυση της σχέσης μεταξύ των κερδών μιας εταιρείας και της τιμής της μετοχής της. Στην οικονομία, μπορεί να βοηθήσει στην

πρόβλεψη της επίδρασης διαφόρων παραγόντων στους οικονομικούς δείκτες. Στην ψυχολογία, μπορεί να χρησιμοποιηθεί για τη μελέτη της σχέσης μεταξύ μεταβλητών όπως η ευφυΐα και η εργασιακή απόδοση.

Είναι σημαντικό να σημειωθεί ότι η γραμμική παλινδρόμηση υποθέτει μια γραμμική σχέση μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών. Αν η σχέση είναι πιο σύνθετη ή μη γραμμική, μπορεί να είναι πιο κατάλληλες εναλλακτικές τεχνικές παλινδρόμησης.

2.2.5. Εφαρμογές Μηχανικής Μάθησης

Ακολουθούν μερικά παραδείγματα εφαρμογών μηχανικής μάθησης [19] [20]:

Πρόβλεψη προτύπων κυκλοφορίας: Η μηχανική μάθηση παίζει καθοριστικό ρόλο στους τομείς της εφοδιαστικής αλυσίδας και των μεταφορών, ιδιαίτερα στην πρόβλεψη προτύπων κυκλοφορίας. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να προβλέψουν με ακρίβεια τα επίπεδα συμφόρησης κυκλοφορίας, αναλύοντας εκτεταμένα ιστορικά δεδομένα κυκλοφορίας, συμπεριλαμβανομένων των καιρικών συνθηκών, της ώρας της ημέρας και άλλων παραγόντων. Αυτές οι προβλέψεις επιτρέπουν τη βελτιστοποίηση της ροής της κυκλοφορίας, τη μείωση των εμποδίων και των καθυστερήσεων και τη βελτίωση των χρόνων ταξιδιού των οχημάτων. Επιπλέον, η μηχανική μάθηση βοηθά στην πρόβλεψη της ζήτησης για δημόσια συγκοινωνία και παρέχει στους οδηγούς ειδοποιήσεις σε πραγματικό χρόνο για την κυκλοφορία και εναλλακτικές διαδρομές. Χρησιμοποιώντας τη δύναμη της μηχανικής μάθησης για την πρόβλεψη των προτύπων κυκλοφορίας, οι πόλεις και οι υπηρεσίες μεταφορών μπορούν να βελτιώσουν την αποδοτικότητα, την ασφάλεια και τη βιωσιμότητα των συστημάτων μεταφορών τους, ενώ παράλληλα βελτιώνουν τη συνολική εμπειρία ταξιδιού για τους πολίτες και μειώνουν τις εκπομπές διοξειδίου του άνθρακα.

Ανίχνευση απάτης: Η μηχανική μάθηση είναι κρίσιμη για την ανίχνευση απάτης σε διάφορους τομείς, όπως οι χρηματοπιστωτικές υπηρεσίες και το ηλεκτρονικό εμπόριο. Αναλύοντας μεγάλους όγκους συναλλακτικών δεδομένων, οι αλγόριθμοι μηχανικής μάθησης μπορούν να εντοπίσουν πρότυπα και ανωμαλίες που υποδεικνύουν πιθανή δόλια δραστηριότητα. Αυτοί οι αλγόριθμοι μαθαίνουν συνεχώς από προηγούμενες περιπτώσεις απάτης, βελτιώνοντας την ακρίβειά τους και την προσαρμοστικότητά τους σε νέες μορφές δόλιας δραστηριότητας. Χρησιμοποιώντας τη μηχανική μάθηση για την ανίχνευση απάτης, οι οργανισμοί μπορούν να ελαχιστοποιήσουν τις χρηματοοικονομικές απώλειες, να προστατεύσουν τη φήμη τους και να ενισχύσουν την εμπιστοσύνη των καταναλωτών. Επιπλέον, η μηχανική μάθηση προσφέρει δυνατότητες ανίχνευσης απάτης σε πραγματικό χρόνο και βελτιστοποιημένες διαδικασίες διερεύνησης. Η μηχανική μάθηση είναι ένα ισχυρό εργαλείο για την πρόληψη έναντι των εξελισσόμενων κυβερνοαπειλών, την προστασία των πόρων και τη διασφάλιση της ασφάλειας των πελατών από εγκληματικές δραστηριότητες.

Αναγνώριση εικόνας: Η μηχανική μάθηση έχει ευρεία εφαρμογή σε πολλούς τομείς, όπως η ασφάλεια, το λιανικό εμπόριο και η υγειονομική περίθαλψη. Σε αυτούς τους τομείς, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται εκτενώς για την επεξεργασία και ταξινόμηση μεγάλων ποσοτήτων οπτικών δεδομένων, όπως ιατρικές εικόνες, φωτογραφίες προϊόντων και υλικό από κάμερες παρακολούθησης, για την αναγνώριση προτύπων και διακριτικών χαρακτηριστικών.

Αναγνώριση φωνής Η μηχανική μάθηση επιτρέπει στους υπολογιστές να κατανοούν και να αναγνωρίζουν την ανθρώπινη ομιλία, μετατρέποντάς την σε κείμενο ή άλλες χρήσιμες μορφές δεδομένων. Αυτή η εξέλιξη έχει οδηγήσει στην ανάπτυξη συστημάτων ελέγχου φωνής και εικονικών βοηθών, επηρεάζοντας τον τρόπο με τον οποίο αλληλεπιδρούμε με την τεχνολογία.

2.3. Ανάλυση Μεγάλων Δεδομένων

2.3.1. Εισαγωγή στην Ανάλυση Μεγάλων Δεδομένων

Η ανάλυση μεγάλων δεδομένων [32] – [35] αναφέρεται στη διαδικασία εξαγωγής ενεργών γνώσεων και σημαντικών προτύπων από τεράστια και πολύπλοκα σύνολα δεδομένων. Περιλαμβάνει τη χρήση προηγμένων εργαλείων, τεχνολογιών και αλγορίθμων για την επεξεργασία, αποθήκευση και ανάλυση δεδομένων με αποδοτικό τρόπο. Χρησιμοποιώντας τη δύναμη της ανάλυσης μεγάλων δεδομένων, οι οργανισμοί μπορούν να ανακαλύψουν κρυφά πρότυπα, τάσεις και συσχετίσεις που μπορούν να οδηγήσουν στην ανάπτυξη της επιχείρησης, τη βελτίωση της λειτουργικής αποδοτικότητας και την ενίσχυση της διαδικασίας λήψης αποφάσεων.

2.3.2. Η Σημασία της Διαχείρισης και Ανάλυσης Μεγάλων Συνόλων Δεδομένων

Η διαχείριση και ανάλυση μεγάλων συνόλων δεδομένων είναι κρίσιμη για πολλούς λόγους:

Πρώτον, τα μεγάλα δεδομένα περιέχουν πολύτιμες πληροφορίες που μπορούν να παρέχουν στις οργανώσεις ανταγωνιστικό πλεονέκτημα. Εξάγοντας γνώσεις από μεγάλα σύνολα δεδομένων, οι επιχειρήσεις μπορούν να εντοπίζουν τις προτιμήσεις των πελατών, τις τάσεις της αγοράς και τις πιθανές ευκαιρίες.

Δεύτερον, τα μεγάλα σύνολα δεδομένων συχνά περιλαμβάνουν ροές δεδομένων σε πραγματικό χρόνο ή σχεδόν σε πραγματικό χρόνο, απαιτώντας από τις οργανώσεις να επεξεργάζονται και να αναλύουν δεδομένα άμεσα. Η ικανότητα να καταγράφουν και να αναλύουν δεδομένα σε πραγματικό χρόνο επιτρέπει στις επιχειρήσεις να λαμβάνουν άμεσες αποφάσεις και να ανταποκρίνονται γρήγορα στις μεταβαλλόμενες συνθήκες της αγοράς.

Τρίτον, τα μεγάλα σύνολα δεδομένων έχουν διάφορες μορφές και δομές, συμπεριλαμβανομένων των δομημένων, ημιδομημένων και μη δομημένων δεδομένων. Αναλύοντας διάφορους τύπους δεδομένων, οι οργανισμοί μπορούν να αποκτήσουν μια ολοκληρωμένη κατανόηση των λειτουργιών τους, των πελατών τους και των αγορών τους.

Τέλος, η αξιοπιστία των δεδομένων είναι ένα σημαντικό ζήτημα στην ανάλυση μεγάλων δεδομένων. Τα μεγάλα σύνολα δεδομένων συχνά περιέχουν θόρυβο, σφάλματα και ασυνέπειες. Η ανάλυση και ο καθαρισμός των δεδομένων βοηθά τους οργανισμούς να διασφαλίσουν την ακρίβεια και την αξιοπιστία των γνώσεων που εξάγονται από αυτά.

2.3.3. Προκλήσεις στην Επεξεργασία Μεγάλων Δεδομένων

Η επεξεργασία μεγάλων δεδομένων παρουσιάζει πολλές προκλήσεις λόγω των χαρακτηριστικών των μεγάλων συνόλων δεδομένων. Ας εξετάσουμε τις κύριες προκλήσεις:

Όγκος Δεδομένων: Τα μεγάλα δεδομένα αναφέρονται σε σύνολα δεδομένων που υπερβαίνουν τις δυνατότητες των παραδοσιακών συστημάτων επεξεργασίας δεδομένων. Η αντιμετώπιση τεράστιων όγκων δεδομένων απαιτεί επεκτάσιμες και κατανεμημένες αρχιτεκτονικές υπολογιστών που μπορούν να διαχειριστούν το φορτίο των δεδομένων.

Ταχύτητα Δεδομένων: Σε πολλές εφαρμογές, τα δεδομένα παράγονται και συλλέγονται με πρωτοφανή ταχύτητα. Η επεξεργασία και ανάλυση δεδομένων σε πραγματικό χρόνο είναι κρίσιμη για να εξάγουν οι οργανώσεις άμεσα γνώσεις και να λαμβάνουν έγκαιρα αποφάσεις.

Ποικιλία Δεδομένων: Τα μεγάλα δεδομένα περιλαμβάνουν διάφορους τύπους δεδομένων, συμπεριλαμβανομένων των δομημένων, ημιδομημένων και μη δομημένων δεδομένων. Η ανάλυση και ενσωμάτωση δεδομένων από διάφορες πηγές και μορφές είναι απαραίτητη για την απόκτηση μιας ολοκληρωμένης εικόνας των δεδομένων.

Αξιοπιστία Δεδομένων: Η αξιοπιστία των δεδομένων αναφέρεται στην ακρίβεια, την αξιοπιστία και την εμπιστοσύνη που μπορούμε να έχουμε στα δεδομένα. Τα μεγάλα σύνολα δεδομένων συχνά περιέχουν θόρυβο, σφάλματα και ασυνέπειες που μπορούν να επηρεάσουν την ποιότητα των γνώσεων που εξάγονται από αυτά. Η διασφάλιση της ποιότητας των δεδομένων είναι ένα σημαντικό μέρος της ανάλυσης μεγάλων δεδομένων.

2.3.4. Εργαλεία και Τεχνολογίες για την Ανάλυση Μεγάλων Δεδομένων

Για τη διαχείριση και την ανάλυση των μεγάλων δεδομένων, έχουν αναπτυχθεί διάφορα εργαλεία και τεχνολογίες στον τομέα της ανάλυσης μεγάλων δεδομένων. Ας εξετάσουμε μερικά από τα βασικά:

- **Hadoop:** Το Hadoop είναι ένα ανοιχτού κώδικα πλαίσιο που παρέχει δυνατότητες κατανεμημένης αποθήκευσης και επεξεργασίας για μεγάλα δεδομένα. Αποτελείται από το Σύστημα Αρχείων Κατανεμημένου Hadoop (HDFS) για την αποθήκευση δεδομένων και το μοντέλο προγραμματισμού MapReduce για παράλληλη επεξεργασία.
- **ApacheSpark:** Το ApacheSpark είναι ένα γρήγορο και γενικής χρήσης κατανεμημένο σύστημα υπολογιστών σχεδιασμένο για επεξεργασία και ανάλυση μεγάλων δεδομένων. Προσφέρει επεξεργασία δεδομένων στη μνήμη, καθιστώντας το σημαντικά ταχύτερο από τα παραδοσιακά συστήματα επεξεργασίας βασισμένα σε δίσκους.
- **Βάσεις δεδομένων NoSQL:** Οι βάσεις δεδομένων NoSQL, όπως το MongoDB και το Cassandra, είναι σχεδιασμένες να διαχειρίζονται μεγάλους όγκους δομημένων, ημιδομημένων και μη δομημένων δεδομένων. Παρέχουν επεκτασιμότητα και ευελιξία για την αποθήκευση και αναζήτηση διαφόρων τύπων δεδομένων.
- **Αλγόριθμοι Μηχανικής Μάθησης:** Οι αλγόριθμοι μηχανικής μάθησης παίζουν κρίσιμο ρόλο στην ανάλυση μεγάλων δεδομένων, επιτρέποντας στις οργανώσεις να

ανακαλύπτουν πρότυπα, να κάνουν προβλέψεις και να αυτοματοποιούν διαδικασίες λήψης αποφάσεων. Αλγόριθμοι όπως τα δέντρα απόφασης, τα τυχαία δάση και τα νευρωνικά δίκτυα χρησιμοποιούνται συχνά για την ανάλυση μεγάλων συνόλων δεδομένων.

2.3.5. Συλλογή και Προετοιμασία Δεδομένων

Πριν από την ανάλυση, τα δεδομένα πρέπει να συλλεχθούν και να προετοιμαστούν για επεξεργασία. Τα ακόλουθα βήματα εμπλέκονται στη συλλογή και προετοιμασία δεδομένων:

- **Εξαγωγή Δεδομένων:** Η εξαγωγή δεδομένων περιλαμβάνει την ανάκτηση δεδομένων από διάφορες πηγές, όπως βάσεις δεδομένων, APIs, web scraping ή συσκευές αισθητήρων. Είναι σημαντικό να συγκεντρωθούν σχετικά και υψηλής ποιότητας δεδομένα για ακριβή ανάλυση.
- **Καθαρισμός Δεδομένων:** Ο καθαρισμός δεδομένων στοχεύει στην απομάκρυνση ασυνεπειών, σφαλμάτων και ακραίων τιμών από το σύνολο δεδομένων. Περιλαμβάνει εργασίες όπως η διαχείριση των ελλιπών τιμών, η επίλυση διπλοτύπων και η τυποποίηση των μορφών δεδομένων.
- **Μετασχηματισμός Δεδομένων:** Ο μετασχηματισμός δεδομένων περιλαμβάνει τη μετατροπή των δεδομένων σε κατάλληλη μορφή για ανάλυση. Μπορεί να περιλαμβάνει εργασίες όπως η κανονικοποίηση δεδομένων, η συγχώνευση ή η μηχανική χαρακτηριστικών.
- **Ενσωμάτωση Δεδομένων:** Η ενσωμάτωση δεδομένων συνδυάζει δεδομένα από πολλαπλές πηγές σε ένα ενιαίο σύνολο δεδομένων. Διασφαλίζει ότι τα δεδομένα είναι συνεπή και μπορούν να αναλυθούν συνολικά.

2.3.6. Αποθήκευση και Διαχείριση Δεδομένων

Η αποτελεσματική αποθήκευση και διαχείριση δεδομένων είναι κρίσιμη για την ανάλυση μεγάλων δεδομένων. Διάφορες προσεγγίσεις χρησιμοποιούνται συνήθως για την αποθήκευση και διαχείριση μεγάλων συνόλων δεδομένων:

- **Κατανεμημένα Συστήματα Αρχείων:** Τα κατανεμημένα συστήματα αρχείων, όπως το HDFS, επιτρέπουν στις οργανώσεις να αποθηκεύουν και να διαχειρίζονται δεδομένα σε πολλαπλές μηχανές σε ένα κατανεμημένο περιβάλλον. Παρέχουν υψηλή επεκτασιμότητα και αντοχή σε σφάλματα.
- **Αποθήκη Δεδομένων (Data Warehouse):** Η αποθήκη δεδομένων περιλαμβάνει την αποθήκευση μεγάλων ποσοτήτων δεδομένων από διάφορες πηγές με οργανωμένο και δομημένο τρόπο. Διευκολύνει την αποδοτική αναζήτηση και ανάλυση δεδομένων.
- **Λίμνη Δεδομένων (Data Lake):** Μια λίμνη δεδομένων είναι ένα κεντρικό αποθετήριο που αποθηκεύει μεγάλους όγκους δομημένων, ημιδομημένων και μη δομημένων δεδομένων στη φυσική τους μορφή. Παρέχει ευελιξία για την εξερεύνηση και ανάλυση δεδομένων.

2.3.7. Επεξεργασία και Ανάλυση Δεδομένων

Η επεξεργασία και ανάλυση μεγάλων συνόλων δεδομένων απαιτεί εξειδικευμένες τεχνικές και προσεγγίσεις. Ας εξετάσουμε μερικές από τις βασικές μεθόδους που χρησιμοποιούνται στην ανάλυση μεγάλων δεδομένων:

- **Επεξεργασία Παρτίδων:** Η επεξεργασία παρτίδων περιλαμβάνει την επεξεργασία δεδομένων σε μεγάλους όγκους σε τακτά χρονικά διαστήματα. Είναι κατάλληλη για εφαρμογές που δεν απαιτούν ανάλυση σε πραγματικό χρόνο και μπορούν να αντέξουν κάποια καθυστέρηση.

- **Επεξεργασία Ροών:** Η επεξεργασία ροών επιτρέπει την ανάλυση δεδομένων σε πραγματικό χρόνο καθώς παράγονται. Χρησιμοποιείται για εφαρμογές που απαιτούν άμεσα αποτελέσματα και γρήγορους χρόνους απόκρισης.
- **Ανάλυση στη Μνήμη:** Η ανάλυση στη μνήμη εκμεταλλεύεται τη δύναμη της RAM για να αποθηκεύει και να επεξεργάζεται δεδομένα σε πραγματικό χρόνο. Παρέχει ταχύτερη πρόσβαση και ανάλυση δεδομένων σε σύγκριση με τα συστήματα που βασίζονται σε δίσκους.

2.3.8. Οπτικοποίηση και Αναφορά Δεδομένων

Η οπτικοποίηση και η αναφορά δεδομένων παίζουν κρίσιμο ρόλο στην επικοινωνία των γνώσεων που προκύπτουν από μεγάλα σύνολα δεδομένων. Αποτελεσματικές τεχνικές οπτικοποίησης βοηθούν τους ενδιαφερόμενους να κατανοούν σύνθετα πρότυπα και τάσεις. Ορισμένες βασικές πτυχές της οπτικοποίησης και αναφοράς δεδομένων περιλαμβάνουν:

- **Οπτικοποίηση Μεγάλων Συνόλων Δεδομένων:** Η οπτικοποίηση μεγάλων συνόλων δεδομένων απαιτεί τεχνικές που μπορούν να διαχειριστούν τον όγκο και την πολυπλοκότητα των δεδομένων. Εργαλεία όπως το Tableau, το PowerBI και το D3.js προσφέρουν ισχυρές δυνατότητες οπτικοποίησης για μεγάλα δεδομένα.
- **Διαδραστικοί Πίνακες Ελέγχου:** Οι διαδραστικοί πίνακες ελέγχου επιτρέπουν στους χρήστες να εξερευνούν και να αλληλεπιδρούν με τα δεδομένα οπτικά. Παρέχουν ένα διαισθητικό περιβάλλον για την εξερεύνηση των δεδομένων και επιτρέπουν στους χρήστες να εμβαθύνουν σε συγκεκριμένες λεπτομέρειες.
- **Εργαλεία Αναφοράς:** Τα εργαλεία αναφοράς, όπως το CrystalReports και το Jaspersoft, επιτρέπουν τη δημιουργία πληροφοριακών και οπτικά ελκυστικών αναφορών. Βοηθούν στη σύνοψη και παρουσίαση των γνώσεων που προκύπτουν από την ανάλυση μεγάλων δεδομένων.

2.3.9. Ασφάλεια και Ιδιωτικότητα Δεδομένων

Καθώς ο όγκος και η ευαισθησία των δεδομένων αυξάνονται, η διασφάλιση της ασφάλειας και της ιδιωτικότητας των δεδομένων γίνεται υψίστης σημασίας. Οι οργανισμοί πρέπει να εφαρμόσουν ισχυρά μέτρα για την προστασία των δεδομένων από μη εξουσιοδοτημένη πρόσβαση και παραβιάσεις. Βασικές εκτιμήσεις για την ασφάλεια και την ιδιωτικότητα των δεδομένων στην ανάλυση μεγάλων δεδομένων περιλαμβάνουν:

- **Κρυπτογράφηση Δεδομένων:** Η κρυπτογράφηση δεδομένων μετατρέπει τα δεδομένα σε μη αναγνώσιμη μορφή, διασφαλίζοντας ότι μόνο οι εξουσιοδοτημένοι χρήστες μπορούν να έχουν πρόσβαση και να τα αποκρυπτογραφούν. Τεχνικές κρυπτογράφησης όπως το AES (Advanced Encryption Standard) χρησιμοποιούνται ευρέως για την ασφαλή φύλαξη των δεδομένων.
- **Έλεγχος Πρόσβασης:** Οι μηχανισμοί ελέγχου πρόσβασης ρυθμίζουν ποιος μπορεί να έχει πρόσβαση και να χειρίζεται δεδομένα. Ο έλεγχος πρόσβασης βάσει ρόλων (RBAC) και ο έλεγχος πρόσβασης βάσει χαρακτηριστικών (ABAC) είναι συνήθεις προσεγγίσεις για τη διαχείριση της πρόσβασης στα δεδομένα σε περιβάλλοντα μεγάλων δεδομένων.
- **Τεχνικές Ανωνυμοποίησης:** Οι τεχνικές ανωνυμοποίησης βοηθούν στην προστασία της ιδιωτικότητας αφαιρώντας ή θολώνοντας τις πληροφορίες που ταυτοποιούν άτομα από το σύνολο δεδομένων.

3. ΜΕΘΟΔΟΛΟΓΙΑ ΠΕΙΡΑΜΑΤΙΚΟΥ ΜΕΡΟΥΣ

3.1. Βασική Φιλοσοφία

Βασικό αντικείμενο του πειραματικού μέρους είναι η εφαρμογή μοντέλων μηχανικής μάθησης πάνω σε σύνολο δεδομένων, το οποίο έχει προκύψει από απαντήσεις σε ερωτήσεις ερωτηματολογίου σχετικά με ακριβή τίτλο: «Καινοτομία – Χρήση Τεχνολογιών Πληροφορικής και Cloud – Επιπτώσεις της Οικονομικής Κρίσης στις Ελληνικές Επιχειρήσεις». Στόχος των συγκεκριμένων μοντέλων μηχανικής μάθησης είναι πρόβλεψη της απάντησης σε συγκεκριμένες ερωτήσεις (θα παρατεθούν στη συνέχεια).

Όσο αφορά το εργαλείο υλοποίησης της πειραματικής διαδικασίας είναι το εργαλείο RapidMiner [36] [37], ένα ισχυρό εργαλείο κατάλληλο για τη διαχείριση μεγάλου όγκου δεδομένων και την εφαρμογή μοντέλων (στην περίπτωση μας προβλεπτικών) μηχανικής μάθησης.

Πιο συγκεκριμένα, τα μοντέλα τα οποία θα εφαρμοστούν, έχουν αναλυθεί σε προηγούμενες ενότητες και είναι τα:

- Δέντρο Αποφάσεων
- Random Forest

Τα παραπάνω μοντέλα αξιολογήθηκαν ως προς τις παρακάτω μετρικές αξιολόγησης:

- Accuracy
- Precision
- Recall

Τέλος, θα παρατεθούν τα αντίστοιχα confusionmatrixes που προέκυψαν από την πραγματοποίηση των πειραματικών διαδικασιών.

3.2. Σύνολο Δεδομένων Πειραματικής Διαδικασίας

Όπως αναφέρθηκε και πολύ συνοπτικά προηγουμένως, το σύνολο δεδομένων δημιουργήθηκε με βάση ερωτηματολόγιο του 2015 με τίτλο: Καινοτομία – Χρήση Τεχνολογιών Πληροφορικής και Cloud – Επιπτώσεις της Οικονομικής Κρίσης στις Ελληνικές Επιχειρήσεις». Το συγκεκριμένο ερωτηματολόγιο περιέχει αρκετές ερωτήσεις ώστε να διαπιστωθεί η επίδραση της οικονομικής κρίσης όσο αφορά την χρήση τεχνολογιών πληροφορικής και cloud και κάποια ενδιαφέροντα αποσπάσματα του παρατίθενται στη συνέχεια:

1.8 Ο αριθμός των κυριότερων ανταγωνιστών σας ανέρχεται σε _____

1.9 Πόσο έντονος είναι ο ανταγωνισμός που αντιμετωπίζετε από άλλες επιχειρήσεις :

(α) ως προς την τιμή (price competition) :

(1) πολύ μικρός (2) μικρός (3) μέτριος (4) μεγάλος (5) πολύ μεγάλος
☐ ☐ ☐ ☐ ☐

(β) ως προς άλλους άξονες ανταγωνισμού (non-price competition), π.χ. ως προς την ποιότητα των προϊόντων/υπηρεσιών, την προσαρμογή τους σε ειδικές ανάγκες των πελατών (customization), την εισαγωγή νέων προϊόντων/υπηρεσιών, την παροχή υποστηρικτικών υπηρεσιών, κλπ :

(1) πολύ μικρός (2) μικρός (3) μέτριος (4) μεγάλος (5) πολύ μεγάλος
☐ ☐ ☐ ☐ ☐

Απαντήσατε τα επόμενα δύο ερωτήματα (1.10 και 1.11) στην κλίμακα 1 έως 5, όπου: 5 = σε πολύ μεγάλο βαθμό, 4 = σε μεγάλο βαθμό, 3 = σε μέτριο βαθμό, 2 = σε μικρό βαθμό και 1 = καθόλου (βάλτε σε κύκλο τον κατάλληλο αριθμό)

Εικόνα 6: Ερωτηματολόγιο - Απόσπασμα 1

2.5 Πραγματοποίησε η επιχείρησή σας κατά την τελευταία τριετία 2012 - 2013 - 2014 δραστηριότητες Έρευνας και Ανάπτυξης (Research and Development – R&D);

ΝΑΙ ☐

ΟΧΙ ☐

Εάν ΝΑΙ: α) Οι δραστηριότητες αυτές Έρευνας και Ανάπτυξης στην παραπάνω τριετία ήσαν:

Περιστασιακές ☐

Συνεχείς ☐

β) Οι ετήσιες δαπάνες σας για Έρευνα & Ανάπτυξη (όπως αυτή ορίστηκε παραπάνω) κατά το 2014 σε τι ποσοστό του κύκλου εργασιών (έσοδα από πωλήσεις) του 2014 ανήλθαν ; _____ %

2.6 Πόσο μεγάλη είναι η έκταση της υπάρχουσας διεθνούς τεχνολογικής γνώσης (τόσο βασικής όσο και εφαρμοσμένης) στην θεματική περιοχή δραστηριοποίησης της επιχείρησής σας, η οποία θα μπορούσε να χρησιμοποιηθεί για την δημιουργία καινοτομιών προϊόντων/υπηρεσιών και μεθόδων/διαδικασιών;

(1) πολύ μικρή (2) μικρή (3) μέτρια (4) μεγάλη (5) πολύ μεγάλη
☐ ☐ ☐ ☐ ☐

Εικόνα 7: Ερωτηματολόγιο - Απόσπασμα 2

2.7 Σε ποιά βαθμό η επιχείρησή σας μπορεί εύκολα και γρήγορα να πραγματοποιήσει τα ακόλουθα;

να αντιδράσει σε εισαγωγές νέων προϊόντων και υπηρεσιών από ανταγωνιστές	1	2	3	4	5
να προσαρμόσει ένα προϊόν ή υπηρεσία της ώστε να καλύψει ειδικές ανάγκες ενός πελάτη	1	2	3	4	5
να ανταποκριθεί σε αλλαγές της ζήτησης ενός προϊόντος ή υπηρεσίας σας από τους πελάτες	1	2	3	4	5
να επεκταθεί σε νέες αγορές στο εσωτερικό της χώρας ή στο εξωτερικό	1	2	3	4	5
να αυξήσει ή να μειώσει την ποικιλία των προϊόντων και υπηρεσιών που προσφέρει	1	2	3	4	5
να εισάγει νέες τιμολογιακές πολιτικές ως απάντηση σε αλλαγές τιμών των ανταγωνιστών	1	2	3	4	5
να εισάγει νέες τεχνολογίες για την παραγωγή των προϊόντων και υπηρεσιών της με χαμηλότερο κόστος και/ή καλύτερη ποιότητα	1	2	3	4	5
να αλλάξει προμηθευτές ώστε να επιτύχει χαμηλότερα κόστη προμηθειών, υψηλότερη ποιότητα, ταχύτερους χρόνους παράδοσης, κλπ.	1	2	3	4	5

2.8 Σε ποιά βαθμό η επιχείρησή σας έχει τις ακόλουθες ικανότητες απόκτησης/διαχείρισης/αξιοποίησης χρήσιμης εξωτερικής γνώσης (σχετικής με την αγορά, π.χ. απαιτήσεις πελατών, προϊόντα/υπηρεσίες ανταγωνιστών, κλπ. ή με νέες τεχνολογίες και υλικά, νομοθεσίες, ευκαιρίες, απειλές) (absorptive capacity);



Μπορεί να αναγνωρίζει/εντοπίζει χρήσιμη για την επιχείρηση γνώση από το εξωτερικό περιβάλλον (π.χ. προμηθευτές, συνεργάτες, ερευνητικά κέντρα και πανεπιστήμια, κλπ.)	1	2	3	4	5
Διαθέτει πρακτικές και διαδικασίες για την εσωτερική διάχυση και ανάλυση της εξωτερικής αυτής γνώσης	1	2	3	4	5
Μπορεί να αφομοιώνει την εξωτερική αυτή γνώση, και να την εντάξει στην υπάρχουσα γνωστική βάση της επιχείρησης, ώστε αυτή να εμπλουτίζεται και να διευρύνεται	1	2	3	4	5
Μπορεί να χρησιμοποιήσει/αξιοποιήσει την εξωτερική αυτή γνώση, σε συνδυασμό με υπάρχουσα εσωτερική γνώση, για την βελτίωση της λειτουργίας και ανταγωνιστικότητας της επιχείρησης, και την δημιουργία καινοτομιών προϊόντων/υπηρεσιών-μεθόδων/διαδικασιών	1	2	3	4	5

Εικόνα 8: Ερωτηματολόγιο - Απόσπασμα 3

5. Επιπτώσεις της Οικονομικής Κρίσης

5.1 Το συνολικό ποσοστό μεταβολής (αύξης ή μείωσης) των πωλήσεών σας κατά την περίοδο της οικονομικής κρίσης 2009-2014 ήταν: (βάλτε X στο κατάλληλο τετράγωνο)

μείωση άνω του 100%	μείωση 80-100%	μείωση 60-80%	μείωση 40-60%	μείωση 20-40%	μείωση 1-20%	αμετάβλητες πωλήσεις	αύξηση 1-20%	αύξηση 20-40%	αύξηση 40-60%	αύξηση 60-80%	αύξηση 80-100%	αύξηση άνω του 100%

5.2 Πόσο επηρέασε η οικονομική κρίση κατά την περίοδο 2009-2014 την επιχείρησή σας (σε σύγκριση με την προηγούμενη πενταετία);

πολύ σοβαρά σοβαρά μέτρια λίγο καθόλου



Εάν η απάντησή σας ήταν 'καθόλου': προχωρείστε το τέλος του ερωτηματολογίου

5.3 Πόσο σημαντικές ήταν οι παρακάτω επιπτώσεις της οικονομικής κρίσης κατά την περίοδο 2009-2014 στην επιχείρησή σας: (βάλτε X στο κατάλληλο τετράγωνο)

	όχι σημαντική	μικρή	μέτρια	σημαντική	πολύ σημαντική
μείωση των πωλήσεων στο εσωτερικό					
μείωση των πωλήσεων στο εξωτερικό					
μείωση των κερδών/αύξηση των ζημιών					
μείωση του προσωπικού					
μείωση των συνολικών επενδύσεων					
μείωση της καινοτομίας (π.χ. εισαγωγής νέων προϊόντων)					
αύξηση χρεών					
περιορισμός ρευστότητας					

Εικόνα 9: Ερωτηματολόγιο - Απόσπασμα 4

5.5 Πόσο σημαντικές ήταν οι επιπτώσεις της οικονομικής κρίσης κατά την περίοδο 2009-2014 στις παρακάτω κατηγορίες επενδύσεων και λειτουργικών δαπανών πληροφορικής στην επιχείρησή σας;

	αύξηση	δεν είχε σημαντική επίπτωση	μικρή μείωση	μέτρια μείωση	σημαντική μείωση	πολύ σημαντική μείωση
Συνολικές hard και soft επενδύσεις πληροφορικής (σε hardware, software, ICT training και ICT consulting)						
Επενδύσεις σε υλικό (hardware)						
Επενδύσεις σε λογισμικό (software)						
Επενδύσεις σε εκπαίδευση προσωπικού σε θέματα πληροφορικής (ICT training)						
Επενδύσεις σε συμβουλευτικές υπηρεσίες πληροφορικής (ICT consulting)						
Συνολικές λειτουργικές δαπάνες πληροφορικής (για προσωπικό και υπηρεσίες)						
Δαπάνες προσωπικού πληροφορικής						
Δαπάνες εξωτερικής ανάθεσης εργασιών πληροφορικής (ICT outsourcing)						
Δαπάνες για υπηρεσίες Cloud Computing						

Εικόνα 10: Ερωτηματολόγιο - Απόσπασμα 5

Το συγκεκριμένο ερωτηματολόγιο διανεμήθηκε σε 363 συμμετέχοντες και έτσι δημιουργήθηκε ένα σύνολο δεδομένων, το οποίο δόθηκε έτοιμο σε μορφή excel και έχει το ακόλουθο μέγεθος:

Πλήθος Γραμμών

363

Πλήθος Στηλών

171

Ακολουθεί ενδεικτικό screenshot από απόσπασμα των δεδομένων:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
ID	Sector	A13	A14	A15A	A15B	A15C	A15D	A16	A16A	A17	A18	A19A	A19B	A110A	A110B	A110C	A111A
1	18	1,992	90	0.33	0.39	0.17	0.11	0	0	2	#NULL!	4	1	3	4	4	1
2	6	1,984	114	0.75	0.15	0.10	0.00	0	0	2	3	4	2	3	5	5	2
3	20	1,989	103	0.05	0.25	0.05	0	0	4	4	3	5	4	4	4	5	3
4	18	1,981	218	0.30	0.10	0.60	0.00	0	0	4	0	4	3	4	5	5	4
5	14	1,998	58	0.22	0.20	0.31	0.27	1	1	1	4	3	3	3	4	2	1
6	14	1,983	272	0.75	0.20	0.05	0.00	1	0	2	4	3	3	1	4	4	1
7	23	1,979	250	1.00	0.00	0.00	0.00	0	0	2	3	5	4	2	5	5	2
8	28	1,955	133	0.32	0.57	0.00	0.11	0	0	5	50	1	2	3	4	4	1
9	19	1,990	54	0.25	0.25	0.50	0.00	1	0	2	7	4	3	4	5	5	1
10	8	1,981	62	0.06	0.06	0.88	0.00	1	0	4	3	5	3	4	5	3	2
11	12	1,984	725	0.35	0.25	0.32	0.08	1	1	3	6	5	4	3	3	3	2
12	19	1,917	50	0.17	0.08	0.50	0.25	0	0	3	2	2	2	2	4	3	1
13	28	1,989	55	0.95	0.03	0.02	0.00	1	1	3	150	4	4	2	5	3	3
14	28	1,993	30	1.00	0.00	0.00	0.00	0	0	3	6	5	4	4	5	2	1
15	28	1,987	95	0.95	0.04	0.01	0.00	1	0	4	27	4	3	3	5	4	3
16	9	1,988	49	0.35	0.02	0.53	0.10	1	1	4	3	3	3	3	5	5	3
17	18	1,998	320	0.10	0.05	0.80	0.05	0	0	4	2	1	1	3	5	3	2
18	1	2,009	18	0.61	0.00	0.39	0.00	1	0	3	5	4	2	1	4	4	3
19	10	2,007	110	0.20	0.05	0.70	0.05	1	1	4	20	5	4	3	5	3	2
20	11	2,003	37	0.18	0.40	0.10	0.32	1	1	4	2	4	3	3	5	4	1
21	20	1,958	62	0.32	0.05	0.63	0.00	0	0	2	5	4	3	3	5	5	4
22	2	1,936	20	0.20	0.20	0.55	0.05	1	1	3	6	5	3	4	3	4	1
23	18	1,987	33	0.06	0.00	0.94	0.00	0	0	4	2	5	5	3	5	3	3
24	18	1,967	28	0.45	0.55	0.00	0.00	0	0	1	15	5	4	5	4	3	1
25	7	1,991	19	0.31	0.00	0.26	0.43	0	0	3	0	1	1	5	4	2	1
26	7	1,946	100	0.50	0.13	0.34	0.03	1	0	4	30	4	3	4	4	4	3
27	22	2,011	4	0.50	0.50	0.00	0.00	0	0	2	10	4	3	2	3	2	1

Εικόνα 11: Απόσπασμα Συνόλου Δεδομένων

Από την αρχή της πειραματικής διαδικασίας έγινε σαφές ότι πρόκειται για ένα σύνολο δεδομένων δύσκολό στη διαχείριση εξαιτίας των παρακάτω παραγόντων:

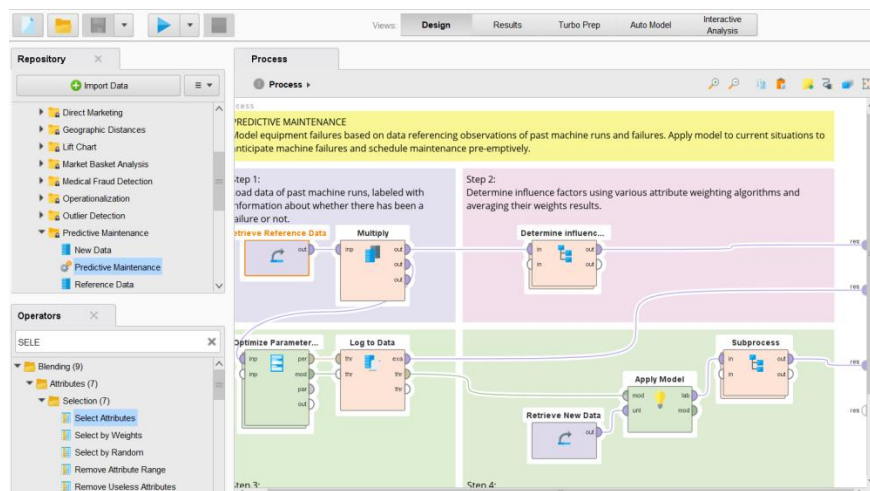
- Μικρό πλήθος γραμμών (εγγραφών)
- Μεγάλο πλήθος στηλών (χαρακτηριστικών)
- Μεγάλο πλήθος ασυμπλήρωτων (NULL) τιμών

Για τον λόγο αυτό αποφασίστηκε να χρησιμοποιηθεί ένα εργαλείο κατάλληλο για τη διαχείριση τέτοιου συνόλου δεδομένων, όπως το RapidMiner, το οποίο θα παρουσιαστεί την επόμενη ενότητα.

3.3. Εργαλείο RapidMiner

3.3.1. Δυνατότητες

Το RapidMiner είναι μια πλατφόρμα επιστήμης δεδομένων που παρέχει ένα περιβάλλον οπτικού προγραμματισμού για την ανάπτυξη και την εφαρμογή εφαρμογών προβλεπτικής ανάλυσης. Είναι μια δημοφιλής επιλογή για επιστήμονες δεδομένων όλων των επιπέδων δεξιοτήτων, αλλά είναι ιδιαίτερα ελκυστική για μη τεχνικούς χρήστες λόγω της φιλικής προς το χρήστη διεπαφής του και του μεγάλου φάσματος δυνατοτήτων του.



Εικόνα 12: Περιβάλλον διεπαφής RapidMiner

Το RapidMiner προσφέρει μια ποικιλία δυνατοτήτων που υποστηρίζουν ολόκληρη τη διαδικασία επιστήμης δεδομένων, από την προετοιμασία δεδομένων μέχρι τη δημιουργία μοντέλων και την επικύρωση. Αυτές οι δυνατότητες περιλαμβάνουν:

- **Προετοιμασία δεδομένων:** Προσφέρει μια ποικιλία operators για τον καθαρισμό δεδομένων, τη διαχείριση και τη μηχανική χαρακτηριστικών.
- **Μηχανική μάθηση:** Προσφέρει μια ποικιλία operators για εποπτευόμενη μάθηση, μη εποπτευόμενη μάθηση και μάθηση ενίσχυσης.
- **Εξόρυξη δεδομένων:** Προσφέρει μια ποικιλία operators για μάθηση κανόνων συσχέτισης, ομαδοποίηση, εξόρυξη κειμένου και ανίχνευση ανωμαλιών.
- **Εφαρμογή μοντέλων:** Προσφέρει μια ποικιλία δυνατοτήτων για την εφαρμογή μοντέλων προβλεπτικής ανάλυσης σε περιβάλλοντα παραγωγής. Το RapidMiner προσφέρει επίσης πολλές δυνατότητες που το καθιστούν ιδιαίτερα ελκυστικό για μη τεχνικούς χρήστες, όπως:
- **Οπτική διεπαφή προγραμματισμού:** Η οπτική διεπαφή προγραμματισμού του RapidMiner καθιστά εύκολη τη δημιουργία και την τροποποίηση ροών εργασίας επιστήμης δεδομένων χωρίς την ανάγκη συγγραφής κώδικα.

- **Έτοιμοι operators:** Το RapidMiner περιλαμβάνει μια βιβλιοθήκη από έτοιμους χειριστές που μπορούν να χρησιμοποιηθούν για την εκτέλεση κοινών εργασιών επιστήμης δεδομένων.
- **Λειτουργία drag-and-drop:** Η λειτουργία drag-and-drop του RapidMiner καθιστά εύκολη τη δημιουργία και την τροποποίηση ροών εργασίας επιστήμης δεδομένων.
- **Διαδραστική απεικόνιση:** Προσφέρει διαδραστικά εργαλεία απεικόνισης που επιτρέπουν στους χρήστες να εξερευνούν τα δεδομένα τους και να απεικονίζουν τα αποτελέσματα της ανάλυσής τους.

3.3.2. Οφέλη

Όσο αφορά τα οφέλη από τη χρήση του RapidMiner, συνοψίζονται στη συνέχεια:

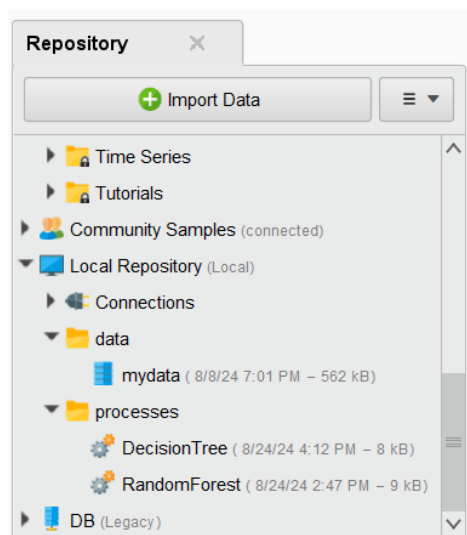
- **Αυξημένη παραγωγικότητα:** Η οπτική διεπαφή προγραμματισμού και οι έτοιμοι operators του RapidMiner μπορούν να βοηθήσουν τους χρήστες να είναι πιο παραγωγικοί, εξαλείφοντας την ανάγκη συγγραφής κώδικα και τη μετάβαση μεταξύ πολλαπλών εργαλείων.
- **Βελτιωμένη ακρίβεια:** Οι δυνατότητες του RapidMiner για προετοιμασία δεδομένων, επιλογή μοντέλων και επικύρωση μοντέλων μπορούν να βοηθήσουν τους χρήστες να βελτιώσουν την ακρίβεια των μοντέλων προβλεπτικής ανάλυσης τους.
- **Μειωμένα κόστη:** Η άδεια ανοιχτού κώδικα του RapidMiner και η δυνατότητά του να αυτοματοποιεί τις ροές εργασίας επιστήμης δεδομένων μπορούν να βοηθήσουν τους χρήστες να μειώσουν το κόστος ανάπτυξης και εφαρμογής εφαρμογών προβλεπτικής ανάλυσης.

3.3.3. Panels

Τέλος, παρατίθενται λεπτομέρειες όσο αφορά τα βασικά panel τα οποία χρησιμοποιήθηκαν κατά την πραγματοποίηση της πειραματικής διαδικασίας:

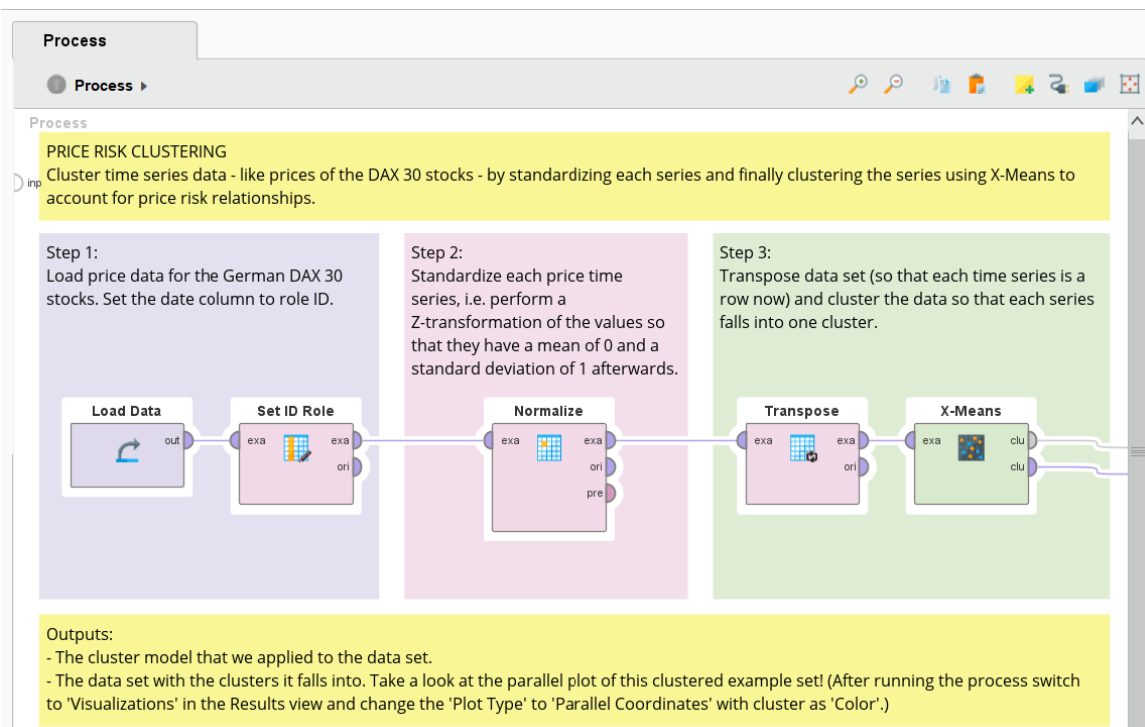
Repository Panel: Το Repository Panel είναι ουσιαστικά η κεντρική περιοχή αποθήκευσης για όλα τα αντικείμενα που δημιουργούνται ή εισάγονται. Περιλαμβάνει τα εξής:

- **Δεδομένα:** Στο αποθετήριο μπορούν να αποθηκευτούν διάφορα σύνολα δεδομένων. Αυτά μπορεί να είναι σε διάφορες μορφές, όπως Excel, CSV, βάσεις δεδομένων ή δεδομένα που δημιουργούνται μέσω των διαδικασιών του RapidMiner.
- **Διαδικασίες:** Εδώ αποθηκεύονται οι διαδικασίες (processes) που έχουν δημιουργηθεί. Μια διαδικασία είναι μια ακολουθία ενεργειών που εκτελούνται σε δεδομένα για να επιτευχθεί ένας συγκεκριμένος στόχος, όπως ο καθαρισμός, η μετατροπή, η ανάλυση ή η δημιουργία μοντέλου.
- **Μοντέλα:** Αφού δημιουργηθεί και εκπαιδευτεί ένα προβλεπτικό μοντέλο, μπορεί να αποθηκευτεί στο αποθετήριο. Τα αποθηκευμένα μοντέλα μπορούν να εφαρμοστούν σε νέα δεδομένα ή να βελτιωθούν περαιτέρω σε μεταγενέστερο στάδιο.
- **Αποτελέσματα:** Τα αποτελέσματα των διαδικασιών, όπως διαγράμματα, στατιστικά στοιχεία ή προβλέψεις, αποθηκεύονται εδώ. Αυτά τα αποτελέσματα μπορούν να προβληθούν οποιαδήποτε στιγμή και να χρησιμοποιηθούν για αναφορές ή παρουσιάσεις.
- **Επεκτάσεις:** Το RapidMiner μπορεί να επεκταθεί με επιπλέον λειτουργικότητα μέσω επεκτάσεων, οι οποίες αποθηκεύονται επίσης στο αποθετήριο μόλις εγκατασταθούν. Η διαχείριση των έργων γίνεται οργανώνοντας αυτά τα αντικείμενα σε φακέλους στον Repository Panel, διευκολύνοντας την πλοήγηση και τη διαχείριση μεγάλου αριθμού αρχείων.



Εικόνα 13: RepositoryPanel

Process Panel: Το Process Panel είναι ο χώρος όπου σχεδιάζονται και δημιουργούνται οι ροές εργασίας ανάλυσης δεδομένων στο RapidMiner Studio. Αυτό το panel αντιπροσωπεύει το χώρο εργασίας ή τον καμβά για τη δημιουργία μιας αναλυτικής διαδικασίας.

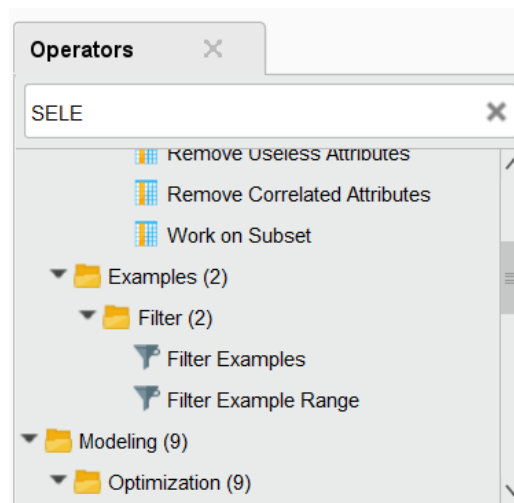


Εικόνα 14: ProcessPanel

Λειτουργεί ως εξής:

- **Σχεδιασμός Ροών Εργασίας:** Μια ροή εργασίας δημιουργείται σύροντας και αποθέτοντας χειριστές από τον Operators Panel στον Process Panel. Οι χειριστές (operators) είναι τα δομικά στοιχεία οποιασδήποτε διαδικασίας στο RapidMiner και μπορούν να εκτελέσουν μια ποικιλία λειτουργιών, από απλές μετατροπές δεδομένων έως πολύπλοκη προβλεπτική ανάλυση.
- **Σύνδεση Operators:** Οι χειριστές συνδέονται με “θύρες” που καθορίζουν τη ροή των δεδομένων από μία ενέργεια στην επόμενη. Αυτές οι συνδέσεις απεικονίζονται ως βέλη, υποδεικνύοντας οπτικά τη σειρά των ενεργειών.
- **Εκτέλεση Διαδικασιών:** Μόλις οι χειριστές συνδεθούν, μπορεί να εκτελεστεί ολόκληρη η διαδικασία ή να προχωρήσει βήμα-βήμα για αποσφαλμάτωση ή κατανόηση των ενδιάμεσων βημάτων.
- **Τροποποίηση Ροών Εργασίας:** Μια ροή εργασίας μπορεί να τροποποιηθεί εύκολα με την προσθήκη, αφαίρεση ή αναδιάταξη χειριστών για τη βελτιστοποίηση ή την προσαρμογή της διαδικασίας ανάλυσης.

Operators Panel: Το Operators Panel είναι μια ολοκληρωμένη βιβλιοθήκη όλων των διαθέσιμων χειριστών στο RapidMiner.

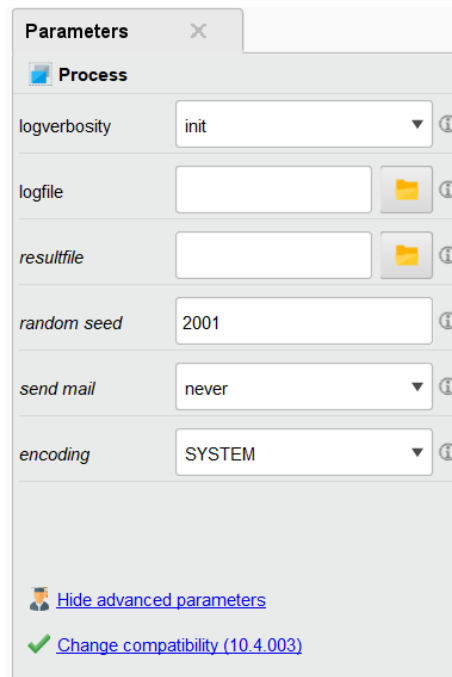


Εικόνα 15: OperatorsPanel

Είναι κατηγοριοποιημένος για να βοηθήσει στην εύρεση του κατάλληλου εργαλείου για την εργασία που απαιτείται:

- **Λειτουργία Αναζήτησης:** Η γραμμή αναζήτησης χρησιμοποιείται για να βρεθούν χειριστές με βάση το όνομα ή τη λειτουργία τους. Αυτό είναι χρήσιμο όταν γνωρίζεται ποια εργασία πρέπει να εκτελεστεί, αλλά όχι το συγκεκριμένο όνομα του χειριστή.
- **Κατηγοριοποίηση:** Οι χειριστές είναι οργανωμένοι σε ομάδες με βάση τη λειτουργία τους, όπως 'Blending', 'Cleaning', 'Modeling', 'Evaluation' και 'Visualization', μεταξύ άλλων. Πληροφορίες Χειριστή: Κάνοντας κλικ σε έναν χειριστή ή περνώντας το ποντίκι πάνω του, εμφανίζονται πληροφορίες σχετικά με το τι κάνει, τις εισόδους που απαιτεί, τα αποτελέσματα που παράγει και τις παραμέτρους διαμόρφωσης που δέχεται.

Parameters Panel: Όταν επιλέγεται ένας χειριστής στο Process Panel, το Parameters Panel εμφανίζει ρυθμίσεις που μπορούν να προσαρμοστούν για να εξατομικεύσουν τη συμπεριφορά του χειριστή, όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 16: ParametersPanel

3.4.Απαντήσεις ερώτησης προς πρόβλεψη

Η ερώτηση, για την πρόβλεψη των απαντήσεων της οποίας θα εφαρμοστούν τα δύο μοντέλα προβλεπτικής ανάλυσης είναι η:

Πόσο σημαντικές ήταν οι επιπτώσεις της οικονομικής κρίσης στις συνολικές hard και soft επενδύσεις πληροφορικής (σε hardware, software, ICT training και ICT consulting)

Στον ακόλουθο πίνακα, παρουσιάζεται η αντιστοιχία των απαντήσεων του ερωτηματολογίου με τις τιμές που είναι καταχωρημένες στο αρχείο excel που δόθηκε:

Τιμή απάντησης σε ερωτηματολόγιο	Τιμή απάντησης σε excel
Αύξηση	1
δεν είχε σημαντική επίπτωση	2
μικρή μείωση	3
μέτρια μείωση	4
σημαντική μείωση	5
πολύ σημαντική μείωση	6

Πίνακας 1:Κλάσεις 1^{ης} παραλλαγής

Επομένως, πρόκειται για ένα πρόβλημα πρόβλεψης 6 κλάσεων (multiclass problem), γεγονός που δυσχεραίνει ακόμη περισσότερο την διαδικασία πρόβλεψης δεδομένων και των δυσκολιών που αναφέρθηκαν προηγουμένως. Η πρόβλεψη των κλάσεων πραγματοποιήθηκε στην πρώτη παραλλαγή των αλγορίθμων που υλοποιήθηκαν.

Προκειμένου να επιτευχθεί υψηλότερο ποσοστό ακρίβειας προβλέψεων, υλοποιήθηκαν δύο επιπλέον παραλλαγές σε κάθε αλγόριθμο, όπου χρησιμοποιούνται 2 κλάσεις. Αρχικά, παρατίθεται ο πίνακας αντιστοίχισης των τιμών των απαντήσεων με τις 2 κλάσεις όσο αφορά τη δεύτερη παραλλαγή:

Τιμή Απάντησης	Κλάση
1	0
2	0
3	0
4	1
5	1
6	1

Πίνακας 2: Κλάσεις 2ης παραλλαγής

Και ο αντίστοιχος πίνακας με τις αντιστοιχήσεις όσο αφορά την δεύτερη παραλλαγή:

Τιμή Απάντησης	Κλάση
1	0
2	0
3	0
4	0
5	1
6	1

Πίνακας 3: Κλάσεις 3ης παραλλαγής

Επομένως, στις 2 τελευταίες παραλλαγές, το πρόβλημα πλέον είναι πρόβλεψης 2 κλάσεων (binary classification problem) και όχι πολλαπλών τάξεων (όπως στην πρώτη παραλλαγή). Αναμενόμενο είναι επομένως στις 2 τελευταίες παραλλαγές να καταγραφούν καλύτερα αποτελέσματα όσο αφορά τις μετρικές αξιολόγησης.

3.5.Μετρικές Αξιολόγησης

Στα πλαίσια της συγκεκριμένης ενότητας, παρατίθενται οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των 2 μοντέλων:

Accuracy

Η accuracy υπολογίζεται ως το ποσοστό των σωστών προβλέψεων σε σχέση με τον συνολικό αριθμό των προβλέψεων.

Weighted Precision

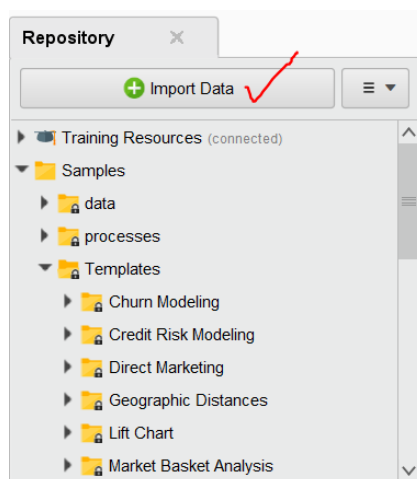
Η μετρική weighted precision λαμβάνει υπόψη τον αριθμό των παρατηρήσεων σε κάθε κατηγορία. Υπολογίζεται ως το άθροισμα της ακρίβειας για κάθε κατηγορία, σταθμισμένο με βάση τη στήριξη (support) της κατηγορίας, δηλαδή τον αριθμό των πραγματικών παρατηρήσεων για κάθε κατηγορία. Η ακρίβεια για μια συγκεκριμένη κατηγορία είναι το ποσοστό των σωστών προβλέψεων για αυτήν την κατηγορία σε σχέση με το συνολικό αριθμό των προβλέψεων που έγιναν για αυτήν την κατηγορία.

Weighted Recall

Η σταθμισμένη ανάκληση υπολογίζεται με τον ίδιο τρόπο όπως η σταθμισμένη ακρίβεια, αλλά λαμβάνει υπόψη την ικανότητα του μοντέλου να ανιχνεύει όλες τις πραγματικές θετικές περιπτώσεις για κάθε κατηγορία. Η ανάκληση για μια συγκεκριμένη κατηγορία είναι το ποσοστό των σωστών προβλέψεων για αυτήν την κατηγορία σε σχέση με το συνολικό αριθμό των πραγματικών περιπτώσεων αυτής της κατηγορίας.

3.6. Εισαγωγή Δεδομένων

Η εισαγωγή των δεδομένων, πραγματοποιήθηκε πολύ απλά επιλέγοντας Import Data στο RepositoryPanel:



Εικόνα 17: Επιλογή ImportData

Η φόρτωση των δεδομένων πραγματοποιήθηκε με επιτυχία όπως φαίνεται και από το παρακάτω screenshot:

Open in Turbo Prep Auto Model Interactive Analysis Filter (363 / 363 examples): all

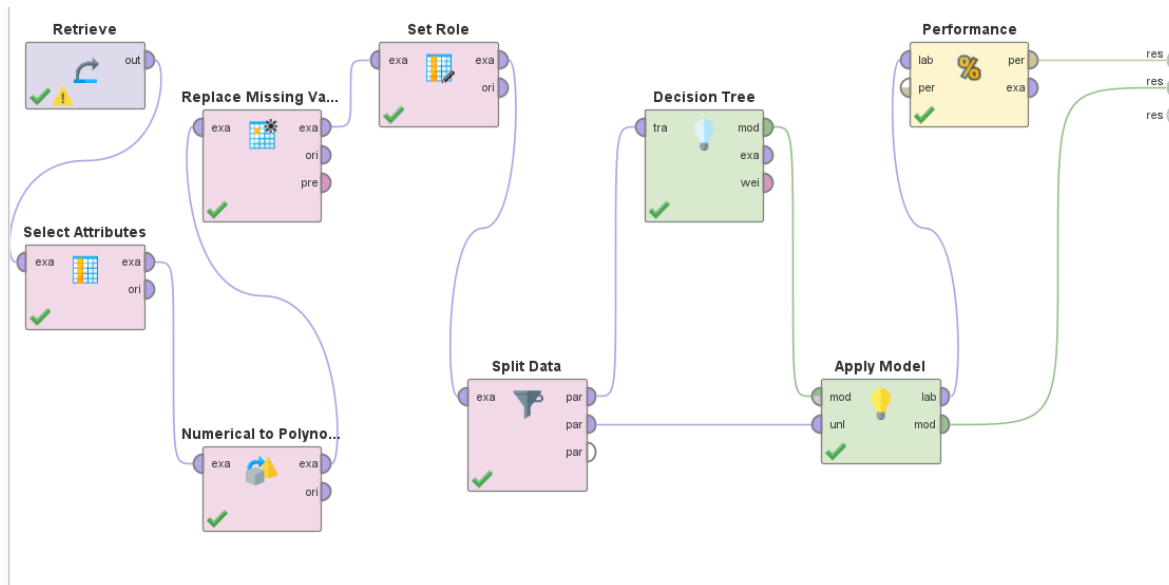
Row No.	ID	Sector	A13	A14	A15A	A15B	A15C	A15D	A16	A16A	A17	A18	A19A
1	1	18	1992	90	0.330	0.390	0.170	0.110	0	0	2	7	4
2	2	6	1984	114	0.750	0.150	0.100	0	0	0	2	3	4
3	3	20	1989	103	0.050	0.650	0.250	0.050	0	0	4	3	5
4	4	18	1981	218	0.300	0.100	0.600	0	0	0	4	0	4
5	5	14	1998	58	0.220	0.200	0.310	0.270	1	0.850	1	4	3
6	6	14	1983	272	0.750	0.200	0.050	0	1	0.090	2	4	3
7	7	23	1979	250	1	0	0	0	0	0	2	3	5
8	8	28	1955	133	0.320	0.570	0	0.110	0	0	5	50	1
9	9	19	1990	54	0.250	0.250	0.500	0	1	0.170	2	7	4
10	10	8	1981	62	0.060	0.060	0.880	0	1	0.010	4	3	5
11	11	12	1984	725	0.350	0.250	0.320	0.080	1	0.900	3	6	5
12	12	19	1917	50	0.170	0.080	0.500	0.250	0	0	3	2	2

Εικόνα 18: Επιτυχής Φόρτωση Δεδομένων

3.7. Εφαρμογή DecisionTree

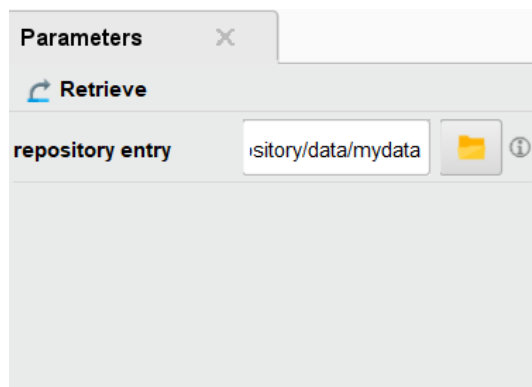
3.7.1. Πρώτη παραλλαγή

Ακολουθεί η σχηματική αναπαράσταση της διαδικασίας που αφορά την εφαρμογή του μοντέλου DecisionTree όσο αφορά την πρώτη παραλλαγή:



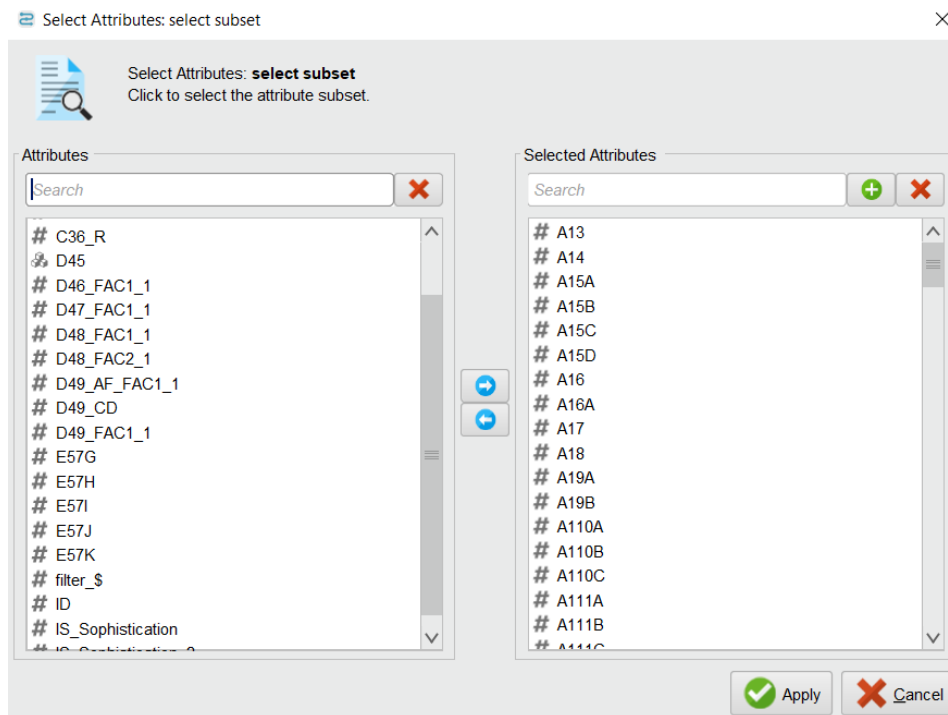
Εικόνα19: DecisionTreeProcess – 1^ηπαραλλαγή

Ο operator Retrieve χρησιμοποιείται για την ανάκτηση του συνόλου δεδομένων και οι ρυθμίσεις του παρουσιάζονται στη συνέχεια:



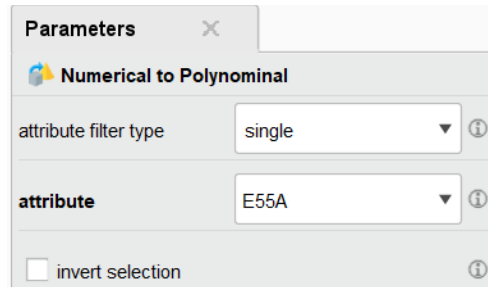
Εικόνα 19: Παράμετροι Retrieve

Ο operator SelectAttributes χρησιμοποιείται για την επιλογή στηλών του συνόλου δεδομένων(χρησιμοποιήθηκαν όλες οι αριθμητικές στήλες μέχρι την E57F). Παρατίθεται απόσπασμα της λίστας των στηλών που τελικά δεν χρησιμοποιήθηκαν:



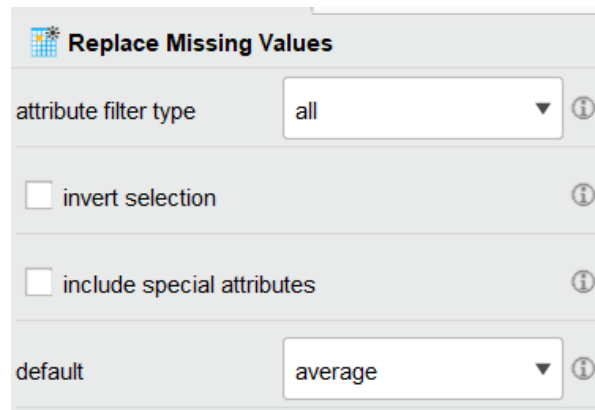
Εικόνα 20: Λίστα αποκλεισμένων Attributes

Ο operator `NumericaltoPolynomial` χρησιμοποιείται για την μετατροπή της στήλης πρόβλεψης (E55A) σε αλφαριθμητική έτσι ώστε στη συνέχεια να μπορεί να χρησιμοποιηθεί ως στήλη πρόβλεψης:



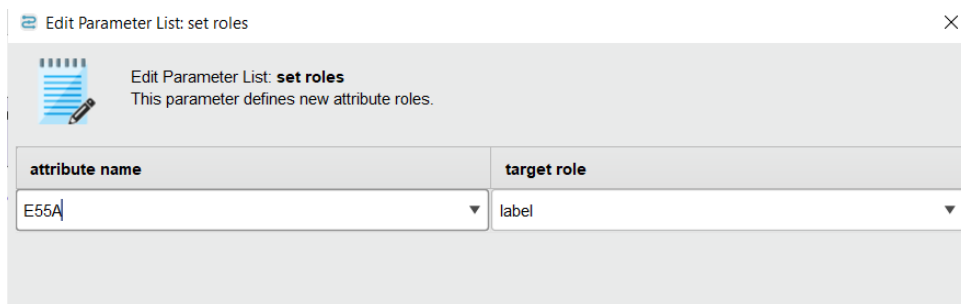
Εικόνα 21: `NumericaltoPolynomial`

Ο operator `ReplaceMissingValues` χρησιμοποιείται για την αντικατάσταση των null τιμών με τον μέσο όρο των τιμών της εκάστοτε στήλης:



Εικόνα23: Παράμετροι `ReplaceMissingValues`

Όσο αφορά τον operator `SetRole` χρησιμοποιήθηκε ώστε να τεθεί η στήλη E55A(που αντιστοιχεί στην απάντηση της ερώτησης που αναλύθηκε προηγουμένως) ως στήλη πρόβλεψης:



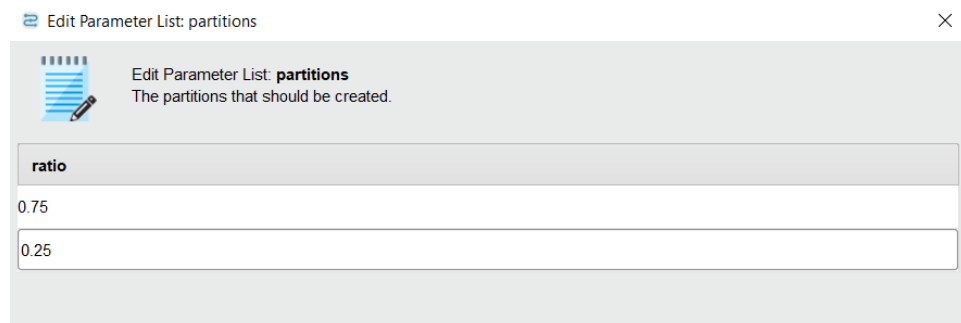
Edit Parameter List: set roles

Edit Parameter List: **set roles**
This parameter defines new attribute roles.

attribute name	target role
E55A	label

Εικόνα 22: Παράμετροι SetRole

Ο operator SplitData χρησιμοποιείται για τον διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης (trainset) και δεδομένα ελέγχου (test) με αναλογία όπως φαίνεται στην συνέχεια:



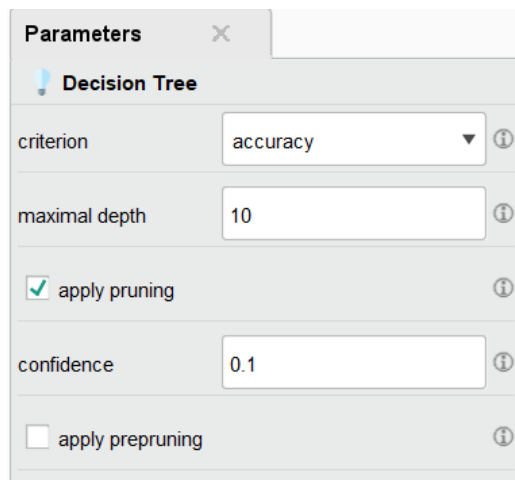
Edit Parameter List: partitions

Edit Parameter List: **partitions**
The partitions that should be created.

ratio
0.75
0.25

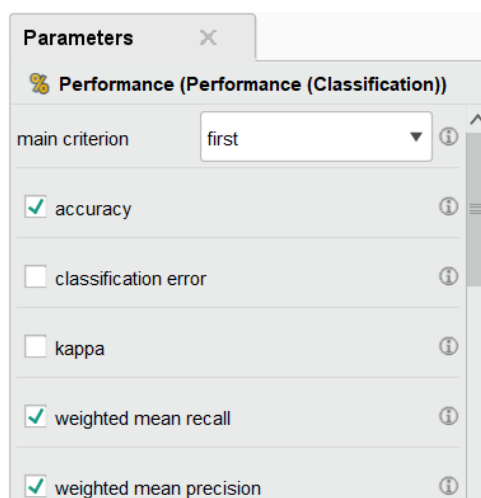
Εικόνα 23: Παράμετροι SplitData

Ακολουθούν οι παράμετροι του operator DecisionTree (ο πλέον βασικός operator):



Εικόνα 24: Παράμετροι DecisionTree

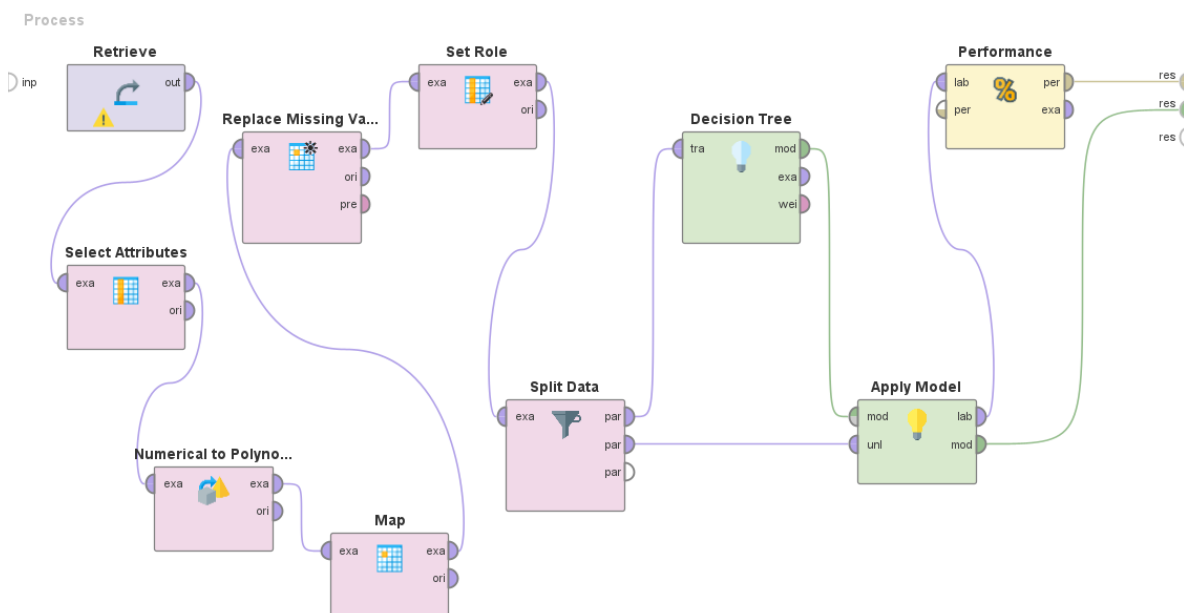
Όσο αφορά τον operator ApplyModel, χρησιμοποιείται για την εφαρμογή του μοντέλου στα δεδομένα ελέγχου. Τέλος, ο τελεστής Performance καθορίζει τις μετρικές αξιολόγησης, όπως φαίνεται παρακάτω:



Εικόνα 25: Παράμετροι Performance

3.7.2. Δεύτερη παραλλαγή

Ακολουθεί η σχηματική αναπαράσταση της διαδικασίας που αφορά την εφαρμογή του μοντέλου DecisionTree όσο αφορά την δεύτερη παραλλαγή:



Εικόνα 26: : DecisionTreeProcess – 2^η παραλλαγή

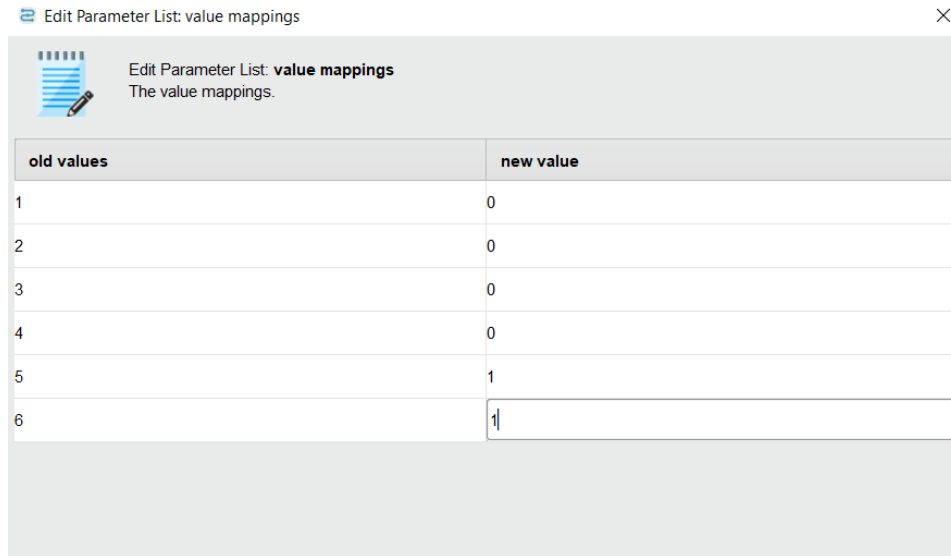
Στην 2^η παραλλαγή χρησιμοποιούνται όλοι οι operators της 1^{ης} παραλλαγής και επιπλέον ο operatorMap, ο οποίος επιτρέπει την αντιστοίχιση των τιμών της στήλης πρόβλεψης με τις δύο επιθυμητές τιμές όπως φαίνεται στη συνέχεια:

Edit Parameter List: value mappings The value mappings.	
old values	new value
1	0
2	0
3	0
4	1
5	1
6	1

Εικόνα 27: DecisionTree - αντιστοιχίσεις τιμών 2ης παραλλαγής

3.7.3. Τρίτη παραλλαγή

Όσο αφορά τη τρίτη παραλλαγή, η σχηματική διάταξη της διαδικασίας είναι ίδια με πριν, με την μόνη διαφορά ότι στον operator Map, χρησιμοποιούνται διαφορετικές αντιστοιχίες:



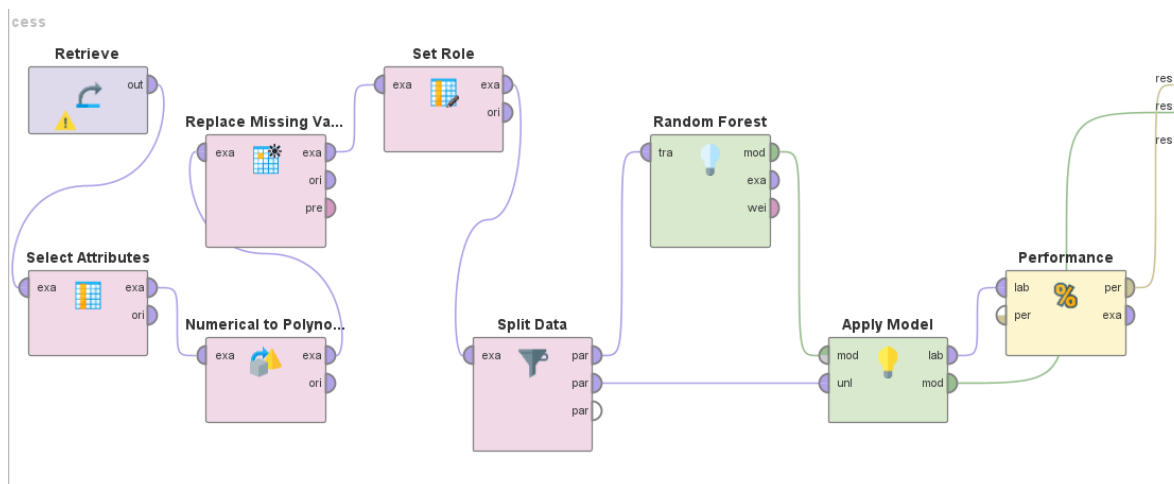
old values	new value
1	0
2	0
3	0
4	0
5	1
6	1

Εικόνα 28: DecisionTree - αντιστοιχίσεις τιμών 3ης παραλλαγής

3.8. Εφαρμογή Random Forest

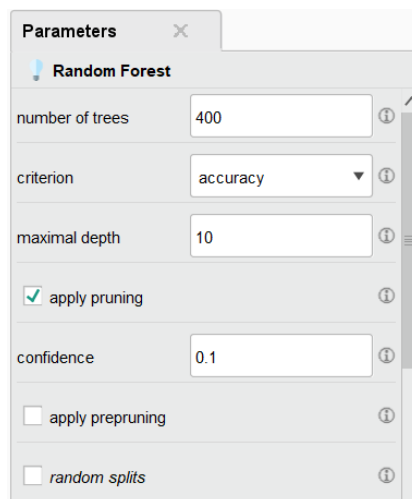
3.8.1. Πρώτη παραλλαγή

Ακολουθεί η σχηματική αναπαράσταση της διαδικασίας που αφορά την εφαρμογή του μοντέλου Random Forest όσο αφορά την πρώτη παραλλαγή:



Εικόνα29: Random Forest Process – 1^η παραλλαγή

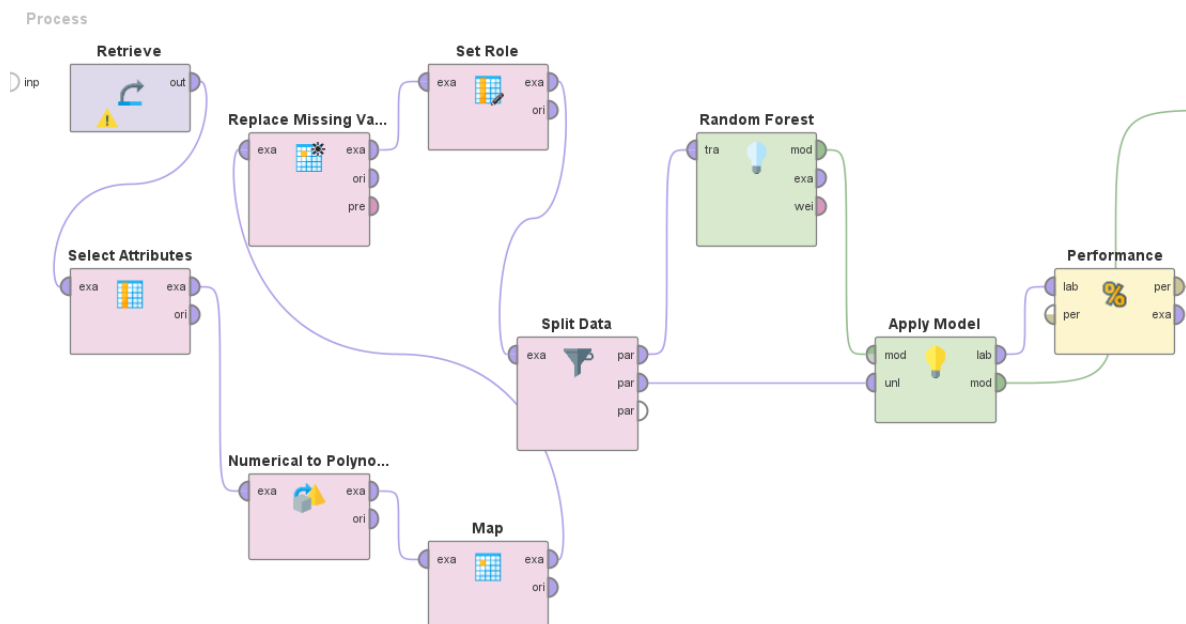
Στην περίπτωση του Random Forest, όλοι οι operators έχουν παραμετροποιηθεί με τον ίδιο ακριβώς τρόπο σε σχέση με προηγούμενως, εκτός βέβαια από τον operator Random Forest, τις παραμέτρους του οποίου παραθέτουμε στην συνέχεια:



Εικόνα 30: Παράμετροι Random Forest

3.8.2. Δεύτερη και τρίτη παραλλαγή

Όσο αφορά τη δεύτερη και τρίτη παραλλαγή, το σχηματικό διάγραμμα είναι το ακόλουθο:



Εικόνα 31: Random Forest Process – 2η και 3η παραλλαγή

Τέλος, αξίζει να αναφερθεί ότι η χρήση του operator Map χρησιμοποιήθηκε με την ίδια λογική όπως και στη περίπτωση του DecisionTree. Μετά την λεπτομερή παρουσίαση της πειραματικής μεθοδολογίας που ακολουθήθηκε, στην επόμενη ενότητα (επόμενη σελίδα) παρουσιάζονται τα αποτελέσματα της πειραματικής διαδικασίας.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΙΚΗΣ ΔΙΑΔΙΚΑΣΙΑΣ

4.1. Αποτελέσματα 1^{ης} παραλλαγής DecisionTree

Τα αποτελέσματα όσο αφορά την μετρική Accuracy:

accuracy: 74.44%

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	29	0	0	4	0	0	87.88%
pred. 3	2	7	2	2	0	0	53.85%
pred. 2	0	3	11	0	0	0	78.57%
pred. 4	2	5	0	6	0	0	46.15%
pred. 6	0	1	0	1	9	0	81.82%
pred. 1	0	1	0	0	0	5	83.33%
class recall	87.88%	41.18%	84.62%	46.15%	100.00%	100.00%	

Εικόνα 32: Αποτελέσματα Accuracy - DecisionTree – 1^η παραλλαγή

Επιπλέον, τα αποτελέσματα όσο αφορά την μετρική Recall:

weighted_mean_recall: 76.64%, weights: 1, 1, 1, 1, 1, 1

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	29	0	0	4	0	0	87.88%
pred. 3	2	7	2	2	0	0	53.85%
pred. 2	0	3	11	0	0	0	78.57%
pred. 4	2	5	0	6	0	0	46.15%
pred. 6	0	1	0	1	9	0	81.82%
pred. 1	0	1	0	0	0	5	83.33%
class recall	87.88%	41.18%	84.62%	46.15%	100.00%	100.00%	

Εικόνα35: ΑποτελέσματαRecall - DecisionTree – 1^η παραλλαγή

Ακολουθούν τα αποτελέσματα όσο αφορά την μετρική Precision:

weighted_mean_precision: 71.93%, weights: 1, 1, 1, 1, 1, 1

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	29	0	0	4	0	0	87.88%
pred. 3	2	7	2	2	0	0	53.85%
pred. 2	0	3	11	0	0	0	78.57%
pred. 4	2	5	0	6	0	0	46.15%
pred. 6	0	1	0	1	9	0	81.82%
pred. 1	0	1	0	0	0	5	83.33%
class recall	87.88%	41.18%	84.62%	46.15%	100.00%	100.00%	

Εικόνα 33: Αποτελέσματα Precision - DecisionTree 1^η παραλλαγή

Τέλος, το μητρώο σύγχυσης (ConfusionMatrix):

```
ConfusionMatrix:
True:   5      3      2      4      6      1
5:     29      0      0      4      0      0
3:      2      7      2      2      0      0
2:      0      3     11      0      0      0
4:      2      5      0      6      0      0
6:      0      1      0      1      9      0
1:      0      1      0      0      0      5
```

Εικόνα 34: Confusion Matrix - Decision Tree– 1^η παραλλαγή

4.2. Αποτελέσματα 2^{ης} παραλλαγής DecisionTree

Τα αποτελέσματα όσο αφορά την μετρική Accuracy:

accuracy: 94.44%

	true 1	true 0	class precision
pred. 1	54	4	93.10%
pred. 0	1	31	96.88%
class recall	98.18%	88.57%	

Εικόνα 35: Αποτελέσματα Accuracy - DecisionTree – 2^η παραλλαγή

Τα αποτελέσματα όσο αφορά την μετρική Recall:

weighted_mean_recall: 93.38%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	54	4	93.10%
pred. 0	1	31	96.88%
class recall	98.18%	88.57%	

Εικόνα 36: Αποτελέσματα Recall - Decision Tree – 2^η παραλλαγή

Τα αποτελέσματα όσο αφορά την μετρική Precision:

weighted_mean_precision: 94.99%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	54	4	93.10%
pred. 0	1	31	96.88%
class recall	98.18%	88.57%	

Εικόνα 37: Αποτελέσματα Precision - DecisionTree – 2η παραλλαγή

Και το αντίστοιχο confusionmatrix:

```
ConfusionMatrix:
True:   1      0
1:     54     4
0:      1    31
```

Εικόνα 38: Confusion Matrix - Decision Tree - 2^η παραλλαγή

4.3. Αποτελέσματα 3^{ης} παραλλαγής DecisionTree

Όσο αφορά τα αποτελέσματα της 3ης παραλλαγής δίνονται στη συνέχεια:

accuracy: 91.11%

	true 1	true 0	class precision
pred. 1	30	4	88.24%
pred. 0	4	52	92.86%
class recall	88.24%	92.86%	

Εικόνα 39: Αποτελέσματα Accuracy - DecisionTree - 3η παραλλαγή

weighted_mean_recall: 90.55%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	30	4	88.24%
pred. 0	4	52	92.86%
class recall	88.24%	92.86%	

Εικόνα 40: Αποτελέσματα Recall - Decision Tree - 3η παραλλαγή

weighted_mean_precision: 90.55%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	30	4	88.24%
pred. 0	4	52	92.86%
class recall	88.24%	92.86%	

Εικόνα 41: Αποτελέσματα Precision - DecisionTree - 3η παραλλαγή

Και το confusion matrix:

```
ConfusionMatrix:
True:    1    0
1:      30    4
0:       4   52
```

Εικόνα 42: Confusion Matrix - Decision Tree - 3η παραλλαγή

4.4. Αποτελέσματα 1^{ης} παραλλαγής Random Forest

Τα αποτελέσματα όσο αφορά την μετρική Accuracy:

accuracy: 67.78%

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	33	6	1	9	3	0	63.46%
pred. 3	0	9	2	3	1	0	60.00%
pred. 2	0	2	10	0	0	1	76.92%
pred. 4	0	0	0	0	0	0	0.00%
pred. 6	0	0	0	1	5	0	83.33%
pred. 1	0	0	0	0	0	4	100.00%
class recall	100.00%	52.94%	76.92%	0.00%	55.56%	80.00%	

Εικόνα 43: Αποτελέσματα Accuracy-Random Forest – 1^η παραλλαγή

Επιπλέον, τα αποτελέσματα όσο αφορά την μετρική Recall:

weighted_mean_recall: 60.90%, weights: 1, 1, 1, 1, 1, 1

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	33	6	1	9	3	0	63.46%
pred. 3	0	9	2	3	1	0	60.00%
pred. 2	0	2	10	0	0	1	76.92%
pred. 4	0	0	0	0	0	0	0.00%
pred. 6	0	0	0	1	5	0	83.33%
pred. 1	0	0	0	0	0	4	100.00%
class recall	100.00%	52.94%	76.92%	0.00%	55.56%	80.00%	

Εικόνα47: Αποτελέσματα Recall - Random Forest- 1^η παραλλαγή

Ακολουθούν τα αποτελέσματα όσο αφορά την μετρική Precision:

weighted_mean_precision: 63.95%, weights: 1, 1, 1, 1, 1, 1

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	33	6	1	9	3	0	63.46%
pred. 3	0	9	2	3	1	0	60.00%
pred. 2	0	2	10	0	0	1	76.92%
pred. 4	0	0	0	0	0	0	0.00%
pred. 6	0	0	0	1	5	0	83.33%
pred. 1	0	0	0	0	0	4	100.00%
class recall	100.00%	52.94%	76.92%	0.00%	55.56%	80.00%	

Εικόνα 44: Αποτελέσματα Precision - DecisionTree 1^η παραλλαγή

Τέλος, το μητρώο σύγχυσης (confusionmatrix):

```
ConfusionMatrix:
True:   5      3      2      4      6      1
5:      33      6      1      9      3      0
3:       0      9      2      3      1      0
2:       0      2     10      0      0      1
4:       0      0      0      0      0      0
6:       0      0      0      1      5      0
1:       0      0      0      0      0      4
```

Εικόνα 45: Confusion Matrix - Random Forest- 1^η παραλλαγή

4.5. Αποτελέσματα 2^{ης} παραλλαγής Random Forest

Τα αποτελέσματα όσο αφορά την μετρική Accuracy:

accuracy: 92.22%

	true 1	true 0	class precision
pred. 1	53	5	91.38%
pred. 0	2	30	93.75%
class recall	96.36%	85.71%	

Εικόνα 46: Αποτελέσματα Accuracy - Random Forest- 2η παραλλαγή

Τα αποτελέσματα όσο αφορά την μετρική Recall:

weighted_mean_recall: 91.04%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	53	5	91.38%
pred. 0	2	30	93.75%
class recall	96.36%	85.71%	

Εικόνα 47: Αποτελέσματα Recall - Random Forest- 2η παραλλαγή

Τα αποτελέσματα όσο αφορά την μετρική Precision:

weighted_mean_precision: 92.56%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	53	5	91.38%
pred. 0	2	30	93.75%
class recall	96.36%	85.71%	

Εικόνα 48: Αποτελέσματα Precision -Random Forest- 2η παραλλαγή

Και το αντίστοιχο confusionmatrix:

```
ConfusionMatrix:
True:   1   0
1:     53   5
0:      2  30
```

Εικόνα53: ConfusionMatrix - Random Forest- 2η παραλλαγή

4.6. Αποτελέσματα 3^{ης} παραλλαγής Random Forest

Όσο αφορά τα αποτελέσματα της 3ης παραλλαγής δίνονται στη συνέχεια:

accuracy: 92.22%

	true 1	true 0	class precision
pred. 1	30	3	90.91%
pred. 0	4	53	92.98%
class recall	88.24%	94.64%	

Εικόνα 49: Αποτελέσματα Accuracy - Random Forest- 3η παραλλαγή

weighted_mean_recall: 91.44%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	30	3	90.91%
pred. 0	4	53	92.98%
class recall	88.24%	94.64%	

Εικόνα 50: Αποτελέσματα Recall - Random Forest- 3η παραλλαγή

weighted_mean_precision: 91.95%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	30	3	90.91%
pred. 0	4	53	92.98%
class recall	88.24%	94.64%	

Εικόνα 51: Αποτελέσματα Precision - Random Forest- 3η παραλλαγή

Και το confusionmatrix:

```
ConfusionMatrix:
True:   1   0
1:     30   3
0:      4  53
```

Εικόνα57: ConfusionMatrix - Random Forest- 3η παραλλαγή

4.7. Σύγκριση Μέσων Αποτελεσμάτων

Προκειμένου να καταστεί εφικτή η σύγκριση των αποτελεσμάτων των αλγορίθμων στην μέση περίπτωση, αρχικά παρατίθεται συγκριτικός πίνακας που περιέχει την μέση τιμή των accuracy, precision και recall από τις 3 παραλλαγές και για τα 2 μοντέλα:

	Μέση τιμή Accuracy	Μέση τιμή Recall	Μέση τιμή Precision
Decision Tree	86.67%	86.86%	85.82%
Random Forest	84.07%	81.12%	82.82%

Πίνακας 4: Συγκριτικές μέσες τιμές μετρικών

Παρατηρείται από τον παραπάνω πίνακα ότι στην περίπτωση του Decision και για τις 3 μετρικές καταγράφηκαν καλύτερα αποτελέσματα (με όχι ωστόσο και τόσο μεγάλη διαφορά).

Αρχικά, αξίζει να αναφερθεί ότι υπήρχαν κλάσεις στην περίπτωση της παραλλαγής 1, στις οποίες σημειώθηκαν εξαιρετικά αποτελέσματα (φθάνοντας το 100%), ενώ άλλες στις οποίες δεν καταγράφηκε καλή συμπεριφορά, μειώνοντας τα συνολικά αποτελέσματα των μετρικών. Συμπερασματικά δεδομένων των δυσκολιών του προβλήματος, τα 2 μοντέλα στην παραλλαγή 1 συμπεριφέρθηκαν αποτελεσματικά.

Επιπλέον, όπως ήταν αναμενόμενο και στους 2 αλγορίθμους, μεταξύ της παραλλαγής 1 και των παραλλαγών 2 και 3, υπάρχει τεράστια διαφορά όσο αφορά και τις 3 μετρικές. Το γεγονός αυτό ήταν αναμενόμενο εξ αρχής, καθώς η μείωση του αριθμού των κλάσεων από 6 σε 2, σημαίνει μεγαλύτερη ευκολία όσο αφορά τη πραγματοποίηση ταξινόμησης των δεδομένων και άρα καλύτερα αποτελέσματα στις μετρικές. Πρόκειται για μια βελτίωση η οποία κυμαίνεται μεταξύ του 20 έως 25%.

Επιπροσθέτως, μεταξύ των παραλλαγών 2 και 3 (που χρησιμοποιούν 2 κλάσεις προς ταξινόμησης) δεν παρατηρείται μεγάλη διαφοροποίηση ως προς τα αποτελέσματα. Ελάχιστα καλύτερη συμπεριφορά παρουσιάζει η παραλλαγή 2, στην οποία υπάρχει ισομερής διαχωρισμός (οι τιμές 1,2,3 αντιστοιχήθηκαν στην κλάση 0, και οι τιμές 4,5,6 αντιστοιχήθηκαν στην κλάση 1).

Συνολικά, η ομαδοποίηση των τιμών ενίσχυσε τα αποτελέσματα και στις 3 μετρικές, γεγονός αρκετά ενθαρρυντικό για την μετέπειτα εξέλιξη της πειραματικής διαδικασίας.

4.8. Αποτελέσματα εκτέλεσης με διαφορετική αναλογία training – testset

Στα πλαίσια της συγκεκριμένης ενότητας παρουσιάζονται τα αποτελέσματα εκτέλεσης και για τις 3 παραλλαγές και για τα δύο μοντέλα με διαφορετική αναλογία (60% trainingset– 40% testset) ως προς τη μετρική accuracy. Ο λόγος που πραγματοποιήθηκε η συγκεκριμένη πειραματική διαδικασία είναι για μπορούν να υπάρξουν κυρίως συγκριτικά αποτελέσματα με την ενότητα που θα ακολουθήσει (automodel).

Αρχικά, τα αποτελέσματα του DecisionTree όσο αφορά τις 3 διαφορετικές παραλλαγές που ακολουθήθηκαν και πριν:

accuracy: 72.92%

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	47	1	0	7	5	0	78.33%
pred. 3	1	13	2	3	0	0	68.42%
pred. 2	0	5	18	1	0	0	75.00%
pred. 4	5	7	0	8	0	0	40.00%
pred. 6	0	0	0	1	10	0	90.91%
pred. 1	0	1	0	0	0	9	90.00%
class recall	88.68%	48.15%	90.00%	40.00%	66.67%	100.00%	

Εικόνα 52: Αποτελέσματα DecisionTree - Παραλλαγή 1 - (60% - 40%)

accuracy: 93.79%

	true 1	true 0	class precision
pred. 1	86	6	93.48%
pred. 0	3	50	94.34%
class recall	96.63%	89.29%	

Εικόνα 53: Αποτελέσματα DecisionTree - Παραλλαγή 2- (60% - 40%)

accuracy: 91.03%

	true 1	true 0	class precision
pred. 1	49	7	87.50%
pred. 0	6	83	93.26%
class recall	89.09%	92.22%	

Εικόνα 54: Αποτελέσματα DecisionTree - Παραλλαγή 3 - (60% - 40%)

Αντίστοιχα, και τα αποτελέσματα του Random Forest:

accuracy: 62.50%

	true 5	true 3	true 2	true 4	true 6	true 1	class precision
pred. 5	52	7	0	15	6	0	65.00%
pred. 3	1	15	7	4	0	0	55.56%
pred. 2	0	4	12	0	1	6	52.17%
pred. 4	0	1	0	0	0	0	0.00%
pred. 6	0	0	0	1	8	0	88.89%
pred. 1	0	0	1	0	0	3	75.00%
class recall	98.11%	55.56%	60.00%	0.00%	53.33%	33.33%	

Εικόνα 55: Αποτελέσματα Random Forest - Παραλλαγή 1 - (60% - 40%)

accuracy: 93.79%

	true 1	true 0	class precision
pred. 1	86	6	93.48%
pred. 0	3	50	94.34%
class recall	96.63%	89.29%	

Εικόνα 56: : Αποτελέσματα Random Forest - Παραλλαγή 2 - (60% - 40%)

accuracy: 91.03%

	true 1	true 0	class precision
pred. 1	47	5	90.38%
pred. 0	8	85	91.40%
class recall	85.45%	94.44%	

Εικόνα 57: : Αποτελέσματα Random Forest - Παραλλαγή 3 - (60% - 40%)

Ακολουθεί και ο αντίστοιχος συγκριτικός πίνακας:

Μοντέλο	Μέση τιμή accuracy
Decision Tree	85.91%
Random Forest	82.44%

Πίνακας 5: Συγκριτικά αποτελέσματα (60% - 40%)

Δύο συνοπτικά συμπεράσματα που προκύπτουν είναι τα παρακάτω:

- Στην περίπτωση του DecisionTree σημειώθηκε ελαφρώς καλύτερη ακρίβεια
- Σε σχέση με τον προηγούμενο διαχωρισμό (75% - 25%) τα αποτελέσματα είναι ελαφρώς χειρότερα

4.9. Εφαρμογή automodel

Το εργαλείο Rapidminer παρέχει την δυνατότητα δημιουργίας και εκτέλεσης αυτοματοποιημένων μοντέλων, που παρέχουν τη δυνατότητα εκτέλεσης πολλαπλών αλγορίθμων πάνω σε σύνολα δεδομένων, έχοντας σαν στόχο να βελτιστοποιηθούν τα αντίστοιχα αποτελέσματα. Η συγκεκριμένη πειραματική διαδικασία ακολουθήθηκε 3 φορές (1 για κάθε παραλλαγή).

Στο σημείο αυτό αξίζει να αναφερθεί ότι:

- Δημιουργήθηκαν 3 διαφορετικά σύνολα δεδομένων σαν είσοδος (1 για κάθε παραλλαγή). Η διαδικασία αυτή πραγματοποιήθηκε με την συγγραφή μια απλής έτσι ώστε στις παραλλαγές 2 και 3 να γίνουν οι σωστές αντιστοιχήσεις των τιμών. Αυτό συνέβη διότι στην περίπτωση του automodel δεν υπάρχει δυνατότητα πραγματοποίησης αντιστοίχισης μέσα από το εργαλείο στην περίπτωση του automodel.
- Η αναλογία διαχωρισμού είναι 60% - 40%. Για τον λόγο αυτό προηγήθηκε και η πειραματική διαδικασία της προηγούμενης ενότητας, ώστε να υπάρξουν συγκριτικά αποτελέσματα ως προς το accuracy.

Επιπλέον, τα μοντέλα τα οποία περιελάμβανε το automodel είναι τα:

- Naïve Bayes
- Generalized Linear Regression
- Logistic Regression
- Last Large Margin
- Deep Learning
- Decision Tree
- Random Forest
- Gradient Boosting
- SVM

Όσο αφορά την εφαρμογή του automodel είναι μια διαδικασία κάποιων βημάτων, που ακολουθούνται μέσα από το εργαλείο και παρατίθενται στη συνέχεια. Αρχικά, πραγματοποιείται φόρτωση του συνόλου των δεδομένων:

Select Data for a New Model

Information

Name: mydata
Number of rows: 363
Number of columns: 171

Attributes / Columns

ID, Sector, A13, A14, A15A, A15B, A15C, A15D, A16, A16A, A17, A18, A19A, A19B, A110A, A110B, A110C, A111A, A111B, A111C, A111D, A112A, A112B, A112C, A112D, B21A, B21B, B22A, B22B, B22C, B23, B23A, B23B, B23C, B24, B25, B25A, B25B, B26A, B26B, B26C, B26D, B26E, B26F, B26G, B26H, B27A, B27B, B27C, B27D, C31A, C31B, C31C, C32A1, C32A2, C32A3, C32A4, C32A5, C32B1, C32B2, C32B3, C32B4, C32B5, C33, C33A, C34, C34A, C35, C35A, C35B, C35C, C35D, C35E, C35F, C36, C37A, C37B, C37C, C37D, C37E, C37F, D41, D42, D43A1, D43A2, D43A3, D43B1, D43B2, D43B3, D44, D45, D46A, D46B, D46C, D46D, D47A, D47B, D47C, D47D, D47E...

IMPORT NEW DATA

Εικόνα 58: AutoModel - Φόρτωση δεδομένων

Ακολουθεί ο καθορισμός του στόχου (predict) και της στήλης πρόβλεψης (E55A):

Predict
Want to predict the values of a column?

Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

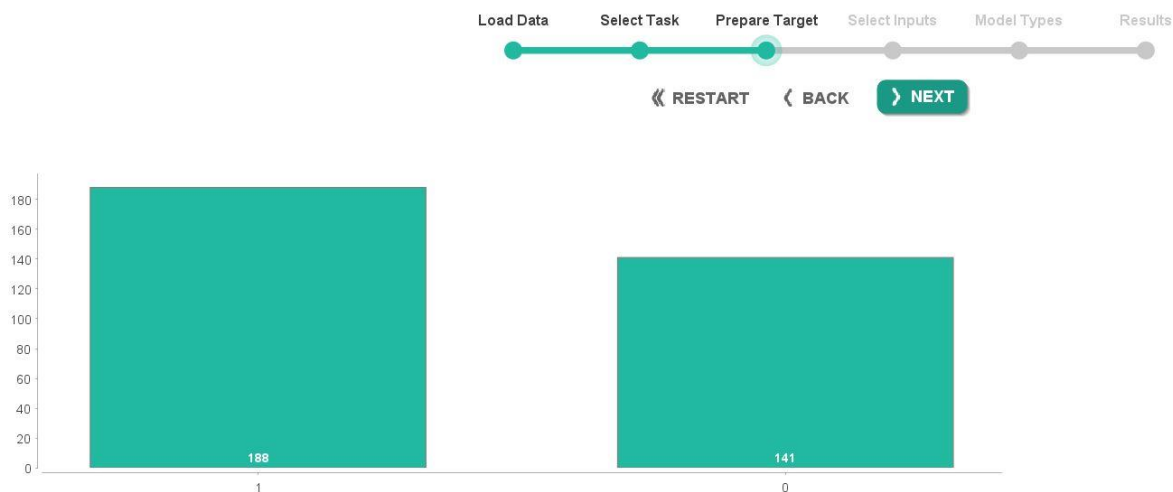
E54D Number	E54E Number	E54F Number	E54G Number	E55A Category	E55B Number	E55C Number	E55D Number	E55E Number	E55F Number	E55G Number
4	3	3	3	5	5	5	4	4	2	3
4	1	3	4	3	2	3	5	3	3	4
1	1	1	1	5	5	5	5	5	5	5
3	2	2	2	2	2	2	5	3	3	3
3	2	2	1	2	2	2	3	3	2	4
4	3	2	2	2	2	3	5	5	4	5
4	2	2	2	4	4	4	4	4	4	5
?	?	?	?	?	?	?	?	?	?	?
1	1	1	1	5	5	5	5	5	5	5

Εικόνα 59: Καθορισμός στόχου και στήλης πρόβλεψης

Στη συνέχεια επιβεβαιώνεται η κατανομή των εγγραφών για κάθε κλάση. Ακολουθούν οι κατανομές και για τις 3 παραλλαγές:



Εικόνα 60: Κατανομή εγγραφών - Παραλλαγή 1



Εικόνα 61: Κατανομή εγγραφών - Παραλλαγή 2



Εικόνα 62: Κατανομή εγγραφών - Παραλλαγή 3

Ακολουθεί η επιλογή των στηλών (όπως και στο προηγούμενο μέρος, μέχρι την E57F):

Selected	Status	Quality	Name ↑	Correlation	ID-ness	Stability
☒	●	CI IT SM T	A110A	1.49%	1.38%	38.48%
☒	●	CI IT SM T	A110B	0.38%	1.38%	57.10%
☒	●	CI IT SM T	A110C	0.76%	1.38%	35.29%
☒	●	CI IT SM T	A111A	1.65%	1.38%	32.58%

Εικόνα 63: Επιλογή στηλών

Και τέλος επιλέγονται τα μοντέλα που θα εκτελεστούν:

Models

- ☒ Naive Bayes
- ☒ Generalized Linear Model
 - ☒ Use Regularization ☐ Calculate p-Values
- ☒ Logistic Regression
- ☒ Fast Large Margin
 - ☒ Automatically Optimize
- ☒ Deep Learning
- ☒ Decision Tree
 - ☒ Automatically Optimize Maximal Depth: 20
- ☒ Random Forest
 - ☒ Automatically Optimize Number of Trees: 20 Maximal Depth: 20

Data Preparation

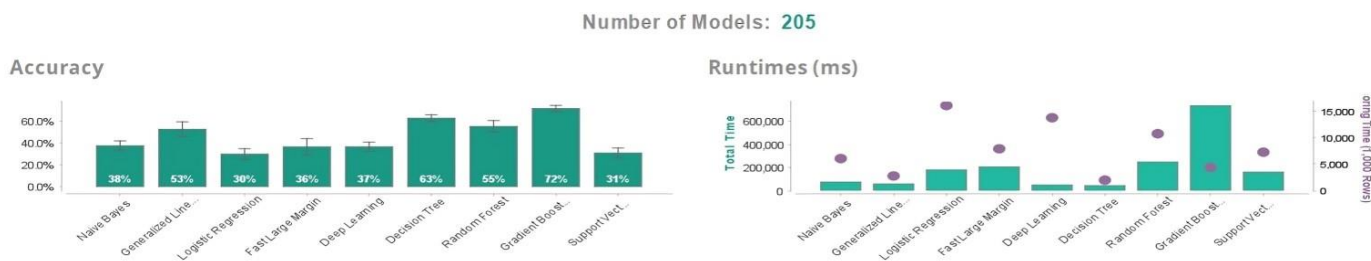
- ☐ Remove Columns with Too Many Values Maximum Number of Values: 50
- ☐ Extract Date Information
- ☐ Extract Text Information
 - Select Text Columns (0)...
 - Number of Extracted Features: 1,000
- ☐ Automatic Feature Selection
 - Additional Minutes (Maximum): 60
 - Final Feature Set should be Accurate
- ☐ Automatic Feature Generation
 - Function Complexity can be Medium

Εικόνα 64: Επιλογή μοντέλων

4.10. Αποτελέσματα automodel

Αρχικά παρατίθενται τα συγκριτικά αποτελέσματα όσο αφορά την 1^η παραλλαγή:

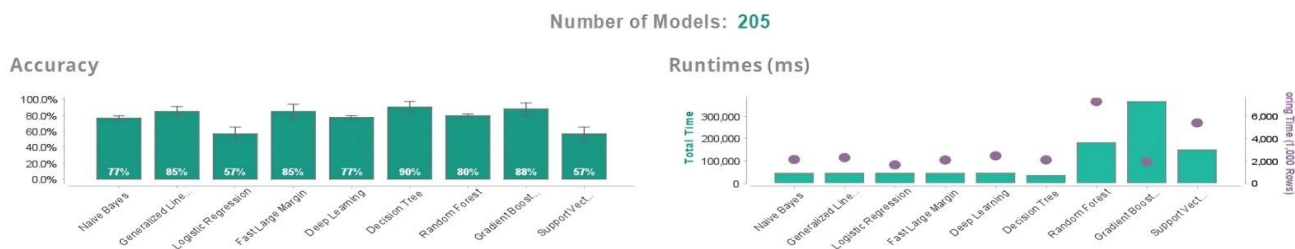
Overview



Εικόνα 65: Αποτελέσματα automodel - Παραλλαγή 1

Ακολουθούν τα αποτελέσματα όσο αφορά τη 2^η παραλλαγή:

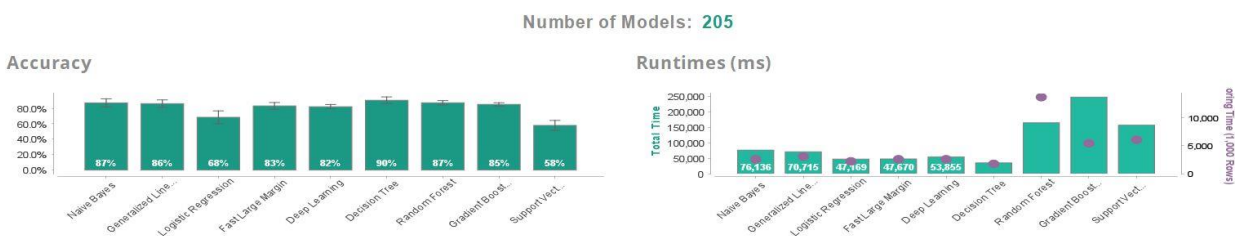
Overview



Εικόνα 66: Αποτελέσματα automodel - Παραλλαγή 2

Και τέλος τα αποτελέσματα όσο αφορά την 3^η παραλλαγή:

Overview



Εικόνα 67: Αποτελέσματα automodel - Παραλλαγή 3

4.11. Σχολιασμός αποτελεσμάτων

Στον παρακάτω πίνακα, παρουσιάζεται το μοντέλο το οποίο πέτυχε τη καλύτερη τιμή ως προς το accuracy σε κάθε παραλλαγή:

Παραλλαγή	Μοντέλο	Τιμή accuracy
1 ^η παραλλαγή	Gradient Boosting	72%
2 ^η παραλλαγή	Decision Tree	90%
3 ^η παραλλαγή	Decision Tree	90%

Πίνακας 6: AutoModel - Μοντέλο με τα καλύτερα αποτελέσματα

Όπως και στο προηγούμενο μέρος της πειραματικής διαδικασίας σημειώθηκαν καλύτερα αποτελέσματα στις παραλλαγές 2 και 3. Τέλος, στους επόμενους πίνακες παρατίθενται τα συγκριτικά αποτελέσματα όσο αφορά τα μοντέλα Decision Tree και Random Forest μεταξύ της τρέχουσας (auto model) και της πειραματικής διαδικασίας της προηγούμενης ενότητας:

Μοντέλο	Μέση τιμή accuracy
Decision Tree (χωρίς auto model)	85.91%
Decision Tree (με auto model)	80.67%

Πίνακας 7: Συγκριτικά αποτελέσματα Decision Tree

Μοντέλο	Μέση τιμή accuracy
Random Forest(χωρίς auto model)	82.44%
Random Forest(με auto model)	74%

Πίνακας 8: Συγκριτικά αποτελέσματα Random Forest

Παρατηρείται από τους παραπάνω πίνακες ότι στην περίπτωση που δεν χρησιμοποιείται το auto model, σημειώθηκαν καλύτερα αποτελέσματα.

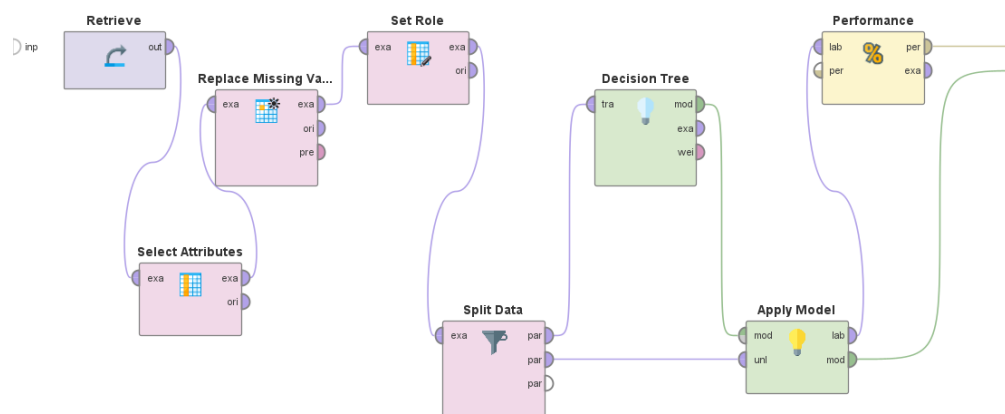
4.12. Νέος στόχος πρόβλεψης

Πραγματοποιώντας σε δεύτερη φάση μια πιο αναλυτική και προσεκτική εκτίμηση της πειραματικής διαδικασίας που ακολουθήθηκε στις προηγούμενες ενότητες, διαπιστώθηκε ότι είναι πολύ πιθανόν να υπάρχουν στήλες οι οποίες έχουν εξάρτηση με την στήλη πρόβλεψης (E55A) και για τον λόγο αυτό μπορούν να ενισχύουν την ποιότητα των αποτελεσμάτων με τρόπο όμως ο οποίος δεν χαρακτηρίζεται από δικαιοσύνη.

Για τον λόγο αυτό δημιουργήθηκε μια νέα μικρότερη έκδοση του συνόλου δεδομένων (newdata.xls), στην οποία ο στόχο πρόβλεψης θα είναι μια τροποποιημένη εκδοχή της στήλης E51 (sales crisis vulnerability), στην οποία η τιμή 1 σημαίνει σοβαρή μείωση των πωλήσεων και η τιμή 0 σημαίνει μη σοβαρή μείωση των πωλήσεων. Επιπλέον αξίζει να σημειωθεί ότι δεν χρησιμοποιήθηκε καμία στήλη της ομάδας E, ενώ από την ομάδα D, διατηρήθηκε η στήλη D41. Οι πειραματικές διαδικασίες που ακολουθήθηκαν είναι:

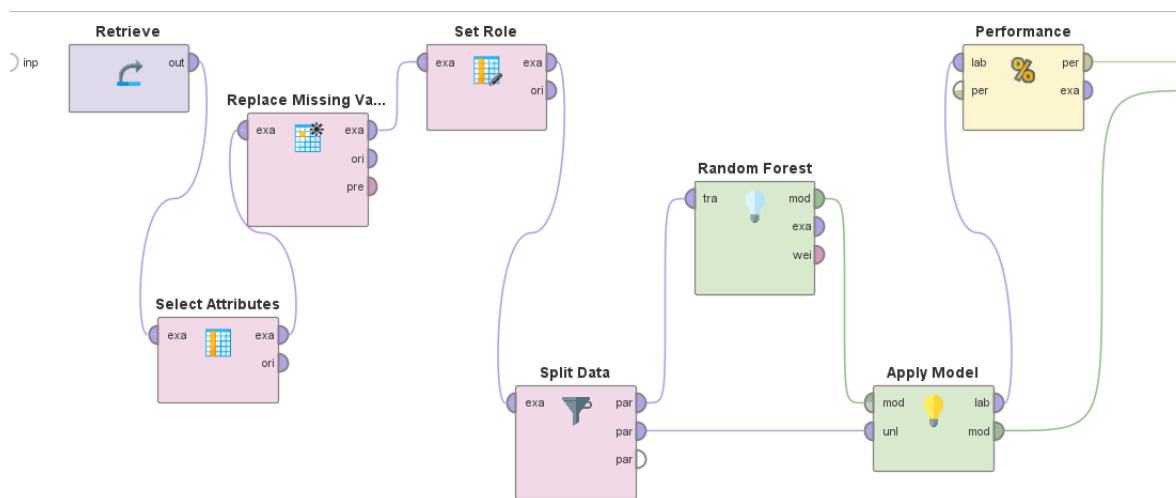
- Decision Tree
- Random Forest
- Auto Model

Ακολουθεί το σχηματικό διάγραμμα για της διαδικασίας για τη περίπτωση του DecisionTree:



Εικόνα 68: DecisionTree - νέος στόχος πρόβλεψης

Έπεται το σχηματικό διάγραμμα για τη περίπτωση του Random Forest:



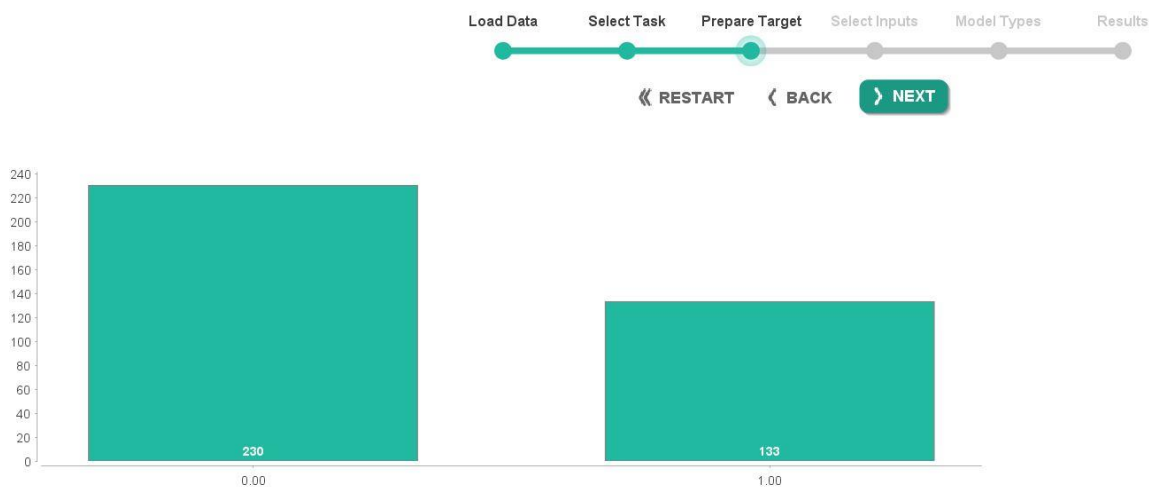
Εικόνα 69: Random Forest - νέος στόχος πρόβλεψης

Επιπλέον, όσο αφορά το AutoModel, ακολουθεί το βήμα επιλογής της στήλης πρόβλεψης:

Predict Want to predict the values of a column? Clusters Want to identify groups in your data? Outliers Want to detect outliers in your data?								
57K umber	A14_R Number	IS_Sophisticat... Number	IS_Sophisticat... Number	C36_R Number	SECT_MAN_... Number	D_LARGE Number	D_MED Number	E51R Category
1	2	4.400	4.500	0.022	1	0	1	1.00
1	2	3.400	3.500	0.035	1	0	1	0.00
1	2	3	3.250	0.019	0	0	1	0.00
1	2	2.400	2.750	0.014	1	0	1	0.00
1	2	3.800	4	0.017	1	0	1	1.00
2	3	2.800	2.500	0.022	1	1	0	0.00
1	3	4.800	4.750	0.800	0	1	0	0.00
2	2	5	5	0.015	0	0	1	0.00
1	2	1.600	1.750	0.019	0	0	1	0.00
3	2	1	1	0	1	0	1	0.00

Εικόνα 70: AutoModel - επιλογή στήλης πρόβλεψης - νέος στόχος πρόβλεψης

Και η αντίστοιχη κατανομή των εγγραφών μεταξύ των 2 κλάσεων της στήλης πρόβλεψης:



Εικόνα 71: Κατανομή εγγραφών κλάσεων - νέος στόχος πρόβλεψης

Όσο αφορά τα αποτελέσματα της πειραματικής διαδικασίας, αρχικά παρατίθενται τα αποτελέσματα του DecisionTree:

accuracy: 72.41%

	true 1.00	true 0.00	class precision
pred. 1.00	29	16	64.44%
pred. 0.00	24	76	76.00%
class recall	54.72%	82.61%	

Εικόνα 72: Αποτελέσματα DecisionTree - νέος στόχος πρόβλεψης

Ακολουθούν τα αποτελέσματα του Random Forest:

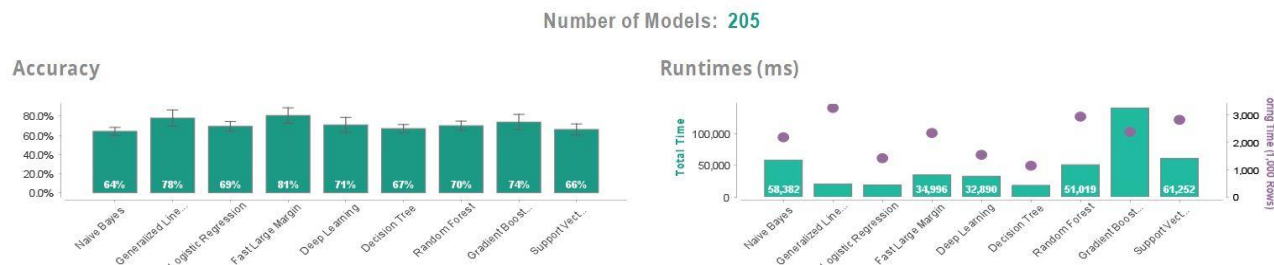
accuracy: 75.86%

	true 1.00	true 0.00	class precision
pred. 1.00	29	11	72.50%
pred. 0.00	24	81	77.14%
class recall	54.72%	88.04%	

Εικόνα 73: Αποτελέσματα Random Forest - νέος στόχος πρόβλεψης

Τέλος, παρατίθενται τα αποτελέσματα που προέκυψαν από την εφαρμογή του AutoModel:

Overview



Εικόνα 74: Αποτελέσματα AutoModel - νέος στόχος πρόβλεψης

Στην περίπτωση του AutoModel, το καλύτερο αποτέλεσμα ακρίβειας επετεύχθη στη περίπτωση του FastLargeMarginμοντέλου και ήταν της τάξης του 81%, ένα αποτέλεσμα άκρως ικανοποιητικό αν αναλογιστεί κανείς ότι πλέον οι στήλες είναι πολύ λιγότερες.

Τέλος, τα συγκριτικά αποτελέσματα όσο αφορά τα μοντέλα DecisionTreeκαι Random Forest:

Μοντέλο	Τιμή accuracy
Decision Tree (χωρίςauto model)	72.41%
Decision Tree (μεauto model)	67%

Πίνακας 9: Συγκριτικά αποτελέσματα DecisionTree–νέος στόχος πρόβλεψης

Μοντέλο	Τιμή accuracy
Random Forest(χωρίςauto model)	75.86%
Random Forest(μεauto model)	70%

Πίνακας 10: Συγκριτικά αποτελέσματα Random Forest–νέος στόχος πρόβλεψης

Και στην προκειμένη περίπτωση, καταγράφηκαν ικανοποιητικές τιμές, με τα μοντέλα τα οποία εφαρμόστηκαν χωρίς τη χρήση automodel να είναι καλύτερα.

4.13. Πρόβλεψη σημαντικότερων μεταβλητών με χρήση Automodel

Χρησιμοποιώντας το εργαλείο RapidMiner και συγκεκριμένα τη λειτουργία AutoModel, έχουμε τη δυνατότητα να αναλύσω τις ανεξάρτητες μεταβλητές και να προσδιορίσω το πόσο επηρεάζουν την εξαρτημένη μεταβλητή του μοντέλου μου. Μέσα από τη διαδικασία της αυτόματης μοντελοποίησης, το RapidMiner αναλύει τις μεταβλητές χρησιμοποιώντας ένα πλήθος αλγορίθμων και μου παρέχει τα weights ή αλλιώς τους συντελεστές βαρύτητας για κάθε ανεξάρτητη μεταβλητή. Αυτοί οι συντελεστές βαρύτητας αντιπροσωπεύουν τη σημαντικότητα κάθε μεταβλητής σε σχέση με το αποτέλεσμα του μοντέλου, επιτρέποντάς μου να εντοπίσω ποιες μεταβλητές επηρεάζουν περισσότερο την εξαρτημένη μεταβλητή. Με βάση αυτά τα αποτελέσματα, μπορώ να επιλέξω να εστιάσω περαιτέρω στις μεταβλητές με τους υψηλότερους συντελεστές βαρύτητας για να βελτιστοποιήσω την ανάλυση και την ακρίβεια των προβλέψεων του μοντέλου μου.

Όπως γνωρίζουμε το automodel παρέχει μία ευρεία γκάμα πολλαπλών αλγορίθμων πάνω σε σύνολα δεδομένων, έχοντας σαν στόχο να βελτιστοποιηθούν τα αντίστοιχα αποτελέσματα.

Με την βοήθεια λοιπόν αυτών των αλγορίθμων θα γίνει ο υπολογισμός για το ποιες από τις ανεξάρτητες μεταβλητές επηρεάζουν περισσότερο την λειτουργία των επιχειρήσεων.

Οι συγκεκριμένες μεταβλητές παρουσιάζουν διαφοροποιήσεις, καθώς η πειραματική διαδικασία εκτελείται κάθε φορά με διαφορετικό αλγόριθμο, όπως οι Decision Tree, Deep Learning, Fast Large Margin, Gradient Boosting, Generalized Linear Regression, Logistic Regression, SVM, Random Forest και Naïve Bayes. Στις επόμενες εικόνες παρουσιάζεται μέρος των αποτελεσμάτων κάθε αλγορίθμου, εστιάζοντας στις δέκα σημαντικότερες ανεξάρτητες μεταβλητές, μαζί με τον συντελεστή βαρύτητας που αντικατοπτρίζει τη σημασία τους στο μοντέλο.

Με βάση τους αλγόριθμους Decision Tree, Deep Learning, Random Forest και Naïve Bayes η σημαντικότερη ανεξάρτητη μεταβλητή είναι η A12, δηλαδή το ερώτημα:

1.12 Πόσο καλές ήσαν οι οικονομικές επιδόσεις της επιχείρησής σας κατά την τελευταία 3ετία (2012-2014) σε σχέση με τους ανταγωνιστές σας, όσον αφορά τα ακόλουθα μεγέθη – απαντήσατε στην κλίμακα 1-5, όπου: 5 = πολύ υψηλότερες από τον μέσο όρο των ανταγωνιστών, 4 = υψηλότερες από τον μέσο όρο, 3 = περίπου στον μέσο όρο, 2 = χαμηλότερες από τον μέσο όρο και 1 = πολύ χαμηλότερες από τον μέσο όρο;

A112A	Κερδοφορία	1 2 3 4 5
A112B	Εξέλιξη πωλήσεων	1 2 3 4 5
A112C	Εξέλιξη μεριδίου αγοράς	1 2 3 4 5
A112D	Απόδοση Επενδεδυμένων Κεφαλαίων (Return on Investment-ROI) = καθαρά κέρδη ως ποσοστό του συνολικού ενεργητικού της επιχείρησης	1 2 3 4 5

X deep_learning		
	A	B
1	Attribute	Weight
2	A112B	0,1810
3	A112D	0,1774
4	A17	0,1699
5	C32B1	0,1382
6	B26B	0,1123
7	B26A	0,1047
8	A15C	0,1026
9	B22A	0,0945
10	B25B	0,0848

X random_forest		
	A	B
1	Attribute	Weight
2	A112B	0,2037
3	A17	0,1898
4	A112A	0,1403
5	B25B	0,1112
6	A15A	0,1086
7	A112C	0,0953
8	A112D	0,0871
9	B27D	0,0850
10	C37B	0,0750

X decision_tree		
	A	B
1	Attribute	Weight
2	A112B	0,1839
3	A17	0,1596
4	B27B	0,0489
5	B27D	0,0454
6	C37B	0,0444
7	A110C	0,0418
8	C35B	0,0368
9	B26G	0,0361
10	C32A2	0,0355

X naive_bayes		
	A	B
1	Attribute	Weight
2	A112A	0,1476
3	A17	0,1231
4	A112C	0,1167
5	SECT_MAN_SEI	0,1125
6	C35B	0,0978
7	A112D	0,0941
8	A15D	0,0921
9	A18	0,0888
10	C37D	0,0880

Εικόνα 75: Αποτελέσματα AutoModel για Decision Tree, Deep Learning, Random Forest και Naive Bayes

Με βάση τους αλγόριθμους Decision Fast Large Margin, Gradient Boosting, Generalized Linear Regression η σημαντικότερη ανεξάρτητη μεταβλητή είναι η A17, δηλαδή το ερώτημα:

1.7 Η ζήτηση των προϊόντων/υπηρεσιών που παράγει η επιχείρησή σας έχει κατά την τελευταία τριετία 2012 – 2013 – 2014:

(1) μειωθεί έντονα (2) μειωθεί (3) παραμένει σταθερή (4) αυξηθεί (5) αυξηθεί έντονα

☐ ☐ ☐ ☐ ☐

X fast_large_m			X gradient_boc			X generalized_		
	A	B		A	B		A	B
1	Attribute	Weight	1	Attribute	Weight	1	Attribute	Weight
2	A17	0,3645	2	A17	0,2319	2	A17	0,4907
3	A112B	0,2079	3	A112B	0,2035	3	A112A	0,2648
4	A112A	0,1919	4	A112A	0,1751	4	A112D	0,1939
5	A112D	0,1866	5	C36_R	0,1206	5	A112B	0,1741
6	C37D	0,1492	6	A112D	0,1140	6	B23C	0,1167
7	A15C	0,0965	7	C37B	0,0833	7	B23	0,1059
8	C31A	0,0948	8	B27D	0,0794	8	B27D	0,1040
9	C32A4	0,0928	9	A112C	0,0783	9	C37F	0,0931
10	C35C	0,0908	10	C37D	0,0674	10	C37D	0,0804

Εικόνα 76: Αποτελέσματα για Decision Fast Large Margin, Gradient Boosting, Generalized Linear Regression

Με βάση τον αλγόριθμο Logistic Regression η σημαντικότερη ανεξάρτητη μεταβλητή είναι η C35A, δηλαδή το ερώτημα:

3.5 Χρησιμοποιεί η επιχείρησή σας social media (π.χ. Facebook, LinkedIn, Twitter, Blogs, Youtube) για την υποστήριξη/ενίσχυση κάποιων επιχειρησιακών λειτουργιών/δραστηριοτήτων της;

ΝΑΙ ☐

ΌΧΙ ☐

Εάν ΝΑΙ: Σε ποιά βαθμό χρησιμοποιείτε social media για τους παρακάτω σκοπούς (βάλτε X στην αντίστοιχη στήλη);

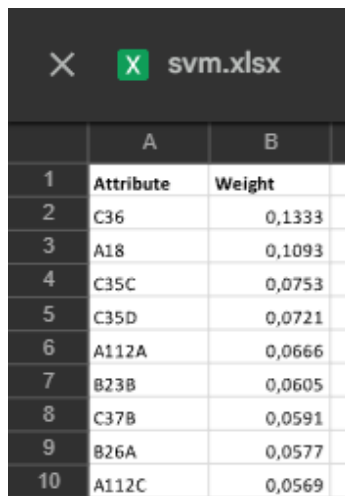
		δεν χρησιμοποιούμε	χρησιμοποιούμε σε μικρό βαθμό	χρησιμοποιούμε σε μεγάλο βαθμό
C35A	για την ανεύρεση προσωπικού (π.χ. μέσω LinkedIn)			
C35B	για την αναζήτηση ιδεών για βελτιώσεις/καινοτομίες προϊόντων/υπηρεσιών			
C35C	για την προώθηση των πωλήσεων			
C35D	για την συλλογή απόψεων, γνώμων, παρατηρήσεων, παρα- πόνων, κλπ. των πελατών για τα προϊόντα/υπηρεσίες σας			
C35E	για την υποστήριξη της εσωτερικής ανταλλαγής πληροφο- ριών/απόψεων μεταξύ των εργαζομένων της επιχείρησης			
C35F	για την υποστήριξη της εξωτερικής ανταλλαγής πληροφο- ριών/απόψεων με άλλες επιχειρήσεις (π.χ. συνεργάτες)			

X logistic-regre		
	A	B
1	Attribute	Weight
2	C35A	0,2015
3	C37D	0,1253
4	C35D	0,1175
5	C32B1	0,1132
6	A17	0,0994
7	C35E	0,0975
8	C13C	0,0956
9	A112C	0,0876
10	C36	0,0875

Εικόνα 77: Αποτελέσματα για Logistic Regression

Με βάση τον αλγόριθμο SVM η σημαντικότερη ανεξάρτητη μεταβλητή είναι η C36, δηλαδή το ερώτημα:

3.6 Αριθμός εξειδικευμένου προσωπικού (υπαλλήλων) πληροφορικής _____



	A	B
1	Attribute	Weight
2	C36	0,1333
3	A18	0,1093
4	C35C	0,0753
5	C35D	0,0721
6	A112A	0,0666
7	B23B	0,0605
8	C37B	0,0591
9	B26A	0,0577
10	A112C	0,0569

Εικόνα 78: Αποτελέσματα για SVM

Συμπεράσματα

Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας, αρχικά πραγματοποιήθηκε η κατάλληλη βιβλιογραφικά ανασκόπηση όσο αφορά τη τεχνητή νοημοσύνη, μηχανική μάθηση και ανάλυση μεγάλων δεδομένων. Από την ανασκόπηση αυτή έγινε σαφής η σημασία και των τριών πεδίων και πόσο μάλλον της συνδυαστικής τους χρήσης.

Όσο αφορά το πειραματικό μέρος, δεν επρόκειτο για ένα απλό πρόβλημα classification, αλλά υπήρχαν αρκετές δυσκολίες εξαιτίας του μικρού πλήθους των εγγραφών, του πολύ μεγάλου πλήθους χαρακτηριστικών, αλλά και των πολλών ελλιπών τιμών. Ωστόσο με την συμβολή του εργαλείου RapidMiner, η πειραματική προσπάθεια στέφτηκε με επιτυχία δεδομένων των ενθαρρυντικών αποτελεσμάτων τα οποία ελήφθησαν.

Σαν μελλοντικές κατευθύνσεις προτείνονται τα παρακάτω:

- Δοκιμή περισσότερων μοντέλων μηχανικής μάθησης
- Προσπάθεια για πρόβλεψη και άλλων στηλών του συνόλου δεδομένων
- Συγκριτικά αποτελέσματα με υλοποίηση σε γλώσσα Python
- Προσπάθεια για εφαρμογή των μοντέλων σε μεγαλύτερο σύνολο δεδομένων (και με λιγότερες ελλιπείς τιμές)

Το βέβαιο είναι ότι εργαλεία, όπως το RapidMiner, θα συνεχίσουν να αποτελούν το ενδιαφέρον της επιστημονικής κοινότητας, καθώς χωρίς να απαιτείται ούτε μια γραμμή κώδικα μπορούν να καταγραφούν θεαματικά αποτελέσματα, συμβάλλοντας με θετικό πρόσημο στην εξέλιξη της επιστημονικής έρευνας.

ΒΙΒΛΙΟΓΡΑΦΙΑ – ΠΗΓΕΣ

- [1]. Dobrev, D. (2012). A definition of artificial intelligence. arXiv preprint arXiv:1210.1568.
- [2]. Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, 58(1), 504-509.
- [3]. Chowdhary, K. R. (2020). Fundamentals of artificial intelligence (pp. 603-649). New Delhi: Springer India.
- [4]. Bhargava, C., & Sharma, P. K. (Eds.). (2021). Artificial intelligence: fundamentals and applications. CRC Press.
- [5]. He, Y. (2019). The importance of artificial intelligence to economic growth. *Korean Journal of Artificial Intelligence*, 7(1), 17-22.
- [6]. Shohsanam, K. (2023). Theoretical Importance of Artificial Intelligence. *Science and innovation*, 2(Special Issue 3), 159-162.
- [7]. Nikolenko, K. (2022). Artificial Intelligence and Society: Pros and Cons of the Present, Future Prospects. *Futurity Philosophy*, 1(2), 54-67.
- [8]. Dombrovsky, V. (2018). Pros and Cons of Artificial Intelligence (Doctoral dissertation, Тернопіль: THEU).
- [9]. Flowers, J. C. (2019, March). Strong and Weak AI: Deweyan Considerations. In *AAAI spring symposium: Towards conscious AI systems* (Vol. 2287, No. 7).
- [10]. Liu, B. (2021). " Weak AI" is Likely to Never Become" Strong AI", So What is its Greatest Value for us?. arXiv preprint arXiv:2103.15294.

- [11]. Martinez, R. (2018). Artificial intelligence: Distinguishing between types & definitions. Nev. LJ, 19, 1015.
- [12]. Dorr, L. (2022). Types of Artificial Intelligence, Explained.
- [13]. Allen, G. (2020). Understanding AI technology. Joint Artificial Intelligence Center (JAIC) The Pentagon United States, 2(1), 24-32.
- [14]. Park, P. S., Goldstein, S., O’Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5).
- [15]. Anyanwu, K. (2011). Overview and Applications of Artificial Intelligence. Bachelor of Engineering (B. Eng), Electrical and Electronic Engineering, University of Technology Owerri.
- [16]. Garg, P. K. (2021). Overview of artificial intelligence. In Artificial intelligence (pp. 3-18). Chapman and Hall/CRC.
- [17]. Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. International Journal of Information Management, 62, 102433.
- [18]. Stahl, B. C., & Stahl, B. C. (2021). Ethical issues of AI. Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies, 35-53.
- [19]. Rebala, G., Ravi, A., Churiwala, S., Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine learning definition and basics. An introduction to machine learning, 1-17.

- [20]. El Naqa, I., & Murphy, M. J. (2015). What is machine learning? (pp. 3-11). Springer International Publishing.
- [21]. Alpaydin, E. (2021). Machine learning. MIT press.
- [22]. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
- [23]. Ayodele, T. O. (2010). Types of machine learning algorithms. New advances in machine learning, 3(19-48), 5-1.
- [24]. Muhamedyev, R. (2015). Machine learning methods: An overview. Computer modelling & new technologies, 19(6), 14-29.
- [25]. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.
- [26]. Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15(1), 713-714.
- [27]. Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37(2.5), 3.
- [28]. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

- [29]. Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(1), 189-194.
- [30]. Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, 47(1), 31-39.
- [31]. Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3), 275-294.
- [32]. Sagioglu, S., & Sinanc, D. (2013, May). Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.
- [33]. Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, 36(6), 1231-1247.
- [34]. Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), 431-448.
- [35]. Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. *Issues in Information Systems*, 16(2).
- [36]. Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.
- [37]. Arunadevi, J., Ramya, S., & Raja, M. R. (2018). A study of classification algorithms using Rapidminer. *International Journal of Pure and Applied Mathematics*, 119(12), 15977-15988.