Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Εφαρμογές Μηχανικής Μάθησης στη δυσλεξία

Διπλωματικη Εργασια

του

ΠΑΝΑΓΙΩΤΗ ΖΩΓΑ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σιόλας
Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Ιούνιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εφαρμογές Μηχανικής Μάθησης στη δυσλεξία

Διπλωματικη Εργασια

του

ΠΑΝΑΓΙΩΤΗ ΖΩΓΑ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σιόλας
Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24η Ιουνίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Α.Γ. Σταφυλοπάτης	Γεώργιος Στάμου	Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2021

(Υπογραφή)

.....

Παναγιωτης Ζωγας
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η δυσλεξία είναι μία μαθησιακή δυσκολία νευροβιολογικής προέλευσης που γίνεται αντιληπτή λόγω δυσκολιών στην εκμάθηση της ανάγνωσης. Σύμφωνα με εκτιμήσεις που βασίζονται σε παραμέτρους και κριτήρια διάγνωσης το 5-10% του πληθυσμού έχει δυσλεξία. Ο εντοπισμός της σε μικρές ηλικίες είναι επιτακτική ανάγκη καθώς έτσι γίνεται η αποτελεσματικότερη αντιμετώπιση της. Η έλλειψη διάγνωσης μπορεί να αποδειχθεί επιζήμια τόσο για την απόδοση του παιδιού στην ανάγνωση όσο και για την ψυχική του υγεία. Το αντικείμενο της παρούσας διπλωματικής είναι με τη χρήση μηχανικής μάθησης να δοθεί ένα πλήθος από λύσεις που αφορούν τη διάγνωση της δυσλεξίας.

Τα δεδομένα με τα οποία γίνεται η διάγνωση της δυσλεξίας συλλέγονται με μία πληθώρα από τρόπους από τους οποίους εξαρτάται και η προεπεξεργασία τους. Στην εργασία αυτή διερευνώνται εκτενώς οι περιπτώσεις συλλογής από την παιχνιδοποίηση ερωτήσεων και την οφθαλμική ανίχνευση. Για τα δεδομένα αυτά εφαρμόστηκε και μία πληθώρα μεθόδων επιλογής χαρακτηριστικών.

Τα μοντέλα τα οποία χρησιμοποιήθηκαν για την ταξινόμηση των δειγμάτων ήταν ταξινομητές της κλασικής μηχανικής μάθησης. Πιο συγκεκριμένα επιλέχθηκαν τα τυχαία δάση, τα δέντρα ενίσχυσης κλίσης, η λογιστική παλινδρόμηση και μηχανές διανυσμάτων υποστήριξης με χρήση διαφόρων συναρτήσεων πυρήνα. Η επιλογή όλων των παραπάνω μοντέλων, εκτός της λογιστικής παλινδρόμησης, έγινε πρώτον λόγω της μη γραμμικότητας τους και δεύτερον λόγω της ικανότητάς τους να αποδίδουν σημασία στα χαρακτηριστικά κατά την εκπαίδευση. Η δεύτερη ιδιότητά τους μπορεί και να διακρίνει ποια είναι τα χαρακτηριστικά που διαχωρίζουν ένα άτομο με δυσλεξία από ένα άλλο. Εδώ αξίζει να σημειωθεί πως την καλύτερη ορθότητα, 92.63%, έδωσε το μοντέλο μηχανών διανυσμάτων υποστήριξης με πυρήνα γκαουσιανής ακτινικής βάσης για το σύνολο δεδομένων της οφθαλμικής ανίχνευσης.

Τέλος έγινε μία σύγκριση των συνόλων δεδομένων, βασισμένη στα αποτελέσματα της σημασίας χαρακτηριστικών, που προέκυψε από την εκπαίδευση των μοντέλων. Αυτή έδειξε πως με τα δεδομένα της οφθαλμικής ανίχνευσης οι κλάσεις των ατόμων με δυσλεξία και χωρίς είναι πιο εύκολα διαχωρίσιμες.

Λέξεις Κλειδιά

Δυσλεξία, Μηχανική Μάθηση, Επιβλεπόμενη μάθηση, Δυαδική ταξινόμηση, Τυχαία δάση, Δέντρα ενίσχυσης κλίσης, Μηχανές διανυσμάτων υποστήριξης, Λογιστική παλινδρόμηση

Abstract

Dyslexia is a learning difficulty of neurobiological origin. It becomes detectable due to hardship while learning how to read. According to speculations based on the parameters and criteria of diagnosis, 5 - 10% of the population has dyslexia. The early identification of this learning difficulty is imperative for its effective treatment. The lack of early treatment could prove damaging for the reading ability of the child and it's mental health. Subject of the present thesis is to give a variety of solutions regarding the screening of dyslexia with the use of machine learning.

The data used for the screening of dyslexia were collected in various ways, and their preprocessing is dependant on the way they are collected. Here are thoroughly examined the data sets collected by eye-tracking and the gamification of questions. On these kinds of data, many methods of feature selection were applied.

The models that were used for the classification of the samples, depending on whether the person had dyslexia or not, were classifiers of traditional machine learning. Specifically random forests, gradient boosting trees, logistic regression and support vector machines with different kernel functions were used for classification. The use of the aforementioned models was due to the fact that all except logistic regression are not linear models, and the fact that after the learning process they can give the feature importance. This second ability of the models can give a better understanding of which features are responsible for separating the subjects with dyslexia from the control. It must be noted that the best accuracy, 92.63%, was given by support vector machines with radial basis kernel function for the eye-tracking data set.

Finally, a comparison between the different data sets was made, based on the feature importance from the model training. This showed that the eye-tracking data were more suitable for discriminating the class of the people with dyslexia from the class of the people without.

Keywords

Dyslexia, Machine Learning, Supervised learning, Binary Classification, Random Forests, Gradient Boosting Trees, Support Vector Machines, Logistic Regression

Ευχαριστίες

Η εκπόνηση της διπλωματικής μου εργασίας σηματοδοτεί το τέλος των προπτυχιακών μου σπουδών στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας κ. Ανδρέα Σταφυλοπάτη για την ευκαιρία που μου έδωσε να ασχοληθώ με αυτό το επιστημονικό πεδίο. Ακόμα θα ήθελα να ευχαριστήσω θερμά τον κ. Γεώργιο Σιόλα για την αρωγή του και τη άρτια συνεργασία κατά τη διάρκεια εκτέλεσης της παρούσας εργασίας. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου και τους φίλους μου που στάθηκαν δίπλα μου κατά την εκτέλεση αυτής της εργασίας αλλά και τους ανθρώπους που με βοήθησαν να αντιμετωπίσω και εγώ αυτή τη μαθησιακή δυσκολία.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

Αντικείμενο της παρούσας εργασίας είναι η διάγνωση της δυσλεξίας, μέσω δεδομένων τα οποία προκύπτουν από οφθαλμική ανίχνευση ή παιχνιδοποίηση, χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Ειδικότερα τα προβλήματα που πραγματεύεται η παρούσα εργασία, εντάσσονται τόσο στο επίπεδο της προεπεξεργασίας των δεδομένων, όσο και στο επίπεδο της εκπαίδευσης των μοντέλων.

Σε επίπεδο προεπεξεργασίας τα προβλήματα που προκύπτουν διαφέρουν ανά σύνολο δεδομένων. Αναφορικά με τα δεδομένα της οφθαλμικής ανίχνευσης παρατηρούνται χρονοσειρές οι οποίες δεν είναι ίδιου μεγέθους. Έτσι είναι απαραίτητο είτε εφαρμόζοντας κάποιες τεχνικές, εδώ εφαρμόστηκε η δυναμική χρονική στρέβλωση, να αποκτήσουν ίδιο μέγεθος, είτε να γίνει μία διανυσματική αναπαράστασή τους εξάγοντας χαρακτηριστικά. Στη συνέχεια αφού γίνει η εξαγωγή ενός πλήθους χαρακτηριστικών πρέπει να εξετασθεί ποια από αυτά είναι χρήσιμα για την εκπαίδευση των μοντέλων. Η συγκεκριμένη πληροφορία δεν είναι χρήσιμη μόνο για την διευκόλυνση του μοντέλου στην εκπαίδευση, αλλά εφόσον πρόκειται για τη διάγνωση μίας μαθησιακής δυσκολίας, δίνει διαισθητικά και μία πληροφορία για το ποια από τα χαρακτηριστικά του συνόλου δεδομένων είναι σημαντικότερα για τη διάγνωση της. Σχετικά με τα δεδομένα που προέκυψαν από την παιχνιδοπίηση ερωτήσεων, το κυριότερο πρόβλημα που παρατηρήθηκε ήταν η έλλειψη ισορροπίας μεταξύ των κλάσεων των ατόμων που είχαν δυσλεξία και αυτών χωρίς. Έτσι έπρεπε αρχικά να εφαρμοστούν κάποιες τεχνικές δειγματοληψίας. Η υπόλοιπη προεπεξεργασία ήταν παρόμοια με αυτή των δεδομένων της οφθαλμικής ανίχνευσης.

Στο επίπεδο εκπαίδευσης των μοντέλων η βιβλιογραφία προτείνει ένα πλήθος ταξινομητών, μεταξύ των οποίων δεν γίνεται σύγκριση ανά πρόβλημα. Επίσης τα δεδομένα που παράχθηκαν με τις παραπάνω δύο μεθόδους δεν είναι γραμμικά διαχωρίσιμα. Έτσι εξετάστηκε ένα πλήθος διαφορετικών ταξινομητών: ένας γραμμικός ως επίπεδο αναφοράς (λογιστική παλινδρόμηση), δύο που αξιοποιούν δενδρικές δομές (τυχαία δάση και δέντρα ενίσχυσης με κλίση) και μηχανές διανυσμάτων υποστήριξης που αν και είναι γραμμικό μοντέλο, μπορεί να γίνει μη γραμμικό με χρήση συναρτήσεων πυρήνα. Η επιλογή των παραπάνω ταξινομητών δεν έγινε μόνο με το κριτήριο της μη γραμμικότητας, αλλά και για την ικανότητά τους να δίνουν πληροφορία για τη

σημασία των χαρακτηριστικών κατά την εκπαίδευση.

Τέλος μετά την εκπαίδευση των μοντέλων έγινε μία συνοπτική σύγκριση των δύο συνόλων δεδομένων, αναφορικά με την καταλληλότητα τους να δώσουν λύση στο πρόβλημα.

1.2 Οργάνωση του τόμου

Η διπλωματική αυτή διακρίνεται σε έξι κεφάλαια. Στο Κεφάλαιο 2 παρουσιάζονται οι συγγενικές εργασίες, δηλαδή αρχικά αναλύεται η δυσλεξία ως μαθησιακή δυσκολία, αλλά και παρουσιάζονται εργασίες που έχουν ήδη γίνει και έχουν παραπλήσιο περιεχόμενο με αυτή. Στο Κεφάλαιο 3 πλαισιώνεται το θεωρητικό υπόβαθρο που απαιτείται για την κατανόηση της εργασίας. Ξεκινά με μία γενική επισκόπηση της μηχανικής μάθησης ως κλάδο της επιστήμης της πληροφορικής και αναφέρονται οι υποκατηγορίες των προβλημάτων που καλείται να λύσει. Στη συνέχεια αναλύονται οι αλγόριθμοι ταξινόμησης που χρησιμοποιήθηκαν για την επίλυση του προβλήματος, και οι τεχνικές που εφαρμόστηκαν για την προεπεξεργασία των δεδομένων. Τέλος αναλύονται οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν για την μέτρηση της απόδοσης των μοντέλων. Στο Κεφάλαιο 4 γίνεται η περιγραφή των συνόλων δεδομένων που χρησιμοποιήθηκαν σε αυτή την εργασία και μία ανάλυση της προεπεξεργασίας που εφαρμόστηκε στο καθένα από αυτά. Ακόμα γίνεται αναφορά στα μοντέλα που εφαρμόστηκαν για κάθε σύνολο δεδομένων και παρουσιάζονται όλες οι τιμές υπερ-παραμέτρων που χρησιμοποιήθηκαν για την εκπαίδευση. Στο τέλος του κεφαλαίου αναφέρονται όλες οι βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος. Στο Κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα της εκπαίδευσης των μοντέλων ανά σύνολο δεδομένων, γίνεται σύγκριση σχετικά με τις τεχνικές προεπεξεργασίας που εφαρμόστηκαν και παρουσιάζονται γραφήματα των σημαντικότερων χαρακτηριστικών των συνόλων δεδομένων σύμφωνα με τα μοντέλα. Στο Κεφάλαιο 6 γίνεται μία συλλογική ανασκόπηση των αποτελεσμάτων που προκύπτουν από το Κεφάλαιο 5 και γίνεται μία σύνδεση αυτών με το πρόβλημα της διάγνωσης της μαθησιακής δυσκολίας. Ακόμα αναφέρονται ορισμένοι περιορισμοί που υπάρχουν από τη φύση του προβλήματος και ενδιαφέρουσες μελλοντικές επεκτάσεις που προτεινονται στην παρούσα εργασία.

Κεφάλαιο 2

Συγγενικές εργασίες

Στο κεφάλαιο αυτό αρχικά γίνεται μία εισαγωγή, η οποία περιγράφει την δυσλεξία ως μαθησιακή δυσκολία και τα προβλήματα που προκύπτουν εξαιτίας της. Στη συνέχεια περιγράφονται τρόποι εντοπισμού της και παρουσιάζονται έρευνες που έχουν ήδη γίνει γύρω από αυτούς. Σκοπός του κεφαλαίου είναι ο αναγνώστης να κατανοήσει τη φύση του προβλήματος.

2.1 Εισαγωγή

Η δυσλεξία είναι μία συγκεκριμένη μαθησιακή δυσκολία νευροβιολογικής προέλευσης. Χαρακτηρίζεται από δυσκολίες που αφορούν την ακριβή και ομαλή αναγνώριση των λέξεων. Πιο συγκεκριμένα, το άτομο παρουσιάζει δυσκολίες στην ανάλυση των λέξεων σε ακουστικές μονάδες συλλαβικής βάσης και στη σύνθεση συλλαβικών ακουστικών μονάδων σε λεκτικά σύνολα με εννοιακό περιεχόμενο. Η διαταραχή αυτή γίνεται αντιληπτή κατά τις δυσκολίες στην εκμάθηση της ανάγνωσης, παρά την επαρκή νοημοσύνη και την εμφανώς αποτελεσματική σχολική συμβολή [1]. Εκτιμάται πώς το 5 - 10% του πληθυσμού έχει δυσλεξία, αλλά αυτές οι εκτιμήσεις βασίζονται στις παραμέτρους και στα κριτήρια της διάγνωσης.

Η διάγνωση της δυσλεξίας είναι ιδιαίτερα δύσκολη σε γλώσσες με εμφανείς ορθογραφίες, όπου το γράφημα και το φώνημα είναι πιο συνεπή [2, 3]. Ακόμα, όταν κάποιος γνωρίζει πως έχει δυσλεξία, μπορεί να αναπτύξει μηχανισμούς ώστε να εξομαλύνει τις αρνητικές επιπτώσεις της [4, 5]. Έτσι ο εντοπισμός και η επαγγελματική υποστήριξη από μικρές ηλικίες είναι οι πιο αποτελεσματικές μέθοδοι υποβοήθησης των παιδιών με δυσκολίες στην ανάγνωση. Η αναμονή για την επίσημη διάγνωση της δυσλεξίας πριν την προσέγγιση για ειδική βοήθεα μπορεί να αποδειχθεί επιζήμια για το παιδί [6]. Ειδικότερα, είναι ιδιαίτερα δύσκολο για το παιδί αν δεν έχει την αρωγή ειδικών, να φτάσει το βαθμό ανάγνωσης της τάξης του, επηρεάζοντας την συνολική απόδοση του στο σχολείο. Τα παραπάνω προκαλούν χαμηλή αυτοπεποίθηση, έλλειψη για κίνητρο όσον αφορά την εκπαίδευση και κατάθλιψη [7].

Η πλήρης διάγνωση της δυσλεξίας είναι μία σύνθετη διαδικασία, που περιλαμβάνει παραμέτρους όπως δημογραφικές πληροφορίες σχετικά με το παιδί, νοημοσύνη, δεξιότητες προφορικού λόγου, ανάγνωσης λέξεων, αποκωδικοποίησης, ορθογραφίας, φωνολογικής επεξεργασίας και κατανόησης προφορικού λόγου σύμφωνα με τον διεθνή σύνδεσμο ατόμων με δυσλεξία [9].

Στην Ελλάδα η διάγνωση της μαθησιακής δυσκολίας στα ΚΔΑΥ μπορεί να καθυστερήσει για μεγάλο χρονικό διάστημα, καθώς τα κέντρα αυτά στην Αθήνα και στη Θεσσαλονίκη καλούνται να εξυπηρετήσουν πληθυσμούς μαθητών μεγαλύτερου όγκου από αυτον που δύνανται [8]. Το ίδιο πρόβλημα αντιμετωπίζουν και άλλες χώρες, όπως για παράδειγμα η Ιρλανδία όπου η αναμονή μπορεί να διαρκέσει έως και εννέα μήνες. Έτσι ενώ η πλήρης διάγνωση της δυσλεξίας είναι απαραίτητη για την αντιμετώπιση της, δεν είναι πρακτικά εφικτή στο σωστό χρόνο σε πολύ μεγάλους πληθυσμούς μαθητών[10]. Στα παραπάνω προβλήματα η μηχανική μάθηση έρχεται να δώσει μία πληθώρα από λύσεις, οι οποίες είναι φθηνότερες και λιγότερο χρονοβόρες.[11]

2.2 Παιχνιδοποίηση

Η παιχνιδοποίηση είναι ίσως ο πιο οικονομικός και σύντομος τρόπος για μία πρώτη μορφή εντοπισμού της δυσλεξίας, καθώς δεν απαιτεί επιπλέον εξοπλισμό πέρα από έναν υπολογιστή ή τάμπλετ. Ο εξεταζόμενος καλείται να παίξει ένα παιχνίδι κατά τη διάρκεια του οποίου συλλέγονται μετρήσεις. Το παιχνίδι έχει δομή τέτοια, ώστε να ελέγχει χαρακτηριστικά του συμμετέχοντα που σχετίζονται με τη δυσλεξία και οι μετρήσεις είναι το πλήθος των κλικ μέχρι να υπάρξει επιτυχία ή αστοχία, ο χρόνος που αφιερώνεται σε κάθε γύρο των παιχνιδιών, οι σωστές απαντήσεις κ.α. Οι Rauschenberger M, Rello L et al [12, 13, 14] προτείνουν δύο διαφορετικές προσεγγίσεις στη δομή των παιχνιδιών.

Η πρώτη προσέγγιση εξαρτάται από τη γλώσσα του εξεταζόμενου. Αποτελείται από μία σειρά παιχνιδιών όπου μέσα από επιλογές γραμμάτων και λέξεων, πολλές φορές με τη συμβολή ακουστικών ερεθισμάτων, αξιολογούνται οι ακόλουθοι δείκτες:

- **Γλωσσικές ικανότητες**, όπως η φωνολογική επίγνωση και η συλλαβική επίγνωση.
- **Ενεργός μνήμη**, όπως οπτική (αλφαβητική), ακουστική (φωνολογική).
- **Εκτελεστικές λειτουργίες**, όπως ενεργοποίηση και προσοχή, διατήρηση προσοχής.
- **Διαδικασίες αντίληψης**, όπως η οπτική διαφοροποίηση και η κατηγοριοποίηση.

Language Skills	**Working Memory**
Alphabetic Awareness	Visual (alphabetical)
Phonological Awareness	Auditory (phonology)
Syllabic Awareness	Sequential (auditory)
Lexical Awareness	Sequential (visual)
Morphological Awareness	**Executive Functions**
Syntactic Awareness	Activation and Attention
Semantic Awareness	Sustained Attention
Orthographic Awareness	Simultaneous Attention
Perceptual Processes	
Visual Discrimination and Categorization	
Auditory Discrimination and Categorization	

Πίνακας 2.1: Γνωστικοί δείκτες.

Η δεύτερη προσέγγιση είναι ανεξάρτητη της γλώσσας του εξεταζόμενου και βασίζεται στην οπτική και ηχητική αντίληψη. Η εμφάνιση ελλείμματος στη γρήγορη ακουστική επεξεργασία, υποδεικνύει ότι τα άτομα με δυσλεξία έχουν πρόβλημα με την επεξεργασία σύντομων ακουστικών σημάτων. Συμπληρωματικά, η δυσλεξία χαρακτηρίζεται και από δυσκολία στην αποκωδικοποίηση οπτικών ερεθισμάτων, για παράδειγμα δυσκολία στην αναγνώριση γραμμάτων. Εκμεταλλευόμενοι αυτά τα χαρακτηριστικά, οι συγγραφείς προτείνουν τη δημιουργία ενός παιχνιδιού με δύο στάδια. Το πρώτο στάδιο είναι το ακουστικό, όπου μέσα από την αντιστοίχιση και κατηγοριοποίηση ηχητικών ερεθισμάτων αποκτώνται πληροφορίες σε σχέση με τις ικανότητες επεξεργασίας του ήχου που έχει ο συμμετέχων, ενώ το δεύτερο στάδιο είναι το οπτικό, όπου μέσα από την αναγνώριση και αντιστοίχιση εικόνων ελέγχονται οι ικανότητες αποκωδικοποίησης.

Μετά τη συλλογή των δεδομένων με τους παραπάνω τρόπους, οι συγγραφείς χρησιμοποίησαν μοντέλα μηχανικής μάθησης για την ταξινόμηση των ατόμων σε δύο ομάδες, η μία με δυσλεξία και η άλλη χωρίς. Τα μοντέλα που δοκιμάστηκαν ήταν μηχανές διανυσματικής υποστήριξης(SVM) και τυχαία δάση(RF), που έπιασαν αντίστοιχα 84.62% και 79.4% ορθότητα. Τέλος παραδείγματα λογισμικού για τη διάγνωση της δυσλεξίας στα Αγγλικά, αποτελούν τα Lexercise Screener [15] και Nessy [16], ενώ το Dytective [17] υποστηρίζει Αγγλικά και Ισπανικά.

2.3 Οφθαλμική ανίχνευση

Σε αντίθεση με την προηγούμενη μέθοδο η ανίχνευση του ματιού απαιτεί την ύπαρξη ενός οφθαλμικού ιχνηλάτη. Το μηχάνημα αυτό, εντοπίζοντας και συνδυάζοντας την κερατοειδή αντανάκλαση και την κόρη του οφθαλμού, υπολογίζει το σημείο που τέμνεται το βλέμμα και η επιφάνεια όρασης [20]. Έτσι, με την ανίχνευση του οφθαλμού μπορεί να μελετηθεί και η κίνηση του. Η κίνηση του ματιού χαρακτηρίζεται από σακκάδες και φάσεις εστίασης. Αναλύοντας, οι φάσεις εστίασης είναι χρονικά διαστήματα κατά τα οποία το μάτι παραμένει ακίνητο, και διαρκούν περίπου 200-300 ms και οι σακκάδες είναι γρήγορες κινήσεις του ματιού που συμβαίνουν ανάμεσα σε φάσεις εστίασης.

Εδώ πρέπει να τονιστεί πως οι κινήσεις των ματιών όπως περιγράφηκαν παραπάνω διαφέρουν ανάμεσα σε τυπικούς αναγνώστες, και σε δυσλεκτικούς. Οι δεύτεροι έχουν μεγαλύτερες και περισσότερες φάσεις εστίασης, μικρότερο μήκος σακκάδων και μεγαλύτερο αριθμό από παλινδρομήσεις, δηλαδή σακκάδες που αναφορικά με ένα κείμενο έχουν κατεύθυνση προς σημείο που έχει ήδη διαβαστεί [21]. Αυτή την ειδοποιός διαφορά ανάμεσα στη συμπεριφορά των ματιών που έχουν οι αναγνώστες αποτελεί το πάτημα για την εφαρμογή των τεχνικών που αναφέρονται παρακάτω.

Συλλέγοντας χρονοσειρές από τις κινήσεις των ματιών κατά το διάβασμα, δηλαδή τη θέση του ματιού ενός αναγνώστη συσχετισμένη σε μία χρονική στιγμή, και εξάγοντας χαρακτηριστικά από αυτές, όπως αριθμό σακκάδων, αριθμό φάσεων εστίασης, μέση διάρκεια φάσεων εστίασης κ.α., οι Nilsson Benfatto M et al [19] και Asvestopoulou T, Manousaki V et al [18] χρησιμοποιούν μηχανική μάθηση για την ταξινόμηση των ατόμων σε ομάδες με δυσλε-

ξία ή χωρίς. Οι πρώτοι χρησιμοποιούν μηχανές διανυσματικής υποστήριξης σε συνδυασμό με επαναλαμβανόμενο αποκλεισμό χαρακτηριστικών (SVM-RFE) και επιτυγχάνουν 95.3% +- 4.6% ορθότητα. Οι δεύτεροι χρησιμοποιούν μηχανές διανυσματικής υποστήριξης, απλοϊκό Μπεϋζιανό ταξινομητή (Naïve Bayesian) και ομαδοποιητή K-μέσων (K-means clustering), σε συνδυασμό με παλινδρόμηση ελάχιστης απόλυτης σμίκρυνσης και επιλογής, εκ των οποίων μεγαλύτερη ορθότητα επιτυγχάνουν οι μηχανές διανυσματικής υποστήριξης 97.10 %.

2.4 Ηλεκτροεγκεφαλογράφημα (EEG)

Έρευνες έχουν δείξει πως τα άτομα με δυσλεξία εμφανίζουν διαφορετικές εγκεφαλικές δομές και εγκεφαλική συμπεριφορά σε σχέση με τα αντίστοιχα άτομα ελέγχου (άτομα χωρις δυσλεξία). Με το ηλεκτροεγκεφαλογράφημα γίνεται εφικτή η παρατήρηση νευρολογικών συμπεριφορών, ενώ το άτομο εκτελεί δραστηριότητες που βοηθούν στη διάγνωση της δυσλεξίας, π.χ. διάβασμα [23]. Στη συνέχεια εξάγοντας στατιστικά χαρακτηριστικά από τα κανάλια ηλεκτροεγκεφαλογραφημάτων, είτε θεωρώντας τα κανάλια ως απλά σήματα (μέση τιμή, τυπική απόκλιση κ.α.)[23], είτε θεωρώντας τα ως κόμβους δικτύου (χαρακτηριστικά μήκος μονοπατιού, βαθμός κόμβου κ.α.)[22], και εφαρμόζοντας μηχανική μάθηση, οι συγγραφείς προσπαθούν να αποδείξουν πως τα σήματα των εγκεφαλικών κυμάτων διαφέρουν ανάμεσα σε άτομα με δυσλεξία και χωρίς.

Οι Rezvani Z, Zare M et al [22] χρησιμοποιούν μηχανές διανυσματικής υποστήριξης, και ταξινόμηση K κοντινότερων γειτόνων (KNN), με τις μηχανές διανυσματικής υποστήριξης γραμμικού πυρήνα να έχουν ορθότητα 95.34%. Ομοίως και οι Perera H et al [23] χρησιμοποιούν μηχανές διανυσματικής υποστήριξης, αλλά με πυρήνα πολυώνυμου τρίτου βαθμού και επιτυγχάνουν μέγιστη ορθότητα 71.88%. Εδώ όμως πρέπει να σημειωθεί πως σκοπός της έρευνας των Perera H et al δεν ήταν η βελτιστοποίηση της απόδοσης του μοντέλου, αλλά η απόδειξη της σχέσης που αναφέρθηκε στην προηγούμενη παράγραφο.

2.5 Μαγνητική τομογραφία (MRI)

Ομοίως με τα ηλεκτροεγκεφαλογραφήματα, η διάγνωση της δυσλεξίας από μαγνητική τομογραφία βασίζεται στις διαφορές στη δομή του εγκεφάλου μεταξύ ατόμων με δυσλεξία και των αντίστοιχων ατόμων ελέγχου. Με το συνδυασμό δομικής μαγνητικής τομογραφίας και απεικόνισης του τανυστή της διάχυσης, οι Cui et al [25] και Płońskiet al [24] εξάγουν για κάθε τμήμα του εγκεφάλου χαρακτηριστικά σχετικά με τις ιδιότητες της λευκής ουσίας όπως όγκο, φλοιικό πάχος κ.α.

Τα παραπάνω χαρακτηριστικά χρησιμοποιήθηκαν από τους Cui et al [25] για την εκπαίδευση μηχανών διανυσματικής υποστήριξης με γραμμικό πυρήνα, και πέτυχαν ορθότητα 83.61%. Ακόμα αποτέλεσμα της έρευνας ήταν και η ανάδειξη των χαρακτηριστικών που ανήκαν στο σύστημα διαβάσματος, το μεταιχμιακό σύστημα και το κινητικό σύστημα ως τα σημαντικότερα για την διάκριση μεταξύ των δύο ομάδων.

Οι Płońskiet al [24] εστίασαν περισσότερο στη δοκιμή διαφορετικών ταξινομητών δοκιμάζοντας λογιστική παλινδρόμηση (LR), μηχανές διανυσματικής υποστήριξης με γραμμικό πυρήνα και τυχαία δάση. Μέγιστη ορθότητα 65% επιτεύχθηκε από τη λογιστική παλινδρόμηση και τα τυχαία δάση.

Εδώ πρέπει να τονιστεί πως ο χαρακτήρας των δύο παραπάνω ερευνών ήταν περισσότερο διερευνητικός ως προς τη φύση του προβλήματος, δηλαδή αν είναι εφικτή η διάγνωση της μαθησιακής δυσκολίας με μηχανική μάθηση από δεδομένα μαγνητικής τομογραφίας.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό περιγράφονται θεωρητικά όλες οι τεχνικές, οι αλγόριθμοι και οι μέθοδοι που χρησιμοποιήθηκαν στην παρούσα εργασία. Αναλύονται θεωρητικές έννοιες και παρουσιάζονται οι τεχνικές προεπεξεργασίας των δεδομένων και οι αλγόριθμοι που εφαρμόστηκαν για την ταξινόμηση. Μετά από αυτό ο αναγνώστης θα είναι σε θέση να καταλάβει τις διατάξεις των μοντέλων που χρησιμοποιήθηκαν στα δεδομένα.

3.1 Μηχανική μάθηση

Η μηχανική μάθηση κατά τον Tom M. Mitchell [27], ορίζεται ως "Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης Π, αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E". Εμπειρικά ο παραπάνω ορισμός γίνεται πιο κατανοητός από τον Arthur Samuel [26] που την ορίζει ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί". Στα πλαίσια της μηχανικής μάθησης αναπτύσσονται αλγόριθμοι που δέχονται ως είσοδο πειραματικά δεδομένα και κατασκευάζουν μοντέλα, προκειμένου να κάνουν προβλέψεις ή να λάβουν αποφάσεις με βάση αυτά. Κατά τη διαδικασία της επαγωγικής μάθησης ο άνθρωπος μαθαίνει μέσα από ερεθίσματα που δέχεται από το περιβάλλον του, δημιουργώντας μία αφαιρετική εικόνα για αυτό που ονομάζεται νοητικό μοντέλο. Εδώ γίνεται φανερό πως με τη μηχανική μάθηση οι ερευνητές προσπαθούν να εισάγουν στον υπολογιστή τον τρόπο που λειτουργούμε εμείς ως σκεπτόμενες οντότητες.

Ως κλάδος της επιστήμης των υπολογιστών η μηχανική μάθηση προέρχεται από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Είναι ένα νέο πεδίο το οποίο εξελίσσεται συνεχώς και βρίσκει εφαρμογές σε πολλούς τομείς όπως η βιοπληροφορική, η οικονομία, το μάρκετινγκ, η χημειοπληροφορική κ.α. Παραδείγματα κλάσεων προβλημάτων που καλείται να λύσει είναι η αντίληψη του υπολογιστή μέσω όρασης, η επεξεργασία της φυσικής γλώσσας, η ταξινόμηση και η ομαδοποίηση μέσα από δεδομένα κτλ. που επαληθεύουν πως αυτή η επιστήμη προσπαθεί να κάνει τον υπολογιστή να λειτουργεί όπως ο άνθρωπος.

Τα προβλήματα της μηχανικής μάθησης μπορούν, ανάλογα με τη φύση του εκπαιδευτικού

σήματος, να διακριθούν σε τρεις κατηγορίες: επιβλεπόμενη μάθηση, μη-επιβλεπόμενη μάθηση και ενισχυτική μάθηση.

- **Επιβλεπόμενη μάθηση:** Κατά τη διάρκεια της εκμάθησης, μία συνάρτηση αντιστοιχεί μία είσοδο σε μία έξοδο, βασισμένη σε ζευγάρια εισόδου - εξόδου [28]. Ο αλγόριθμος αναλύει τα δεδομένα εκπαίδευσης και παράγει ένα μοντέλο, το οποίο μπορεί να χρησιμοποιηθεί για την αντιστοίχιση νέων δεδομένων. Αυτό επιτυγχάνεται με τη βελτιστοποίηση μίας συνάρτησης κόστους που συσχετίζει την είσοδο με την έξοδο. Προβλήματα επιβλεπόμενης μάθησης είναι η ταξινόμηση, η πρόγνωση και η διερμηνεία.
- **Μη-επιβλεπόμενη μάθηση:** Στην μη-επιβλεπόμενη μάθηση ο αλγόριθμος μαθαίνει μοτίβα από δεδομένα χωρίς ετικέτες, δηλαδή από δεδομένα που σε αντίθεση με πριν η έξοδος δεν παρέχεται στον αλγόριθμο. Σε αυτό το είδος μηχανικής μάθησης η μηχανή χτίζει μία εσωτερική αναπαράσταση του κόσμου της [29]. Δύο προβλήματα μη-επιβλεπόμενης μάθησης αποτελούν η ανάλυση συσχετισμών και η ομαδοποίηση.
- **Ενισχυτική μάθηση:** Με την ενισχυτική μάθηση το πρόβλημα μοντελοποιείται ως πράκτορες που λαμβάνουν αποφάσεις σε ένα περιβάλλον, το οποίο πολλές φορές αλλάζει κατάσταση. Σκοπός τους είναι να μεγιστοποιηθεί το συνολικό κέρδος από τις επιβραβεύσεις που λαμβάνουν παίρνοντας μία απόφαση [30]. Η παρούσα κατηγορία προσπαθεί να βρει την χρυσή τομή ανάμεσα στην εξερεύνηση του χώρου διερεύνησης του προβλήματος, και στην εκμετάλλευση της γνώσης που έχει συσσωρευθεί.

3.2 Αλγόριθμοι Ταξινόμησης

Σε αυτή την ενότητα περιγράφονται θεωρητικά οι αλγόριθμοι που χρησιμοποιήθηκαν για την ταξινόμηση των δεδομένων σε άτομα με δυσλεξία και σε άτομα χωρίς.

Με κάθε έναν από τους αλγορίθμους που αναφέρονται παρακάτω, κάποιος μπορεί να εξάγει συμπεράσματα και για το κατά πόσο σημαντικά ήταν τα χαρακτηριστικά του συνόλου δεδομένων στην ταξινόμηση.

3.2.1 Δενδρικές δομές

Δέντρα απόφασης για ταξινόμηση

Τα δέντρα απόφασης(Decision Trees) διαχωρίζουν τον χώρο των δεδομένων σε ορθογωνικές δομές και εκπαιδεύουν ένα απλό μοντέλο, όπως είναι μία σταθερά, σε κάθε μία από αυτές. Συμπληρώνοντας, κάθε διαμέριση αντιστοιχεί σε έναν κόμβο ενός δυαδικού δέντρου, οι ακμές που οδηγούν στα παιδιά του είναι υποχώροι, και τα παιδιά του είναι διαμερίσεις αυτών των υποχώρων. Θεωρώντας N παρατηρήσεις της μορφής $(x_{i1}, x_{i2}, ..., x_{ip}, y_i)$ όπου i = 1,2,...,N, p το σύνολο των εισόδων, y_i μία από K κλάσεις και M διαμερίσεις $R_1, ..., R_m, ..., R_M$ του χώρου δεδομένων από το δέντρο, ορίζουμε:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

Όπου I η δείκτρια συνάρτηση και $\hat{p}_{mk}$ η ποσότητα των παρατηρήσεων της κλάσης k στη διαμέριση m προς το πλήθος των στοιχείων της διαμέρισης.

Όσες παρατηρήσεις βρίσκονται στην διαμέριση m ταξινομούνται με βάση την πλειονότητα των παρατηρήσεων που αυτή περιέχει, δηλαδή:

$$k(m) = \arg\max_k \hat{p}_{mk}$$

Εδώ αξίζει να αναφερθεί η διαδικασία εύρεσης των διαμερίσεων $R_1, ..., R_M$. Θεωρώντας μία συνάρτηση $Q_m(T)$ ως την συνάρτηση απώλειας του δέντρου για τη διαμέριση m, μία μεταβλητή j και ένα σημείο διαμέρισης s ορίζονται οι υποχώροι:

$$R_1(j, s) = \{X \mid X_j \leq s\} \text{ και } R_2(j, s) = \{X \mid X_j < s\}$$

Αναζητούμε το ζεύγος των j και s που ικανοποιούν την:

$$\min_{j,s}[Q_1(T) + Q_2(T)]$$

Για κάθε j η επιλογή του s μπορεί να γίνει εύκολα, ψάχνοντας όλα τα δεδομένα εισόδου, και άρα η επιλογή του ζεύγους (j,s) είναι εύκολη. Αναδρομικά επαναλαμβάνουμε σε κάθε μία από τις περιοχές που χωρίσαμε.

Στην παραπάνω διαδικασία, για μεγάλο μέγεθος δέντρου, ελλοχεύει ο κίνδυνος της υπερεκπαίδευσης στα δεδομένα. Έτσι η βιβλιογραφία προτείνει την εξής στρατηγική για την αποφυγή του παραπάνω προβλήματος. Αρχικά η διαδικασία διαμερισμού του χώρου δεδομένων σταματά όταν οι κόμβοι φύλλα έχουν έναν ελάχιστο αριθμό από στοιχεία, π.χ. 5 και έτσι δημιουργείται ένα δέντρο,έστω T_0. Τότε χρησιμοποιείται μία τεχνική κόστους - πολυπλοκότητας για το κλάδεμα του δέντρου. Ορίζουμε ένα υποδέντρο $T \subset T_0$, που προκύπτει αν κλαδέψουμε το δέντρο, δηλαδή αν αφαιρέσουμε έναν οποιοδήποτε αριθμό κόμβων του. Αν ένας κόμβος φύλλο m όπως παραπάνω, αντιστοιχεί σε μία περιοχή R_m, και συνολικά υπάρχουν $\mid T \mid$ περιοχές στο υποδέντρο, τότε το κριτήριο κόστους - πολυπλοκότητας παίρνει τη μορφή:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha \mid T \mid$$

Εξετάζονται πολλές τιμές για το $\alpha \geq 0$, το οποίο ερμηνεύεται ως η ανταλλαγή μεταξύ του μεγέθους του δέντρου, και της ικανότητας του να μάθει τα δεδομένα. Τέλος το $T \subset T_0$ δέντρο που επιλέγεται, πρέπει ελαχιστοποιεί το $C_\alpha(T)$ για το εκάστοτε α.

Οι διαφορετικές $Q_m(T)$ που προτείνονται για ταξινόμηση είναι:

$$Misclassification\ error : \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{m,k(m)}$$

$$Gini\ index : \sum_{k \neq k'} \hat{p}_{m,k}\hat{p}_{m,k'} = \sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$$

$$Cross - entropy\ or\ deviance : -\sum_{k=1}^{K} \hat{p}_{m,k} \log \hat{p}_{m,k}$$

Ενθυλάκωση

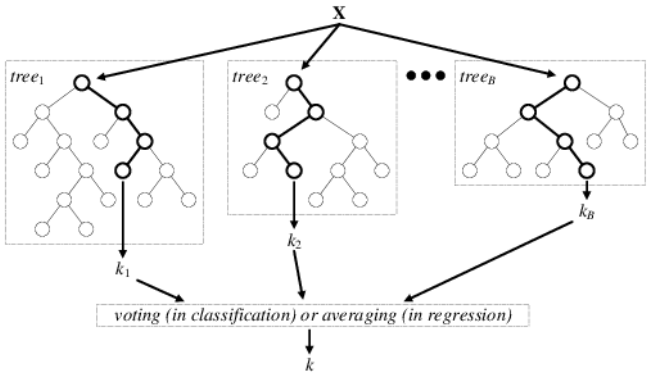
Για τη διαδικασία της ενθυλάκωσης (Bagging), γίνεται αρχικά η δημιουργία ενός προκαθορισμένου πλήθους B συνόλων δεδομένων εκπαίδευσης, με βάση το αρχικό. Κάθε σύνολο δημιουργείται δειγματοληπτικά με επανατοποθέτηση από το αρχικό (bootstrap samples), και καθένα από αυτά χρησιμοποιείται για την εκπαίδευση ενός μοντέλου. Τελικά όταν το νέο μοντέλο δεχτεί δεδομένα, το αποτέλεσμα που παράγει (στην περίπτωση της ταξινόμησης) είναι η κατηγορία που ψήφισε η πλειοψηφία των μοντέλων. Σκοπός της διαδικασίας είναι η μείωση της διασποράς του αποτελέσματος του μοντέλου, μέσα από τον συνδυασμό μοντέλων που έχουν θόρυβο, αλλά είναι αμερόληπτα.

Τυχαία δάση

Συνδυάζοντας τις παραπάνω δύο μεθόδους προκύπτει ο εξής αλγόριθμος:

Algorithm 1 Τυχαία δάση

1. **Για** $b = 0$ μέχρι το B:
   (α) **Δημιούργησε** ένα νέο σύνολο εκπαίδευσης από το αρχικό παίρνοντας δείγματα με επανατοποθέτηση.
   (β) **Δημιούργησε** ένα δέντρο απόφασης και εκπαίδευσέ το στο νέο σύνολο δεδομένων.
2. **Επέστρεψε** το νέο συνολικό μοντέλο που προέκυψε.
3. **Για κάθε** πρόβλεψη σε νέο σημείο βγάλε ως αποτέλεσμα την κλάση που προέκυψε στην πλειοψηφία των δέντρων.



Σχήμα 3.1: Τυχαία δάση.

Το μοντέλο του βήματος 2 ονομάζεται τυχαίο δάσος (Random Forest) καθώς αποτελείται από πολλά δέντρα αποφάσεων, εκπαιδευμένα σε τυχαία υποσύνολα του συνόλου εκπαίδευσης.

Δέντρα ενίσχυσης

Διατηρώντας τη σημειολογία των δέντρων αποφάσεων, σε κάθε περιοχή του χώρου ενός δέντρου απόφασης R_j αντιστοιχούμε μία σταθερά γ_j ώστε οι προβλέψεις να γίνονται ως:

$$x \in R_j \Rightarrow f(x) = \gamma_j$$

Και το δέντρο πλέον ορίζεται ως:

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j)$$

Όπου J το πλήθος των τελικών περιοχών διαμέρισης του δέντρου και $\Theta = \{R_j, \gamma_j\}_{j=1}^{J}$ μεταπαράμετροι. Αυτές οι παράμετροι βρίσκονται ελαχιστοποιώντας το εμπειρικό ρίσκο:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=0}^{J} \sum_{x \in R_j} L(y_i, \gamma_j)$$

Όπου $L(y_i, \gamma_j)$ η συνάρτηση σφάλματος μεταξύ των πραγματικών τιμών y_i και των προβλέψεων του μοντέλου γ_j. Εδώ αναφέρουμε τον αλγόριθμο AdaBoost ο οποίος είναι ο εξής:

Algorithm 2 AdaBoost

1. **Αρχικοποίησε** τα βάρη $w_i = 1/N, i = 1, 2, ..., N$, όπου N ο αριθμός των δεδομένων εκπαίδευσης.
2. **Για** $m = 1$ μέχρι το M:
   (α) **Εκπαίδευσε** το μοντέλο ταξινόμησης G_m στα δεδομένα εκπαίδευσης με βάρη w_i
   (β) **Υπολόγισε** το σφάλμα
   $$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$$
   (ς) **Υπολόγισε** $\alpha_m = \log((1 - err_m)/err_m)$
   (δ) **Θέσε**
   $$w_i \leftarrow w_i \cdot e^{[\alpha_m \cdot I(y_i \neq G_m(x_i))]}, i = 1, 2, ..., N$$
3. **Επέστρεψε** ως αποτέλεσμα $G(x) = sign[\sum_{m=1}^{M} \alpha_m G_m(x)]$

Αν θεωρήσουμε πως τα μοντέλα ταξινόμησης $G_m(x)$ γενικεύονται ως:

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m),$$

όπου β_m οι συνιστώσες διαστολής, και $b(x; \gamma_m)$ απλούστερες συναρτήσεις που λαμβάνουν εισόδους πολλών μεταβλητών x και έχουν παραμέτρους γ_m, ο AdaBoost γενικεύεται ως:

Algorithm 3 Γενίκευση AdaBoost

1. **Αρχικοποίησε** $f_0(x) = 0$.
2. **Για κάθε** $m = 1$ μέχρι M:
   (α) **Υπολόγισε**
   $$(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma))$$
   (β) **Θέσε** $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.

Εφαρμόζοντας αυτούς τους αλγορίθμους στα δέντρα έχουμε,

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m)$$

Και σε κάθε βήμα θα πρέπει να λυθεί:

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \Theta_m))$$

Με βάση το παραπάνω $\hat{\Theta}_m$ το $\hat{\gamma}_{jm}$ που προκύπτει είναι:

$$\hat{\gamma}_{jm} = \arg\min_{\gamma_{jm}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm})$$

Για την ταξινόμηση δεδομένων που ανήκουν σε K κλάσεις, ως συνάρτηση σφάλματος προτείνεται η πολυωνυμική απόκλιση K-κλάσεων:

$$L(y, p(x)) = -\sum_{k=1}^{K} I(y = G_k) f_k(x) + \log(\sum_{l=1}^{K} e^{f_l(x)})$$

όπου G_k συμβολίζεται η κλάση k, και για το $p_k(x)$:

$$p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}$$

Στα παραπάνω μπορεί να προστεθεί και η χρήση της κλίσης (Gradient), για την βελτιστοποίηση των υπολογισμών. Αυτή η μέθοδος αποτελεί μία προσέγγιση της παραπάνω διαδικασίας κατά την οποία θεωρούμε το σφάλμα από την χρήση της f(x) για την πρόβλεψη του y ως:

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i))$$

Έτσι ο νέος στόχος που προκύπτει είναι η ελαχιστοποίηση της L(f) ως προς την f.

Αντιμετωπίζοντας το παραπάνω ως αριθμητική βελτιστοποίηση:

$$\hat{f} = \arg\min_{f} L(f)$$

όπου οι παράμετροι f είναι τιμές που προσεγγίζουν την f(x) σε κάθε κάθε σημείο x του συνόλου εκπαίδευσης, και λύνεται ως:

$$f_M = \sum_{m=0}^{M} h_m$$

Κατά τη διαδικασία της πτώσης της παραγώγου επιλέγεται $h_m = -\rho_m g_m$ όπου ρ_m βαθμωτό και g_m η παράγωγος της L(f) υπολογισμένη στο $f = f_{m-1}$, και συμβολίζουμε για το πρόβλημα με K κλάσεις:

$$-g_{ikm} = [\frac{\partial L(y_i, f_1(x_i), ..., f_K(x_i))}{\partial f_k(x_i)}]_{f(x_i)=f_{m-1}(x_i)} = I(y_i = G_k) - p_k(x_i)$$

$$\rho_m = \arg\min_{\rho} L(f_{m-1} - \rho g_m)$$

$$f_m = f_{m-1} - \rho_m g_m$$

Λαμβάνοντας υπόψη όλα τα παραπάνω ο αλγόριθμος για τα δέντρα ενίσχυσης κλίσης (Gradient Tree Boosting Algorithm) για ταξινόμηση είναι:

Algorithm 4 Δέντρα ενίσχυσης κλίσης

1. **Αρχικοποίησε** $f_{k0}(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$
2. **Για** $m = 1$ μέχρι M:
   (α) **Για** $k = 1$ μέχρι K:
      i. **Για** $i = 1$ μέχρι N:
         A. **Υπολόγισε**: $g_{ikm} = -[\frac{\partial L(y_i, f_1(x_i), ..., f_K(x_i))}{\partial f_k(x_i)}]_{f(x_i)=f_{k,m-1}(x_i)}$
      ii. **Εκπαίδευσε** ένα δέντρο απόφασης στα g_{ikm}, το οποίο θα δώσει περιοχές $R_{jkm}, j = 1, 2, ..., J_{km}$
      iii. **Για** $j = 1$ μέχρι J_{km}:
         A. **Υπολόγισε**: $\gamma_{jkm} = \arg\min_{\gamma} \sum_{x_i \in R_{jkm}} L(y_i, f_{k,m-1}(x_i) + \gamma)$
      iv. **Ενημέρωσε** $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm})$
3. **Επέστρεψε** ως αποτέλεσμα $\hat{f}(x) = \arg\max_k p_k(x)$, όπου $f_k(x) = f_{kM}(x)$

Συμπερασματικά στην παραπάνω διαδικασία εκπαιδεύονται MxK δέντρα απόφασης. Σκοπός κάθε δέντρου στην επανάληψη m είναι η διόρθωση του σφάλματος του προηγούμενου, και τελικά η διόρθωση του συνολικού σφάλματος.

3.2.2 Λογιστική Παλινδρόμηση

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic regression) μοντελοποιεί τις εκ των υστέρων πιθανότητες Κ κλάσεων μέσω γραμμικών συναρτήσεων του x. Ο υπολογισμός αυτός γίνεται ως εξής:

$$\log \frac{Pr(G=1 \mid X=x)}{Pr(G=K \mid X=x)} = w_{10} + w_1^T x$$

$$\log \frac{Pr(G=2 \mid X=x)}{Pr(G=K \mid X=x)} = w_{20} + w_2^T x$$

$$\dots$$

$$\log \frac{Pr(G=K-1 \mid X=x)}{Pr(G=K \mid X=x)} = w_{(K-1)0} + w_{(K-1)}^T x$$

Το μοντέλο εκφράζεται μέσα από λογαριθμικές πιθανότητες, με παρονομαστή την πιθανότητα της τελευταίας κλάσης. Μετά από λίγους υπολογισμούς οι πιθανότητες αυτές προκύπτουν ως:

$$Pr(G=k \mid X=x) = \frac{e^{w_{k0}+w_k^T x}}{1+\sum_{l=1}^{K-1} e^{w_{l0}+w_l^T x}}, k = 1, ..., K-1$$

$$Pr(G=K \mid X=x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{w_{l0}+w_l^T x}}$$

Όπου $\theta = \{w_{10}, w_1^T, ..., w_{(K-1)0}, w_{(K-1)}^T\}$ οι παράμετροι του προβλήματος και $G = 1, 2, ..., K-1$ οι διαφορετικές κλάσεις.

Για λόγους απλότητας παρουσιάζεται η συνάρτηση σφάλματος του προβλήματος δύο κλάσεων, όπου $y = 1$ όταν $g = 1$ και $y = -1$ όταν $g = 2$. Ακόμα συμβολίζεται $z_1 = w_{l0} + w_l^T x$. Τότε οι παραπάνω τύποι γίνονται:

$$Pr(Y=1 \mid Z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} = \sigma(z)$$

$$Pr(Y=-1 \mid Z) = \frac{1}{1+e^z} = \sigma(-z)$$

και άρα:

$$Pr(Y \mid Z) = \sigma(yz)$$

Τα μοντέλα της λογιστικής παλινδρόμησης συνήθως εκπαιδεύονται χρησιμοποιώντας ως συνάρτηση κόστους την αρνητική της μέγιστης πιθανοφάνειας. Αυτό προκύπτει από το γεγονός ότι κατά την εκπαίδευση πρέπει να μεγιστοποιείται η πιθανότητα ενώ μειώνεται η συνάρτηση κόστους. Αρχικά παρουσιάζεται η συνάρτηση κόστους για το πρόβλημα πολλών κλάσεων, και στη συνέχεια για των δύο.

$$L(\theta) = -\sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)$$

$$L(\theta) = -\sum_{i=1}^{N} \{y_i \log p(y_i \mid z) + (1-y_i) \log(1-p(y_i \mid z))\}$$

$$= -\sum_{i=1}^{N}\{y_i \log(\frac{1}{1+e^{-y_i z}}) + (1-y_i)\log(\frac{1}{1+e^{-y_i z}})\}$$
$$= \sum_{i=1}^{N}\log(1+e^{-y_i z}) = \sum_{i=1}^{N}\log(1+e^{-y_i(w_{l0}+w_l^T x)})$$

Κανονικοποίηση

Πριν τη διαδικασία της εκπαίδευσης στα δεδομένα, αντί να βελτιστοποιηθεί η παραπάνω συνάρτηση, μπορεί να προστεθεί ένας παραπάνω περιορισμός στο πόσο μπορούν να μεγαλώσουν οι συντελεστές, ώστε να μην υπάρξει υπερεκπαίδευση. Αυτή η διαδικασία λέγεται κανονικοποίηση (Regularization) και διακρίνουμε τρεις διαφορετικές περιπτώσεις ανάλογα με τη συνάρτηση που χρησιμοποιείται, οι οποίες παρουσιάζονται παρακάτω.

- L1(Lasso): Αυτή η κανονικοποίηση έχει την ιδιότητα της επιλογής χαρακτηριστικών, μετατρέποντας συντελεστές σε 0 και ελαχιστοποιείται η συνάρτηση κόστους:

$$\min_{w,c} \|w\|_1 + C\sum_{i=1}^{N} log(1+e^{-y_i(w^T x_i+w_0)})$$

- L2(Ridge): Αυτή η κανονικοποίηση τιμωρεί πολύ μεγάλους συντελεστές, αλλά δεν τους μηδενίζει και ελαχιστοποιείται η συνάρτηση κόστους:

$$\min_{w,c} \frac{1}{2}w^T w + C\sum_{i=1}^{N} log(1+e^{-y_i(w^T x_i+w_0)})$$

- Elastic-Net: Πρόκειται για έναν συνδυασμό των παραπάνω δύο τεχνικών, και η βαρύτητα που δίνεται σε κάθε μία ρυθμίζεται με το ρ, ελαχιστοποιείται η συνάρτηση κόστους:

$$\min_{w,c} \frac{1-\rho}{2}w^T w + \rho\|w\|_1 + C\sum_{i=1}^{N} log(1+e^{-y_i(w^T x_i+w_0)})$$

Τεχνικές μαθηματικής βελτιστοποίησης

Για την ελαχιστοποίηση της συνάρτησης κόστους χρησιμοποιήθηκαν τέσσερις διαφορετικές τεχνικές μαθηματικής βελτιστοποίησης.

Κατάβαση συντεταγμένων (Coordinate Descent), πρόκειται για την πιο απλή λύση στο πρόβλημα της βελτιστοποίησης. Η γενική ιδέα του αλγορίθμου είναι πως σε κάθε επανάληψη επιλέγεται μία συντεταγμένη μέσω ενός κανόνα επιλογής συντεταγμένων, και η συνάρτηση ελαχιστοποιείται πάνω σε αυτή, κρατώντας σε σταθερή τιμή τις υπόλοιπες συντεταγμένες. Για την επιλογή του βήματος με το οποίο ελαχιστοποιείται η συντεταγμένη αρκεί μία γραμμική αναζήτηση στις τιμές της.

Μαθηματικά το πρόβλημα διατυπώνεται ως: έστω μία συνάρτηση $F(x)$, εδώ $L(\theta)$, με αρχικές τιμές $x^0 = (x_1^0, x_2^0, ..., x_n^0)$. Για την ελαχιστοποίηση της, στον γύρο $k+1$ του αλγορίθμου για την συντεταγμένη i, το x^{k+1} ορίζετε μέσω του x^k λύνοντας το πρόβλημα:

$$x_i^{k+1} = \arg\min_{yin\mathbb{R}} f(x_1^{k+1}, ..., x_{i-1}^{k+1}, y, x_{i+1}^{k}, ..., x_i^n)$$

Το οποίο επαναλαμβάνεται για κάθε μεταβλητή, και έτσι σε κάθε επανάληψη ισχύει:

$$F(x^0) \geq F(x^1) \geq F(x^2) \geq ...$$

Ο αλγόριθμος παίρνει τη μορφή:

Algorithm 5 Κατάβαση συντεταγμένων

1. **Διάλεξε** τα αρχικά $x^0 = (x_1^0, x_2^0, ..., x_n^0)$
2. **Μέχρι να υπάρξει σύγκλιση** ή μέχρι ένα μέγιστο αριθμό από επαναλήψεις:
   (α) **Διάλεξε** έναν δείκτη συντεταγμένης.
   (β) **Διάλεξε** ένα βήμα α.
   (ς) **Ενημέρωσε** το x_i ως $x_i - \alpha \frac{\partial F}{\partial x_i}(x)$.

Εδώ πρέπει να σημειωθεί ότι η ποσότητα που μειώνεται κάθε συντεταγμένη είναι ανάλογη της κλίσης, καθώς μεγαλύτερη κλίση σημαίνει πως το σημείο βρίσκεται πιο μακριά από το σημείο που ελαχιστοποιείται η συνάρτηση σφάλματος. Τέλος, η μέθοδος αυτή υστερεί σε δύο σημεία. Το πρώτο είναι πως αν η συνάρτηση προς εξέταση δεν είναι λεία, τότε ο αλγόριθμος μπορεί να κολλήσει σε κάποιο μη στατικό σημείο της συνάρτησης, και το δεύτερο είναι πως είναι δύσκολη στην παραλληλοποίηση, δεδομένου ότι η ελαχιστοποίηση γίνεται για μία συντεταγμένη τη φορά. Οι μέθοδοι που αναφέρονται παρακάτω έχουν παρόμοια δομή αλγορίθμου, με διαφορές στην ανανέωση των συντεταγμένων, και στο βήμα.

Η μέθοδος του Νεωτον για βελτιστοποίηση χρησιμεύει στην εύρεση ριζών, μίας διαφορίσιμης συνάρτησης F, δηλαδή τα σημεία όπου $F(x) = 0$. Εφαρμόζοντας τη στην παράγωγο μίας συνάρτησης που είναι δύο φορές διαφορίσιμη, μπορούν να βρεθούν τα ακρότατα της.

Χρησιμοποιώντας την ίδια σημειολογία με την προηγούμενη μέθοδο, αυτή προσεγγίζει τη συνάρτηση προς ελαχιστοποίηση με ένα πολυώνυμο Taylor δευτέρου βαθμού κοντά στο αρχικό σημείο x^0.

$$f(x^0 + t) \approx f(x^0) + \nabla f(x^0)t + \frac{1}{2}\nabla^2 f(x^0)t^2$$

$$= f(x^0) + \nabla f(x^0)t + \frac{1}{2}H_f(x^0)t^2 = f(x^0) + f'(x^0)t + \frac{1}{2}f''(x^0)t^2$$

Το ελάχιστο βρίσκεται ως:

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}(f(x^0) + f'(x^0)t + \frac{1}{2}f''(x^0)t^2) = f'(x^0) + f''(x^0)t$$

$$t = -[f''(x^0)]^{-1}f'(x^0)$$

Όπου ∇, H_f ο πίνακας πρώτης παραγώγου και ο Χεσιανός πίνακας αντίστοιχα.

Σε κάθε επόμενη επανάληψη του αλγορίθμου, οι μεταβλητές μεταβάλλονται ως:

$$x^{k+1} = x^k - \gamma[f''(x_k)]^{-1}f'(x_k)$$

Όπου $0 < \gamma \leq 1$ είναι ένα μικρό βήμα.

Όπως και πριν, και αυτή η μέθοδος έχει κάποιες αδυναμίες. Η πρώτη είναι πως αν ο Χεσιανός πίνακας δεν είναι αντιστρέψιμος, η μέθοδος δεν μπορεί να δουλέψει. Στη συνέχεια, η μέθοδος μπορεί να μην συγκλίνει ποτέ, αλλά μπορεί να ακολουθεί έναν κύκλο σημείων. Τέλος μπορεί να συγκλίνει σαγματικο σημείο, αντί τοπικού ελάχιστου.

Για να γίνει κατανοητή η μέθοδος της στοχαστικής κατάβασης μέσης παραγώγου (Stochastic Average Gradient descent), περιγράφεται συνοπτικά η απλή μέθοδος της κατάβασης παραγώγου (Gradient descent). Θεωρώντας πως μία συνάρτηση $F(x)$ πολλών μεταβλητών είναι διαφορίσιμη κοντά σε ένα σημείο α, τότε η $F(x)$ μειώνεται με το μεγαλύτερο ρυθμό, αν κάποιος μετακινηθεί από α στη διεύθυνση της αρνητικής παραγώγου στο α. Ομοίως με προηγουμένως, κάθε βήμα του αλγορίθμου ορίζεται ως:

$$x^{k+1} = x^k - \gamma_k \nabla F(x^k)$$

όπου αν το γ_k είναι αρκετά μικρό, τότε:

$$F(x^0) \geq F(x^1) \geq F(x^2) \geq ...$$

Μία τιμή του γ_k που προτείνεται ώστε ο αλγόριθμος να συγκλίνει είναι:

$$\gamma_k = \frac{\left|(x^k - x^{k-1})^T [\nabla F(x^k) - \nabla F(x^{k-1})]\right|}{\|\nabla F(x^k) - \nabla F(x^{k-1})\|^2}$$

Σύμφωνα με τα παραπάνω η μέθοδος της στοχαστικής κατάβασης παραγώγου αποτελεί μία στοχαστική προσέγγιση της κατάβασης παραγώγου όπου αντικαθιστά την πραγματική παράγωγο, υπολογισμένη σε όλο το σύνολο δεδομένων, με μία προσέγγιση της, υπολογισμένη σε ένα υποσύνολο του συνόλου δεδομένων. Ενώ πριν για τον υπολογισμό της παραγώγου της συνάρτησης σφάλματος χρησιμοποιούνταν και τα N σημεία του συνόλου εκπαίδευσης, δηλαδή η $F(x) = \frac{1}{N}\sum_{i=1}^{N} F_i(x)$, και η ανανέωση του βάρους μπορούσε να γραφτεί και ως:

$$x^{k+1} = x^k - \gamma_k \nabla F(x^k) = x^k - \frac{\gamma_k}{N}\sum_{i=1}^{N} \nabla F_i(x^k)$$

Τώρα γίνεται:

$$x^{k+1} = x^k - \gamma_k \nabla F_i(x^k)$$

Όπου F_i είναι η συνάρτηση, στο ioστό τυχαίο στοιχείο του συνόλου εκπαίδευσης.

Επέκταση της στοχαστικής κατάβασης παραγώγου είναι η στοχαστική κατάβαση μέσης παραγώγου, όπου σε κάθε βήμα πάλι επιλέγεται ένα τυχαίο i και υπολογίζεται η παράγωγος σε αυτό, αλλά εδώ λαμβάνονται υπόψη και όλες οι τιμές της παραγώγου που έχουν ήδη υπολογιστεί. Πιο συγκεκριμένα αρχικοποιούμε τα βάρη όπως στην κατάβαση συντεταγμένων, και θέτουμε $g_i^0 = x^0, i = 1, 2, ..., N$, στη συνέχεια σε κάθε βήμα ακολουθεί ο υπολογισμός της παραγώγου της συνάρτησης για ένα τυχαίο i, και οι τιμές των υπόλοιπων F_i παραμένουν ίδιες. Η ανανέωση γίνεται ως:

$$x^{k+1} = x^k - \frac{\gamma_k}{N}\sum_{i=1}^{N} \nabla F_i(x^k)$$

Παρατίθεται και η μέθοδος SAGA παραλλαγή της προηγούμενης, όπου η ανανέωση γίνεται ως:

$$x^{k+1} = x^k - \gamma_k(\nabla F_i^k - \nabla F_i^{k-1} + \frac{1}{N}\sum_{i=1}^{N} \nabla F_i(x^k))$$

Και ο αλγόριθμος Broyden–Fletcher–Goldfarb–Shanno επιλέγει την κατεύθυνση της κατάβασης μετασχηματίζοντας την παράγωγο, με πληροφορίες της καμπυλότητας. Σε αντίθεση με τη μέθοδο του Newton δεν απαιτείται η αντιστροφή του Χεσιανού πίνακα. Η δομή του αλγορίθμου διαφέρει ελαφρώς από τους προηγούμενους, και παρουσιάζεται:

Algorithm 6 Broyden–Fletcher–Goldfarb–Shanno

1. **Διάλεξε** τα αρχικά $x^0 = (x_1^0, x_2^0, ..., x_n^0)$ και μία προσέγγιση του Χεσιανού πίνακα B_0
2. **Μέχρι να υπάρξει σύγκλιση** ή μέχρι ένα μέγιστο αριθμό από επαναλήψεις:
   (α) **Διάλεξε** μία διεύθυνση p_k λύνοντας $B_k p_k = -\nabla F(x^k)$.
   (β) **Διάλεξε** ένα βήμα $\alpha_k = \arg\min_\alpha F(x^k + \alpha p_k)$.
   (ς) **Θέσε** $s_k = a_k p_k$ και ενημέρωσε $x^{k+1} = x^k + s_k$

$$y_k = \nabla F(x^{k+1}) - \nabla F(x^k)$$

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k}$$

3.2.3 Μηχανές Διανυσμάτων Υποστήριξης

Μηχανές Διανυσμάτων Υποστήριξης

Έστω ότι κάποια γνωστά σημεία δεδομένων ανήκουν σε δύο γραμμικά διαχειρίσιμες κλάσεις, και σκοπός είναι η ταξινόμηση κάθε νέου σημείου Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) αντιμετωπίζουν το πρόβλημα, βλέποντας κάθε σημείο σαν ένα διάνυσμα p διαστάσεων, και χωρίζοντας αυτά τα σημεία με ένα $(p-1)$ διάστατο υπερεπίπεδο. Σκοπός τους είναι να βρουν αυτό το υπερεπίπεδο που μεγιστοποιεί την απόσταση αυτού από το κοντινότερο σημείο κάθε κλάσης και κατ΄ επέκταση δημιουργεί τη μεγαλύτερη διαμέριση μεταξύ των δύο κλάσεων.

Το παραπάνω πρόβλημα γενικεύεται και σε περισσότερες κλάσεις, οι οποίες μπορεί να μην είναι γραμμικά διαχωρίσιμες μεταξύ τους. Για τη λύση του προβλήματος αυτού, όπως και σε πολλά άλλα γραμμικά μοντέλα, προτείνεται η χρήση συναρτήσεων πυρήνα (kernel functions), οι οποίες μεταφέρουν την αρχική πληροφορία σε έναν νέο υπερχώρο, στον οποίο οι κλάσεις είναι γραμμικά διαχωρίσιμες. Όπως και όλες οι προηγούμενες μέθοδοι, έτσι και αυτή ανήκει στην κατηγορία μοντέλων της επιβλεπόμενης μάθησης.

Για τη μαθηματική πλαισίωση του προβλήματος, έστω ότι έχουμε δεδομένα εκπαίδευσης της μορφής $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ όπου για τη απλοποίηση του προβλήματος $y_i \in \{-1, 1\}$.

Τα y_i είναι οι ετικέτες των σημείων και δηλώνουν σε πια κλάση ανήκουν τα p-διάστατα x_i. Όπως αναφέρθηκε πριν, τη λύση στο πρόβλημα δίνει το υπερεπίπεδο το οποίο μεγιστοποιεί την απόσταση από τα κοντινότερα σημεία x_i κάθε κλάσης, δηλαδή το υπερεπίπεδο μέγιστου περιθωρίου.

Κάθε υπερεπίπεδο γράφεται ως, $w^T x - b = 0$, όπου w κάθετο διάνυσμα προς το υπερεπίπεδο. Η παράμετρος $\frac{b}{\|w\|}$ καθορίζει την απόσταση του υπερεπιπέδου, από την αρχή των αξόνων στη διεύθυνση του w. Τώρα διακρίνουμε δύο υποπεριπτώσεις.

Η πρώτη αφορά δεδομένα τα οποία είναι γραμμικά διαχωρίσιμα, και λέγεται σκληρού περιθωρίου. Σε αυτήν υπάρχουν δύο παράλληλα υπερεπίπεδα τα οποία διαχωρίζουν τα δεδομένα των δύο κλάσεων, ώστε η απόσταση μεταξύ τους να είναι το δυνατόν μεγαλύτερη. Το υπερεπίπεδο μέγιστου περιθωρίου βρίσκεται στη μέση της απόστασης αυτών.

Τα υπερεπίπεδα που αναφέρθηκαν παραπάνω περιγράφονται ως $w^T x - b = 1$, όπου κάθε σημείο πάνω από αυτό το όριο ανήκει στην κλάση 1, και ως $w^T x - b = -1$, όπου κάθε σημείο κάτω από αυτό το όριο ανήκει στην κλάση -1. Αφού η απόσταση μεταξύ τους είναι $\frac{2}{\|w\|}$ για τη μεγιστοποίηση της πρέπει να ελαχιστοποιηθεί το $\|w\|$. Ακόμα για τα δεδομένα ισχύει, $w^T x_i - b \leq -1$ ή $w^T x_i - b \geq 1$ για κάθε $i = 1, 2, ..., N$, το οποίο γράφεται και ως $y_i(w^T x_i - b) \geq 1$.

Το πρόβλημα μετατρέπεται στην ελαχιστοποίηση του $\|w\|$, τηρώντας $y_i(w^T x_i - b) \geq 1$. Ο τρόπος που γίνεται η ταξινόμηση στη συνέχεια είναι $x \mapsto sgn(w^T x_i - b)$. Από τα παραπάνω προκύπτει πως το υπερεπίπεδο μέγιστου περιθωρίου εξαρτάται μόνο από τα x_i τα οποία είναι πιο κοντά σε αυτό, και ονομάζονται διανύσματα υποστήριξης.

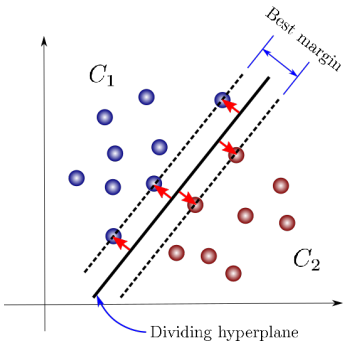
Η δεύτερη περίπτωση ονομάζεται μαλακού περιθωρίου καθώς εδώ οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες. Εδώ τη λύση δίνει η συνδεδεμένη συνάρτηση σφάλματος (hinge loss), $\max(0, 1 - y_i(w^T x_i - b))$. Αυτή η συνάρτηση γίνεται 0 όταν ισχύει η συνθήκη $y_i(w^T x_i - b) \geq 1$, δηλαδή το x_i βρίσκεται εκτός περιθωρίου. Για δεδομένα εντός του περιθωρίου το σφάλμα είναι ανάλογο της απόστασης από τα άκρα του. Εδώ το πρόβλημα μετατρέπεται στην ελαχιστοποίηση του:

$$[\frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - y_i(w^T x_i - b))] + \lambda \|w\|^2$$

με το λ να είναι η ανταλλαγή μεταξύ του περιθωρίου, και της βεβαίωσης ότι τα x_i βρίσκονται έξω από αυτό.

Τα χαρακτηριστικά που κάνουν τις μηχανές διανυσμάτων υποστήριξης να ξεχωρίζουν είναι:

- Η αποτελεσματικότητα τους σε πολυδιάστατους χώρους, ακόμα και όταν ο αριθμός των διαστάσεων είναι μεγαλύτερος από το πλήθος των δειγμάτων.
- Η χρήση ενός υποσυνόλου των σημείων εκπαίδευσης, χρησιμοποιούνται μόνο τα σημεία - διανύσματα υποστήριξης.
- Η απουσία επιρροής από τοπικά ελάχιστα.
- Η ευελιξία τους, υπάρχουν πολλές συναρτήσεις πυρήνα οι οποίες προσαρτώνται στη συνάρτηση επιλογής.



Σχήμα 3.2: Μηχανές Διανυσμάτων Υποστήριξης.

Συναρτήσεις πυρήνα

Η ταξινόμηση που αναφέρθηκε παραπάνω αφορά γραμμικά όρια. Όμως όπως αναφέρθηκε προηγουμένως οι μηχανές διανυσμάτων υποστήριξης υποστηρίζουν και μη γραμμικά. Αυτό επιτυγχάνεται επεκτείνοντας το χώρο των διαστάσεων χρησιμοποιώντας μη γραμμικές επεκτάσεις.

Γενικά τα γραμμικά όρια σε έναν επεκτεταμένο χώρο διαστάσεων, βοηθούν στον καλύτερο διαχωρισμό των κλάσεων και μεταφράζονται ως μη γραμμικά στον πραγματικό χώρο. Θεωρητικά το πλήθος των διαστάσεων στο νέο υπερχώρο είναι πολύ μεγάλο, ακόμα και άπειρο. Εδώ όμως πρέπει να σημειωθεί πως δουλεύοντας σε έναν τέτοιο χώρο ελλοχεύει ο κίνδυνος της υπερεκπαίδευσης.

Ο μηχανισμός με τον οποίο οι μηχανές διανυσμάτων υποστήριξης διατηρούν τον υπολογιστικό φόρτο στις μεγαλύτερες διαστατικότητες σε λογικά πλαίσια, είναι πως το εσωτερικό γινόμενο σε ζευγάρια διανυσμάτων εισόδου θα πρέπει να γίνεται εύκολα, ορίζοντας τα με βάση συναρτήσεις πυρήνα που εξαρτώνται από το πρόβλημα. Θα πρέπει να ισχύει

$$\sum_i \alpha_i k(x_i, x) = constant$$

Η συνάρτησης πυρήνα πραγματοποιεί εσωτερικά και το εσωτερικό γινόμενο των διανυσμάτων. Τέτοιες συναρτήσεις είναι:

- Γραμμική:
$$k(x_i, x_j) = (x_i \cdot x_j)$$
- Πολυωνυμική d βαθμού:
$$k(x_i, x_j) = (x_i \cdot x_j + \theta)^d$$
- Γκαουσιανή ακτινικής βάσης:
$$k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$
- Σιγμοειδής:
$$k(x_i, x_j) = \tanh(kx_i \cdot x_j + c)$$

3.3 Τεχνικές προεπεξεργασίας δεδομένων

Σε αυτή την ενότητα πλαισιώνεται το θεωρητικό υπόβαθρο για όλες τις τεχνικές που χρησιμοποιήθηκαν για την προεπεξεργασία των δεδομένων. Μέσα σε αυτές περιλαμβάνονται:

- Τεχνικές επιλογής χαρακτηριστικών: Στατιστικό Kolmogorov–Smirnov, Ανάλυση Διακύμανσης με τιμές F, επιλογή χαρακτηριστικών μέσω μοντέλων, Γενετικοί Αλγόριθμοι.
- Τεχνικές μείωσης της διαστατικότητας: Ανάλυση κύριων συνιστωσών.
- Δειγματοληψία: Υποδειγματοληψία, Υπερδειγματοληψία.
- Τεχνικές κλιμάκωσης: Κανονική κλιμάκωση, Κλιμάκωση ελαχίστου μεγίστου.
- Δυναμική χρονική στρέβλωση.
- Εντοπισμός σακκάδων και φάσεων εστίασης.

3.3.1 Στατιστική Επιλογή Χαρακτηριστικών

Για την διαδικασία επιλογής των χαρακτηριστικών χρησιμοποιήθηκαν κάποια στατιστικά τεστ, τα οποία εξετάζουν την ομοιότητα της κατανομής των χαρακτηριστικών σε σχέση με τις κλάσεις που ανήκει κάθε δείγμα και επιστρέφουν μία τιμή η οποία υποδεικνύει το πόσο μοιάζουν. Τα τεστ αυτά είναι:

- Στατιστικό Kolmogorov–Smirnov: Χρησιμοποιείται για να εξετάσει αν δύο δείγματα ανήκουν στην ίδια κατανομή. Η τιμή του είναι μία ποσοτικοποίηση της απόστασης μεταξύ των αθροιστικών συναρτήσεων κατανομής των δύο δειγμάτων. Ο τύπος του δείκτη είναι:
$$D = \max_x \mid F(x) - G(x) \mid$$
Όπου $F(x), G(x)$ οι αθροιστικές συναρτήσεις κατανομής των δύο δειγμάτων.
- Ανάλυση Διακύμανσης (ANOVA) με τιμές F: Με την ανάλυση διακύμανσης εξετάζεται αν υπάρχουν διαφορές στις μέσες τιμές δύο ή περισσότερων πληθυσμών. Αν η τιμή F είναι μικρότερη ή κοντά στο 1, τότε οι πληθυσμοί προέρχονται από κανονικές κατανομές με ίδιο μέσο όρο. Διαφορετικά τα δείγματα προέρχονται από πληθυσμούς με διαφορετικούς μέσους όρους. Η τιμή F είναι:
$$F = \frac{s_b^2}{s_w^2}$$
$$s_b^2 = \frac{\sum_{j=1}^{m} n_j (\bar{x}_j - \bar{X})^2}{m-1}$$
$$s_w^2 = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n-m}$$
Όπου m είναι το πλήθος των πληθυσμών, n το συνολικό πλήθος των δειγμάτων, n_j το συνολικό πλήθος των δειγμάτων του πληθυσμού j, $\bar{x}_j$ ο μέσος των δειγμάτων του πληθυσμού j και $\bar{X}$ ο ολικός μέσος.

3.3.2 Επιλογή χαρακτηριστικών μέσω μοντέλων

Για την επιλογή χαρακτηριστικών χρησιμοποιήθηκαν και δύο μέθοδοι, οι οποίες εκμεταλλεύονται τα βάρη των γραμμικών μοντέλων ή μοντέλων που παρέχουν κάποια πληροφορία για την χρησιμότητα των χαρακτηριστικών στην ταξινόμηση.

Η πρώτη είναι η απλή επιλογή των k-καλύτερων χαρακτηριστικών, μέσω της πληροφορίας που παρέχει το μοντέλο.

Η δεύτερη είναι η αναδρομική επιλογή χαρακτηριστικών, κατά την οποία επιλέγεται ένα βήμα μείωσης των χαρακτηριστικών μέχρι να μείνουν τα k-καλύτερα. Σε αυτή το μοντέλο αρχικά εκπαιδεύεται σε όλα τα χαρακτηριστικά, και παράγει την πληροφορία που αναφέρθηκε παραπάνω για το καθένα. Στη συνέχεια αφαιρεί τα χειρότερα σύμφωνα με το βήμα που δόθηκε και επαναλαμβάνει μέχρι να μείνουν k. Σε αυτή τη μέθοδο δίνεται και μία επέκταση, όπου μέσω μίας διαδικασίας διασταυρωμένης επικύρωσης επιλέγεται αυτόματα η υπερπαράμετρος k. Ο τρόπος με τον οποίο τα μοντέλα παρέχουν πληροφορία για την σημαντικότητα των χαρακτηριστικών περιγράφεται παρακάτω.

- Λογιστική παλινδρόμηση: Τα βάρη των χαρακτηριστικών στη συνάρτηση απόφασης όπως περιγράφηκε στην αντίστοιχη ενότητα.

- Τυχαία δάση και δέντρα ενίσχυσης με κλίση: Η σημαντικότητα των χαρακτηριστικών εκφράζεται μέσω της συνάρτησης σφάλματος των δέντρων, για παράδειγμα μέσω του δείκτη Gini.

- Μηχανές διανυσμάτων υποστήριξης: Τα βάρη των χαρακτηριστικών στη συνάρτηση απόφασης όπως περιγράφηκε στην αντίστοιχη ενότητα. Σε αυτή τη μέθοδο ο προσδιορισμός της σημαντικότητας είναι εφικτός μόνο στην περίπτωση του γραμμικού πυρήνα.

3.3.3 Γενετικοί Αλγόριθμοι

Εισαγωγή

Οι γενετικοί αλγόριθμοι αποτελούν μία ευριστική εξερεύνηση του χώρου τιμών ενός προβλήματος. Είναι εμπνευσμένοι από την Δαρβινική θεωρία της φυσικής επιλογής. Η διαδικασία της φυσικής επιλογής ξεκινά επιλέγοντας τα καταλληλότερα άτομα από έναν πληθυσμό. Αυτά παράγουν απογόνους, που διατηρούν τα χαρακτηριστικά των γονέων, και εντάσσονται στην επόμενη γενιά. Οι απόγονοι γονέων που έχουν καλύτερα χαρακτηριστικά, αναμένεται να έχουν καλύτερα χαρακτηριστικά από απογόνους ατόμων με χειρότερα και έτσι θα έχουν μεγαλύτερη πιθανότητα να επιβιώσουν και να μεταβιβάσουν τα χαρακτηριστικά τους. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να βρεθούν άτομα με τα καλύτερα χαρακτηριστικά. Η φυσική επιλογή όπως περιγράφηκε αντικατοπτρίζεται στον αλγόριθμο, μαζί με άλλους παράγοντες όπως ο ελιτισμός, οι μεταλλάξεις, και η διασταύρωση. Κάθε στάδιό του θα αναλυθεί ξεχωριστά.

Algorithm 7 Γενετικός Αλγόριθμος

1. **Αρχικοποίησε** έναν τυχαίο αρχικό πληθυσμό.
2. **Υπολόγισε** την ικανότητα (fitness) κάθε ατόμου.
3. **Μέχρι** τη σύγκλιση ή μετά από έναν αριθμό από γενιές:
   (α) **Διάλεξε** τα άτομα που θα αναπαραχθούν.
   (β) **Υλοποίησε** τη διασταύρωση (crossover).
   (ς) **Εφάρμοσε** τη μετάλλαξη (mutation).
   (δ) **Υπολόγισε** την ικανότητα (fitness) κάθε νέου ατόμου.
4. **Επέστρεψε** τα καλύτερα άτομα του τελευταίου πληθυσμού.

Αρχικοποίηση

Το πρώτο στάδιο του αλγορίθμου. Σε αυτό δημιουργείται από τον υπολογιστή σύμφωνα με μία παράμετρο μεγέθους πληθυσμού, ένα σύνολο από γονιδιώματα, όπου κάθε ένα από αυτά αντιστοιχεί σε ένα άτομο του πληθυσμού.

Το κάθε γονιδίωμα αναπαρίσταται από ένα διάνυσμα, όπου κάθε συνιστώσα του παίρνει τιμές στον χώρο εξερεύνησης που έχει οριστεί. Το μέγεθος του γονιδιώματος είναι αυτό που χρειάζεται η συνάρτηση ικανότητας, η οποία περιγράφεται αναλυτικά πιο κάτω, και κάθε επιμέρους μεταβλητή του ονομάζεται γονίδιο.

Η αρχικοποίηση του πληθυσμού γίνεται τις περισσότερες φορές με τυχαίο τρόπο από τον υπολογιστή και προσδιορίζει το μεγαλύτερο μέρος του χώρου που θα διερευνηθεί. Συνήθως η αρχικοποίηση δεν είναι εντελώς τυχαία, αλλά γίνεται με τυχαίο τρόπο, στη γειτονιά των τιμών που αναμένεται η βέλτιστη λύση.

Επιλογή ατόμων προς αναπαραγωγή

Σε αυτό το στάδιο, από την εκάστοτε γενιά, επιλέγονται τα άτομα που θα αναπαραχθούν για να παράγουν την επόμενη γενιά. Η ιδέα αυτή βασίζεται στο γεγονός ότι άτομα με καλύτερα χαρακτηριστικά θα παράγουν καλύτερους απογόνους. Η επιλογή των ατόμων γίνεται μέσω της συνάρτησης ικανότητας η οποία προσδιορίζει το κατά πόσο τα χαρακτηριστικά ενός γονιδιώματος είναι καλύτερα από ένα άλλο. Τα άτομα προς αναπαραγωγή μπορεί να είναι τυχαία, με μεγαλύτερο βάρος όμως σε αυτά με καλύτερα χαρακτηριστικά ή μπορεί και να είναι απλά αυτά με τα καλύτερα χαρακτηριστικά. Κάποιες μέθοδοι εφαρμόζουν την παραπάνω διαδικασία σε ένα τυχαίο δείγμα του πληθυσμού κάθε γενιάς, καθώς ο υπολογισμός της καταλληλότητας όλου του πληθυσμού μπορεί να είναι πολύ υπολογιστικά χρονοβόρος.

Η συνάρτηση ικανότητας όπως αναφέρθηκε και προηγουμένως, προσδιορίζει το κατά πόσο ένα γονιδίωμα έχει θεμιτά χαρακτηριστικά. Έτσι είναι άμεσα συνδεδεμένη με τη φύση του προ-

βλήματος προς επίλυση. Θα πρέπει να είναι ικανή να διαχειρίζεται και γονιδιώματα τα οποία δεν είναι έγκυρα για τους περιορισμούς του προβλήματος, θέτοντας την τιμή τους στην ελάχιστη ή μέγιστη τιμή, αν πρόκειται για πρόβλημα μεγιστοποίησης ή ελαχιστοποίησης αντίστοιχα. Την τιμή αυτής της συνάρτησης προσπαθεί να βελτιστοποιήσει ο γενετικός αλγόριθμος. Σε περιπτώσεις που δεν είναι εφικτή η αποτύπωση της συνάρτησης σε μία έκφραση, για τον υπολογισμό της μπορεί να γίνει μία προσομοίωση του προβλήματος για κάθε γονιδίωμα προς εξέταση. Εδώ φαίνεται το πόσο υπολογιστικά βαριά μπορεί να γίνει η αποτίμηση της ικανότητας ολόκληρου του πληθυσμού.

Σε αυτό το στάδιο μπορεί να ενταχθεί και ο ελιτισμός ως υπερ-παράμετρος. Ο ελιτισμός προσδιορίζει το αν και πόσα άτομα που έχουν τα καλύτερα χαρακτηριστικά, θα περάσουν αυτούσια από μία γενιά στην επόμενη. Ανακεφαλαιώνοντας, το στάδιο αυτό είναι αρμόδιο για την βελτιστοποίηση του προβλήματος με βάση το σύνολο τιμών που ορίζει ο αρχικός πληθυσμός. Ο ελιτισμός αν ενταχθεί, είναι η παράμετρος που εγγυάται ότι η ποιότητα των καλύτερων αποτελεσμάτων δεν θα φθίνει από γενιά σε γενιά. Πρέπει όμως να σημειωθεί, πως τα άτομα που περνούν αυτούσια, προσθέτουν μία προκατάληψη στα αποτελέσματα και μπορεί να μην επιτρέψουν την εξερεύνηση μεγαλύτερου χώρου τιμών για δεδομένο μέγεθος πληθυσμού.

Γενετικοί τελεστές

Πρόκειται για τη διασταύρωση και τη μετάλλαξη. Η εφαρμογή γενετικών τελεστών στον πληθυσμό που επιλέχθηκε προς αναπαραγωγή, είναι ο τρόπος με τον οποίο δημιουργείται η νέα γενιά και διαμορφώνει τον τρόπο με τον οποίο εξερευνάται ο χώρος τιμών. Για τη δημιουργία ενός ή περισσότερων παιδιών επιλέγονται δύο ή περισσότεροι γονείς από τον πληθυσμό προς διασταύρωση. Τα γονιδιώματα αυτών θα υποστούν διασταύρωση και μετάλλαξη, δημιουργώντας τα νέα άτομα, με χαρακτηριστικά των γονέων. Οι τελεστές λέγονται γενετικοί καθώς εφαρμόζονται στα γονιδιώματα.

Τα άτομα του πληθυσμού που προέρχονται από ελιτισμό δεν υποβάλλονται σε αυτούς τους μετασχηματισμούς, εκτός αν έχουν επιλεγεί και για αναπαραγωγή. Σε αυτή την περίπτωση αυτά περνούν αυτούσια, αλλά περνούν και οι απόγονοι αυτών.

Διασταύρωση

Διασταύρωση ή ανασυνδυασμός, είναι ο γενετικός τελεστής που συνδυάζει τα γονιδιώματα των γονέων για να προκύψουν οι απόγονοι. Αποτελεί έναν στοχαστικό τρόπο για την παραγωγή νέων λύσεων στο πρόβλημα προς διερεύνηση και μιμείται την διασταύρωση που γίνεται στα πραγματικά γονιδιώματα. Είναι ο συντελεστής που προσδιορίζει τον τρόπο με τον οποίο γίνεται η διερεύνηση του χώρου τιμών που προέκυψαν από τους γονείς και κατ΄ επέκταση, από τον αρχικό πληθυσμό. Τέλος αν ο ρυθμός διασταύρωσης είναι πολύ μεγάλος, ο αλγόριθμος μπορεί να οδηγηθεί σε πρόωρη σύγκλιση.

Στους παραδοσιακούς γενετικούς αλγόριθμους τα γονιδιώματα προσομοιώνονται με πίνακες δυαδικών τιμών. Επειδή τα προβλήματα που κλήθηκε να λύσει η παρούσα εργασία έχουν

αυτή τη φύση, παρουσιάζονται διαφορετικοί τρόποι για διασταύρωση τέτοιων τιμών.

- Διασταύρωση ενός σημείου: Επιλέγεται τυχαία ένα σημείο, σημείο διασταύρωσης, στους πίνακες - γονιδιώματα των γονέων. Από εκείνο το σημείο και μετά οι δυαδικές τιμές στους υποπίνακες ανταλλάσσονται. Έτσι προκύπτουν δύο παιδιά, καθένα από τα οποία έχει μέχρι το σημείο διασταύρωσης το γονιδίωμα του ενός γονέα και έπειτα από εκείνο το σημείο, του άλλου.

- Διασταύρωση k σημείων: Επιλέγονται τυχαία k σημεία διασταύρωσης και ανά δύο σημεία γίνεται ανταλλαγή του κάθε υποπίνακα μεταξύ των γονέων. Όπως και πριν, τα δύο παιδιά που προκύπτουν έχουν τμήματα γονιδιωμάτων από κάθε γονέα.

- Ομοιόμορφη διασταύρωση: Για τη δημιουργία του γονιδιώματος ενός παιδιού κάθε bit επιλέγεται τυχαία από έναν γονέα με ίδια πιθανότητα. Η μέθοδος μπορεί να γενικευτεί και στην τυχαία επιλογή με άνιση πιθανότητα μεταξύ γονέων. Έτσι το παιδί καταλήγει να έχει περισσότερα χαρακτηριστικά από έναν γονέα.

Μετάλλαξη

Όπως αναφέρθηκε, η διασταύρωση καλύπτει τη διερεύνηση του χώρου τιμών που έχει προκύψει από τους γονείς. Έτσι λύσεις οι οποίες είναι βέλτιστες και βρίσκονται εκτός αυτού του χώρου μένουν ανεξερεύνητες. Λύση σε αυτό το πρόβλημα δίνει ο τελεστής της μετάλλαξης. Αυτός χρησιμοποιείται για τη διατήρηση γενετικής ποικιλομορφίας από μία γενιά στην επόμενη και είναι ανάλογος της βιολογικής μετάλλαξης.

Με τον τελεστή μετάλλαξης ο γενετικός αλγόριθμος είναι ικανός να διερευνήσει μεγαλύτερο μέρος του χώρου τιμών του προβλήματος. Η μετάλλαξη συμβαίνει με μία πιθανότητα η οποία δεν πρέπει να παίρνει πολύ μεγάλη τιμή, καθώς τότε η αναζήτηση θα γίνεται τυχαία. Αν η τιμή είναι πολύ μικρή, τότε ο αλγόριθμος οδηγείται σε μείωση της γενετικής ποικιλομορφίας. Εδώ για τη δυαδική αναπαράσταση γονιδίων υπάρχουν δύο τεχνικές.

- Μετάλλαξη bit σειράς: Κάθε bit στη σειρά ενός γονιδιώματος έχει μία πιθανότητα $\frac{1}{l}$ να αλλάξει, όπου l το μήκος του πίνακα γονιδιώματος

- Αναποδογύρισμα bit: Ο τελεστής μετάλλαξης αλλάζει από 0 σε 1 και από 1 σε 0 την τιμή των bit ολόκληρου του γονιδιώματος.

3.3.4 Ανάλυση Κύριων Συνιστωσών

Η ανάλυση κύριων συνιστωσών(Principal Component Analysis) αντιθέτως με τις μεθόδους που αναφέραμε προηγουμένως δεν εφαρμόζεται για την επιλογή χαρακτηριστικών. Είναι μία μέθοδος που εφαρμόζεται για τη μείωση της διαστατικότητας των δεδομένων. Σύμφωνα με αυτή οι νέες συνιστώσες (χαρακτηριστικά) δημιουργούνται ως γραμμικός συνδυασμός των αρχικών. Για τον υπολογισμό των νέων χαρακτηριστικών, μπορεί να χρησιμοποιηθεί η αποσύνθεση μοναδικής τιμής (singular value decomposition).

Πιο συγκεκριμένα, έστω ότι ο πίνακας X είναι ο πίνακας των χαρακτηριστικών, έστω ένας πίνακας βαρών των συνιστωσών W, και ο πίνακας των νέων συνιστωσών T. Ο μετασχηματισμός που θέλουμε να πετύχουμε είναι:

$$T = XW$$

Ο πίνακας X είναι μεγέθους nxp, όπου n ο αριθμός των παρατηρήσεων και p ο αριθμός των γνωρισμάτων. Αν θέλουμε να προβάλουμε τον X από p σε l διαστάσεις με $l \leq p$, τότε ο T είναι nxl και ο W θα πρέπει να είναι pxl. Οι στήλες του W είναι ιδιοδιανύσματα του $X^T X$ και ο W έχει μέγεθος pxp αλλά επιλέγονται απλά οι l πρώτες στήλες.

Για τον παραπάνω μετασχηματισμό με την αποσύνθεση μοναδικής τιμής, ο X εκφράζεται ως:

$$X = U\Sigma W^T$$

Ο Σ είναι ένας nxp ορθογώνιος διαγώνιος πίνακας θετικών τιμών και λέγεται οι μοναδικές τιμές του X, ο U είναι ένας nxn ορθογώνιος πίνακας, οι στήλες του οποίου είναι ορθογώνια μοναδιαία διανύσματα μεγέθους n και λέγεται οι αριστερές μοναδικές τιμές του X, ο W είναι ένας pxp ορθογώνιος πίνακας, οι στήλες του οποίου είναι ορθογώνια μοναδιαία διανύσματα μεγέθους p και λέγεται οι δεξιές μοναδικές τιμές του X. Για τους U και W ισχύει $U^T U = I, W^T W = I$. Και ο παραπάνω μετασχηματισμός είναι:

$$T = XW$$

$$= U\Sigma W^T W$$

$$= U\Sigma$$

Όπως και πριν και εδώ στον Σ επιλέγονται οι l πρώτες στήλες.

3.3.5 Δειγματοληψία

Τα μοντέλα μηχανικής μάθησης πολλές φορές αποτυγχάνουν ή δίνουν λανθασμένα καλή επίδοση, σε σύνολα δεδομένων ταξινόμησης τα οποία έχουν ανισόρροπες κατανομές κλάσεων. Αυτό συμβαίνει γιατί πολλοί αλγόριθμοι είναι σχεδιασμένοι ώστε να λειτουργούν σε δεδομένα όπου οι κλάσεις έχουν παρόμοιο αριθμό από παρατηρήσεις. Έτσι όταν μία κλάση έχει πολύ μικρό αριθμό δεδομένων, αυτά λανθασμένα αγνοούνται προκειμένου να βελτιστοποιηθεί η απόδοση.

Τη λύση σε αυτό το πρόβλημα δίνει η δειγματοληψία των δεδομένων, με την οποία το σύνολο δεδομένων εκπαίδευσης μετασχηματίζεται ώστε να υπάρχει ισορροπία μεταξύ των κλάσεων. Οι μέθοδοι δειγματοληψίας που εξετάζονται είναι η τυχαία υποδειγματοληψία και η τυχαία υπερδειγματοληψία.

- Τυχαία υποδειγματοληψία: Με τη μέθοδο αυτή επιλέγονται δείγματα, χωρίς αντικατάσταση, με τυχαίο τρόπο από την κλάση με τα περισσότερα δείγματα, μειώνοντας με αυτό τον τρόπο τα δείγματα της.

- Τυχαία υπερδειγματοληψία: Με τη μέθοδο αυτή γίνεται δειγματοληψία με αντικατάσταση με τυχαίο τρόπο, στην κλάση με τα περισσότερα δείγματα, αυξάνοντας με αυτό τον τρόπο τα δείγματα της.

3.3.6 Τεχνικές κλιμάκωσης

Οι αλγόριθμοι της μηχανικής μάθησης αντιμετωπίζουν τα σύνολα δεδομένων ως απλά νούμερα. Έτσι κάποια μοντέλα όταν βλέπουν χαρακτηριστικά με μεγάλες τιμές, οδηγούνται στην υπόθεση ότι αυτές οι συνιστώσες έχουν κάποια υπεροχή για το τελικό αποτέλεσμα. Έτσι αυτά τα χαρακτηριστικά αποκτούν, πολλές φορές λανθασμένα, μεγαλύτερη σημασία για την εκπαίδευση του μοντέλου. Ο μετασχηματισμός των δεδομένων με τον οποίο λύνεται αυτό το πρόβλημα ονομάζεται κλιμάκωση. Η κλιμάκωση των χαρακτηριστικών μπορεί να οδηγήσει και σε γρηγορότερη σύγκλιση των αλγορίθμων βελτιστοποίησης. Εδώ παρουσιάζονται δύο είδη.

- Κανονική κλιμάκωση: Με αυτόν τον τρόπο κλιμάκωσης οι τιμές x_i των χαρακτηριστικών μετατρέπονται ως,

$$x'_i = \frac{x_i - \bar{x}}{\sigma}$$

  Όπου $\bar{x}$ ο μέσος των τιμών του κάθε χαρακτηριστικού και σ η διασπορά τους. Ο μετασχηματισμός αυτός υπό την προϋπόθεση ότι τα χαρακτηριστικά ακολουθούν Γκαουσιανή κατανομή, τα μετασχηματίζει ώστε να ακολουθούν κανονική κατανομή.

- Κλιμάκωση ελαχίστου μεγίστου: Ομοίως με πριν οι τιμές μετατρέπονται ως,

$$x'_i = \frac{x_i - \min x}{\max x - \min x}$$

  Και τα χαρακτηριστικά μεταφέρονται στο [0,1].

3.3.7 Δυναμική χρονική στρέβλωση

Δυναμική χρονική στρέβλωση (Dynamic time warping) είναι ένας αλγόριθμος στην ανάλυση χρονοσειρών, ο οποίος χρησιμεύει για την μέτρηση της ομοιότητας δύο χρονοσειρών και υπολογίζει το βέλτιστο ταίριασμα μεταξύ αυτών σύμφωνα με κάποιους περιορισμούς και κανόνες.

- Κάθε δείκτης της πρώτης ακολουθίας πρέπει να αντιστοιχεί σε κάποιο δείκτη της δεύτερης, και ανάποδα.

- Ο πρώτος δείκτης της μίας ακολουθίας αντιστοιχεί στον πρώτο της άλλης.

- Ο τελευταίος δείκτης της μίας ακολουθίας αντιστοιχεί στον τελευταίο της άλλης.

- Το ζευγάρωμα των δεικτών από την μία ακολουθία στην άλλη πρέπει να είναι μονότονα αυξανόμενο.

Δεδομένων των παραπάνω κανόνων, βέλτιστο ταίριασμα των ακολουθιών θεωρείται αυτό που έχει το ελάχιστο κόστος. Το κόστος υπολογίζεται ως το άθροισμα των απόλυτων διαφορών των τιμών κάθε ζεύγους δεικτών. Ψευδοκώδικας για τον αλγόριθμο δίνεται παρακάτω.

Algorithm 8 Δυναμική χρονική στρέβλωση

1. X: Η πρώτη χρονοσειρά με $X = [x_1, ..., x_N]$
2. Y: Η δεύτερη χρονοσειρά με $Y = [y_1, ..., y_M]$
3. w: παράμετρος τοπικότητας $\mid N - M \mid \leq w$
4. **Θέσε** $w = \max(w, \mid N - M \mid)$
5. **Αρχικοποίησε** κάθε τιμή ενός πίνακα αποστάσεων D με τιμές $d_{i,j}, i = 1, .., N, j = 1, ..M$ στο άπειρο και βάλε $d_{0,0} = 0$.
6. **Για** i από 1 μέχρι N:
   (α) **Για** j από $\max(1, i - w)$ μέχρι $\max(M, i + w)$:
      i. **Θέσε** $d_{i,j} = 0$
7. **Για** i από 1 μέχρι N:
   (α) **Για** j από $\max(1, i - w)$ μέχρι $\max(M, i + w)$:
      i. **Θέσε** $L = \mid x_i - y_j \mid$
      ii. **Θέσε** $d_{i,j} = L + \min(d_{i-1,j}, d_{i,j-1}, d_{i-1,j-1})$
8. **Επέστρεψε** $d_{N,M}$

Στον παραπάνω αλγόριθμο το w ρυθμίζει τη μέγιστη επιτρεπτή απόσταση μεταξύ δύο δεικτών που εξετάζονται για αντιστοίχιση. Ακόμα στη σχέση $d_{i,j} = L + \min(d_{i-1,j}, d_{i,j-1}, d_{i-1,j-1})$ τα $d_{i-1,j}, d_{i,j-1}, d_{i-1,j-1}$ διαισθητικά εξηγούνται ως το κόστος του ταιριάσματος αν και ο προηγούμενός δείκτης i αντιστοιχεί στον j, αν ο i αντιστοιχεί στον j και στον προηγούμενό του και αν ο i αντιστοιχεί μόνο στον j.

3.3.8 Εύρεση περιοδικότητας μέσω του γραφήματος ισχύος του φάσματος

Η ισχύς ενός σήματος ορίζεται ως:

$$P = \lim_{T \to \infty} \frac{1}{T} \int_{t_0 - \frac{T}{2}}^{t_0 + \frac{T}{2}} |x(t)|^2 \, dt$$

Για ένα οποιοδήποτε t_0. Ο όρος $x(t)$ μπορεί να αντικατασταθεί από τον $\hat{x}(f)$ όπου είναι ο μετασχηματισμός Fourier του $x(t)$ για συχνότητα f. Ο μετασχηματισμός αυτός δίνεται ως:

$$\hat{x}(f) = \int_{-\infty}^{+\infty} e^{-i2\pi ft} x(t) dt$$

Ορίζοντας ως $x_T(t) = x(t)w_T(t)$ όπου το $w_T(t)$ είναι μονάδα εντός μίας σχετικής περιόδου και μηδέν αλλού, η ισχύς συνδυάζοντας όλα τα παραπάνω γράφεται ως:

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{+\infty} |x_T(f)|^2 \, df$$

Έτσι η πυκνότητα της φασματικής ισχύος είναι:

$$S_{xx}(f) = \lim_{T\to\infty} \frac{1}{T} |\hat{x}_T(f)|^2$$

Τέλος εφόσον το σήμα είναι διακριτό ο τελικός τύπος υπολογισμού της πυκνότητας της φασματικής ισχύος είναι:

$$S_{xx}(f) = \lim_{N\to\infty} \frac{(\Delta t)^2}{T} \left| \sum_{n=-N}^{N} x_n e^{-i2\pi f n \Delta t} \right|^2$$

Όπου $x_n = x(n\Delta t)$ τα δείγματα του σήματος, για μία συνολική περίοδο δειγματοληψίας $T = (2N+1)\Delta t$ και $-N \leq n \leq N$.

Υπολογίζοντας για διάφορες συχνότητες την πυκνότητα φασματικής ισχύος, δημιουργείται το διάγραμμα ισχύος φάσματος. Μέσα από αυτό επιλέγοντας τη συχνότητα για την οποία μεγιστοποιείται η φασματική ισχύς, κάποιος μπορεί να βρει την κύρια περίοδο του σήματος.

3.3.9 Εντοπισμός σακκάδων και φάσεων εστίασης

Η κίνηση του ματιού, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, χαρακτηρίζεται από σακκάδες και φάσεις εστίασης. Μέσα σε αυτή τη συμπεριφορά όμως παρεμβάλεται και θόρυβος από μικροκινήσεις του ματιού και σφάλματα του εξοπλισμού. Έτσι ο διαχωρισμός των σακκάδων από τα σημεία εστίασης γίνεται πιο αξιόπιστα μέσω της ταχύτητας και της επιτάχυνσης των κινήσεων του ματιού. Ανάλογα με τα δεδομένα προς ανάλυση ορίζεται ένας μέγιστος αριθμός ταχύτητας και επιτάχυνσης που μπορεί να έχει το μάτι σε μία φάση εστίασης και ένα ελάχιστο διάστημα που πρέπει να περάσει μετά από μία σακκάδα για να υπάρξει επόμενη.

Η θέση του ματιού σε δύο άξονες αποτυπώνεται σε δύο χρονοσειρές, κάθε μία από τις οποίες περιέχει τη θέση του ματιού πάνω στον αντίστοιχο άξονα σε μία χρονική στιγμή. Με βάση αυτές τις χρονοσειρές, αρχικά υπολογίζεται η ευκλείδεια απόσταση του ενός δείγματος από το επόμενο και σε κάθε σημείο διαιρώντας με την χρονική διαφορά που λήφθηκαν τα δείγματα υπολογίζεται η ταχύτητα του ματιού. Η επιτάχυνση ορίζεται απλώς ως η διαφορά μεταξύ των δειγμάτων της ταχύτητας. Στη συνέχεια για κάθε σημείο εξετάζεται αν η ταχύτητα και η επιτάχυνση ξεπερνούν της δοθείσες τιμές. Στα σημεία που τις ξεπερνούν για πρώτη φορά θεωρείται πως ξεκινά σακκάδα, και όταν σταματήσουν να τις ξεπερνούν θεωρείται πως η σακκάδα έληξε. Η παραπάνω διαδικασία γίνεται με σεβασμό στο ελάχιστο διάστημα που πρέπει να παρεμβάλεται μεταξύ σακκάδων. Τα διαστήματα στα οποία δεν υπάρχουν σακκάδες θεωρούνται φάσεις εστίασης.

3.4 Μετρικές Αξιολόγησης

Τα μοντέλα ταξινόμησης εκπαιδεύονται με σκοπό των διαμοιρασμό των δεδομένων σε κλάσεις. Μετά την εκπαίδευσή τους θα πρέπει κάποιος να είναι σε θέση να αξιολογήσει την ικανότητα των μοντέλων στο να εκτελούν αυτή την εργασία με επιτυχία. Γι΄ αυτό το σκοπό έχει οριστεί ένα σύνολο από ευρέως χρησιμοποιούμενες μετρικές οι οποίες εκφράζουν την ακρίβεια των προβλέψεων ενός μοντέλου σε σχέση με τις πραγματικές τιμές των δεδομένων αξιολόγησης. Έτσι γίνεται αισθητό πως αυτές δεν σχετίζονται με τη συνάρτηση κόστους που αναφέρθηκε στα μοντέλα ταξινόμησης. Τα προβλήματα που λύνει αυτή η διπλωματική εργασία είναι δυαδικής ταξινόμησης και οι τύποι που παρατίθενται ενδείκνυνται για αυτά.

Για τη δυαδική ταξινόμηση των δεδομένων ορίζονται ως: αληθώς θετική (True Positive) μία πρόβλεψη που ήταν θετική και η πραγματική τιμή των δεδομένων ήταν θετική, ψευδώς θετική (False Positive) μία πρόβλεψη που ήταν θετική και η πραγματική τιμή των δεδομένων ήταν αρνητική, αληθώς αρνητική (True Negative) μία πρόβλεψη που ήταν αρνητική και η πραγματική τιμή των δεδομένων ήταν αρνητική, ψευδώς αρνητική (False Negative) μία πρόβλεψη που ήταν αρνητική και η πραγματική τιμή των δεδομένων ήταν θετική. Συμβολίζεται TP το σύνολο των True Positive τιμών, FP το σύνολο των False Positive τιμών, TN το σύνολο των True Negative τιμών, FN το σύνολο των False Negative τιμών. Οι μετρικές ορίζονται ως:

- Ορθότητα (Accuracy): Εκφράζει το ποσοστό επιτυχίας του μοντέλου, δηλαδή το πλήθος των δειγμάτων που το μοντέλο ταξινόμησε σωστά προς το σύνολο όλων των δειγμάτων.

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

- Ακρίβεια (Precision): Εκφράζει το ποσοστό των επιτυχημένων προβλέψεων μίας κλάσης προς τις συνολικές προβλέψεις της κλάσης.

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

- Ανάκληση (Recall): Εκφράζει το ποσοστό των επιτυχημένων προβλέψεων μίας κλάσης προς το πραγματικό πλήθος των παρατηρήσεων που ανήκουν σε αυτή.

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

- F1 score: Αποτελεί τον αρμονικό μέσο των Precision και Recall.

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Κεφάλαιο 4

Σύνολα Δεδομένων

Στο κεφάλαιο αυτό περιγράφονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την παρούσα εργασία, η προεπεξεργασία τους, καθώς και τα μοντέλα τα οποία εφαρμόστηκαν στο καθένα. Με το συνδυασμό του παρόντος και του προηγούμενου κεφαλαίου, ο αναγνώστης θα είναι σε θέση να κατανοήσει σε βάθος τις τεχνικές οι οποίες εφαρμόστηκαν για την προεπεξεργασία των δεδομένων, αλλά και πώς αυτές οδηγούν στην καλύτερη εκπαίδευση των μοντέλων.

Τα δεδομένα που εξετάσθηκαν ήταν δεδομένα που προέκυψαν από παιχνιδοποίηση και οφθαλμική ανίχνευση, όπως αναφέρονται στις ενότητες 2.2 και 2.3 αντίστοιχα. Σε αυτό το κεφάλαιο γίνεται πιο αισθητή, εμπειρικά, η σύνδεση των μεθόδων συλλογής δεδομένων στην διάγνωση της δυσλεξίας. Εξαιτίας της διαφοράς στη φύση τους, ακόμα, αναφέρονται και οι διαφορετικές μέθοδοι προεπεξεργασίας που εφαρμόστηκαν σε αυτά.

4.1 Σύνολα Δεδομένων

4.1.1 Δεδομένα παιχνιδοποίησης

Περιγραφή

Όπως περιγράφεται από τους Rello L, Baeza-Yates R et al. [13], οι συμμετέχοντες στην καταγραφή των δεδομένων ήταν άτομα σε ηλικίες 7 έως 17 χρονών. Τα άτομα με δυσλεξία επιλέχθηκαν μέσω επικοινωνίας με κέντρα διάγνωσης και συνδέσμων ατόμων με δυσλεξία. Τα άτομα αυτά θα έπρεπε να έχουν επίσημη γνωμάτευση της μαθησιακής δυσκολίας. Τα άτομα χωρίς δυσλεξία επιλέχθηκαν από σχολεία και ήταν μόνο αυτά που δεν είχαν προβλήματα με τη γλώσσα σε καταγραφές του σχολείου. Η μητρική γλώσσα των ατόμων που συμμετείχαν στο πείραμα ήταν τα Ισπανικά.

Το σύνολο των δεδομένων αποτελούνταν από 3644 εγγραφές, μία για κάθε συμμετέχοντα. Οι 392 ήταν από άτομα επισήμως διαγνωσμένα με δυσλεξία και κάθε εγγραφή είχε 196 χαρακτηριστικά καθώς και έναν χαρακτηρισμό του αν το άτομο είχε δυσλεξία ή όχι. Τα πρώτα τέσσερα χαρακτηριστικά ήταν δημογραφικά. Πιο συγκεκριμένα το πρώτο αφορούσε το φύλο και έπαιρνε τιμές θηλυκό και αρσενικό. Το δεύτερο αφορούσε το πλήθος των μητρικών γλωσ-

σών και έπαιρνε τιμή ναι, αν ο συμμετέχων είχε μόνο την Ισπανική γλώσσα ως μητρική και όχι αν είχε περισσότερες γλώσσες ως μητρικές. Το τρίτο αφορούσε το αν ο συμμετέχων είχε μη προβιβάσιμο βαθμό σε κάποιο γλωσσικό σχολικό μάθημα και έπαιρνε τιμές ναι και όχι. Το τελευταίο ήταν η ηλικία του συμμετέχοντος και ήταν από 7 έως 17 χρονών.

Τα υπόλοιπα χαρακτηριστικά προέκυψαν μέσω μετρικών από 32 ερωτήσεις - παιχνίδια, κάθε μία εκ των οποίων εξέταζε διαφορετικές γλωσσικές ικανότητες του ατόμου, όπως αυτές περιγράφηκαν στην ενότητα 2.2. Κάθε ερώτηση αποτελούνταν από έναν ή περισσότερους γύρους με διαφορετικές παραλλαγές του ίδιου παιχνιδιού.

Εδώ παρουσιάζονται οι διαφορετικές ερωτήσεις και οι κατηγορίες που αυτές ανήκουν:

- Ερωτήσεις 1-4: Σε αυτές ο συμμετέχων ακούει ένα γράμμα από τα e,g,b και d και το αντιστοιχεί σε ένα γράμμα μεταξύ άλλων που είναι ορθογραφικά και φωνολογικά παρόμοια, μέσα σε ένα συγκεκριμένο χρονικό διάστημα. Αυτές οι ερωτήσεις εξετάζουν την αλφαβητική επίγνωση, τη φωνολογική επίγνωση και την οπτική διαφοροποίηση και κατηγοριοποίηση.

- Ερωτήσεις 5-9: Σε αυτές ο συμμετέχων ακούει συλλαβές και πρέπει να τις αντιστοιχίσει με το πως είναι αυτές γραμμένες. Αυτές οι ερωτήσεις εξετάζουν τη φωνολογική επίγνωση, τη συλλαβική επίγνωση, την ακουστική διαφοροποίηση και κατηγοριοποίηση.

- Ερωτήσεις 10-13: Σε αυτές ο συμμετέχων καλείται να αντιστοιχίσει μία λέξη που ακούει, με την ορθογραφία της. Ανάμεσα στης λέξεις που έχει να επιλέξει ο συμμετέχων, πέρα από τη σωστή είναι λέξεις που μοιάζουν ορθογραφικά σε αυτή που ακούει ή ψευδολέξεις. Αυτές οι ερωτήσεις εξετάζουν τη λεκτική επίγνωση, την ακουστική ενεργό μνήμη και την ακουστική διαφοροποίηση και κατηγοριοποίηση.

- Ερωτήσεις 14-17: Σε αυτές ο συμμετέχων καλείται να βρει όσο το δυνατόν περισσότερα διαφορετικά γράμματα μπορεί, μέσα σε μία δεδομένη χρονική περίοδο, για παράδειγμα E/F, g/q, u/n, c/o, b/d, d/p, b/q κ.α. Αυτές οι ερωτήσεις εξετάζουν την οπτική διαφοροποίηση και κατηγοριοποίηση και τις εκτελεστικές λειτουργίες.

- Ερωτήσεις 18-21: Σε αυτές ο συμμετέχων καλείται να ακούσει ψευδολέξεις και να διαλέξει το πως γράφονται ανάμεσα σε άλλες ψευδολέξεις. Για παράδειγμα ακούει pamata και έχει να διαλέξει μία από τις mapata, matapa, pamada κ.α. Αυτές οι ερωτήσεις εξετάζουν την οπτική ενεργό μνήμη, την ακολουθιακή ακουστική ενεργό μνήμη και την ακουστική διαφοροποίηση και κατηγοριοποίηση.

- Ερωτήσεις 22-23: Σε αυτές ο συμμετέχων καλείται να συμπληρώσει τα κενά σε λέξεις που λείπουν γράμματα ή να αφαιρέσει ένα γράμμα από μία λέξη που έχει περισσότερα από ένα. Αυτές οι ερωτήσεις εξετάζουν την ορθογραφική, λεξική και φωνολογική επίγνωση.

- Ερώτηση 24: Σε αυτή δίνεται στο συμμετέχοντα μία πρόταση, στην οποία η λάθος ορθογραφία σε μία λέξη, δημιουργεί και λάθος στο νόημα της πρότασης. Αυτή εξετάζει τη μορφολογική και σημασιολογική επίγνωση.

- Ερώτηση 25: Σε αυτή δίνεται στον συμμετέχοντα μία πρόταση στην οποία υπάρχει μία συντακτικά λάθος λέξη, η οποία δημιουργεί και λάθος στο νόημα της πρότασης. Αυτή εξετάζει τη συντακτική επίγνωση.

- Ερώτηση 26: Σε αυτή δίνεται στον συμμετέχοντα μία λέξη που έχει ένα λάθος γράμμα, και καλείται να το διορθώσει επιλέγοντας ανάμεσα σε κάποια γράμματα που του δίνονται. Αυτή εξετάζει τη φωνολογική, λεκτική και ορθογραφική επίγνωση.

- Ερωτήσεις 27-28: Σε αυτές δίνονται στον συμμετέχοντα μπερδεμένα είτε γράμματα είτε συλλαβές και καλείται να κάνει αναγραμματισμό αυτών, ώστε να δημιουργήσει μία λέξη χωρίς να αφήσει κάποιο από αυτά. Αυτές εξετάζουν την φωνολογική, ορθογραφική, συλλαβική και λεξική επίγνωση.

- Ερώτηση 29: Σε αυτή δίνεται στον συμμετέχοντα μία πρόταση χωρίς κενά και καλείται να διαχωρίσει τις λέξεις. Αυτή εξετάζει την φωνολογική, ορθογραφική και λεξική επίγνωση.

- Ερώτηση 30: Σε αυτή δίνεται στον συμμετέχοντα για τρία δευτερόλεπτα μία ακολουθία από γράμματα και αυτός καλείται να τα ξαναγράψει. Αυτή εξετάζει την ακολουθιακή οπτική ενεργό μνήμη, και την οπτική διαφοροποίηση και κατηγοριοποίηση.

- Ερωτήσεις 31-32: Σε αυτές ο συμμετέχων ακούει τέσσερις λέξεις και τέσσερις ψευδολέξεις και καλείται να τις γράψει. Αυτές εξετάζουν την λεκτική, ορθογραφική και φωνολογική επίγνωση και την ακουστική ενεργό μνήμη.

Για κάθε μία από τις παραπάνω ερωτήσεις έγινε η συλλογή έξι χαρακτηριστικών ώστε να ποσοτικοποιηθεί η επίδοση του κάθε ατόμου. Τα χαρακτηριστικά αυτά προκύπτουν από μετρικές της αλληλεπίδρασης του ανθρώπου με τον υπολογιστή και ήταν τα:

- Αριθμός των κλικ (Clicks).

- Αριθμός σωστών απαντήσεων (Hits).

- Αριθμός λάθος απαντήσεων (Misses).

- Σκορ το οποίο ορίζεται ως:

$$\frac{Hits}{\#set\,of\,exercises}$$

- Ακρίβεια που ορίζεται ως:

$$\frac{Hits}{Clicks}$$

- Λόγος αστοχιών που ορίζεται ως:

$$\frac{Misses}{Clicks}$$

Δειγματοληψία

Το παρόν σύνολο δεδομένων έχει αρκετές ιδιομορφίες, οι οποίες καθιστούν αναγκαία την προεπεξεργασία του. Η πρώτη και κυριότερη που κάποιος μπορεί να παρατηρήσει είναι η έλλειψη ισορροπίας στο διαχωρισμό των δειγμάτων σε κλάσεις (άτομα με δυσλεξία και άτομα χωρίς). Όπως αναφέρθηκε στην προηγούμενη ενότητα στα 3644 άτομα μόνο τα 392 είχαν δυσλεξία. Δηλαδή ο λόγος των δειγμάτων κάθε κλάσης προς τα συνολικά δείγματα είναι 10.8% για τα άτομα με δυσλεξία και 89.2% για τα άτομα χωρίς. Έτσι έγινε αναγκαία η χρήση της δειγματοληψίας, συνδυασμός τυχαίας υποδειγματοληψίας και υπερδειγματοληψίας για την εξισορρόπηση των δεδομένων στις κλάσεις με και χωρίς δυσλεξία. Το σύνολο στο οποίο έγινε η δειγματοληψία ήταν μόνο αυτό της εκπαίδευσης του μοντέλου. Οι παράμετροι για το ρυθμό δειγματοληψίας παρουσιάζονται στο τέλος της ενότητας με αυτές όλων των μοντέλων της προεπεξεργασίας.

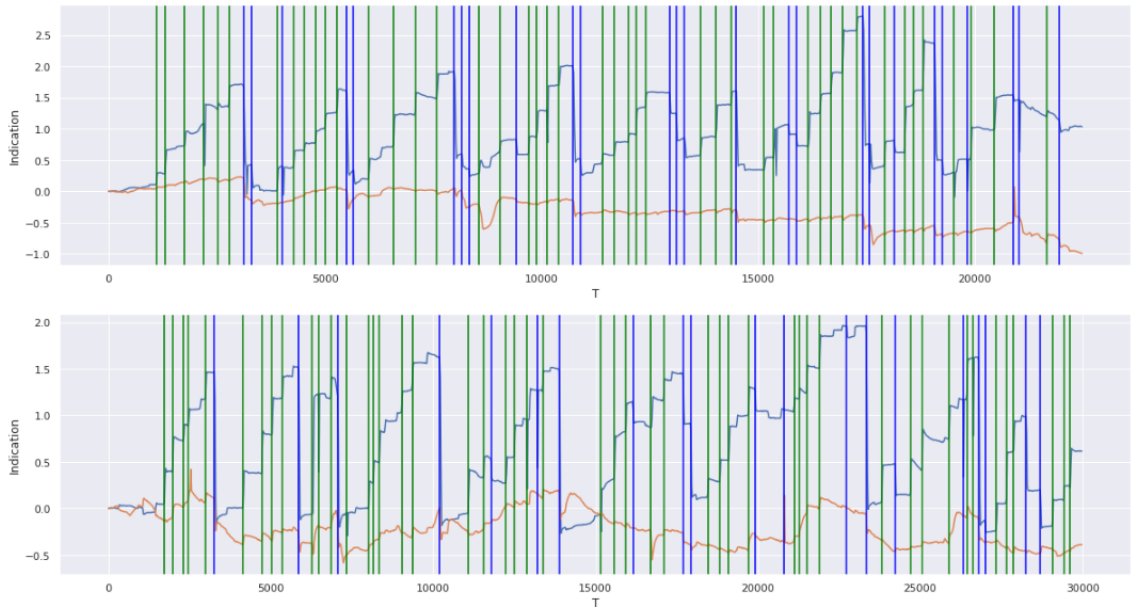
4.1.2 Δεδομένα οφθαλμικής ανίχνευσης

Περιγραφή

Σύμφωνα με τους Nilsson Benfatto M, Öqvist Seimyr G et al [19], τα άτομα που συμμετείχαν στην έρευνα, συμμετείχαν και στο ερευνητικό σχέδιο Kronoberg(1989-2010) με αντικείμενο την ανάπτυξη στην ανάγνωση σε παιδιά σχολικής ηλικίας, στη Σουηδία. Οι ηλικίες τους ήταν 9 με 10 χρονών από έναν αρχικό πληθυσμό 193 ατόμων, που συμμετείχαν στα πλαίσια μίας οφθαλμολογικής εξέτασης. Αυτή στόχευε στην εύρεση κάποιας διαφοράς ανάμεσα στους πληθυσμούς υψηλού και χαμηλού ρίσκου κατά την ανάγνωση. Για την επιλογή των ατόμων κάθε κλάσης (χαμηλού και υψηλού ρίσκου), θα έπρεπε να πληρούνται κάποια κριτήρια. Από τον συνολικό αριθμό των παιδιών που συμμετείχαν, επιλέχθηκαν 185, 97 υψηλού ρίσκου για την εμφάνιση της δυσλεξίας και 88 χαμηλού ρίσκου άτομα ελέγχου.

Αρχικά, τα άτομα υψηλού ρίσκου, έπρεπε να έχουν τη Σουηδική γλώσσα ως κύρια, βρίσκονταν στο πέμπτο εκατοστημόριο από το σύνολο των ατόμων που συμμετείχαν σε δύο κανονικοποιημένα τεστ αποκωδικοποίησης και είχαν επανειλημμένη δυσκολία στο διάβασμα σύμφωνα με την εκτίμηση του δάσκαλου της εκάστοτε τάξης. Άτομα με νοητική υστέρηση αφαιρέθηκαν από το σύνολο δεδομένων. Το παραπάνω δεν έρχεται σε αντιπαράθεση με την επικρατούσα σημερινή άποψη ότι η δυσλεξία εμφανίζεται σε όλες τις νοητικές δυνατότητες και είναι τα δείγματα χαμηλότερης ποιότητας μίας κανονικής κατανομής της ικανότητας στην ανάγνωση. Τα άτομα χαμηλού ρίσκου επιλέχθηκαν με βάση την επίδοσή τους στην ανάγνωση. Αυτή θα έπρεπε να είναι στο μέσο ή άνω του μέσου.

Κατά τη διάρκεια της οφθαλμικής εξέτασης για περίπου 5 λεπτά δόθηκαν στους συμμετέχοντες 8 γραμμές κειμένου, που περιείχαν 10 προτάσεις με μέσο πλήθος λέξεων 4.6 λέξεις ανά πρόταση. Από την παραπάνω διαδικασία προέκυψαν 185 χρονοσειρές, με ένα δείγμα ανά 200 ms και τιμές τις συντεταγμένες του κάθε ματιού LX, LY, RX και RY. Τα L και R ορίζουν το αριστερό και το δεξί μάτι και τα X και Y την κίνηση στον οριζόντιο και κάθετο άξονα.



Σχήμα 4.1: Δεδομένα οφθαλμικής ανίχνευσης.

Οι κάθετες γραμμές συμβολίζουν σημεία στα οποία έχει γίνει σακκαδική κίνηση. Είναι πράσινες αν αυτή είναι προς τα δεξιά και μπλε αν αυτή είναι προς τα αριστερά. Οι υπόλοιπες στιγμές θεωρούνται ως φάσεις εστίασης. Η πιο ανοιχτή μπλε γραμμή συμβολίζει της κίνηση του ματιού στον άξονα X ενώ η πορτοκαλί στον άξονα Υ. Το πρώτο γράφημα ανήκει σε άτομο ελέγχου, ενώ το δεύτερο σε άτομο υψηλού ρίσκου.

Δυναμική χρονική στρέβλωση

Στο σύνολο δεδομένων που περιγράφηκε οι χρονοσειρές δεν έχουν όλες το ίδιο μήκος. Έτσι πριν την εξαγωγή χαρακτηριστικών από τις χρονοσειρές εφαρμόστηκε η δυναμική χρονική στρέβλωση των χρονοσειρών, προκειμένου να αποκτήσουν όλες ίδιο αριθμό δειγμάτων όπως αναφέρθηκε στην ενότητα 3.3.7.

Επιλέγοντας τη χρονοσειρά με τα περισσότερα δείγματα έγινε με τον τρόπο που περιγράφηκε η αντιστοίχιση των σημείων της με τα σημεία κάθε άλλης χρονοσειράς. Το παραπάνω είχε ως αποτέλεσμα την επιμήκυνση όλων των χρονοσειρών σε χρονοσειρές με ίσο πλήθος δειγμάτων.

Ο παραπάνω μετασχηματισμός των δεδομένων θεωρήθηκε πως δεν επιβάλει μεγάλο θόρυβο στα δεδομένα καθώς οι χρονοσειρές που εξετάζονται έχουν περιοδικότητα, με περίοδο το χρόνο ανάγνωσης μίας γραμμής από τον αναγνώστη. Έτσι αναγνώστες που διαβάζουν γρήγορα και στρωτά το κείμενο θα έχουν και πιο ομαλές τιμές σε κάθε περίοδο τους, σε αντίθεση με αναγνώστες που παρουσιάζουν θόρυβο μέσα σε μία περίοδο.

Τα πειράματα που ακολούθησαν εφαρμόστηκαν ξεχωριστά στα δεδομένα με και χωρίς δυναμική χρονική στρέβλωση.

Εξαγωγή χαρακτηριστικών

Τα μοντέλα που αναφέρονται στη βιβλιογραφία δεν είναι ικανά να διαχειριστούν τα δεδομένα στην μορφή που αναφέρεται παραπάνω. Για να γίνει εφικτό αυτό πρέπει να γίνει η μετατροπή των χρονοσειρών σε διανύσματα ίσου μεγέθους μέσω της εξαγωγής χαρακτηριστικών από αυτές.

Όπως αναφέρθηκε και στην ενότητα 2.3 η κίνηση του ματιού χαρακτηρίζεται από σακκαδικές κινήσεις και φάσεις εστίασης. Ακόμα αναφέρθηκε πως η διαφοροποίηση μεταξύ ατόμων με δυσλεξία και χωρίς, έγκειται στο γεγονός ότι παρουσιάζουν διαφορές σε αυτές τις δύο συνιστώσες της κίνησης του οφθαλμού. Έτσι, εντοπίζοντας τα σημεία που υπάρχουν σακκαδικές κινήσεις και φάσεις εστίασης, καθώς και προσδιορίζοντας την κατεύθυνση που συμβαίνουν αυτές, κάποιος μπορεί να εξάγει στατιστικά χαρακτηριστικά, για παράδειγμα αριθμό εμπρόσθιων σακκάδων, μέση απόσταση που διανύεται κλπ. ώστε να δημιουργήσει ένα διάνυσμα που χαρακτηρίζει τη συμπεριφορά των οφθαλμών του αναγνώστη κατά τη διάρκεια της ανάγνωσης ενός κειμένου.

Εφαρμόζοντας τη διαδικασία που ορίστηκε στην ενότητα 3.3.9 για την εύρεση των σακκάδων, έγινε η μετατροπή της σχετικής θέσης του ματιού (οι συντεταγμένες που αναφέρθηκαν) στην ευκλείδεια απόσταση από δείγμα σε δείγμα και κατ΄ επέκταση σε ταχύτητα και επιτάχυνση. Ως ταχύτητα εντοπισμού σακκαδικών κινήσεων επιλέχθηκαν οι 0.5 μονάδες ανά δευτερόλεπτο. Οι μονάδες αυτές είναι σχετικές καθώς έχουν να κάνουν με το πλέγμα στο οποίο ο οφθαλμικός ιχνηλάτης τοποθετεί τις συντεταγμένες του ματιού. Μόλις γίνει ο διαχωρισμός των κινήσεων σε σακκαδικές και φάσεις εστίασης, αφαιρώντας την πρώτη από την τελευταία θέση κάθε συμβάντος βρίσκεται η κατεύθυνση του. Έτσι προκύπτουν εμπρόσθιες σακκαδικές κινήσεις, παλινδρομήσεις και εμπρόσθιες φάσεις εστίασης και φάσεις εστίασης με πίσω φορά, ανάλογα με την σακκαδική κίνηση που προηγούνταν.

Μετά από την παραπάνω διαδικασία εύρεσης και κατηγοριοποίησης των συμβάντων έγινε εξαγωγή των εξής χαρακτηριστικών για κάθε είδος συμβάντος:

- Χρονική διάρκεια του συμβάντος.
- Απόσταση που διανύθηκε σε αυτό.
- Μέση θέση του ματιού κατά το συμβάν.
- Τυπική απόκλιση του ματιού κατά το συμβάν.
- Μέγιστη απόσταση ανάμεσα σε οποιεσδήποτε δύο θέσεις του συμβάντος.
- Αθροιστική απόσταση ανάμεσα σε όλες της διαδοχικές θέσεις του συμβάντος.

Όλες οι παράμετροι εκτός από την πρώτη, μετρήθηκαν ξεχωριστά στον οριζόντιο και κατακόρυφο άξονα, και για τη μέση θέση των δύο ματιών $\frac{L+R}{2}$, αλλά και για τη διαφορά τους $L - R$. Έτσι κάθε συμβάν χαρακτηρίζεται από 84 μετρικές. Για κάθε μία από αυτές τις 84 μετρικές υπολογίστηκε ο μέσος και η τυπική απόκλιση. Τέλος πέρα από τα παραπάνω που πρότειναν οι Nilsson Benfatto M, Öqvist Seimyr G et al προστέθηκαν άλλα 3 χαρακτηριστικά

που αφορούσαν την περιοδικότητα των χρονοσειρών στον άξονα Χ. Αυτά ήταν οι περίοδοι που προκύπτουν από τις τρεις πιο ισχυρές συχνότητες του γραφήματος ισχύος του φάσματος του σήματος, όπως ορίστηκε στην ενότητα 2.3.9. Εδώ πρέπει να σημειωθεί πως ο υπολογισμός της περιοδικότητας έγινε πριν τη διαδικασία της δυναμικής χρονικής στρέβλωσης. Τα παραπάνω δημιούργησαν ένα διάνυσμα 171 χαρακτηριστικών στο σύνολο.

Ο λόγος που χρησιμοποιήθηκε η περιοδικότητα ως χαρακτηριστικό, είναι το γεγονός ότι η κίνηση του ματιού στον άξονα Χ χαρακτηρίζεται από περιοδικότητα, καθώς εμφανίζει μέγιστα και ελάχιστα σε σχετικά ίσα χρονικά διαστήματα ανάλογα με τη θέση της λέξης στη γραμμή που διαβάζει ο αναγνώστης. Αποκτά ελάχιστο στην αρχή της γραμμής και μέγιστο στο τέλος. Ακόμα περιέχει και πληροφορία για την ταχύτητα ανάγνωσης, καθώς οι γρήγοροι αναγνώστες έχουν μικρότερη περίοδο για την ανάγνωση γραμμής. Τα χαρακτηριστικά στο σύνολο τους περιέχουν πληροφορία για την χρονική διάρκεια κάθε συμβάντος, την τάξη μεγέθους του, την κατεύθυνση, τη σταθερότητα και τη συμμετρία στα μάτια.

4.2 Προεπεξεργασία

Στην προηγούμενη ενότητα αναφέρθηκαν μεθοδολογίες που αφορούσαν την προεπεξεργασία των συνόλων δεδομένων ξεχωριστά, εξαιτίας της διαφοράς των χαρακτηριστικών τους. Προσθετικά στις παραπάνω τεχνικές προεπεξεργασίας χρησιμοποιήθηκαν τεχνικές κλιμάκωσης και μείωσης διαστατικότητας.

Οι τεχνικές κλιμάκωσης, κανονική κλιμάκωση και κλιμάκωση ελαχίστου μεγίστου χρησιμοποιήθηκαν και στα δύο σύνολα δεδομένων. Η χρησιμότητα τους έγκειται στο γεγονός πως και τα δύο σύνολα περιέχουν χαρακτηριστικά τα οποία έχουν διαφορετική τάξη μεγέθους μεταξύ τους. Έτσι χωρίς την κλιμάκωση το μοντέλο λανθασμένα μπορεί να θεωρούσε πως κάποιο χαρακτηριστικό έχει μεγαλύτερη ισχύ από τα άλλα και τελικά να οδηγούνταν σε λάθος συμπεράσματα. Ακόμα, όταν έχει γίνει κλιμάκωση των δεδομένων, οι τεχνικές αριθμητικής βελτιστοποίησης μπορεί να φτάσουν σε σύγκλιση πιο γρήγορα. Η κλιμάκωση έγινε σε κάθε χαρακτηριστικό ξεχωριστά.

Σε επόμενο στάδιο της κλιμάκωσης έγινε η μείωση της διαστατικότητας. Στο δεύτερο σύνολο δεδομένων για 185 παρατηρήσεις έχουν προκύψει 171 χαρακτηριστικά. Έτσι καθίσταται αναγκαία η μείωση των διαστάσεων για την εκπαίδευση των μοντέλων. Στο πρώτο σύνολο δεδομένων παρατηρήθηκε πως πολλά από τα χαρακτηριστικά που χρησιμοποιήθηκαν δεν παρουσιάζουν σημαντικές διαφορές ανάμεσα στις δύο κλάσεις. Έτσι πρέπει να γίνει μείωση της διαστατικότητας για να μειωθεί και ο θόρυβος που μπορεί να προκύπτει από αυτά τα χαρακτηριστικά.

Για τη μείωση της διαστατικότητας εφαρμόστηκαν δύο τεχνικές. Η πρώτη ήταν με χρήση της ανάλυσης κύριων συνιστωσών και η δεύτερη με επιλογή των χαρακτηριστικών. Για την επιλογή χαρακτηριστικών έγινε χρήση στατιστικών μεθόδων, γενετικών αλγορίθμων και η επιλογή μέσω μοντέλων όπως αυτές αναφέρθηκαν στο προηγούμενο κεφάλαιο.

Πιο αναλυτικά για την εφαρμογή κάθε μίας από τις παραπάνω τεχνικές:

- Για τη σωστή εφαρμογή της ανάλυσης κύριων συνιστωσών έπρεπε τα δεδομένα να έχουν

υποστεί κανονική κλιμάκωση πριν γίνει η εισαγωγή τους στον αλγόριθμο.

- Για τη χρήση στατιστικών μεθόδων, επειδή εσωτερικά σε αυτές γίνεται η σύγκριση δύο πληθυσμών, με και χωρίς δυσλεξία, δεν είχε γίνει κλιμάκωση πριν εφαρμοστούν, καθώς αυτή θα αλλοίωνε τα στατιστικά τεστ.

- Για την επιλογή χαρακτηριστικών μέσω μοντέλων εφαρμόστηκε κάθε είδους προεπεξεργασία που αναφέρθηκε παραπάνω προτού εκπαιδευτούν τα μοντέλα.

- Για την χρήση γενετικών αλγορίθμων, έγινε πριν τη χρήση τους κανονική κλιμάκωση των χαρακτηριστικών. Στη συνέχεια κάθε γονίδιο στο γονιδίωμα αντιστοιχούσε σε ένα χαρακτηριστικό και έπαιρνε τιμές 0 (το χαρακτηριστικό δεν συμπεριλαμβάνεται στην εκπαίδευση) και 1 (το χαρακτηριστικό συμπεριλαμβάνεται στην εκπαίδευση). Τέλος η συνάρτηση ικανότητας επέστρεφε στο πρώτο σύνολο δεδομένων το F1 score και στο δεύτερο την ορθότητα, όπως αυτές προέκυπταν από την αξιολόγηση του μοντέλου. Περισσότερες λεπτομέρειες για την αξιολόγηση των μοντέλων αναφέρονται στην αντίστοιχη ενότητα.

Μετά το στάδιο αυτό τα δεδομένα είναι σε θέση να χρησιμοποιηθούν για την εκπαίδευση των μοντέλων.

4.3 Μοντέλα

Η ανάλυση των μοντέλων, των οποίων έγινε η εφαρμογή στα δεδομένα, έγινε στην προηγούμενη ενότητα. Πιο συγκεκριμένα τα μοντέλα που χρησιμοποιήθηκαν και στα δύο σύνολα δεδομένων ήταν τα τυχαία δάση, τα δέντρα ενίσχυσης με κλίση, η λογιστική παλινδρόμηση και οι μηχανές διανυσμάτων υποστήριξης. Πριν γίνει ακόμα η λήψη αποτελεσμάτων, με τη θεωρητική ανασκόπηση των μοντέλων και με σεβασμό προς τα δεδομένα, μπορούμε να εξάγουμε κάποιες υποθέσεις για την επίδοση αυτών.

Αρχικά, αναμένουμε πως η χειρότερη επίδοση θα προκύψει από το μοντέλο λογιστικής παλινδρόμησης και από τη χρήση μηχανών διανυσμάτων υποστήριξης με γραμμικό πυρήνα. Αυτή η υπόθεση βασίζεται στο ότι τα δύο σύνολα που εξετάζονται σε κάθε περίπτωση (άτομα με ή χωρίς δυσλεξία για το πρώτο σύνολο δεδομένων και άτομα με υψηλό ρίσκο και χαμηλό για την εμφάνιση δυσλεξίας του δεύτερου συνόλου δεδομένων) δεν είναι γραμμικά διαχωρίσιμα. Έτσι τα δύο μοντέλα που αναφέρθηκαν εφόσον είναι από τη φύση τους γραμμικά, δηλαδή προσπαθούν να προσδιορίσουν το υπερεπίπεδο που διαχωρίζει βέλτιστα τις δύο κατηγορίες, αδυνατούν να εντοπίσουν την πολύπλοκη σχέση μεταξύ των χαρακτηριστικών που οδηγεί στον διαχωρισμό της μίας κλάσης από την άλλη. Τέτοιες σχέσεις εντοπίζονται από τα τυχαία δάση, τα δέντρα ενίσχυσης με κλίση και τις μηχανές διανυσμάτων υποστήριξης με μη γραμμικό πυρήνα, τα οποία αποτελούν μη γραμμικά μοντέλα. Η χρήση των γραμμικών μοντέλων έγινε για τη σύγκριση της επίδοσής τους με τα μη γραμμικά αλλά και για λόγους που αναφέρονται παρακάτω.

Άλλη μία υπόθεση που μπορεί να γίνει, είναι πως στο δεύτερο σύνολο δεδομένων αναμένεται να εμφανίσουν καλύτερη επίδοση τα μοντέλα μηχανών διανυσμάτων υποστήριξης, καθώς όπως αναφέρθηκε και στην αντίστοιχη ενότητα του προηγούμενου κεφαλαίου, είναι αποτελεσματικά σε πολυδιάστατους χώρους ακόμα και όταν ο αριθμός των διαστάσεων ξεπερνά το πλήθος των δειγμάτων.

Η επιλογή των μοντέλων που χρησιμοποιήθηκαν δεν έγινε μόνο με βάση την επίδοση που ήταν αναμενόμενο πως θα είχαν. Εξίσου σημαντική παράμετρο για την επιλογή τους αποτέλεσε και η δυνατότητα επεξήγησης της σημασίας των χαρακτηριστικών, από αυτά. Εδώ φαίνεται και η κύρια αιτία της επιλογής των γραμμικών μοντέλων, καθώς μέσα από τα βάρη αυτών μπορεί να εξαχθεί πληροφορία για τη σημασία των χαρακτηριστικών στην εκπαίδευση του μοντέλου.

Στο πρώτο πρόβλημα που εξετάζεται, η σημασία των χαρακτηριστικών μπορεί να ερμηνευθεί και ως η ισχύς που έχουν οι μετρικές των ερωτήσεων, αλλά και οι ίδιες οι ερωτήσεις και τα δημογραφικά χαρακτηριστικά για τον εντοπισμό της δυσλεξίας μέσω παιχνιδοποίησης. Στο δεύτερο η σημασία των χαρακτηριστικών υποδεικνύει το ποιες και σε ποιο βαθμό οι στατιστικές ιδιότητες των χαρακτηριστικών της συμπεριφοράς του ματιού, χρησιμεύουν για τον προσδιορισμό του ρίσκου εμφάνισης δυσλεξίας.

Η εξαγωγή της σημασίας των χαρακτηριστικών δεν είναι σημαντική μόνο για την φυσική ερμηνεία των αποτελεσμάτων, αλλά και για τη μείωση της διαστατικότητας με την επιλογή χαρακτηριστικών. Τα γραμμικά μοντέλα ενώ μπορεί να μην έχουν τη δυνατότητα να διαχωρίσουν με το καλύτερο δυνατό τρόπο τις δύο κλάσεις, έχουν τη δυνατότητα να δώσουν σημασία στα

χαρακτηριστικά και άρα να βοηθήσουν στη μείωση της διαστατικότητας.

Σε αυτό το σημείο μπορεί να επισημανθεί και η σπουδαιότητα των μοντέλων που βασίζονται σε δέντρα απόφασης και οι μηχανές διανυσμάτων υποστήριξης. Τα πρώτα είναι μη γραμμικά μοντέλα τα οποία υποστηρίζουν και την εξαγωγή της σημασίας των χαρακτηριστικών. Αυτό τα καθιστά ικανά να ανταποκριθούν στη διπλή φύση του προβλήματος (ταξινόμηση των δειγμάτων σε κλάσεις, αλλά και ερμηνεία των αποτελεσμάτων). Ομοίως οι μηχανές διανυσμάτων υποστήριξης, αν και γραμμικά μοντέλα στη βασική θεωρία τους, με τη χρήση συναρτήσεων πυρήνα μπορούν να ανταποκριθούν και σε μη γραμμική ταξινόμηση. Ακόμα με τη χρήση γραμμικού πυρήνα καθίσταται δυνατή και η εξαγωγή της σημασίας των χαρακτηριστικών από αυτά. Το μοντέλο λογιστικής παλινδρόμησης αν και γραμμικό, ενδεχομένως να εμφανίσει καλή επίδοση σε χαμηλότερες διαστάσεις, που οι κλάσεις στα σύνολα δεδομένων διαχωρίζονται γραμμικά με μικρή απόκλιση και είναι ικανό να δώσει μία καλή εκτίμηση για επιλογή χαρακτηριστικών.

Τα μοντέλα εφαρμόστηκαν ως τα τελευταία στάδια (components) αρχιτεκτονικών σωλήνωσης (pipeline). Οι αρχιτεκτονικές αυτές είχαν μία γενική δομή ως δειγματοληψία, κλιμάκωση, εξαγωγή χαρακτηριστικών, εφαρμογή μοντέλου. Ανάλογα με το σύνολο δεδομένων και τις τεχνικές που χρησιμοποιήθηκαν, κάποια από αυτά τα στάδια παραλήφθηκαν ή έγινε αλλαγή στη σειρά τους. Οι αρχιτεκτονικές που εφαρμόστηκαν και οι παράμετροι των μοντέλων που εξετάστηκαν φαίνονται στους αντίστοιχους πίνακες.

Η αξιολόγηση των μοντέλων έγινε με τη χρήση της μεθοδολογίας πλήρους διασταυρωμένης επικύρωσης 10 τμημάτων (10-fold cross-validation). Με τη μεθοδολογία αυτή τα σύνολα δεδομένων χωρίστηκαν σε δέκα τμήματα. Για κάθε ένα από αυτά το μοντέλο εκπαιδεύεται σε όλα τα άλλα και προβλέπει το τμήμα υπό εξέταση για την αξιολόγηση του. Τέλος ως μετρική επίδοσης λαμβάνεται ο μέσος όλων των επαναλήψεων. Για το πρώτο σύνολο δεδομένων έγινε αναγκαία η χρήση της στρωματοποιημένης (stratified) πλήρους διασταυρωμένης επικύρωσης καθώς δεν υπήρχε ισορροπία μεταξύ των δειγμάτων στις κλάσεις.

Στάδια σωλήνωσης			
Δειγματοληψία	Τυχαία Υποδειγματοληψία	sampling_strategy	0
			0.25
			0.5
			0.75
			1
	Τυχαία Υπερδειγματοληψία	sampling_strategy	0
			0.25
			0.5
			0.75
			1
Κλιμάκωση	Κανονική κλιμάκωση		
	Κλιμάκωση ελαχίστου μεγίστου		
Επιλογή χαρακτηριστικών	Γενετικός Αλγόριθμος(Μηχανές Διανυσμάτων Υποστήριξης)	population_size	100
			250
			500
			750
			1000
		mutation_probability	0.1
			0.15
			0.2
		elit_ratio	0.1
			0.15
			0.2
		crossover_probability	0.5
		crossover_type	uniform
	Στατιστική Επιλογή Χαρακτηριστικών	Στατιστικό Kolmogorov–Smirnov	
		Ανάλυση Διακύμανσης με τιμές F	
	Επιλογή χαρακτηριστικών μέσω μοντέλων	Λογιστική παλινδρόμηση	
		Τυχαία δάση και δέντρα ενίσχυσης με κλίση	
		Μηχανές διανυσμάτων υποστήριξης	

Στάδια σωλήνωσης			
Ταξινομητές	Μηχανές Διανυσμάτων Υποστήριξης	C	0.5
			1
			5
			10
			15
			20
			30
		kernel	linear
			rbf
			poly
			sigmoid
		gamma	scale
			0.0001
			0.001
			0.01
			0.1
			1
	Λογιστική Παλινδρόμηση	solver	sag
			liblinear
			saga
			newton-cg
			lbfgs
		penalty	l1
			l2
			elasticnet
		C	0.5
			1
			5
			10
			15
			20
			30
		l1_ratio	0
			0.2
			0.5
			0.7
			1
	Δέντρα ενίσχυσης κλίσης	learning_rate	0.0001
			0.001
			0.01
			0.1
		n_estimators	100
			200
			300
			400
			500
	Τυχαία δάση	n_estimators	100
			200
			300
			400
			500
		criterion	gini
			entropy

Πίνακας 4.1: Στάδια σωλήνωσης.

Εδώ φαίνονται όλα τα στάδια σωλήνωσης με την σειρά που εφαρμόστηκαν στα δεδομένα. Η σειρά δεν είναι απόλυτη καθώς αρχικά σε πολλές από τις δοκιμές δεν έγινε η χρίση όλων των σταδίων και στη συνέχεια ανάλογα με την τεχνική που εφαρμόστηκε στο εκάστοτε στάδιο μπορεί αυτό να μετακινούνταν πιο κοντά στην αρχή της σωλήνωσης. Για την επιλογή μέσο μοντέλων οι παράμετροι ήταν οι ίδιοι με τους απλούς ταξινομητές. Τέλος το πλήθος των χαρακτηριστικών που επιλέχτηκαν έπαιρνε τιμές 30,60,85,100,130,168.

4.4 Εργαλεία

Η υλοποίηση της παρούσας εργασίας έγινε σε Python3. Η γλώσσα αυτή επιλέχθηκε καθώς παρέχει μία μεγάλη ποικιλία από βιβλιοθήκες για τη στατιστική επεξεργασία των δεδομένων, πράξεων πινάκων και αλγορίθμων μηχανικής μάθησης. Οι βιβλιοθήκες που χρησιμοποιήθηκαν ήταν οι NumPy, Scipy, Statsmodels, DTAIDistance, Imbalanced-learn, Pandas, Scikit-learn, LightGBM, Geneticalgorithm, PyGaze και Seaborn, των οποίων η χρήση αναλύεται συνοπτικά.

- NumPy: Περιέχει συναρτήσεις και κλάσεις που υποστηρίζουν πράξεις πολυδιάστατων πινάκων, όπως και συναρτήσεις στο πεδίο της γραμμικής άλγεβρας. Οι πίνακες οι οποίοι δημιουργούνται από αυτήν παρουσιάζουν καλύτερη επίδοση από τις λίστες της Python σε επίπεδο μνήμης και ταχύτητας.

- Scipy: Αποτελεί μία επέκταση των δυνατοτήτων της NumPy, καθώς έχει γραφτεί με τη χρήση αυτής, αλλά υποστηρίζει παραπάνω επιστημονικές μεθόδους και υπολογισμούς όπως η επεξεργασία σημάτων.

- Statsmodels: Αυτή η βιβλιοθήκη προσφέρει μία γκάμα από στατιστικές μεθόδους, μοντέλα και τεστ αλλά και δυνατότητες στατιστικής εξερεύνησης των δεδομένων.

- Pandas: Το pandas είναι μια βιβλιοθήκη για το χειρισμό και την ανάλυση δεδομένων. Συγκεκριμένα, προσφέρει δομές δεδομένων και λειτουργίες για χειρισμό αριθμητικών πινάκων και χρονοσειρών. Με την χρήση του pandas γίνεται πιο εύκολα η διαχείριση των χαρακτηριστικών σε συνδυασμό με το όνομά τους, καθώς και ένα πρώτο είδος επεξεργασίας αυτών.

- DTAIDistance: Πρόκειται για μία βιβλιοθήκη μεθόδων εύρεσης απόστασης μεταξύ χρονοσειρών. Σε αυτήν υλοποιείται η δυναμική χρονική στρέβλωση.

- Scikit-learn: Είναι μία βιβλιοθήκη ανοιχτού κώδικα της Python που περιέχει την υλοποίηση των περισσότερων από τους αλγόριθμους μηχανικής μάθησης που αναφέρθηκαν. Ακόμα περιέχει μεθόδους και αλγόριθμους κλιμάκωσης, επιλογής χαρακτηριστικών και γενικότερα μείωση της διαστατικότητας και τη δημιουργία αρχιτεκτονικών σωλήνωσης, όπως και τεχνικές αξιολόγησης μοντέλων. Η υλοποίηση όλων των αλγορίθμων μηχανικής μάθησης που χρησιμοποιήθηκαν εκτός των δέντρων ενίσχυσης κλίσης περιέχεται σε αυτή τη βιβλιοθήκη.

- Imbalanced-learn: Η βιβλιοθήκη αυτή βασίζεται στην προηγούμενη και παρέχει εργαλεία για την εκπαίδευση μοντέλων σε μη ισορροπημένες κλάσεις. Αυτή η βιβλιοθήκη παρείχε τις μεθόδους της δειγματοληψίας.

- LightGBM: Είναι μία βιβλιοθήκη που περιέχει την υλοποίηση μηχανών ενίσχυσης κλίσης για μηχανική μάθηση που αναπτύχθηκε αρχικά από τη Microsoft. Βασίζεται σε αλγόριθμους δέντρων αποφάσεων και χρησιμοποιείται για την κατάταξη, την ταξινόμηση και άλλες εργασίες.

- Geneticalgorithm: Πρόκειται για μία βιβλιοθήκη στην οποία υλοποιούνται γενετικοί αλγόριθμοι.
- PyGaze: Είναι ένα ανοιχτού κώδικα εργαλείο με σκοπό την εύκολη ανάπτυξη κώδικα για πειράματα οφθαλμικής ανίχνευσης. Περιέχει μεθόδους για την εύρεση σακκάδων και φάσεων εστίασης.
- Seaborn: Η βιβλιοθήκη αυτή χρησιμοποιήθηκε καθώς παρέχει ένα πλήθος μεθόδων για την οπτικοποίηση και παρουσίαση των δεδομένων.

Κεφάλαιο 5

Αποτελέσματα και συμπεράσματα

Μετά από το θεωρητικό πλαίσιο και την ανάλυση της κάθε μεθόδου που αναφέρθηκαν στα προηγούμενα κεφάλαια, καθώς και κατόπιν της γεφύρωσης αυτών των μεθόδων με το πρόβλημα που κλήθηκε να λύσει η παρούσα εργασία, φυσικό είναι να παρατεθούν στον αναγνώστη και τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν για τη διάγνωση της μαθησιακής δυσκολίας.

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα των αλγορίθμων που χρησιμοποιήθηκαν και γίνεται μία σύγκριση των αλγορίθμων μεταξύ τους. Στη συνέχεια γίνεται ένας σχολιασμός των μεθόδων και του κατά πόσον αυτές είναι ικανές να εφαρμοστούν για τη διάγνωση της δυσλεξίας.

Για το παρόν πρόβλημα σημαντικότερη μετρική αποτελεί η ανάκληση και ακολουθεί η ακρίβεια της κλάσης των ατόμων με δυσλεξία. Το παραπάνω προκύπτει από το γεγονός ότι είναι προτιμότερο να γίνει περαιτέρω διερεύνηση από κάποιον ειδικό στην περίπτωση ενός ατόμου που δεν έχει δυσλεξία, παρά να μην δοθεί σημασία στην περίπτωση κάποιου ατόμου που έχει δυσλεξία. Έτσι οι μετρικές που παρουσιάζονται, εκτός της ορθότητας (που δεν ορίζεται ανά κλάση), είναι για την κλάση των ατόμων με δυσλεξία.

5.1 Δεδομένα οφθαλμικής ανίχνευσης

5.1.1 Μετρικές αξιολόγησης

Στην παρούσα υποενότητα οι μετρικές αξιολόγησης είναι αυτές των καλύτερων από τους συνδυασμούς των αλγορίθμων που εφαρμόστηκαν. Λόγω του πλήθους των συνδυασμών δεν γίνεται η προβολή όλων των αποτελεσμάτων.

Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Τυχαία δάση	86.12%	87.84%	87.93%	87.77%
Λογιστική παλινδρόμηση	82.71%	83.02%	84.18%	84.11%
Μηχανές Διανυσμάτων Υποστήριξης	85.87%	86.23%	87.09%	86.44%

Πίνακας 5.1: Μετρικές αξιολόγησης ταξινομητών.

Ο πρώτος πίνακας που παρουσιάζεται αφορά την εκπαίδευση των μοντέλων στα δεδομένα, χωρίς καμία προεπεξεργασία εκτός της εξαγωγής χαρακτηριστικών. Οι μετρικές παρουσιάζουν σχετικά υψηλές τιμές με μέγιστη να εμφανίζεται στα τυχαία δάση, με 86.12% στην ορθότητα, 87.84% στο F1 score, 87.93% στην ακρίβεια και 87.77% στην ανάκληση.

Δυναμική χρονική στρεύλωση				
Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Τυχαία δάση	83.97%	83.18%	86.11%	81.77%
Λογιστική παλινδρόμηση	78.07%	77.28%	80.92%	75.22%
Μηχανές Διανυσμάτων Υποστήριξης	72.92%	69.19%	80.57%	63.0%

Πίνακας 5.2: Μετρικές αξιολόγησης ταξινομητών με δυναμική χρονική στρεύλωση.

Στον δεύτερο πίνακα παρουσιάζονται τα μοντέλα με χρήση δυναμικής χρονικής στρέβλωσης όπως αυτή αναφέρθηκε στην αντίστοιχη υποενότητα. Παρατηρείται πως τα αποτελέσματα είναι χαμηλότερα από την απλή εκπαίδευση των μοντέλων στα δεδομένα, με μέγιστα πάλι να εμφανίζονται στα τυχαία δάση, 83.97% στην ορθότητα, 83.18% στο F1 score, 86.11% στην ακρίβεια και 81.77% στην ανάκληση. Το παραπάνω φαινόμενο συμβαίνει καθώς με την τεχνική της δυναμικής χρονικής στρέβλωσης πιθανόν να υπάρχει αλλοίωση στο θόρυβο του ματιού που οφείλεται στη δυσλεξία. Έτσι χάνεται πληροφορία η οποία θα βοηθούσε τα μοντέλα να διακρίνουν τα άτομα στις δύο ομάδες. Παραπάνω τα αποτελέσματα των μοντέλων που εκπαιδεύτηκαν με δυναμική χρονική στρέβλωση των δεδομένων δεν παρουσιάζονται καθώς είναι κατώτερα των αντίστοιχων μοντέλων που εκπαιδεύτηκαν με τα δεδομένα χωρίς προεπεξεργασία.

Μετά τη χρήση προεπεξεργασίας(Πίνακας 5.3) μπορούμε να παρατηρήσουμε πώς η επίδοση των ταξινομητών έχει μικρή αύξηση. Για κάθε ταξινομητή οι αντίστοιχες μετρικές αξιολόγησης έχουν πολύ κοντινές τιμές, με τις μηχανές διανυσμάτων υποστήριξης με χρήση γενετικών αλγορίθμων να παρουσιάζουν τις καλύτερες 92.63% στην ορθότητα, 92.43% στο F1 score, 95.63% στην ακρίβεια και 90% στην ανάκληση.

5.1.2 Σημαντικότερα χαρακτηριστικά

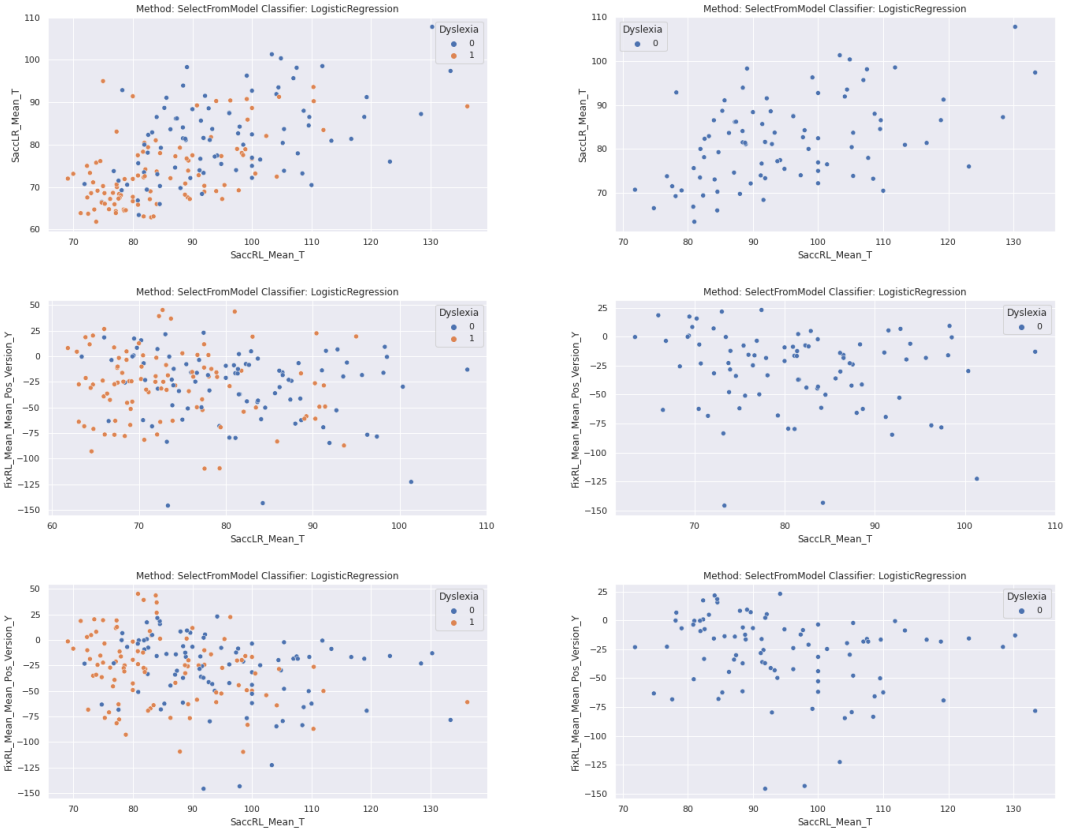
Όπως έχει είδη αναφερθεί στις διαφορετικές αρχιτεκτονικές σωλήνωσης που εξετάστηκαν υπήρχε ένα στάδιο επιλογής χαρακτηριστικών. Παρακάτω παρατίθενται τα διαγράμματα των τριών σημαντικότερων χαρακτηριστικών από τις μεθόδους που παρουσιάστηκαν στους παραπάνω πίνακες. Κάθε σημείο που εμφανίζεται σε αυτά αντιστοιχεί σε ένα άτομο. Τα άτομα με δυσλεξία παρουσιάζονται με πορτοκαλί, ενώ τα άτομα χωρίς με μπλε. Στη αριστερή στήλη των διαγραμμάτων παρουσιάζονται και οι δύο κλάσεις, ενώ στη δεξιά μόνο η κλάση των ατόμων χωρίς δυσλεξία. Τα διαγράμματα των χαρακτηριστικών έγιναν ανά δύο καθώς έτσι φαίνεται η κατανομή των σημείων σε κάθε άξονα, αλλά και τυχόν υποχώροι που μπορεί να λαμβάνει υπόψη του το εκάστοτε μοντέλο.

Τα αποτελέσματα της επιλογής χαρακτηριστικών μέσω του γενετικού αλγορίθμου δεν

Ορθότητα						
Σωλήνωση		Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
	Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Τυχαία δάση	91.34%	91.53%	94.4%	89.88%
Κανονική κλιμάκωση	Αναδρομική επιλογή χαρακτηριστικών μέσω μηχανών διανυσμάτων υποστήριξης	Λογιστική παλινδρόμηση	91.87%	91.79%	95.77%	88.88%
	Γενετικός αλγόριθμος	Μηχανές Διανυσμάτων Υποστήριξης	92.63%	92.43%	95.63%	90%

Πίνακας 5.3: Μετρικές αξιολόγησης σωλήνωσεων.

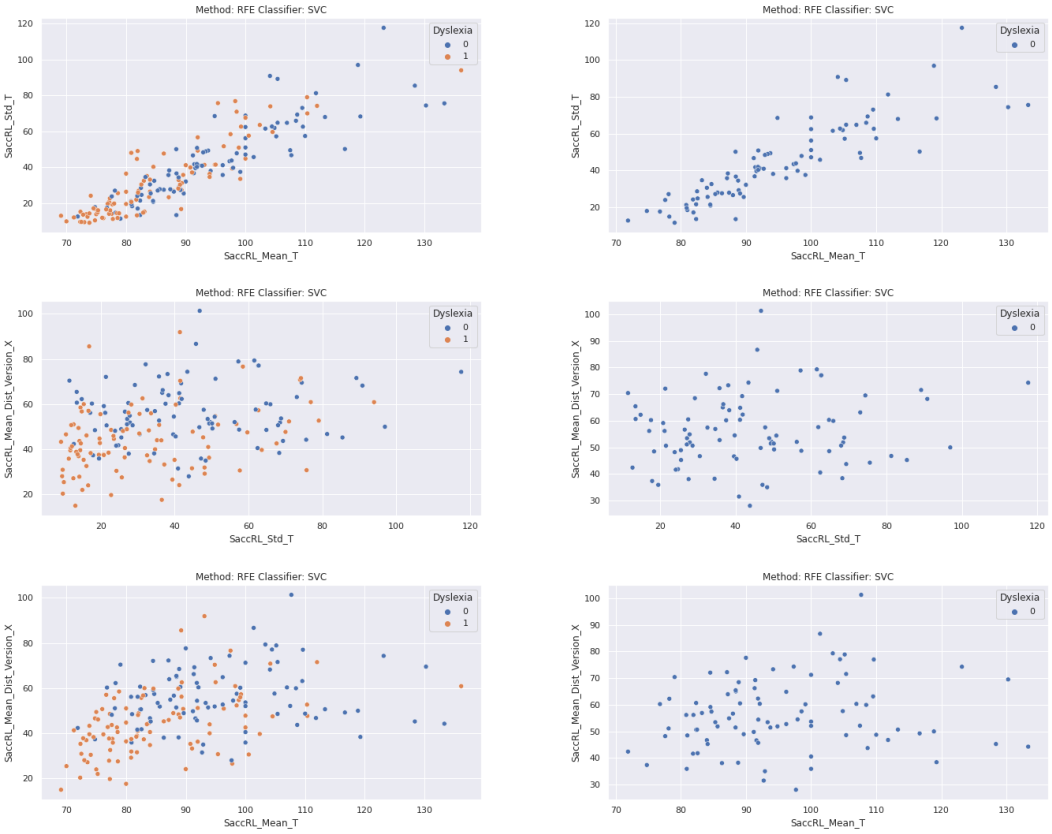
μπορούν να οπτικοποιηθούν καθώς ο αλγόριθμος δεν επιστρέφει κάποια πληροφορία για το κατά πόσο αυτά είναι σημαντικά για την εκπαίδευση των μοντέλων ή κάποια στατιστική μετρική.



Σχήμα 5.1: Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης.

Από τα παραπάνω διαγράμματα κάποιος εύκολα μπορεί να διακρίνει πως τα χαρακτηριστικά που κρίνουν τα μοντέλα ως τα τρία σημαντικότερα δημιουργούν ήδη χωρία όπου οι δύο κλάσεις αν και δεν είναι γραμμικά διαχωρίσιμες είναι εύκολα διακριτές μεταξύ τους. Η διαπίστωση αυτή δικαιολογεί και το γιατί η λογιστική παλινδρόμηση, αν και γραμμικό μοντέλο καταφέρνει να έχει παρόμοια επίδοση με τα υπόλοιπα μοντέλα.

Τα χαρακτηριστικά που τα παραπάνω μοντέλα έκριναν ως σημαντικότερα ήταν η μέση διάρκεια των σακκάδων από τα δεξιά προς τα αριστερά, η μέση διάρκεια των σακκάδων από τα αριστερά προς τα δεξιά, η τυπική απόκλιση της διάρκειας των σακκάδων από τα δεξιά προς τα αριστερά και η μέση απόσταση που διανύεται από τις σακκάδες με διεύθυνση από δεξιά πως τα αριστερά ως προς την μέση θέση των δύο ματιών. Τα παραπάνω χαρακτηριστικά ήταν αναμενόμενο να εμφανιστούν ως σημαντικότερα καθώς είναι και αυτά που σύμφωνα με τη θεωρία διαχωρίζουν τους αναγνώστες ελέγχου από τους δυσλεκτικούς. Συμπληρωματικά με τα παραπάνω εμφανίζεται η μέση τιμή των μέσων θέσεων των φάσεων εστίασης με διεύθυνση από δεξιά προς τα αριστερά στον κάθετο άξονα.



Σχήμα 5.2: Αναδρομική επιλογή χαρακτηριστικών μέσω μηχανών διανυσμάτων υποστήριξης.

5.2 Δεδομένα παιχνιδοποίησης

5.2.1 Μετρικές αξιολόγησης

Στην παρούσα υποενότητα οι μετρικές αξιολόγησης είναι αυτές των καλύτερων από τους συνδυασμούς των αλγορίθμων που εφαρμόστηκαν. Λόγω του πλήθους των συνδυασμών δεν γίνεται η προβολή όλων των αποτελεσμάτων.

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των ταξινομητών που εκπαιδεύτηκαν στα δεδομένα χωρίς καμία επεξεργασία.

Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Τυχαία δάση	89.81%	14.04%	76.88%	7.92%
Λογιστική παλινδρόμηση	89.46%	26.65%	53.05%	18.09%
Δέντρα ενίσχυσης κλίσης	90.53%	30.64%	72.58%	19.62%
Μηχανές Διανυσμάτων Υποστήριξης	89.21%	0.0%	0.0%	0.0%

Πίνακας 5.4: Μετρικές αξιολόγησης ταξινομητών.

Όπως φαίνεται από τον πίνακα όλοι οι ταξινομητές έχουν υψηλή ορθότητα, κοντά στο 90% (90.53% από τα δέντρα ενίσχυσης κλίσης), αυτό συμβαίνει καθώς η κλάση των ατόμων

χωρίς δυσλεξία είναι πολύ μεγαλύτερη από αυτή των ατόμων με δυσλεξία. Έτσι οι ταξινομητές θεωρώντας την πλειονότητα των ατόμων ως μη δυσλεκτικά επιτυγχάνουν υψηλή ορθότητα. Όμως αν ρίξει κάποιος μία ματιά στις μετρικές ακρίβεια, ανάκλιση και F1 score της κλάσης των ατόμων με δυσλεξία μπορεί να διακρίνει ότι οι μετρικές είναι αρκετά χαμηλές.

Αρχικά η μετρική του μεγαλύτερου ενδιαφέροντος που είναι η ανάκληση είναι αρκετά χαμηλή με μέγιστο 19.62% από τα δέντρα ενίσχυσης με κλίση και 18.09% από την λογιστική παλινδρόμηση. Δηλαδή από το σύνολο των ατόμων με δυσλεξία ο ταξινομητής κατάφερε να τοποθετήσει στη σωστή κλάση μόνο το 19%. Η χειρότερη περίπτωση που παρατηρείται είναι των μηχανών διανυσμάτων υποστήριξης που είναι 0. Το παραπάνω σημαίνει πως ο ταξινομητής τοποθέτησε όλα τα δείγματα στην κλάση των ατόμων χωρίς δυσλεξία.

Επόμενη μετρική ενδιαφέροντος είναι η ακρίβεια. Σε αυτή παρατηρούνται σχετικά υψηλές τιμές έως και 76.88% από τα τυχαία δάση και 72.58% από τα δέντρα ενίσχυσης με κλίση. Το παραπάνω ερμηνεύεται ως ότι από τα δείγματα που ταξινομήθηκαν στην κλάση των ατόμων με δυσλεξία, τα 76.88% ή 72.58% ταξινομήθηκαν σωστά.

Το F1 score είναι ο αρμονικός μέσος των παραπάνω δύο. Μέγιστη τιμή σε αυτόν έχουν τα δέντρα ενίσχυσης με κλίση τα οποία εμφάνισαν υψηλότερες τιμές και στις άλλες δύο.

Ορθότητα

Στον πίνακα 5.5 η βελτιστοποίηση του μοντέλου έχει γίνει ως προς την ορθότητα. Με μία πρώτη ματιά δεν έχει γίνει σημαντική βελτίωση των μετρικών αξιολόγησης. Το μέγιστο της ορθότητας έχει μείνει κοντά στο 90%, της ακρίβειας στο 77% και της ανάκλησης στο 21%. Σε αντίθεση όμως με την απλή εκπαίδευση των μοντέλων στα δεδομένα, εδώ όλοι οι ταξινομητές εκτός των μηχανών διανυσμάτων υποστήριξης έχουν ακρίβεια πάνω από 66.66% (λογιστική παλινδρόμηση) και ανάκληση πάνω από 19.37% (λογιστική παλινδρόμηση).

Όπως και πάνω έτσι και εδώ οι μηχανές διανυσμάτων υποστήριξης εμφανίζουν την χαμηλότερη τιμή στην ανάκληση αλλά την υψηλότερη στην ακρίβεια. Αυτό σημαίνει πως τοποθέτησαν τα λιγότερα άτομα στην κλάση των ατόμων με δυσλεξία, αλλά τα περισσότερα από αυτά ήταν σωστά τοποθετημένα. Τα δέντρα ενίσχυσης με κλίση μαζί με τα τυχαία δάση εμφάνισαν το μεγαλύτερο F1 score.

Ακόμα παρατηρείται πως το επίπεδο της κλιμάκωσης δεν παίζει σημαντικό ρόλο στο τελικό αποτέλεσμα. Το παραπάνω προκύπτει από το γεγονός ότι αρχικά για παρόμοια αποτελέσματα μετρικών έχουμε και τις δύο μορφές κλιμάκωσης που έχουν αναφερθεί. Τέλος στην περίπτωση των δέντρων ενίσχυσης με κλίση τα αποτελέσματα και χωρίς κλιμάκωση είναι παρόμοια.

Αναφορικά με την επιλογή χαρακτηριστικών παρατηρείται ότι τα μοντέλα που δίνουν τα καλύτερα αποτελέσματα είναι τα δέντρα ενίσχυσης με κλίση και η λογιστική παλινδρόμηση, ενώ η μέθοδος επιλογής είναι η αναδρομική επιλογή χαρακτηριστικών.

Στο επίπεδο της δειγματοληψίας για τα τυχαία δάση με τη χρήση υπερδειγματοληψίας δόθηκε η μεγαλύτερη ορθότητα.

Ακρίβεια και Ανάκληση

Από τους πίνακες 5.6 και 5.7 προκύπτουν τα εξής. Η πρώτη και κύρια παρατήρηση που αφορά τις παραπάνω μετρικές βρίσκεται στο επίπεδο της δειγματοληψίας. Η ακρίβεια βελτιστοποιείται χωρίς καθόλου δειγματοληψία, ενώ η ανάκληση με υποδειγματοληψία. Αυτό συμβαίνει καθώς αφαιρώντας δείγματα της κλάσης των ατόμων με δυσλεξία (η κλάση με τα περισσότερα δείγματα), γίνεται πιο εύκολο για τα μοντέλα να προβλέψουν άτομα τα οποία έχουν δυσλεξία, καθώς η δεύτερη κλάση αποκτά μεγαλύτερη ισχύ. Συνέπεια της παραπάνω διαδικασίας είναι, ότι δείγματα τα οποία ανήκουν στην κλάση των ατόμων χωρίς δυσλεξία που βρίσκονται κοντά στα δείγματα των ατόμων με δυσλεξία πλέον ταξινομούνται ως δυσλεκτικά, έτσι παρατηρείται πτώση της ακρίβειας και αύξηση της ανάκλησης.

Επόμενη παρατήρηση είναι πως και εδώ η κλιμάκωση δεν παίζει ιδιαίτερο ρόλο στα αποτελέσματα καθώς και πως σε επίπεδο επιλογής χαρακτηριστικών τα κύρια μοντέλα είναι τα δέντρα ενίσχυσης κλίσης και η λογιστκή παλινδρόμηση. Σε αντίθεση με την προηγούμενη μετρική, καλά αποτελέσματα δίνει και η επιλογή χαρακτηριστικών μέσω τυχαίων δασών και η κύρια μέθοδος είναι η απλή επιλογή των χαρακτηριστικών μέσω μοντέλων.

Τέλος οι ταξινομητές που δίνουν τα καλύτερα αποτελέσματα, αρχικά για την μετρική της ακρίβειας, είναι οι μηχανές διανυσμάτων υποστήριξης, με ακρίβεια 87.5% και τα τυχαία δάση 84.02%. Τα δεύτερα εμφανίζουν και μεγαλύτερη ανάκληση 11.75% από τις μηχανές διανυσμάτων υποστήριξης 6.37%. Η ορθότητα όλων των μοντέλων δεν έχει μεγάλη απόκλιση από την βελτιστοποίηση σε επίπεδο ακρίβειας και είναι γύρω στο 90%. Το F1 score εμφανίζει μέγιστο στα δέντρα ενίσχυσης κλίσης 33.6%.

Σε επίπεδο ανάκλησης όλα τα μοντέλα έχουν παρόμοιες τιμές από 69.89% (μηχανές διανυσμάτων υποστήριξης) μέχρι 73.23% (δέντρα ενίσχυσης κλίσης). Η ακρίβεια είναι αρκετά χαμηλότερη από αυτή που αναφέρθηκε πριν, από 30.02%(λογιστική παλινδρόμηση) έως 34.55%(τυχαία δάση). Αντιθέτως όμως με προηγουμένως το F1 score είναι καλύτερο από πριν, 42.18%(λογιστική παλινδρόμηση) έως 46.79%(τυχαία δάση), όμως εμφανίζεται χαμηλότερη ορθότητα από όλα τα μοντέλα γύρω στο 80%.

F1 score

Στον πίνακα 5.8 παρατηρείται συμπληρωματικά σε ότι έχει αναφερθεί στην παρούσα υποενότητα, η εμφάνιση υποδειγματοληψίας μαζί με υπερδειγματοληψία, καθώς και η επιλογή χαρακτηριστικών με μηχανές διανυσμάτων υποστήριξης.

Όλα τα μοντέλα παρουσιάζουν παρόμοια F1 score, 47.03% (τυχαία δάση) έως 52.58% (δέντρα ενίσχυσης κλίσης). Όπως και στην προηγούμενη ενότητα έτσι και εδώ παρατηρείται μία ανταλλαγή μεταξύ ακρίβειας και ανάκλησης. Το υψηλότερο F1 score αλλά και η ενδιάμεση τιμή ανάμεσα στης δύο μετρικές δίνουν τα δέντρα ενίσχυσης κλίσης 52.58% με ακρίβεια 50.13% και ανάκληση 56.11%. Αυτό σημαίνει πως προέβλεψαν σωστά λίγο πάνω από τα μισά άτομα που είχαν δυσλεξία, και από τα συνολικά άτομα που προέβλεψαν ως δυσλεκτικά, τα μισά ήταν σωστά. Η ορθότητα για το παραπάνω αποτέλεσμα ήταν 89.21%, ίδια με το αν είχε θέσει ο ταξινομητής όλα τα άτομα ως μη δυσλεκτικά.

Ορθότητα							
Σωλήνωση			Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Τυχαία Υπερδειγματοληψία	Κλιμάκωση ελαχίστου μεγίστου	Αναδρομική επιλογή χαρακτηριστικών μέσω δέντρων ενίσχυσης κλίσης	Τυχαία δάση	90.88%	34.8%	76.64%	22.68%
	Κλιμάκωση ελαχίστου μεγίστου	Αναδρομική επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Λογιστική παλινδρόμηση	90.28%	29.71%	66.66%	19.37%
	Κανονική κλιμάκωση	Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Δέντρα ενίσχυσης κλίσης	90.8%	33.6%	77.37%	21.67%
	Κανονική κλιμάκωση	Αναδρομική επιλογή χαρακτηριστικών μέσω δέντρων ενίσχυσης κλίσης	Μηχανές Διανυσμάτων Υποστήριξης	89.95%	15.93%	81.75%	8.92%

Πίνακας 5.5: Μετρικές αξιολόγησης σωλήνωσεων με βελτιστοποίηση ως προς την ορθότητα.

Ακρίβεια						
Σωλήνωση		Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Κανονική κλιμάκωση	Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Τυχαία δάση	90.23%	20.23%	84.02%	11.75%
Κλιμάκωση ελαχίστου μεγίστου	Απλή επιλογή χαρακτηριστικών μέσω δέντρων ενίσχυσης κλίσης	Λογιστική παλινδρόμηση	89.95%	21.82%	68.24%	13.26%
Κανονική κλιμάκωση	Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Δέντρα ενίσχυσης κλίσης	90.8%	33.6%	77.37%	21.67%
Κανονική κλιμάκωση	Αναδρομική επιλογή χαρακτηριστικών μέσω τυχαίων δασών	Μηχανές Διανυσμάτων Υποστήριξης	89.76%	11.75%	87.5%	6.37%

Πίνακας 5.6: Μετρικές αξιολόγησης σωλήνωσεων με βελτιστοποίηση ως προς την ακρίβεια.

Ανάκληση							
Σωλήνωση			Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Τυχαία Υποδειγματοληψία	Κλιμάκωση ελαχίστου μεγίστου	Απλή επιλογή χαρακτηριστικών μέσω δέντρων ενίσχυσης κλίσης	Τυχαία δάση	82.21%	46.79%	34.55%	72.7%
Τυχαία Υποδειγματοληψία	Κανονική κλιμάκωση	Απλή επιλογή χαρακτηριστικών μέσω τυχαίων δασών	Λογιστική παλινδρόμηση	78.81%	42.18%	30.02%	71.43%
Τυχαία Υποδειγματοληψία	Κλιμάκωση ελαχίστου μεγίστου	Απλή επιλογή χαρακτηριστικών μέσω τυχαίων δασών	Δέντρα ενίσχυσης κλίσης	81.36%	45.72%	33.29%	73.23%
Τυχαία Υποδειγματοληψία	Κλιμάκωση ελαχίστου μεγίστου	Αναδρομική επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Μηχανές Διανυσμάτων Υποστήριξης	80.65%	43.84%	32.06%	69.89%

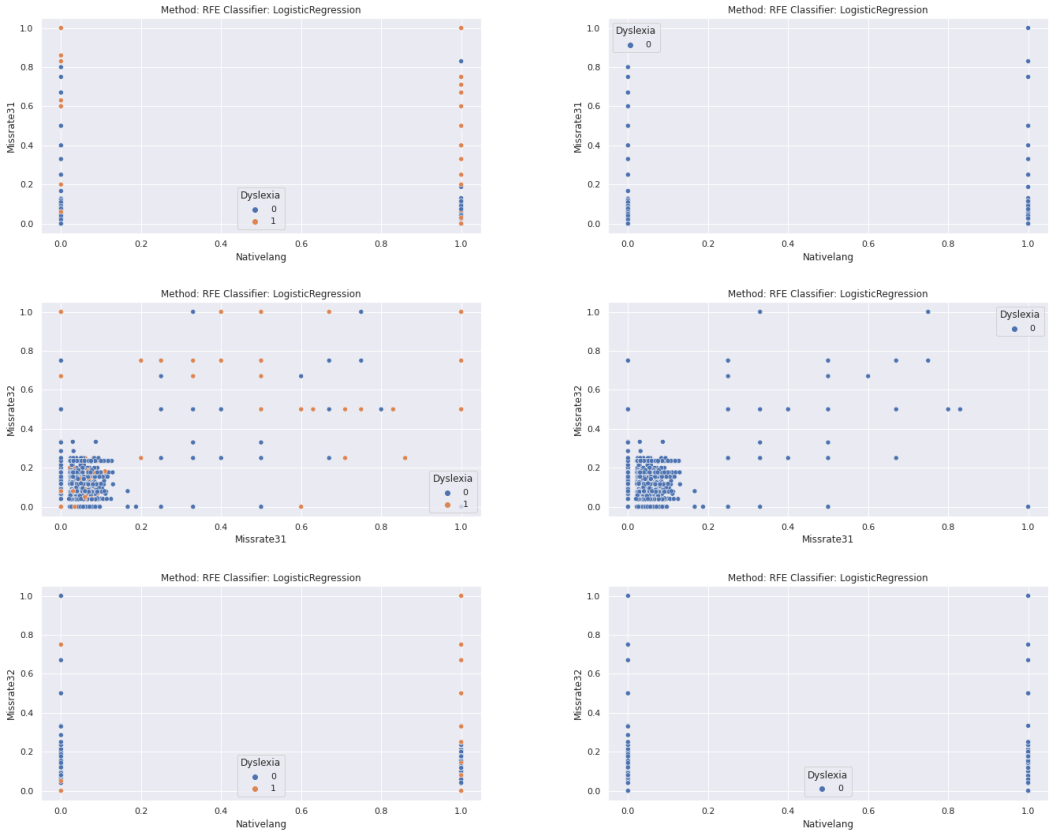
Πίνακας 5.7: Μετρικές αξιολόγησης σωλήνωσεων με βελτιστοποίηση ως προς την ανάκληση.

F1							
Σωλήνωση			Ταξινομητής	Ορθότητα	F1	Ακρίβεια	Ανάκληση
Τυχαία Υποδειγματοληψία	Κανονική κλιμάκωση	Αναδρομική επιλογή χαρακτηριστικών μέσω μηχανών διανυσμάτων υποστήριξης	Τυχαία δάση	83.01%	47.03%	35.53%	69.89%
Τυχαία Υπερδειγματοληψία	Κλιμάκωση ελαχίστου μεγίστου	Αναδρομική επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Λογιστική παλινδρόμηση	86.44%	48.44%	41.11%	59.18%
	Τυχαία Υπερδειγματοληψία	Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης	Δέντρα ενίσχυσης κλίσης	89.21%	52.58%	50.13%	56.11%
Τυχαία Υπερδειγματοληψία	Κανονική κλιμάκωση	Αναδρομική επιλογή χαρακτηριστικών μέσω τυχαίων δασών	Μηχανές Διανυσμάτων Υποστήριξης	87.98%	49.59%	45.11%	55.35%

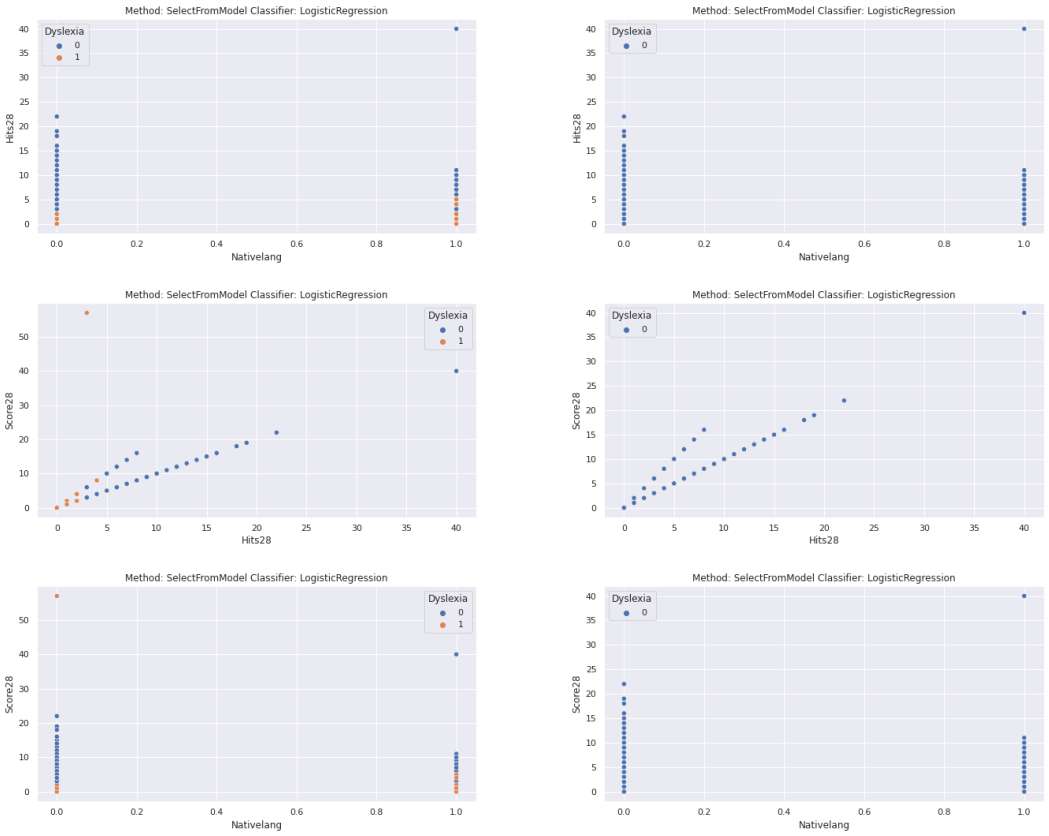
Πίνακας 5.8: Μετρικές αξιολόγησης σωλήνωσεων με βελτιστοποίηση ως προς το F1 score.

5.2.2 Σημαντικότερα χαρακτηριστικά

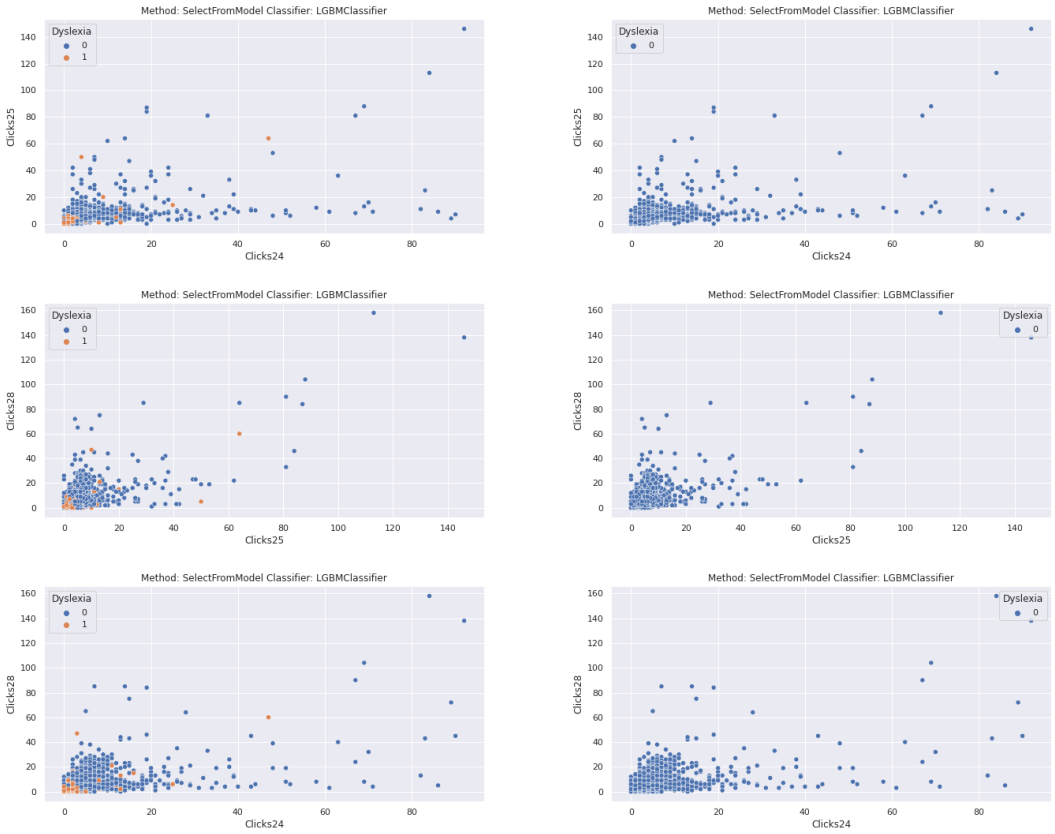
Όπως έχει ήδη αναφερθεί, στις αρχιτεκτονικές σωλήνωσης που εξετάστηκαν υπήρχε ένα στάδιο επιλογής χαρακτηριστικών. Παρακάτω παρουσιάζονται τα διαγράμματα των τριών σημαντικότερων χαρακτηριστικών από τις μεθόδους που παρουσιάστηκαν στους παραπάνω πίνακες. Κάθε σημείο που εμφανίζεται σε αυτά αντιστοιχεί σε ένα άτομο. Τα άτομα με δυσλεξία παρουσιάζονται με πορτοκαλί, ενώ τα άτομα χωρίς με μπλε. Στη αριστερή στήλη των διαγραμμάτων παρουσιάζονται και οι δύο κλάσεις, ενώ στη δεξιά μόνο η κλάση των ατόμων χωρίς δυσλεξία. Τα διαγράμματα των χαρακτηριστικών έγιναν ανά δύο, καθώς έτσι φαίνεται η κατανομή των σημείων σε κάθε άξονα, αλλά και τυχόν υποχώροι που μπορεί να λαμβάνει υπόψη του το εκάστοτε μοντέλο.



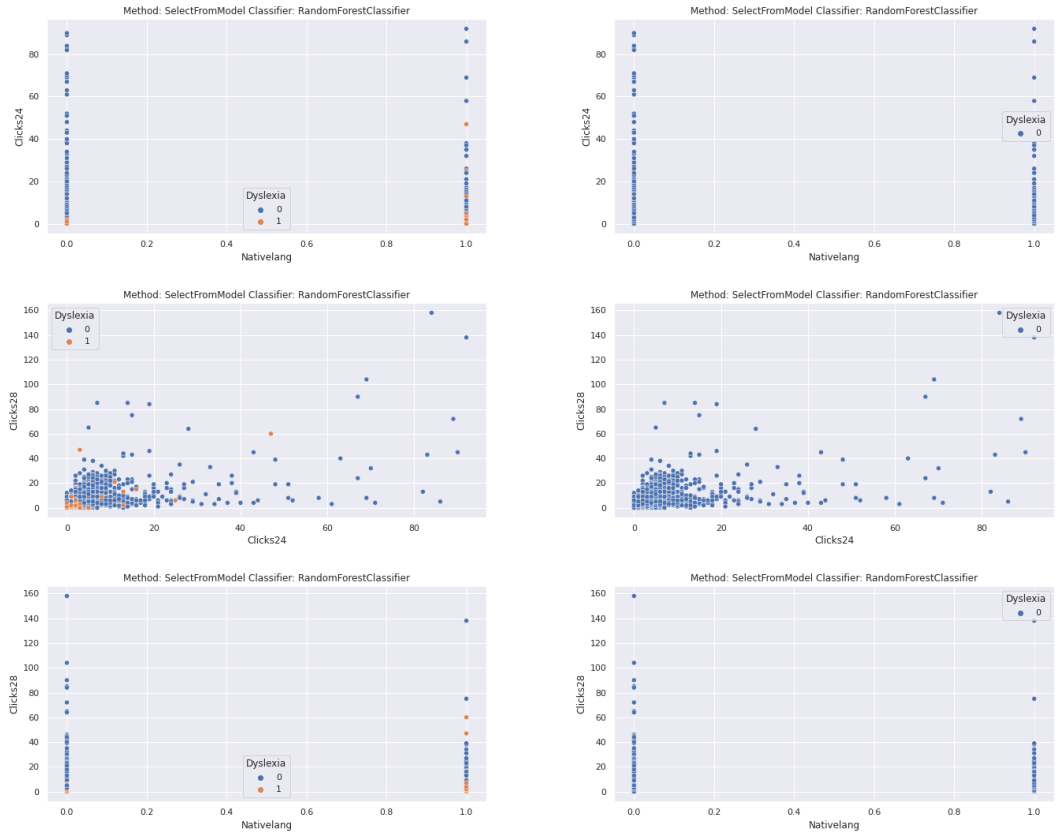
Σχήμα 5.3: Αναδρομική επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης.



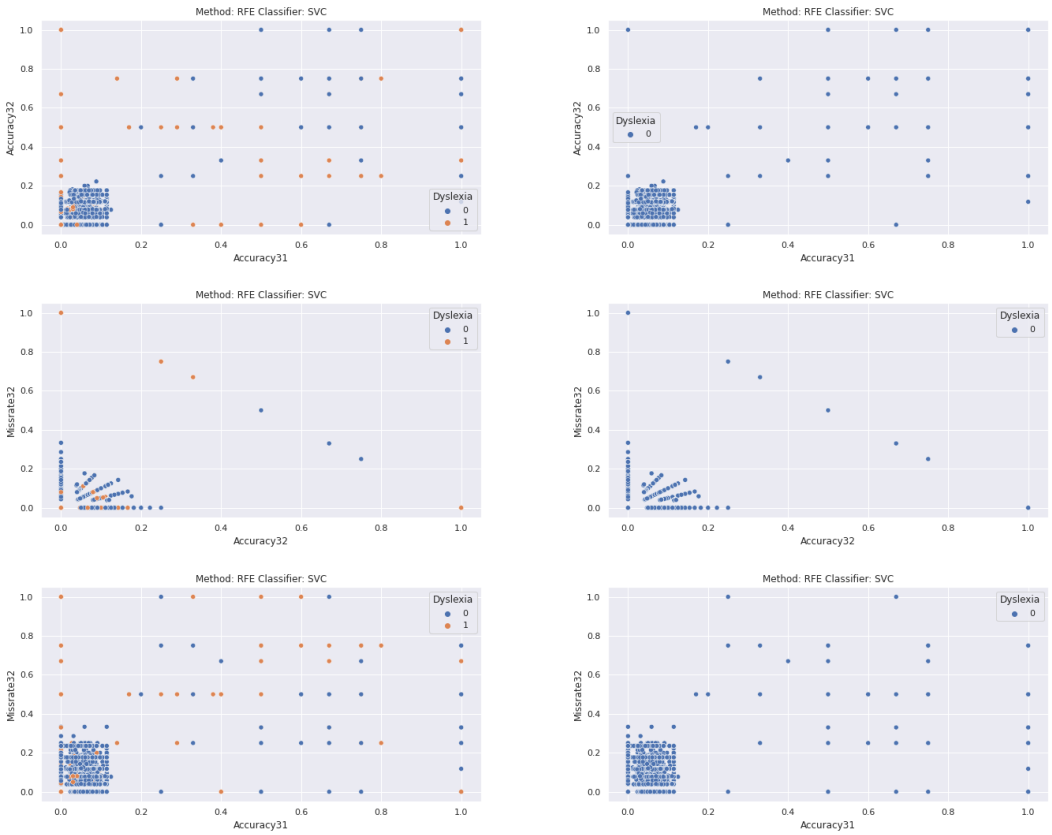
Σχήμα 5.4: Απλή επιλογή χαρακτηριστικών μέσω λογιστικής παλινδρόμησης.



Σχήμα 5.5: Αναδρομική και απλή επιλογή χαρακτηριστικών μέσω δέντρων ενίσχυσης κλίσης.



Σχήμα 5.6: Αναδρομική και απλή επιλογή χαρακτηριστικών μέσω τυχαίων δασών.



Σχήμα 5.7: Αναδρομική και απλή επιλογή χαρακτηριστικών μέσω μηχανώ διανυσμάτων υποστήριξης.

Από τα παραπάνω διαγράμματα εύκολα γίνεται εμφανής ο λόγος για τον οποίο η δειγματοληψία είναι σημαντική για την εκπαίδευση των μοντέλων, αλλά και ο λόγος για τον οποίο αύξηση στην ανάκληση οδηγεί σε μείωση της ακρίβειας και αντιστρόφως. Τα δείγματα των ατόμων με δυσλεξία (στη γενική περίπτωση) δεν παρουσιάζονται ως θόρυβος σε σχέση με τα δείγματα των ατόμων ελέγχου όπως ήταν η αρχική υπόθεση.

Όπως φαίνεται από τα διαγράμματα τα δείγματα αυτά δεν είναι εύκολα διαχωρίσιμα από τα δείγματα ελέγχου καθώς τα δύο σύνολα είναι αλληλοεπικαλυπτόμενα. Σπανίως φαίνεται να υπάρχουν σημεία ατόμων με δυσλεξία που αν αφαιρεθούν δεν βρίσκεται στη θέση τους δείγμα ατόμου χωρίς δυσλεξία. Το παραπάνω φαινόμενο ισχύει για τα τρία σημαντικότερα χαρακτηριστικά που προέκυψαν από την επιλογή μέσω μοντέλων και ενδεχομένως να ισχύει και για τα υπόλοιπα.

Με την υποδειγματοληψία των ατόμων ελέγχου ελευθερώνονται περιοχές στον πολυδιάστατο χώρο, οι οποίες πλέον ανήκουν στα άτομα με δυσλεξία. Το ίδιο αποτέλεσμα αλλά με διαφορετικό τρόπο επιτυγχάνεται με την υπερδειγματοληψία. Με αυτήν, αναλόγως το μοντέλο, τα άτομα της κλάσης με δυσλεξία αποκτούν μεγαλύτερο βάρος και έτσι μία ευρύτερη περιοχή δίνεται σε αυτά.

Αναφορικά με την ακρίβεια και την ανάκληση, μεγαλύτερη ανάκληση σημαίνει πως εντοπίστηκαν περισσότερα άτομα με δυσλεξία. Αυτό συμβαίνει γιατί τα μοντέλα δίνουν τελικά

μεγαλύτερη περιοχή σε αυτή την κλάση. Με αυτό τον τρόπο όμως άτομα χωρίς δυσλεξία ταξινομούνται ως δυσλεκτικά. Έτσι εφόσον αυτά είναι αναλογικά περισσότερα από τα άτομα με δυσλεξία μειώνεται η ακρίβεια.

Τα παραπάνω διαγράμματα ακόμα δείχνουν πως τα σημαντικότερα χαρακτηριστικά για τον διαχωρισμό των κλάσεων είναι η μητρική γλώσσα, ο λόγος αστοχιών και η ακρίβεια των ερωτήσεων 31-32, ο αριθμός των κλικ ερωτήσεων 24 και 25, τα κλικ ο αριθμός των σωστών απαντήσεων και το σκορ της ερώτησης 28.

Η μητρική γλώσσα, αν ο συμμετέχων στο πείραμα είχε την ισπανική ως μοναδική μητρική γλώσσα, αποτελεί δημογραφικό χαρακτηριστικό το οποίο αν και χρησιμοποιήθηκε στην εκπαίδευση ώστε να γίνει ευκολότερη σύγκλιση των μοντέλων δεν περιέχει πληροφορία για το ποια είναι τα χαρακτηριστικά των ικανοτήτων του ατόμου που το καθορίζουν ως δυσλεκτικό. Ωστόσο σύμφωνα με τα υπόλοιπα αποτελέσματα και σε συνδυασμό με την ανάλυση των ερωτήσεων της προηγούμενης ενότητας μπορούμε να διακρίνουμε τα φυσικά χαρακτηριστικά του ατόμου που το ταξινομούν ως δυσλεκτικό. Αυτά είναι η λεκτική, ορθογραφική, φωνολογική, μορφολογική, σημασιολογική και συντακτική επίγνωση.

Κεφάλαιο 6

Επίλογος

Το παρόν κεφάλαιο είναι το τελευταίο της εργασίας. Σε αυτό παρατίθενται στον αναγνώστη τα συμπεράσματα που εξάγονταιαπό αυτήν, αλλά και ιδέες για μελόντικές επεκτάσεις.

6.1 Συμπεράσματα

Στην παρούσα εργασία εξετάστηκαν εκτενώς οι αλγόριθμοι ταξινόμησης της μηχανικής μάθησης όπως αυτοί έχουν αναλυθεί στις προηγούμενες ενότητες αλλά και ένα πλήθος αρχιτεκτονικών σωλήνωσης που εφαρμόστηκαν για την βελτίωση της επίδοσής τους. Όπως ήταν αναμενόμενο η επίδοση των μοντέλων εξαρτάται τόσο από τα δεδομένα που τους προσφέρονται για εκπαίδευση, όσο και από τις τεχνικές προεπεξεργασίας που εφαρμόζονται σε αυτά τα δεδομένα.

Για το πρόβλημα των δεδομένων οφθαλμικής ανίχνευσης αρχικά παρατηρούμε πως η τεχνική δυναμικής χρονικής στρέβλωσης, λειτουργεί αρνητικά στην διαδικασία της εκπαίδευσης σύμφωνα με τους πρώτους πίνακες της αντίστοιχης ενότητας. Τα χαρακτηριστικά που εξάγονται για την εκπαίδευση των μοντέλων είναι στατιστικού χαρακτήρα π.χ. μέση τιμή, τυπική απόκλιση κ.λ.π. Έτσι όταν η παραπάνω διαδικασία εφαρμόζεται πριν την εξαγωγή τους, τα οδηγεί στο να έχουν μία κανονικότητα, καθώς εξ ορισμού η τεχνική δυναμικής χρονικής στρέβλωσης μεταφέρει τις χρονοσειρές στο ίδιο μήκος, ενώ προσπαθεί να διατηρήσει τη συμπεριφορά που αυτές έχουν, αφαιρώντας τυχόν θόρυβο. Ο θόρυβος αυτός όμως, όπως διαπιστώνεται από τα αποτελέσματα, είναι απαραίτητο να υπάρχει ώστε να μεταφερθεί στα εξαγόμενα χαρακτηριστικά. Αυτός είναι τελικά που θα οδηγήσει στο διαχωρισμό των ατόμων με δυσλεξία και χωρίς.

Με μία πρώτη ματιά στον πρώτο πίνακα της ενότητας 5.1.1 παρατηρούμε πως την καλύτερη επίδοση εμφανίζουν τα μοντέλα τυχαία δάση με 86.12% ορθότητα και οι μηχανές διανυσμάτων υποστήριξης με 85.87% ορθότητα. Η λογιστική παλινδρόμηση αν και τελευταία στην κατάταξη, δεν είναι μακριά από τα υπόλοιπα με 85.87% ορθότητα. Το παραπάνω αποτέλεσμα αν και μη αναμενόμενο, καθώς η λογιστική παλινδρόμηση είναι γραμμικό μοντέλο το οποίο καλείται να διαχωρίσει δύο μη γραμμικά διαχωρίσιμα σύνολα, γίνεται κατανοητό από τα γραφήματα της ενότητας 5.1.2. Σε αυτά φαίνεται ότι κρατώντας τα κατάλληλα χαρακτηριστικά ή δίνοντας

μεγαλύτερο βάρος σε αυτά και βρίσκοντας έναν καλό διαχωρισμό του χώρου εκπαίδευσης, μπορεί να γίνει γραμμικός διαχωρισμός των δύο συνόλων με λίγες απώλειες.

Στη συνέχεια αναφορικά με τις τεχνικές αρχιτεκτονικών σωλήνωσης που χρησιμοποιήθηκαν για την προεπεξεργασία των δεδομένων, παρατηρούμε πως με τη σωστή προεπεξεργασία των δεδομένων, όλα τα μοντέλα παρουσιάζουν παρόμοια μέγιστη απόδοση. Στην κορυφή βρίσκονται οι μηχανές διανυσμάτων υποστήριξης με χρήση γενετικών αλγορίθμων με 92.63% ορθότητα. Ακολουθούν η λογιστική παλινδρόμηση και τα τυχαία δάση με 91.87% και 91.34% ορθότητα αντίστοιχα. Εδώ αξίζει να σημειωθούν δύο συμπεράσματα. Πρώτο, ότι με τη χρήση γενετικών αλγορίθμων η ορθότητα αυξάνεται σε σχέση με τις υπόλοιπες μεθόδους, αλλά οι ίδιοι δεν μπορούν να δώσουν τη σημασία των χαρακτηριστικών. Αυτό μπορεί να γίνει σε ένα δεύτερο στάδιο με τη χρήση του μοντέλου στο οποίο έγινε η βελτιστοποίηση. Δεύτερο, ότι η λογιστική παλινδρόμηση έχει την ίδια επίδοση με τα υπόλοιπα μοντέλα με τη χρήση της επιλογής των χαρακτηριστικών. Η χρήση της κλιμάκωσης παρατηρείται πως δεν παίζει ιδιαίτερα σημαντικό ρόλο για την επίδοση των μοντέλων.

Αναφορικά με το πρόβλημα της παιχνιδοποίησης, από τον πρώτο πίνακα της ενότητας 5.2.1 μπορούμε να παρατηρήσουμε πως τα μοντέλα δεν είναι σε θέση να διαχωρίσουν τις δύο κλάσεις. Το παραπάνω συμπέρασμα εξάγεται από το γεγονός ότι ενώ η ορθότητα είναι κοντά στο 90%, για την κλάση των ατόμων με δυσλεξία, η ανάκληση και η ακρίβεια είναι χαμηλές, με την ανάκληση 19.62% μέγιστο και την ακρίβεια 76.88%. Η χειρότερη περίπτωση είναι των μηχανών διανυσμάτων υποστήριξης με 0% σε όλες τις μετρικές που αφορούν την κλάση της δυσλεξίας. Το παραπάνω σημαίνει πως το μοντέλο προέβλεψε όλα τα άτομα ως μη έχοντα δυσλεξία. Με μία ματιά στα γραφήματα της ενότητας 5.2.2 διαπιστώνεται πως τα δύο σύνολα δεν διαχωρίζονται εύκολα καθώς στις διαστάσεις που προβλέπουν τα μοντέλα ως σημαντικότερες, τα δύο σύνολα αλληλοεπικαλύπτονται σε πολλά (αναλογικά με το μέγεθός τους) σημεία.

Συγκρίνοντας τα αντίστοιχα μοντέλα από τον πίνακα της προηγούμενης παραγράφου με αυτά του πίνακα της βελτιστοποίησης, ως προς την ορθότητα παρατηρούμε πως δεν υπάρχει ουσιαστική βελτίωση των αποτελεσμάτων. Ωστόσο αυτή που υπάρχει μπορεί να αποδοθεί στην ύπαρξη ενός στάδιου κλιμάκωσης και ενός επιλογής χαρακτηριστικών. Για την κλιμάκωση φαίνεται πως δεν υπάρχει κάποιο είδος που παίρνει τα πρωτεία, ωστόσο στην επιλογή χαρακτηριστικών μπορούμε να καταλάβουμε πως την καλύτερη επίδοση έχουν η λογιστική παλινδρόμηση και τα δέντρα ενίσχυσης με κλίση.

Με την σύγκριση των αποτελεσμάτων στα οποία έγινε βελτιστοποίηση ως προς την ακρίβεια και την ανάκληση παρατηρούμε πως υπάρχει μία αντιστρόφως ανάλογη σχέση μεταξύ των δύο μετρικών ανάλογα με την προεπεξεργασία. Η επιλογή χαρακτηριστικών παραμένει απαραίτητη και στις δύο περιπτώσεις με την λογιστική παλινδρόμηση και τα δέντρα ενίσχυσης με κλίση να είναι και πάλι αυτά που εμφανίζονται περισσότερο. Βεβαίως στην περίπτωση της ακρίβειας εμφανίζεται μέγιστο 87.5% από τις μηχανές διανυσμάτων υποστήριξης, αλλά με πολύ μικρή ανάκληση 6.37% με χρήση τυχαίων δασών για την επιλογή χαρακτηριστικών. Το στάδιο της κλιμάκωσης παραμένει, και εδώ φαίνεται για την βελτιστοποίησης της ακρίβειας να χρησιμοποιείται περισσότερο η κανονική κλιμάκωση, ενώ στην βελτιστοποίηση της ανάκλησης η

κλιμάκωση ελαχίστου μεγίστου. Τέλος για την βελτιστοποίηση της ανάκλησης χρησιμοποιείται και ένα στάδιο υποδειγματοληψίας. Μέγιστο για την ανάκληση παρουσιάζουν τα δέντρα ενίσχυσης με κλίση στο 73.23% με 33.29% ακρίβεια. Στην περίπτωση της βελτιστοποίησης ως προς την ανάκληση έχουμε πτώση της ορθότητας στο 80%, αλλά κάτι τέτοιο είναι προτιμητέο σε σχέση με την χαμηλή ανάκληση, 20%, καθώς είναι προτιμότερο να γίνει λανθασμένη ταξινόμηση ενός ατόμου ως δυσλεκτικό, παρά ένα άτομο που είχε δυσλεξία να ταξινομηθεί ως μη δυσλεκτικό.

Τέλος στην περίπτωση της βελτιστοποίησης ως προς το F1 score δεν μπορούμε να εξάγουμε σημαντικά συμπεράσματα καθώς οι αντίστοιχες ένδειξεις στην ανάκληση και στην ακρίβεια είναι μικρές, για την ακρίβεια 35% - 50% και για την ανάκληση 55% - 69.89%.

Με μία ανασκόπηση των παραπάνω παραγράφων κάποιος μπορεί να συμπεράνει πως η διάγνωση της μαθησιακής δυσκολίας γίνεται πιο εύκολα με τα δεδομένα οφθαλμικής ανίχνευσης. Η δυσλεξία όπως είχε αναφερθεί και στο δεύτερο κεφάλαιο είναι μία μαθησιακή δυσκολία που με πιο απλά λόγια αφορά την δυσκολία στην ανάγνωση. Έτσι από τον ίδιο τον ορισμό της, δεδομένα όπως της οφθαλμικής ανίχνευσης βρίσκονται πιο κοντά στη ρίζα του προβλήματος, ενώ δεδομένα που προκύπτουν από την παιχνιδοποίηση εισάγουν πολύ θόρυβο καθώς όχι μόνο στηρίζονται πολύ στις ερωτήσεις τις οποίες χρησιμοποιούν για τη διάγνωση, αλλά εξαρτώνται και από τη γλώσσα στην οποία διατυπώνονται αυτές οι ερωτήσεις. Ακόμα μπορεί να επηρεάζονται και από άλλες μαθησιακές δυσκολίες που σχετίζονται με την ορθογραφία. Τα παραπάνω δεν πλαισιώνονται μόνο θεωρητικά, αλλά είναι εύκολα διακριτά από τα γραφήματα του προηγούμενου κεφαλαίου στην παρουσίαση των χαρακτηριστικών και από τα αποτελέσματα της εκπαίδευσης των μοντέλων. Ωστόσο η μέθοδος της παιχνιδοποίησης δεν πρέπει να απορριφθεί, αλλά πρέπει να πλαισιωθεί καλύτερα, καθώς είναι αυτή που μπορεί να γίνει πιο εύκολα προσιτή στον κόσμο.

6.2 Περιορισμοί

Στην παρούσα εργασία έγινε μία εκτενής μελέτη τόσο των μοντέλων μηχανικής μάθησης με σκοπό την διάγνωση της δυσλεξίας, όσο και μία σύγκριση μεταξύ δύο μεθόδων συλλογής δεδομένων που χρησιμοποιούνται για την διάγνωσή της. Εδώ λοιπόν πρέπει να σημειωθεί ότι οι μέθοδοι αυτοί δεν πρέπει να χρησιμοποιούνται ως μοναδικό εργαλείο για τη διάγνωση της μαθησιακής δυσκολίας, καθώς δεν λαμβάνουν υπόψη τις σημαντικές παραμέτρους που καλούνται να αντιμετωπίσουν οι ειδικοί:

- Δεν λαμβάνουν υπόψη την νοημοσύνη του υπό διάγνωση ατόμου. Στην πράξη τα άτομα προς εξέταση υποβάλονται και σε τεστ νοημοσύνης ώστε να αποφευχθούν άλλοι παράγοντες που προκαλούν μείωση των φωνολογικών ικανοτήτων.
- Η δυσλεξία σπανίως υπάρχει μόνη της και συχνά γίνεται λάθος η διάγνωσή της ως μία η περισσότερες διαφορετικές μαθησιακές δυσκολίες. Συνήθως μπορεί να συνυπάρχει με τις: δυσπραξία, δυσορθογραφία, διαταραχή ελλειμματικής προσοχής και υπερκινητικότητας (ΔΕΠΥ), ειδική γλωσσική διαταραχή (ΕΓΔ) κ.λ.π. Πιο συγκεκριμένα το 40%

των ανθρώπων με δυσλεξία έχουν και δυσπραξία, και το 18% με 42% έχουν ΔΕΠΥ. Οι παραπάνω μαθησιακές δυσκολίες μπορεί να εμφανίζονται μόνες τους ή σε συνδυασμούς μεταξύ τους και είναι επιτακτική η ανάγκη της διάγνωσης και του διαχωρισμού τους από ειδικό. Τέλος, επιπρόσθετος θόρυβος στα δεδομένα που προκύπτουν από τις παραπάνω μεθόδους, μπορεί να προστεθεί εκτός από τη συνύπαρξη άλλων μαθησιακών δυσκολιών και από την κόπωση ή απλά την έλλειψη προσοχής του ατόμου που συμμετέχει.

- Βασική παραδοχή για τη συλλογή των δεδομένων ήταν η διάγνωση από ειδικούς. Ωστόσο αυτή η εκτίμηση μπορεί να διαφέρει από ειδικό σε ειδικό.

- Τα τεστ που έγιναν δεν μπορούν να διαχωρίσουν τους διαφορετικούς βαθμούς δυσλεξίας ή να συμπεριλάβουν την ιστορία του υποκειμένου, συνιστώσες οι οποίες λαμβάνονται υπόψη στη διάγνωση από ειδικό.

6.3 Μελλοντικές επεκτάσεις

Με βάση το κεφάλιο 2 και τη μορφή των δεδομένων, στην παρούσα εργασία προτείνεται μια πληθώρα επεκτάσεων. Μία πρώτη επέκταση είναι η γενίκευση του προβλήματος, ώστε να μπορεί να προβλέψει τη σοβαρότητα της μαθησιακής δυσκολίας (multiclass classification), με σκοπό να είναι πιο εύκολο να δοθεί η απαραίτητη προσοχή. Στη συνέχεια θα μπορούσε να γίνει μία σύγκριση όλων των γνωστών μεθόδων διάγνωσης, όπως με ηλεκτροεγκεφαλογράφημα ή με μαγνητική τομογραφία, για τις οποίες δεν υπάρχουν αρκετά δεδομένα. Ακόμα, αναφορικά με την παιχνιδοποίηση θα μπορούσε να γίνει με μία διαδικασία A\B testing η δημιουργία ενός συστήματος, το οποίο θα συλλέγει δεδομένα και μετρικές που βρίσκονται πιο κοντά στη διάγνωση της νόσου. Ακόμα θα μπορούσε να γίνει και η συλλογή νέων μετρικών με τη μορφή χρονοσειρών ώστε να φαίνεται πιο ολοκληρωμένα η συμπεριφορά του χρήστη κατά τη συμπλήρωση των ερωτήσεων. Τέλος δεδομένου ότι τα περισσότερα δεδομένα όπως της οφθαλμικής ανίχνευσης, του ηλεκτροεγκεφαλογραφήματος και της μαγνητικής τομογραφίας είναι ήδη στη μορφή χρονοσειρών θα μπορούσε να γίνει η χρήση μίας μεγάλης πληθώρας τεχνικών που προσφέρουν τα βαθιά νευρωνικά δίκτυα για την διαχείριση τέτοιου είδους δεδομένων (αναδρομικά νευρωνικά δίκτυα). Εμπόδιο σε αυτές τις επεκτάσεις αποτελεί η έλλειψη δεδομένων, τόσο από πλευράς όγκου, όσο και από πλευράς διαθεσιμότητας.

Βιβλιογραφία

[1] Lyon GR, Shaywitz SE, Shaywitz BA. A definition of dyslexia. *Annals of Dyslexia.*, 53(1):1-14, 2003.

[2] Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM. Specific reading disability(dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry.*, 45(1):2-40, 2004.

[3] Brunswick N. In: McDougall S, de Mornay Davies P. Unimpaired reading development and dyslexia across different languages. *Reading and dyslexia in different orthographies. Psychology Press.*, p. 131-154, 2010.

[4] Krafnick AJ, Flowers DL, Napoliello EM, Eden GF. Gray matter volume changes following reading intervention in dyslexic children. *Neuroimage.*, 57(3):733-741, 2011.

[5] Gabrieli JD. Dyslexia: a new synergy between education and cognitive neuroscience. *Science.*, 325(5938):280-283, 2009.

[6] Vaughn S, Cirino PT, Wanzek J, Wexler J, Fletcher JM, Denton CD et al. Response to intervention for middle school students with reading difficulties: effects of a primary and secondary intervention. *School Psych Rev.*, 39(1): 3–21, 2010.

[7] Alexander-Passe N. How dyslexic teenagers cope: An investigation of self-esteem, coping and depression. *Dyslexia.*, 12(4):256-275, 2006.

[8] Anastasiou D, Polychronopoulou S. Identification and Overidentification of Specific Learning Disabilities (Dyslexia) in Greece. *Learning Disability Quarterly.*, 32(2):55-69, 2009.

[9] International Dyslexia Association.

[10] Dyslexia Association of Ireland.

[11] Kaisar S. Developmental dyslexia detection using machine learning techniques : A survey. *ICT Express.*, 6(3):181-184, 2020.

[12] Rello L, Romero E, Rauschenberger M, Ali A, Williams K, Bigham JP, White NC. Screening dyslexia for English using HCI measures and machine learning. *Proceedings*

of the 2018 international conference on digital health - DH'18. ACM Press., 80–84, 2018.

[13] Rello L, Baeza-Yates R, Ali A, Bigham JP, Serra M. Predicting risk of dyslexia with an online gamified test. *arXiv preprint arXiv:1906.03168*, 1:1-13, 2019.

[14] Rauschenberger M, Rello L, Baeza-Yates R, Bigham, Bigham JP. Towards language independent detection of dyslexia with a web-based game. *In W4A'18. Lyon, France. https://doi.org/10.1145/3192714.3192816*, 2018.

[15] Lexercise. Dyslexia test - Online from Lexercise. http://www.lexercise.com/tests/dyslexia-test, 2016.

[16] Nessy. Dyslexia screening - Nessy UK. https://www.nessy.com/uk/product/dyslexia-screening/, 2011.

[17] Rello et al. Dytective: Diagnosing risk of dyslexia with a game. In Proc. Pervasive Health'16, Cancun, Mexico, 2016.

[18] Asvestopoulou T, Manousaki V, Psistakis A, Andreadakis V, Aslanides I, Papadopouli M. Dyslexml: Screening tool for dyslexia using machine learning. *ArXiv abs/1903.06274*, 1–6, 2019.

[19] Nilsson Benfatto M, Öqvist Seimyr G, Ygge J, Pansell T, Rydberg A, et al. Screening for Dyslexia Using Eye Tracking during Reading *PLoS ONE 11(12): e0165508. https://doi.org/10.1371/journal.pone.0165508*, 2016.

[20] Optotech Ltd. Greece. Τι είναι το Eye Tracking; https://www.eyetracking.gr/faq/52-faq-1?lang=el .

[21] Smyrnakis I, Andreadakis V, Selimis V, Kalaitzakis M, Bachourou T, Kaloutsakis G, Kymionis GD, Smirnakis S, Aslanides IM. RADAR: A novel fast-screening method for reading difficulties with special focus on dyslexia. *PLoS ONE*, 2017.

[22] Rezvani Z, Zare M, Žarić G, Bonte M, Tijms J, Van der Molen M, González GF. Machine learning classification of dyslexic children based on eeg local network features. *bioRxiv https://doi.org/10.1101/569996*, 1-23, 2019.

[23] Perera H, Shiratuddin MF, Wong KW, Fullarton K. Eeg signal analysis of writing and typing between adults with dyslexia and normal controls. *Int. J. Interact. Multimedia Artif. Intell.*, 5(1):62-67, 2018.

[24] Płoński P, Gradkowski W, Altarelli I, Monzalvo K, van Ermingen-Marbach M, Grande M, Heim S, Marchewka A, Bogorodzki P, Ramus F, et al. Multi-parameter machine learning approach to the neuroanatomical basis of developmental dyslexia. *Hum. Brain Mapp.*, 38(2):900-908, 2017.

[25] Cui Z, Xia Z, Su M, Shu H, Gong G. Disrupted white matter connectivity underlying developmental dyslexia: a machine learning approach. *Hum. Brain Mapp.*, 37(4):1443-1458, 2016.

[26] Samuel AL. Some studies in machine learning using the game of checkers. *IBM Journal of research and development.*, 3(3):210-229, 1959.

[27] Mitchell T. Machine Learning, McGraw Hill, p.2, 1997.

[28] Stuart JR, Peter N. Artificial Intelligence: A Modern Approach. *Third Edition, Prentice Hall ISBN 9780136042594.*, 2010.

[29] Hinton G, Sejnowski T. Unsupervised Learning: Foundations of Neural Computation. *MIT Press. ISBN 978-0262581684.*, 1999.

[30] Hu J, Niu H, Carrasco J, Lennox B, Arvin F. Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology.*, 69(12):14413-14423, 2020.

[31] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning (2nd ed.). *Springer.*, ISBN 0-387-95284-5, 2008.

Παράρτημα Α΄

Βέλτιστες Υπερ-παράμετροι

Πίνακας Α΄.1: Βέλτιστες υπερ-παράμετροι για τα δεδομένα οφθαλμικής ανίχνευσης.

Στάδια σωλήνωσης			
Επιλογή χαρακτηριστικών	Γενετικός Αλγόριθμος(Μηχανές Διανυσμάτων Υποστήριξης)	population_size	100
		mutation_probability	0.1
		elit_ratio	0.2
		crossover_probability	0.5
		crossover_type	uniform
	Επιλογή χαρακτηριστικών μέσω μοντέλων	n	85
Ταξινομητές	Μηχανές Διανυσμάτων Υποστήριξης	C	1
		kernel	rbf
		gamma	scale
	Λογιστική Παλινδρόμηση	solver	lbfgs
		penalty	l2
		C	1
		l1_ratio	0
	Τυχαία δάση	n_estimators	200
		criterion	gini

Πίνακας Α΄.2: Βέλτιστες υπερ-παράμετροι για τα δεδομένα παιχνιδοποίησης.

Στάδια σωλήνωσης			
Δειγματοληψία	Τυχαία Υποδειγματοληψία	sampling_strategy	0.75
	Τυχαία Υπερδειγματοληψία	sampling_strategy	0.5
Επιλογή χαρακτηριστικών	Επιλογή χαρακτηριστικών μέσω μοντέλων	n	100
Ταξινομητές	Μηχανές Διανυσμάτων Υποστήριξης	C	1
		kernel	rbf
		gamma	scale
	Λογιστική Παλινδρόμηση	solver	lbfgs
		penalty	l2
		C	1
		l1_ratio	0
	Δέντρα ενίσχυσης κλίσης	learning_rate	0.01
		n_estimators	200
	Τυχαία δάση	n_estimators	200
		criterion	gini

Πίνακας Α΄.3: Υπερ-παράμετροι για τα δεδομένα παιχνιδοποίησης.

Στάδια σωλήνωσης			
Δειγματοληψία	Τυχαία Υποδειγματοληψία	sampling_strategy	0.75
	Τυχαία Υπερδειγματοληψία	sampling_strategy	0.5
Επιλογή χαρακτηριστικών	Επιλογή χαρακτηριστικών μέσω μοντέλων	n	100
Ταξινομητές	Μηχανές Διανυσμάτων Υποστήριξης	C	1
		kernel	rbf
		gamma	scale
	Λογιστική Παλινδρόμηση	solver	lbfgs
		penalty	l2
		C	1
		l1_ratio	0
	Δέντρα ενίσχυσης κλίσης	learning_rate	0.01
		n_estimators	200
	Τυχαία δάση	n_estimators	200
		criterion	gini

Πίνακας Α΄.4: Ειδικές υπερ-παράμετροι για τα μοντέλα επιλογής χαρακτηριστικών.

Μοντέλα			
Ταξινομητές	Μηχανές Διανυσμάτων Υποστήριξης	kernel	linear
	Λογιστική Παλινδρόμηση	solver	liblinear
		penalty	l1