Έλεγχος Γνησιότητας Τραπεζογραμματίων με μεθόδους Μηχανικής Μάθησης

Αντώνιος Μαυρίδης
Τμήμα Ψηφιακών Συστημάτων
Πανεπιστήμιο Πειραιώς
Πειραιάς Ελλάδα
antmav0@gmail.com

ΠΕΡΙΛΗΨΗ

Τα τραπεζογραμμάτια είναι ένας τύπος διαπραγματεύσιμης γραμμής και είναι ένα από τα πιο σημαντικά περιουσιακά στοιχεία μιας χώρας. Η χρηματοπιστωτική αγορά καθίσταται θύμα πολλαπλών εγκληματικών οργανώσεων που παράγουν πλαστά τραπεζογραμμάτια και τα εισάγουν στο χρηματοπιστωτικό σύστημα. Για το ανθρώπινο μάτι η σωστή επιλογή μεταξύ πλαστών και γνήσιων τραπεζογραμματίων είναι σχεδόν αδύνατη. Αυτός είναι ο λόγος για τον οποίο χρειαζόμαστε τεχνικές μηχανικής μάθησης. Η ανάλυση δεδομένων βοηθά στην αντιμετώπιση του προβλήματος. Αυτή η μελέτη προτείνει τεχνικές μηχανικής μάθησης για την αξιολόγηση του ελέγχου ταυτότητας των τραπεζογραμματίων. Οι αλγόριθμοι εποπτευόμενης μηχανικής μάθησης και κατηγοριοποιητές όπως της Λογιστικής Παλινδρόμησης και των Μηχανών Διανυσμάτων Υποστήριξης (SVM) χρησιμοποιούνται για την διαφοροποίηση μεταξύ γνήσιων και πλαστών τραπεζογραμματίων. Η μελέτη δείχνει την σύγκριση αυτών των μεθόδων στην κατηγοριοποίηση των τραπεζογραμματίων.

ΕΝΝΟΙΕΣ CCS

• Υπολογιστικές Μεθοδολογίες • Μηχανική Μάθηση • Εκμάθηση Παραδειγμάτων • Εποπτευόμενη Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Μηχανική Μάθηση, Εποπτευόμενη Μάθηση, Κατηγοριοποίηση, Μηχανές Διανυσμάτων Υποστήριξης, Λογιστική Παλινδρόμηση, Έλεγχος Ταυτότητας Τραπεζογραμματίων.

1. ΕΙΣΑΓΩΓΗ

Η μηχανική μάθηση είναι ένα πεδίο μελέτης το οποίο ασχολείται με αλγορίθμους που μαθαίνουν από παραδείγματα. Η Κατηγοριοποίηση είναι μια διεργασία που απαιτεί την χρήση αλγορίθμων μηχανικής μάθησης για την αντιστοίχιση μιας ετικέτας στις εγγραφές του συνόλου δεδομένων. Η Κατηγοριοποίηση αναφέρεται σε ένα προγνωστικό πρόβλημα μοντελοποίησης, όπου μια κλάση προβλέπεται για ένα δοσμένο παράδειγμα δεδομένων εισαγωγής. Οι αλγόριθμοι προγνωστικής κατηγοριοποίησης αξιολογούνται με βάση τα αποτελέσματά τους. Υπάρχουν διαφορετικοί τύποι μεθόδων κατηγοριοποίησης μια εκ των οποίων ονομάζεται Δυαδική Κατηγοριοποίηση. Η Δυαδική Κατηγοριοποίηση αναφέρεται σε εκείνες τις διεργασίες κατηγοριοποίησης που η μεταβλητής στόχος λαμβάνει μόνο δύο τιμές. Συνήθως οι εργασίες δυαδικής κατηγοριοποίησης περιλαμβάνουν μια κλάση η οποία αποτελεί την φυσιολογική κατάσταση, και μια άλλη κατηγορία που είναι η μη φυσιολογική κατάσταση. Στην κλάση με την κανονική κατάσταση εκχωρείται η ετικέτα 0 και στην κλάση με την ανώμαλη κατάσταση εκχωρείται η ετικέτα 1. Οι πιο δημοφιλείς αλγόριθμοι που μπορούν χρησιμοποιηθούν σε ένα πρόβλημα δυαδικής κατηγοριοποίηση είναι οι εξής: Λογιστική Παλινδρόμηση, K-Κοντινότεροι Γείτονες (K-NN), Δέντρα απόφασης, Μηχανές Διανυσμάτων Υποστήριξης, Απλοϊκός Bayes. Μερικοί αλγόριθμοι είναι ειδικά σχεδιασμένοι για τα προβλήματα δυαδικής κατηγοριοποίησης όπως η Λογιστική Παλινδρόμηση και οι Μηχανές Διανυσμάτων Υποστήριξης.

Χρησιμοποιούμε λοιπόν την παραπάνω ιδέα για την αντιμετώπιση της εργασίας κατηγοριοποίησης τραπεζογραμματίων. Το πρόβλημα του προσδιορισμού ενός νομίσματος ως ψεύτικου ή πραγματικού είναι ένα πρόβλημα διαλογής, το οποίο στοχεύει στον αυτόματο διαχωρισμό των πλαστών τραπεζογραμματίων από τα γνήσια.

Η μελέτη στοχεύει στην εκπαίδευση των κατηγοριοποιητών, χρησιμοποιώντας ένα σύνολο εκπαίδευσης που επιτρέπει στον εκάστοτε αλγόριθμο να προσεγγίσει προσεκτικά μια μαθηματική συνάρτηση που θα είχε δημιουργήσει τις ετικέτες των δεδομένων εκπαίδευσης. Αυτό βοηθά τους κατηγοριοποιητές να ταξινομήσουν ένα σημείο δοκιμής με την κατάλληλη ετικέτα κλάσης. Με άλλα λόγια η τεχνική αυτή βρίσκει τη λύση στο πρόβλημά μας.

2. ΠΕΡΙΓΡΑΦΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση των μοντέλων προέρχεται από το αποθετήριο μηχανικής μάθησης UCI [1].

Τα δεδομένα εξήχθησαν από εικόνες που ελήφθησαν από γνήσια και πλαστά χαρτονομίσματα. Για την ψηφιοποίηση, χρησιμοποιήθηκε μια βιομηχανική κάμερα που χρησιμοποιείται συνήθως για έλεγχο εκτύπωσης. Οι τελικές εικόνες έχουν 400x400

pixel. Λόγω του φακού και της απόστασης από το αντικείμενο που ερευνήθηκε, αποκτήθηκαν εικόνες σε κλίμακα του γκρι με ανάλυση περίπου 660 dpi. Το εργαλείο Wavelet Transform χρησιμοποιήθηκε για την εξαγωγή χαρακτηριστικών από εικόνες. Τα χαρακτηριστικά εξήχθησαν από τις εικόνες μετά την υποβολή τους σε μετασχηματισμό κυμάτων και τα εξαγόμενα χαρακτηριστικά είναι Variance, Skewness, Curtosis και Entropy. Το σύνολο δεδομένων αποτελείται μόνο από αυτές τις τέσσερις δυνατότητες και την ετικέτα κλάσης.

Το σύνολο δεδομένων σχεδιάστηκε για να διακρίνει τα πλαστά από τα γνήσια τραπεζογραμμάτια. Το σύνολο δεδομένων έχει 1372 εγγραφές. Η μεταβλητή στόχος περιέχει δύο τιμές: 0 και 1 όπου το 0 αντιπροσωπεύει τα γνήσια χαρτονομίσματα και το 0 τα πλαστά. Οι στατιστικές παράμετροι των δεδομένων φαίνονται στον Πίνακα 2 και μια σύντομη περιγραφή των γνωρισμάτων των δεδομένων παρουσιάζεται στον Πίνακα 1.

Πίνακας 1. Περιγραφή Συνόλου Δεδομένων

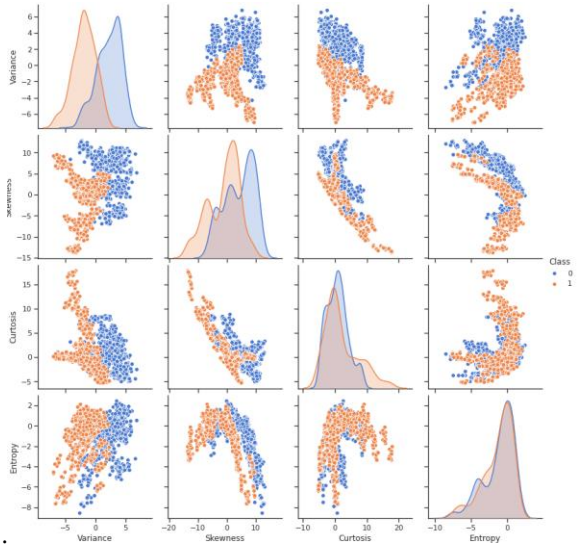
Όνομα Γνωρίσματος	Τύπος Τιμής	Περιγραφή
Variance	Συνεχή	Το Variance βρίσκει πώς κάθε εικονοστοιχείο διαφέρει από τα γειτονικά εικονοστοιχεία και τα ταξινομεί σε διαφορετικές περιοχές.
Skewness	Συνεχή	Το Skewness είναι το μέτρο της έλλειψης συμμετρίας.
Curtosis	Συνεχή	Το Kurtosis είναι ένα μέτρο για το αν τα δεδομένα είναι heavy tailed ή light-tailed σε σχέση με μια κανονική κατανομή.
Entropy	Συνεχή	Η εντροπία εικόνας είναι μια ποσότητα που χρησιμοποιείται για να περιγράψει την ποσότητα των πληροφοριών που πρέπει να κωδικοποιηθούν, από έναν αλγόριθμο συμπίεσης.
Class	Ακέραια	Η κλάση περιέχει δύο τιμές, την 0 που αντιπροσωπεύει γνήσια χαρτονομίσματα και την 1 που αντιπροσωπεύει ψεύτικα χαρτονομίσματα.

Πίνακας 2. Πίνακας που δείχνει τις στατιστικές παραμέτρους των γνωρισμάτων του συνόλου δεδομένων των τραπεζογραμματίων.

	Variance	Skewness	Curtosis	Entropy
Count	1372	1372	1372	1372
Mean	0.433	1.922	1.397	-1.191
Std	2.842	5.869	4.310	2.101
Min	-7.042	-13.773	-5.286	-8.548
25%	-1.773	-1.708	-1.574	-2.413
50%	0.496	2.319	0.616	-0.586
75%	2.821	6.814	3.179	0.394
Max	6.824	12.951	17.927	2.449

Σύμφωνα με το διάγραμμα διασποράς όλων των γνωρισμάτων που φαίνεται στην Εικόνα 1, συνάγουμε τα εξής:

1. Η διακύμανση (Variance) και των δύο κλάσεων κατανέμεται κανονικά.
2. Και οι δυο κλάσης παρουσιάζουν δεξιά κλίση.
3. Το Curtosis της κλάσης 0 είναι προσεγγιστικά κανονικά κατανεμημένο με δεξιά κλίση (right skewed), ομοίως και της κλάσης 1 έχει δεξιά κλίση αλλά παρουσιάζει ακραίες τιμές.
4. Η εντροπία και των δύο κλάσεων τυχαίνει να ακολουθεί μια κατανομή αριστερά-στρεβλωμένη (left skewed), υποδεικνύοντας την ύπαρξη ακραίων τιμών

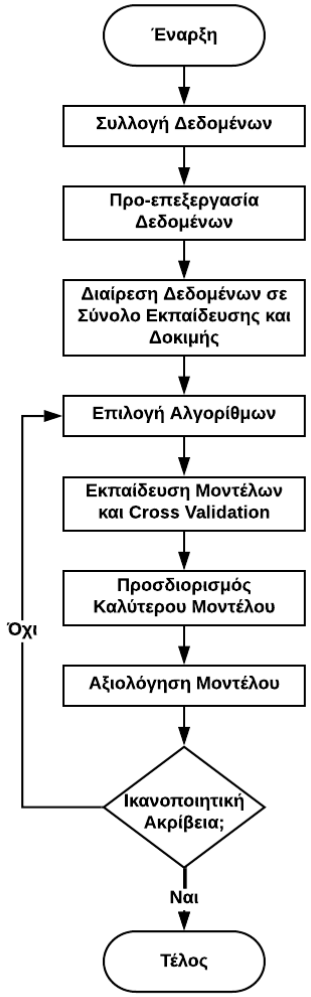


Εικόνα 1: Διάγραμμα διασποράς μεταξύ όλων των ζευγών των γνωρισμάτων.

Σε όλα τα πειράματα που πραγματοποιήθηκαν, το 80% των παραδειγμάτων του συνόλου των δεδομένων χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου και το υπόλοιπο 20% χρησιμοποιήθηκε για τη δοκιμή του μοντέλου.

3. ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ - ΜΕΘΟΔΟΛΟΓΙΑ

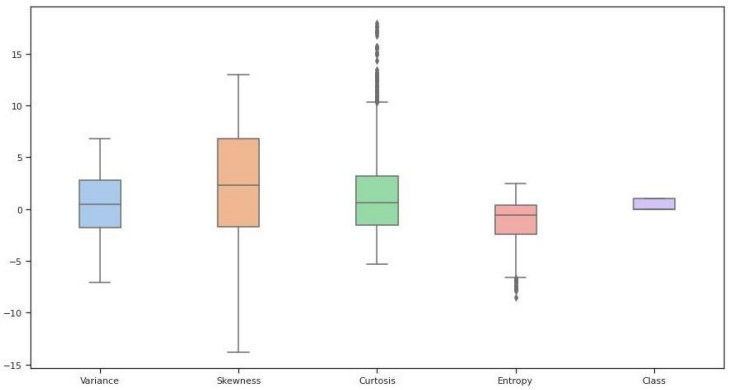
Η ροή εργασίας των διεξαγόμενων πειραμάτων που πραγματοποιήθηκαν παρουσιάζεται στο Σχήμα. 1.



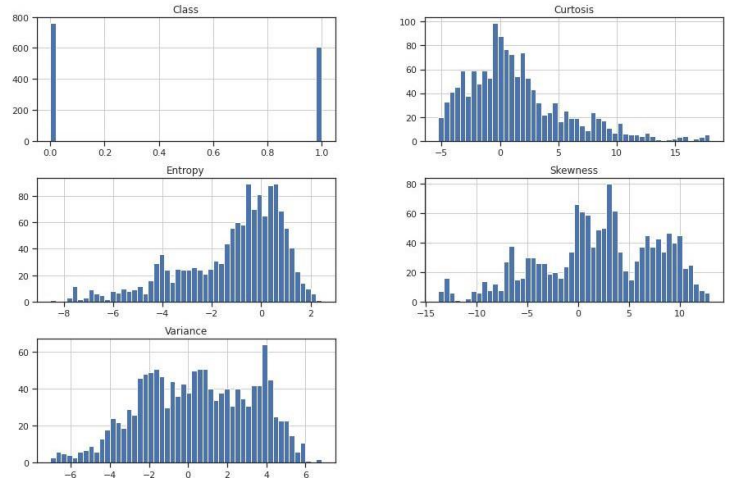
Σχήμα 1: Η ροή εργασίας των πειραμάτων που πραγματοποιήθηκαν.

Το πρώτο βήμα ήταν η συλλογή των δεδομένων. Το επόμενο βήμα ήταν η προ-επεξεργασία των δεδομένων και η διερευνητική τους ανάλυση. Στην συνέχεια χωρίσαμε τα δεδομένα σε αναλογία 80:20 σε σύνολο εκπαίδευσης και δοκιμής αντίστοιχα. Έπειτα επιλέξαμε ένα σύνολο αλγορίθμων και τους αξιολογήσαμε με βάση την ακρίβεια τους, χρησιμοποιώντας την τεχνική Διασταυρωτικής Επικύρωσης (Cross Validation). Αυτό είχε ως απώτερο σκοπό την αποφυγή της Υπέρ-προσαρμογής (Over-Fitting) των κατηγοριοποιητών στα δεδομένα μας, και παράλληλα την εστίαση του ενδιαφέροντός μας στους κατηγοριοποιητές με την μεγαλύτερη ακρίβεια. Από τα αποτελέσματα που προέκυψαν επιλέξαμε τους δύο κατηγοριοποιητές με την υψηλότερη ακρίβεια και εκ νέου τους εκπαιδεύσαμε και τους αξιολογήσαμε με βάση τα σύνολα εκπαίδευσης και δοκιμής.

Όσο αναφορά την προ-επεξεργασία των δεδομένων, εφαρμόσαμε τυποποίηση του συνόλου δεδομένων (Standardizing the Data Set) και στην συνέχεια προχωρήσαμε στην διερευνητική ανάλυσή τους. Το διάγραμμα διασποράς για όλους τους συνδυασμούς των γνωρισμάτων του συνόλου δεδομένων φαίνεται στην Εικόνα 1. Ομοίως, τόσο το Θηκόγραμμα όσο και το Ιστόγραμμα για όλα τα χαρακτηριστικά του συνόλου παρουσιάζονται αντίστοιχα στην Εικόνα 2 και Εικόνα 3.



Εικόνα 2: Θηκόγραμμα όλων των γνωρισμάτων.



Εικόνα 3: Ιστόγραμμα όλων των γνωρισμάτων.

Από την διερευνητική ανάλυση συμπεραίνουμε τα εξής:

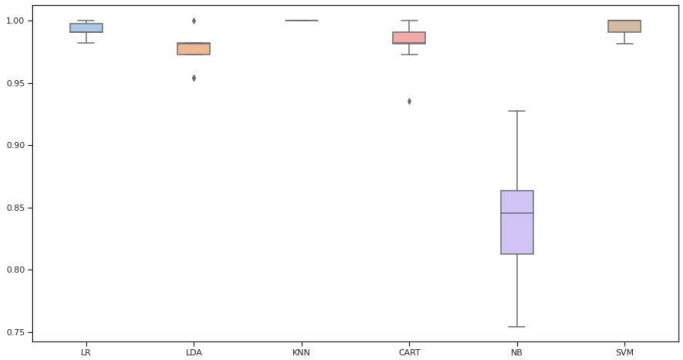
1. Το διάγραμμα διασποράς της Εικόνας 1 αποκαλύπτει ότι τα δεδομένα δεν διαχωρίζονται γραμμικά στις δύο

διαστάσεις (2D), επομένως θα πρέπει να γίνουν κατάλληλες επιλογές κατά την επιλογή του κατηγοριοποιητή.

2. Το Θηκόγραμμα της Εικόνας 2 και το Ιστόγραμμα της Εικόνας 3 αποκαλύπτουν ότι το Variance και το Skewness πλησιάζουν την κανονική κατανομή.

4. ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ

Μόλις ορίσαμε το πρόβλημά μας και προ-επεξεργαστήκαμε τα δεδομένα μας, έπρεπε να εφαρμόσουμε αλγόριθμους μηχανικής, οι οποίοι θα οδηγούσαν στην λύση του προβλήματος. Έτσι ορίσαμε ένα test harness. Το test harness είναι τα δεδομένα που θα εκπαιδεύσουμε και θα δοκιμάσουμε σε έναν αλγόριθμο, αλλά και το μέτρο απόδοσης που θα χρησιμοποιήσουμε για την αξιολόγηση της απόδοσής του. Το μέτρο απόδοσης είναι ο τρόπος που θέλουμε να αξιολογήσουμε μια λύση στο πρόβλημα. Είναι η μέτρηση που θα κάνουμε για τις προβλέψεις που έγιναν από ένα εκπαιδευμένο μοντέλο στο σύνολο δεδομένων δοκιμής. Το κύριο μέτρο απόδοσης που πρόκειται να αξιολογήσουμε είναι η ακρίβεια ταξινόμησης (classification accuracy). Η μέθοδος που θα χρησιμοποιήσουμε στο test harness ονομάζεται Cross Validation. Αρχικά περιλαμβάνει το διαχωρισμό του συνόλου δεδομένων σε πολλές ομάδες παρόμοιου μεγέθους, οι οποίες ονομάζονται πτυχές (folds). Στη συνέχεια, το μοντέλο εκπαιδεύεται σε όλες τις πτυχές, εκτός από μια που παραμένει εκτός και έπειτα το μοντέλο αυτό δοκιμάζεται στην «εξωτερική» πτυχή. Η διαδικασία επαναλαμβάνεται έτσι ώστε κάθε πτυχή να έχει την ευκαιρία να μείνει έξω και να ενεργεί ως το σύνολο δεδομένων δοκιμής. Τέλος, τα μέτρα απόδοσης υπολογίζονται κατά μέσο όρο σε όλες τις πτυχές για να εκτιμηθεί η ικανότητα του αλγορίθμου στο πρόβλημα. Στην περίπτωσή μας εφαρμόσαμε 10 cross validation. Αυτό εκπαίδευσε τον κατηγοριοποιητή στο 80% των δεδομένων και τον δοκίμασε στο υπόλοιπο 20%, παρέχοντας ένα μέσο μέτρο για την ακρίβεια κατηγοριοποίησης σε 10 επαναλήψεις με διαφορετικά δείγματα σε κάθε διαχωρισμό. Για να επιλέξουμε την καταλληλότερη μέθοδο, δοκιμάσαμε διαφορετικούς κατηγοριοποιητές και αλγόριθμους όπως Logistic Regression, Linear Discriminant Analysis, Gaussian Naïve Bayes, Support Vector Machines (SVM), K-Nearest Neighbor και Decision Tree.



Εικόνα 4: Θηκόγραμμα για τις ακρίβειες κατηγοριοποίησης όλων των αλγορίθμων.

Πίνακας 3: Αποτελέσματα αλγορίθμων μετά το Cross Validation.

	Mean	**Standard Deviation**
Logistic Regression	0.9917	0.0063
Linear Discriminant Analysis	0.9762	0.0131
K-Nearest Neighbor	1.0	0.0
Decision Tree	0.9817	0.0142
Gaussian Naïve Bayes	0.8413	0.0494
Support Vector Machines	0.9954	0.0061

Τελικά, παρατηρούμε ότι η Λογιστική Παλινδρόμηση (Logistic Regression) και οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είχαν την υψηλότερη ακρίβεια στα δεδομένα μας, οπότε στην συνέχεια θα προχωρήσουμε σε περαιτέρω ανάλυση και αξιολόγηση των δύο μεθόδων αυτών.

5. ΑΛΓΟΡΙΘΜΟΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

5.1 Μηχανές Διανυσμάτων Υποστήριξης

Μία μηχανή διανυσμάτων υποστήριξης (Support Vector Machines - SVM) είναι ένα διαχωριστικό μοντέλο κατηγοριοποίησης, το οποίο εκπαιδεύει γραμμικά ή μη γραμμικά όρια απόφασης στο χώρο των χαρακτηριστικών, με σκοπό να διαχωρίσει τις κατηγορίες. Πέρα από τη μεγιστοποίηση της διαχωρισιμότητας ανάμεσα σε δύο κατηγορίες, μία μηχανή διανυσμάτων υποστήριξης παρέχει ισχυρές δυνατότητες προσαρμογής, δηλαδή έχει τη δυνατότητα να ελέγχει την πολυπλοκότητα του μοντέλου, έτσι ώστε να εξασφαλίζεται καλή απόδοση γενίκευσης. Λόγω της μοναδικής της ικανότητας να προσαρμόζει την εκπαίδευση, η SVM μπορεί να εκπαιδεύει ιδιαίτερα εκφραστικά μοντέλα, χωρίς να εμφανίζονται προβλήματα υπερ-προσαρμογής. Η SVM έχει ισχυρές ρίζες στη στατιστική θεωρία μάθησης και βασίζεται στην αρχή της δομικής ελαχιστοποίησης κινδύνου. Μία άλλη μοναδική πτυχή της SVM είναι ότι αναπαριστά τα όρια απόφασης χρησιμοποιώντας ένα μόνο υποσύνολο των δειγμάτων εκπαίδευσης, τα οποία είναι δυσκολότερο να κατηγοριοποιηθούν. Το σύνολο αυτό είναι γνωστό ως διανύσματα υποστήριξης. Επομένως, είναι ένα διαχωριστικό μοντέλο, το οποίο επηρεάζεται μόνο από τα στιγμιότυπα εκπαίδευσης που βρίσκονται κοντά στα όρια των δύο κατηγοριών, σε αντιδιαστολή με τη μάθηση της γενικής κατανομής κάθε κατηγορίας [2].

5.2 Λογιστική Παλινδρόμηση

Τα μοντέλα κατηγοριοποίησης, τα οποία αποδίδουν απευθείας ετικέτες κατηγοριών, χωρίς να υπολογίζουν υπό συνθήκη πιθανότητες, ονομάζονται διαχωριστικά μοντέλα. Ένα από τα πιο γνωστά μοντέλα της κατηγορίας αυτής ονομάζεται λογιστική

παλινδρόμηση, το οποίο εκτιμά απευθείας την πρόγνωση ενός στιγμιοτύπου δεδομένων x, χρησιμοποιώντας τις τιμές των χαρακτηριστικών του [2]. Η λογιστική παλινδρόμηση μοιάζει πολύ με τη γραμμική παλινδρόμηση, αλλά δεν μπορεί να πραγματοποιηθεί μέσω αυτής, λόγω της τιμής η οποία υπολογίζεται για μια μεταβλητή. Στη γραμμική παλινδρόμηση, η τιμή της μεταβλητής, η οποία υπολογίζεται, είναι συνεχής, σε αντίθεση με την τιμή της λογιστικής παλινδρόμησης όπου η τιμή είναι διακριτή. Σε αντίθεση με τη γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση μειώνει το εύρος των προβλέψεων και η τιμή που λαμβάνει είναι πάντα μεταξύ 0 και 1.
Για την ανάλυση της λογιστικής παλινδρόμησης, υπολογίζεται αρχικά ο λόγος πιθανότητας που ονομάζεται απόδοση (odds). Εάν θεωρήσουμε το p ως την πιθανότητα επιτυχίας εμφάνισης του συμβάντος και το 1-p την πιθανότητα αποτυχίας εμφάνισης του συμβάντος, τότε ο λόγος πιθανότητας υπολογίζεται από τον τύπο:

$$\text{odds} = \frac{\text{p}}{\text{1−p}} \quad (1)$$

Τέλος, ορίζεται η συνάρτηση logit, που είναι ο φυσικός λογάριθμος της αναλογίας πιθανότητας, ώστε να μπορεί να ενσωματωθεί στο μοντέλο παλινδρόμησης και συμβολίζεται ως:

$$logit(p) = ln(\frac{p}{1-p}) \quad (2)$$

Το αποτέλεσμα του μοντέλου λογιστικής παλινδρόμησης ερμηνεύεται ως λογάριθμος απόδοσης για τη δυνατότητα συμμετοχής στην κλάση.

6. ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

Μια μήτρα σύγχυσης είναι μια τεχνική για τη σύνοψη της απόδοσης ενός αλγορίθμου κατηγοριοποίησης. Ο αριθμός των σωστών και λανθασμένων προβλέψεων συνοψίζεται με τιμές μέτρησης και κατανέμεται ανά κλάση. Στο πρόβλημα των δύο κατηγοριών προσπαθούμε να κάνουμε διακρίσεις μεταξύ των παρατηρήσεων με ένα συγκεκριμένο αποτέλεσμα (πλασματικά τραπεζογραμμάτια), και των κανονικών παρατηρήσεων (γνήσια τραπεζογραμμάτια). Με αυτόν τον τρόπο, μπορούμε να ορίσουμε τη σειρά συμβάντων ως «θετική – positive» και τη σειρά χωρίς συμβάντα ως «αρνητική – negative». Στη συνέχεια, μπορούμε να αντιστοιχίσουμε τη στήλη συμβάντων (event) των προβλέψεων ως «αληθές-true» και των μη συμβάντων (no-event) ως «ψευδές-false». Αυτό μας δίνει:

- "**True Positive**" (TP) για σωστά προβλεπόμενες τιμές συμβάντων (ο αριθμός των περιπτώσεων που αναγνωρίζονται σωστά ως γνήσια τραπεζογραμμάτια).
- "**False Positive**" (FP) για εσφαλμένα προβλεπόμενες τιμές συμβάντων (ο αριθμός των περιπτώσεων που προσδιορίζονται εσφαλμένα ως γνήσια τραπεζογραμμάτια).
- "**True Negative**" (TN) για σωστά προβλεπόμενες τιμές χωρίς συμβάντα (ο αριθμός των περιπτώσεων που αναγνωρίζονται σωστά ως ψεύτικα τραπεζογραμμάτια).
- "**False Negative**" (FN) για εσφαλμένα προβλεπόμενες τιμές χωρίς συμβάντα (ο αριθμός των περιπτώσεων που προσδιορίστηκε εσφαλμένα ως ψεύτικες σημειώσεις).

Μπορούμε να συνοψίσουμε μια μήτρα σύγχυσης ως εξής:

Πίνακας 4: Μήτρα Σύγχυσης.

	Event	**No- event**
Event	TP	FP
No- event	FN	TN

Ακολουθούν οι μήτρες σύγχυσης των αλγορίθμων που χρησιμοποιήσαμε.

Πίνακας 5: Μήτρα Σύγχυσης για την Λογιστική Παλινδρόμηση.

	Event	**No- event**
Event	154	4
No- event	0	117

Πίνακας 6: Μήτρα Σύγχυσης για SVM.

	Event	**No- event**
Event	158	0
No- event	0	117

Οι παραπάνω μήτρες συγχίσεις μπορούν να μας βοηθήσουν να υπολογίσουμε πιο προηγμένες μετρικές αξιολόγησης, οι οποίες παρουσιάζονται παρακάτω.

Οι ακόλουθες μετρικές έχουν χρησιμοποιηθεί για τη μέτρηση της απόδοσης των μοντέλων που εφαρμόστηκαν.

- Accuracy: Το accuracy ενός τεστ είναι η ικανότητά του να διαφοροποιεί σωστά τα πραγματικά και τα πλαστά τραπεζογραμμάτια.
  Accuracy = (TP +TN) / (TP+TN + FP + FN).
- Precision: Το precision μετράει τον αριθμό των χαρτονομισμάτων που ο κατηγοριοποιητής χαρακτήρισε ως γνήσιος είναι πραγματικά γνήσιος.
  Precision = TP / (TP + FP).
- Recall: Το Recall μετράει το πλήθος των θετικών κλάσεων που ο κατηγοριοποιητής προέβλεψε σωστά.
  Recall = TP / (TP + FN).
- F1–Measure: Το F1-score βοηθάει να μετρήσουμε το Recall και το Precision την ίδια στιγμή.
  $\text{F1– Measure} = 2\frac{Precision * Recall}{Precision + Recall}$.

Ακολουθούν οι μετρήσεις αξιολόγησης των αλγορίθμων που χρησιμοποιήσαμε.

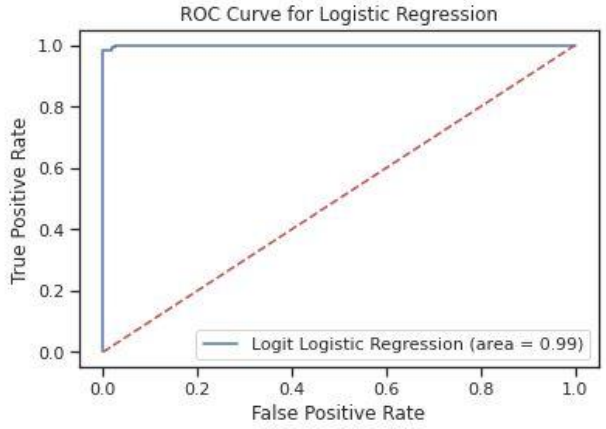
Πίνακας 7: Μετρικές Αξιολόγησης Αλγορίθμων.

	Accuracy	**Precision**	**Recall**	**F1-Measure**
Logistic Regression	98.54	100	97	99
SVM	100	100	100	100

Επιπλέον, μια άλλη μέτρηση απόδοσης για το πρόβλημα είναι η καμπύλη ROC - AUC. Η καμπύλη ROC είναι μια δισδιάστατη γραφική παράσταση ενός κατηγοριοποιητή στην οποία απεικονίζεται το ποσοστό ψευδών θετικών (false positives) στον άξονα Χ και το ποσοστό των αληθώς θετικών (true positives) στον άξονα Υ. Το γράφημα ROC απεικονίζει τις σχετικές αντισταθμίσεις (tradeoffs) που κάνει ένας κατηγοριοποιητής ανάμεσα στα οφέλη (αληθινά θετικά) και τα κόστη (ψευδώς θετικά).

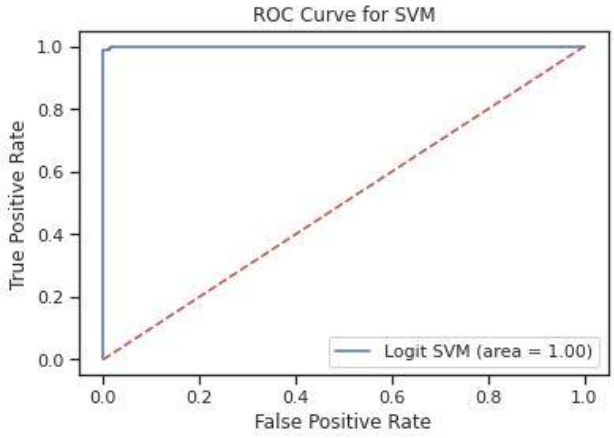
Η AUC αντιπροσωπεύει το βαθμό ή μέτρο διαχωρισμού. Όσο υψηλότερη είναι η AUC, τόσο καλύτερα το μοντέλο μπορεί να προβλέπει στιγμιότυπο της κλάσης 0 ως 0 και στιγμιότυπο της κλάσης 1 ως 1. Με άλλα λόγια, όσο υψηλότερο είναι το AUC, τόσο καλύτερα το μοντέλο διακρίνει μεταξύ γνήσιων και πλαστών τραπεζογραμματίων. Ένα εξαιρετικό μοντέλο έχει AUC κοντά στο 1 που σημαίνει ότι έχει καλό μέτρο διαχωρισμού. Ένα φτωχό μοντέλο έχει AUC κοντά στο 0 που σημαίνει ότι έχει το χειρότερο μέτρο διαχωρισμού.

Παρακάτω στην Εικόνα 5 και Εικόνα 6 παρουσιάζονται οι καμπύλες ROC – AUC των κατηγοριοποιητών που χρησιμοποιήσαμε για να λύσουμε το πρόβλημά μας.



Εικόνα 5: Καμπύλη ROC Λογιστικής Παλινδρόμησης.

Η μετρική AUC για τον κατηγοριοποιητή της Λογιστικής Παλινδρόμησης είναι 0.999621.



Εικόνα 6: Καμπύλη ROC για SVM.

Η μετρική AUC για τον κατηγοριοποιητή της Μηχανής Διανυσμάτων Υποστήριξης (SVM) είναι 0.999830.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην φάση της εκπαίδευσης τα μοντέλα εκπαιδεύτηκαν με το 80% των δεδομένων, δηλαδή με 1097 δείγματα από τα οποία τα 609 δείγματα ήταν γνήσια και 488 ήταν ψεύτικα. Για την δοκιμή των μοντέλων, εξετάστηκαν 275 δείγματα, όπου τα 153 ήταν από γνήσια χαρτονομίσματα και τα 122 από πλαστά. Στον πειραματισμό μας αξιολογήσαμε την απόδοση δύο μοντέλων, της Λογιστικής Παλινδρόμησης και της Μηχανής Διανυσμάτων Υποστήριξης (SVM). Η επιλογή των δύο μοντέλων αυτών ήταν αποτέλεσμα μιας σειράς συγκρίσεων διαφορετικών αλγορίθμων μεταξύ των οποίων ανταγωνίστηκαν οι εξής αλγόριθμοι: Logistic Regression, Linear Discriminant Analysis, Gaussian Naïve Bayes, Support Vector Machines (SVM), K-Nearest Neighbor και Decision Tree. Τα πρώτα αποτελέσματα έπειτα και από την εφαρμογή της μεθόδου Cross Validation έδειξαν ότι μοντέλα της Λογιστικής Παλινδρόμησης και των SVM προβλέπουν με μεγαλύτερη ακρίβεια τα δεδομένα μας. Έπειτα ακολούθησε περαιτέρω διερευνητική αξιολόγηση των δύο μοντέλων αυτών. Παρατηρήσαμε ότι η απόδοση του SVM είναι καλύτερη από της Λογιστικής Παλινδρόμησης και οι αντίστοιχες μετρικές των δύο αυτών κατηγοριοποιητών παρουσιάζονται στον Πίνακα 7. Επιπλέον η μετρική AUC για τα SVM είναι υψηλότερη από εκείνη της Λογιστικής Παλινδρόμησης. Σύμφωνα με τα αποτελέσματά μας, συμπεραίνουμε ότι η Μηχανή Διανυσμάτων Υποστήριξης επιτυγχάνει εξαιρετική κατηγοριοποίηση και καλύτερη από αυτή της Λογιστικής Παλινδρόμησης.

8. ΣΥΝΟΨΗ

Συνοψίζοντας αφού αναλύσαμε διάφορες τεχνικές που χρησιμοποιούνται για την ανίχνευση πλαστών τραπεζογραμματίων, αυτή η μελέτη παρουσιάζει τον έλεγχο γνησιότητας τραπεζογραμματίων χρησιμοποιώντας δύο τεχνικές εποπτευόμενης μάθησης. Εκτελέστηκαν εκτενή πειράματα στο σύνολο δεδομένων χρησιμοποιώντας και τα δύο μοντέλα για να βρούμε το καλύτερο μοντέλο κατηγοριοποίησης

χαρτονομισμάτων. Το ROC και άλλες μετρικές υπολογίστηκαν για να συγκρίνουμε τις επιδόσεις και των δύο τεχνικών. Τα αποτελέσματα έδειξαν ότι η Μηχανή Διανυσμάτων Υποστήριξης (SVM) ξεπερνά την Λογιστική Παλινδρόμηση και δίνει μεγαλύτερο ποσοστό επιτυχίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] https://archive.ics.uci.edu/ml/datasets/banknote+authentication
[2] P. N. Tan, M. Steinbach, A. Karpatne and V. Kumar, 2005 “Introduction to Data Mining”