ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση τεχνικών μείωσης της επίδρασης των ανισοκατανεμημένων δεδομένων κατά την ανίχνευση κακόβουλων ηλεκτρονικών συναλλαγών

ΠΕΦΑΝΗΣ ΝΙΚΗΤΑΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

κ. ΝΙΚΟΛΑΟΣ ΔΟΥΛΑΜΗΣ (καθηγητής Ε.Μ.Π.)

ΟΚΤΩΒΡΙΟΣ 2022

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου και την ευγνωμοσύνη μου στον επιβλέποντα καθηγητή μου κ. Νικόλαο Δουλάμη για την εμπιστοσύνη που μου έδειξε και την πολύτιμη βοήθειά του ώστε να διεκπεραιωθεί με τον καλύτερο τρόπο η παρούσα διπλωματική εργασία. Επίσης, στους υποψήφιους διδάκτορές του κ. Σταύρο Συκιώτη και κ. Ιάσονα Κατσαμένη, οι οποίοι επίσης συνέβαλαν με τη βοήθειά τους και τον χρόνο που διέθεσαν. Τέλος, ένα ευχαριστώ στην οικογένειά μου, τους φίλους μου και τους συναδέλφους μου για όλη την υποστήριξη και κατανόηση που μου παρείχαν κατά τη διάρκεια των σπουδών μου.

Περιεχόμενα

Περίληψη

Η ραγδαία ανάπτυξη της τεχνολογίας κατά τη διάρκεια των τελευταίων ετών έχει συντελέσει στην ψηφιοποίηση ολοένα και περισσότερων υπηρεσιών. Ο τραπεζικός κλάδος σίγουρα έχει συνταχθεί πλήρως με αυτήν τη νέα τάση, προσφέροντας τη δυνατότητα πλέον στον τελικό χρήστη να διεκπεραιώνει μια μεγάλη πληθώρα συναλλαγών ηλεκτρονικά. Η ταχύτατη αυτή ανάπτυξη της τεχνολογίας εγκυμονεί ωστόσο και κάποιους κινδύνους, με σημαντικότερο εξ αυτών να αποτελούν οι κακόβουλες συναλλαγές. Με τον όρο αυτό εννοούμε αυτές τις συναλλαγές που έγιναν από τρίτους εις βάρος των τελικών χρηστών με σκοπό την εξαπάτηση και οικονομική εκμετάλλευση αυτών. Στο πλαίσιο αυτής της εργασίας θα ασχοληθούμε εκτενώς με το συγκεκριμένο πρόβλημα, παρουσιάζοντας μεθόδους για την ανάπτυξη ενός συστήματος ικανού να ανιχνεύει αυτόματα τις κακόβουλες ηλεκτρονικές συναλλαγές. Μελετώντας αυτό το πρόβλημα προέκυψε ένα επιπλέον ζήτημα, αυτό των ανισοκατανεμημένων δεδομένων, καθώς η πλειοψηφία των συναλλαγών ανήκει στην κατηγορία των έγκυρων και όχι των κακόβουλων. Συνεπώς, εκτός από το αρχικό πρόβλημα, αναλύθηκαν και μέθοδοι με τις οποίες θα μπορούσαν να εξαλειφθούν οι συνέπειες ενός ανισοκατανεμημένου συνόλου δεδομένων στην εκπαίδευση ενός συστήματος αυτόματης αναγνώρισης των κακόβουλων συναλλαγών. Αναλύθηκαν τεχνικές για τη μείωση των φαινομένων αυτών, οι οποίες δοκιμάστηκαν σε διαφορετικές αρχιτεκτονικές μηχανικής μάθησης. Η διαφοροποίηση τους έγκειται στο γεγονός ότι αρχικά μεταχειρίζονται τα δεδομένα εκπαίδευσης με τελείως διαφορετικό τρόπο. Έτσι μπορούμε να εξάγουμε γενικότερα συμπεράσματα για την εφαρμογή των αλγορίθμων αυτών στο συνολικό πεδίο των μεθόδων μηχανικής μάθησης.

Λέξεις κλειδιά: μέθοδοι εξομάλυνσης, ανισοκατανεμημένο σύνολο δεδομένων, νευρωνικά δίκτυα, ηλεκτρονικές συναλλαγές, μοντέλα μηχανικής μάθησης

Abstract

The rapid development of technology during the last years has led to the digitalization of more and more services. Banking is one of the leading sectors towards this direction by offering the end user the ability to handle a wide variety of transactions online. However, this rapid development of technology poses some risks, with malicious transactions being the most important among them. By this term we mean those transactions made by third parties at the expense of end users with the purpose of deceiving and financially exploiting them. In the context of this paper we will deal extensively with this problem, presenting methods for developing a system capable of automatically detecting malicious electronic transactions. While studying this problem the additional issue of unevenly distributed data arose, since the majority of transactions fall into the category of valid rather than malicious. Therefore, in addition to the initial issue, methods have been also analyzed on how the consequences of an unevenly distributed data set could be eliminated while training a system for automatic identification of malicious transactions. Several techniques of reducing these effects have been analyzed and tested on different machine learning architectures. Their differentiation lies on the fact that they initially treat the training set data in a completely different way. Thus we can draw more general conclusions for the application of these algorithms in the overall field of machine learning methods.

Key words: smoothing methods, imbalanced dataset, neural networks, electronic transactions, machine learning models

1. Εισαγωγή

Η αντιμετώπιση όλων των προβλημάτων στον πραγματικό κόσμο δεν μπορεί να είναι κοινή. Για κάποια από αυτά η επίλυση μπορεί να είναι μια αρκετά ακριβής διαδικασία λόγω καλά ορισμένων συναρτήσεων και μεθόδων. Παραδείγματος χάριν, ο υπολογισμός της ταχύτητας ενός κινούμενου οχήματος αποτελεί μια τέτοια περίπτωση ή ο υπολογισμός της ελάχιστης διαδρομής σε έναν γράφο από ένα σημείο εκκίνησης σε έναν κόμβο στόχο. Σε αμφότερες τις περιτώσεις υπάρχουν συναρτήσεις ή αλγόριθμοι ούτως ώστε η λύση τους να επιτευχθεί με ελάχιστο ή και μηδενικό σφάλμα.

Ωστόσο υπάρχουν και προβλήματα για τα όποια δεν υπάρχουν ή δεν γίνεται να διατυπωθούν τέτοιες μέθοδοι επίλυσής τους. Η απάντηση στο ερώτημα αν σε μια φωτογραφία υπάρχει ή όχι ένα μήλο αποτελεί μια τέτοια χαρακτηριστική περίπτωση. Μέχρι στιγμής δεν υπάρχει κάποιος αλγόριθμος ή ακριβής συνάρτηση που να αποκρίνεται με βεβαιότητα για την ύπαρξη ή μη του εν λόγω φρούτου (π.χ. κάποιος κανόνας ότι αν περισσότερα από 50 pixels έχουν χρώμα στην οικογένεια του κόκκινου τότε περιέχεται ειδάλλως όχι). Η μαθηματική λοιπόν περιγραφή και επίλυση του συγκεκριμένου ζητήματος παραμένει δύσκολη, μολονότι για ένα ανθρώπινο ον η απάντηση είναι ιδιαιτέρως εύκολη και διαισθητική. Αυτό συμβαίνει διότι οι άνθρωποι μαθαίνουμε αντιμετωπίζουμε τέτοιου είδους προβλήματα διαισθητικά, κατόπιν πολλών υποδείξεων και παρατήρησης αρχίζουμε να αντιλαμβανόμαστε και να αναγνωρίζουμε τα διάφορα αντικείμενα.

Κάπως έτσι λειτουργούν και οι αλγόριθμοι μηχανικής μάθησης. Λαμβάνουν ως είσοδο ένα σύνολο από επισημειώμενα δεδομένα και προσαρμόζονται με τέτοιο τρόπο ώστε να μπορούν να προβλέπουν σωστά, στο σύνολο, αυτό το πρόβλημα. Επιστρέφοντας στο προαναφερθέν παράδειγμα, αν επιθυμούμε ένα σύστημα να είναι ικανό να απαντάει στην ερώτηση πότε μία εικόνα περιέχει ένα μήλο και πότε όχι, θα εισάγαμε ένα πλήθος από εικόνες , κάποιες εκ των οποίων θα περιείχαν και κάποιες όχι, και το σύστημα στη συνέχεια θα «εκπαιδευόταν» με τέτοιο τρόπο ώστε να μπορεί να κατηγοριοποιεί σωστά τα δεδομένα αυτά. Το προαναφερθέν σύνολο καλείται σύνολο εκπαίδευσης, ενώ το αντίστοιχο πρόβλημα ονομάζεται πρόβλημα ταξινόμησης επειδή το σύστημα καλείται να ξεχωρίζει τα δεδομένα μεταξύ των κλάσεων. Στο σημείο αυτό να σημειώσουμε ότι αυτό δεν είναι το μοναδικό πρόβλημα που μπορούν να προσφέρουν λύση τα συστήματα μηχανικής μάθησης, ωστόσο σε αυτό θα γίνει εκτενής αναφορά παρακάτω.

Πρώτο βήμα για την αντιμετώπιση του συγκεκριμένου ζητήματος ειναι η εύρεση ενός επαρκούς συνόλου με επισημειωμένα δεδομένα όλων των κλάσεων. Το επόμενο ερώτημα που προκύπτει είναι κατά πόσο υπάρχουν επαρκή δεδομένα από όλες τις κλάσεις, παρεμφερή μάλιστα και σε πλήθος. Με άλλα λόγια, αν το σύνολο δεδομένων που θα χρησιμοποιηθεί ως βάση είναι ομοιόμορφα κατανεμημένο. Για να γίνει πιο εύκολα κατανοητό το συγκεκριμένο ζήτημα θα αναφέρουμε το εξής, πιο δύσκολο πρόβλημα από τον τομέα της ιατρικής. Ας υποθέσουμε ότι

επιθυμούμε την κατασκευή ενός συστήματος ικανού να ανιχνεύει την πιθανή ύπαρξη όγκου σε μια ακτινογραφία ή τομογραφία. Αρχικά θα πρέπει να αναπτυχθεί ένα σύνολο δεδομένων με εικόνες και ετικέτες, στο οποίο ωστόσο ιδανικά θα ήταν επιθυμητό η κατανομή των ετικετών να είναι ομοιόμορφη, που στην περίπτωσή μας σημαίνει ίδιο ή όσο το δυνατόν πλησιέστερο πλήθος εικόνων που περιέχουν και δεν περιέχουν αντίστοιχα κάποιας μορφής όγκο. Γίνεται άμεσα αντιληπτό εντούτοις ότι η συγκεκριμένη απαίτηση δεν μπορεί να ικανοποιηθεί. Κατά συνέπεια, η εκπαίδευση ενός τέτοιου συστήματος καθίσταται εξαρχής δύσκολη. Ενδεχομένως την ίδια δυσκολία να αντιμετωπίζαμε και αν είχαμε έναν άνθρωπο ως «σύστημα». Ένας σχετικά άπειρος ιατρός δεν θα είναι σε θέση να ξεχωρίσει και να προβλέψει με ακρίβεια την ύπαρξη ενός όγκου παρατηρώντας μόνο τομογραφίες ατόμων που είναι καθαρές, υγιών ανθρώπων δηλαδή. Ακόμα και αν έχει δει ορισμένες εικόνες, αν αυτές είναι ελάχιστες, για παράδειγμα 1%, τότε είναι πολύ δύσκολο να μάθει να λύνει αποδοτικά το πρόβλημα αν δε χρησιμοποιήσει άλλη γνώση (βιβλία, εμπειρίες άλλων κ.α.).

Μπορούμε λοιπόν να πούμε με βεβαιότητα ότι αν το πρόβλημα που εξετάζουμε δυσκολεύει τόσο ένα ανθρώπινο ον, είναι αναμενόμενο τα αποτελέσματα εκπαίδευσης ενός συστήματος μηχανικής μάθησης να είναι ακόμα πιο αποθαρρυντικά. Και στην πράξη αυτό είναι που συμβαίνει, όπως αναφέρεται και στη βιβλιογραφία [4, 5, 6]. Το πρόβλημα της ανισοκατανομής των δεδομένων εκπαίδευσης αποτελεί μέχρι και σήμερα ανοιχτό ερευνητικό πεδίο με νέες μεθόδους για τη μείωση της επίδρασης του να προτείνονται στη βιβλιογραφία.

Συνοψίζοντας, στο πλαίσιο της παρούσας εργασίας θα μελετήσουμε και παρουσιάσουμε διάφορες μεθόδους για τη βελτίωση της απόδοσης των συστημάτων που είναι «εκπαιδευμένα» έχοντας δεχθεί ως είσοδο ανισοκατανεμημένα δεδομένα.

2. Fraud Detection

Όπως προαναφέρθηκε, στην παρούσα εργασία θα εξετάσουμε το πως μπορούμε να ανιχνεύσουμε αυτόματα μια τραπεζική απάτη[3]. Η επιλογή έγινε λόγω του ιδιαίτερου χαρακτηριστικού που παρουσιάζει το εν λόγω εγχείρημα, και το οποίο είναι η άνιση κατανομή του συνόλου δεδομένων. Βοήθησε επίσης και το ότι πρόκειται για ένα σύνολο δεδομένων το όποιο έχει μελετηθεί εκτενώς στη βιβλιογραφία[3]. Στο πλαίσιο των πάρα πολλών καθημερινών τραπεζικών συναλλαγών που εκτελούνται ελάχιστες είναι αυτές που είναι προϊόν απάτης. Αν και αυτό είναι ευτυχές γεγονός, εντούτοις αυτό οδηγεί στην ανισοκατανομή των συνόλων δεδομένων με τραπεζικές συναλλαγές, καθώς αυτά περιέχουν πάρα πολλές έγκυρες και ελάχιστες μη έγκυρες (απάτες) συναλλαγές.

Αυτό καθιστά τη σημαντική διαδικασία της εκπαίδευσης ενός συστήματος αυτόματης ανίχνευσης μη έγκυρων συναλλαγών ιδιαίτερα απαιτητική, ούτως ώστε να μπορούν αυτές να προβλεφθούν εγκαίρως και με ακρίβεια. Επίσης καθίσταται υποχρεωτική η προσπάθεια για εξάλειψη της ανισοκατανομής των δεδομένων..

Μελετώντας το υφιστάμενο πρόβλημα πρέπει να ληφθούν υπόψη και άλλοι παράγοντες κια χαρακτηριστικά. Θα πρέπει να απαντήσουμε στο ερώτημα αν θα ήταν ένα ωφέλιμο ένα σύστημα ανίχνευσης που θα χαρακτήριζε όλες τις συναλλαγές ως έγκυρες και καμία ως απάτη, κάτι το οποίο είναι ιδιαιτέρως πιθανό να συμβεί στην περίπτωση που μελετάμε με το τόσο ανισοκατανεμημένο σύνολο δεδομένων. Την ίδια ερώτηση οφείλουμε να απαντήσουμε και για την αντίθετη περίπτωση, αυτή κατά την οποία το σύστημα ταξινομεί όλες τις συναλλαγές ως μη έγκυρες - απάτες -, αυτό θα είχε κάποιο νόημα; Ενδεχομένως να είχε για τον τελικό χρήστη της τράπεζας, ο οποίος τουλάχιστον θα ήταν εξασφαλισμένος σε μια απόπειρα εξαπάτησής του, ωστόσο αυτή η περίπτωση είναι καταστροφική για την τράπεζα, αφού κάθε περίπτωση θα απαιτούσε εξακρίβωση, με ό,τι αυτό συνεπάγεται σε κατανάλωση πόρων, χρόνου κλπ. Συνεπώς, όλα συνηγορούν στο εξής δίλημμα: Είναι προτιμότερο ένα σύστημα που εντοπίζει όλες τις έγκυρες, και ας έχει εσφαλμένα ταξινομήσει μέσα σε αυτές και μερικές απάτες, ή το αντίθετο;

¨Όσον αφορά τον τελικό χρήστη γίνεται εύκολα αντιληπτό ότι η απάντηση είναι προφανής, καθώς επιθυμητό είναι ένα σύστημα το οποίο θα βρίσκει πάντα αν υπάρχει η περίπτωση απάτης και ας βγάζει μερικές φορές κάποια έγκυρη συναλλαγή ως άκυρη. Τα νούμερα ωστόσο μάλλον έχουν αντίθετη άποψη. Αν η ερώτηση διαπιστωθεί διαφορετικά σίγουρα η διαισθητική απάντηση θα είναι διαφορετική. Αν η επιλογή έπρεπε να γίνει ανάμεσα σε δύο συστήματα όπου το ένα έχει 99.9% ακρίβεια και το άλλο 50% ποιο θα επιλέγαμε; Στο υπό εξέταση ζήτημα δεν μπορεί να απαντηθεί τόσο εύκολα επειδή είναι ιδιαιτέρως κρίσιμο το πότε κάνει λάθος το σύστημα. Είναι πολύ προτιμότερο έγκυρες συναλλαγές να ταξινομούνται ως απάτες παρά το ανάποδο. Δεν πρέπει ωστόσο να

ξεχάσουμε τον γενικό κανόνα από τον οποίο δεν πρέπει να παρεκκλίνει το σύστημά μας. Ιδανικά επιθυμητό είναι αυτό το σύστημα που δεν θα του ξεφύγει καμία συναλλαγή που αποτελεί προϊόν απάτης.

Αυτή αποτελεί και τη βασική αρχή που ακολουθήθηκε στη μεθοδολογία μας και η οποία θα παρουσιαστεί αναλυτικά παρακάτω.

Το σύνολο δεδομένων που αναλύθηκε στο πλαίσιο της παρούσας εργασίας είναι το Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine[1] και το όποιο ανακτήθηκε μέσω του αποθετηρίου kaggle[2]. Αυτό περιέχει συναλλαγές μέσω πιστωτικής κάρτας οι οποίες έγιναν το Σεπτέμβριο του 2013 από Ευρωπαίους κατόχους. Το χρονικό έυρος μέσα στο οποίο πραγματοποιήθηκαν οι συναλλαγές είναι δύο μέρες και βρέθηκαν 492 μη- έγκυρες από τις συνολικές 284,807 συναλλαγές. Από αυτό προκύπτει ότι το ποσοστό της θετικής κλάσης, δηλαδή οι συναλλαγές που τελικά ήταν απάτες, ανέρχεται μόλις στο 0.172% του συνόλου, καθιστώντας το εν λόγω dataset ως ένα από τα πλέον ανισοκατανεμημένα της περιοχής. Αυτό θα μας βοηθήσει να αναδείξουμε πιο παραστατικά τα προβλήματα που περιγράψαμε και αφορούν ένα ανισοκατανεμημένο σύνολο, από την άλλη όμως θα κάνει και την προσπάθεια βελτίωσης των αποτελεσμάτων πιο δύσκολη.

Για ευνόητους λόγους τα δεδομένα των καταναλωτών που παρέχονται είναι ανώνυμα. Έτσι, για κάθε συναλλαγή παρέχεται ένα διάνυσμα 29 τιμών τα οποία έχουν εξαχθεί μέσω της μεθόδου PCA. Επίσης παρατίθεται το ποσό ο χρόνος εκτέλεσης της συναλλαγής. Όσον αφορά στην ετικέτα για την εγκυρότητα της κάθε συναλλαγής, η τιμή 0 υποδηλώνει μια έγκυρη και η τιμή 1 μια μη έγκυρη - απάτη.

Στη συνέχεια παρουσιάζονται τα πεδία του συνόλου δεδομένων καθώς και οι πέντε πρώτες σειρές του συνόλου των δεδομένων.

```
Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10',
       'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20',
       'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount',
       'Class'],
      dtype='object')
```

Εικόνα 1 Τα πεδία του συνόλου Credit Card Fraud Detection

[1] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
[2] https://www.kaggle.com/

	scaled_amount	scaled_time	V1	V2	V3	V4	V5	V6
0	1.783274	-0.994983	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388
1	-0.269825	-0.994983	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361
2	4.983721	-0.994972	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499
3	1.418291	-0.994972	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203
4	0.670579	-0.994960	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921

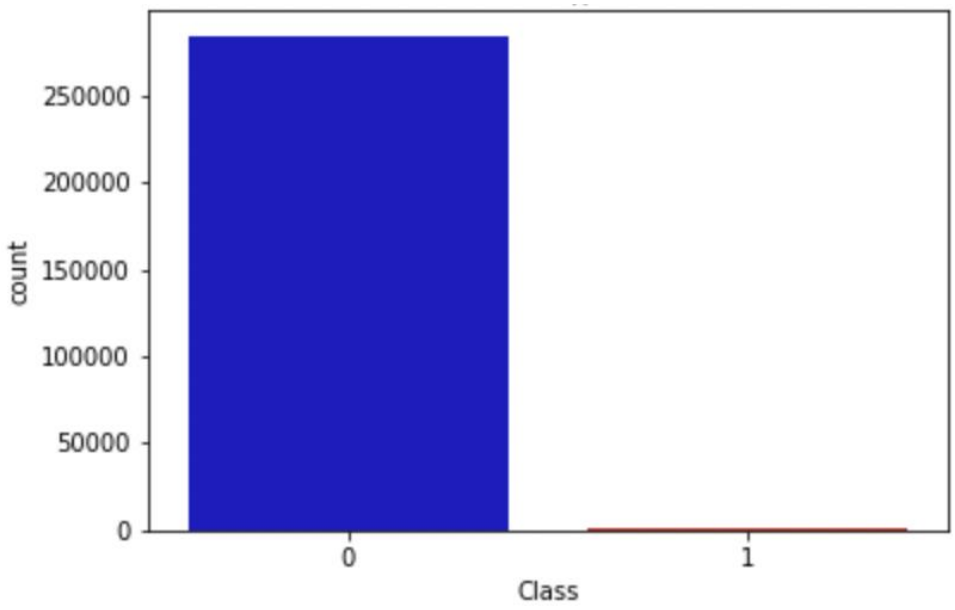
V7	V8	V9	V10	V11	V12	V13	V14
0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.991390	-0.311169
-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.489095	-0.143772
0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.717293	-0.165946
0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.178228	0.507757	-0.287924
0.592941	-0.270533	0.817739	0.753074	-0.822843	0.538196	1.345852	-1.119670

V15	V16	V17	V18	V19	V20	V21	V22
1.468177	-0.470401	0.207971	0.025791	0.403993	0.251412	-0.018307	0.277838
0.635558	0.463917	-0.114805	-0.183361	-0.145783	-0.069083	-0.225775	-0.638672
2.345865	-2.890083	1.109969	-0.121359	-2.261857	0.524980	0.247998	0.771679
-0.631418	-1.059647	-0.684093	1.965775	-1.232622	-0.208038	-0.108300	0.005274
0.175121	-0.451449	-0.237033	-0.038195	0.803487	0.408542	-0.009431	0.798278

V21	V22	V23	V24	V25	V26	V27	V28	Class
-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	0
-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	0
0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	0
-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	0
-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	0

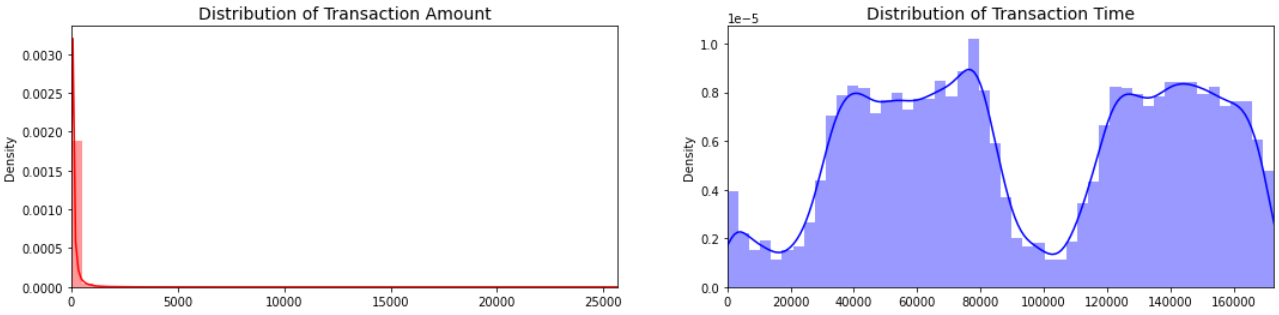
Εικόνα 2 Τα πεδία του συνόλου Credit Card Fraud Detection μαζί με τις 5 πρώτες γραμμές του.

Θα ακολουθήσει μία ανασκόπηση των δεδομένων τα οποία δεν έχουν παραχθεί με τη μέθοδο PCA [7]. Στο παρακάτω διάγραμμα αποτυπώνονται τα στιγμιότυπα κάθε κλάσης, κάνοντας δια της γραφικής μεθόδου ακόμα πιο κατανοητό το πόσο έντονο είναι το πρόβλημα των ανισοκατανεμημένων δεδομένων. Παρατηρούμε πόσο μεγαλύτερη είναι η κλάση 0 (έγκυρες) σε σχέση με την κλάση 1(απάτες).



Εικόνα 3 Το ιστόγραμμα του συνόλου δεδομένων. Από το διάγραμμα αυτό φαίνεται το πρόβλημα της ανισοκατανομής των κλάσεων για το συγκεκριμένο σύνολο δεδομένων.

Στα επόμενα δύο διαγράμματα παρουσιάζονται οι κατανομές των μεταβλητών amount και time στα δεδομένα. Η κατανομή του χρόνου είναι σχεδόν τυχαία μιας και δεν μπορούν να συσχετιστούν μεταξύ τους οι χρονικές στιγμές των συναλλαγών, ενώ τα ποσά (amount) είναι τα περισσότερα μικρά και κυμαίνονται από 500 μέχρι 1000.



Εικόνα 4 Η κατανομή των πεδίων "Amount" και "Time" για το Credit Card Fraud Detection.

2.1. Ανισοκατανεμημένα Δεδομένα

Όπως αναφέρεται παραπάνω, όταν τα δεδομένα ενός συνόλου εκπαίδευσης δεν είναι ομοιόμορφα κατανεμημένα μεταξύ των κλάσεων τότε η εκπαίδευση ενός συστήματος μηχανικής μάθησης καθίσταται αρκετά δύσκολη. Ωστόσο, η ανισοκατανομή των δεδομένων δεν αποτελεί το μοναδικό πρόβλημα.

Ένα ακόμα ζήτημα είναι αυτό της αξιολόγησης των συστημάτων. Για παράδειγμα, όπως παρουσιάστηκε και προηγουμένως, το ποσοστό της θετικής κλάσης επί του συνόλου των δεδομένων αποτελεί το 0.172%. Κατά συνέπεια, ένα σύστημα το οποίο θα αποκρινόταν πάντα, έστω τυφλά, ότι ανιχνεύει την κλάση μηδέν, θα είχε μεν επιτυχία της τάξης του 100 - 0.172 % = 99.828%, ουσιαστικά όμως αυτό δεν θα είχε καμία απολύτως σημασία. Ένα τέτοιο σύστημα δεν θα είχε εκπαιδευτεί σωστά, δεν θα ήταν αξιόπιστο. Αυτό το ενδεχόμενο προφανώς και δεν είναι αποδεκτό, μιας και για να κρίνουμε τις μεθόδους που παρουσιάζουμε θα πρέπει τα μοντέλα να αξιολογηθούν σωστά. Επίσης, αν ένα τέτοιο μοντέλο σαν του προηγούμενου παραδείγματος περάσει την παραγωγή, βασισμένο σε δεδομένα που προκύπτουν από λανθασμένη αξιολόγηση, τότε αποτελεί σοβαρό κίνδυνο.

Σε αυτό το σημείο ας ορίσουμε το ποσοστό επιτυχίας μιας ταξινόμησης δεδομένων. Αυτό θα ισούται με το πηλίκο του πλήθους των σωστών προβλέψεων προς τις συνολικές που αφορούν το εξεταζόμενο δείγμα. Πρόκειται για μια αρκετά ικανοποιητική μετρική όσον αφορά σε ισοκατανεμημένα δεδομένα, όχι όμως και στην περίπτωση ενός ανισοκατανεμημένου σετ. Αυτός είναι και ο λόγος που στην βιβλιογραφία προτείνονται εναλλακτικές μετρικές όπως το **precision**, το **recall** και η **f1**. Στη συνέχεια ακολουθεί αναλυτική παρουσίαση αυτών.

2.2. Μετρικές F1, recall, precision

Αρχικά θα εισαχθεί ο πίνακας σύγχυσης (confusion matrix), μαζί με τα πεδία του, ούτως ώστε να γίνουν αυτές οι μετρικές όσο το δυνατόν πιο κατανοητές. Εύλογα δημιουργείται η απορία τι είναι ο πίνακας σύγχυσης. Πρόκειται για μια συγκεντρωτική παρουσίαση των αποτελεσμάτων πρόβλεψης ενός προβλήματος ταξινόμησης. Σε αυτόν τον πίνακα παρουσιάζονται οι σωστές και λανθασμένες προβλέψεις και διαχωρίζονται ανά κατηγορία. Ουσιαστικά μέσω αυτού του πίνακα παρουσιάζονται τα σφάλματα στα οποία υποπίπτει το εξεταζόμενο μοντέλο, καθώς και ο τύπος αυτών. Οι τιμές του πίνακα αυτού είναι οι ακόλουθες:

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Εικόνα 5 Τυπική μορφή confusion matrix

Ο εν λόγω πίνακας χωρίζεται σε στήλες και γραμμές όπου οι τιμές negative, positive στις γραμμές αφορούν στην πραγματική τιμή των δειγμάτων ενώ οι αντίστοιχες τιμές στις στήλες αυτών που έγιναν predicted από το μοντέλο. Εν συνεχεία ορίζουμε ως:

- **True Negative**: Ο αριθμός των δειγμάτων που προβλέφθηκαν σωστά ως αρνητικά (ήταν negative και προβλέφθηκαν σαν negative)

- **False Negative**: Ο αριθμός των δειγμάτων που είναι θετικά - δηλαδή απάτη - και προβλέφθηκαν λανθασμένα ως αρνητικά - δηλαδή ως έγκυρες αλλαγές. Στόχος είναι αυτή η περίπτωση να τείνει στο μηδέν ή, αν είναι δυνατόν, ιδανικά να εξαλειφθεί.

- **True Positive**: Ο αριθμός των δειγμάτων που ήταν θετικά και προβλέφθηκαν σωστά ως θετικά. Πρόκειται για τις αυτές τις συναλλαγές που όντως ήταν απάτη και ο αλγόριθμος κατάφερε να τις εντοπίσει. Ιδανικά, θα θέλαμε ο αριθμός αυτός να είναι όσο το δυνατόν μεγαλύτερος, να μην ξεφύγει καμία από τις απάτες, ο αλγόριθμος να είναι τόσο αποδοτικός που να τις ανιχνεύσει όλες.

- **False Positive**: Ο αριθμός των δειγμάτων που ήταν αρνητικά και προβλέφθηκαν λανθασμένα ως θετικά. Στην περίπτωση μας αυτό σημαίνει ότι η συναλλαγή ήταν έγκυρη συναλλαγή και προβλέφθηκε λανθασμένα ως απάτη. Τα λάθη αυτά δεν είναι τόσο σημαντικά για το συγκεκριμένο πρόβλημα όσο αυτά της περίπτωσης του false negative.

Με βάση αυτόν τον πίνακα μπορούμε εύκολα να δούμε σε ποια σημεία το σύστημα αποτυγχάνει και να εστιάσουμε σε αυτά. Ακολουθεί η ανάλυση τριών μετρικών, οι οποίες αξιοποιούν τις τιμές των πεδίων του προαναφερθέντος πίνακα για να αξιολογήσουν το μοντέλο ταξινόμησης. Με αυτόν τον τρόπο προκύπτουν χρήσιμα συμπεράσματα.

- **Ακρίβεια (Precision)**: Ορίζεται ως ο λόγος των δειγμάτων που είναι θετικά και τα οποία ο αλγόριθμος προβλέπει σωστά ως θετικά προς το συνολικό πλήθος των δειγμάτων που προβλέφθηκαν ως θετικά, είτε αυτό έγινε ορθώς είτε ψευδώς. Ο μαθηματικός τύπος της συγκεκριμένης μετρικής είναι ο ακόλουθος:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Πρακτικά η μετρική αυτή μας δείχνει πόσα δείγματα που ο αλγόριθμος πρόβλεψε ως θετικά είναι σωστά. Η ακρίβεια είναι μια χρήσιμη μετρική όταν το κόστος των ψευδώς θετικών (false positives) είναι υψηλό. Χαρακτηριστική περίπτωση αποτελεί ο εντοπισμός spam e-mail. Εν προκειμένω, ένα ψευδώς ανίχνευση θετικού δείγματος σημαίνει ότι ένα email που δεν είναι spam θα καταλήξει στην ανεπιθύμητη αλληλογραφία. Ο χρήστης φυσικά δεν το θέλει αυτό, καθώς έτσι μπορεί να χαθεί ένα κρίσιμο για αυτόν μήνυμα. Οπότε σε ένα τέτοιο σύστημα θέλουμε η τιμή της ακρίβειας να είναι υψηλή.

- **Ανάκληση (Recall):** Ορίζεται ως ο λόγος των δειγμάτων που προβλέφθηκαν σωστά ως θετικά προς το σύνολο των δειγμάτων που όντως ήταν θετικά, είτε προβλέφθηκαν είτε όχι.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Ουσιαστικά η εν λόγω μετρική μας υποδεικνύει πόσα από τα θετικά δείγματα ανιχνεύθηκαν σωστά από τον αλγόριθμο. Η χρησιμότητα της συγκεκριμένης μετρικής φαίνεται όταν το κόστος των ψευδώς αρνητικών είναι υψηλό. Το πρόβλημα που εξετάζουμε στο πλαίσιο της παρούσας εργασίας αποτελεί μια τέτοια περίπτωση. Ένα σύστημα στον τομέα της ιατρικής που κάνει αυτόματη διάγνωση ασθενειών αποτελεί μια δεύτερη. Επιστρέφοντας στο ζήτημα ανίχνευσης κακόβουλων ηλεκτρονικών συναλλαγών, σε ένα σύστημα αυτόματης ανίχνευσης αυτών η ταξινόμηση ενός θετικού δείγματος, δηλαδή μιας κακόβουλης συναλλαγής, στις έγκυρες (όχι απάτη) μπορεί να επιφέρει σημαντικές συνέπειες στον χρήστη.

Από τις δύο προαναφερθείσες μετρικές παρατηρούμε το εξής: κατασκευάζοντας ένα σύστημα ικανό να προβλέπει περισσότερα δείγματα ως θετικά θα έχει ως αποτέλεσμα την αύξηση της τιμής της μετρικής ανάκλησης, επειδή μέσα σε αυτά θα προβλέπει και περισσότερα true positives. Αντιθέτως, αυτό θα οδηγήσει σε μείωση της τιμής της πρώτης μετρικής που παρουσιάσαμε, καθώς θα υπάρξει αύξηση της τιμής του παρονομαστή, όπως γίνεται εύκολα αντιληπτό από τη μαθηματική έκφραση της μετρικής precision. Προκειμένου να υπάρξει μια ισορροπία και άμβλυνση αυτού του φαινομένου ορίζεται η ακόλουθη μετρική F1 με τον ακόλουθο τύπο:

- **F1**: ο σταθμισμένος μέσος μεταξύ της ακρίβειας και της ανάκλησης,

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

και χρησιμοποιείται όπως προείπαμε όταν θέλουμε να κρατήσουμε μία ισορροπία μεταξύ των δύο μετρικών precision και recall. Επειδή η ακρίβεια δεν αποτελεί καλή μετρική για την αξιολόγηση συστημάτων εκπαιδευμένων σε μη ισορροπημένα δεδομένα, μπορεί κάλλιστα να χρησιμοποιηθεί και η F1.

Ερμηνεύοντας τις παραπάνω μετρικές και τα αποτελέσματα αυτών προκύπτουν τα εξής χρήσιμα συμπεράσματα:

- υψηλές τιμές ανάκλησης και ακρίβειας: η κατηγορία ανιχνεύεται ικανοποιητικά από το μοντέλο
- χαμηλή ανάκληση + υψηλή ακρίβεια: το μοντέλο δεν μπορεί να ανιχνεύσει ικανοποιητικά την κατηγορία, ωστόσο είναι αρκετά αξιόπιστο όταν το επιτυγχάνει
- υψηλή ανάκληση + χαμηλή ακρίβεια: η κατηγορία είναι καλά ανιχνευμένη, αλλά το μοντέλο περιλαμβάνει επίσης δεδομένα που ανήκουν σε άλλες κατηγορίες
- χαμηλή ανάκληση + χαμηλή ακρίβεια: η κατηγορία δεν ανιχνεύεται ικανοποιητικά από το μοντέλο

3. Μέθοδοι Εξομάλυνσης

Στη συγκεκριμένη παράγραφο θα παρουσιάσουμε κάποιες τεχνικές με τις οποίες μπορεί να μειωθεί το πρόβλημα της ανισοκατανομής των δεδομένων το οποίο αναφέραμε προηγουμένως. Η εκπαίδευση οποιουδήποτε συστήματος αναγνώρισης και ταξινόμησης των συναλλαγών σε έγκυρες και μη γίνεται πιο δύσκολο, όταν η συντριπτική πλειοψηφία των διαθέσιμων δεδομένων ανήκει στη μια κλάση (στην περίπτωσή μας στην αρνητική - έγκυρες συναλλαγές).

Πιο συγκεκριμένα θα μελετήσουμε τις παρακάτω τεχνικές κατά σειρά: υποδειγματοληψία, υπερδειγματοληψία, μέθοδο Smote, Over-Undersampling καθώς και μια regularization τεχνική. Το θεωρητικό κομμάτι της εκάστοτε τεχνικής παρουσιάζεται παρακάτω ενώ στο κεφάλαιο των πειραμάτων αναλύονται τα αποτελέσματα των μεθόδων αυτών.

3.1.Μέθοδος Υποδειγματοληψίας

Η πρώτη από αυτές ονομάζεται τεχνική της υποδειγματοληψίας. Σκοπός της είναι η μείωση της επίδρασης των ανισοκατανεμημένων δεδομένων στην εκπαίδευση του συστήματος [16]. Οι δύο κύριοι άξονες για τον μετασχηματισμό ενός μη ισορροπημένου συνόλου δεδομένων είναι η αφαίρεση παραδειγμάτων από την κλάση πλειοψηφίας (εν προκειμένω των έγκυρων συναλλαγών), που ονομάζεται υποδειγματοληψία, και η αντιγραφή παραδειγμάτων από την κατηγορία μειοψηφίας ώστε αυτά να αυξηθούν, η οποία καλείται υπερδειγματοληψία. Στην πρώτη περίπτωση πρακτικά μειώνουμε το πλήθος των δεδομένων, των περιπτώσεων δηλαδή που θα εισαχθούν στο σύστημα προς εκπαίδευση, επιλέγοντας να αφαιρέσουμε γεγονότα από την πλειοψηφική τάξη. Αυτό συμβάλλει στη μείωση της ανισοκατανομής. Αντιθέτως, η δεύτερη τεχνική αφορά στον πολλαπλασιασμό των δειγμάτων με την επανάληψη δειγμάτων που ανήκουν στην κατηγορία με τα λιγότερα δείγματα, την κλάση μειοψηφίας. Στη συνέχεια θα ακολουθήσει εκτενέστερη ανάλυση αυτής της τεχνικής.

Ὅπως προείπαμε, η εν λόγω τεχνική της τυχαίας υποδειγματοληψίας περιλαμβάνει τη διαγραφή περιπτώσεων, από την κλάση του συνόλου εκπαίδευσης με το μεγαλύτερο πλήθος, τα οποία επιλέγονται τυχαία. Προκύπτει με αυτόν τον τρόπο ένα μετασχηματισμένο πακέτο δεδομένων, στο οποίο η μια τάξη (μειοψηφίας και στην περίπτωσή μας οι απάτες) έχει παραμείνει άθικτη, ενώ

στην άλλη έχει περικοπεί το πλήθος των δεδομένων της. Δυνητικά αυτή η διαδικασία μπορεί να επαναλαμβάνεται έως ότου πετύχουμε το επιθυμητό αποτέλεσμα, ήτοι ίδιο πλήθος δεδομένων σε αμφότερες τις κλάσεις. Η χρησιμοποίηση αυτής της τεχνικής προϋποθέτει ότι το διαθέσιμο σύνολο δεδομένων θα περιλαμβάνει ικανοποιητικό αριθμό περιπτώσεων της κλάσης μειοψηφίας.

Ωστόσο, αν κάποιος θέλει να εφαρμόσει τη συγκεκριμένη μέθοδο θα πρέπει να λάβει υπόψη του και τα μειονεκτήματα αυτής. Παραδείγματος χάρη, ελλοχεύει πάντα ο κίνδυνος ανάμεσα στα παραδείγματα της πλειοψηφικής τάξης που θα επιλεγούν για διαγραφή να βρίσκονται και κάποια χρήσιμα, κάτι το οποίο δεν μπορεί να ελεγχθεί. Αυτά τα δεδομένα που θα διαγραφούν μπορεί να είναι ιδιαιτέρως χρήσιμα, να περιέχουν κάποια κρίσιμα πληροφορία απαραίτητη στον χρήστη για τη λήψη αποφάσεων. Δεδομένου ότι τα παραδείγματα από την πλειοψηφική τάξη διαγράφονται τυχαία, δεν υπάρχει τρόπος να εντοπιστούν ή να διατηρηθούν τα "καλά" ή αυτά που είναι περισσότερα πλούσια σε πληροφορίες. Όποτε αυτό μπορεί να δημιουργήσει σημαντικά προβλήματα κατά την εκμάθηση των μοντέλων καθώς και να έχουμε διαφορετική απόδοση κάθε φορά μετά την εφαρμογή της συγκεκριμένης μεθόδου. Γίνεται επομένως εύκολα κατανοητό ότι ένα σύστημα ταξινόμησης με «καλύτερο» σύνολο δεδομένων, εννοώντας αυτό στο οποίο συμπτωματικά θα έχουν διατηρηθεί παραδείγματα με περισσότερη πληροφορία, θα υπερτερεί όσον αφορά στην απόδοσή του συγκριτικά με ένα άλλο σύστημα που θα έχουν αφαιρεθεί τα κρίσιμα δεδομένα.

Επομένως, αν χρησιμοποιήσουμε ένα σετ δεδομένων με πλήθος χιλίων δειγμάτων και, υποθέτοντας ότι η ασθενής (μειοψηφίας) κλάση αριθμεί μόνο εκατό εξ αυτών, τότε εφαρμόζοντας τη μέθοδο της υποδειγματοληψίας θα προκύψει ένα σύνολο με διακόσια δείγματα, με κάθε κλάση να αποτελείται με τα μισά εξ αυτών. Η τυχαία αυτή επιλογή των δειγμάτων που θα «κοπούν» θα καθορίσει την απόδοση του συστήματος ταξινόμησης. Στο σημείο αυτό να προσθέσουμε ότι στο παρόν σύνολο δεδομένων ο αριθμός των δειγμάτων που ανήκουν στην κλάση μειοψηφίας είναι πολύ μικρός, ενδεχομένως και όχι επαρκής για τη σωστή μοντελοποίηση των κλάσεων.

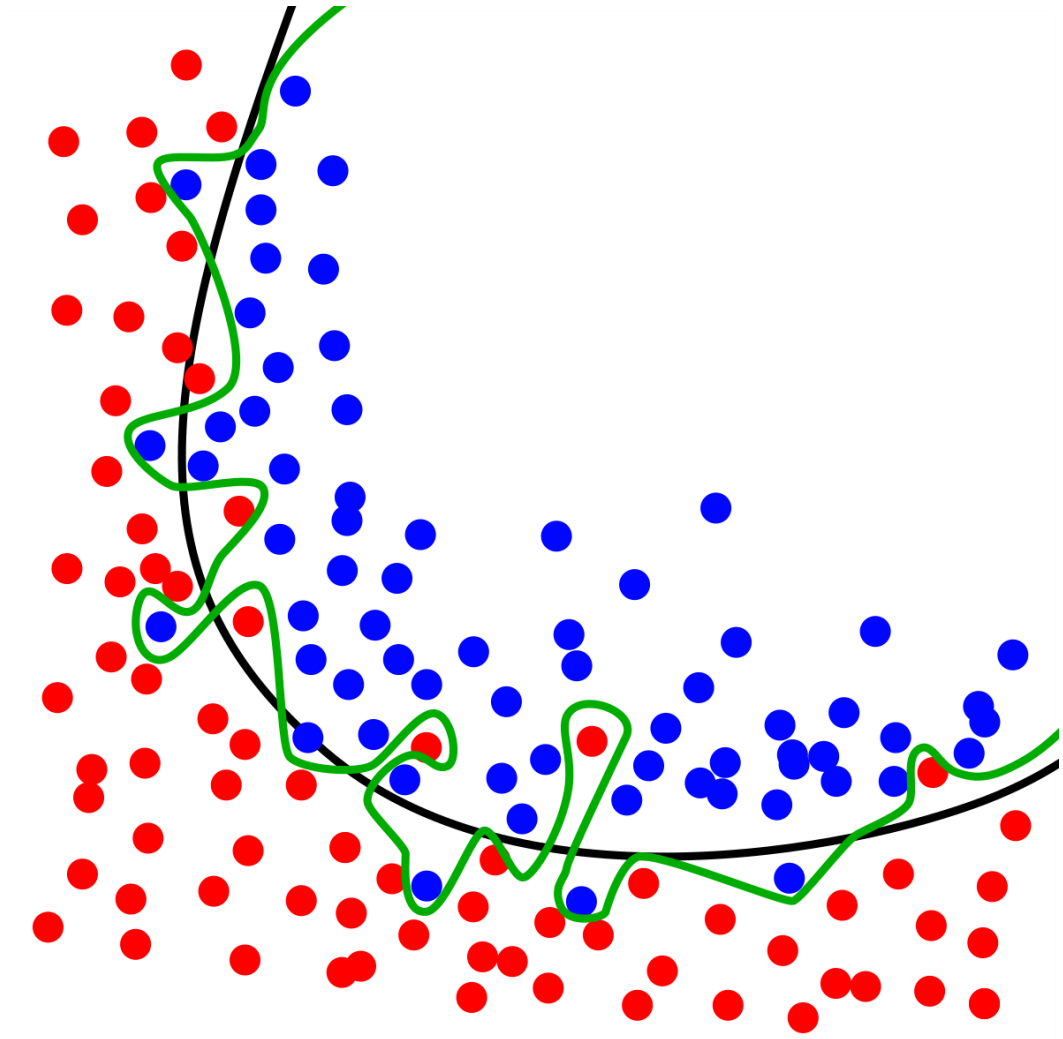
3.2.Μέθοδος Υπερδειγματοληψίας

Σε συνέχεια των προηγουμένων θα παρουσιάσουμε τη μέθοδο της υπερδειγματοληψίας (oversampling). Εν αντιθέσει με την προηγούμενη τεχνική, εδώ δεν αφαιρούμε αλλά προσθέτουμε παραδείγματα. Πιο συγκεκριμένα, ορισμένα παραδείγματα που ανήκουν στην κλάση με το μικρότερο πλήθος (μειοψηφίας) αντιγράφονται και πάλι με τυχαίο τρόπο και προσθέτονται στο σύνολο των δεδομένων εκπαίδευσης. Τα παραδείγματα του συνόλου δεδομένων εκπαίδευσης επιλέγονται τυχαία

με αντικατάσταση. Αυτό σημαίνει ότι παραδείγματα από την κατηγορία μειοψηφίας μπορούν να επιλεγούν και να προστεθούν στο νέο «πιο ισορροπημένο» εκπαιδευτικό σύνολο δεδομένων πολλές φορές. Επιλέγονται δηλαδή από το αρχικό σύνολο δεδομένων εκπαίδευσης, προστίθενται στο νέο σύνολο δεδομένων εκπαίδευσης και στη συνέχεια επιστρέφονται ή «αντικαθίστανται» στο αρχικό σύνολο δεδομένων και όλη αυτή η διαδικασία τους επιτρέπει να επιλεγούν ξανά.

Για να δούμε όμως σε ποιες περιπτώσεις αυτή η μέθοδος μπορεί να χαρακτηριστεί ως αποδοτική. Αυτό μπορεί να συμβεί όταν χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης που η απόδοσή τους εξαρτάται από μια ανισοκατανεμημένη κατανομή δεδομένων και όπου πολλά διπλά παραδείγματα για μια δεδομένη τάξη μπορούν να επηρεάσουν την προσαρμογή του μοντέλου. Χαρακτηριστικές τέτοιες περιπτώσεις αποτελούν οι αλγόριθμοι που μαθαίνουν επαναληπτικά τους συντελεστές, όπως τα τεχνητά νευρωνικά δίκτυα που χρησιμοποιούν στοχαστική κλίση καθόδου. ΄Αλλη περίπτωση επίσης αποτελούν τα μοντέλα που αναζητούν καλούς διαχωρισμούς των δεδομένων, όπως μηχανές υποστήριξης διανυσμάτων και δέντρα αποφάσεων. Τα συγκεκριμένα μάλιστα μοντέλα επιδεικνύουν εξαιρετική απόδοση σε αρκετά προβλήματα. Αντίστοιχα μοντέλα χρησιμοποιήθηκαν και στο πλαίσιο της παρούσας διπλωματικής για να αναδειχθεί η χρησιμότητα της μεθόδου σε αυτά.

Σε ορισμένες περιπτώσεις, η αναζήτηση μιας ισορροπημένης κατανομής για ένα σύνολο δεδομένων με σοβαρή ανισορροπία μπορεί να προκαλέσει την υπερβολική προσαρμογή των επηρεαζόμενων αλγορίθμων στην κατηγορία μειοψηφίας, οδηγώντας σε αυξημένο σφάλμα μη γενίκευσης (overfitting). Αυτό οφείλεται στο γεγονός ότι υπάρχει η πιθανότητα το δίκτυο να αρχίσει απλά να απομνημονεύει τα δείγματα που διαβάζει, τα όποια επαναλαμβάνονται πολλαπλές φορές, αντί να μάθει πως να τα διαχωρίζει. Ουσιαστικά λοιπόν θα προκύψει καλύτερο αποτέλεσμα όσον αφορά συνολικά στην εκπαίδευση του συστήματος, ωστόσο στο συγκεκριμένο μετασχηματισμένο σετ δεδομένων το αποτέλεσμα δεν θα είναι το αναμενόμενο. Για να γίνει καλύτερα κατανοητό αυτό που περιγράφουμε παρουσιάζεται το παρακάτω σχήμα:



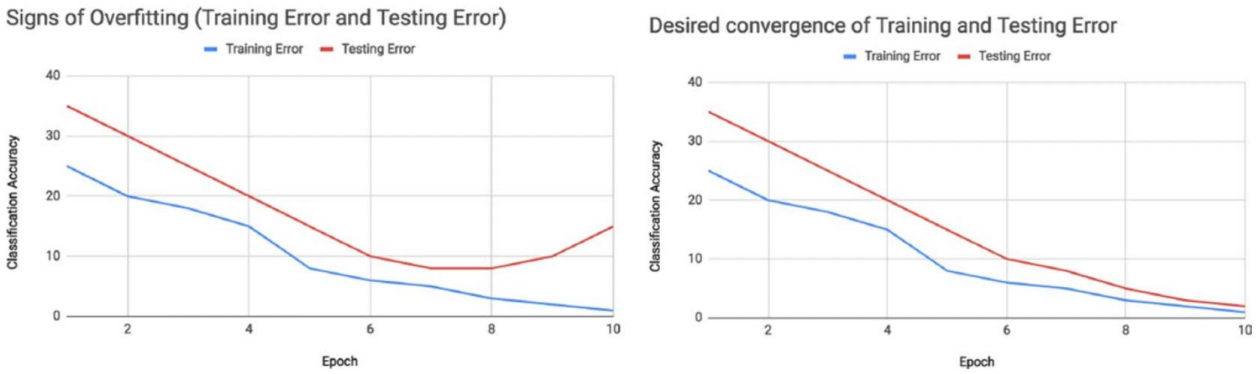
Εικόνα 6 Ένα παράδειγμα όπου φαίνονται δύο διαχωριστικές γραμμές διαφορετικής πολυπλοκότητας. Από την κατανομή των δεδομένων και το είδος της διαχωριστικής γραμμής συμπεραίνουμε ότι η πράσινη γραμμή είναι αποτέλεσμα υπερεκπαίδευσης.

Στο παραπάνω σχήμα απεικονίζονται δεδομένα, τα οποία διαχωρίζονται σε δύο κλάσεις, αυτή με τα μπλε και αντίστοιχα αυτή με τα κόκκινα. Παρατηρώντας το σύνολο των δεδομένων βλέπουμε ότι η πράσινη γραμμή αποτελεί τη βέλτιστη λύση, καθώς χωρίζει τα δείγματα με τέτοιο τρόπο ώστε να μην υπάρχουν δεδομένα της μιας κλάσης που να περιλαμβάνονται στην άλλη. Η συγκεκριμένη λύση λοιπόν θα ήταν 100% επιτυχής. Από την άλλη η πράσινη γραμμή είναι μια αρκετά περίπλοκη γραμμή. Η μαύρη αντιθέτως μπορεί να ταξινομεί λανθασμένα κάποια από τα δείγματα, ωστόσο κάποια από αυτά μπορεί να είναι ακραίες περιπτώσεις και όχι αντιπροσωπευτικές του υπό εξέταση προβλήματος. Για να γίνει πιο κατανοητό το παραπάνω επιχείρημα, ας αναλογιστούμε το παράδειγμα με έναν πιλότο ενός αεροσκάφους και ο οποίος αντιμετωπίζει το δίλημμα αν θα πετάξει ή όχι, με δεδομένο ότι οι καιρικές συνθήκες είναι πάρα πολύ κακές (πολύ δυνατός άνεμος, καταιγίδες κλπ). Θα μπορούσε να υπάρχει μια περίπτωση που πέταξε παρά τις δυσμενείς καιρικές συνθήκες, επειδή ήταν ας πούμε μια έκτακτη ανάγκη (πχ αεροδιακομιδή ασθενούς με έμφραγμα από απομακρυσμένο νησί). Δεν θα θέλαμε ωστόσο αυτό να αποτελεί γενικό κανόνα, ούτε αυτή η περίπτωση θα ήταν

αντιπροσωπευτική του προβλήματος. Αντίστοιχα λοιπόν ένα νευρωνικό δίκτυο θα μπορούσε να «διαβάσει» πολλές φορές ένα δείγμα ακραίο και να επηρεαστεί η διαχωριστική γραμμή που αναφέραμε παραπάνω. Το σύστημα θα οδηγηθεί σε overfitting.

Το θετικό είναι πως το φαινόμενο του overfitting μπορεί εύκολα να εντοπιστεί, όταν για παράδειγμα το σφάλμα της εκπαίδευσης είναι σημαντικά μικρότερο από αυτό του testing.

Για να έχουμε μια καλύτερη εικόνα της επιρροής αυτής της μεθόδου στην εκπαίδευση του αλγορίθμου, είναι σημαντικό να συγκρίνουμε την απόδοση του μοντέλου τόσο στα σύνολα δεδομένων εκπαίδευσης όσο και στα δοκιμαστικά σύνολα δεδομένων μετά από υπερδειγματοληψία, και εν συνεχεία να συγκρίνουμε τα αποτελέσματα με τον ίδιο αλγόριθμο στο αρχικό σύνολο δεδομένων. Στο παρακάτω διάγραμμα απεικονίζεται ένα παράδειγμα [18] για το πως μπορούμε να διακρίνουμε το overfitting από τις καμπύλες εκμάθησης και ελέγχου.



Εικόνα 7 Αριστερά: το σφάλμα επικύρωσης αυξάνεται καθώς ο ρυθμός εκπαίδευσης συνεχίζει να μειώνεται. Το μοντέλο προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης και έχει κακή απόδοση. Δεξιά: σύγκλιση μεταξύ του σφάλματος εκπαίδευσης και δοκιμής

Η γραφική παράσταση στα αριστερά δείχνει ένα σημείο καμπής όπου το σφάλμα του test set αρχίζει να αυξάνεται καθώς ο ρυθμός εκπαίδευσης συνεχίζει να μειώνεται. Πρακτικά, αυτό οφείλεται στο γεγονός ότι το μοντέλο ταξινόμησης δεν έχει μάθει ουσιαστικά να διαχωρίζει, απλά έχει «αποστηθίσει» τα δεδομένα και την κλάση που ανήκουν. Η αυξημένη προπόνηση έχει προκαλέσει το μοντέλο να προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης και να έχει κακή απόδοση στο σετ δοκιμών σε σχέση με το σετ εκπαίδευσης. Αντίθετα, στα δεξιά παρατηρούμε ότι υπάρχει μια σύγκλιση μεταξύ του σφάλματος εκπαίδευσης και δοκιμής.

Η αύξηση του αριθμού των παραδειγμάτων για την κατηγορία μειοψηφίας, ειδικά εάν η λιγότερα συχνή κλάση ήταν σοβαρή, μπορεί επίσης να οδηγήσει σε αξιοσημείωτη αύξηση του υπολογιστικού κόστους κατά την προσαρμογή του μοντέλου, ειδικά αν σκεφτεί κανείς ότι το μοντέλο βλέπει ξανά τα ίδια παραδείγματα στο σύνολο δεδομένων εκπαίδευσης ξανά και ξανά.

Παρακάτω παρουσιάζονται σε ένα σχήμα οι μέθοδοι υπερ και υπο δειγματοληψίας, όπου απεικονίζονται και οι βασικές διαφορές ανάμεσά τους:

Εικόνα 8 Μια σχηματική αναπαράσταση των μεθόδων της υπερ και υποδειγματοληψίας όπου φαίνεται ξεκάθαρα ο διαφορετικός τρόπος με τον όποιο επιδρούν στα δεδομένα.

Στην προαναφερθείσα περίπτωση με τα 1000 δείγματα, το νέο σύνολο δεδομένων μετά την εφαρμογή της υπερδειγματοληψίας θα αριθμεί συνολικά 1800 δείγματα, με τα χίλια εξ αυτών να αποτελούν και τα επιπρόσθετα οκτακόσια να είναι ουσιαστικά 8 φορές επανάληψη των δειγμάτων που ανήκουν στην κλάση μειοψηφίας. Κατά συνέπεια, από τα 1800 δείγματα, τα 900 θα αποτελούν μοναδικά δείγματα της “πλειοψηφικής” κλάσης, ενώ στα υπόλοιπα 900 θα περιέχονται τα αρχικά 100 της κλάσης μειοψηφίας επί εννέα φορές. Εν κατακλείδι, το σύνολο που θα προκύψει θα έχει ίδιο αριθμό δειγμάτων και στις δύο κλάσεις, χωρίς να έχει χαθεί κάποια πληροφορία. Ωστόσο, κάποια δείγματα θα περιλαμβάνονται αρκετές φορές.

3.3.Μέθοδος Over & Undersampling

Είδαμε προηγουμένως ότι υπάρχουν δύο μέθοδοι αντιμετώπισης της μεροληψίας που υπάρχει ανάμεσα στις δύο κλάσεις ενός συνόλου δεδομένων, αυτές της υποδειγματοληψίας και υπερδειγματοληψίας. Η κάθε μια δίνει μια λύση στο πρόβλημα, δημιουργώντας ωστόσο η κάθε μια και έναν νέο προβληματισμό. Η μεν πρώτη λόγω της πιθανότητας να χαθεί μέρος σημαντικής πληροφορίας, η δε δεύτερη λόγω των πολλών επαναλήψεων των ίδιων δεδομένων που οδηγεί στο φαινόμενο του overfitting.

Αν όμως συνδυάσουμε τις δύο αυτές τεχνικές προκύπτει μια άλλη μέθοδος over/under sampling, όπου εφαρμόζοντας και τις δύο προαναφερθείσες τεχνικές μπορούμε να βελτιώσουμε την ανισοκατανομή του συνόλου δεδομένων. Στην συγκεκριμένη τεχνική εφαρμόζονται ταυτόχρονα μια

υπερδειγματοληψία στην κλάση μειοψηφίας και μια υποδειγματοληψία στην κλάση με το μεγαλύτερο πλήθος δεδομένων.

Ο συνδυασμός αυτών των δύο μεθόδων ενδέχεται να έχει ως αποτέλεσμα πολύ καλύτερη συνολική απόδοση του συστήματος, σε σύγκριση πάντα με την απόδοση που θα είχε αν εφαρμοζόταν μεμονωμένα είτε η μια είτε η άλλη μέθοδος.

Στην πράξη, εάν είχαμε ένα σύνολο δεδομένων με κατανομή κλάσης 1:100, θα μπορούσαμε εφαρμόζοντας την υπερδειγματοληψία να αυξήσουμε την αναλογία σε 1:10 (αντιγράφοντας παραδείγματα από την κατηγορία μειοψηφίας) και, στη συνέχεια, να εφαρμόσουμε υποδειγματοληψία για να βελτιώσουμε περαιτέρω την αναλογία σε 1:2 με διαγραφή ορισμένων δεδομένων κλάση πλειοψηφίας. Αυτό θα μπορούσε να εφαρμοστεί με μη ισορροπημένη μάθηση, χρησιμοποιώντας ένα RandomOverSampler με στρατηγική δειγματοληψίας ρυθμισμένο στο 0,1 (10%) και, στη συνέχεια, χρησιμοποιώντας ένα RandomUnderSampler με στρατηγική δειγματοληψίας στο 0,5 (50%).

3.4. Μέθοδος Smote

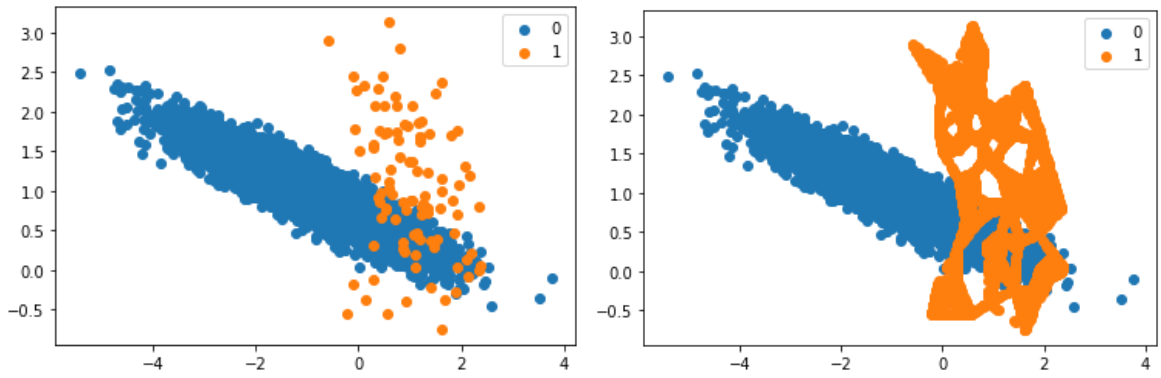
Στις προηγούμενες παραγράφους παρουσιάσαμε, για την αντιμετώπιση των συνεπειών που επιφέρει η εκπαίδευση συστημάτων με την χρήση μη ισορροπημένων συνόλων δεδομένων, τις τεχνικές της υπερδειγματοληψίας της μειοψηφικής κλάσης και της υποδειγματοληψίας της πλειοψηφικής κλάσης. Η απλούστερη προσέγγιση περιλαμβάνει την αντιγραφή παραδειγμάτων στην τάξη μειοψηφίας, αν και αυτά τα παραδείγματα, όπως έχει αναφερθεί, δεν προσθέτουν νέες πληροφορίες στο μοντέλο. Αντί αυτής της προσέγγισης θα μελετήσουμε αν μπορούμε να συνθέσουμε δείγματα με βάση τα ήδη υφιστάμενα, αυξάνοντας με αυτόν τον τρόπο και πάλι τα δεδομένα μας στην κλάση μειοψηφίας. Αυτή η τεχνική αποκαλείται Τεχνική Υπερδειγματοληψίας Συνθετικής Μειονότητας (Synthetic Minority Oversampling Technique ή SMOTE).

Ένα ζήτημα με την ανισοκατανεμημένη ταξινόμηση είναι ότι υπάρχουν πολύ λίγα παραδείγματα της τάξης μειοψηφίας ώστε ένα μοντέλο να εκπαιδευτεί σωστά. Ειδικά στο συγκεκριμένο σύνολο δεδομένων που εξετάζουμε στο πλαίσιο της παρούσας εργασίας το πρόβλημα είναι ιδιαίτερα έντονο, με τα δείγματα της κλάσης μειοψηφίας να αποτελούν ουσιαστικά μόνο το 0.172% του συνολικού dataset. Η λειτουργία της τεχνικής SMOTE είναι η ακόλουθη: επιλέγει παραδείγματα που βρίσκονται κοντά στον χώρο χαρακτηριστικών, τα ενώνει μεταξύ τους με μια κοινή γραμμή και σχεδιάζει ένα νέο δείγμα σε ένα σημείο κατά μήκος αυτής της γραμμής.

Ειδικότερα, αρχικά επιλέγεται ένα τυχαίο παράδειγμα από την κλάση μειοψηφίας. Στη συνέχεια βρίσκεται το k από τους πλησιέστερους γείτονες για αυτό το παράδειγμα (συνήθως k=5). Επιλέγεται ένας τυχαία επιλεγμένος γείτονας και δημιουργείται ένα συνθετικό παράδειγμα σε ένα τυχαία επιλεγμένο σημείο μεταξύ των δύο παραδειγμάτων στο χώρο χαρακτηριστικών. Αυτή η διαδικασία μπορεί να χρησιμοποιηθεί για τη δημιουργία όσων συνθετικών παραδειγμάτων χρειάζονται για την τάξη μειοψηφίας. Όπως αναφέρεται και στη βιβλιογραφία [19], συνίσταται αρχικά η εφαρμογή της μεθόδου της τυχαίας υποδειγματοληψίας, ώστε να μειωθεί το πλήθος παραδειγμάτων στην κλάση πλειοψηφίας και, στη συνέχεια, η εφαρμογή της τεχνικής SMOTE για την υπερδειγματοληψία της μειοψηφίας ώστε να εξισορροπηθεί εν τέλει η κατανομή των κλάσεων.

Η προσέγγιση αυτή είναι αποτελεσματική επειδή δημιουργούνται νέα συνθετικά παραδείγματα από την τάξη μειοψηφίας που είναι λογικά, δηλαδή είναι σχετικά κοντά σε χώρο χαρακτηριστικών με υπάρχοντα παραδείγματα από την τάξη μειοψηφίας. Ένα γενικό μειονέκτημα της συγκεκριμένης μεθόδου είναι ότι αυτή η σύνθεση νέων δεδομένων γίνεται χωρίς ουσιαστικά να λαμβάνεται υπόψη η πλειοψηφική τάξη, με αποτέλεσμα πιθανώς διφορούμενα παραδείγματα εάν υπάρχει ισχυρή επικάλυψη για τις κλάσεις.

Προκειμένου να γίνει πιο εύκολα κατανοητή η λειτουργία της συγκεκριμένης μεθόδου ακολουθεί ένα παράδειγμα επαύξησης των δεδομένων για ένα πρόβλημα δύο διαστάσεων. Αριστερά, στο πρώτο σχήμα βλέπουμε το αρχικό σύνολο δεδομένων, ενώ στο δεξιό σχήμα το επαυξημένο σύνολο κατόπιν της σύνθεσης δεδομένων και εφαρμογής της υπό εξέταση μεθόδου.



Εικόνα 9 Ο τρόπος με τον οποίο η μέθοδος Smote συνθέτει δεδομένα. Παρουσιάζεται το σύνολο δεδομένων πριν και μετά την εφαρμογή της μεθόδου

Όπως παρατηρούμε στο δεύτερο σχήμα, η τεχνική αυτή έχει προσθέσει αρκετά σημεία που βρίσκονται ανάμεσα σε πραγματικά κοντινά σημεία του αρχικού συνόλου δεδομένων. Ωστόσο, αυτό που επίσης είναι αντιληπτό είναι πως έχουν εμφανιστεί σημεία της κλάσης 1 πάνω σε σημεία της κλάσης 0, οπότε έχουμε την περίπτωση της αλληλοεπικάλυψης που αναφέραμε και παραπάνω. Το συγκεκριμένο φαινόμενο μπορεί να χειροτερεύσει την απόδοση και να μην έχουμε τα επιθυμητά αποτελέσματα, μιας και έτσι ενισχύεται ο θόρυβος που μπορεί να προέρχεται από outliers.

3.5. Μέθοδος κανονικοποίησης (Regularization)

Στην εποπτευόμενη μηχανική εκμάθηση (supervised learning), τα μοντέλα εκπαιδεύονται σε ένα υποσύνολο δεδομένων, γνωστό και ως δεδομένα εκπαίδευσης. Ο σκοπός είναι να υπολογιστεί η κλάση κάθε σημείου από τα δεδομένα εκπαίδευσης. Εντούτοις, η υπερπροσαρμογή (overfitting) συμβαίνει, όπως αναφέρεται και παραπάνω, όταν το μοντέλο αποστηθίζει τα δεδομένα εκπαίδευσης αντί να γενικεύει και να μαθαίνει τη γενικευμένη διαχωριστική επιφάνεια που περιγράφουν τα δεδομένα. Αυτό μπορεί να οφείλεται σε διάφορους λόγους, με έναν πιθανό να είναι το μικρό πλήθος δεδομένων εκπαίδευσης (σε μια ή σε περισσότερες κλάσεις), κάτι το οποίο ισχύει και στο dataset που εξετάζουμε σε αυτήν την εργασία. ό

Παρόλα αυτά, υπάρχει τρόπος ώστε το μοντέλο να μην υπερεκπαιδευτεί, με διασημότερο εξ αυτών να αποτελεί η μέθοδος της κανονικοποίησης (Regularization). Η κανονικοποίηση προσθέτει ουσιαστικά μια ποινή καθώς αυξάνεται η πολυπλοκότητα του μοντέλου. Οι L1 και L2 είναι οι δύο πιο γνωστές συναρτήσεις που χρησιμοποιούνται στη μηχανική μάθηση και χρησιμοποιούνται όπως είναι λογικό κατά τη διαδικασία εκπαίδευσης. Παρακάτω παρουσιάζουμε τους μαθηματικούς τύπους αυτών:

- L1 (Least Absolute Deviations ή LAD): χρησιμοποιείται για την ελαχιστοποίηση του σφάλματος που είναι το άθροισμα όλων των απόλυτων διαφορών μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής.

$$L1LossFunction = \sum_{i=1}^{n} \left| y_{true} - y_{predicted} \right|$$

- L2 (μέσο τετραγωνικό σφάλμα ή Least Square Error ή LSE): χρησιμοποιείται για την ελαχιστοποίηση του σφάλματος που είναι το άθροισμα όλων των τετραγωνικών διαφορών μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής.

$$L2LossFunction = \sum_{i=1}^{n} (y_{true} - y_{predicted})^2$$

Γενικά, η συνάρτηση L2 προτιμάται στην πλειοψηφία των περιπτώσεων. Ωστόσο, όταν ανάμεσα στο πλήθος των δεδομένων υπάρχουν και ακραία στοιχεία, υπάρχει η πιθανότητα η συνάρτηση απώλειας L2 να μην είναι το ίδιο αποδοτική. Ο λόγος πίσω από αυτήν την κακή απόδοση βρίσκεται αν δούμε την μαθηματική έκφραση αυτής. Καθώς υπολογίζονται οι τετραγωνικές διαφορές των σφαλμάτων, εάν υπάρχουν πολλές ακραίες τιμές στο σύνολο δεδομένων η L2 ενισχύει την επίδραση αυτών, αντί να την μειώνει. Εξαιτίας αυτού, επιλέξαμε να χρησιμοποιήσουμε ως μέθοδο κανονικοποίησης τη συνάρτηση L1, εφόσον το σύνολο μας περιλαμβάνει ακραίες τιμές.

4. Μοντέλα Μηχανικής Μάθησης

Τα τελευταία χρόνια, η Τεχνητή Νοημοσύνη γνωρίζει ιλιγγιώδη ανάπτυξη σε ποικίλους τομείς, καθιστώντας τις μηχανές ικανές να μαθαίνουν από την εμπειρία. Η Μηχανική Μάθηση είναι ο τομέας που διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Ήδη από το 1959, ο Άρθουρ Σάμουελ την είχε ορίσει ως "*Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί*". Αρκετά χρόνια αργότερα, το 1997, ο Tom M. Mitchell πρότεινε έναν άλλο ορισμό σύμφωνα με τον οποίον: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία Ε ως προς μια κλάση εργασιών Τ και ένα μέτρο επίδοσης Ρ, αν η επίδοσή του σε εργασίες της κλάσης Τ, όπως αποτιμάται από το μέτρο Ρ, βελτιώνεται με την εμπειρία Ε»

Οι διάφοροι τύποι μηχανικής μάθησης που έχουν αναπτυχθεί χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και διαφέρουν ως προς τον τύπο δεδομένων που είναι διαθέσιμα για εκπαίδευση, τη σειρά και τη μέθοδο με την οποία λαμβάνονται τα δεδομένα εκπαίδευσης και τα δεδομένα ελέγχου που χρησιμοποιούνται για την αξιολόγηση του αλγορίθμου μάθησης. Όπως αντίστοιχα συμβαίνει και με τους τρόπους που μαθαίνει ένα ανθρώπινο ον, έτσι και στον τομέα της μηχανικής μάθησης αναπτύσσονται τρεις τρόποι μάθησης, οι οποίοι είναι:

- Η επιβλεπόμενη μάθηση (supervised learning)
- Η μη επιβλεπόμενη μάθηση (unsupervised learning)
- Η ενισχυτική μάθηση (reinforcement learning)

Ως επιβλεπόμενη μάθηση (supervised learning) ορίζεται η διαδικασία που ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα: ταξινόμησης (classification), πρόγνωσης (prediction) και διερμηνείας (interpretation). Στο πλαίσιο της παρούσας εργασίας θα εστιάσουμε περισσότερο σε αυτόν τον τρόπο, παρουσιάζοντας εκτενέστερα συχνά χρησιμοποιούμενους αλγορίθμους.

Στη μη επιβλεπόμενη μάθηση (unsupervised learning) ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Ο εν λόγω τρόπος μάθησης χρησιμοποιείται σε περιπτώσεις ανάλυσης συσχετισμών (association analysis) και ομαδοποίησης (clustering).

Τέλος, στην περίπτωση της ενισχυτικής μάθησης (reinforcement learning) ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον.

Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

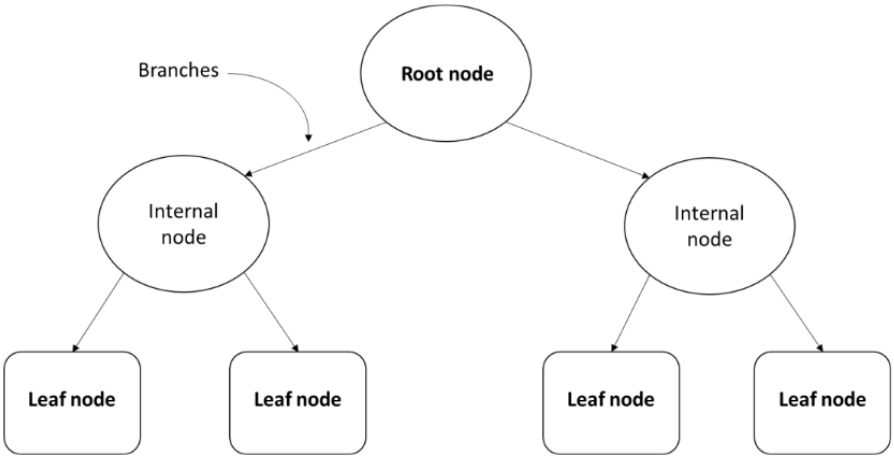
4.1. Δέντρο απόφασης (Decision Tree)

Το Decision Tree Learning ανήκει στην κατηγορία της επιβλεπόμενης μάθησης (supervised learning). Πρόκειται για μια εποπτευόμενη μαθησιακή προσέγγιση που χρησιμοποιείται στη στατιστική, την εξόρυξη δεδομένων και τη μηχανική μάθηση. Πιο συγκεκριμένα, ένα δέντρο απόφασης ταξινόμησης ή παλινδρόμησης χρησιμοποιείται ως μοντέλο πρόβλεψης για την εξαγωγή συμπερασμάτων σχετικά με ένα σύνολο παρατηρήσεων.

Ας δούμε ποια είναι η διαφορά ανάμεσα στα δέντρα ταξινόμησης και παλινδρόμησης. Ως δέντρα ταξινόμησης αποκαλούνται αυτά στα οποία η μεταβλητή στόχος μπορεί να λάβει ένα διακριτό σύνολο τιμών. Αντίθετα, όταν η μεταβλητή στόχος μπορεί να λάβει συνεχείς (πχ πραγματικούς) αριθμούς, τότε μιλάμε για δέντρα απόφασης παλινδρόμησης. Σε αυτές τις δομές δέντρων, τα φύλλα αντιπροσωπεύουν ετικέτες κλάσεων και τα κλαδιά αντιπροσωπεύουν συνδέσμους χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες κλάσεων.

Τα δέντρα αποφάσεων είναι από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης, λόγω της ευκρίνειας και της απλότητάς τους [8]. Όσον αφορά στην ευκρίνεια, αυτή η κατηγορία αλγορίθμων μας επιτρέπει να δούμε τον τρόπο λειτουργίας τους, βοηθώντας να κατανοήσουμε πως προκύπτει το αποτέλεσμα. Πρόκειται για καθαρή, «διάφανη» μέθοδο αυτό που εννοούμε μιλώντας για ευκρίνεια. Εν αντιθέσει με ένα βαθύ νευρωνικό δίκτυο, το οποίο δεν είναι τόσο ευκρινές μιας και δεν μας αφήνει να δούμε τον τρόπο λειτουργίας του. Θα μιλήσουμε ωστόσο εκτενέστερα για αυτά σε επόμενη παράγραφο. Αυτή η ευκρίνεια που περιγράφουμε καθιστά τους συγκεκριμένους αλγορίθμους δημοφιλείς και συνιστά τη χρήση τους σε εφαρμογές για την ανάλυση αποφάσεων καθώς και στην εξόρυξη γνώσης. Στην ανάλυση αποφάσεων, ένα decision tree μπορεί να χρησιμοποιηθεί για να αναπαραστήσει οπτικά τις επιλογές και τον τρόπο λήψης αποφάσεων. Όσον αφορά στη δεύτερη διαδεδομένη περίπτωση χρήσης τους, ένα δέντρο αποφάσεων περιγράφει δεδομένα (αλλά το δέντρο ταξινόμησης που προκύπτει μπορεί να είναι μια είσοδος για την μετέπειτα ανάλυση του από έναν αντίστοιχο αλγόριθμο λήψης αποφάσεων). Η εκμάθηση του δέντρου αποφάσεων είναι μια μέθοδος που χρησιμοποιείται συνήθως στην εξόρυξη δεδομένων [9]. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που να προβλέπει την τιμή μιας μεταβλητής στόχου με βάση πολλές μεταβλητές εισόδου.

Ένα δέντρο αποφάσεων μπορεί να ταξινομεί δεδομένα με πολύ απλό τρόπο. Για τους σκοπούς της παρούσης θεματολογίας, έστω ότι έχουμε πεπερασμένα διακριτά χαρακτηριστικά εισόδου και υπάρχει ένα μοναδικό χαρακτηριστικό στόχου που ονομάζεται "κλάση της ταξινόμησης". Ένα δέντρο αποφάσεων ταξινόμησης ή πιο απλά δέντρο ταξινόμησης αποτελείται από έναν ριζικό κόμβο (root node), τα κλαδιά (branches), τους εσωτερικούς κόμβους, δηλαδή αυτούς που βρίσκονται μεταξύ του ριζικού και των φύλλων, και τους τερματικούς κόμβους ή φύλλα (terminal nodes ή leaves).

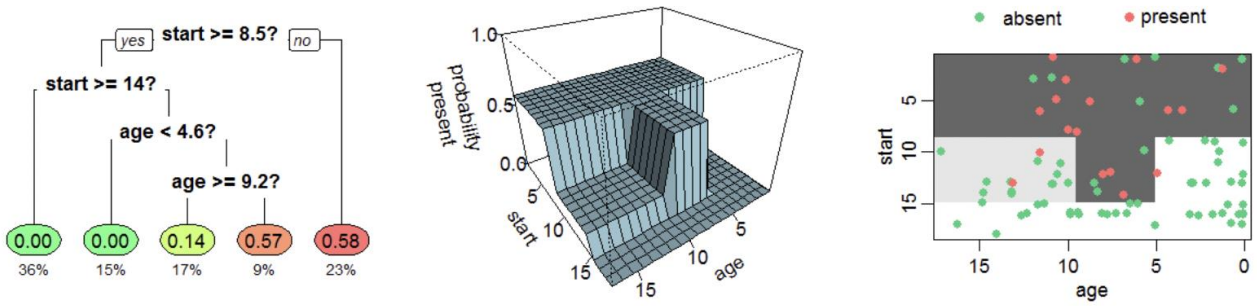


Εικόνα 10 Απεικόνιση ενός δέντρου απόφασης

Σε ένα δέντρο ταξινόμησης κάθε εσωτερικός (μη φύλλο) κόμβος επισημαίνεται με ένα χαρακτηριστικό εισόδου. Τα τόξα που προέρχονται από έναν κόμβο που έχει επισημανθεί με ένα χαρακτηριστικό εισόδου επισημαίνονται με καθεμία από τις πιθανές τιμές του χαρακτηριστικού στόχου ή το τόξο οδηγεί σε έναν δευτερεύοντα κόμβο απόφασης σε ένα διαφορετικό χαρακτηριστικό εισόδου. Κάθε φύλλο του δέντρου επισημαίνεται με μια κλάση ή μια κατανομή πιθανοτήτων στις κλάσεις. Αυτό σημαίνει ότι το σύνολο δεδομένων έχει ταξινομηθεί από το δέντρο είτε σε μια συγκεκριμένη κατηγορία είτε σε ένα υποσύνολο κλάσης μέσω μιας κατανομής πιθανοτήτων.

Ακολούθως, ένα δέντρο κατασκευάζεται ως εξής; Ο ριζικός (αρχικός) κόμβος root node αποτελεί το σύνολο των χαρακτηριστικών εισόδου και διαχωρίζεται σε υποσύνολα (τα οποία αποτελούν τα διαδοχικά παιδιά). Ο διαχωρισμός γίνεται βάσει ενός συνόλου κανόνων διαχωρισμού που βασίζεται σε χαρακτηριστικά ταξινόμησης. Αυτή η διαδικασία είναι γνωστή ως αναδρομική κατάτμηση, εκτελείται δε αναδρομικά σε κάθε παραγόμενο υποσύνολο. Η αναδρομή ολοκληρώνεται όταν το υποσύνολο σε έναν κόμβο έχει όλες τις ίδιες τιμές της μεταβλητής στόχου ή όταν ο διαχωρισμός δεν προσθέτει πλέον αξία στις προβλέψεις. Αυτή η διαδικασία επαγωγής δέντρων αποφάσεων από πάνω προς τα κάτω (TDIDT) [10] είναι ένα παράδειγμα ενός άπληστου αλγορίθμου και είναι μακράν η πιο κοινή στρατηγική για την εκμάθηση δέντρων αποφάσεων από δεδομένα [9, 10].

Στην εξόρυξη δεδομένων, τα δέντρα αποφάσεων μπορούν επίσης να περιγραφούν ως ο συνδυασμός μαθηματικών και υπολογιστικών τεχνικών που βοηθούν στην περιγραφή, την κατηγοριοποίηση και τη γενίκευση ενός δεδομένου συνόλου δεδομένων.



Εικόνα 11 Οι διαφορετικοί τύποι με τον όποιον μπορεί ένας αλγόριθμος τύπου decision tree να ταξινομήσει τα δεδομένα εισόδου του.

Τα δέντρα αποφάσεων που χρησιμοποιούνται για την εξόρυξη δεδομένων είναι:

- Δέντρο ταξινόμησης (Classification Tree): Όταν το προβλεπόμενο αποτέλεσμα είναι η κλάση (διακεκριμένη) στην οποία ανήκουν τα δεδομένα.
- Δέντρο παλινδρόμησης (Regression Tree): Όταν το αποτέλεσμα – στόχος είναι πραγματικός αριθμός (π.χ. η τιμή ενός σπιτιού ή η διάρκεια παραμονής ενός ασθενούς σε ένα νοσοκομείο).

Στο πλαίσιο της παρούσας εργασίας θα ασχοληθούμε με το δέντρο ταξινόμησης, καθώς μας ενδιαφέρει ο διαχωρισμός των δειγμάτων σε 2 κλάσεις (απάτη και έγκυρη συναλλαγή) οι οποίες είναι διακριτές και όχι συνεχείς.

Ο όρος ανάλυση δέντρου ταξινόμησης και παλινδρόμησης (Classification and regression trees ή CART) [11] είναι ένας γενικός όρος που χρησιμοποιείται για να αναφέρεται σε οποιαδήποτε από τις παραπάνω διαδικασίες. Τα δέντρα που χρησιμοποιούνται για παλινδρόμηση και τα δέντρα που χρησιμοποιούνται για ταξινόμηση έχουν κάποιες ομοιότητες – αλλά και κάποιες διαφορές, όπως η διαδικασία που χρησιμοποιείται για τον προσδιορισμό του σημείου διαχωρισμού.

Ορισμένες τεχνικές, που συχνά ονομάζονται μέθοδοι συνόλου, κατασκευάζουν περισσότερα από ένα δέντρα αποφάσεων:

- **Ενισχυμένα δέντρα (Boosted trees)**: Η μέθοδος αυτή δημιουργεί σταδιακά ένα σύνολο εκπαιδεύοντας κάθε νέο στιγμιότυπο ενισχύοντας τις περιπτώσεις εκπαίδευσης που είχαν προηγουμένως εσφαλμένα μοντελοποιηθεί. Χαρακτηριστικό παράδειγμα είναι το AdaBoost.

Αυτά μπορούν να χρησιμοποιηθούν για προβλήματα τύπου παλινδρόμησης και τύπου ταξινόμησης.[7][8]

- **Bootstrap aggregated (ή bagged) δέντρα αποφάσεων**: μια πρώιμη μέθοδος συνόλου, δημιουργεί πολλαπλά δέντρα απόφασης επαναλαμβάνοντας επαναληπτικά δεδομένα εκπαίδευσης με αντικατάσταση και ψηφίζοντας τα δέντρα για μια συναινετική πρόβλεψη. Ένας τυχαίος ταξινομητής δασών είναι ένας συγκεκριμένος τύπος συγκέντρωσης bootstrap.
- **Δάσος περιστροφής (Rotation forest)**: στο οποίο κάθε δέντρο απόφασης εκπαιδεύεται εφαρμόζοντας πρώτα την ανάλυση κύριου συστατικού (PCA) σε ένα τυχαίο υποσύνολο των χαρακτηριστικών εισόδου [8, 10].

Κάποιοι αλγόριθμοι δέντρων αποφάσεων είναι οι εξής:

- ID3 (Iterative Dichotomiser 3)
- C4.5 (διάδοχος του ID3)
- CART (Classification And Regression Tree) Δέντρο ταξινόμησης και παλινδρόμησης[6]
- Αυτόματη ανίχνευση αλληλεπίδρασης Chi-square (Chi-square automatic interaction detection - CHAID): Εκτελεί διαχωρισμούς πολλαπλών επιπέδων κατά τον υπολογισμό των δέντρων ταξινόμησης [12].
- MARS: επεκτείνει τα δέντρα αποφάσεων για να χειρίζονται καλύτερα τα αριθμητικά δεδομένα.
- Δέντρα συμπερασμάτων υπό όρους (Conditional Inference Trees). Προσέγγιση που βασίζεται σε στατιστικά στοιχεία που χρησιμοποιεί μη παραμετρικές δοκιμές ως διαχωριστικά κριτήρια, διορθωμένα για πολλαπλές δοκιμές για την αποφυγή υπερβολικής προσαρμογής. Αυτή η προσέγγιση έχει ως αποτέλεσμα την αμερόληπτη επιλογή προβλέψεων και δεν απαιτεί κλάδεμα.[13]

Το ID3 και το CART εφευρέθηκαν ανεξάρτητα την ίδια περίπου εποχή (μεταξύ 1970 και 1980), ωστόσο ακολουθήθηκε μια παρόμοια προσέγγιση για την εκμάθηση ενός δέντρου αποφάσεων από πλειάδες εκπαιδεύσεων .

Έχει επίσης προταθεί η αξιοποίηση των εννοιών της ασαφούς θεωρίας συνόλων για τον ορισμό μιας ειδικής έκδοσης του δέντρου αποφάσεων, γνωστής ως Δέντρο Ασαφών Αποφάσεων (FDT) [14]. Σε αυτόν τον τύπο ασαφούς ταξινόμησης, γενικά, ένα διάνυσμα εισόδου x συσχετίζεται με πολλές κλάσεις, καθεμία με διαφορετική τιμή εμπιστοσύνης. Τα ενισχυμένα σύνολα FDT έχουν επίσης

διερευνηθεί πρόσφατα, και έχουν δείξει επιδόσεις συγκρίσιμες με εκείνες άλλων πολύ αποτελεσματικών ασαφών ταξινομητών.

4.2.Multilayer Perceptron

Κάθε αρχιτεκτονική μηχανικής μάθησης αποτελείται από κομμάτια των οποίων η εσωτερική δομή όπως π.χ. οι παράμετροι του ή οι μεταβλητές ενός υπερεπιπέδου είναι εκπαιδεύσιμη, δηλαδή μπορεί να προσαρμόζεται στα δεδομένα. Η προσαρμογή αυτή στην ουσία αποτελεί την εκπαίδευση του αλγόριθμου και έχει σαν στόχο πρακτικά την ελαχιστοποίηση ενός κριτηρίου πχ της απόστασης από τα δεδομένα. Εντούτοις, οι διάφοροι αλγόριθμοι μηχανικής μάθησης είναι πιθανόν να διαφέρουν αρκετά στην εσωτερική τους δομή.

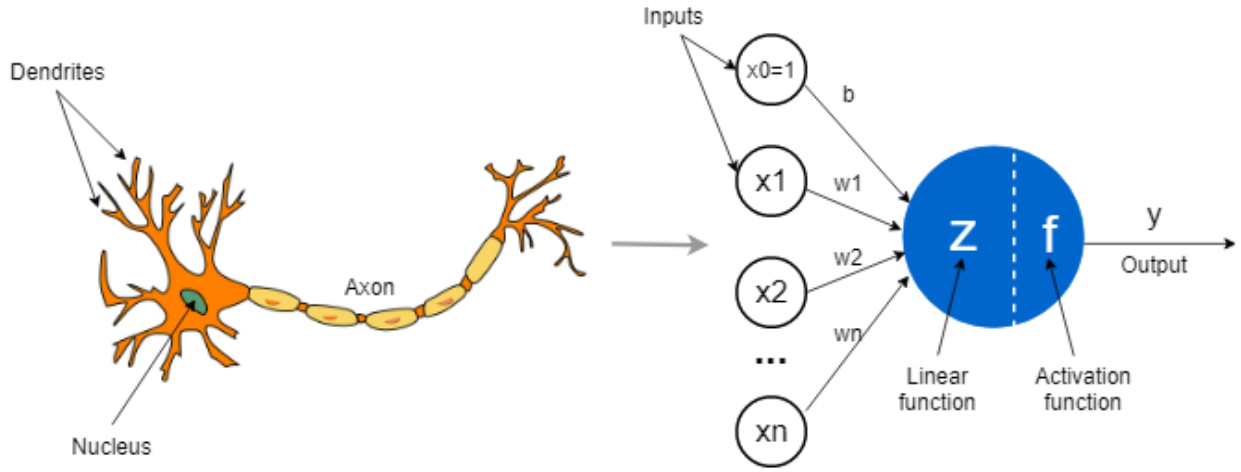
Εκτός από τα δέντρα απόφασης, αλγόριθμο μηχανικής μάθησης αποτελεί και ένα νευρωνικό δίκτυο. Ένα τέτοιο δίκτυο μπορεί να αποτελείται από πλήθος νευρώνων ικανών ο οποίος αποτελείται από νευρώνες οι οποίοι μπορούν να προσαρμόζονται – εκπαιδεύονται. Υπάρχουν διαφόρων τύπων νευρώνες οι οποίοι μπορούν να χρησιμοποιηθούν. Κάθε ένας από αυτούς μπορεί να επιφέρει καλύτερες επιδόσεις ανάλογα με το πρόβλημα. Για παράδειγμα, ένα μοντέλο τύπου LSTM (long short-term memory) [15] είναι πολύ αποδοτικό στην επίλυση προβλημάτων χρονοσειρών όπως ανάλυση κειμένων, μετοχών κ.α. Αντίθετα, ένα συνελικτικό δίκτυο παρουσιάζει καλύτερα αποτελέσματα σε ένα πρόβλημα ανάλυσης εικόνων. Στο πλαίσιο της παρούσας εργασίας θα ασχοληθούμε με τα πλήρως συνδεδεμένα δίκτυα, τα οποία είναι τα πρώτα που εισήχθησαν και τα πλέον διαδεδομένα, μιας και αποτελούν κομμάτι όλων των προαναφερθέντων αρχιτεκτονικών.

Οι τεχνητοί νευρώνες (ονομάζονται επίσης Perceptrons, Units ή Nodes) είναι τα πιο απλά στοιχεία ή δομικά στοιχεία σε ένα νευρωνικό δίκτυο. Είναι εμπνευσμένα από βιολογικούς νευρώνες που βρίσκονται στον ανθρώπινο εγκέφαλο.

Σε αυτό το άρθρο θα συζητήσουμε πώς τα perceptron εμπνέονται από βιολογικούς νευρώνες, θα σχεδιάσουμε τη δομή ενός perceptron, θα συζητήσουμε τις δύο μαθηματικές συναρτήσεις μέσα σε ένα perceptron και, τέλος, θα εκτελέσουμε μερικούς υπολογισμούς μέσα σε ένα perceptron.

Μπορούμε να θεωρήσουμε έναν τεχνητό νευρώνα ως ένα μαθηματικό μοντέλο εμπνευσμένο από έναν βιολογικό νευρώνα. Στο παρακάτω σχήμα φαίνεται η θεωρικά αποδεκτή εικόνα ενός

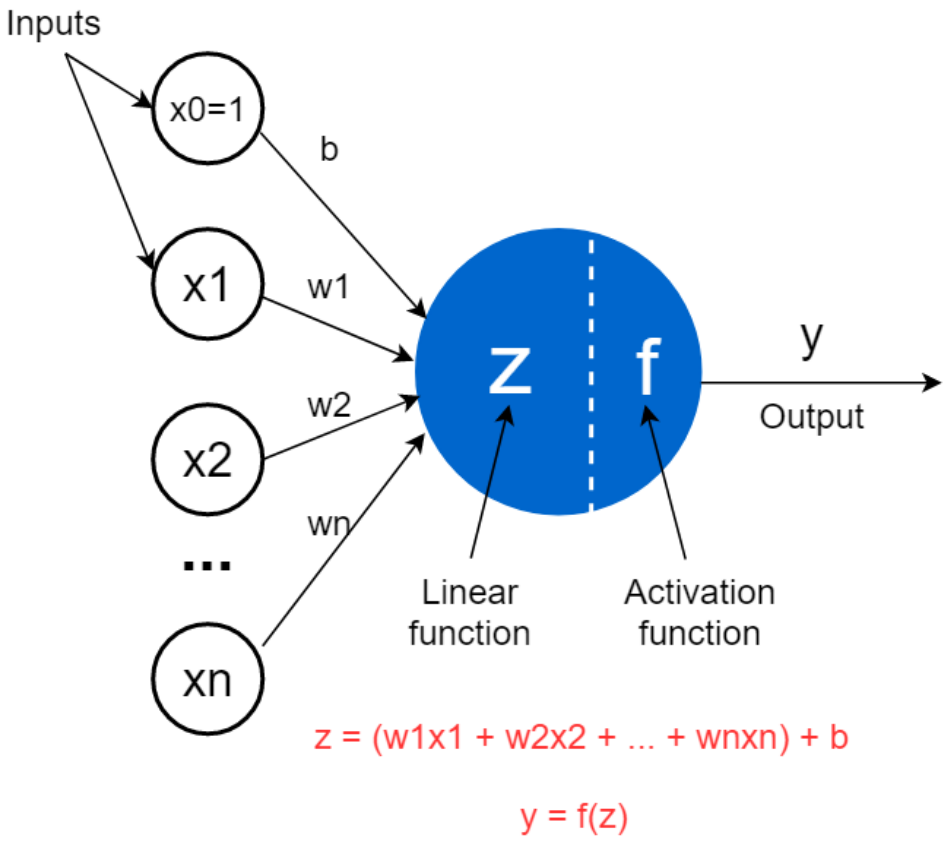
βιολογικού νευρώνα και δίπλα η μαθηματική μοντελοποίησή του που χρησιμοποιείται από τα νευρωνικά δίκτυα.



Εικόνα 12 Η γραφική αναπαράσταση ενός βιολογικού (αριστερά) και ενός τεχνητού νευρώνα (δεξιά). Από το σχήμα αυτό φαίνονται οι ομοιότητες μεταξύ τους και ο τρόπος με τον όποιος ο τεχνητός προσομοιάζει τον βιολογικό νευρώνα

Στην παραπάνω εικόνα παρατηρούμε την αντιστοιχία που υπάρχει ανάμεσα σε έναν βιολογικό και έναν τεχνητό νευρώνα. Στον βιολογικό (αριστερό μέρος της εικόνας) η λήψη των ερεθισμάτων - σημάτων εισόδου γίνεται μέσω των δενδριτών (μικρές ίνες). Αντίστοιχα, στον τεχνητό (δεξί μέρος της εικόνας) η λήψη γίνεται από άλλα perceptrons μέσω νευρώνων εισόδου με διάφορες τιμές. Οι συνδέσεις ανάμεσα στους δενδρίτες και τον βιολογικό νευρώνα ονομάζονται συνάψεις, στον δε τεχνητό οι συνδέσεις ουσιαστικά είναι τα διαφορετικά βάρη της κάθε εισόδου. Με την έννοια βάρη εννοούμε τη διαφορετική σπουδαιότητα της κάθε εισόδου. Και στις δύο περιπτώσεις παράγεται μια έξοδος. Η διαφορά είναι ότι στην περίπτωση του βιολογικού νευρώνα παράγεται μια έξοδος με βάση τα σήματα που λαμβάνουν οι δενδρίτες, ενώ στον τεχνητό το perceptron κάνει τους υπολογισμούς βάσει των τιμών εισόδου και παράγει μια έξοδο. Σε έναν βιολογικό νευρώνα, το σήμα εξόδου μεταφέρεται από τον άξονα. Ομοίως, ο άξονας σε ένα perceptron είναι η τιμή εξόδου που θα είναι η είσοδος για τα επόμενα perceptrons.

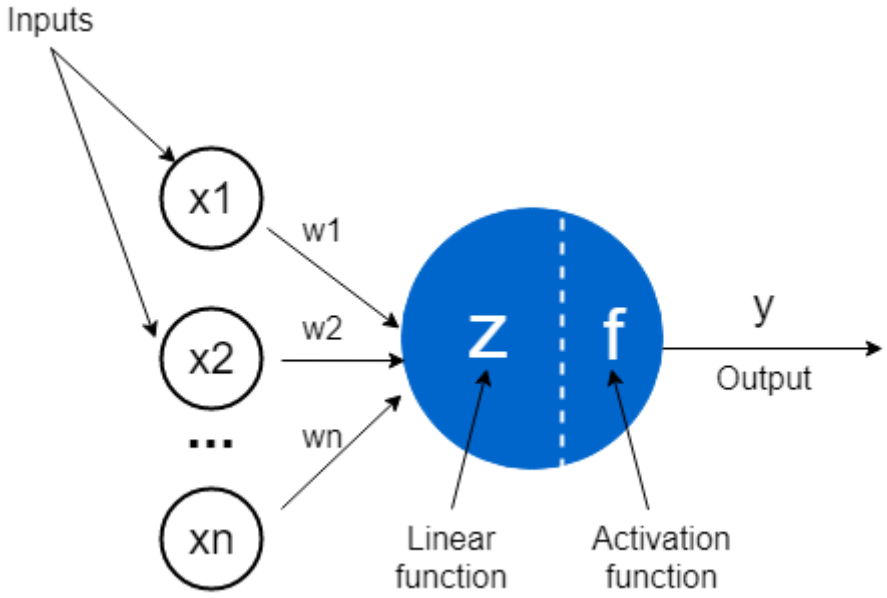
Η παρακάτω εικόνα δείχνει μια λεπτομερή δομή ενός perceptron. Σε ορισμένα περιβάλλοντα, η μεροληψία - bias, b μπορεί επίσης να συμβολίζεται με w_0. Η είσοδος x_0 παίρνει πάντα την τιμή 1. Συνεπώς έχουμε ότι $b * 1 = b$.



Εικόνα 13 Λεπτομερής παρουσίαση ενός τεχνητού νευρώνα.

Η έξοδος του νευρώνα προκύπτει από την ανωτέρω αλγεβρική εξίσωση. Ουσιαστικά, σε ένα perceptron εισάγονται τα $x_1, x_2, x_3, \ldots, x_n$, τα οποία όπως αναφέραμε και προηγουμένως έχουν τα αντίστοιχα βάρη, $w_1, w_2, w_3, \ldots, w_n$ κλπ. Στη συνέχεια, στα γινόμενα Wn*Xn προστίθεται και η μεροληψία b, οπότε υπολογίζεται η γραμμική συνάρτηση z. Με την εφαρμογή σε αυτής μιας συνάρτησης ενεργοποίησης f (activation function) προκύπτει η έξοδος του νευρώνα.

Όταν σχεδιάζεται ένα perceptron συνήθως αγνοείται η μονάδα προκατάληψης (b) για ευκολία και έτσι το διάγραμμα απλοποιείται σημαντικά όπως παρακάτω. Παρόλα αυτά, στους υπολογισμούς εξακολουθούμε να εξετάζουμε τη μονάδα μεροληψίας.



Συνοψίζοντας, ένα perceptron αποτελείται συνήθως από δύο μαθηματικές συναρτήσεις:

- Τη γραμμική συνάρτηση Perceptron: Αυτό ονομάζεται επίσης γραμμικό μέρος του perceptron και συμβολίζεται με z. Η έξοδος του είναι το σταθμισμένο άθροισμα των εισόδων συν τη μονάδα πόλωσης και υπολογίζεται ως εξής:

  $z = w_1 * x_1 + w_2 * x_2 + \ldots + w_n * x_n + b$

  όπου:

  - Τα $x_1, x_2, \ldots, x_n$ είναι είσοδοι που λαμβάνουν αριθμητικές τιμές. Μπορεί να υπάρχουν πολλές (πεπερασμένες) είσοδοι για έναν μόνο νευρώνα. Μπορούν να είναι ακατέργαστα δεδομένα εισόδου ή έξοδοι άλλων perceptrons.
  - Τα w1, w2, …, wn είναι βάρη που λαμβάνουν αριθμητικές τιμές ανάλογα με τη σπουδαιότητα της κάθε εισόδου.
  - $w_1 * x_1 + w_2 * x_2 + \ldots + w_n * x_n$ το σταθμισμένο άθροισμα των εισόδων.
  - Το b ονομάζεται όρος πόλωσης ή μονάδα πόλωσης και παίρνει επίσης μια αριθμητική τιμή. Προστίθεται στο σταθμισμένο άθροισμα των εισόδων. Ο σκοπός της συμπερίληψης ενός όρου μεροληψίας είναι να μετατοπιστεί η συνάρτηση ενεργοποίησης κάθε perceptron για να μην ληφθεί μηδενική τιμή. Στην περίπτωση δηλαδή που όλες οι είσοδοι x1, x2, …, xn είναι 0, το z θα ισούται με την τιμή της πόλωσης.

Τα βάρη και οι προκαταλήψεις ονομάζονται παράμετροι σε ένα μοντέλο νευρωνικών δικτύων. Μέσω της διαδικασίας εκπαίδευσης ενός νευρωνικού δικτύου υπολογίζονται οι εν λόγω παράμετροι.

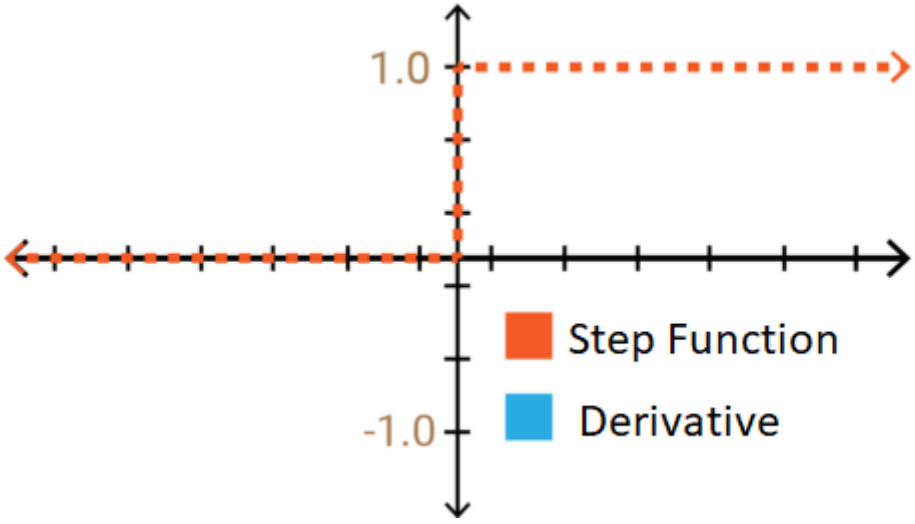
Κατ' αναλογία με ένα μοντέλο γραμμικής παλινδρόμησης, τα βάρη αντιστοιχούν στους συντελεστές και ο όρος μεροληψίας είναι γνωστός ως τομή.

- Τη μη γραμμική (συνάρτηση ενεργοποίησης) συνάρτηση Perceptron: Ορίζεται επίσης και ως μη γραμμική συνιστώσα του perceptron και συμβολίζεται με f. Εφαρμόζεται στο z για να λάβουμε την έξοδο y ανάλογα με τον τύπο της συνάρτησης ενεργοποίησης που χρησιμοποιούμε. Πρακτικά μια μη-γραμμική συνάρτηση ενεργοποίησης είναι μια συνάρτηση με πολλαπλές μοίρες. Τα τεχνητά νευρωνικά δίκτυα έχουν σχεδιαστεί ως καθολικές προσεγγίσεις συναρτήσεων και προορίζονται να λειτουργήσουν σε αυτόν τον στόχο. Αυτό σημαίνει ότι πρέπει να έχουν τη δυνατότητα να υπολογίζουν και να μαθαίνουν οποιαδήποτε συνάρτηση. Χάρη στις μη γραμμικές λειτουργίες ενεργοποίησης, μπορεί να επιτευχθεί ισχυρότερη εκμάθηση των δικτύων. Υπάρχουν ποικίλες συναρτήσεις ενεργοποίησης, με την κάθε μία να έχει σημεία που υπερτερεί ή μειονεκτεί συγκριτικά με τις υπόλοιπες συναρτήσεις. Στη συνέχεια παρουσιάζονται ορισμένες από αυτές.

4.2.1. Συναρτήσεις ενεργοποίησης:

Βηματική Συνάρτηση:

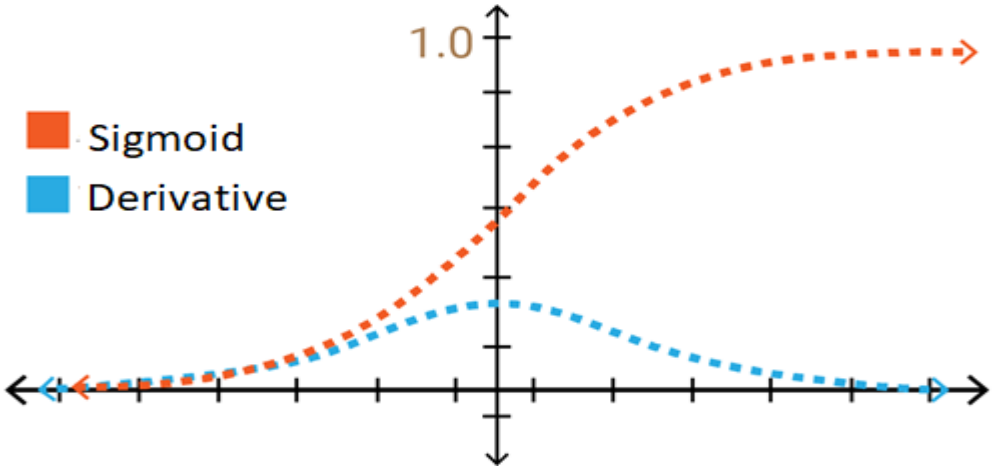
Είναι μια συνάρτηση που παίρνει μια δυαδική τιμή και χρησιμοποιείται ως δυαδικός ταξινομητής. Κατά συνέπεια, αυτό το χαρακτηριστικό έχει ως αποτέλεσμα να συναντάται πιο συχνά σε στρώματα εξόδου. Δεν συνίσταται η χρήση της σε κρυφά επίπεδα (αυτά που αποκαλούμε εσωτερικά επίπεδα) επειδή δεν είναι παραγωγίσιμη στο 0, αν και για την εκπαίδευση αυτό δεν αποτελεί σημαντικό πρόβλημα. Το σημαντικότερο όμως μειονέκτημα της συγκεκριμένης συνάρτησης είναι η μηδενική τιμή παραγωγού της σε όλα τα υπόλοιπα σημεία. Αυτό σημαίνει ότι η συνάρτηση δεν θα επιστρέφει κάποια κατεύθυνση για την εκπαίδευση αλλά σταθερά μηδέν. Αυτό καθιστά αδύνατη την εκπαίδευση ενός νευρωνικού δικτύου.



Εικόνα 14 Η γραφική αναπαράσταση της βηματικής συνάρτησης για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της.

Σιγμοειδής (sigmoid) συνάρτηση:

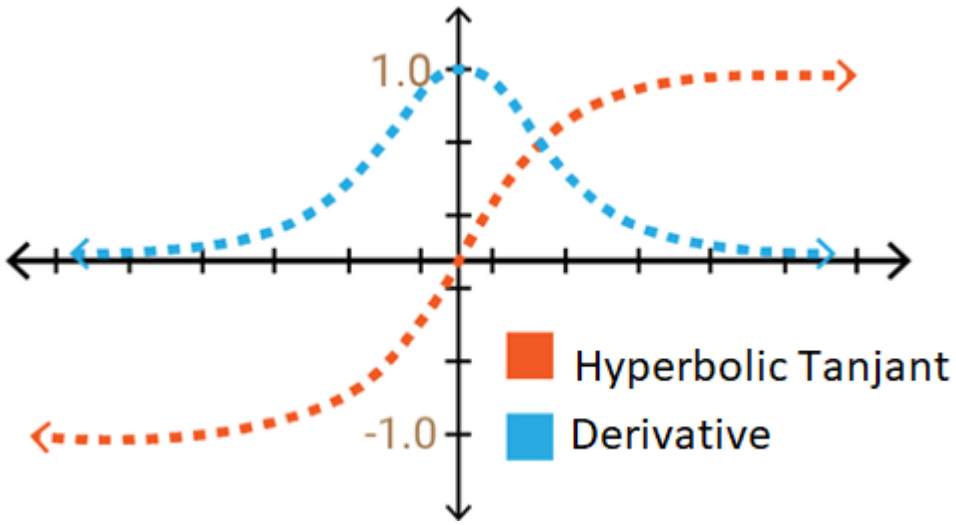
Η σιγμοειδής συνάρτηση είναι αρκετά παρόμοια με τη βηματική, με τη μόνη διαφορά ότι είναι παντού παραγωγίσιμη όπως φαίνεται και στο παρακάτω διάγραμμα, όποτε μπορεί να χρησιμοποιηθεί κατά την εκπαίδευση ενός νευρωνικού. Παρόλα αυτά το πρόβλημα της συγκεκριμένης συνάρτησης έγκειται στην παράγωγό της, η όποια είναι πολύ μικρή τόσο σαν μέγιστη τιμή όσο και στα άκρα της, όπου και παίρνει τιμές κοντά στο 0. Αφού οι τιμές αυτές όμως είναι τόσο μικρές στα άκρα τότε αυτό πρακτικά σημαίνει ότι αν ένας νευρώνας έχει κάποια ακραία τιμή (π.χ. -5) τότε η παράγωγος θα είναι πολύ κοντά στο 0 με αποτέλεσμα κατά την εκπαίδευσή τα βάρη του νευρώνα να μην ανανεώνονται σχεδόν καθόλου. Το πρόβλημα αυτό στην βιβλιογραφία αναφέρεται ως vanishing gradient και είναι ένα αρκετά σύνηθες φαινόμενο.



Εικόνα 15 Η γραφική αναπαράσταση της σιγμοειδούς συνάρτησης για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της.

Υπερβολική εφαπτομένη (tanh)**:**

Η συγκεκριμένη συνάρτηση παρουσιάζει αρκετές ομοιότητες με τη σιγμοειδή που παρουσιάστηκε παραπάνω. Η διαφορά τους έγκειται στο πεδίο τιμών τους, με αυτό της υπερβολικής εφαπτομένης να κυμαίνεται από -1 έως +1. Πλεονεκτεί συγκριτικά με τη σιγμοειδή συνάρτηση, καθώς η παράγωγός της είναι πιο απότομη, πράγμα που σημαίνει ότι μπορεί να πάρει μεγαλύτερες τιμές. Αυτό σημαίνει ότι θα είναι πιο αποτελεσματικό γιατί έχει μεγαλύτερο εύρος για ταχύτερη εκμάθηση. Παραμένει ωστόσο το πρόβλημα των κλίσεων στα άκρα της συνάρτησης και κατά συνέπεια το φαινόμενο του vanishing gradient μειώνεται, αλλά δεν εξαλείφεται.



Εικόνα 16 Η γραφική αναπαράσταση την συνάρτηση της υπερβολικής εφαπτομένης για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της

ReLU (Rectified Linear Unit):

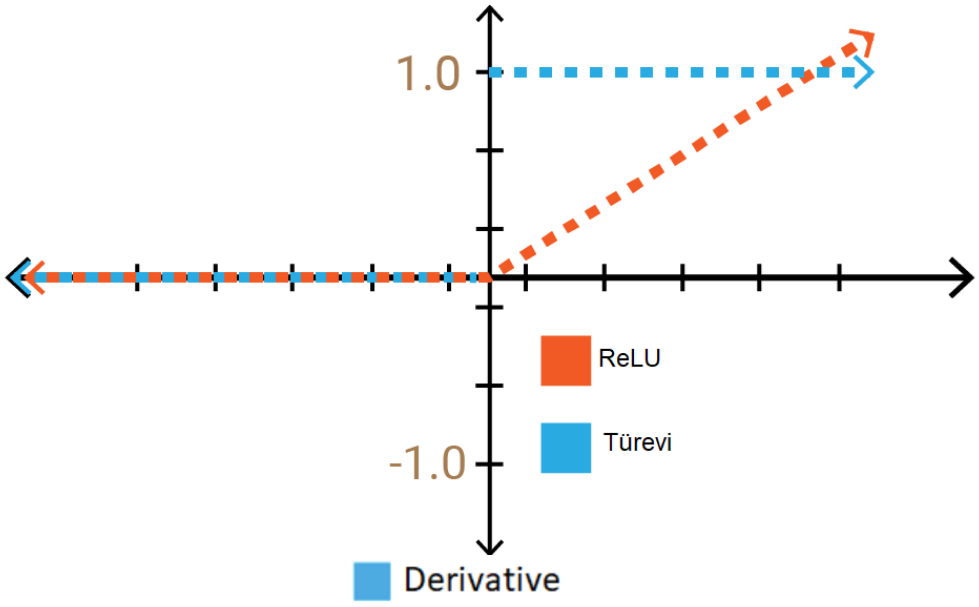
Πρακτικά η ReLU ορίζεται ως εξής:

$$ReLU = max(0, x)$$

και έτσι ουσιαστικά αποτελείται από 2 γραμμικά μέρη. Όπως και η βηματική, έτσι και η ReLU δεν είναι παραγωγίσιμη στο 0 αλλά αυτό δεν αποτελεί ιδιαίτερο πρόβλημα για την εκπαίδευση ενός νευρωνικού δικτύου.

Το βασικό πλεονέκτημα της παρούσας συνάρτησης έναντι των υπολοίπων είναι η σταθερή τιμή της παραγωγού για τις τιμές που είναι μεγαλύτερες του μηδενός.

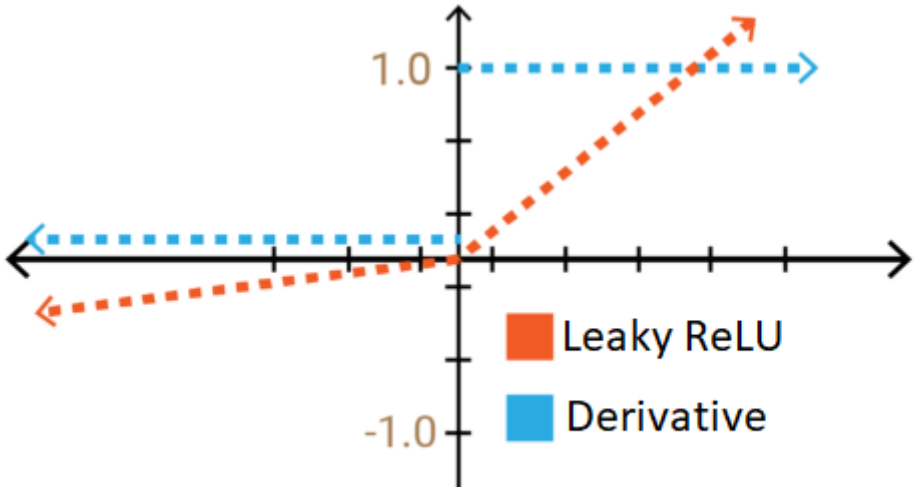
Έτσι εξαλείφεται το πρόβλημα του vanishing gradient που αναφέρθηκε παραπάνω για τις τιμές αυτές. Παρ' όλα αυτά για τιμές μικρότερες του μηδενός η παράγωγος είναι σταθερή και ίση με το μηδέν. Οπότε για αυτές τις τιμές το πρόβλημα παραμένει. Γενικά η ReLU είναι η πλέον διαδεδομένη και ευρέως χρησιμοποιούμενη συνάντηση ενεργοποίησης σε νευρωνικά δίκτυα λόγω των αποτελεσμάτων που επιφέρει. Επίσης, θεωρείται η πιο κοντινή στην πραγματική συνάρτηση ενεργοποίησης των βιολογικών νευρωνικών δικτύων.



Εικόνα 17 Η γραφική αναπαράσταση της συνάρτησης ReLU και της παραγώγου της. Η παράγωγος ορίζεται παντού πλην του 0 με μέγιστη τιμή το 1 για κάθε x > 0, σε αντίθεση με τις παραπάνω συναρτήσεις που είχαν μέγιστη τιμή μόνο σε ένα σημείο.

Leaky-ReLU Function:

Η συνάρτηση αυτή είναι ίδια με την προηγούμενη με τη μόνη διαφορά να εντοπίζεται στις τιμές που είναι μικρότερες του μηδενός. Δίνουμε μια μικρή κλίση στις τιμές που είναι μικρότερες του 0 ώστε να αποφύγουμε το πρόβλημα του vanishing gradient που αντιμετώπιζε σε αυτήν την περιοχή η συνάρτηση ReLU. Ένα ενδεικτικό εύρος τιμών της κλίσης είναι ανάμεσα σε 0.01 και 0.1.



Εικόνα 18 Η γραφική αναπαράσταση της Leaky ReLU και της παραγώγου της. Ορίζεται σε όλο το πεδίο τιμών πλην του 0, με μέγιστη τιμή το 1 για κάθε x > 0. Σε αντίθεση με την ReLU, η παράγωγος δεν είναι 0 για κάθε x < 0 αλλά έχει μια μικρή θετική τιμή

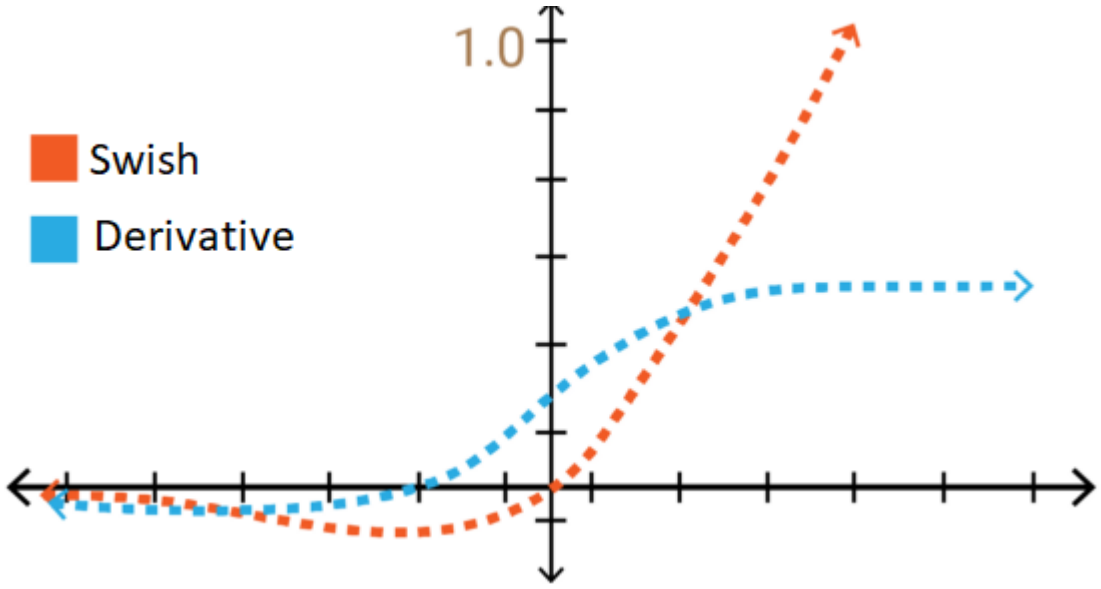
Συνάρτηση Softmax**:**

Έχει δομή παρόμοια με τη σιγμοειδή λειτουργία. Όπως και με την ίδια την Sigmoid, αποδίδει αρκετά καλά όταν χρησιμοποιείται ως ταξινομητής. Η πιο σημαντική διαφορά είναι ότι προτιμάται στο επίπεδο εξόδου των μοντέλων βαθιάς μάθησης, ειδικά όταν είναι απαραίτητο να ταξινομηθούν περισσότερα από δύο. Επιτρέπει τον προσδιορισμό της πιθανότητας ότι η είσοδος ανήκει σε μια συγκεκριμένη κλάση παράγοντας τιμές στην περιοχή 0-1. Άρα εκτελεί μια πιθανολογική ερμηνεία των αποτελεσμάτων ενός νευρωνικού δικτύου και για αυτόν ακριβώς τον λόγο χρησιμοποιείται σχεδόν πάντα ως συνάρτηση ενεργοποίησης στο τελευταίο επίπεδο.

Συνάρτηση Swish:

Η εν λόγω συνάρτηση παρουσιάζει αρκετές ομοιότητες με την ReLU, διαφέροντας κυρίως στην αρνητική περιοχή (όπως και η Leaky ReLU). Το κομμάτι αυτό αντί να έχει σταθερή τιμή ίση με 0 (όπως στην περίπτωση τηςReLU), ή μια μονότονη μικρή αρνητική κλίση (ισχύει στην περίπτωση τηςLeaky ReLU), αντιθέτως παρουσιάζει μια μεταβλητή κλίση και έτσι στις πολύ αρνητικές τιμές τείνει στο 0. Αξίζει να σημειωθεί ότι η έξοδος της συνάρτησης swish μπορεί να μειωθεί ακόμα και όταν η είσοδος αυξάνεται. Η αναλυτική περιγραφή της συνάρτησης αυτής στο αρνητικό μέρος παρουσιάζεται παρακάτω:

$$f(x) = 2 * x * sigmoid(beta * x)$$



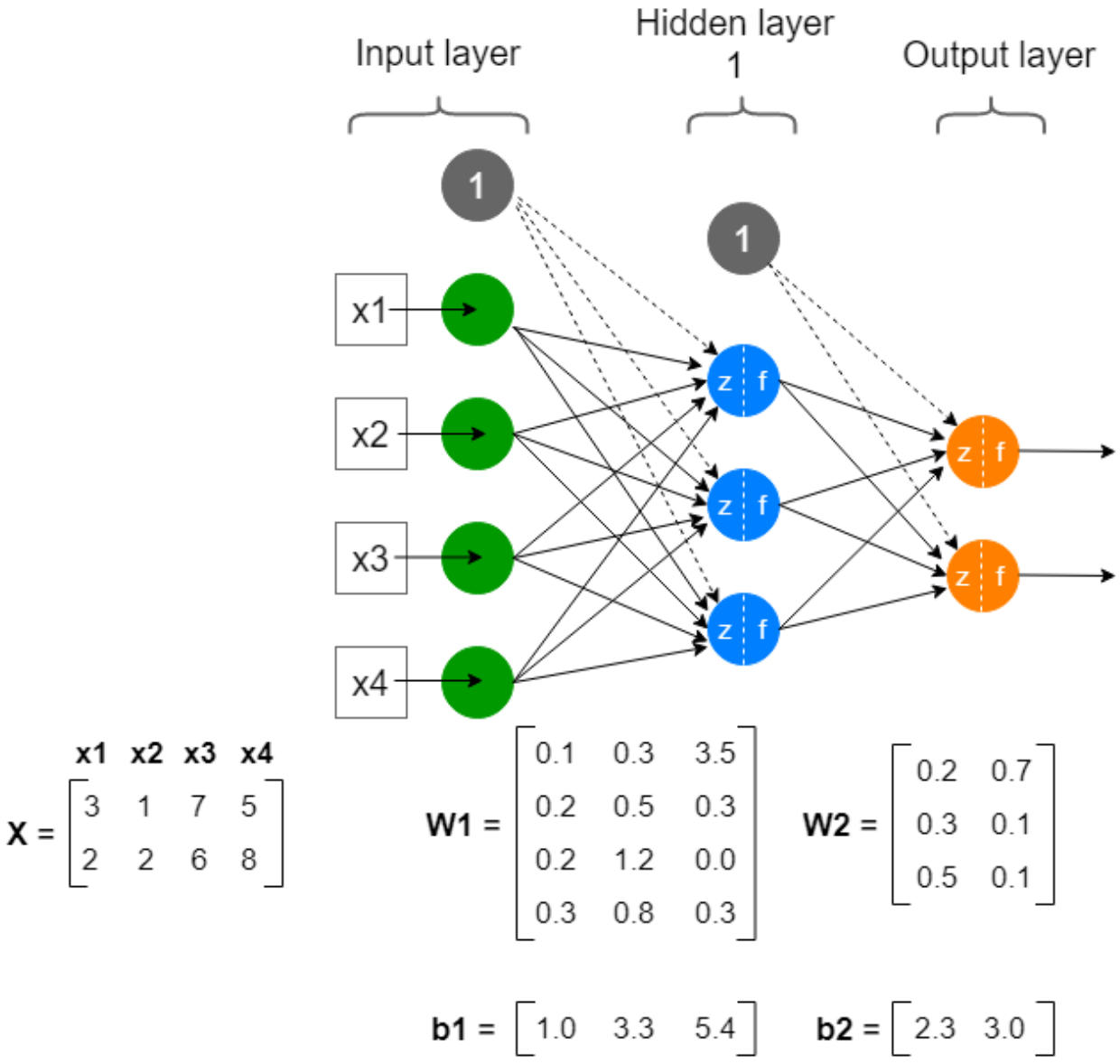
Εικόνα 19 Η γραφική αναπαράσταση της συνάρτηση Swish για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της. Σε αυτό φαίνεται ότι η παράγωγός της ορίζεται σε όλο το πεδίο τιμών

Η μεταβλητή b είναι μια εκμαθήσιμη παράμετρος. Στην παραπάνω εξίσωση αν $b = 0$ τότε το σιγμοειδές τμήμα είναι πάντα $1/2$ και η $f(x)$ είναι γραμμική. Από την άλλη πλευρά, εάν η βήτα είναι μια πολύ μεγάλη τιμή, το σιγμοειδές κομμάτι είναι σχεδόν διψήφια συνάρτηση (0 για $\chi < 0.1 \kappa\alpha\iota\, x > 0$. Έτσι η $f(x)$ συγκλίνει στη συνάρτηση ReLU. Επομένως, η τυπική συνάρτηση Swish επιλέγεται ως $b = 1$. Με αυτόν τον τρόπο, παρέχεται μια ήπια παρεμβολή (που συσχετίζει τα σετ μεταβλητών τιμών με μια συνάρτηση στο δεδομένο εύρος και την επιθυμητή ακρίβεια).

Συνοψίζοντας, σκοπός μιας συνάρτησης ενεργοποίησης είναι να εισαγάγει μη γραμμικότητα στο δίκτυο. Ένα νευρωνικό δίκτυο που δε διαθέτει συνάρτηση ενεργοποίησης θα μπορούσε να μοντελοποιήσει μόνο γραμμικές σχέσεις – συναρτήσεις. Εντούτοις, τα περισσότερα προβλήματα στον πραγματικό κόσμο είναι μη γραμμικά. Επομένως, είναι ιδιαίτερα σημαντικό τα νευρωνικά δίκτυα να μπορούν να περιγράψουν μη-γραμμικές συναρτήσεις από το οποίο απορρέει και η αναγκαιότητα για τη μελέτη των συναρτήσεων ενεργοποίησης.

Ένα perceptron από μόνο του δεν είναι αρκετά ικανό να μοντελοποιήσει πολύπλοκες μη γραμμικές σχέσεις ή ακόμα και απλές σχέσεις στα δεδομένα εισόδου. Επομένως, πολλαπλά perceptron στοιβάζονται μαζί για να σχηματίσουν μια πολύπλοκη δομή (δίκτυο/αρχιτεκτονική) που ονομάζεται Τεχνητό Νευρωνικό Δίκτυο ή Artificial Neural Network (ANN), το οποίο είναι αρκετά ικανό να μοντελοποιήσει οποιαδήποτε μη γραμμική σχέση σύμφωνα με το Θεώρημα Καθολικής Προσέγγισης. Ένα νευρωνικό δίκτυο είναι επίσης γνωστό ως Perceptron πολλαπλών επιπέδων

(Multilayer Perceptron MLP). Αυτό συμβαίνει επειδή μια συλλογή από perceptrons στοιβάζεται σε πολλαπλά στρώματα μέσω των οποίων πραγματοποιούνται οι συνδέσεις μεταξύ των perceptrons. Επιπλέον, ένα MLP είναι ένα πλήρως (πυκνά) συνδεδεμένο νευρωνικό δίκτυο (Fully Convolutional Neural Network FCNN) στο οποίο κάθε κόμβος σε κάθε επίπεδο συνδέεται με όλους τους άλλους κόμβους στο επόμενο επίπεδο, όπως στο παρακάτω διάγραμμα. Οι κόμβοι που βρίσκονται στο ίδιο επίπεδο δεν μοιράζονται καμία σύνδεση.



Εικόνα 20 Στο ανωτέρω σχήμα απεικονίζεται η δομή ενός τεχνητού νευρωνικού δικτύου που διαθέτει ένα μόνο κρυφό επίπεδο, μαζί με ένα παράδειγμα τιμών των παραμέτρων κάθε επιπέδου

Το παραπάνω δίκτυο έχει ένα επίπεδο εισόδου (input layer) , ένα επίπεδο εξόδου (output layer) και ένα κρυφό στρώμα (hidden layer). Ένα τεχνητό νευρωνικό δίκτυο με μόνο ένα κρυφό στρώμα αποκαλείται και Shallow Neural Network ή ρηχό νευρωνικό δίκτυο. Προκύπτει έτσι και ο

ορισμός για το βαθύ νευρωνικό δίκτυο ή Deep Neural Network, το οποίο ορίζεται ως το δίκτυο που διαθέτει δύο ή περισσότερα κρυφά επίπεδα.

Ως τελευταίο βασικό ορισμό αναφέρουμε τα νευρωνικά δίκτυα τροφοδοσίας (Feed Forward Neural Network FFNN). Ο όρος "τροφοδοσία προς τα εμπρός" σημαίνει ότι τα δεδομένα μετακινούνται από την είσοδο στην έξοδο μέσω επιπέδων προς μία (προς τα εμπρός) κατεύθυνση. Σε ένα νευρωνικό δίκτυο τροφοδοσίας προς τα εμπρός, η μόνη ανατροφοδότηση συμβαίνει όταν τα δεδομένα εισόδου φτάσουν στο επίπεδο εξόδου και δεν υπάρχει ανάδραση ενώ τα δεδομένα κινούνται μέσω των ενδιάμεσων στρωμάτων.

Ανακεφαλαιώνοντας, ένα νευρωνικό δίκτυο αποτελείται από τα εξής επίπεδα:

- Το επίπεδο εισόδου: Το επίπεδο εισόδου αποτελείται από νευρώνες εισόδου που λαμβάνουν εισόδους, x1, x2, κ.λπ. Οι νευρώνες εισόδου ενός MLP (FCNN) δεν εκτελούν κανέναν υπολογισμό και απλώς βγάζουν ό,τι τους δίνεται. Με άλλα λόγια, διατηρούν απλώς τα δεδομένα εισόδου. Επομένως, οι νευρώνες εισόδου ενός MLP (FCNN) είναι απλώς κόμβοι. Υπάρχει επίσης ένας νευρώνας πόλωσης και πάντα βγάζει την τιμή 1. Υπάρχουν διάφοροι τύποι επιπέδων εισόδου που χρησιμοποιούμε στο Keras. Τα πιο δημοφιλή είναι τα Dense (FC-Fully Connected), Convolutional και Recurrent. Το σχήμα των δεδομένων εισόδου εξαρτάται από τον τύπο του επιπέδου εισόδου που χρησιμοποιούμε στα νευρωνικά μας δίκτυα. Σημαίνει ότι πρέπει να αναδιαμορφώσουμε τα δεδομένα εισόδου ανάλογα με τον τύπο του επιπέδου εισόδου. Οι νευρώνες εισόδου παίρνουν πάντα αριθμούς. Εάν παρέχουμε εικόνες, βίντεο, κείμενα ή ομιλία για δεδομένα εισαγωγής, θα μετατραπούν σε αριθμούς. Αυτοί οι αριθμοί είναι διατεταγμένοι σε πολυδιάστατους πίνακες (ονομάζονται επίσης τανυστές στην ορολογία βαθιάς μάθησης).

  Σε έναν τύπο στρώματος εισόδου πυκνού (FC), ο αριθμός των νευρώνων εισόδου, δηλαδή το μέγεθος του επιπέδου εισόδου, είναι ίσος με τον αριθμό των χαρακτηριστικών (στήλες) των δεδομένων εισόδου. Αυτά τα χαρακτηριστικά δωρίζονται από τα x1, x2, x3, κ.λπ. Αυτές είναι οι είσοδοι στο πρώτο κρυφό στρώμα και συνδέονται πλήρως με τους νευρώνες στο κρυφό στρώμα μέσω των βαρών που μετρούν το επίπεδο σπουδαιότητας. Το μέγεθος του επιπέδου εισόδου είναι μια υπερπαράμετρος που πρέπει να καθορίσουμε στο μοντέλο μας πριν από την εκπαίδευση. Η βέλτιστη τιμή του δεν μαθαίνεται από δεδομένα. Αντίθετα, οι παράμετροι μαθαίνουν τις βέλτιστες τιμές τους από δεδομένα κατά τη διάρκεια της εκπαιδευτικής διαδικασίας.

- Το κρυφό επίπεδο (hidden layer): Αυτό είναι το επίπεδο μεταξύ των επιπέδων εισόδου και εξόδου. Όπως αναφέραμε και προηγουμένως, σε ένα ρηχό νευρωνικό δίκτυο συναντάμε μόνο ένα κρυφό στρώμα. Σε βαθιά νευρωνικά δίκτυα όπου υπάρχουν δύο ή περισσότερα κρυφά επίπεδα, ο αριθμός των κρυφών επιπέδων θα πρέπει να προσδιορίζεται από τον προγραμματιστή. Ο αριθμός των κρυφών επιπέδων μετρά το βάθος ενός νευρωνικού δικτύου. Όσο αυξάνει ο αριθμός των κρυφών επιπέδων σε ένα νευρωνικό δίκτυο, τόσο ικανότερο γίνεται να μοντελοποιήσει πιο σύνθετες σχέσεις Ωστόσο, χρειάζεται περισσότερος χρόνος για την εκπαίδευση και επίσης το μοντέλο θα τείνει να ταιριάζει υπερβολικά με τα δεδομένα.

- Το στρώμα εξόδου: Το στρώμα εξόδου αποτελείται από νευρώνες εξόδου που κάνουν την τελική πρόβλεψη. Πρέπει να προσδιορίσουμε το μέγεθος του επιπέδου εξόδου, δηλαδή τον αριθμό των νευρώνων στο επίπεδο εξόδου. Η αξία του εξαρτάται από το είδος του προβλήματος που καλείται να λύσει το νευρωνικό.

4.3.SVM (support vector machines)

Ένα άλλο, επίσης γνωστό, μοντέλο μηχανικής μάθησης είναι τα Support Vector Machines (SVM), το οποίο χρησιμοποιείται πολύ συχνά σε προβλήματα ταξινόμησης. Ο λόγος της εκτεταμένης του χρήσης σε αυτά είναι η αποτελεσματικότητά του σε μεγάλο πλήθος περιπτώσεων. Γενικότερα, μπορούν να χρησιμοποιηθούν είτε σε προβλήματα ταξινόμησης είτε σε regression problems, ωστόσο για τους σκοπούς της παρούσης μας αφορά η πρώτη περίπτωση. Η επιτυχία τους λοιπόν οφείλεται στο γεγονός πως μπορούν να διαχειριστούν τόσο συνεχή όσο και κατηγοριοποιημένα δεδομένα. Σκοπός τους είναι να κατασκευάζουν ένα πολυδιάστατο (multidimensional) υπερεπίπεδο για τον διαχωρισμό των δεδομένων ανά κλάση. Το υπερεπίπεδο αυτό μαθαίνεται αυτόματα μελετώντας τα δεδομένα και προσπαθώντας να ελαχιστοποιήσει το σφάλμα ταξινόμησης σε αυτά. Η κυρία αρχή λειτουργία τους στηρίζεται στην εύρεση του υπερεπιπέδου με τη μεγαλύτερη δυνατή απόσταση - κενό μεταξύ των κλάσεων maximum marginal hyperplane (MMH).

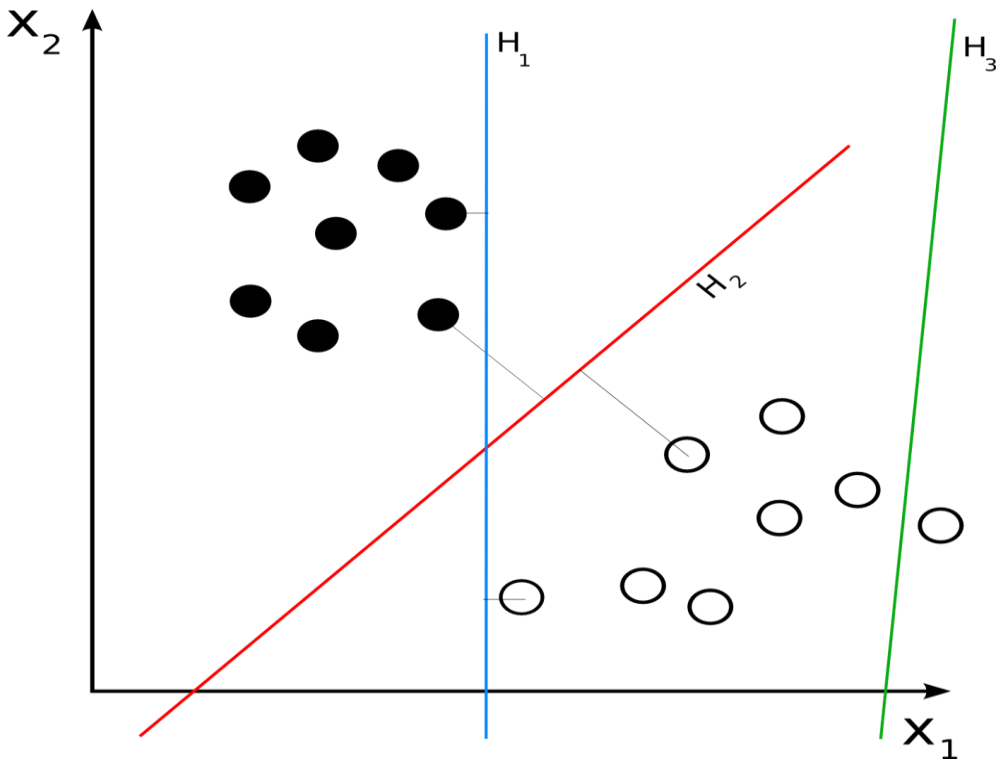
Για την καλύτερη κατανόηση του προβλήματος στο παρακάτω διάγραμμα παρουσιάζεται ένα απλό παράδειγμα διαχωρισμού δεδομένων δύο διαστάσεων σε δύο κλάσεις. Υποθέτουμε πως τα μαύρα δεδομένα ανήκουν στην κλάση 0 και τα λευκά στην κλάση 1. Αρχικά η γραμμή H3 μπορούμε εύκολα να καταλάβουμε ότι δεν είναι κατάλληλη για να επιτύχει τον διαχωρισμό των δεδομένων στις δύο αυτές κλάσεις. Παρόλα αυτά τα επίπεδα H1 και H2 μπορούν να ταξινομήσουν τα δεδομένα με

την ίδια ακριβώς ακρίβεια. Οπότε εδώ τίθεται το ερώτημα του τρόπου με τον οποίον θα επιλέξουμε την κατάλληλη διαχωριστική επιφάνεια.

Ο κύριος στόχος μας είναι να διαχωρίσουμε το σύνολο δεδομένων με τον καλύτερο δυνατό τρόπο. Η απόσταση μεταξύ των κοντινότερων δεδομένων τα οποία ανήκουν σε διαφορετικές κλάσεις ονομάζεται στην βιβλιογραφία ως κενό (margin). Σκοπός του αλγορίθμου SVM συνεπώς είναι να επιλέξει ένα διαφορετικό επίπεδο το οποίο θα αφήνει το μεγαλύτερο δυνατό margin μεταξύ των κοντινότερων δεδομένων. Η εύρεση του επιπέδου αυτού γίνεται ακολουθώντας τα παρακάτω βήματα:

1. Αρχικά δημιουργεί πολλαπλές επιφάνειες οι οποίες μπορούν να διαχωρίσουν το πρόβλημα με τον καλύτερο δυνατό τρόπο. Αυτές οι επιφάνειες στο παραπάνω παράδειγμα είναι η H1, H2. Η H3 δεν αποτελεί τέτοια υπερ-επιφάνεια μιας και δεν έχει το ελάχιστο δυνατό σφάλμα ταξινόμησης, όπως έχουν οι H1, H2.
2. Στη συνέχεια, από αυτές επιλέγει την διαχωριστική επιφάνεια η οποία μεγιστοποιεί το κενό μεταξύ των κοντινών δεδομένων, στο παράδειγμά μας μεταξύ των H1, H2 επιλέγει την H2.

Ο αλγόριθμος αυτός όπως αναφέρεται παραπάνω έχει επιδείξει εξαιρετικά αποτελέσματα σε πληθώρα προβλημάτων. Στο δικό μας πρόβλημα που προσπαθούμε να ανιχνεύσουμε τις μη έγκυρες ηλεκτρονικές συναλλαγές, χρησιμοποιούνται ως είσοδος οι τιμές που έχουν παραχθεί από την PCA, δηλαδή το διάνυσμα V1-V28, καθώς και οι κανονικοποιημένες τιμές των πεδίων amount και time, ενώ έξοδο του συστήματος αποτελεί η πρόβλεψη για το αν η συναλλαγή ήταν έγκυρη ή απάτη.

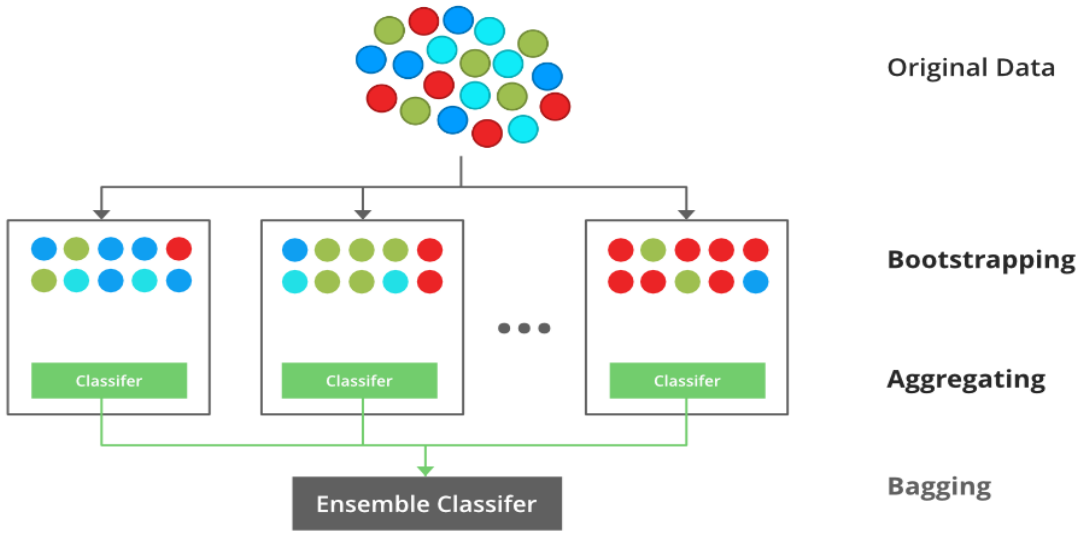


Εικόνα 21 Ο τρόπος υπολογισμού του υπερεπιπέδου διαχωρισμού σύμφωνα με την μέθοδο SVM.

4.4.XGBoost

Ο XgBoost αποτελεί αρκτικόλεξο του Extreme Gradient Boosting. Οι ταξινομητές τύπου Gradient boosting classifiers είναι ένα σύνολο από αλγορίθμους μηχανικής μάθησης οι όποιοι λειτουργούν συνδυασμένα ώστε να μπορέσουν να επιτύχουν μια πιο σίγουρη πρόβλεψη. Συγκεκριμένα, αντί κάθε φορά να εκπαιδεύεται ένα μοντέλο για την πρόβλεψη, χρησιμοποιούνται πολλαπλά τα όποια εκπαιδεύονται παράλληλα, με διαφορετικά βάρη εκκίνησης, διαφορετικές παραμέτρους και ίσως και διαφορετικές εσωτερικές αρχιτεκτονικές. Συνήθως χρησιμοποιούνται Decision Trees τα όποια έχουν εξηγηθεί παραπάνω αναλυτικότερα. Τα μοντέλα αυτού του τύπου έχουν αποκτήσει μεγάλη προσοχή τελευταία μιας και έχουν καταφέρει εκπληκτικά αποτελέσματα σε αρκετά σύνθετα tasks και σε ευρύ φάσμα.

Μια οπτική απεικόνιση της μεθόδου αυτής παρουσιάζεται στο παρακάτω σχήμα. Σε αυτό φαίνεται η χρήση διαφορετικών μοντέλων και η ενσωμάτωση τους για την βελτίωση των τελικών αποτελεσμάτων της ταξινόμησης.

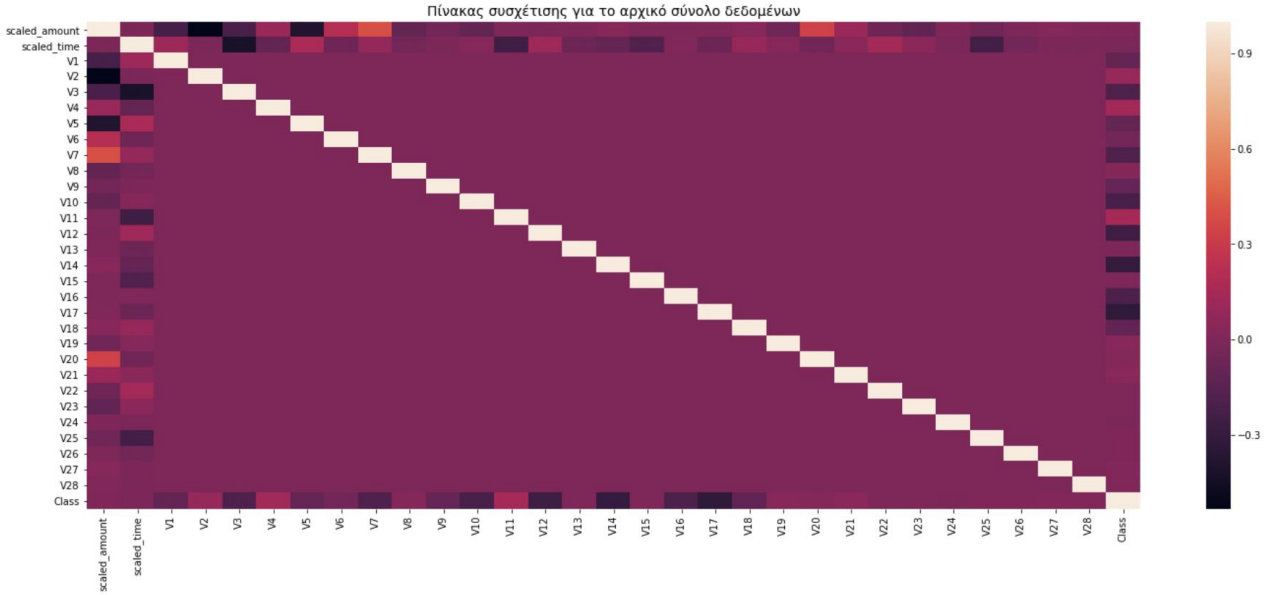


Εικόνα 22 Η γραφική αναπαράσταση των boosted μεθόδων ταξινόμησης και ο τρόπος με τον οποίον χρησιμοποιούν πολλαπλούς ταξινομητές για την τελική πρόβλεψη.

5. Πειράματα

Πριν ξεκινήσουμε τη σύγκριση των αποτελεσμάτων στο επίπεδο ταξινόμησης, μελετήσαμε γενικότερα την ποιότητα των παραγόμενων δεδομένων, ώστε να αντιληφθούμε αν υπήρξε κάποιου είδους βελτίωση συγκριτικά με το αρχικό, ιδιαίτερα ανισοκατανεμημένο σετ δεδομένων. Αυτό επιχειρήθηκε να γίνει με ποικίλους τρόπους. Σε πρώτο στάδιο, προκειμένου να είμαστε σε θέση να καταλάβουμε αν βελτιώθηκε η ποιότητα των δεδομένων, έγινε αντιπαραβολή των νέων με τα αρχικά.

Ο πρώτος τρόπος σύγκρισης που επιλέξαμε ήταν αυτός των πινάκων συσχέτισης. Συγκεκριμένα υπολογίσαμε την ανά ζεύγη συσχέτιση (pairwise correlation) με χρήση της μεθόδου Pearson. Στο διάγραμμα που ακολουθεί φαίνεται η συσχέτιση ανά στήλη για τα αρχικά, μη-ισορροπημένα δεδομένα.

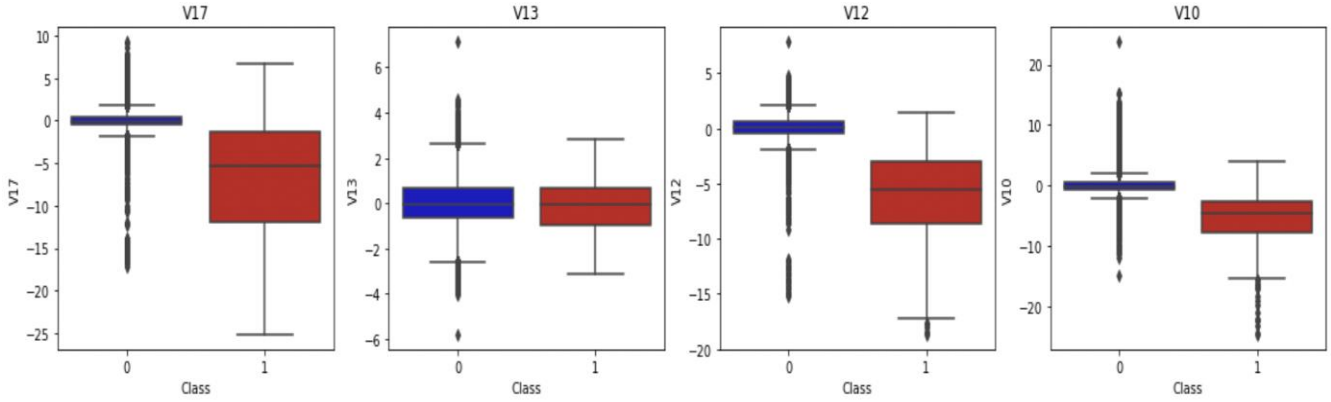


Εικόνα 23 Ο πίνακας συσχέτισης του αρχικού συνόλου δεδομένων.

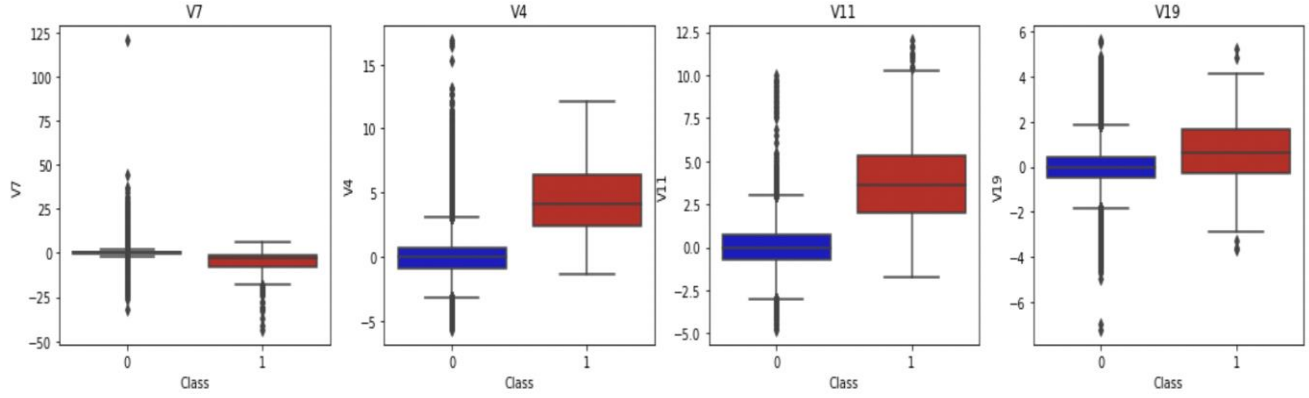
Σε κάθε πεδίο αυτού του πίνακα εμφανίζεται η συσχέτιση που έχει το πεδίο που αναφέρεται στην αντίστοιχη στήλη με το πεδίο που αναφέρεται στην αντίστοιχη γραμμή. Οπότε από το διάγραμμα αυτό μας αφορά κυρίως η γραμμή (ή στήλη) του πεδίου Class (τελευταία στήλη ή τελευταία γραμμή). Σε αυτό φαίνεται πόσο συσχετισμένο είναι κάθε άλλο πεδίο των δεδομένων με το αν η συγκεκριμένη συναλλαγή είναι απάτη ή όχι.

Από το διάγραμμα αυτό φαίνεται ότι τα πεδία V17, V13, V12 και V10 είναι αρνητικά συσχετισμένα με την κλάση των δειγμάτων. Αντίθετα υπάρχει θετική συσχέτιση μεταξύ αυτού και των πεδίων V7, V4, V11 και V19. Οι συσχετίσεις αυτές φαίνονται πιο έντονα στα παρακάτω

διαγράμματα. Στο πρώτο εμφανίζονται τα boxplot των V17, V13, V12, V10 σε σχέση με την κλάση (απάτη ή όχι) ενώ στο επόμενο αυτών που έχουν θετική συσχέτιση.

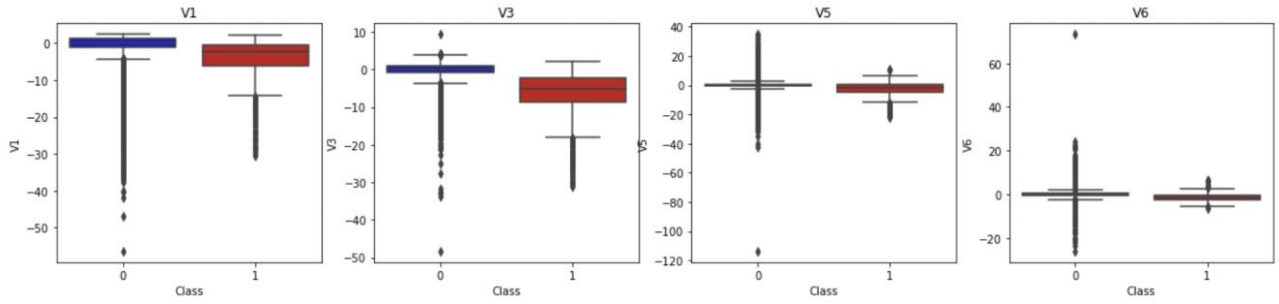


Εικόνα 24 Τα bar plot των συσχετίσεων για τις μεταβλητές V17, V13, V12 και V10 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

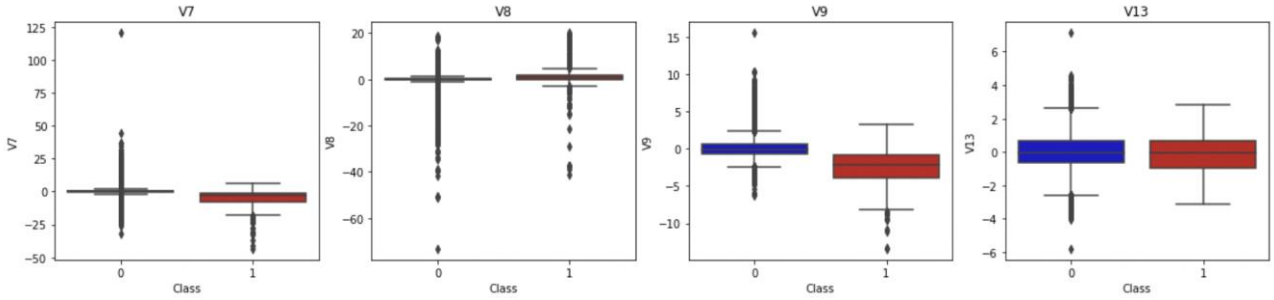


Εικόνα 25 Τα bar plot των συσχετίσεων για τις μεταβλητές V7, V4,, V11 και V19 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

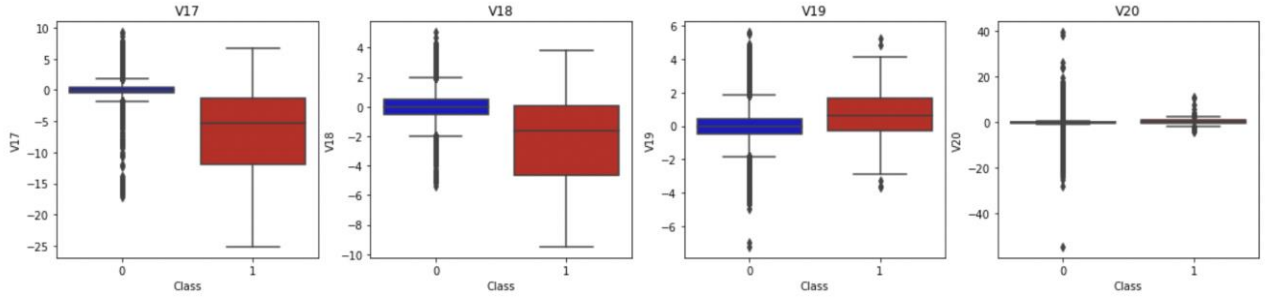
Στα παραπάνω δυο σχήματα εμφανίζονται τα διαγράμματα για τα πεδία που εμφανίζουν αρνητική και θετική συσχέτιση με την κλάση. Προκειμένου να γίνει πιο κατανοητή η σχέση ανάμεσα στα πεδία αυτά και τον τύπο της συναλλαγής, παρακάτω παρουσιάζονται και τα αντίστοιχα διαγράμματα για όλα τα υπόλοιπα πεδία.



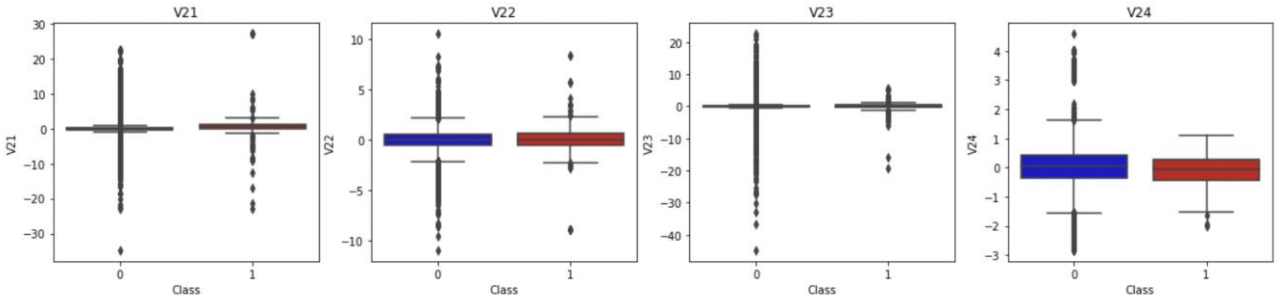
Εικόνα 26 Τα bar plot των συσχετίσεων για τις μεταβλητές V1, V3, V5 και V6 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



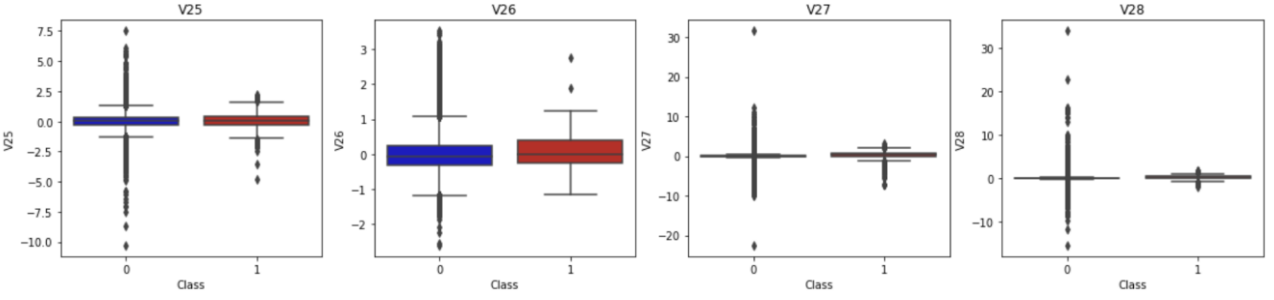
Εικόνα 27 Τα bar plot των συσχετίσεων για τις μεταβλητές V7, V8, V9 και V13 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



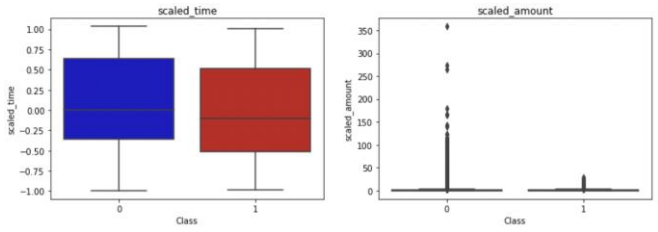
Εικόνα 28 Τα bar plot των συσχετίσεων για τις μεταβλητές V17, V18, V19 και V20 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



Εικόνα 29 Τα bar plot των συσχετίσεων για τις μεταβλητές V21, V22, V23 και V24 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



Εικόνα 30 Τα barplot των συσχετίσεων για τις μεταβλητές V25, V26, V27 και V28 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

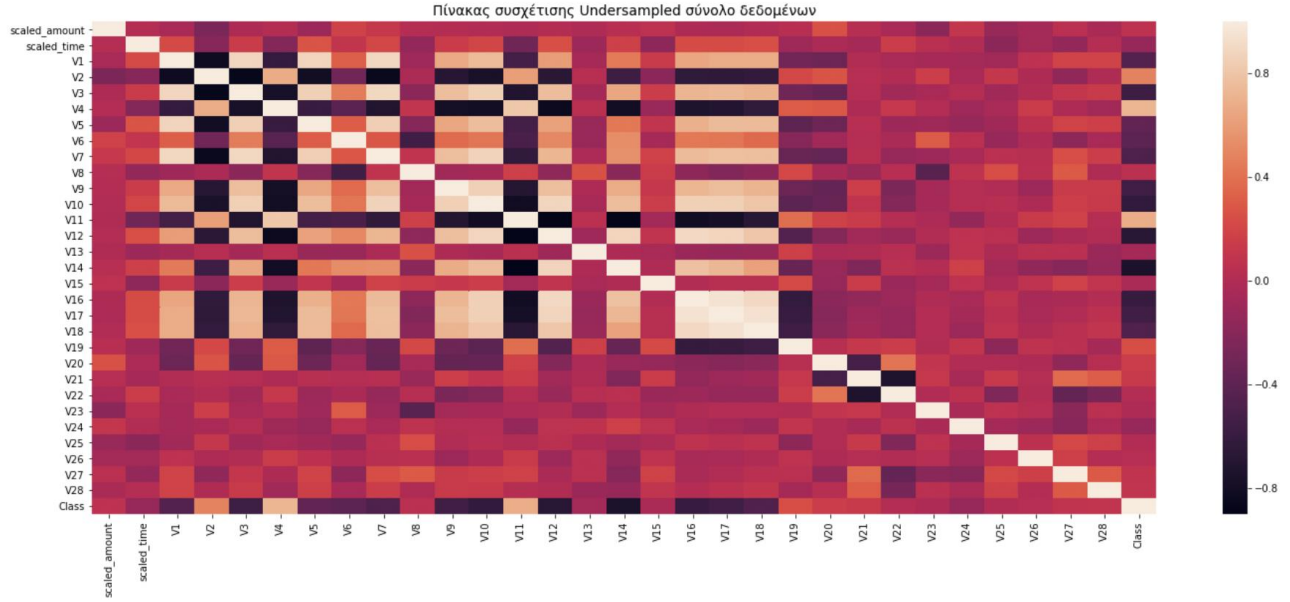


Εικόνα 31 Τα bar plot των συσχετίσεων για τις μεταβλητές scaled_time και scaled_amount και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

Από τα τελευταία διαγράμματα φαίνεται ότι τα πεδία που δεν έχουν συσχέτιση με την κλάση εμφανίζουν ανάλογη συμπεριφορά ανεξάρτητα από το αν μια συναλλαγή είναι απάτη ή όχι (κατανέμονται στις ίδιες θέσεις). Αυτό κάνει τα πεδία αυτά να μην είναι κατάλληλα (από μόνα τους) για να υποδείξουν μοτίβα σχετικά με την κατηγοριοποίηση των δεδομένων. Παρόλα αυτά, δεν είναι σωστό να απορρίψουμε τα πεδία αυτά από την είσοδο για κυρίως 2 λόγους. Ο πρώτος είναι ότι από μόνο του ένα πεδίο μπορεί να μην είναι ικανό να διαχωρίσει τα δεδομένα αλλά σε συνδυασμό να επιτυγχάνει πολύ καλή ακρίβεια. Χαρακτηριστικό παράδειγμα τέτοιας περίπτωσης αποτελεί ο γνωστός σε όλους δείκτης μάζας σώματος. Για τον υπολογισμό του συγκεκριμένου δείκτη λαμβάνονται υπόψη το ύψος και το βάρος ενός ατόμου, Κάθε ένα από αυτά ξεχωριστά δεν αρκεί για να κατατάξουμε ένα άτομο σε κάποια κατηγορία, ωστόσο ο συνδυασμός τους είναι μια από τις πιο διάσημες παγκοσμίως μετρικές. Το δεύτερο πρόβλημα είναι το πρόβλημα της γενίκευσης το όποιο υπάρχει καθολικά όταν μεταχειριζόμαστε ανισοκατανεμημένα δεδομένα. Για αυτό το μικρό πλήθος μη-έγκυρων συναλλαγών μπορεί μια μεταβλητή να μην έχει κάποια συσχέτιση, αλλά για ένα άλλο σύνολο δεδομένων - πιο γενικευμένο μπορεί να επηρεάζει σημαντικά την πρόβλεψη. Για παράδειγμα, αν η μια από τις μεταβλητές αυτές ήταν η χώρα και στο σύνολο μας είχαμε μόνο χώρες της ΕΕ τότε το πιο πιθανό ήταν το πεδίο αυτό να είχε ελάχιστη συσχέτιση με την κλάση της πρόβλεψης. Αν όμως στο σύνολο των δεδομένων ενσωματωνόταν και αυτό της Ινδίας, η όποια αποτελεί το μεγαλύτερο σημείο εκκίνησης παράνομων συναλλαγών (λόγω των VPN και της νομοθεσίας της scam-centers κ.α.),, τότε η χώρα εκκίνησης της συναλλαγής θα είχε μεγάλη συσχέτιση με την εγκυρότητά της. Οπότε αξίζει να μελετήσουμε τις συσχετίσεις που υπάρχουν στο σύνολο δεδομένων, κυρίως για να εξετάσουμε κατά πόσο κάθε μέθοδος εξισορρόπησης τις αλλάζει και όχι για να απορρίψουμε κάποιο πεδίο.

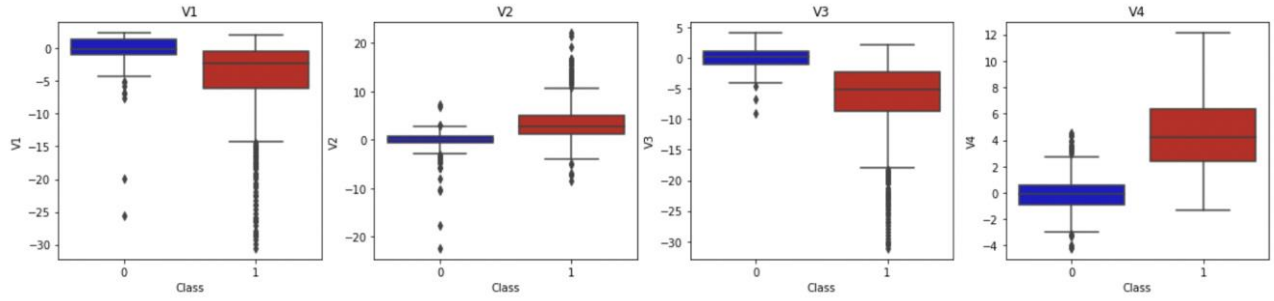
Στην συνέχεια αξίζει να μελετήσουμε τις αντίστοιχες συσχετίσεις για τα δεδομένα που παράγονται από κάθε μια από τις μεθόδους που χρησιμοποιήσαμε και να μελετήσουμε κατά πόσο αυτές επηρέασαν τις αρχικές τιμές που παρουσιάστηκαν παραπάνω. Θεωρητικά τα ισοκατανεμημένα δεδομένα θέλουμε να ακολουθούν την κατανομή του αρχικού συνόλου των δεδομένων. Στο

παρακάτω διάγραμμα εμφανίζεται ο πίνακας συσχέτισης για τα δεδομένα που παράχθηκαν μέσω της μεθόδου Undersampling:

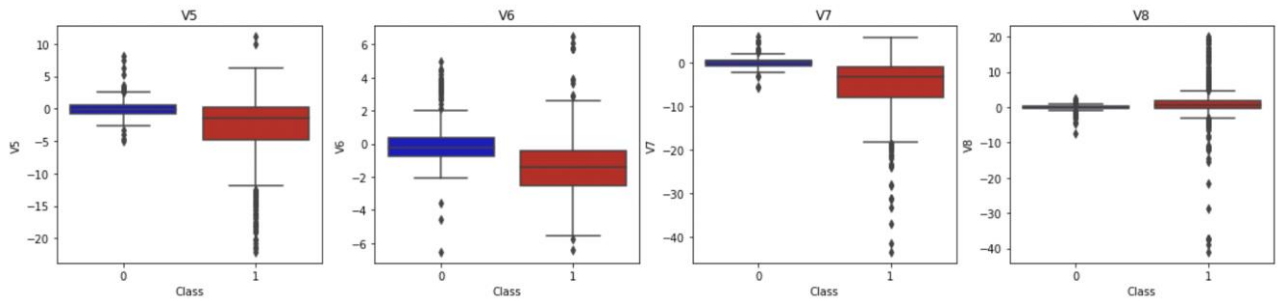


Εικόνα 32 Ο πίνακας συσχέτισης για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.

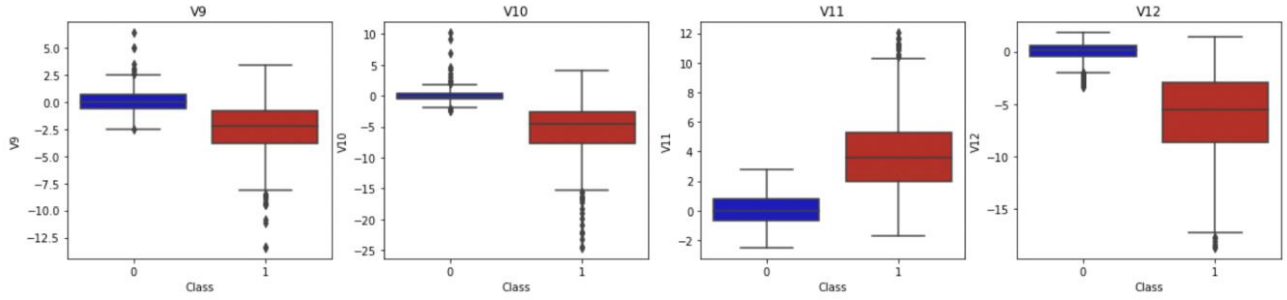
Από αυτό το διάγραμμα παρατηρείται μεγάλη διαφορά στις συσχετίσεις των πεδίων συγκριτικά με αυτές του ανισοκατανεμημένου συνόλου δεδομένων. Παρακάτω εμφανίζονται και τα αντίστοιχα box plots.



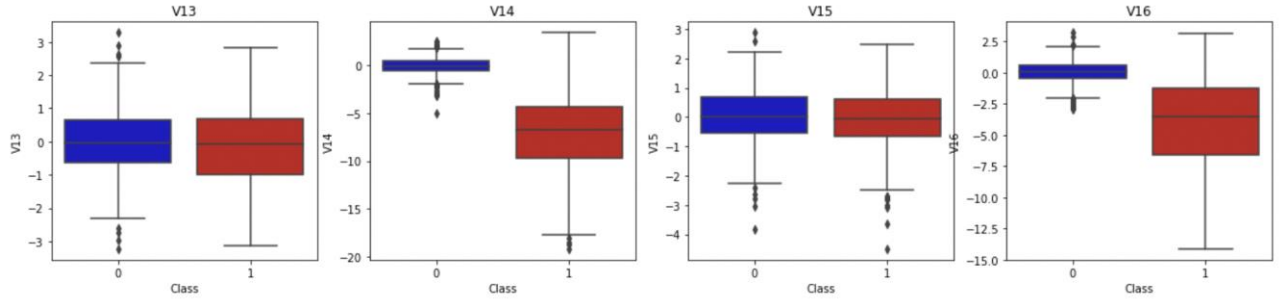
Εικόνα 33 Τα bar plot των συσχετίσεων για τις μεταβλητές V1, V2, V3 και V4 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας



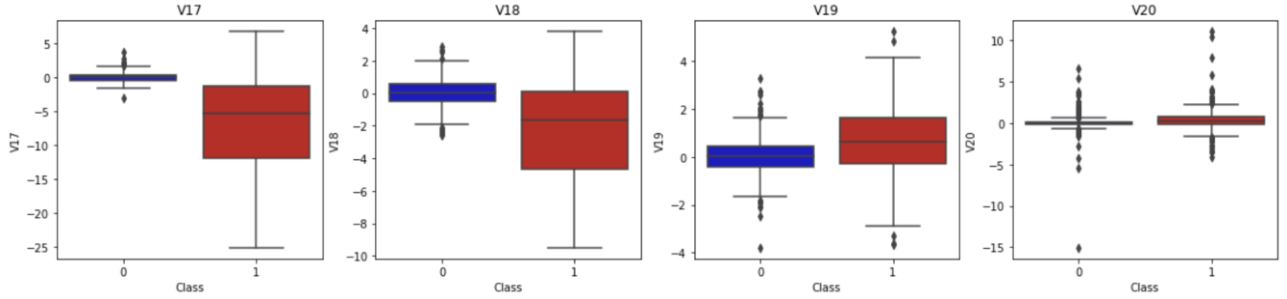
Εικόνα 34 Τα bar plot των συσχετίσεων για τις μεταβλητές V5, V6, V7 και V8 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.



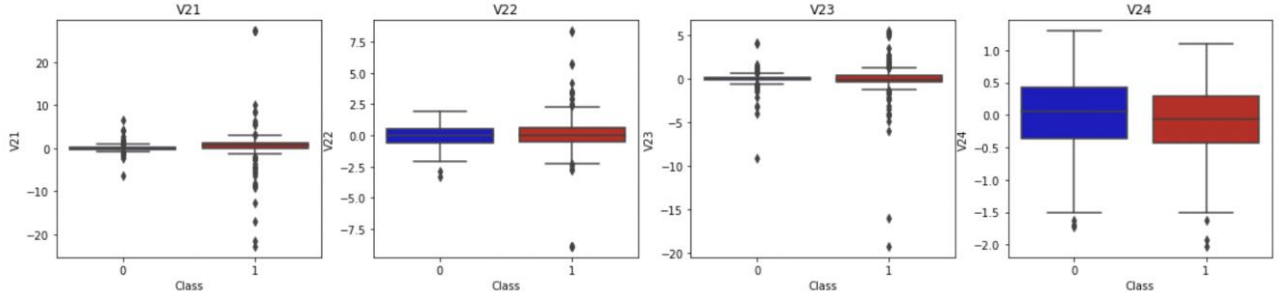
Εικόνα 35 Τα bar plot των συσχετίσεων για τις μεταβλητές V9, V10, V11 και V12 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.



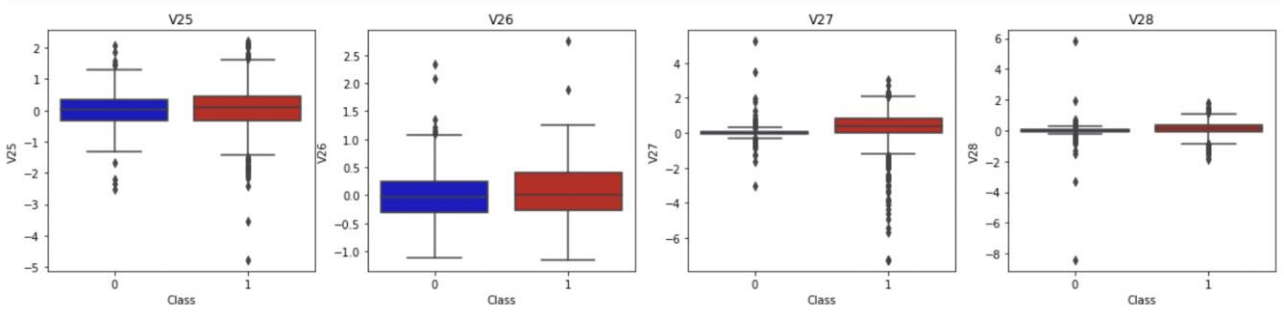
Εικόνα 36 Τα bar plot των συσχετίσεων για τις μεταβλητές V13, V14, V15 και V16 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.



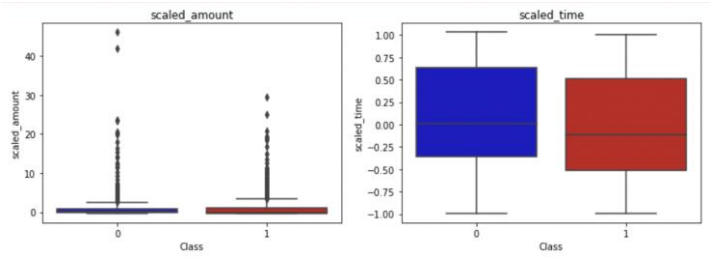
Εικόνα 37 Τα bar plot των συσχετίσεων για τις μεταβλητές V17, V18, V19 και V20 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας



Εικόνα 38 Τα bar plot των συσχετίσεων για τις μεταβλητές V21, V22, V23 και V24 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.

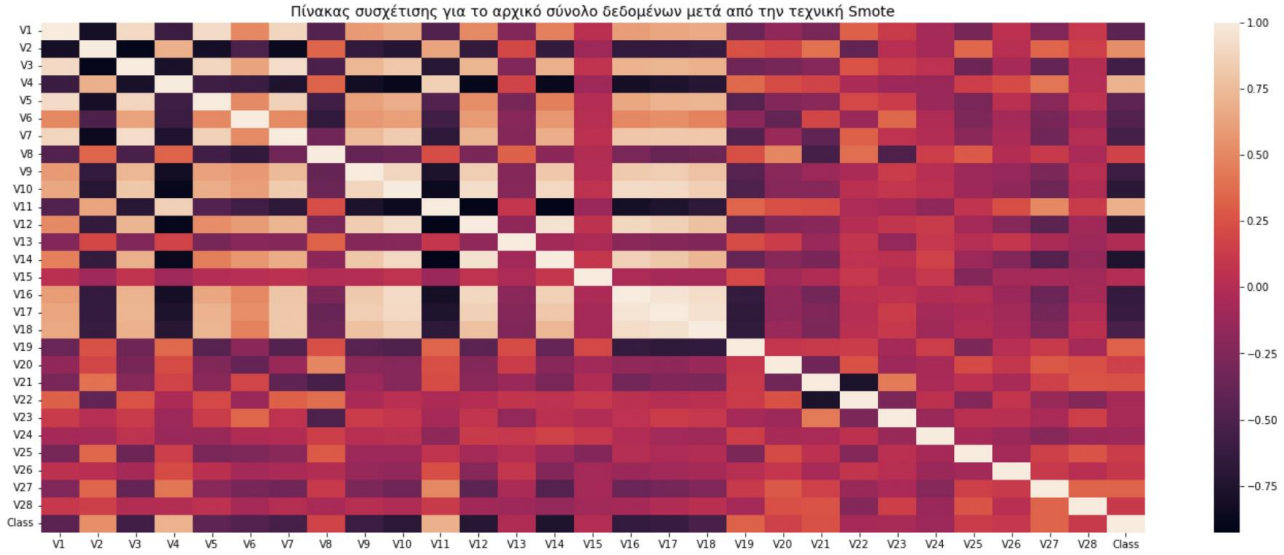


Εικόνα 39 Τα bar plot των συσχετίσεων για τις μεταβλητές V25, V26, V27 και V28 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.

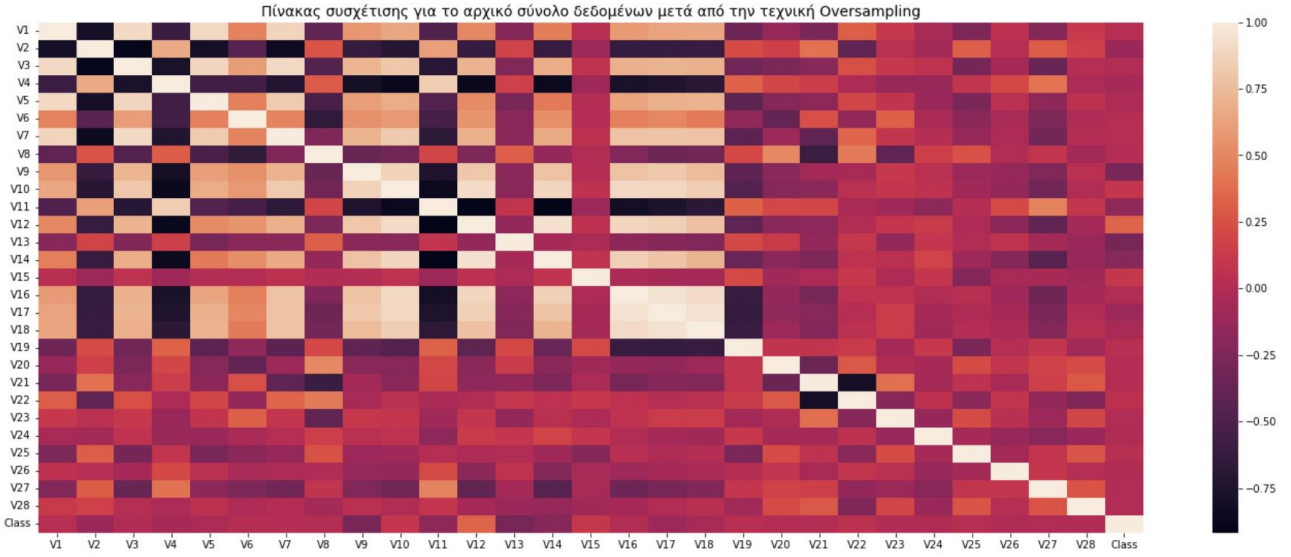


Εικόνα 40 Τα bar plot των συσχετίσεων για τις μεταβλητές scaled_amount και scaled_time και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας.

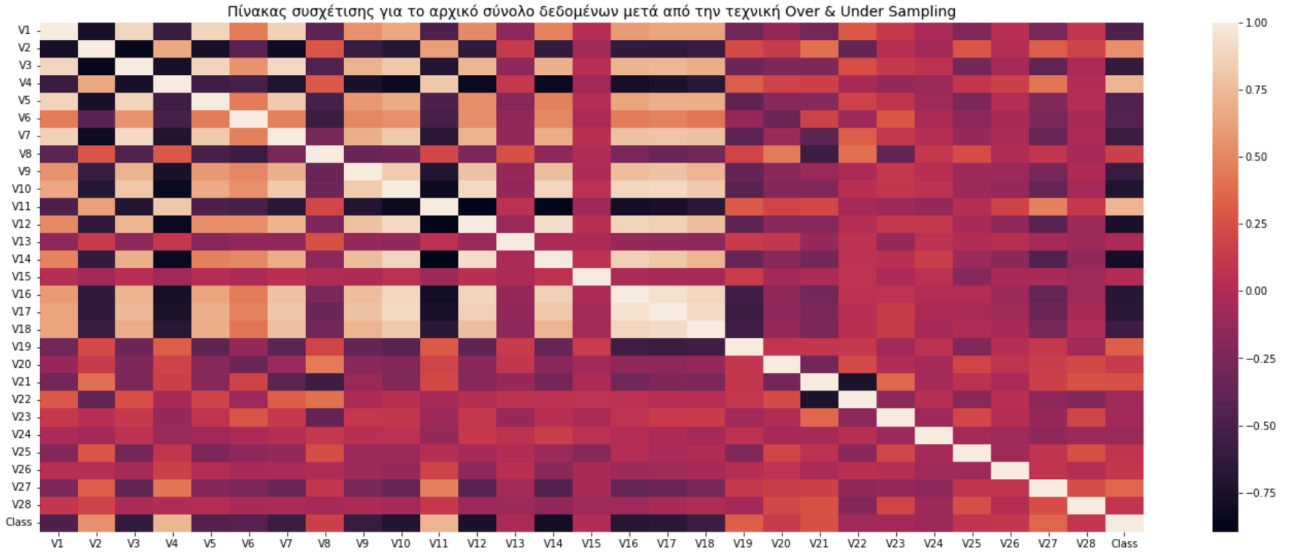
Παρακάτω παρουσιάζονται τα αντίστοιχα διαγράμματα συσχετίσεων για κάθε μια από τις υπόλοιπες μεθόδους.



Εικόνα 41 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου smote.



Εικόνα 42 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου oversampling.

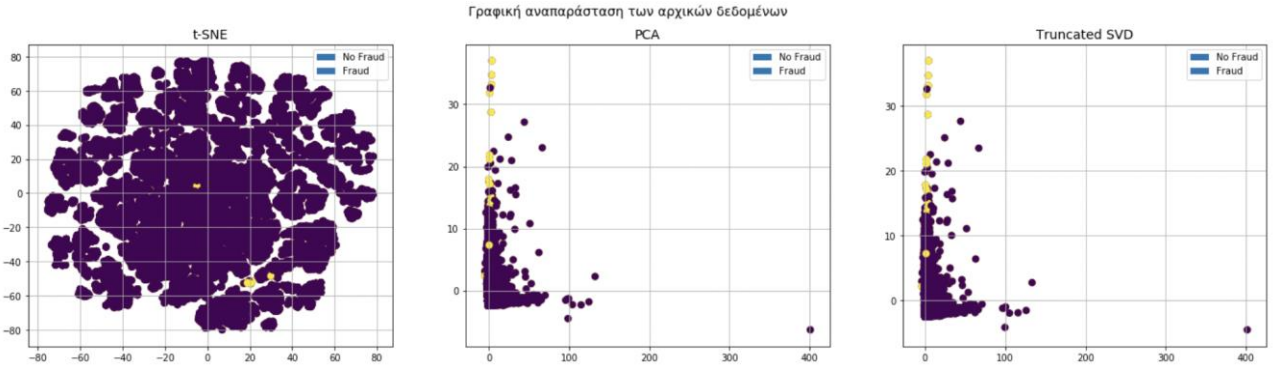


Εικόνα 43 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου Over & Undersampling.

Το επόμενο βήμα είναι να δούμε κατά πόσο μπορούμε να οπτικοποιήσουμε τα δεδομένα ώστε να τα καταλάβουμε λίγο καλύτερα. Σε αυτήν την προσπάθεια καλούμαστε να αντιμετωπίσουμε τη δυσκολία σχεδιασμού των δεδομένων στο χώρο εξαιτίας των πολλών διαστάσεών τους. Κατά συνέπεια, οφείλουμε να μειώσουμε τις διαστάσεις σε μόλις δύο ώστε να ώστε να γίνει εφικτή η οπτικοποίησή τους στον χώρο. Με το ανάλογο κόστος φυσικά, που δεν είναι άλλο από την προφανή απώλεια πληροφορίας. Όπως γίνεται εύκολα κατανοητό, είναι αδύνατο να μειωθούν οι μεταβλητές χωρίς να χάσουμε μέρος της πληροφορίας. Αν ίσχυε το αντίθετο, αυτό θα σήμαινε ότι η ποιότητα των δεδομένων είναι ενδεχομένως χαμηλή.

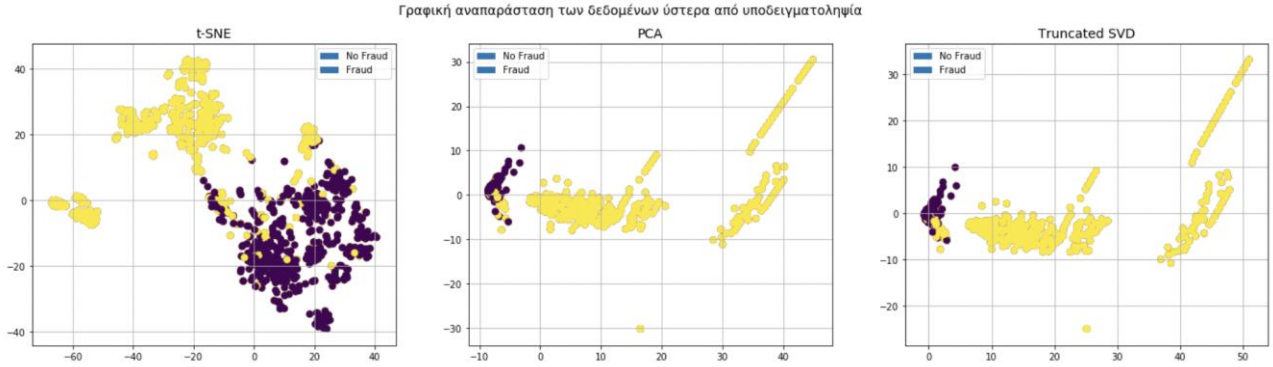
Προκειμένου να επιτύχουμε την εν λόγω μείωση θα εφαρμόσουμε τρεις διαφορετικές τεχνικές και πιο συγκεκριμένα τις PCA, t-SNE και Truncated SVD. Στη συνέχεια θα

οπτικοποιήσουμε τα αποτελέσματά τους. Παρακάτω απεικονίζονται τα αποτελέσματα από την εφαρμογή των συγκεκριμένων τεχνικών:



Εικόνα 44 Η γραφική αναπαράσταση των δειγμάτων του αρχικού συνόλου δεδομένων έπειτα από την εφαρμογή των τριών μεθόδων μείωσης της διαστατικότητάς τους.

Τα δεδομένα τα όποια έχουν υποστεί υποδειγματοληψία εμφανίζονται στο παρακάτω διάγραμμα.

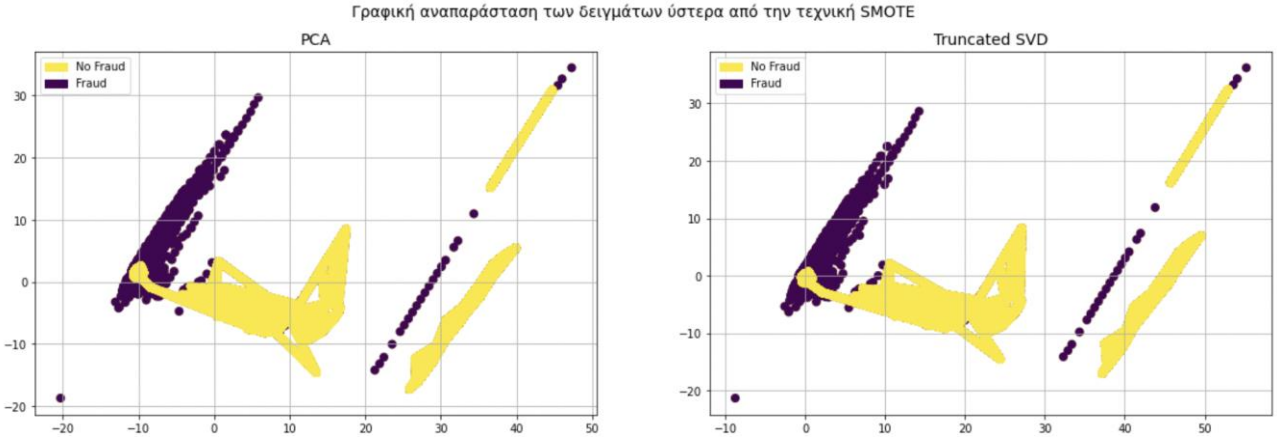


Εικόνα 45 Η γραφική αναπαράσταση των δειγμάτων του συνόλου δεδομένων μετά την εφαρμογή της μεθόδου υποδειγματοληψίας και των 3 μεθόδων μείωσης της διαστατικότητάς τους.

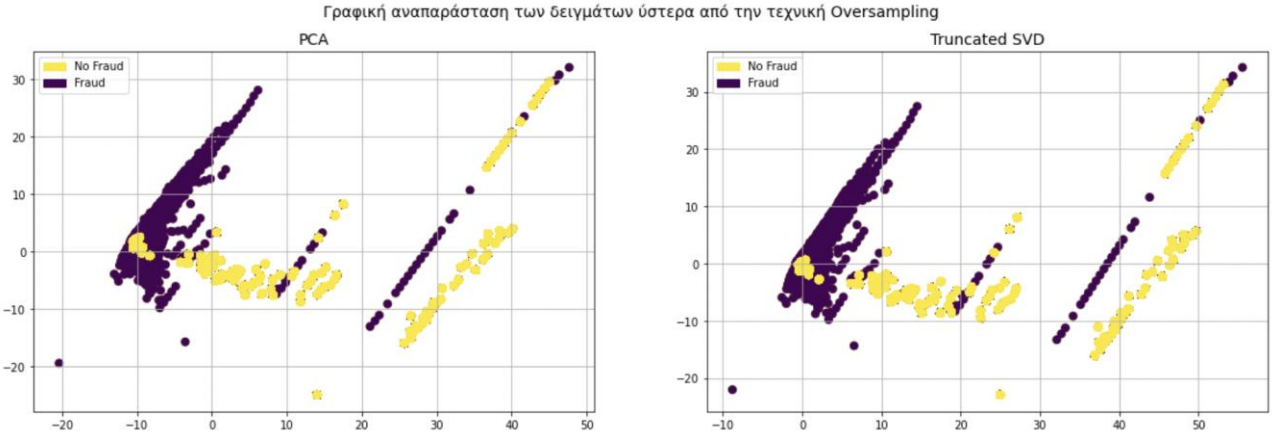
Από το παραπάνω σχήματα εύκολα διακρίνεται ότι τα δεδομένα τα οποία έχουν υποστεί υποδειγματοληψία μπορούν να διαχωριστούν καλύτερα από τα αρχικά. Αυτό όμως δεν οφείλεται σε κάποια τεχνοτροπία της μεθόδου αλλά κατά κύριο λόγο στο γεγονός ότι οι μέθοδοι μείωσης των διαστάσεων έτρεξαν με λιγότερα δεδομένα. Αυτό έχει σαν αποτέλεσμα η συντριπτική πλειοψηφία των δεδομένων από έγκυρες συναλλαγές να απουσιάζει και να φαίνονται ουσιαστικά σαν να υπάρχουν 2 διακριτές περιοχές. Βέβαια, στα αρχικά δεδομένα βλέπουμε ότι ορισμένες από τις μη έγκυρες συναλλαγές μπορούν να διαχωριστούν με σχετικά καλή ακρίβεια όπως φαίνεται από τις

μεθόδους PCA και Truncated SVD. Τα αποτελέσματα της t-sne δεν παρέχουν αυτή την πληροφορία, αφού τα αρχικά δεδομένα δεν μπορούν να διακριθούν κάπως σε γειτονιές ανά κλάση.

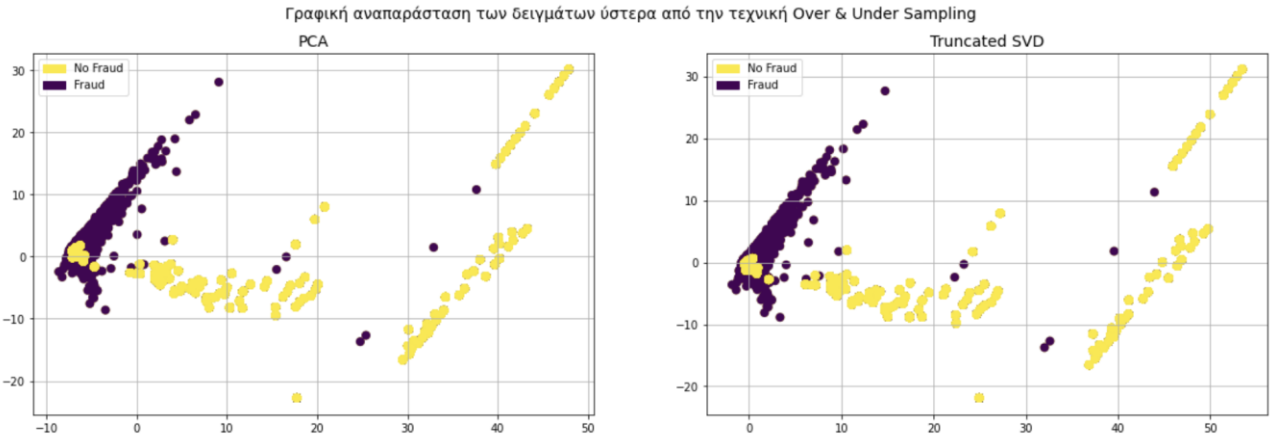
Παρακάτω παρουσιάζονται οι αντίστοιχες γραφικές αναπαραστάσεις ύστερα από την εφαρμογή κάθε μιας μεθόδου.



Εικόνα 46 Η γραφική αναπαράσταση των δειγμάτων του συνόλου δεδομένων μετά την εφαρμογή της μεθόδου smote έπειτα από την εφαρμογή 3 μεθόδων μείωσης της διαστατικότητάς τους.



Εικόνα 47 Η γραφική αναπαράσταση των δειγμάτων του συνόλου δεδομένων μετά την εφαρμογή της μεθόδου της υπερδειγματοληψίας και των 3 μεθόδων μείωσης της διαστατικότητάς τους.



Εικόνα 48 Η γραφική αναπαράσταση των δειγμάτων του συνόλου δεδομένων μετά την εφαρμογή της μεθόδου της υπερ & υποδειγματοληψίας έπειτα από την εφαρμογή 3 μεθόδων μείωσης της διαστατικότητάς τους.

Από τα παραπάνω διαγράμματα παρατηρούμε ότι ουσιαστικά οι μόνες αλλαγές μεταξύ των αναπαραστάσεων της υποδειγματοληψίας, υπερδειγματοληψίας και του συνδυασμού τους είναι η προσθήκη ή η αφαίρεση σημείων. Από την άλλη πλευρά, στην γραφική απεικόνιση των σημείων που προέρχονται από την τεχνική Smote φαίνεται ότι έχουν προστεθεί σημεία, δημιουργώντας πολλαπλές γραμμές όπως ακριβώς παρουσιάστηκε και αναλύθηκε και στο παράδειγμα στο κεφάλαιο που αναλύεται η συγκεκριμένη μέθοδος.

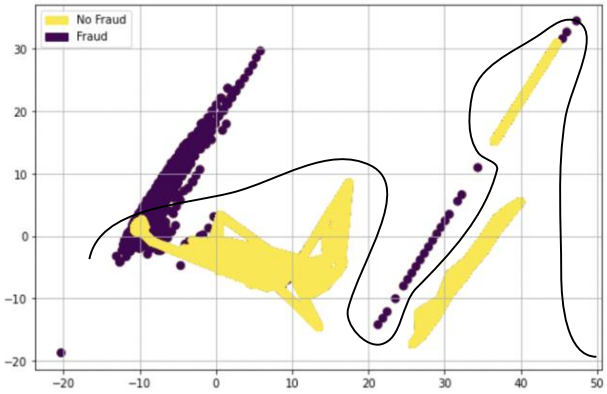
Η αξιολόγηση της αποδοτικότητας των μεθόδων αυτών έγινε με τη δοκιμή των δεδομένων πριν και ύστερα από την εφαρμογή κάθε μεθόδου σε διαφορετικές αρχιτεκτονικές ταξινόμησης. Αξίζει να σημειωθεί ότι το σύνολο δοκιμής (test set) παρέμεινε σταθερό και αμετάβλητο για κάθε ένα πείραμα. Οπότε στο σύνολο δοκιμής, αν ένα σύστημα το όποιο έχει εκπαιδευτεί στο αρχικό σύνολο δεδομένων έχει απόδοση a ενώ ένα σύστημα το όποιο έχει εκπαιδευτεί σε δεδομένα που έχουν υποστεί κάποια μέθοδο έχει απόδοση b τότε αν:

- $a > b$, τότε η μέθοδος δεν βοήθησε την εκπαίδευση του συστήματος
- $a < b$, τότε η μέθοδος βελτίωσε την απόδοση της ταξινόμησης
- $a == b$, τότε η μέθοδος άφησε ανεπηρέαστα τα αποτελέσματα

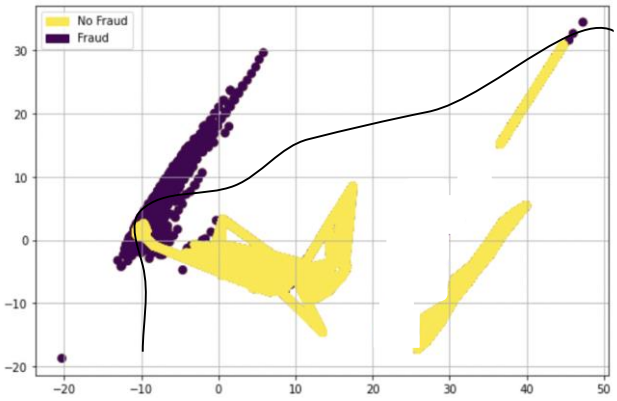
Ο παραπάνω διαχωρισμός παρόλο που είναι διαισθητικά κατανοητός δεν είναι απόλυτα σωστός. Αυτό οφείλεται στο γεγονός ότι οι περισσότερες από αυτές τις μεθόδους είναι στατιστικές όποτε η απόδοση ενός συστήματος μπορεί να είναι μειωμένη, όχι επειδή η μέθοδος δεν είναι καλή αλλά

επειδή η συγκεκριμένη εκτέλεση κάποιας μεθόδου χειροτέρεψε την ποιότητα του νέου συνόλου δεδομένων.

Παρακάτω φαίνονται οι γραφικές αναπαραστάσεις πιθανών αποτελεσμάτων της μεθόδου υποδειγματοληψίας για κάποιο σύνολο δεδομένων μαζί με τις αντίστοιχες διαχωριστικές γραμμές που θα μπορούσαν να έχουν, όπως την φαντάζεται ο συγγραφέας. Στο πρώτο σχήμα είναι η γραφική αναπαράσταση κάποιου συνόλου δεδομένων, ενώ στα 2 επόμενα τα αποτελέσματα από δυο διαφορετικά παραδείγματα.



Εικόνα 49 Η καμπύλη ταξινόμησης για ένα τεχνητό σύνολο δεδομένων



Εικόνα 50 Καμπύλη ταξινόμησης για ένα τεχνητό σύνολο δεδομένων και πως διαφέρει ανά υποσύνολο του πραγματικού κόσμου.

Από τα παραπάνω σχήματα είναι ξεκάθαρο ότι ανάλογα με τα δεδομένα που θα αλλάξει κάθε μέθοδος μπορεί να αλλάξει σημαντικά το αποτέλεσμα του συστήματος που εκπαιδεύεται. Αν για παράδειγμα οποιαδήποτε αρχιτεκτονική δεν έχει δει πότε τα δεδομένα που σβήνει η μέθοδος 2 είναι αδύνατο να παράγει μια διαχωριστική όμοια με αυτή που παράγεται από τα αρχικά δεδομένα. Οπότε το συγκεκριμένο σύστημα είναι καταδικασμένο να αποτύχει ανεξάρτητα από την πολυπλοκότητα του ή την δομή του. Παρ' όλα αυτά, αν η μέθοδος επηρεάσει - σβήσει κάποια άλλα δεδομένα τότε η απόδοση του αναμένεται να είναι σίγουρα καλύτερη όπως φαίνεται από το τρίτο σχήμα. Οπότε για

να μην οδηγηθούμε σε εσφαλμένες εκτιμήσεις σχετικά με την απόδοση των μεθόδων ακολουθήσαμε την εξής πειραματική διαδικασία:

1. Κρατήσαμε ένα σταθερό σύνολο δεδομένων ελέγχου, στο οποίο θα εξεταστούν όλοι οι αλγόριθμοι
2. Για κάθε μέθοδο:
    a. Εφαρμόσαμε στα δεδομένα εκπαίδευσης
    b. Εκπαιδεύσαμε κάθε έναν από τους αλγορίθμους ταξινόμησης
    c. Ελέγξαμε την απόδοση των αλγορίθμων στο σύνολο ελέγχου
    d. Επαναλάβαμε την παραπάνω διαδικασία 20 φορές
3. Κρατήσαμε σαν μετρική για κάθε μέθοδο και κάθε αρχιτεκτονική τον μέσο όρο της απόδοσης για κάθε μια από τις 20 επαναλήψεις.

Στην βιβλιογραφία έχουν αναφερθεί ανάλογες διαδικασίες σε αντίστοιχα προβλήματα με μικρότερο αριθμό επαναλήψεων από 20 αλλά εδώ επιλέχθηκε ώστε να μειώσουμε όσο το δυνατόν περισσότερο την επίδραση της τύχης από τα πειράματά μας. Παρακάτω παρουσιάζονται τα αποτελέσματα των μεθόδων αφού επαναλάβαμε την διαδικασία που περιγράφεται παραπάνω για κάθε μια μέθοδο και κάθε μια διαφορετική αρχιτεκτονική.

Μέθοδος/Μετρική	Accuracy	Precision	Recall	F1
Αρχικό Σύνολο Δεδομένων	99.83%	97.22%	41.95%	58.61%
Υπερδειγματοληψία	96%	4.8%	91.1%	9.11%
SMOTE	96.4%	4.22%	91.11%	8.06%
Υποδειγματοληψία	88.5%	1.16%	77.77%	2.28%
Υπερ-Υποδειγματοληψία	98.7%	9.34%	64.51%	16.31%
Κανονικοποιήση	99.83%	59.66%	77.12%	67.27%

6. Συμπεράσματα

Στην παρούσα εργασία μελετήθηκε το πρόβλημα της ανίχνευσης απάτης σε τραπεζικές συναλλαγές. Καθημερινά λαμβάνουν χώρα εκατομμύρια συναλλαγές, ελάχιστες από αυτές όμως αποτελούν απάτη. Συνεπώς, το σύνολο των δεδομένων είναι ακραία ανισοκατανεμημένο. Συγκεκριμένα, από το σύνολο των συναλλαγών μόνο περίπου το 0.172% αποτελούν απάτη. Αυτό καθιστά εξαιρετικά δύσκολη την εκπαίδευση ενός μοντέλου μηχανικής μάθησης για την αναγνώριση μη έγκυρων συναλλαγών και αυτό γιατί αυτά κατά κύριο λόγο θα έχει μάθει να προβλέπει όλες τις συναλλαγές ως έγκυρες μιας και αυτό έχει δει σε μεγαλύτερο βαθμό. Σκοπός της παρούσας εργασίας είναι η μελέτη διάφορων μεθόδων για τη μείωση της επίδρασης που έχει αυτή η ανισοκατανομή στο μοντέλο εκμάθησης. Συγκεκριμένα μελετήθηκαν οι μέθοδοι της υποδειγματοληψίας, υπερδειγματοληψίας, smote, υπερ και υποδειγματοληψίας καθώς και αυτή της κανονικοποίησης. Μεταξύ αυτών, η τελευταία έδειξε τα καλύτερα αποτελέσματα. Σε αυτό συνέβαλε το γεγονός ότι μέσω της μεθόδου της κανονικοποίησης το μοντέλο είναι σε θέση να γενικεύσει καλύτερα. Αυτό είναι ορατό από το γεγονός ότι το τελικό μοντέλο καταφέρνει καλύτερες και τις 2 κλάσεις εξόδου καθώς επίσης έχει και ισορροπημένα αποτελέσματα μεταξύ της ακρίβειας και της ανάκλησης. Αντίθετα, οι προηγούμενες μέθοδοι κατάφερναν είτε καλά αποτελέσματα σε μια μόνο κλάση είτε υψηλή ακρίβεια ή υψηλή ανάκληση, με αποτέλεσμα ο μέσος του να είναι μικρός. Καταλήξαμε σε κάτι αναμενόμενο, αφού από την υφιστάμενη βιβλιογραφία προτείνεται επίσης ως πιο αποτελεσματική τεχνική για τη γενίκευση των νευρωνικών δικτύων η μέθοδος της κανονικοποιήσης. Από την άλλη η μέθοδος της υπερδειγματοληψίας οδηγούσε όλα τα μοντέλα σε υπερεκπαίδευση, ενώ η μέθοδος της υποδειγματοληψίας μείωνε σημαντικά την ποιότητα του συνόλου εκπαίδευσης. Ο συνδυασμός των μεθόδων επίσης δεν κατάφερε σημαντικά αποτελέσματα μιας και πάλι οδηγούσε τα συστήματα στην υπερεκπαίδευση. Τέλος η τεχνική smote, είχε αρκετά κοντινά αποτελέσματα με την υπερδειγματοληψία.

7. Προτάσεις

Από την παραπάνω ανάλυση φάνηκε ότι η μέθοδος της κανονικοποίησης είναι η καταλληλότερη για την εκπαίδευση συστημάτων μηχανικής μάθησης με ανισοκατανεμημένα σύνολα, μεταξύ του συνόλου των μεθόδων που εξετάστηκαν. Η ανάλυση αυτή έγινε στο σύνολο δεδομένων Credit Card Fraud Detection και εξετάστηκε σε 4 διαφορετικές αρχιτεκτονικές τέτοιων συστημάτων. Παρόλο τα ξεκάθαρα καλύτερα αποτελέσματα που επιφέρει η μέθοδος της κανονικοποίησης πράγμα που συμφωνεί με τα σχετικά αποτελέσματα της βιβλιογραφίας, είναι απαραίτητο να μελετηθεί το συγκεκριμένο πρόβλημα σε περισσότερα σύνολα δεδομένων. Επίσης είναι σημαντικό να δοκιμαστεί κατά πόσο τα αποτελέσματα συμφωνούν και σε περισσότερα συστήματα διαφορετικού μεγέθους. Με αυτόν τον τρόπο μπορούν να εξαχθούν συμπεράσματα σχετικά με το πόσο κάθε σύστημα μπορεί να επηρεαστεί από τα προβλήματα κάθε μεθόδου. Για παράδειγμα ένα μικρό νευρωνικό δίκτυο το οποίο δεν έχει την δυνατότητα να αποστηθίσει όλο το σύνολο δεδομένων μπορεί να επωφεληθεί περισσότερο με τη μέθοδο της υπερδειγματοληψίας και όχι από αυτή της κανονικοποιήσης. Τέλος είναι σημαντικό να εξεταστούν και περισσότερες και πιο σύγχρονες μέθοδοι για την επίλυση των προβλημάτων που επιφέρει η ανισοκατανομή των δεδομένων εκπαίδευσης σε ένα σύστημα μηχανικής μάθησης, όπως τεχνικές σύνθεσης δεδομένων με την χρήση δημιουργικών νευρωνικών δικτύων.

Βιβλιογραφία

[1] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[2] Kingma, Durk P., and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions." *Advances in neural information processing systems* 31 (2018).

[3] Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." *Journal of Network and Computer Applications* 68 (2016): 90-113.

[4] Cao, Kaidi, et al. "Learning imbalanced datasets with label-distribution-aware margin loss." *Advances in neural information processing systems* 32 (2019).

[5] Tanha, Jafar, et al. "Boosting methods for multi-class imbalanced data classification: an experimental review." *Journal of Big Data* 7.1 (2020): 1-47.

[6] Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning." The Journal of Machine Learning Research 18.1 (2017): 559-563.

[7] Paul, Liton Chandra, Abdulla Al Suman, and Nahid Sultan. "Methodological analysis of principal component analysis (PCA) method." *International Journal of Computational Engineering & Management* 16.2 (2013): 32-38.

[8] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry* 27.2 (2015): 130.

[9] Adnan, Md Nasim, and Md Zahidul Islam. "Forex++: A new framework for knowledge discovery from decision forests." *Australasian Journal of Information Systems* 21 (2017).

[10] Liu, Han, and Alexander Gegov. "Induction of modular classification rules by information entropy based rule generation." *Innovative Issues in Intelligent Systems*. Springer, Cham, 2016. 217-230.

[11] Zimmerman, Richard K., et al. "Classification and Regression Tree (CART) analysis to predict influenza in primary care patients." BMC infectious diseases 16.1 (2016): 1-11.

[12] Atti, Astri, and D. Dodo. "Chi-Square Automatic Interaction Detection (Chaid) Analysis for Home Quality Status Segmentation." *American Journal of Engineering Research* 7.4 (2018): 183-188.

[13] Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. "ctree: Conditional inference trees." *The comprehensive R archive network* 8 (2015).

[14] Ramya, K., Yuvaraja Teekaraman, and KA Ramesh Kumar. "Fuzzy-based energy management system with decision tree algorithm for power security system." *International Journal of Computational Intelligence Systems* 12.2 (2019): 1173-1178.

[15] Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." *Neural computation* 31.7 (2019): 1235-1270.

[16] Tsai, Chih-Fong, et al. "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection." *Information Sciences* 477 (2019): 47-54.

[17] Kobyzev, Ivan, Simon JD Prince, and Marcus A. Brubaker. "Normalizing flows: An introduction and review of current methods." *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020): 3964-3979.

[18] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." *Journal of big data* 6.1 (2019): 1-48.

[19] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

[20] Mitchell, T. (1997). *Machine Learning*, McGraw Hill, *Machine Learning*, McGraw Hill, p.2

Ευρετήριο εικόνων