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

Εμπλουτισμός δεδομένων μέσω δημιουργικών νευρωνικών δικτύων για την ανίχνευση κακόβουλων ηλεκτρονικών συναλλαγών

ΧΡΥΣΟΥΛΑ ΦΙΛΑΝΔΡΙΑΝΟΥ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

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

ΟΚΤΩΒΡΙΟΣ 2022

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν στην εκπόνηση της παρούσας διπλωματικής εργασίας:

Τον επιβλέποντα καθηγητή μου κ. Νικόλαο Δουλάμη για την εμπιστοσύνη που μου έδειξε και μου ανέθεσε το συγκεκριμένο θέμα καθώς και για τη βοήθειά του, λύνοντας κάθε απορία με αμεσότητα και συμβάλλοντας στην ολοκλήρωση της παρούσης εργασίας.

Τους υποψήφιους διδάκτορές του κ. Σταύρο Συκιώτη και κ. Ιάσονα Κατσαμένη, οι οποίοι επίσης συνέβαλαν με τη βοήθειά τους στην ολοκλήρωση της εργασίας μου.

Την οικογένειά μου και τους φίλους μου για όλη την υποστήριξη και βοήθεια που μου παρείχαν.

Πίνακας Περιεχομένων

Περιεχόμενα

Περίληψη

Τα τελευταία χρόνια με την ραγδαία αύξηση της τεχνολογίας όλο και περισσότερες υπηρεσίες ψηφιοποιούνται. Ένας από τους σημαντικότερους τομείς που αλλάζει εποχή είναι αυτός των τραπεζικών συστημάτων, καθώς όλο και περισσότερες τράπεζες ψηφιοποιούν τις υπηρεσίες τους και δίνουν την δυνατότητα στους χρήστες να κάνουν γρήγορα και εύκολα συναλλαγές μέσω ηλεκτρονικών συσκευών. Σε αυτή όμως την προσπάθεια ελλοχεύουν και κίνδυνοι. Ένας εκ των σημαντικότερων είναι αυτός των κακόβουλων συναλλαγών, δηλαδή συναλλαγές οι οποίες δεν έγιναν από τους ίδιους τους χρήστες αλλά από τρίτους, με σκοπό την οικονομική εκμετάλλευσή τους. Στην παρούσα εργασία μελετήθηκε το συγκεκριμένο πρόβλημα καθώς και η δυνατότητα ανάπτυξης ενός συστήματος για την αυτόματη αναγνώριση αυτών των κακόβουλων συναλλαγών. Κατά τη διάρκεια της μελέτης του προβλήματος αυτού αναδύθηκε ένα έμφυτο ζήτημα το οποίο είναι η ανισοκατανομή των δεδομένων. Καθημερινά γίνονται πάρα πολλές έγκυρες συναλλαγές, σε αντίθεση με τις απάτες που αποτελούν μικρό μερίδιο. Για τη μελέτη του αρχικού προβλήματος αναλύθηκαν και οι τρόποι με τους οποίους μπορεί να μειωθούν οι συνέπειες που επιφέρουν τα ανισοκατανεμημένα δεδομένα στην εκπαίδευση ενός τέτοιου συστήματος. Αναλύθηκαν πέντε παραδοσιακές τεχνικές, χωρίς σύνθεση νέων δεδομένων, για τη μείωση των φαινομένων αυτών και δοκιμάστηκαν σε πέντε διαφορετικές αρχιτεκτονικές μηχανικής μάθησης. Μέσω αυτής της ανάλυσης αναδείχθηκε το γεγονός ότι για να λυθεί ουσιαστικά το πρόβλημα της ανισοκατανομής των δεδομένων απαιτούνται περισσότερα δεδομένα από την κλάση μειονότητας. Για αυτόν τον λόγο μελετήθηκε η περιοχή των δημιουργικών νευρωνικών δικτύων και κατασκευάστηκε ένα state-of-the-art τέτοιο σύστημα για τη σύνθεση δεδομένων. Μελετήθηκε ο τρόπος λειτουργίας πολλαπλών τέτοιων αλγορίθμων που έχουν προταθεί στην βιβλιογραφία καθώς και τα θετικά και τα αρνητικά που επιφέρει ο κάθε ένας. Σκοπός των αλγορίθμων αυτών είναι η δημιουργία στιγμιότυπων για την εξομάλυνση της κατανομής των κλάσεων. Μέσω των τεχνικών αυτών καταφέραμε να επιτύχουμε καλύτερά αποτελέσματα ταξινόμησης συγκριτικά με τις παραδοσιακές τεχνικές.

Λέξεις κλειδιά: δημιουργικά νευρωνικά δίκτυα, μηχανική μάθηση, ανίχνευση απάτης, ανισοκατανεμημένα δεδομένα, εμπλουτισμός δεδομένων

Abstract

In recent years, more and more services have been digitalized due to the continuous technological change and evolution. One of the most important sectors that is changing the era is that of banking systems, as more and more banks are enabling users to make transactions quickly and easily through electronic devices. However, this rapid technological evolution has also its drawbacks. One of the most important is that of malicious transactions, i.e. transactions which were not made by the users themselves but by third parties, with the aim of their financial exploitation. In this thesis, we will attempt to study this specific problem as well as the possibility of developing a system for the automatic identification of these malicious transactions. However, in our attempt to study of this problem we had to deal with an inherent issue that emerged, the uneven distribution of the data. There are a lot of valid transactions taking place every day, in contrast with the malicious ones (also called frauds) which make up a small share. The ways in which the consequences of unevenly distributed data can be reduced in the training of such a system have been also analyzed. Five traditional methods, without new data synthesis, have been analyzed and tested on different machine learning architectures in order to reduce the previously mentioned problem.

Throughout this analysis it was made very clear that, in order to effectively solve the problem of unequal distribution of data, more data from the minority class is required. For this reason, the area of creative neural networks was studied and a state-of-the-art system for data synthesis was constructed. The mode of operation of several such algorithms that have been proposed in the literature was studied, as well as the advantages and disadvantages of each one. The purpose of these algorithms is to create snapshots in order to make the distribution of classes more even. Through these techniques we managed to achieve better classification results compared to traditional ones.

Key words: generative adversarial networks, machine learning, fraud detection, imbalanced dataset, data enrichment

1. Εισαγωγή

Μερικά από τα προβλήματα του πραγματικού κόσμου είναι εύκολο να λυθούν ενώ κάποια άλλα όχι. Επίσης, για πολλά προβλήματα υπάρχουν καλά ορισμένες συναρτήσεις ή μέθοδοι επίλυσης οι οποίες είναι ικανές να υπολογίσουν τη λύση αυτών με αρκετά μεγάλη ακρίβεια. Τέτοιο πρόβλημα για παράδειγμα είναι ο υπολογισμός του χρόνου ταξιδιού ή της τελικής ταχύτητας ενός αντικειμένου αν αφεθεί από συγκεκριμένο ύψος. Παράδειγμα αυτής της κατηγορίας αποτελεί και ο υπολογισμός της ελάχιστης διαδρομής σε έναν γράφο από ένα σημείο εκκίνησης σε έναν κόμβο στόχο. Και στα 2 παραπάνω προβλήματα υπάρχουν συναρτήσεις ή αλγόριθμοι για την εύρεση των λύσεων με ελάχιστο ή και μηδενικό σφάλμα.

Παρόλα αυτά, υπάρχουν και προβλήματα για τα όποια δεν υπάρχουν ή δεν γίνεται να διατυπωθούν τέτοιες μέθοδοι επίλυσής τους. Παραδείγματος χάριν, το να εντοπίσουμε σε μια σκήνη αν υπάρχει ένα πορτοκάλι ή όχι. Για αυτό το πρόβλημα δεν έχει διατυπωθεί μέχρι στιγμής κάποιος αλγόριθμος ή κάποια φόρμουλα (π.χ. αν περισσότερα από 50 pixels έχουν χρώμα στην οικογένεια του πορτοκαλί τότε περιέχεται ειδάλλως όχι). Παρόλο που το συγκεκριμένο πρόβλημα είναι πολύ εύκολο και διαισθητικό για έναν άνθρωπο, εντούτοις καθίσταται πολύ δύσκολη η μαθηματική περιγραφή της λύσης του. Αυτό οφείλεται στο γεγονός ότι ο τρόπος που ο άνθρωπος μαθαίνει να λύνει αυτό το πρόβλημα δεν είναι μαθηματικός αλλά διαισθητικός, παρατηρώντας ουσιαστικά πολλές σκηνές-εικόνες που περιέχουν πορτοκάλι ενώ κάποιος άλλος άνθρωπος του το υποδεικνύει.

Ακολουθώντας αυτήν ακριβώς τη λογική λειτουργούν και οι αλγόριθμοι μηχανικής μάθησης. Δέχονται στην είσοδο ένα σύνολο από επισημειωμένα δεδομένα και προσαρμόζονται με τέτοιον τρόπο ώστε να μπορούν να προβλέπουν σωστά, στο σύνολο, αυτό το πρόβλημα. Στο παραπάνω παράδειγμα, αν θέλαμε ένα σύστημα να μάθει να ξεχωρίζει πότε μία εικόνα περιέχει ένα πορτοκάλι και πότε όχι, θα δίναμε ένα σύνολο από εικόνες που μερικές από αυτές θα περιείχαν ενώ άλλες όχι και το σύστημα θα έπρεπε να προσαρμοστεί με τέτοιον τρόπο ώστε να μπορεί να ξεχωρίζει σωστά τα δεδομένα αυτά. Το προαναφερθέν σύνολο ονομάζεται σύνολο εκπαίδευσης, ενώ το αντίστοιχο πρόβλημα ονομάζεται πρόβλημα ταξινόμησης επειδή το σύστημα καλείται να ξεχωρίζει τα δεδομένα μεταξύ και των κλάσεων. Εδώ να τονίσουμε ότι αυτό δεν είναι το μοναδικό πρόβλημα στο οποίο μπορούν να δώσουν λύση σήμερα τα συστήματα μηχανικής μάθησης, ωστόσο αναφέρεται εκτεταμένα γιατί αποτελεί το κύριο μέρος της παρούσας εργασίας.

Συνεπώς, για την επίλυση ενός προβλήματος ταξινόμησης το πρώτο βήμα αποτελεί η συλλογή ενός συνόλου δεδομένων με επισημειωμένα δεδομένα όλων των κλάσεων. Τι γίνεται όμως στις περιπτώσεις όπου το σύνολο των δεδομένων δεν κατανέμεται ομοιόμορφα; Το πρόβλημα εδώ μπορεί να γίνει σαφές ευκολότερα αν μελετηθεί ένα διαισθητικά πιο δύσκολο πρόβλημα από το

παραπάνω. Έστω ότι θέλουμε να κατασκευάσουμε ένα ιατρικό σύστημα το οποίο να μπορεί να διακρίνει σε μια ακτινογραφία ή τομογραφία αν περιέχεται κάποιος όγκος ή όχι. Το πρώτο βήμα θα ήταν η ανάπτυξη ενός συνόλου δεδομένων με εικόνες και ετικέτες. Η κατανομή των ετικετών στο σύνολο δεδομένων θα θέλαμε ιδανικά να είναι ομοιόμορφη δηλαδή να υπάρχει ίσος ή τουλάχιστον παρόμοιος αριθμός εικόνων μεταξύ των κλάσεων. Αν η συνθήκη αυτή δεν πληρείται, όπως είναι και το λογικό, η εκπαίδευση θα δυσκολέψει πάρα πολύ. Αντίστοιχο πρόβλημα θα υπήρχε και αν αντί για σύστημα θέλαμε να εκπαιδευτεί ένας άνθρωπος. Έστω ότι ένας νέος γιατρός θέλει να μάθει να ξεχωρίζει τέτοια δείγματα και στην πορεία της εμπειρίας του βλέπει μόνο εικόνες που δεν περιέχουν όγκο. Είναι δυνατόν υπό αυτές τις συνθήκες ο γιατρός να προβλέψει ότι μία εικόνα έχει όγκο; 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.).

Εφόσον λοιπόν το πρόβλημα αυτό δυσχεραίνει τόσο την εκπαίδευση ενός ανθρώπου, αναμένουμε τα αποτελέσματα στην εκπαίδευση ενός συστήματος μηχανικής μάθησης να είναι ακόμη χειρότερα. 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.

Σκοπός της παρούσας εργασίας είναι η μελέτη διαφόρων μεθόδων για τη βελτίωση της απόδοσης των συστημάτων τα οποία έχουν εκπαιδευτεί σαν ανισοκατανεμημένα δεδομένα. Συγκεκριμένα μελετώνται διάφορες τέτοιες τεχνικές και η επίδραση που επιφέρουν σε μοντέλα διαφόρων αρχιτεκτονικών για την καλύτερη δυνατή ανάλυση των αποτελεσμάτων.

2. Fraud Detection

Το πρόβλημα που θα μελετηθεί στην παρούσα εργασία είναι αυτό της αυτόματης εύρεσης απάτης για ένα τραπεζικό σύστημα [3]. Το συγκεκριμένο πρόβλημα επιλέχθηκε επειδή αποτελεί ένα από τα χαρακτηριστικά tasks στα οποία υπάρχει άνιση κατανομή των δεδομένων και είναι ένα σύνολο δεδομένων το όποιο έχει μελετηθεί εκτενώς στη βιβλιογραφία [3]. Καθημερινά σε ένα τραπεζικό σύστημα γίνονται εκατομμύρια έγκυρες συναλλαγές αλλά - ευτυχώς - ελάχιστες από αυτές αποτελούν απάτη. Οπότε όλα τα σύνολα δεδομένων που έχουν αναπτυχθεί για την αυτόματη αναγνώριση των μη έγκυρων συναλλαγών έχουν πάρα πολλά δεδομένα για έγκυρες συναλλαγές και ελάχιστα για μη.

Η σωστή εκπαίδευση ενός τέτοιου συστήματος αποτελεί ιδιαιτέρα σημαντική διαδικασία η όποια πρέπει να μελετηθεί διεξοδικά και σε βάθος ώστε να μπορούν να προβλεφθούν έγκυρα και έγκαιρα οι άπατες. Οπότε σε αυτό το πρόβλημα η απαλοιφή του προβλημάτων που επιφέρει η ανισοκατανομή των δεδομένων αποτελεί σημαντικό πεδίο που πρέπει να μελετηθεί.

Το παρόν πρόβλημα έχει επίσης επιπλέον ιδιαίτερα χαρακτηριστικά τα οποία πρέπει να λάβουμε υπόψη για τη σωστή μελέτη του. Αν υποθέσουμε ότι έχουμε ένα σύστημα το όποιο δεν ταξινομεί καμία συναλλαγή ως απάτη, τότε το σύστημα αυτό μπορούμε να πούμε με σιγουριά ότι δεν είναι χρήσιμο. Το πρόβλημα αυτό είναι αρκετά πιθανό να το έχουμε λόγο της δομής των δεδομένων. Από την άλλη αν είχαμε ένα σύστημα που έβγαζε όλες τις συναλλαγές ως άπατες και έπρεπε να κινηθεί μια διαδικασία για την εξακρίβωση τους, τότε ούτε αυτό θα ήταν τόσο χρήσιμο. Το δεύτερο ωστόσο στον τελικό χρήστη, θα είχε μια χρησιμότητα γιατί θα μπορούσε να τους σώσει από μια απάτη, παρόλα αυτά δεν θα ήταν καθόλου χρήσιμο στο τραπεζικό σύστημα. Έτσι οδηγούμαστε στο τελικό ερώτημα το όποιο είναι το εξής: Είναι προτιμότερο ένα σύστημα το όποιο βρίσκει σωστά πάντα όλες τις έγκυρες συναλλαγές και ας ταξινομεί μερικές άπατες ως έγκυρες συναλλαγές ή το αντίθετο;

Η απάντηση για έναν χρήστη είναι ξεκάθαρη, 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 όμως δεν λένε το ίδιο. Αν η ερώτηση διαπιστωθεί διαφορετικά σίγουρα η διαισθητική απάντηση θα είναι διαφορετική. Αν είχαμε να επιλέξουμε μεταξύ 2 συστημάτων όπου το ένα έχει 99.9% ακρίβεια και ενός άλλου με 50% ποιο θα διαλέγαμε; Την συγκεκριμένη ερώτηση, για το πρόβλημα που μελετάμε δεν μπορούμε να την απαντήσουμε γιατί μας αφορά ιδιαίτερα το πότε κάνει λάθος το σύστημα. Θα θέλαμε τα λάθη να γίνονται σε έγκυρες συναλλαγές οι όποιες ταξινομήθηκαν σαν άπατες και όχι το αντίθετο. Η γενική αρχή που θέλουμε να

πληροί το σύστημα μας είναι η εξής: Θέλουμε να αναπτύξουμε ένα σύστημα το όποιο να μην ταξινομεί λανθασμένα άπατες.

Αυτό είναι ένα χαρακτηριστικό το όποιο αποτέλεσε ιδιαιτέρα σημαντικό πυλώνα για τη μεθοδολογία που ακολουθήσαμε και έτσι οι περιορισμοί που εισήγαμε για να ικανοποιήσουμε αυτήν την συνθήκη αναλύονται εκτενέστερα παρακάτω.

Στην παρούσα εργασία αναλύθηκε το σύνολο δεδομένων Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine[1] το όποιο ανακτήθηκε μέσω του αποθετηρίου kaggle[2]. Το σύνολο περιέχει συναλλαγές μέσω πιστωτικής κάρτας οι οποίες έγιναν το Σεπτέμβριο του 2013 από ευρωπαίους κάτοχους. Στο σύνολο παρουσιάζονται συναλλαγές οι οποίες έγιναν μέσα σε δύο ημέρες και βρέθηκαν 492 μη- έγκυρες από τις συνολικές 284,807 συναλλαγές. Δηλαδή τα στιγμιότυπα της θετικής κλάσης - απάτη αποτελούν μόνο το 0.172% του συνόλου. Αυτό καθιστά το dataset ένα από τα πιο ανισοκατανεμημένα της περιοχής, γεγονός το όποιο θα αναδείξει εύκολα τα προβλήματα που αναλύθηκαν, ενώ παράλληλα θα δυσκολέψει κατά πολύ τη βελτίωση των αποτελεσμάτων με την χρήση των μεθόδων.

Για λόγους προστασίας των καταναλωτών τα δεδομένα που παρέχονται είναι ανώνυμα. Έτσι, για κάθε συναλλαγή παρέχεται ένα διάνυσμα 29 τιμών τα οποία έχουν εξαχθεί μέσω της μεθόδου PCA μαζί με το ποσό και την χρονική στιγμή της συναλλαγής. Επίσης, για κάθε συναλλαγή παρέχεται και μία ετικέτα η οποία υποδηλώνει την εγκυρότητα της, με τιμές μηδέν για έγκυρη και ένα για απάτη.

Παρακάτω παρουσιάζονται τα πεδία του συνόλου δεδομένων καθώς και 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

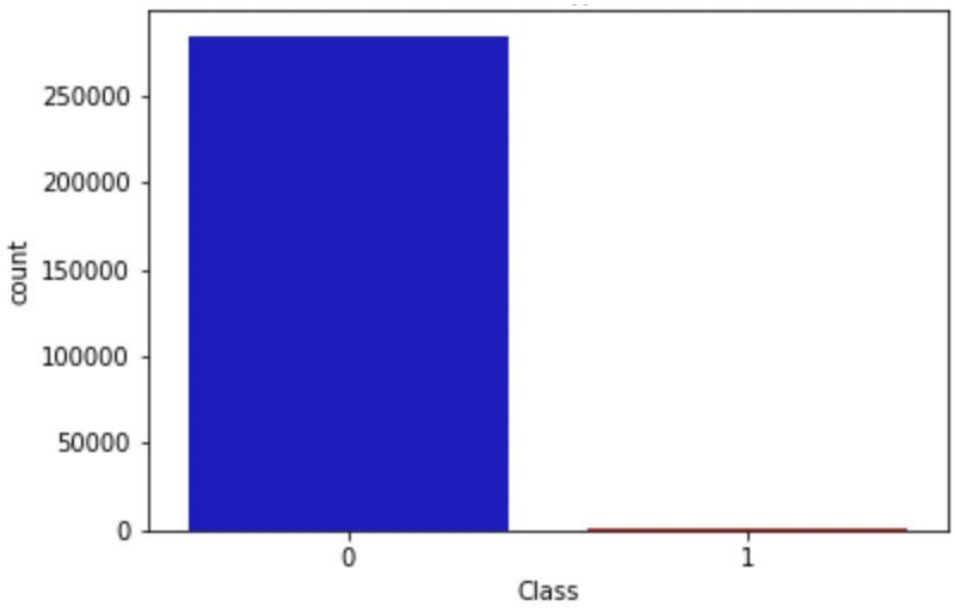
V7	V8	V9	V10	V11	V12	V13	V14
0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.991390	-0.311169
-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.489095	-0.143772
0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.717293	-0.165946
0.237609	0.377436	-1.387024	-0.054952	-0.226487	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

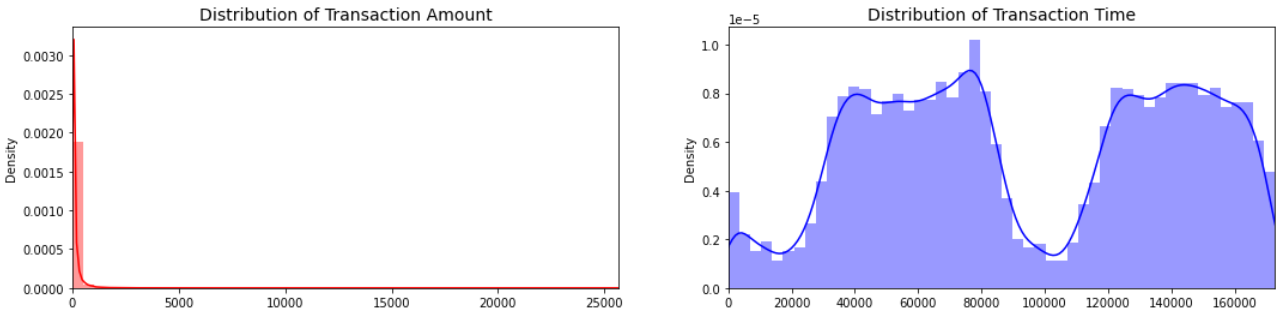
Παρακάτω θα γίνει μία ανασκόπηση των δεδομένων τα οποία δεν έχουν παραχθεί μέσω της PCA μεθόδου [7]. Στο πρώτο διάγραμμα φαίνεται ο αριθμός του στιγμιότυπων τα οποία ανήκουν σε κάθε κλάση. Οπότε πέρα από τα στατιστικά που παρουσιάστηκαν παραπάνω στο διάγραμμα φαίνεται και οπτικά πόσο εντονότερη είναι η ύπαρξη της κλάσης 0 - μη απάτη.



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

Από το παραπάνω διάγραμμα φαίνεται και γραφικά το πρόβλημα τις ανισοκατανομής των δεδομένων.

Στα επόμενα δύο διαγράμματα παρουσιάζονται οι κατανομές των μεταβλητών amount και time στα δεδομένα. Όπως είναι λογικό, η κατανομή του χρόνου είναι περίπου τυχαία αφού δεν υπάρχει κάποια συσχέτιση μεταξύ των χρόνων που γίνονται οι συναλλαγές, ενώ τα ποσά (amount) είναι τα περισσότερα μικρά της τάξεως από 500 μέχρι 1000.



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

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. Παρόλα αυτά, η ανισοκατανομή των δεδομένων δεν επιφέρει μόνο προβλήματα κατά την εκπαίδευση.

Σοβαρά προβλήματα δημιουργούνται και κατά την αξιολόγηση των συστημάτων. Για παράδειγμα, όπως προαναφέρθηκε, το ποσοστό της θετικής κλάσης επί του συνόλου των δεδομένων είναι 0.172%. Οπότε, αν είχαμε ένα σύστημα το οποίο απαντούσε πάντα την κλάση μηδέν τότε αυτόματα το ποσοστό επιτυχίας του θα ήταν 100 - 0.172 % = 99.828%, ενώ πρακτικά δεν έχει μάθει τίποτα. Αυτό αποτελεί σοβαρό πρόβλημα μιας και αν δεν κάνουμε σωστή αξιολόγηση των μοντέλων δεν θα μπορέσει να καταγραφεί η επίδραση των μεθόδων που μελετάμε. Επίσης, αν ένα τέτοιο μοντέλο περάσει την παραγωγή, βασισμένοι σε δεδομένα που προκύπτουν από λανθασμένη αξιολόγηση, τότε αποτελεί σοβαρό κίνδυνο.

Το ποσοστό επιτυχίας ταξινόμησης ορίζεται ως το πηλίκο της διαίρεσης του αριθμού των σωστών προβλέψεων προς το πλήθος των συνολικών δειγμάτων. Η μετρική αυτή είναι πολύ καλή για ισοκατανεμημένα δεδομένα αλλά σίγουρα δεν είναι κατάλληλη για ανομοιόμορφο δεδομένα. Γι' αυτό στην βιβλιογραφία προτείνονται άλλες μετρικές όπως το **precision**, το **recall** και η **f1**, οι οποίες παρουσιάζονται εκτενέστερα παρακάτω.

Μετρικές F1, recall, precision

Για την καλύτερη δυνατή εξήγηση των μετρικών αυτών πρώτα θα εισαχθεί ο πίνακας σύγχυσης (confusion matrix), μαζί με τα πεδία του. Ένας πίνακας σύγχυσης είναι μια σύνοψη των αποτελεσμάτων πρόβλεψης σε ένα πρόβλημα ταξινόμησης. Ο αριθμός των σωστών και εσφαλμένων προβλέψεων συνοψίζεται με τιμές μέτρησης και αναλύεται ανά κατηγορία. Ο πίνακας σύγχυσης πρακτικά υποδεικνύει τα σημεία στα οποία ένα μοντέλο κάνει λανθασμένες προβλέψεις και δίνει πληροφορίες όχι μόνο για τα σφάλματα αλλά και για τους τύπους αυτών. Οι τιμές του πίνακα αυτού παρουσιάζονται παρακάτω:

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

Εικόνα 5 Η τυπική μορφή του πίνακα σύγχυσης

Ουσιαστικά χωρίζουμε τον πίνακα σε στήλες και γραμμές όπου το negative, positive στις στήλες αφορά την πραγματική τιμή των δειγμάτων ενώ στις γραμμές αυτών που έγιναν predicted από το μοντέλο. Βάσει αυτών ορίζονται 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 - έγκυρες αλλαγές. Ο αριθμός αυτός θέλουμε να είναι όσο το δυνατόν πιο κοντά στο 0.

- **True Positive**: Ο αριθμός των δειγμάτων που ήταν θετικά και προβλέφθηκαν σωστά. Στην παρούσα περίπτωση είναι ο αριθμός των συναλλαγών που ήταν απάτη και ο αλγόριθμος κατάφερε να τις εντοπίσει. Οπότε θέλουμε ο αριθμός αυτός να είναι όσο το δυνατόν μεγαλύτερος, να εντοπίζονται όσο το δυνατόν περισσότερες άπατες.

- **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.

Με βάση αυτόν τον πίνακα μπορούμε εύκολα να δούμε σε ποια σημεία το σύστημα αποτυγχάνει και να εστιάσουμε σε αυτά. Σε συνέχεια αυτής της μεθοδολογίας παρακάτω εισάγονται 3 μετρικές που χρησιμοποιούν τα πεδία του παραπάνω πίνακα και μέσω αυτών μπορούν να βγουν συμπεράσματα για την λειτουργία ενός μοντέλου ταξινόμησης.

- **Ακρίβεια (Precision)**: Ορίζεται ως ο λόγος των δειγμάτων που είναι θετικά και προβλέφθηκαν σωστά ως προς το πλήθος δειγμάτων που προβλέφθηκαν ως θετικά. Παρακάτω παρουσιάζεται η μαθηματική έκφραση της σχέσης αυτής:

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

Εικόνα 6 Η μαθηματική σχέση που περιγράφει την ακρίβεια

Πρακτικά η μετρική αυτή μας δείχνει πόσα δείγματα που ο αλγόριθμος πρόβλεψε ως θετικά είναι σωστά. Η ακρίβεια είναι καλή μετρική όταν το κόστος των ψευδώς θετικών (true positives) είναι υψηλό. Ένα τέτοιο παράδειγμα είναι ο εντοπισμός spam e-mail. Σε αυτό το πρόβλημα ένα ψευδώς θετικό δείγμα σημαίνει ότι δεν είναι spam αλλά εμείς το κατατάξαμε ως τέτοιο. Αυτό το λάθος είναι ιδιαίτερα σημαντικό, γιατί ο χρήστης μπορεί να χάσει σημαντική πληροφορία. 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 προς το πλήθος των θετικών δειγμάτων.

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

Εικόνα 7 Η μαθηματική σχέση που περιγράφει την ανάκληση

Πρακτικά η μετρική αυτή μας δείχνει πόσα από τα θετικά δείγματα ο αλγόριθμος τα βρήκε σωστά. Η ανάκληση είναι σημαντική μετρική όταν το κόστος των ψευδώς αρνητικών είναι υψηλό. Ένα τέτοιο πρόβλημα είναι αυτό που μελετάμε στην παρούσα εργασία. Σε ένα σύστημα ανίχνευσης απάτης η ταξινόμηση ενός θετικού δείγματος, δηλαδή μιας κακόβουλης συναλλαγής, ως έγκυρης μπορεί να επιφέρει σημαντικές συνέπειες στον χρήστη. Το ίδιο ισχύει και σε ένα ιατρικό σύστημα το οποίο κάνει αυτόματη διάγνωση ασθενειών.

Από τις παραπάνω μετρικές βλέπουμε ότι αν έχουμε έναν αλγόριθμο ο οποίος προβλέπει περισσότερα δείγματα ως θετικά, τότε επειδή μέσα σε αυτά θα προβλέπει και περισσότερα

true positives η τιμή της ανάκλησης θα αυξηθεί. Παρόλα αυτά η τιμή της ακρίβειας θα μειωθεί γιατί θα αυξηθεί ο παρονομαστής. Οπότε μεταξύ αυτών των μετρικών θέλουμε να υπάρχει μία ισορροπία. Γι' αυτό το λόγο ορίζεται η F1, για την οποία ισχύει:

- **F1**: ορίζεται ως ο σταθμισμένος μέσος μεταξύ της ακρίβειας και της ανάκλησης.

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

Εικόνα 8 Η μαθηματική σχέση που περιγράφει την μετρική F1, η όποια ουσιαστικά αποτελεί τον σταθμισμένο μέσο της ακρίβειας και της ανάκλησης.

Η συνάρτηση αυτή χρησιμοποιείται όταν θέλουμε να κρατήσουμε μία ισορροπία μεταξύ της ακρίβειας και της ανάκλησης. Εξαιτίας του γεγονότος ότι το accuracy δεν αποτελεί καλή μετρική για την αξιολόγηση συστημάτων εκπαιδευμένων σε μη ισορροπημένα δεδομένα, μπορεί να αντικατασταθεί από την F1.

Πρακτικά οι τιμές των μετρικών αυτών και ο τρόπος με τον όποιον αυτές μεταφράζονται είναι ο εξής:

- υψηλή ανάκληση + υψηλή ακρίβεια: η κατηγορία αντιμετωπίζεται τέλεια από το μοντέλο
- χαμηλή ανάκληση + υψηλή ακρίβεια: το μοντέλο δεν μπορεί να εντοπίσει καλά την κατηγορία, αλλά είναι πολύ αξιόπιστο όταν το κάνει
- υψηλή ανάκληση + χαμηλή ακρίβεια: η κατηγορία είναι καλά ανιχνευμένη, αλλά το μοντέλο περιλαμβάνει επίσης σημεία άλλων κατηγοριών σε αυτήν
- χαμηλή ανάκληση + χαμηλή ακρίβεια: η κατηγορία δεν αντιμετωπίζεται καλά από το μοντέλο

3. Μοντέλα Μηχανικής Μάθησης

3.1. Decision Tree

Το Decision Tree Learning είναι μια εποπτευόμενη μαθησιακή προσέγγιση που χρησιμοποιείται στη στατιστική, την εξόρυξη δεδομένων και τη μηχανική μάθηση. Σε αυτόν τον φορμαλισμό, ένα δέντρο απόφασης ταξινόμησης ή παλινδρόμησης χρησιμοποιείται ως μοντέλο πρόβλεψης για την εξαγωγή συμπερασμάτων σχετικά με ένα σύνολο παρατηρήσεων.

Τα μοντέλα δέντρων όπου η μεταβλητή στόχος μπορεί να λάβει ένα διακριτό σύνολο τιμών ονομάζονται δέντρα ταξινόμησης. 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 απόφασης όπου η μεταβλητή στόχος μπορεί να λάβει συνεχείς τιμές (συνήθως πραγματικούς αριθμούς) ονομάζονται δέντρα παλινδρόμησης.

Τα δέντρα αποφάσεων είναι από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης, δεδομένης της ευκρίνειας και της απλότητάς τους [8]. Με τον όρο ευκρίνεια εννοείται ότι αυτή η οικογένεια μοντέλων είναι διαφανής, δηλαδή μπορεί να μας επιστρέψει τους κανόνες με τους οποίους αυτά λειτουργούν, σε αντίθεση για παράδειγμα με ένα βαθύ νευρωνικό δίκτυο (που παρουσιάζεται παρακάτω) από το όποιο δεν μπορούμε να εξάγουμε τον τρόπο λειτουργίας του. Για αυτό και αυτοί οι αλγόριθμοι χρησιμοποιούνται εκτενώς 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. Στην εξόρυξη δεδομένων, 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).

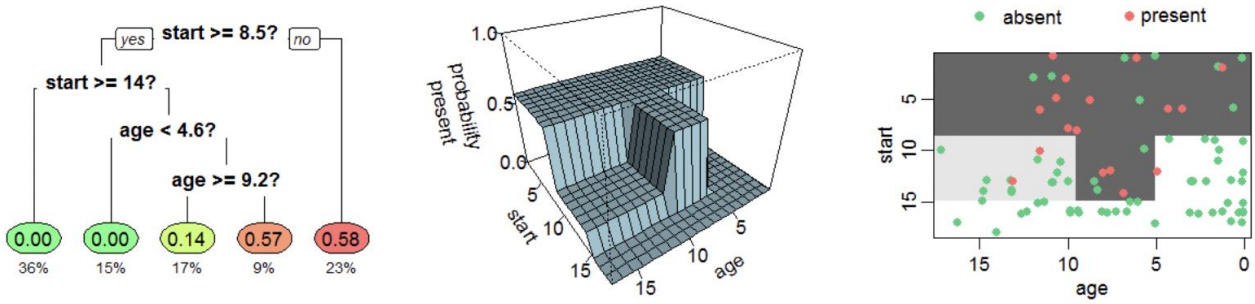
Η εκμάθηση του δέντρου αποφάσεων είναι μια μέθοδος που χρησιμοποιείται συνήθως στην εξόρυξη δεδομένων [9]. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που να προβλέπει την τιμή μιας μεταβλητής στόχου με βάση πολλές μεταβλητές εισόδου.

Ένα δέντρο αποφάσεων είναι μια απλή αναπαράσταση για την ταξινόμηση παραδειγμάτων. Για αυτήν την ενότητα, ας υποθέσουμε ότι όλα τα χαρακτηριστικά εισόδου έχουν πεπερασμένα διακριτά χαρακτηριστικά εισόδου και υπάρχει ένα μοναδικό χαρακτηριστικό στόχου που ονομάζεται "κλάση της ταξινόμησης". Ένα δέντρο αποφάσεων ή ένα δέντρο ταξινόμησης είναι ένα δέντρο στο οποίο 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, υποδηλώνοντας ότι το σύνολο δεδομένων έχει ταξινομηθεί από το δέντρο είτε σε μια συγκεκριμένη κατηγορία είτε σε μια συγκεκριμένη κατανομή πιθανοτήτων (η οποία στρέφεται προς ορισμένα υποσύνολα κλάσεων).

Ένα δέντρο χτίζεται με τον διαχωρισμό του συνόλου των χαρακτηριστικών εισόδου, που αποτελεί τον κόμβο ρίζας του δέντρου, σε υποσύνολα (τα οποία αποτελούν τα διαδοχικά παιδιά). Ο διαχωρισμός βασίζεται σε ένα σύνολο κανόνων διαχωρισμού που βασίζεται σε χαρακτηριστικά ταξινόμησης. Αυτή η διαδικασία επαναλαμβάνεται σε κάθε παραγόμενο υποσύνολο με αναδρομικό τρόπο που ονομάζεται αναδρομική κατάτμηση. 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 9 Οι διαφορετικοί τύποι με τους οποίους 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 προβλεπόμενο αποτέλεσμα μπορεί να θεωρηθεί 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 όποιες είναι διακεκριμένες.

Ο όρος ανάλυση δέντρου ταξινόμησης και παλινδρόμησης (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].

Μια ειδική περίπτωση ενός δέντρου αποφάσεων είναι μια λίστα αποφάσεων, που είναι ένα δέντρο απόφασης μονής όψης, έτσι ώστε κάθε εσωτερικός κόμβος έχει ακριβώς 1 κόμβο φύλλου και ακριβώς 1 εσωτερικό κόμβο ως παιδί (εκτός από τον πιο κάτω κόμβο, του οποίου μοναχοπαίδι είναι ένας μονόφυλλος κόμβος). Αν και λιγότερο εκφραστικές, οι λίστες αποφάσεων είναι αναμφισβήτητα πιο κατανοητές από τα δέντρα γενικών αποφάσεων λόγω της πρόσθετης αραιότητας, και επειδή επιτρέπουν την επιβολή μη άπληστων μεθόδων μάθησης και μονοτονικών περιορισμών.

Οι αξιοσημείωτοι αλγόριθμοι δέντρων αποφάσεων περιλαμβάνουν:

- ID3 (Iterative Dichotomiser 3)
- C4.5 (διάδοχος του ID3)
- CART (Classification And Regression Tree) Δένδρο 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 εκπαίδευσης.

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.

3.2. 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 όπως παρουσιάστηκε παραπάνω στον svm είναι εκπαιδεύσιμη, δηλαδή μπορεί να προσαρμόζεται στα δεδομένα. Η προσαρμογή αυτή στην ουσία αποτελεί την εκπαίδευση του αλγόριθμου και έγκειται ουσιαστικά στην ελαχιστοποίηση ενός κριτηρίου πχ της απόστασης από τα δεδομένα. Παρόλα αυτά η εσωτερική δομή κάθε αλγορίθμου, ενώ έχει κοινό στόχο, μπορεί να διαφέρει σημαντικά.

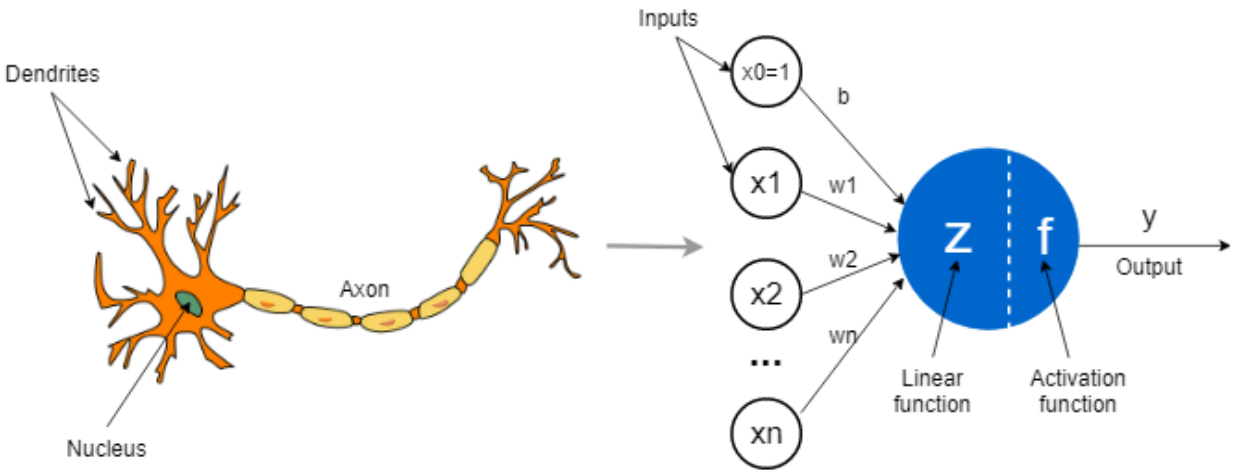
Ένα νευρωνικό δίκτυο είναι ουσιαστικά ένας αλγόριθμος μηχανικής μάθησης ο οποίος αποτελείται από νευρώνες οι οποίοι μπορούν να προσαρμόζονται – εκπαιδεύονται. Υπάρχουν διαφόρων τύπων νευρώνες οι οποίοι μπορούν να χρησιμοποιηθούν. Κάθε ένας από αυτούς μπορεί να

επιφέρει καλύτερες επιδόσεις ανάλογα με το πρόβλημα. Για παράδειγμα, ένα μοντέλο τύπου LSTM [15] είναι πολύ αποδοτικό στην επίλυση προβλημάτων χρονοσειρών όπως ανάλυση κειμένων, μετοχών κ.α. Αντίθετα ένα συνελικτικό δίκτυο έχει πολύ καλά αποτελέσματα στην ανάλυση 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.

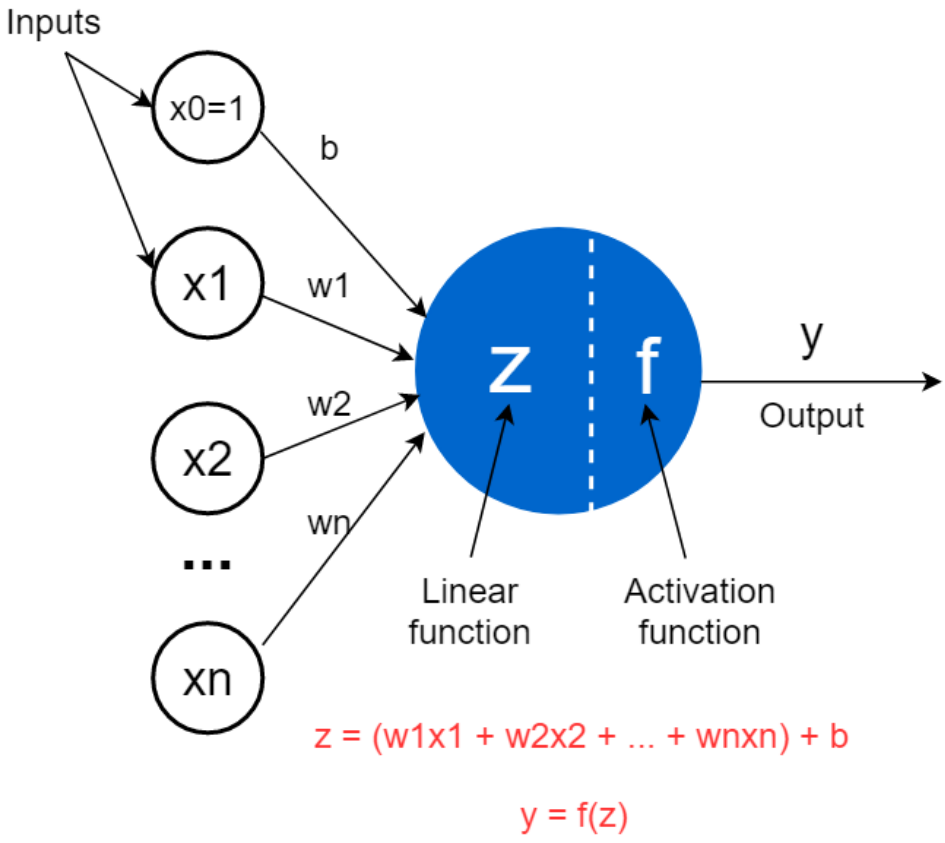
Αξίζει να σημειωθεί πώς οι τεχνητοί νευρώνες (perceptrons) εμπνέονται από βιολογικούς νευρώνες. Μπορείτε να θεωρήσετε έναν τεχνητό νευρώνα ως ένα μαθηματικό μοντέλο εμπνευσμένο από έναν βιολογικό νευρώνα. Στο παρακάτω σχήμα φαίνεται η θεωρικά αποδεκτή εικόνα ενός βιολογικού νευρώνα και δίπλα η μαθηματική μοντελοποίησή του που χρησιμοποιείται από τα νευρωνικά δίκτυα.



Εικόνα 10 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.

Ένας βιολογικός νευρώνας λαμβάνει τα σήματα εισόδου του από άλλους νευρώνες μέσω δενδριτών (μικρών ινών). Ομοίως, ένα perceptron λαμβάνει τα δεδομένα του από άλλα perceptron μέσω νευρώνων εισόδου που λαμβάνουν αριθμούς. Τα σημεία σύνδεσης μεταξύ των δενδριτών και των βιολογικών νευρώνων ονομάζονται συνάψεις. Ομοίως, οι συνδέσεις μεταξύ των εισόδων και των perceptrons ονομάζονται βάρη. Μετρούν το επίπεδο σπουδαιότητας κάθε εισόδου. Σε έναν βιολογικό νευρώνα, ο πυρήνας παράγει ένα σήμα εξόδου με βάση τα σήματα που παρέχονται από τους δενδρίτες. Ομοίως, ο πυρήνας (χρωματισμένος με μπλε) σε ένα perceptron εκτελεί ορισμένους υπολογισμούς με βάση τις τιμές 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 perceptron.

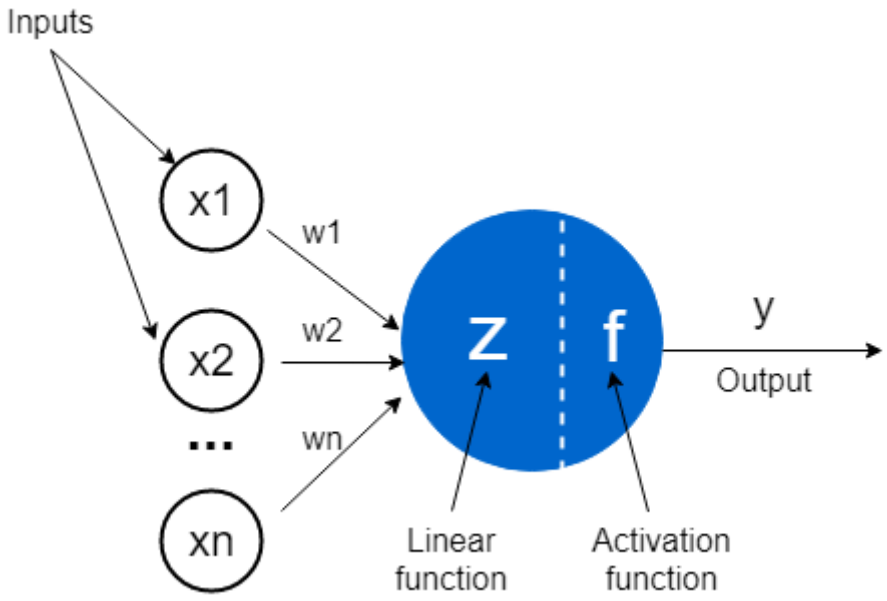
Η παρακάτω εικόνα δείχνει μια λεπτομερή δομή ενός perceptron. Σε ορισμένα περιβάλλοντα, η μεροληψία - bias, b συμβολίζεται με w_0. Η είσοδος x_0 παίρνει πάντα την τιμή 1. Άρα, $b * 1 = b$.



Εικόνα 11 Μια πιο ενδελεχής παρουσίαση ενός τεχνητού νευρώνα μαζί με την συνάρτηση εξόδου του.

Με κόκκινο αναγράφεται η αλγεβρική εξίσωση για τον υπολογισμό της εξόδου του νευρώνα. Πρακτικά, ένα perceptron παίρνει τις εισόδους, $x_1, x_2, x_3, \ldots, x_n$, τις πολλαπλασιάζει με τα βάρη, $w_1, w_2, w_3, \ldots, w_n$, προσθέτει τον όρο πόλωσης, b και μετά υπολογίζει τη γραμμική συνάρτηση z στην οποία εφαρμόζεται μια συνάρτηση ενεργοποίησης f. για να υπολογιστεί η έξοδος y.

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



Ένα perceptron αποτελείται συνήθως από δύο μαθηματικές συναρτήσεις.

- 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 μπορεί να υπολογιστεί 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.

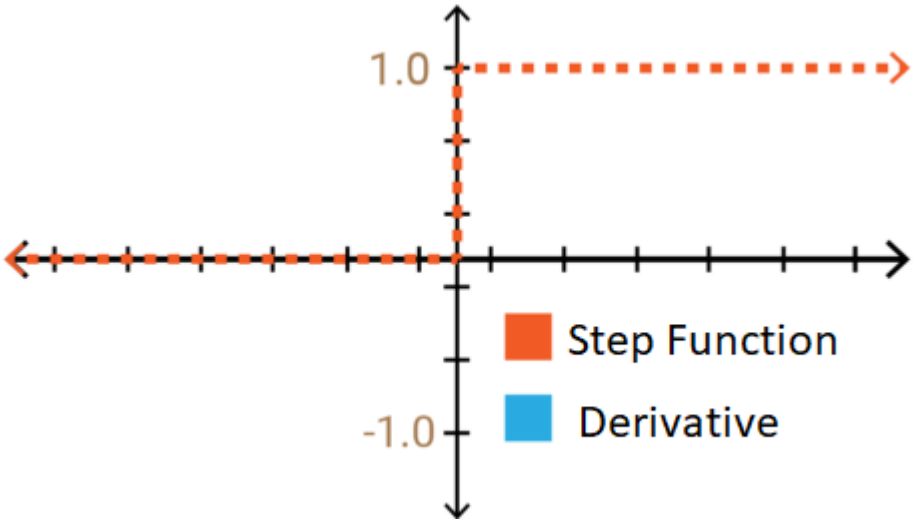
- Τα w1, w2, …, wn είναι βάρη που λαμβάνουν αριθμητικές τιμές και ελέγχουν το επίπεδο σημασίας κάθε εισόδου. Όσο μεγαλύτερη είναι η τιμή, τόσο πιο σημαντική είναι η είσοδος.
- $w_1 * x_1 + w_2 * x_2 + \ldots + w_n * x_n$ ονομάζεται το σταθμισμένο άθροισμα των εισόδων.
- Το b ονομάζεται όρος πόλωσης ή μονάδα πόλωσης που παίρνει επίσης μια αριθμητική τιμή. Προστίθεται στο σταθμισμένο άθροισμα των εισροών. Ο σκοπός της συμπερίληψης ενός όρου μεροληψίας είναι να μετατοπιστεί η συνάρτηση ενεργοποίησης κάθε perceptron για να μην ληφθεί μηδενική τιμή. Με άλλα λόγια, εάν όλες οι είσοδοι x1, x2, …, xn είναι 0, το z είναι ίσο με την τιμή της πόλωσης.

Τα βάρη και οι προκαταλήψεις ονομάζονται παράμετροι σε ένα μοντέλο νευρωνικών δικτύων. Οι βέλτιστες τιμές για αυτές τις παραμέτρους βρίσκονται κατά τη διαδικασία εκμάθησης (εκπαίδευσης) του νευρωνικού δικτύου.

Μπορείτε επίσης να σκεφτείτε την παραπάνω συνάρτηση z ως ένα μοντέλο γραμμικής παλινδρόμησης στο οποίο τα βάρη είναι γνωστά ως συντελεστές και ο όρος μεροληψίας είναι γνωστός ως τομή. Αυτή είναι απλώς η ορολογία που χρησιμοποιείται για να προσδιορίσει το ίδιο πράγμα σε διαφορετικά πλαίσια.

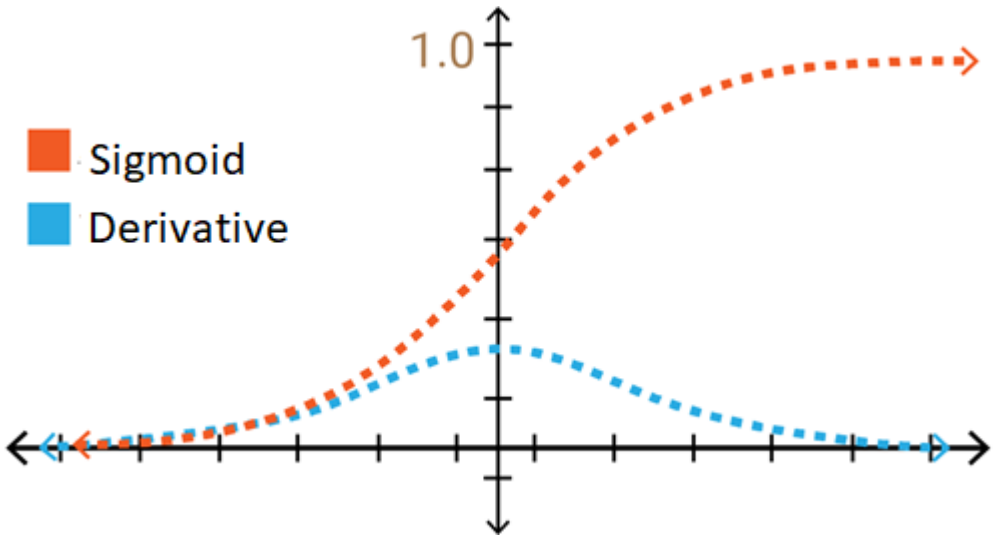
- Μη γραμμική (ενεργοποίηση) συνάρτηση Perceptron: Αυτό ονομάζεται επίσης μη γραμμική συνιστώσα του perceptron και συμβολίζεται με f. Εφαρμόζεται στο z για να λάβουμε την έξοδο y με βάση τον τύπο της συνάρτησης ενεργοποίησης που χρησιμοποιούμε. Πρακτικά μια μη-γραμμική συνάρτηση ενεργοποίησης είναι μια 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. Στην βιβλιογραφία έχουν προταθεί διάφορες συναρτήσεις ενεργοποίησης κάθε μια με τα δικά της πλεονεκτήματα και μειονεκτήματα σε σχέση με τις υπόλοιπες.
  - **Βηματική Συνάρτηση**: Είναι μια συνάρτηση που παίρνει μια δυαδική τιμή και χρησιμοποιείται ως δυαδικός ταξινομητής. Επομένως, γενικά προτιμάται στα στρώματα εξόδου. Δεν συνιστάται η χρήση του σε κρυφά επίπεδα (εσωτερικά επίπεδα) επειδή δεν είναι παραγωγίσιμη στο 0 αλλά για την εκπαίδευση αυτό δεν

αποτελεί σημαντικό πρόβλημα. Το σημαντικότερο όμως πρόβλημα που προκύπτει με την παρούσα συνάρτηση 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 δυνατή την εκπαίδευση ενός νευρωνικού δικτύου.



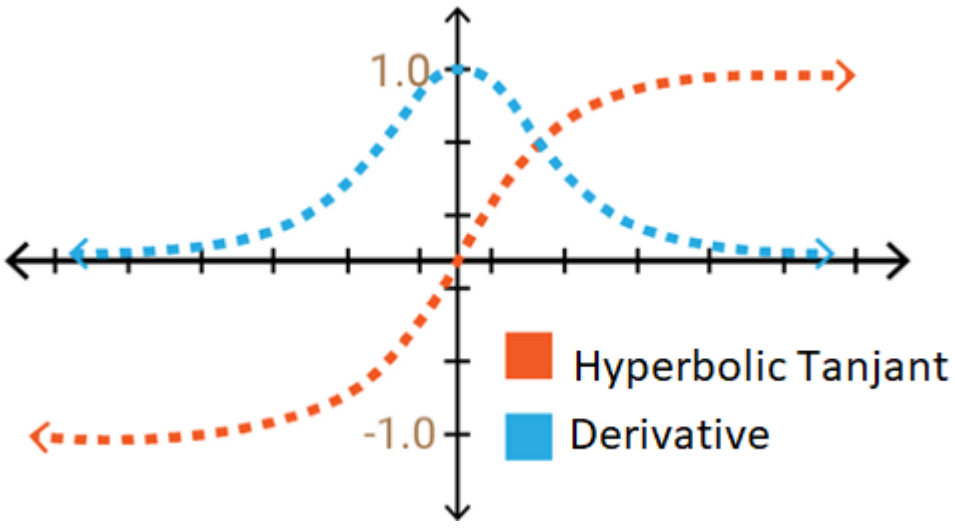
Εικόνα 12 Η γραφική αναπαράσταση της βηματικής συνάρτησης για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της.

- **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 13 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 έχει παρόμοια δομή με τη σιγμοειδή που παρουσιάστηκε παραπάνω. Ωστόσο, διαφέρουν στο πεδίο τιμών τους όπου η tanh είναι (-1, + 1). Το πλεονέκτημα έναντι της σιγμοειδούς συνάρτησης είναι ότι 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. Αλλά και πάλι, το πρόβλημα των κλίσεων στα άκρα της συνάρτησης συνεχίζεται απλά μειώνεται το φαινόμενο του vanishing gradient.



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

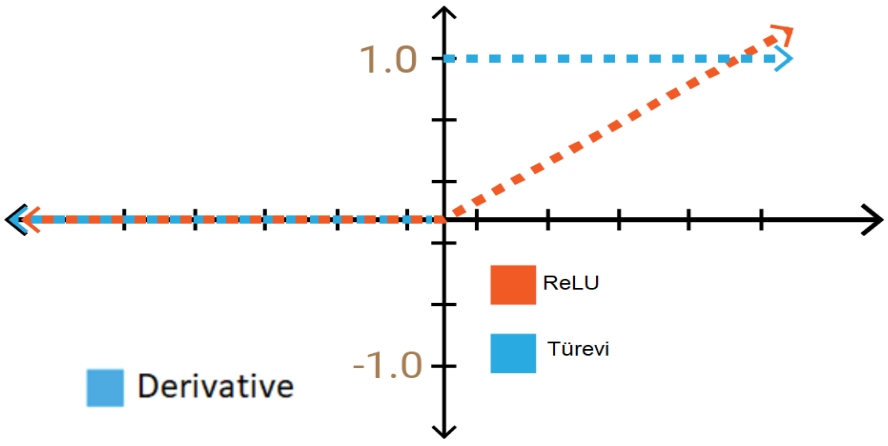
Στο παραπάνω σχήμα απεικονίζεται η συνάρτηση, η οποία είναι παραγωγίσιμη σε όλο το πεδίο τιμών, με τη μέγιστη τιμή της παραγώγου να φτάνει το 1, σε αντίθεση με την παράγωγο της σιγμοειδούς η όποια έφτανε μέχρι το 0.25.

- **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.

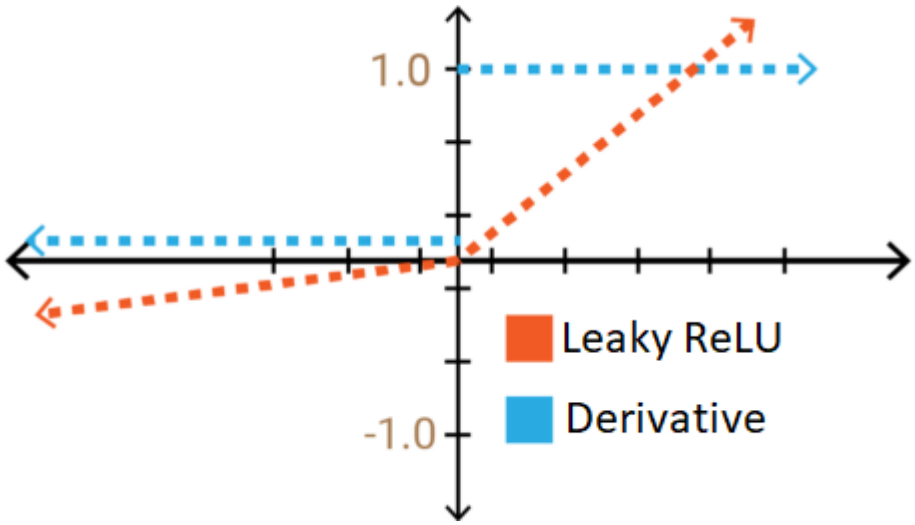
Το βασικό πλεονέκτημα της παρούσας συνάρτησης έναντι των υπολοίπων είναι η σταθερή τιμή της παραγωγού για τις τιμές που είναι μεγαλύτερες του μηδενός. Έτσι εξαλείφεται το πρόβλημα του vanishing gradient που αναφέρθηκε παραπάνω για τις τιμές αυτές. Παρ' όλα αυτά για τιμές μικρότερες του μηδενός η παράγωγος 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 15 Η γραφική αναπαράσταση της συνάρτησης ReLU για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της.

Στο παραπάνω σχήμα φαίνεται ότι η συνάρτηση είναι παραγωγίσιμη σε όλο το πεδίο τιμών πλην του 0 και η μέγιστη τιμή της παραγώγου φτάνει το 1, τιμή την όποια διατηρεί για κάθε x > 0, σε αντίθεση με τις προηγούμενες 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. Εκεί για να αποφύγουμε όσα αναφέρθηκαν παραπάνω δίνεται μία μικρή κλίση ώστε η παράγωγος να έχει μία μικρή τιμή διάφορη του μηδενός και να μειώνεται το πρόβλημα του vanishing gradient. Η τιμή της κλίσης συνήθως κυμαίνεται από 0.01 έως 0.1.



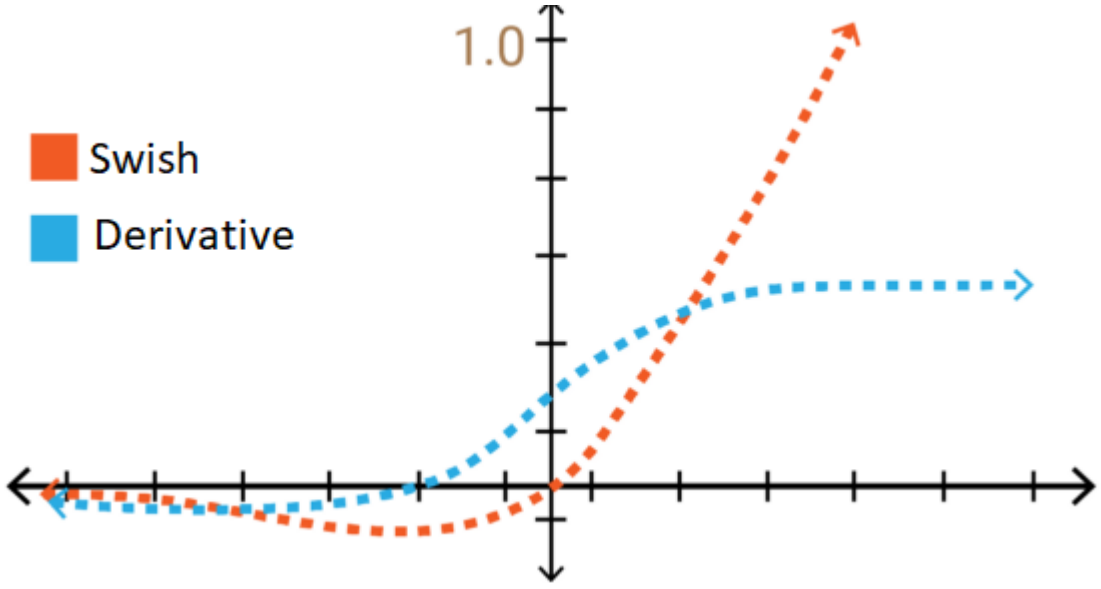
Εικόνα 16 Η γραφική αναπαράσταση την συνάρτηση Leaky ReLU για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της

Σε αυτό φαίνεται ότι η παράγωγος της είναι παραγωγίσιμη σε όλο το πεδίο τιμών πλην του 0 και η μέγιστη τιμή της φτάνει το 1 τιμή την όποια διατηρεί για κάθε x > 0, σε αντίθεση με τις παραπάνω συναρτήσεις που είχαν μέγιστη τιμή μόνο σε ένα σημείο. Επιπλέον, σε αντίθεση με την απλή ReLU η τιμή της παραγώγου 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 συνάρτηση αυτή είναι αρκετά παρόμοια με την ReLU, με την κυριότερη διαφορά τους να έγκειται στην αρνητική περιοχή (όπως και η Leaky ReLU). Το κομμάτι αυτό αντί να έχει σταθερή τιμή ίση με 0 (όπως στην ReLU) ή μια μονότονη μικρή αρνητική κλίση (όπως στην Leaky ReLU) έχει μεταβλητή κλίση και έτσι στις πολύ αρνητικές τιμές τείνει στο 0. Αξίζει να σημειωθεί ότι η έξοδος της συνάρτησης swish μπορεί να πέσει ακόμα και όταν η είσοδος αυξάνεται. Η αναλυτική περιγραφή της συνάρτησης αυτή στο αρνητικό μέρος παρουσιάζεται παρακάτω:



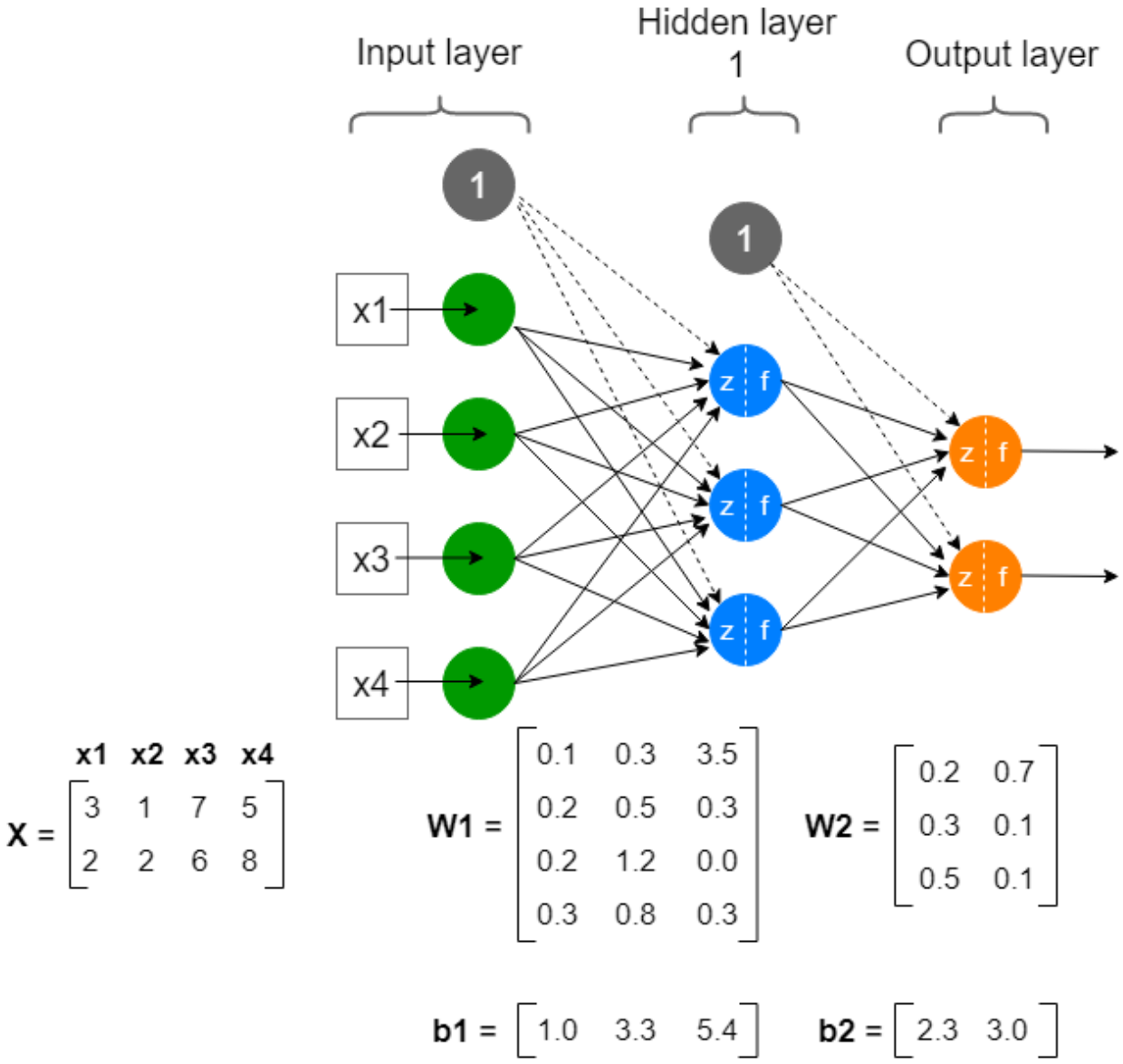
$$f(x) = 2 * x * sigmoid(beta * x)$$

Εικόνα 17 Η γραφική αναπαράσταση της συνάρτησης Swish για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της. Σε αυτό φαίνεται ότι η παράγωγος της είναι παραγωγίσιμη 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 $\chi < 0.1$ $και$ $x > 0$. 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).

Εν κατακλείδι ο σκοπός μιας συνάρτησης ενεργοποίησης είναι να εισαγάγει μη γραμμικότητα στο δίκτυο. Χωρίς συνάρτηση ενεργοποίησης, ένα νευρωνικό δίκτυο μπορεί να μοντελοποιήσει μόνο γραμμικές σχέσεις. Στα δεδομένα του πραγματικού κόσμου οι περισσότερες συναρτήσεις είναι μη-γραμμικές. 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.

Ωστόσο, ένα μόνο perceptron 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 (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 (MPL). 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 (FCNN) στο οποίο κάθε κόμβος σε κάθε επίπεδο συνδέεται με όλους τους άλλους κόμβους στο επόμενο επίπεδο, όπως στο παρακάτω διάγραμμα. Ωστόσο, οι κόμβοι σε ένα μόνο επίπεδο δεν μοιράζονται καμία σύνδεση.



Εικόνα 18 Η γραφική αναπαράσταση της συνάρτησης Swish για μια τιμή. Στο παρόν διάγραμμα εμφανίζεται τόσο η συνάρτηση όσο και η πρώτη παράγωγός της. Σε αυτό φαίνεται ότι η παράγωγος της είναι παραγωγίσιμη σε όλο το πεδίο τιμών

Το παραπάνω δίκτυο έχει ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και ένα κρυφό στρώμα. Ένα ANN με μόνο ένα κρυφό στρώμα είναι γνωστό ως Shalow Neural Network. Το παραπάνω δίκτυο έχει μόνο ένα κρυφό στρώμα και είναι, επομένως, ένα παράδειγμα ρηχού νευρωνικού δικτύου. Αντίθετα, ένα ANN με δύο ή περισσότερα κρυφά επίπεδα είναι γνωστό ως Deep Neural Network.

Ως τελευταίο βασικό ορισμό σε αυτήν την ανάρτηση αναφέρουμε τα νευρωνικά δίκτυα τροφοδοσίας (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.

Ουσιαστικά λοιπόν 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. Σε ένα ρηχό νευρωνικό δίκτυο, υπάρχει 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. Είναι επίσης μια υπερπαράμετρος. Ο αριθμός των κρυφών επιπέδων μετρά το βάθος ενός νευρωνικού δικτύου. Καθώς ο αριθμός των κρυφών επιπέδων αυξάνεται, τα νευρωνικά δίκτυα μπορούν να μοντελοποιήσουν πιο σύνθετες σχέσεις στα δεδομένα. 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. Αυτή είναι επίσης μια υπερπαράμετρος. Η αξία του εξαρτάται από το είδος του προβλήματος που καλείται να λύσει το νευρωνικό.

3.3.SVM

Τα Support Vector Machines είναι από τα πλέον διαδεδομένα μοντέλα μηχανικής μάθησης για προβλήματα ταξινόμησης. Αυτό οφείλεται κυρίως στις πολύ καλές επιδόσεις που έχουν επιτύχει σε πληθώρα προβλημάτων. Γενικότερα, μπορούν να χρησιμοποιηθούν και σε προβλήματα ταξινόμησης και σε regression problems, αλλά στο πλαίσιο της παρούσας εργασίας θα τα χρησιμοποιήσουμε ως classification models. Τα svm μπορούν με την ίδια ευκολία να χειριστούν τόσο συνεχή όσο και κατηγοριοποιημένα δεδομένα. Σκοπός τους είναι να κατασκευάζουν ένα multidimensional υπερεπίπεδο για το διαχωρισμό των δεδομένων ανά κλάση. Το υπερεπίπεδο αυτό μαθαίνεται αυτόματα μελετώντας τα δεδομένα προσπαθώντας να ελαχιστοποιήσει το σφάλμα ταξινόμησης σε αυτά. Η κυρία αρχή λειτουργία τους στηρίζεται στην εύρεση του υπερεπιπέδου με το μεγαλύτερη δυνατή απόσταση - κενό μεταξύ των κλάσεων maximum marginal hyperplane (MMH).

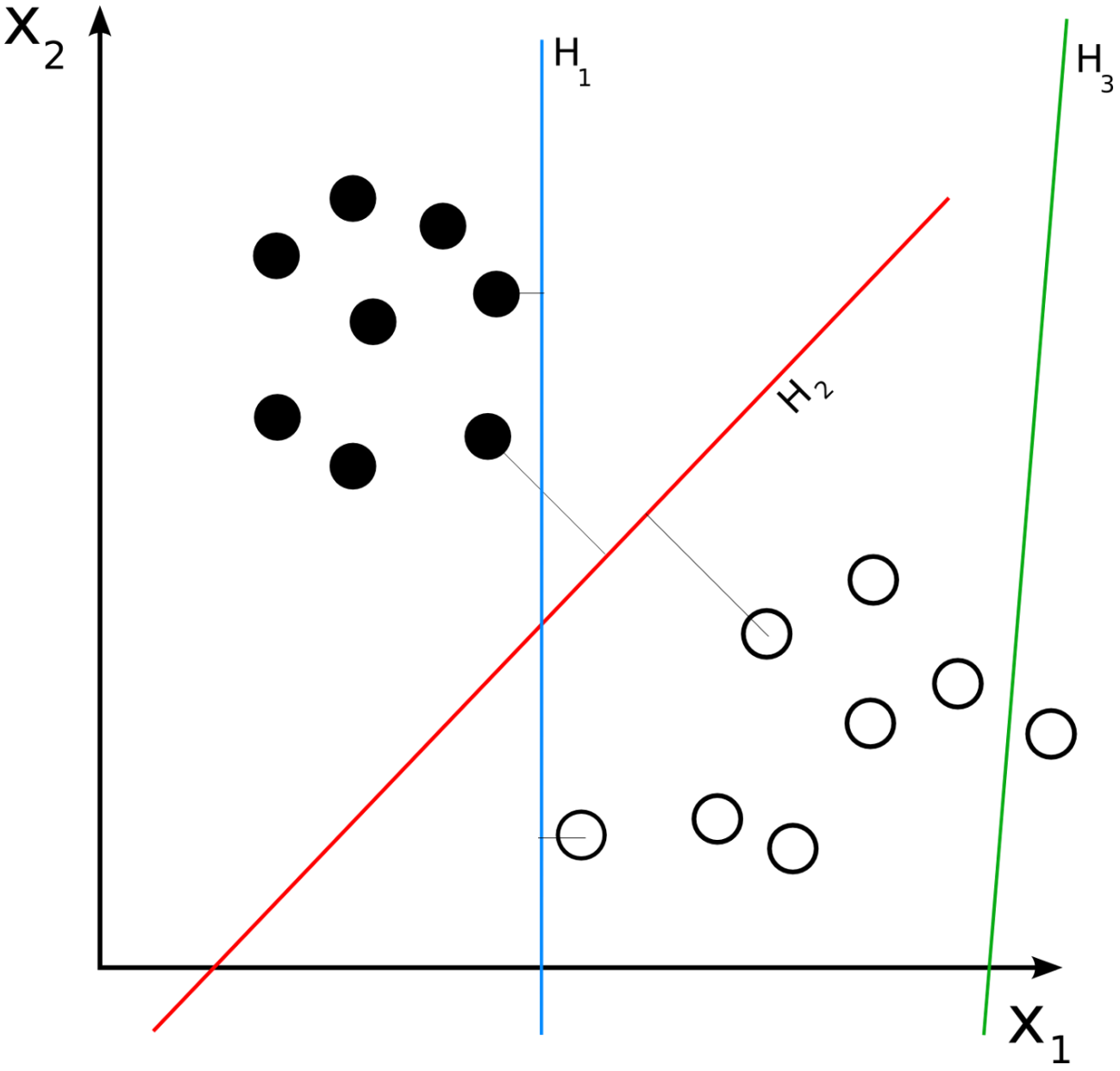
Για την καλύτερη κατανόηση του προβλήματος στο παρακάτω διάγραμμα παρουσιάζεται ένα απλό παράδειγμα διαχωρισμού δεδομένων δύο διαστάσεων σε δύο κλάσεις. Έστω ότι τα μαύρα δεδομένα ανήκουν στην κλάση 0 και τα λευκά στην κλάση 1. Αρχικά η γραμμή H3 μπορούμε εύκολα να καταλάβουμε ότι δεν είναι κατάλληλη 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 προβλήματα σε πληθώρα προβλημάτων. Στο παρόν πρόβλημα χρησιμοποιείται με είσοδο τις τιμές που έχουν παραχθεί από την PCA, δηλαδή το διάνυσμα V1-V28, καθώς και τις κανονικοποιημένες τιμές των πεδίων amount και time, ενώ σαν έξοδο προβλέπει αν η συναλλαγή αποτελεί απάτη ή όχι.

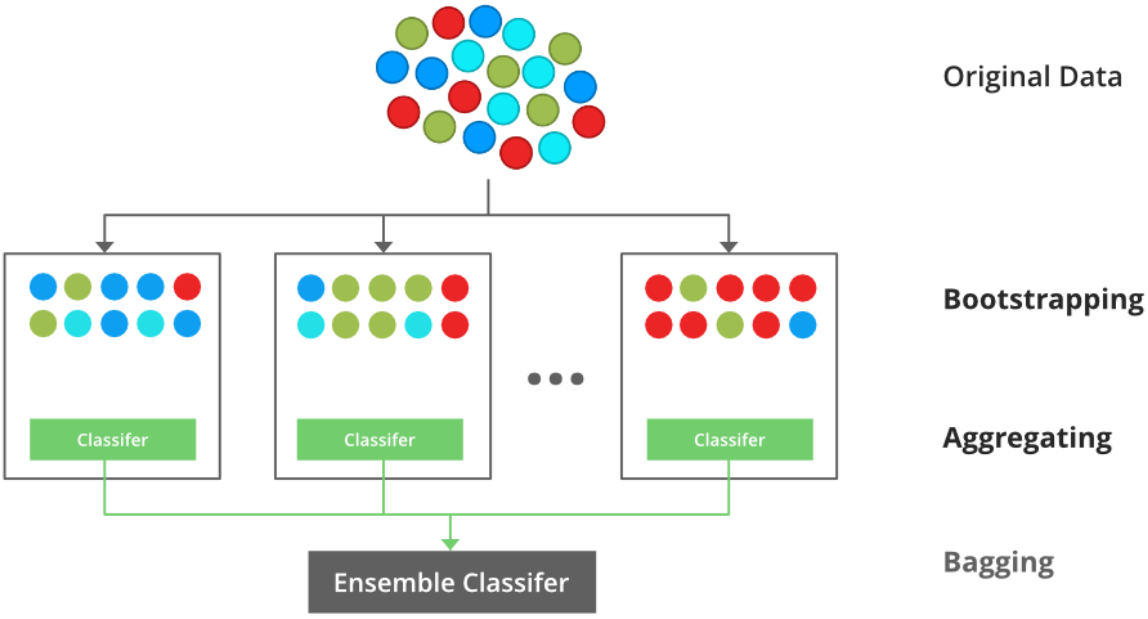


Εικόνα 19 Ο τρόπος υπολογισμού του υπερεπιπέδου διαχωρισμού σύμφωνα με την μέθοδο SVM.

3.4.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 20 . Η γραφική αναπαράσταση των boosted μεθόδων ταξινόμησης και ο τρόπος με τον όποιο χρησιμοποιούν πολλαπλούς ταξινομητές για την τελική πρόβλεψη.

4. Μέθοδοι Εξομάλυνσης

Όπως έχει αναφερθεί και παραπάνω το πρόβλημα που καλούμαστε να αντιμετωπίσουμε στην συγκεκριμένη εργασία ξεπερνά αυτό ενός απλού binary classification. Ο λόγος που το συγκεκριμένο πρόβλημα είναι πιο δύσκολο είναι η κατανομή των δεδομένων ανά κλάση. Τα δεδομένα περιέχουν κυρίως δείγματα έγκυρων συναλλαγών και ελάχιστά δεδομένα από απάτη. Η επίδραση αυτής της ανισοκατανομής θα είναι ριζική στην εκπαίδευση οποιουδήποτε συστήματος αναγνώρισης μη-έγκυρων συναλλαγών.

Γι' αυτόν τον λόγο, στο συγκεκριμένο κεφάλαιο θα μελετήσουμε θεωρητικά διάφορες μεθόδους μείωσης των επιδράσεων αυτής της ανισοκατανομής. Συγκεκριμένα θα μελετήσουμε 6 μεθόδους: την Υποδειγματοληψία, την υπερδειγματοληψία, την μέθοδο Smote, την Over-Undersampling, μια regularization τεχνική καθώς και μια τεχνική παραγωγής δεδομένων με την χρήση αντιστρέψιμων νευρωνικών δικτύων. Το θεωρητικό κομμάτι κάθε μιας τέτοια μεθόδου παρουσιάζεται παρακάτω ενώ στο κεφάλαιο των πειραμάτων αναλύονται τα αποτελέσματα των μεθόδων αυτών.

4.1 Υποδειγματοληψία

Μια προσέγγιση για την αντιμετώπιση του προβλήματος της ανισοκατανομής των δεδομένων είναι η τυχαία υποδειγματοληψία του συνόλου εκπαίδευσης [16]. Οι δύο κύριες προσεγγίσεις για την τυχαία επαναδειγματοληψία ενός μη ισορροπημένου συνόλου δεδομένων είναι η διαγραφή παραδειγμάτων από την κλάση πλειοψηφίας, που ονομάζεται υποδειγματοληψία, και η αντιγραφή παραδειγμάτων από την κατηγορία μειοψηφίας, που ονομάζεται υπερδειγματοληψία. Στην πρώτη κατηγορία ουσιαστικά μειώνεται το σύνολο των δεδομένων διαγράφοντας μέρη του συνόλου, ώστε τελικά να μην υπάρχουν επιδράσεις της ανισοκατανομής στο σύστημα κατά την εκπαίδευση. Η δεύτερη τεχνική αφορά στον πολλαπλασιασμό των δειγμάτων με την επανάληψη δειγμάτων που ανήκουν στην κατηγορία με τα λιγότερα δείγματα. Επιπλέον ανάλυση της δεύτερης μεθόδου θα γίνει παρακάτω.

Η τυχαία υποδειγματοληψία περιλαμβάνει την τυχαία επιλογή παραδειγμάτων από την πλειοψηφική τάξη προς διαγραφή από το σύνολο δεδομένων εκπαίδευσης.
Αυτό έχει ως αποτέλεσμα τη μείωση του αριθμού των παραδειγμάτων στην πλειοψηφική τάξη στη μετασχηματισμένη έκδοση του συνόλου δεδομένων εκπαίδευσης. Αυτή η διαδικασία μπορεί να

επαναληφθεί μέχρι να επιτευχθεί η επιθυμητή κατανομή κλάσης, όπως ίσος αριθμός παραδειγμάτων για κάθε κλάση.

Αυτή η προσέγγιση μπορεί να είναι πιο κατάλληλη για εκείνα τα σύνολα δεδομένων όπου υπάρχει ανισορροπία κλάσης και επαρκής αριθμός παραδειγμάτων στην κατηγορία μειοψηφίας. Ένας περιορισμός της υποδειγματοληψίας είναι ότι διαγράφονται παραδείγματα από την πλειοψηφική τάξη που μπορεί να είναι χρήσιμα, σημαντικά ή ίσως κρίσιμα για την τοποθέτηση ενός ισχυρού ορίου απόφασης. Δεδομένου ότι τα παραδείγματα διαγράφονται τυχαία, δεν υπάρχει τρόπος να εντοπιστούν ή να διατηρηθούν "καλά" ή περισσότερα πλούσια σε πληροφορίες παραδείγματα από την πλειοψηφική τάξη. 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. Αν για παράδειγμα ένα σύνολο κρατήσει “σημαντικά” δείγματα για την εκπαίδευση τότε η απόδοση του μοντέλου ταξινόμησης αναμένεται να είναι σημαντικά υψηλότερη από το ίδιο μοντέλο που θα εκπαιδευτεί σε δεδομένα τα όποια έχουν αφαιρεθεί τα συγκεκριμένα δείγματα.

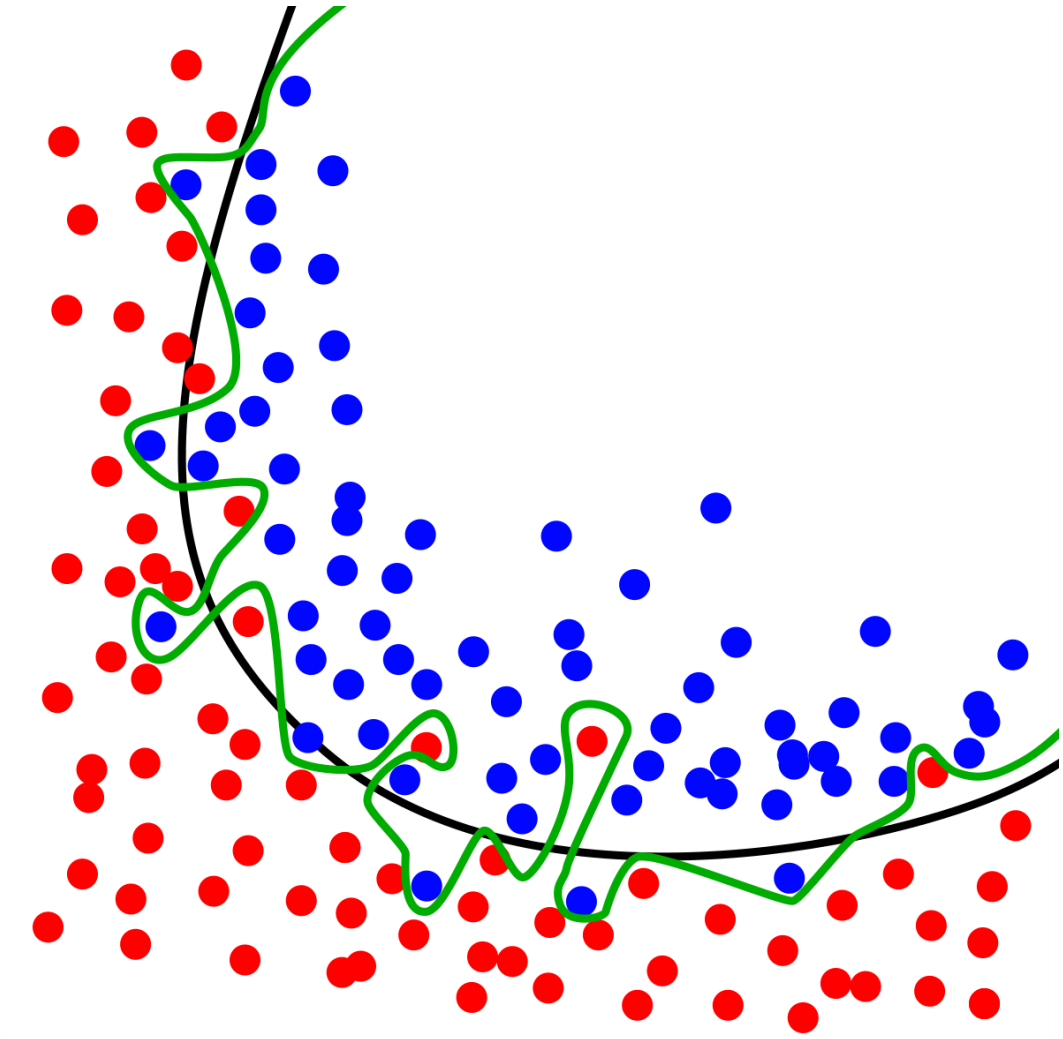
Κατά συνέπεια, αν έχουμε ένα σύνολο δεδομένων με συνολικά 1000 δείγματα όπου τα 100 μόνο ανήκουν σε μια κατηγορία τότε μετά την εφαρμογή της συγκεκριμένης μεθόδου το σύνολο θα έχει 200 δείγματα, εκ των οποίων 100 της λιγότερα συχνής κλάσης και άλλα 100 που ανήκουν στη δεύτερη κλάση τα όποια επιλέχθηκαν τυχαία. Η επιλογή αυτών των 100 όπως αναφέρθηκε και παραπάνω μπορεί να καθορίσει σημαντικά την απόδοση της μεθόδου μοντελοποίησης. Επιπλέον, στο παρόν σύνολο δεδομένων ο αριθμός των δειγμάτων που ανήκουν στη λιγότερα συχνή κλάση είναι πολύ μικρός, και ίσως μη επαρκής για τη σωστή μοντελοποίηση των κλάσεων.

4.2 Υπερδειγματοληψία

Η τυχαία υπερδειγματοληψία περιλαμβάνει την τυχαία αντιγραφή παραδειγμάτων από την τάξη μειοψηφίας και την προσθήκη τους στο σύνολο δεδομένων εκπαίδευσης. Παραδείγματα από το σύνολο 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 από το αρχικό σύνολο δεδομένων εκπαίδευσης, προστίθενται στο νέο σύνολο δεδομένων εκπαίδευσης και στη συνέχεια επιστρέφονται ή «αντικαθίστανται» στο αρχικό σύνολο δεδομένων, επιτρέποντάς τους να επιλεγούν ξανά.

Αυτή η τεχνική μπορεί να είναι αποτελεσματική για τους αλγόριθμους μηχανικής μάθησης που επηρεάζονται από μια ανισοκατανεμημένη κατανομή δεδομένων και όπου πολλά διπλά παραδείγματα για μια δεδομένη τάξη μπορούν να επηρεάσουν την προσαρμογή του μοντέλου. Αυτό μπορεί να περιλαμβάνει αλγορίθμους που μαθαίνουν επαναληπτικά τους συντελεστές, όπως τα τεχνητά νευρωνικά δίκτυα που χρησιμοποιούν στοχαστική κλίση καθόδου. Μπορεί επίσης να επηρεάσει μοντέλα που αναζητούν καλούς διαχωρισμούς των δεδομένων, όπως μηχανές υποστήριξης διανυσμάτων και δέντρα αποφάσεων. Αξίζει να σημειωθεί ότι τέτοια μοντέλα έχουν επιδείξει εξαιρετικά αποτελέσματα σε πληθώρα προβλημάτων. Επίσης αντίστοιχα μοντέλα έχουν χρησιμοποιηθεί στην παρούσα διπλωματική για να αναδειχθεί η χρησιμότητα της μεθόδου σε αυτά.

Μπορεί να είναι χρήσιμο να προσαρμοστεί η συχνότητα επανάληψης των δεδομένων της κλάσης στόχου. 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 γενίκευσης (overfitting). Αυτό πρακτικά σημαίνει ότι το δίκτυο μπορεί να απομνημονεύσει τα δείγματα τα όποια επαναλαμβάνονται πολλαπλές φορές αντί να μάθει πως να τα διαχωρίζει. Το αποτέλεσμα μπορεί να είναι καλύτερη απόδοση στο σύνολο δεδομένων εκπαίδευσης, αλλά χειρότερη απόδοση στο συγκρατημένο ή δοκιμαστικό σύνολο δεδομένων. Ένα παράδειγμα αυτού του φαινομένου φαίνεται στο παρακάτω σχήμα.

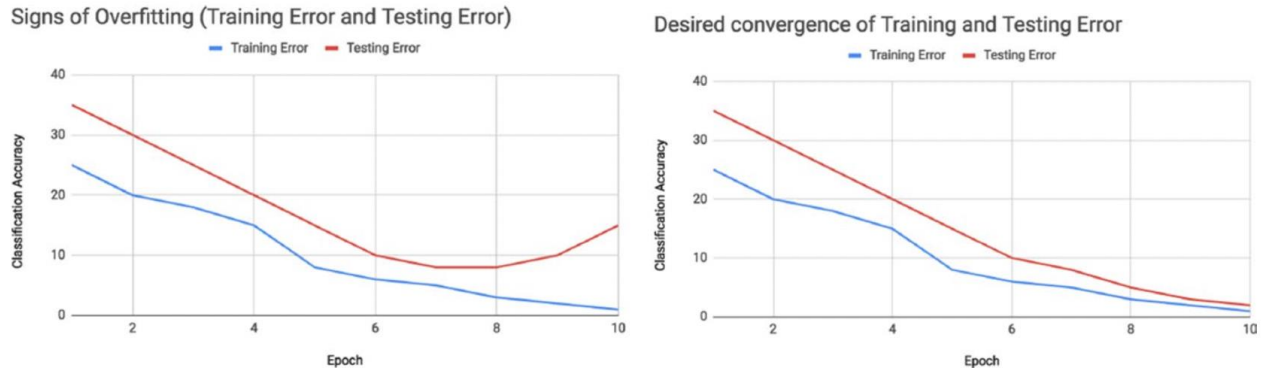


Εικόνα 21 Ένα παράδειγμα όπου φαίνονται 2 διαχωριστικές γραμμές διαφορετικής πολυπλοκότητας. Από την κατανομή των δεδομένων και το είδος της διαχωριστικής γραμμής υποθέτουμε ότι η πράσινη γραμμή είναι αποτέλεσμα υπερεκπαίδευσης.

Στο παραπάνω σχήμα έχουμε ένα σύνολο δεδομένων το οποίο χωρίζεται σε 2 κλάσεις, τα κόκκινα και τα μπλε δείγματα. Αν προσπαθήσουμε να ζωγραφίσουμε μια γραμμή (όχι ευθεία) η όποια να χωρίζει τα δείγματα όσο το δυνατόν καλύτερα εξετάζοντας μόνο την επίδοση, θα ήταν πράσινη, η όποια θα είχε 100% επιτυχία. Πρακτικά όμως η συγκεκριμένη γραμμή δεν είναι η βέλτιστη, γιατί είναι αρκετά περίπλοκη για να περιγράψει τον διαχωρισμό, μιας και τα δείγματα που ταξινομούνται λάθος από τη μαύρη γραμμή μπορεί να είναι κάποιες ακραίες περιπτώσεις οι οποίες δεν περιγράφουν ωστόσο το πρόβλημα. Ας υποθέσουμε για παράδειγμα ότι έχουμε τα δείγματα τα όποια εξετάζουν αν ένα αεροπλάνο θα πετάξει ή όχι βάσει των καιρικών συνθηκών. Αν ένα αεροπλάνο είχε πετάξει με 10 μποφόρ κάποια στιγμή, εξαιτίας του γεγονότος ότι έκανε μια σοβαρή αεροδιακομιδή ή επειδή δεν είχε γίνει σωστή πρόβλεψη, δεν θα θέλαμε σε καμία των περιπτώσεων να ακολουθείται αυτή η απόφαση κάθε φορά, απλά και μόνο επειδή υπάρχει ένα δείγμα. Συνεπώς αν αυτά τα δείγματα το νευρωνικό δίκτυο τα δει πολλαπλές φορές, τότε θα παίξουν σοβαρό ρόλο στη διαμόρφωση της διαχωριστικής γραμμής με αποτέλεσμα το σύστημα να μην γενικεύσει και να κάνει

overfitting. Το καλό με αυτό όμως είναι ότι μπορούμε εύκολα και γρήγορα να εντοπίσουμε αυτά τα φαινόμενα. Αν το σφάλμα της εκπαίδευσης είναι σημαντικά μικρότερο από αυτό του testing τότε έχουμε μια πολύ καλή ένδειξη ότι το δίκτυο έχει κάνει overfitting.

Ως εκ τούτου, για να αποκτήσουμε μια εικόνα για την επίδραση 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 να συγκρίνουμε τα αποτελέσματα με τον ίδιο αλγόριθμο στο αρχικό σύνολο δεδομένων. Στο παρακάτω διάγραμμα εμφανίζεται ένα παράδειγμα [18] για το πως μπορούμε να διακρίνουμε το overfitting από τις καμπύλες εκμάθησης και ελέγχου.



Εικόνα 22 Αριστερά απεικονίζεται ένα σημείο καμπής όπου το σφάλμα επικύρωσης αρχίζει να αυξάνεται καθώς ο ρυθμός εκπαίδευσης συνεχίζει να μειώνεται. Αντίθετα, δεξιά απεικονίζεται ένα μοντέλο με την επιθυμητή σχέση 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. 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, η γραφική παράσταση στα δεξιά δείχνει ένα μοντέλο με την επιθυμητή σχέση 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.

Μια σχηματική αναπαράσταση των μεθόδων υπερ και υπο δειγματοληψίας όπου διαφαίνονται και οι βασικές διαφορές παρουσιάζεται στο παρακάτω σχήμα:

Εικόνα 23 Μια σχηματική αναπαράσταση των μεθόδων της υπερ και υποδειγματοληψίας όπου φαίνεται ξεκάθαρα ο διαφορετικός τρόπος με τον όποιο επιδρούν στα δεδομένα.

Στο παραπάνω παράδειγμα των 1000 δειγμάτων το νέο σύνολο μετά την εφαρμογή της υπερδειγματοληψίας θα έχει συνολικά 1800 δείγματα. Τα 1000 αποτελούν τα ίδια με το αρχικό και τα υπόλοιπα 800 είναι η επανάληψη 8 φορών των δειγμάτων της λιγότερο συχνής κλάσης. Οπότε από τα 1800 δείγματα, τα 900 θα αποτελούν τα μοναδικά δείγματα της "συχνότερης" κλάσης και τα υπόλοιπα 900 θα είναι 9 φορές τα ίδια δείγματα της λιγότερα συχνής κλάσης. Έτσι, τελικά πάλι το σύνολο μετά την εφαρμογή της μεθόδου θα είναι ισομοιρασμένο και δεν θα έχει χαθεί πληροφορία, απλά μερικά δείγματα θα έχουν επαναληφθεί πολλαπλές φορές.

4.3 Over & UnderSampling

Όπως αναφέρθηκε και προηγουμένως, τα δείγματα που ανήκουν στη λιγότερο συχνή κλάση είναι πολύ μικρά σε σχέση με τη δεύτερη. Αυτό στην υπερδειγματοληψία θα δημιουργήσει το πρόβλημα ότι τα δείγματα αυτά θα επαναλαμβάνονται πάρα πολλές φορές με αποτέλεσμα να οδηγούν τον αλγόριθμο μοντελοποίησης σε overfitting πολύ πιο εύκολα και συχνά. Από την άλλη, για τον ίδιο λόγο η μέθοδος undersampling θα μειώσει σημαντικά την ποσότητα της πληροφορίας που περιέχει το dataset. Συνεπώς, και οι δυο αυτές μέθοδοι μπορεί να επιλύουν το πρόβλημα της ανισοκατανομής αλλά κάθε μια εισάγει διαφορετικά προβλήματα στο τελικό σύνολο δεδομένων.

Μια ενδιαφέρουσα μέθοδος μπορεί να προκύψει συνδυάζοντας τις δύο προηγούμενες μεθόδους της υπερδειγματοληψίας και της υποδειγματοληψίας. Παραδείγματος χάρη, μια μικρή ποσότητα υπερδειγματοληψίας μπορεί να εφαρμοστεί στην τάξη μειοψηφίας για να βελτιωθεί η μεροληψία προς αυτά τα παραδείγματα, ενώ εφαρμόζεται επίσης μια μικρή ποσότητα υποδειγματοληψίας στην πλειοψηφική τάξη για να μειωθεί η μεροληψία σε αυτήν την κατηγορία.

Αυτό μπορεί να οδηγήσει σε βελτιωμένη συνολική απόδοση σε σύγκριση με την εκτέλεση της μιας ή της άλλης τεχνικής μεμονωμένα.

Πιο συγκεκριμένα, εάν είχαμε ένα σύνολο δεδομένων με κατανομή κλάσης 1:100, θα μπορούσαμε πρώτα να εφαρμόσουμε υπερδειγματοληψία για 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 παραδειγμάτων από την πλειοψηφική τάξη.

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%).

4.4 Smote

Όπως έχει ήδη προαναφερθεί και παραπάνω, μια προσέγγιση για την αντιμετώπιση των συνεπειών που επιφέρει η εκπαίδευση συστημάτων με την χρήση μη ισορροπημένων συνόλων δεδομένων είναι η υπερδειγματοληψία της μειοψηφικής κλάσης ή η υποδειγματοληψία 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. Για τον λόγο αυτό αξίζει να μελετηθεί η δυνατότητα σύνθεσης δειγμάτων από τα υπάρχοντα παραδείγματα. Αυτή είναι μια μέθοδος αύξησης δεδομένων για την κατηγορία μειοψηφίας και αναφέρεται ως Τεχνική Υπερδειγματοληψίας Συνθετικής Μειονότητας ή (Synthetic Minority Oversampling Technique ή SMOTE).

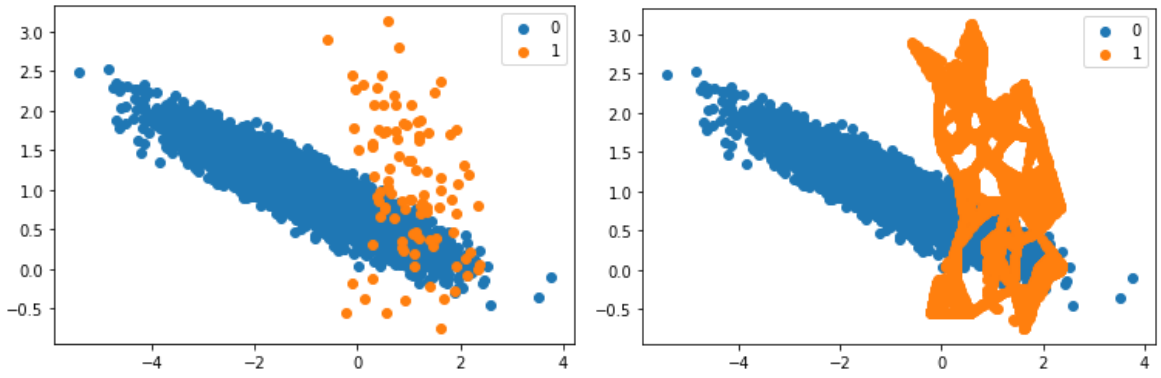
Ένα πρόβλημα με την ανισόρροπη ταξινόμηση είναι ότι υπάρχουν πολύ λίγα παραδείγματα της τάξης μειοψηφίας ώστε ένα μοντέλο να μάθει αποτελεσματικά το όριο απόφασης. Ειδικά στο συγκεκριμένο σύνολο δεδομένων, το πρόβλημα είναι ιδιαίτερα έντονο με τα δείγματα της κλάσης μειοψηφίας να αποτελούν ουσιαστικά μόνο το 0.172% του συνολικού συνόλου δεδομένων. Η τεχνική SMOTE λειτουργεί επιλέγοντας παραδείγματα που βρίσκονται κοντά στον χώρο χαρακτηριστικών, σχεδιάζοντας μια γραμμή μεταξύ των παραδειγμάτων στο χώρο χαρακτηριστικών και σχεδιάζοντας ένα νέο δείγμα σε ένα σημείο κατά μήκος αυτής της γραμμής.

Συγκεκριμένα, αρχικά επιλέγεται ένα τυχαίο παράδειγμα από την κατηγορία 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 για την τάξη μειοψηφίας απαιτούνται. Όπως περιγράφεται από τους συγγραφείς [19], πρώτα προτείνεται η χρήση τυχαίας υποδειγματοληψίας για την περικοπή του αριθμού των παραδειγμάτων στην κλάση

πλειοψηφίας και, στη συνέχεια, η εφαρμογή του SMOTE για την υπερδειγματοληψία της μειοψηφίας ώστε να εξισορροπηθεί η κατανομή της κλάσης.

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 εν λόγω προσέγγισης είναι ότι τα συνθετικά παραδείγματα δημιουργούνται χωρίς να λαμβάνεται υπόψη η πλειοψηφική τάξη, με αποτέλεσμα πιθανώς διφορούμενα παραδείγματα εάν υπάρχει ισχυρή επικάλυψη για τις κλάσεις.

Για να γίνει κατανοητός ο τρόπος με τον όποιο δουλεύει η συγκεκριμένη τεχνική παρακάτω παρουσιάζουμε ένα παράδειγμα της επαύξησης των δεδομένων για ένα πρόβλημα δύο διαστάσεων. Στο πρώτο σχήμα φαίνεται το αρχικό σύνολο δεδομένων ενώ στο επόμενο το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου SMOTE.



Εικόνα 24 Ο τρόπος με τον όποιον η μέθοδος Smote συνθέτει δεδομένα. Αριστερά είναι ένα αρχικό σύνολο δεδομένων ενώ δεξιά είναι τα δεδομένα μετά την εφαρμογή της μεθόδου αυτής.

Όπως φαίνεται από το δεύτερο σχήμα, η τεχνική αυτή έχει προσθέσει σημεία τα όποια βρίσκονται ανάμεσα σε πραγματικά κοντινά σημεία του αρχικού συνόλου δεδομένων. Η τεχνική αυτή όπως φαίνεται μπορεί να μην έχει πάντα τα καλύτερα δυνατά αποτελέσματα. Για παράδειγμα μέσω της SMOTE, έχουν προστεθεί σημεία της κλάσης 1 επάνω στα δείγματα της κλάσης 0. Αυτό μπορεί να ενισχύει τον θόρυβο που μπορεί να προέρχεται από outliers.

4.5 Regularization

Στην εποπτευόμενη μηχανική εκμάθηση, τα μοντέλα εκπαιδεύονται σε ένα υποσύνολο δεδομένων, γνωστό και ως δεδομένα εκπαίδευσης. Ο στόχος είναι να υπολογιστεί η κλάση κάθε σημείου από τα δεδομένα εκπαίδευσης. Τώρα, η υπερπροσαρμογή 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 συμβαίνει για διάφορες

αιτίες, εκ των όποιων μια είναι ότι μπορεί να έχουμε λίγα δεδομένα εκπαίδευσης (σε μια ή σε περισσότερες κλάσεις) όπως συμβαίνει στο παρόν σύνολο δεδομένων.

Στο επίπεδο του μοντέλου υπάρχουν ορισμένοι τρόποι για την αποφυγή της υπερεκπαίδευσης. Η πλέον γνωστή οικογένεια μεθόδων είναι η κανονικοποίηση (Regularization). Η κανονικοποίηση προσθέτει βασικά μια ποινή καθώς αυξάνεται η πολυπλοκότητα του μοντέλου. Οι L1 και L2 είναι οι δύο πιο γνωστές τέτοιες συναρτήσεις που χρησιμοποιούνται στη μηχανική μάθηση και χρησιμοποιούνται όπως είναι λογικό κατά τη διαδικασία εκπαίδευσης, περιγράφονται δε μαθηματικά ως εξής:

- 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.

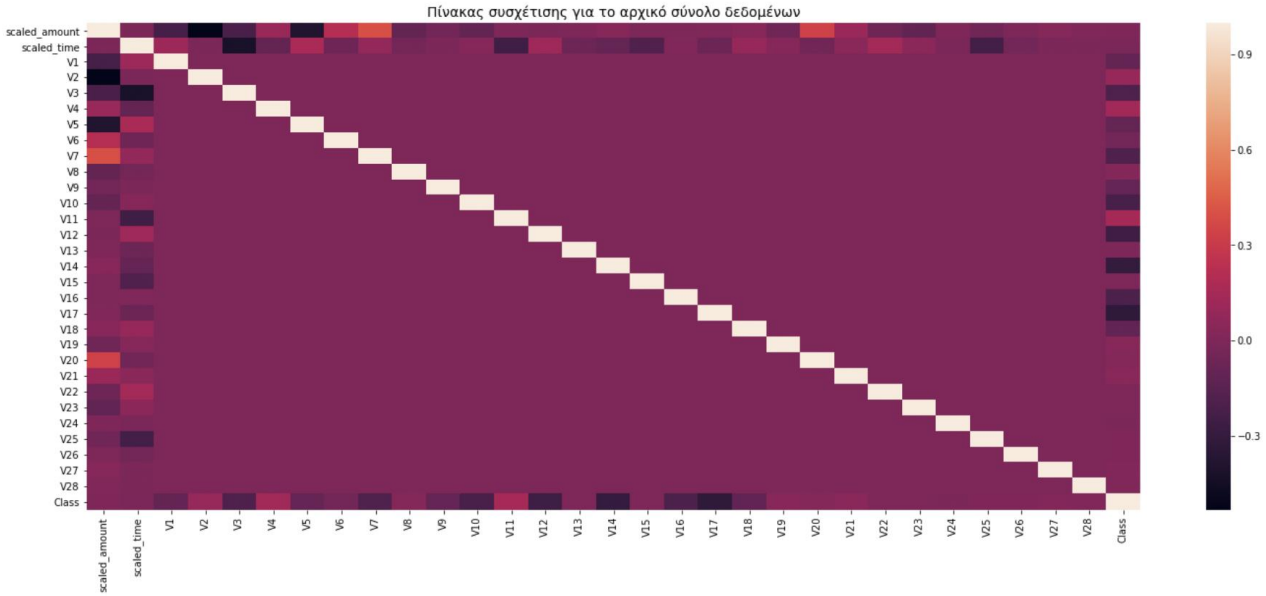
$$L2LossFunction = \sum_{i=1}^{n} (y_{true} - y_{predicted})^2$$

Γενικά, η λειτουργία απώλειας L2 προτιμάται στις περισσότερες περιπτώσεις. Αλλά όταν τα ακραία στοιχεία υπάρχουν στο σύνολο δεδομένων, τότε η συνάρτηση απώλειας L2 δεν έχει καλή απόδοση. Ο λόγος πίσω από αυτήν την κακή απόδοση είναι ότι εάν το σύνολο δεδομένων έχει ακραίες τιμές, τότε λόγω της εξέτασης των τετραγωνικών διαφορών οδηγεί σε πολύ μεγαλύτερο σφάλμα, όποτε ουσιαστικά ενισχύει την επίδραση των σημείων αυτών αντί να την ελαχιστοποιεί. Για αυτό και στην παρούσα εργασία έχουμε χρησιμοποιήσει σαν μέθοδο κανονικοποίησης το σφάλμα L1, μιας και το σύνολο μας περιλαμβάνει ακραίες τιμές.

5. Πειράματα

Αρχικά πριν την σύγκριση των αποτελεσμάτων στο επίπεδο ταξινόμησης μελετήσουμε γενικότερα την ποιότητα των παραγόμενων δεδομένων. Η μέτρηση αυτή της ποιότητας θα γίνει με διάφορους τρόπους. Αρχικά η ανάλυση της ποιότητας των νέων δεδομένων πρέπει να γίνει σε αντιπαραβολή με αυτή των αρχικών, έτσι ώστε να συγκρίνουμε αν οι μέθοδοι αυτοί βελτιώνουν ή όχι την ποιότητα των δεδομένων.

Η πρώτη μέθοδος που θα χρησιμοποιήσουμε για να προσπαθήσουμε να βγάλουμε συμπεράσματα σχετικά με τα δεδομένα μας είναι οι πίνακες συσχέτισης. Συγκεκριμένα υπολογίσαμε την ανά ζεύγη συσχέτιση (pairwise correlation) με χρήση της μεθόδου Pearson. Παρακάτω εμφανίζεται το διάγραμμα της συσχέτισης ανά στήλη για τα αρχικά, μη-ισορροπημένα δεδομένα.

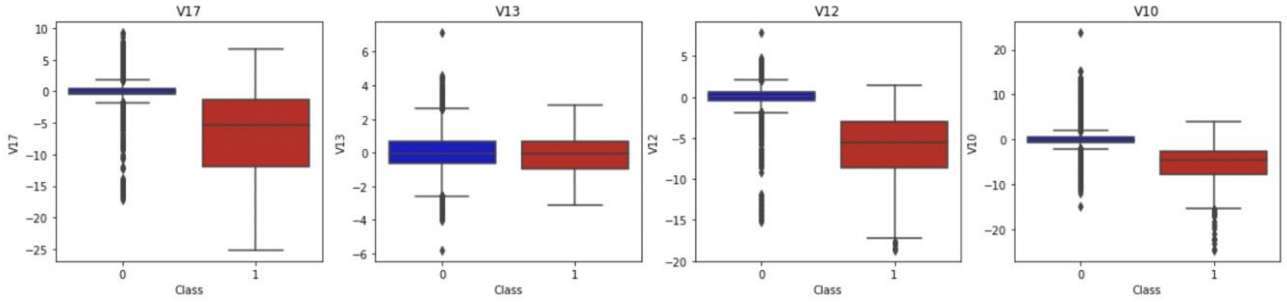


Εικόνα 25 Ο πίνακας συσχέτισης του αρχικού συνόλου δεδομένων.

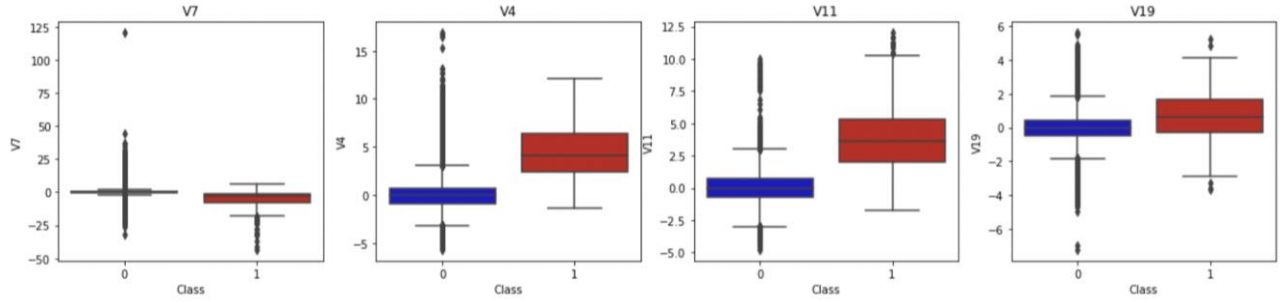
Σε κάθε πεδίο αυτού του πίνακα εμφανίζεται η συσχέτιση που έχει το πεδίο που αναφέρεται στην αντίστοιχη στήλη με το πεδίο που αναφέρεται στην αντίστοιχη γραμμή. Οπότε από το διάγραμμα αυτό μας αφορά κυρίως η γραμμή (ή στήλη) του πεδίου Class (τελευταία στήλη ή τελευταία γραμμή). Σε αυτό φαίνεται πόσο 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.

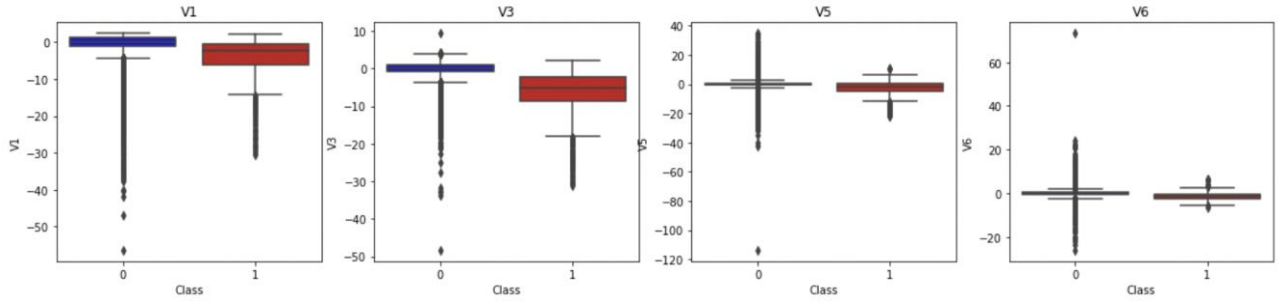


Εικόνα 26 Τα barplot των συσχετίσεων για τις μεταβλητές V17, V13, V12 και V10 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

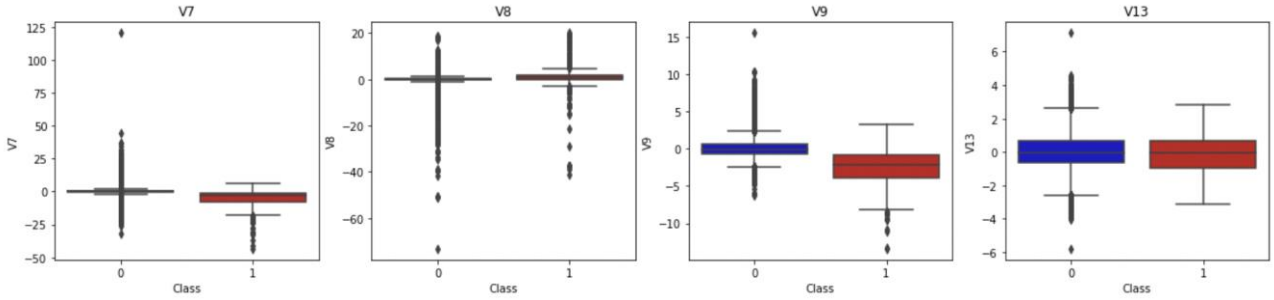


Εικόνα 27 Τα barplot των συσχετίσεων για τις μεταβλητές V7, V4, V11 και V19 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

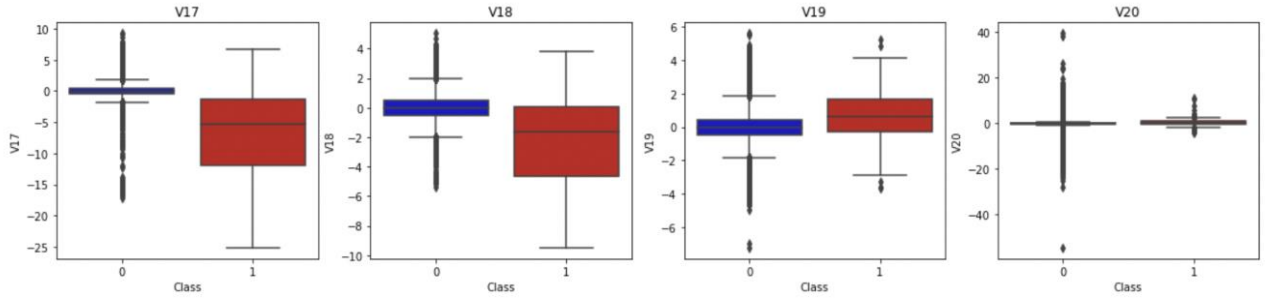
Στα παραπάνω δυο σχήματα εμφανίζονται τα διαγράμματα για τα πεδία που εμφανίζουν αρνητική και θετική συσχέτιση με την κλάση. Για να γίνει ξεκάθαρη σχέση που υπάρχει μεταξύ των πεδίων αυτών με τον τύπο της συναλλαγής παρακάτω παρουσιάζονται και τα αντίστοιχα διαγράμματα για όλα τα υπόλοιπα πεδία.



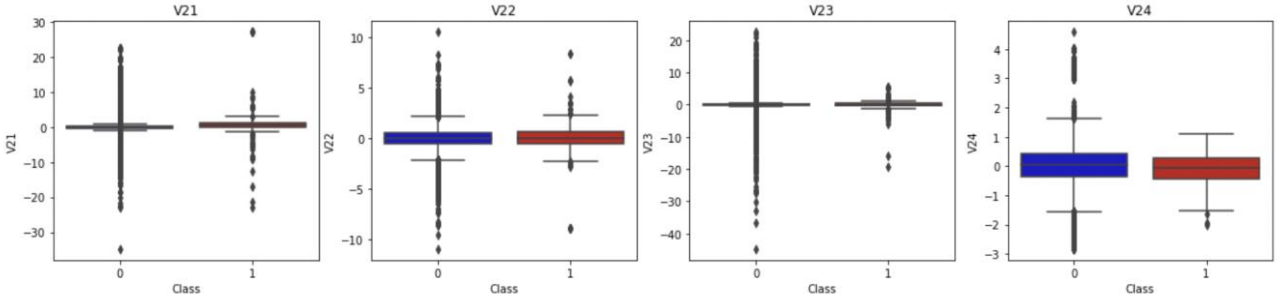
Εικόνα 28 Τα barplot των συσχετίσεων για τις μεταβλητές V1, V3, V5 και V6 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων



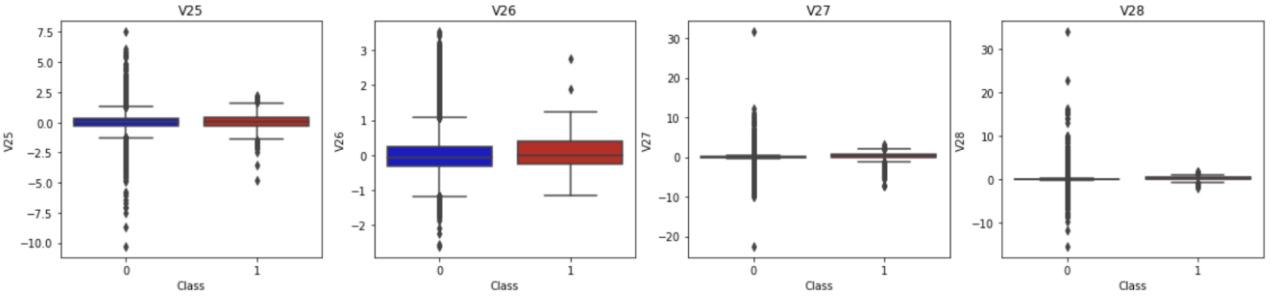
Εικόνα 29 Τα barplot των συσχετίσεων για τις μεταβλητές V7, V8, V9 και V13 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



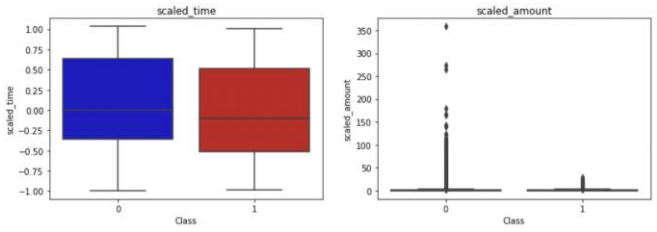
Εικόνα 30 Τα barplot των συσχετίσεων για τις μεταβλητές V17, V18, V19 και V20 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



Εικόνα 31 Τα barplot των συσχετίσεων για τις μεταβλητές V21, V22, V23 και V24 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.



Εικόνα 32 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.

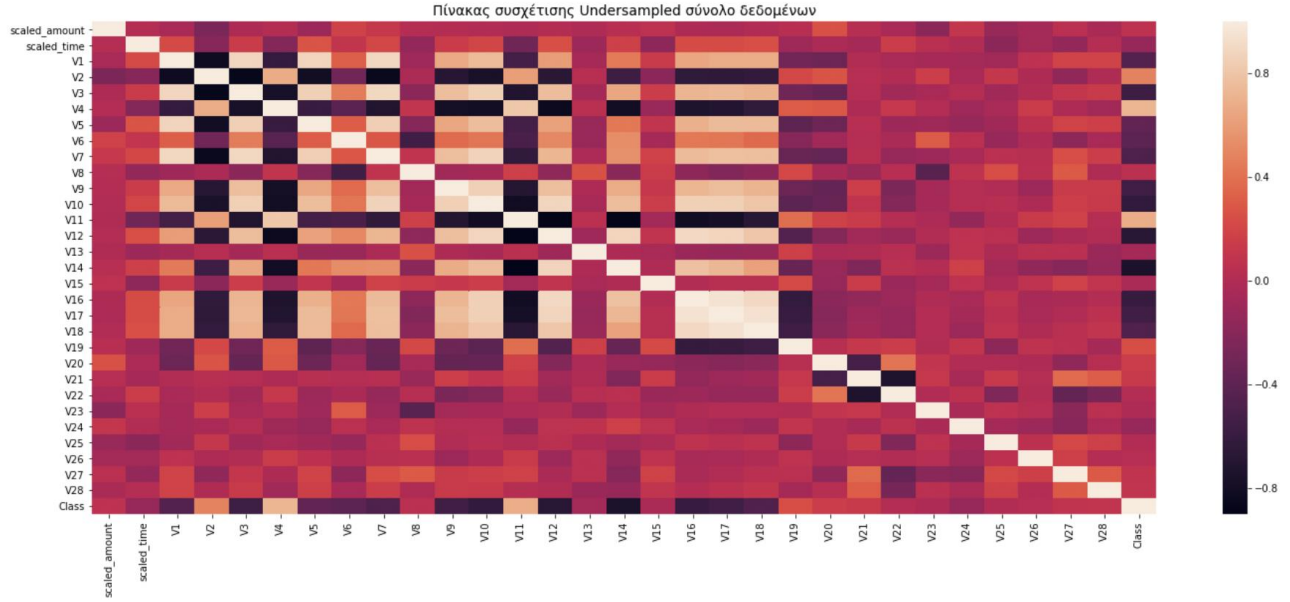


Εικόνα 33 Τα barplot των συσχετίσεων για τις μεταβλητές scaled_time και scaled_amount και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το αρχικό σύνολο δεδομένων.

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. Ένα παράδειγμα που μπορεί να συμβαίνει αυτό είναι το πρόβλημα ταξινόμησης του δείκτη μάζας σώματος. Από μόνα τους το ύψος ή βάρος δεν μπορούν κατατάξουν σωστά ένα άτομο αλλά ο συνδυασμός τους είναι μια από τις γνωστές μετρικές που χρησιμοποιείται σε όλο τον κόσμο. 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.

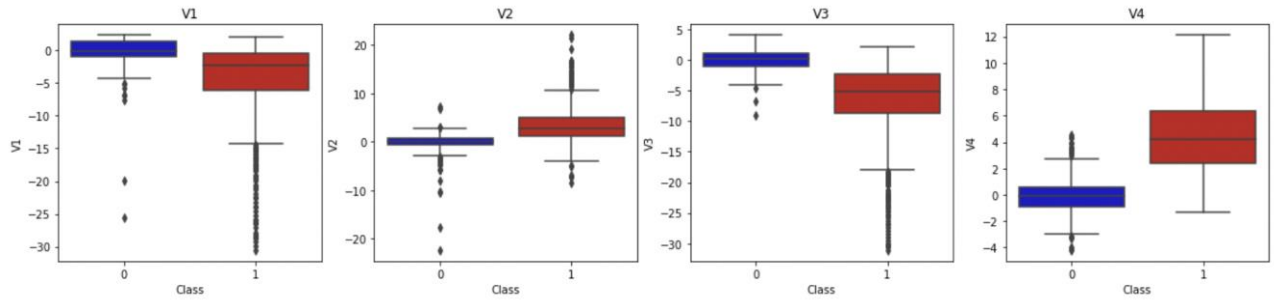
Στην συνέχεια αξίζει να μελετήσουμε τις αντίστοιχες συσχετίσεις για τα δεδομένα που παράγονται από κάθε μια από τις μεθόδους που χρησιμοποιήσαμε και να μελετήσουμε κατά πόσο αυτές επηρέασαν τις αρχικές τιμές που παρουσιάστηκαν παραπάνω. Θεωρητικά τα ισοκατανεμημένα δεδομένα θέλουμε να ακολουθούν την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

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

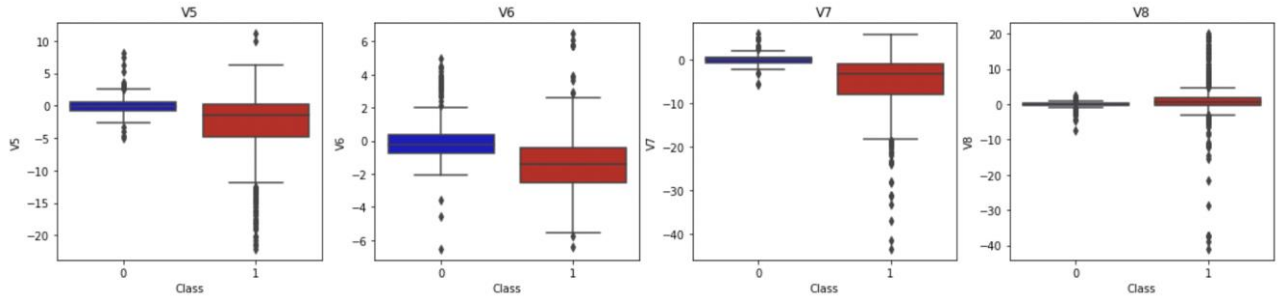


Εικόνα 34 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου undersampling.

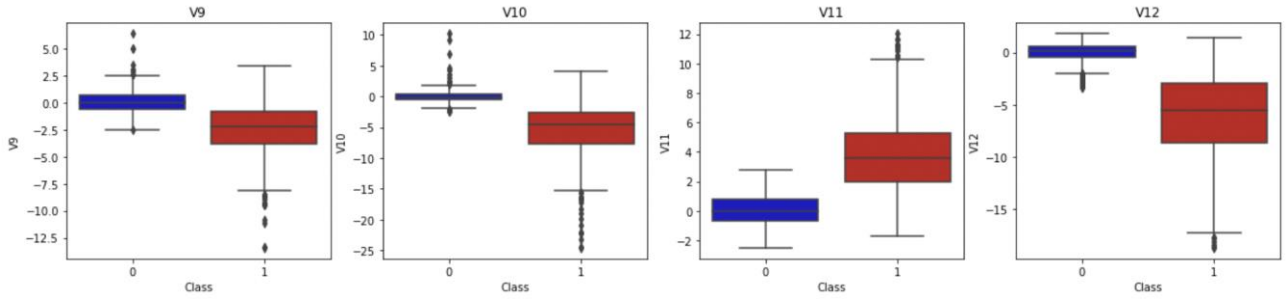
Από αυτό το διάγραμμα παρατηρείται μεγάλη διαφορά στις συσχετίσεις των πεδίων συγκριτικά με αυτές του ανισοκατανεμημένου συνόλου δεδομένων. Παρακάτω εμφανίζονται και τα αντίστοιχα box plots.



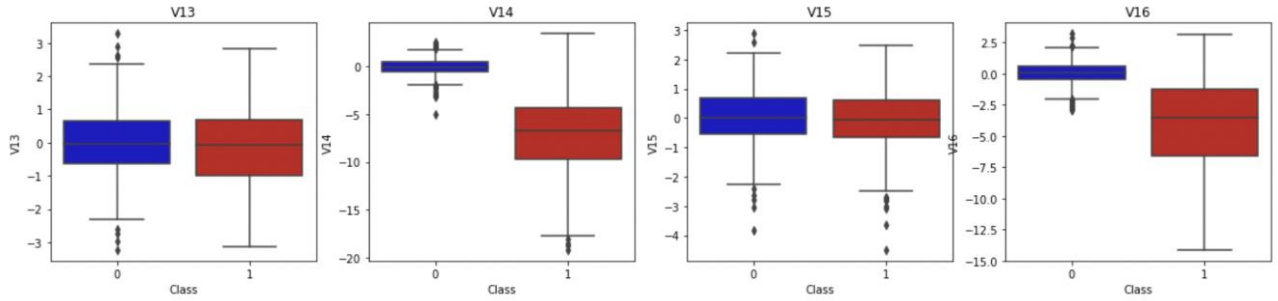
Εικόνα 35 Τα barplot των συσχετίσεων για τις μεταβλητές V1, V2, V3 και V4 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling



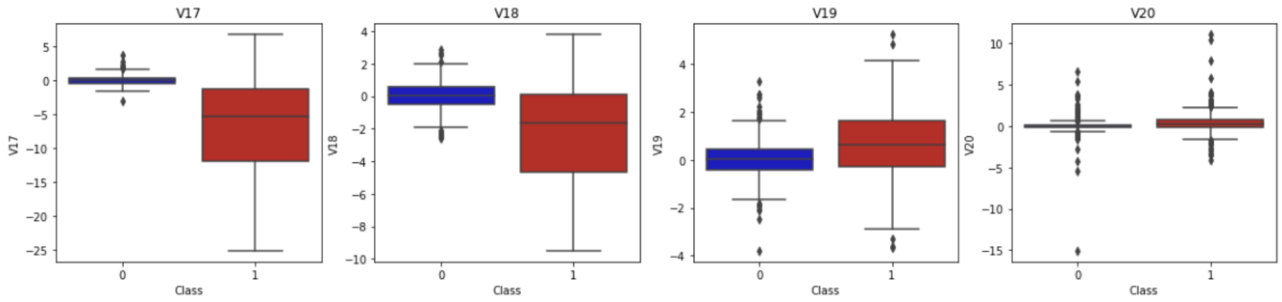
Εικόνα 36 Τα barplot των συσχετίσεων για τις μεταβλητές V5, V6, V7 και V8 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling.



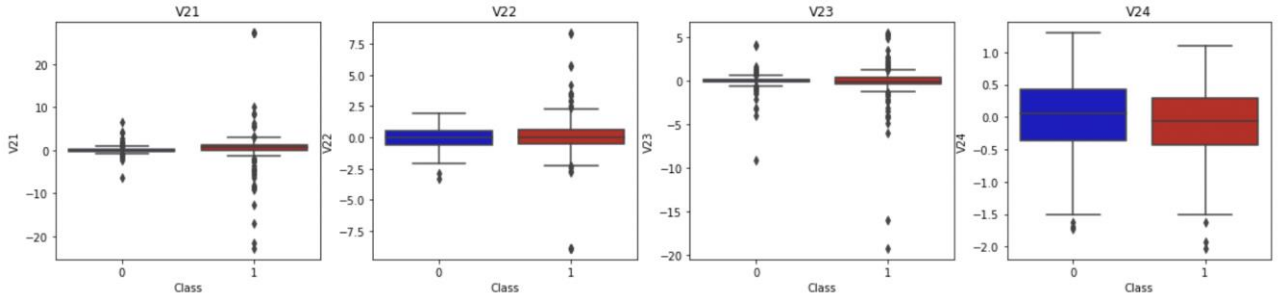
Εικόνα 37 Τα barplot των συσχετίσεων για τις μεταβλητές V9, V10, V11 και V12 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling.



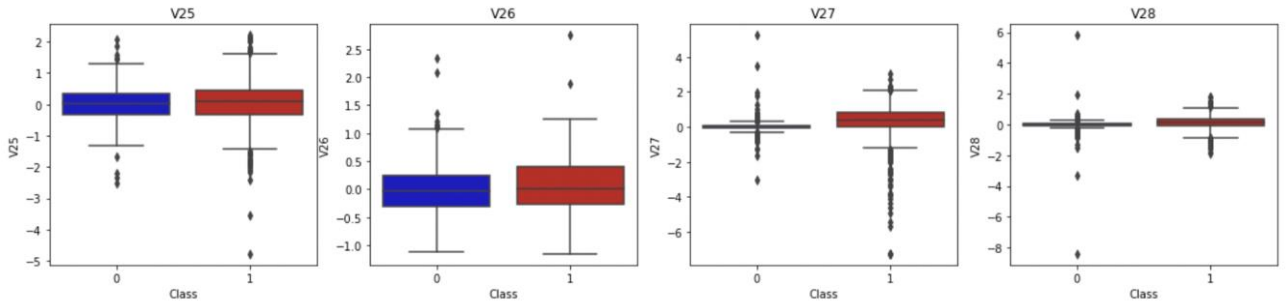
Εικόνα 38 Τα barplot των συσχετίσεων για τις μεταβλητές V13, V14, V15 και V16 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling.



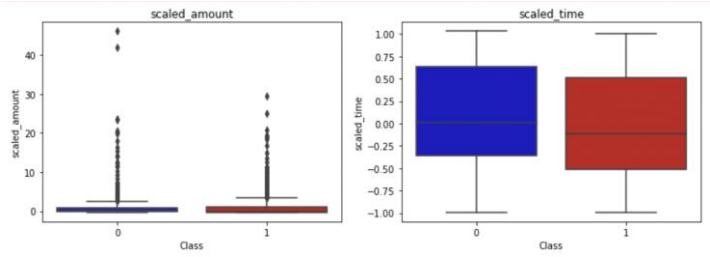
Εικόνα 39 Τα barplot των συσχετίσεων για τις μεταβλητές V17, V18, V19 και V20 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling.



Εικόνα 40 Τα barplot 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 undersampling.

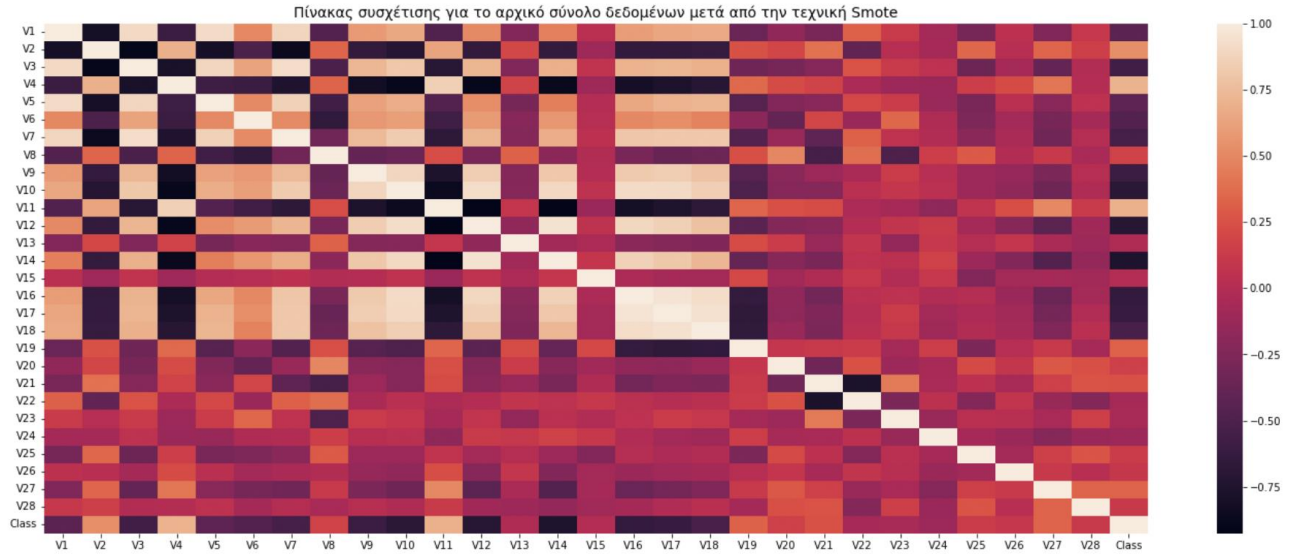


Εικόνα 41 Τα barplot των συσχετίσεων για τις μεταβλητές V25, V26, V27 και V28 και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling.



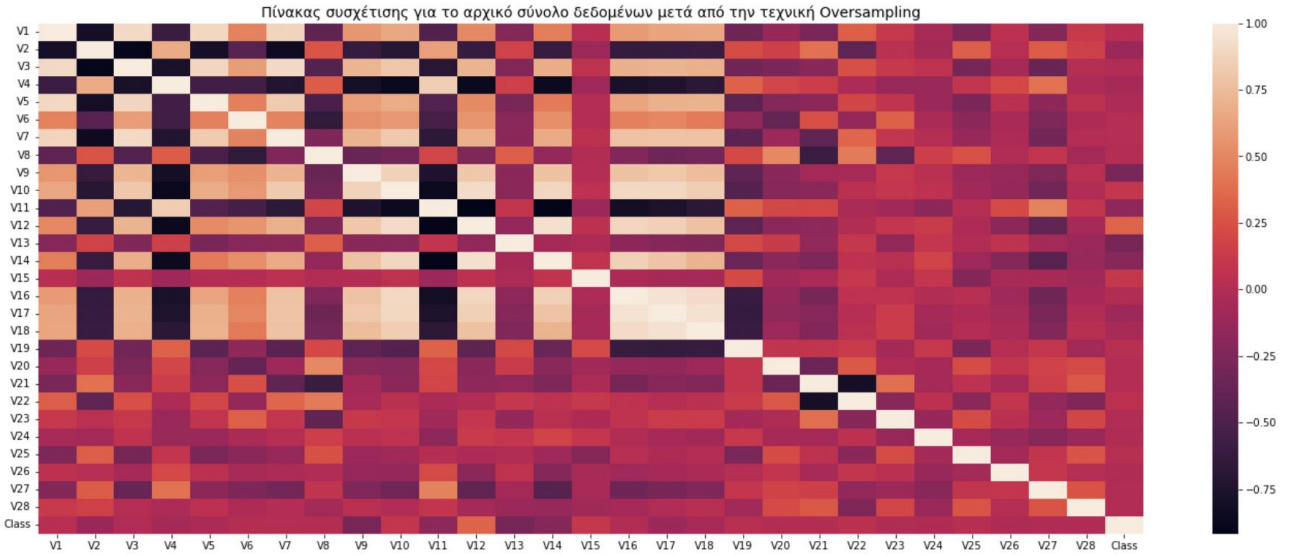
Εικόνα 42 Τα barplot των συσχετίσεων για τις μεταβλητές scaled_amount και scaled_time και τις κλάσεις του δείγματος (απάτη-όχι απάτη) για το σύνολο δεδομένων μετά την εφαρμογή της μεθόδου undersampling.

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

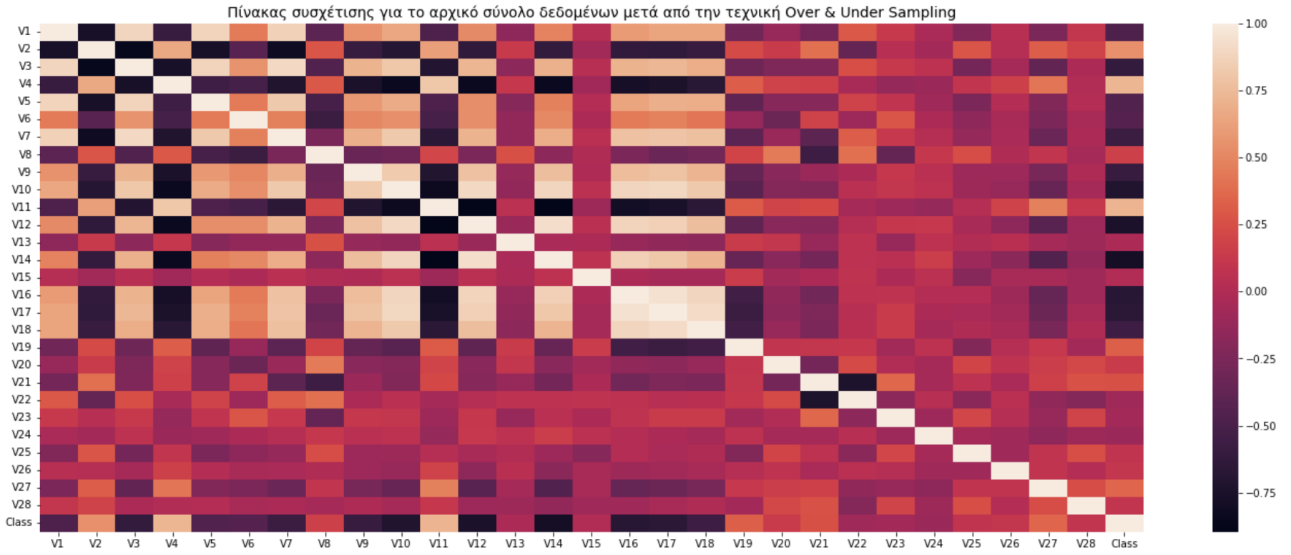


Εικόνα 43 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου smote.

Figure 43. Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου smote.



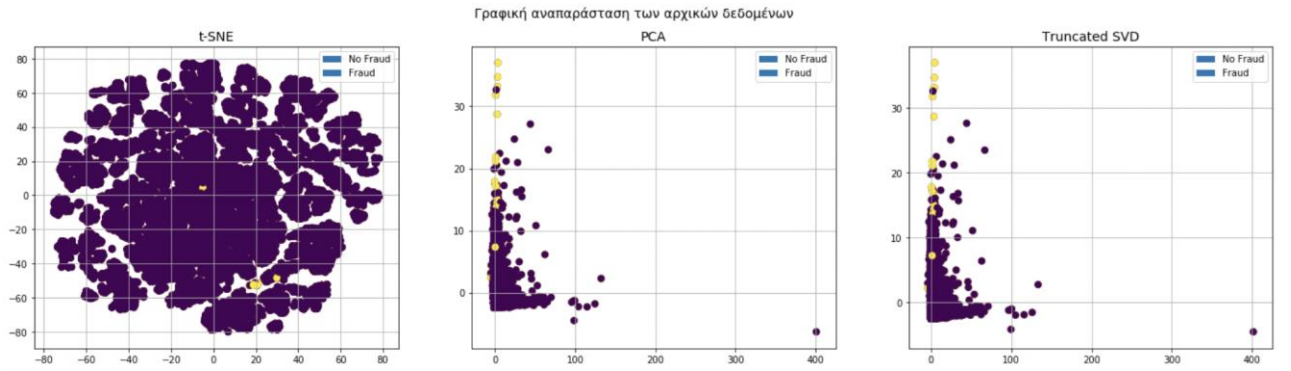
Εικόνα 44 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου oversampling.



Εικόνα 45 Ο πίνακας συσχέτισης για το σύνολο δεδομένων έπειτα από την εφαρμογή της μεθόδου over & undersampling.

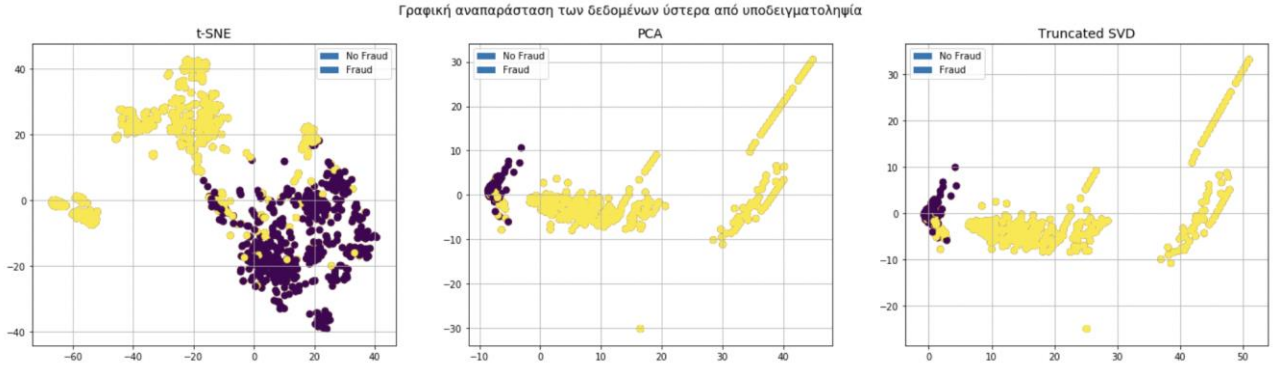
Το επόμενο βήμα είναι να δούμε κατά πόσο μπορούμε να οπτικοποιήσουμε τα δεδομένα ώστε να τα καταλάβουμε λίγο καλύτερα. Τα δεδομένα μας όμως είναι πολλαπλών διαστάσεων, πράγμα που καθιστά αδύνατο να τα σχεδιάσουμε στον χώρο. Για αυτόν τον λόγο πρώτα θα επιχειρήσουμε να μειώσουμε την διάσταση των δεδομένων σε 2 ώστε να μπορούμε να τα οπτικοποιήσουμε στον χώρο. Αυτό προφανώς δεν γίνεται χωρίς την απώλεια κάποιας πληροφορίας. Αν μπορούσαμε από όλες τις μεταβλητές να κρατήσουμε 2 για να τις σχεδιάσουμε στον χώρο και δεν χάναμε κάποια πληροφορία τότε αυτό θα ήταν ιδιαίτερα ανησυχητικό για την ποιότητα των δεδομένων μας.

Για να μειώσουμε τη διαστατικότητα των δεδομένων μας θα χρησιμοποιήσουμε 3 διαφορετικές μεθόδους και θα οπτικοποιήσουμε τα αποτελέσματά τους. Οι μέθοδοι που θα χρησιμοποιήσουμε είναι η PCA, η t-SNE και η Truncated SVD. Τα αποτελέσματα των παραπάνω μεθόδων για τα αρχικά δεδομένα εμφανίζονται στο παρακάτω διάγραμμα:



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

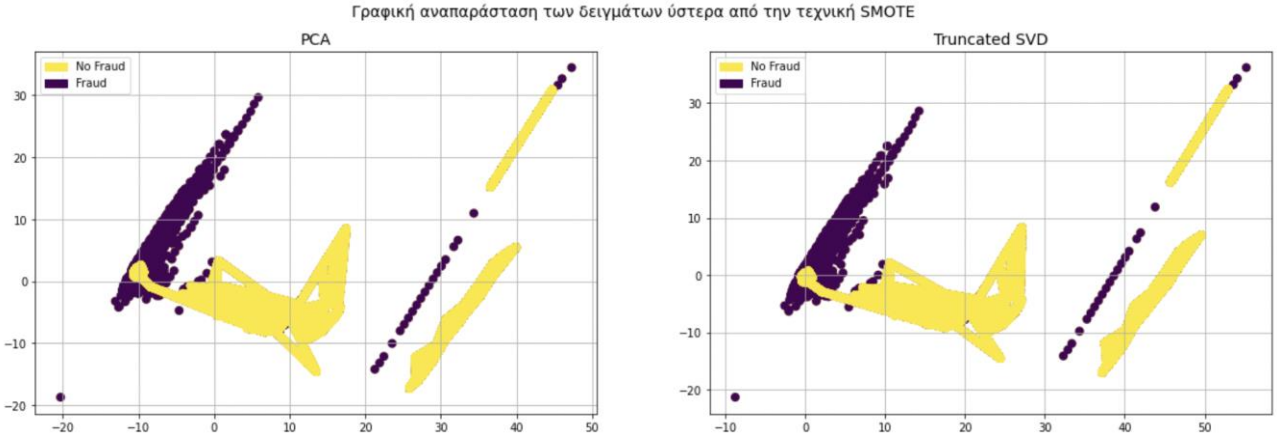
Τα δεδομένα τα οποία έχουν υποστεί υποδειγματοληψία εμφανίζονται στο παρακάτω διάγραμμα.



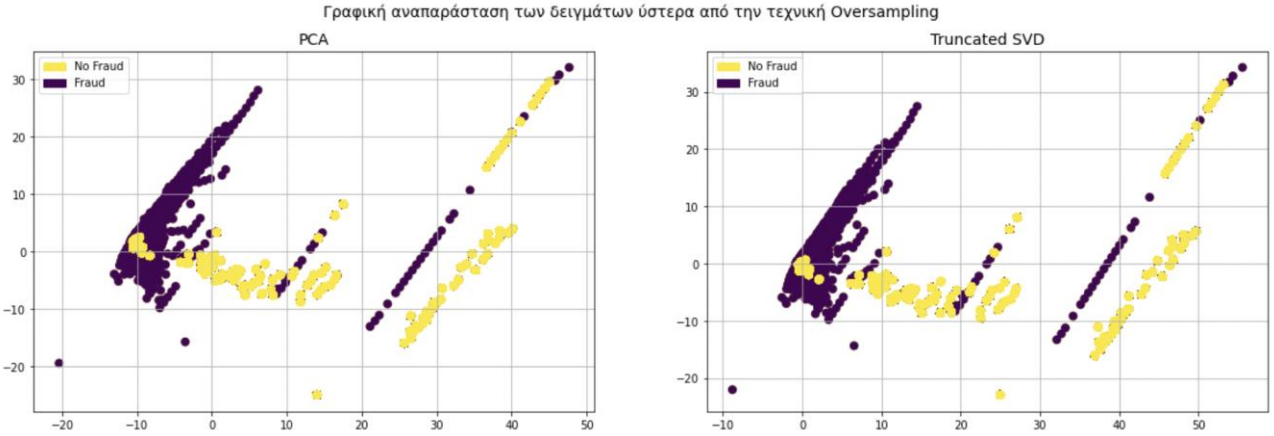
Εικόνα 47 Η γραφική αναπαράσταση των δειγμάτων του αρχικού συνόλου δεδομένων έπειτα από την εφαρμογή 3 μεθόδων μείωσης της διαστατικότητάς τους.

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 της διαστατικότητας 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.

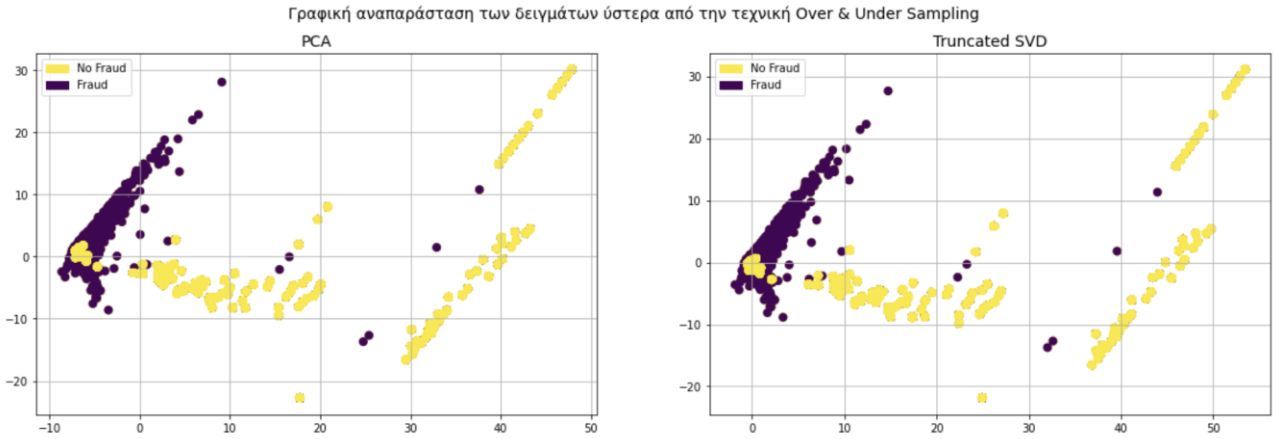
Παρακάτω παρουσιάζονται οι αντίστοιχες γραφικές αναπαραστάσεις ύστερα από την εφαρμογή κάθε μιας μεθόδου.



Εικόνα 48 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 49 Η γραφική αναπαράσταση των δειγμάτων του συνόλου δεδομένων μετά την εφαρμογή της μεθόδου της υπερδειγματοληψίας έπειτα από την εφαρμογή 3 μεθόδων μείωσης της διαστατικότητας τους.



Εικόνα 50 Η γραφική αναπαράσταση των δειγμάτων του συνόλου δεδομένων μετά την εφαρμογή της μεθόδου της υπερ & υποδειγματοληψίας έπειτα από την εφαρμογή 3 μεθόδων μείωσης της διαστατικότητας τους.

Από τα διαγράμματα των 2 μεθόδων βλέπουμε 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 αναπαράσταση 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.

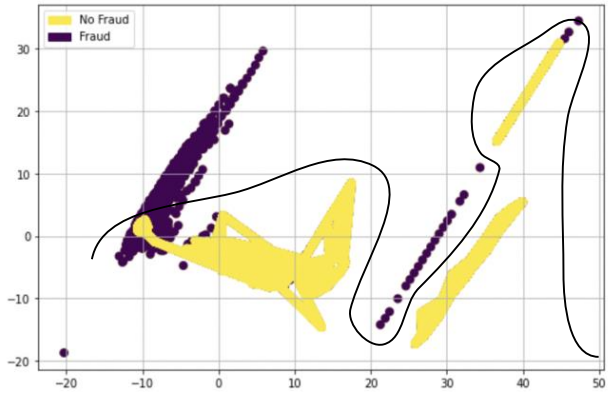
Η αξιολόγηση της αποδοτικότητας των μεθόδων αυτών έγινε με τη δοκιμή των δεδομένων πριν και ύστερα από την εφαρμογή κάθε μεθόδου σε διαφορετικές αρχιτεκτονικές ταξινόμησης. Αξίζει να σημειωθεί ότι το σύνολο δοκιμής (test set) παρέμεινε σταθερό και αμετάβλητο για κάθε ένα πείραμα. Οπότε στο σύνολο δοκιμής, 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:

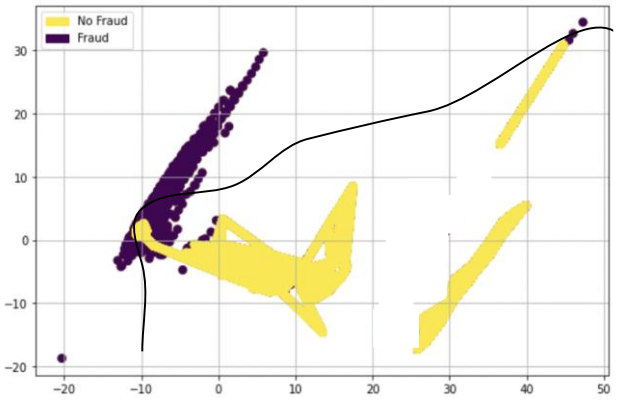
- $a > b$, τότε η μέθοδος δεν βοήθησε την εκπαίδευση του συστήματος
- $a < b$, τότε η μέθοδος βελτίωσε την απόδοση της ταξινόμησης
- $a == b$, τότε η μέθοδος άφησε ανεπηρέαστα τα αποτελέσματα

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 δύο διαφορετικά παραδείγματα.



Εικόνα 51 Η καμπύλη ταξινόμησης για ένα τεχνητό σύνολο δεδομένων.



Εικόνα 52 Η 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 την μέθοδο 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.

Στη 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 περιγράφηκε 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%

Εικόνα 53 Συγκριτικός πίνακας της μέσης απόδοσης κάθε αλγορίθμου (εκτός συνθετικούς οι όποιοι εμφανίζονται παρακάτω) για όλα τα μοντέλα.

6. Δημιουργικά Νευρωνικά Δίκτυα

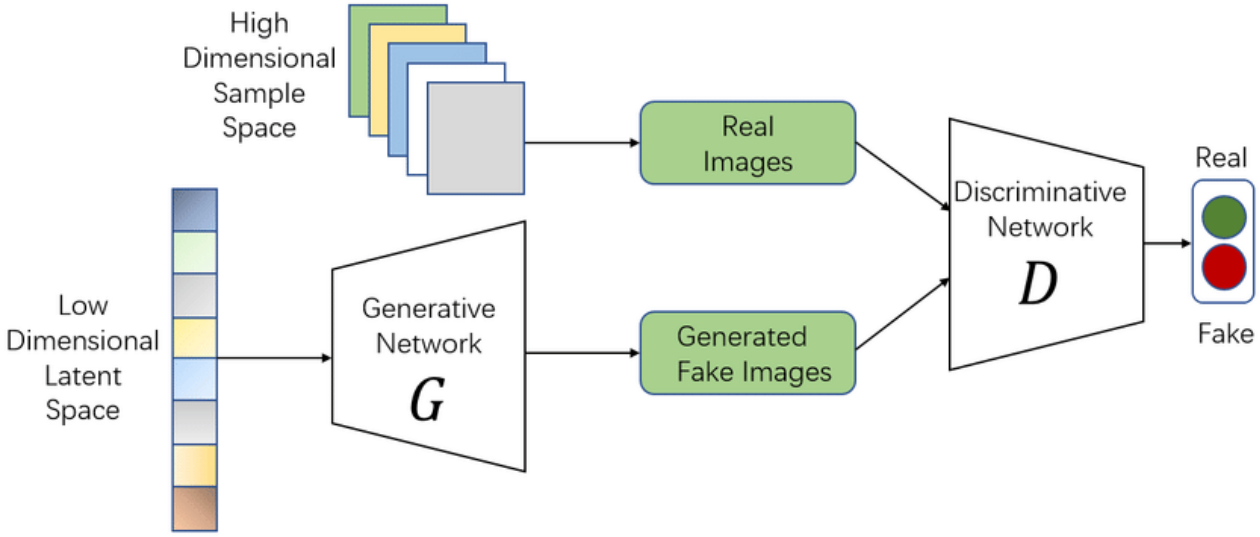
Ο τελευταίος τρόπος που θα δοκιμάσουμε να λύσουμε το πρόβλημα της ανισοκατανομής των δεδομένων είναι με τη χρήση δημιουργικών νευρωνικών δικτύων. Αυτή η οικογένεια νευρωνικών δικτύων είναι ευρέως γνωστή για τα αποτελέσματά της σε σημείο που πλέον η ερευνά στο πεδίο έχει στραφεί στην ανάπτυξη αλγορίθμων διαχωρισμού πραγματικών από συνθετικά δεδομένα.

Η πλέον γνωστή αρχιτεκτονική αυτής της οικογένειας είναι τα Generative Adversarial Networks (GANs) η αρχιτεκτονική των όποιων παρουσιάζεται στο παρακάτω σχήμα. Η αρχιτεκτονική αυτή αποτελείται από 2 νευρωνικά δίκτυα: τον generator, ο όποιος δέχεται στην είσοδο ένα διάνυσμα θορύβου (μικρό σε μέγεθος) και τον discriminator, ο όποιος δέχεται στην είσοδο μια εικόνα και πρέπει να απαντήσει αν είναι πραγματική ή συνθετική. Η εκπαίδευση του αλγορίθμου γίνεται μέσω του αλγορίθμου min-max και είναι ανταγωνιστική. Αυτό συμβαίνει γιατί σκοπός του discriminator είναι να μάθει να ξεχωρίζει όσο το δυνατόν καλύτερα συνθετικές από πραγματικές εικόνες, ενώ του generator να παράγει όσο το δυνατόν εικόνες πιο κοντινές στην κατανομή των δεδομένων εκπαίδευσης. Κατά συνέπεια, όσο καλύτερος γίνεται ο generator τόσο μεγαλώνει το loss στον discriminator όποτε μαθαίνει να ξεχωρίζει καλύτερα τις εικόνες και το αντίθετο. Μετά την εκπαίδευση ο discriminator δεν χρησιμοποιείται άλλο και για τη σύνθεση δεδομένων χρησιμοποιείται μόνο ο generator στον όποιο δίνεται σαν είσοδος ένα τυχαίο διάνυσμα θορύβου.

Σε αυτή την αρχιτεκτονική αξίζει να σημειωθούν δυο βασικές σχεδιαστικές επιλογές. Η πρώτη αφορά στο γεγονός ότι ο χρήστης δεν έχει κανένα έλεγχο στη σύνθεση των δεδομένων. Για παράδειγμα, αν έχουμε ένα δημιουργικό δίκτυο το όποιο παράγει εικόνες από προφίλ ατόμων, τότε κατά τη δημιουργία δεν μπορούμε να επικοινωνήσουμε με κάποιον τρόπο στο δίκτυο να παράγει άτομα με μπλε μάτια. Στην εξερεύνηση του χώρου εισόδου του generator για τον έλεγχο της παραγωγής έχουν προταθεί διάφοροι αλγόριθμοι, σκοπός των όποιων είναι να προσπαθήσουν να κατευθύνουν τη σύνθεση. Ο πλέον γνωστός τέτοιος αλγόριθμος είναι αυτός του GANSPACE ο οποίος χρησιμοποιεί την μέθοδο PCA για να ενώσει ορισμένα χαρακτηριστικά εισόδου με συγκεκριμένα χαρακτηριστικά των εικόνων. Για παράδειγμα, προσπαθεί να εξάγει κανόνες της μορφής: οι θέσεις της εισόδου 120-125 καθορίζουν το χρώμα των ματιών του προσώπου, ενώ οι 127-140 το χρώμα των μαλλιών.

Το δεύτερο σημείο το όποιο αξίζει να σημειωθεί είναι ο τρόπος εκπαίδευσης αυτών των δικτύων, ο όποιος αποτέλεσε και την κινητήρια δύναμη των αρχιτεκτονικών αυτών. Ο generator δεν εκπαιδεύεται άμεσα στην κατανομή των δεδομένων αλλά εκπαιδεύεται έμμεσα μέσω του loss του generator. Οπότε αυτός δεν μαθαίνει να παράγει εικόνες ελαχιστοποιώντας το loss-likelihood του dataset αλλά μεγιστοποιώντας το error του discriminator. Στην εξερεύνηση του χώρου εισόδου του

generator για τον έλεγχο της παραγωγής έχουν προταθεί διάφοροι αλγόριθμοι, σκοπός των όποιων είναι να προσπαθήσουν να κατευθύνουν την σύνθεση.



Εικόνα 54 Η τυπική αρχιτεκτονική ενός GAN (τύπος δημιουργικού νευρωνικού δικτύου).

Όπως αναφέρθηκε και προηγουμένως, ο discriminator δέχεται σαν είσοδο ένα διάνυσμα τυχαίου θορύβου το όποιο συνήθως είναι πολύ μικρότερο από το παραγόμενο αντικείμενο. Στο παραπάνω παράδειγμα, με τις εικόνες, ο generator μπορεί να δέχεται σαν είσοδο ένα διάνυσμα 128 - 512 διαστάσεων ενώ η παραγόμενη εικόνα να είναι 256*256*3 = 196.608 διαστάσεων.
Οπότε σκοπός του generator είναι να εμπλουτίσει τον θόρυβο αυτό με τιμές ούτως ώστε να δημιουργηθεί μια ρεαλιστική εικόνα.

Σε αυτό το σημείο προκύπτει το εξής ερώτημα: Θα μπορούσαμε να κάνουμε το αντίστροφο, δηλαδή σε ένα τέτοιο δίκτυο να δίνουμε σαν είσοδο μια εικόνα και να παίρνουμε ένα dense διάνυσμα χαρακτηριστικών; Αυτή η τεχνική θα ήταν ιδιαιτέρα χρήσιμη γιατί θα είχαμε έναν τρόπο να μεταφέρουμε διανύσματα υψηλού επιπέδου σε έναν πυκνό χώρο στον όποιον θα περιέχεται αποκλειστικά η χρήσιμη πληροφορία (για τα ‘μάτια’ του generator). Αν είχαμε αυτά τα πυκνά χαρακτηριστικά θα μπορούσαμε να εκπαιδεύσουμε τον ταξινομητή μας σε αυτόν τον χώρο αντί σε ένα χώρο με πολύ μεγαλύτερες διαστάσεις και ίσως περιττή (redundant) πληροφορία.

Σε αυτό το πρόβλημα έχουν προταθεί διάφοροι heuristics αλγόριθμοι για την αντιστροφή τέτοιων δικτύων. Οι πιο γνωστοί χρησιμοποιούν ένα τρίτο δίκτυο είτε παράλληλα με την εκπαίδευση είτε εκ των υστέρων (post-hoc) για να κάνει αυτή την αντιστροφή. Άλλες μέθοδοι προσεγγίζουν το πρόβλημα ως βελτιστοποίησης (optimazation problem) και χρησιμοποιούν ανάλογες τεχνικές για να αντιστρέψουν κάθε δείγμα. Η αλήθεια όμως είναι ότι η αντιστροφή των δικτύων αυτών είναι ένα υπολογιστικά πολύ δύσκολο πρόβλημα. Συγκεκριμένα, καλούμαστε να απαντήσουμε στην ερώτηση

αν ένα δείγμα έχει παραχθεί από ένα δίκτυο ανήκει στην κλάση NP-Hard. Οπότε το να βρούμε και την είσοδο η όποια δημιουργεί το δείγμα αυτό είναι σίγουρα τουλάχιστον αντίστοιχα δύσκολο.

Σκοπός μας είναι να καταφέρουμε να δημιουργήσουμε δεδομένα που θεωρούνται απάτες, τα όποια θα καταφέρουν να διατηρούν την κατανομή των αρχικών δεδομένων και παράλληλα να βελτιώσουν τα αποτελέσματα των δεδομένων μας. Ιδανικά για τον σκοπό αυτόν θα θέλαμε 2 δημιουργικά δίκτυα, ένα για να παράγει δεδομένα απάτης και ένα κανονικά. Αυτό όμως στην περίπτωση μας που έχουμε δεδομένα τα όποια δεν είναι ισοκατανεμημένα δεν θα έχει τα επιθυμητά αποτελέσματα και ο λόγος είναι ότι το δίκτυο το όποιο θα παράγει δεδομένα απάτης θα εκπαιδευτεί σε ελάχιστα δεδομένα σε σχέση με το δεύτερο δίκτυο. Έτσι αναμένεται η απόδοσή του να είναι πολύ μικρότερη σε σχέση με το αρχικό και τελικά τα καινούργια δεδομένα να μπερδέψουν το σύστημα που κάνει το classification. Οπότε πρέπει να σχεδιαστεί μια διαφορετική στρατηγική.

Ιδανικά θα θέλαμε ένα δημιουργικό νευρωνικό δίκτυο στο οποίο θα έχουμε έλεγχο στη σύνθεση π.χ. να δίνεται σαν είσοδος η κλάση που θέλουμε να ανήκει το παραγόμενο δείγμα. Τέτοιου είδους νευρωνικά δίκτυα έχουν προταθεί στη βιβλιογραφία, όπως το cycle gan, και επιτυγχάνουν εξαιρετικά αποτελέσματα, αλλά όταν τα δεδομένα εισόδου είναι μεγάλα σε όγκο για όλες τις παραγόμενες κλάσεις, πράγμα που στην περίπτωση μας δεν ισχύει. Το Cycle-GAN εισάγει μια πρόσθετη λειτουργία στο loss του (Cycle-Consistency) για να βοηθήσει στη σταθεροποίηση της εκπαίδευσης ενός GAN. Ο όρος αυτός εφαρμόζεται κυρίως σε εφαρμογές μετάφρασης εικόνας σε εικόνα. Ωστόσο, το CycleGAN μαθαίνει να μεταφράζει από μια κλάση εικόνων σε μια άλλη, όπως για παράδειγμα από άλογα σε ζέβρες. Αυτό υλοποιείται μέσω συναρτήσεων απώλειας συνέπειας προς τα εμπρός και προς τα πίσω. Μια γεννήτρια παίρνει εικόνες αλόγων και μαθαίνει να τις χαρτογραφεί σε ζέβρες έτσι ώστε ο χρήστης να μην μπορεί να πει αν ήταν αρχικά μέρος του συνόλου ζέβρας ή όχι, όπως συζητήθηκε παραπάνω. Μετά από αυτό, οι ζέβρες που δημιουργούνται από εικόνες αλόγων περνούν μέσω ενός δικτύου που τις μεταφράζει ξανά σε άλογα. Ένας δεύτερος διαχωριστής καθορίζει εάν αυτή η επαναμεταφρασμένη εικόνα ανήκει στο σύνολο αλόγων ή όχι. Και οι δύο αυτές απώλειες διάκρισης αθροίζονται για να σχηματίσουν την απώλεια συνέπειας κύκλου.

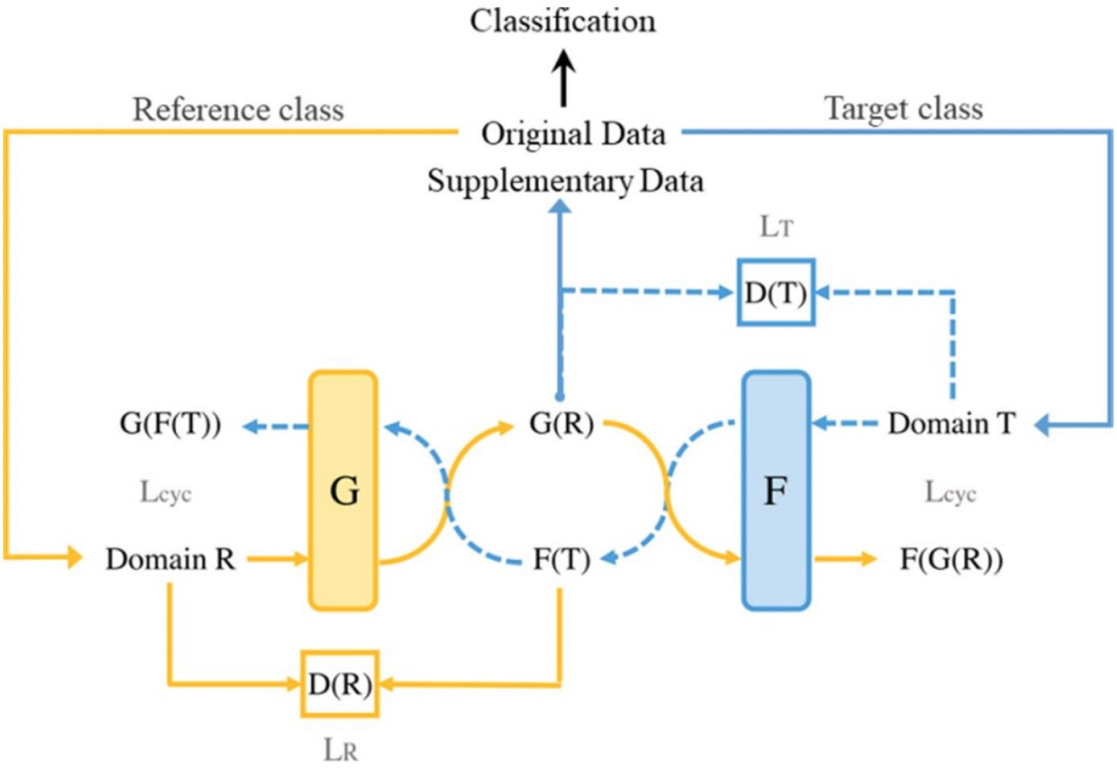
Η χρήση των CycleGAN δοκιμάστηκε στο έργο της Ταξινόμησης Συναισθημάτων, χρησιμοποιώντας το σύνολο δεδομένων αναγνώρισης συναισθημάτων, FER2013, Facial Expression Recognition Database, όπου κατασκεύασαν έναν ταξινομητή CNN για να αναγνωρίσουν 7 διαφορετικά συναισθήματα: θυμό, αηδία, φόβο, χαρούμενο, λύπη, έκπληξη και ουδέτερο. Αυτές οι κατηγορίες είναι μη ισοκατανεμημένες και το CycleGAN χρησιμοποιείται ως μέθοδος έξυπνης υπερδειγματοληψίας.

Τα CycleGAN έμαθαν μια μη ζευγαρωμένη μετάφραση εικόνας σε εικόνα μεταξύ τομέων. Ένα παράδειγμα των κλάσεων σε αυτό το πρόβλημα είναι από “neutral” σε “disgust”. Το CycleGAN

μαθαίνει να μεταφράζει μια εικόνα που αντιπροσωπεύει μια ουδέτερη εικόνα σε μια εικόνα που αντιπροσωπεύει το συναίσθημα αηδίας όπως φαίνεται στο παρακάτω σχήμα.

Εικόνα 55 Ένα δείγμα για τον τρόπο εναλλαγής συναισθήματος σε εικόνες μέσω της χρήσης ενός CycleGan.

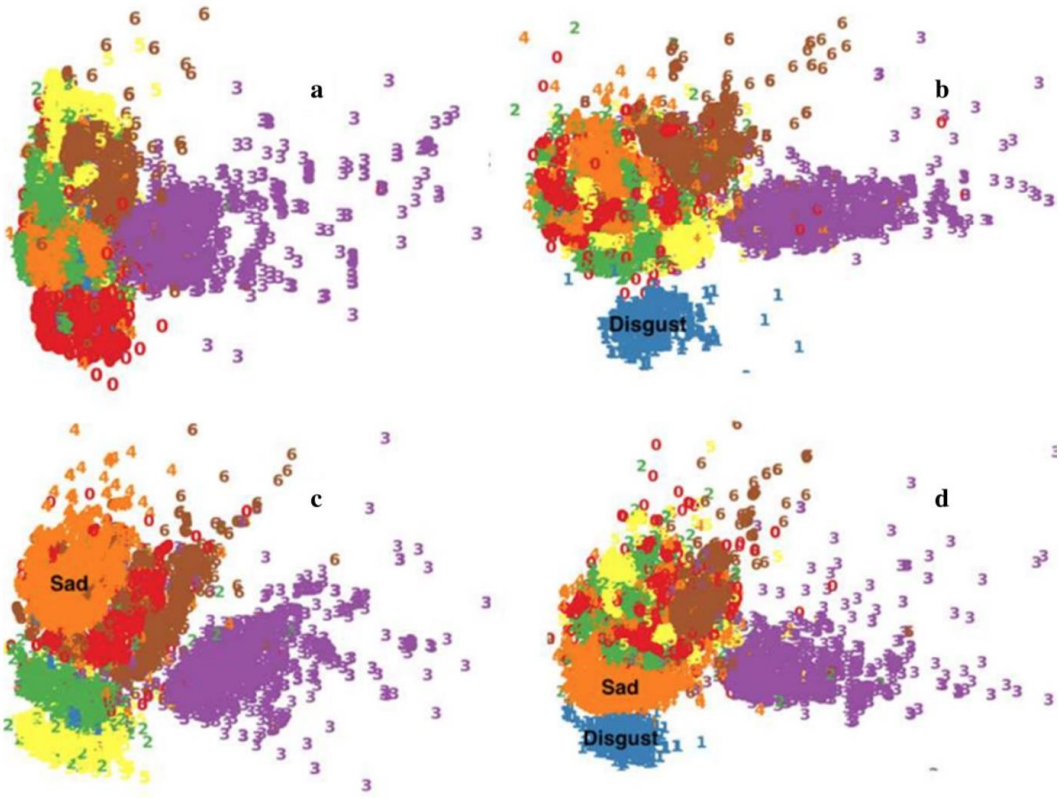
Η αρχιτεκτονική ενός τέτοιου μοντέλου εμφανίζονται στο παρακάτω διάγραμμα.



Εικόνα 56 Η αρχιτεκτονική για την σύνθεση δειγμάτων διαφορετικής κλάσης με βάση κάποιο στιγμιότυπο εισόδου μέσω της χρήσης ενός CycleGan.

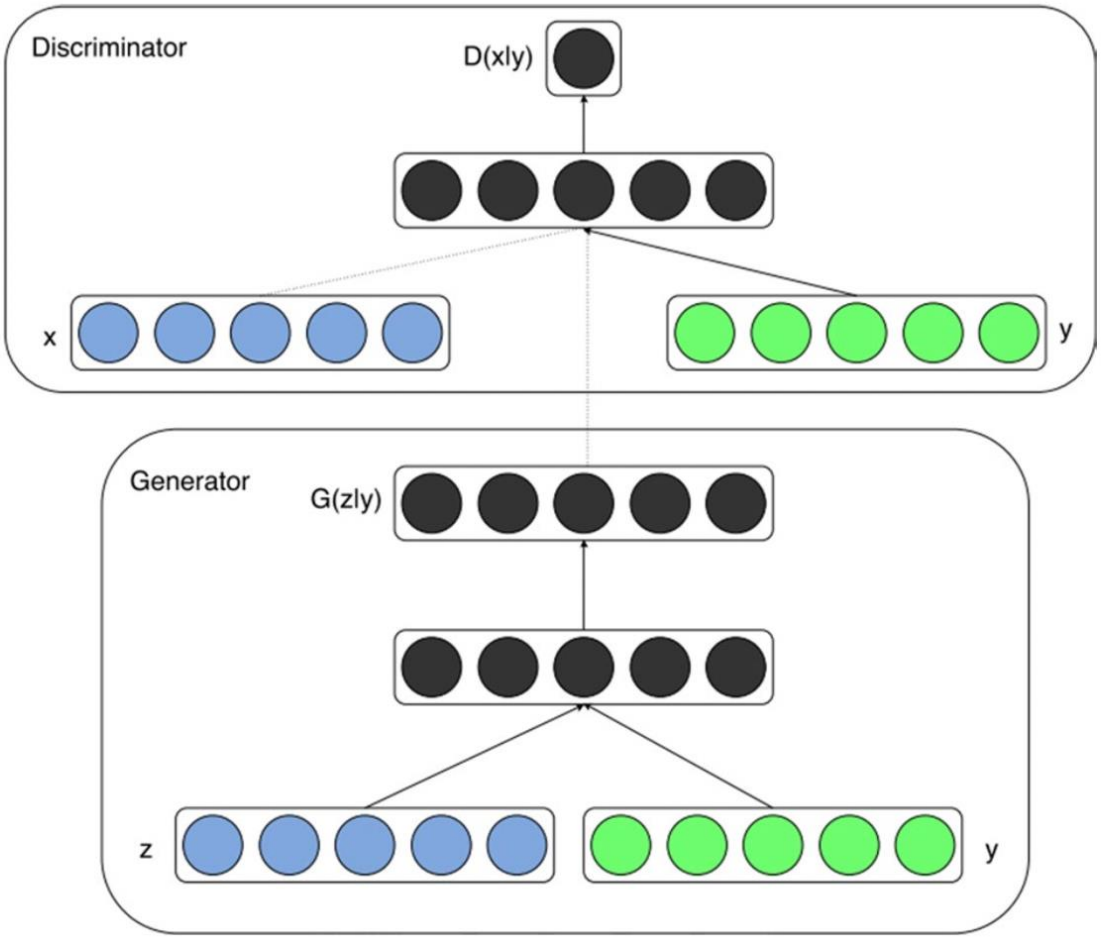
Η χρήση του CycleGAN για τη μετάφραση εικόνων από τις άλλες 7 τάξεις στις τάξεις μειοψηφίας ήταν πολύ αποτελεσματική στη βελτίωση της απόδοσης του μοντέλου CNN στην αναγνώριση συναισθημάτων. Χρησιμοποιώντας αυτές τις τεχνικές, η ακρίβεια βελτιώθηκε κατά 5-10%. Για την περαιτέρω κατανόηση της αποτελεσματικότητας της προσθήκης παρουσιών που δημιουργούνται από το GAN, χρησιμοποιείται μια οπτικοποίηση t-SNE. Το t-SNE [20] είναι μια τεχνική οπτικοποίησης που μαθαίνει να χαρτογραφεί μεταξύ διανυσμάτων υψηλών διαστάσεων σε

χώρο χαμηλής διάστασης για να διευκολύνει την απεικόνιση των ορίων απόφασης, όπως φαίνεται στο παρακάτω σχήμα.



Εικόνα 57 Ο τρόπος με τον όποιον ένα gan διαχωρίζει εσωτερικά του τα στιγμιότυπα των διαφορετικών κλάσεων.

Μια άλλη ενδιαφέρουσα αρχιτεκτονική GAN για χρήση στην Επαύξηση Δεδομένων είναι τα Conditional GAN [21]. Τα υπό όρους GAN προσθέτουν ένα διάνυσμα υπό όρους τόσο στη γεννήτρια όσο και στη διάταξη διάκρισης προκειμένου να αμβλύνουν τα προβλήματα με την κατάρρευση της λειτουργίας. Εκτός από την εισαγωγή ενός τυχαίου διανύσματος z στη γεννήτρια, τα υπό όρους GAN εισάγουν επίσης ένα διάνυσμα y το οποίο θα μπορούσε να είναι κάτι σαν μια κωδικοποιημένη ετικέτα κλάσης με ένα hot, π.χ. [0 0 0 1 0]. Αυτή η ετικέτα κλάσης στοχεύει μια συγκεκριμένη κλάση για τη γεννήτρια και τη ταξινόμηση. Η αρχιτεκτονική ενός τέτοιου δικτύου εμφανίζεται στο παρακάτω διάγραμμα.



Εικόνα 58 Η τυπική αρχιτεκτονική ενός conditional gan.

Μια διαφορετική λύση είναι η κατασκευή ενός δημιουργικού νευρωνικού δικτύου το όποιο θα εκπαιδευτεί σε όλα τα δεδομένα και στην συνέχεια να ελέγξουμε την παραγωγή post-hoc επεμβαίνοντας στον λανθάνοντα χώρο (latent space). Για παράδειγμα, αν εκπαιδεύσουμε ένα τέτοιο δίκτυο για το όποιο γνωρίζουμε ότι από ένα αρχικό σημείο στον latent space μπορούμε να παράγουμε

δεδομένα προς μια κλάση απλά μετακινώντας το αρχικό σημείο προς μια κατεύθυνση τότε θα μπορούσαμε να συνθέσουμε δεδομένα απάτης χωρίς να χρειαζόμαστε πολλά δεδομένα.

Για την τελευταία λύση πρέπει να έχουμε ένα αρχικό σημείο όμως για το όποιο να ξέρουμε ακριβώς ποιο είναι το διάνυσμα στον latent space που το δημιουργεί καθώς και την κλάση που ανήκει. Έτσι το πρόβλημα αυτό ανάγεται στο πρόβλημα της αντιστροφής, το όποιο από μόνο του όπως έχει αναφερθεί προηγουμένως είναι αρκετά σύνθετο.

Παρ' όλα αυτά υπάρχει μια οικογένεια δημιουργικών νευρωνικών δικτυών τα οποία είναι εκ-φύσεως αντιστρέψιμα, τα όποια ονομάζονται Normalizing Flows (μοντέλα ροής) [17]. Τα πλέον γνωστά τέτοια δίκτυα είναι τα realNVP και Glow, όπου ουσιαστικά το ένα είναι επέκταση του προηγούμενου.

Normalizing flows [17] είναι αλυσίδες συναρτήσεων οι όποιες είναι αντιστρέψιμες ή η αλγεβρική αναστροφή τους μπορεί να υπολογιστεί. Για παράδειγμα η συνάρτηση:

$$f(x) = x + 2$$

είναι μια αντιστρέψιμη συνάρτηση, γιατί για κάθε έξοδο μπορούμε να υπολογίσουμε την μια και μοναδική είσοδο που την δημιούργησε. Για τον ίδιο λόγο η συνάρτηση:

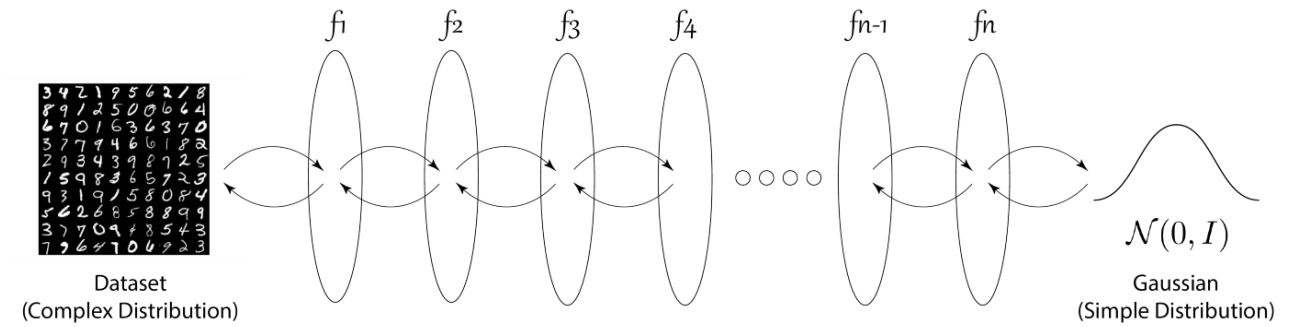
$$f(x) = x^2$$

δεν είναι αντιστρέψιμη συνάρτηση, γιατί δεν είναι 1 προς 1, μπορεί περισσότερα από ένα σημεία να αντιστοιχίζονται στην ίδια έξοδο. Στα περισσότερα νευρωνικά δίκτυα όπου η συνάρτηση

ενεργοποίησης, έστω και σε ένα σημείο είναι η $relu$, δεν είναι αντιστρέψιμη, γιατί η συγκεκριμένη συνάρτηση δεν είναι αντιστρέψιμη. Επισημαίνεται ότι η συνάρτηση αυτή ορίζεται όπως παρακάτω:

$relu(x)$ = x if x > 0 otherwise 0

Οι αντιστρέψιμες συναρτήσεις στην βιβλιογραφία αναφέρονται και ως bijective functions. Οπότε ένα τέτοιο δίκτυο αποτελεί την αλυσίδα τέτοιων συναρτήσεων και έχει τη δομή που παρουσιάζεται στο παρακάτω σχήμα.



Εικόνα 59 Η τυπική αρχιτεκτονική ενός invertible flow και ο τρόπος με τον όποιο μετασχηματίζουν μια απλή σε μια σύνθετη κατανομή.

Αυτό το νευρωνικό δίκτυο προσπαθεί να αντιστοιχίσει κάθε μια από τις εικόνες του mnist σε ένα σημείο της γκαουσιανής κατανομής. Τα δεδομένα του mnist είναι ουσιαστικά τα pixel και αποτελούν σύνθετα δεδομένα τα όποια γίνονται match με τα δεδομένα μιας απλής κατανομής όπως της γκαουσιανής. Βέβαια, λόγω του ότι κάθε τέτοια συνάρτηση είναι αντιστρέψιμη, μπορούμε να μεταβούμε από τα δεδομένα της σύνθετης κατανομής σε αυτά της απλής με την ίδια ευκολία. Μια βασική διαφορά ενός τέτοιου δικτύου από ένα απλό gan είναι ότι το gan εκπαιδεύεται με είσοδο ένα τυχαίο σημείο (απλή κατανομή) και το δίκτυο κατασκευάζει ένα σύνθετο δεδομένο (π.χ. μια εικόνα). Αντίθετα, ένα normalizing flow κατά την εκπαίδευση δέχεται στην είσοδο ένα σύνθετο δεδομένο και προσπαθεί να το αντιστοιχίσει σε ένα σημείο μιας απλής κατανομής, (δηλαδή γίνεται ακριβώς το

αντίστροφο). Επομένως η σύνθεση ενός δεδομένου σε ένα τέτοιο δίκτυο γίνεται ακολουθώντας την αντίθετη κατεύθυνση από αυτή της εκπαίδευσης.

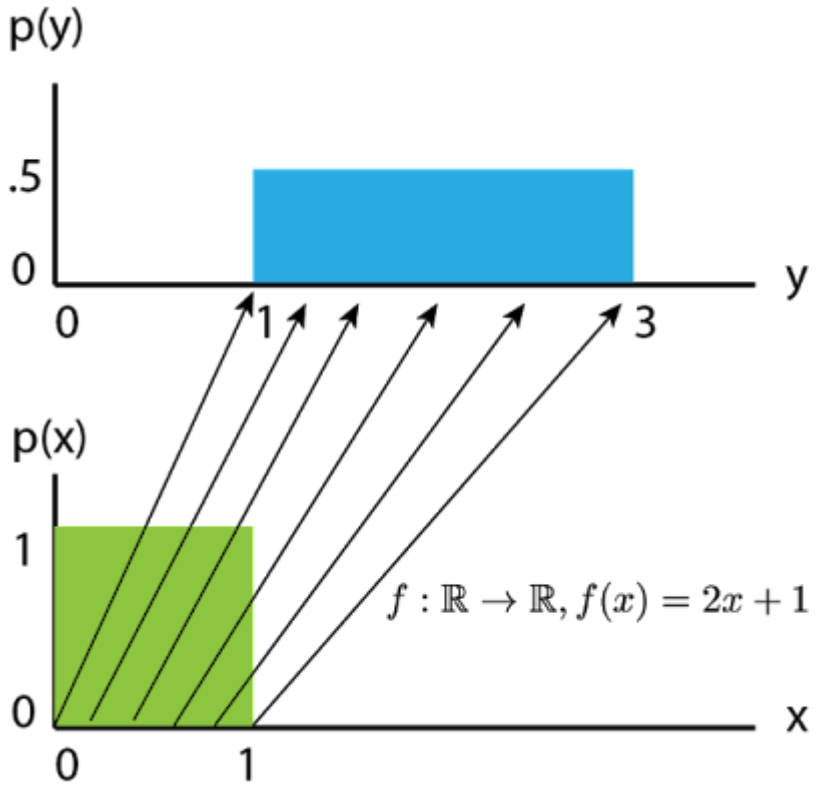
Τα δίκτυα αυτά εκπαιδεύονται χρησιμοποιώντας το negative loglikelihood loss όπου η συνάρτηση p(z) είναι η συνάρτηση πιθανότητας. Η παρακάτω loss function προκύπτει χρησιμοποιώντας τον τύπο αλλαγής μεταβλητών.

$$logp_\theta(x) \;=\; logp_\theta(z) \;+\; log|det(\frac{dz}{dx})| \;=\; logp_\theta(z) \;+\; \sum_{i\,=\,1}^{K} \;\; log|det\,(\frac{dh_i}{dh_{i-1}})|$$

Κάθε τέτοια συνάρτηση (bijector) είναι υπεύθυνη για τον μετασχηματισμό μιας αρχικής κατανομής σε μια άλλη. Γι' αυτό τελικά το δίκτυο σαν σύνολο εκπαιδεύεται στο να μετασχηματίζει μια σύνθετη κατανομή (αυτή των δεδομένων) σε μια απλούστερη. Για παράδειγμα, αν έχουμε δεδομένα τα οποία έρχονται τυχαία από ένα κουτί και σε αυτά εφαρμόσουμε τον μετασχηματισμό της παρακάτω συνάρτησης, η όποια είναι αντιστρέψιμη:

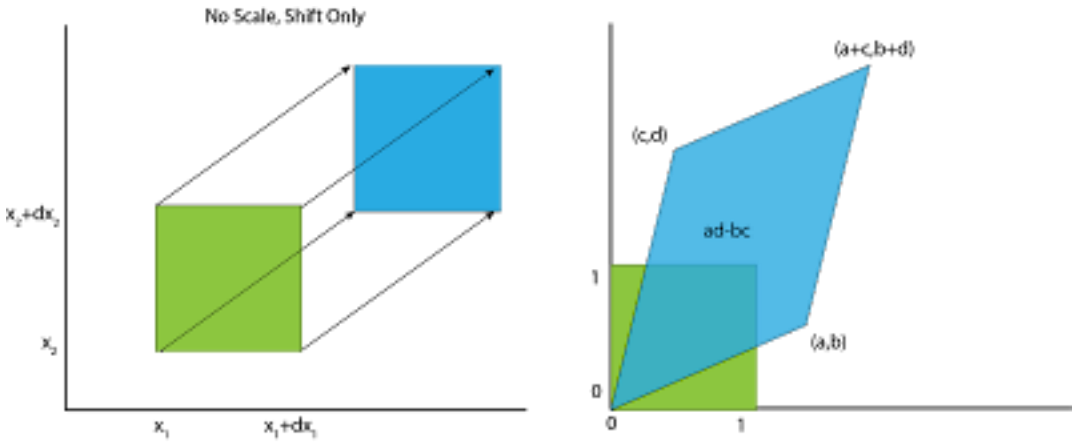
$$f(x) \;=\; 2x \;+\; 1$$

τότε τα μετασχηματισμένα δεδομένα θα είναι όπως αυτά που εμφανίζονται στο παρακάτω διάγραμμα.



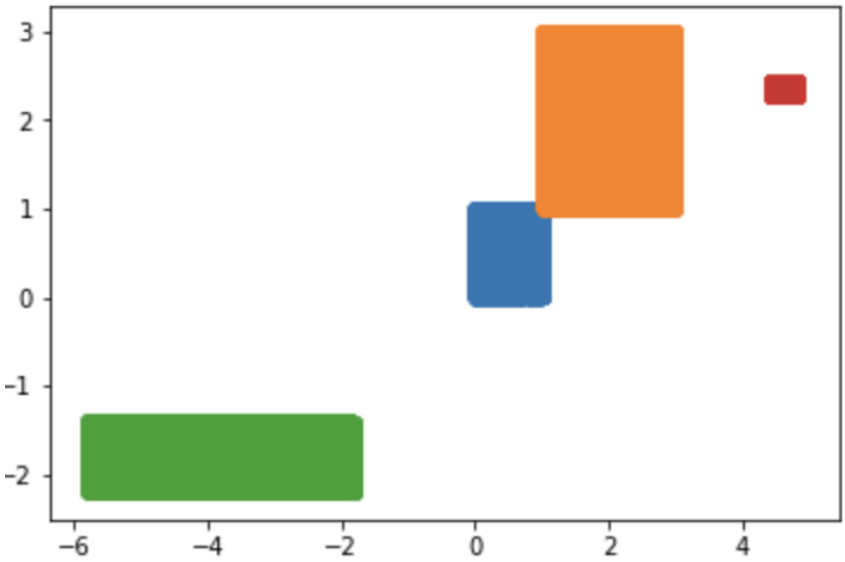
Εικόνα 60 Ο τρόπος με τον όποιον μια αρχική κατανομή μπορεί να μετασχηματισθεί σε μια πιο σύνθετη.

Όπως φαίνονται τα δεδομένα αρχικά κλιμακώνονται (με τον παράγοντα 2) και στην συνέχεια μετακινούνται (κατά 1). Οπότε τα δεδομένα αυτά αρχικά περνούν από scale και στην συνέχεια από μια shifting συνάρτηση. Αν είχαμε μόνο μια από τις 2 συναρτήσεις τα νέα δεδομένα θα είχαν την παρακάτω μορφή.



Εικόνα 61 Ο τρόπος με τον όποιον μια αρχική κατανομή μπορεί να μετασχηματισθεί σε μια πιο σύνθετη.

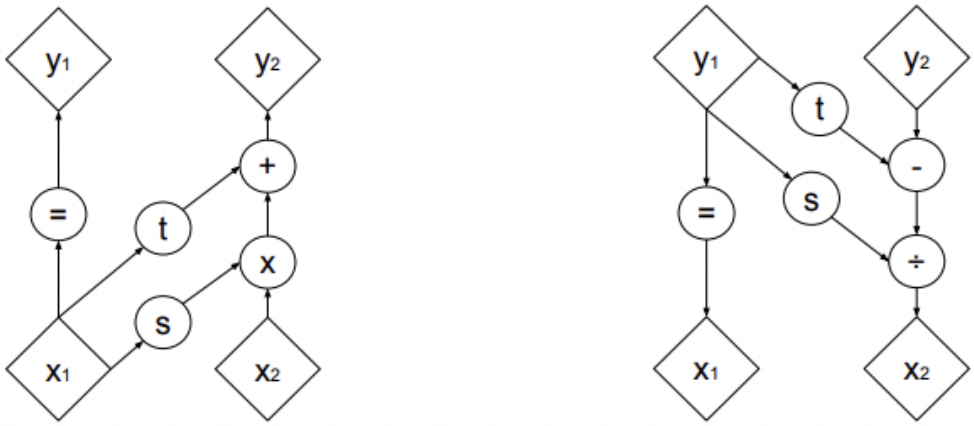
Ακολουθώντας τέτοιους διαδοχικούς μετασχηματισμούς μπορούμε να μετασχηματίσουμε τα αρχικά μας δεδομένα σε δεδομένα διαφορετικών κατανομών. Στο παρακάτω σχήμα έχουμε αρχίσει από τα δεδομένα στο μπλε κουτί, δηλαδή δεδομένα φραγμένα από το 0 μέχρι 1 και στις 2 διαστάσεις και έχουμε εφαρμόσει 3 φορές τέτοιους μετασχηματισμούς (scale και shift). Η έξοδος από το πρώτο βήμα του μετασχηματισμού είναι στο πορτοκαλί κουτί, του επόμενου βήματος στο πράσινο και του τελευταίου στο κόκκινο. Εδώ οι παράγοντες του πολλαπλασιασμού και της πρόσθεσης είναι σταθεροί και ανεξάρτητοι από το σημείο εισόδου αλλά αν αυτό δεν ίσχυε τότε θα είχαμε μια πιο σύνθετη συνάρτηση μετατροπής. Αυτό θα γινόταν αν για παράδειγμα το scale και το shift παράγονταν από μια συνάρτηση ή ένα νευρωνικό δίκτυο, όπως φαίνεται στο παραπάνω σχήμα (αριστερό) όπου έχουμε πολλαπλασιάσει τα διανύσματα με τον πίνακα [[a, b],[c, d]].



Εικόνα 62 Ένα τεχνητό παράδειγμα για την επίδειξη του τρόπου με τον όποιον μια αρχική κατανομή μπορεί να μετασχηματισθεί σε μια πιο σύνθετη.

Με παρόμοιο τρόπο δουλεύει και η αρχιτεκτονική realNVP. Για μια είσοδο, αρχικά σπάει το διάνυσμα εισόδου στην μέση, όποτε αν έστω x το διάνυσμα εισόδου έχουμε τα x1, x2 όπου η ένωση τους ξαναδημιουργεί το x. Το πρώτο μισό διάνυσμα περνάει όπως είναι στην έξοδο ενώ στο δεύτερο εφαρμόζονται ένας scaling και ένας shifting bijector. Η αντιστροφή για το πρώτο μισό είναι μια απλή διαδικασία μιας και η είσοδος και η έξοδος είναι η ίδια ενώ για να αντιστρέψουμε το δεύτερο μισό αρκεί να κάνουμε την πρόσθεση, αφαίρεση και τον πολλαπλασιασμό και διαίρεση και έχουμε το αρχικό μας διάνυσμα. Αξίζει σε αυτό να σημειωθεί ότι δεν χρειάζεται τα συστήματα που παράγουν το scale και το shift να είναι αντιστρέψιμα, αυτά μπορεί να είναι όσο σύνθετα θέλουμε, γιατί μας

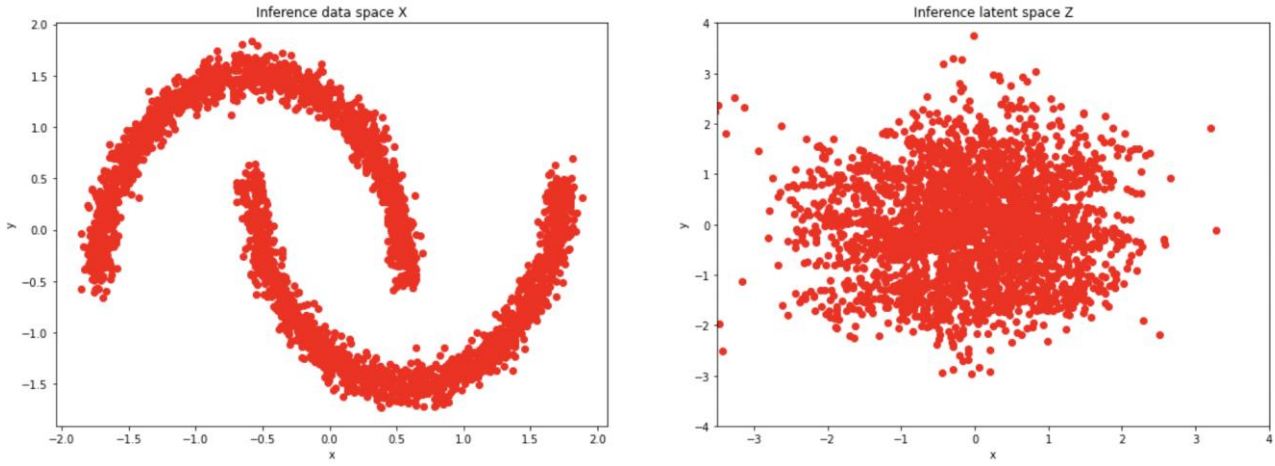
αρκεί που είναι αντιστρέψιμες οι πράξεις που γίνονται στα δεδομένα. Τέλος, σε αυτό το σημείο δημιουργείται το ερώτημα τι γίνεται με το μισό διάνυσμα το όποιο περνάει αυτούσιο στην έξοδο. Για να αποφύγουμε ένα κομμάτι της εξόδου να είναι ίδιο με την είσοδο στο επόμενο βήμα αντιστρέφουμε τα μέρη που γίνεται η επεξεργασία. Με αυτόν τον τρόπο στο επόμενο βήμα το δεύτερο κομμάτι θα περάσει αυτούσιο ενώ το πρώτο (το όποιο είναι το ίδιο με την είσοδο) θα περάσει από τους νέους bijectors. Έτσι, ένα δίκτυο τύπου realNVP ουσιαστικά είναι μια αλυσίδα από shifting και scaling bijectors.



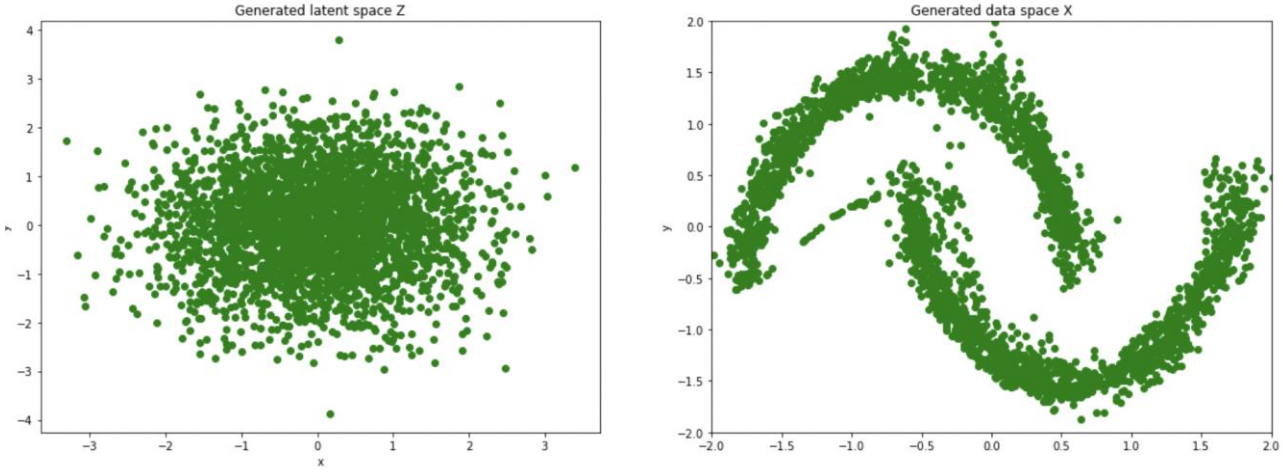
Εικόνα 63 Η βασική δομική μονάδα ενός μοντέλου τύπου RealNVP.

Παρακάτω παρουσιάζεται ένα σχήμα με τον μετασχηματισμό μιας γκαουσιανής κατανομής σε δεδομένα τύπου moons μέσω την εκπαίδευσης ενός μοντέλου realNVP 6 επιπέδων. Για αυτό το μοντέλο το scaling και το shifting είναι 2 μοντέλα mlp 6 επιπέδων με relu και tanh συναρτήσεις

ενεργοποίησης. Από τα σχήματα αυτά βλέπουμε ότι το realNVP έχει μάθει αρκετά καλά να δημιουργεί δεδομένα της αρχικής κατανομής.

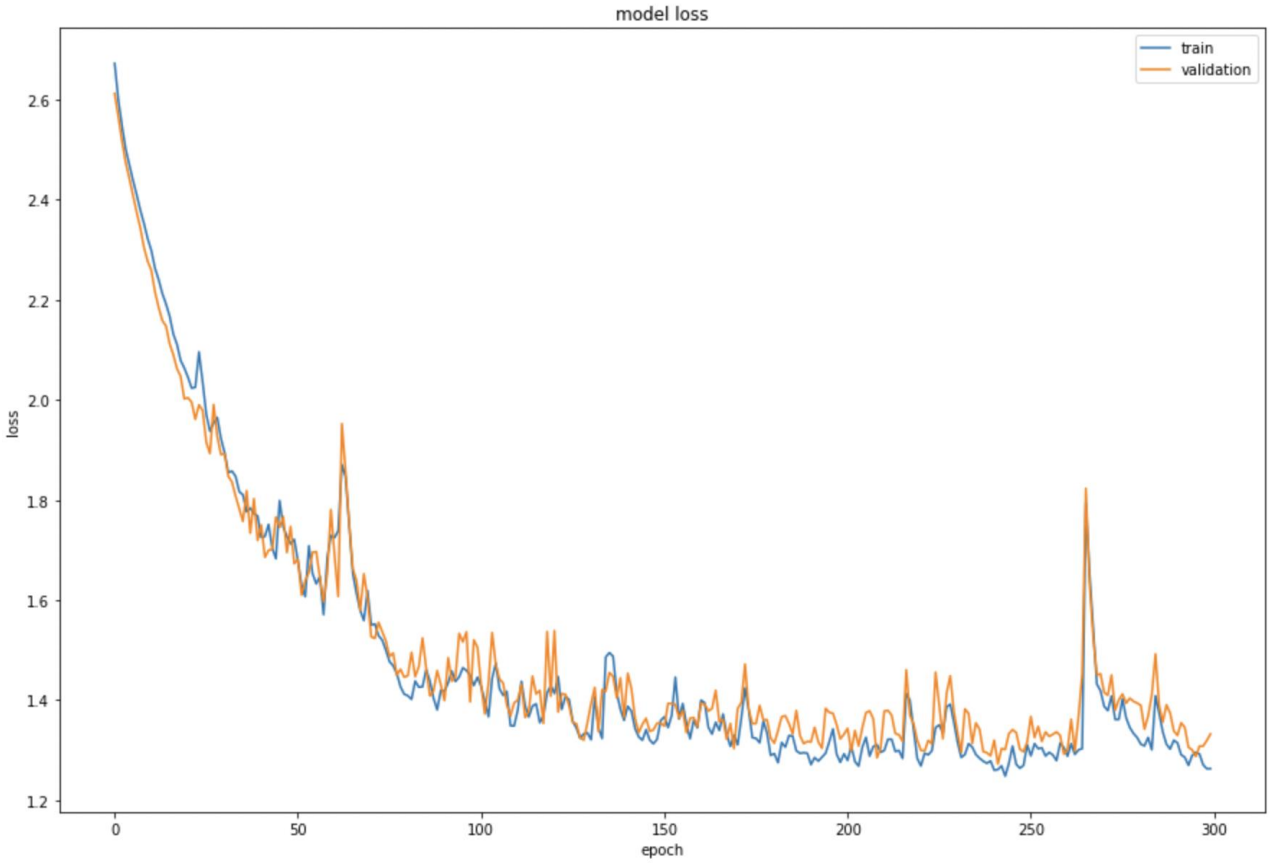


Εικόνα 64 Ο μετασχηματισμός μιας απλής σε μια σύνθετη κατανομή μέσω ενός νευρωνικού δικτύου τύπου flow.



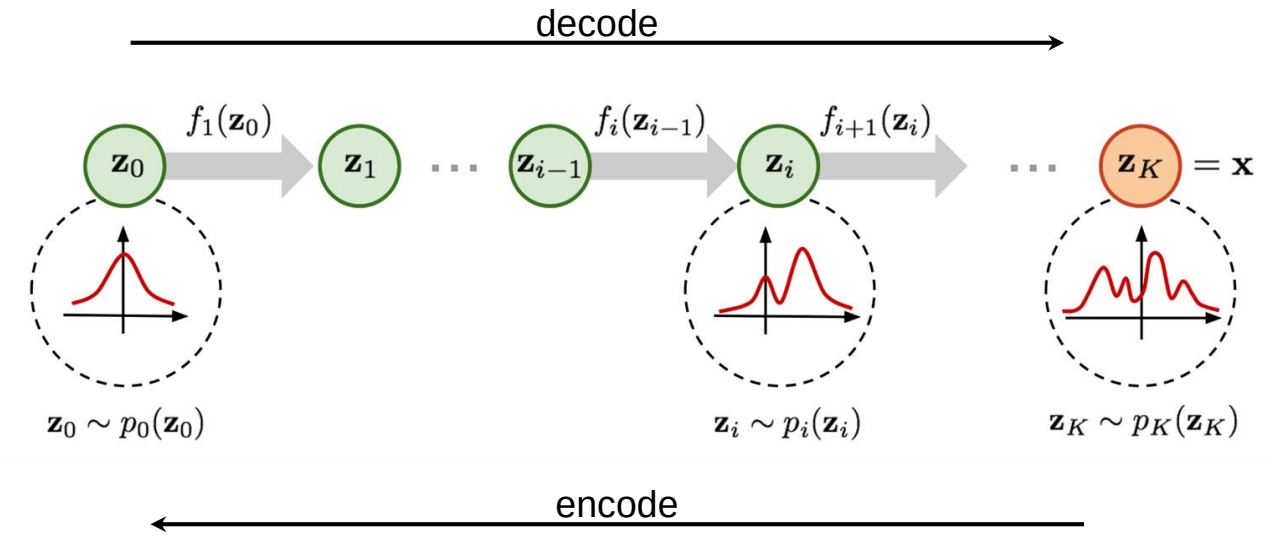
Εικόνα 65 Ο μετασχηματισμός μιας απλής σε μια σύνθετη κατανομή μέσω ενός νευρωνικού δικτύου τύπου flow.

Με τον ίδιο τρόπο εκπαιδεύσαμε και ένα μοντέλο για την παραγωγή δεδομένων συναλλαγών του συνόλου δεδομένων που μελετάμε. Παρακάτω εμφανίζεται η συνάρτηση εκπαίδευσης του μοντέλου για τα training και τα validation sets αντίστοιχα.



Εικόνα 66 Το train και validation error της εκπαίδευσης του νευρωνικού που χρησιμοποιήθηκε κατά τη σύνθεση.

Η αρχιτεκτονική του μοντέλου που χρησιμοποιήθηκε παρουσιάζεται στο παρακάτω σχήμα.



Εικόνα 67 Το train και validation error της εκπαίδευσης του νευρωνικού που χρησιμοποιήθηκε κατά τη σύνθεση.

Σε αυτό συμβολίζουμε τις 2 πράξεις του μοντέλου ως εξής:

- Αποκωδικοποίηση (decode): Η μετάβαση από ένα διάνυσμα τυχαίου θορύβου σε ένα σύνθετο διάνυσμα το όποιο ακολουθεί την κατανομή εκπαίδευσης
- Κωδικοποίηση (encode): Η μετάβαση από ένα δείγμα υψηλής πολυπλοκότητας σε ένα χαμηλότερης

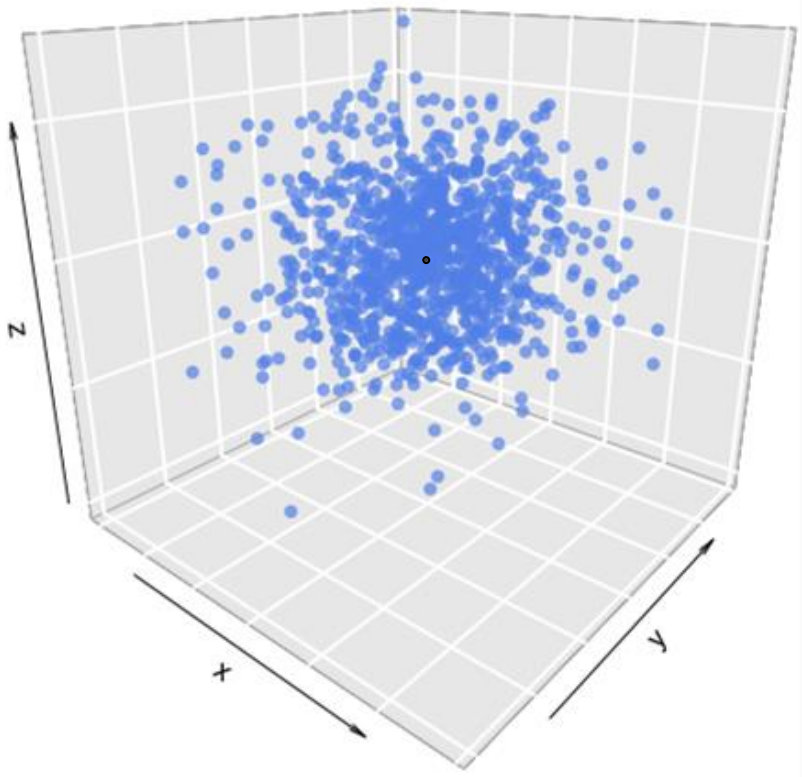
Το μοντέλο αυτό κατασκευάστηκε χρησιμοποιώντας το framework tensorflow[3].

Μέσω του μοντέλου αυτού έχουμε την δυνατότητα τόσο να παράγουμε νέα σημεία όσο και για ένα δοσμένο σημείο να πάρουμε τον αρχικό θόρυβο που το δημιουργεί μιας και το μοντέλο είναι αντιστρέψιμο. Παρόλα αυτά, για τα δεδομένα που παράγουμε τυχαία δεν μπορούμε να ξέρουμε την

[3] https://www.tensorflow.org/

κλάση τους, μιας και το μοντέλο έχει εκπαιδευτεί τόσο σε δεδομένα απάτης όσο και μη. Εντούτοις, στην βιβλιογραφία έχουν προταθεί τρόποι για την παραγωγή επισημειωμένων δεδομένων.

Αρχικά, επειδή το μοντέλο είναι αντιστρέψιμο μπορούμε εύκολα να έχουμε το σημείο το όποιο δημιουργεί ένα σημείο του συνόλου εκπαίδευσης. Για την συγκεκριμένη συναλλαγή γνωρίζουμε την ετικέτα της όποτε μπορούμε να παράγουμε κοντινά σημεία σε αυτό και να τους δώσουμε την ετικέτα του αρχικού σημείου. Στην μέθοδο αυτή κάνουμε την υπόθεση ότι τα δεδομένα που παράγονται κοντά σε ένα σημείο ανήκουν στην ίδια κλάση. Η υπόθεση αυτή όμως δεν επιβεβαιώνεται ούτε θεωρητικά αλλά ούτε πειραματικά, μιας και ανά task αυτό μπορεί να διαφέρει σημαντικά. Πρακτικά στο παρακάτω σχήμα φαίνεται γραφικά ο τρόπος με τον όποιον λειτουργεί αυτή η μέθοδος. Σε αυτό, το μαύρο σημείο αποτελεί το διάνυσμα θορύβου για κάποιο σημείο του συνόλου εκπαίδευσης. Με βάση αυτό δειγματοληπτούμε τον χώρο γύρω του μέχρι κάποια απόσταση, δηλαδή κρατώντας μια ακτίνα γύρω του και δίνουμε στα σημεία που παράγονται από αυτά τα διανύσματα την ετικέτα του αρχικού. Στον αλγόριθμο αυτόν σημαντική υπερπαράμετρο αποτελεί η τιμή της ακτίνας. Για μικρές τιμές της ακτίνας η βεβαιότητα σχετικά με την ετικέτα αυξάνεται αλλά παράλληλα μικραίνει η διαφορετικότητα των δειγμάτων. Από την άλλη, για μεγάλες τιμές της ακτίνας αυξάνεται σημαντικά η διαφορετικότητα, ωστόσο η απομάκρυνση από το αρχικό δείγμα μειώνει την πιθανότητα η ετικέτα να παραμείνει ίδια με αυτή του αρχικού δείγματος.



Εικόνα 68 Μια γραφική αναπαράσταση της δειγματοληψίας γύρω από ένα σημείο σε έναν χώρο 3 διαστάσεων. Το μαύρο σημείο είναι το κέντρο και περιφερικά του βρίσκονται τα τυχαία σημεία.

Άλλη μέθοδος για την παραγωγή δειγμάτων μέσω των αντιστρέψιμων νευρωνικών δικτύων είναι η σύνθεση δειγμάτων με παρόμοιο τρόπο με αυτή της τεχνικής SMOTE. Για δύο δείγματα τα όποια ανήκουν στην ίδια κλάση, μεταφερόμαστε στο latent space και στην συνέχεια υποθέτουμε ότι όλα τα δείγματα που παράγονται από σημεία που βρίσκονται ανάμεσα στα σημεία της ευθείας που τα ενώνει ανήκουν στην ίδια κλάση. Παρόλο που αυτή η τεχνική μπορεί διαισθητικά να μοιάζει ότι θα λειτουργήσει, δεν υπάρχει κάποια θεωρητική τεκμηρίωση ότι τα σημεία που βρίσκονται ανάμεσα στην γραμμή αυτή όντως θα έχουν ανήκουν στην ίδια κλάση. Συνεπώς είναι απαραίτητο να βρεθεί ένας τρόπος αξιοποίησης αυτών των μοντέλων για την σύνθεση αντίστοιχων δεδομένων.

Στο [2] προτείνεται μια μέθοδος για τον χειρισμό του latent space τέτοιων μοντέλων. Συγκεκριμένα μπορούμε να εκπαιδεύσουμε ένα μοντέλο που βασίζεται σε ροή, χωρίς ετικέτες, και στη συνέχεια να χρησιμοποιήσουμε την εκμάθηση λανθάνουσας αναπαράστασης (latent space) για τον χειρισμό των χαρακτηριστικών της εισόδου. Ένα πιο διαισθητικό παράδειγμα αυτού του χειρισμού θα ήταν για ένα τέτοιο μοντέλο το όποιο παράγει πρόσωπα, να μπορούμε να κατευθύνουμε την παραγωγή και να προσθέσουμε στο ίδιο πρόσωπο μπλε μάτια ή ξανθά μαλλιά. Άλλα σημασιολογικά χαρακτηριστικά θα μπορούσε να είναι το στυλ μιας εικόνας, το ύψος ενός μουσικού ήχου ή το συναίσθημα μιας πρότασης κειμένου. Δεδομένου ότι τα μοντέλα που βασίζονται σε ροή (flow based models) έχουν έναν τέλειο κωδικοποιητή, μπορούμε να κωδικοποιήσουμε τις εισόδους και να υπολογίσουμε το μέσο λανθάνον διάνυσμα εισόδων (mean latent space vector) με και χωρίς το χαρακτηριστικό. Στο παράδειγμα που αναφέρεται παραπάνω θα πρέπει να βρούμε το διάνυσμα εισόδου για όλες τις εικόνες με μπλε μάτια και να πάρουμε το μέσο διάνυσμα σαν σημείο που αναπαριστά την περιοχή αυτή. Στην συνέχεια πρέπει να κάνουμε το ίδιο για τις υπόλοιπες εικόνες (χωρίς μπλε μάτια) και να πάρουμε το αντίστοιχο διάνυσμα. Η διανυσματική κατεύθυνση μεταξύ

των δύο μπορεί στη συνέχεια να χρησιμοποιηθεί για να χειριστεί μια αυθαίρετη είσοδο προς αυτό το χαρακτηριστικό. Ο αλγόριθμος για αυτόν τον χειρισμό παρουσιάζεται παρακάτω.

- Είσοδος:
  a. ένα χαρακτηριστικό το όποιο θέλουμε να προστεθεί σε κάποιο δείγμα έστω f_{new}
  b. ένα δείγμα του πραγματικού κόσμου x χωρίς κάποιο χαρακτηριστικό (π.χ. μια εικόνα ενός ατόμου με καφέ μάτια ή μια έγκυρη συναλλαγή). Έστω $F = \{f_1, f_2, \ldots, f_n\}$ το σύνολο των χαρακτηριστικών του δείγματος x τότε πρέπει να ισχύει: $f_{new} \notin \square$
  c. Ένα σύνολο δεδομένων D με επισημειώμενά τα δείγματα του με βάση το χαρακτηριστικό αυτό (π.χ. ένα σύνολο με πρόσωπα όπου ξέρουμε το χρώμα των ματιών κάθε ατόμου ή ένα σύνολο με συναλλαγές όπου ξέρουμε αν είναι απάτη ή όχι). Το σύνολο αυτό μπορεί να είναι τόσο το σύνολο εκπαίδευσης ή και γενικότερα οποιοδήποτε σύνολο δεδομένων με ανάλογη κατανομή με αυτή του συνόλου εκπαίδευσης
- Έξοδος:
  a. Ένα νέο δείγμα x' με σύνολο χαρακτιρηστικά $F' = \{f_1, f_2, \ldots, f_n\} \cup \{f_{new}\}$

1. Για κάθε ένα δείγμα x_i με σύνολο χαρακτηριστικών F_i με $f_{new} \in F_i$
   a. Παίρνουμε το διάνυσμα εισόδου για την είσοδο αυτή $z_i = encode(x_i)$
2. Υπολογίζουμε το: $Z_{positive} = \text{average}(z_i)$ που είναι το διάνυσμα αναπαράστασης των δειγμάτων που έχουν το συγκεκριμένο χαρακτηριστικό
3. Για κάθε ένα δείγμα x_j με σύνολο χαρακτηριστικών F_j με $f_{new} \notin F_j$:
   a. Παίρνουμε το διάνυσμα εισόδου για την είσοδο αυτή $z_j = encode(x_j)$
4. Υπολογίζουμε το: $z_{negative} = \text{average}(z_j)$ που είναι το δίανυσμα αναπαράστασης των δειγμάτων που δεν έχουν το συγκεκριμένο χαρακτηριστικό
5. Υπολογίζουμε το διάνυσμα: $z_{manipulate} = z_{positive} - z_{negative}$
6. Παίρνουμε το διάνυσμα εισόδου για το δείγμα εισόδου x, συμβολίζεται με $z = encode(x)$
7. Επιστρέφουμε το δείγμα εξόδου το οποίο περιέχει το χαρακτηριστικό: $x_{manipulated} = \text{decode}(z_{input} + \text{alpha} * z_{manipulate})$

Η υπερπαράμετρος $\text{alpha} \in [-1,1]$ και ελέγχει κατά πόσο το νεο δείγμα θα περιέχει ή όχι το χαρακτηριστικό. Η παραπάνω διαδικασία απαιτεί σχετικά μικρό όγκο δεδομένων με ετικέτα και μπορεί να γίνει αφού το μοντέλο έχει εκπαιδευτεί (δεν χρειάζονται ετικέτες κατά την εκπαίδευση).

Προηγούμενη εργασία με χρήση του GAN's απαιτεί ξεχωριστή εκπαίδευση ενός κωδικοποιητή. Οι προσεγγίσεις που χρησιμοποιούν μοντέλα τύπου VAE's μόνο εγγυώνται ότι ο αποκωδικοποιητής και ο κωδικοποιητής είναι συμβατοί για δεδομένα κατά τη διανομή. Άλλες προσεγγίσεις περιλαμβάνουν την άμεση εκμάθηση της συνάρτησης που αντιπροσωπεύει τον μετασχηματισμό, όπως το Cycle-GAN, ωστόσο απαιτούν επανεκπαίδευση για κάθε μετασχηματισμό. Η μέθοδος αυτή έχει επιβεβαιωθεί πειραματικά από το [2] και γι' αυτό επιλέχθηκε σαν μέθοδος παραγωγής και στην παρούσα εργασία. Το χαρακτηριστικό το όποιο θα προτεθεί είναι αν μια συναλλαγή είναι απάτη ώστε να αυξηθούν τα δείγματα που δεν είναι έγκυρες συναλλαγές. Πρακτικά η μέθοδος αυτή στο πρόβλημα μας, όπου σαν δείγματα εισόδου θα έχουμε τις έγκυρες συναλλαγές, θα αφαιρεί από κάθε είσοδο τα χαρακτηριστικά τα όποια τα καθιστούν την συγκεκριμένη συναλλαγή έγκυρη και θα προσθέτει χαρακτηριστικά τα οποία την κάνουν απάτη. Σε αυτό το σημείο αξίζει να σημειωθεί ότι για την εκπαίδευση του μοντέλου παραγωγής δεν χρησιμοποιήθηκε το test set, όποτε το δίκτυο δεν μπορεί να γνωρίζει την κατανομή των δεδομένων αυτών. Ένα παράδειγμα όπου φαίνονται ξεκάθαρα οι δυνατότητες του αλγορίθμου αυτού όπως παρουσιάζεται στο [2] είναι το παρακάτω διάγραμμα. Περισσότερα παραδείγματα του αλγορίθμου αυτού βρίσκεται στην online πλατφόρμα[4] που εισάγεται στο [2].

Εικόνα 69 Ένα παράδειγμα επεξεργασίας ενός στιγμιότυπου προς μια συγκεκριμένη κατεύθυνση μέσω της τεχνικής που παρουσιάζεται στο [2].

[4] https://openai.com/blog/glow/

Εικόνα 70 Ένα παράδειγμα επεξεργασίας ενός στιγμιότυπου προς μια συγκεκριμένη κατεύθυνση μέσω της τεχνικής που παρουσιάζεται στο [2].

Εικόνα 71 Ένα παράδειγμα επεξεργασίας ενός στιγμιότυπου προς μια συγκεκριμένη κατεύθυνση μέσω της τεχνικής που παρουσιάζεται στο [2].

Στις παραπάνω εικόνες το σημείο εκκίνησης είναι το ίδιο και η μόνη διαφορά έγκειται στο χρώμα των μαλλιών. Στο πρώτο set εικόνων είναι μαύρα ενώ στο τελευταίο ξανθά. Μέσω της παραμέτρου alpha μπορούμε να ελέγξουμε πόσο ξανθό θα είναι το χρώμα των μαλλιών το οποίο φαίνεται από το ενδιάμεσο παράδειγμα όπου είναι σε μια απόχρωση ενδιάμεση του μαύρου και του ξανθού. Στο παραπάνω παράδειγμα διαφαίνεται και ένα πρόβλημα του αλγορίθμου αυτού, το όποιο σχετίζεται με την πόλωση που μπορεί να εισάγει ως προς τα χαρακτηριστικά. Για παράδειγμα, αν κάθε εικόνα που έχει ένα χαρακτηριστικό f_{new} πάντα συνδιάζεται και με ένα δευτερεύον επιπλέον χαρακτηριστικό f' τότε ουσιαστικά στον latent space είναι σαν να υπάρχει ένα κοινό χαρακτηριστικό $f_{union} =$

$\{f_{new}\} \cup \{f'\}$. Έτσι κάθε φορά που προστίθεται η αφαιρείται το f_{new}, αναγκαστικά προστίθεται η αφαιρείται και το f'. Αυτό είναι εμφανές στο παραπάνω παράδειγμα αν υποθέσουμε ότι:

- $f_{new} = \{\xi\alpha\nu\theta\acute{\alpha}\ \mu\alpha\lambda\lambda\acute{\iota}\alpha\}$
- $f' = \{\alpha\pi\acute{o}\chi\rho\omega\sigma\eta\ \delta\acute{\varepsilon}\rho\mu\alpha\tau o\varsigma\}$

Από τις εικόνες αυτές φαίνεται ότι όσο σκουραίνουν τα μαλλιά του εικονιζόμενου τόσο σκουραίνει και το δέρμα ενώ όσο γίνονται πιο ξανθή η απόχρωση των μαλλιών τόσο παραπάνω φωτίζεται και το πρόσωπο. Το φαινόμενο αυτό είναι λογικό να συμβαίνει μιας και αν όλες οι εικόνες που μελετάμε ή που έχει δει το δίκτυο κατά την εκπαίδευση έχουν πάντα από κοινού δύο χαρακτηριστικά τότε το χαρακτηριστικό αυτό στα "μάτια" του δικτύου είναι ένα. Στην περίπτωση με τις συναλλαγές παρατηρήθηκε ότι σχεδόν σε όλα τα δείγματα το amount των συναλλαγών μειωνόταν όταν μεταβαίναμε από μια έγκυρη συναλλαγή σε μια όχι. Η παρατήρηση αυτή έρχεται να επιβεβαιωθεί από το γεγονός ότι ο μέσος όρων των συναλλαγών που δεν είναι έγκυρες είναι σημαντικά μικρότερος των έγκυρων, όπως φαίνεται και στο παρακάτω σχήμα.

Εικόνα 72 Ένα παράδειγμα επεξεργασίας ενός στιγμιότυπου προς μια συγκεκριμένη κατεύθυνση μέσω της τεχνικής που παρουσιάζεται στο [2].

Ακολουθώντας τον αλγόριθμο αυτό παρήγαμε δείγματα μη έγκυρων συναλλαγών από έγκυρες για την ενίσχυση του συνόλου και στην συνέχεια ακολουθήσαμε την ίδια πειραματική διαδικασία για την αξιολόγηση του νέου συνόλου. Παρακάτω παρουσιάζεται ένας συγκεντρωτικός

πίνακας με τα αποτελέσματα για όλες τις μεθόδους και τα πειράματα που έγιναν στην παρούσα διπλωματική εργασία.

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%
Αντιστρέψιμα δημιουργικά νευρωνικά δίκτυα	**99.99%**	**77.12%**	**81.43%**	**79.21%**

Εικόνα 73 Συγκριτικός πίνακας της μέσης απόδοσης κάθε αλγορίθμου για όλα τα μοντέλα. Από τον πίνακα αυτόν φαίνεται ότι η καλύτερη απόδοση επιτεύχθηκε για την μέθοδο σύνθεσης δεδομένων μέσω ενός νευρωνικού δικτύου τύπου Inversible flow.

7. Συμπεράσματα

Στην παρούσα εργασία μελετήθηκε το πρόβλημα της ανίχνευσης απάτης σε τραπεζικές συναλλαγές. Καθημερινά λαμβάνουν χώρα εκατομμύρια συναλλαγές, ελάχιστες από αυτές όμως αποτελούν απάτη. Οπότε 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 εγκύρων συναλλαγών και αυτό γιατί αυτά κατά κύριο λόγο θα έχει μάθει να προβλέπει όλες τις συναλλαγές ως έγκυρες μιας και αυτό έχει δει στο μεγαλύτερο βαθμό. Οπότε 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 εξής μέθοδοι: Υπερδειγματοληψία, υποδειγματοληψία, μέθοδος smote, ο συνδυασμός των μεθόδων της υπερ και υποδειγματοληψίας, η κανονικοποιήση, ενώ τελευταία μελετήθηκαν οι συνθετικές μέθοδοι. Οι μέθοδοι αυτές δοκιμάστηκαν σε 5 διαφορετικού τύπου ταξινομητές ώστε να μελετηθεί εκτενώς ο τρόπος επίδρασής τους. Τέλος μελετήθηκαν μέθοδοι για τη σύνθεση δεδομένων με σκοπό την παραγωγή στιγμιότυπων για την υποβοήθηση κατά την εκπαίδευση. Με αυτή τη μέθοδο έχουμε τη δυνατότητα να παράγουμε αυθαίρετα μεγάλο αριθμό νέων δεδομένων τα όποια θα μπορέσουν με αυτόν τον τρόπο να μειώσουν τις συνέπειες της ανισοκατανομής των δεδομένων.

Συγκεκριμένα έγινε μια εκτενής μελέτη της σχετικής βιβλιογραφίας όπου διαπιστώθηκε ότι αντίστοιχες μέθοδοι χρησιμοποιούνται συχνά και σε πληθώρα προβλημάτων, πέρα από την εκπαίδευση ταξινομητών. Παρόλα αυτά, σε αντίστοιχα προβλήματα χρησιμοποιούνται πολύ μεγαλύτερα σύνολα δεδομένων τα όποια δεν εμφανίζουν τόσο έντονο το φαινόμενο της ανισοκατανομής. Η μέθοδος που αναπτύχθηκε στην παρούσα εργασία αφορά την ανάπτυξη ενός αντιστρέψιμου δημιουργικού δικτύου. Με αυτό μπορούμε δειγματοληπτικά να παράγουμε δεδομένα γύρω από ένα συγκεκριμένο σημείο της εισόδου, έχοντας έτσι μια σχετική βεβαιότητα όσο αναφορά την ετικέτα κάθε παραγόμενου δείγματος.

Η συγκεκριμένη τεχνική επέφερε τα καλύτερα αποτελέσματα και βοήθησε όσο καμία άλλη από αυτές που μελετήθηκαν στην εκπαίδευση και των 5 ταξινομητών που εκπαιδεύτηκαν.

Επόμενα Βήματα

Στη συνέχεια πρόκειται να μελετηθούν εκτενέστερα οι συνθετικές μέθοδοι με σκοπό την βελτίωση της εκπαίδευσης ταξινομητών. Το πρώτο βήμα είναι η ενσωμάτωση περισσότερων μοντέλων σύνθεσης, διαφορετικού τύπου αλλά και πολυπλοκότητας. Έπειτα, οι μέθοδοι αυτοί θα εφαρμοστούν σε περισσότερα μοντέλα ώστε να υπάρξει μια πληρέστερη αναφορά των αποτελεσμάτων τόσο σε διαφορετικές αρχιτεκτονικές όσο και σε διαφορετικής πολυπλοκότητας μοντέλα. Επίσης σημαντικό βήμα για τη συνέχιση της παρούσας μελέτης είναι η δοκιμή του σε διαφορετικά σύνολα δεδομένων πέρα από το fraud detection.

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] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[21] Li, Yitong, et al. "Storygan: A sequential conditional gan for story visualization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Ευρετήριο Εικόνων