Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Μέθοδοι συμπλήρωσης ελλιπών τιμών σε ηλεκτρονικά ιατρικά δεδομένα με χρήση τεχνικών βαθιάς μάθησης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

του

ΨΥΧΟΓΥΙΟΥ ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων: Δημήτριος Ασκούνης
Καθηγητής

Αθήνα, Μάρτιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Μέθοδοι συμπλήρωσης ελλιπών τιμών σε ηλεκτρονικά ιατρικά δεδομένα με χρήση τεχνικών βαθιάς μάθησης

Μελέτη και υλοποίηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΨΥΧΟΓΥΙΟΥ ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων: Δημήτριος Ασκούνης
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Μαρτίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Χρυσόστομος Δούκας
Καθηγητής	Καθηγητής	Αν. Καθηγητής

Αθήνα, Μάρτιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Κωνσταντίνος Ψυχογυιός

9 Μαρτίου 2023

Περίληψη

Τα δεδομένα από ιατρικά εργαστήρια σε ηλεκτρονική μορφή (EHR) γίνονται ολοένα και πιο διαδεδομένα καθώς μεγάλες φαρμακευτικές εταιρείες συνοψίζουν χρόνια έρευνας με αυτόν τον τρόπο. Ακόμα, πολλά νοσοκομεία έχουν πλέον υιοθετήσει την ψηφιοποίηση των δεδομένων των ασθενών τους και σε αρκετές περιπτώσεις τα έχουν διαθέσει στην επιστημονική κοινότητα. Όμως, λόγω της φύσης των δεδομένων αυτών παρουσιάζεται συχνά το πρόβλημα των απουσιαζόντων τιμών. Τα δεδομένα αυτά χρησιμοποιούνται συχνά για πρόβλεψη η κατάτμησή όπου οι απουσιάζοντες τιμές αποδίδονται με κάποια τεχνική καταλογισμού η αφαιρούνται από το σύνολο. Πολλές φορές η τεχνική που χρησιμοποιείται είναι η απλή αντικατάσταση με τη μέση τιμή και την πιο συχνά εμφανιζόμενη τιμή για τις συνεχείς και τις κατηγορικές μεταβλητές αντίστοιχα.

Στόχος της διπλωματικής εργασίας είναι η εφαρμογή συγχρόνων τεχνικών βαθιάς μάθησης όπως γενετικά ανταγωνιστικά δίκτυα και αυτοκωδικοποίητες σε πραγματικά ηλεκτρονικά ιατρικά δεδομένα για την ακριβή απόδοση των απουσιαζόντων τιμών με σκοπό τη βελτίωση της διαδικασίας της πρόβλεψης πάνω σε αυτά.

Λέξεις Κλειδιά

ΠΑΔ, ΗΦΥ, Απουσιάζουσες τιμές, Αποτίμηση απουσιαζόντων τιμών

Abstract

Electronic health records are becoming adopted more and more by big pharmaceutical companies aggregating years of research. Moreover, many hospitals have converted their patient's data within the scope of digitalization. In both cases, it is regular for the data to be publicly shared for research purposes. However, due to the nature of these kind of data missingness is very common. EHR data are commonly used for classification and prediction where missing values are either deleted (Case deletion) or imputed. Regarding the imputation case the mean,mode imputation is widely adopted.

This diploma thesis aims to apply modern deep-learning approaches such as GANS and Autoencoders to the problem of missing value imputation. The evaluation will be based both on the accuracy of the imputed data and on the post-processing accuracy for the classification task

Keywords

GAN, EHR, Missing values, Missing data imputation

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Δημήτριο Ασκούνη για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης. Επίσης ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Λουκά Ηλία για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς και τους φίλους μου μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Μάρτιος 2023

Κωνσταντίνος Ψυχογυιός

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Στο σύγχρονο ψηφιοποιημένο ιατρικό περιβάλλον η σύλλεξη, αποθήκευση και αξιοποίηση δεδομένων ειναι ζωτικής σημασίας. Στο πλάισιο αυτό, πολλά νοσοκομεία και σχετικοί οργανισμοί καταγράφουν και αποθηκεύουν δεδομένα σχετικα με ασθενείς, ιατρικές έρευνες κ.τλ. Τα δεδομένα αυτά μπορεί να είναι σε μορφή φωτογραφιών (π.χ. MRI scans), σε μορφή πίνακα (Tabular) κ.α. συμπεριλμβάνοντας η όχι την χρονική συνιστώσα .

Ο ηλεκτρονικός φάκελος υγείας (ΗΦΥ ή EHR) είναι ένα έγγραφο που περιέχει ιατρικές πληροφορίες, π.χ. εργαστηριακές μετρήσεις, για έναν ασθενή και αποθηκεύεται ηλεκτρονικά. Έτσι, μπορεί να διαμοιραστεί σε πολλαπλές εγκαταστάσεις και να έχει γρήγορη πρόσβαση από τους ασθενείς ή το ιατρικό προσωπικό. Ένας ηλεκτρονικός φάκελος υγείας χρησιμοποιείται κυρίως για σκοπούς καθορισμού στόχων και σχεδιασμού της φροντίδας των ασθενών, τεκμηρίωσης της παροχής φροντίδας και αξιολόγησης των αποτελεσμάτων της φροντίδας [1]. Τα δεδομένα αυτά παρέχουν ευκαιρίες για τη βελτίωση της φροντίδας των ασθενών, την ενσωμάτωση μέτρων απόδοσης στην κλινική πρακτική και τη διευκόλυνση της κλινικής έρευνας [2]. Ένα παράδειγμα έρευνας που βασίζεται στους ΗΦΥ είναι η πρόβλεψη του καρδιαγγειακού κινδύνου με τη χρήση μεθόδων παλινδρόμησης μηχανικής μάθησης [3]. Ένα τέτοιο μοντέλο, μπορεί να χρησιμοποιηθεί ως σύστημα υποστήριξης αποφάσεων για να βοηθήσει τους γιατρούς και το προσωπικό να διαχειρίζονται τους ασθενείς και να ενεργούν προληπτικά.

Ωστόσο, είναι πολύ συνηθισμένο για αυτού του είδους τα δεδομένα να έχουν ένα ποσοστό απουσιαζόντων τιμών [4]. Τα ελλιπή δεδομένα εμφανίζονται όταν οι τιμές των μεταβλητών ενδιαφέροντος δεν έχουν μετρηθεί ή καταγραφεί για όλα τα υποκείμενα του δείγματος. Τα δεδομένα μπορεί να λείπουν για διάφορους λόγους [5], όπως: (ι) άρνηση του ασθενούς να απαντήσει σε συγκεκριμένες ερωτήσεις, π.χ. ο ασθενής δεν αναφέρει δεδομένα σχετικά με το εισόδημα- (ιι) απώλεια του ασθενούς για παρακολούθηση- (ιιι) σφάλμα του ερευνητή ή μηχανικό σφάλμα, π.χ. βλάβη του πιεσόμετρου- και (ι) γιατροί που δεν παραγγέλνουν ορισμένες εξετάσεις για ορισμένους ασθενείς, π.χ. δεν παραγγέλλεται εξέταση χοληστερόλης για ορισμένους ασθενείς. Οι ελλείπουσες τιμές μπορούν να οριστούν από τρεις κύριους μηχανισμούς που είναι η πλήρης τυχαία έλλειψη (MCAR), η τυχαία έλλειψη (MAR) και η μη τυχαία έλλειψη (MNAR) [6]. Η πρώτη περίπτωση (MCAR) συμβαίνει όταν η έλλειψη που παρουσιάζεται σε ένα ΗΦΥ ακολουθεί ένα εντελώς τυχαίο μοτίβο. Το δεύτερο (MAR) υποδεικνύει ότι η έλλειψη σε μια μεταβλητή σχετίζεται με μια άλλη μεταβλητή. Η τρίτη περίπτωση (MNAR) δείχνει ότι η έλλειψη σε μια μεταβλητή εξαρτάται από την ίδια τη μεταβλητή. Αξίζει

επίσης να σημειωθεί ότι εντός ενός συνόλου δεδομένων κλινικών ΗΦΥ μπορεί να υπάρχουν περισσότερα από ένα ελλείποντα πρότυπα τη δεδομένη στιγμή με διαφορετικά ποσοστά έλλειψης.
Στο πλαίσιο της κλινικής έρευνας, τα δεδομένα που λείπουν συνήθως αντιμετωπίζονται ανεπαρκώς [7]. Η πιο συνηθισμένη προσέγγιση είναι η ανάλυση πλήρους περίπτωσης όπου οι γραμμές που περιέχουν ελλιπείς τιμές είτε στις μεταβλητές πρόβλεψης είτε στις μεταβλητές έκβασης απορρίπτονται. Αυτή η επιλογή είναι εξαιρετικά προβληματική, καθώς οδηγεί σε μικρότερο σύνολο δεδομένων και σε ένα μοντέλο που δεν είναι σε θέση να γενικεύσει καλά. Επίσης, η μέθοδος αυτή παράγει συχνά αποτελέσματα και σφάλματα που μπορεί να είναι μικρά για το πλήρες υποσύνολο δεδομένων, αλλά στην πραγματικότητα είναι αισιόδοξα. Επιπλέον, διαφορετικές μελέτες μπορεί να χρησιμοποιούν διαφορετικά υποσύνολα του ίδιου συνόλου δεδομένων, π.χ. αντί για γραμμές μπορεί να παραλείπονται στήλες ή ένας συνδυασμός και των δύο, και έτσι η επιλογή αυτή δυσχεραίνει τις συγκρίσεις. Μια άλλη προσέγγιση για την επίλυση αυτού του ζητήματος είναι ο απλός καταλογισμός με τον αλγόριθμο mean, mode (πιο συχνό) ή KNN [8]. Αυτές οδηγούν σε ένα πλήρες σύνολο δεδομένων, αλλά είναι πολύ απλές και συνεπώς καταλογίζουν τιμές που δεν είναι ρεαλιστικές. Για παράδειγμα, όσον αφορά τα δεδομένα σε επίπεδο ασθενών με καρδιαγγειακές παθήσεις (CVD), υπάρχει συνήθως ισχυρή συσχέτιση μεταξύ των αντίστοιχων μεταβλητών, π.χ. συστολική και διαστολική αρτηριακή πίεση, η οποία θα πρέπει να ενσωματωθεί στο μοντέλο υπολογισμού ελλειπουσών τιμών. Αυτό είναι κάτι που οι μονομεταβλητές στατιστικές προσεγγίσεις και οι απλοί αλγόριθμοι παλινδρόμησης αδυνατούν να υπολογίσουν οδηγώντας σε ανακριβή αποτελέσματα [9]. Αυτές οι συσχετίσεις υπάρχουν φυσικά στα περισσότερα σύνολα ιατρικών δεδομένων όπου έχουν διεξαχθεί εξετάσεις για τον ίδιο ασθενή, έχουν πραγματοποιηθεί εργαστηριακές μετρήσεις για μια συγκεκριμένη εργασία κ.λπ. Ορισμένες κάπως πιο σύνθετες και με καλύτερα αποτελέσματα μέθοδοι είναι η Missforest (MF) [10] και ο αλγόριθμος πολλαπλών αποτιμήσεων με αλυσιδωτές εξισώσεις (multivariate imputation by chained equations - MICE) [11]. Παρόλο που αυτές οι μέθοδοι είναι πιο εξελιγμένες, εξακολουθούν να μην έχουν την ικανότητα να αναλύουν πλήρως τις πολύπλοκες σχέσεις που καθορίζουν τα σύνολα δεδομένων ΗΦΥ [12, 13]. Αυτό είναι κάτι που είναι πιο σοβαρό σε διαχρονικές μελέτες, όπου οι πληροφορίες σχετικά με μια ελλιπή τιμή θα πρέπει να συσχετίζονται με προηγούμενες τιμές του ίδιου ασθενούς.

1.1 Αντικείμενο της διπλωματικής

Το βασικό ζήτημα αντιμετωπίζει η παρούσα διπλωματική είναι η απόδοση ρεαλιστικών τιμών σε απουσιάζουσες με σύγχρονες (state-of-the-art) μεθόδους έχοντας απώτερο σκοπό την επίτευξη καλύτερης απόδοσης στο κομμάτι της πρόβλεψης. Για να αντιμετωπιστούν αυτοί οι περιορισμοί, στην παρούσα εργασία συγκρίνουμε διάφορες μεθόδους υπολογισμού ελλιπών δεδομένων. Συγκεκριμένα, προτείνουμε δύο προσεγγίσεις βαθιάς μάθησης που βασίζονται στις DAE και GAN. Η πρώτη προσέγγιση βασίζεται σε μια DAE που χρησιμοποιεί KNN για προ-υπολογισμό. Χρησιμοποιώντας αυτό το μοντέλο ως βάση, εφαρμόζουμε διάφορες αλλαγές όσον αφορά τόσο την αρχιτεκτονική όσο και τη διαδικασία εκπαίδευσης,

οι οποίες αποδίδουν σημαντικά ακριβέστερα αποτελέσματα. Προσαρμόζουμε τη συνάρτηση απωλειών, βελτιστοποιούμε την αρχιτεκτονική των στρωμάτων (π.χ. προσθήκη ομαλοποίησης παρτίδων, επιλογή βέλτιστου αριθμού στρωμάτων) και αναθεωρούμε τη διαδικασία εκπαίδευσης αλλάζοντας συχνά τις τιμές που προεπιβάλλονται (με KNN) και τις τιμές που πρέπει να προστεθούν (δείκτες) αποτρέποντας το μοντέλο από το να συγκλίνει σε ένα τοπικό ελάχιστο. Η βελτίωση που εφαρμόζεται στη διαδικασία εκπαίδευσης είναι ανεξάρτητη από τα μοντέλα και μπορεί να χρησιμοποιηθεί και για διαφορετικές μεθόδους τεκμαίρεσης. Όσον αφορά την προσέγγιση GAN, βασιστήκαμε επίσης στην υπάρχουσα αρχιτεκτονική κάνοντας βελτιώσεις που αφορούν τη συγκεκριμένη περίπτωση που μελετάμε. Πιο συγκεκριμένα, χρησιμοποιούμε ως γεννήτρια μια DAE με προ-υπολογισμό KNN και εφαρμόζουμε τις προαναφερθείσες προσαρμογές στη διαδικασία εκπαίδευσης. Για την αξιολόγηση των μοντέλων μας χρησιμοποιούμε τέσσερα δημόσια διαθέσιμα σύνολα δεδομένων ΗΦΥ. Τέλος, τα προτεινόμενα μοντέλα αξιολογούνται τόσο για τις εργασίες υπολογισμού όσο και για τις εργασίες πρόβλεψης μετά τον υπολογισμό. Μελετάμε το τελευταίο για να διερευνήσουμε αν η επιλογή πιο ισχυρών μεθόδων εμφύτευσης θα οδηγήσει σε υψηλότερες επιδόσεις πρόβλεψης ή όχι. Αυτό είναι πολύ σημαντικό, δεδομένου ότι η πρόβλεψη είναι συνήθως ο κοινός στόχος των ερευνητών και των επαγγελματιών όταν εφαρμόζουν τεχνικές μηχανικής μάθησης σε δεδομένα ΗΦΥ. Οι κύριες συνεισφορές μας μπορούν να συνοψιστούν ως εξής:

- Επεκτείνουμε και βελτιώνουμε τα υπάρχοντα μοντέλα υπολογισμού ελλειπουσών τιμών που βασίζονται σε DAEs και GANs.
- Αξιολογούμε διεξοδικά διάφορες μεθόδους υπολογισμού ελλιπών τιμών για διάφορα ποσοστά ελλιπών στοιχείων.
- Αποδεικνύουμε ότι τα σύνολα δεδομένων ΗΦΥ για καρδιακές παθήσεις (και ιατρικά γενικά) απαιτούν πιο εξελιγμένες μεθόδους υπολογισμού ελλειπουσών τιμών, ιδίως όταν το ποσοστό ελλειπουσών τιμών είναι υψηλό.
- Δείχνουμε ότι οι ισχυρές μέθοδοι βαθιάς μάθησης μπορούν να αντιμετωπίσουν τόσο αριθμητικά όσο και κατηγορικά δεδομένα.
- Διεξάγουμε πρόβλεψη μετά τον υπολογισμό και δείχνουμε ότι ο ακριβέστερος υπολογισμός οδηγεί σε μεγαλύτερη απόδοση στο έργο πρόβλεψης.

Με την χρήση αυτών τον μοντέλων λοιπόν στόχος της διπλωματικής είναι η συγκριτική μελέτη για διάφορα ποσοστά απουσίας μεταβλητών (5 εως 50 τις εκατό).

1.2 Οργάνωση του τόμου

Η εργασία αυτή είναι οργανωμένη σε επτά κεφάλαια: Στο Κεφάλαιο 2 δίνεται το θεωρητικό υπόβαθρο των βασικών τεχνολογιών που σχετίζονται με τη διπλωματική αυτή. Αρχικά περιγράφονται μερικά γενικά πράγματα για την μηχναική μάθηση και τους μηχανισμούς πίσω απο τις απουσιάζουσες τιμές. Στην συνέχεια αναλύεται το υπόβαθρο για τις τεχνικές που χρησιμοποιήθηκαν στα πλαίσια αυτής της διπλωματικής. Στο Κεφάλαιο 3 γίνεται μια λεπτομερής περιγραφή των μεθόδων που χρησιμοποιήθηκαν καθώς και των εισαγλωμενων

μεθόδων βαθειών νευρωνικών δικτύων. Επίσης σε αυτό δίνεται και μια περιγραφή του συνόλου δεδομένων που χρησιμοποιήθηκε. Στο Κεφάλαιο 4 αναλύονται τα αποτελέσματα των μεθόδων που αναφέρθηκαν στο προηγούμενο κεφάλαιο. Εδώ τα αποτελέσματα είναι και για την αποτίμηση τιμών αλλα και για την πρόβλεψη μετά την αποτίμηση. Τέλος, στο Κεφάλαιο 5 γίνεται μία περίληψη της διπλωματικής και δίνονται μελλοντικές κατευθύνσεις.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

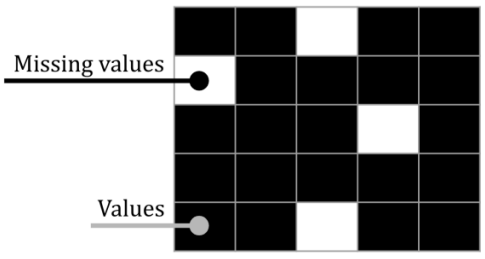
Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζεται λεπτομερώς το πρόβλημα των απουσιαζόντων τιμών, οι τρόποι αντιμετώπισης του που υπάρχουν στη βιβλιογραφία αλλά και το υπόβαθρο για τις μεθόδους που χρησιμοποιήθηκαν στην παρούσα διπλωματική, δηλαδή οι μέθοδοι KNN, Missforest, Autoencoders, GANS. Πιο συγκεκριμένα, γίνεται μια εισαγωγή στους μηχανισμούς απουσιαζόντων τιμών και στις μεθόδους που έχουν χρησιμοποιηθεί από τους ερευνητές στην βιβλιογραφία. Έπειτα, γίνεται μια γενική περιγραφή της μηχανικής μάθησης η οποία στη συνέχεια γίνεται πιο στοχευμένη στο υπόβαθρο σχετικό με τις μεθόδους αποτίμησης απουσιαζόντων τιμών που χρησιμοποιήθηκαν.

2.1 Τύποι μηχανισμών απουσιαζόντων τιμών

2.1.1 MCAR

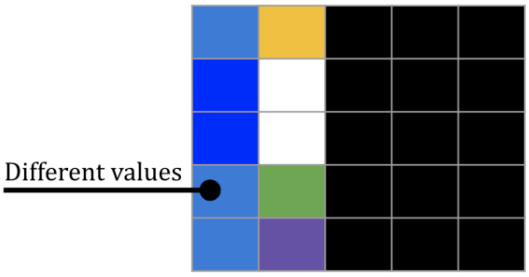
Ο μηχανισμός αυτός ορίζει ότι το γεγονός ότι μια μεταβλητή απουσιάζει είναι ανεξάρτητο από την ίδια την μεταβλητή η από οποιαδήποτε άλλη μεταβλητή του συνόλου δεδομένων. Για παράδειγμα, εάν έχουμε εργαστηριακές μετρήσεις σχετικά με καρδιαγγειακά νοσήματα, π.χ. αρτηριακή πίεση και χοληστερόλη, ορισμένοι ασθενείς μπορεί να έχουν ελλιπείς τιμές στη χοληστερόλη επειδή δεν μπόρεσαν να επισκεφθούν το εργαστήριο τη συγκεκριμένη ημέρα λόγω απεργίας των μέσων μαζικής μεταφοράς. Αυτό απεικονίζεται στην εικόνα 2.1.



Σχήμα 2.1: *Απουσιάζοντες τιμές με τον μηχανισμό MCAR. Τα μαύρα κουτάκια απεικονίζουν τις υπάρχοντες τιμές ενώ τα άσπρα τις απουσιάζουσες.*

2.1.2 MAR

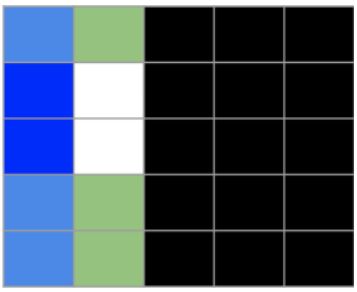
Ο μηχανισμός αυτός ορίζει ότι το γεγονός ότι μια μεταβλητή απουσιάζει είναι ανεξάρτητο από την ίδια την μεταβλητή αλλά μπορεί να εξαρτάται άλλη μεταβλητή του συνόλου δεδομένων. Ένα παράδειγμα θα μπορούσε να είναι ότι η έλλειψη δεδομένων για τη διαστολική αρτηριακή πίεση σχετίζεται με χαμηλή συστολική αρτηριακή πίεση. Αυτό απεικονίζεται στην εικόνα 2.2.



Σχήμα 2.2: *Απουσιάζοντες τιμές με τον μηχανισμό MAR. Τα μαύρα κουτάκια απεικονίζουν τις υπάρχουτες τιμές ενώ τα άσπρα τις απουσιάζουσες.Βλέπουμε ότι η μεταβλητή της δεύτερης στήλης μπορεί να καθοριστεί αλλά όχι πλήρως απο τις τιμές της πρώτης στήλης.*

2.1.3 MNAR

Ο μηχανισμός αυτό· ορίζει ότι το γεγονός ότι μια μεταβλητή απουσιάζει μπορεί να είναι εξαρτημένο και από την ίδια την μεταβλητή αλλά και από άλλες μεταβλητές του συνόλου δεδομένων. Μια τέτοια περίπτωση είναι όταν τα άτομα με υψηλή χοληστερόλη δεν επισκέπτονται το νοσοκομείο για να κάνουν εργαστηριακές εξετάσεις. Αυτό απεικονίζεται στην εικόνα 2.3.



Σχήμα 2.3: *Απουσιάζοντες τιμές με του μηχανισμό MCAR. Τα μαύρα κουτάκια απεικονίζουν τις υπάρχουτες τιμές ενώ τα άσπρα τις απουσιάζουσες.Στην περίπτωση αυτή οι στήλες 1 και 2 είναι συσχετιζόμενες και σκούρο μπλε στην στήλη 1 σημαίνει απουσιάζουσα τιμή στην στήλη 2.*

2.2 Μηχανική μάθηση

2.2.1 Ορισμός Μηχανικής μάθησης

Machine Learning / Μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με την υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών. Έχει ισχυρούς δεσμούς με την μαθηματική βελτιστοποίηση, η οποία της παρέχει μεθόδους, την θεωρία και τομείς εφαρμογής.

Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. Η Μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, όπου η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων, γνωστή και ως μη επιτηρούμενη μάθηση[1].

2.2.2 Βασικά Είδη Μηχανικής Μάθησης

1. Επιβλεπόμενη μάθηση

[1] https://www.csc.com.gr

Η Επιβλεπόμενη μάθηση είναι μία κατηγορία μηχανικής μάθησης, στόχος της οποίας είναι ο χαρακτηρισμός δεδομένων με βάση κάποια δεδομένα εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο παραδειγμάτων τα οποία χρησιμοποιούνται για εκπαίδευση μοντέλων. Στην επιβλεπόμενη μάθηση, κάθε παράδειγμα αποτελείται από ένα σύνολο εισόδου (συνήθως ένα διάνυσμα από χαρακτηριστικά) και μια επιθυμητή τιμή εξόδου. Οι Αλγόριθμοι επιβλεπόμενης μάθησης αναλύουν τα δεδομένα εκπαίδευσης και παράγουν ένα μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει νέα παραδείγματα. Το βέλτιστο σενάριο επιτρέπει στον αλγόριθμο να καθορίσει σωστά την ετικέτα της κατηγορίας για άγνωστα μέχρι τώρα παραδείγματα. Για να επιτευχθεί αυτό, απαιτείται ο αλγόριθμος μάθησης να γενικεύει από τα δεδομένα εκπαίδευσης σε αθέατες καταστάσεις με ένα ¨λογικό¨ τρόπο.

2. Μη επιβλεπόμενη μάθηση

Η μη επιβλεπόμενη μάθηση αποτελεί κατηγορία της μηχανικής μάθησης, στόχος της οποίας είναι η ανακάλυψη πιθανής δομής που μπορεί να κρύβεται πίσω από μη χαρακτηρισμένα δεδομένα. Εφόσον τα παραδείγματα τα οποία χρησιμοποιούνται δεν είναι χαρακτηρισμένα, δεν υπάρχει σφάλμα ή σήμα ανταμοιβής για να αξιολογηθούν οι πιθανές λύσεις. Αυτό είναι που διακρίνει την μη-επιβλεπόμενη μάθηση από την επιβλεπόμενη μάθηση και την ενισχυτική (ημι-επιβλεπόμενη) μάθηση.

3. Ενισχυτική μάθηση

Η ενισχυτική μάθηση (ρεινφορσεμεντ λεαρνινγ) στην επιστήμη των υπολογιστών είναι ένας γενικός όρος που έχει δοθεί σε μια οικογένεια τεχνικών στις οποίες το σύστημα μάθησης προσπαθεί να μάθει μέσα από την άμεση αλληλεπίδραση με το περιβάλλον. Εφαρμόζεται στον έλεγχο κίνησης ρομπότ, στη βελτιστοποίηση εργασιών σε εργοστάσια, στη μάθηση επιτραπέζιων παιχνιδιών, κτλ. Η έννοια της ενισχυτικής μάθησης είναι εμπνευσμένη από τα αντίστοιχα ανάλογα της μάθησης με επιβράβευση και τιμωρία που συναντώνται ως μοντέλα μάθησης των έμβιων όντων. Σκοπός του συστήματος μάθησης είναι να μεγιστοποιήσει μια συνάρτηση του αριθμητικού σήματος ενίσχυσης (ανταμοιβή), για παράδειγμα την αναμενόμενη τιμή του σήματος ενίσχυσης στο επόμενο βήμα. Το σύστημα δεν καθοδηγείται από κάποιον εξωτερικό επιβλέποντα για το ποια ενέργεια θα πρέπει να ακολουθήσει αλλά πρέπει να ανακαλύψει μόνο του ποιες ενέργειες είναι αυτές που θα του αποφέρουν το μεγαλύτερο κέρδος.

2.2.3 Κλασσικοί αλγόριθμοι Μηχανικής μάθησης

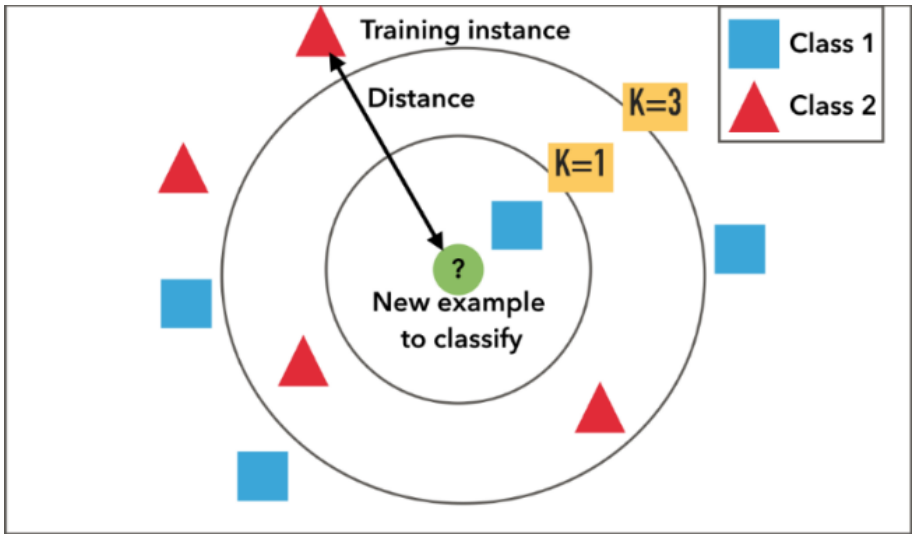
Στην συνέχεια αναλύουμε το υπόβαθρο για κάποιους βασικούς αλγορίθμους μηχανικής μάθησης (επιβλεπομένης η μη) οι οποίοι αποτελούν την βάση για τις μεθόδους αποτίμησης απουσιαζόντων τιμών που θα χρησιμοποιηθούν και θα αναλυθούν. Στο παρούσα παράγραφο γίνεται μαι εισαγωγή στις πιο κλασσικές μεθόδους μηχανικής μάθησης και ενώ παρακάτω αναλύονται και πιο σύνθετες βασισμένες σε νευρωνικά δίκτυα.

Κ-κοντινότεροι γείτονες KNN

Ο αλγόριθμος αυτός είναι μια μη παραμετρική μέθοδος μηχανικής μάθησης η οποία βασίζεται στην απόσταση. Με το μη παραμετρικος εννοειται ότι δεν υπάρχουν παράμετροι που μπορύν να βελτιστοποιηθούν κατα την εκπαίδευση του μοντέλου. Μπορεί να χρησιμοποιηθεί και για ταξινόμηση αλλά και για οπισθοδρόμηση. Η βασική ιδέα είναι η χρήση των k κοντινότερων γειτόνων προκειμένου να να βρεθεί η πλειοψηφία μια τιμής σε περίπτωση ταξινόμησης και η μέση τιμή στην περίπτωση της οπισθοδρώμησης. Η πιο σύνηθες συνάρτηση που χρησιμοποιείται για την έυρεση της απόστασης στον χώρο είναι η ευκλείδια απόσταση:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2.1}$$

Όπου τα q και p είναι 2 δείγματα ενός συνόλου δεδομένων. Γραφικά ο αλγόριθμος φαίνεται στο Σχ 2.4[2]. Εδώ βλέπουμε ένα νέο δείγμα (πράσινο) προς ταξινόμηση. Στην περίπτωση όπου το Κ είναι 1 ταξινομείται στην κλάση 1 όπως βλέπουμε στον μικρό κύκλο ενώ στην περίπτωση όπου το Κ είναι ίσο με 3 ταξινομείται στην κλάση 2 (μεγάλος κύκλος).



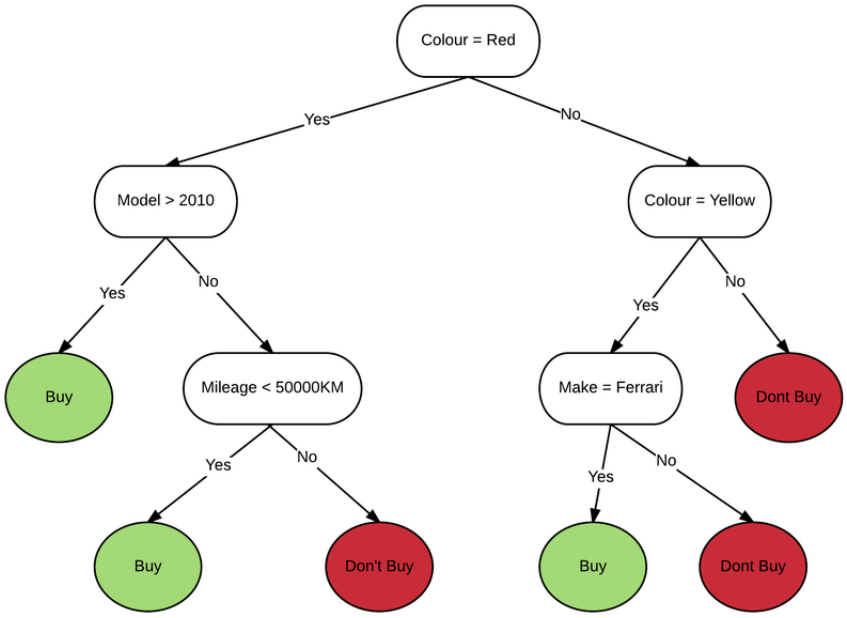
Σχήμα 2.4: *Ο αλγόριθμος k κοντινότεροι γείτονες.*

Δέντρο απόφασης

Τα δέντρα αποφάσεων είναι μια μέθοδος μηχανικής μάθησης με επίβλεψη που μπορεί να χρησιμοποιηθεί τόσο για την ταξινόμηση όσο και για την παλινδρόμηση. Η βασική ιδέα πίσω από τα δέντρα αποφάσεων είναι απλή: τα παραδείγματα που ανήκουν σε διαφορετικές κλάσεις έχουν τουλάχιστον μία διαφορετική τιμή σε ένα από τα χαρακτηριστικά τους [14]. Γι' αυτό το λόγο η μέθοδος λειτουργεί ταξινομώντας τα παραδείγματα με βάση τις τιμές των χαρακτηριστικών. Η ταξινόμηση γίνεται με μια συστηματικά διατεταγμένη σειρά ερωτήσεων

[2]https://medium.datadriveninvestor.com/knn-algorithm-and-implementation-from-scratch-b9f9b739c28f?gi=03edf822a8d2

έτσι ώστε κάθε ερώτηση να ρωτά ένα χαρακτηριστικό και να διακλαδίζεται με βάση την τιμή του χαρακτηριστικού [15]. Εκτός από τη μηχανική μάθηση, τα δέντρα αποφάσεων χρησιμοποιούνται για παράδειγμα στην εξόρυξη δεδομένων και στην επιχειρησιακή έρευνα [16]. Ένα παράδειγμα φαίνεται στο σχήμα 2.5[3].



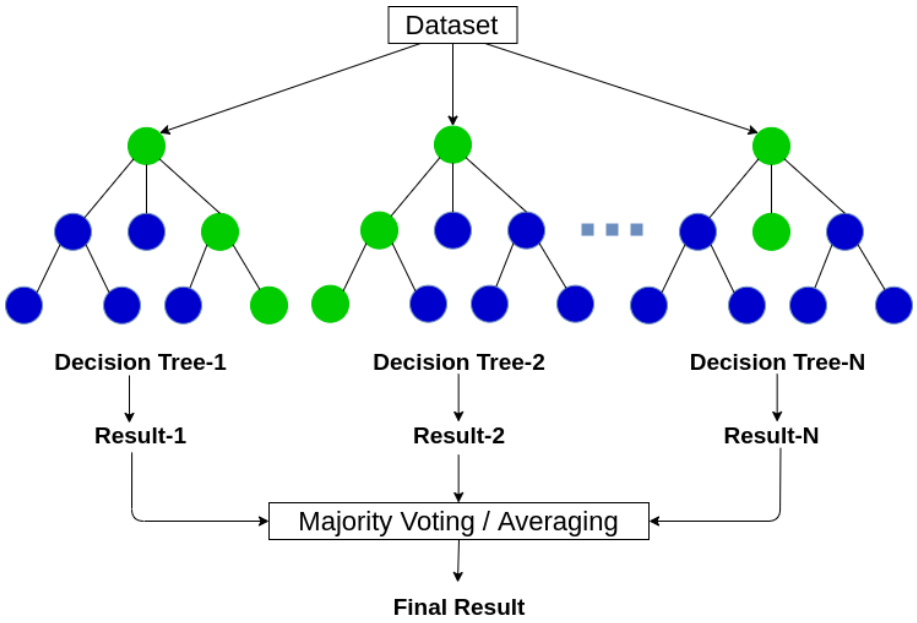
Σχήμα 2.5: *Παράδειγμα Δέντρου απόφασης.*

Τυχαίο Δάσος

Για να μπορέσουμε να δώσουμε μια περιγραφή του τυχαίου δάσους, πρέπει πρώτα να συζητήσουμε τα δέντρα απόφασης. Τα δέντρα αποφάσεων ή (στην περίπτωσή μας) τα δέντρα ταξινόμησης είναι μοντέλα που βασίζονται στην κατάτμηση του χώρου χαρακτηριστικών και στην αποθήκευση μιας κατανομής επί των ετικετών κλάσης για κάθε περιοχή. Αυτό μπορεί να επαναληφθεί χρησιμοποιώντας ένα δέντρο, το οποίο με τη σειρά του συνεπάγεται ότι η συνάρτηση υπόθεσης ή προβλεπτικός παράγοντας για αυτόν τον αλγόριθμο μάθησης έχει τη μορφή δέντρου. Κάθε φύλλο αυτού του δέντρου θα αντιστοιχεί στη συνέχεια σε κάθε περιοχή και του αποδίδεται η αντίστοιχη πιθανότητα σχετική με τις κλάσεις. Με βάση αυτή την περιγραφή, μπορούμε να χρησιμοποιήσουμε τα δέντρα αποφάσεων για να επινοήσουμε προγνωστικούς δείκτες για κάθε περίπτωση δοκιμής. Παρόλο που τα δέντρα αποφάσεων είναι εύκολα στην ερμηνεία και είναι διαισθητικά, οι προβλέψεις τους δεν είναι τόσο ακριβείς όσο για άλλους τύπους προβλεπτών, όπως αυτοί που βασίζονται στη λογιστική παλινδρόμηση. Επιπλέον, η δομή του δέντρου είναι πολύ ευαίσθητη στα παρεχόμενα δεδομένα. Αυτό σημαίνει ότι μικρές αλλαγές στα δεδομένα μπορεί να έχουν δραστικές επιπτώσεις στην προκύπτουσα δενδρική δομή. Προκειμένου να αντιμετωπιστεί αυτό το ζήτημα, χρησιμοποιείται

[3]https://venngage.com/blog/what-is-a-decision-tree/

ο αλγόριθμος τυχαία δάση. Αυτό το μοντέλο μπορούν να θεωρηθούν ως συνδυασμός πολλαπλών δέντρων απόφασης. Πιο συγκεκριμένα, σχεδιάζονται με την εκπαίδευση διαφορετικών δέντρων σε διαφορετικά υποσύνολα δεδομένων που επιλέγονται τυχαία με αντικατάσταση. Η πρόβλεψη που προκύπτει από ένα τυχαίο δάσος παράγεται από τον συνδυασμό αυτών των δέντρων απόφασης. Σχηματικά αυτό φαίνεται στο Σχήμα 2.6.[4]



Σχήμα 2.6: *Ο αλγόριθμος τυχαία δάση.*

2.2.4 Νευρωνικά Δίκτυα

Στην παράγραφο αυτή αρχικά γίνεται μια γενική εισαγωγή στα φυσικά άλλα και στα τεχνητά νευρωνικά δίκτυα Στην συνέχεια γίνεται η ανάλυση πιο σύνθετων μεθόδων βασισμένων σε αυτά οι οποίες αποτελούν την βάση για τις μεθόδους αποτίμησης απουσιαζόντων τιμών με χρήση βαθιάς μάθησης που θα παρουσιαστούν παρακάτω.

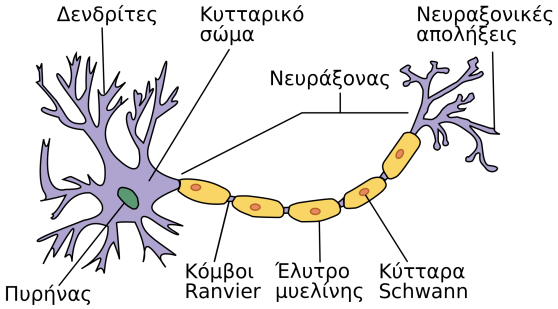
Φυσικά νευρωνικά Δίκτυα

Νευρωνικό δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων 2.7.[5] Στην περίπτωση βιολογικών νευρώνων, πρόκειται για ένα τμήμα νευρικού ιστού. Στην περίπτωση τεχνητών νευρώνων, πρόκειται για ένα αφηρημένο αλγοριθμικό κατασκεύασμα το οποίο εμπίπτει στον τομέα της υπολογιστικής νοημοσύνης, όταν στόχος του νευρωνικού δικτύου είναι η επίλυση κάποιου υπολογιστικού προβλήματος, ή της υπολογιστικής νευροεπιστήμης, όταν στόχος είναι η υπολογιστική προσομοίωση της λειτουργίας των βιολογικών νευρωνικών δικτύων με βάση κάποιο μαθηματικό μοντέλο τους.

Το παρόν άρθρο αφορά τα τεχνητά νευρωνικά δίκτυα, τους τεχνητούς νευρώνες και τα μοντέλα βιολογικών νευρώνων της υπολογιστικής νευροεπιστήμης. Για τα βιολογικά νευρωνικά

[4]https://medium.com/@curryrowan/the-complete-guide-to-random-forests-part-2-934eabf35534
[5]https://en.wikipedia.org/wiki/Neuron

δίκτυα δείτε τα άρθρα νευρώνας και Κεντρικό Νευρικό Σύστημα[6].



Σχήμα 2.7: *Σχηματικό διάγραμμα ενός τυπικού νευρώνα*

Τεχνητά νευρωνικά δίκτυα

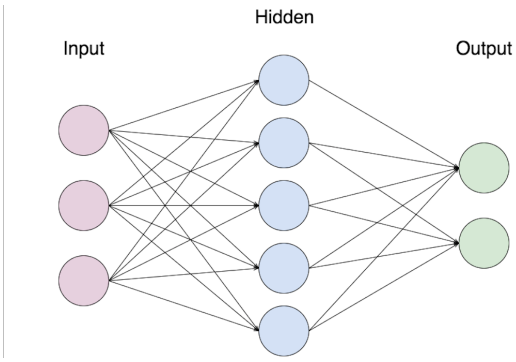
Νευρωνικό δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων μονάδων επεξεργασίας που ονομάζουμε Νευρώνες. Στους υπολογιστές είναι ένα υπολογιστικό μοντέλο που χρησιμοποιείται για την επίλυση κάποιου υπολογιστικών προβλημάτων.
Το νευρωνικό δίκτυο ονομάζεται δίκτυο καθώς αποτελείται απο υπολογιστικούς κόμβους που συνδέονται μεταξύ τους. Κάθε υπολογιστικός κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων (από άλλους νευρώνες είτε από κάποια άλλη είσοδος), εκτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η έξοδος από αυτόν τον κόμβο μπορεί είτε να αποτελέσει μέρος της συνολικής εξόδου του ΤΝΔ είτε να διοχετευτεί σε άλλους κόμβους Γενικά (αν και δεν ειναι υποχρεωτική η διάκριση αυτή) θεωρούμε πως υπάρχουν τριών ειδών νευρώνες
1. Οι νευρώνες εισόδου, των οποίων η εργασία είναι να διοχετεύσουν στους υπολογιστικούς νευρώνες την είσοδο του προβλήματος (πχ πρότυπα).
2. Οι νευρώνες εξόδου, χρησιμοποιούνται για να παρουσιάσουν στο περιβάλλον την απάντηση του ΤΝΔ σε κάποιο πρόβλημα, όπως για παράδειγμα την εκτίμηση της κατηγορίας ενός προβλήματος κατηγοριοποίησης.
3. Οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες, οι οποίοι πολλαπλασιάζουν κάθε είσοδό που δέχονται από νευρώνες εισόδου ή από άλλους νευρώνες επεξεργασίας με μια τιμή συσχετισμένη με αυτούς που ονομάζεται βάρος.
Το συνολικό αποτέλεσμα εισάγεται σε μια συνάρτηση που θα ονομάζουμε συνάρτηση ενεργοποίησης και είτε παρουσιάζεται στην έξοδο είτε δίδεται σε κάποιον άλλο νευρώνα επεξεργασίας. Ένα σχηματικό παράδειγμα ενός απλού ΤΝΔ παρουσιάζεται στην εικόνα 2.8. Στο σχήμα αυτό υπάρχουν δύο νευρώνες εισόδου, τέσσερις νευρώνες επεξεργασίας και ένας νευρώνας εξόδου. Αν θέλουμε να εκφράσουμε την έξοδο ενός νευρώνα επεξεργασίας k θα μπορούσαμε να χρησιμοποιήσουμε την παρακάτω εξίσωση:

[6]https://en.wikipedia.org/wiki/Neural_network

$$y_k = f(\sum_{i=0}^{d} w_i x_i + \vartheta_k) \tag{2.2}$$

όπου:

1. x_i είναι η i είσοδος από τις d που έχει το πρόβλημα.
2. w_i είναι το βάρος της διασύνδεσης με την i είσοδο.
3. ϑ_k είναι η πόλωση για τον νευρώνα k. Η τιμής της πόλωσης συνήθως είναι ανεξάρτητη από το πρόβλημα.
4. Η συνάρτηση $f(x)$ είναι η συνάρτηση ενεργοποίησης και στην βιβλιογραφία χρησιμοποιείται μια πληθώρα συναρτήσεων ενεργοποίησης που παρουσιάζονται στην επόμενη ενότητα.

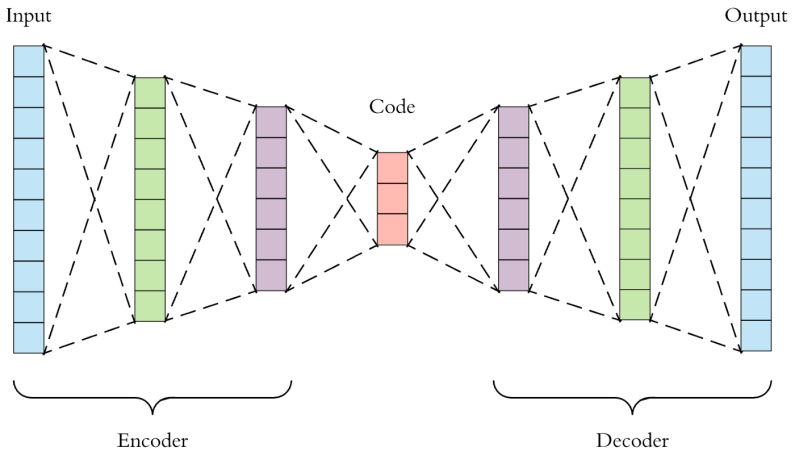


Σχήμα 2.8: *Σχηματικό διάγραμμα ενός τεχνιτού νευρωνικού δικτύου*

Αυτοκωδικοποιητές

Ο αυτοκωδικοποιητής είναι μια μέθοδος μη επιβλεπομένης μάθησης όπου ο στόχος είναι η αναπαραγωγή της εισόδου στην έξοδο,δηλαδή $y_i = x_i$. Αποτελείται από 2 ακολουθιακά νευρωνικά δίκτυα τα οποία συνήθως είναι ίδιας αρχιτεκτονικής και ονομάζονται κωδικοποιητής και αποκωδικοποιητής (Εικόνα 2.9).
Ο αυτοκωδικοποιητής προσπαθεί να μάθει μια συνάρτηση της μορφής $h_{W,b} = x$, δηλαδή μια εκτίμηση της συνάρτησης ταυτότητας σύμφωνα με την οποία η έξοδος θα είναι ίδια με την είσοδο.Η διαδικασία αυτή αρχικά φαντάζει τετριμμένη ωστόσο με τον να βάζουμε περιορισμούς στο δίκτυο μπορούν να ανακαλυφθούν ενδιαφέρον μοτίβα του συνόλου δεδομένων.Τέτοιος περιορισμοί μπορεί να είναι η μείωση του αριθμού τον νευρώνων σε κάποια στήλη του δικτύου έτσι ώστε αυτή να είναι μικρότερη απο την είσοδο. Με τον τρόπο αυτό, προκειμένου να γίνει σωστά η μεταφορά της εισόδου στην έξοδο ο αυτοκωδικοποιητής πρέπει να μάθει τις

συσχετίσεις μεταξύ των μεταβλητών.



Σχήμα 2.9: *Σχηματικό διάγραμμα ενός αυτοκωδικοποιητή.*

Λόγω της ικανότητας των αυτοκωδικοποιητών να ανακαλύπτουν κρυφές σχέσεις μεταξύ των χαρακτηριστικών είναι ιδιαίτερα χρήσιμοι σε προβλήματα που αφορούν ανίχνευση ανωμαλιών, μείωση διαστατικοτητας, αποτίμηση απουσιαζόντων τιμών κ.α.

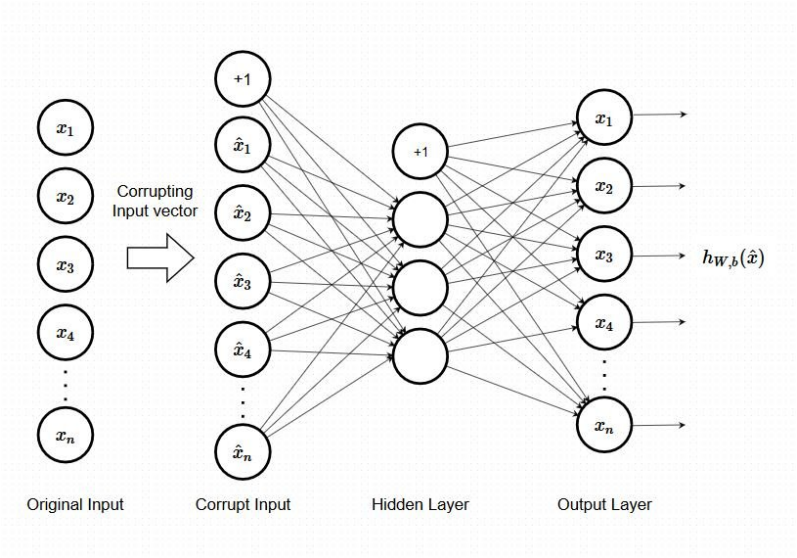
Αυτοκωδικοποιητές αποθορυβοποίησης

Οι αυτοκωδικοποιητές αποθορυβοποίησης δέχονται ως είσοδο ένα αλλοιωμένο με θόρυβο σύνολο δεδομένων και πρέπει να το μετατρέψουν στο πραγματικό και ολόκληρο σύνολο (Εικόνα 2.10). Με τον τρόπο αυτό μαθαίνουν να αφαιρούν τον θόρυβο από τα δεδομένα[7]. Για να γίνει Αυτό με επιτυχία πρέπει να μάθει ο αυτοκωδικοποιητής τις κρυφές συσχετίσεις μεταξύ των χαρακτηριστικών.
Η διαδικασία την εκπαίδευσης έχει ως εξής:

- Η αρχική είσοδος x μετατρέπεται σε διεφθαρμένη $\tilde{x}$ με στοχαστική αντικατάσταση $\tilde{x} \sim q_d d(\tilde{x}|x)$.
- Η διεφθαρμένη είσοδος απεικονίζεται στην κρυφή αναπαράσταση όπως και στον κανονικό αυτοκωδικοποιητή, $h(x) = f_\theta(\tilde{x}) = s(W\tilde{x} + b)$.
- Απο την κρυφλη αναπαράσταση το μοντέλο ανακατασκευάζει $z = g_{\backslash\theta}(h)$

Οι παράμετροι θ και $\backslash\theta$ εκπαιδεύονται έτσι ώστε να ελαχιστοποιούν το σφάλμα ανακατασκευής. Ο θόρυβος μπορεί να είναι οποιασδήποτε μορφής όπως π.χ. Γκαουσιανός η θόρυβος salt and pepper και εφαρμόζεται σε κάθε δείγμα πριν την εκπαίδευση.

[7] https://en.wikipedia.org/wiki/Autoencoder

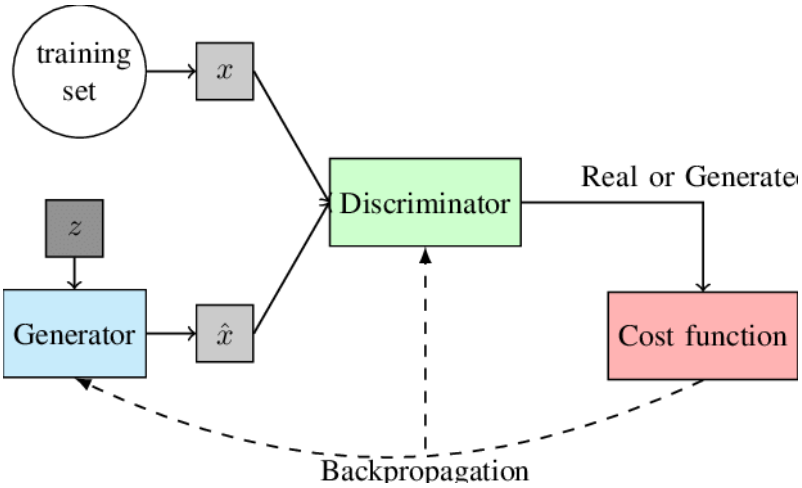


Σχήμα 2.10: *Σχηματικό διάγραμμα ενός αυτοκωδικοποιητή αποθορυβοποίησης.*

Γενετικά ανταγωνιστικά δίκτυα

Τα Παραγωγικά Αντιπαλικά Δίκτυα (ΠΑΔ) (Εικόνα 2.11), γνωστά επίσης ως Αντιπαλικά Δίκτυα, Παραγωγικά Ανταγωνιστικά Δίκτυα, Παραγωγικά Δίκτυα Αντιπάλων και Αναγεννητικά Ανταγωνιστικά Δίκτυα (στα αγγλικά Generative Adversarial Networks - GAN) είναι μια κατηγορία συστημάτων μηχανικής μάθησης που εφευρέθηκε από τον Ian Goodfellow και τους συναδέλφους του το 2014. Βασίζονται στην λογική της αντιπαλικής μάθησης. Δύο νευρωνικά δίκτυα διαγωνίζονται σε ένα παίγνιο (με την έννοια της θεωρίας παιγνίων, συχνά αλλά όχι πάντα με τη μορφή ενός παιγνίου μηδενικού αθροίσματος). Δοθέντος ενός συνόλου εκπαίδευσης, αυτή η τεχνική μαθαίνει να δημιουργεί νέα δεδομένα με τα ίδια στατιστικά στοιχεία. Για παράδειγμα, ένα αντιπαλικό δίκτυο εκπαιδευμένο σε φωτογραφίες μπορεί να δημιουργήσει νέες φωτογραφίες που φαίνονται τουλάχιστον επιφανειακά αυθεντικές στους ανθρώπινους παρατηρητές, έχοντας πολλά ρεαλιστικά χαρακτηριστικά. Αν και αρχικά προτάθηκαν αμιγώς ως μορφή παραγωγικού μοντέλου για εφαρμογές μη επιβλεπόμενη μάθηση, τα ΠΑΔ έχουν επίσης αποδειχθεί χρήσιμα για την ημι-εποπτευόμενη μάθηση, την πλήρως εποπτευόμενη μάθηση και ενισχυτική μάθηση. Σε ένα σεμινάριο του 2016, ο Yann LeCun περιέγραψε τα GAN ως ¨την πιο έξυπνη ιδέα στη μηχανική μάθηση τα τελευταία είκοσι χρόνια¨.

Το παραγωγικό δίκτυο δημιουργεί υποψηφίους ενώ το διαχωριστικό δίκτυο τους αξιολογεί. Ο διαγωνισμός λειτουργεί με όρους διανομών δεδομένων. Το παραγωγικό δίκτυο μαθαίνει να προβάλλει από έναν λανθάνοντα χώρο σε μια επιθυμητή κατανομή δεδομένων, ενώ το διαχωριστικό δίκτυο διακρίνει τους παραγμένους υποψήφιους από την πραγματική κατανομή. Ο στόχος εκπαίδευσης του παραγωγικού δικτύου είναι η αύξηση του ποσοστού σφάλματος του διακριτικού δικτύου (δηλ. να ¨ξεγελάσει¨ το διαχωριστικό δίκτυο με την παραγωγή νέων υποψηφίων που το διαχωριστικό δίκτυο πιστεύει ότι προήλθαν από την πραγματική κατανομή).



Σχήμα 2.11: *Σχηματικό διάγραμμα ενός γενετικού ανταγωνιστικού δικτύου.*

Ένα γνωστό σύνολο δεδομένων χρησιμεύει ως το αρχικό σύνολο εκπαίδευσης για τον διαχωριστή. Το διαχωριστικό μοντέλο εκπαιδεύεται παρουσιάζοντάς το με δείγματα από το σύνολο δεδομένων εκπαίδευσης, μέχρι να επιτευχθεί αποδεκτή ακρίβεια. Το παραγωγικό μοντέλο εκπαιδεύεται βάσει του κατά πόσο καταφέρνει να ξεγελάσει το διαχωριστικό. Συνήθως, το παραγωγικό σπέρνεται με τυχαία εισερχόμενη δειγματοληψία από προκαθορισμένο λανθάνοντα χώρο (π.χ. πολλαπλή κανονική κατανομή). Στη συνέχεια, οι παραγμένοι υποψήφιοι αξιολογούνται από το διαχωριστικό. Οπισθοδιάδοση εφαρμόζεται και στα δύο δίκτυα έτσι ώστε το παραγωγικό μοντέλο να παράγει καλύτερες εικόνες, ενώ το διαχωριστικό γίνεται πιο εξειδικευμένο στην ανίχνευση συνθετικών εικόνων. Το παραγωγικό μοντέλο είναι συνήθως ένα νευρωνικό δίκτυο ενώ το διαχωριστικό είναι ένα συνελικτικό νευρικό δίκτυο[8]. Η εκπαίδευση των δικτύων αυτών βασίζεται στην παρακάτω εξίσωση και είναι ουσιαστικά μια εκδοχή του παιχνιδιού μεγίστου ελαχίστου :

$$\min_G \max_D \mathbb{E}_{x\sim p_{\delta\alpha\tau\alpha}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_\zeta(z)}[1 - \log D(G(z))] \tag{2.3}$$

Ο Γεννήτορας προσπαθεί να μεγιστοποιήσει αυτήν την τιμή ενώ ο παραγωγός επιδιώκει να την ελαχιστοποιήσει.

[8]https://en.wikipedia.org/wiki/Generative_adversarial_network

Κεφάλαιο 3

Σχετική δουλειά στην βιβλιογραφία

Υπάρχουν πολλοί τρόποι αντιμετώπισης του προβλήματος των απουσιαζόντων τιμών και η κατάλληλη λύση κάθε φορά μπορεί να είναι διαφορετική. Μερικοί από τους βασικούς λόγους που μπορεί να καθορίσουν την επιλογή αυτή είναι το ποσοστό απουσιαζόντων τιμών, ο μηχανισμός απουσιαζόντων τιμών που χαρακτηρίζει τις μεταβλητές του συνόλου δεδομένων αλλά και η ίδια η μορφή των δεδομένων (είτε αυτή είναι μορφής εικόνας είτε είναι μορφής πίνακα).

3.1 Διαγραφή (Case deletion)

Ο πιο απλό τρόπος αντιμετώπισης των απουσιαζόντων τιμών είναι η διαγραφή των εγγραφών που δεν είναι πλήρης και ονομάζεται (List-wise) [17, 18, 19, 20, 21] διαγραφή. Η μέθοδος αυτή μπορεί να οδηγήσει στην απώλεια μεγάλου μέρους του συνόλου δεδομένων προσδίδοντας (bias) στην ανάλυση. Η (pair-wise)[17] στοχεύει στο να ελαχιστοποιήσει την απώλεια δεδομένων που αναφέρθηκε πριν.Στην περίπτωση αυτή μια εγγραφή διαγράφεται μόνο όταν περιέχει απουσιάζοντες τιμές σε κάποια μεταβλητή η οποία χρησιμοποιείται στο μοντέλο. Ωστόσο, στην περίπτωση αυτή αν υπάρχουν πολλά διαφορετικά στατιστικά προς υπολογισμό στο σύνολο δεδομένων και κάθε ένα από αυτά αφορά διαφορετικό σύνολο μεταβλητών μπορεί τελικά κάθε ένα από αυτά να έχει υπολογιστεί από διαφορετικό υποσύνολο. Ακόμα, και στις 2 περιπτώσεις είναι απαραίτητο να ισχύει ο μηχανισμός (MCAR) ως προς την απουσία των δεδομένων.

3.2 Μέθοδοι απόδοσης τιμών (imputation)

Μια πιο σωστή λύση για το προαναφερθέν πρόβλημα είναι η απόδοση ρεαλιστικών τιμών στις απουσιάζουσες με χρήση στατιστικών μεθόδων η με πιο σύνθετων αλγορίθμων (π.χ. νευρωνικά δίκτυα).

3.2.1 Λύσεις βασισμένες στην στατιστική

Στην περίπτωση αυτή οι απουσιάζουσες τιμές αντικαθίστανται με άλλες ρεαλιστικές που προκύπτουν από τα στατιστικά και τις συσχετίσεις του συνόλου δεδομένων. Σε αυτήν την περίπτωση, μια από τις πιο διαδεδομένες τεχνικές είναι η αντικατάσταση με την μέση, πιο συχνή η την διάμεση τιμή της ίδιας της μεταβλητής [22]. Η περίπτωση αυτή έχει το πλεονέκτημα του ότι διατηρούνται τα στατιστικά(όπως π.χ. η μέση τιμή ανά μεταβλητή) του συνόλου δεδομένων. Ωστόσο, με αυτήν την τεχνική δεν λαμβάνεται υπ'όψιν η σχέση που μπορεί να υπάρχει μεταξύ των μεταβλητών και συνήθως τα αποτελέσματα είναι biased.
Μία άλλη μέθοδος είναι η οποία εφαρμόζεται σε δεδομένα με χρονική εξάρτηση είναι η αντικατάσταση με την πιο πρόσφατη τιμή που παρατηρήθηκε. Για παράδειγμα, έστω ότι μελετάμε κάποιους ιατρικούς δείκτες για ασθενείς με διάστημα μελέτης 2 χρόνων όπως το BMI (Body Mass Index). Αν ο δείκτης αυτός για έναν ασθενή απουσιάζει για τον δεύτερο μήνα της μελέτης θα χρησιμοποιηθεί ο τελευταίος δείκτης του πρώτου μήνα ως τιμή αντικατάστασης. Είναι φανερό ωστόσο ότι για μεταβλητές με μεγάλη διακύμανση η για πολλές συνεχόμενες απουσιάζουσες τιμές η μέθοδος αυτή δεν δίνει καλά αποτελέσματα.
Οι Mir, Adil Aslam κ.α. [23] δοκιμάζουν διάφορες μεθόδους απόδοσης απουσιαζόντων τιμών μεταξύ των οποίων είναι και η λύση της μέσης η της πιο συχνά εμφανιζόμενης τιμής. Για την αξιολόγηση των μεθόδων χρησιμοποιούν ένα πραγματικό σύνολο δεδομένων με 15692 παρατηρήσεις. Τα αποτελέσματα δείχνουν ότι αυτή η μέθοδος έχει τα χειρότερα αποτελέσματα και για το στάδιο της αποτίμησης τιμών αλλά και στο κομμάτι που αφορά την πρόβλεψη.
Οι Gupta κ.α. [24] αξιολόγησαν μοντέλα μηχανικής μάθησης για το πρόβλημα της πρόβλεψης καρδιακής προσβολής χρησιμοποιώντας το σύνολο δεδομένων της καρδιακής μελέτης Framingham και το σύνολο δεδομένων Heart για το UCI Machine Learning Repository. Σε ότι αφορά την προεπεξεργασία και ειδικότερα τις ελλείπουσες τιμές, η προσέγγισή τους ήταν να προσδώσουν με τη χρήση του μέσου όρου ή της διαμέσου, όπου η τελευταία προτιμήθηκε για χαρακτηριστικά με λοξές κατανομές.
Οι Guo κ.α. [25] ανέπτυξαν μια προσέγγιση βαθιάς μάθησης για το πρόβλημα της πρόβλεψης της καρδιακής ανεπάρκειας χρησιμοποιώντας συνθετικά δεδομένα EHR. Στην ανάλυσή τους, επέλεξαν να απορρίψουν τα χαρακτηριστικά με περισσότερες από 50% των καταχωρίσεων να λείπουν και να προσδώσουν τα υπόλοιπα χρησιμοποιώντας το μέσο όρο και το συχνότερο για αριθμητικά και κατηγορικά χαρακτηριστικά αντίστοιχα.
Ακόμα μέθοδοι βασισμένες σε βαθιά νευρωνικά δίκτυα και πιο συγκεκριμένα αυτοκωδικοποιητές και γενετικά αντιπαλικά νευρωνικά δίκτυα έχουν εφαρμοστεί σε ερευνητικές δουλειές ως κομμάτια μια συνολικής προσέγγισης. Η προσέγγιση αυτή αφορά την παλινδρόμηση η την κατάτμησή με την χρήση δεδομένων Ηλεκτρονικού φακέλου υγείας [26, 27].

3.2.2 Λύσεις βασισμένες σε μεθόδους μηχανικής μάθησης

Ως προς την μηχανική μάθηση, ο αλγόριθμος KNN [28] μπορεί να εφαρμοστεί σε δεδομένα με απουσιάζουσες τιμές όπου αυτές αγνοούνται κατά τον υπολογισμό της απόστασης μεταξύ των δειγμάτων. Ο αλγόριθμος αυτός υπολογίζει την απόσταση μεταξύ των δειγμάτων και αντικαθιστά κάθε απουσιάζουσα τιμή με τον μέσο όρο από τα K δείγματα τα οποία βρίσκονται στον χώρο πιο κοντά (για τα οποία η συγκεκριμένη μεταβλητή δεν είναι απουσι-

άζουσα). Συχνά, οδηγεί σε καλά αποτελέσματα άλλα είναι ευαίσθητος στην επιλογή του K αλλά και στην επιλογή της μετρικής. Οι Anil Jadhav κ.α. [29] δοκίμασαν διάφορες τεχνικές αποτίμησης απουσιαζόντων τιμών όπως ο αλγόριθμος των κ-κοντινότερων γειτόνων και η αποτίμηση με βάση την μέση τιμή. Για να αξιολογήσουν τους αλγορίθμους αυτούς χρησιμοποίησαν 5 δημόσια σύνολα δεδομένων απο το γνωστό αποθετήριο UCI. Τα πειράματα έδειξαν ότι η μέθοδος KNN είχε την μεγαλύτερη απόδοση πετυχαίνοντας συστηματικά μικρότερο μέσο τετραγωνισμένο λάθος για διάφορα ποσοστά απουσιαζόντων τιμών.
Επιπλέον, οι κλασσικοί αλγόριθμοι παλινδρόμησης μπορούν να χρησιμοποιηθούν για το πρόβλημα αυτό όπου η εξαρτώμενη μεταβλητή είναι αυτή για την οποία παρουσιάζονται απουσιάζουσες τιμές. Οι τιμές που δημιουργούνται ωστόσο χαρακτηρίζονται από στατιστική αβεβαιότητα και για τον λόγο αυτό συχνά η διαδικασία επαναλαμβάνεται αρκετές φορές. Στο πλαίσιο αυτό, ο αλγόριθμος MICE (Multiple Imputation by Chained Equations) [30] είναι ένας από τους πιο διαδεδομένος στον τομέα και επιφέρει ικανοποιητικά αποτελέσματα. Η μέθοδος αυτή προβλέπει την τιμή για τις απουσιάζουσες τιμές συμπληρώνοντας αρχικά στις άλλες μεταβλητές όλες τις άλλες απουσιάζουσες τιμές και έπειτα χρησιμοποιώντας κάποιο προβλεπτικό μοντέλο (π.χ. Τυχαία δάση). Η διαδικασία αυτή επαναλαμβάνεται για ένα συγκεκριμένο αριθμό επαναλήψεων η μέχρι να ικανοποιηθεί κάποιο κριτήριο.

3.2.3 Λύσεις βασισμένες σε βαθιά μάθηση

Οι τελευταίες εξελίξεις στον τομέα αφορούν το πεδίο της βαθειάς μάθησης όπου γενετικά ανταγωνιστικά δίκτυα[31, 32] και αυτοκωδικοποιητές[33] έχουν προσαρμοστεί με συγκεκριμένο τρόπο για την αντιμετώπιση αυτού του προβλήματος. Για παράδειγμα, Οι Boseong ο Seo κ.ά. [34] δοκίμασαν έναν κωδικοποιητή αποθορυβοποίησης με προ-εισαγωγή KNN με δεδομένα αερίου. Η σύγκριση έγινε με άλλες κοινές προσεγγίσεις υπολογισμού ελλιπών τιμών. Τα αποτελέσματα έδειξαν ότι η λύση του αυτόματου κωδικοποιητή πέτυχε την καλύτερη απόδοση στο κομμάτι της απόδοσης απουσιαζόντων τιμών.
Οι Sungkyu Park κ.ά. [35] συγκέντρωσαν δεδομένα EHR από φορητές συσκευές με σκοπό την πρόβλεψη με βάση τη μηχανική μάθηση. Σε αυτό το σύνολο δεδομένων το ποσοστό των ελλιπών δεδομένων ήταν 2,83%. Για να το προσδώσουν αυτό αξιολόγησαν αρχικά τις μεθόδους υπολογισμού ελλιπών τιμών σε ένα πλήρες υποσύνολο των δεδομένων. Οι μέθοδοι ήταν: (i) GAIN, (ii) KNN, (iii) mean, mode. Τα αποτελέσματα δείχνουν ότι η προσέγγιση της βαθιάς μάθησης υπερτερεί των απλούστερων μεθόδων με σημαντική διαφορά.
Οι Weinan Dong κ.α. [36] αξιολόγησαν σύγχρονες μεθόδους υπολογισμού ελλιπών τιμών, όπως οι GAIN, MICE και Missforest. Χρησιμοποίησαν δύο πραγματικά σύνολα δεδομένων για να υποστηρίξουν τους ισχυρισμούς τους. Τα αποτελέσματα δείχνουν ότι η προσέγγιση βαθιάς μάθησης επιτυγχάνει καλύτερες επιδόσεις και αυτό είναι κάτι που είναι πιο εμφανές όταν το ποσοστό έλλειψης είναι υψηλό.

3.3 Ευρήματα μελέτης σχετικής βιβλιογραφίας

Όταν πρόκειται για εργασίες που βασίζονται σε δεδομένα EHR, παρατηρούμε ότι πολλοί ερευνητές δεν δίνουν μεγάλη έμφαση στο μέρος του υπολογισμού των ελλειπόντων τιμών, επιλέγοντας συχνά να απορρίψουν αυτές τις εγγραφές. Μια πολύ συνηθισμένη προσέγγιση είναι επίσης ο καταλογισμός με τη χρήση απλών μεθόδων, όπως ο μέσος όρος, ο τρόπος και ο KNN, οι οποίες είναι γνωστό ότι αγνοούν τις πολύπλοκες σχέσεις που καθορίζουν τα σύνολα ιατρικών δεδομένων. Όσον αφορά την έρευνα που χρησιμοποιεί την καρδιακή μελέτη Framingham, δεν βρήκαμε καμία εργασία που να χρησιμοποιεί σύνθετους αλγορίθμους για το μέρος του υπολογισμού των ελλειπόντων τιμών, παρόλο που οι ελλείψεις σε αυτό το σύνολο δεδομένων είναι σημαντικές. Βλέπουμε επίσης ότι οι ερευνητές δεν αξιολογούν τις στρατηγικές υπολογισμού χρησιμοποιώντας πρόβλεψη μετά τον υπολογισμό.
Ως εκ τούτου, η εργασία μας διαφέρει σημαντικά από τις ερευνητικές εργασίες που αναφέρθηκαν παραπάνω, καθώς χρησιμοποιούμε σύγχρονες μεθόδους βαθιάς μάθησης, κάνουμε βελτιώσεις όσον αφορά την περίπτωσή μας χωρίς απώλεια γενικότητας και επιπλέον δοκιμάζουμε τις μεθόδους υπολογισμού μας για το έργο της πρόβλεψης.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Διατύπωση προβλήματος και μεθόδων

Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά τα βήματα που ακολουθήθηκαν για την υλοποίηση του συστήματος. Αρχικά γίνεται μια μαθηματική περιγραφή του προβλήματος των απουσιαζόντων τιμών καθώς και την πρόβλεψη που ακολουθεί αυτήν την διεργασία. Στα πλαίσια αυτής της διπλωματικής θα εξεταστούν και τα δυο αυτά σενάριο για μια πιο ολοκληρωμένη σύγκριση των μεθόδων. Έπειτα περιγράφεται η επιλογή του συνόλου δεδομένων και τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν. Για το σύνολο δεδομένων δίνονται πληροφορίες σχετικές με το μέγεθος το ποσοστό απουσιαζόντων τιμών άλλα και τον τύπο των γνωρισμάτων. Στη συνέχεια αναλύεται η υλοποίηση και οι βασικοί αλγόριθμοι που θα εξεταστούν. Σε αυτούς καταλογίζονται οι έτοιμοι αλγόριθμοι από την βιβλιογραφία καθώς και δυο προτεινόμενες μέθοδοι που αναπτύχθηκαν στα πλαίσια αυτής τής διπλωματικής.

4.1 Διατύπωση προβλήματος

4.1.1 Αποτίμηση απουσιαζίντων τιμών

Το πρόβλημα των απουσιαζόντων τιμών συμβαίνει όταν σε ένα σύνολο δεδομένων κάποιες μεταβλητές έχουν κενές τιμές. Ας υποθέσουμε μια τυχαία μεταβλητή $X = (X_1, X_2, .., X_N) \in X^N$ όπου το X αναπαριστά τον χώρο που ανήκει κάθε δείγμα με κατανομή $P(X)$. Ας υποθέσουμε επίσης ένα διάνυσμα μάσκας $M = (M_1, M_2, .., M_N)$ όπου κάθε M_i παίρνει τιμές στο $\{0, 1\}$ και $M_i = 1$ σημαίνει ότι η τιμή υπάρχει σε αντίθεση με το $M_i = 0$ που δείχνει ότι η συγκεκριμένη τιμή απουσιάζει. Έχοντας d στιγμιότυπα των X και M ορίζουμε ένα σύνολο δεδομένων (X^i, M^i) για $i = 0, 1, .., d$. Έπειτα, το $(\hat{X}^i, M^i)$ προκύπτει αντικαθιστώντας κάθε μεταβλητή j συνδεόμενη με ένα δείγμα i αν $M_{i,j} = 0$ με μια αρχική τιμή (πολλές φορές τυχαίος θόρυβος η 0). Σε μια τέτοια περίπτωση, δοθέντος ενός μοντέλου IMP ο στόχος είναι η κατασκευή ενός συνόλου δεδομένων $\tilde{X}^i = IMP(\hat{X}^i, M^i)$ για κάθε $i = 0, 1, .., d$. Κάθε αποδιδόμενο δείγμα θα πρέπει να δίνεται βάση της κατανομής $P(X|\tilde{X} = \tilde{X}^i)$ καθώς θέλουμε τα αποδιδόμενα δεδομένα να έχουν παρόμοια κατανομή με το πραγματικό σύνολο δεδομένων.
Το αποτέλεσμα είναι ένα σύνολο δεδομένων $\bar{X}$ όπου για κάθε δείγμα i έχουμε :

$$\bar{X}^i = X^i \odot M^i + (1 - M^i) \odot \tilde{X}^i \tag{4.1}$$

4.1.2 Πρόβλεψη μετά την αποτίμηση

Με την χρήση διαφορετικών αλγορίθμων αποτίμησης απουσιαζόντων τιμών A_i φορ $i = 0, 1, ..S$ στο αρχικό σύνολο δεδομένων X έχουμε S διαφορετικά σύνολα δεδομένων $\bar{D_1}, \bar{D_2}, .., \bar{D_S}$. Για κάθε ένα από αυτά τα σύνολα χρησιμοποιούμε μια προβλεπτική μέθοδο (π.χ. τυχαία δάση) για την πρόβλεψη της μεταβλητής σχετιζόμενη με την καρδιακή ασθένεια (δυαδική ταξινόμηση).

4.2 Δεδομένα

Το σύνολο ηλεκτρονικών δεδομένων υγείας που χρησιμοποιούμε είναι το Framingham heart study[1]. Πρόκειται για ένα σύνολο δεδομένων με διαδοχικές χρονικά μετρήσεις σε 4434 ασθενείς της πόλης Framingham των ΗΠΑ οι οποίο είτε πάσχουν είτε όχι απο καρδιακή ασθένεια. Για κάθε ασθενή, έχουν διεξαχθεί εργαστηριακά τεστ σε κάθε συνεδρία και έχουν μετρηθεί κάποιες μεταβλητές σχετιζόμενες με την καρδιακή ασθένεια όπως η συστολική πίεση στο αίμα, η διαστολική πίεση στο αίμα και το ζάχαρο. Ακόμα για κάθε ασθενή υπάρχει μια δυαδική μεταβλητή η οποία υποδηλώνει την ύπαρξη καρδιακής ασθένειας. Συνολικά, το σύνολο δεδομένων έχει 39 μεταβλητές αλλά για τα πλαίσια αυτής της διπλωματικής 15 θα χρησιμοποιηθούν. Από αυτές, 8 είναι συνεχείς και 7 κατηγορικές, όπου όλες οι κατηγορικές παίρνουν την τιμή 0 η την τιμή 1(δυαδικές). Το σύνολο αυτό επίσης έχει και μεταβλητό ποσοστό απουσιαζόντων τιμών από μεταβλητή σε μεταβλητή όπου μπορεί να είναι μεταξύ των 0 - 13%. Πιο συγκεκριμένα, πληροφορίες φαίνονται στον πίνακα 4.1.

4.3 Μέθοδοι

Στην παράγραφο αυτή αναλύουμε τις μεθόδους που αναπτύχθηκαν βασισμένες σε ήδη υπάρχοντες μεθόδους βαθιών νευρωνικών δικτύων. Οι επιλογές που κάναμε αφορούν την διαδικασία την εκπαίδευσης αλλά και την αρχιτεκτονική των δικτύων.

4.3.1 Simple

Το μοντέλο αυτό είναι η απλή στατιστική προσέγγιση του πιο συχνού, του μέσου καταλογισμού. Καταλογίζουμε τις κατηγορικές ελλείπουσες τιμές χρησιμοποιώντας την πιο συχνή κλάση και τις αριθμητικές μεταβλητές χρησιμοποιώντας τον μέσο όρο που προκύπτει από την αντίστοιχη στήλη.

4.3.2 KNN imputer

Ο KNN ο οποίος υπολογίζει τα δεδομένα που λείπουν λαμβάνοντας υπόψη την απόσταση μεταξύ των διανυσμάτων του δείγματος στο χώρο του συνόλου δεδομένων. Για κάθε χαρακτηριστικό που λείπει, εξετάζει τα K πλησιέστερα δείγματα που έχουν παρατηρηθεί για αυτό το χαρακτηριστικό και υπολογίζει τον μέσο όρο των τιμών τους όσον αφορά τα αριθμητικά δεδομένα. Όταν πρόκειται για κατηγορικά δεδομένα, το αποτέλεσμα είναι η πιο συχνή κλάση

[1] https://www.framinghamheartstudy.org/fhs-for-researchers/

Πίνακας 4.1: *Σύνολο δεδομένων Framingham heart study.*

Feature	Description	Type	Missing
Sex	Male, Female	Categorical	0
Totchol	Serum Total Cholesterol (mg/dL)	Numerical	409
Age	Age in ears	Numerical	0
SysBP	Systolic Blood Pressure (mmHg)	Numerical	0
Cursmoke	Current smoking at exam	Categorical	0
Cigpday	Number of cigarettes smoked each day	Numerical	79
Bmi	Serum Total Cholesterol (mg/dL)	Numerical	52
Diabetes	Is Diabetic	Categorical	0
Bpmeds	Use of Anti -hyp medication at exam	Categorical	52
Heartrate	Heart rate beats/min	Numerical	6
Glucose	Casual serum glucose (mg/dL)	Numerical	1440
Prevhyp	Prevalent hypertention	Categorical	0
Prevstrk	Prevalent Stroke	Categorical	0
DiaBP	Diastolic Blood Pressure (mmHg)	Numerical	0
CVD	Cardiovascular disease	Categorical	0

των K πλησιέστερων γειτόνων. Στην περίπτωσή μας για $K = 5$, δεδομένου ενός δείγματος $S(X, Y, 0)$ και των 5 πλησιέστερων γειτόνων του $N_5 = \{(X_i, Y_i, 1) | i = 1, 2, ..., 5\}$ ορίζουμε:

$$Y = \begin{cases} argmax_z\{\sum_{(X_i Y_i, 1) \in N_5)} 1(Y_i = z)\} & \text{Αν το } Y \text{ είναι κατηγορικό} \\ \frac{1}{5}\sum_{i=1}^{5} Y_i & \text{Αν το } Y \text{ είναι αριθμητικό} \end{cases} \tag{4.2}$$

όπου z μπορεί να είναι είτε 0 είτε 1 αφού έχουμε μόνο δυαδικές τιμές και $1(Y_i = z)$ είναι μια συνάρτηση που επιστρέφει 0 αν $(Y_i = z)$ και αλλιώς επιστρέφει 0. Η μετρική για τη μέτρηση της απόστασης μεταξύ δύο σημείων p και q είναι η ευκλείδεια:

$$d(p, q) = \sum_{i=1}^{n}(q_i - p_i)^2 \tag{4.3}$$

όπου n είναι ο αριθμός των μεταβλητών για κάθε σημείο δεδομένων.

4.3.3 Missforest

Αξιοποιούμε τη μέθοδο που πρότειναν οι Stekhoven Daniel J. και Bühlmann Peter[37]. Σε αυτό το μοντέλο, αρχικά όλες οι στήλες, εκτός από μία, αποδίδονται με τη χρήση μέσου και τρόπου απόδοσης. Κατά συνέπεια, ο αλγόριθμος τυχαία δάση χρησιμοποιείται για την πρόβλεψη των τιμών που λείπουν στη στήλη που είχε αποκλειστεί προηγουμένως. Οι προβλεπόμενες τιμές χρησιμοποιούνται στη συνέχεια στην θέση των κενών. Αυτή η διαδικασία επαναλαμβάνεται στα δεδομένα σε έναν βρόχο όπου κάθε επανάληψη βασίζεται στην προηγούμενη βελτιώνοντας τις αποδιδόμενες τιμές. Η διαδικασία συνεχίζεται έως ότου η διαφορά μεταξύ του αποδιδόμενου πίνακα M_{new}^{imp} και M_{old}^{imp} δεν αυξάνεται. Στην περίπτωσή μας χρησιμοποιούμε επίσης μέγιστες επαναλήψεις = 20 και αριθμό δέντρων = 100. Η προαναφερθείσα διαφορά για αριθμητικά χαρακτηριστικά N είναι :

$$\delta_N = \frac{\sum_{j \in N}(M_{new}^{imp} - M_{old}^{imp})^2}{\sum_{j \in N}(M_{new}^{imp})^2} \tag{4.4}$$

και για την κατηγορική Φ:

$$\delta_F = \frac{\sum_{j \in F}\sum_{i=1}^{i=n} I_{M_{new}^{imp} \neq M_{old}^{imp}}}{F_{NA}} \tag{4.5}$$

όπου F_{NA} είναι ο αριθμός των τιμών που λείπουν σε κατηγορικές μεταβλητές.

4.3.4 Το μοντέλο NAA

Χρησιμοποιούμε το μοντέλο αυτόματου κωδικοποιητή που προτάθηκε από τους Aidos και Tom[33], το οποίο ονομάζεται αυτόματος κωδικοποιητής με επίγνωση της γειτονιάς (NAA). Σε αυτή την εργασία, χρησιμοποιείται ένας υπερπλήρης DAE για το πρόβλημα του υπολογισμού ελλιπών τιμών, όπου ο προ-υπολογισμός γίνεται με KNN. Το μέρος του προ-υπολογισμού διεξάγεται για ολόκληρο το σύνολο δεδομένων πριν από την εκπαίδευση του μοντέλου χρησιμοποιώντας $k = 5$. Με τη χρήση του KNN ως μεθόδου προ-υπολογισμού είναι πιο δύσκολο να συγκλίνει σε τοπικά ελάχιστα κατά τη διάρκεια των πρώτων εποχών της εκπαίδευσης, καθώς η αρχική εκτίμηση είναι αξιοπρεπής.

4.3.5 Βελτιωμένο Μοντέλο αυτοκωδικοποιητή (I-NAA)

Το μοντέλο που αναπτύξαμε είναι ένας αυτοκωδικοποιητής αποθορυβοποίησης όπου χρησιμοποιείται ο αλγόριθμος KNN για την πρώτη αποτίμηση τιμών. Ως προς την αρχιτεκτονική, επιλέγουμε να μειώνεται ο αριθμός νευρώνων του κωδικοποιητή καθώς πηγαίνουμε από το ένα επίπεδο στο άλλο και αντίστοιχα να αυξάνεται του αυτοκωδικοποιητή. Αυτή η επιλογή έγινε καθώς ένα τέτοιο μοντέλο επιβάλλει περιορισμού στα δεδομένα είσοδο (με την μείωση της διαστατικότητας) και συνεπώς το δίκτυο μαθαίνει τις σχέσεις του διέπουν το σύνολο δεδομένων. Σχετικά με την διαδικασία εκπαίδευσης, ας θεωρήσουμε ένα αντίγραφο D_{copy} του αρχικού συνόλου δεδομένων D. Αρχικά, εισάγουμε απουσιάζουσες τιμές στο D_{copy} και μετά τις αποτιμούμε με την χρήση του αλγορίθμου KNN καταλήγοντας σε ένα σύνολο $D_{imputed}$. Έπειτα, παρτίδες του $D_{imputed}$ δίνονται ως είσοδος στο δίκτυο το οποίο πρέπει να μάθει να τις μετατρέπει στις αντίστοιχες παρτίδες του D. Με τους επιπλέον περιορισμούς που εφαρμόζονται στο δίκτυο το αποτέλεσμα της διαδικασίας αυτής είναι ότι αυτο μαθαίνει τις σχέσεις που διέπουν τα χαρακτηριστικά του συνόλου.
Παρόλα αυτά, αν για όλη την διαδικασία της εκπαίδευσης το $D_{imputed}$ παραμείνει σταθερό ο αυτοκωδικοποιητής μπορεί να μάθει να αντιγράφει τις συγκεκριμένες τιμές στις αντίστοιχες του D και όχι τις συσχετίσεις του συνόλου. Επειδή λοιπόν εμείς θέλουμε να γίνει αυτό αλλά και να λάβουμε υπόψιν την χωρική γειτονία (KNN) μια τιμής κατά την πρώτη αποτίμηση χρησιμοποιούμε τον αλγόριθμο KNN για μια πρώτη αποτίμηση αλλά αλλάζουμε κάθε N εποχές την τιμή του K σε μια που δεν έχει δωθεί στο παρελθόν. Εμπειρικά η τιμή του N επιλέγεται να είναι ίση με 10.
Με την ίδια λογική, αν οι τιμές που πρέπει να αποτιμηθούν, π.χ. X_{12} και X_{23}, παραμείνουν ίδιες για όλη την διάρκεια της εκπαίδευσης το μοντέλο ίσως μάθει να αποτιμά σωστά μόνο αυτές τις τιμές και όχι γενικά οποιαδήποτε τιμή μπορεί να λείπει. Για τον λόγο αυτό, αλλάζουμε τις τιμές που είναι κενές στην αρχή κάθε εποχής. Αυτό αποτρέπει το μοντέλο από την σύγκλιση σε κάποιο τοπικό ελάχιστο. Επίσης κατασκευάζουμε μια συνάρτηση απωλειών σχετική με την περίπτωση μας όπου για τα N κατηγορικά γνωρίσματα χρησιμοποιείται η ρίζα του μέσου τετραγωνισμένου λάθους και για τα C κατηγορικά η δυαδική διασταυρούμενη εντροπία. Πιο συγκεκριμένα για τα συνεχή:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2 \tag{4.6}$$

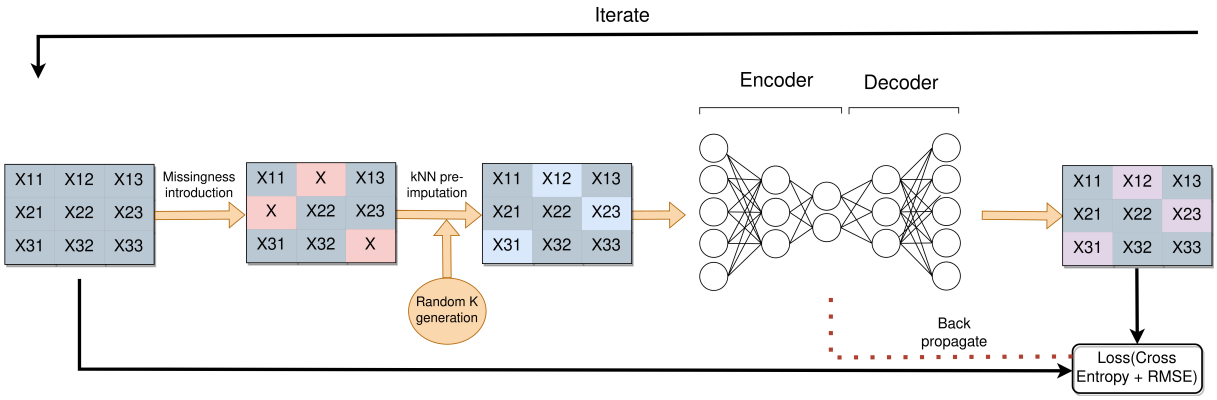
Για τα κατηγορικά:

$$BCE = -\frac{1}{N}\sum_{i=N+1}^{N+C} y_i \cdot log(p(y_i) + (1 - y_i) \cdot log(1 - p(1 - y_i)) \tag{4.7}$$

οπότε μαζί

$$Loss = RMSE + BCE \tag{4.8}$$

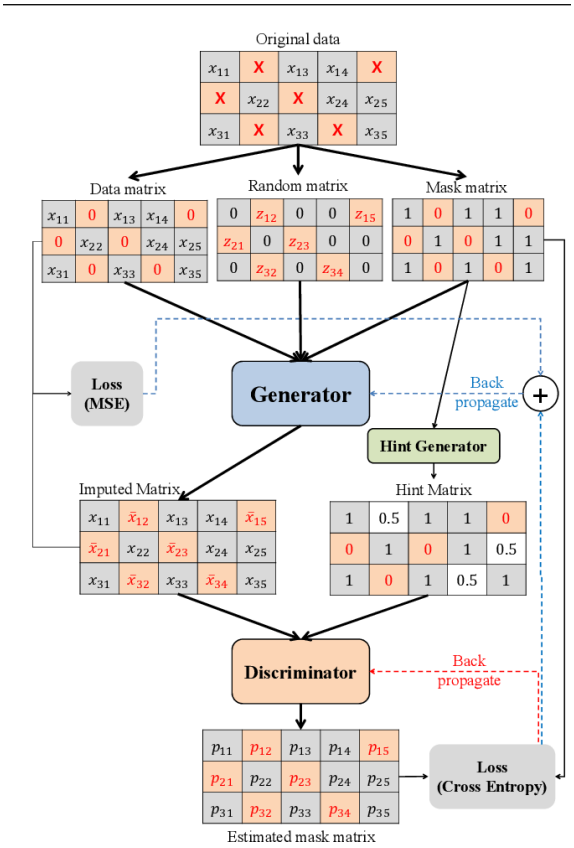
Η αρχιτεκτονική και η μεθοδολογία που περιγράφηκε φαίνονται στην εικόνα 4.1.



Σχήμα 4.1: *Μεθοδολογία και αρχιτεκτονική του αυτοκωδικοποιητή.*

4.3.6 Το μοντέλο GAIN

Το μοντέλο αυτό [38] βασίζεται στην αρχική αρχιτεκτονική των γενετικών ανταγωνιστικών δικτύων κάνοντας κάποιες αλλαγές έτσι ώστε η μέθοδος να προσαρμοστεί στο πρόβλημα των απουσιαζόντων τιμών. Γραφικά το μοντέλο αυτό φαίνεται στο σχήμα 4.2.



Σχήμα 4.2: *Σχηματικό διάγραμμα του μοντέλου Gain.*

Διαχωριστής/Γεννήτορας

Ο γεννήτορας δέχεται ως είσοδο τα $\tilde{X}, M, Z$ όπου M ο πίνακας που καθορίζει τις απουσιάζουσες και μη τιμές, Z ο πίνακας θορύβου, $\tilde{X}$ ο πίνακας του συνόλου δεδομένων με τις απουσιάζουσες τιμές. Ως έξοδο δίνει έναν πίνακα $\bar{X}$ με τις απουσιάζουσες τιμές συμπληρωμένες. Τελικά για το αποτέλεσμα του γεννήτορα έχουμε :

$$X = M \odot \tilde{X} + (1 - M) \odot \bar{X} \tag{4.9}$$

Ο διαχωριστής D εκπαιδεύεται ανταγωνιστικά σε σύγκριση με τον γεννήτορα.Η διαδικασία αυτή ωστόσο γίνεται διαφορετικά σε σχέση με τα κλασσικά γενετικά ανταγωνιστικά δίκτυα. Στην μέθοδο αυτή αντί η έξοδος του γεννήτορα να διαχωρίζεται ολόκληρη ως ψεύτικη η αληθινή, διαχωρίζεται σε μερικά μέρη τα οποία είναι αληθή και μερικά που είναι ψευδή. Ο διαχωριστής λοιπόν προσπαθεί να καταλάβει ποιες μεταβλητές είναι πραγματικές και ποιές ήταν αρχικά απουσιάζουσες και έπειτα παραχθείσες από τον γεννήτορα. Ουσιαστικά λοιπόν ως έξοδο προσπαθεί να παράγει τον πίνακα M. Έτσι, ο διαχωριστής εκπαιδεύεται με βάση :

$$min_D - \sum_{i=1}^{batchsize} L_D(M_i, \bar{M}_i, H_i) \tag{4.10}$$

Όσον αφορά τον γεννήτορα ορίζουμε:

$$L_G(M, \bar{M}, H) = - \sum_{i:H_i=0} (1 - M_i) \cdot log(M_i) \tag{4.11}$$

η οποία διαισθητικά είναι μια τιμή που μετράει πόσο συχνά ο γεννήτορας ξεγελάει τον διαχωριστή. Δεύτερον, ορίζουμε:

$$L_M(\tilde{X}, \bar{X}) = - \sum_{i=1}^{d} M_i \cdot Diff(\tilde{X}_i, \bar{X}_i) \tag{4.12}$$

όπου d είναι το μέγεθος της διάστασης του συνόλου δεδομένων και :

$$Diff(\tilde{X}, \bar{X}) \begin{cases} (\tilde{X}_i - \bar{X}_i)^2 & \text{αν } X_i \text{ είναι αριθμητικό} \\ -X_i log(\bar{X}_i) & \text{αν } X_i \text{ είναι δυαδικό} \end{cases} \tag{4.13}$$

Η εξίσωση 4.12 από την άλλη πλευρά μετρά πόσο ακριβής είναι ο γεννήτορας στην αναδημιουργία των παρατηρούμενων συνιστωσών της εισόδου. Τέλος, ο γεννήτορας εκπαιδεύεται ώστε να ελαχιστοποιεί την εξίσωση που ορίζεται παρακάτω:

$$min_G \sum_{i=1}^{batchsize} L_G(M_i, \bar{M}_i, H_i) + aL_{M_i}(\tilde{X}_i, \bar{X}_i) \tag{4.14}$$

όπου a είναι μια παράμετρος κλιμάκωσης.

Μηχανισμός hint

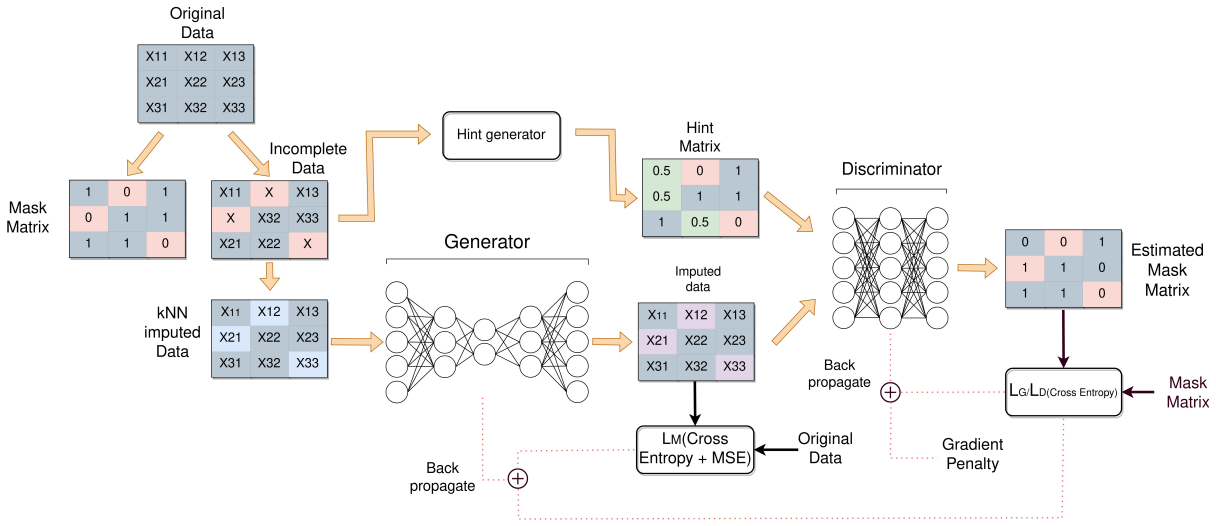
Στην αρχική δημοσίευση μαζί με την έξοδο του γεννήτορα ο διαχωριστής δέχεται και κάποια επιπλέον πληροφορία σχετικά με τον πίνακα μάσκας που πρέπει να βρεί. Ο πίνακας αυτός ουσιαστικά φανερώνει κάθε φορά ένα ποσοστό από τον πίνακα μάσκας αφήνοντάς το υπόλοιπο ως άγνωστο για να προβλεφθεί από τον διαχωριστή. Για παράδειγμα, αν το ηιντ έχει ποσοστό 90%, τότε το ποσοστό αυτό του πίνακα μάσκας είναι γνωστό και το υπόλοιπον 10% πρέπει να προβλεφθεί από τον διαχωριστή. Όσο μεγαλύτερο είναι λοιπόν τόσο πιο εύκολο είναι για τον διαχωριστή να προβλέψει σωστά τον πίνακα M. Στην δημοσίευση αποδεικνύεται θεωρητικά και πρακτικά ότι ο μηχανισμός αυτός είναι απαραίτητος προκειμένου να εκπαιδευτεί σωστά το δίκτυο.

4.3.7 Βελτιωμένο γενετικό μοντέλο (I-GAIN)

Βασισμένοι στο προαναφερθέν μοντέλο GAIN υλοποιούμε κάποιες βελτιώσεις οι οποίες αποδίδουν καλύτερα αποτελέσματα. Αρχικά χρησιμοποιούμε ομαλοποίηση παρτίδας πρίν από κάθε επίπεδο νευρώνων[39] και στο δίκτυο του γεννήτορα αλλά και στο δίκτυο του διαχωριστή. Ακόμα όπως και με την περίπτωση του αυτοκωδικοποιητή, χρησιμοποιούμε τον αλγόριθμο KNN για μια πρώτη αποτίμηση προς αποφυγή της σύγκλισης σε τοπικά ελάχιστα. Οι μεθοδολογίες εκπαίδευσης που εισήχθηκαν στον παράγραφο του αυτοκωδικοποιητή εφαρμόζονται και εδώ. Χρησιμοποιείται επίσης η ίδια συνάρτηση απωλειών για τον γεννήτορα. Επίσης, οι συγγραφείς του GAIN προτείνουν μια απλή αρχιτεκτονική 3 επιπέδων για τον γεννήτορα. Αντίθετα, εμείς χρησιμοποιούμε μια πιο σύνθετη με 5 επίπεδα τα οποία έχουν μορφή αυτοκωδικοποιητή καθώς αυτός αποδείχθηκε ικανός να μάθει καλά την κατανομή ενός συνόλου δεδομένων. Μια τελευταία προσθήκη που κάναμε είναι η εισαγωγή παραγωγικού πέναλτι στην συνάρτηση απωλειών του διαχωριστή για πιο ομαλή εκπαίδευση όπως αυτό ορίζεται στην [40]:

$$GP = \frac{a}{batchsize} \sum_{i=1}^{batchsize} (\|\nabla \bar{X}_i \cdot D(\bar{X}_i)\|_2 - 1)^2 \tag{4.15}$$

όπου το a είναι παράμετρος κλιμάκωσης. Η προσέγγιση αυτή φαίνεται στην εικόνα . 4.3.



Σχήμα 4.3: *Μεθοδολογία και αρχιτεκτονική του βελτιωμένου GAIN.*

Κεφάλαιο 5

Αποτελέσματα

Στο κεφάλαιο αυτό περιγράφουμε τα αποτελέσματα για τα μοντέλα και την μεθοδολογία που προαναφέρθηκε. Για κάθε μοντέλο θα εξετάσουμε τις περιπτώσεις της αποτίμησης απουσιαζόντων τιμών και της μετά-αποτίμησης πρόβλεψη. Η δεύτερη περίπτωση είναι ουσιαστικά η πρόβλεψη με κάποιον αλγόριθμο ταξινόμησής για κάθε ολοκληρωμένο σύνολο δεδομένων που προκύπτει από κάθε μέθοδο αποτίμησης απουσιαζόντων τιμών.

5.1 Αποτίμηση απουσιαζόντων τιμών

5.1.1 Τρόπος διεξαγωγής πειραμάτων

Για να αξιολογήσουμε τις μεθόδους μας, πρώτα επιλέγουμε ένα υποσύνολο από κάθε αρχικό σύνολο δεδομένων που δεν έχει κενές τιμές. Στη συνέχεια, εισάγουμε τυχαία κενά σε αυτά τα σύνολα δεδομένων και συγκρίνουμε τις εκτιμώμενες τιμές με τις πραγματικές. Αυτό το βήμα εκτελείται χρησιμοποιώντας τη βιβλιοθήκη pyampute[41] το οποίο είναι μία γλώσσα python της συλλογής R ampute[42]. Εντός αυτής, υλοποιείται μία πολυμεταβλητή διαδικασία αποτίμησης που επιτρέπει την εισαγωγή διαφορετικών μηχανισμών κενών ατομικά ή ομαδοποιημένα. Χρησιμοποιώντας αυτήν την βιβλιοθήκη, εισάγουμε κενά ποσοστών 10%, 20%, 30%, 40% και 50% για τον μηχανισμό απουσιαζόντων τιμών MCAR. Για να αξιολογήσουμε αυτά τα μοντέλα, χρησιμοποιούμε διαίρεση 5-πλευρών διασταυρώσεων και βγάζουμε τον μέσο όρο για τα αποτελέσματα μεταξύ των 5 πλευρών που κρατάμε. Η παραπάνω διαδικασία επαναλαμβάνεται 10 φορές και βρίσκουμε και πάλι τον μέσο όρο των αποτελεσμάτων. Συνολικά, εκπαιδεύουμε και αξιολογούμε κάθε μοντέλο $5 \times 10 = 50$ φορές.

5.1.2 Μετρικές

Στην παράγραφο αυτή αναλύουμε τις μετρικές που θα χρησιμοποιήσουμε για την αξιολόγηση των μεθόδων που αναφέρθηκαν τόσο για την αποτίμηση απουσιαζόντων τιμών όσο και για την πρόβλεψη μετά. Επειδή το σύνολο δεδομένων μας περιέχει συνεχή και κατηγορικά δεδομένα είναι αναγκαία η χρήση 2 μετρικών αντίστοιχα. Για λόγους που αναφέρονται στην συνέχεια επιλέγονται η κανονικοποιημένη ρίζα του τετραγωνισμένου λάθους και η περιοχή κάτω από την χαρακτηριστική καμπύλη για την μελέτη της επίδοσης των μεθόδων ως προς την αποτίμηση των συνεχών και των κατηγορικών γνωρισμάτων αντίστοιχα. Τονίζεται ξανά, ότι τρέχουμε πολλές φορές το ίδιο πείραμα για κάθε μοντέλο και η μετρική που

παρουσιάζεται τελικά είναι ο μέσος όρος αυτής από τα διαφορετικά τρεξίματα.

Κανονικοποιημένη ρίζα του τετραγωνισμένου λάθους (NRMSE)

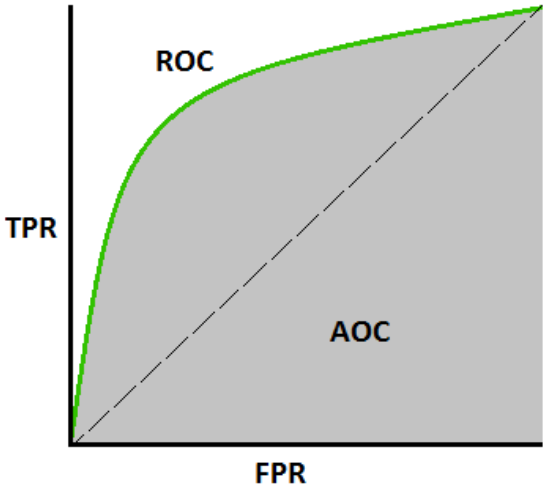
Αυτή η μετρική χρησιμοποιείται στο στάδιο της αποτίμησης για τις συνεχείς μεταβλητές. Χρησιμοποιούμε την κανονικοποιημένη έκδοση του μέσου τετραγωνισμένου λάθους καθώς θέλουμε για κάθε ποσοστό απουσιαζόντων τιμών να βγάλουμε έναν μέσο όρο για όλα τα συνεχή γνωρίσματα. Σε περίπτωση που αυτά λοιπόν δεν ήταν κανονικοποιημένα τα γνωρίσματα με μεγαλύτερες τιμές θα υπερίσχυαν. Έχουμε λοιπόν:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{Real_i - Imputed_i}{\sigma_i}\right)^2} \tag{5.1}$$

όπου όμως τα οι τιμές των Real και Imputed είναι κανονικοποιημένες στο $[0, 1]$.

Περιοχή κάτω απο την χαρακτηριστική καμπύλη (AUROC)

Η μετρική αυτή πρόκειται για το εμβαδό της περιοχής κάτω από την καμπύλη που δημιουργείται όταν ορίσουμε μια γραφική με άξονες τον πραγματικό ρυθμό τον προβλεφθέντων πραγματικών και τον πραγματικό ρυθμό των προβλεφθέντων αρνητικών για το πρόβλημα της δυαδικής ταξινόμησης. Ακόμα, η μετρική αυτή αξιοποιεί τις πιθανότητες πρόβλεψης και δεν χρειάζεται να δέχεται στρογγυλοποιημένα αποτελέσματα. Γραφικά η μετρική αυτή φαίνεται στην εικόνα 5.1.



Σχήμα 5.1: *Περιοχή κάτω απο τη υχαρακτηριστική καμπύλη AUROC.*

5.1.3 Αποτελέσματα

Τα αποτελέσματα των προτεινόμενων μας μεθόδων που αναφέρονται στο Τμήμα 4.3 εμφανίζονται στο Σχήμα 5.2 για τα αριθμητικά χαρακτηριστικά και στο Σχήμα 5.3 για τα

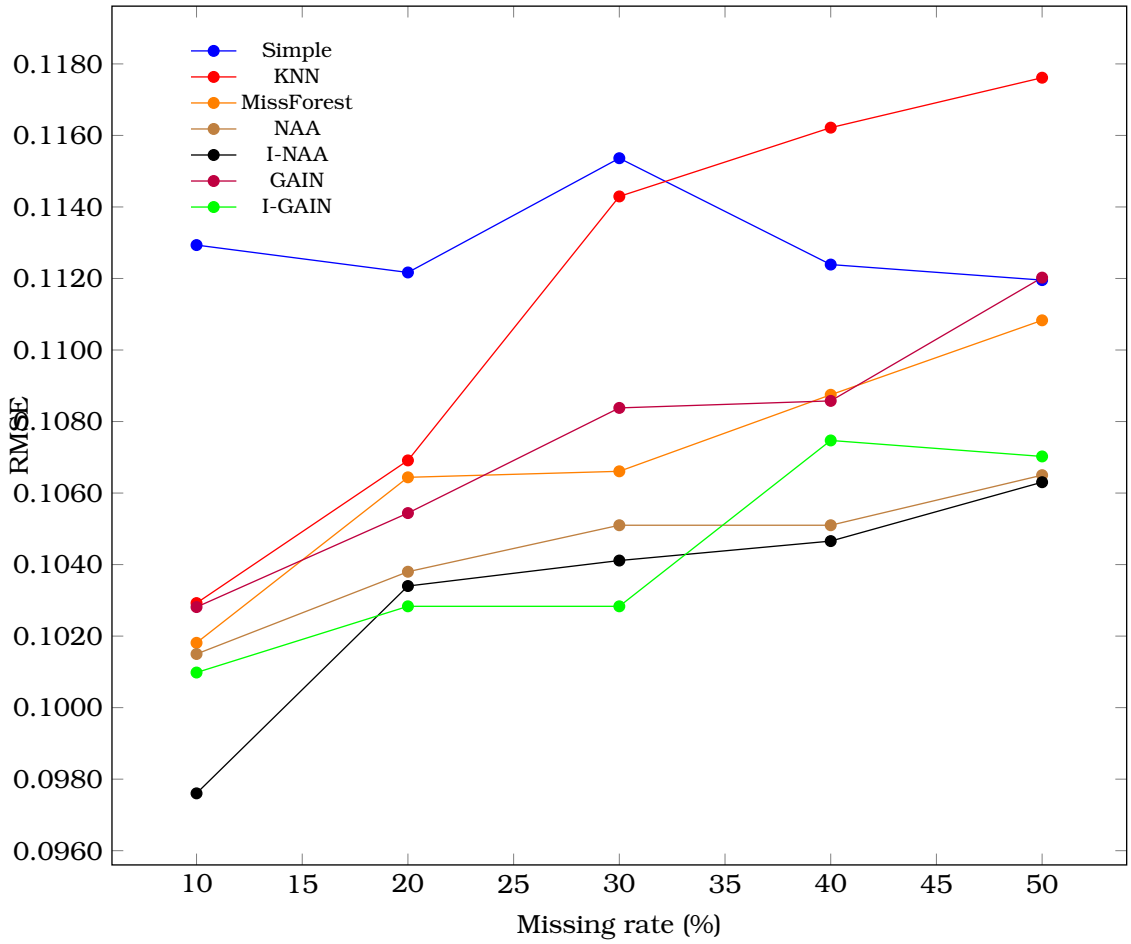
κατηγορικά χαρακτηριστικά.
Όσον αφορά τα αριθμητικά χαρακτηριστικά (Σχ. 5.2), παρατηρούμε ότι οι μέθοδοι που παρουσιάσαμε, δηλαδή ο I-NAA και I-GAIN, είναι οι προσεγγίσεις με τις καλύτερες επιδόσεις, επιτυγχάνοντας τις χαμηλότερες βαθμολογίες RMSE για όλα τα ποσοστά απουσίας. Πιο συγκεκριμένα, ο I-NAA αποδίδει καλύτερα από τον NAA για όλα τα ποσοστά έλλειψης, με τη μεγαλύτερη διαφορά να παρατηρείται στο ποσοστό έλλειψης 10% που αντιστοιχεί σε 0,04. Σημειώνεται επίσης ότι το I-GAIN βελτιώνει τις επιδόσεις του GAIN όσον αφορά όλα τα ποσοστά έλλειψης επιτυγχάνοντας έως και 0,06 χαμηλότερο RMSE για ποσοστό έλλειψης 30%. Βλέπουμε επίσης ότι οι προτεινόμενες προσεγγίσεις βαθιάς μάθησης υπερτερούν έναντι των τυπικών προσεγγίσεων, δηλαδή των Simple και KNN, για όλα τα ποσοστά έλλειψης. Παρατηρούμε επιπλέον ότι ο MissForest επιτυγχάνει καλύτερη απόδοση από τις μεθόδους Simple και KNN, ενώ επιτυγχάνει χαμηλότερη απόδοση από τις μεθόδους I-NAA, I-GAIN και NAA. Τέλος, η επιδείνωση της απόδοσης με την αύξηση των ποσοστών έλλειψης είναι χαμηλότερη για τις προσεγγίσεις βαθιάς μάθησης, δηλαδή 0,0087 και 0,015 RMSE για τους αλγορίθμους I-NAA και KNN αντίστοιχα.
Όσον αφορά τα κατηγορικά χαρακτηριστικά (Σχ. 5.3), παρατηρούμε ότι τα μοντέλα που εισήγαμε, δηλαδή τα I-NAA και I-GAIN, αποδίδουν τα καλύτερα αποτελέσματα AUROC. Συγκεκριμένα, το I-NAA υπερτερεί έναντι του NAA κατά 1,57-4,50%, ενώ το I-GAIN βελτιώνει την απόδοση του GAIN κατά 1,30-6,00%. Παρατηρούμε επίσης ότι η επιδείνωση αυτών των μεθόδων με την αύξηση του υψηλού ποσοστού είναι αργή και οι επιδόσεις είναι αποδεκτές ακόμη και για 50% ελλείψεις. Παρατηρούμε επίσης ότι ο MissForest αποδίδει καλύτερα από την GAIN στα περισσότερα ποσοστά έλλειψης, αλλά αποδίδει χειρότερα από όλες τις άλλες προσεγγίσεις βαθιάς μάθησης. Επιπλέον, παρατηρούμε ότι οι τυπικές προσεγγίσεις, δηλαδή η Simple και KNN, έχουν κακές επιδόσεις. Για να είμαστε πιο ακριβείς, η απόδοση του KNN κυμαίνεται από 55,50% έως 73,50%, ενώ το μοντέλο Simple αποδίδει σταθερό AUROC 50,00%, καθώς προβλέπει μόνο μία κλάση. Παρατηρούμε επίσης ότι η απόδοση που επιτυγχάνεται από το KNN παρουσιάζει πτώση καθώς αυξάνεται το ποσοστό απουσίας η οποία είναι η μεγαλύτερη συγκριτικά με τις υπόλοιπες μεθόδους.

5.2 Πρόβλεψη μετά την αποτίμηση

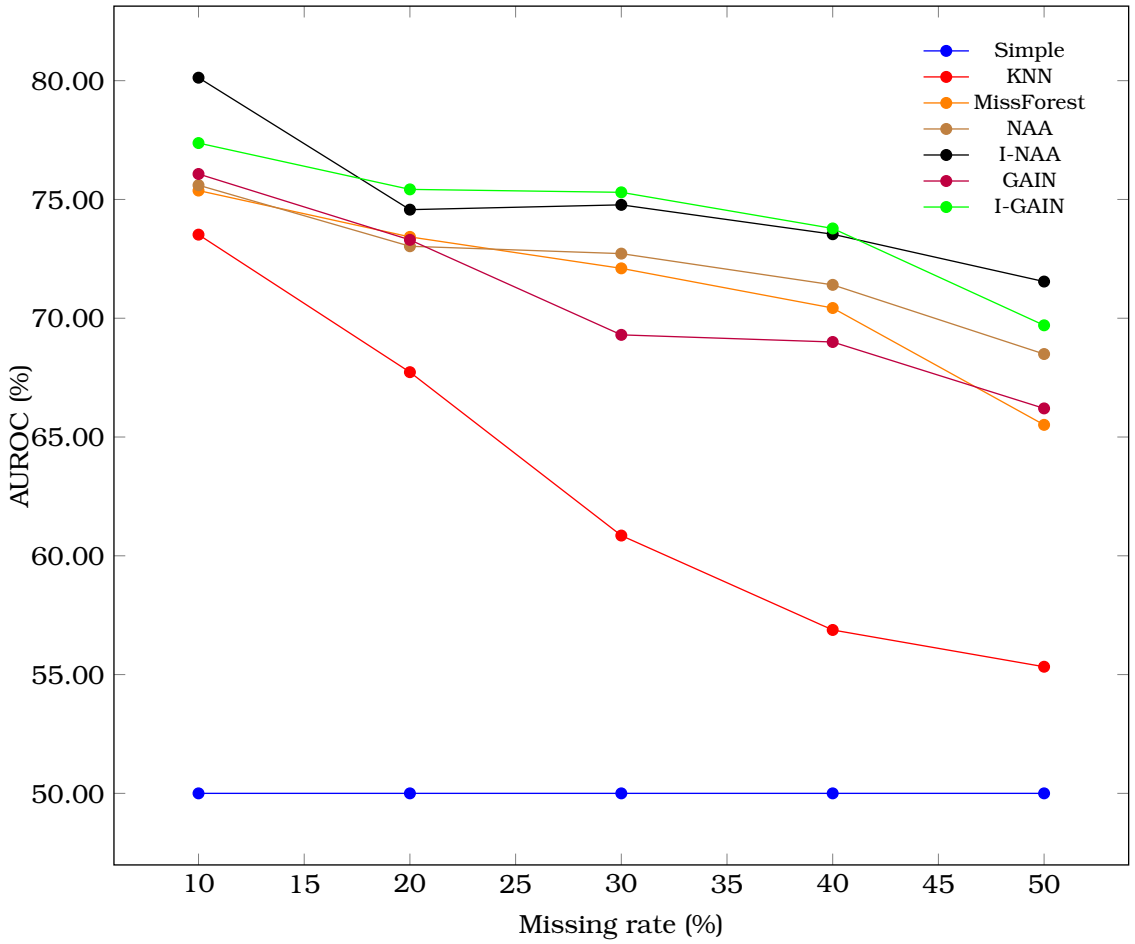
Στην παράγραφό αυτή αναλύουμε τα αποτελέσματα της πρόβλεψης μετά την αποτίμηση. Για κάθε μοντέλο αποτίμησης έχουμε κρατήσει το σύνολο δεδομένων με τις κενές τιμές συμπληρωμένες και ελέγχουμε την απόδοση ενός προβλεπτικού μοντέλου.

5.2.1 Τρόπος διεξαγωγής πειραμάτων

Για να αξιολογήσουμε περαιτέρω τα μοντέλα μας, διεξάγουμε πρόβλεψη μετά την αποτίμηση για να δούμε αν οι καλύτερες βαθμολογίες AUROC και RMSE οδηγούν επίσης σε καλύτερες επιδόσεις για το βήμα της πρόβλεψης. Για να το κάνουμε αυτό για κάθε μοντέλο, αποθηκεύουμε το πλήρες, υπολογισμένο σύνολο δεδομένων και εκτελούμε ένα απλό Random Forest για την πρόβλεψη της μεταβλητής CVD χρησιμοποιώντας 5-πτυχή διασταυρωμένη επικύρωση και SMOTE[42] για το σύνολο εκπαίδευσης κάθε πτυχής (δεδομένου ότι



Σχήμα 5.2: *Μέση τιμή της ρίζας του τετραγωνισμένου λάθους για διαφορετικά ποσοστά κενών τιμών.*



Σχήμα 5.3: *Περιοχή κάτω απο την χαρακτηριστική καμπύλη για διαφορετικά ποσοστά κενών τιμών.*

το σύνολο δεδομένων είναι εξαιρετικά μη ισορροπημένο).

5.2.2 Μετρική

Για το στάδιο αυτό χρησιμοποιούμε την γνωστή μετρική Σκορ-F1. Ο λόγος που επιλέγεται αυτή η μετρική είναι διότι η μεταβλητή προς πρόβλεψη είναι εξαιρετικά μη ισορροπημένη και η απλή ακρίβεια δίνει λάθος αποτελέσματα. Το Σκορ-F1 συνδυάζει την ακρίβεια και την ανάκληση ενός ταξινομητή σε μια ενιαία μετρική, λαμβάνοντας τον αρμονικό μέσο όρο τους. Χρησιμοποιείται κυρίως για τη σύγκριση της απόδοσης δύο ταξινομητών. Ας υποθέσουμε ότι ο ταξινομητής A έχει υψηλότερη ανάκληση και ο ταξινομητής B έχει υψηλότερη ακρίβεια. Σε αυτή την περίπτωση, τα αποτελέσματα F1 και για τους δύο ταξινομητές μπορούν να χρησιμοποιηθούν για να προσδιοριστεί ποιος από τους δύο παράγει καλύτερα αποτελέσματα.

Το Σκορ-F1 ενός μοντέλου ταξινόμησης υπολογίζεται ως εξής:

$$\frac{2(P \cdot R)}{P + R} \tag{5.2}$$

όπου P = η ακρίβεια, R = η ανάκληση του μοντέλου ταξινόμησης.

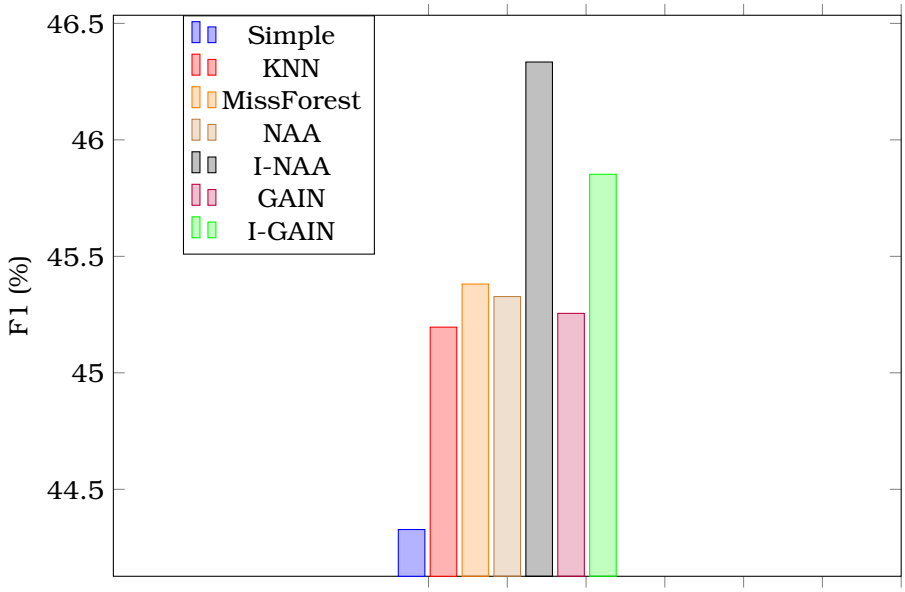
5.2.3 Αποτελέσματα

Για κάθε σύνολο δεδομένων που παράγεται από μια μέθοδο υπολογισμού ελλιπών τιμών, εκτελούμε πρόβλεψη μετά τον υπολογισμό με τα αποτελέσματα που απεικονίζονται στο Σχήμα 5.4.

Παρατηρούμε ότι οι I-NAA και I-GAIN παράγουν τα καλύτερα αποτελέσματα F1-Σκορ μετά την εισαγωγή. Πιο συγκεκριμένα, το I-NAA είναι το μοντέλο με τις καλύτερες επιδόσεις, ξεπερνώντας τις άλλες προσεγγίσεις κατά 0,48-2,43%. Παρατηρούμε επίσης ότι το I-GAIN υπερτερεί του GAIN και στην εργασία μετά τον υπολογισμό κατά 0,60%. Επιπλέον, το Miss-Forest παράγει το τρίτο καλύτερο F1-Σκορ πρόβλεψης με 45,38%. Ακόμα, παρατηρούμε ότι η προσέγγιση Simple έχει τη χειρότερη επίδοση, επιτυγχάνοντας βαθμολογία F1 44,10%. Από την άλλη ο αλγόριθμος KNN παράγει λίγο καλύτερο αποτέλεσμα σε σχέση με την μέθοδο Simple το οποίο είναι όμως και πάλι χαμηλά. Γενικά παρατηρείται συμπεριφορά ανάλογη με αυτή στο κομμάτι της απόδοσης κενών τιμών. Αυτό είναι λογικό και αναμενόμενο καθώς ένα πιο πλήρες σύνολο δεδομένων το οποίο περιέχει πιο ρεαλιστικές τιμές μπορεί να χρησιμοποιηθεί καλύτερα από έναν ταξινομητή για να γίνει μαι πιο σωστή πρόβλεψη.

5.3 Συνολικός σχολιασμός

Συνολικά, μπορούμε να δούμε ότι οι μέθοδοι συμπεριφέρονται με παρόμοιο τρόπο και για την αποτίμηση απουσιαζόντων τιμών άλλα και για την πρόβλεψη μετά την αποτίμηση. Πιο



Σχήμα 5.4: *Σκορ-F1 για τα διάφορα σύνολα δεδομένων που προέκυψαν απο την αποτίμηση απουσιαζόντων τιμών.*

συγκεκριμένα, βλέπουμε ότι και για τα δυο πειράματα οι μέθοδοι Simple και KNN έχουν τα χειρότερα αποτελέσματα το οποίο είναι και αναμενόμενο καθώς είναι απλές και δεν μπορούν να χρησιμοποιήσουν σωστά τις σχέσεις μεταξύ των γνωρισμάτων του συνόλου δεδομένων. Από την άλλη, η μέθοδος Missforest έχει αξιόλογα αποτελέσματά και για τα δυο πειράματα το οποίο δικαιολογεί και την ευρέα χρήση της από την επιστημονική κοινότητα για το πρόβλημα αυτό. Ωστόσο ακόμα και αν η προαναφερθέν μέθοδος αποφέρει καλά αποτελέσματα, είναι φανερό ότι οι τεχνικές που βασίζονται στην μηχανική μάθηση είναι πιο αποτελεσματικές δίνοντας καλύτερα αποτελέσματα και για τα δυο πειράματα. Αν και όλες οι μέθοδοι βαθιάς μάθησης πετυχαίνουν καλύτερα αποτελέσματα, οι βελτιώσεις που υλοποιήθηκαν στα πλαίσια της διπλωματικής είναι πιο ακριβής και για τις δυο περιπτώσεις το οποίο επιβεβαιώνει ότι οι αρχιτεκτονικές προσθήκες αλλά και οι προσθήκες που αφορούν την διαδικασία εκπαίδευσης είναι χρήσιμες. Μπορούμε συνολικά να πούμε ότι το στάδιο της αποτίμησης απουσιαζόντων τιμών είναι σημαντικό και η χρήση πιο σύγχρονων και αποτελεσματικών μεθόδων αποφέρει καλύτερα αποτελέσματα με ένα σύνολο δεδομένων το οποίο ανταποκρίνεται καλύτερα στα πραγματικά δεδομένα.

Μέρος III

Επίλογος

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Στην παρούσα εργασία, μελετήσαμε την περίπτωση του υπολογισμού ελλιπών τιμών σε ένα σύνολο δεδομένων για καρδιαγγειακά νοσήματα. Βασιστήκαμε στις υπάρχουσες αρχιτεκτονικές βαθιάς μάθησης εισάγοντας βελτιώσεις που ταιριάζουν στη συγκεκριμένη περίπτωση. Τις αξιολογήσαμε τόσο για την απόδοση κενών τιμών όσο και για τις επιδόσεις μετά την απόδοση. Όσον αφορά την διαδικασία του υπολογισμού ελλιπών τιμών, πειραματιστήκαμε με διάφορα ποσοστά ελλιπών τιμών και δείξαμε ότι οι προτεινόμενες προσεγγίσεις υπερτερούν έναντι των σύγχρονων προσεγγίσεων, επιτυγχάνοντας κανονικοποιημένες βαθμολογίες RMSE και AUROC έως και 0,095, 82,00% αντίστοιχα. Ακόμα, είδαμε ότι όσο το ποσοστό των ελλιπών τιμών γίνεται μεγαλύτερο η διαφορά αυτή είναι όλο και πιο έντονη. Αυτό είναι λογικό καθώς όσο αυξάνεται το ποσοστό αυτό αυξάνεται και η δυσκολία της απόδοσης τιμών. Για την εργασία πρόβλεψης (πρόβλεψη μετά τον υπολογισμό), τα ευρήματα έδειξαν ότι οι εισαγόμενες προσεγγίσεις μας πέτυχαν τα καλύτερα αποτελέσματα F1 για τον υπολογισμό με διαφορά 2,50% σε σύγκριση με άλλες προσεγγίσεις. Αυτό αποδεικνύει ότι η αξιοποίηση πιο σύνθετων και δυνατών μεθόδων για την αντιμετώπιση των κενών τιμών είναι σημαντική γιατί οδηγεί σε καλύτερα προβλεπτικά μοντέλα που είναι και ο πιο σημαντικός στόχος.

6.2 Μελλοντικές Επεκτάσεις

Στο μέλλον, σκοπεύουμε να αξιολογήσουμε τις μεθόδους μας σε περισσότερα σύνολα ιατρικών δεδομένων. Επιπλέον, δεδομένου ότι αυτές οι μέθοδοι εκπαιδεύονται και δοκιμάζονται στο ίδιο σύνολο δεδομένων, έχουν ουσιαστικά προσαρμοστεί σε μια συγκεκριμένη κατανομή ασθενών. Για το λόγο αυτό, ένα μοντέλο που εκπαιδεύτηκε στο Framingham θα μπορούσε να χρησιμοποιηθεί για τον υπολογισμό ελλιπών τιμών σε διαφορετικά δεδομένα καρδιακών παθήσεων για την αξιολόγηση της απόδοσης μεταξύ των συνόλων δεδομένων. Επιπλέον, θα πρέπει να γίνει και μαι επιπλέον μελέτη για τους υπόλοιπους μηχανισμούς ελλιπών τιμών ώστε να αξιολογηθούν οι προτεινόμενες μέθοδοι σε μεγαλύτερο βάθος και σε ένα ευρύτερο φάσμα σεναρίων.

Βιβλιογραφία

[1] Häyrinen, Kristiina and Saranto, Kaija and Nykänen, Pirkko. *Definition, Structure, Content, Use and Impacts of Electronic Health Records: A Review of the Research Literature*. *International journal of medical informatics*, 77:291–304, 2008.

[2] Cowie, Martin and Blomster, Juuso and Curtis, Lesley and Duclaux, Sylvie and Ford, Ian and Fritz, Fleur and Goldman, Samantha and Janmohamed, Salim and Kreuzer, Jörg and Leenay, Mark and Michel, Alexander and Ong, Seleen and Pell, Jill and Southworth, Mary and Stough, Wendy and Thoenes, Martin and Zannad, Faiez and Zalewski, Andrew. *Electronic health records to facilitate clinical research*. *Clinical research in cardiology : official journal of the German Cardiac Society*, 106, 2017.

[3] Kennedy, Edward and Wiitala, Wyndy and Hayward, Rodney and Sussman, Jeremy. *Improved Cardiovascular Risk Prediction Using Nonparametric Regression and Electronic Health Record Data*. *Medical care*, 51, 2012.

[4] Yilong Zhang and Zachary Zimmer and Lei Xu and Raymond L. H. Lam and Susan Huyck and Gregory Golm. *Missing Data Imputation With Baseline Information in Longitudinal Clinical Trials*. *Statistics in Biopharmaceutical Research*, 14(2):242–248, 2022.

[5] Wells, Brian J and Chagin, Kevin M and Nowacki, Amy S and Kattan, Michael W. *Strategies for handling missing data in electronic health record derived data*. *EGEMS (Washington, DC)*, 1(3):1035, 2013.

[6] Rubin DB. *Multiple imputation for nonresponse in survey*. John Wiley & Sons, 2004.

[7] SWJ Nijman and AM Leeuwenberg and I Beekers and I Verkouter and JJL Jacobs and ML Bots and FW Asselbergs and KGM Moons and TPA Debray. *Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review*. *Journal of Clinical Epidemiology*, 142:218–229, 2022.

[8] Duy Le, Tan and Beuran, Razvan and Tan, Yasuo. *Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare*. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018.

[9] Ayilara, Olawale and Zhang, Lisa and Sajobi, Tolu and Sawatzky, Richard and Bohm, Eric and Lix, Lisa. *Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry*. *Health and Quality of Life Outcomes*, 17, 2019.

[10] Stekhoven, Daniel J. and Bühlmann, Peter. *MissForest—non-parametric missing value imputation for mixed-type data*. *Bioinformatics*, 28(1):112–118, 2011.

[11] Azur, Melissa and Stuart, Elizabeth and Frangakis, Constantine and Leaf, Philip. *Multiple Imputation by Chained Equations: What is it and how does it work?* *International journal of methods in psychiatric research*, 20:40–9, 2011.

[12] Gondara, Lovedeep and Wang, Ke. *MIDA: Multiple Imputation Using Denoising Autoencoders* *Advances in Knowledge Discovery and Data Mining*, Cham, 2018. Springer International Publishing.

[13] Kogan, Emily and Twyman, Kathryn and Heap, Jesse and Milentijevic, Dejan and Lin, Jennifer and Alberts, Mark. *Assessing stroke severity using electronic health record data: A machine learning approach*. *BMC Medical Informatics and Decision Making*, 20, 2020.

[14] Kotsiantis, S. B. *Supervised Machine Learning: A Review of Classification Techniques*. *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*, NLD, 2007. IOS Press.

[15] Wu, Xindong and Kumar, Vipin and Quinlan, Ross and Ghosh, Joydeep and Yang, Qiang and Motoda, Hiroshi and Mclachlan, G. and Ng, Shu Kay Angus and Liu, Bing and Yu, Philip and Zhou, Zhi-Hua and Steinbach, Michael and Hand, David and Steinberg, Dan. *Top 10 algorithms in data mining*. *Knowledge and Information Systems*, 14, 2007.

[16] Rokach, Lior and Maimon, Oded. *Data mining with decision trees. Theory and applications*, τόμος 69, 2008.

[17] Graham, John W. *Missing Data Analysis: Making It Work in the Real World*. *Annual Review of Psychology*, 60(1):549–576, 2009.

[18] Mohan, Senthilkumar and Thirumalai, Chandrasegar and Srivastava, Gautam. *Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques*. *IEEE Access*, 7:81542–81554, 2019.

[19] Bashir, Saba and Khan, Zain Sikander and Khan, Farhan Hassan and Anjum, Aitzaz and Bashir, Khurram. *Improving heart disease prediction using feature selection approaches*. *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2019.

[20] Safial Islam Ayon and Md. Milon Islam and Md. Rahat Hossain. *Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques*. *IETE Journal of Research*, 0(0):1–20, 2020.

[21] Jegan, Chitra. *Heart Attack Prediction System Using Fuzzy C Means Classifier*. *IOSR Journal of Computer Engineering*, 14:23–31, 2013.

[22] Young, William and Weckman, Gary and Holland, William. *A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits*. *Theoretical Issues in Ergonomics Science*, 12:15–43, 2011.

[23] Mir, Adil Aslam and çelebi, Fatih Vehbi and Rafique, Muhammad and Hussain, Lal and Almasoud, Ahmed S. and Alajmi, Masoud and Al-Wesabi, Fahd N. and Hilal, Anwer Mustafa. *An Improved Imputation Method for Accurate Prediction of Imputed Dataset Based Radon Time Series*. *IEEE Access*, 10:20590–20601, 2022.

[24] Gondara, Lovedeep and Wang, Ke. *MIDA: Multiple Imputation Using Denoising Autoencoders* *Advances in Knowledge Discovery and Data Mining*, Cham, 2018. Springer International Publishing.

[25] Guo, Aixia and Foraker, Randi E. and MacGregor, Robert M. and Masood, Faraz M. and Cupps, Brian P. and Pasque, Michael K. *The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation*. *Frontiers in Digital Health*, 2, 2020.

[26] Brett K. Beaulieu-Jones and Jason H. Moore and et al. *Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders*. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22:207–218, 2017.

[27] Park, Sungkyu and Li, Cheng-Te and Han, Sungwon and Hsu, Cheng and Lee, Sang Won and Cha, Meeyoung. *Learning Sleep Quality from Daily Logs*. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, New York, NY, USA, 2019. Association for Computing Machinery.

[28] Shichao Zhang. *Nearest neighbor selection for iteratively kNN imputation*. *Journal of Systems and Software*, 85(11):2541–2552, 2012.

[29] Anil Jadhav and Dhanya Pramod and Krishnan Ramanathan. *Comparison of Performance of Data Imputation Methods for Numeric Dataset*. *Applied Artificial Intelligence*, 33(10):913–933, 2019.

[30] Buuren, Stef and Groothuis-Oudshoorn, Catharina. *MICE: Multivariate Imputation by Chained Equations in R*. *Journal of Statistical Software*, 45, 2011.

[31] Taunk, Kashvi and De, Sanjukta and Verma, Srishti and Swetapadma, Aleena. *A Brief Review of Nearest Neighbor Algorithm for Learning and Classification*, 2019.

[32] Psychogyios, Konstantinos and Ilias, Loukas and Askounis, Dimitris. *Comparison of Missing Data Imputation Methods using the Framingham Heart study dataset*. *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2022.

[33] Aidos, Helena and Tomás, Pedro. *Neighborhood-aware autoencoder for missing value imputation*, 2021.

[34] Boseong Seo and Jaekyung Shin and Taejin Kim and Byeng D. Youn. *Missing data imputation using an iterative denoising autoencoder (IDAE) for dissolved gas analysis*. *Electric Power Systems Research*, 212:108642, 2022.

[35] Park, Sungkyu and Li, Cheng-Te and Han, Sungwon and Hsu, Cheng and Lee, Sang Won and Cha, Meeyoung. *Learning Sleep Quality from Daily Logs*. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, New York, NY, USA, 2019. Association for Computing Machinery.

[36] Dong, Weinan and Fong, Daniel and Yoon, Jin-sun and Wan, Eric and Bedford, Laura and Tang, Eric and Lam, Cindy. *Generative adversarial networks for imputing missing data for big data clinical research*. *BMC Medical Research Methodology*, 21, 2021.

[37] Stekhoven, Daniel J. and Bühlmann, Peter. *MissForest—non-parametric missing value imputation for mixed-type data*. *Bioinformatics*, 28(1):112–118, 2011.

[38] Yoon, Jinsung and Jordon, James and van der Schaar, Mihaela. *GAIN: Missing Data Imputation using Generative Adversarial Nets* *Proceedings of the 35th International Conference on Machine Learning*, τόμος 80 στο *Proceedings of Machine Learning Research*. PMLR, 2018.

[39] Ioffe, Sergey and Szegedy, Christian. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* *Proceedings of the 32nd International Conference on Machine Learning*, τόμος 37 στο *Proceedings of Machine Learning Research*, Lille, France, 2015. PMLR.

[40] Gulrajani, Ishaan and Ahmed, Faruk and Arjovsky, Martin and Dumoulin, Vincent and Courville, Aaron. *Improved Training of Wasserstein GANs*. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, 2017. Curran Associates Inc.

[41] Schouten, Rianne M and Zamanzadeh, Davina and Singh, Prabhant. *pyampute: a Python library for data amputation*, 2022.

[42] Rianne Margaretha Schouten and Peter Lugtig and Gerko Vink. *Generating missing values for simulation purposes: a multivariate amputation procedure*. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

Παραρτήματα

Παράρτημα Α΄

Λίστα δημοσιεύσεων

Στα πλαίσια της διπλωματικής αυτής αναπτύχθηκε η εξής δημοσίευση:

- K. Psychogyios, L. Ilias and D. Askounis, "Comparison of Missing Data Imputation Methods using the Framingham Heart study dataset," 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 2022, pp. 1-5, doi: 10.1109/BHI56158.2022.9926882.

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels and according to World Health Organization is the leading cause of death worldwide. EHR data regarding this case, as well as medical cases in general, contain missing values very frequently. The percentage of missingness may vary and is linked with instrument errors, manual data entry procedures, etc. Even though the missing rate is usually significant, in many cases the missing value imputation part is handled poorly either with case-deletion or with simple statistical approaches such as mode and median imputation. These methods are known to introduce significant bias, since they do not account for the relationships between the dataset's variables. Within the medical framework, many datasets consist of lab tests or patient medical tests, where these relationships are present and strong. To address these limitations, in this paper we test and modify state-of-the-art missing value imputation methods based on Generative Adversarial Networks (GANs) and Autoencoders. The evaluation is accomplished for both the tasks of data imputation and post-imputation prediction. Regarding the imputation task, we achieve improvements of 0.20, 7.00% in normalised Root Mean Squared Error (RMSE) and Area Under the Receiver Operating Characteristic Curve (AUROC) respectively. In terms of the post-imputation prediction task, our models outperform the standard approaches by 2.50% in F1-score.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
κ.α.	και άλλα
ΤΝΔ	Τεχνητό νευρωνικό δίκτυο
ΗΥΦ	Ηλεκτρονικός φάκελος υγείας
ΕΜΠ	Εθνικό μετσόβιο πολυτεχνείο
ΓΑΝ	Γενετικό αναγωνιστικό δίκτυο
ΠΑΔ	Παραγωγικά αντιπαλικά δίκτυα
GAN	Generative adversarial network
GAIN	Generative adversarial imputation network
NAA	Neighborhood aware autoencoder
DAE	Denoising autoencoder
EHR	Electronic health record
KNN	k-nearest neighbors
SOTA	State of the art
MICE	multivariate imputation by chained equations
CVD	cardiovascular disease
MF	Missforest
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
απόδοση	imputation
αμεταβλητότητα	idempotency
σύγχρονος	state of the art
αντιμεταθετικότητα	commutativity
βάση δεδομένων	database
γνώρισμα	attribute
διαπροσωπεία	interface
διαφορά	difference
σκορ	score
ακρίβεια	accuracy