NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Predicting Post-Traumatic Stress Disorder (PTSD) symptoms in women suffering from breast cancer using machine learning

DIPLOMA THESIS

Konstantinos N. Rizavas

Supervisor:	Theodora Varvarigou
	Professor ECE, NTUA
In-charge:	Georgios Stamatakos
	Research Professor ICCS, ECE, NTUA
Advisory Committee:	Theodora Varvarigou, Georgios Stamatakos, Panagiotis Frangos,
	Professor ECE, NTUA
Honorary Member:	Emmanuel Protonotarios, Professor Emeritus, ECE, NTUA
of the Advisory Committee	

Athens, November 2023

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Predicting Post-Traumatic Stress Disorder (PTSD) symptoms in women suffering from breast cancer using machine learning

DIPLOMA THESIS

Konstantinos N. Rizavas

Supervisor: Theodora Varvarigou
Professor NTUA

In-charge: Georgios Stamatakos
Research Professor ICCS-NTUA

Approved by the examination committee on 6 November 2023

(Signature)	*(Signature)*	*(Signature)*
Theodora Varvarigou	*Georgios Stamatakos*	*Panagiotis Frangos*
Professor NTUA	*Research Professor ICCS-NTUA*	*Professor NTUA*

Athens, November 2023

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Konstantinos Rizavas

Electrical and Computer Engineer, NTUA

Περίληψη

Στο διάστημα των τελευταίων χρόνων, το φαινόμενο του καρκίνου του μαστού συνεχώς αυξάνεται σε συχνότητα, αλλά η θνησιμότητα της ασθένειας μειώνεται, χάρη στη αδιάκοπη πρόοδο της σύγχρονης ιατρικής και των τεχνολογικών μέσων. Ερχόμαστε αντιμέτωποι με μία νέα πραγματικότητα, στην οποία πρέπει να βοηθήσουμε τις γυναίκες που επέζησαν από καρκίνο του μαστού να ανακάμψουν ψυχολογικά και να επανενταχθούν ομαλά στην κοινωνία και στο εργατικό δυναμικό. Στο πλαίσιο αυτών των άνευ προηγουμένου συνθηκών, η παρούσα διπλωματική εργασία, τμήμα του χρηματοδοτούμενου από την Ευρωπαϊκή Ένωση προγράμματος BOUNCE, ερευνά τις δυνατότητες της χρήσης αλγορίθμων μηχανικής μάθησης στην πρόβλεψη συμπτωμάτων Διαταραχής Μετατραυματικού Στρες (ΔΜΣ) σε γυναίκες που πάσχουν από πρώιμο καρκίνο του μαστού, με τελικό στόχο την δημιουργία του ιδανικού μοντέλου και της αντίστοιχης μεθοδολογίας.

Τα δεδομένα που χρησιμοποιήθηκαν συγκεντρώθηκαν στα τέσσερα ογκολογικά νοσοκομεία – πρότυπα: IEO (Μιλάνο, Ιταλία), HUS (Ελσίνκι, Φινλανδία), HUJI (Ιερουσαλήμ, Ισραήλ) και Champalimaud (Λισαβόνα, Πορτογαλία). Οι μέθοδοι προεπεξεργασίας που χρησιμοποιήθηκαν έλαβαν υπόψιν τους την αναμενόμενα έντονη ανισορροπία των ιατρικών μας δεδομένων, τον περιορισμένο αριθμό δειγμάτων και τον αυξημένο αριθμό χαρακτηριστικών, καθώς και τους ταξινομητές μηχανικής μάθησης που σκοπεύαμε να χρησιμοποιήσουμε. Η εκπαίδευση των μοντέλων αξιοποίησε την τεχνική του επαναλαμβανόμενου cross-validation για να επιλέξει τις καλύτερες υπερπαραμέτρους του κάθε μοντέλου και η επίδοση των καλύτερων μοντέλων ελέγχθηκε σε ένα σύνολο δεδομένων ελέγχου που κρατήσαμε ξεχωριστά για να προσομοιάσουμε άγνωστα δεδομένα του πραγματικού κόσμου.

Τα πειράματα διενεργήθηκαν στα πλαίσια μιας «αφαιρετικής μελέτης» η οποία είχε ως στόχο την αναγνώριση των σημαντικών τμημάτων της διαδικασίας προεπεξεργασίας και μοντελοποίησης, καθώς και την ταυτοποίηση των χαρακτηριστικών εκείνων που υποδεικνύουν μεγάλη πιθανότητα παρουσίασης συμπτωμάτων ΔΜΣ και συνεπώς επηρεάζουν πολύ έντονα τα μοντέλα μας, οδηγώντας τα σε καλύτερες προβλέψεις. Η προκύπτουσα διαδικασία δοκιμάστηκε σε δεδομένα που προέρχονταν από νοσοκομεία των οποίων δεδομένα δεν είχαν συμπεριληφθεί στην εκπαίδευση, με σκοπό να ελεγχθεί η ικανότητα γεωγραφικής γενίκευσης των μοντέλων μας. Τα αποτελέσματα αυτής της έρευνας είναι πολλά υποσχόμενα και τονίζουν εμφατικά τις δυνατότητες της μηχανικής μάθησης στον κλάδο της ιατρικής και συγκεκριμένα στην πρόβλεψη ασθενειών και ψυχολογικών διαταραχών.

Λέξεις-κλειδιά:

Πρόγραμμα BOUNCE, Καρκίνος του Μαστού, ΔΜΣ, Μηχανική Μάθηση

Abstract

Over the course of the last few years, the phenomenon of breast cancer is constantly increasing in frequency, but the mortality of the disease is decreasing, thanks to the continuous advance of modern medicine and technological tools. A new reality dawns upon us, in which we need to help the surviving women BOUNCE back psychologically and reintegrate them smoothly in our society and workforce. Amidst these unprecedented circumstances, this diploma thesis, part of the European Union – funded BOUNCE project, researches the potential of using machine learning algorithms to try and accurately predict Post Traumatic Stress Disorder (PTSD) symptoms in women suffering from early breast cancer and ultimately aims to create the optimal model and associated methodology.

The data used were gathered at the four oncology centers – pilots: IEO (Milan, Italy), HUS (Helsinki, Finland), HUJI (Jerusalem, Israel) and Champalimaud (Lisbon, Portugal). The preprocessing methods used took into account the expected heavy imbalance of our medical data, the limited number of samples and the high number of features to consider, as well as the machine learning classifiers to be used. The model training leveraged repeated cross-validation in order to tune their hyper-parameters and the best models were evaluated on a separately-held test set to simulate unknown real-world data.

The experiments conducted were part of an ablation study that tried to identify the important aspects of our preprocessing and modelling procedure, and also pinpoint the important features that indicate high probability of developing PTSD symptoms and therefore greatly impact our models, leading them to better predictions. The resulting procedure was tested when being used on data from completely different hospitals to check its geographical generalizability. The outcome of this study demonstrates considerable promise and highlights the potential of machine learning in the field of medicine and more specifically in predicting diseases and psychological disorders.

Keywords

BOUNCE project, Breast Cancer, PTSD, Machine Learning

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την Καθηγήτρια κα. Θεοδώρα Βαρβαρίγου για την ευκαιρία που μου έδωσε να εκπονήσω την διπλωματική μου εργασία μαζί της. Επίσης, θα ήθελα να ευχαριστήσω την υπόλοιπη τριμελή επιτροπή, το Διευθυντή Ερευνών - Research Professor κ. Γεώργιο Σταματάκο και τον Καθηγητή κ. Παναγιώτη Φράγκο.

Ιδιαίτερα, θα ήθελα να ευχαριστήσω θερμά τον κ. Γεώργιο Σταματάκο για την εμπιστοσύνη και την κατεύθυνση της διπλωματικής και την ερευνήτρια του ΕΠΙΣΕΥ Δρ. Ελένη Κολοκοτρώνη για την συνεχή βοήθεια και καθοδήγηση καθ’ όλη την διάρκεια της εκπόνησης της εργασίας μου.

Θα ήθελα επίσης να ευχαριστήσω την Ευρωπαϊκή Επιτροπή και την Κοινοπραξία του ευρωπαϊκού ερευνητικού προγράμματος BOUNCE (https://www.bounce-project.eu/, https://cordis.europa.eu/project/id/777167) για την ευκαιρία να αξιοποιήσω μέρος των κλινικών δεδομένων του.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου για την υποστήριξη τους καθ’ όλη την διάρκεια των σπουδών μου.

Κωνσταντίνος Ριζάβας,

Αθήνα, Οκτώβριος 2023

Acknowledgements

I would like to thank Professor Theodora Varvarigou for the opportunity she offered me to carry out my diploma thesis under her supervision. I would also like to thank the members of the advisory committee Research Professor Georgios Stamatakos and Professor Panagiotis Frangos.

In particular, I would like to thank Research Prof. Georgios Stamatakos for the trust he showed to me and the direction of my diploma thesis, as well as Dr Eleni Kolokotroni, ICCS-NTUA for her continuous support and guidance throughout the implementation of the thesis work.

I would also like to acknowledge the European Commission (EC) and the EC funded BOUNCE project consortium for the opportunity they offered me to have access to part of the clinical data collected (https://www.bounce-project.eu/, https://cordis.europa.eu/project/id/777167).

Last but not least, I would like to thank my family and my friends for their precious support during my studies.

Konstantinos Rizavas

Athens, October 2023

Contents

Figures

Tables

Εκτεταμένη περίληψη στα Ελληνικά

Εισαγωγή

Η αντιμετώπιση του καρκίνου του μαστού είναι μια όλο και μεγαλύτερη πρόκληση, καθώς στη σύγχρονη εποχή τα κρούσματα συνεχώς πληθαίνουν. Παρά την συνεχώς αυξανόμενη εμφάνιση του φαινομένου, χάρη στη συνεχή πρόοδο της σύγχρονης ιατρικής και των συνεχώς βελτιούμενων τεχνολογικών μέσων που έχουμε στη διάθεση μας, η θνησιμότητα της ασθένειας έχει μειωθεί δραματικά. Αυτή η ευχάριστη τροπή, δημιουργεί μία νέα πραγματικότητα, την ανάγκη ομαλής επανένταξης των επιζώντων γυναικών στην κοινωνία και στις τάξεις του εργατικού δυναμικού. Ο στόχος του ευρωπαϊκού προγράμματος BOUNCE, τμήμα του οποίου απαρτίζει και η παρούσα διπλωματική είναι να αυξήσουν την ανθεκτικότητα των γυναικών που πάσχουν από πρώιμο καρκίνο του μαστού στις αναπόφευκτες ψυχοσωματικές συνέπειες της ασθένειας, της διάγνωσης της και της επιδιωκόμενης θεραπείας. Πιο συγκεκριμένα, ο στόχος της παρούσας διπλωματικής είναι να χρησιμοποιήσει τα δεδομένα που συλλέχθηκαν, στο πλαίσια του προγράμματος BOUNCE, στα τέσσερα ογκολογικά νοσοκομεία – πρότυπα, το IEO (Μιλάνο, Ιταλία), το HUS (Ελσίνκι, Φινλανδία), το HUJI (Ιερουσαλήμ, Ισραήλ) και το Champalimaud (Λισαβόνα, Πορτογαλία) για να προβλέψει την πιθανή ανάπτυξη συμπτωμάτων διαταραχής μετατραυματικού στρες.

0.1.1 Βασικές κλινικές πτυχές

Ο καρκίνος είναι μια συλλογή ασθενειών που χαρακτηρίζονται από την ανεξέλεγκτη αύξηση κακοηθών κυττάρων στον οργανισμό. Οι κακοήθεις όγκοι, γνωστοί και ως καρκίνοι, αναπτύσσονται όταν κύτταρα αρχίσουν να αλλοιώνονται γενετικά και εξελίσσονται σε κύτταρα που διαιρούνται ακανόνιστα. Αυτή η ανεξέλεγκτη ανάπτυξη κυττάρων μπορεί να σχηματίσει όγκους ή να εξαπλωθεί σε άλλα μέρη του σώματος μέσω της διάδοσης μεταστάσεων. Ο καρκίνος είναι μια σοβαρή ασθένεια που μπορεί να επηρεάσει κάθε μέρος του σώματος, και οι τύποι του καρκίνου είναι πολλοί. Οι ακριβείς αιτίες του καρκίνου εξακολουθούν να ερευνώνται, αλλά πολλοί παράγοντες, όπως γενετικοί, περιβαλλοντικοί και συμπεριφορικοί, μπορεί να συμβάλλουν στην ανάπτυξή του. Η πρόληψη, η πρόβλεψη, η διάγνωση και η αντιμετώπιση του καρκίνου απασχολούν έναν τεράστιο αριθμό ερευνητών γιατρών παγκοσμίως και βρίσκονται στην καρδιά της σύγχρονης ιατρικής.

Πιο συγκεκριμένα, ο καρκίνος του μαστού είναι μια μορφή καρκίνου που προκαλείται στον ιστό του μαστού και συναντάται κατά κύριο λόγο στις γυναίκες, αλλά μπορεί να επηρεάσει και άνδρες. Ο καρκίνος του μαστού μπορεί να παρουσιαστεί σε διάφορες μορφές και στάδια. Συνήθως, ανιχνεύεται με την ευαίσθητη μέθοδο του μαστογράφου. Οι παράγοντες κινδύνου για τον καρκίνο του μαστού περιλαμβάνουν τη γενετική προδιάθεση, την ηλικία, την οικογενειακή ιστορία, την έκθεση σε ορμονικές αλλαγές, καθώς και περιβαλλοντικούς παράγοντες. Η θεραπεία του καρκίνου του μαστού εξαρτάται από το στάδιο της νόσου και περιλαμβάνει συνήθως χειρουργική επέμβαση, ακτινοθεραπεία, χημειοθεραπεία, ακόμη και θεραπεία με στόχο τα ορμονικά κύτταρα, ανάλογα με τον τύπο του καρκίνου.

Η διαταραχή μετατραυματικού στρες (ΔΜΣ ή PTSD) είναι μια ψυχολογική διαταραχή που μπορεί να αναπτυχθεί όταν ένα άτομο έχει υποστεί ένα έντονο και τραυματικό γεγονός ή σειρά τραυματικών γεγονότων. Η ΔΜΣ προκαλείται συνήθως όταν το άτομο έχει υποστεί βία, σεξουαλική κακοποίηση, κάποιο ατύχημα, φυσικές καταστροφές ή γεγονότα που απειλούν τη ζωή του, όπως φυσικά είναι και τα σοβαρά προβλήματα υγείας. Τα κύρια συμπτώματα της ΔΜΣ περιλαμβάνουν επανεμφανιζόμενες ενοχλητικές θεάσεις, ήχους, ή αναμνήσεις από το τραυματικό γεγονός, εφιάλτες, υπερευαισθησία, αποφυγή συγκεκριμένων ερεθισμάτων, θλίψη και αϋπνίες. Από τα παραπάνω είναι προφανές ότι μπορεί να προκαλέσει σοβαρά προβλήματα στην καθημερινότητα του πάσχοντος ατόμου. Η αντιμετώπιση της ΔΜΣ συνήθως περιλαμβάνει ψυχοθεραπεία, καθώς και - σε ορισμένες περιπτώσεις – φαρμακευτική αγωγή για τη διαχείριση των συμπτωμάτων. Η υποστήριξη από επαγγελματίες ψυχικής υγείας και η υποστήριξη από την οικογένεια και τους φίλους είναι σημαντικά στοιχεία της ανάρρωσης για τα άτομα με ΔΜΣ.

0.1.2 Πτυχές ανάλυσης δεδομένων

Η παρούσα διπλωματική χρησιμοποιεί την τεχνική της επιβλεπόμενης μηχανικής μάθησης για να προβλέψει την εμφάνιση συμπτωμάτων διαταραχής μετατραυματικού στρες. Πρόκειται για μία από τις βασικές τεχνικές μηχανικής μάθησης και αναφέρεται σε ένα τύπο αλγορίθμου κατά τον οποίο το μοντέλο εκπαιδεύεται χρησιμοποιώντας ένα σύνολο δεδομένων που απαρτίζονται από κάποια είσοδο – ένα διάνυσμα χαρακτηριστικών και μια έξοδο – επιθυμητό αποτέλεσμα. Σκοπός του αλγορίθμου είναι να μάθει να προβλέπει τις ετικέτες ή τα αποτελέσματα για νέα δεδομένα που δεν έχει ξανά δει. Το μοντέλο μηχανικής μάθησης αναλύει αυτά τα δεδομένα και προσπαθεί να εξάγει πρότυπα και συσχετίσεις ανάμεσα στις εισόδους και τις ετικέτες, χρησιμοποιώντας τον εκάστοτε αλγόριθμο. Η επιβλεπόμενη μάθηση αποτελεί έναν από τους βασικούς τύπους μηχανικής μάθησης και χρησιμοποιείται ευρέως σε πολλές εφαρμογές όπως αναγνώριση προτύπων, αυτόματη μετάφραση, ανίχνευση αναφορών στα μέσα κοινωνικής δικτύωσης και άλλες.

Πολύ διαδεδομένες είναι οι τεχνικές bagging και boosting που χρησιμοποιούνται για την βελτίωση υπάρχοντων ταξινομητών. Οι δύο τεχνικές είναι φαινομενικά παρόμοιες, αφού χρησιμοποιούν και οι δύο πολλούς ταξινομητές (όπως πχ δέντρα αποφάσεων, που θα αναλυθούν παρακάτω) για να λάβουν τις τελικές τους αποφάσεις. Ωστόσο, έχουν αρκετά διαφορετικές ιδέες που αντιπροσωπεύουν:

- Bagging:

Κατασκευάζονται πολλοί απλοί ταξινομητές, ανεξάρτητοι μεταξύ τους, χρησιμοποιώντας διάφορες ιδέες που εξασφαλίζουν την διαφορετικότητα τους. Η πρόβλεψη του κάθε μοντέλου έχει το ίδιο βάρος. Το αποτέλεσμα είναι η μείωση του variance, δηλαδή της ικανότητας του αρχικού μοντέλου (δέντρο απόφασης) να υπερπροσαρμόζεται στα δεδομένα.

- Boosting:

Κατασκευάζονται πολλοί (συνήθως σχετικά απλοί – όπως μικρά δέντρα απόφασης) ταξινομητές, ο καθένας από τους οποίους εξαρτάται από τους προηγούμενους. Συνήθως, οι επόμενοι ταξινομητές δίνουν περισσότερη αξία στην ορθότερη πρόβλεψη των δειγμάτων που οι προηγούμενοι ταξινομητές απέτυχαν να ταξινομήσουν. Το αποτέλεσμα είναι η μείωση του bias των απλοϊκών ταξινομητών, δηλαδή της ανικανότητας τους να προσαρμόζονται επαρκώς στα δεδομένα.

Οι αλγόριθμοι μηχανικής μάθησης που δοκιμάστηκαν σε αυτή την διπλωματική είναι:

- **Δέντρα αποφάσεων:**

Η διαδικασία λειτουργίας ενός δέντρου αποφάσεων είναι αρκετά απλή. Ξεκινά από την ρίζα που αντιπροσωπεύει το σύνολο των δεδομένων εκπαίδευσης. Στη συνέχεια, το δέντρο αποφάσεων διαιρεί τα δεδομένα σε διαφορετικούς υπό-κόμβους βάσει των χαρακτηριστικών που παρέχονται. Κάθε κόμβος του δέντρου περιέχει μια ερώτηση ή μια συνθήκη που σχετίζεται με τα δεδομένα, και, ακολουθώντας τις ακμές του δέντρου, τα δεδομένα καταλήγουν σε φύλλα (leaf nodes) όπου λαμβάνεται η τελική απόφαση. Τείνουν να υπερπροσαρμόζονται στα δεδομένα, το οποίο σημαίνει ότι έχουν χειρότερη ικανότητα γενίκευσης και πρόβλεψης άγνωστων δεδομένων. Ωστόσο, υπάρχουν τεχνικές αποφυγής της υπερπροσαρμογής τους, μέσω «κλαδέματος» ή τεχνικών πρόωρου σταματημού. Επίσης, το μοντέλο μπορεί να εφαρμοστεί τόσο σε αριθμητικά όσο και σε κατηγορηματικά δεδομένα.

- **Random forest**:

Ο αλγόριθμος random forest είναι το προφανές επόμενο βήμα και το αποτέλεσμα της χρήσης τεχνικών bagging πάνω στα δέντρα αποφάσεων. Αντί να χρησιμοποιήσουμε ένα μεγάλο δέντρο απόφασης, δημιουργούμε πολλά μικρότερα, λαμβάνοντας υπόψιν μόνο ένα υποσύνολο των χαρακτηριστικών μας κάθε φορά. Κάτι τέτοιο οδηγεί συνήθως σε καλύτερη ικανότητα γενίκευσης του αλγορίθμου μας σε άγνωστα δεδομένα, που είναι και το κύριο ζητούμενο και ο απώτερος σκοπός σε προβλήματα μηχανικής μάθησης.

- **Adaboost**:

Ο αλγόριθμος adaboost είναι μια εφαρμογή τεχνικών boosting πάνω σε απλά – μικρά δέντρα αποφάσεων. Χρησιμοποιεί ως ταξινομητή – βάση απλοϊκά δέντρα αποφάσεων και σε κάθε επόμενο βήμα (μέχρι κάποια συνθήκη τερματισμού) το νέο δέντρο απόφασης στοχεύει στην διόρθωση των λαθών των προηγούμενων δέντρων.

- **XGBoost**:

Ο αλγόριθμος xgboost είναι κι αυτός μια εφαρμογή τεχνικών boosting πάνω σε απλά – μικρά δέντρα αποφάσεων, αλλά ακολουθεί διαφορετική λογική. Χρησιμοποιεί ως ταξινομητή – βάση απλοϊκά δέντρα αποφάσεων και σε κάθε βήμα (μέχρι κάποια συνθήκη τερματισμού) προσπαθεί να προβλέψει τη διαφορά – λάθος των μέχρι τώρα προβλέψεων και να χρησιμοποιήσει αυτή την πρόβλεψη για να «κινήσει» την πρόβλεψη του προς την σωστή κατεύθυνση και, στο τέλος του αλγορίθμου, στην σωστή πρόβλεψη.

- **Μηχανές διανυσμάτων στήριξης (ΜΔΣ ή SVM):**

Ο αλγόριθμος των μηχανών διανυσμάτων στήριξης είναι από τους πιο καλά εξερευνημένους και θεμελιωμένους αλγορίθμους μηχανικής μάθησης. Η εφαρμογή του ακόμα και σε χώρους υψηλής διαστατικότητας δεδομένων – όπως ο δικός μας – έχει καλά αποτελέσματα, αλλά αυτό δεν σημαίνει ότι η χρήση τεχνικών μείωσης της διαστατικότητας των δεδομένων δεν βελτιώνει περαιτέρω τα αποτελέσματα, γι’ αυτό και θα εξερευνηθεί κατά την πειραματική διαδικασία. Επιπλέον, είναι ο μοναδικός από τους επιλεγμένους αλγορίθμους που δεν μπορεί να διαχειριστεί κατηγορηματικά δεδομένα, που μας οδηγεί αναπόφευκτα στην χρήση της τεχνικής one-hot encoding για την μετατροπή κατηγορηματικών χαρακτηριστικών σε αριθμητικά. Πιο συγκεκριμένα, το κατηγορηματικό χαρακτηριστικό «μετατρέπεται» σε έναν πλήθος αριθμητικών χαρακτηριστικών που λαμβάνουν τις τιμές {0, 1}, ίσο σε αριθμό με τις διαφορετικές διακριτές τιμές του χαρακτηριστικού.

- **Ταξινομητής ψηφοφορίας**

Βάζουμε τους πέντε παραπάνω ταξινομητές μας σε ψηφοφορία και η ετικέτα με τις περισσότερες ψήφους επιλέγεται ως η προβλεπόμενη ετικέτα του δείγματος. Στα θετικά ενός τέτοιου ταξινομητή είναι πιο σταθερές προβλέψεις και πιθανόν καλύτερη επίδοση (όταν οι ταξινομητές κυμαίνονται σε παρόμοια επίπεδα αποτελεσμάτων), ενώ στα αρνητικά η ανάγκη περισσότερων υπολογιστικών πόρων και χρόνου πρόβλεψης, αφού πρέπει να εκπαιδευτούν όλα τα προαναφερθέντα μοντέλα.

0.2 Εφαρμογή μηχανικής μάθησης στα πολύπλευρα δεδομένα του προγράμματος BOUNCE

Ο τελικός στόχος είναι η χρήση των δεδομένων που συγκεντρώθηκαν στα τέσσερα ογκολογικά κέντρα – πρότυπα για την πρόβλεψη της ανάπτυξης συμπτωμάτων μετατραυματικού στρες στις γυναίκες που υποφέρουν από πρώιμο καρκίνο του μαστού. Για την παραπάνω πρόβλεψη χρησιμοποιήθηκαν τα δεδομένα των μηνών 0 (διάγνωση καρκίνου και πιθανή χειρουργική επέμβαση) και 3, όλων των τύπων: ιατρικά, ψυχοκοινωνικά, κοινωνικά – δημογραφικά καθώς και δεδομένα του τρόπου ζωής των γυναικών. Πέρα από τα δεδομένα που συλλέχθηκαν στα νοσοκομεία, μέρος των δεδομένων προήλθε και από την ηλεκτρονική πλατφόρμα Noona, ένα λογισμικό για προσωπική καταγραφή πληροφοριών από τους ίδιους τους ασθενείς.

0.2.1 Προεπεξεργασία των δεδομένων

Η προεπεξεργασία των δεδομένων είναι το πιο δύσκολο κομμάτι της εκπαίδευσης μοντέλων μηχανικής μάθησης. Κάθε σύνολο δεδομένων έχει τα δικά του ιδιαίτερα χαρακτηριστικά και τις δικές του δυσκολίες. Το δικό μας σύνολο δεδομένων χαρακτηρίζεται από σχετικά λίγες παρατηρήσεις (αριθμό δεδομένων), αλλά πολλά χαρακτηριστικά (μεταβλητές που λαμβάνουμε υπόψιν μας για κάθε παρατήρηση). Τα βήματα προεπεξεργασίας που ακολουθήθηκαν παρατίθενται παρακάτω:

0. Εξάλειψη των τεσσάρων χαρακτηριστικών που οι ιατροί – ερευνητές υπέδειξαν:

Οι ιατροί υπέδειξαν ότι κάποια από τα χαρακτηριστικά έχουν πολύ μεγάλη λογική – ιατρική συσχέτιση με άλλα, πιο εμφανώς σημαντικά χαρακτηριστικά και η ύπαρξη όλων μάλλον θα μπέρδευε τα μοντέλα περισσότερα από ότι θα βοήθαγε την εκμάθηση τους. Τα ονόματα των τεσσάρων χαρακτηριστικών είναι:

- Emot_Fun_QLQ30.0
- Emot_Fun_QLQ30.3
- nccn_distress_thermometer.0
- nccn_distress_thermometer.3

Το βήμα αυτό εκτελέσθηκε στην αρχή της προεπεξεργασίας, με στόχο να μην επηρεάσουν τα υπόλοιπα τμήματα της, ενώ στο τέλος σκοπεύαμε να μην τα συμπεριλάβουμε.

1. Εξάλειψη παρατηρήσεων χωρίς ετικέτα:

Είναι προφανές ότι παρατηρήσεις χωρίς την αντίστοιχη ετικέτα τους δεν μπορούν να συνεισφέρουν στα μοντέλα επιβλεπόμενης μηχανικής μάθησης που αξιοποιεί η παρούσα διπλωματική.

2. Εξάλειψη χαρακτηριστικών που λείπουν από πολλές παρατηρήσεις:

Δεν μας πειράζει να διώξουμε μερικά χαρακτηριστικά δεδομένου ότι έχουμε μεγάλο αριθμό. Συνεπώς, επιλέχθηκε το κατώφλι 0.25, δηλαδή αν λείπουν περισσότερες από 1 στις 4 τιμές διώχνουμε το χαρακτηριστικό.

3. Εξάλειψη των παρατηρήσεων από τις οποίες λείπουν πολλά χαρακτηριστικά:

Σε αντίθεση με το προηγούμενο βήμα, μιας και διαθέτουμε μικρό αριθμό παρατηρήσεων, θα προτιμούσαμε να μην διώξουμε πολλές από αυτές, εκτός αν έχουν πραγματικά πολύ λίγη πληροφορία. Γι’ αυτό, επιλέχθηκε το πιο συγκρατημένο κατώφλι 0.5, δηλαδή αν λείπουν περισσότερα από τα μισά χαρακτηριστικά, διώχνουμε αυτή την παρατήρηση.

4. Εξάλειψη των χαρακτηριστικών με σχεδόν μηδενική διακύμανση:

Καθώς τα αριθμητικά χαρακτηριστικά έχουν διαφορετικά εύρη τιμών, δεν μπορούμε πραγματικά να τα συγκρίνουμε και να τα κατηγοριοποιήσουμε ως έχοντα χαμηλή διακύμανση. Από την άλλη, τα κατηγορηματικά χαρακτηριστικά δύνανται να αξιολογηθούν ως τέτοια αν:

- Η πιο συνήθης τιμή είναι πολύ πιο συνηθισμένη και από την δεύτερη πιο συνήθη τιμή και
- Ο αριθμός διαφορετικών τιμών του χαρακτηριστικού δεν υπερβαίνει ένα κατώφλι

5. Εξάλειψη υψηλά συσχετισμένων χαρακτηριστικών:

Δύο υψηλά συσχετισμένα χαρακτηριστικά περιέχουν κατά μεγάλο ποσοστό την ίδια πληροφορία και η συμπερίληψη και των δύο συχνά «μπερδεύει» τα μοντέλα. Επιλέχθηκε η συσχέτιση Spearman που είναι μη-παραμετρική και μπορεί να διακρίνει μονοτονικές σχέσεις μεταξύ μεταβλητών και όχι μόνο γραμμικές. Επίσης, μπορεί να εφαρμοστεί σε αριθμητικά και κατηγορηματικά δεδομένα, σε αντίθεση με την πιο συνηθισμένη συσχέτιση Pearson και είναι λιγότερο ευαίσθητη σε outliers, δηλαδή σπάνιες τιμές πολύ έξω από το σύνηθες εύρος τιμών. Χρησιμοποιήθηκε το συνηθισμένο κατώφλι συσχέτισης 0.8.

6. Χωρισμός των δεδομένων σε τμήματα εκπαίδευσης/ελέγχου:

Δοκιμάζουμε δύο διαφορετικές μεθόδους χωρισμού των δεδομένων. Σε πρώτη φάση τα χωρίζουμε τυχαία με ποσοστά 70% στο τμήμα εκπαίδευσης και 30% στο τμήμα ελέγχου. Σε δεύτερη φάση, πειραματιζόμαστε για το αν τα μοντέλα μας μπορούν να γενικεύσουν τις προβλέψεις τους σε δεδομένα προερχόμενα από διαφορετικά νοσοκομεία και γεωγραφικές περιοχές, από αυτά στα οποία εκπαιδεύθηκαν. Γι’ αυτό, χωρίζουμε τα δεδομένα μας χρησιμοποιώντας ως τεστ ελέγχου κάθε φορά τα δεδομένα που προέρχονται από ένα από τα τέσσερα νοσοκομεία – πρότυπα.

7. Πρόβλεψη ελλειπουσών τιμών:

Δεδομένου ότι κάποια από τα μοντέλα που χρησιμοποιούμε δεν μπορούν να διαχειριστούν την ύπαρξη ελλειπουσών τιμών στα δεδομένα μας, πρέπει να χρησιμοποιήσουμε κάποια τεχνική πρόβλεψης τους. Επιλέχθηκε η τεχνική *Multiple Imputation Chained Equations (MICE)*, κατά την οποία κάθε ημιτελές χαρακτηριστικό προβλέπεται από ένα ξεχωριστό μοντέλο. Ο αλγόριθμος MICE επιτρέπει την αντιμετώπιση ελλειπουσών τιμών σε δεδομένα με πολλαπλές μεταβλητές, λαμβάνοντας υπόψη τις συσχετίσεις μεταξύ τους. Αυτό βοηθά στη διατήρηση της δομής των δεδομένων και στην ανακατασκευή των ελλειπουσών τιμών με τρόπο πιο ρεαλιστικό από απλές αντικαταστάσεις με μέσους όρους ή κατώφλια. Άξιο αναφοράς είναι ότι στην δημιουργία των μοντέλων πρόβλεψης των τιμών χρησιμοποιούνται μόνο τα δεδομένα εκπαίδευσης, αλλά προβλέπονται οι ελλείπουσες τιμές και των δεδομένων ελέγχου.

Παράγουμε 5 διαφορετικά σύνολα δεδομένων χρησιμοποιώντας τις προβλέψεις μας. Ύστερα, χρησιμοποιούμε τα μοντέλα μας και στα 5 αυτά σύνολα και για κάθε ταξινομητή παίρνουμε τον μέσο όρο της πρόβλεψής του στα διαφορετικά σύνολα. Με αυτόν τον τρόπο είμαστε πιο σίγουροι ότι, αφού τα μοντέλα μας μπορούν να προβλέψουν καλά χρησιμοποιώντας όλα τα διαφορετικά σύνολα δεδομένων μας που προσπαθούν να προσεγγίσουν το «πραγματικό» σύνολο δεδομένων αν δεν υπήρχαν ελλείπουσες τιμές, τότε το μοντέλο μας πράγματι μαθαίνει την κρυμμένη «δομή» των δεδομένων μας και όχι ένα από σύνολα που με κάποια τυχαιότητα δημιουργήσαμε.

8. Χρήση τεχνικών μείωσης της διαστατικότητας των δεδομένων μας:

Οι βασικοί λόγοι που μας οδηγούν στην χρήση τέτοιων τεχνικών είναι ευκολότεροι υπολογισμοί, ευκολότερο μάζεμα δεδομένων από τους ασθενείς, καθώς και πιθανή βελτίωση των αποτελεσμάτων των μοντέλων μας, αφού συχνά αυτά υποφέρουν από την λεγόμενη «κατάρα της διαστατικότητας», δηλαδή έχουν υποδεέστερα αποτελέσματα όταν εκπαιδεύονται σε σύνολα δεδομένων με λίγες παρατηρήσεις και πολλά – συγκριτικά – χαρακτηριστικά. Η τεχνική που χρησιμοποιήσαμε ονομάζεται *recursive feature elimination*, δηλαδή επαναληπτική εξάλειψη χαρακτηριστικών. Καταρχάς, επιλέγεται ένα μοντέλο που μπορεί εσωτερικά να υπολογίσει σημαντικότητα χαρακτηριστικών, όπως για παράδειγμα το random forest που χρησιμοποιήσαμε εμείς. Έπειτα, για έναν επιλεγμένο αριθμό από πιθανούς αριθμούς χαρακτηριστικών S_i που θέλουμε να δοκιμάσουμε, κρατάμε τα S_i πιο σημαντικά χαρακτηριστικά και ελέγχουμε την επίδοση τους σε σχέση με τον αρχικό αριθμό χαρακτηριστικών και τους υπόλοιπους που επιλέξαμε. Στο τέλος κρατάμε αυτόν με την καλύτερη επίδοση, εντός κάποιων ορίων ανοχής. Αυτή η τεχνική εφαρμόστηκε σε καθένα από τα 5 σύνολα δεδομένων που προβλέψαμε στο προηγούμενο βήμα και η ένωση των επιλεγμένων χαρακτηριστικών είναι αυτή που εν τέλει χρησιμοποιήθηκε.

9. One-hot μετασχηματισμός των κατηγορικών χαρακτηριστικών:

Όπως προαναφέρθηκε, κάποια μοντέλα όπως τα SVM δεν μπορούν να διαχειριστούν κατηγορικά δεδομένα, αλλά μόνο αριθμητικά. Σε τέτοιες περιπτώσεις είναι απαραίτητη η μετατροπή αυτών των δεδομένων, κατά την οποία το κατηγορηματικό χαρακτηριστικό «μετατρέπεται» σε έναν πλήθος αριθμητικών χαρακτηριστικών που λαμβάνουν τις τιμές {0, 1}, ίσο σε αριθμό με τις διαφορετικές διακριτές τιμές του χαρακτηριστικού.

10. Κανονικοποίηση χαρακτηριστικών:

Η κανονικοποίηση των δεδομένων είναι μια διαδικασία κατά την οποία επιδιώκουμε να φέρουμε τα διαφορετικά και ετερογενή δεδομένα που διαθέτουμε στο ίδιο εύρος τιμών. Το μοναδικό μοντέλο που για να λειτουργήσει αποτελεσματικά χρειάζεται να κανονικοποιήσουμε τα δεδομένα πριν του τα δώσουμε είναι το SVM, καθώς σκοπεύουμε να χρησιμοποιήσουμε ευκλείδεια απόσταση για να μετρήσουμε τις αποστάσεις των σημείων του. Αν δεν τα κανονικοποιήσουμε, στην πραγματικότητα δίνουμε περισσότερη σημασία στα δεδομένα που έχουν μεγαλύτερη κλίμακα, σε σχέση με αυτά που έχουν μικρότερη. Τα υπόλοιπα μοντέλα βασίζονται στα δέντρα απόφασης, στα οποία κάθε κόμβος αναπαριστά μία ερώτηση, η ουσία της οποίας δεν αλλάζει βάσει της κλίμακας των δεδομένων και συνεπώς δεν χρειάζονται κανονικοποίηση.

11. Εξισορρόπηση των συνόλων δεδομένων:

Δεν πρέπει να ξεχνάμε ότι πλέον έχουμε 5 διαφορετικά σύνολα δεδομένων που χειριζόμαστε ταυτόχρονα. Καθένα από αυτά λοιπόν για να εκπαιδευτούν πάνω του μοντέλα χρειάζεται να γίνει εξισορρόπηση των ετικετών του, χρησιμοποιώντας μία εκ των τεχνικών downsampling, oversampling ή ROSE (τεχνητού oversampling). Η μη εξισορρόπηση είναι δυνατή, αλλά όπως θα φανεί στα πειράματα οδηγεί σε πολύ χειρότερη επίδοση των μοντέλων. Πιο συγκεκριμένα, μιας και η εκπαίδευση των μοντέλων μας γίνεται με την τεχνική του repeated cross-validation, για κάθε repeat και για κάθε fold, κατά την οποία περίπτωση το μοντέλο δοκιμάζεται σε ένα συγκεκριμένο σύνολο εκπαίδευσης και ελέγχου, χρειάζεται να εφαρμοστεί η επιλεγμένη τεχνική εξισορρόπησης.

Η μετρική που χρησιμοποιείται καθ’ όλη τη διάρκεια της εργασίας για την αποτίμηση των επιδόσεων του είναι το **f2-score**, που πρόκειται για τον ελαφρώς τροποποιημένο αρμονικό μέσο των precision και recall, με περισσότερη βαρύτητα στο recall. Το recall μας λέει τι ποσοστό από την θετική κλάση ορθώς εντοπίσαμε, ενώ το precision το ποσοστό από τα δείγματα που προβλέψαμε ότι ανήκαν στην θετική κλάση που όντως ανήκαν στην θετική κλάση.

0.2.2 Αποτελέσματα πειραμάτων

Μετά την επιλογή των υπερπαραμέτρων με τις προαναφερθείσες τεχνικές, ελέγχουμε την επίδοση των ταξινομητών μας στο σύνολο δεδομένων ελέγχου, που χρησιμοποιείται για να προσομοιώσει άγνωστα πραγματικά δεδομένα. Παίρνουμε το αποτέλεσμα κάθε ταξινομητή για καθένα από τα 5 διαφορετικά σύνολα δεδομένων που έχουμε και υπολογίζουμε τον μέσο όρο των επιδόσεών του.

Στο πρώτο πείραμα, ελέγξαμε την αξία του $11^{ου}$ βήματος της προεπεξεργασίας και της εξισορρόπησης των δεδομένων μας. Δοκιμάστηκαν η μη εξισορρόπηση των δεδομένων και η χρήση τριών διαφορετικών τεχνικών εξισορρόπησης, downsampling, oversampling ή ROSE (τεχνητού oversampling). Τα πειράματα έδειξαν ότι η ενδεικνυόμενη τεχνική είναι αυτή του downsampling, από άποψη επίδοσης, αλλά και από άποψη χρόνου και υπολογιστικού φόρτου.

Table 1: Σύγκριση επιδόσεων f2 χρησιμοποιώντας διαφορετικές τεχνικές εξισορρόπησης

Sampling method *Classifier*	*None*	*Downsampling*	*Oversampling*	*ROSE*
Decision tree	0.110	0.344	0.322	**0.308**
Random forest	**0.190**	0.390	0.100	0.262
SVM	0.000	0.332	0.000	0.090
Adaboost	0.124	0.378	**0.320**	0.214
XGBoost	0.128	0.356	0.104	0.272
Voting	0.098	**0.392**	0.088	0.236

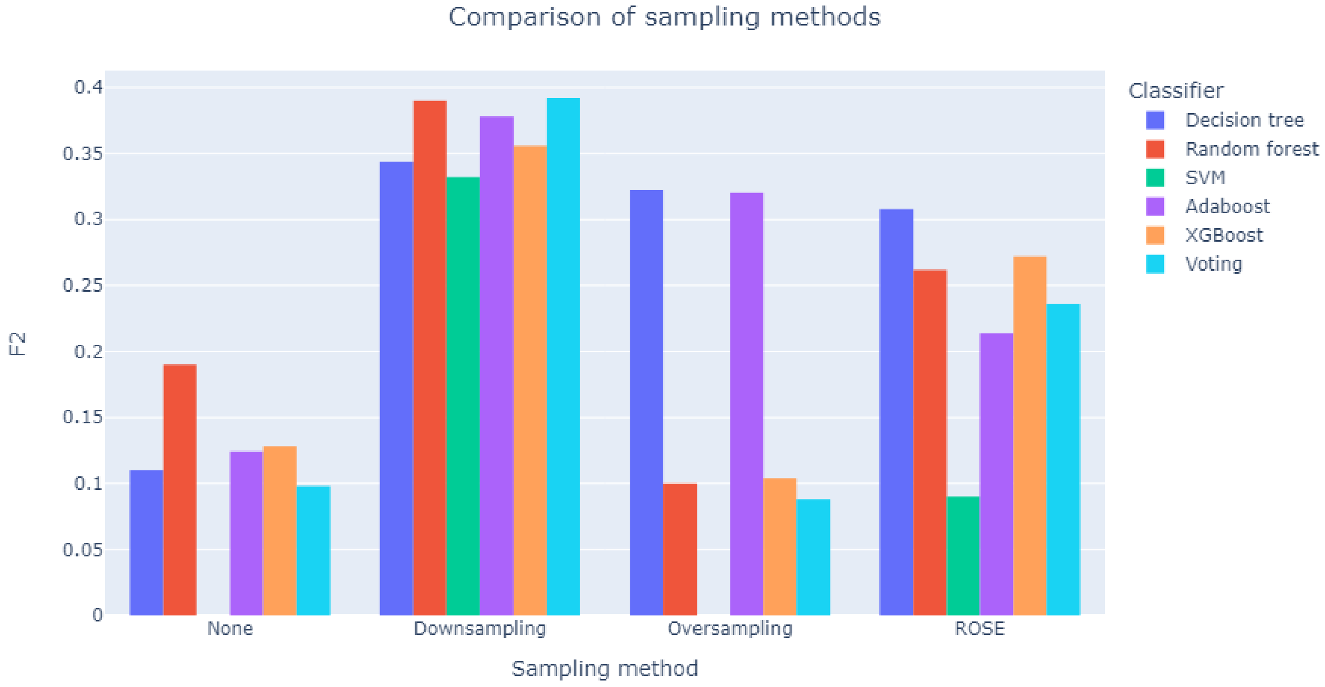


Figure 1: Σύγκριση τεχνικών εξισορρόπησης

Στο δεύτερο πείραμα, ελέγξαμε το προαιρετικό βήμα του preprocessing κατά το οποίο αγνοούμε τα τέσσερα χαρακτηριστικά που επέδειξαν οι ιατροί – ερευνητές ως λογικά συσχετισμένα με τα υπόλοιπα σημαντικά ψυχολογικά – ιατρικά χαρακτηριστικά. Τα αποτελέσματα μαρτυρούν μικρές διαφορές, αλλά ελαφρώς καλύτερα αποτελέσματα όταν τα αγνοούμε και συνεπώς αυτό κάνουμε.

Table 2: Σύγκριση επιδόσεων f2 αγνοώντας ή όχι τα υποδεδειγμένα χαρακτηριστικά

Ignoring features *Classifier*	*No*	*Yes*
Decision tree	0.344	0.380
Random forest	0.390	0.382
SVM	0.332	0.330
Adaboost	0.378	0.396
XGBoost	0.356	0.372
Voting	**0.392**	**0.396**

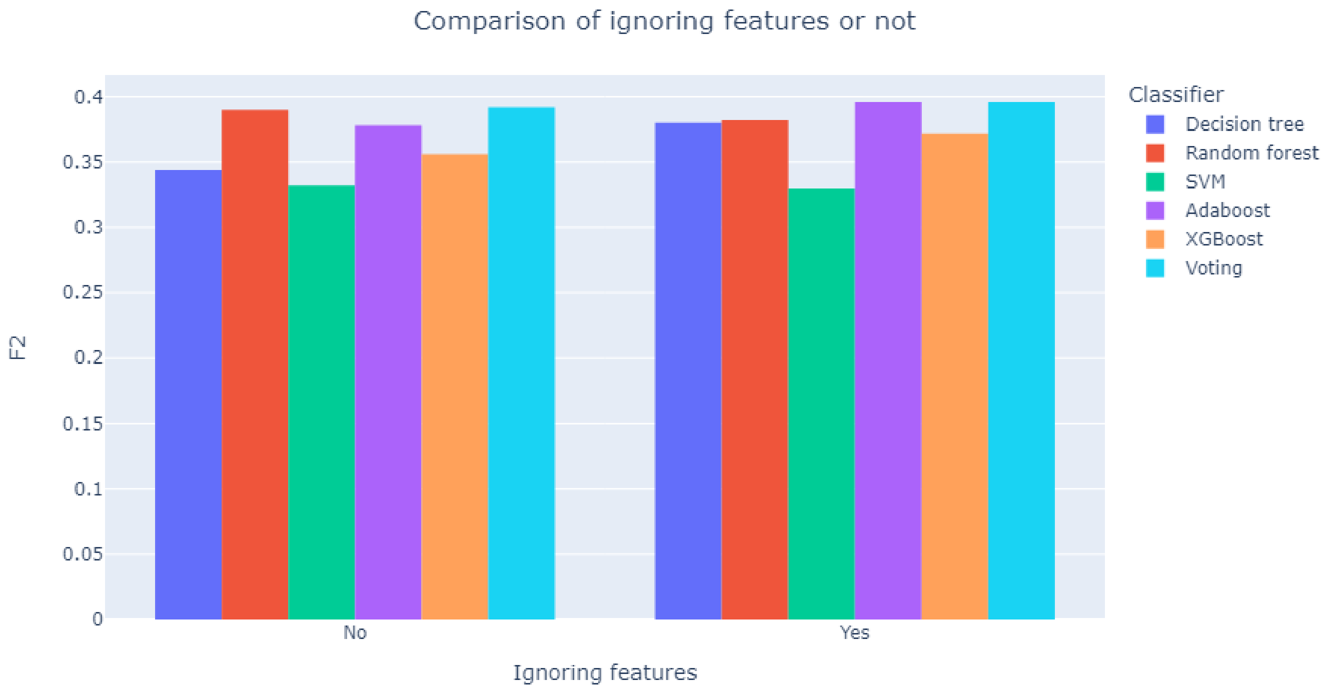


Figure 2: Σύγκριση αγνόησης των λογικά συσχετισμένων χαρακτηριστικών

Στο τρίτο πείραμα, χρησιμοποιούμε την τεχνική RFE για μείωση των χαρακτηριστικών μας όπως αναφέραμε στο βήμα 8 της προεπεξεργασίας. Αφού εντοπίζουμε τα πιο σημαντικά χαρακτηριστικά βάσει του RFE, χρησιμοποιούμε μόνο αυτά για να εκπαιδεύσουμε τους ταξινομητές μας. Παρατηρούμε σε γενικές γραμμές ελαφρώς αυξημένες επιδόσεις (ειδικά στο SVM), το οποίο είναι πολύ σημαντικό εύρημα, αφού μας επιτρέπει να συγκεντρώσουμε τα 27 χαρακτηριστικά που επιλέχθηκαν αντί για τα αρχικά 172. Κάτι τέτοιο προφανώς είναι μεγάλη διευκόλυνση για τους ιατρούς που αναπόφευκτα μετέχουν στην διαδικασία συγκέντρωσης των δεδομένων από τους ασθενείς.

Table 3: Σύγκριση επιδόσεων f2 χρησιμοποιώντας ή όχι RFE

Features used / *Classifier*	*All*	*Reduced*
Decision tree	0.380	0.336
Random forest	0.382	0.400
SVM	0.330	**0.436**
Adaboost	0.396	0.364
XGBoost	0.372	0.356
Voting	**0.396**	0.400

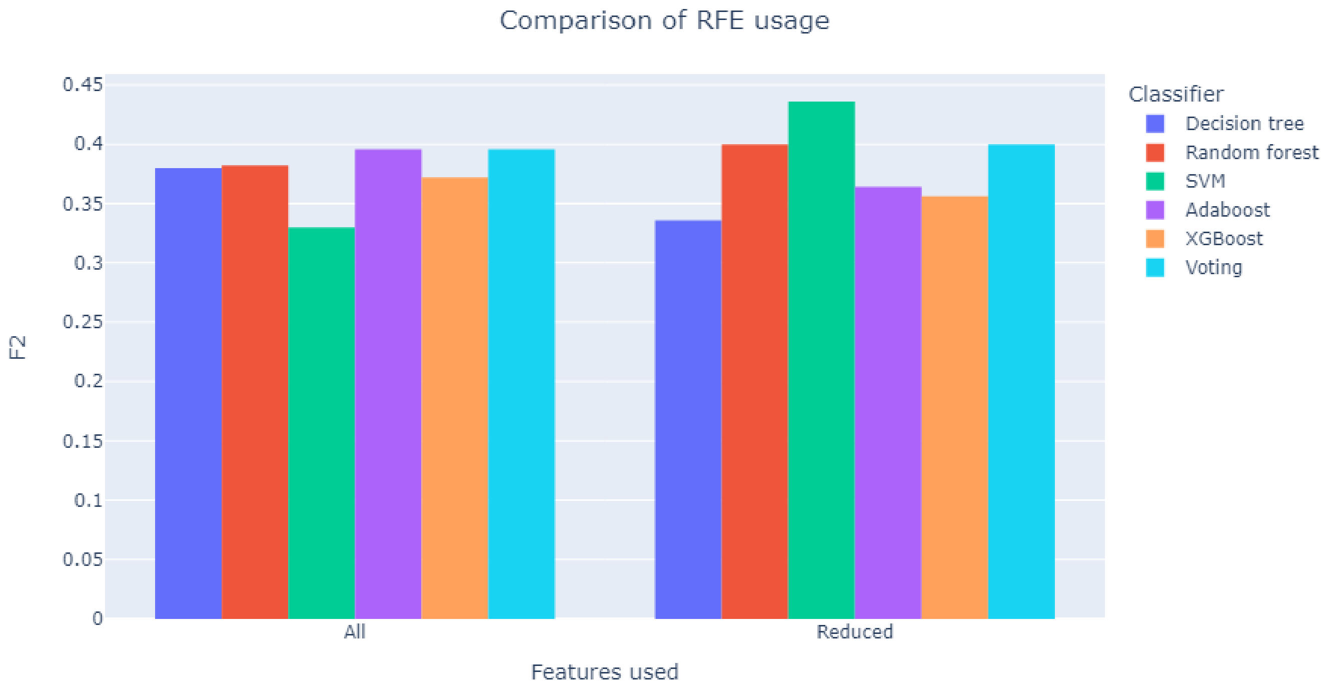


Figure 3: Σύγκρισης χρήσης RFE

Στο τέταρτο και τελευταίο ***πείραμα***, δοκιμάσαμε την γενίκευση των αποτελεσμάτων μας αν τα μοντέλα μας δεν έχουν εκπαιδευτεί σε δεδομένα από τα νοσοκομεία των οποίων τους ασθενείς θέλουμε να προβλέψουμε. Τα αποτελέσματα είναι αρκετά κοντά μεταξύ τους, το οποία μας δίνει αισιοδοξία για την χρήση των μοντέλων μας και σε άλλα νοσοκομεία, αφού μαρτυρά ότι αυτά κατάφεραν να μάθουν επαρκώς τη σχέση μεταξύ χαρακτηριστικών και ετικέτας που χαρακτηρίζει γενικά τους ασθενείς και όχι συγκεκριμένα τους ασθενείς ενός νοσοκομείου μιας συγκεκριμένης πόλης. Όπως εξηγήθηκε κατά την αναλυτική παρουσίαση των πειραμάτων, η μετρική f2 στη συγκεκριμένη περίπτωση μας «παραπλανεί» ελαφρώς λόγω των διαφορετικών ποσοστών δειγμάτων θετικής κλάσης στα διαφορετικά σύνολα δεδομένων ελέγχου που προκύπτουν, γι' αυτό παραθέτουμε την μετρική AUC:

Table 4: Σύγκριση επιδόσεων AUC στα δεδομένα των διαφορετικών νοσοκομείων

Testing on *Classifier*	*Random*	*IEO*	*HUS*	*HUJI*	*CHAMP*
Decision tree	0.636	0.618	0.626	0.720	0.640
Random forest	0.690	**0.736**	0.700	0.804	0.700
SVM	**0.734**	0.656	**0.732**	0.786	0.692
Adaboost	0.660	0.710	0.702	0.774	0.648
XGBoost	0.654	0.706	0.656	0.782	**0.728**
Voting	0.692	0.696	0.702	**0.830**	0.708

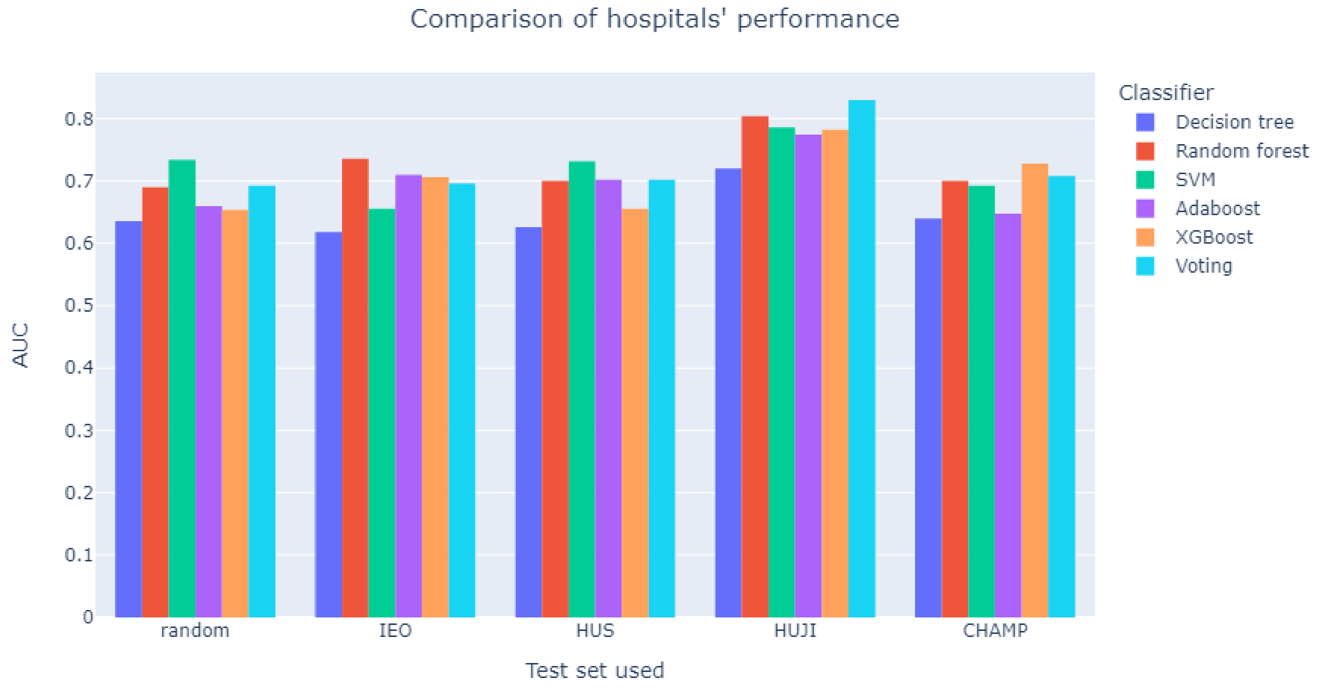


Figure 4: Σύγκριση επιδόσεων AUC στα διαφορετικά νοσοκομεία

0.3 Συμπεράσματα

Η παρούσα διπλωματική θεμελίωσε μία σειρά βημάτων και αποφάσεων για την δημιουργία του ιδανικού – υπό τις συνθήκες δεδομένων – μοντέλου για την πρόβλεψη συμπτωμάτων μετατραυματικού στρες σε γυναίκες με πρώιμο καρκίνο του μαστού. Η προεπεξεργασία των δεδομένων που προτάθηκε οδηγεί σε καλές επιδόσεις, αν συνδυαστεί με την κατάλληλη τεχνική εξισορρόπησης. Η χρήση των μειωμένων χαρακτηριστικών διευκολύνει πολύ την συγκέντρωση των δεδομένων και έχει πολύ καλά αποτελέσματα. Οι προβλέψεις των μοντέλων γενικεύονται γεωγραφικά και δεν εξαρτώνται από τοπικά χαρακτηριστικά ή ιδιορρυθμίες – προβλήματα των νοσοκομείων στη συλλογή των δεδομένων. Δεν υπήρξε ξεκάθαρα καλύτερος ταξινομητής, μεταξύ όσων δοκιμάστηκαν. Αν έπρεπε να διαλέξουμε, θα επιλέγαμε τον SVM ή τον random forest ως μεμονωμένους ή τον voting αν είχαμε τον απαραίτητο χρόνο και υπολογιστικούς πόρους για να εκπαιδεύσουμε και τους πέντε μεμονωμένους ταξινομητές, αφού προσφέρει μεγαλύτερη σταθερότητα στις προβλέψεις του.

Chapter 1. Introduction

1.1 A brief outline of the BOUNCE project

Coping with breast cancer more and more becomes a major socio-economic challenge not least due to its constantly increasing incidence in the developing world. There is a growing need for novel strategies to improve understanding and capacity to predict resilience of women to the variety of stressful experiences and practical challenges related to breast cancer. This is a necessary step toward efficient recovery through personalized interventions. BOUNCE brings together modelling, medical, and social sciences experts to advance current knowledge on the dynamic nature of resilience as it relates to efficient recovery from breast cancer. It takes into consideration clinical, cancer-related biological, lifestyle, and psychosocial parameters in order to predict individual resilience trajectories throughout the cancer continuum and eventually increase resilience in breast cancer survivors and help them remain in the workforce and enjoy a better quality of life. BOUNCE delivers a unified clinical model of modifiable factors associated with optimal disease outcomes and deploys a prospective multi-center clinical pilot at four major oncology centers (in Italy, Finland, Israel and Portugal), where a total of 660 women will be recruited in order to assess its clinical validity against crucial patient outcomes (illness progression, wellbeing, and functionality). The advanced computational tools to be employed validate indices of patients' capacity to bounce back during the highly stressful treatment and recovery period following diagnosis of breast cancer. The overreaching goal of BOUNCE is to incorporate elements of a dynamic, predictive model of patient outcomes in building a decision-support system used in routine clinical practice to provide physicians and other health professionals with concrete, personalized recommendations regarding optimal psychosocial support strategies [1].

Cancer is devastatingly proving to be increasing in frequency in the developing world. Breast cancer in particular is the most frequent type of cancer found in women and is currently on the rise. Nevertheless, thanks to the continuous advance of modern medicine and the ever-evolving technological tools, the mortality of the disease is predictably decreasing, even with invasive type cancer. This pleasant achievement subsequently creates an important problem; the survivors have gone through many difficulties, possibly including surgery, chemotherapy and other cures that are hard to just leave in the past, resulting to many possible psychosomatic marks. Even the news of facing breast cancer, as well as waiting for the results of the diagnostic testing are enough to create a lot of stress and anxiety. Also, the psychological state of a woman just diagnosed with breast cancer is different from that of a woman who is coping after her surgery and they should all be treated accordingly.

The goal of the BOUNCE project is to amplify the resilience of breast cancer patients and help them **bounce** back. More specifically, the goal of the BOUNCE project is to create a decision-support system to be used in clinical practice that provides health professionals with trustworthy, personalized interventions, as well as the exact timing these interventions should take place. Its goal is not to create a standalone decision, but to assist in pointing early whether each patient seems to have a reasonable possibility of developing psychological problems due to the stressful situation. Afterwards, further examination will be carried out by a medical expert of the field. The main focus is on assisting patients susceptible to mental tumbling early, so as to facilitate the patients' recovery and smooth reintegration in the community and into the workforce. It goes without saying that the most important factor in the psychological state of a patient is the disease prediction and therapeutic outcome. That is the exact reason why the resilience predictive model will be coupled with an in-silico breast cancer predictive model of acceptable validity. In other words, there is a double goal; the project wants to be able to influence both clinicians, by giving them assistance in crucial treatment decisions, as well as mental health professionals to deal with the psychological problems that simultaneously arise during the therapy. The above trajectories are modeled using machine learning for the psychological one and mostly mechanistic modelling (combined with some machine learning) for the medical one. The mechanistic multiscale models to be used will help integrating already acquired knowledge of crucial mechanisms of tumor growth to the machine learning models.

The end-product is a clinically validated trajectory prediction of the patient's clinical relapse, physical/psychological well-being and functionality. The study aims to pinpoint not only the modifiable characteristics that lead to augmented/declined resilience and optimal disease outcomes, but also the time points at which these characteristics have an important impact on the patient and use this knowledge to provide customized and well-timed interventions. The model is developed for breast cancer patients, but is generalizable for many chronic illnesses, taking into account that the logic behind all the modeling and the combined trajectories does not apply only to our specific case [2].

In the framework of the BOUNCE project, several important aspects of the resilience of women with early breast cancer have been addressed, analyzed and modelled. Some of these analyses are included in the following papers: [3], [4], [5], [6], [7], [8], [9], [10], [11] and [12]

1.2 Clinical questions addressed

The part of the project I participated in involves the prediction of the possible psychological turmoil of the breast cancer patients as soon as possible, as well as analyzing the most important characteristics that lead to the decision of whether the patient is likely to suffer emotionally. Therefore:

- If these characteristics are modifiable, they can be tampered with in order to try and boost the resilience of the patients.
- If they are not modifiable, they can serve as important indexes of early discernment of the expected psychological trajectory of the patients.

More specifically, a supervised machine learning approach was chosen, using data from the earlier, and thus most important, months after the diagnosis. Data from months zero and three (M0, M3) were used towards predicting the Post Traumatic Stress Disorder (PTSD) symptoms of the patients. More information about the exact label considered will be discussed in the next chapter.

1.3 Organization of diploma thesis

In the second chapter of this diploma thesis, the basic clinical aspects that played a key part during the whole thesis are discussed. First of all, the biggest health problem that plagues our modern society is issued, cancer, discussing the factors that contribute to its existence and also the ways of dealing with it that we have in our modern-era medical arsenal. We specifically elaborate about breast cancer and its unique problems and solutions. Afterwards, we define Post Traumatic Stress Disorder (PTSD), present its symptoms and the questionnaire we used to make it the variable to predict for this thesis. Finally, we take a look at the Hospital Anxiety and Depression Scale (HADS) questionnaire that will play an important part.

The third chapter focuses on the data-analysis aspects that are necessary for understanding the contents of this thesis. After a brief discussion of the types of machine learning, we focus on the one we used, supervised learning, and its metrics. Furthermore, we analyze how the classifiers that are used in this thesis function, their advantages and disadvantages and why they were chosen for the problem in hand.

In the fourth chapter, we explore related work to discover if and how our research has value and offers a new point of view to previous research on the subject. We present research papers that attempted to predict PTSD or similar psychological variables using other methods. Afterwards we present research papers that utilized machine learning in similar problems with good results. In the final paragraph, we discuss why using machine learning in our problem of predicting PTSD makes sense and the important differentiation that makes it novel research in the field.

The fifth chapter showcases the suggested methodology we followed towards solving the problem in hand. It includes the data preprocessing steps, the development of the predictive models and the evaluation of said models.

In the sixth chapter, we present all our findings, further analyze all the steps and the ideas behind them. We include figures and tables to explain our experiments and use them to deduce our results.

The seventh and final chapter is devoted in a discussion of our results, the possible impact that they could have in the scientific community and further work that could be done, using this thesis as a stepping stone towards reaching new heights.

Chapter 2. Basic clinical aspects

2.1 Cancer

Cancer is a puzzling and frightening set of diseases that have afflicted multicellular living beings for more than 200 million years, and there is evidence of cancers among ancestors of modern humans going back well over a million years. Unlike infectious diseases, parasites, and many environmental diseases, cancer is not primarily caused by some entity that is foreign to our bodies. Its agents of destruction are human cells that have, as it were, slipped their reins, and have been recruited and to some extent transformed into pathological organisms or the building blocks of tumors [13].

Understanding how cancer fundamentally works and what causes it (or them) is a very hard problem towards the "solution" of which a staggering amount of resources has been devoted, with mixed results. Some forms of cancer, e.g. childhood leukemia has seen its survival rate rise from 10% to 80%. However, there are other types for which we have made barely any progress if at all. There is no good answer to the question of whether it is one disease or many, and there is little point to asking the question. Cancer is a disease (or group of them) that involves abnormal cell growth with the potential to invade other parts of the body. It is considered a genetic disease, due to genes and more specifically their mutations playing an important role in developing cancer. Nevertheless, it is not still clear whether mutations are effects or causes of cancer, seeing as many external – environmental factors play decisive factor as well, by causing mutations or disrupting cellular mechanisms.

There is no easy part about cancer; it is almost impossible both to prevent and to cure. The complexity and diversity of cancer, occurring as it does in different organs and cell types with associated intratumor heterogeneity, implies the need for a multitude of tests for early detection coupled with treatments tailored to specific types of cancers. Successful cancer prevention is not a trivial challenge. It requires considerable commitment to implementation at national level through strategies that reach all segments of society. Solutions cannot be aimed only at individuals, but must be supported by legislative and regulatory measures. Some exposures, notably reduction in exposure to air pollution, require international agreements in order to be truly effective [14].

Nevertheless, there are some widely accepted factors regarding lifestyle choices that are linked to cancer and according general principles that minimize the risks as much as possible [15]:

- **Smoking:**

Cancer-causing substances in tobacco catalyze the formation of DNA adducts, subsequently resulting in the accumulation of somatic mutations. Since tobacco was established as a carcinogen linked to lung cancer in the 1950s, there has been a large body of evidence showing that tobacco use increases the risk of more than 15 types of cancer. The International Agency for Research on Cancer (IARC) also concludes that passive smoke is carcinogenic.

- **Physical activity:**

Numerous epidemiological studies are conducted each year to examine whether physical activity reduces the risk of several types of cancers. The evidence so far is strong for breast, colon, and endometrial cancers and limited for lung, liver, and esophageal cancers. The American Cancer Society recommends 150 – 300 minutes of moderate intensity physical activity or 75 – 150 minutes of vigorous intensity physical activity per week.

- **Diet:**

Diet is an important risk factor for cancer, having a role in energy balance and in other biological mechanisms independent of body weight. However, in this case it is more difficult to associate nutrition with cancer, most likely due to the long latency between exposure and outcome, the complexity of dietary components and nutrients, as well as the inevitable measurement errors. There are a lot of advice given by the World Cancer Research Fund (WCRF) in collaboration with the American Institute for Cancer Research (AICR) such as inclusion of greater dietary fiber and whole grain intake, following a high calcium diet and avoiding red meat. All of these are considered as probable steps that reduce the risk of colorectal cancer.

- **Alcohol consumption:**

According to WCRF alcohol consumption increases the risk of at least 6 types of cancer; esophagus, breast (which we will discuss further), colorectum, stomach, liver, and mouth, pharynx and larynx. Reducing alcohol consumption is listed as one of the World Health Organization (WHO) Best Buys for controlling noncommunicable diseases (NCDs).

The knowledge of the problems arising from the aforementioned behaviors is referred to as health literacy and raising awareness is considered one of the primary ways of handling the prevention of cancer. Added to these risk factors is the problem of air pollution; complex components of particulate matter exhibit high carcinogenic potential through several mechanisms. This is especially obvious in countries like China suffering from the worst air pollution. A report from the National Cancer Center of China attributed 14.4% of lung cancer deaths to particulate matter ($PM_{2.5}$) air pollution. The Global Burden of Disease (GBD) study estimated that the number of cancer deaths related to ambient particulate matter pollution increased by more than 300% from 1990 to 2017. Obviously, this is the toughest risk factor mentioned and dealing with it is much harder. Legislation needs to be passed concerning high pollution industries, promoting public transport etc.

The two figures below aid the visualization of all the aforementioned risk factors:

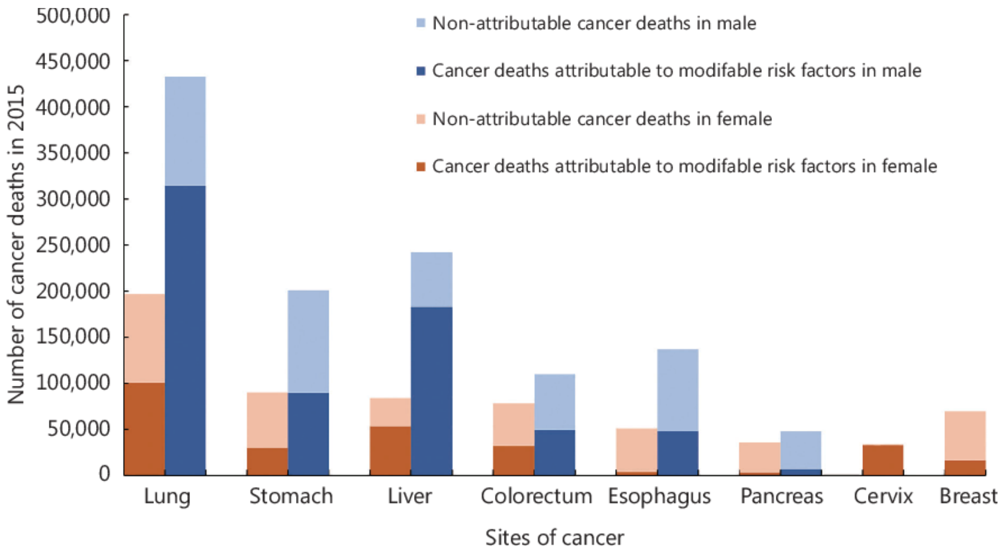
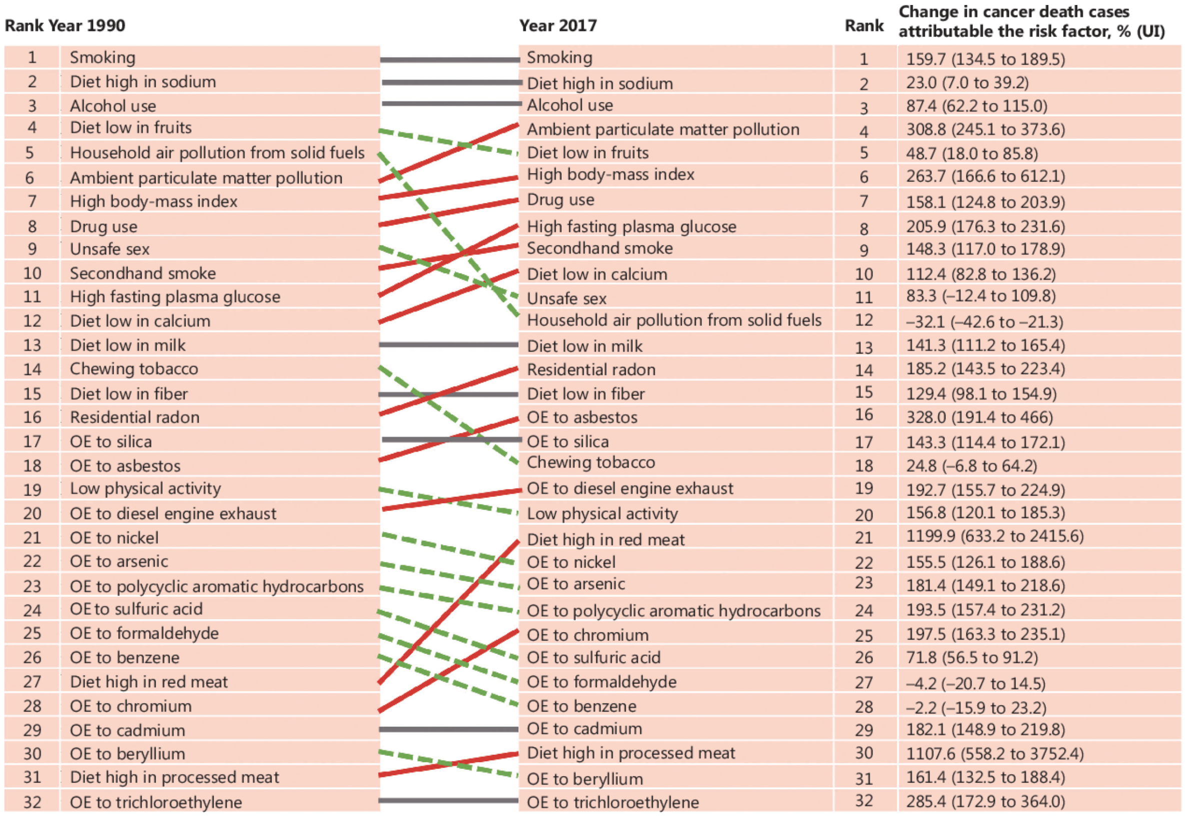


Figure 5: The proportion of cancer deaths that could be attributed to modifiable risk factors in common types of cancer deaths among males and females in China, from [15]



Rank	Year 1990	Year 2017	Rank	Change in cancer death cases attributable the risk factor, % (UI)
1	Smoking	Smoking	1	159.7 (134.5 to 189.5)
2	Diet high in sodium	Diet high in sodium	2	23.0 (7.0 to 39.2)
3	Alcohol use	Alcohol use	3	87.4 (62.2 to 115.0)
4	Diet low in fruits	Ambient particulate matter pollution	4	308.8 (245.1 to 373.6)
5	Household air pollution from solid fuels	Diet low in fruits	5	48.7 (18.0 to 85.8)
6	Ambient particulate matter pollution	High body-mass index	6	263.7 (166.6 to 612.1)
7	High body-mass index	Drug use	7	158.1 (124.8 to 203.9)
8	Drug use	High fasting plasma glucose	8	205.9 (176.3 to 231.6)
9	Unsafe sex	Secondhand smoke	9	148.3 (117.0 to 178.9)
10	Secondhand smoke	Diet low in calcium	10	112.4 (82.8 to 136.2)
11	High fasting plasma glucose	Unsafe sex	11	83.3 (–12.4 to 109.8)
12	Diet low in calcium	Household air pollution from solid fuels	12	–32.1 (–42.6 to –21.3)
13	Diet low in milk	Diet low in milk	13	141.3 (111.2 to 165.4)
14	Chewing tobacco	Residential radon	14	185.2 (143.5 to 223.4)
15	Diet low in fiber	Diet low in fiber	15	129.4 (98.1 to 154.9)
16	Residential radon	OE to asbestos	16	328.0 (191.4 to 466)
17	OE to silica	OE to silica	17	143.3 (114.4 to 172.1)
18	OE to asbestos	Chewing tobacco	18	24.8 (–6.8 to 64.2)
19	Low physical activity	OE to diesel engine exhaust	19	192.7 (155.7 to 224.9)
20	OE to diesel engine exhaust	Low physical activity	20	156.8 (120.1 to 185.3)
21	OE to nickel	Diet high in red meat	21	1199.9 (633.2 to 2415.6)
22	OE to arsenic	OE to nickel	22	155.5 (126.1 to 188.6)
23	OE to polycyclic aromatic hydrocarbons	OE to arsenic	23	181.4 (149.1 to 218.6)
24	OE to sulfuric acid	OE to polycyclic aromatic hydrocarbons	24	193.5 (157.4 to 231.2)
25	OE to formaldehyde	OE to chromium	25	197.5 (163.3 to 235.1)
26	OE to benzene	OE to sulfuric acid	26	71.8 (56.5 to 91.2)
27	Diet high in red meat	OE to formaldehyde	27	–4.2 (–20.7 to 14.5)
28	OE to chromium	OE to benzene	28	–2.2 (–15.9 to 23.2)
29	OE to cadmium	OE to cadmium	29	182.1 (148.9 to 219.8)
30	OE to beryllium	Diet high in processed meat	30	1107.6 (558.2 to 3752.4)
31	Diet high in processed meat	OE to beryllium	31	161.4 (132.5 to 188.4)
32	OE to trichloroethylene	OE to trichloroethylene	32	285.4 (172.9 to 364.0)

OE: Occupational exposure

Figure 6: Rank changes in cancer deaths attributable to 32 modifiable risk factors, and percentage changes in cancer deaths attributable to these risk factors in China from 1990 to 2017, from [15]

There are many types of cancer treatment the compatibility of which depend on the type of cancer and its stage. The most usual types include [16]:

- **surgery**, which attempts to completely remove the tumor and the affected cells
- **chemotherapy**, which employs the use of drugs to destroy cancer cells
- **radiation therapy**, which uses high-energy x-ray or other particles to destroy cancer cells
- **immunotherapy**, which uses substances made either by the body or in a laboratory in order to boost the immune system and help the body find and destroy the cancer cells itself
- **targeted therapy**, which uses drugs to target specific genes and proteins that help cancer cells survive and grow. This type of therapy targets either the tissue environment that cancer cells grow in or cells related to cancer growth, like blood vessels
- **hormone therapy**, which removes, blocks or adds specific hormones to the body

Some people will only get one treatment but most people have a combination of them, such as surgery with chemotherapy and/or radiation therapy [17].

2.2 Breast cancer

Breast cancer remains a worldwide public health dilemma and is currently the most common tumor in the globe. Breast cancer is life-threatening disease in females and the leading cause of mortality among women population. Amongst all the malignant diseases, breast cancer is considered as one of the leading causes of death in post-menopausal women accounting for 23% of all cancer deaths. There is a huge difference in breast cancer survival rates worldwide, with an estimated 5-year survival of 80% in developed countries to below 40% for developing countries. Developing countries face resource and infrastructure constraints that challenge the objective of improving breast cancer outcomes by timely recognition, diagnosis and management [18].

To further analyze the types of breast cancer a small reference must be done to the anatomy of the breast. Both males and females have breasts. The breast is made up of fatty tissue called adipose tissue. The female's breasts usually contain more glandular tissue than that of the males. Female breasts contain 12 – 20 lobes which are further divided into smaller lobules. These lobes and lobules are connected via milk ducts. The adipose tissue of the breast is supplied by a network of nerves, blood vessels, lymph vessels, lymph nodes, and is also composed of fibrous connective tissue and ligaments [18].

Breast cancer is divided into invasive and non-invasive types:

- **Non-invasive** means that the cancer has not spread further away from the lobule or ducts where it originated from. The most common examples are ductal carcinoma in situ and lobular carcinoma in situ (in situ = in place), which appear when atypical cells develop within the milk ducts or the breast lobules respectively, but have not extended to close proximity tissue or outside.

- **Invasive** means that the abnormal cells from within the lobules or the milk ducts split out into close proximity of breast tissue. Cancer cells can pass through the breast to different parts of the body through immune system or the systemic circulation. The most common organs to which these cells spread are brain, bones, lungs and liver. Once again, the most common examples are infiltrating ductal carcinoma and infiltrating lobular carcinoma.

Breast cancer is the second leading cause of cancer deaths among women. The development of breast cancer is a multi-step process involving multiple cell types, and its prevention remains challenging in the world. Early diagnosis of breast cancer is one of the best approaches to prevent this disease. In some developed countries, the 5-year relative survival rate of breast cancer patients is above 80% due to early prevention. In the recent decade, great progress has been made in the understanding of breast cancer as well as in the development of preventative methods. The pathogenesis and tumor drug-resistant mechanisms are revealed by discovering breast cancer stem cells, and many genes are found related to breast cancer [19].

The most important risk factors are considered to be, in decreasing importance:

1. **Aging**:

Breast cancer is highly related to increasing age. It is the leading cause of cancer death for women aged 20 to 59 years [20]. It is highly suggested for women over the age of 40 or older to have a mammography (an x-ray imaging method used to examine the breast for the early detection of cancer and other breast diseases).

2. **Family history**:

About a quarter of breast cancer cases tend to be related to family history. Women whose mother or sister had breast cancer are more likely to have breast cancer by a relative factor of 1.75 compared to other women, which increases to 2.5 to women with two or more first-degree relatives who had it [21].

3. **Reproductive factors**:

Reproductive factors such as early menarche, late menopause, late age at first pregnancy and low parity can increase the breast cancer risk. Each 1-year delay in menopause increases the risk of breast cancer by 3%. Each 1 – year delay in menarche or each additional birth decreases the risk of breast cancer by 5% or 10%, respectively [19].

4. **Estrogen**:

Both endogenous and exogenous estrogens are associated with the risk of breast cancer. The endogenous estrogen is usually produced by the ovary in premenopausal women and ovariectomy can reduce the risk of breast cancer [22]. The main sources of exogenous estrogen are the oral contraceptives and the hormone replacement therapy (HRT).

5. **Lifestyle**:

As mentioned in the above section, lifestyle choices are a major risk factor for every type of cancer. Following a proper and balanced lifestyle usually fends off successfully most types of cancer for a while. Likewise for breast cancer, excessive alcohol and too much dietary fat intake increases the risk of breast cancer. More specifically, alcohol consumption can elevate the level of estrogen-related hormones in the blood and trigger the estrogen receptor pathways. Although the relationship between smoking and breast cancer risk remains controversial, mutagens from cigarette smoke have been detected in the breast fluid from non-lactating women [19].

The only real prevention method is screening. Not primary tumors but the tumor metastasis causes over 90% of cancer deaths [23]. However, like in most types of cancer, if breast cancer is diagnosed as a primary tumor or at an early stage of metastasis, the breast tumor could be removed by surgery and the chemotherapy could work effectively. Mammography as mentioned before is also an effective screening method to obtain a good image of the state of the breasts.

The next important thing we should be discussing is the treatments that are considered the most effective right now, always taking into account the stage of the cancer:

- **Surgery**:

As should be expected, surgery is one of the most important ways to deal with breast cancer. Modified radical mastectomy has traditionally been the standard of care for early-stage invasive breast cancers. However, breast-conserving surgery has been favored more recently. This therapy involves removing the tumor without removing excess healthy breast tissue, with the outcome of a breast that is more aesthetically acceptable to the patient than the outcome from radical mastectomy [24].

- **Radiation therapy**:

Typically, whole-breast radiation is performed following breast-conserving surgery to treat subclinical disease. Radiation therapy is expensive and time-consuming, and shorter therapies can be more appealing, but five-year results appear favorable.

- **Chemotherapy**:

Chemotherapy is the standard of care for women with node-positive cancer or with a tumor larger than 1 cm. Factors such as age and comorbidities also influence the decision to use chemotherapy. A systematic review of 12 studies demonstrated disease free and overall survival advantages when using a taxane-containing regimen for pre-menopausal and post-menopausal women with early-stage breast cancer [25].

- **Endocrine therapy**:

Endocrine therapies such as Selective Estrogen Receptor Modulators (SERMs), aromatase inhibitors, and gonadotropin-releasing hormone antagonists, prevent estrogen production or block estrogen, thereby preventing stimulation of an estrogen-sensitive tumor. Endocrine therapy is not effective against cancers that are lacking hormone receptors [24]. Nevertheless, more than 70% of breast cancers are Estrogen Receptor (ER) – positive breast cancers and therefore the aforementioned drugs do a very effective job. SERMs are compounds that act as either agonists or antagonists of estrogen receptors. One of the most famous SERMs is tamoxifen (TAM), which has been used to treat breast cancer for more than 30 years [19].

- **Tissue-targeted therapy**:

Monoclonal antibodies like trastumuzab (Herceptin) improve disease-specific and overall survival when added to anthracyclines and paclitaxel (Taxol) chemotherapy in women with node-positive and high risk, node-negative breast cancers overexpressing HER2, a protein encoded by the ERBB2 gene [24].

Overall, the fact remains that breast cancer is one of the most common types of cancer that infest our modern era. The mortality of the disease is thankfully dropping, however the survivors need extra care to be able to readjust into our society smoothly.

2.3 Post-Traumatic Stress Disorder (PTSD)

Post-Traumatic Stress Disorder develops in some people who have experienced a shocking, scary, or dangerous event. It is natural to feel afraid during and after a traumatic situation. Fear is a part of the body's "fight-or-flight" response, which helps us avoid or respond to potential danger [26]. The most common situation is for people to gradually recover from their trauma and get over the initial symptoms, whichever they might be. However, there is a number of people who don't and continue to experience problems over longer periods of time and suffer from their trauma for prolonged time.

There are many traumatic events that may lead to PTSD, like physical or sexual assault, abuse, an accident or a disaster. Any harmful, mentally or physically, event and of course any life-threatening situation, as in our case a disease, can heavily impact the victim's psyche and cause PTSD. Anyone can develop PTSD, at any point in time, if there is enough traumatic stimulation and therefore it may affect their mental, physical, social and spiritual well-being.

Symptoms of PTSD usually fall into one of the following four categories:

- **Intrusion**:

Intrusive thoughts such as repeated, involuntary memories, distressing dreams or flashbacks of the traumatic event. In some occasions flashbacks may become so vivid, that people feel like they are reliving the traumatic experience all over again.

- **Avoidance**:

Avoiding any reminders of the traumatic event, such as people, places, objects and activities that may seem related to it. People actively avoid remembering the event and are hesitant to talk about it or their feelings concerning it.

- **Alterations in cognition and mood**:

Inability remembering important aspects of the traumatic event, negativity that leads to distorted thoughts about the cause and the consequences of the event and also about the way of viewing oneself or others leading to wrongful blaming and emotions such as ongoing fear, horror, anger, guilt or shame. Also, significant fall in general mental well-being, failing to enjoy previously pleasant activities and experience positive emotions.

- **Alterations in arousal and reactivity**:

Arousal and reactive symptoms may include being irritable and having angry outbursts; behaving recklessly or in a self-destructive way, being overly watchful of one's surroundings in a suspecting way, being easily startled, or having problems concentrating or sleeping [27].

Many people who are exposed to a traumatic event experience similar symptoms in the days following the event. For it to be considered as PTSD, symptoms should last for more than a month and must cause significant distress and problems in the person's everyday life. The symptoms may even occur later along the line and they could persist for months or even years. PTSD is closely related to other conditions such as depression, memory problems and other physical and mental health problems.

Not everyone who experiences a traumatic event develops PTSD and not everyone who suffers from PTSD needs psychiatric treatment. In fact, most people's symptoms subside over time with the help of the people's natural support system (family and friends). Nevertheless, there are some people that do need psychiatric help to escape from the psychological pit in which they have found themselves into. Psychiatrists and other mental health professionals use various effective (research-proven) methods to help people recover from PTSD. Both talk therapy (psychotherapy) and medication provide effective evidence-based treatments for PTSD. One category of psychotherapy, cognitive behavior therapies (CBT), in particular is very effective. Cognitive processing therapy, prolonged exposure therapy and stress inoculation therapy are only some of the types of CBT used to treat PTSD. As for medication, some antidepressants such as SSRIs and SNRIs (selective serotonin re-uptake inhibitors and serotonin-norepinephrine re-uptake inhibitors), are commonly used to treat the core symptoms of PTSD. They are used either alone or in combination with psychotherapy. Other medications may be used to lower anxiety and physical agitation, or treat the nightmares and sleep problems that trouble many people with PTSD [27].

PTSD was considered the label we wanted to predict in this thesis, for patients that suffered from breast cancer. Since we intended to use supervised learning, we first needed a way to assess PTSD symptoms. Although there are several screening tools available for use in assessing PTSD symptoms, the PTSD CheckList (PCL) is the most widely used self-report assessment instrument of PTSD symptoms. More specifically, we used the PTSD checklist for DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, 5th edition), which is a 20-item self-report measure that assesses the presence and severity of PTSD symptoms.

Respondents are asked to rate how bothered they have been by each of 20 items in the past month on a **5-point Likert scale** ranging from 0-4. **Items are summed to provide a total severity score (range = 0-80)**:

- 0 = Not at all
- 1 = A little bit
- 2 = Moderately
- 3 = Quite a bit
- 4 = Extremely

DSM-5 symptom cluster severity scores can be obtained by summing the scores for the items within a given cluster, i.e., cluster B (items 1-5), cluster C (items 6-7), cluster D (items 8-14), and cluster E (items 15-20). A provisional PTSD diagnosis can be made by treating each item rated as 2 = "Moderately" or higher as a symptom endorsed, then following the DSM-5 diagnostic rule which requires at least: 1 B item (questions 1-5), 1 C item (questions 6-7), 2 D items (questions 8-14), 2 E items (questions 15-20). Preliminary validation work is sufficient to make initial cut-point score suggestions, but this information may be subject to change. **A total score of 33 or higher suggests the patient may benefit from PTSD treatment** and is the cutoff we used in this thesis to determine whether a patient has PTSD symptoms.

2.4 HADS – Hospital Anxiety and Depression Scale

Another important questionnaire used in this thesis is HADS. There is a need to assess the contribution of mood disorder, especially anxiety and depression, in order to understand the experience of suffering in the setting of medical practice [28]. HADS is a self-assessment scale that has been developed and found to be a reliable instrument for detecting states of depression and anxiety in the setting of a hospital or a medical clinic.

Depression level is assessed according to the questions: "Do you take as much interest in things as you used to? Do you laugh as readily? Do you feel cheerful? Do you feel optimistic about the future?". **Anxiety level** is assessed by the questions: "Do you feel tense and wound up? Do you worry a lot? Do you have panic attacks? Do you feel something awful is about to happen?". The questionnaire responses are analysed in the light of the results of this estimation of the **severity of both anxiety and of depression**. This enabled a reduction of the number of items in the questionnaire to just seven reflecting anxiety and seven reflecting depression (of the seven depression items five reflected aspects of reduction in pleasure response) within the BOUNCE project. Each item is answered by the patient on a four-point (0–3) response category so the possible scores ranged from 0 to 21 for anxiety and 0 to 21 for depression. A score of 0 to 7 for either subscale could be regarded as being in the normal range, a score of 11 or higher indicating probable presence ('caseness') of the mood disorder and a score of 8 to 10 being just suggestive of the presence of the respective state. Further work done within the BOUNCE project indicated that the two subscales, anxiety and depression, were independent measures. Our expectation is that the results of these scales, which will be used as features in our models, will prove very important in the final prediction.

Chapter 3. Data-analysis aspects

3.1 Introduction

Machine learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through learning from data, without being explicitly programmed. In other words, machine learning allows computers to automatically learn and make predictions or decisions based on patterns and information which is inherently found in the data. The main idea is that for more complex problems, it is easier for us to create a model that **learns** how to solve them based on existing data, than creating a model with specific instructions-steps to solve them. Therefore, machine learning algorithms are the easy choice when the development of a specific algorithm that solves the problem is very hard or impossible, but are catching on to simpler everyday problems as well due to their efficiency. Machine learning has a wide range of applications, including natural language processing, computer vision, recommendation systems, autonomous vehicles, and more. It has become an essential tool in various industries for solving complex problems and making data-driven decisions.

3.2 Machine learning categories

Just like people can learn in many ways, the same principle applies to computers. The developed algorithms allow the successful handling of many different problems in many different fields.

3.2.1 Supervised learning

Supervised learning is a fundamental and widely used approach in machine learning, and it has numerous real-world applications where making predictions or classifications based on available data is crucial. The model is trained on labeled data, with a clear input-output relationship. The principle is simple; the model is fed data that is already correctly classified and tries to adjust its internal hyperparameters to find the best possible mapping from inputs to outputs. The mathematical goal is to minimize the difference between the labels and the predicted outputs, in order to better classify unseen similar data in the future. The model takes the input data and produces an output based on the learned relationship.

There are two types of supervised learning:

- <u>Classification</u>:

  In these tasks, the goal is to categorize input data into predefined classes or categories. For example, classifying emails as spam or not spam, or recognizing digits in handwritten digits recognition.

- Regression:

  In these tasks, the goal is to predict a continuous numeric value. For example, we might want to predict house prices based on features like square footage, number of bedrooms or maybe today's temperature based on features like the month, the day's prevailing winds.

To assess the performance of a supervised learning model, it is typically evaluated on a separate dataset (the test set) that was not used during training. Common evaluation metrics for classification tasks include:

- *Accuracy:*

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

It works well only if there are equal number of samples belonging to each class, otherwise it is easy to be misleaded.

- *Precision:*

This metric attempts to answer the question "what proportion of positive identifications was actually correct".

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

where:

TP = True Positives, meaning the samples that where correctly identified as belonging to the positive class

NP = Negative Positives, meaning the samples that were incorrectly identified as belonging to the positive class

and similarly

TN = True Negatives, meaning the samples that where correctly identified as belonging to the negative class

FN = False Negatives, meaning the samples that were incorrectly identified as belonging to the negative class

- *Recall*:

This metric attempts to answer the question “what proportion of actual positives was identified correctly”.

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

We can observe that, by definition, recall seems like a fitting metric for medical problems such as ours, since maximizing it would mean that we succeeded in finding as many positive cases (e.g. PTSD victims) as possible. However, this is not actually true, because we could achieve a perfect recall of 1.0 even if we predicted all the cases as positive. Nevertheless, it is fairly obvious that such a model would be useless. That is why recall and precision are naïve metrics and should be considered together to be able to make a decision. In practice, they are rarely used as standalone metrics.

- *F_β-score*:

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

A factor indicating how much more important recall is than precision. For example, if we consider recall to be twice as important as precision, we can set β = 2. The standard F-score is equivalent to setting β = 1 and is the harmonic mean between precision and recall. F-score is a very common and useful metric for many problems. In medical fields, it might be smarter to use the F2 score, to give more importance towards the recall as explained above.

- *Area Under Curve (AUC)*:

A Receiver Operating Characteristic Curve (ROC) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

  - True Positive Rate (TPR), synonym of recall:

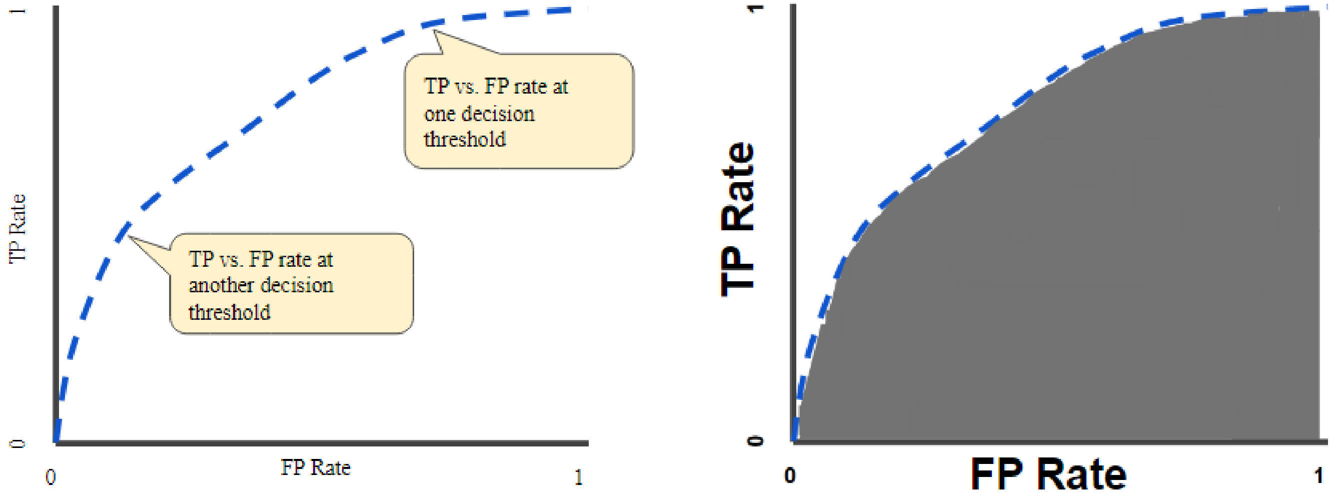
$$TPR = \frac{TP}{TP + FN}$$

  - False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

An ROC plots TPR vs FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both FP and TP.

Figure 7: Example of ROC curve and AUC



AUC measures the entire two-dimensional area underneath the entire ROC curve. It provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0, whereas one whose predictions are 100% correct has an AUC of 1.0.

AUC is a good metric because it is scale-invariant, which means it measures how well predictions are ranked, rather than their absolute values. Furthermore, it is classification-threshold-invariant, which means it measures the quality of the model’s predictions irrespective of what classification threshold is chosen [29]. The intuition behind ROC AUC is that it measures how well a binary classifier can distinguish or separate between the positive and negative classes. It reflects the probability that the model will **correctly rank** a randomly chosen positive instance **higher** than a random negative one.

In regression problems, the most commonly used metrics are:

- *Mean Absolute Error (MAE)*:

$$Mean\ Absolute\ Error = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j|$$

where:

N = number of samples

y = ground truth

$\hat{y}$ = prediction

It gives us the measure of how far the predictions were from the actual output. However, they don't give us any idea of the direction of the error i.e., whether we are under predicting the data or over predicting the data.

- *Mean Squared Error (MSE):*

$$Mean\ Squared\ Error = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - \widehat{y_j}\right)^2$$

It is quite similar to MAE, the only difference being that MSE takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient. Also, as we take the square of the error, the effects of larger errors become more pronounced than smaller errors, hence the model can now focus more on the larger errors.

3.2.2 Unsupervised learning

Unsupervised machine learning is a type of machine learning where the algorithm is trained on unlabeled data, meaning the data does not have predefined categories or target labels. In contrast to supervised learning, where the algorithm learns to make predictions based on labeled examples, unsupervised learning seeks to find patterns, relationships, or structures within the data without explicit guidance. The primary goal of unsupervised learning is to discover the underlying structure or distribution in the data. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition. This can involve tasks such as:

- **Clustering**:

Clustering is a data mining technique which groups unlabeled data based on their similarities or differences. The algorithm does this without prior knowledge of what the clusters should represent. K-means clustering and hierarchical clustering are common examples of unsupervised clustering techniques.

- **Dimensionality reduction**:

Dimensionality reduction techniques aim to reduce the number of features (dimensions) in the data while preserving its important characteristics. Principal Component Analysis (PCA) and t-SNE (t-Distributed Stochastic Neighbor Embedding) are examples of unsupervised dimensionality reduction methods.

- **Density estimation**:

Density estimation methods attempt to model the underlying probability distribution of the data. Gaussian Mixture Models (GMMs) and kernel density estimation are examples of techniques used for density estimation.

Unsupervised learning has various real-world applications, such as customer segmentation in marketing, anomaly detection in cybersecurity, topic modeling in natural language processing, and image compression in computer vision. It's especially useful when you want to explore and understand the inherent structure or patterns within your data, even when you don't have labeled examples to train a supervised model.

3.2.3 Reinforcement learning

Reinforcement Learning (RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences. It is similar to supervised learning in the sense that it also uses mapping between input and output, but unlike supervised learning, where the feedback provided is the correct label for the task, reinforcement learning uses rewards and punishments as signals for positive and negative behavior. It differs mainly in respect to goals. While the goal in supervised learning is to find similarities and differences between data points, in the case of reinforcement learning the goal is to find a suitable action model that would maximize the **total cumulative reward** of the agent. The figure below illustrates the **action-reward feedback loop** of a generic RL model:

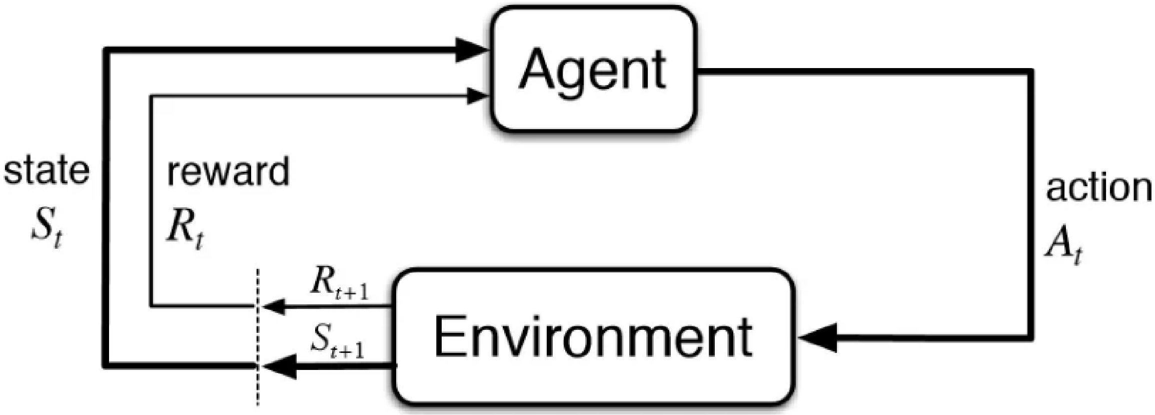


Figure 8: Action-reward feedback loop, from [30]

Here are some key components and concepts of reinforcement learning:

- *Agent*:

The learning algorithm or system that makes decisions and interacts with the environment.

- *Environment:*

The external system or context in which the agent operates. The environment responds to the actions taken by the agent and provides feedback in the form of rewards and state changes.

- *State*:

A representation of the current situation or configuration of the environment that the agent can observe. States can be discrete or continuous, depending on the problem.

- *Action*:

The move or decision that the agent makes at each time step. Actions can also be discrete or continuous.

- *Policy*:

The strategy or mapping from states to actions that the agent uses to make decisions. The policy defines the agent's behavior.

- *Reward*:

A scalar value that the agent receives from the environment after taking an action in a particular state. The goal of the agent is to maximize the cumulative reward over time.

- *Value function*:

The expected cumulative reward an agent can achieve starting from a specific state and following a given policy. It quantifies the goodness of being in a particular state.

- *Q-function*:

Similar to the value function, but it considers both a specific state and a specific action. The Q-function measures the expected cumulative reward of taking a specific action in a specific state and then following a policy.

Reinforcement learning algorithms typically involve the agent taking actions, receiving rewards, and updating its policy or strategy to improve its decision-making over time. The agent aims to learn an optimal policy or set of actions that maximize the expected cumulative reward. RL has found applications in a wide range of fields, including robotics, autonomous systems, game playing (e.g., AlphaGo), recommendation systems, and natural language processing. It's particularly useful in situations where the optimal decision-making strategy is not known in advance, and the agent must learn through interaction with its environment.

3.3 Classifiers

3.3.1 Decision tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. It is a supervised machine learning algorithm that is primarily used in data mining and data analysis. Decision trees are constructed by recursively partitioning the dataset into subsets based on the values of input features, ultimately leading to a decision or prediction.

It is very natural and intuitive to classify a pattern through a sequence of questions, in which the next question asked depends on the answer to the current question. Such a sequence of questions is displayed in a directed decision tree or simply tree, where by convention the first or **root node** is displayed at the top, connected by successive (directional) links or **branches** to other nodes. These are similarly connected until we reach terminal or **leaf** nodes, which have no further links. The classification of a particular pattern begins at the root node, which asks for the value of a particular property of the pattern. The different links from the root node correspond to the different possible values. Based on the answer we follow the appropriate link to a subsequent or descendent node. It is most common that the links are mutually distinct and exhaustive, i.e., one and only one link will be followed. The next step is to make the decision at the appropriate subsequent node, which can be considered the root of a sub-tree. We continue this way until we reach a leaf node, which has no further question. Each leaf node bears a category label and the test pattern is assigned the category of the leaf node reached [31]. Below we can observe an example decision tree:

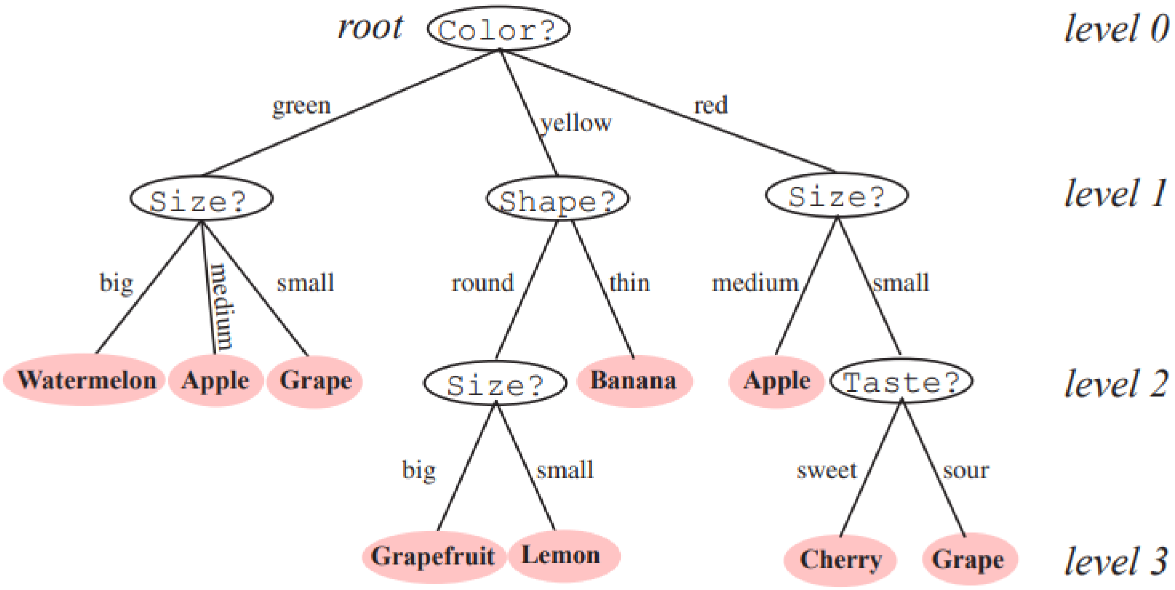


Figure 9: Example decision tree, from [31]

It would be ideal if all the samples in each subset had the same category label. In that case, we would say that each subset was pure, and could terminate that portion of the tree. Usually, however, there is a mixture of labels in each subset, and thus for each branch we will have to decide either to stop splitting and accept an imperfect decision, or instead select another property and grow the tree further. Some well-known decision tree algorithms include Iterative Dichotomiser 3 (ID3), a classic algorithm that uses entropy and information gain for splitting decisions, C4.5 (an extension of ID3 that can handle both categorical and continuous data) and CART, A versatile decision tree algorithm that can be used for both classification and regression tasks.

Classification And Regression Trees (CART) is a general framework that can be instantiated in various ways to produce different decision trees. First of all, concerning the number of splits per node, we can use any number, but because of the universal expressive power of binary trees and the comparative simplicity in training, we choose two splits per node. The fundamental principle underlying tree creation is that of simplicity: we prefer decisions that lead to a simple, compact tree with few nodes. To this end, we seek a property test T at each node N that makes the data reaching the immediate descendent nodes as "pure" as possible. In formalizing this notion, it turns out to be more convenient to define the impurity, rather than the purity of a node. Several different mathematical measures of impurity have been proposed, all of which have basically the same behavior. Let i(N) denote the impurity of a node N. In all cases, we want i(N) to be 0 if all of the patterns that reach the node bear the same category label, and to be large if the categories are equally represented. The most popular measures are:

- **Entropy impurity**:

$$i(N) = -\sum_{j} P(\omega_j) * \log_2 P(\omega_j)$$

where $P(\omega_j)$ is the fraction of patterns at node N that are in category ω_j

By the well-known properties of entropy, if all the patterns are of the same category, the impurity is 0; otherwise it is positive, with the greatest value occuring when the different classes are equally likely.

- **Gini impurity**:

$$i(N) = 1 - \sum_{j} P^2(\omega_j)$$

This is just the expected error rate at node N if the category label is selected randomly from the class distribution present at N. This criterion is more strongly peaked at equal probabilities than is the entropy impurity.

In CART also arises the problem of deciding when to stop splitting during the training of a binary tree. If we continue to grow the tree fully until each leaf node corresponds to the lowest impurity, then the data has typically been overfit. Conversely, if splitting is stopped too early, then the error on the training data is not sufficiently low and hence performance may suffer. A common method is to set a (small) threshold value in the reduction in impurity; splitting is stopped if the best candidate split at a node reduces the impurity by less than that pre-set amount. We can also stop splitting when a node represents fewer samples than a small percentage of the data, e.g. 5%.

Decision trees have advantages such as simplicity and interpretability, but they can be prone to overfitting, especially when they grow too deep. To mitigate this issue, pruning techniques and ensemble methods are often used in conjunction with decision trees to improve their predictive performance.

3.3.2 Random forest

Random forest, first suggested by Leo Breiman [32] is an ensemble learning technique used in machine learning for both classification and regression tasks. It is a powerful and versatile algorithm that improves the performance and robustness of decision trees. Instead of relying on a single decision tree, a Random Forest combines the predictions of multiple decision trees to make more accurate and stable predictions, following the ideas of ***bagging***:

1. We start by creating multiple random subsets of the original dataset, called “bootstrapped” datasets. In this not all samples are included and some samples can be included more than once.
2. For each bootstrapped dataset, we construct a decision tree, without pruning it, which means we allow it to overfit. The key difference in creating this decision tree is that we consider only a random subset of features at each node. This introduces an element of randomness and decorrelates the trees, making them more diverse.
3. After constructing all the decision trees, each tree makes a prediction on the test data. In the case of classification, each tree "votes" for a class, and in regression, each tree provides a numeric prediction. The final prediction is typically determined by majority voting (for classification) or averaging (for regression) across all the trees.

Random forests have a unique way of identifying how well they were trained. Every time we create a bootstrapped dataset, we inevitably leave some samples out of this dataset. These are called the out-of-bag samples. We can use these samples to check the robustness and accuracy of the random forest we trained. For a single sample, we aggregate the decisions of every tree for which it is considered an out-of-bag sample and see if we made a correct prediction or not. Doing this for every out-of-bag sample, we compute the **out-of-bag accuracy** of our random forest.

The main advantages of random forests are:

- **Improved accuracy**:

Random Forests generally provide better predictive accuracy than individual decision trees, especially when the dataset is noisy or complex.

- **Reduced Overfitting**:

By averaging predictions from multiple trees and introducing randomness in feature selection, Random Forests are less prone to overfitting compared to single decision trees.

- **Robustness**:

They are robust to outliers and can handle missing values without a lot of data preprocessing.

- **Variable Importance**:

Random Forests can measure the importance of each feature in the classification or regression task, which can be useful for feature selection.

- **Parallelization**:

Training and prediction in Random Forests can be easily parallelized, making them suitable for large datasets and distributed computing environments.

Random Forests are widely used in various applications, including image classification, natural language processing, finance, and bioinformatics. They are considered one of the most effective and versatile machine learning algorithms available and are often a first choice for many real-world predictive modeling tasks.

3.3.3 Adaboost

The idea of ***boosting*** is that instead of creating all the classifiers that we mean to aggregate at the same time, like in bagging, we create them serially. We create the first classifier, e.g. a small (not overfit) decision tree as best as we can and then we use the deviation between its output and the desired results to create the best “additional” classifier. That means that the second classifier’s main focus is to deliver better results on the samples that the first classifier failed to, the third one’s on the samples that the first two classifiers failed and so on.

Adaboost, short for adaptive boosting, is another ensemble learning technique used in machine learning for classification and regression tasks first introduced in [33]. The main steps of adaboost are:

1. **Initialization:**

Initially, each data sample in the training set is assigned an equal weight. These weights represent the importance of each sample.

2. **Weak learner training**:

We start by training a weak learner (e.g., a decision tree with limited depth) on the training data. The weak learner is trained to minimize the classification error, but it doesn't have to be highly accurate on its own.

3. **Weighted error**:

After training the weak learner, AdaBoost calculates the weighted error rate of the model. This error rate is based on how well the model performs on the training data, with more weight given to data points that were misclassified. At the first iteration, as mentioned during the initialization, all samples are considered equally important.

4. **Update weights**:

AdaBoost adjusts the weights of the data points to focus more on the misclassified data points. Data points that were misclassified by the weak learner are assigned higher weights, while correctly classified data points are assigned lower weights.

5. **Repeat:**

Steps 2 to 4 are repeated for a predetermined number of iterations or until a certain level of accuracy is achieved. In each iteration, a new weak learner is trained on the data with updated weights.

6. **Combine Weak Learners:**

AdaBoost combines the predictions of all the weak learners by assigning a weight to each learner based on its accuracy. More accurate learners are given higher weights, and less accurate learners are given lower weights.

7. **Final prediction**:

To make predictions on new data, AdaBoost combines the weighted predictions of all the weak learners. In classification tasks, it uses a weighted majority vote, and in regression tasks, it uses a weighted average.

Adaboost as a classifier has several advantages including:

- **Improved accuracy**:

It usually achieves higher accuracy compared to using an individual decision tree.

- **Automatic feature selection:**

AdaBoost can implicitly select important features by assigning higher weights to them.

- **Simplicity**:

Creating these weak learners doesn’t require a lot of resources but can still achieve a very strong performance.

- **Avoids overfitting**:

Adaboost is less prone to overfitting compared to training a simpler model, such as a decision tree.

3.3.4 Extreme gradient boosting

Extreme Gradient Boosting (XGB) is, as indicated by the name, another **boosting** algorithm that is known for its speed and performance in both classification and regression tasks. It was designed in order to enhance the performance of the classic boosting algorithm and is also leverages decision trees as weak learners.

The idea of the classic gradient boosting algorithm is similar to adaboost, albeit a little different. Gradient boosting creates a tree that makes a simple prediction at first and then tries to create another one that optimally predicts the pseudo-residuals, which are the deviations of the predictions from the correct values (either we are talking about regression or classification). Let's take a better look at the exact steps it takes:

1. **Initialization:**

Gradient boosting starts with an initial prediction, often a simple one, which can be a constant value (e.g., the mean of the target variable for regression tasks) or a default class label (e.g., the majority class for classification tasks).

2. **Building weak learners:**

A weak learner, such as a decision tree with limited depth (also known as a "stump"), is trained on the dataset to predict the residuals or errors of the initial prediction. In other words, the weak learner focuses on what the current model gets wrong.

3. **Weighted addition of models:**

The predictions from the weak learner are scaled by a small learning rate (also known as a "shrinkage" factor) and added to the current model. This addition updates the current model to make it better at capturing the data's patterns.

4. **Update residuals:**

The residuals (the differences between the actual target values and the current model's predictions) are updated based on the new predictions. The next weak learner is then trained on these updated residuals.

5. **Repeat:**

Steps 2 to 4 are repeated for a predefined number of iterations (boosting rounds) or until a stopping criterion is met. In each iteration, a new weak learner is trained to improve the model's predictions.

6. **Final prediction**:

The final prediction is made by aggregating the predictions of all the weak learners.

XGBoost is an advanced and highly optimized implementation of gradient boosting. While XGBoost shares the fundamental concept of gradient boosting it incorporates several key differences and optimizations:

- **Regularization techniques**:

XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms in its objective function. This helps prevent overfitting by penalizing complex models.

- **Parallelization and speed:**

XGBoost is designed for efficiency and speed. It utilizes techniques like parallelization during tree construction, which makes it significantly faster than classic gradient boosting, especially for large datasets.

- **Handling missing values:**

XGBoost has built-in capabilities to handle missing values in the dataset. It uses a technique called "Sparsity Aware Split Finding" to efficiently handle missing data, which can be a challenging aspect in traditional gradient boosting.

- **Integration with GPU**:

XGBoost can be easily integrated with GPUs, taking advantage of their parallel processing capabilities to further speed up training.

In summary, XGBoost builds upon the principles of classic gradient boosting but introduces several optimizations and features that make it faster, more accurate, and easier to use for a wide range of machine learning tasks. Its combination of regularization techniques, parallelization, and customizable features has made it a popular choice in many real-world applications.

3.3.5 Support Vector Machines

Support Vector Machines (SVMs) is a powerful machine learning algorithm used for classification and regression tasks. They are particularly well-suited for binary classification problems, but can also be extended to handle multi-class classification. SVMs aim to find the optimal hyperplane that best separates data points belonging to different classes while maximizing the margin between these classes.

Here are the key concepts and components of Support Vector Machines:

- **Hyperplane:**

Hyperplane is a decision boundary that separates data points into two classes. For a two-dimensional dataset, a hyperplane is a straight line, but in higher dimensions, it becomes a hyperplane.

- **Margin:**

Margin is the distance between the hyperplane and the nearest data points from each class. SVMs' goal is to maximize the margin because a larger margin implies a better generalization to unseen data and a lower risk of overfitting.

- **Support vectors:**

They are the closest data points to the deciding hyperplane and as such the most critical in defining the margin. These points influence the position and orientation of the hyperplane.

- **Kernel**:

SVMs can use a kernel function to map data points into higher-dimensional space, where they might (with the proper kernel function) become linearly separable, even if they were not in the original feature space. Common kernel functions include linear, polynomial, radial basis function (RBF) and sigmoid kernels.

- **C Parameter (Regularization)**:

The C parameter in SVM is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors. A smaller C emphasizes a larger margin (potentially more errors), while a larger C allows for fewer errors (potentially a smaller margin). The choice of C influences the model's bias-variance trade-off.

SVMs come with many advantages:

- Effective in high-dimensional spaces
- Versatile due to different kernel functions for handling non-linear data
- Good generalization performance when the margin is well-defined
- Robust to overfitting, especially when using a soft-margin approach (C-parameter)

And also with some disadvantages:

- Computationally expensive for large datasets with a lot of features
- Sensitivity to the choice of hyperparameters, especially the kernel and C parameter
- Don't inherently provide probability estimates

SVMs have been widely used in various applications, including image classification, text classification, bioinformatics, and more. Their ability to find optimal hyperplanes in complex feature spaces makes them a valuable tool in machine learning.

Chapter 4. Related work

4.1 Predicting psychological consequences of breast cancer

Šimunović and Ljubotina in [34] used sociodemographic, medical, religious and self-perception psychological data to try and discern patterns that relate to PTSD symptoms in Croatian patients who had been treated following breast cancer diagnosis and were in remission at the time of the study. They identified PTSD symptoms on a patient using the PCL-C (civilian version), because the most recent PCL-5 does not have an official cutoff. The criteria they set for a patient to be identified as suffering from PTSD were:

- PCL-C score above the official cutoff of 50 as recommended in validation studies of PCL-C scale on samples of cancer patients [35]
- Cluster's criterion; at least one symptom of cluster B, at least three symptoms of cluster C and at least two symptoms of cluster D

They also used a customized questionnaire to determine the stress scale of the patients.

In order to predict PTSD symptom severity, they calculated bivariate correlation between all the features and the PCL-C total scores to select the most optimal potential predictors. Categorical variables where one-hot coded to be integrated in the above procedure. In the end, hierarchical regression analysis was performed using the selected variables to predict PTSD symptom severity.

Gibbons et al in [36] used sociodemographic, medical data, as well as self-assessment questionnaires towards predicting anxiety, depression and overall cancer-related distress. The data was gathered using a national breast cancer screening service in a large university affiliated hospital serving a large geographical area in Ireland. They used the HADS scale to measure anxiety and depression and a questionnaire adapted from previous research to measure cancer-related distress. They used Pearson Product Moment correlations (statistical method) to examine the relationships between the predictors and outcome variables, as well as to identify medical and demographic factors to control for in the regressions. They also conducted hierarchical multiple regressions to examine the influence of illness perceptions and coping on cancer-related distress, anxiety and depression in women with breast cancer. In conclusion, this study did not used medical variables to predict the three labels discussed. It used illness perception to try and predict them, because individuals do not always have adequate knowledge of the medical indices of their disease, and hence these variables would not then necessarily predict psychological adjustment.

Tsaras et al in [37] also used sociodemographic and medical data, in order to predict anxiety and depression in breast cancer patients, in an oncology public hospital of Greece. They also used the Patient Health Questionnaire-2 (PHQ-2) and the Generalized Anxiety Disorder-2 (GAD-2) questionnaire as brief screening instruments for depression and anxiety. The first goal was to identify the correlation between the variables and the outcomes (depression and anxiety) by using univariate logistic regression. Afterwards, the selected variables were included in a multivariate logistic regression model in order to further identify the best predictors of depression and anxiety risk.

Perez-Tejada et al in [38] focused their research in using monoamine levels to predict anxiety and depression in female breast cancer survivors recruited from the Onkologikoa Fundazioa Hospital in Spain. More specifically, they used high-performance liquid chromatography (HPLC) to analyze the patients' blood and determine the levels of dopamine (DA), noradrenaline (NA), serotonin (5-HT) and kynurenine (KYN). Consequently, they used regression analysis to study the correlation between the predictors and anxiety/depression which were measured using the HADS.

Lam et al in [39] attempted to predict satisfaction levels amongst Chinese women suffering from advanced breast cancer, while awaiting or receiving chemotherapy. The labels considering at baseline were:

- health system and information unmet (HSI) needs
- psychological distress
- physical symptom distress
- patient satisfaction

Patient satisfaction was measured again at 1.5-, 3-, 6- and 12-months post-baseline. They used latent growth curve (LGC) analysis for assessing the change of patient satisfaction over the 12 months' follow-up. Several variables, including health system and information needs and physical symptom distress consistently predict subsequent psychosocial distress. Consequently, they used hierarchical multiple regression analysis to examine if baseline health system information needs, physical symptom distress, anxiety and depression also predicted patient satisfaction one-year post-baseline.

4.2 Machine learning in similar problems

Kalafi et al in [40] used machine learning and deep learning to predict Malaysian breast cancer survivors. The features consisted of demographic and clinical characteristics on which data preprocessing steps were used; samples with missing values were dropped, the most frequent label (0 = alive) was downsampled to match the least frequent and a random forest model was fit to determine which features are the most important towards predicting patient survivability. The models used to predict survivability included decision tree, random forest, SVM, and Multi-Layer Perceptron (MLP). The best results came from MLP, random forest, decision tree and SVM in that order.

Ge et al in [41] experimented towards predicting the psychological state of Chinese undergraduate students during the COVID-19 pandemic using machine learning. The data included sociodemographic information and a mental health questionnaire (College Students Mental Health Screening Scale). The General Anxiety Disorder-7 was used as an anxiety measure and the Insomnia Severity Index-7 as an insomnia measure. An important note is that the data was gathered at the time of enrollment, but was used for predictions during the COVID outbreak – at a later date, making this a longitudinal study. XGBoost was used to predict probable anxiety and insomnia, with the metrics being used included the area under the receiver-operation curve (AUC), sensitivity (recall), specificity and accuracy. XGBoost coefficient were also used to assess positive or negative relation between predictors and labels.

According to Montazeri et al in [42] machine learning has commonly been used to try and predict schizophrenia and bipolar disorders. The most used algorithms in these studies are SVM, random forest (RF), gradient boosting (GB), logistic regression and decision trees in that order. RFs algorithms demonstrated significantly higher accuracy and sensitivity than SVM and GB. GB demonstrated significantly higher specificity than SVM and RF.

4.3 Combining above ideas

Reviewing the aforementioned work, we can come to the conclusion that relatively to our problem in hand, we only found one research in [34] trying to predict PTSD symptoms of female breast cancer patients. Similar searches considering anxiety, depression and satisfaction levels among breast cancer patients instead, have tried to focus mainly on identifying the most correlated predictor variables. Wherever predictions were made, they were usually made using relatively simple regression models after previously choosing the best predictors. There is no research that is dedicated solely on predicting the PTSD symptoms of breast cancer patients, as we intend to do. Just like in [39], we will also use anxiety and depression measured using questionnaires as part of our predictors and we also expect them to be among the most important ones, being very closely associated to the PTSD symptoms we attempt to predict.

Also, a very important part of the novelty of this thesis comes from the heterogeneity of the data. Our data were collected in four different oncology centers – pilots: IEO (Milan, Italy), HUS (Helsinki, Finland), HUJI (Jerusalem, Israel) and Champalimaud (Lisbon, Portugal). No other study included data from multiple geographical locations so far from each other. We will also try to discern how well our predictions can generalize to unseen data from other regions by using the "leave-one-out" method; we will train our models on the data from 3 out of 4 oncology centers and then use them to predict the data from the 4th one (repeating the process for all the centers as test sets). Furthermore, our data includes sociodemographic, medical, psychosocial and lifestyle variables to be used as predictors, which cover every possible aspect, compared to previous analyses that focused only on sociodemographic and medical ones.

Finally, concerning our models, we intend to try out decision tree, random forest, adaboost, xgboost and SVM. As shown above, most of the related work that has gone into predicting similar psychological labels, has tested out successfully decision trees, random forest, SVM and gradient boosting. We also include adaboost in our line-up of models which similarly to gradient boosting is derived from the logic of boosting trees and could prove to be a valuable tool. More about how each of these algorithms works have already been mentioned in paragraph 3.3.

Chapter 5. Methodology

5.1 Introduction

The methodology we use for this thesis is pretty straightforward. The goal is to use the data gathered by the BOUNCE associates across the many months that the patients were tracked in order to predict whether the patients are susceptible to PTSD symptoms and to do so as soon as possible. Towards this goal, we use the data gathered during the months M0 and M3, where M0 is considered the baseline; the month when the diagnosis of the breast cancer happened and possible surgery to try and deal with it. The label considered is PTSD symptoms at M6. To do so, we first preprocess the data to get them in uniform and in proper state to be fed to the models that we are going to use. Afterwards, we use the machine learning algorithms we selected as the most appropriate for our problem on our preprocessed data. Finally, we evaluated the models' results comparing them and selecting the best one.

5.2 Data preprocessing

The most important part of every machine learning solution is the data. Most of the work towards a successful result is usually done before feeding the data in any model. The preprocessing steps we follow are:

0. Eliminating some features that the medical experts hinted (optional step)
1. Eliminating samples with no label
2. Eliminating features with too many missing values across samples
3. Eliminating samples with too many missing features
4. Eliminating near-zero variance features
5. Eliminating highly-correlated features
6. Splitting the dataset into train/test
7. Imputing the missing values of our two datasets
8. Using feature reduction methods to lower the dimensionality of our high-dimensional data
9. One-hot encoding the categorical features to be able to use them in the models that cannot handle them otherwise
10. Normalizing the features when needed (depends on the model)
11. Balance the training data

5.3 Machine learning models

After cleaning up and preprocessing the data we are ready to feed them to our machine learning models. The models we decided to try out and compare are:

- **Decision tree:**

Decision trees are a great baseline model. They tend to overfit, but there are many ways of pruning them or early stopping to deal with their high variance problem, resulting in a model with pretty good performance. It also works with all types of data, so things like one-hot encoding are not necessary. They are pretty fast, but also easy to interpret.

- **Random forest:**

Random forest is the obvious next step and the result of using the bagging method on decision trees. Instead of a big, somewhat overfit decision tree, we create many smaller ones, taking into account a subset of features each time, which generally leads to better generalization to unknown data which is always the most important thing when training models.

- **Adaboost**:

Adaboost is an algorithm based on an idea to improve random forest. It also uses many weak learners, like small decision trees (stumps), but creates subsequent ones based on the errors of previous ones.

- **XGBoost:**

XGBoost is another algorithm focused on improving the ideas of random forest. It also constructs many weak learners, like stumps, with the idea that every subsequent one aims to reduce the errors that the previous ones failed to.

- **Support vector machines**:

Support vector machines are considered because it is a well-researched and straightforward machine learning algorithm. It is also effective in high dimensional spaces; that means it is still effective in cases where the number of dimensions is greater than the number of samples, as happens in our case.

- **Voting classifier:**

We also try using a voting classifier: having an odd number of classifiers we use is an ideal circumstance to check the performance of a hard voting classifier that can possibly enhance our results.

More advantages of these machine learning algorithms were discussed in section 3.3, in which we analyzed more in-depth the way they work, as well as their advantages and disadvantages.

The model training involves running a grid-search on the important hyperparameters of each model, while performing a repeated cross-validation on the training data, to decide on the optimal hyperparameters of the model (algorithm) for our problem.

5.4 Results evaluation

After doing all the experiments using the aforementioned models, we have to deduce our results. First, we compare our models' performance on the test set using our selected metric, f2 score, to choose the best overall model. We try to explain the results and also employ the assistance of explainable AI towards that goal. In the end, we decide on the optimal hyperparameters, preprocessing procedure and best models.

Chapter 6. Experimental procedure and Results

The experimental analysis was conducted using the programming language R and all the code used to create the following results can be found on my GitHub repository. During the explanation, after explaining the variables, we will be referring to them the way they are in the code, in order to be easy to associate the explanation to the corresponding part.

6.1 Data provision

The aim of the BOUNCE project is to take into consideration heterogeneous multi-scale data gathered at four oncology centers – pilots: IEO (Milan, Italy), HUS (Helsinki, Finland), HUJI (Jerusalem, Israel) and Champalimaud (Lisbon, Portugal). These data can be categorized as:

- **clinical/biological/genetic:**

Genetic risk factors, epidemiological factors, type and timing of treatment and medication, patient-reported symptoms, tumor biology and type, basic laboratory tests, no. of visits to various carers and emergency units, survival, etc.

- **psychosocial:**

Life events, and stressors, health-related Quality of Life, perceived social support, counselling and support sessions received, Depression, Coping (CERQ) Flexibility and Posttraumatic Growth, Distress Thermometer, etc.

- **socio-demographic:**

Age, gender, family history, family status, working status, level of education, insurance status, absence from work, no. of disability pensions, etc.

- **lifestyle:**

Alcohol consumption, smoking (past or current), physical exercise, etc.

Furthermore, additional data were collected from HUS patients through Noona, a self-monitoring tool for cancer patients, mainly focusing on patient-reported information: pain, fatigue and weakness, changes in mood or emotions, stomach and bowel symptoms, respiratory symptoms, reduced muscle strength or numbness in legs, mental performance, changes in general state of health and many more.

The aforementioned data were gathered by the members of the BOUNCE project and formulated the dataset that was used in this diploma thesis. The heterogeneity of these data is clear and is part of the novelty this thesis has to offer.

6.2 Data analysis

The specific features (and labels) that we used to carry out our experiments are all included in the appendix at the end of the thesis.

6.3 Data preprocessing

The preprocessing of the data is usually the most important and the hardest part of training a model. The reason it is the most important is that every model has its own weaknesses and without good data, even the best model can't learn much. The reason it is the hardest is that every dataset has different features that need unique care. Nevertheless, in order to have a fair comparison between different classifiers we enacted a unified preprocessing procedure to all our data, keeping in mind that the result should be usable by all our chosen classifiers.

Initial observations:

The initial dataset has shape (excluding the label):

732 rows x 177 columns

Note: *Rows correspond to samples and columns to features and we will refer to them interchangeably in the rest of the analysis.*

That means we have a relatively small number of samples, which is also pretty imbalanced due to the data's nature (as most medical data, labels 1 are very scarce). Also, we observe a pretty high dimensionality of features which is generally bad when training machine learning models, many of which suffer from the known "curse of dimensionality". The aforementioned observations incline us to be a bit careful about dropping more samples, but allows us to be less reserved when dropping features.

The preprocessing steps we followed to clean the data for the training were:

0. Eliminating some features that the medical experts hinted: (optional step, second experiment)

It was noted by the medical experts that some features have very high logical correlation with some other, more prominent ones and the existence of all of them would more likely confuse the model, than assist its learning. The names of those features are:

- Emot_Fun_QLQ30.0
- Emot_Fun_QLQ30.3
- nccn_distress_thermometer.0
- nccn_distress_thermometer.3

This step had to be conducted at the very beginning, so as not to let the features we are going to remove interact with the other features during the rest of our preprocessing, since we are going to remove them in the end.

1. Eliminating samples with no label:

The label we considered was based on the PTSD PCL-5 questionnaire for the month M6. Although one of the most important problems of our data is their imbalance, we have no choice but to drop the samples that have no label. The number of unusable samples is 154, bringing our dataset down to 578 samples.

2. Eliminating features with too many missing values across samples:

Features that have a very high number of missing values usually do not offer much information and it is better to remove them and drop the dimensionality along the way. Literature does not offer a standard optimal threshold for such a procedure, as everything is data dependent and some very important features might be useful even with very high missing value percentage. Therefore, we decided to use *missing_samples_threshold* = 0.25, which means we allow up to ¼ of the data to be missing until we decide to drop the feature – column. We used a relatively small threshold since the dimensionality of our data is already pretty high and, as mentioned earlier, we would not mind reducing them a bit.

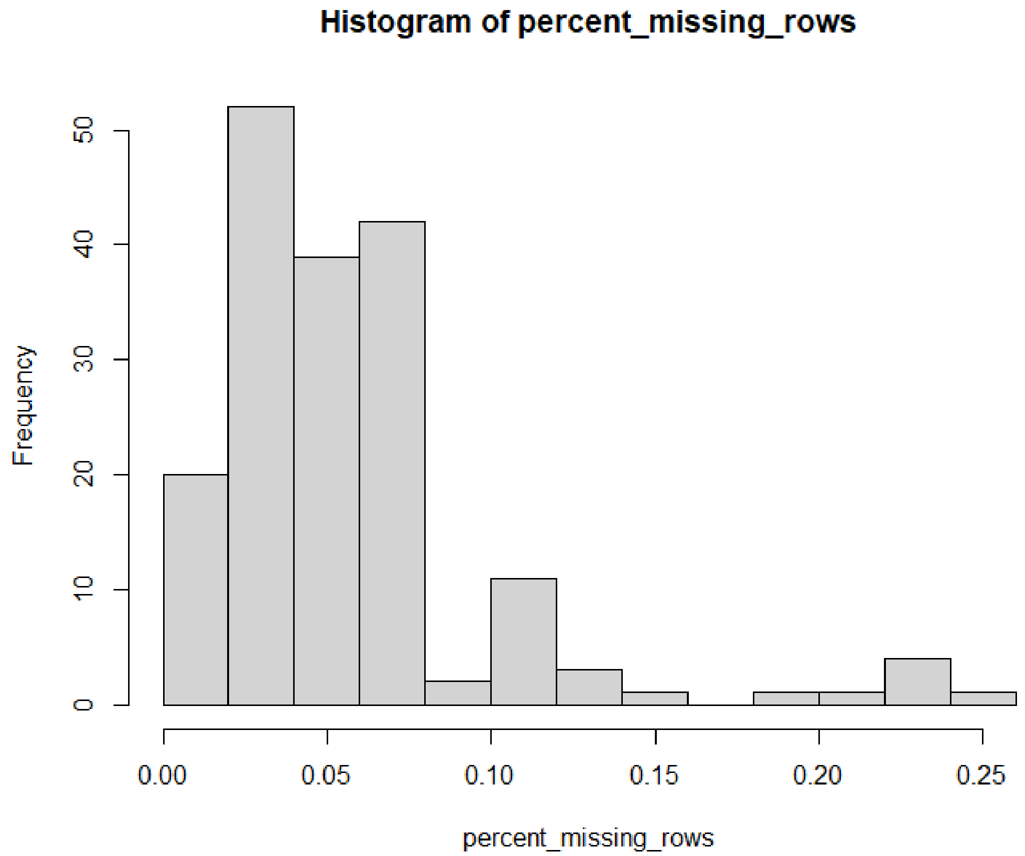


Figure 10: Histogram of percentage of missing values per column

However, observing the resulting histogram, we see that our chosen threshold leads to no columns being dropped.

3. Eliminating samples with too many missing features:

The idea is the same as in step 2; samples with too many missing values – features will most likely be of little use to us. Nevertheless, as we mentioned earlier, we do not wish to reduce our samples even more so a more conservative *missing_features_threshold* = 0.5 was chosen, which means we allow up to ½ (half) of the columns to be missing before deciding to drop a sample.

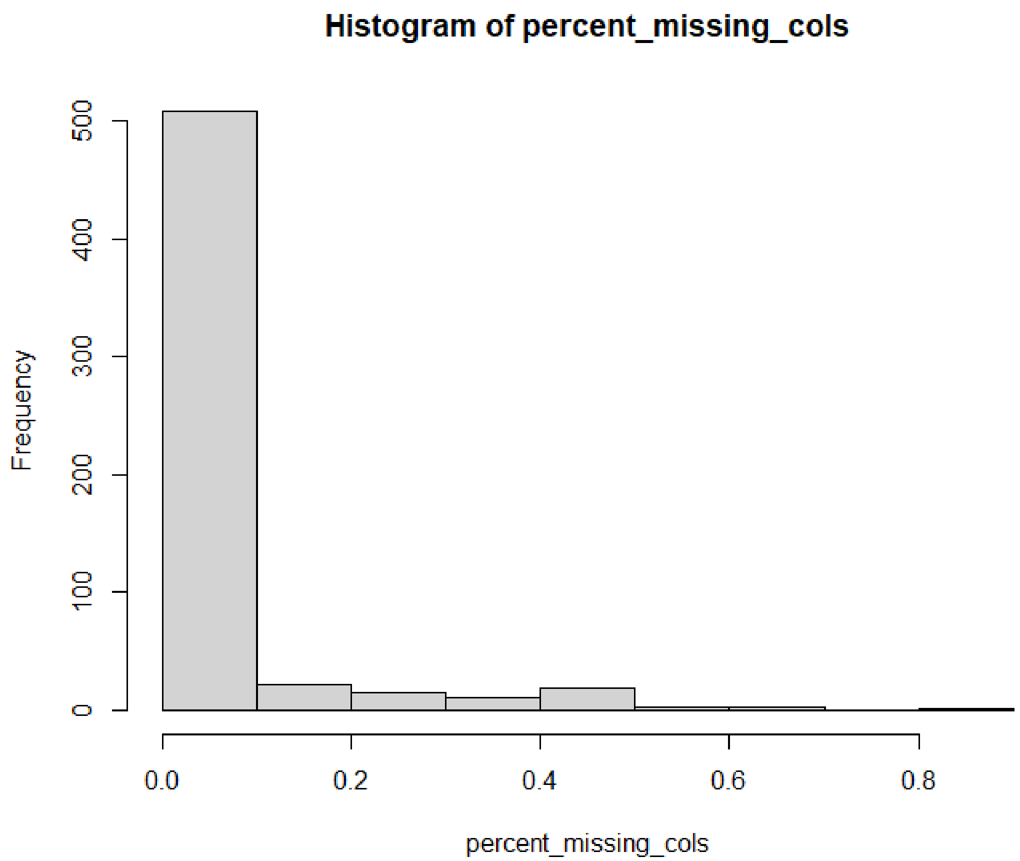


Figure 11: Histogram of percentage of missing values per row

Observing the resulting histogram, we can see that only a very small number of samples have more than 50% missing values and need to be dropped. The exact number is 5.

4. Eliminating near-zero variance features:

Ideally, we want to eliminate all "low" variance features. Numeric features have ranges of values that could differ very much. For example, feature A could range from 0.1 to 1, while feature B could range from 10,000 to 100,000. There is not a rigorous method to determine if a feature has a "low" variance. Comparison with other features is difficult due to the different scales, excluding the scenarios with literally zero variance (feature is constant), which is a pathological behavior that is rarely met in practice. Categorical features, however, are easier to filter out in case of near-zero variance. We filter out the categorical variables if both of the following criteria are met:

- The ratio of the most common value to the second most common is higher than *freqCut*
- The ratio of the unique values to the total number of samples is lower than *uniqueCut*

The combination of the above two criteria makes sure that the variance of the categorical feature is "near zero", which means it carries little information. For our case, we chose *freqCut* = 95/5 and *uniqueCut* = 10.

5. Eliminating highly-correlated features:

If two features are very highly correlated, they will obviously impart nearly the same information to our model. Including both, however, won't increment additional information, but will actually weaken the model, because it infuses it with noise. We choose Spearman correlation, which is a non-parametric measure that assesses the strength and direction of a monotonic relationship between two variables. It does not assume linearity and is appropriate for both continuous and ordinal data, which are included in our problem. It is also less sensitive to outliers compared to Pearson correlation (the classic linearity correlation). Using a relatively conservative *cor_threshold* = 0.8, we find the following correlation pairs and eliminate one of the two (the first one in this case accordingly):

Table 5: Correlated features

Dropped	**Kept**
M3_mMOS_social_support_total	M3_mMOS_instumental_support
baseline_neutrofiles	baseline_leukocytes
M0_Flexibility_PACT	M0_Forward_PACT
Sex_Enjoy_BR23.0	Sex_Funct_BR23.0
Sex_Enjoy_BR23.3	Sex_Funct_BR23.3

We can easily observe even based on just the names that the dropped features do seem to have obvious logical correlation with their pairs we kept. For the exact features' meaning, see the appendix.

***Note**:*

The next step in the preprocessing step is to impute the missing values, as many models can't handle their existence. However, it is incorrect to use all the data towards this purpose, as doing that would leak information from the test set. The idea of the test set is to be used as sample "unseen" data, which means that it should only be used to evaluate the performance of a concrete (with set hyperparameters) model. Therefore, they cannot participate in the imputation process and that's why at this point we have to split the dataset.

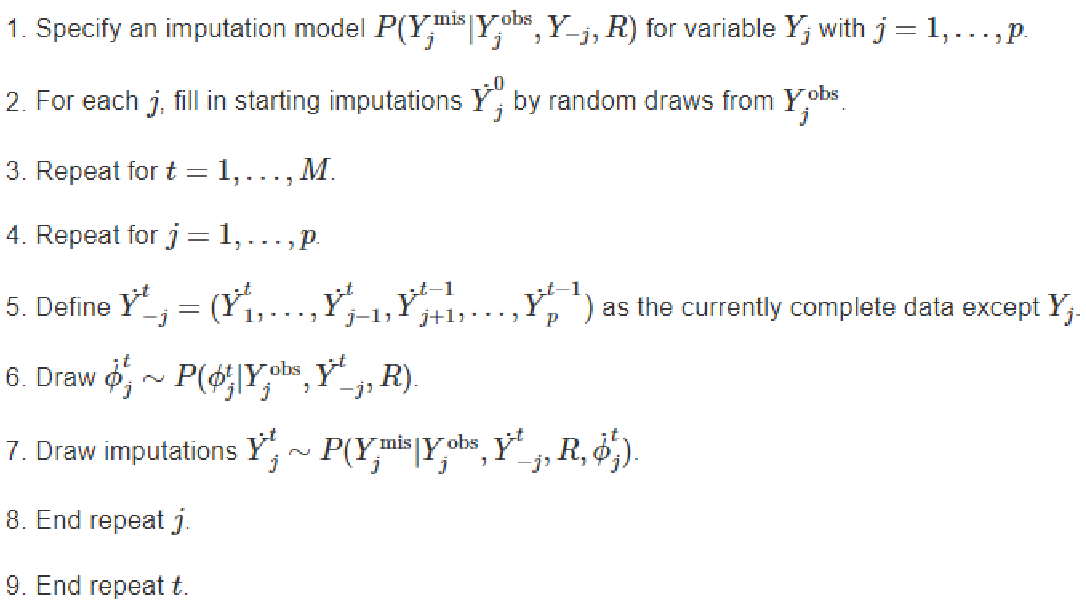
6. Splitting the dataset into train/test:

We will try two different approaches while splitting the data:

- Split the data randomly with *train_size* = 0.7. The threshold seemed like an easy decision since a 60-40 split would leave very few samples in the train set, whereas an 80-20 split would leave very few samples in the test set and even fewer positive samples, which would result in vast differences in the performance of different runs, as even one more positive sample correctly classified would have a very big effect on our metrics.
- Split the dataset using the data collected by one of the four oncology centers – pilots as the testing data and the rest as training data. This approach aims to discern whether our models can generalize and handle well "local" data which may vary from the ones they have been trained with. We look into this approach later, in the fourth experiment.

7. Imputing the missing values of our two datasets:

In order to impute the dataset's missing values, we used the Multiple Imputation Chained Equations (MICE) algorithm and the corresponding R package. The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data. In addition, MICE can impute continuous two-level data, and maintain consistency between imputations by means of passive imputation [43]. The exact mathematical steps of the MICE imputation can be observed in the figure below:



```
1. Specify an imputation model $P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, Y_{-j}, R)$ for variable $Y_j$ with $j = 1, \ldots, p$.
2. For each $j$, fill in starting imputations $\dot{Y}_j^0$ by random draws from $Y_j^{\text{obs}}$.
3. Repeat for $t = 1, \ldots, M$.
4. Repeat for $j = 1, \ldots, p$.
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ as the currently complete data except $Y_j$.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat $j$.
9. End repeat $t$.
```

Figure 12: MICE algorithm, from [44]

The algorithm starts with a random draw from the observed data, and imputes the incomplete data in a variable-by-variable fashion. One iteration consists of one cycle through all Y_j.

The number of iterations M = 5 was chosen. That means we generate 5 different imputed datasets. We treat each of the imputed datasets as an approximation of the “real” dataset which would have no missing values. Hence, since we cannot obtain the real dataset, we use an imputation method to get as close as possible, but there will always be some divergence – noise in the predictions. By imputing many different datasets, using our classifiers on all of them and aggregating the results, we make sure that the final classifier that we choose will be more robust to that said “noise”. Therefore, the following steps from now on are executed for every one of the 5 imputed datasets separately.

8. Using feature reduction methods to lower the dimensionality of our high-dimensional data (optional):

The main reasons we want to reduce the dimensionality of our features are:

- **Computational efficiency**:

Training models with fewer features is computationally less intensive, which can lead to faster training times and lower resource requirements, making it more practical for real-time or large-scale applications.

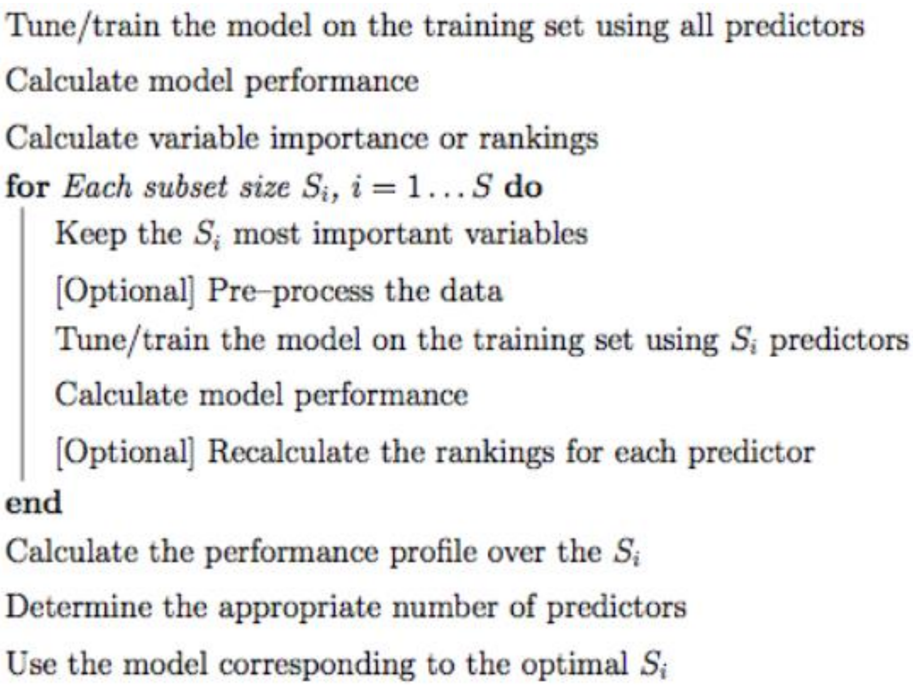
- **Easier data gathering**:

Especially in our context of gathering data in hospital environments, many potential problems could arise in gathering the data; anxious patients could very likely not be willing to complete a ton of questionnaires and other surveys or feel like they don’t want to answer some particular ones for their personal reasons. Being able to make a good prediction with fewer features is a much-wanted property of our modeling analysis.

- **Improved performance**:

Sometimes really high-dimensional data can lead to overfitting, where a model learns noise in the data rather than the true underlying patterns. Feature reduction can sometimes reduce the risk of overfitting by focusing the model on the most relevant information.

More specifically, the feature reduction method we used is **recursive feature elimination** (RFE). The following figure shows the basic RFE algorithm:

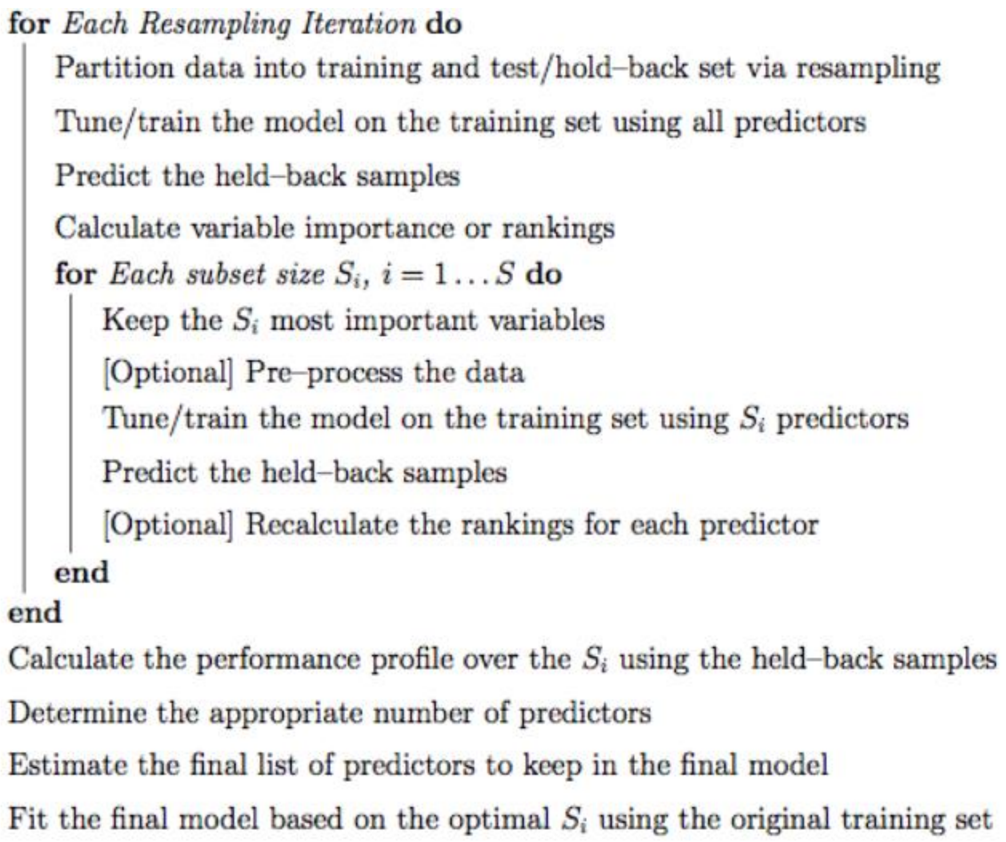


```
Tune/train the model on the training set using all predictors
Calculate model performance
Calculate variable importance or rankings
for Each subset size $S_i$, $i = 1 \ldots S$ do
    Keep the $S_i$ most important variables
    [Optional] Pre–process the data
    Tune/train the model on the training set using $S_i$ predictors
    Calculate model performance
    [Optional] Recalculate the rankings for each predictor
end
Calculate the performance profile over the $S_i$
Determine the appropriate number of predictors
Use the model corresponding to the optimal $S_i$
```

Figure 13: RFE algorithm, from [45]

The idea is that we use a model that can inherently calculate variable importance rankings. Therefore, we calculate model performance and variable importance. Then, for the selected list of feature sizes S_i we want to check, we keep the S_i top-ranked features and use them to fit the model and calculate model performance.

Since feature selection is part of the model building process, resampling methods (e.g. cross-validation, the bootstrap) should factor in the variability caused by feature selection when calculating performance. To get performance estimates that incorporate the variation due to feature selection, it is suggested that the steps in the previous algorithm be encapsulated inside an outer layer of resampling:



```
for *Each Resampling Iteration* do
    Partition data into training and test/hold–back set via resampling
    Tune/train the model on the training set using all predictors
    Predict the held–back samples
    Calculate variable importance or rankings
    for *Each subset size* $S_i$, $i = 1 \ldots S$ do
        Keep the $S_i$ most important variables
        [Optional] Pre–process the data
        Tune/train the model on the training set using $S_i$ predictors
        Predict the held–back samples
        [Optional] Recalculate the rankings for each predictor
    end
end
Calculate the performance profile over the $S_i$ using the held–back samples
Determine the appropriate number of predictors
Estimate the final list of predictors to keep in the final model
Fit the final model based on the optimal $S_i$ using the original training set
```

Figure 14: RFE algorithm with resampling, from [45]

This does provide better estimates of performance, but it is more computationally burdensome. Another complication to using resampling is that multiple lists of the "best" predictors are generated at each iteration. At first this may seem like a disadvantage, but it does provide a more probabilistic assessment of predictor importance than a ranking based on a single fixed data set. At the end of the algorithm, a consensus ranking can be used to determine the best predictors to retain [45].

In our experiments, we used the second version of the algorithm, including cross-validation in the process and more specifically repeated k-fold cross-validation. That means we split the training dataset in k folds and use one as the validation set each time, for k times. Then we repeat the process until we believe that the result has converged. We used **5-fold** cross-validation and repeated it **10 times**. The model we fit to follow the procedure was **random forest**, which is the most common algorithm used for this procedure.

The metric tried to maximize during our RFE, as well as during the model evaluation later is the **f2-score**, which was judged to be the most appropriate for our problem. The AUC is also a good metric and as we will see in almost all cases follows the same results as those of the f2. Those metrics were further analyzed here.

We also used a tolerance parameter for the results of the model; if the usage of fewer predictors has lower performance than the best one achieved (using more predictors) within a tolerance margin, then we pick the fewer predictors. This small twist allows us to deal with the case in which using the full number of features proves to be the best option (which is a possible result), supposing there is not a significant drop in the next best performance.

The next figure shows an example RFE result for one of the imputed datasets:

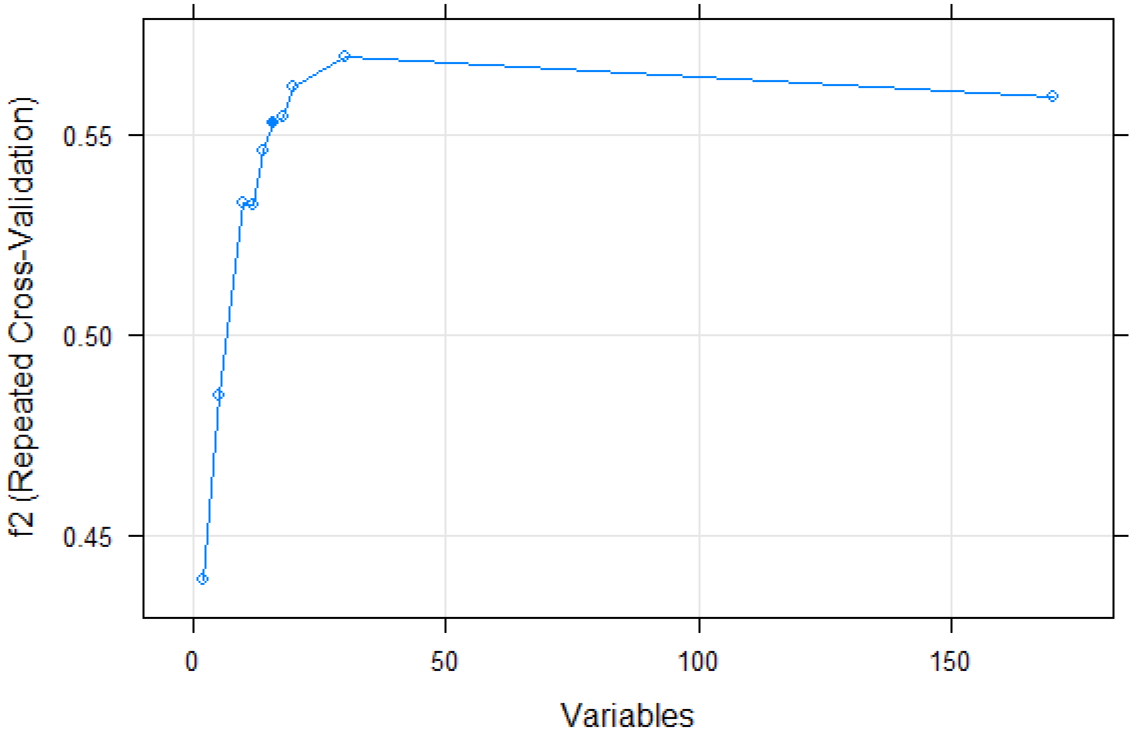


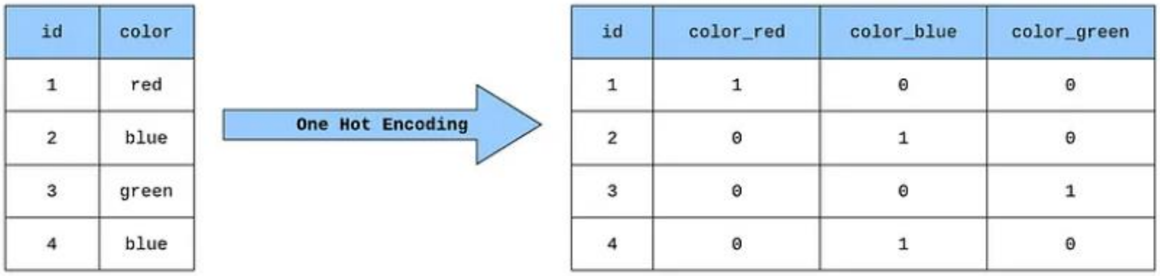
Figure 15: RFE results

We can observe that there are four different numbers of features that outperformed our chosen one, but because of the tolerance we just mentioned, we choose 16 as our optimal number instead of 30. The tolerance cutoff we chose was *rfe_tol* = 3.5 as a percentage or 0.035 in our scale.

In the third experiment, we use both the original features and the reduced ones to determine if and how much the feature reduction affected our performance.

9. One-hot encoding the categorical features to be able to use them in the models that cannot handle them otherwise:

Some models like the SVM cannot handle non-numeric data. For those cases, we use the one-hot encoding/transformation to create numeric columns:



id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Figure 16: Example of one-hot encoding, from [46]

A categorical feature with N different values is encoded into N different features each taking values only in {0, 1}. It is obvious that only one of the new features can be 1.

For models that need one-hot encoding the data beforehand, we used the one-hot transformation on the original data. For the reduced features experiment, we don't use RFE on the already one-hot encoded data; instead, we take the reduced features and only if they contain categorical data, we use the one-hot transformation on them.

10. Normalizing the features when needed (depends on the model):

The only model of the ones we have chosen that really needs to have normalized features is SVM. Since we intend to use Euclidean distance for measuring the distance of points in the SVM algorithm, by not normalizing the data we essentially give some features with larger scales more importance than those with smaller ones.

All the other models are tree-based which means the decisions at any node are taken by only considering a single feature and seeing how much the question in hand (e.g. age > 17 or male = YES) lowers the weighted sum of the impurities of the child nodes. Further discussion about the impurities of trees was conducted here, in paragraph 3.3.1.

11. Balance the training data:

The last, but also one of the most important steps of our preprocessing is using a sampling method to balanced our extremely imbalanced dataset. Out of our 573 samples, only 43 of them are positive (label = 1), which is roughly 7.5%.

We try out four different sampling approaches (as part of our first experiment):

- **No sampling**:

See how the models perform on the original unbalanced dataset.

- **Downsampling:**

Using this method, we keep all the samples from the minority class (1) and pick at random an even number of samples from the majority class (0).

- **Oversampling:**

Inversely, we keep all the samples from the majority class (0) and pick at random samples from the minority class (1) to include them more than once in the final training dataset, until we achieve an even number of samples from the two classes.

- **ROSE:**

ROSE uses smoothed bootstrapping to draw artificial samples from the feature space neighborhood around the minority class. It is a smarter method of oversampling, by creating synthetic data.

All these sampling methods are coupled inside our repeated cross-validation logic described earlier. During every resampling iteration, which means every repeat of every k-fold split, the relevant training set is sampled using one of the aforementioned methods, so that the classes are balanced and the classifier learns to focus on both classes similarly.

6.4 Modeling and results

6.4.1 Classifiers’ discovery spaces

As mentioned before, the classifiers we tried out in our experiments are decision tree, random forest, adaboost, xgboost and SVM. In the cases of xgboost and SVM that cannot handle categorical variables, we used the one-hot encoded features.

Every classifier has its own distinct hyperparameters that need to be tuned, according to the training data. We tried out every combination of the said hyperparameters by using grid search repeated cross-validation; for every element of our grid of hyperparameters (every combination) – which means we have a concrete model, we evaluate the model’s performance using repeated k-fold cross validation and choose the combination with the highest performance. In our experiments, we used **5 folds** and **5 repeats**. The below tables show the hyperparameter spaces we explored:

Table 6: Decision tree hyperparameter space

Complexity parameter (CP)	{0.001, 0.005, 0.01, 0.05, 0.1}

Complexity parameter is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue. We could also say that tree construction does not continue unless it would decrease the overall lack of fit by a factor of cp. It is the only tuning parameter the R caret package we used allows us to experiment with.

Table 7: Random forest hyperparameter space

mtry	**All features** {default – 10, default – 5, default, default + 5, default + 10, default + 20}
	Reduced features {default, default + 2, default + 4, default + 6}

where

$$default = \sqrt{\#features}$$

Mtry is the total number of features that our random forest considers at any split to best perform the said split. It is the only tuning parameter the R caret package we used allows us to experiment with. Experiments have shown that the best value for this variable, considered as **default** nowadays, is the square root of the total number of features. Therefore, we tried some numbers around it to maybe fine-tune this parameter a little bit for our problem.

Table 8: SVM hyperparameter space

C	{0.1, 0.5, 1, 2, 5, 10}
sigma	{0.1, 0.5, 1, 2, 5}

C: the cost of constraints violation. The C-constant of the regularization term in the Lagrange formulation. Defaults to 1.

sigma: inverse kernel width for the Radial Basis kernel function

Table 9: Adaboost hyperparameter space

mfinal	{25, 50, 100}
maxdepth	{1, 3, 5}
coeflearn	Breiman

mfinal: an integer, the number of iterations for which boosting is run or the number of trees to use. Defaults to 100.

maxdepth: controls the depth of the trees we create in the process. In adaboost, weak learners are optimal, so we decided to keep them relatively small.

coeflearn: controls the weight updating coefficient alpha. Defaults to Breiman.

Table 10: XGBoost hyperparameter space

nrounds	{25, 50, 100}
eta	{0.05, 0.1, 0.3}
gamma	0
max_depth	{4, 6, 8}
min_child_weight	1
subsample	{0.8, 1}
colsample_bytree	{0.8, 1}

nrounds: the maximum number of trees to create

eta: also known as learning_rate. Step size shrinkage used in update to prevent overfitting.

gamma: also known as min_split_loss. Minimum loss reduction required to make a further partition on a leaf node of the tree. Defaults to 0, which means we don’t use this criterion to suppress the tree.

max_depth: maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. Defaults to 6.

min_child_weight: minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. Defaults to 1.

colsample_bytree: the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed. Defaults to 1, which means we consider all the features at every tree.

subsample: subsample ratio of the training instances. Defaults to 1, which means we use all our data at every boosting iteration.

We also included a hard-voting classifier: the prediction is the most common prediction among the rest of our classifiers (taking advantage of the fact that we have an odd number of classifiers).

6.4.2 Models' results

After choosing the best concrete model (with set hyperparameters) using the above procedure, we evaluate its performance on the test set, which is used to simulate unknown real-world data. We have a concrete model for every different imputed dataset as we already mentioned. The figures we show below are the aggregate results when taking into consideration all the imputed datasets for each classifier.

We can observe that both metrics considered follow similar patterns (as we will notice for all of the graphs). Nevertheless, we use f2 to rank the best classifier, since we also tried to maximize f2 while training.

Experiment 1: Sampling method

The goal of this experiment is to try out the different sampling methods discussed in the final preprocessing step and see how much they impact our results.

- No sampling:

Table 11: Classifier performance without sampling

Classifier	F2	AUC
Decision tree	0.110	0.532
Random forest	**0.190**	0.570
SVM	0.000	0.500
Adaboost	0.124	0.534
XGBoost	0.128	0.544
Voting	0.098	0.536

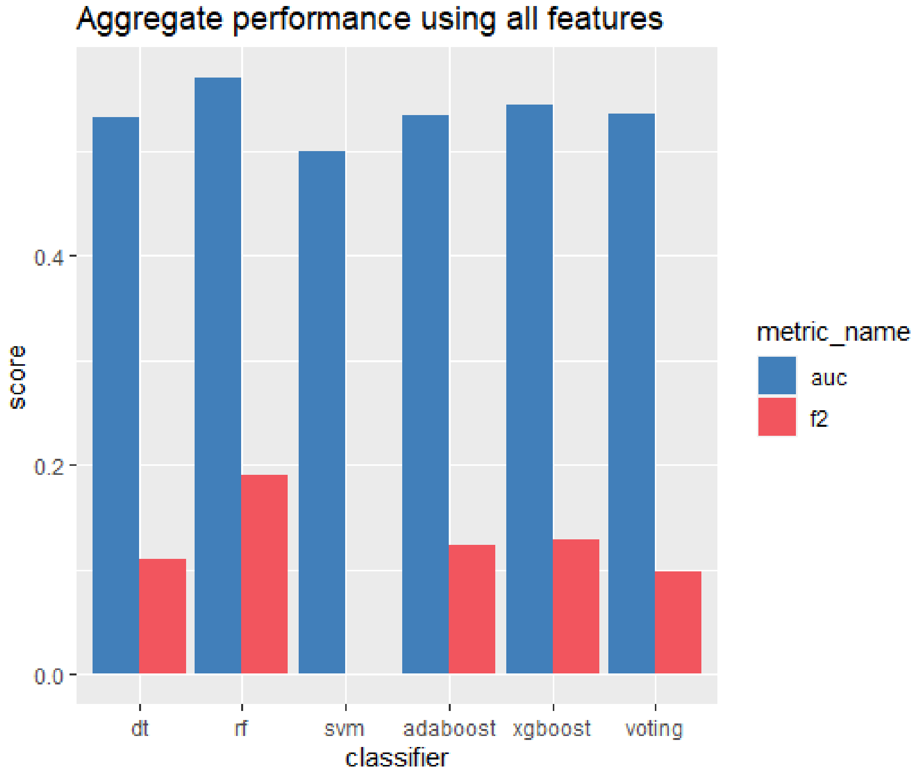


Figure 17: Classifier performance without sampling

It is obvious that our performance leaves a lot to be desired. Nevertheless, that was to be expected, given that the performance of the classifiers relies heavily on our data. Our extremely imbalanced data are too hard to handle for the classifiers who unwittingly give all our samples the same importance, leading to classifying almost all samples as 0 and resulting to disappointing f2 scores.

- Downsampling:

Table 12: Classifier performance using downsampling

Classifier	*F2*	*AUC*
Decision tree	0.344	0.646
Random forest	0.390	0.688
SVM	0.332	0.612
Adaboost	0.378	0.672
XGBoost	0.356	0.656
Voting	**0.392**	0.686

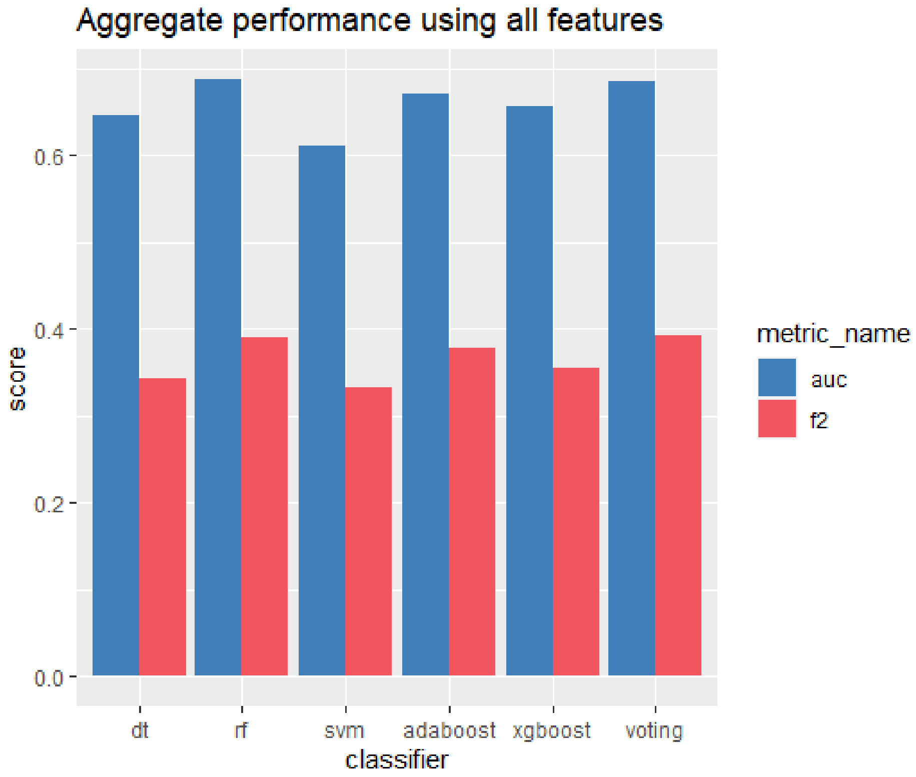


Figure 18: Classifier performance using downsampling

We can see that downsampling the data before training has a very big impact on improving performance. Every classifier performs much better and our scores are in the desired scope. Furthermore, since downsampling highly reduces the data we use at every training, it is much faster than the rest of the methods.

- Oversampling:

Table 13: Classifier performance using oversampling

Classifier	F2	AUC
Decision tree	0.322	0.626
Random forest	0.100	0.536
SVM	0.000	0.500
Adaboost	**0.320**	0.632
XGBoost	0.104	0.518
Voting	0.088	0.510

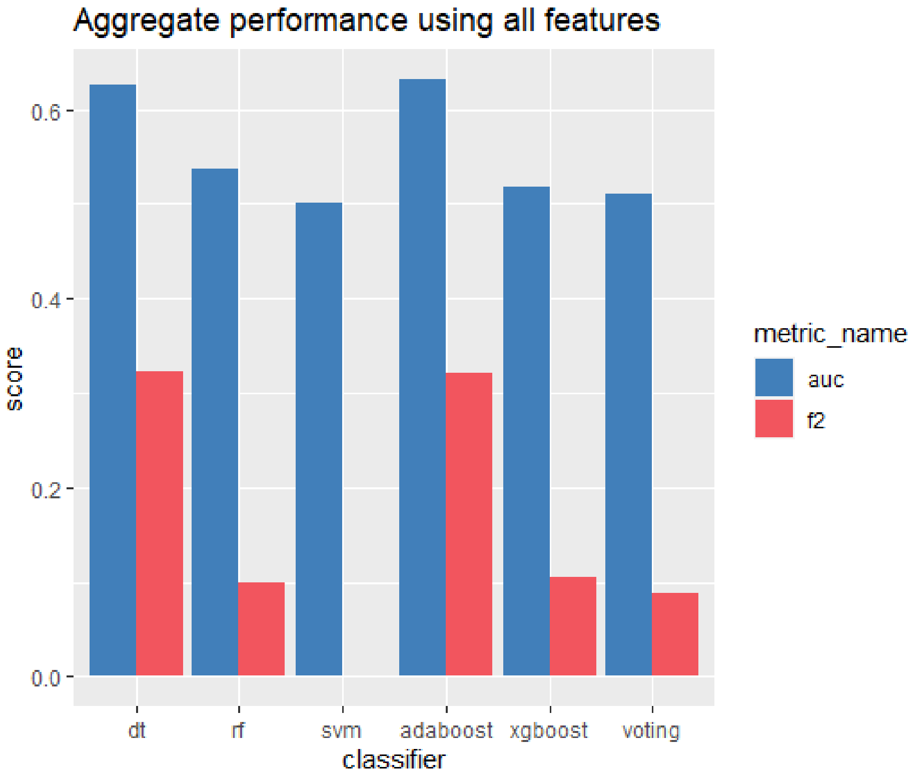


Figure 19: Classifier performance using oversampling

A couple of good performing classifiers, but overall worse scores. Also, as mentioned above, oversampling practically doubles the data that is used for every training, therefore doubling the total training time.

- ROSE:

Table 14: Classifier performance using ROSE

Classifier	*F2*	*AUC*
Decision tree	**0.308**	0.610
Random forest	0.262	0.592
SVM	0.090	0.512
Adaboost	0.214	0.566
XGBoost	0.272	0.502
Voting	0.236	0.572

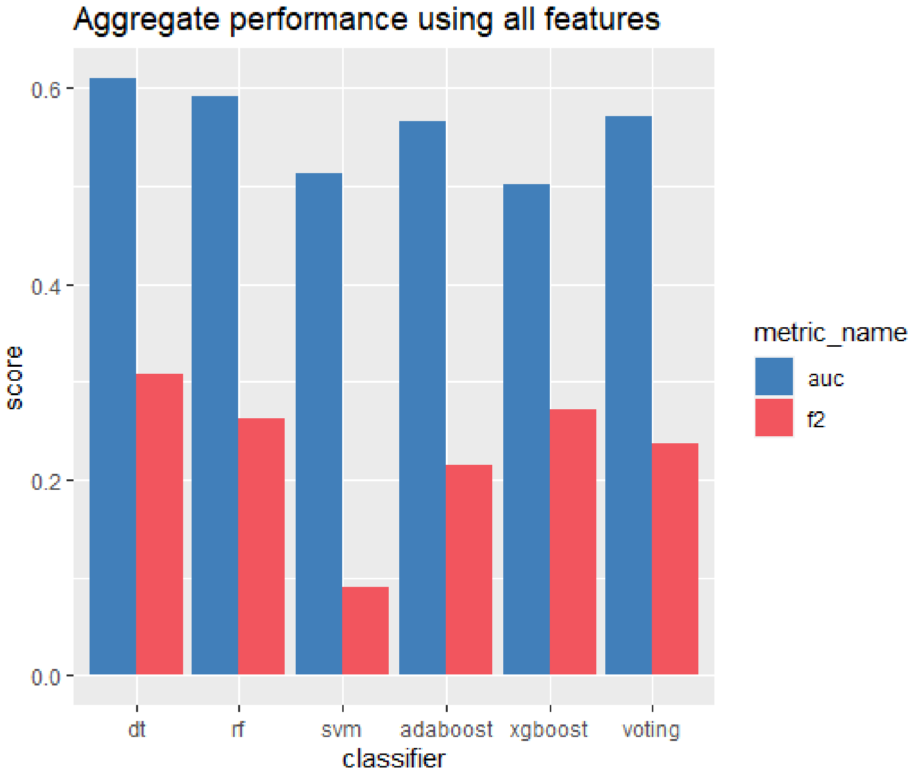


Figure 20: Classifier performance using ROSE

We can deduce that it is a more stable oversampling method, as it produces data that most classifiers handle decently (except SVM), but still yields worse performances than downsampling.

Conclusion:

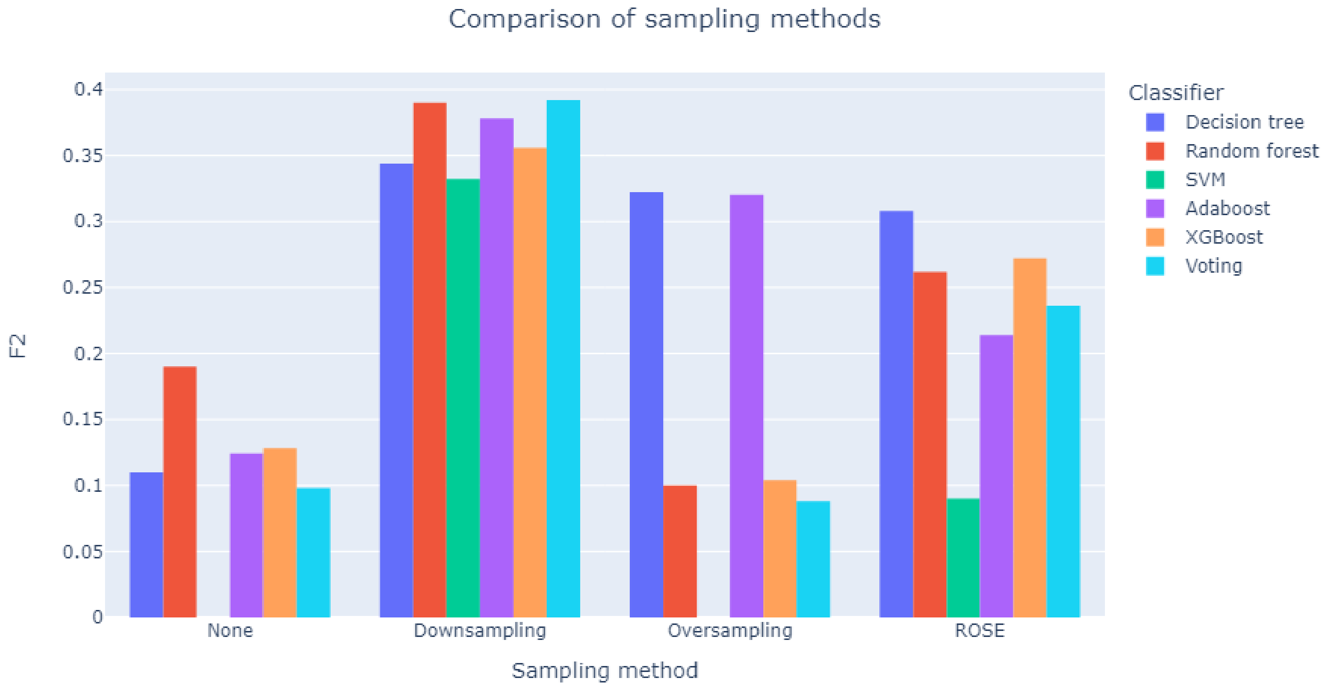


Figure 21: Comparison of sampling methods

Downsampling is the best sampling method as indicated both by results and by run times. Therefore, from now on we will only use downsampling for balancing the training data. Furthermore, when using RFE later on, we fit a random forest model to our data which also needs sampling. We will use downsampling as the balancing method for the RFE as well, in order to critically – as shown by this experiment – improve performance.

Experiment 2: Ignoring hinted features

In this experiment, we try out whether ignoring the four features hinted by the medical experts – as mentioned during the preprocessing – leads to better results.

- Using the original features:

Table 15: Classifier performance using all the features

Classifier	F2	AUC
Decision tree	0.344	0.646
Random forest	0.390	0.688
SVM	0.332	0.612
Adaboost	0.378	0.672
XGBoost	0.356	0.656
Voting	**0.392**	0.686

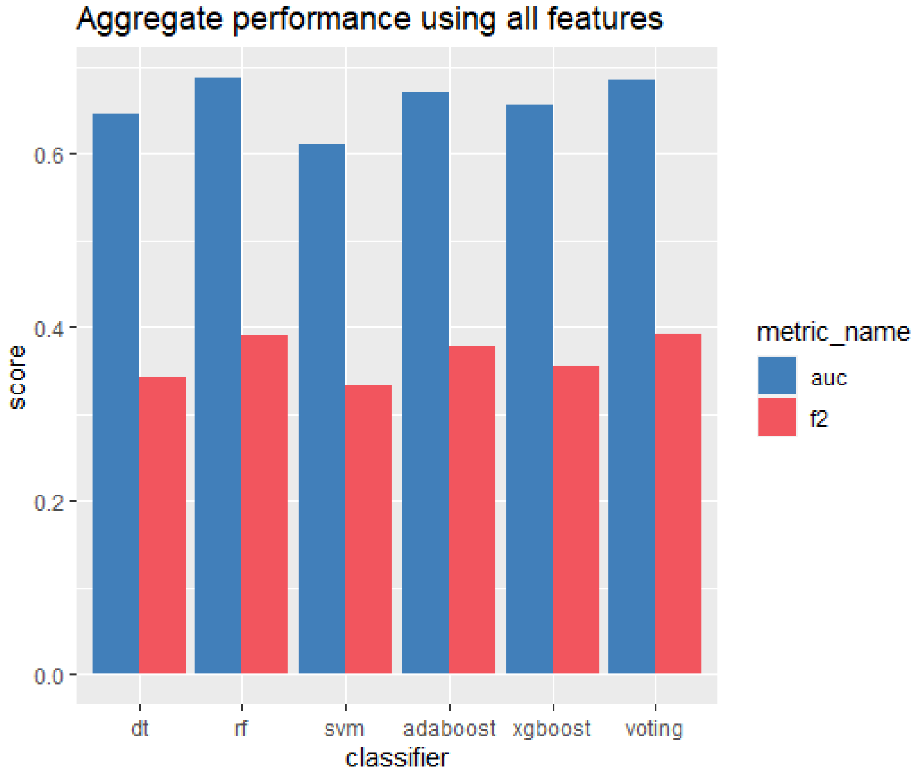


Figure 22: Classifier performance using all the features

- Ignoring the four features:

Table 16: Classifier performance ignoring hinted features

Classifier	*F2*	*AUC*
Decision tree	0.380	0.684
Random forest	0.382	0.676
SVM	0.330	0.638
Adaboost	0.396	0.688
XGBoost	0.372	0.688
Voting	**0.396**	0.692

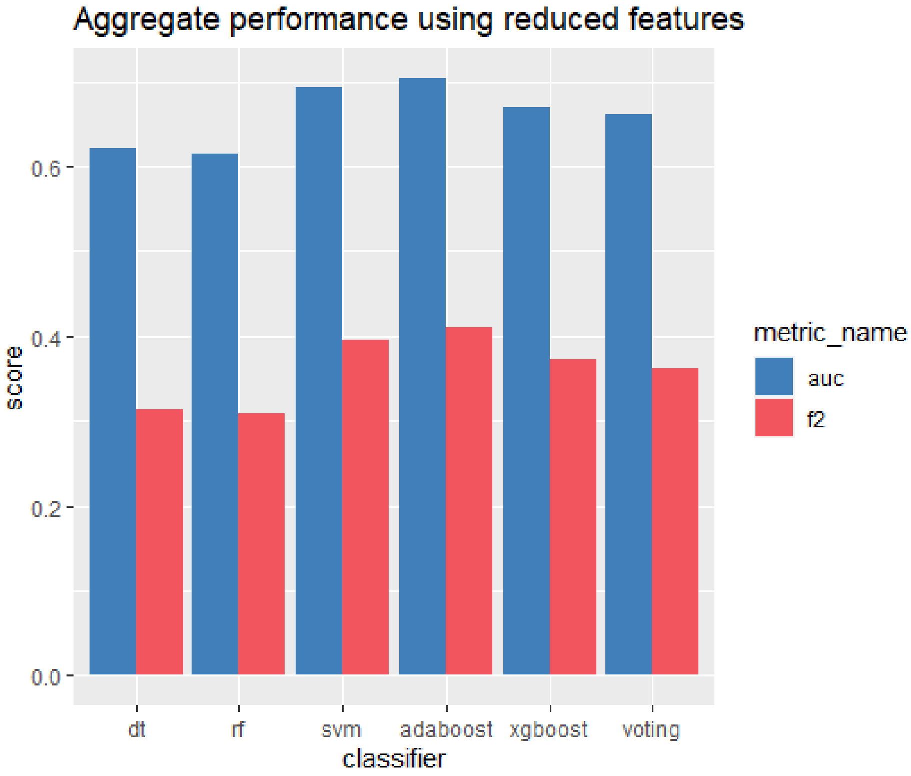


Figure 23: Classifier performance ignoring hinted features

We can see that ignoring these features seems to improve our performance ever so slightly. The difference is minuscule and could be attributed to the random processes that occur, e.g., the dataset's imputations, the randomness of the sampling method etc. Nevertheless, since there is no performance reduction, we can safely discard the mentioned features.

Conclusion:

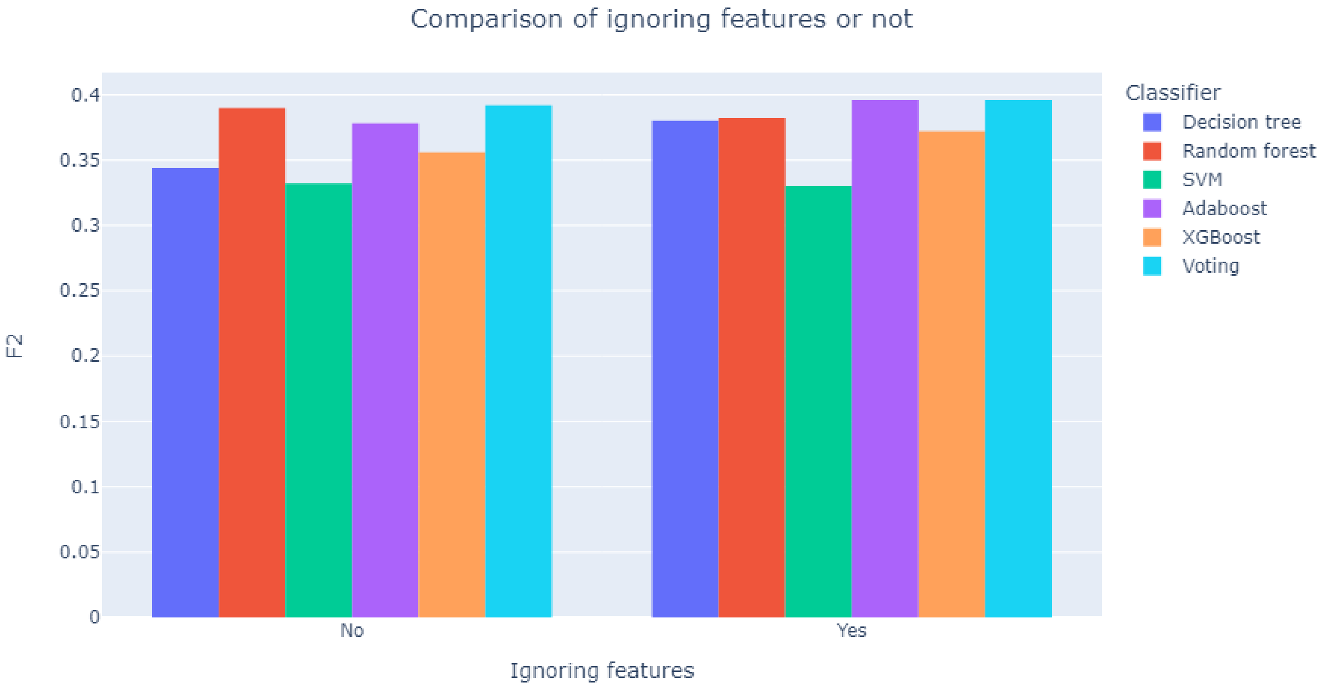


Figure 24: Comparison of ignoring features or not

Ignoring the aforementioned features leads to slightly better results and therefore from now on we will not include them in the data used for training.

Experiment 3: Using feature reduction

In this experiment we test out if reducing our features using RFE, as discussed during the preprocessing, enhances performance.

RFE is performed on every imputed dataset separately and for every one of them some important features are chosen. We consider the union of the features chosen each time to be the features judged as important and therefore use only them to train our classifiers.

In the following table we present which features were chosen as important, as well as how many times (out of the 5 RFE corresponding to the 5 imputed datasets) they were chosen.

Table 17: Chosen by RFE frequency table

Variable	Times chosen out of 5
Anxiety_HADS.3	5
Depression_HADS.3	5
Anxiety_HADS.0	5
Depression_HADS.0	5
Negative_affect_PANAS.3	5
Future_Persp_Image_BR23.3	5
M0_comprehensibility_SOC	5
M0_optimism_LOT	5
M0_meaningfulness_SOC	5
perceived_suppport_1_item.0	5
M0_manageability_SOC	5
M0_Forward_PACT	4
M3_mMOS_emotional_support	4
Cogn_Fun_QLQ30.3	4
M3_MAC_helpless	3
M3_MAC_anxious_preoc	3
perceived_suppport_1_item.3	3
M3_mMOS_instumental_support	2
M0_rumination_CERQ	2
Global_QLQ30.0	1
Body_Image_BR23.3	1
Side_Effects_BR23.3	1
M0_fear_of_recur_FCRI	1
baseline_thrombocytes	1
M0_resilience_CDRISC	1
Insomnia_QLQ30.3	1
M3_FARE_family_coping	1

We can see that the features that were chosen every (or almost every) time are all features that are obviously logically related to PTSD symptoms and we expected them to play an important part in predicting our outcome. For the exact meaning of every one of these features see the appendix at the end.

In order to more accurately depict the relationship between our chosen features and the predicted label, we employ an explainable AI tool from R called DALEX and use it randomly on the random forest classifier trained on one of our five imputed datasets using the aforementioned chosen features. To refrain from including too much information, we only include the resulting accumulated – local dependency profiles for the variables chosen all 5 times. The idea of partial dependency plots (to which accumulated – local dependency plots are an upgraded – more accurate version) is that they showcase how the expected value of our model's prediction changes with the variable in consideration.

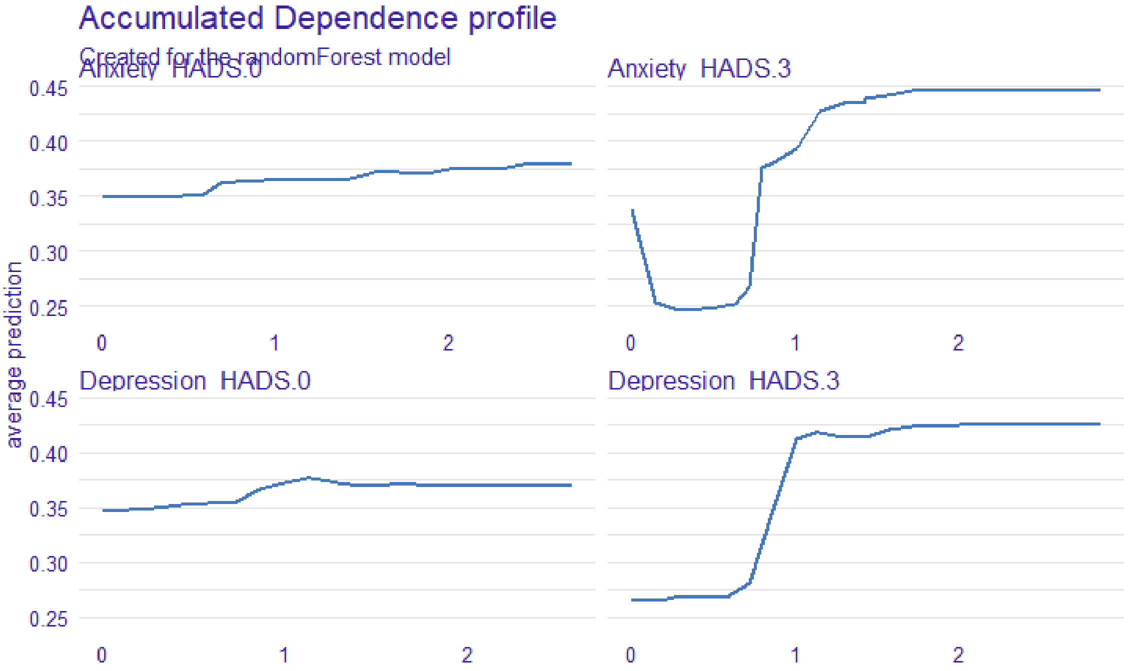


Figure 25: Accumulated dependency profiles (part 1)

As we expected, higher results on the HAD scale result to higher average prediction, which is even more prominent for the M3 results, as is to be expected since it is the closer (timewise) test result to our label.

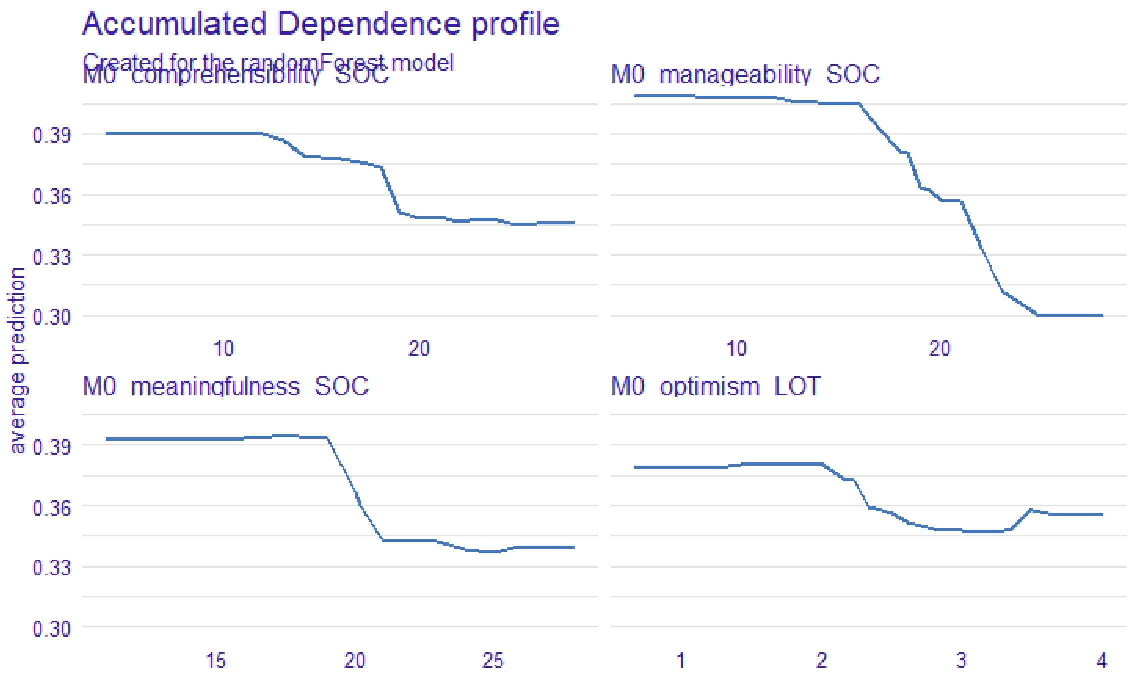


Figure 26: Accumulated dependency profiles (part 2)

The Sense Of Coherence (SOC) questionnaire proves to be a very important one:

- Comprehensibility: the stimuli deriving from one's internal and external environments in the course of living are structured, predictable, and explicable
- Manageability: the resources are available to one to meet the demands posed by these stimuli
- Meaningfulness: these demands are challenges, worthy of investment and engagement

As expected, when there is a better understanding around the patient's view of the disease (and generally) and therefore a more methodical and calm reaction to the relating problems, we can assume stronger mental health on the patient's part and smaller possibility of PTSD symptoms.

Optimism measured using the Life Orientation Test (LOT) also proves to be an important variable to consider, as should be expected, since there is an obvious logical correlation between it and our PTSD label.

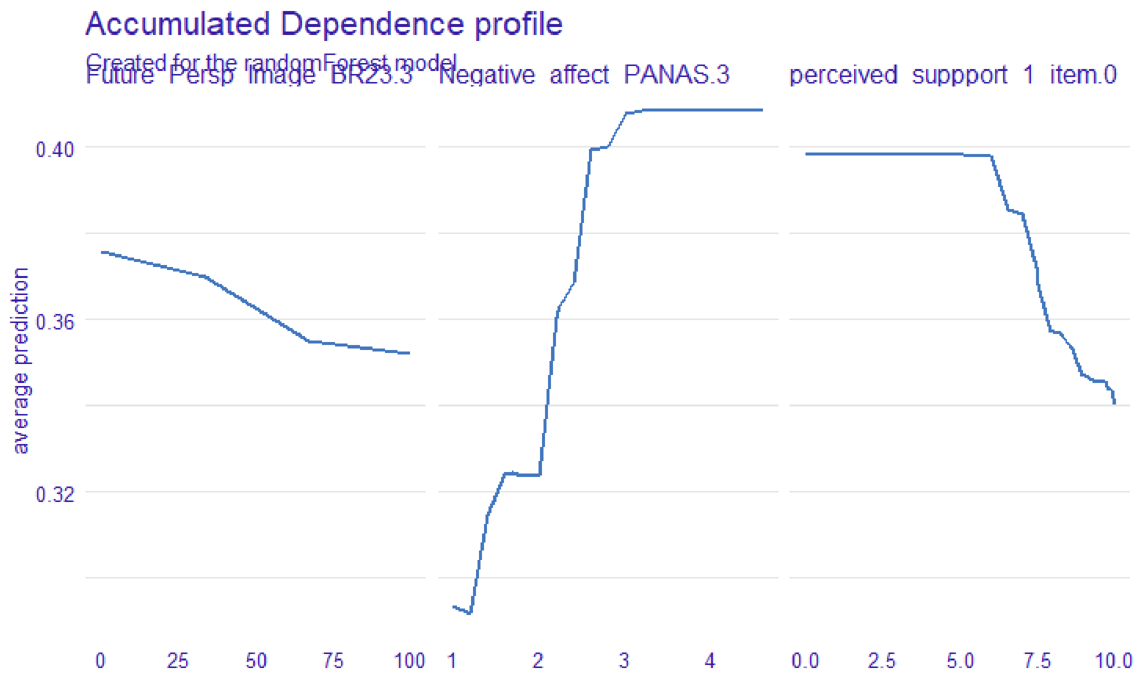


Figure 27: Accumulated dependency profiles (part 3)

Again, we can see that all the important variables relate to self-perception and mental health. Negative affect (Positive And Negative Affectivity – Short form) measure negative feelings and the more intense they are, the more likely we are to predict PTSD. If the patient feels – perceives more support, she obviously feels less threatened and is less likely to develop PTSD symptoms. Finally, a better future perspective image directly relates to mental health and lowers the probability of PTSD.

All of these results make sense and indicate that our analysis is in the right direction.

Now let's go back to our experiment and compare the performances using all the features and using only the reduced features chosen by the RFE:

Table 18: Performance comparison when using RFE

	All features		*Reduced features*	
Classifier	**F2**	**AUC**	**F2**	**AUC**
Decision tree	0.380	0.684	0.336	0.636
Random forest	0.382	0.676	0.400	0.690
SVM	0.330	0.638	**0.436**	0.734
Adaboost	0.396	0.688	0.364	0.660
XGBoost	0.372	0.688	0.356	0.654
Voting	**0.396**	0.692	0.400	0.692

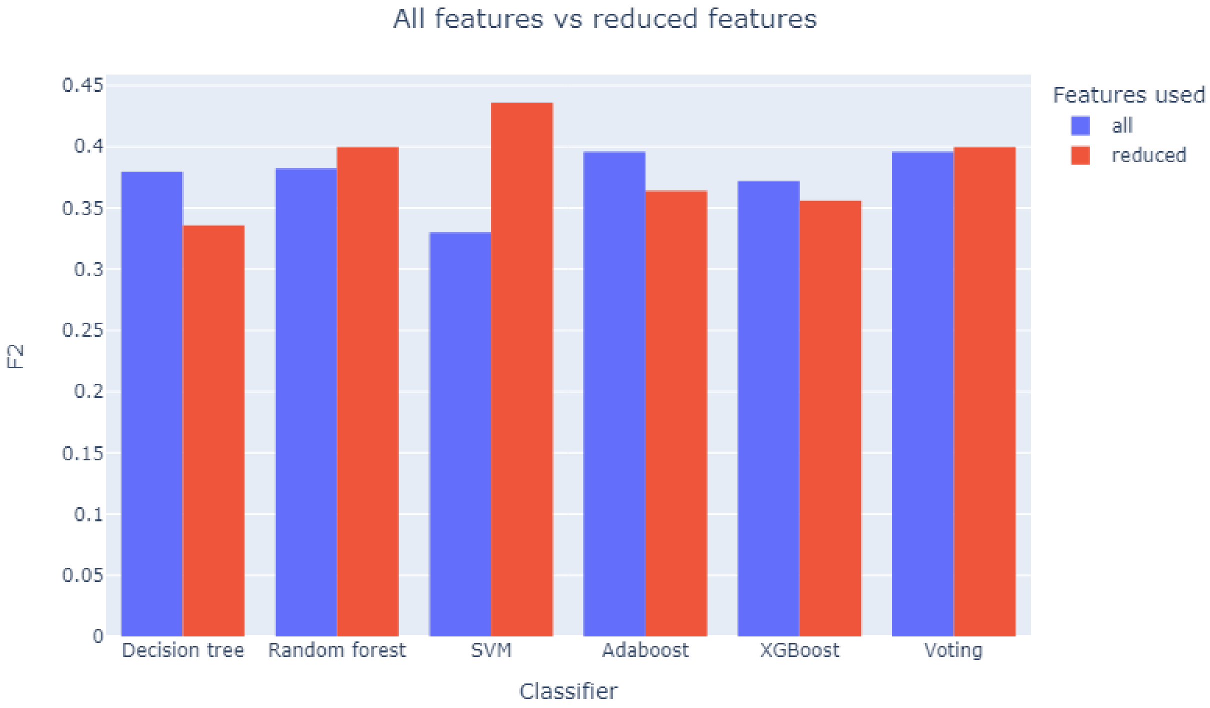


Figure 28: Performance comparison when using RFE

We can observe similar performances for most classifiers and a dramatic performance enhancement for our SVM. This was an expected outcome, since SVM is a classifier that suffers from the curse of dimensionality; this means that as the number of features (dimensions) increases, the amount of data needed to effectively train the model also increases exponentially. In high-dimensional spaces, SVMs may struggle due to overfitting or increased computational complexity. Lower-dimensional features can mitigate this issue. Since we have few data at our disposal and a relatively high dimensional space, we expected a performance improvement when performing feature reduction.

Conclusion:

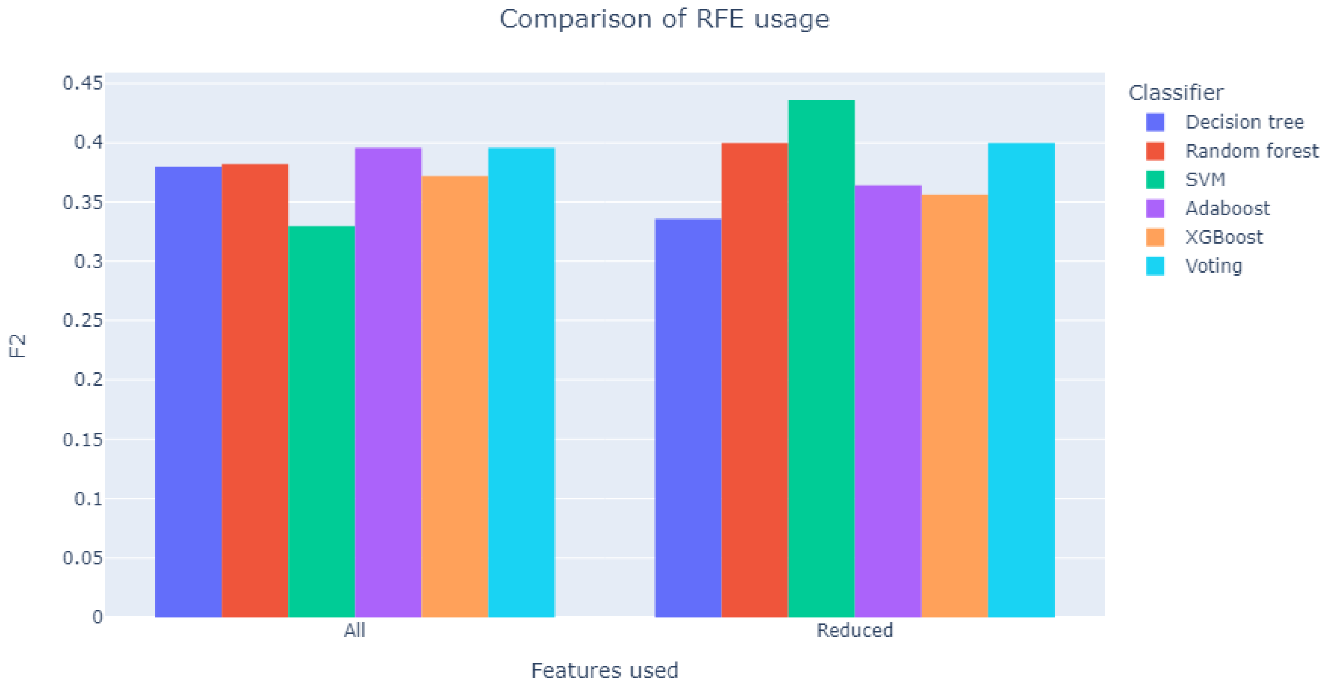


Figure 29: Comparison of RFE usage

Using the *reduced features* that we decided by performing RFE on our number of imputed datasets does not seem to lead to a downgrade in performance in most classifiers and highly enhances the performance of our SVM. Furthermore, using the chosen 27 features is obviously a big assistance to the medics involved in the data gathering process, as it is much easier and well-targeted than needing to gather the original 172 features. Therefore, from now on we will use these 27 features. We could even try to keep fewer of these, maybe only the ones that were chosen at least 3 times, but we already have a small enough number of features so we decided against it.

Comparing the classifiers' performance, we can see that overall, the top-performing ones seem to be *random forest*, *SVM* and *adaboost*. The *voting* classifier performs very well too, now that we have tuned the hyper-parameters to make our classifiers' performance more stable, but has the downside that we need to train all our five classifiers and therefore takes more time and resources to train. If we opt for performance, we should choose the voting classifier, whereas if we want faster calculations for our model, we could choose one of the three mentioned classifiers.

Experiment 4: Model performance on different hospital data

In this experiment, we intend to test how well our models' performance generalizes to data from different geographical locations. We split into train/test data by using as a test set each time the data that originated from one of the four oncology centers – pilots: IEO (Milan, Italy), HUS (Helsinki, Finland), HUJI (Jerusalem, Israel) and Champalimaud (Lisbon, Portugal).

Table 19: Classifier f2 performance on data from different hospitals

Testing on *Classifier*	*Random*	*IEO*	*HUS*	*HUJI*	*CHAMP*
Decision tree	0.336	0.420	0.268	0.494	0.316
Random forest	0.400	**0.532**	0.376	0.598	0.394
SVM	**0.436**	0.460	**0.390**	0.564	0.368
Adaboost	0.364	0.512	0.374	0.554	0.326
XGBoost	0.356	0.504	0.316	0.566	**0.426**
Voting	0.400	0.494	0.380	**0.620**	0.400

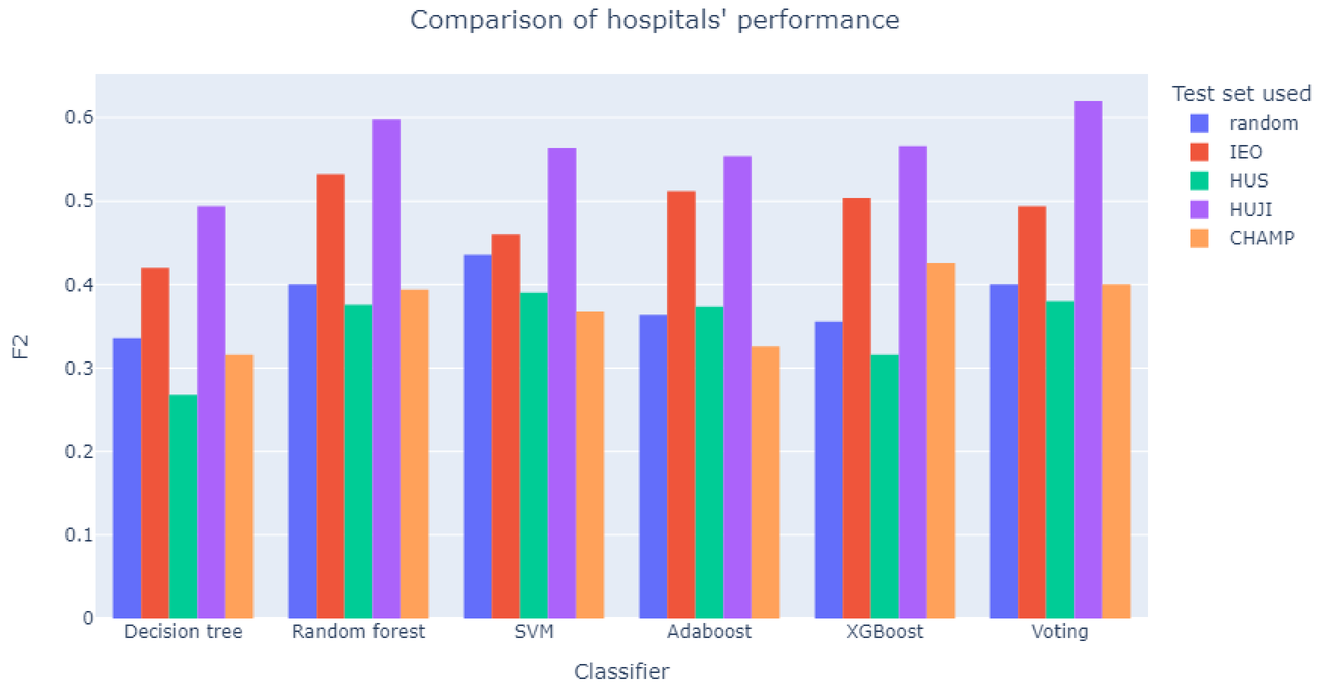


Figure 30: Classifier f2 performance on data from different hospitals

We can clearly see some interesting differences in the performance of the classifiers for the different hospitals. More specifically, when considering f2 score, predictions on data from the IEO and HUJI hospitals yield much better results than when randomly split, whereas predictions on data from HUS and CHAMP generally yield worse results than when randomly split. This could indicate towards different data quality across hospitals, but we decided to consider the AUC metric separately this time for better clarity:

Table 20: Classifier AUC performance on data from different hospitals

Testing on *Classifier*	*Random*	*IEO*	*HUS*	*HUJI*	*CHAMP*
Decision tree	0.636	0.618	0.626	0.720	0.640
Random forest	0.690	**0.736**	0.700	0.804	0.700
SVM	**0.734**	0.656	**0.732**	0.786	0.692
Adaboost	0.660	0.710	0.702	0.774	0.648
XGBoost	0.654	0.706	0.656	0.782	**0.728**
Voting	0.692	0.696	0.702	**0.830**	0.708

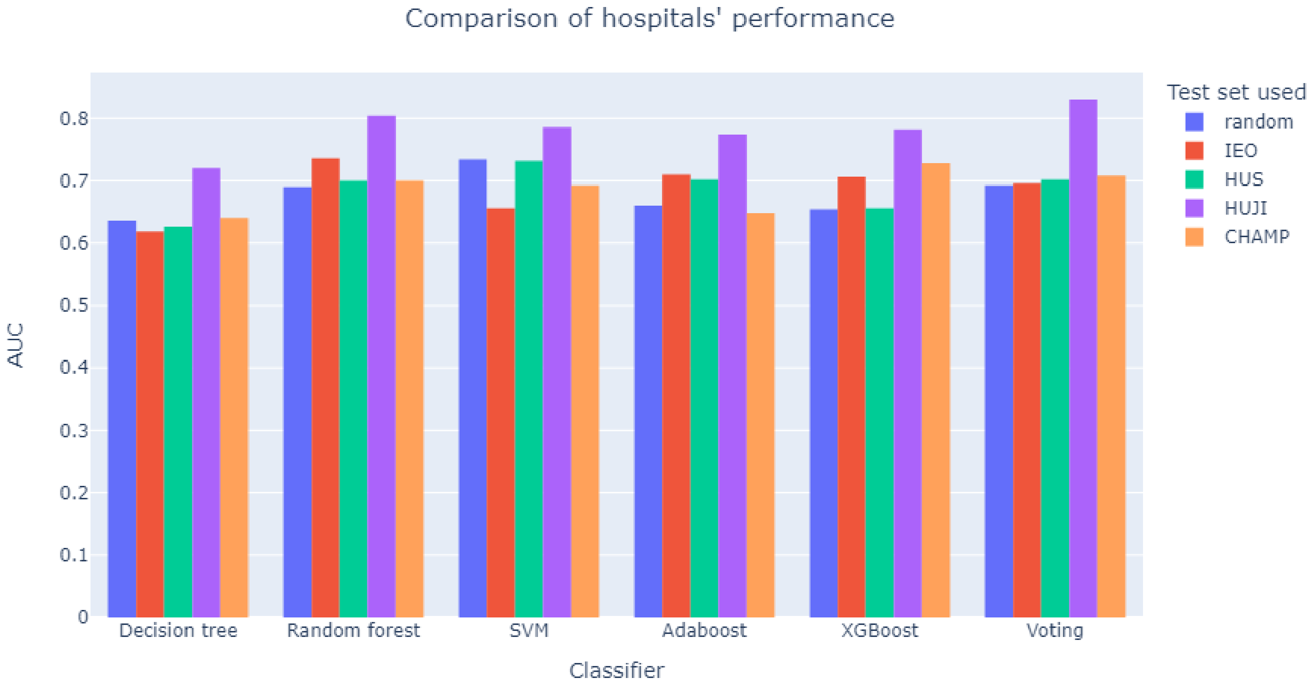


Figure 31: Classifier AUC performance on data from different hospitals

When AUC is considered, results seem much closer across hospitals. Upon closer inspection, the reason behind this mismatch is the unique nature of the f2 score along with the imbalance percentages of labels across the data from the different hospitals, as shown in the following table:

Table 21: Correlation between class percentage and best f2 score

Test data from	Number of positive class samples	Positive class label percentage (%)	Best f2 score
random	12	7	0.436
IEO	12	11	0.532
HUS	11	5	0.390
HUJI	12	10	0.620
CHAMP	8	6	0.426

It is clear that the higher the positive class percentage in our test set, the higher the f2 score we achieve. The reasoning behind this is that the f2 score only considers how many of the actual 1's we accurately predicted (recall) and how many of the predicted 1's indeed had label 1 (precision). However, when the test set consist of more data with label 0, even if the percentage of correctly classified negative samples remains roughly the same, our precision suffers and so does our f2 score. Our previous analysis is a good example of why a lot of metrics should be considered simultaneously, since all of them have their own advantages and disadvantages.

Conclusion:

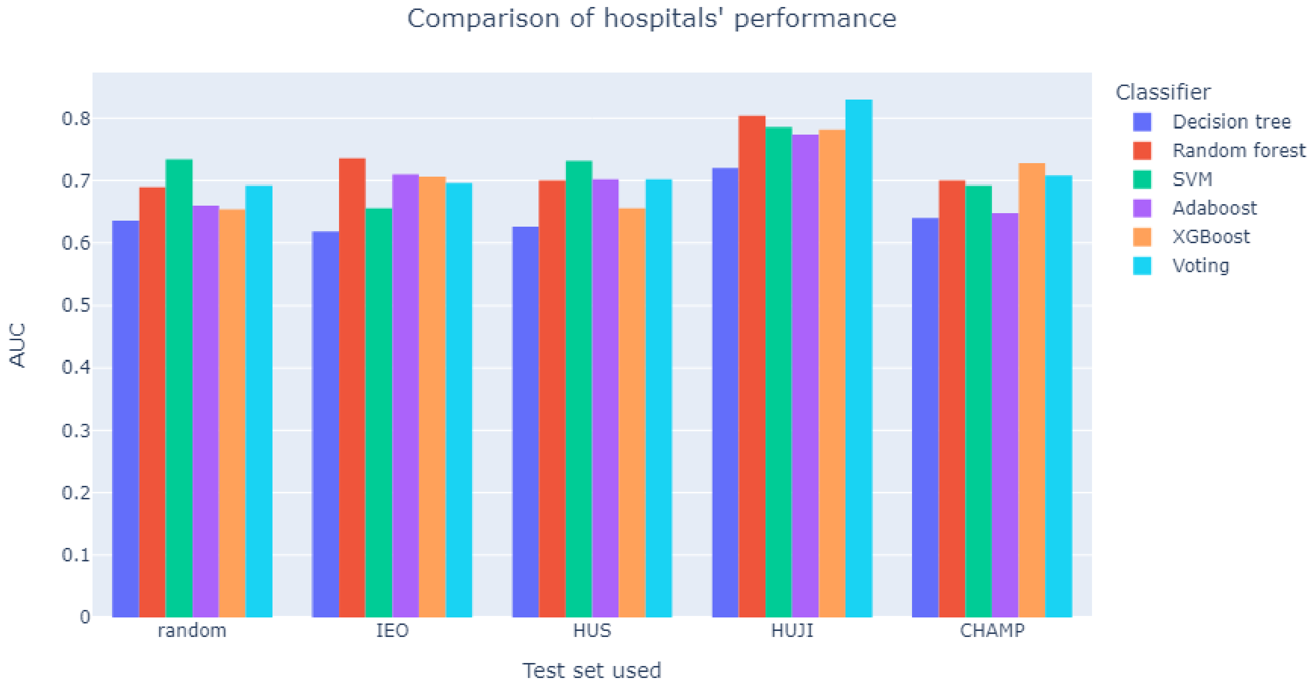


Figure 32: Comparison of performance on different hospitals

Our classifiers performance seems to be effectively generalizable across data from different hospitals and geographical locations.

Chapter 7. Discussion

7.1 Conclusion

In this diploma thesis, we organized a series of steps and decision in order to create the most optimal – under the limitations of our dataset – model for predicting PTSD symptoms in women with early breast cancer. The preprocessing steps that we suggested led to good performance, assuming it is combined with the appropriate sampling method, downsampling. The usage of the reduced features created using the RFE algorithm helps a lot the medical staff in charge of collecting this information and provides very good results, on par and even better than the full set of features. The models' prediction seems to generalize very well geographically and therefore do not seem to depend/focus on local characteristics or indicate any problem in the data collection methods across the hospitals being discussed. There was no clear "best" classifier. If a decision was to be made, we would choose SVM or random forest as single classifiers, or if we had the necessary time and resources for training all five classifiers, we would choose our voting classifier which was consistently among the top performing and is – by definition – the most stable classifier assuming there is not a huge disparity (which there isn't) in the different models' performance.

7.2 Impact

After a retraining is done using the complete dataset, the model we choose (together with our imputation model) is ready to be used for making predictions of PTSD symptoms in early breast cancer patients. Of course, the model is not to be used as a standalone. The goal is for it to be a supplementary tool for the medical experts to make a decision and intervene early to assist the women in need. It could be integrated in medical software and used for making decisions in real time, since the inference of the created models is very fast.

7.3 Further work

There are several steps that could be taken towards better understanding of our results and bettering our models:

- Our data is very limited, so more data should be gathered that will lead to more accurate models. For now, since they are so limited, our test set is quite small, which results in big fluctuations in our test set results, due to the randomness of some of our processes (e.g., sampling methods) – which becomes even more prominent due to our usage of the f2-score which is very sensitive when we have so few positive samples. To somewhat deal with this problem, instead of using a single test set, we could use an outer layer of cross-validation; we split our original dataset in k-folds and follow the same procedure we used for every fold (and possibly a few repeats as well), to gain a more accurate assessment of our models' performance. Of course, this requires a lot more computational time and resources.
- The accumulated local dependency profiles we used hold a lot of information towards understanding more about our features' correlation with our label, but also between them. More complex plots can be created including multiple variables and possibly the label that better capture their internal relationships.
- The entire process could be fully automated and integrated into a black-box model that, upon receiving the dataset and the specified label, preprocesses the data, trains the models and evaluates them, presenting to you the final best models, their performance and some explainability plots. This black-box could be used in other similar medical problems as well, it is not restricted to our specific breast cancer – PTSD problem.
- We should also try to use the labels of other months (M12, M18) to see if we can similarly predict PTSD symptoms further from baseline using our early months' data (M0 and M3). There may be some patterns we can deduce from the early months, while using data from later months can be impossible, since many patients do not stick to their scheduled medical exams, check-ups and possible questionnaires.
- Another approach towards predicting our label could be deep learning and Multi-Layer Perceptrons (MLP). It seems less intuitive for our problem and will most likely be harder to interpret and explain the results, but it is still very much a valid approach that could yield good results.

References

[1] *BOUNCE project*. https://www.bounce-project.eu/

[2] G. Pettini *et al.*, “Predicting Effective Adaptation to Breast Cancer to Help Women BOUNCE Back: Protocol for a Multicenter Clinical Pilot Study,” *JMIR Res. Protoc.*, vol. 11, no. 10, p. e34564, Oct. 2022, doi: 10.2196/34564.

[3] S. Almeida *et al.*, “132P The psychological impact of the COVID-19 pandemic on patients with early breast cancer,” *Ann. Oncol.*, vol. 32, p. S79, May 2021, doi: 10.1016/j.annonc.2021.03.146.

[4] R. Pat-Horenczyk *et al.*, “Trajectories of Quality of Life among an International Sample of Women during the First Year after the Diagnosis of Early Breast Cancer: A Latent Growth Curve Analysis,” *Cancers*, vol. 15, no. 7, Art. no. 7, Jan. 2023, doi: 10.3390/cancers15071961.

[5] E. C. Karademas *et al.*, “Changes over time in self-efficacy to cope with cancer and well-being in women with breast cancer: a cross-cultural study,” *Psychol. Health*, vol. 0, no. 0, pp. 1–14, 2023, doi: 10.1080/08870446.2023.2202205.

[6] E. C. Karademas *et al.*, “The mutual determination of self-efficacy to cope with cancer and cancer-related coping over time: a prospective study in women with breast cancer,” *Psychol. Health*, vol. 0, no. 0, pp. 1–14, 2022, doi: 10.1080/08870446.2022.2038157.

[7] E. C. Karademas *et al.*, “Cognitive, emotional, and behavioral mediators of the impact of coping self-efficacy on adaptation to breast cancer: An international prospective study,” *Psychooncology.*, vol. 30, no. 9, pp. 1555–1562, 2021, doi: 10.1002/pon.5730.

[8] E. C. Karademas *et al.*, “The Interplay Between Trait Resilience and Coping Self-efficacy in Patients with Breast Cancer: An International Study,” *J. Clin. Psychol. Med. Settings*, vol. 30, no. 1, pp. 119–128, Mar. 2023, doi: 10.1007/s10880-022-09872-x.

[9] P. Poikonen-Saksela *et al.*, “A graphical LASSO analysis of global quality of life, sub scales of the EORTC QLQ-C30 instrument and depression in early breast cancer,” *Sci. Rep.*, vol. 12, no. 1, Art. no. 1, Feb. 2022, doi: 10.1038/s41598-022-06138-2.

[10] L. Vehmanen *et al.*, “Associations between Physical Exercise, Quality of Life, Psychological Symptoms and Treatment Side Effects in Early Breast Cancer,” *Breast J.*, vol. 2022, p. e9921575, Nov. 2022, doi: 10.1155/2022/9921575.

[11] G. Stamatakos *et al.*, “In-silico systems for well-being: Artificial Intelligence based analysis of psychological, mental, functional and quality of life aspects of life after breast cancer treatment,” presented at the VPH2020 (Virtual Physiological Human Conference 2020), Paris, France (online), Aug. 2020, pp. 573–574.

[12] G. Stamatakos *et al.*, “Artificial intelligence based in silico models for the prediction of resilience related psychological, psychiatric and functional trajectories in women with early breast cancer,” presented at the VPH 2022 (Virtual Physiological Human Conference 2022), Porto, Portugal, Sep. 2022, p. 112.

[13] D. M. Hausman, “What Is Cancer?,” *Perspect. Biol. Med.*, vol. 62, no. 4, pp. 778–784, 2019, doi: 10.1353/pbm.2019.0046.

[14] C. P. Wild *et al.*, “Cancer Prevention Europe,” *Mol. Oncol.*, vol. 13, no. 3, pp. 528–534, Mar. 2019, doi: 10.1002/1878-0261.12455.

[15] D. Sun *et al.*, “Cancer burden in China: trends, risk factors and prevention,” *Cancer Biol. Med.*, vol. 17, no. 4, pp. 879–895, Nov. 2020, doi: 10.20892/j.issn.2095-3941.2020.0387.

[16] “How Cancer is Treated,” Cancer.Net. Accessed: Sep. 12, 2023. [Online]. Available: https://www.cancer.net/navigating-cancer-care/how-cancer-treated

[17] “Treatment for Cancer - NCI.” Accessed: Sep. 12, 2023. [Online]. Available: https://www.cancer.gov/about-cancer/treatment

[18] A. M, I. M, D. M, and K. Au, “Awareness and current knowledge of breast cancer,” *Biol. Res.*, vol. 50, no. 1, Oct. 2017, doi: 10.1186/s40659-017-0140-9.

[19] Y.-S. Sun *et al.*, “Risk Factors and Preventions of Breast Cancer,” *Int. J. Biol. Sci.*, vol. 13, no. 11, pp. 1387–1397, 2017, doi: 10.7150/ijbs.21635.

[20] “Cancer Statistics, 2021 - PubMed.” Accessed: Sep. 13, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33433946/

[21] H. R. Brewer, M. E. Jones, M. J. Schoemaker, A. Ashworth, and A. J. Swerdlow, “Family history and risk of breast cancer: an analysis accounting for family structure,” *Breast Cancer Res. Treat.*, vol. 165, no. 1, pp. 193–200, Aug. 2017, doi: 10.1007/s10549-017-4325-2.

[22] Endogenous Hormones and Breast Cancer Collaborative Group *et al.*, “Sex hormones and risk of breast cancer in premenopausal women: a collaborative reanalysis of individual participant data from seven prospective studies,” *Lancet Oncol.*, vol. 14, no. 10, pp. 1009–1019, Sep. 2013, doi: 10.1016/S1470-2045(13)70301-2.

[23] S. Valastyan and R. A. Weinberg, “Tumor metastasis: molecular insights and evolving paradigms,” *Cell*, vol. 147, no. 2, pp. 275–292, Oct. 2011, doi: 10.1016/j.cell.2011.09.024.

[24] K. L. Maughan, M. A. Lutterbie, and P. S. Ham, “Treatment of breast cancer,” *Am. Fam. Physician*, vol. 81, no. 11, pp. 1339–1346, Jun. 2010.

[25] T. Ferguson, N. Wilcken, R. Vagg, D. Ghersi, and A. K. Nowak, “Taxanes for adjuvant treatment of early breast cancer,” *Cochrane Database Syst. Rev.*, no. 4, p. CD004421, Oct. 2007, doi: 10.1002/14651858.CD004421.pub2.

[26] “Post-Traumatic Stress Disorder,” National Institute of Mental Health (NIMH). Accessed: Sep. 15, 2023. [Online]. Available: https://www.nimh.nih.gov/health/topics/post-traumatic-stress-disorder-ptsd

[27] “What is Posttraumatic Stress Disorder (PTSD)?” Accessed: Sep. 15, 2023. [Online]. Available: https://www.psychiatry.org:443/patients-families/ptsd/what-is-ptsd

[28] A. S. Zigmond and R. P. Snaith, “The hospital anxiety and depression scale,” *Acta Psychiatr. Scand.*, vol. 67, no. 6, pp. 361–370, Jun. 1983, doi: 10.1111/j.1600-0447.1983.tb09716.x.

[29] “Classification: ROC Curve and AUC | Machine Learning,” Google for Developers. Accessed: Sep. 16, 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[30] S. Bhatt, “Reinforcement Learning 101,” Medium. Accessed: Sep. 17, 2023. [Online]. Available: https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

[31] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[32] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[33] Y. Freund and R. E. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” in *Computational Learning Theory*, P. Vitányi, Ed., in Lecture

Notes in Computer Science. Berlin, Heidelberg: Springer, 1995, pp. 23–37. doi: 10.1007/3-540-59119-2_166.

[34] M. Šimunović and D. Ljubotina, “Prevalence, Structure and Predictors of Posttraumatic Stress Disorder Symptoms in Croatian Patients Following Breast Cancer,” *Psychiatr. Danub.*, vol. 32, no. 2, pp. 187–196, 2020, doi: 10.24869/psyd.2020.187.

[35] K. N. DuHamel *et al.*, “Construct validity of the posttraumatic stress disorder checklist in cancer survivors: analyses based on two samples,” *Psychol. Assess.*, vol. 16, no. 3, pp. 255–266, Sep. 2004, doi: 10.1037/1040-3590.16.3.255.

[36] A. Gibbons, A. Groarke, and K. Sweeney, “Predicting general and cancer-related distress in women with newly diagnosed breast cancer,” *BMC Cancer*, vol. 16, no. 1, p. 935, Dec. 2016, doi: 10.1186/s12885-016-2964-z.

[37] K. Tsaras *et al.*, “Assessment of Depression and Anxiety in Breast Cancer Patients: Prevalence and Associated Factors,” *Asian Pac. J. Cancer Prev. APJCP*, vol. 19, no. 6, pp. 1661–1669, Jun. 2018, doi: 10.22034/APJCP.2018.19.6.1661.

[38] J. Perez-Tejada, A. Labaka, O. Vegas, A. Larraioz, A. Pescador, and A. Arregi, “Anxiety and depression after breast cancer: The predictive role of monoamine levels,” *Eur. J. Oncol. Nurs. Off. J. Eur. Oncol. Nurs. Soc.*, vol. 52, p. 101953, Jun. 2021, doi: 10.1016/j.ejon.2021.101953.

[39] W. W. T. Lam *et al.*, “Factors predicting patient satisfaction in women with advanced breast cancer: a prospective study,” *BMC Cancer*, vol. 18, no. 1, p. 162, Feb. 2018, doi: 10.1186/s12885-018-4085-3.

[40] E. Y. Kalafi, N. a. M. Nor, N. A. Taib, M. D. Ganggayah, C. Town, and S. K. Dhillon, “Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data,” *Folia Biol. (Praha)*, vol. 65, no. 5–6, pp. 212–220, 2019.

[41] F. Ge, D. Zhang, L. Wu, and H. Mu, “Predicting Psychological State Among Chinese Undergraduate Students in the COVID-19 Epidemic: A Longitudinal Study Using a Machine Learning,” *Neuropsychiatr. Dis. Treat.*, vol. 16, pp. 2111–2118, 2020, doi: 10.2147/NDT.S262004.

[42] M. Montazeri, M. Montazeri, K. Bahaadinbeigy, M. Montazeri, and A. Afraz, “Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review,” *Health Sci. Rep.*, vol. 6, no. 1, p. e962, Jan. 2023, doi: 10.1002/hsr2.962.

[43] S. V. Buuren and K. Groothuis-Oudshoorn, “**mice** : Multivariate Imputation by Chained Equations in *R*,” *J. Stat. Softw.*, vol. 45, no. 3, 2011, doi: 10.18637/jss.v045.i03.

[44] *https://stefvanbuuren.name/fimd/sec-FCS.html*. Accessed: Sep. 26, 2023. [Online]. Available: https://stefvanbuuren.name/fimd/sec-FCS.html

[45] M. Kuhn, *20 Recursive Feature Elimination | The caret Package*. Accessed: Sep. 26, 2023. [Online]. Available: https://topepo.github.io/caret/recursive-feature-elimination.html#rfe

[46] G. Novack, “Building a One Hot Encoding Layer with Tensorflow,” Medium. Accessed: Sep. 26, 2023. [Online]. Available: https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39

Appendix

Variable	Type	Coding	Meaning
mrn	factor		Patient code
Depression_HADS.0	numeric (double)	0-3	Hospital Anxiety and Depression Scale, depression (M0)
Depression_HADS.3	numeric (double)	0-3	Hospital Anxiety and Depression Scale, depression (M3)
education_2	factor	2 levels 1: ≤12 years 2: >12 years	Education level
number_of_children	numeric (integer)		The number of children
m0_income_3	factor	3 levels 1: low (HUS&HUJI:0-1500 IEO&CHAMP:0-1000) 2: average (HUS&HUJI:1501-3500 IEO&CHAMP:1001-3000) 3: high (HUS&HUJI:>3500 IEO&CHAMP:>3000)	The level of income (M0)
m0_sick_leave_days	numeric (integer)		The number of sick leave days (M0)
m0_do_you_smoke	factor	3 levels 1: Yes 2: No 3: I only smoked in the past	Smoking (M0)
m0_drinking_EK	factor	3 levels 0: No Drinking 1: Drinking in Moderation 2: Heavy drinking	The level of drinking (M0) Heavy drinking: consuming more than 3 drinks on any day or more than 7 drinks per week
m0_BMI	numeric (double)		Body Mass Index results (M0)

m0_diet	factor	3 levels 0: No diet 1: Mediterranean/Vegetarian type 2: Special	The type of diet of a patient (M0) Mediterranean/Vegetarian type: low calories, low carb, Mediterranean, vegetarian, no red meat Special: e.g. protein only, vegan, gluten free, diary free, FODMAP-free, Macrobiotic, Ketogenic, paleo, Lactose-free
m3_employment_status	factor	6 levels 1: Employed full time 2: Employed part time 3: Housewife 4: Retired 5: Self-employed 6: Unemployed	The employment status (M3)
m3_sick_leave_days	numeric (integer)		The number of sick leave days (M3)
m3_mental_health_support	factor	2 levels 0: No 1: Yes	Whether or not the patient receives mental health aid (M3)
m3_mental_health_support_times	numeric (integer)		How many times she received it (M3)
m3_do_you_do_any_activities_to_support_your_wellbeing	factor	2 levels 0: No 1: Yes	Doing activities to support wellbeing (M3)
m3_used_services_to_support_wellbeing	factor	2 levels 0: No 1: Yes	Using services to support wellbeing (M3)
m3_domestic_help_during_last_three_months	factor	2 levels 0: No 1: Yes	Domestic help during last 3 months (M3)

m3_domestic_help_days	numeric (integer)		Number of domestic help days (M3)
M0_optimism_LOT	numeric (double)	0-4	Life Orientation Test Revised (LOT-R) Optimism (M0)
M0_comprehensibility_SOC	numeric (double)	4-28	Sense Of Coherence test comprehensibility results (M0)
M0_manageability_SOC	numeric (double)	4-28	Sense Of Coherence test manageability results (M0)
M0_meaningfulness_SOC	numeric (double)	4-28	Sense Of Coherence test meaningfulness results (M0)
M0_fear_of_recur_FCRI	numeric (double)	0-4	Fear of Cancer Recurrence Inventory - short form total (average) score(M0)
M0_Forward_PACT	numeric (double)	1-7	Perceived Ability to Cope with Trauma, Forward focus (M0)
M0_Trauma_PACT	numeric (double)	1-7	Perceived Ability to Cope with Trauma, Trauma focus (M0)
M0_Flexibility_PACT	numeric (double)	2-14	Perceived Ability to Cope with Trauma, flexibility score (M0)
M0_self_blame_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, self-blame (M0)
M0_other_blame_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire other-blame (M0)
M0_rumination_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, rumination (M0)
M0_catastrophizing_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation

			Questionnaire, catastrophizing (M0)
M0_perspective_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, perspective (M0)
M0_pos_refus_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, positive refocusing (M0)
M0_pos_reapp_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, positive reappraisal (M0)
M0_acceptance_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, acceptance (M0)
M0_planning_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, planning (M0)
M0_negative_overall_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, negative overall (M0)
M0_positive_overall_CERQ	numeric (double)	1-5	Cognitive Emotion Regulation Questionnaire, positive overall (M0)
M0_mindfulness_MAAS	numeric (double)	1-6	Mindful Attention Awareness Scale, mindfulness (M0)
M0_resilience_CDRISC	numeric (double)	0-4	Connor Davidson Resilience Scale, resilience-average score (M0)
M0_coping_with_cancer_CBI	numeric (double)	1-9	Cancer Behavior Inventory, coping self-efficacy (M0)

M3_PTGI_relating_to_others	numeric (double)	0-5	Post Traumatic Growth Inventory, relating to others (M3)
M3_PTGI_new_possibilities	numeric (double)	0-5	Post Traumatic Growth Inventory, new possibilities (M3)
M3_PTGI_personal_strength	numeric (double)	0-5	Post Traumatic Growth Inventory, personal strength (M3)
M3_PTGI_spiritual_change	numeric (double)	0-5	Post Traumatic Growth Inventory, spiritual change (M3)
M3_PTGI_appreciation_of_life	numeric (double)	0-5	Post Traumatic Growth Inventory, appreciation of life (M3)
M6_ptsd_PCL5	numeric (double)	0-100	Post-Traumatic Stress Disorder (PTSD) related questionnaire, total symptom severity score (M6)
M6_clusterB_PCL5	numeric (integer)	0-25	PTSD related questionnaire, symptom cluster B severity score (M6)
M6_clusterC_PCL5	numeric (integer)	0-10	PTSD related questionnaire, symptom cluster C severity score (M6)
M6_clusterD_PCL5	numeric (double)	0-35	PTSD related questionnaire, symptom cluster D severity score (M6)
M6_clusterE_PCL5	numeric (double)	0-30	PTSD related questionnaire, symptom cluster E severity score (M6)
M12_ptsd_PCL5	numeric (double)	0-100	PTSD related questionnaire,

			total symptom severity score (M12)
M12_clusterB_PCL5	numeric (double)	0-25	PTSD related questionnaire, symptom cluster B severity score (M12)
M12_clusterC_PCL5	numeric (integer)	0-10	PTSD related questionnaire, symptom cluster C severity score (M12)
M12_clusterD_PCL5	numeric (double)	0-35	PTSD related questionnaire, symptom cluster D severity score (M12)
M12_clusterE_PCL5	numeric (double)	0-30	PTSD related questionnaire, symptom cluster E severity score (M12)
M6_DSMdiagnosis_PCL	factor	2 levels 0: Negative 1: Positive	PTSD related questionnaire, a provisional PTSD diagnosis (M6)
M12_DSMdiagnosis_PCL	factor	2 levels 0: Negative 1: Positive	PTSD related questionnaire, a provisional PTSD diagnosis (M12)
M18_ptsd_PCL5	numeric (double)	0-100	PTSD related questionnaire, total symptom severity score (M18)
M18_clusterB_PCL5	numeric (double)	0-25	PTSD related questionnaire, symptom cluster B severity score (M18)
M18_clusterC_PCL5	numeric (integer)	0-10	PTSD related questionnaire, symptom cluster C severity score (M18)
M18_clusterD_PCL5	numeric (double)	0-35	PTSD related questionnaire, symptom cluster D severity score (M18)

M18_clusterE_PCL5	numeric (double)	0-30	PTSD related questionnaire, symptom cluster E severity score (M18)
M18_DSMdiagnosis_PCL	factor	2 levels 0: Negative 1: Positive	PTSD related questionnaire, a provisional PTSD diagnosis (M18)
M3_MAC_helpless	numeric (double)	1-4	Mini-Mental Adjustment to Cancer Scale, helpless (M3)
M3_MAC_anxious_preoc	numeric (double)	1-4	Mini-Mental Adjustment to Cancer Scale, anxious preoccupation (M3)
M3_MAC_fighting	numeric (double)	1-4	Mini-Mental Adjustment to Cancer Scale, fighting (M3)
M3_MAC_avoidance	numeric (double)	1-4	Mini-Mental Adjustment to Cancer Scale, avoidance (M3)
M3_MAC_fatalism	numeric (double)	1-4	Mini-Mental Adjustment to Cancer Scale, fatalism (M3)
M3_FARE_commun_cohesion	numeric (double)	1-7	FAmily REsilience Questionnaire, communication and cohesion (M3)
M3_FARE_family_coping	numeric (double)	1-7	FAmily REsilience Questionnaire, perceived family coping (M3)
general_se_1_item.0	numeric (double)	0-10	A general self-efficacy item
perceived_suppport_1_item.0	numeric (double)	0-10	A general perceived support item
LifeEvents_012.0	factor	3 levels 0: None 1: One event 2: Two or more events	Negative

Anxiety_HADS.0	numeric (double)	0-3	Hospital Anxiety and Depression Scale, anxiety (M0)
Positive_affect_PANAS.0	numeric (double)	1-5	Positive Affect Negative Affect Schedule (M0)
Negative_affect_PANAS.0	numeric (double)	1-5	Positive Affect Negative Affect Schedule (M0)
Global_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, global health status / QoL (M0)
Phys_Fun_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, physical functioning (M0)
Role_Fun_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, role functioning (M0)
Cogn_Fun_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, cognitive functioning (M0)
Soc_Fun_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, social functioning (M0)
Fatigue_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, fatigue (M0)
Nausea_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, nausea and vomiting (M0)
Pain_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life

			Questionnaire, pain (M0)
Dyspnoea_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, dyspnea (M0)
Insomnia_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, insomnia (M0)
Apetite_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, appetite loss (M0)
Constipation_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, constipation (M0)
Diarrhoea_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, diarrhea (M0)
Financial_QLQ30.0	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, financial impact (M0)
Body_Image_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, body image (M0)
Side_Effects_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, systemic therapy side effects (M0)
Breast_Symptoms_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, breast symptoms (M0)

Arm_Symptoms_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, arm symptoms (M0)
Future_Persp_Image_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, future perspective (M0)
Sex_Funct_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, sexual functioning (M0)
Upset_Hair_Image_BR23B.0	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, upset by hair loss (M0)
mos.3	numeric (integer)	1-5	the MOS Adherence to medical advice scale (M3)
general_se_1_item.3	numeric (double)	0-10	A general self-efficacy item
perceived_suppport_1_item.3	numeric (double)	0-10	A general perceived support item
single_item_cope1.3	numeric (integer)	1-5	Single item: what has done to cope: Tried to relax
single_item_cope2.3	numeric (integer)	1-5	Single item: what has done to cope: Distracted yourself
single_item_cope3.3	numeric (integer)	1-5	Single item: what has done to cope: Prayed
single_item_cope4.3	numeric (integer)	1-5	Exercised or used physical activity
single_item_cope5.3	numeric (integer)	1-5	Single item: what has done to cope: Tried to look at the more

			positive sides of your experience
single_item_cope6.3	numeric (integer)	1-5	Single item: what has done to cope: Burst to tears or lashed out
single_item_cope7.3	numeric (integer)	1-5	Single item: what has done to cope: Talked to somebody important
single_item_cope8.3	numeric (integer)	1-5	Single item: what has done to cope: Asked for help somebody important
single_item_cope9.3	numeric (integer)	1-5	Single item: what has done to cope: Tried to see your experience rather as a challenge
single_item_cope10.3	numeric (integer)	1-5	Single item: what has done to cope: Talked to your physician about your concerns
bipq1.3	numeric (integer)	0-10	Illness Perception Questionaire - Brief form Self-control beliefs
bipq2.3	numeric (integer)	0-10	Illness Perception Questionaire - Brief form Treatment control beliefs
LifeEvents_012.3	factor	3 levels 0: None 1: One event 2: Two or more events	Negative Life Events
Anxiety_HADS.3	numeric (double)	0-3	Hospital Anxiety and Depression Scale, anxiety (M3)
Positive_affect_PANAS.3	numeric (double)	1-5	Positive Affect Negative Affect Schedule (M3)

Negative_affect_PANAS.3	numeric (double)	1-5	Positive Affect Negative Affect Schedule (M3)
Global_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, global health status / QoL (M3)
Phys_Fun_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, physical functioning (M3)
Role_Fun_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, role functioning (M3)
Cogn_Fun_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, cognitive functioning (M3)
Soc_Fun_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, social functioning (M3)
Fatigue_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, fatigue (M3)
Nausea_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, nausea and vomiting (M3)
Pain_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, pain (M3)
Dyspnoea_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, dyspnea (M3)

Insomnia_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, insomnia (M3)
Apetite_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, appetite loss (M3)
Constipation_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, constipation (M3)
Diarrhoea_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, diarrhea (M3)
Financial_QLQ30.3	numeric (double)	0-100	QLQ-C30 - EORTC Quality of Life Questionnaire, financial impact (M3)
Body_Image_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, body image (M3)
Side_Effects_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, systemic therapy side effects (M3)
Breast_Symptoms_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, breast symptoms (M3)
Arm_Symptoms_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, arm symptoms (M3)

Future_Persp_Image_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, future perspective (M3)
Sex_Funct_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, sexual functioning (M3)
Upset_Hair_Image_BR23B.3	numeric (double)	0-100	QLQ-BR23 - EORTC quality of life questionnaire breast cancer module, upset by hair loss (M3)
FORTH_baseline_chronic_illness	factor	2 levels 0: No 1: Yes	Existence of chronic illness
FORTH_preexisting_illnesses	numeric (int)		Number of pre-existing illnesses
FORTH_preexisting_mentalillness	factor	2 levels 0: No 1: Yes	Pre-existence of mental illness
FORTH_preexisting_metabolicillness	factor	2 levels 0: No 1: Yes	Pre-existence of metabolic illness
FORTH_m3_performancestatus	factor	4 levels 0: Fully active, able to carry on all pre-disease performance without restriction 1: Restricted in physically strenuous activity, but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work	ECOG Performance status (M3)

		2: Ambulatory and capable of all selfcare, but unable to carry out any work activities; up and about more than 50% of waking hours 3: Capable of only limited selfcare; confined to bed or chair more than 50% of waking hours 4: Completely disabled; cannot carry on any selfcare; totally confined to bed or chair 5: Dead	
FORTH_m3_illness_events0.1.2	factor	3 levels 0: None 1: One event 2: Two or more events	Illness events (M3)
FORTH_M0_psychotropics	factor	2 levels 0: No 1: Yes	Consumption of psychotropic drugs (M0)
FORTH_M3_psychotropics	factor	2 levels 0: No 1: Yes	Consumption of psychotropic drugs (M3)
FORTH_M3_Mental_health_support	factor	2 levels 0: No 1: Yes	Mental health support (M3)
m0_exercise_012	factor	3 levels 0: None 1: Low/moderate 2: Heavy	Level of exercise (M0) moderate aerobic activity: walking, cycling, etc. heavy aerobic activity: running, HIT training, etc.

			Heavy level of exercise: ≥200 min/week of moderate aerobic activity or ≥100 min/week of heavy aerobic activity or an equivalent combination of moderate- and heavy-intensity activity or ≥5 times of week muscle strength activity or (any aerobic activity & 4 times/week of muscle strength activity or (≥150 min/week of moderate aerobic activity or ≥75 min/week of heavy aerobic activity or an equivalent combination) & ≥2 times/week of muscle strength activity or (≥180 min/week of moderate aerobic activity or ≥90 min/week of heavy aerobic activity or an equivalent combination) & 1

			time/week of muscle strength activity or (≥100 min/week of moderate aerobic activity or ≥50 min/week of heavy aerobic activity or an equivalent combination) & 3 time/week of muscle strength activity)
M3_mMOS_instumental_support	numeric (double)	1-5	Modified Medical Outcomes Study Social Support Survey, instrumental support
M3_mMOS_emotional_support	numeric (double)	1-5	Modified Medical Outcomes Study Social Support Survey, emotional support
M3_mMOS_social_support_total	numeric (double)	1-5	Medical Outcome Study, total (average) social support
mastectomy	factor	2 levels 0: No 1: Yes	Had mastectomy procedure (all patients have undergone surgery: lumbetomy mastectomy lum)
surgery.3	factor	2 levels 0: No 1: Yes	Had surgery before M3
antiher2_treatment	factor	2 levels 0: No 1: Yes	Had anti-HER2 treatment

antiher2_regimen	factor	3 levels 0: None 1: Trastutzumab 2: Trastutzumab plus Pertuzumab	Anti-HER2 regimen followed
cancer_stage	factor	3 levels 1: Stage 1 2: Stage 2 3: Stage 3	Cancer stage
cancer_grade	factor	3 levels 1: Grade 1 2: Grade 2 3: Grade 3	Cancer grade
baseline_pr	numeric (double)	0-100	Progesterone receptors positivity (percentage score) (M0)
baseline_ki67	numeric (double)	0-100	ki67 percentage score (M0)
baseline_pt	numeric (double)		Tumor size in mm
baseline_pn	factor	4 levels 1: N0 2: N1 3: N2 4: N3	Node (N) describes whether the cancer has spread to the lymph nodes (M0)
baseline_histological_type	factor	3 levels 1: Ductal 2: Lobular 3: Other	Histological type classification (M0)
family_history	factor	2 levels 0: No 1: Yes	Family history of cancer (first degree relatives)
baseline_hormone_replacement_pre_treatment	factor	2 levels 0: No 1: Yes	Had hormone replacement therapy before treatment (M0)
m0_menopausalstatuspre	factor	3 levels 1: premenopausal 2: perimenopausal 3: postmenopausal	Menopausal status pretreatment (M0)
M0_performancestatus	factor	4 levels	ECOG performance status (M0)

		0: Fully active, able to carry on all pre-disease performance without restriction 1: Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work 2: Ambulatory and capable of all selfcare but unable to carry out any work activities; up and about more than 50% of waking hours 3: Capable of only limited selfcare; confined to bed or chair more than 50% of waking hours 4: Completely disabled; cannot carry on any selfcare; totally confined to bed or chair 5: Dead	
baseline_creatinine	numeric (double)		Creatinine (M0)
baseline_alt	numeric (double)		Alanine aminotransferase (M0)
baseline_hb	numeric (double)		Haemoglobin (M0)
baseline_leukocytes	numeric (double)		Leykocytes (M0)
baseline_neutrofiles	numeric (double)		Neutrofiles (M0)
baseline_thrombocytes	numeric (integer)		Thrombocytes (M0)
chemo0_type	factor	3 levels 0: No chemo	Had and type of chemotherapy

		1: neo-adjuvant chemo 2: adjuvant chemo	
radiotherapy	factor	2 levels 0: No 1: Yes	Had radiotherapy
TamoxifenVSother	factor	4 levels 0: No endocrine treatment 1: Tamoxifen 2: AI (letrozol, anastrozole, exemestane) 3: Ovarian suppression plus tamoxifen or AI	Endocrine (hormonal) treatment regimen AI: aromatase inhibitors
chemo.3	factor	4 levels 0: No chemotherapy at M3 1: During chemotherapy at M3 2: Within 3 months after the end of chemotherapy at M3 3: More than three months after the end of chemotherapy at M3	Chemotherapy status at M3
radio.3	factor	4 levels 0: No radiotherapy at M3 1: During radiotherapy at M3 2: Within 3 months after the end of radiotherapy at M3 3: More than three months after the end of radiotherapy at M3	Radiotherapy status at M3
antiher2.3	factor	4 levels 0: No antiher2 treatment at M3	Anti-HER2 treatment status at M3

		1: During antiher2 treatment at M3 2: Within 3 months after the end of antiher2 treatment at M3 3: More than three months after the end of antiher2 treatment at M3	
endocrine.3	factor	2 levels 0: No endocrine at M3 1: During endocrine treatment	Endocrine treatment status at M3
care_team	factor	4 levels 1: CHAMP 2: IEO 3: HUS 4: HUJI	Clinical site (hospital)
age_baseline	numeric (integer)		Age at baseline (M0)
m0_marital	factor	4 levels 1: Married 2: Separated/divorced or Widowed 3: Common-law partner 4: Single or Engaged	Marital status (M0)
m0_employment	factor	2 levels 1: Employed full time, Self-employed, Retired 2: Employed part time, Housewife, Unemployed	Employment status (M0)
baseline_ki67class	factor	2 levels 0: <20% 1: ≥20%	ki67 class (M0)

baseline_erclass	factor	2 levels 0: Negative 1: Positive	ER-negative or ER-positive breast cancer (M0) ER -> Estrogen Receptor A cutoff of 1% has been used
baseline_prclass	factor	2 levels 0: Negative 1: Positive	PR-negative or PR-positive breast cancer (M0) PR -> Progesterone Receptor A cutoff of 1% has been used Reference: Validity of 1% Hormonal Receptor Positivity Cutoff by the ASCO/College of American Pathologists Guidelines at the Georgia Cancer Center Firas Kreidieh, Ramses F. Sadek, Li Fang Zhang, Aaron Gopal, Jean-Pierre Blaize, David Yashar, Reena Patel, Hiral S. Patel, Shou-Ching Tang, and Houssein Abdul Sater JCO Precision Oncology 2022 :6
baseline_LuminalB_f	factor	2 levels 0: No 1: Yes	Whether tumor is luminal B (ER positive, PR any and HER2 positive) (M0)

baseline_subtypes	factor	5 levels 1: Luminal A 2: Luminal B-like (HER2 negative) 3: Luminal B-like (HER2 positive) 4: Her2-positive 5: Triple-negative	Luminal A: ER+, PR+, HER2-, low Ki67 (<20%) Luminal B-like (HER2 negative): ER+, PR+/-, HER2-, high Ki67 (≥20%) or ER+, PR-, HER2-, Ki67 any Luminal B-like (HER2 positive): ER+, PR+/-, HER2-, any Ki67 HER2-positive (non luminal): ER-, PR-, HER2+, any Ki67 Triple-negative: ER-, PR-, HER2-, any ki67 References: 1. Karihtala P and Jukkola A (2020) High Parity Predicts Poor Outcomes in Patients With Luminal B-Like (HER2 Negative) Early Breast Cancer: A Prospective Finnish Single-Center Study. Front. Oncol. 10:1470. doi: 10.3389/fonc.2020.01470 2. Senkus, E., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rutgers, E., Zackrisson, S., Cardoso, F., & ESMO

			Guidelines Committee (2015). Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology : official journal of the European Society for Medical Oncology, 26 Suppl 5, v8–v30. https://doi.org/10.1093/annonc/mdv298
baseline_NLR	numeric (double)		Neutrophil to leukocyte ratio (M0) Comment: lymphocytes were not provided, abbreviation is misleading
nccn_distress_thermometer.0	numeric (double)	0-10	NCCN distress thermometer (M0)
Emot_Fun_QLQ30.0	numeric (double)	0-100	Quality of Life Questionnaire (M0)
nccn_distress_thermometer.3	numeric (double)	0-10	NCCN distress thermometer (M3)
Emot_Fun_QLQ30.3	numeric (double)	0-100	Quality of Life Questionnaire (M3)
Sex_Enjoy_BR23.0	numeric (double)	0-100	QLQ-BR23 - EORTC Quality of life questionnaire breast cancer module, sexual enjoyment (M0)
Sex_Enjoy_BR23.3	numeric (double)	0-100	QLQ-BR23 - EORTC Quality of life questionnaire breast cancer module, sexual enjoyment (M3)