Αξιολόγηση της Επίδοσης Τεχνικών Μηχανικής Μάθησης στην Ταξινόμηση του Συνόλου Δεδομένων Wisconsin Breast Cancer

ΓΚΙΡΚΙΖΑΣ ΡΩΜΑΝΟΣ

Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας, Λαμία, Ελλάδα

Περίληψη

Η παρούσα εργασία πραγματεύεται μία συγκριτική μελέτη μεταξύ διάφορων ταξινομητών και συγκεκριμένα της επίδοσής τους όσον αφορά την αποτελεσματικότητα διάγνωσης Καρκίνου του Μαστού με δεδομένα που προέρχονται από το σετ δεδομένων "Wisconsin Breast Cancer (Original)" [1]. Συγκεκριμένα, οι ταξινομητές που χρησιμοποιήθηκαν και αξιολογήθηκαν είναι οι Naïve Bayes (NB), MultilayerPerceptron (MLP), K-nearest neighbors (K-nn), Sequential Minimal Optimization (SMO), C4.5 και Random Forest (RF). Τα αποτελέσματα φανερώνουν μία ελαφριά υπεροχή του NB αλγορίθμου στην ταξινόμηση του συγκεκριμένου dataset όσον αφορά τα πολλαπλά μέτρα αξιολόγησης σε σχέση με τους υπόλοιπους ταξινομητές. Η εργασία αυτή περιλαμβάνει την απλή και άμεση μετάβαση από τη θεωρία στην πράξη μέσω της υλοποίησης ταξινόμησης με τις παραπάνω μεθόδους και η παρουσίαση της αποτελεσματικότητάς τους στην επίλυση του προαναφερθέντος προβλήματος.

Λέξεις-κλειδιά: *Καρκίνος του Μαστού, Ταξινόμηση, Wisconsin Breast Cancer Dataset (WBCD)*

I. Εισαγωγη-Υποβαθρο

Ο Καρκίνος συγκαταλέγεται, ως γνωστόν, στις πιο θανατηφόρες ασθένειες του κόσμου. Ειδικότερα, ο καρκίνος του μαστού προκάλεσε τον θάνατο σε 685.000 γυναίκες παγκοσμίως το 2020, ύστερα από 2.3 εκατομμύρια διαγνώσεις, σύμφωνα με στοιχεία του Παγκοσμίου Οργανισμού Υγείας [7]. Η διάσταση, καθώς και η σοβαρότητα των αποτελεσμάτων αυτού του καρκινικού είδους απαιτεί την άμεση εύρεση νέων και τη βελτιστοποίηση των ήδη υπαρχόντων μέσων υγείας για την πρόληψη, διάγνωση και αντιμετώπισή του. Σε αυτό το πλαίσιο καθίστανται ιδιαιτέρως χρήσιμοι οι αλγόριθμοι Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης, οι οποίοι με την ικανότητα που έχουν να επεξεργάζονται και να κατηγοριοποιούν δεδομένα, καθώς και με μια πληθώρα άλλων δυνατοτήτων, αποτελούν σημαντική βοήθεια για τους ειδικούς που τείνουν να τους εμπιστεύονται όλο και περισσότερο [8],[9]. Έτσι, ο στόχος αυτής της εργασίας είναι να πραγματοποιηθεί μία συγκριτική ανάλυση μεταξύ 6 επιλεγμένων ταξινομητών. Στα πλαίσια της εργασίας, μελετήθηκε μία σειρά συναφών εργασιών και επιστημονικών μελετών, προκειμένου να αναφερθούν τα αποτελέσματα αντίστοιχων πειραμάτων πάνω στο ίδιο dataset. Ο Ahmed *et al.* [2] το 2020 κατέληξε στο συμπέρασμα ότι με ακρίβεια της τάξης του 97.27%, ο NB αποτελεί την καταλληλότερη μέθοδο ταξινόμησης. Ο Mohammed *et al*. το 2020 [3] απέδειξε ότι ο SMO είναι αυτός που αποδίδει καλύτερα (99.56%), ενώ παρόμοιο αποτέλεσμα βρήκε το 2018 ο Obaid *et al.* [4] με το Support Vector Machine να είναι ο πιο ακριβής ταξινομητής (98.1%). Η Rosly *et al.* [5], επίσης το 2018, συνδύασαν τη μέθοδο bagging με τον ταξινομητή NB προκειμένου να πετύχουν ακρίβεια 97.51% στην ταξινόμηση των δεδομένων, ενώ σε παλαιότερη βιβλιογραφία [6] στην οποία μελετήθηκαν 3 μοντέλα νευρωνικών δικτύων 1) Adaptive Resonance Theory based neural network (ART), 2) Self Organizing Map based neural network (SOM) και 3) Back Propagation Neural network (BPN) επιτεύχθηκε από τη Mumtaz *et al*. ακρίβεια διάγνωσης 91-100% με την τρίτη τεχνική, δυσαναλόγως του πλήθους δειγμάτων που χρησιμοποιήθηκαν.

Στη συνέχεια, θα περιγραφεί η μεθοδολογία της εργασίας, δηλαδή από τι αποτελείται το dataset και θα αναλυθεί η πειραματική διαδικασία (2η ενότητα), ακολούθως θα παρουσιαστούν τα αποτελέσματα αυτής (3η ενότητα), και τέλος βρίσκονται τα συμπεράσματα που εξάγονται και οι προτάσεις μου για μελλοντικές εργασίες ή βελτιώσεις (4η ενότητα).

II. Μεθοδολογια

Το WBCD αποτελούνταν από 701 από τις κλινικές υποθέσεις του Dr.Wolberg από τις οποίες αφαιρέθηκαν οι 2 από το αρχικό σύνολο, αφήνοντάς το με 699 δείγματα. Αυτές οι καταγραφές λάβανε χώρα από το 1989 μέχρι το 1991.Το repository του UCI, Breast Cancer Wisconsin (Original), αποτελεί ένα σετ ταξινόμησης 2 κλάσεων, όπου το πρόβλημα είναι ο καρκίνος του μαστού και τα ενδεχόμενα αποτελέσματα είναι το να είναι μη-καρκινογόνος ή benign, ή το να είναι καρκινογόνος ή malignant. Ο διαχωρισμός μεταξύ των 2

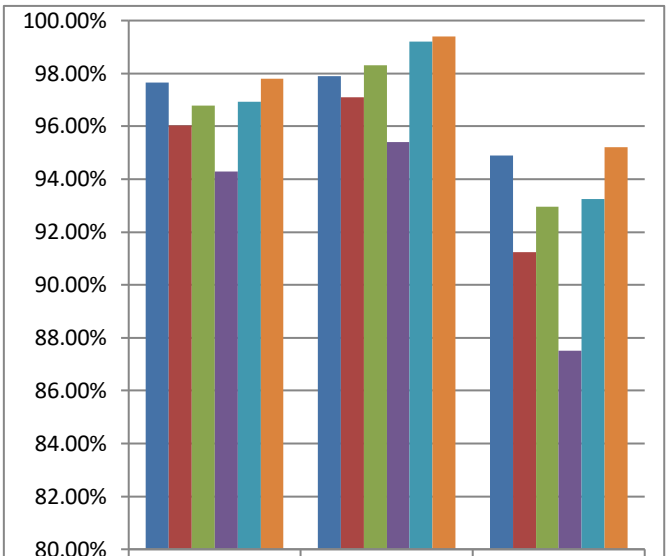
κατηγοριών παρατηρείται μέσω της καταγραφής των τιμών 10 χαρακτηριστικών, τα οποία είναι τα εξής: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses και τέλος class (2 για benign και 4 για malignant). Κάθε ένα από αυτά, εκτός της κλάσης, λαμβάνει μία τιμή από 1 έως 10. Επιπλέον, υπάρχει και ένας χαρακτηριστικός κωδικός για την κάθε περίπτωση (sample code number).

Πριν την έναρξη των πειραμάτων πήρα την απόφαση να αφαιρέσω 16 διανύσματα από το σύνολο του dataset, λόγω του γεγονότος ότι είχαν το καθένα από αυτά ελλιπή τιμή για το χαρακτηριστικό Bare_nuclei και όντας ιατρικά δεδομένα θεώρησα σωστό να μην τα λάβω υπόψιν στον υπολογισμό των κλάσεων ή να βάλω στη θέση τους τη μέση τιμή του υπόλοιπου σετ. Από τα 16 αυτά instances, τα 2 ήταν κλάσης malignant και τα 14 κλάσης benign. Επιπλέον, η τιμή του Bare_nuclei είναι ακέραιος μεταξύ 1 και 10, ενώ ο μέσος όρος των υπολοίπων δεκαδικός. Συμπερασματικά, το σετ το οποίο ταξινομήθηκε στα πλαίσια αυτής της εργασίας αποτελείται από 683 διανύσματα, τα οποία όμως δεν περιέχουν ελλιπείς τιμές. Τα 444 ανήκουν στην κλάση benign, ενώ τα υπόλοιπα 239 στην κλάση malignant, πράγμα που σημαίνει ότι το εν λόγω dataset είναι ελαφρώς imbalanced. Θα μπορούσε σε μια μελλοντική μελέτη να διερευνηθεί η αποτελεσματικότητα των ταξινομητών στο ίδιο dataset έπειτα από την επιλογή κάποιου φίλτρου με το οποίο να πραγματοποιούνταν η εξισορρόπηση των δεδομένων.

Η συγκριτική μελέτη της εργασίας πραγματοποιήθηκε με την αρωγή των εργαλείων της WEKA, καθώς διαθέτει μια πληθώρα ταξινομητών, από τους οποίους εγώ επέλεξα να χρησιμοποιήσω τον IBk (K-nn), ρυθμισμένο να λαμβάνει υπόψιν τους 3 κοντινότερους γείτονες, τον Naïve Bayes, το MultilayerPerceptron, τον J-48 (C4-5), τον Random Forest και τον SMO. Η επιλογή των συγκεκριμένων ταξινομητών έγινε με γνώμονα την χρήση τους από την βιβλιογραφία η οποία μελετήθηκε. Κατά τη διάρκεια της μελέτης επιλέχθηκαν διαφορετικές τιμές για κάποια από τα χαρακτηριστικά των ταξινομητών. Συγκεκριμένα, πειραματίστηκα με το πλήθος των κοντινών γειτόνων στον k-nn, όπως προανέφερα, με το kernel και το c στο SMO, με βέλτιστα αποτελέσματα να εμφανίζονται χρησιμοποιώντας RBFKernel και c=1, ενώ στο MLP δοκίμασα διαφορετικό πλήθος hidden layer, διαπιστώνοντας ότι με 1 κρυφό επίπεδο είναι περισσότερο τελεσφόρος. Η εκτέλεση των πειραμάτων έγινε με την αυτόματη επιλογή της WEKA για cross-validation 10 folds.

III. Αποτελεσματα

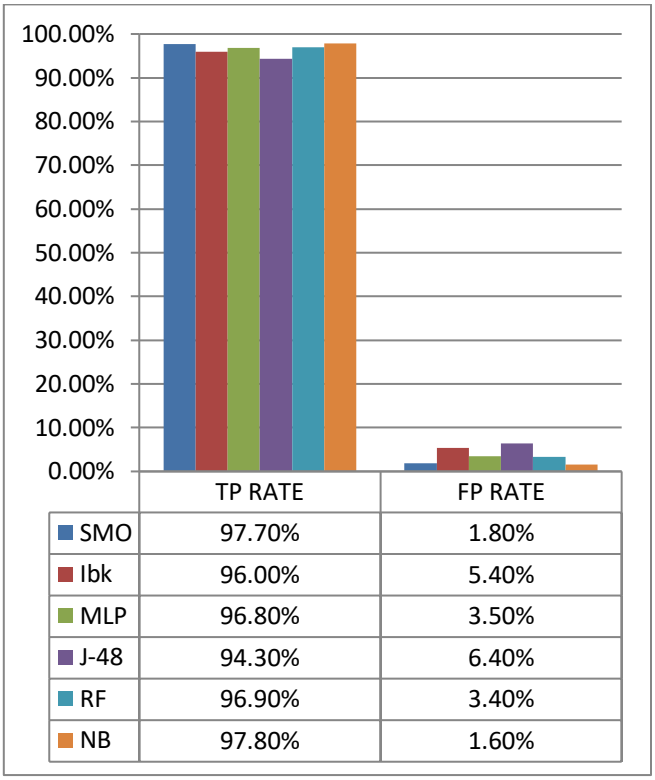
Τα αποτελέσματα που φαίνονται στην εικόνα 1, φανερώνουν ότι με βάση 3 βασικά μέτρα αξιολόγησης των ταξινομητών, το Accuracy, την καμπύλη ROC και το Kappa statistic, ο Naïve Bayes επικρατεί των άλλων 5 ταξινομητών, έχοντας ποσοστό 97.80%, 99.40% και 95.22% στα αντίστοιχα μέτρα. Όσον αφορά την «ακρίβεια», ακολουθεί ο SMO με ποσοστό πολύ κοντά σε αυτό του NB (97.66%) και ο Random Forest με 96.93%. Ακολουθεί ο MLP με ποσοστό 96.78%. Τη δεύτερη καλύτερη απόδοση στην καμπύλη ROC την πέτυχε ο RF με ποσοστό ελάχιστα μικρότερο του NB, 99.20%, ενώ στην τρίτη θέση βρέθηκε ο MLP με 98.30%. Τέλος, ο SMO έχει και το δεύτερο καλύτερο ποσοστό Kappa statistic με 94.90% και το τρίτο καλύτερο παρατηρείται από τον RF με 93.26%.



	Accuracy	ROC	Kappa statistic
SMO	97.66%	97.90%	94.90%
Ibk	96.04%	97.10%	91.25%
MLP	96.78%	98.30%	92.95%
J-48	94.29%	95.40%	87.51%
RF	96.93%	99.20%	93.26%
NB	97.80%	99.40%	95.22%

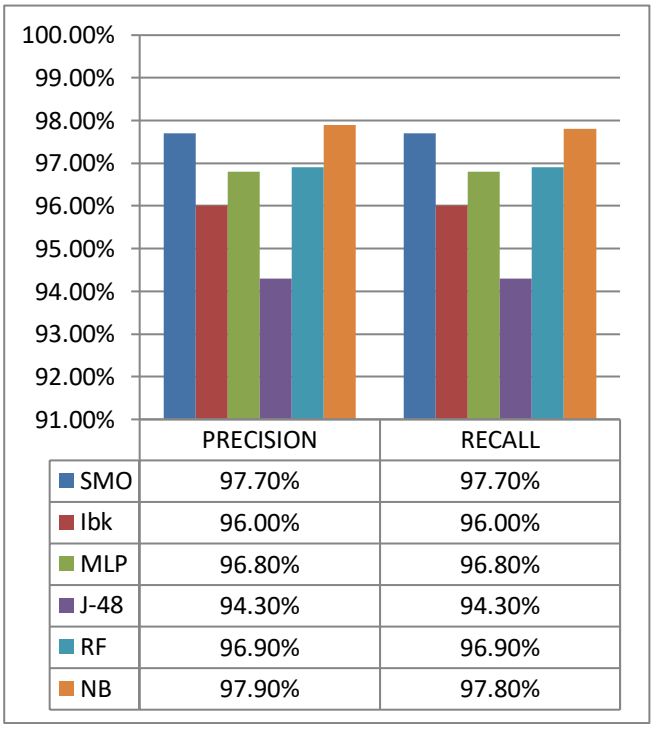
Εικόνα 1

Στην εικόνα 2 και 3 απεικονίζονται οι καταγραφές κάποιων άλλων μέτρων αξιολόγησης για τους ταξινομητές που εξέτασα, και συγκεκριμένα του Recall, του Precision, του FalsePositive Rate και του TruePositive Rate. Όπως και πριν, έτσι και σ'αυτά τα μέτρα σύγκρισης ο ταξινομητής Naïve Bayes αποδεικνύεται πιο τελεσφόρος από τους υπόλοιπους ταξινομητές έχοντας 97.80% TP Rate, μόλις 1.60% FP Rate, 97.90% Precision και 97.8O% Recall, ενώ ο δεύτερος τη τάξει ταξινομητής αποδεικνύεται εκ νέου ο SMO με ποσοστά 97.70%, 1,80%, 97,70% και 97,70% στα αντίστοιχα μέτρα αξιολόγησης. Στον αντίποδα, μπορούμε να παρατηρήσουμε πως ο J48 έχει τη χειρότερη επίδοση σε κάθε μία από τις διαφορετικές αξιολογήσεις.



	TP RATE	FP RATE
SMO	97.70%	1.80%
Ibk	96.00%	5.40%
MLP	96.80%	3.50%
J-48	94.30%	6.40%
RF	96.90%	3.40%
NB	97.80%	1.60%

Εικόνα 2



	PRECISION	RECALL
SMO	97.70%	97.70%
Ibk	96.00%	96.00%
MLP	96.80%	96.80%
J-48	94.30%	94.30%
RF	96.90%	96.90%
NB	97.90%	97.80%

Εικόνα 3

Στους διπλανούς confusion matrices μπορούμε να εξακριβώσουμε την αξιοπιστία των αποτελεσμάτων που εξήχθησαν από τα μέτρα αξιολόγησης που αναφέρθηκαν.
Ο NB έχει τα περισσότερα ορθά predicted διανύσματα όσον αφορά την κλάση malignant (4) με 237/239, και την τρίτη καλύτερη όσον αφορά την κλάση benign (2) με 431/444.

Ο RF βρίσκεται κοντά έχοντας 230/239 και 432/444 αντίστοιχα. Ο IBk έχει την καλύτερη απόδοση ταξινόμησης για την κλάση benign με 434/444, αλλά υστερεί στην ταξινόμηση της κλάσης malignant με μόλις 222/239. Ο SMO κινείται στο ίδιο περίπου επίπεδο με τον NB με μία καλή πρόγνωση της μη καρκινογόνου κλάσης 431/444, αλλά πολύ καλύτερη επίδοση στην πρόγνωση της καρκινογόνου 236/239. Ο MLP έχει την 3η καλύτερη επίδοση (μαζί με τον RF) όσον αφορά την πρόγνωση των καρκινογόνων διανυσμάτων 230/239 και μία εξίσου καλή πρόγνωση σε σχέση με τους προαναφερθέντες ταξινομητές στην πρόγνωση των μη-καρκινογόνων 431/444. Τελευταίος σε επιδόσεις έρχεται ο J48 με 422/444 και 222/239 αντίστοιχα.

		Predicted	
	class	**2**	**4**
Actual	**2**	431	13
	4	2	237

Πίνακας 1: NB Confusion Matrix

		Predicted	
	class	**2**	**4**
Actual	**2**	431	13
	4	9	230

Πίνακας 2: MLP Confusion Matrix

		Predicted	
	class	**2**	**4**
Actual	**2**	431	13
	4	3	236

Πίνακας 3: SMO Confusion Matrix

		Predicted	
	class	**2**	**4**
Actual	**2**	434	10
	4	17	222

Πίνακας 4: IBk Confusion Matrix

		Predicted	
	class	**2**	**4**
Actual	**2**	422	22
	4	17	222

Πίνακας 5: J48 Confusion Matrix

		Predicted	
	class	**2**	**4**
Actual	**2**	432	12
	4	9	230

Πίνακας 6: RF Confusion Matrix

IV. ΣΥΜΠΕΡΑΣΜΑΤΑ

Ο καρκίνος του μαστού αποτελεί μία από τις μείζονες θανατηφόρες ασθένειες πλήττοντας μεγάλο αριθμό γυναικών παγκοσμίως καθιστώντας την πρόγνωση, την διάγνωση και την πάταξή του επιτακτική ανάγκη της κοινωνίας. Η μηχανική μάθηση και οι αλγόριθμοι της, οι οποίοι χρησιμοποιούνται συχνά από τους επιστήμονες και ερευνητές του κλάδου έχουν τη δυνατότητα να δώσουν μια ανάσα. Κατέληξα στο γεγονός ότι στην συγκεκριμένη πειραματική μελέτη ο ισχυρότερος από τους 6 ταξινομητές που συγκρίθηκαν στην πρόγνωση καρκίνου του μαστού στα δείγματα του WBCD είναι ο Naïve Bayes. Σε επέκταση αυτής της εργασίας θα μπορούσε στο μέλλον να δοκιμαστεί η αποτελεσματικότητα του συνδυασμού των αποδοτικότερων από τους ταξινομητές της παρούσας εργασίας καθώς και η περαιτέρω προεπεξεργασία των δεδομένων, με σκοπό την αύξηση της ακρίβειας πρόγνωσης,.

ΑΝΑΦΟΡΕΣ

[1] Wolberg,WIlliam. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Ahmed, M. T., Imtiaz, M. N. and Karmakar, A. (2020). Analysis of Wisconsin Breast cancer original dataset using data mining and machine learning algorithms for breast cancer prediction. Journal of Science, Technology and Environment Informatics, 09(02), 665-672

[3] Mohammed, S.A., Darrab, S., Noaman, S.A., Saake, G. (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan, Y., Shi, Y., Tuba, M. (eds) Data Mining and Big Data. DMBD 2020. Communications in Computer and Information Science, vol 1234. Springer, Singapore. https://doi.org/10.1007/978-981-15-7205-0_10.

[4] Ibrahim Obaid, Omar & Mohammed, Mazin & Abd Ghani, Mohd Khanapi & Mostafa, Salama & Al-Dhief, Fahad. (2018). Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer. International Journal of Engineering and Technology. 7. 160-166. 10.14419/ijet.v7i4.36.23737.

[5] Rosly, R., Makhtar, M., Awang, M.K., Awang, M.I., & Rahman, M.N. (2018). Analyzing performance of classifiers for medical datasets. *International journal of engineering and technology*, *7*, 136.

[6] K. Mumtaz, S. A. Sheriff and K. Duraiswamy, "Evaluation of three neural network models using Wisconsin breast cancer database," 2009 International Conference on Control, Automation, Communication and Energy Conservation, Perundurai, India, 2009, pp. 1-7.

[7] WHO. (2023, July 12). “Breast Cancer”. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[8] WHO. (2021, June 28). “WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use”. Retrieved from https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use .

[9] WHO. (2023, October 19). “WHO outlines considerations for regulation of artificial intelligence for health”. Retrieved from https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health .