Prediction of Cervical Cancer

(Πρόβλεψη του καρκίνου του τραχήλου της μήτρας)

Σταύρος Γαρδέλης
Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων
ΠΜΣ "Μεγάλα Δεδομένα και Αναλυτική"
Μηχανική Μάθηση: Μέθοδοι και Αλγόριθμοι
Πειραιάς, Ελλάδα
E-mail : st_gardelis@hotmail.com

Τζαβάρας Ιωάννης
Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων
ΠΜΣ "Μεγάλα Δεδομένα και Αναλυτική"
Μηχανική Μάθηση: Μέθοδοι και Αλγόριθμοι
Πειραιάς, Ελλάδα
E-mail : tzavaras95@gmail.com

Περίληψη: Ο καρκίνος του τραχήλου της μήτρας είναι από τους συχνότερους γυναικολογικούς καρκίνους παγκοσμίως. Αποτελεί τη 2η αιτία θανάτου από καρκίνους των γυναικών [1]. Στο παρόν έγγραφο παρουσιάζεται η πρόβλεψη του αποτελέσματος της βιοψίας ενός ασθενή , λαμβάνοντας υπόψη συγκεκριμένα χαρακτηριστικά και δεδομένα ασθενών που αφορούν δημογραφικές πληροφορίες, συνήθειες καθώς και ιστορικά ιατρικά αρχεία.

Τα δεδομένα συλλέχθηκαν από το νοσοκομείο «Hospital Universitario de Caracas» της Βενεζουέλας . Το δείγμα περιέχει 32 χαρακτηριστικά από 858 ασθενείς. Λόγω του ότι υπάρχουν χαμένες τιμές χρησιμοποιήθηκαν τεχνικές για τη συμπλήρωσή τους, είτε απαλοιφή συγκεκριμένων χαρακτηριστικών που ούτως ή άλλως δεν επηρέαζαν το τελικό αποτέλεσμα. Πραγματοποιήθηκαν machine learning τεχνικές , όπως οι παρακάτω αλγόριθμοι: KNN, Bayesian classifiers , νευρωνικά δίκτυα(MLPClassifier) και Support Vector Machines (SVM).

Keywords : Cervical cancer , machine learning , neural networks, Bayesian classifiers, KNN, SVM, Over-sampling, Under-sampling, Feature selection, Lasso, Pearson Correlation

I. ΕΙΣΑΓΩΓΗ

Η μήτρα είναι το όργανο μέσα στο οποίο αναπτύσσεται ένα έμβρυο κατά τη διάρκεια της εγκυμοσύνης. Στο κατώτερο τμήμα της υπάρχει ένα στενό πέρασμα μήκους 2-3 εκατοστών που αποτελεί τον τράχηλο της μήτρας και βρίσκεται στο σημείο που τελειώνει ο κόλπος. Συνήθως οι όγκοι που αναπτύσσονται στην περιοχή αυτή ξεκινούν από τη λεγόμενη μεταβατική ζώνη. Πρόκειται για μια περιοχή όπου συναντάται το εσωτερικό μέρος του τραχήλου της μήτρας (ενδοτράχηλος), που βρίσκεται πιο κοντά στο σώμα της μήτρας, με το εξωτερικό μέρος του τραχήλου της μήτρας (εξωτράχηλος). Ο καρκίνος του τραχήλου της μήτρας αναπτύσσεται αργά και για πολλά χρόνια. Συνήθως πριν την ανάπτυξή του καρκίνου, στους ιστούς της περιοχής συμβαίνουν αλλαγές σε κυτταρικό επίπεδο (που ονομάζονται δυσπλασίες ή προκαρκινικές βλάβες).

Σε αναπτυσσόμενα κράτη, ο καρκίνος του τραχήλου της μήτρας αποτελεί τον 2ο σε συχνότητα καρκίνο στις γυναίκες μετά τον καρκίνο του μαστού. Η συχνότητα του καρκίνου του τραχήλου είναι 6-10/100.000 με ετήσιο ρυθμό θανάτων 3/100.0001. Στις αναπτυσσόμενες χώρες αναφέρονται περισσότερα από 450.000 περιστατικά κάθε χρόνο [2]. Ο συνολικός κίνδυνος εμφάνισης κατά τη διάρκεια της ζωής μιας γυναίκας είναι γύρω στο 5% σε περιοχές της Αφρικής, της Ινδίας και της Λατινικής Αμερικής και 1% στην Ευρώπη και τη Βόρεια Αμερική.

Οι γυναίκες χαμηλότερων κοινωνικοοικονομικών τάξεων παρουσιάζουν τη μεγαλύτερη συχνότητα προσβολής κυρίως λόγω της μη επαρκούς πρόσβασης σε συστήματα μαζικού ελέγχου (Screening) [3]. Το πλέον διαδεδομένο τεστ Παπανικολάου είναι ένα είδος διαγνωστικού ελέγχου. Ανακαλύφθηκε από τον ιατρό Γεώργιο Παπανικολάου στο πανεπιστήμιο του Cornell τη δεκαετία του 1940 και αποτελεί μια από τις σημαντικότερες ανακαλύψεις του 20ου αιώνα στην πρόληψη ασθενειών. Ανακαλύπτει κυτταρικές ανωμαλίες που αν δεν αντιμετωπιστούν έγκαιρα οδηγούν σε καρκίνο.

Τα τελευταία 60 χρόνια που έχει υιοθετηθεί η εξέταση αυτή, στην Ευρώπη και στις Ηνωμένες Πολιτείες Αμερικής, έχει μειωθεί το ποσοστό θνησιμότητας στο 50%. Παρόλα αυτά στον υπόλοιπο κόσμο παραμένει πολύ υψηλό αυτό το ποσοστό[4] . Το 88% θανάτων από τον καρκίνο του τραχήλου της μήτρας, παρατηρούνται από χώρες με μειωμένους πόρους για πρόληψη (τεστ ΠΑΠ) και θεραπεία. Τα υψηλότερα ποσοστά εμφανίζονται σε περιοχές με υψηλής μετάδοσης του ιού ανθρωπίνων θηλωμάτων (HPV) και του ιού της ανθρώπινης ανοσοανεπάρκειας (HIV), συγκεκριμένα της Λατινικής Αμερικής, της νοτίως της Σαχάρας Αφρικής και της Νότιας Ασίας .

Λόγω των μη επαρκών πόρων στις μη αναπτυσσόμενες χώρες, δημιουργήθηκε η ανάγκη για ανίχνευση του συγκεκριμένου καρκίνου από απλά χαρακτηριστικά, όπως το κάπνισμα, η διατροφή , η ηλικίας πρώτης σεξουαλικής επαφής, αντισυλληπτικές μέθοδοι και άλλα. Εδώ και κάποια χρόνια έχουν αρχίσει οι προσπάθειες πρόβλεψης του καρκίνου μέσω λογισμικών μηχανικής μάθησης. Στο παρόν έγγραφο παρουσιάζεται ένα προγνωστικό μοντέλο βάσει απλών χαρακτηριστικών που αναφέρθηκαν παραπάνω. Η συγκεκριμένη μεθοδολογία στηρίζεται στο πλεονέκτημα ότι ξέρουμε τα αποτελέσματα για κάθε ασθενή και έτσι η ακρίβεια της πρόβλεψης μπορείς να είναι πιο ακριβής. Στηρίζεται σε μεθόδους επιβλεπόμενης μηχανικής μάθησης (Supervised machine learning).

A. Περιγραφή παρόμοιων εργασιών

Παρόμοιες εργασίες, οι οποίες αφορούν το cervical cancer dataset αναφέρουν το κύριο πρόβλημα του dataset το οποίο είναι ανισορροπία του δείγματος(imbalanced data). Οι Al-Wesabi, Avishek Choudhury και Daehan Won στην εργασία τους [5] αναφέρουν την αναγκαιότητα για επεξεργασία του αρχικού dataset και αντιμετωπίζουν το

πρόβλημα των imbalanced data με τεχνικές over και under sampling καθώς και την επιλογή χαρακτηριστικών με την μέθοδο Sequential Forward Selection (SFS). Σε άλλη δημοσίευση [6] οι Kelwin Fernandes, Davide Chicco, Jaime S. Cardoso και Jessica Fernandes προτείνουν ένα μία τεχνική μηχανικής μάθησης για μείωση διαστάσεων και αναπτύσσουν ένα μοντέλο πρόβλεψης το οποίο έχει εφαρμογή και σε άλλα παρόμοια σύνολα δεδομένων.file.

II. ΜΕΘΟΔΟΛΟΓΙΑ

Η παρούσα εργασία μπορεί να περιγραφή από τέσσερα βασικά μέρη. Αρχικά η προ-επεξεργασία του συνόλου δεδομένων καθώς και η ανάλυση των χαρακτηριστικών, η παρουσίαση των μετρικών με τις οποίες αξιολογήσαμε τα δεδομένα, η τροποποίηση του συνόλου δεδομένων με τεχνικές standardization, over-sampling, under-sampling και η εφαρμογή και σύγκριση τεσσάρων κατηγοριοποιητών KNN, SVM, NN's, GNB, και τέλος η ανακάλυψη των κυριότερων χαρακτηριστικών του καρκίνου της μήτρας με τις τεχνικές LASSO, Pearson correlation στο σύνολο δεδομένων μας καθώς και η σύγκριση των αποτελεσμάτων με παρόμοιες εργασίες.

IMBALANCED DATA

Τα δεδομένα μας στην κλάση στόχο χαρακτηρίζονται από ανισορροπία (Class Imbalance Problem). Δηλαδή, υπάρχει μεγάλη συχνότητα εμφάνισης της μίας εκ των δύο τιμών που εμφανίζονται στην κλάση στόχο. Σε τέτοιες περιπτώσεις, πολλοί αλγόριθμοι κατηγοριοποίησης έχουν μικρό ποσοστό πρόβλεψης της λιγότερο εμφανιζόμενης κλάσης-τιμής στην κλάση στόχο. Υπάρχουν πολλές διαφορετικές προσεγγίσεις σε ένα τέτοιο πρόβλημα όπως η εφαρμογή cost-sensitive αλγορίθμων. Επιπλέον, σε περιπτώσεις όπου η κατανομή των τιμών στο train και test σύνολο δεδομένων (για παράδειγμα στο train σύνολο να έχουμε μεγάλη ανισορροπία ενώ στο test σύνολο δεδομένων να είναι πιο ισορροπημένα ή αντίστροφα) ενδείκνυνται οι τεχνικές over-sampling(δημιουργία αντιγράφων από τις εγγραφές) και under-sampling(επιλογή υποσυνόλου του συνόλου δεδομένων).[10]

Τα δεδομένα που χρησιμοποιούνται στην παρούσα εργασία περιέχουν ποσοστό True class (θεωρώντας ως True class την περίπτωση στην κλάση 'Biopsy' να υπάρχει η τιμή 1) μόλις 6,41% της συνολικής κλάσης.

Επιπλέον αφού τα δεδομένα μας είναι δεδομένα ιατρικού τύπου, μεγαλύτερο βάρος δόθηκε στην αύξηση της τιμής της μετρικής Sensitivity. Για το λόγο αυτό ενδιαφερόμαστε για το True Positive Rate, ο αλγόριθμός μας δηλαδή να προβλέπει όλες τις True class τιμές ή αλλιώς το ποσοστό των τιμών που κατηγοριοποιήθηκαν ως θετικές στο σύνολο των θετικών να είναι στο 100%.

III. ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

Το Cervical cancer (Risk Factors) Data Set περιέχει τα παρακάτω 32 χαρακτηριστικά , όπως επίσης και 4 κλάσεις, από τις οποίες θα επικεντρωθούμε κυρίως στη βιοψία.

TABLE I. ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Attribute	**Χαρακτηριστικό**	**Τύπος**
Age	Ηλικία	(int)
Number of sexual partners	Αριθμός ερωτικών συντρόφων	(int)
First sexual intercourse (age)	Ηλικία πρώτης σεξουαλικής επαφής	(int)
Num of pregnancies	Αριθμός εγκυμοσύνων	(int)
Smokes	Κάπνισμα	(bool)
Smokes (years)	Χρόνια καπνίσματος	(bool)
Smokes (packs/year)	Κάπνισμα πακέτα ανά χρόνια	(bool)
Hormonal Contraceptives	Αντισυλληπτικά	(bool)
Hormonal Contraceptives (years)	Αντισυλληπτικά χρόνια	(int)
IUD	Σπιράλ (τρόπος αντισύλληψης)	(bool)
IUD (years)	Σπιράλ χρόνια	(int)
STDs	Σεξουαλικώς Μεταδιδόμενα Νοσήματα (ΣΜΝ)	(bool)
STDs (number)	Σεξουαλικώς Μεταδιδόμενα Νοσήματα (ΣΜΝ) αριθμός	(int)
STDs:condylomatosis	Κονδυλώματα	(bool)
STDs:cervical condylomatosis	Κονδυλώματα στον τράχηλο της μήτρας	(bool)
STDs:vaginal condylomatosis	Κολπικά κονδυλώματα	(bool)
STDs:vulvo perineal condylomatosis	Κιδοιο περινεϊκή κονδυλωμάτωση	(bool)
STDs:syphilis	Σύφιλη	(bool)
STDs:pelvic inflammatory disease	Φλεγμονώδης νόσος της πυέλου	(bool)
STDs:genital herpes	ΕΡΠΗΣ γεννητικών οργάνων	(bool)
STDs:molluscum contagiosum	Μολυσματική τέρμινθος	(bool)
STDs:AIDS	AIDS	(bool)
STDs:HIV	Ιός της Ανθρώπινης Ανοσοανεπάρκειας	(bool)
STDs:Hepatitis	Ηπατίτιδα	(bool)
STDs:HPV – Human papilloma virus	Ιός των ανθρωπίνων θηλωμάτων	(bool)
STDs: Number of diagnosis	Αριθμός διάγνωσης	(int)
STDs: Time since first diagnosis	Ηλικία πρώτης διάγνωσης	(int)
STDs: Time since last diagnosis	Ηλικία τελευταίας διάγνωσης	(int)
Dx:Cancer Oncotype DX	Τεστ Ογκότυπος (Ανάλυση γονιδίων)	(bool)
Dx:CIN	Τεστ για "Ενδοεπιθηλιακή νεοπλασία του τραχήλου"	(bool)
Dx:HPV	Τεστ για "Ίός των ανθρωπίνων θηλωμάτων"	(bool)
Dx	Τεστ για επανεμφάνιση του καρκίνου	(bool)
Hinselmann: target variable	**Κολποσκόπηση**	**(bool)**
Schiller: target variable	**Schiller τεστ**	**(bool)**
Cytology: target variable	**κυτταρολογική εξέταση Παπανικολάου**	**(bool)**
Biopsy: target variable	**Βιοψία**	**(bool)**

Από το σύνολο δεδομένων λείπουν 3622 τιμές (ερωτήματα τα οποία δεν απαντήθηκαν) το οποίο αντιστοιχεί στο 11,7% του συνόλου δεδομένων (συνολικά 30888 τιμές). Όλα τα χαρακτηριστικά είναι αριθμητικά και οι κενές τιμές συμπληρώθηκαν με την μέση τιμή κάθε χαρακτηριστικού.

Από το σύνολο δεδομένων μας διαγράφηκαν τα χαρακτηριστικά:

1. STDs: Time since last diagnosis και
2. STDs: Time since first diagnosis

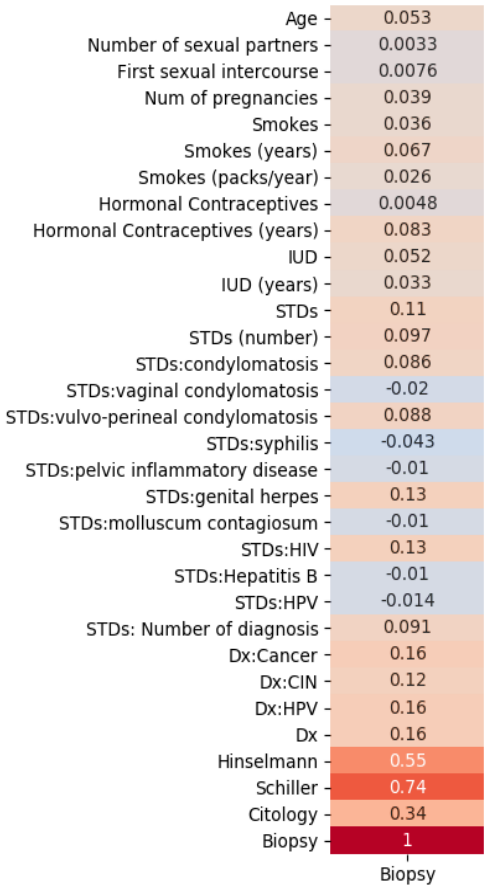
   (91,7% κενές τιμές),
3. STDs: AIDS και
4. STDs: cervical condylomatosis (μοναδική τιμή το 0).

Επίσης, από το σύνολο δεδομένων διαγράφηκαν 98 εγγραφές διότι περισσότερο από το 50% του δείγματος ήταν κενές τιμές).

Επιπλέον πληροφορίες που αντλούμε από το σύνολο δεδομένων είναι ότι :

- Οι ηλικίες των ερωτηθέντων κυμαίνεται από 13 έως 84 ετών με μέσο όρο ηλικίας 27 χρονών,
- Πάνω από το 93% των ερωτηθέντων ήταν έγκυος τουλάχιστον μία φορά,
- Κάτω του 14% είναι καπνιστές,
- Το 62,8% είχαν πρώτη σεξουαλικής επαφή πριν τα 18 τους.

Στον παρακάτω πίνακα παρουσιάζονται οι συσχετίσεις των χαρακτηριστικών με το target class (Biopsy).



Εικόνα 1. Correlation Matrix

IV. Περιγραφη μετρικων

Για την σύγκριση των κατηγοριοποιητών χρησιμοποιήθηκαν διάφορες μετρικές. Κυριότερες από αυτές ήταν accuracy, precision, sensitivity, AUC και f1-score. Οι μετρικές αυτές προέρχονται από τον υπολογισμό των:

TABLE II. PRDEICTION / LABEL (TP/FP/TN/FN)

	Prediction	**Label**	
TP (True Positives)	+	+	
FP (False Positives)	+	-	(Type error I)
TN (True Negatives)	-	-	
FN (False Negative)	-	+	(Type error II)

A. *Accuracy*

Ως Accuracy (Ακρίβεια) θεωρείτε ο βαθμός σωστών προβλέψεων στο σύνολο όλων των προβλέψεων. Η ακρίβεια μπορεί να οδηγήσει σε λάθος συμπεράσματα για το μοντέλο μας όταν υπάρχει ανισορροπία στο σύνολο τιμών της κλάσης πρόβλεψης (το μοντέλο δεν θα προβλέπει ποτέ ή σχεδόν ποτέ την τιμή κλάσης με λίγες εμφανίσεις).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. *Precision*

Εκφράζει το ποσοστό των τιμών που κατηγοριοποιήθηκαν ως θετικές και είναι όντως θετικές.

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

C. *Sensitivity*

Μετρική γνωστή και ως Recall, true positive rate ή hit rate. Εκφράζει το ποσοστό των τιμών που κατηγοριοποιήθηκαν ως θετικές, στο σύνολο των θετικών τιμών.

$$Sensitivity = Recall = \frac{TP}{(TP + FN)}$$

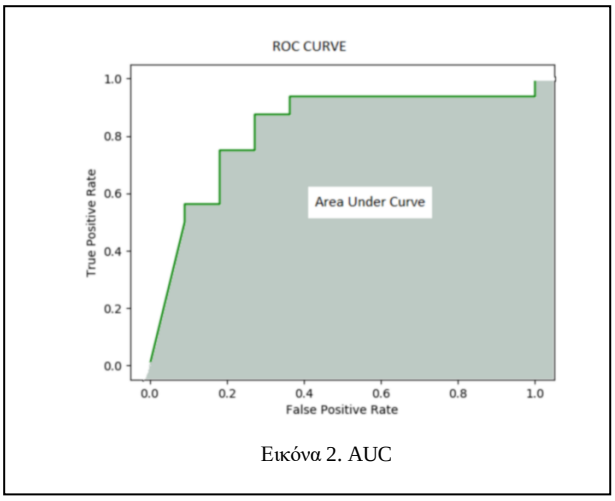
D. *F1-score*

Είναι ο αρμονικός μέσος των τιμών Recall και Precision.

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

E. AUC

Η μετρική AUC (**A**rea **U**nder The **C**urve) αφορά την περιοχή κάτω από την καμπύλη ROC curve. Η ROC ανάλυση εξετάζει την σχέση μεταξύ ευαισθησίας (sensitivity) και ακρίβειας (specificity) ενός binary κατηγοριοποιητή. Η καμπύλη ROC είναι το βασικό εργαλείο που χρησιμοποιεί η ROC ανάλυση.



Εικόνα 2. AUC

V. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

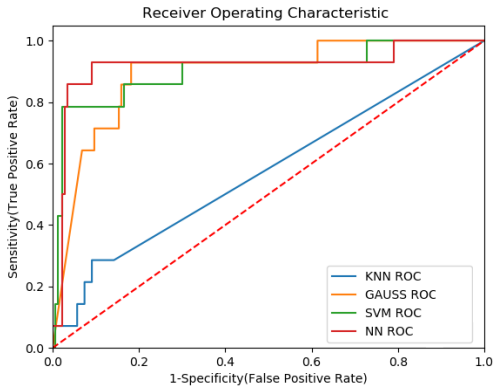
Ο παρακάτω πίνακας παρουσιάζει τα αποτελέσματα των κατηγοριοποιητών KNN (με αριθμό γειτόνων= 5), SVM (με kernel = 'linear'), GNB, NN's (MLPClassifier: 'lbfgs' με πλήθος κρυφών επιπέδων 30) στο σύνολο δεδομένων.

TABLE III. ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΕΣ

Classifiers	**KNN (k=5)**	**SVM**	**GAUSSIAN**	**NN's**
Accuracy	92.63	93.00	10.52	92.63
Recall / Sensitivity	06.66	0,851	92.30	45.00
Precision	100	0,639	06.62	75.00
F1 -score	12.50	0,727	12.37	56.20
ROC accuracy	60.91	89.43	87.22	82.58

Παρατηρούμε ότι τα αποτελέσματα των μοντέλων μας έχουν επηρεαστεί από την ανισορροπία του συνόλου δεδομένων μας (imbalanced data). Πιο συγκεκριμένα, ο αλγόριθμος KNN προβλέπει σχεδόν όλες τις τιμές ώς 0, ενώ ο Gaussian κατηγοριοποιητής προβλέπει σχεδόν όλα τα αποτελέσματα ως 1.

Παρακάτω ακολουθεί η καμπύλη ROC των αποτελεσμάτων για τους τέσσερις κατηγοριοποιητές.



Εικόνα 3. ROC Καμπύλη

ΑΝΤΙΜΕΤΩΠΙΣΗ IMBALANCED DATA

Για την αντιμετώπιση του προβλήματος εφαρμόσαμε over-sampling και under-sampling αφού πραγματοποιήσαμε κανονικοποίηση των δεδομένων.

A. Over-sampling

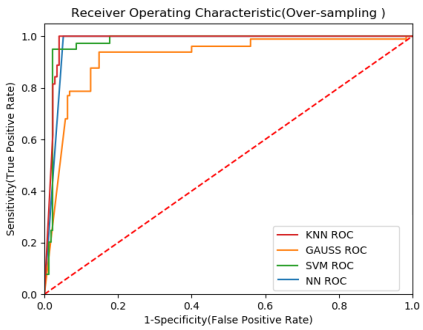
Με την χρήση του RandomOverSampler της κλάσης imblearn της βιβλιοθήκης sklearn το σύνολο δεδομένων μας αυξήθηκε σε 1411 εγγραφές.

TABLE IV. OVER SAMPLING

Classifiers	**KNN (k=5)**	**SVM**	**GAUSSIAN**	**NN's**
Accuracy	95.46	94.05	56.90	97.45
Recall / Sensitivity	100	94.68	98.94	100
Precision	91.75	94.17	55.00	95.28
F1 -score	95.69	94.42	70.69	97.58
ROC accuracy	99.14	95.47	94.01	97.49

Παρατηρούμε ότι τα νευρωνικά δίκτυα μας δίνουν τα καλύτερα αποτελέσματα. Πιο συγκεκριμένα, προβλέπουν όλες τις True τιμές (recall 100%).

Παρακάτω παρατηρούμε την καμπύλη ROC των κατηγοριοποιητών μετα από over-sampling.



Εικόνα 4. ROC Καμπύλη μετά από over-Sampling

B. *Under-Sampling*

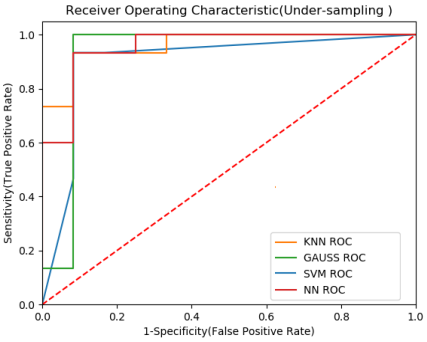
Με την χρήση του RandomUnderSampler της κλάσης imblearn της βιβλιοθήκης sklearn το σύνολο δεδομένων μας μειώθηκε σε μόλις 108 εγγραφές. Παρά την μείωση του όγκου των δεδομένων το νέο σύνολο δεδομένων περιλαμβάνει ίσο αριθμό true και false στην κλάση στόχο (πλήρως ισορροπημένη κλάση ως προς την συχνότητα εμφάνισης των τιμών).

TABLE V. UNDER SAMPLING

classifiers	**KNN (k=5)**	**SVM**	**GAUSSIAN**	**NN's**
Accuracy	62.96	92.29	66.67	88.88
Recall / Sensitivity	71.49	84.61	30.77	84.85
Precision	38.47	100	100	91.85
F1 -score	50.03	91.66	47.01	92.85
ROC accuracy	71.27	93.40	94.50	92.50

Παρατηρούμε ότι οι κατηγοριοποιητές SVM και Gauss όσες τιμές προβλέπουν ως 1 (True) είναι 1 (True) (precision 100%) .

Παρακάτω φαίνεται η καμπύλη ROC των κατηγοριοποιητών μετά από under-sampling.



Εικόνα 5. ROC Καμπύλη μετά από under -Sampling

Παρατηρώντας τα αποτελέσματα μπορούμε εύκολα να διαπιστώσουμε ότι ο SVM εκπαιδεύεται και δίνει καλύτερα αποτελέσματα σε σχέση με τους υπόλοιπους τρεις αλγορίθμους. Η τεχνική Under-sampling μας δίνει εμφανώς χειρότερα αποτελέσματα.

C. *Επιλογή χαρακτηριστικών με την μέθοδο LASSO*

Η επιλογή συγκεκριμένων χαρακτηριστικών είναι μια πολύ σημαντική διαδικασία στα προβλήματα μηχανικής μάθησης. Με την επιλογή συγκεκριμένων χαρακτηριστικών επιτυγχάνουμε την μείωση του overfitting, την αντοχή του μοντέλου μας στον θόρυβο και καλύτερα αποτελέσματα[9].

Η μέθοδος LASSO (Least Absolute Shrinkage and Selection Operator) είναι ένας τύπος γραμμικής παλινδρόμησης ο οποίος χρησιμοποιεί μία μέθοδο «μαζέματος» των τιμών του δείγματος «γύρω» από μία κεντρική τιμή(π.χ. μέσος όρος). Με την χρήση της μεθόδου Lasso από την βιβλιοθήκη sklearn.linear_model με παράμετρο alpha=0.015 επιλέχθηκαν 11 χαρακτηριστικά (δείκτες : 8, 13, 15, 18, 20, 25, 26, 27, 28, 29, 30). Η εκπαίδευση των αλγορίθμων με τα συγκεκριμένα χαρακτηριστικά παρουσιάζεται στον παρακάτω πίνακα.

TABLE VI. LASSO

classifiers	**KNN (k=5)**	**SVM**	**GAUSSIAN**	**NN's**
Accuracy	91.05	93.15	89.47	90.52
Recall/ Sensitivity	31.25	81.25	87.50	37.50
Precision	45.45	56.52	43.75	42.85
F1 -score	37.03	66.67	58.33	39.99
ROC accuracy	75.61	90.98	85.65	80.99

D. *Επιλογή χαρακτηριστικών με χρήση συσχέτισης Pearson*

Επιλέχθηκαν 11 χαρακτηριστικά με την μέθοδο corr() της κλάσης Pandas (δείκτες: 11, 12, 18, 20,24, 25, 26, 27, 28, 29, 30). Τα αποτελέσματα των αλγορίθμων που εκπαιδευτήκαν με τα συγκεκριμένα χαρακτηριστικά παρουσιάζονται στον παρακάτω πίνακα.

TABLE VII. PEARSON

classifiers	**KNN (k=5)**	**SVM**	**GAUSSIAN**	**NN's**
Accuracy	92.10	94.21	92.10	92.10
Recall/ Sensitivity	66.66	94.44	75.00	66.66
Precision	50.00	62.92	42.85	50.00
F1 -score	57.14	75.55	54.54	57.14

ROC accuracy	87.58	94.55	94.94	90.51

Σε παρόμοια εργασία μελετήθηκε η επιλογή υποσυνόλου χαρακτηριστικών με την μέθοδο SFS (Sequential Forward Selection). Σύγκριση με τα αποτελέσματα εργασίας με την εργασία [5] για τους αλγορίθμους SVM , KNN.

TABLE VIII. ΣΥΓΚΡΙΣΗ ΜΕ ΕΡΓΑΣΙΑ [5]

	Results [5]		Pearson Correlation		LASSO	
	KNN	SVM	KNN	**SVM**	KNN	SVM
Sensitivity	**100.0**	91.07	66.66	**94.21**	31.25	81.25
Precision	**81.06**	90.53	50.00	**94.44**	45.45	56.52
F1 -score	**89.54**	90.81	57.41	**75.55**	37.03	66.67

Παρατηρούμε ότι στο υποσύνολο των χαρακτηριστικών που επιλέχθηκε με την μέθοδο SFS ο KNN δίνει καλύτερα αποτελέσματα από τα αποτελέσματα που δίνει όταν εκπαιδεύεται στο σύνολο χαρακτηριστικών που επιλέχθηκε από την συσχέτιση Pearson ή από την μέθοδο LASSO. Τα αποτελέσματα αυτά θα μπορούσαμε να τα χαρακτηρίσουμε αναμενόμενα μιας και το Lasso είναι ένα γραμμικό μοντέλο ενώ ο KNN είναι ένας μη γραμμικός κατηγοριοποιητής.

VI. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εργασία αυτή παρουσιάζει και αναλύει την σύγκριση μεταξύ τεσσάρων διαφορετικών κατηγοριοποιητών μηχανικής μάθησης με σκοπό την ανάδειξή του καλύτερου μοντέλου πρόβλεψης για το σύνολο δεδομένων Cervical cancer (Risk Factors) Data Set.

Κατά την εκτέλεση της εργασίας παρατηρήθηκε το φαινόμενο της μεγάλης ανισορροπίας των δεδομένων (μόλις το 6,41% ήταν 1-True class). Ακολουθήσαμε προσεγγίσεις οι οποίες περιείχαν τεχνικές αύξησης και μείωσης των εγγραφών των δεδομένων (over/under sampling) καθώς και τεχνικές για την επιλογή υποσυνόλου των χαρακτηριστικών με τις μεθόδους LASSO και Pearson correlation. Η μέθοδος Over-sampling δίνει καλύτερα αποτελέσματα από την μέθοδο Under-sampling με τα νευρωνικά δίκτυα να μας δίνουν τα καλύτερα αποτελέσματα.

Η σύγκριση των κατηγοριοποιητών μετά την εφαρμογή μεθόδων μείωσης διαστάσεων έδειξε ότι η επιλογή χαρακτηριστικών με την μέθοδο LASSO μειώνει την αποτελεσματικότητα του αλγορίθμου KNN, ενώ η επιλογή χαρακτηριστικών με την μέθοδο Pearson Correlation μας έδωσε τα καλύτερα αποτελέσματα όσο αφορά σε Sensitivity (ευαισθησία) με τον αλγόριθμο SVM.

Η μέθοδος SFSS όπως παρουσιάστηκε στην εργασία [5] παραμένει η τεχνική που μεγάλο βάρος δόθηκε στην μετρική recall- Sensitivity μιας και τα δεδομένα μας είναι ιατρικής φύσης. Άλλες μετρικές οι οποίες χρησιμοποιήθηκαν αλλά δεν αναφέρονται στην εργασία είναι η Logarithmic loss.

Τέλος, η μελέτη του συνόλου λαμβάνοντας ως target class και τις τέσσερις κλάσεις Hinselmann, Schiller, Cytology, Biopsy θα μπορούσε να αποτελέσει ένα ξεχωριστό πρόβλημα multi-class κατηγοριοποίησης.

VII. ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Ιός των ανθρώπινων θηλωμάτων(HPV) , Πανεπιστήμιο Αθηνών Αρεταίειο Νοσοκομείο , http://www.aretaieio-obgyn.com/el/hpv.html?start=1*)*

[2] Κανελλόπουλος Δημήτριος , "Καρκίνος του τραχήλου της μήτρας. Νεότερα δεδομένα", ΤΟ ΒΗΜΑ ΤΟΥ ΑΣΚΛΗΠΙΟΥ, Τόμος 16, Τεύχος 4, Ειδικό άρθρο 2017

[3] Kauffman RP, Griffin SJ, Lund JD, Tullar PE. 2013. Current recommendations for cervical cancer screening: do they render the annual pelvic examination obsolete? Medical Principles and Practice 22(4):313–322 DOI 10.1159/000346137.

[4] Arbyn M, Castellsague X, de Sanjose S, Bruni L, Saraiya M, Bray F, Ferlay J: Worldside burden of cervical cancer in 2008.

[5] Classification of Cervical Cancer Dataset ,Abstract ID: 2423 ,Y. M. S. Al-Wesabi, Avishek Choudhury, Daehan Won ,Binghamton University, USA.

[6] Supervised deep learning embeddings for the prediction of cervical cancer diagnosis Kelwin Fernandes, Davide Chicco, Jaime S. Cardoso, and Jessica Fernandes Instituto de Engenharia de Sistemas.

[7] Computational Methods of Feature Selection, Huan Liu, Hiroshi Motoda, CRC Press, Boca Raton, FL (2007), 440 pp, ISBN 978-1-58488-878-9

[8] Enhanced Classification Model for Cervical Cancer Dataset based on Cost Sensitive Classifier Hayder K. Fatlawi Information Technology Research and Development Center , University of Kufa, Najaf, Iraq

[9] On the Consistency of Feature Selection With Lasso for Non-linear Targets, Yue Zhang,Soumya Ray, Weihong Guo.

[10] [10] Encyclopedia Machine, Claude Sammut, Geoffrey I. Webb, Springer, Boston, MA(2010)

[11] Learning from Imbalanced Data Sets, Fernández Hilario, A., García López, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F,(2018) ISBN 978-3-319-98074-4

[12] Class Imbalance Problem in Data Mining: Review, Mr.Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik, 2013 ISSN 2277-5420

[13] Balancing of Imbalanced Data Classification Using Enhanced Fuzzy and SMOTE Technique, Chancha, pertik Garg, 2019, ISSN(Online): 2320-9801

[14] Python Data Science Handbook: Essential Tools for Working with Data 1st Edition, Jake VanderPlas, ISBN-10: 9781491912058

[15] Khalid, S., Khalil, T., and Nasreen, S., 2014, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning," In Science and Information Conference (SAI), IEEE, Aug 27, 372-378

[16] Wu, W. and H. Zhou, 2017, "Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches", IEEE Access, 5:p. 25189-25195

[17] Feature Selection for Classification M. Dash , H. Liu , Department of Information System & Computer Science, National University of Singapore, Singapore 119260, 21 March 1997