ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ(ΠΜΣ)

Εργαστήριο Βιομαθηματικών

«Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική»

«Research Methodology in Biomedicine, Biostatistics and Clinical Bioinformatics»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΘΟΔΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΓΟΝΙΔΙΑΚΑ ΔΕΔΟΜΕΝΑ

MACHINE LEARNING APPROACHES IN GENOMIC DATA

Τάψας Απόστολος

Τριμελής εξεταστηκή επιτροπή:

Επιβλέπων Καθηγητής: ΔΡ. Ευάγγελος Ευαγγέλου, Επίκουρος Καθηγητής
Τμήμα Ιατρικής, Πανεπιστήμιο Ιωαννίνων.

ΔΡ. Στεφανίδης Ιωάννης, Καθηγητής Παθολογίας-Νεφρολογίας
Τμήμα Ιατρικής, Πανεπιστήμιο Θεσσαλίας.

Χρυσούλα Δοξάνη
Τμήμα Ιατρικής, Πανεπιστήμιο Θεσσαλίας.

Διπλωματική Εργασία υποβληθείσα στο Τμήμα Ιατρικής του Πανεπιστημίου Θεσσαλίας ως μέρος των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική.

Λάρισα 2018

Machine Learning Approaches in Genomic Data

Tapsas Apostolos
Tripartite committee
Supervisor: Dr. Evaggelos Evangeloy
Dr. Ioannis Stefanidis
PhD candidate Xrusoula Doxani

Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική

Research Methodology in Biomedicine, Biostatistics and Clinical Bioinformatics

Department of Biomathematics
School of Medicine
University of Thessaly

A thesis submitted for the degree of

Master of Science

Larissa 2018

Machine Learning Approaches in Genomic Data

Tapsas Apostolos [1*], Evangelos Evangelou[2*]

Abstract

Στόχοι: Στόχος είναι η εφαρμογή μεθόδων μηχανικής μάθησης σε μια πραγματική μελέτη περίπτωσης με σκοπό την εύρεση του καλύτερου ομαδοποιητή.

Δεδομένα και Μέθοδοι: Τα δεδομένα της UK Biobank περιέχουν γονότυπους για 488.377 συμμετέχοντες. Από τους συμμετέχοντες αυτούς, 438,427 έχουν γονοτυπηθεί σε 825.427 πολυμορφισμούς από το Affymetrix UK Biobank Axiom Array chip ενώ οι υπόλοιποι 49.950 γονοτυπήθηκαν σε 807.411 πολυμορφισμούς από το Affymetrix Affymetrix UK BiLEVE Axiom Array chip της μελέτης UK BiLEVE. Η καταγραφή των μη τυποποιημένων πολυμορφισμών (imputation) πραγματοποιήθηκε κεντρικά από την UK Biobank χρησιμοποιώντας μια πλατφόρμα αναφοράς που συγχώνευσε τις πλατφόρμες UK10K and 1000 Genomes Phase 3 καθώς επίσης και την Consortium Reference Consortium (HRC) πλατφόρμα. Από τους 488.377 συμμετέχοντες της UK Biobank με διαθέσιμα γενετικά δεδομένα αποκλείστηκαν τα άτομα μη-ευρωπαϊκής προέλευσης και με τη χρήση των κεντρικά παρεχόμενων δεδομένων συγγενείας της UK Biobank, αποκλείστηκαν από την ανάλυση οποιαδήποτε ζεύγη συγγενών 1ου και 2ου βαθμού. Εντοπίστηκαν τα περιστατικά μελανώματος, μέσω της σύνδεσης με τα κεντρικά μητρώα του Εθνικού Κέντρου Υγείας του Ηνωμένου Βασιλείου (NHS). Τα κεντρικά μητρώα του NHS παρέχουν πληροφορίες σχετικά με τις καταχωρήσεις καρκίνου και τους θανάτους που κωδικοποιούνται σύμφωνα με τη 10η αναθεώρηση της Διεθνούς Ταξινόμησης των Νοσημάτων (κωδικοί ICD-9 και ICD-10 (C43) (WHO, http://www.who.int/classifications/icd/en/). Καταλήξαμε σε 2,871 περιπτώσεις καρκίνου από άτομα ευρωπαϊκής προέλευσης, των οποίων η πρώτη διάγνωση καρκίνου ήταν μελάνωμα. Από τους υπόλοιπους συμμετέχοντες της UK Biobank επιλέχθηκαν 378,624 υγιείς μάρτυρες οι οποίοι δεν είχαν ποτέ διαγνωσθεί με καρκίνο ή είχαν αναφέρει την ύπαρξη καρκίνου μέσω αυτοαναφοράς και δεν είχαν καταχωρηθεί ποτέ στο εθνικό μητρώο καρκίνου. Στην συνέχεια επιλέχθηκαν 564 ασθενείς οι οποίοι είχαν διαγνωστεί με μελάνωμα και 564 υγιείς ασθενείς, με σκοπό την εκπαίδευση ενός τεχνητού νευρωνικό δικτιού το οποίο θα προβλέπει αν ένας νέος ασθενείς επρόκειτο να αναπτύξει μελάνωμα.

Αποτελέσματα: Τα ποσοστά επιτυχίας πρόβλεψης των δύο δικτύων που εκπαιδευτήκαν Multilayer Perceptron -MLP και Support Vector Machine -SVM ήταν 68.99% και 80.71% αντίστοιχα όταν για την εκπαίδευση τους έγινε η χρήση μόνο των φαινοτυπικών δεδομένων. Με την εισαγωγή των γενετικών δεδομένων στην εκπαίδευση των μοντέλων, το ποσοστό επιτυχίας πρόβλεψης αυξήθηκε στα 72.1% για το μοντέλο MLP και 81.67% για το μοντέλο SVM.

Συμπεράσματα: Το σημαντικότερο αποτέλεσμα της παρούσας εργασίας είναι πως με την προσθήκη της γενετικής πληροφορίας στα μοντέλα επιτευχθήκαν μεγαλύτερα ποσοστά ακρίβειας.

Το μοντέλο που προτείνετε ως καλύτερη επιλογή είναι η μηχανή υποστήριξης διανυσμάτων(Support Vector Machine-SVM) χωρίς αυτό να συνεπάγετε με την απόρριψη ενός μοντέλου τεχνητού νευρωνικού δικτύου(Multilayer Perceptron), καθώς η απόδοση της ακρίβειας του εξαρτάτε από την αρχιτεκτονική του.

Keywords

Μεγάλα Δεδομένα – Ανάλυσης Δεδομένων – Εξόρυξη Δεδομένων – Μηχανική Μάθηση – Μάθηση σε Βάθος – Τεχνητά Νευρωνικά Δίχτυα – Python

[1] *Department of Bio-mathematics, University of Thessaly, Larissa, Greece*

[2] *Department of Health, University of Ioannina, Ioannina, Greece*

***Corresponding authors**: atapsas@uth.gr, vangelis@cc.uoi.gr

Machine Learning Approaches in Genomic Data

Tapsas Apostolos [1]*, Evangelos Evangelou[2]*

Abstract

Purpose: The aim is to apply mechanical learning methods to a real case study in order to find the best classifier.

Data and Methods: UK biobank data contain 488.377 genotypes. From 438.427 has been genotyped to 804.427 polymorphisms from AffymetrixUK Biobank Axiom Array chip and the other 49.950 has been genotyped to 807.411 polymorphism from Affymetrix Affymetrix UK BiLEVE Axiom Array chip for the UK BiLEVE study. The imputation register was held from UK Biobank with UK10K 1000 Genomes phase 3 platform and Consortium Reference Consortium platform. Of the 488,377 British Biobank participants with available genetic data, people of non-European origin were excluded and using centrally-provided data from the English family of Biobank, exclusions from the analysis were any pairs of 1st and 2nd tier grades. Instances of melanoma have been identified by linking to the central registries of the National Health Center of the United Kingdom (NHS). NHS core registers provide information on cancer registries and deaths coded in accordance with the 10th revision of the International Classification of Diseases (ICD-9 (172) and ICD-10 (C43) (WHO, http: // www .who.int / classifications / icd / el /). We ended up with 2,871 cases of cancer from people of European origin whose first diagnosis of cancer was melanoma. Of the remaining UK Biobank participants, 378,624 healthy controls were selected who had never been diagnosed with cancer or had reported self-reported cancer and had never been registered with the national cancer registry. Then, 564 patients who were diagnosed with melanoma and 564 healthy patients were selected to study an artificial neural network that would predict whether a new patient was going to develop melanoma.

Results:The accuracy score for multilayer perceptron-MLP model and support vector machine-SVM model is 68.99 and 80.71 without the genomic data. With the genomic data input the accuracy score is 72.1 for multilayer perceptron-MLP and 81.67 for support vector machine SVM model.

Conclusion: The most important result of this study is that with the addition of genetic information to the models, greater precision has been made. The model recommend as the best choice is the Support Vector Machine (SVM) without this leading to the rejection of an Artificial Neural Network model (Multilayer Perceptron), as accuracy is dependent on its architecture.

Keywords

Big Data – Genomic Data – Data Analysis – Data Mining – Machine Learning – Deep Learning – Neurall Networks – Python

[1] *Department of Bio-mathematics, University of Thessaly, Larissa, Greece*

[2] *Department of Health, University of Ioannina, Ioannina, Greece*

***Corresponding authors**: atapsas@uth.gr, vangelis@cc.uoi.gr

Εισαγωγή

Είναι γενικώς αποδεκτό πώς ζούμε, εξελισσόμαστε και αναπτυσσόμαστε στο σύγχρονο περιβάλλον της κοινωνίας της πληροφορίας. Στην εποχή αυτή έχουμε την δυνατότητα πρόσβασης σε γνώσεις στις οποίες θα ήταν δύσκολο ή αδύνατο να έχουμε πρόσβαση στο παρελθόν. Η δυνατότητα αυτή προκύπτει από την ανταλλαγή και αποθήκευση δεδομένων σε μεγάλες βάσεις δεδομένων με την μορφή ακατέργαστης πληροφορίας, σαν επακόλουθο της ραγδαίας εξέλιξης της τεχνολογίας κατά τον τελευταίο αιώνα.

Η ανταλλαγή και αποθήκευση, αυτής της τόσο μεγάλης συγκομιδής δεδομένων (Big Data) παρατηρείτε όχι μόνο στο κοινωνικό τομέα (Social Media)[11], αλλά και στον επιστημονικό.
Επιστημονικά κέντρα όπως π.χ η NASA[9] και το CERN[10] αποθηκεύουν μετρήσεις με την μορφή πληροφορίας.

Ακόμη, μια συγκομιδή δεδομένων του βελησνεκούς αυτού συναντάται και στο πεδίο της Ιατρική.
Οι κλασικές μέθοδοι της στατιστικής αδυνατούν να επεξεργαστούν αυτό τον τεράστιο όγκο δεδομένων[12], με απώτερο αποτέλεσμα να προκύπτει άμεσα η ανάγκη ανάπτυξης νέων μεθόδων επεξεργασίας των παραπάνω δεδομένων.

Η ανάγκη αυτή δημιούργησε ένα νέο τομέα στην επιστήμη της Πληροφορικής, την Εξόρυξη Δεδομένων. Η Εξόρυξη Δεδομένων αποτελεί την συνεργασία διάφορων επιστημονικών κλάδων όπως αυτοί της Ιατρικής, των Μαθηματικών και Πληροφορικής. Φαίνεται να είναι πολλά υποσχόμενη καθώς εφαρμόζει μια σειρά νέων τεχνικών όπως η Μηχανική Μάθηση[13].

Η ανάγκη για εφαρμογή της εξόρυξης δεδομένων σε διάφορους κλάδους της Υγείας αυξάνεται με τον ίδιο ρυθμό αύξησης των δεδομένων που αποθηκεύονται καθημερινά στις βιοτράπεζες.

Παραδείγματος χάριν στις βιοτράπεζες συναντούμε δεδομένα Γενετικής πληροφορίας(Genomic Data)[14] τα οποία αποτελούν ένα μεγάλο όγκο ακατέργαστης πληροφορίας.

Η Συνεργασία της επιστήμης της Πληροφορικής με την Μοριακή Βιολογια για την επεξεργασία αυτών των δεδομένων δημιούργησε την Βιο-πληροφορική.

Εφαρμογές Μηχανικής Μάθησης στην Ιατρική

Η μηχανική μάθηση βρίσκει ευρεία εφαρμογή στον επιστημονικό κλάδο την Ιατρική, μερικές από αυτές τις εφαρμογές είναι:

- Ιατρική διάγνωση [23].
- Πρόγνωση και πρόβλεψη του καρκίνου[21][22].
- Ομαδοποίηση του DNA[19][20].

.

Υπόβαθρο

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση αποτελεί μια κλασική στατιστική μέθοδο ομαδοποίησης ή πρόβλεψης τιμών[16]. Η λογιστική παλινδρόμηση εφαρμόζεται όταν δεν ικανοποιούνται οι υποθέσεις της γραμμικής παλινδρόμησης οι οποίες είναι οι εξής:

- Ομοσκεδαστικότητα[16].
- Κανονικότητα[16].
- Γραμμικότητα[16].

Συνεπώς το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο το οποίο χρησιμοποιείται όταν η εξαρτημένη μεταβλητή (Υ) είναι δίτιμη[16], γνωστή και ως μεταβλητή Bernoulli, τα σφάλματα του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης Υ είναι διακριτή ή δίτιμη[16] και ορίζετε ως εξής:

$$Y_i = E(Y_i) + e_i[16] \tag{1}$$

$$E(Y_i) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)}+1} = \frac{1}{e^{-(\beta_0+\beta_1 x)}+1}[16] \tag{2}$$

Επιπροσθέτως, η λογιστική παλινδρόμηση δίνει την δυνατότητα πρόβλεψης πιθανοφάνειας και αυτό συμβαίνει διότι η εξαρτημένη μεταβλητή (Υ) είναι δίτιμη, συνεπώς η συνάρτηση πυκνότητας πιθανότητας θα είναι:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}[16] \tag{3}$$

Η από κοινού συνάρτηση πιθανότητας είναι:

$$g(Y_1 \ldots Yn) = \prod_i^n \pi_{i=1}^{Y_i} f_i(Y_i) = \prod_i^n \pi_{i=1}^{Y_i} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}[16] \tag{4}$$

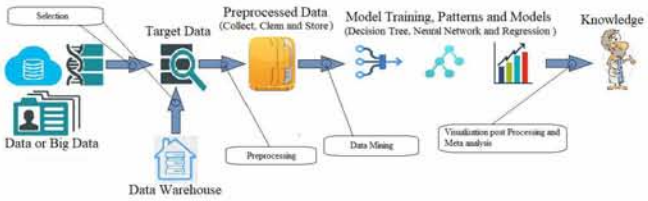
Λογαριθμίζοντας την παραπάνω συνάρτησης προκύπτει η παρακάτω σχέση:

$$lnL = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 x_i) - \sum_{i=1}^{n} ln(1+e^{b_0 b_1 x_i}) \tag{5}$$

Η σχέση(5) ονομάζεται λογαριθμική συνάρτηση πιθανοφάνειας ή ιδιότητα υπολογισμού της πιθανότητα εμφάνισης ενός συμβάντος που εμφανίζει η παραπάνω συνάρτηση[16], εφαρμόζεται σε μια πληθώρα ιατρικών προβλημάτων[15].

Διαδικασία Εξόρυξης Δεδομένων

Η Εξόρυξη Δεδομένων αποτελείται από μία σειρά ενεργειών οι οποίες ακολουθούνται για τη δημιουργία γνώσης.
Η διαδικασία αυτή αποτελείται από τέσσερα στάδια τα οποία παρουσιάζονται στο Σχήμα1.[1].



Σχήμα **1.** Διαδικασία Εξόρυξης Δεδομένων

Περιγραφή Μοντέλων

Η ιδέα των τεχνητών νευρωνικών δικτύων αναπτύχθηκε από τους McCulloch και Pits[1] το 1940 οι οποίοι περιέγραψαν ένα μοντέλο το οποίο αποτελεί το πιο απλό νευρωνικό δίκτυο, γνωστό και ως Perceptron. Οι McCulloch και Pitts περιέγραψαν αυτό το δίκτυο σαν μία διαδικασία πολλαπλασιασμού των εισόδων με ένα συντελεστή βάρους και στην συνέχεια η εισαγωγή τους σε μία συνάρτηση ενεργοποίησης η οποία εξάγει την τελική κατάσταση του νευρώνα[1].

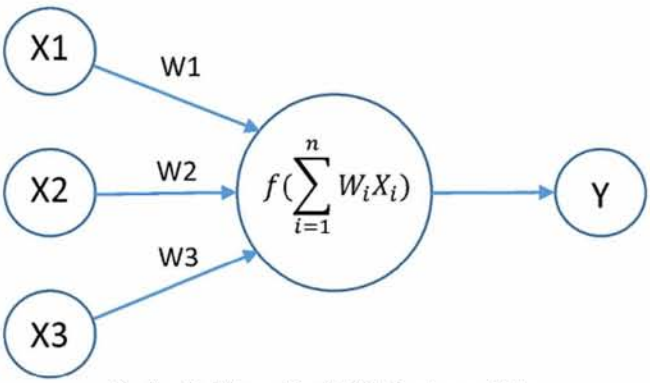
Αναλυτικότερα αν x1, x2,.. , xn είναι οι είσοδοι του νευρώνα τότε το άθροισμα που δέχεται ο νευρώνας δίνετε από την σχέση:

$$u = \sum w_i * x_i - \theta [1] \tag{6}$$

Όπου wi τα συναπτικά βάρη του νευρώνα, ϑ το κατώφλι ενεργοποίησης και u η διέγερση του νευρώνα. Εν συνεχεία το παραπάνω αποτέλεσμα εισάγεται στην συνάρτηση ενεργοποίησης, και εξάγεται η τελική τιμή(κατάσταση του νευρώνα)y όπως περιγράφεται στην παρακάτω σχέση:

$$y = f(u)[1] \tag{7}$$

Οι συναρτήσεις ενεργοποίησης που αναφέρθηκαν ποικίλουν. Μερικές από αυτές είναι: η βηματική συνάρτηση(step function), σιγμοειδής συνάρτηση (sigmoid), υπερβολική εφαπτομένη συνάρτηση(hyper-bolic tangent) και τέλος η συνάρτηση κατωφλίου (threshold function)[1].



Σχήμα **2.** Μοντέλο McCulloch και Pitts

Το μοντέλο McColloch και Pitts είναι αρκετά κοντά σε ένα μοντέλο απλής γραμμικής παλινδρόμησης1.
Έστω ένας νευρώνας μίας εισόδου, στον οποίο εισέρχονται οι τιμές της μεταβλητής x πολλαπλασιαζόμενες με το αντίστοιχο βάρος τους w1 με συνάρτηση ενεργοποίησης την f(x)=x. Η παραπάνω περιγραφή εκφράζεται μαθηματικά με βάση την σχέση(1) ως:

$$y = x * w1 - \theta \tag{8}$$

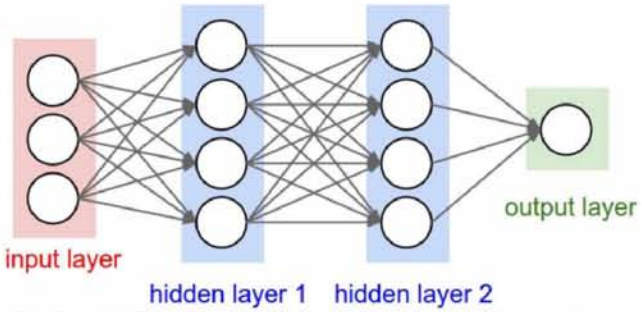
Στην παραπάνω σχέση οι τιμές των w1 και ϑ είναι σταθερές οπότε μπορεί να ορισθούν και ως έκφραση μιας ευθείας με τις εξής ιδιότητες:

- Μπορεί να προβλεφθεί η τιμή της εξαρτημένης μεταβλητής y.
- Η γενική μορφή της σχέσης(3) μπορεί να ομαδοποιηθεί.

Τέλος, αν στην παραπάνω σχέση εισαχθούν περισσότερες ανεξάρτητες μεταβλητές τότε η σχέση αυτή θεωρείται ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

Οι δυνατότες του μοντέλου ενός νευρώνα περιορίζονται σε προβλήματα του δυσδιάστατου χώρου ή σε προβλήματα των οποίων οι πειραματικές μονάδες περιγράφονται από δύο χαρακτηριστικά[1].

Για την άρση λοιπόν αυτού του περιορισμού γίνεται η χρήση περισσότερων νευρώνων. Η χρήση κρυφών νευρώνων για παράδειγμα δημιουργεί μια αρχιτεκτονική πολλών στρωμάτων και νευρώνων γνωστή ως δίκτυο perceptron πολλών στρωμάτων(Multi-Layer Perceptron-MLP)[1].



Σχήμα **3.** Μοντέλο Multi-Layer Perceptron 2 στωμάτων

Το σφάλμα στον κόμβο εξόδου j του n-οστού δεδομένου εκφράζεται απο την σχέση:

$$e_i(n) = d_j(n) - y_j(n)[1] \tag{9}$$

Όπου d είναι η μεταβλητή στόχος και y η έξοδος του νευρώνα.

Σε ένα δίκτυο perceptron πολλών στρωμάτων τα βάρη των κόμβων ρυθμίζονται βάσει διορθώσεων που ελαχιστοποιούν το σφάλμα σε ολόκληρη την έξοδο απο τήν σχέση:

$$\varepsilon(n) = 1/2 \sum_j e_j^2(n)[1] \tag{10}$$

Ένα χαρακτηριστικό των MLP είναι ότι οι νευρώνες οποιουδήποτε

στρώματος l τροφοδοτούν αποκλειστικά το-υς νευρώνες του επόμενου στρώματος l+1[1]. Η δυνατότητ-α αναπαράστασης διαχωριστικών επιφανειών που παρέχουν τα MLP μπορεί να δώσει λύσεις σε μη γραμμικά διαχωρίσιμα προβλήματα[1].

Εκπαίδευση του μοντέλου

Με τον όρο εκπαίδευση ενός δικτύου πολλών στρωμάτων εννοείται η διαδικασία ρύθμισης των βαρών του έτσι ώστε να ικανοποιείται κάποιο κριτήριο καταλληλόλητας[1]. Αυτό που κάνει ενδιαφέρουσα την εκπαίδευση ενός MLP είναι πως το κατάλληλο σε μέγεθος δίκτυο μπορεί να εκπαιδευθεί και να μάθει οποιαδήποτε συνάρτηση με οποιαδήποτε επιθυμητή ποιότητα προσέγγισης [1].

Ο παραπάνω ισχυρισμός είναι γνωστός και ως ιδιότητα του καθολικού προσεγγιστή[1].

Τέλος, τα βασικά προβλήματα που παρουσιάζει το παραπάνω μοντέλο είναι :

- Πρόβλημα αρχιτεκτονικής, δηλαδή πόσους νευρώνες και πόσα κρυφά επίπεδα θα πρέπει να διαθέτει το μοντέλο για τη βέλτιστη δυνατή λύση του εκάστοτε προβλήματος[1].
- Ποια συνάρτηση ενεργοποίησης είναι η καταλληλότερη[1].
- Βραδεία εκπαίδευση[1].

Μηχανές Διανυσμάτων Υποστήριξης

Αν θεωρηθεί ένα γραμμικά διαχωρίσιμο πρόβλημα δύο ομάδων αυτό σημαίνει πως υπάρχουν άπειρά ζεύγη w,w0 που μπορούν να διαχωρίσουν τις δύο κλάσεις. Για την επιλογή της καλύτερης δυνατής λύσης χρησιμοποιείται το κριτήριο του περιθωρίου ταξινόμησης (margin) γ το οποίο ορίζεται ως το άθροισμα των γ0 και γ1 των κλάσεων όπου:

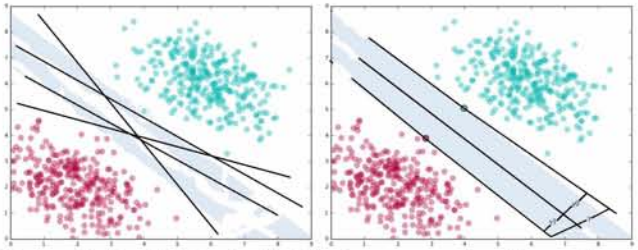
$$\gamma_0 = min_{x \in C0} \frac{|W^T x + w_0|}{||W||} = min_{x \in C0} \frac{-(W^T x + w_0)}{||W||} [1] \quad (11)$$

$$\gamma_1 = min_{x \in C1} \frac{|W^T x + w_0|}{||W||} = min_{x \in C1} \frac{(W^T x + w_0)}{||W||} [1] \quad (12)$$

$$\gamma = \gamma_0 + \gamma 1 [1] \quad (13)$$

Τα πρότυπα x της κλάσης C0 και x' της κλάσης C1 με τα οποία επιτυγχάνεται η ελάχιστη απόσταση καλούνται διανύσματα υποστήριξης(support vectors) όπως φαίνεται στο παρακάτω σχήμα.

Η επέκταση της παραπάνω περιγραφής του μοντέλου σε μη γραμμικά διαχωρίσιμες κλάσεις μπορεί να γίνει χρησιμοποιώντας κάποιο κατάλληλο μη γραμμικό μετασχηματισμό Φ(.)[1]



Σχήμα **4.** Small-Margin Vs Maximum-margin. Hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

Στην περίπτωση αυτή η βέλτιστη διαχωριστική επιφάνεια περιγράφεται απο την παρακάτω σχέση:

$$g^*(x) = w^T \Phi(x) + w_0 = \sum_{i=1}^{P} \lambda_i di \Phi(x_i)^T \Phi(x) + w_0 [1][1] \quad (14)$$

Εάν στην παραπάνω σχέση τα εσωτερικά γινόμενα αντικατασταθούν απο κ(x,y)ως:

$$k(x,y) = \Phi(x)^T \Phi(x) [1] \quad (15)$$

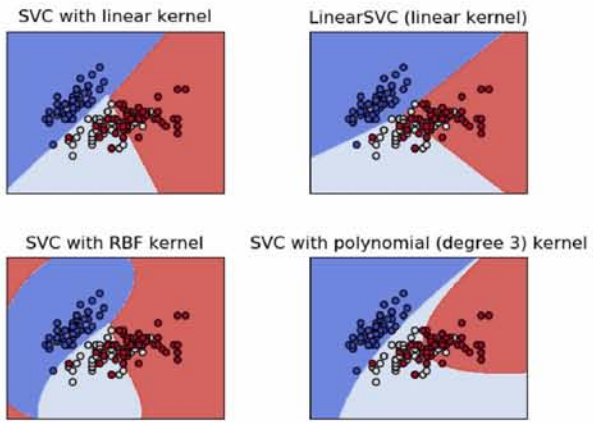
Τότε η σχέση (9) ορίζεται ως:

$$g^*(x) = w^T \Phi(x) + w_0 = \sum_{i=1}^{P} \lambda_i di k(x,y) + w_0 [1] \quad (16)$$

Η Συνάρτηση k(...) ονομάζεται συνάρτηση πυρήνα (kernel). Οι συνηθέστερες Συναρτήσεις πυρήνα είναι οι εξής:

- Γκαουσιανή RBF
- Πολυωνυμική
- Σιγμοειδής

Όπως φαίνεται στο παρακάτω σχήμα:



Σχήμα **5.** Support Vector Machine Kernels

Μέθοδοι

Εργαλεία

Το εργαλείο που χρησιμοποιήθηκε για τη συγγραφή των προγραμμάτων προεπεξεργασίας και εκπαίδευσης των μοντέλων είναι το περιβάλλον **Anaconda**, το οποίο περιέχει το επιστημονικό εργαλείο spyder 3.2.8 της γλώσσας προγραμματισμού Python 3.6.

Μέγεθος και Χαρακτηριστικά του δείγματος

Τα δεδομένα τα οποία συλλέχθηκαν από τη βιοτράπεζα του Ηνωμένου Βασιλείου(http://www.ukbiobank.ac.uk) αγγίζουν τα 322.579 περιστατικά τα οποία βρισκόντουσαν σε δύο αρχεία. Το αρχείο με όνομα "phenotypic dataset" το οποίο περιείχε τα φαινοτυπικά χαρακτηριστικά των 322.579 περιπτώσεων, και το αρχείο με όνομα 20snps file 2 το οποίο περιείχε τους πολυμορφισμούς των υπαίτιων για το μελάνωμα[8] όπως φαίνεται στον παρακάτω πίνακα.

SNP	numcode	Region	Gene	EffectAllele	EAF	Beta	P-value
rs1241086	150856153	1q21.3	ARNT/SETDB1	G	0,63	0,13	5.2 x 10-13
rs1858550	226608104	1q42.12	PARP1	C	0,66	0,143	1.7 x 10-13
rs6750047	38276549	2p22.2	RMDN2 (CYP1B1)	A	0,43	0,088	2.9 x 10-7
rs7582362	202176294	2q33-q34	CASP8	A	0,27	0,113	8.9 x 10-9
rs380286	1320247	5p15.33	TERT/CLPTM1L	A	0,44	0,152	1.6 x 10-17
rs250417	33952378	5p13.2	SLC45A2	C	0,97	0,891	2.3 x 10-12
rs6914598	21163919	6p22.3	CDKAL1	C	0,32	0,108	3.5 x 10-8
rs1636744	16984280	7p21.1	AGR3	T	0,4	0,105	7.1 x 10-9
rs7852450	21825075	9p21	CDKN2A/MTAP	T	0,52	0,212	4.6 x 10-32
rs1073922	109060830	9q31.2	TMEM38B (RAD23B, TAL2)	T	0,24	0,12	7.1 x 10-11
rs2995264	105668843	10q24.33	OBFC1	G	0,09	0,144	8.5 x 10-7
rs498136	69367118	11q13.3	CCND1	A	0,32	0,116	1.5 x 10-12
rs1393350	89011046	11q14-q21	TYR	A	0,27	0,198	2.5 x 10-25
rs7300822	108187689	11q22-q23	ATM	G	0,86	0,188	1.4 x 10-12
rs4778138	28335820	15q13.1	OCA2	A	0,84	0,18	3.1 x 10-9
rs1259663	54115829	16q12.2	FTO	A	0,16	0,143	1.8 x 10-9
rs7557060	89846677	16q24.3	MC1R	C	0,08	0,6	6.2 x 10-92
rs6088372	32586748	20q11.2-q12	ASIP	T	0,14	0,267	3.4 x 10-25
rs408825	42743496	21q22.3	MX2	T	0,6	0,141	3.2 x 10-15
rs2092180	38571563	22q13.1	PLA2G6	A	0,53	0,116	2.1 x 10-11

Σχήμα **6.** Λίστα πολυμορφισμών (SNP) που σχετίζονται στατιστικά σημαντικά με τον κίνδυνο ανάπτυξης μελανώματος.

Τα χαρακτηριστικά των φαινοτυπικών δεδομένων είναι τα εξής:

- Φύλο (Sex)
- Ηλικία (Age)
- Ώρες έκθεσης στον ήλιο το καλοκαίρι (Time summer)
- Ώρες έκθεσης στον ήλιο το χειμώνα (Time winter)
- Ηλιακά εγκαύματα (Sunburns)
- Χρήση αντηλιακού (Sun Protection)
- Αν έχει γίνει solarium (Solarium)
- Χρώμα δέρματος (Skin colour)
- Ευκολία μαυρίσματος (Ease tanning)
- Χρώμα μαλλιών (Hair col)
- Γήρανση προσώπου (Facial ageing)
- Ανάπτυξη μελανώματος (Class)

Προεπεξεργασία

Για να εισαχθούν τα δεδομένα στο τελικό μοντέλο θα πρέπει να υποστούν κάποια προ επεξεργασία με σκοπό την εξισορρόπηση των δεδομένων[3].
Η προεπεξεργασία αποτελείται από εξής τρία βασικά βήματα:

- Την εξάλειψή ελλειπών τιμών (missing values), διότι προκύπτουν προβλήματα υπολογισμού[2].
- Τον έλεγχο της ισορροπίας των δεδομένων, δηλαδή αν οι κλάσεις των δεδομένων είναι ίσες ως προς τον αριθμό τους και εάν όχι τότε θα πρέπει να εξισορροπηθούν, διότι προκύπτουν προβλήματα ειδίκευσης του μοντέλου σε μία κλάση (το μοντέλο προβλέπει καλά μόνο την μία κλάση) [3]
- Τη μείωση, αν είναι δυνατό, των διαστάσεων των δεδομένων ή αλλιώς την εξάλειψή περιττών χαρακτηριστικών του δείγματος (dataset)[3].

Χρησιμοποιήσαμε ένα προσθετικό (additive) μοντελο κληρονομικότητας για την αναλύση των δεδομένών. Τέλος, θα πρέπει να ακολουθήσει ο έλεγχος εξάλειψης χαρακτηριστικών ή αλλιώς η μείωση διαστάσεων των δεδομέ-νων.

Ανάλυσης Κυρίων Συνιστωσών

Μία τέτοια μέθοδος είναι η Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis-PCA) η οποία προϋποθέτει τα παρακάτω βήματα[2][4]:

- Τον υπολογισμό του πίνακα συνδιακυμάνσεων S των χαρακτηριστικών D[2][4].

$$S = cov(D) \tag{17}$$

Όπου

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}[(x_i\bar{x}) * (y_i\bar{y})] \tag{18}$$

- Τον υπολογισμό των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα συνδιακυμάνσεων[2][4].

$$S * eigenvectors = eigenvalues * eigenvectors \tag{19}$$

Ο πίνακας των ιδιοτιμών eigenvalues δείχνει την «συμμετοχή» του κάθε χαρακτηριστικού σε όλα τα δεδομένα. Εάν η τιμή αυτή είναι πάνω από 1 τότε η μέθοδος προτείνει πως το χαρακτηριστικό αυτό θα πρέπει να ληφθεί υπόψιν (Kaiser, 1960).

Αν η τιμή του χαρακτηριστικού είναι κάτω από 1 τότε το χαρακτηριστικό αυτό δεν πρέπει να ληφθεί υπόψιν (Kaiser, 1960).

Αποτελέσματα

Τα δεδομένα αποθηκευτήκαν στο αρχείο με όνομα alleles και συγχωνευτήκαν με τα δεδομένα του αρχείου phenotypic dataset.

Τα δεδομένα τα οποία περιείχαν ελλειπούσες τιμές αφαιρέθηκαν με αποτέλεσμα το νέο μέγεθος του δείγματος να είναι 180.536 περιπτώσεις εκ των οποίων οι 564 αποτελούν περιπτώσεις μελανώματος.

Στην συνέχεια προστέθηκαν άλλες 564 υγιείς περιπτώσεις στο τελικό αρχείο με όνομα dataset με τελικό μέγεθος δείγματος 1016 περιπτώσεις 499 ευρωπαίων γυναικών και 517 ευρωπαίων ανδρών με μέσο όρο ηλικίας 57 έτη.

Η επιλογή αυτή έγινε με σκοπό την αποφυγή της εξειδίκευσης του μοντέλου καθώς η μεγάλη ανισορροπία των περιπτώσεων(μελανώματος και μη μελανώματος) θα οδηγούσε σε απόρριψη της μικρότερης πληθυσμιακής ομάδας από το μοντέλο.[5]

Στον παρακάτω πίνακα φαίνονται οι ιδιοτιμές των δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων.

Πίνακας ιδιοτιμών	
Χαρακτηριστικό	Ιδιοτιμή
sex	62.65
age	16.73
time summer	13.95
time winter	7.23
sunburns	0.2
sun protection	0.2
solarium	1.4
skin col	1.21
ease tanning	0.67
hair col	0.80
facial ageing	0.94

Πίνακας 1. Πίνακας ιδιοτιμών του κάθε χαρακτηριστικού.

Αν και το κριτήριο Kaiser προβλέπει πως τα χαρακτηριστικά, ηλιακά εγκαύματα, η χρήση του αντηλιακού, η ευκολία μαυρίσματος και το χρώμα τον μαλλιων θα πρεπει να παραληφθουν, παρόλα αυτά διατηρήθηκαν στο μοντέλο λόγω της γνωστής προγνωστικής τους αξίας στην ανάπτυξη του μελανώματος.

Η πρώτη προσέγγιση του προβλήματος έγινε με την κλασική μέθοδο της λογιστικής παλινδρόμησης.

Η υψηλότερη ακρίβεια πρόβλεψης μελανώματος χρησιμοποιώντας μόνο τα φαινοτυπικά χαρακτηριστικά(αρχείο με όνομα phenotypic dataset) των περιπτώσεων είναι 58.66% (σχήμα 7).

Με την πρόσθεση των γενετικών δεδομένων (αρχείο phenotypic dataset και alleles) στο μοντέλο, η υψηλότερη ακρίβεια πρόβλεψης ανέρχεται στο 60.31%(σχήμα 7).

Προσεγγίζοντας το πρόβλημα με ένα multilayer perceptron-MLP για την ομαδοποίηση των δεδομένων.

Η διαδικασία της εκπαίδευσης ενός MLP αποτελείται από δοκιμές οι οποίες γίνονται με σκοπό την επίτευξη της καλύ-τερης δυνατής ακρίβειας της πρόβλεψης. Οι δοκιμές πραγματοποιήθηκαν σε ενα MLP ενός και δύο κρυφών στρωμάτων με μέγιστο αριθμό νευρώνων τους χίλιους.

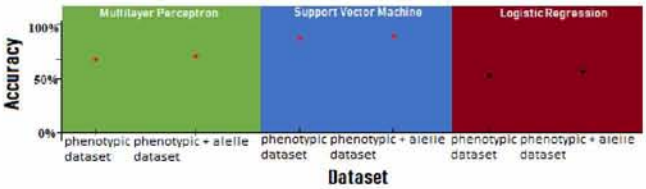
Η υψηλότερη ακρίβεια πρόβλεψης μελανώματος χρησιμοποιώντας μόνο τα φαινοτυπικά χαρακτηριστικά(αρχείο με όνομα phenotypic dataset) των περιπτώσεων είναι 69% (σχήμα 7), με ένα κρυφό επίπεδο 498 νευρώνων.

Με την πρόσθεση των γενετικών δεδομένων (αρχείο phenotypic dataset και alleles) στο μοντέλο, η υψηλότερη ακρίβεια πρόβλεψης ανέρχεται στο 72.1%(σχήμα 7) ομοίως, με ένα κρυφό επίπεδο 498 νευρώνων.

Η τρίτη προσέγγιση του προβλήματος είναι η εκπαίδευση ενός SVM.

Η υψηλότερη ακρίβεια πρόβλεψης μελανώματος χρησιμοποιώντας μόνο τα φαινοτυπικά χαρακτηριστικά(αρχείο με όνομα phenotypic dataset) των περιπτώσεων ανέρχεται στο ποσοστό του 80.71%
(σχήμα 7).

Με την πρόσθεση των γενετικών δεδομένων (αρχείο phenotypic dataset και alleles) στο μοντέλο, η υψηλότερη ακρίβεια πρόβλεψης ανέρχεται στο 81.67%(σχήμα 7).



Σχήμα **7.** Ποσοστά Ακρίβειας Μοντέλων

Συμπεράσματα

Το σημαντικότερο αποτέλεσμα είναι η αύξηση του ποσοστού πρόβλεψης των μοντέλων με την εισαγωγή της γενετικής πληροφορίας.

Το ποσοστό ακρίβειας πρόβλεψης του μοντέλου γραμμικής παλινδρόμησης ήταν το μικρότερο απο τα μοντέλα και οι τιμές του ήταν κοντά στα προβλεπόμενα αποτελέσματα σύμφωνα και με τη βιβλιογραφία[15][17][18].

Τα αποτελέσματα δείχνουν ότι το SVM αποτελεί την αποδοτικότερη λύση συγκριτικά με τις άλλες δυο προσεγγίσεις όπως τεκμηριώνεται και στην επιστημονική βιβλιογραφία [6][7][17][18].

Η σωστή επιλογή της συνάρτησης πύρηνα(kernel) είναι εξαιρετικά σημαντική για την επίτευξη ορθών αποτελεσμάτων[7].

Στην παρούσα έρευνα τα δεδομένα κατά κύριο λόγο αποτελούνται από διακριτές τιμές και ο χώρος παραμέτρων ήταν ευκολά προσβάσιμος. Παρόλα αυτά όμως η «εικόνα» των δεδομένων παραμένει άγνωστη με αποτέλεσμα να υποθέτουμε πως οι κλάσεις των δεδομένων δεν είναι γραμμικά διαχωρίσιμες και κατ' επέκταση να χρησιμοποιηθεί η ανάλογη συνάρτηση πυρήνα (kernel).

Οι περιορισμοί της έρευνας είναι: το μέγεθος του δείγματος το οποίο προέκυψε από την εξάλειψη των ελλείπων τιμών.

Η υπολογιστική δυνατότητα που δόθηκε για τις δοκιμές των αρχιτεκτονικών του MLP, καθώς η υπολογιστική δύναμη επιτρέπει ένα περιορισμένο χώρο δοκιμών με αποτέλεσμα η αρχιτεκτονική του MLP που χρησιμοποιήθηκε ίσως να μην είναι η αποδοτικότερη δυνατή, καθώς αν υπήρχε η υπολογιστική δυνατότητα για την επίτευξη περισσότερων δοκιμών σε μεγαλύτερη πληθώρα αρχιτεκτονικών να είχε προκύψει μια καλύτερη αρχιτεκτονική η οποία συνεπάγεται αύξηση της ακρίβειας του μοντέλου.

Ευχαριστίες

Η παρούσα έρευνα δεν θα ήταν δυνατόν να εκπονηθεί χωρίς την βοήθεια του επιβλέποντα Κ. Ευάγγελο Ευαγγέλου επίκουρου καθηγητή του τμήματος Ιατρικής, του Πανεπιστημίου Ιωαννίνων, ο οποίος με τις γνώσεις του και την διάθεση του για βοήθεια ήταν δίπλα μου σε οτιδήποτε χρειάστηκα σαν οδηγός σε αυτή την προσπάθεια.

Θερμές ευχαριστίες θα πρέπει να δοθούν και στον υποψήφιο διδάκτορα Κ.Γιώργο Ντρίτσο για την εύρεση και αποστολή των δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων.

Βιβλιογραφία

[1] Κωσταντίνος Διαμαντάρας. Τεχνητά Νευρωνικά Δίκτυα. Κλειδάριθμος,2007. Αθήνα.

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Εισαγωγή στην Εξόρυξη Δεδομένων. ΤΖΙΟΛΑ,2010. Θεσσαλονίκη.

[3] Mr.Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik, 2013, Class Imbalance Problem in Data Mining: Review, International Journal of Computer Science and Network (IJCSN), Volume 2, Issue 1, pp 4-5.

[4] Krystyna Kuźniar, Maciej Zajac, 2015, Some methods of pre-processing input data for neural networks, Institute of Fundamental Technological Research, Polish Academy of Sciences, Computer Assisted Methods in Engineering and Science, 22: pp 141–151

[5] Mike Wasikowski, Member and Xue-wen Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.

[6] E.A.Zanaty, Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification, Egyptian Informatics Journa, 13,177,183,2012.

[7] N.Barabino1, M.Pallavicini, A.Petrolini, M.Pontil, A.Verri, Support Vector Machines vs Multi-LayerPerceptron in Particle Identi cation,ESANN'1999 proceedings - European Symposium on Artificial Neural Networks Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 257-262.

[8] Law MH, Bishop DT, Lee JE, Brossard M, Martin NG, Moses EK, Song F8, Barrett JH, Kumar R, Easton DF, Pharoah PDP, Swerdlow AJ, Kypreou KP, Taylor JC, Harland M, Randerson-Moor J, Akslen LA, Andresen PA, Avril MF, Azizi E, Scarrà GB, Brown KM, Debniak T, Duffy DL, Elder DE, Fang S, Friedman E, Galan P, Ghiorzo P, Gillanders EM, Goldstein AM, Gruis NA, Hansson J, Helsing P, Hočevar M, Höiom V, Ingvar C, Kanetsky PA, Chen WV; GenoMEL Consortium; Essen-Heidelberg Investigators; SDH Study Group; Q-MEGA and QTWIN Investigators; AMFS Investigators; ATHENS Melanoma Study Group, Landi MT, Lang J, Lathrop GM, Lubiński J, Mackie RM, Mann GJ, Molven A, Montgomery GW, Novaković S, Olsson H, Puig S, Puig-Butille JA, Qureshi AA, Radford-Smith GL, van der Stoep N, van Doorn R, Whiteman DC, Craig JE, Schadendorf D, Simms LA, Burdon KP, Nyholt DR, Pooley KA, Orr N, Stratigos AJ, Cust AE, Ward SV, Hayward NK, Han J, Schulze HJ, Dunning AM, Bishop JAN, Demenais F, Amos CI, MacGregor S, Iles MM, Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma,Epub,987-995. doi: 10.1038/ng.3373 Sep.2015

[9] Piyush Mehrotra, L. Harper Pryor, F. Ron Bailey, Marc Cotnoir, Supporting "Big Data" Analysis and Analytics at the NASA Advanced Supercomputing (NAS) Facility,NASA Advanced Supercomputing (NAS) Division, NASA Ames Research Center, Moffett Field, CA 94035, Version 1, January 29, 2014.

[10]Tapio Niemi, Jukka K. Nurminen, Jukka, K. NurminenJuha-Matti LiukkonenAri-Pekka Hameri, Ari-Pekka Hameri,Towards Green Big Data at CERN,November 2017

[11] Savita Kumari,impact of big data and social media on society,March 2016

[12] Chun Wang, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan, Statistical methods and computing for big data,Stat Interface. 2016; 9(4): 399–414., Sep 29

[13] Yanqing Zhang, Jagath C. Rajapakse, Machine Learning in Bioinformatics, A JOHN WILEY SONS, INC PUBLICATIONS, 2009, New Jersey.

[14] C. Lathe III (OpenHelix), Jennifer M. Williams (OpenHelix), Mary E. Mangan (OpenHelix) Donna Karolchik (University of California, Santa Cruz Genome Bioinformatics Group,Genomic Data Resources: Challenges and Promises, Challenges and Promises. Nature Education 1(3):2,2008.

[15] Kypreou KP, Stefanaki I, Antonopoulou K, Karagianni F, Ntritsos G, Zaras A, Nikolaou V, Kalfa I, Chasapi V, Polydorou D, Gogas H, Spyrou GM, Bertram L, Lill CM, Ioannidis JP, Antoniou C, Evangelou E, Stratigos AI.J Invest Dermatol., Prediction of Melanoma Risk in a Southern European Population Based on a Weighted Genetic Risk Score, 2016 Mar;136(3):690-5. doi: 10.1016/j.jid.2015.12.007. Epub 2015 Dec 14.

[16] Joseph M. Hilbe, Practical Guide to Logistic Regression, Chapman and Hall/CRC, 2015.

[17] Diego Alejandro Salazar, Jorge Iván Vélez, Juan Carlos Salazar , Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?, Revista Colombiana de Estadística Número especial en Bioestadística, volumen 35, no. 2, pp. 223 a 237.Junio 2012.

[18] Abdulhamit Subasi Ergun Erçelebi, Classification of EEG signals using neural network and logistic regression, Computer Methods and Programs in Biomedicine Volume 78, Issue 2, May 2005, Pages 87-99.

[19] Ngoc Giang Nguyen, Vu Anh Tran, Duc Luu Ngo, Dau Phan, Favorisen Rosyking Lumbanraja, Mohammad Reza Faisal, Bahriddin Abapihi, Mamoru Kubo, Kenji Satou, DNA Sequence Classification by Convolutional Neural Network,27 April 2016.

[20] Wang JT, Rozen S, Shapiro BA, Shasha D, Wang Z, Yin M, New techniques for DNA sequence classification., 1999 Summer;6(2):209-18.

[21] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noeld, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,Procedia Computer Science Volume 83, 2016, Pages 1064-1069.

[22] Konstantin aKourou, Themis P.Exarchos, Konstantinos P.Exarchos, Michalis V.Karamouzis, Dimitrios I.Fotiadis, Machine learning applications in cancer prognosis and prediction,Computational and Structural Biotechnology Journal Volume 13, 2015, Pages 8-17.

[23] Igor Kononenk, Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in Medicine Volume 23, Issue 1, August 2001, Pages 89-109.