Πανεπιστήμιο Δυτικής Μακεδονίας
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

Ανάλυση Τεχνικών Προγνωστικής Μοντελοποίησης στη Διαχείριση Χρόνιων Καρδιακών Παθήσεων

Φοιτητής: Αχιλλέας Μουκούλης
ΑΕΜ: 527
Επιβλέπων: Καθηγητής Παντελής Αγγελίδης, Καθηγητής Π.Δ.Μ

Περιεχόμενα

Η παρουσίαση ακολουθεί την παρακάτω δομή:

- Προγνωστική Ανάλυση
- Βασικές Έννοιες
- Εργαλεία που χρησιμοποιήθηκαν
- Τα Δεδομένα μας
- Πειραματική Διαδικασία
- Συμπεράσματα
- Μελλοντική Επέκταση

Την παρουσίαση μαζί με το κείμενο και τους κώδικες μπορείτε να τα βρείτε στο: https://github.com/AxilleasMoukoulis

Προγνωστική Ανάλυση

Τι είναι;

Αποτελεί επέκταση της **Εξόρυξης Δεδομένων**

Εφαρμόζει γνωστές τεχνικές ανάλυσης δεδομένων ώστε να αποτελέσει εργαλείο **“πρόβλεψης του μέλλοντος”**

Τι κάνει;

Παρέχει τη δυνατότητα **πρόληψης** και **αντιμετώπισης** προσδοκώμενων αποτελεσμάτων

Παρέχει **σχέσεις** μεταξύ διαφόρων **παραγόντων** ώστε να αντιμετωπιστεί το ρίσκο μίας επιλογής με βάση ένα συγκεκριμένο πλήθος παραγόντων που την **επηρεάζουν**.

Τα μοτίβα που ανακαλύπτονται μέσα από παρελθοντικά δεδομένα και δεδομένα συναλλαγών, μπορούν να εφαρμοστούν ώστε να προσδιοριστεί το ρίσκο μίας μελλοντικής απόφασης ή οι ευκαιρίες που μπορούν να προκύψουν μέσα από αυτή

Διαδικασία Προγνωστικής Ανάλυσης

1) Προσδιορισμός Προβλήματος

Προσδιορισμός του αποτελέσματος, του χρόνου διεκπεραίωσης, των δεδομένων

2) Συλλογή Δεδομένων

Προετοιμασία λήψης των δεδομένων από καθορισμένες πηγές

3) Ανάλυση Δεδομένων

Εξαγωγή χρήσιμων πληροφοριών και συμπερασμάτων

4) Στατιστική Ανάλυση

Χρήση στατιστικών μοντέλων για την επικύρωση των υποθέσεων

5) Μοντελοποίηση Δεδομένων

Επιλογή του καταλληλότερου προγνωστικού μοντέλου για τα δεδομένα.

6) Εφαρμογή Παρακολούθηση

Real time πρόγνωση και παρακολούθηση των αποτελεσμάτων

Πλεονεκτήματα στην Υγεία

Προγνωστικά μοντέλα εφαρμόζονται για κάθε νέο ασθενή ώστε ο γιατρός να έχει άμεση εκτίμηση της κατάστασης

- Μεγαλύτερη ακρίβεια στις διαγνώσεις
- Βελτίωση της προληπτικής Ιατρικής και της δημόσιας Υγείας
- Εξατομικευμένη Ιατρική
- Φαρμακοβιομηχανία
- Πλεονεκτήματα για τους ασθενείς

Συμπέρασμα

Στις αναπτυγμένες χώρες η προγνωστική ανάλυσης αποτελεί το επόμενο βήμα της Ιατρικής

- Καλύτερη πληροφόρηση των ασθενών
- Ο ρόλος των ιατρών θα αλλάξει
- Νοσοκομεία – Ασφαλιστικές

Οι αλλαγές είναι καθ' οδών και θα φέρουν αποδοτικότερη εξάσκηση της Ιατρικής και μείωση των ασθενειών.

Βασικές Έννοιες

Δεδομένα

Ως δεδομένα ορίζουμε ένα σύνολο από τιμές ποιοτικών και ποσοτικών μεταβλητών που είναι σε θέση να επεξεργαστεί ένα υπολογιστικό σύστημα

Εξόρυξη Δεδομένων

Η διαδικασία της ανάλυσης δεδομένων από διαφορετικές οπτικές γωνίες η οποία συνοψίζεται σε χρήσιμη πληροφορία

Μεγάλα Δεδομένα

Σύνολα δεδομένων τα οποία είναι τόσο μεγάλα σε μέγεθος ή πολυπλοκότητα ώστε οι γνωστές τεχνικές ανάλυσης και τα λογισμικά, δεν είναι σε θέση να τα διαχειριστούν

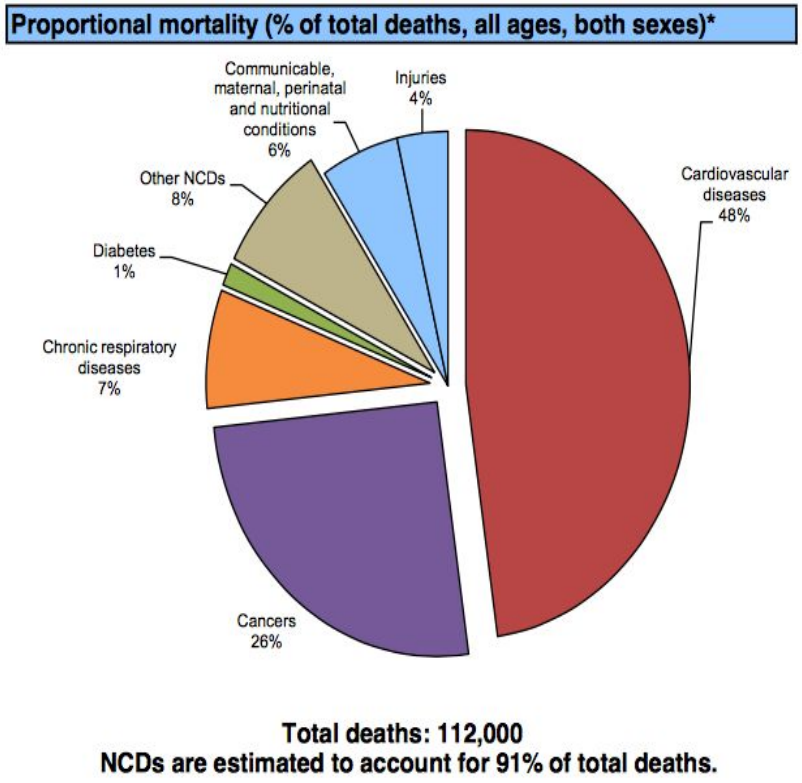
Μεγάλα Δεδομένα με λίγα λόγια

Χαρακτηρίζονται από 5 βασικά χαρακτηριστικά (5 V's)

- **Volume**
- **Variety**
- **Velocity**
- **Variability**
- **Veracity**

Εργαλεία που χρησιμοποιήθηκαν

- R / R Studio
- Sublime Text Editor
- Google Docs



Καρδιακές Παθήσεις

Στη χώρα μας για το 2014

- Το 48% των θανάτων προέρχεται από κάποια καρδιακή πάθηση
- Συνολικά 53.760 θάνατοι
- Πηγή: http://www.who.int/nmh/countries/grc_en.pdf

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non -anginal pain 4 = asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic result: 0 = normal 1 = having ST-T abnormality 2 = showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 to 3
Thal	Discrete	3 = normal 6 = fixed defect 7 = reversible defect
Diagnosis (num)	Discrete	Diagnosis classes: 0 = healthy 1 = patient who is subject to possible heart disease

Τα δεδομένα μας

- Cleveland Clinic Foundation Heart Rate Disease Dataset
- 303 καταγεγραμμένα περιστατικά
- 14 στήλες με χαρακτηριστικά
- Στόχος, η πρόβλεψη ύπαρξης καρδιακής πάθησης ή όχι
- Πηγή: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Πειραματική Διαδικασία

Μελετήθηκαν οι παρακάτω τεχνικές για τα δεδομένα:

- **Association Rules**
- **K Nearest Neighbors Classifier**
- **Naive Bayes Classifier**
- **Artificial Neural Networks**
- **Decision Trees (CART - C4.5 - Random Forest)**
- **General Boosted Regression**

Κανόνες Συσχέτισης

Ανιχνεύει μοτίβα που επαναλαμβάνονται μέσα στα δεδομένα και εξάγει κανόνες που επαληθεύουν τα δεδομένα με κάποιο ποσοστό εμπιστοσύνης

	lhs		rhs	support	confidence	lift
[1]	{cp=asymptomatic,slope=flat,thal=reversable defect}	=>	{num=Disease}	0.1551155	0.9591837	2.090882
[2]	{cp=asymptomatic,exang=Yes,thal=reversable defect}	=>	{num=Disease}	0.1650165	0.9433962	2.056468
[3]	{cp=asymptomatic,fbs=< 120 mg/dl,thal=reversable defect}	=>	{num=Disease}	0.1947195	0.9076923	1.978639
[4]	{cp=asymptomatic,exang=Yes,slope=flat}	=>	{num=Disease}	0.1617162	0.9074074	1.978018
[5]	{sex=male,cp=asymptomatic,exang=Yes}	=>	{num=Disease}	0.1848185	0.9032258	1.968902
[6]	{sex=male,cp=asymptomatic,thal=reversable defect}	=>	{num=Disease}	0.1980198	0.8955224	1.952110
[7]	{sex=male,cp=asymptomatic,fbs=< 120 mg/dl,thal=reversable defect}	=>	{num=Disease}	0.1650165	0.8928571	1.946300
[8]	{sex=male,cp=asymptomatic,fbs=< 120 mg/dl,exang=Yes}	=>	{num=Disease}	0.1584158	0.8888889	1.937650
[9]	{sex=male,exang=Yes,thal=reversable defect}	=>	{num=Disease}	0.1518152	0.8846154	1.928334
[10]	{sex=male,cp=asymptomatic,slope=flat}	=>	{num=Disease}	0.1683168	0.8793103	1.916770
[11]	{sex=male,cp=asymptomatic,restecg=probable or definite left ventricular hypertrophy}	=>	{num=Disease}	0.1650165	0.8620690	1.879186
[12]	{fbs=< 120 mg/dl,slope=flat,thal=reversable defect}	=>	{num=Disease}	0.1749175	0.8548387	1.863425
[13]	{fbs=< 120 mg/dl,exang=Yes,slope=flat}	=>	{num=Disease}	0.1551155	0.8545455	1.862786
[14]	{cp=asymptomatic,fbs=< 120 mg/dl,exang=Yes}	=>	{num=Disease}	0.1914191	0.8529412	1.859289
[15]	{sex=male,slope=flat,thal=reversable defect}	=>	{num=Disease}	0.1650165	0.8474576	1.847336

Συμπεράσματα

- Η τιμή του lift δεν είναι ικανοποιητική
- Η τιμή του support θα μπορούσε να ήταν μεγαλύτερη
- Απαιτείται μεγαλύτερο dataset
- Οι σημαντικότεροι κανόνες τείνουν προς τη σωστή κατεύθυνση

Κ Πλησιέστεροι Γείτονες

- Κάθε δείγμα αναπαρίσταται ως σημείο δεδομένων D διαστάσεων
- D το πλήθος των χαρακτηριστικών
- Υπολογίζεται η απόσταση σε σχέση με τα σημεία εκπαίδευσης
- Καθορίζεται η ετικέτα της κατηγορίας του δείγματος ελέγχου
- Επαναλαμβάνεται για όλα τα σημεία ελέγχου

Αποτελέσματα

- Συνολικά: 297 εγγραφές συνολικά
- Training set: 250 εγγραφές - 84% του συνόλου
- Testing set: 47 εγγραφές - 16% του συνόλου
- Με τον παραπάνω διαχωρισμό επιτεύχθηκε η καλύτερη απόδοση
- Ποσοστό επιτυχίας: 79%

Αποτελέσματα

	Σύνολο	Κατηγοριοποίηση	Ποσοστό
0	**22**	**21**	**95%**
1	**10**	**6**	**60%**
2	**9**	**7**	**77%**
3	**4**	**2**	**50%**
4	**2**	**1**	**50%**
Αποτέλεσμα	**47**	**37**	**79%**

Συμπεράσματα

- Το μεγαλύτερο ποσοστό επιτυχίας για την πολυπληθέστερη κατηγορία
- Καλύτερα αποτελέσματα με μεγαλύτερο dataset
- Περισσότερες εγγραφές για τις κατηγορίες 3 και 4
- Η πολυωνυμική κατηγοριοποίηση δεν είναι εύκολη διαδικασία

Naive Bayes Classifier

- Υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα, δεδομένης μίας ετικέτας Y
- $P(X \mid Y = y) = \prod P(X_i \mid Y = y)$
- Αρκεί να εκτιμηθεί η υπό συνθήκη πιθανότητα για κάθε X_i δοθέντος του Y
- Για να κατηγοριοποιήσει μία εγγραφή ελέγχου, υπολογίζει την εκ των υστέρων πιθανότητα για κάθε Y
- $P(Y \mid X) = P(Y) \prod P(X_i \mid Y) / P(X)$

Αποτελέσματα

- Συνολικά: 297 εγγραφές συνολικά
- Διωνυμική Κατηγοριοποίηση
- 0: Δεν υπάρχει καρδιακή πάθηση - 1: Υπάρχει καρδιακή πάθηση
- Training set: 286 εγγραφές - 90% του συνόλου
- Testing set: 29 εγγραφές - 10% του συνόλου
- Με τον παραπάνω διαχωρισμό επιτεύχθηκε η καλύτερη απόδοση
- Ποσοστό επιτυχίας: ~86%

Αποτελέσματα

Training set	Test set	Ποσοστό
70%	30%	73%
80%	20%	69%
90%	10%	86%

Συμπεράσματα

- Κατάλληλη τεχνική για σύνθετα σύνολα δεδομένων
- Καλύτερα αποτελέσματα με μεγαλύτερο dataset
- Τα δεδομένα να ακολουθούν κανονική κατανομή

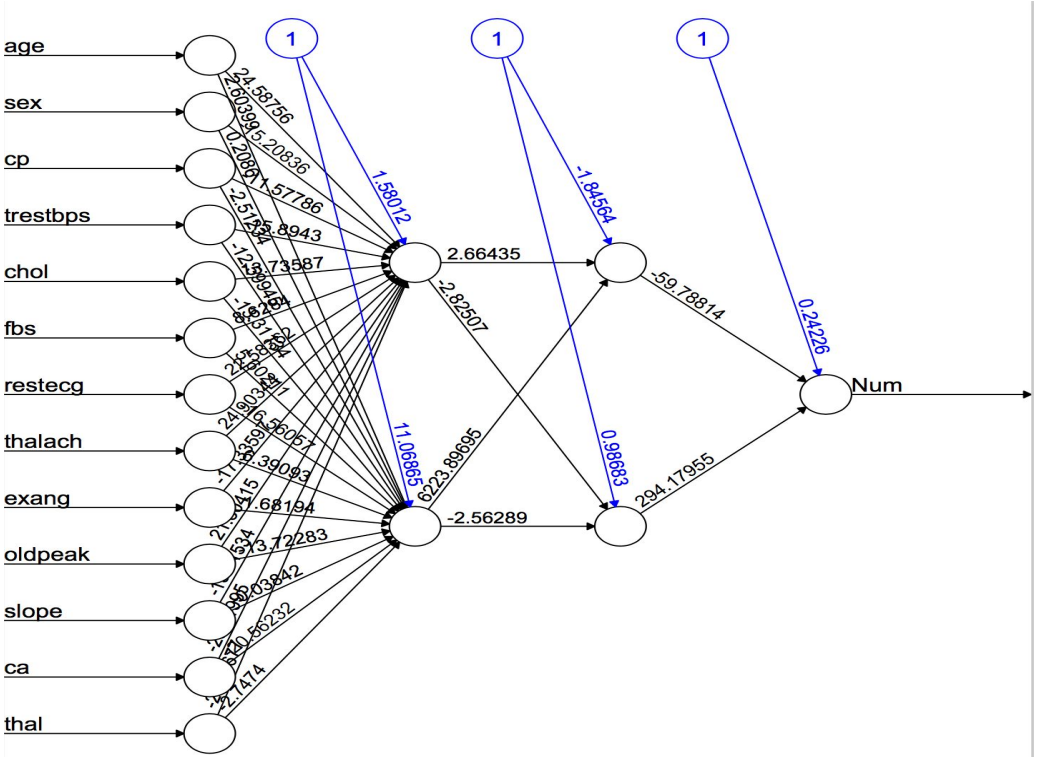
Τεχνητό Νευρωνικό Δίκτυο

- **Hidden Layers - Hidden Nodes**
- Οι κόμβοι του ενός επιπέδου συνδέονται με μόνο με τους κόμβους του επόμενου
- Συναρτήσεις ενεργοποίησης
- Ικανό να μοντελοποιήσει πολύπλοκες σχέσεις μεταξύ μεταβλητών εισόδου - εξόδου

Αποτελέσματα

- Συνολικά: 297 εγγραφές συνολικά
- Διωνυμική Κατηγοριοποίηση
- **Training set:** 80% - **Testing set:** 20% του συνόλου
- Με τον παραπάνω διαχωρισμό επιτεύχθηκε η καλύτερη απόδοση
- **Scaling** των δεδομένων
- Ποσοστό επιτυχίας: ~83%

Αποτελέσματα



Συμπεράσματα

- Ένα νευρωνικό δίκτυο έχει το προτέρημα να βελτιώνεται
- Περισσότερα δεδομένα = Μεγαλύτερο ποσοστό επιτυχίας
- Αν κάθε κατηγορία είχε αρκετά δεδομένα, τότε θα είχε επιτευχθεί και η πολυωνυμική κατηγοριοποίηση

Δένδρα Απόφασης

- Το δένδρο απόφασης μεγαλώνει αναδρομικά
- Αν όλες η καταγραφές ανήκουν στην ίδια κατηγορία, τότε ο κόμβος είναι ένα φύλλο
- Αντίθετα, επιλέγεται μία συνθήκη ελέγχου
- Οι καταγραφές χωρίζονται σε μικρότερα υποσύνολα, κόμβοι παιδιά
- Αναδρομική εκτέλεση σε κάθε παιδί

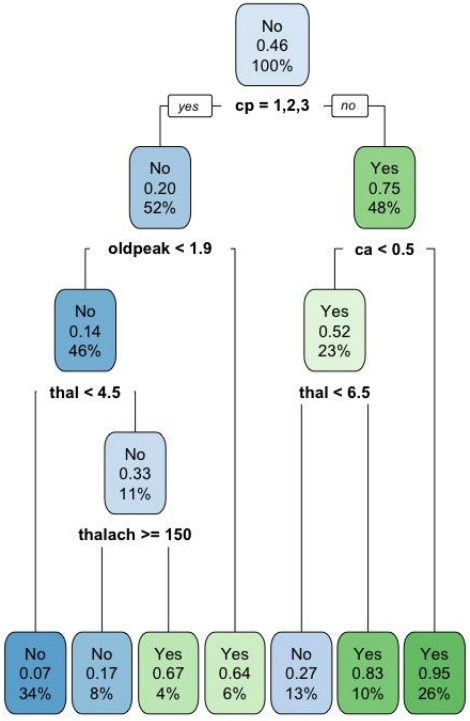
Δένδρα Απόφασης

- Τρεις τεχνικές - **CART, C4.5,** και **Random Forest**
- Διωνυμική Κατηγοριοποίηση
- **Training set:** 80% - **Testing set:** 20% του συνόλου
- Αφαίρεση ορισμένων χαρακτηριστικών

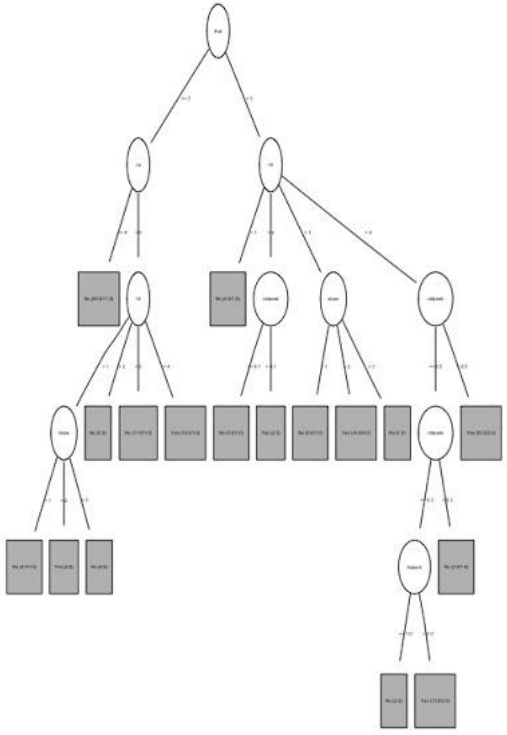
Αποτελέσματα

Αλγόριθμος	Απόδοση	Πακέτο
CART	**81,35%**	**rpart**
C4.5	**79,66%**	**RWeka**
Random Forest	**79,66%**	**randomForest**

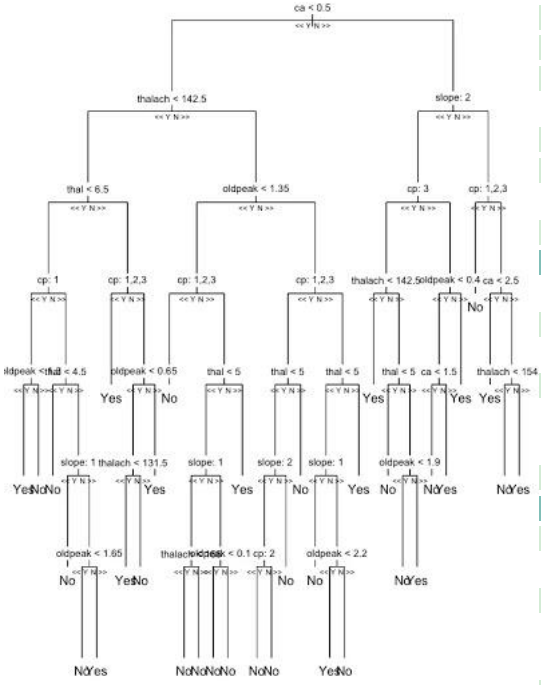
CART


No
0.46
100%
yes
cp = 1,2,3
no
No
0.20
52%
Yes
0.75
48%
oldpeak < 1.9
ca < 0.5
No
0.14
46%
Yes
0.52
23%
thal < 4.5
thal < 6.5
No
0.33
11%
thalach >= 150
No
0.07
34%
No
0.17
8%
Yes
0.67
4%
Yes
0.64
6%
No
0.27
13%
Yes
0.83
10%
Yes
0.95
26%

C4.5



RF


ca < 0.5
thalach < 142.5
slope: 2
thal < 6.5
oldpeak < 1.35
cp: 3
cp: 1,2,3
cp: 1
cp: 1,2,3
cp: 1,2,3
cp: 1,2,3
thalach < 142.5
oldpeak < 0.4
ca < 2.5
No
oldpeak < 4.5
oldpeak < 0.65
thal < 5
thal < 5
thal < 5
thal < 5
ca < 1.5
thalach < 154
Yes
No
Yes
Yes
Yes
slope: 1
thalach < 131.5
slope: 1
slope: 2
slope: 1
oldpeak < 1.9
oldpeak < 1.65
oldpeak < 0.1
cp: 2
oldpeak < 2.2

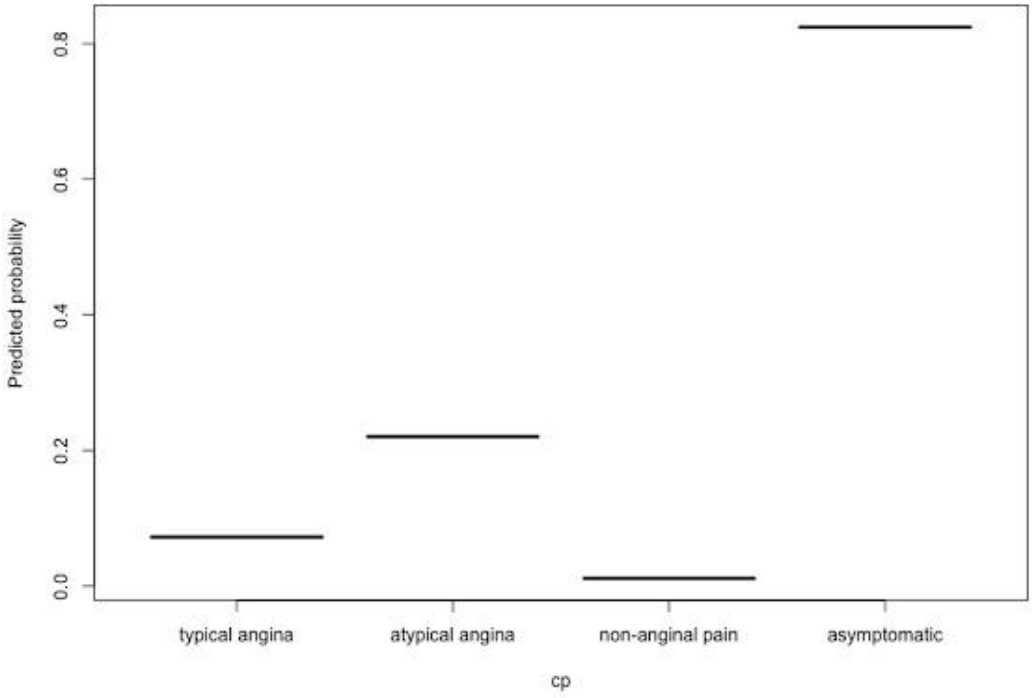
Συμπεράσματα

- Σχεδόν ίδιο ποσοστό επιτυχημένης κατηγοριοποίησης
- Δεν ήταν αντιπροσωπευτικά τα δεδομένα εκπαίδευσης
- Αριθμητικά και κατηγορικά χαρακτηριστικά
- Μη αποδεκτές τεχνικές για πραγματική εφαρμογή με αυτά τα δεδομένα

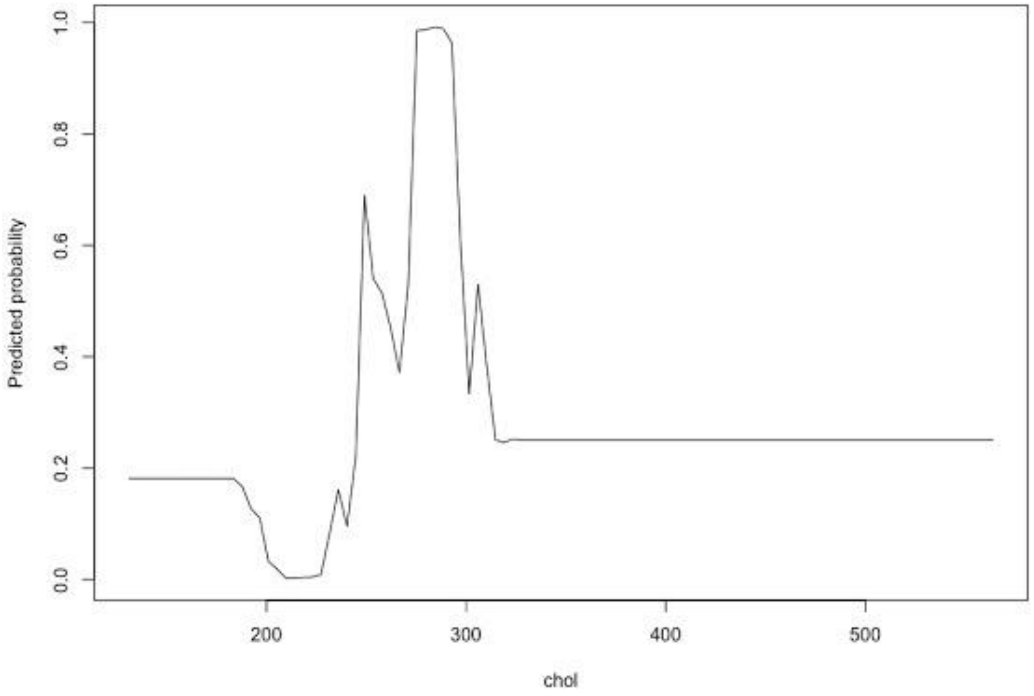
General Boosted Regression

- Τεχνική δένδρου απόφασης
- Σύνολο από μεθόδους που εξάγουν το αποτέλεσμα από ένα πλήθος μοντέλων
- Εφαρμόζεται το ένα μοντέλο μετά το άλλο ώστε να ελαχιστοποιηθούν τα επιμέρους σφάλματα
- Θυμίζει τον ανθρώπινο τρόπο εκμάθησης

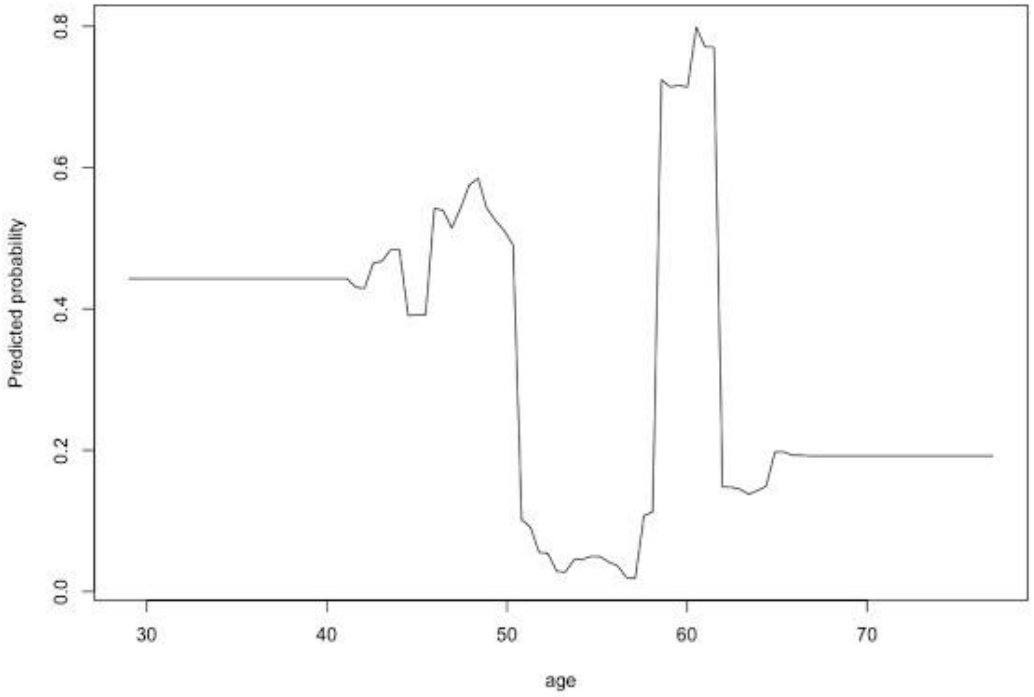
Διάγραμμα 1


Predicted probability
0.0
0.2
0.4
0.6
0.8
typical angina
atypical angina
non-anginal pain
asymptomatic
cp

Διάγραμμα 2



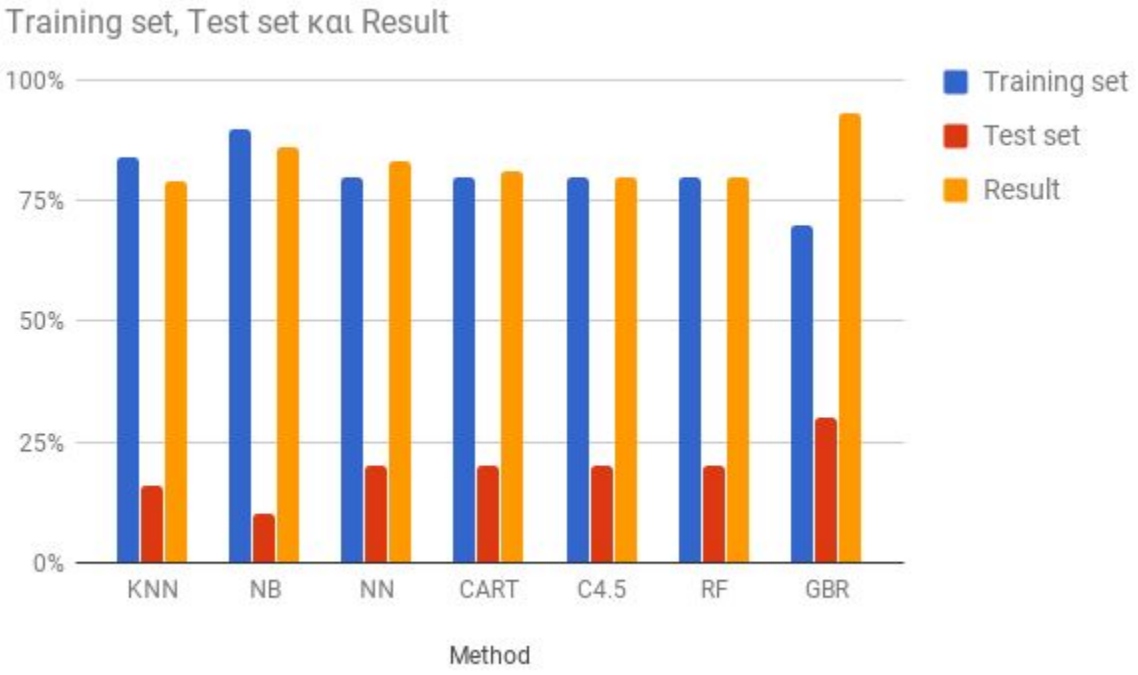
Διάγραμμα 3



Αποτελέσματα

Αλγόριθμος	Training set	Test set	Ποσοστό
KNN	**84%**	**16%**	**79%**
Naive Bayes	**90%**	**10%**	**86%**
Neural Network	**80%**	**20%**	**83%**
CART	**80%**	**20%**	**81%**
C4.5	**80%**	**20%**	**80%**
Random Forest	**80%**	**20%**	**80%**
GBR	**70%**	**30%**	**93%**

Αποτελέσματα



Συμπεράσματα

- **General Boosted Regression**
- Έλεγχος των αποτελεσμάτων από ειδικό
- Μεγαλύτερο πλήθος δεδομένων = Καλύτερη απόδοση
- Η πολυωνυμική κατηγοριοποίηση θα ήταν δυνατή με αρκετά δεδομένα για όλες τις κατηγορίες
- Ευκολότερη πρόσβαση σε ιατρικά δεδομένα

Μελλοντική Επέκταση

- **Online** υπηρεσία συμβουλευτικής πρόβλεψης
- Υπηρεσία - ιατρικό εργαλείο
- Συλλογή ανώνυμων ιατρικών δεδομένων για βελτίωση
- Μελέτη και βελτίωση των μοντέλων
- Ευκολότερη πρόσβαση σε ιατρικά δεδομένα

Σας ευχαριστώ!

Ερωτήσεις - Απορίες?

Επικοινωνία:
axilleasmoukoulis@gmail.com
https://github.com/AxilleasMoukoulis

“Numbers have an important story to tell. They rely on you to give them a voice.

-Stephen Few