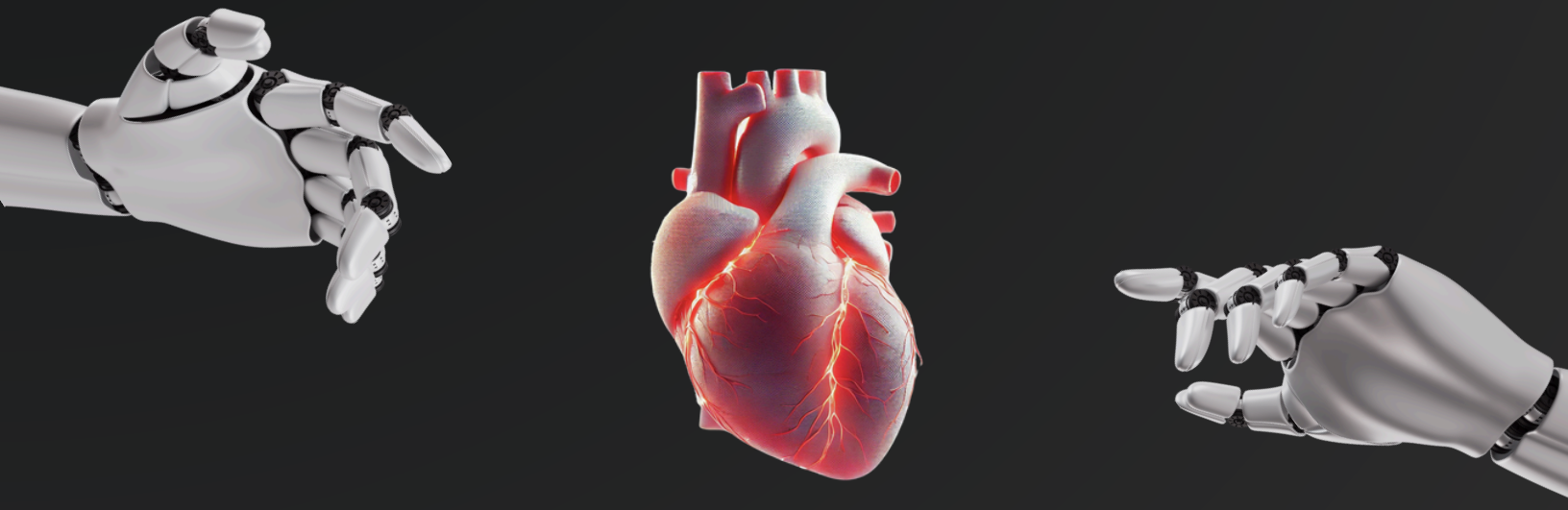
KIORAI MARIOS
University of Macedonia



Predicting Heart Attack with machine learning

Heart attack prediction

Kiorai Marios

2024-07-14

Αυτό το project χρησιμοποιεί μηχανική μάθηση για την ανάπτυξη ενός προβλεπτικού μοντέλου που εκτιμά τον κίνδυνο στεφανιαίας νόσου (CHD) σε ένα διάστημα 10 ετών, μια χρόνια κατάσταση που συνδέεται στενά με το έμφραγμα. Βασίζεται στο σύνολο δεδομένων της Μελέτης Καρδιάς Framingham και αναλύει παράγοντες, όπως τα επίπεδα χοληστερόλης, την αρτηριακή πίεση, την ηλικία, το φύλο, τις συνήθειες καπνίσματος και τη γλυκόζη. Αναπτύχθηκαν και συγκρίθηκαν δύο προβλεπτικά μοντέλα: ένα μοντέλο Δέντρου Ταξινόμησης και Παλινδρόμησης (CART) και ένα μοντέλο λογιστικής παλινδρόμησης.

Data: Framingham Coronary Disease Prediction Analysis

Goal: Predict the 10-year coronary heart disease (CHD) risk using machine learning models

1) Loading Data and Libraries and preparing data for analysis

```
## Load necessary libraries library(caTools) library(rpart)
library(rpart.plot) library(ROCR) library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
## Import data from CSV file
Data <- read.csv("C:/Users/mario/Downloads/framingham.csv")
## Cleaning data by removing NA values
CleanData <- na.omit(Data)
## Split data into Train/Test sets
set.seed(964)
split <- sample.split(CleanData$TenYearCHD, SplitRatio = 0.65)
train <- subset(CleanData, split == TRUE)
test <- subset(CleanData, split == FALSE)
cat("Training rows:", nrow(train), "Testing rows:", nrow(test), "\n")
```

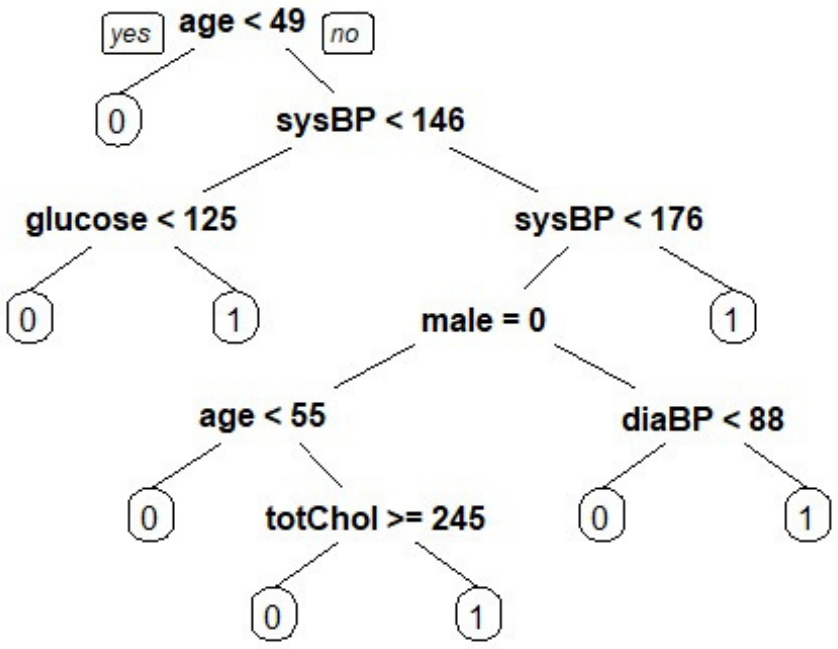
```
## Training rows: 2378 Testing rows: 1280
```

2a) Building CART model for prediction

```
## Build CART model
framinghamTree <- rpart(TenYearCHD ~ age + male + education +
currentSmoker + cigsPerDay + BPMeds + prevalentStroke + prevalentHyp +
totChol + sysBP + diaBP + BMI + heartRate + glucose,

                        data = train,
                        method = "class",
                        parms = list(prior = c(0.7, 0.3)))

## Visualize the decision tree
prp(framinghamTree)
```



Το παραπάνω δέντρο απόφασης δείχνει πως διάφοροι παράγοντες όπως η ηλικία, η συστολική αρτηριακή πίεση, το επίπεδο γλυκόζης και το φύλο συμβάλλουν στην πρόβλεψη του κινδύνου καρδιακής νόσου σε 10 χρόνια (TenYearCHD). Κάθε διακλάδωση στο δέντρο αντιπροσωπεύει μια απόφαση βασισμένη σε μία από αυτές τις μεταβλητές, και κάθε τελικός κόμβος (φύλλο) δίνει την πρόβλεψη για το αν υπάρχει ή όχι κίνδυνος καρδιακής νόσου.

2b) Calculating accuracy for the CART model

```
predictCART <- predict(framinghamTree, newdata = test, type="class")
table(test$TenYearCHD, predictCART)
```

```
## ## predictCART
##       0   1
##   0 995  90
##   1 144  51

## Υπολογισμός ακρίβειας

(978+53)/(978+53+142+107)
## [1] 0.8054687
```

Έχουμε 978 True Negatives, 107 False Positives, 142 False Negatives and 53 True Positives. Έχουμε άρα ακρίβεια(accuray) του μοντέλου ίση με 80.6%, που σημαίνει ότι προβλέπει σωστά το αποτέλεσμα για το 80.6% των περιπτώσεων στο σύνολο δοκιμασίας.

3) Building Logistic Regression model for prediction

```
## Build Logistic Regression model
logisticModel <- glm(TenYearCHD ~ age + male + education +
currentSmoker + cigsPerDay + BPMeds + prevalentStroke + prevalentHyp +
totChol + sysBP + diaBP + BMI + heartRate + glucose,
                     data = train,
                     family = binomial)

## Display model summary to check coefficients
summary(logisticModel)
##
## Call:
## glm(formula = TenYearCHD ~ age + male + education + currentSmoker +
##     cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + totChol +
##     sysBP + diaBP + BMI + heartRate + glucose, family = binomial,
##     data = train)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.148e+00 8.724e-01 -9.340 < 2e-16 ***
## age                6.227e-02 8.241e-03   7.557 4.13e-14 ***
## male               4.622e-01 1.344e-01   3.440 0.000582 ***
## education         -2.838e-02 6.075e-02 -0.467 0.640335
## currentSmoker      8.568e-02 1.919e-01   0.446 0.655490
## cigsPerDay         2.461e-02 7.554e-03   3.258 0.001123 **
## BPMeds             3.375e-01 2.929e-01   1.152 0.249317
## prevalentStroke    8.557e-01 5.943e-01   1.440 0.149930
## prevalentHyp       1.692e-01 1.747e-01   0.968 0.332961
## totChol            1.514e-03 1.494e-03   1.013 0.310914
## sysBP              1.893e-02 4.899e-03   3.863 0.000112 ***
## diaBP             -7.304e-03 8.041e-03  -0.908 0.363725
## BMI                2.078e-02 1.580e-02   1.316 0.188225
## heartRate         -5.874e-03 5.293e-03  -1.110 0.267068
## glucose            8.491e-03 1.977e-03   4.295 1.75e-05 ***
```

```
## --- ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1 ## ## (Dispersion parameter for binomial family taken to be
1) ## ##


       Null deviance: 2028.7 on 2377 degrees of freedom
## Residual deviance: 1787.6 on 2363 degrees of freedom ## AIC: 1817.6
## ## Number of Fisher Scoring iterations: 5
```

```
## Predict probabilities for both models
PredictROC_CART <- predict(framinghamTree, newdata = test, type =
"prob")[,2]
PredictROC_Logistic <- predict(logisticModel, newdata = test, type =
"response")

## Combine predictions into a data frame for visualization
predictions_df <- data.frame(
 TrueLabel = test$TenYearCHD,
 CART_Prob = PredictROC_CART,
 Logistic_Prob = PredictROC_Logistic
)
```

Logistic Regression model Summary

Η σχέση που χρησιμοποιήθηκε στην παλινδρόμηση δείχνει την πρόβλεψη καρδιακού προβλήματος με βάση την τις παραάτω μεταβλητές.

- **Άνδρας:** Θετικός συντελεστής, που δείχνει ότι το να είναι κανείς άνδρας συνδέεται με υψηλότερο κίνδυνο.
- **Ηλικία:** Θετικός συντελεστής, που υποδεικνύει ότι ο κίνδυνος αυξάνεται με την ηλικία.
- **CigsPerDay(Τσιγάρα ανά ημέρα):** Θετικός συντελεστής, που δείχνει ότι το κάπνισμα περισσότερων τσιγάρων την ημέρα συνδέεται με υψηλότερο κίνδυνο καρδιακής νόσου.
- **sysBP(Συστολική Αρτηριακή Πίεση):** Θετική επίδραση, που υποδηλώνει ότι η υψηλότερη πίεση αίματος σχετίζεται με αυξημένο κίνδυνο καρδιακής νόσου.
- **Γλυκόζη:** Θετικός συντελεστής, που υποδεικνύει ότι ένα υψηλότερο επίπεδο γλυκόζης σχετίζεται επίσης με αυξημένο κίνδυνο.

4) Building Linear Regression model for prediction

```
## Build Linear Regression model
linearModel <- lm(TenYearCHD ~ age + male + education + currentSmoker +
cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP
+ diaBP + BMI + heartRate + glucose,
                  data = train)
```

```r
## Display summary to see the effect of variables
summary(linearModel)
```

```
##
## Call:
## lm(formula = TenYearCHD ~ age + male + education + currentSmoker +
##     cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + totChol +
##     sysBP + diaBP + BMI + heartRate + glucose, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8176 -0.1858 -0.1064 -0.0146 1.0941
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.5884649 0.0958215  -6.141 9.58e-10 ***
## age               0.0070504 0.0009459   7.453 1.27e-13 ***
## male              0.0468699 0.0154727   3.029 0.00248 **
## education        -0.0035069 0.0069651  -0.504 0.61466
## currentSmoker     0.0106516 0.0222302   0.479 0.63188
## cigsPerDay        0.0018487 0.0009435   1.959 0.05019 .
## BPMeds            0.0351416 0.0427353   0.822 0.41098
## prevalentStroke 0.1740049 0.0925900     1.879 0.06033 .
## prevalentHyp      0.0201190 0.0217513   0.925 0.35508
## totChol          -0.0001425 0.0001724  -0.827 0.40846
## sysBP             0.0029349 0.0006292   4.665 3.27e-06 ***
## diaBP            -0.0015073 0.0010214  -1.476 0.14016
## BMI               0.0019884 0.0019244   1.033 0.30159
## heartRate        -0.0007293 0.0006069  -1.202 0.22956
## glucose           0.0014782 0.0002782   5.313 1.18e-07 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3408 on 2363 degrees of freedom
## Multiple R-squared: 0.1059, Adjusted R-squared: 0.1006
## F-statistic:    20 on 14 and 2363 DF, p-value: < 2.2e-16
```

Linear Regression model Summary

Η σχέση που χρησιμοποιήθηκε στην παλινδρόμηση δείχνει την πρόβλεψη καρδιακού προβλήματος με βάση την τις παραάτω μεταβλητές.

- **Άνδρας:** αυξάνει τον κίνδυνο CHD κατά περίπου 0,73% ανά έτος.
- **Ηλικία:** αυξάνει τον κίνδυνο CHD κατά περίπου 4,8%.
- **Τσιγάρα ανά ημέρα(CigsPerDay):** αυξάνουν τον κίνδυνο κατά 0,2% για κάθε τσιγάρο που καπνίζεται ημερησίως.
- **Συστολική Αρτηριακή Πίεση (sysBP):** αυξάνει τον κίνδυνο κατά 0,26% για κάθε μονάδα.

Γλυκόζη: αυξάνει τον κίνδυνο CHD κατά περίπου 0,15% ανά μονάδα.

Όλοι οι παράγοντες είναι στατιστικά σημαντικοί. Ωστόσο, το R-squared (10%) του μοντέλου εξηγεί μόνο ένα μικρό μέρος του κινδύνου CHD, υποδεικνύοντας ότι άλλοι μη μετρημένοι παράγοντες συμβάλλουν στον υπόλοιπο κίνδυνο οι οποίοι δεν υπάρχουν στο σύνολο των δεδομένων μας. Άλλοι πιθανοί παράγοντες που μπορεί να επηρεάζουν τον κίνδυνο CHD:

- **Χοληστερόλη(Cholesterol)**
- **Διατροφή και σωματική δραστηριότητα(Diet and Physical Activity)**
- **Δείκτης Μάζας Σώματος(BMI)**
- **Οικογενειακό Ιστορικό Καρδιαγγειακής Νόσου(Family History)**
- **Στρες και Ψυχολογικοί Παράγοντες(Stress and Psychological Factors)**

4) Calculating correlations

```
## Calculate correlation matrix
cor_matrix <- cor(CleanData[, c("age", "cigsPerDay", "sysBP",
"glucose", "male")])
## Display correlation matrix
cor_matrix
```

```
##
male
## age
0.024411605
## cigsPerDay -0.1892953 1.00000000 -0.09478131 -0.053726433
0.330322451
## sysBP
0.045358037
## glucose
0.002939977
## male
1.000000000
```

Correlation Summary

- **Ηλικία και sysBP:** έχουν μέτρια θετική συσχέτιση (38,8%), υποδεικνύοντας
- ότι τα άτομα μεγαλύτερης ηλικίας τείνουν να έχουν υψηλότερη πίεση.
- **Ηλικία:** αυξάνει τον κίνδυνο CHD κατά περίπου 4,8%.
- **Άνδρας και τσιγάρα :** είναι μέτρια συσχετισμένα (33%), με τους άνδρες να καπνίζουν περισσότερο κατά μέσο όρο.

  Others: Οι υπόλοιπες μεταβλητές έχουν αδύναμες συσχετίσεις, υποδεικνύοντας περιορισμένη επίδραση μεταξύ τους.

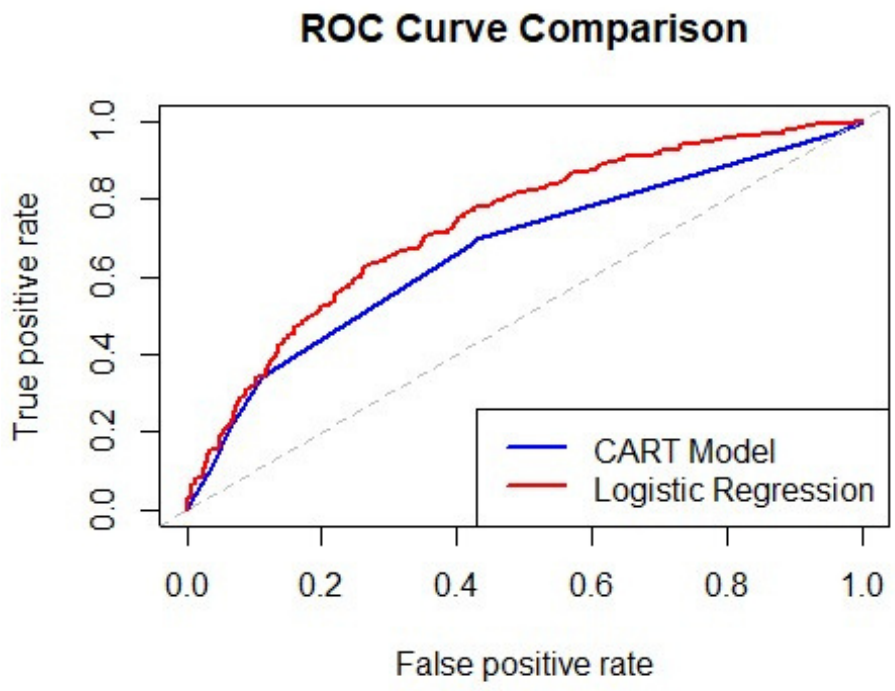
6) ROC Curve Comparison

```
## ROC Curve for CART Model
pred_CART <- prediction(PredictROC_CART, test$TenYearCHD)
```

```r
perf_CART <- performance(pred_CART, "tpr", "fpr")
plot(perf_CART, main = "ROC Curve Comparison", col = "blue", lwd = 2)
abline(a = 0, b = 1, col = "gray", lty = 2)

## ROC Curve for Logistic Regression Model
pred_Logistic <- prediction(PredictROC_Logistic, test$TenYearCHD)
perf_Logistic <- performance(pred_Logistic, "tpr", "fpr")
plot(perf_Logistic, col = "red", lwd = 2, add = TRUE)
## Add Legend
legend("bottomright", legend = c("CART Model", "Logistic Regression"),

       col = c("blue", "red"), lwd = 2)
```



Η καμπύλη ROC παραπάνω συγκρίνει την απόδοση των δύο μοντέλων μηχανικής μάθησης (CART και λογιστικής παλινδρόμησης) ως προς την ακρίβεια. Το μοντέλο λογιστικής παλινδρόμησης (κόκκινη γραμμή) βρισκεται πάνω από το μοντέλο CART (μπλε γραμμή) και αυτό μας δείχνει ότι η λογιστική παλινδρόμηση είναι πιο αποτελεσματική στον διαχωρισμό ασθενών με υψηλό και χαμηλό κίνδυνο. Μια τιμή AUC πιο κοντά στο 1 δείχνει καλύτερη απόδοση, και η λογιστική παλινδρόμηση έχει υψηλότερη AUC.