UNIVERSITY OF PATRAS

SCHOOL OF ENGINEERING

Department of Computer Engineering & Informatics

Division of Applications and Foundations of Computer Science

Pattern Recognition Laboratory

Εργαστηριακή Άσκηση για το μάθημα Θεωρία Αποφάσεων

2019-2020

Παναγιώτης Κωνσταντίνος Αθανασόπουλος Α.Μ:1058112

Ερώτημα 1.
Τα χαρακτηριστικά κάθε δείγματος είναι 9 και είναι τα εξής:
Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome

Το αρχείο περιέχει 768 δείγματα εκπαίδευσης.

Τα δείγματα των ασθενών που πάσχουν απο διαβήτη είναι 268 ενώ τα δείγματα των ασθενών που δεν πάσχουν από διαβήτη είναι 500.

Ερώτημα 2.

Ο αφελής ταξινομητής Bayes Naive Bayes είναι μια απλή τεχνική για την κατασκευή ταξινομητών: μοντέλα που αποδίδουν ετικέτες κλάσης σε στιγμιότυπα προβλημάτων που αντιπροσωπεύονται ως φορείς των τιμών χαρακτηριστικών, όπου οι ετικέτες κλάσης προέρχονται από κάποια πεπερασμένη ομάδα. Δεν υπάρχει ένας και μόνος αλγόριθμος για την κατάρτιση αυτών των ταξινομητών, αλλά μια οικογένεια αλγορίθμων βασισμένη σε μια κοινή αρχή: όλοι οι απλοί ταξινομητές Bayes υποθέτουν ότι η αξία ενός συγκεκριμένου στοιχείου είναι ανεξάρτητη από την αξία οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης. Για παράδειγμα, ένας καρπός μπορεί να θεωρηθεί μήλο αν είναι κόκκινο, στρογγυλό και περίπου 10 cm σε διάμετρο. Ένας αφελής ταξινομητής Bayes θεωρεί ότι κάθε ένα από αυτά τα χαρακτηριστικά συμβάλλει ανεξάρτητα στην πιθανότητα ότι αυτός ο καρπός είναι ένα μήλο, ανεξάρτητα από οποιαδήποτε πιθανή συσχέτιση μεταξύ των χαρακτηριστικών χρώματος, στρογγυλότητας και διαμέτρου.

Για ορισμένους τύπους μοντέλων πιθανοτήτων, οι αφελείς ταξινομητές Bayes μπορούν να εκπαιδευτούν πολύ αποτελεσματικά σε μια εποπτευόμενη μαθησιακή ρύθμιση. Σε πολλές πρακτικές εφαρμογές, η εκτίμηση παραμέτρων για τα αφημένα μοντέλα Bayes χρησιμοποιεί τη μέθοδο της μέγιστης πιθανότητας. Με άλλα λόγια, μπορεί κανείς να εργαστεί με το αφελές μοντέλο Bayes χωρίς να δεχτεί τη Bayesian πιθανότητα ή χρησιμοποιώντας οποιαδήποτε Bayesian μεθόδους.

Παρά το αφελές σχεδιασμό τους και τις φαινομενικά υπεραπλουστευμένες υποθέσεις, οι αφελείς ταξινομητές Bayes εργάστηκαν αρκετά καλά σε πολλές πολύπλοκες πραγματικές καταστάσεις. Το 2004, μια ανάλυση του Bayesian ταξινομικού προβλήματος έδειξε ότι υπάρχουν σοβαροί θεωρητικοί λόγοι για την προφανώς μη εφαρμόσιμη αποτελεσματικότητα των αφελών ταξινομητών Bayes. Παρόλα αυτά, μια συνολική σύγκριση με

άλλους αλγορίθμους ταξινόμησης το 2006 έδειξε ότι η ταξινόμηση Bayes υπερέχει από άλλες προσεγγίσεις, όπως ενισχυμένα δέντρα ή τυχαία δάση.

Ένα πλεονέκτημα του αφελούς Bayes είναι ότι απαιτεί μόνο ένα μικρό αριθμό δεδομένων εκπαίδευσης για την εκτίμηση των παραμέτρων που είναι απαραίτητες για την ταξινόμηση.

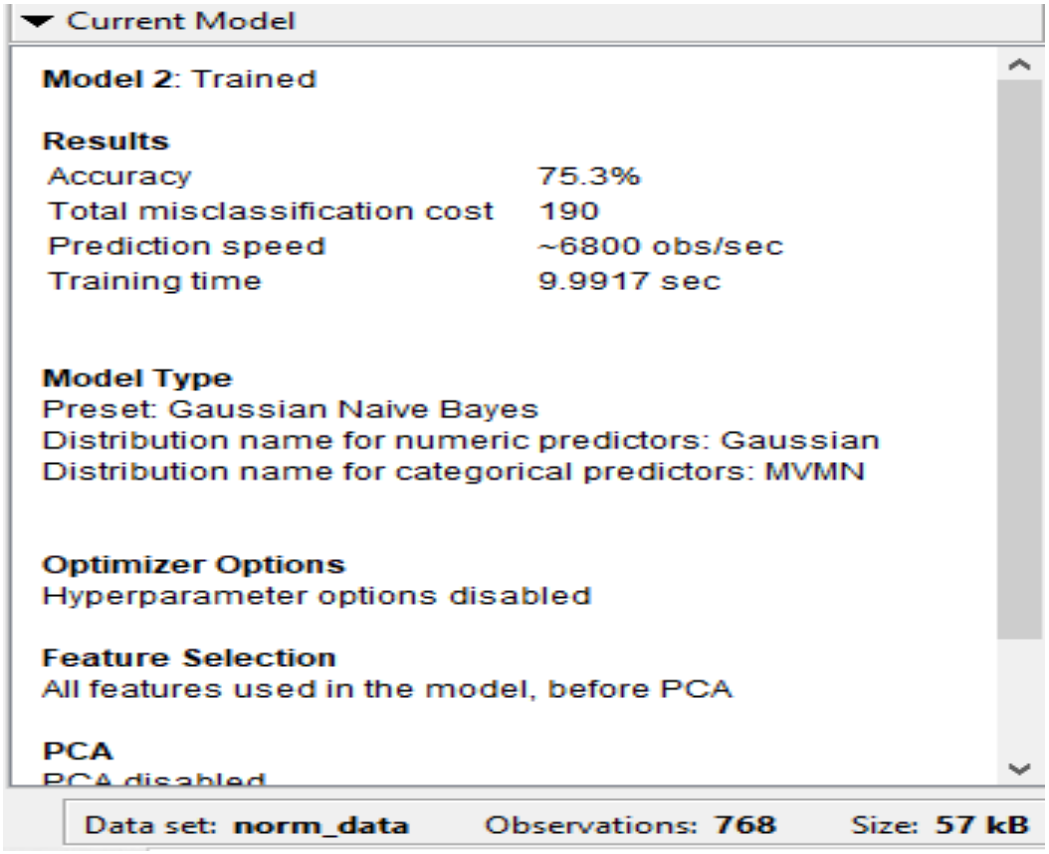
Gaussian αφελής Bayes αλγόριθμος

Όταν πρόκειται για συνεχή δεδομένα, μια τυπική παραδοχή είναι ότι οι συνεχείς τιμές που σχετίζονται με κάθε κατηγορία κατανέμονται σύμφωνα με μια κανονική (ή Gaussian) κατανομή. Για παράδειγμα, υποθέτουμε ότι τα δεδομένα κατάρτισης περιέχουν ένα συνεχές χαρακτηριστικό x. Κατατάσσουμε πρώτα τα δεδομένα από την κλάση και στη συνέχεια υπολογίζουμε το μέσο όρο και τη διακύμανσή του x σε κάθε κατηγορία.Έστω μ_κ ο μέσος όρος των τιμών στο x που σχετίζονται με την τάξη C_k και σ_κ^2 να είναι η διορθωμένη Bessel διακύμανση των τιμών σε x που σχετίζονται με την κατηγορία C_k. Ας υποθέσουμε ότι έχουμε συγκεντρώσει κάποια τιμή παρατήρησης u. Στη συνέχεια, η πιθανότητα *διανομής* της u δεδομένης μιας τάξης C_k, p(x = u| C_k), μπορεί να υπολογιστεί με σύνδεση u στην εξίσωση για κανονική κατανομή παραμετροποιημένη από μ_κ και σ_κ^2. Αυτό είναι, p(x = u| C_k) =

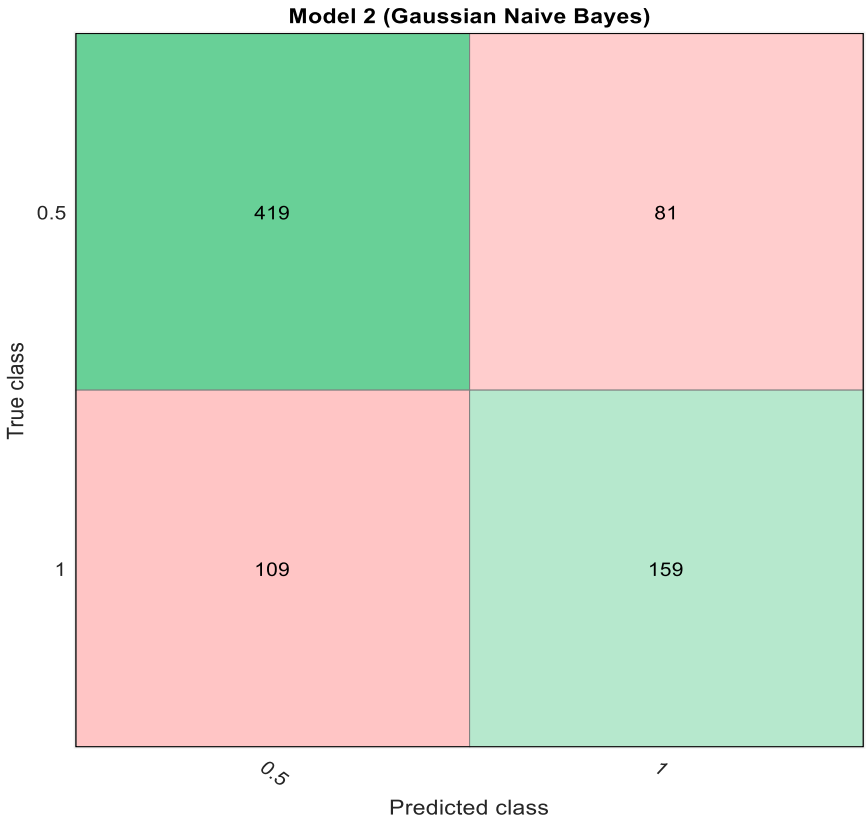
$$\frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Ερώτημα 3.

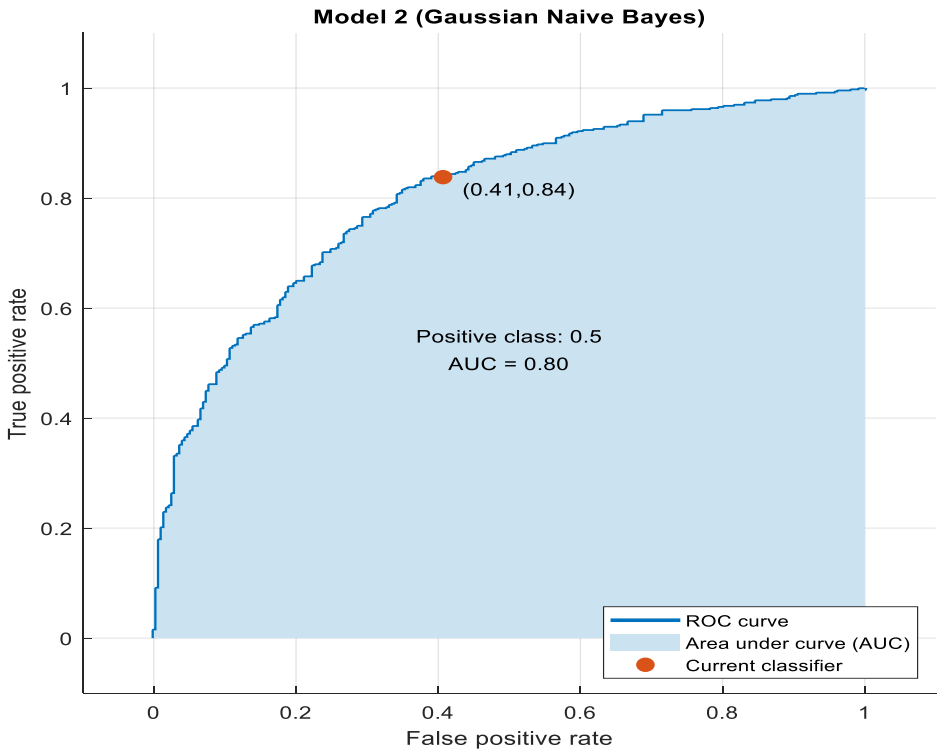
Εκπαιδεύοντας τα δεδομένα με την μέθοδο 5 fold cross validation η απόδοση της μεθόδου που προκύπτει είναι 75,3% η οποία είναι αρκετά καλή.



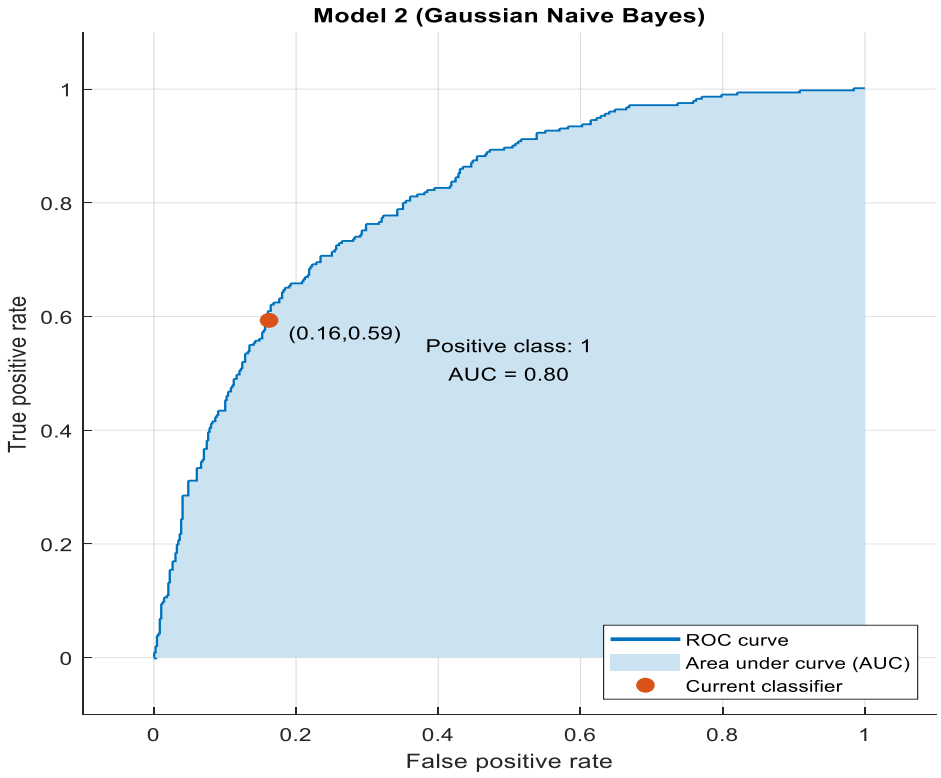
Confusion Matrix



ROC Curve για την κλάση 0,5



ROC Curve για την κλάση 1



Από το confusion matrix:

TP(true positive)=419, TN(true negative)=159, FN(false negative)=81, FP(false positive)=109

Sensitivity=$\frac{TP}{TP+FN}=\frac{419}{419+81}=\frac{419}{500}$=0,838

Specificity=$\frac{TN}{TN+FP}=\frac{159}{159+109}=\frac{159}{268}$=0,593

Επομένως,η μετρική του γεωμετρικού μέσου (Geometric Mean) της ευαισθησίας (Sensitivity) και της ειδίκευσής (Specificity) είναι:

Geometric Mean = sqrt (Sensitivity * Specificity)=sqrt(0,838*0,593)=sqrt(0,496934)= 0.7049

Ερώτημα 4.

- Για τον ταξινομητή κ κοντινότερου γείτονα

Εκπαιδεύοντας τα δεδομένα με την μέθοδο 5 fold cross validation η απόδοση της μεθόδου που προκύπτει είναι 72,7% η οποία είναι αρκετά καλή.

Current Model

Model 4: Trained

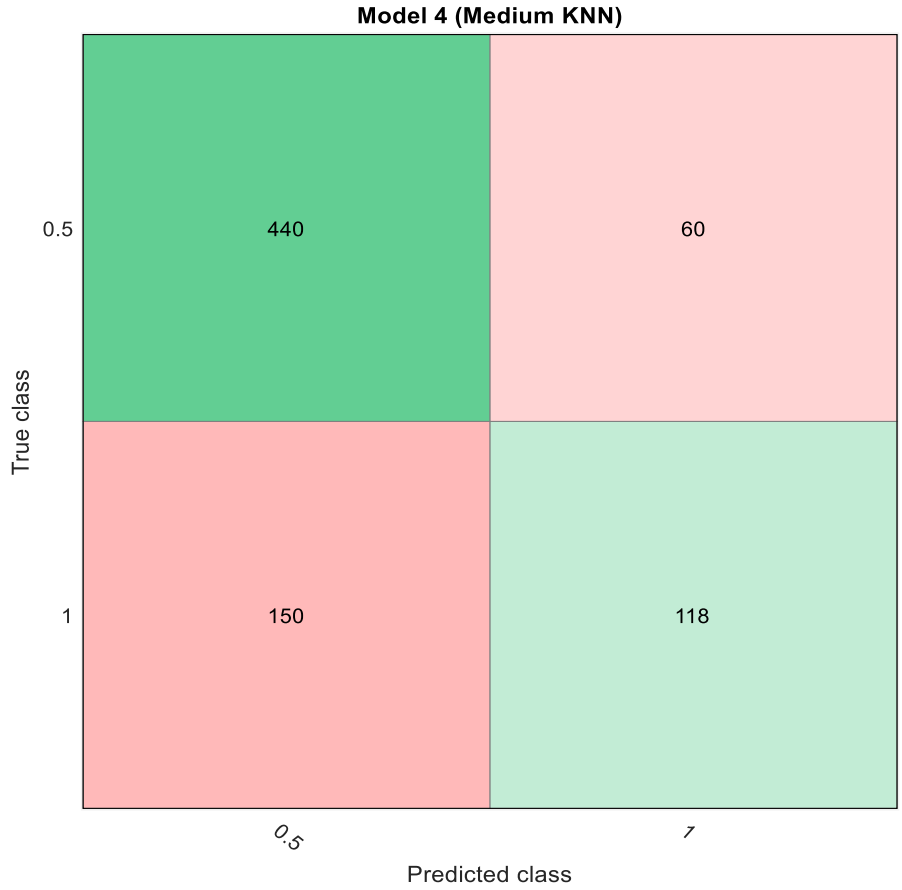
Results

Accuracy	72.7%
Total misclassification cost	210
Prediction speed	~3700 obs/sec
Training time	14.679 sec

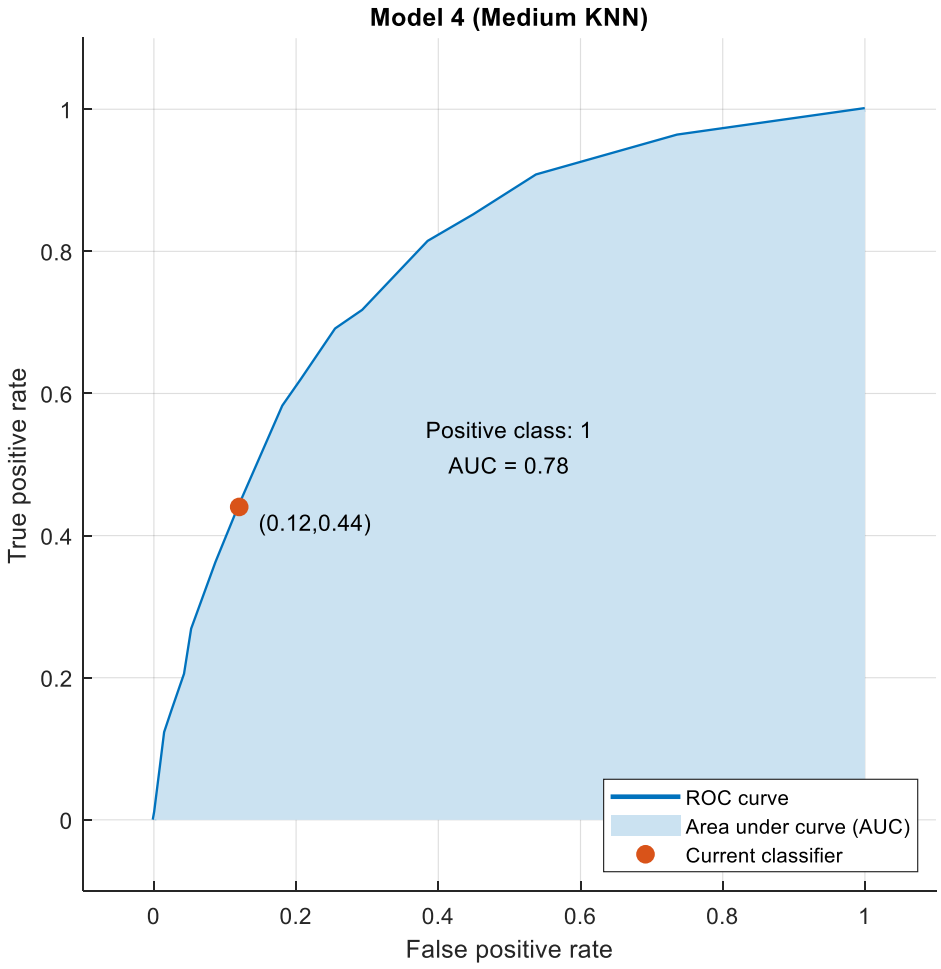
Model Type

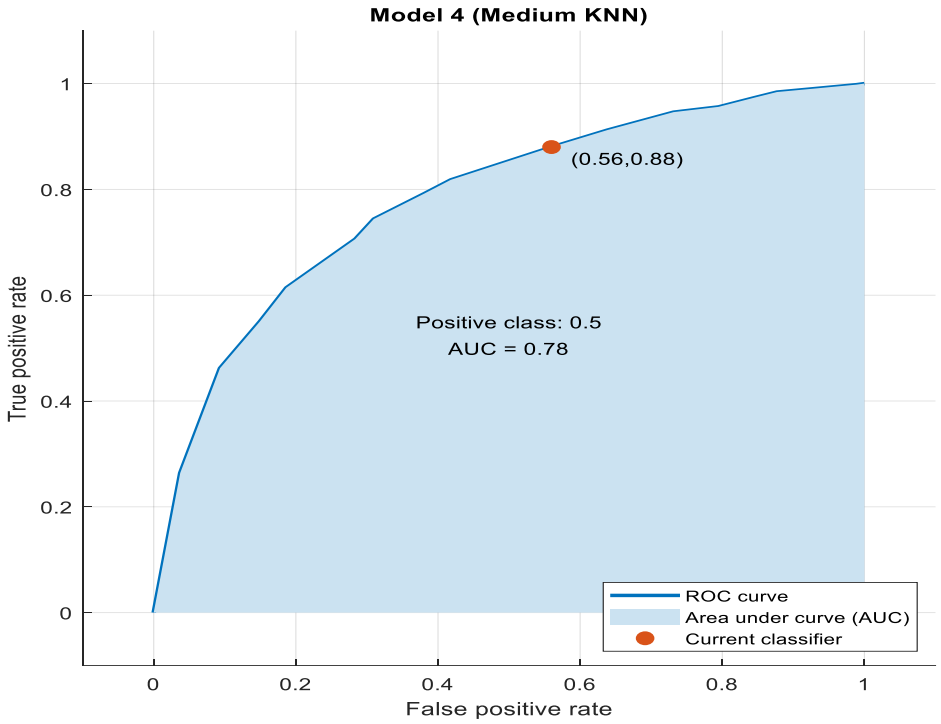
Preset: Medium KNN
Number of neighbors: 10
Distance metric: Euclidean
Distance weight: Equal
Standardize data: true

Confusion Matrix



ROC Curve για την κλάση 0,5





Από το confusion matrix:

TP(true positive)=440, TN(true negative)=118, FN(false negative)=60, FP(false positive)=150

Sensitivity=$\frac{TP}{TP+FN}=\frac{440}{440+60}=\frac{419}{500}$=0,838

Specificity=$\frac{TN}{TN+FP}=\frac{118}{118+150}=\frac{118}{268}$=0,44

Επομένως,η μετρική του γεωμετρικού μέσου (Geometric Mean) της ευαισθησίας (Sensitivity) και της ειδίκευσής (Specificity) είναι:

Geometric Mean = sqrt (Sensitivity * Specificity)=sqrt(0,838*0,44)=sqrt(0,36872)= 0.6072

- Για τον ταξινομητή Support Vector Machines με Radial Basis Function kernel function με βήμα 5

Εκπαιδεύοντας τα δεδομένα με την μέθοδο 5 fold cross validation η απόδοση της μεθόδου που προκύπτει είναι 75,8% η οποία είναι αρκετά καλή.

Current Model

Model 6: Trained

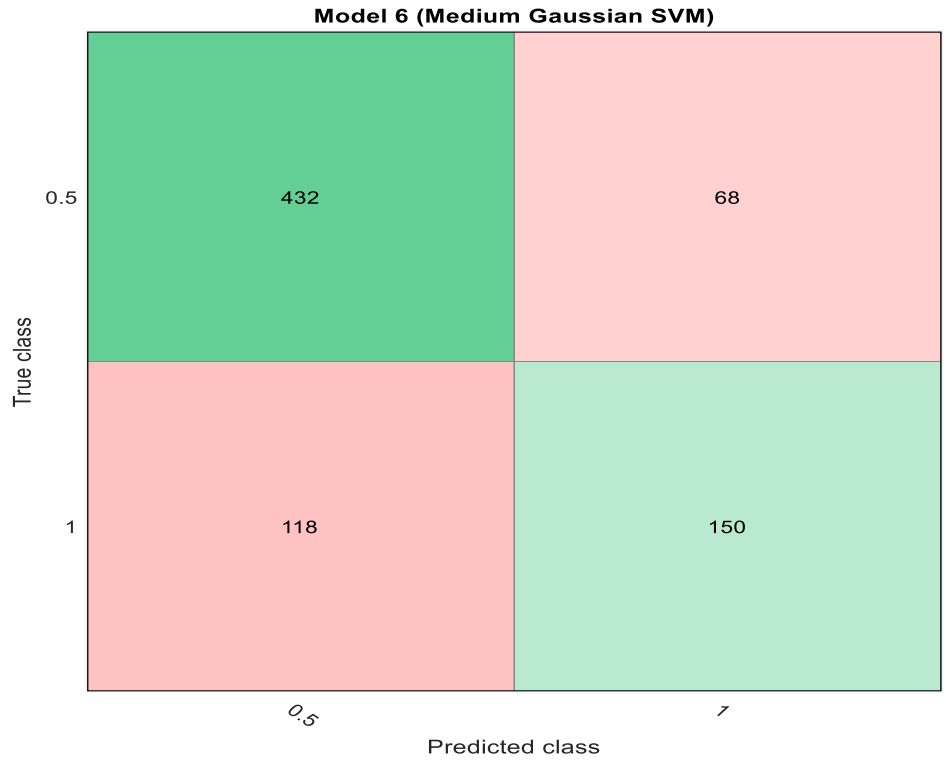
Results

Accuracy	75.8%
Total misclassification cost	186
Prediction speed	~11000 obs/sec
Training time	3.2982 sec

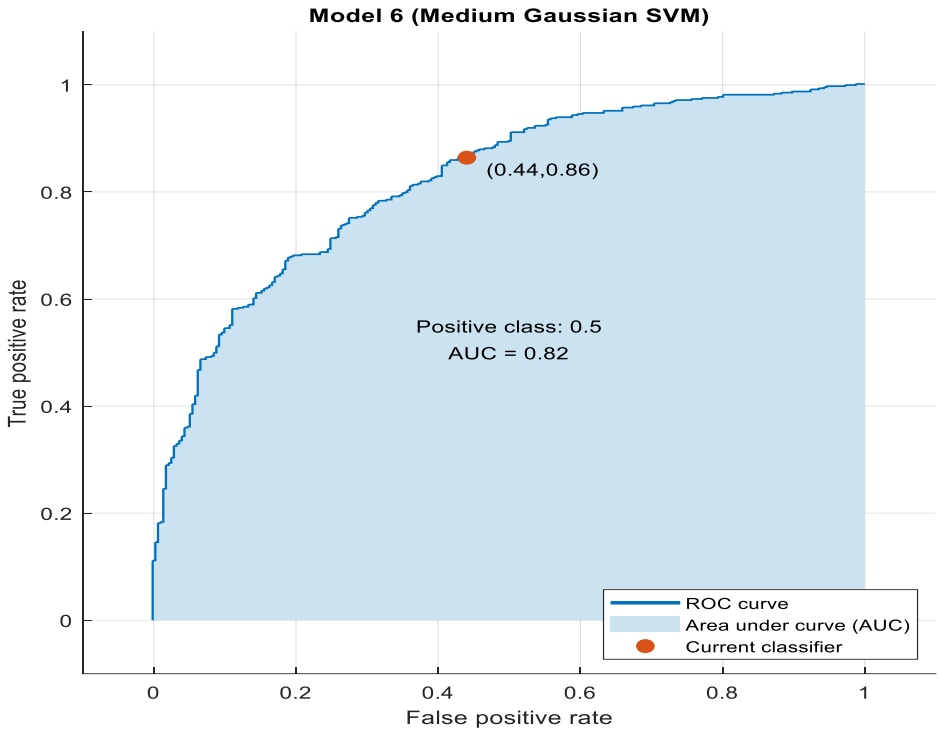
Model Type

Preset: Medium Gaussian SVM
Kernel function: Gaussian
Kernel scale: 2.8
Box constraint level: 1
Multiclass method: One-vs-One
Standardize data: true

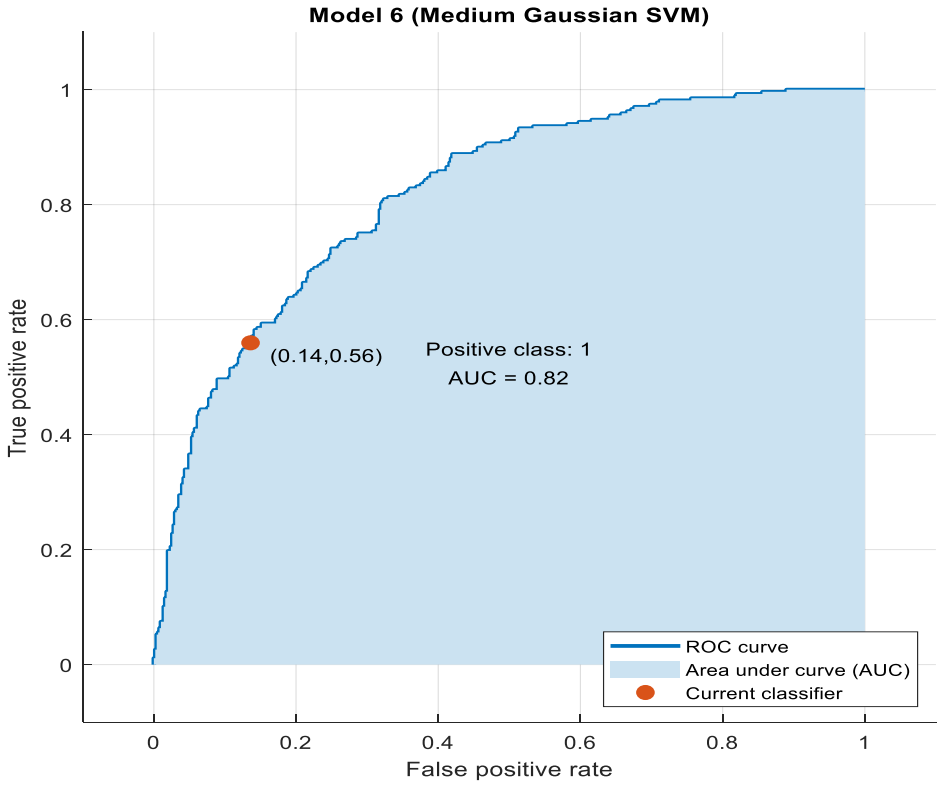
Confusion Matrix



ROC Curve για την κλάση 0,5



ROC Curve για την κλάση 1



Από το confusion matrix:

TP(true positive)=432, TN(true negative)=150, FN(false negative)=68, FP(false positive)=118

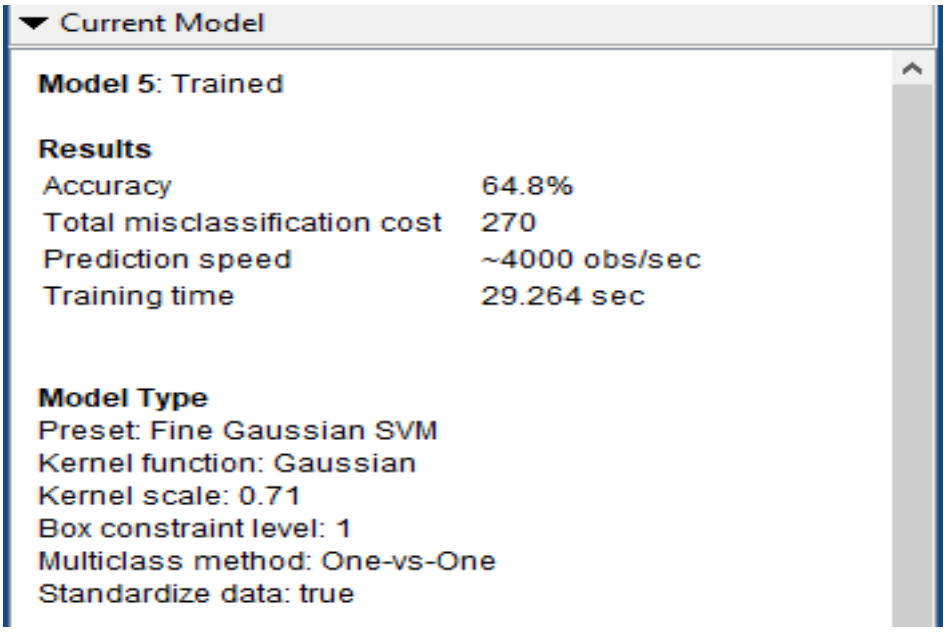
Sensitivity=$\frac{TP}{TP+FN}=\frac{432}{432+68}=\frac{432}{500}$=0,864

Specificity=$\frac{TN}{TN+FP}=\frac{150}{150+118}=\frac{150}{268}$=0,559

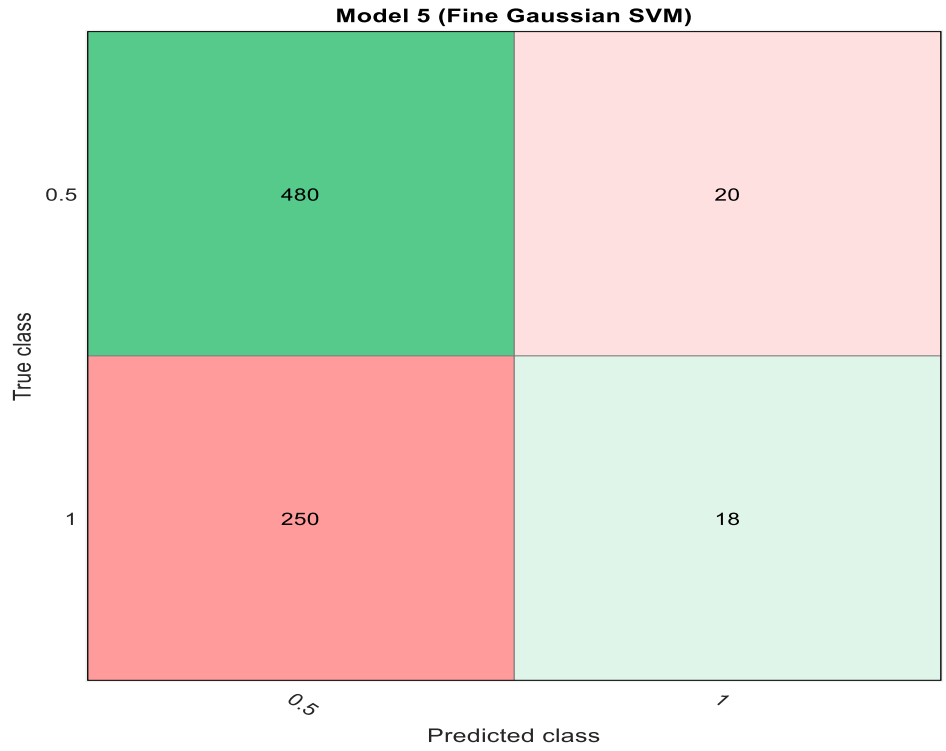
Επομένως,η μετρική του γεωμετρικού μέσου (Geometric Mean) της ευαισθησίας (Sensitivity) και της ειδίκευσής (Specificity) είναι:
Geometric Mean = sqrt (Sensitivity * Specificity)=sqrt(0,864*0,559)=sqrt(0,482976)= 0.695

- Για τον ταξινομητή Support Vector Machines με Radial Basis Function kernel function με βήμα 0,5

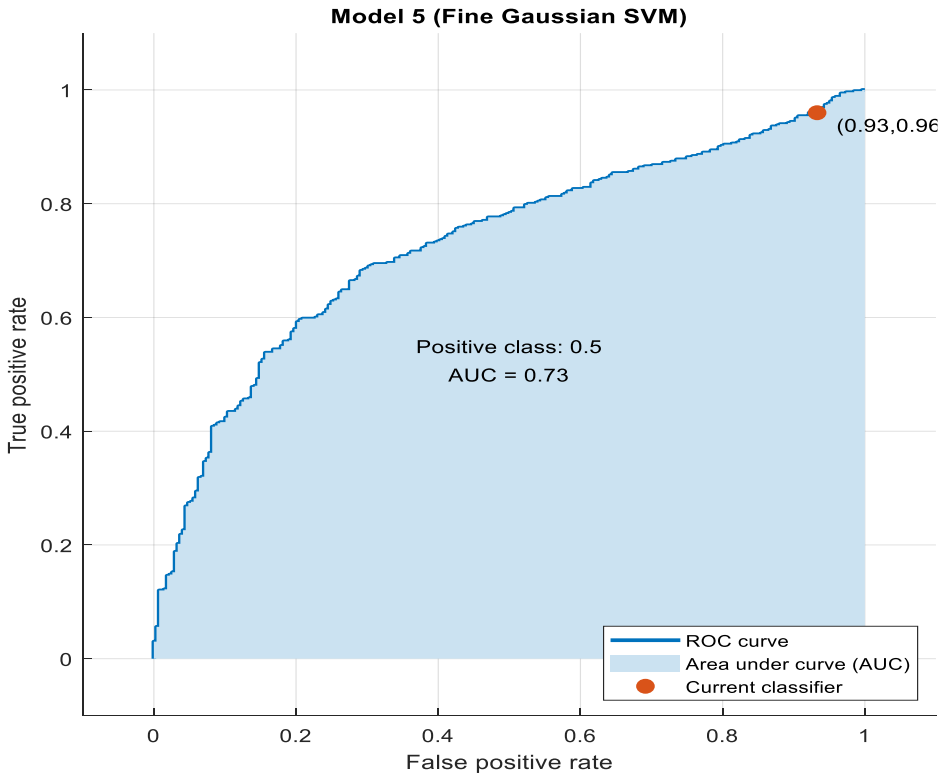
Εκπαιδεύοντας τα δεδομένα με την μέθοδο 5 fold cross validation η απόδοση της μεθόδου που προκύπτει είναι 64,8% η οποία είναι μέτρια.



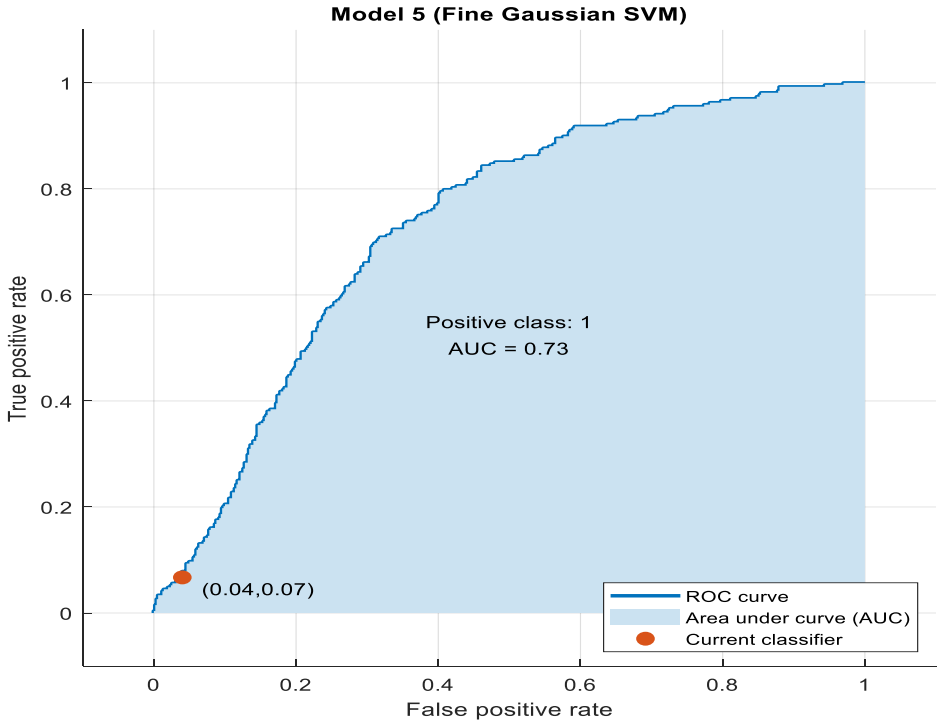
Confusion Matrix



ROC Curve για την κλάση 0,5



ROC Curve για την κλάση 1



Από το confusion matrix:

TP(true positive)=480, TN(true negative)=18, FN(false negative)=20, FP(false positive)=250

Sensitivity=$\frac{TP}{TP+FN}=\frac{480}{480+20}=\frac{480}{500}$=0,96

Specificity=$\frac{TN}{TN+FP}=\frac{18}{18+250}=\frac{18}{268}$=0,067

Επομένως,η μετρική του γεωμετρικού μέσου (Geometric Mean) της ευαισθησίας (Sensitivity) και της ειδίκευσής (Specificity) είναι:

Geometric Mean = sqrt (Sensitivity * Specificity)=sqrt(0,96*0,067)=sqrt(0,06432)= 0.2536

Επομέμως,ο ταξινομιτής με την καλύτερη απόδοση είναι ο ταξινομητής Support Vector Machines με Radial Basis Function kernel function με βήμα 5 με απόδοση 75,8%.

Όμως,ο ταξινομιτής με την μεγαλύτερη μετρική του γεωμετρικού μέσου είναι ο Naïve Bayes με γεωμετρικό μέσο 0,7049.

Η μετρική του γεωμετρικού μέσου είναι καλύτερη από την μετρίκη του αριθμητικού μέσου για ένα dataset με δεδομένα που παρουσιάζουν συσχετισμό.

Επομένως,προτείνεται ο ταξινομητής που έχει καλύτερη απόδοση με την μετρική του γεωμετρικού μέσου δηλαδή ο ταξινομητής Naïve Bayes με γεωμετρικό μέσο 0,7049.

Κάθε μέθοδος ταξινόμησης δίνει διαφορετικά αποτελέσματα γιατί εφαρμόζει διαφορετικό αλγόριθμο για την εκπαίδευση των στοιχείων.