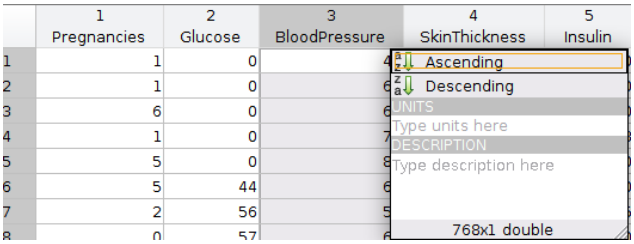
Εργαστηριακή Άσκηση για το μάθημα Θεωρία Αποφάσεων 2019-2020 Μέρος Β’

ΠΑΝΑΓΙΩΤΗΣ ΧΡΙΣΤΟΠΟΥΛΟΣ 1054409 ΕΤΟΣ Δ’

Ερώτημα 1:

Σύμφωνα με το δοθέν αρχείο csv, υπάρχουν 9 χαρακτηριστικά δειγμάτων (Αριθμός Εγκυμοσυνών, Δείκτης Γλυκόζης, Αρτηριακή Πίεση, Πάχος Δέρματος, Δείκτης Ινσουλίνης. Δείκτης Μάζας Σώματος/BMI, Συνάρτηση Κληρονομικότητας Σακχαρώδους Διαβήτη, Ηλικία και Αποτέλεσμα) για 768 δείγματα εκπαίδευσης. Ωστόσο, δεν έχουμε πληροφορίες όλων των δειγμάτων για όλες τις κατηγορίες. Με τη βοήθεια του Ascending Sort της Matlab μετά την εισαγωγή του αρχείου, μπορούμε να βρούμε τις μηδενικές πληροφορίες. Αποδεχόμαστε ότι οι αριθμοί των εγκυμοσυνών, της ηλικίας και του αποτελέσματος είναι αυθεντικοί, οπότε έχουμε τον εξής αριθμό ελλιπών πληροφοριών στις υπόλοιπες 8 κατηγορίες:



- Γλυκόζη: 5 μηδενικά Δείγματα
- Αρτηριακή Πίεση: 35 μηδενικά Δείγματα
- Πάχος Δέρματος: 227 μηδενικά Δείγματα
- Δείκτης Ινσουλίνης: 374 μηδενικά Δείγματα
- Δείκτης Μάζας Σώματος/BMI: 11 μηδενικά Δείγματα
- Συνάρτηση Κληρονομικότητας Σακχαρώδους Διαβήτη: 0 Δείγματα

Αριθμός Εγκυμοσυνών	Δείκτης Γλυκόζης	Αρτηριακή Πίεση:	Πάχος Δέρματος	Δείκτης Ινσουλίνης	Δείκτης Μάζας	Συνάρτηση Κληρονομικότητας	Ηλικία	Αποτέλεσμα
768	763	733	541	394	757	768	768	768

Κανονικοποιώντας τις τιμές στη Matlab στο πεδίο τιμών [-1,1] προκύπτει ο πίνακας N (ενδεικτικά στιγμιότυπα με βάση τον φθίνοντα αριθμό εγκυμοσυνών)

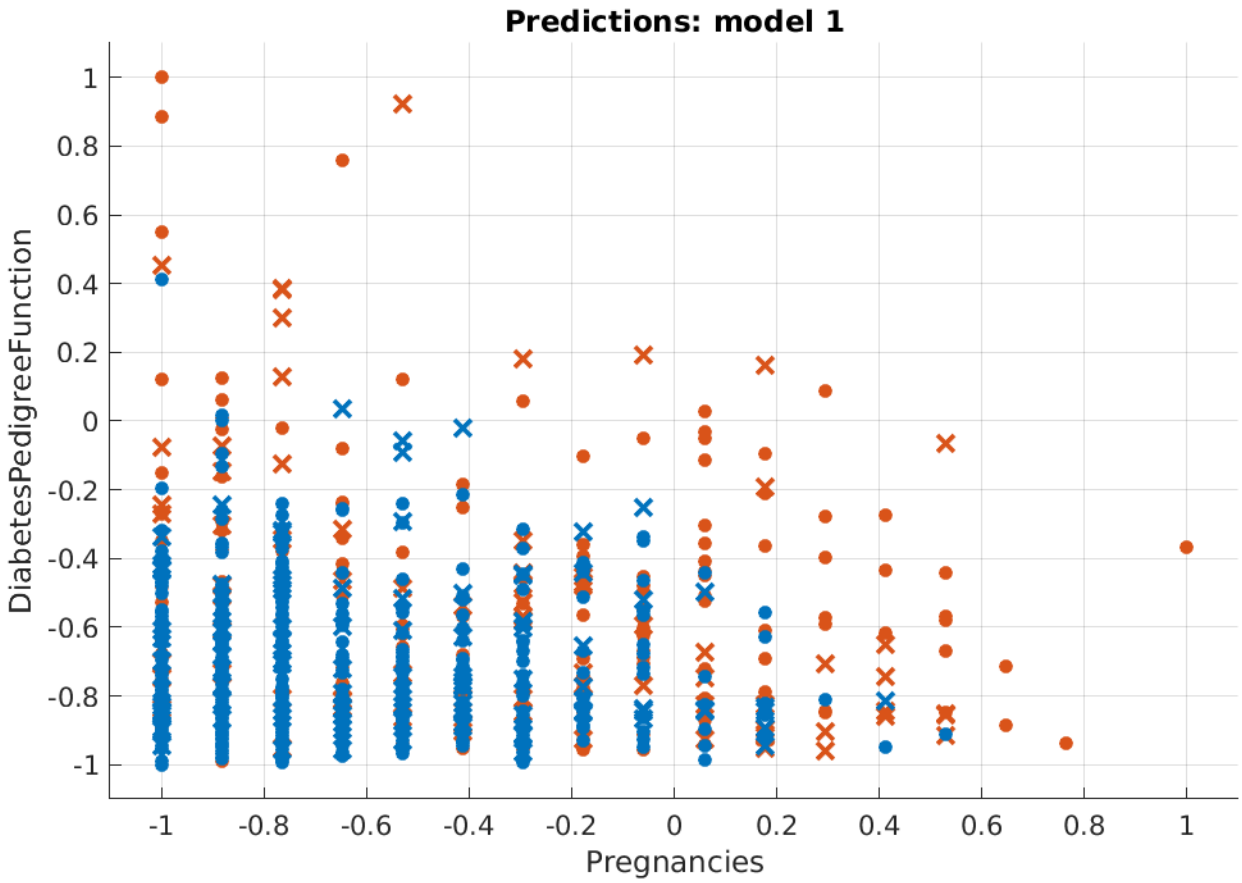
```
N = normalize(diabetes,'range',[-1,1])
```

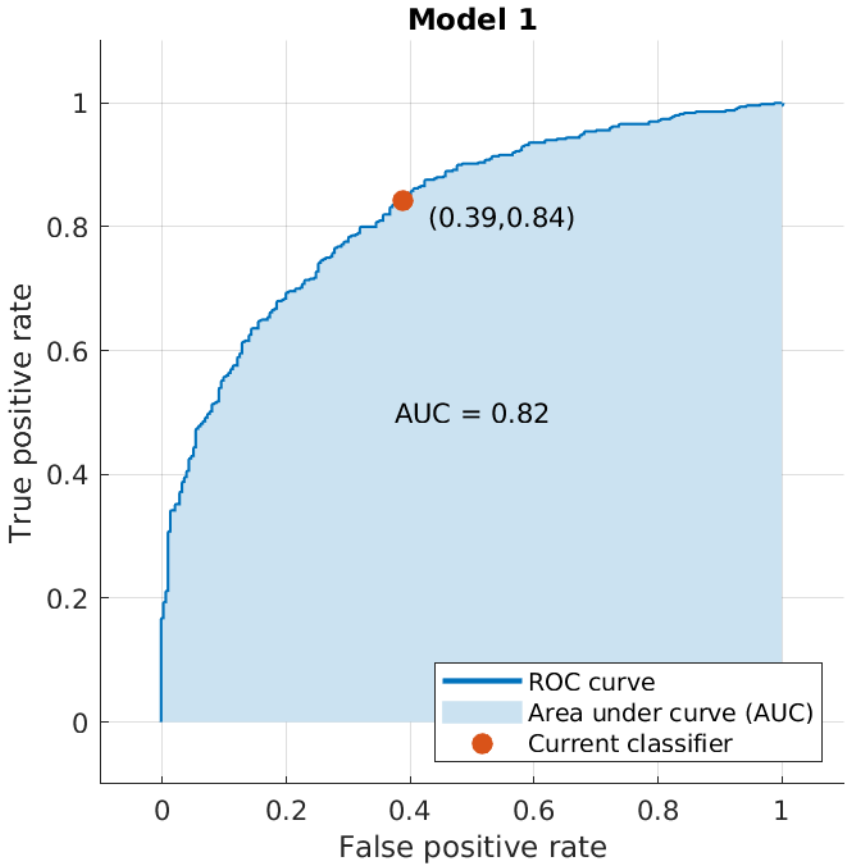
	1 Pregnancies	2 Glucose	3 BloodPressure	4 SkinThickness	5 Insulin
1	17	163	72	41	114
2	15	136	70	32	110
3	14	100	78	25	184
4	14	175	62	30	0
5	13	145	82	19	110
6	13	126	90	0	0
7	13	106	72	54	0

	1 Pregnancies	2 Glucose	3 BloodPressure	4 SkinThickness	5 Insulin
1	1	0.6382	0.1803	-0.1717	-0.7305
2	0.7647	0.3668	0.1475	-0.3535	-0.7400
3	0.6471	0.0050	0.2787	-0.4949	-0.5650
4	0.6471	0.7588	0.0164	-0.3939	-1
5	0.5294	0.4573	0.3443	-0.6162	-0.7400
6	0.5294	0.0653	0.1475	-1	-1
7	0.5294	0.0653	0.1803	0.0909	-1

Ερώτημα 2:

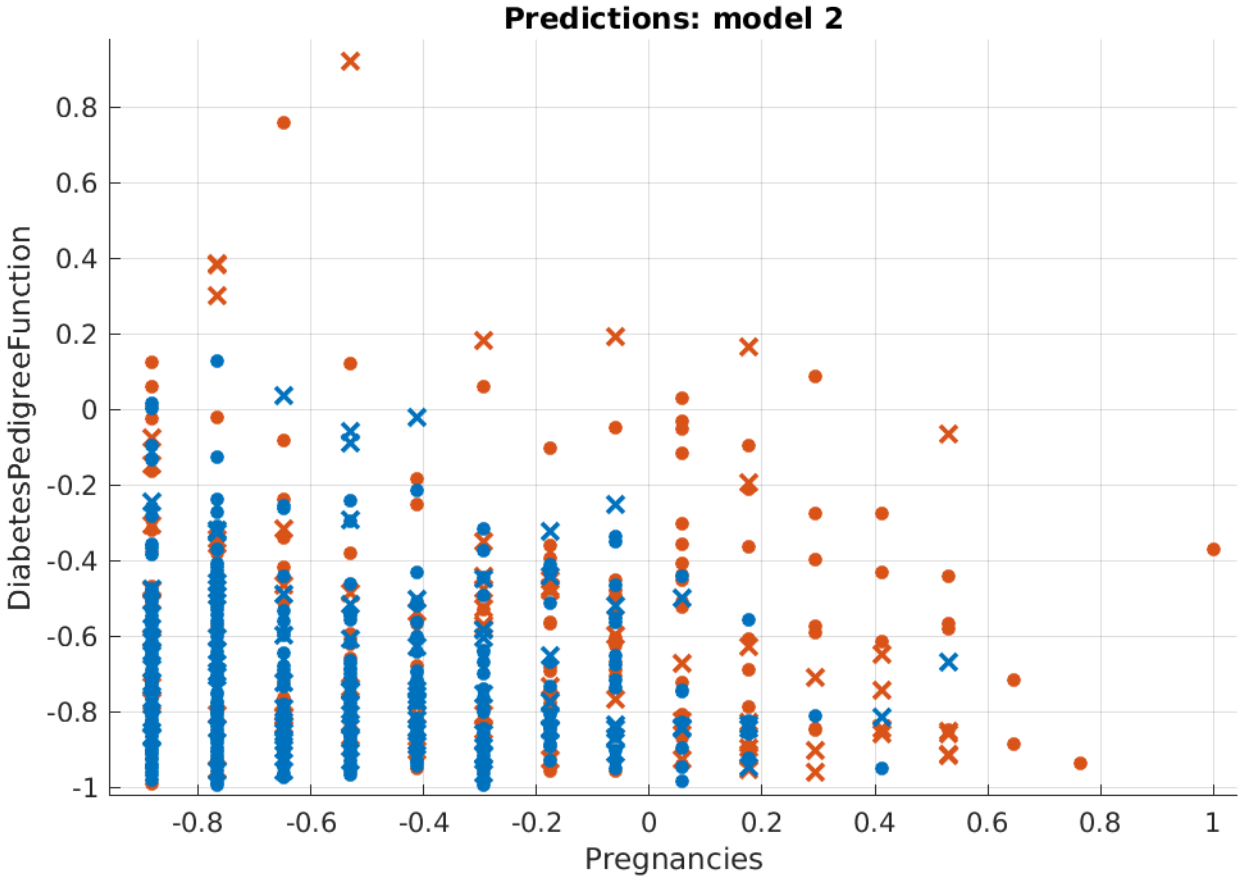
Ο αφελής ταξινομητής Bayes είναι βασισμένος στα θεωρήματα του Thomas Bayes για την εκ των υστέρων πιθανότητα υποθέτοντας (αφέλεια) ότι τα δείγματα του πειράματος είναι ανεξάρτητα μεταξύ τους. Ο ταξινομητής αυτός χρησιμοποιείται εκτενώς στη μηχανική μάθηση λόγω της ευκολίας, της ταχύτητας και του μικρου αριθμού δειγμάτων εκπαίδευσης που χρειάζεται. Το παρακάτω παράδειγμα βασίζεται στην κανονικοποιημένη μορφή του πίνακα diabetes(N) χωρίς validation με βάση το αποτέλεσμα ταξινομημένο κατά τους άξονες συνάρτησης κληρονομικότητας και αριθμού εγκυμοσυνών. Ο παρακάτω πίνακας έχει 72,7 % επιτυχία πρόβλεψης σύμφωνα με τη Matlab.

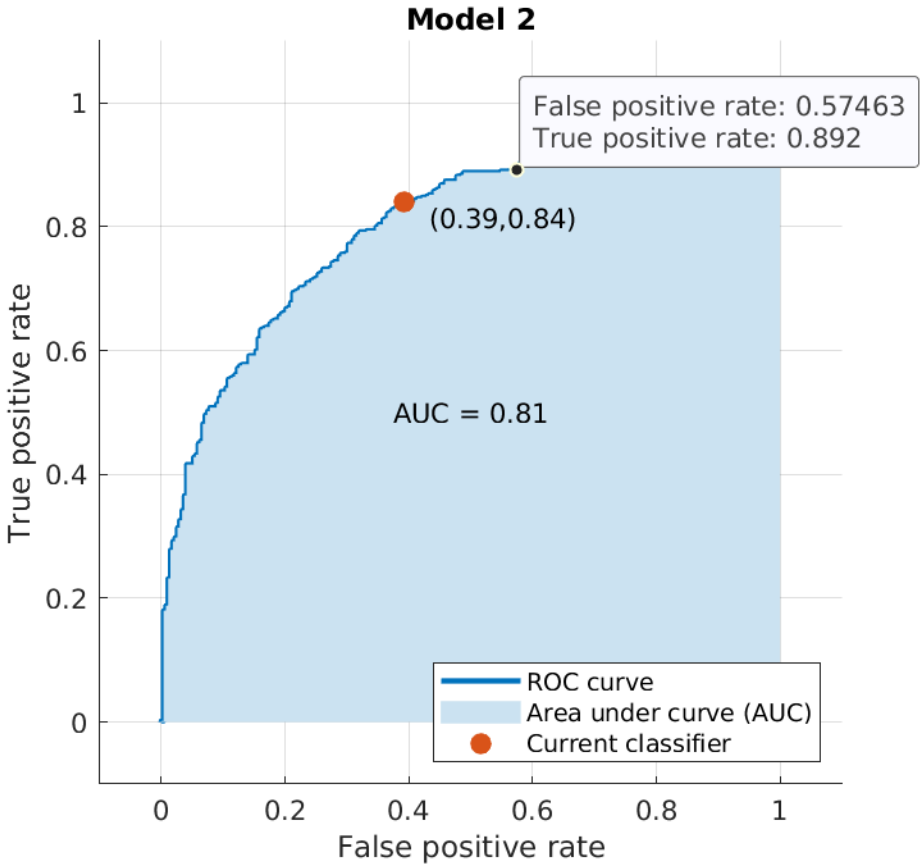




Ερώτημα 3:

Το 5-fold cross-validation ενός ταξινομητή Bayes προστατεύει το πείραμα από την υπερφόρτωση διαμερίζοντας τα δεδομένα σε 5 πτυχές (folds) και υπολογίζοντας την επιτυχία του πειράματος σε κάθε fold ξεχωριστά. Παραθέτοντας τα κανονικοποιημένα δεδομένα του αρχείου diabetes.csv στη Matlab έχουμε τα εξής αποτελέσματα:

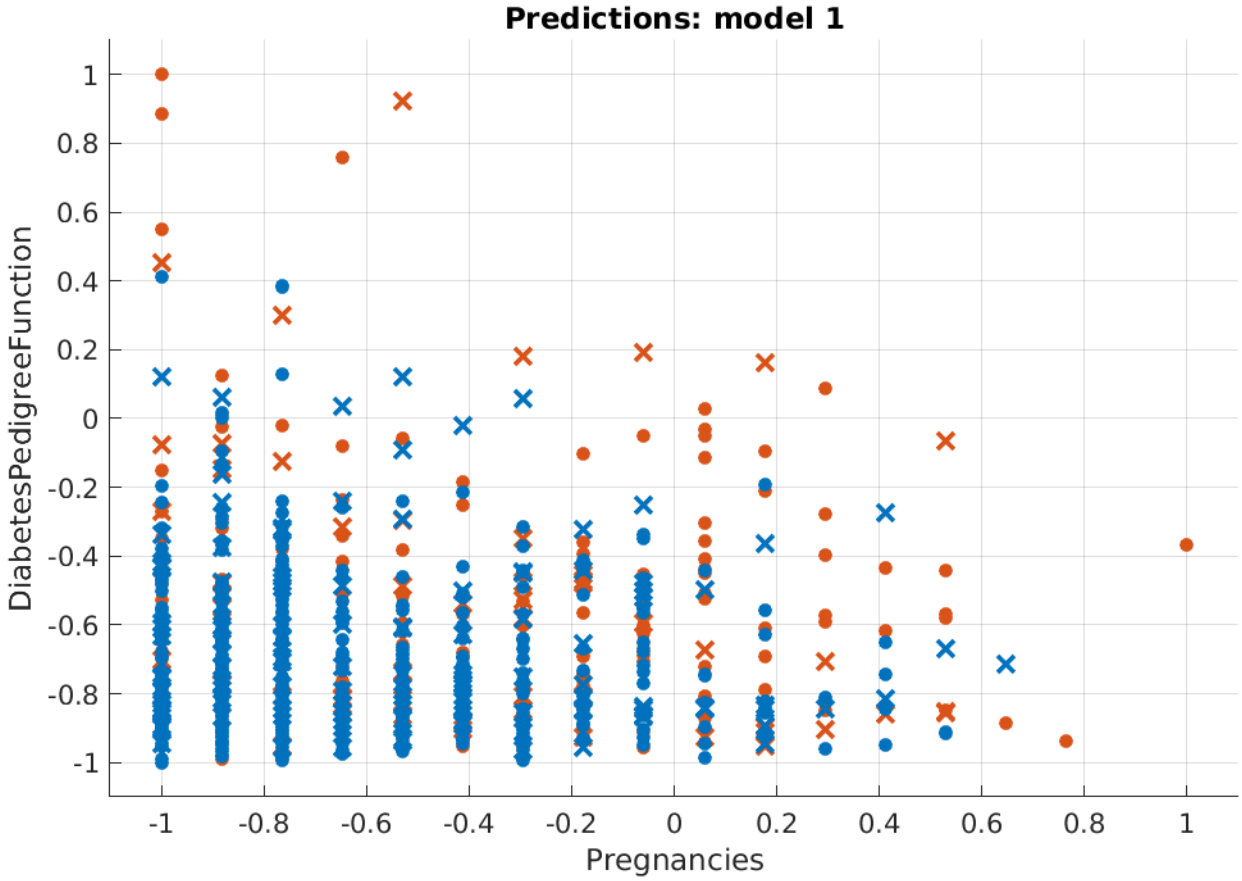

Predictions: model 2
DiabetesPedigreeFunction
Pregnancies
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

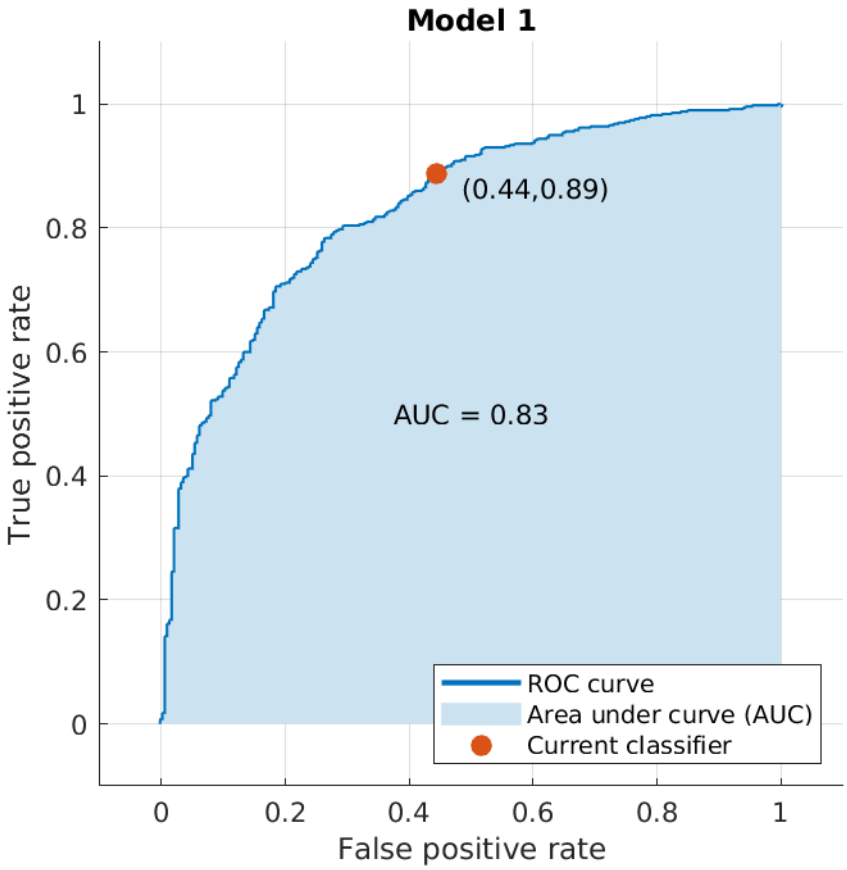


Οι πίνακες αυτοί έχουν ποσοστό επιτυχίας 75,9% , σαφώς υψηλότερο από τον προηγούμενο ταξινομητή.

Ερώτημα 4.1:

Μεταβάλουμε την τιμή της C σύμφωνα με την άσκηση και συνεχίζουμε συγκρίνοντας τους γεωμετρικούς μέσους.Το C στα γραμμικά SVM μεταβάλλει τον βαθμό που θέλουμε να αποφύγουμε τη λάθος πρόβλεψη. Ειδικότερα, ένα μεγάλο C θα κάνει τον ταξινομητή μας να πάρει σαν γραμμή απόφασης ένα μικρότερο υπερεπίπεδο, αν αυτό διαχωρίζει καλύτερα τα στοιχεία μας. Αντίθετα,με μικρές τιμές του C, ο ταξινομητής θα ψάξει για κάποιο πολύ μεγάλο υπερεπίπεδο, ακόμα και αν αυτό βάζει πολλά στοιχεία σε λάθος κλάση. Όπως φαίνεται στο συνημμένο αρχείο κώδικα SVM.m, οι ταξινομήσεις προκύπτουν ως εξής:

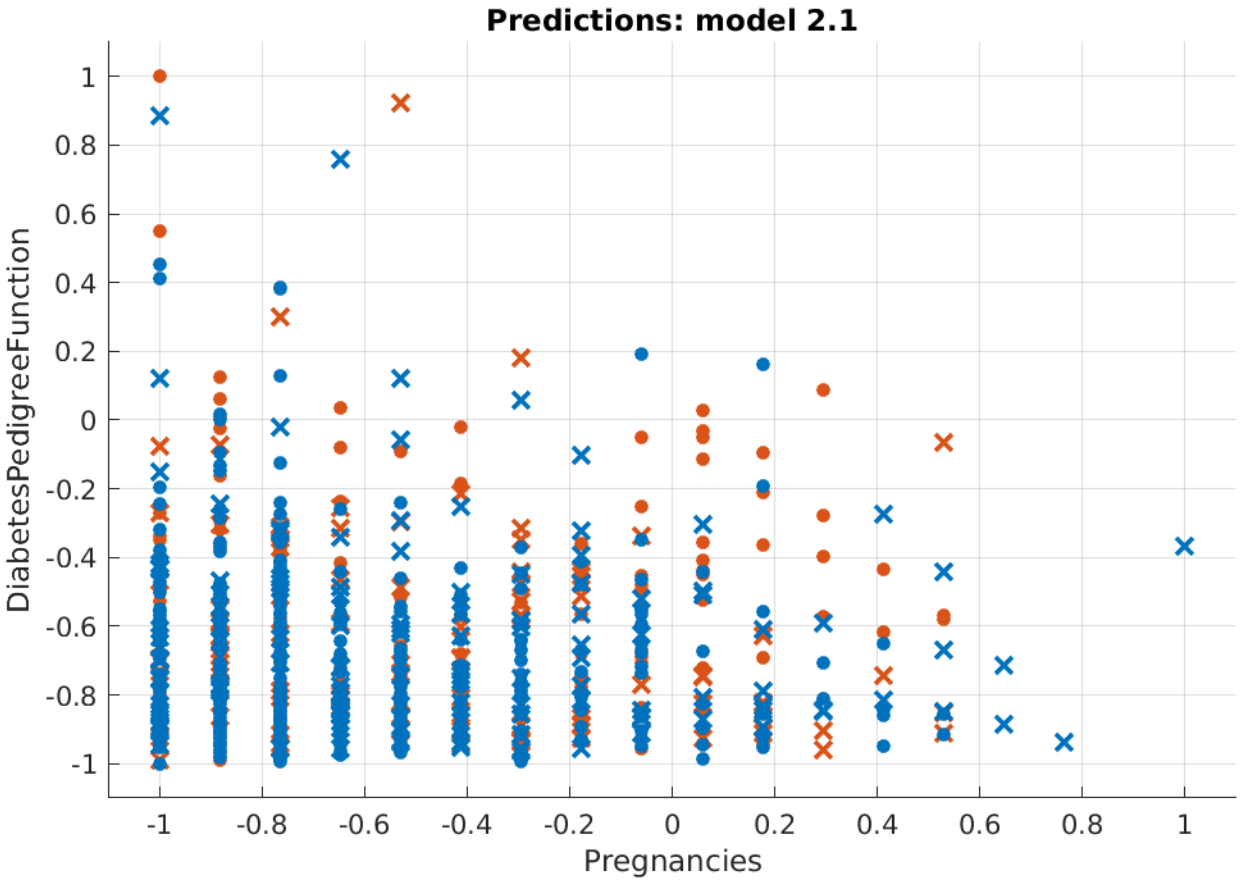

Predictions: model 1
DiabetesPedigreeFunction
Pregnancies
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

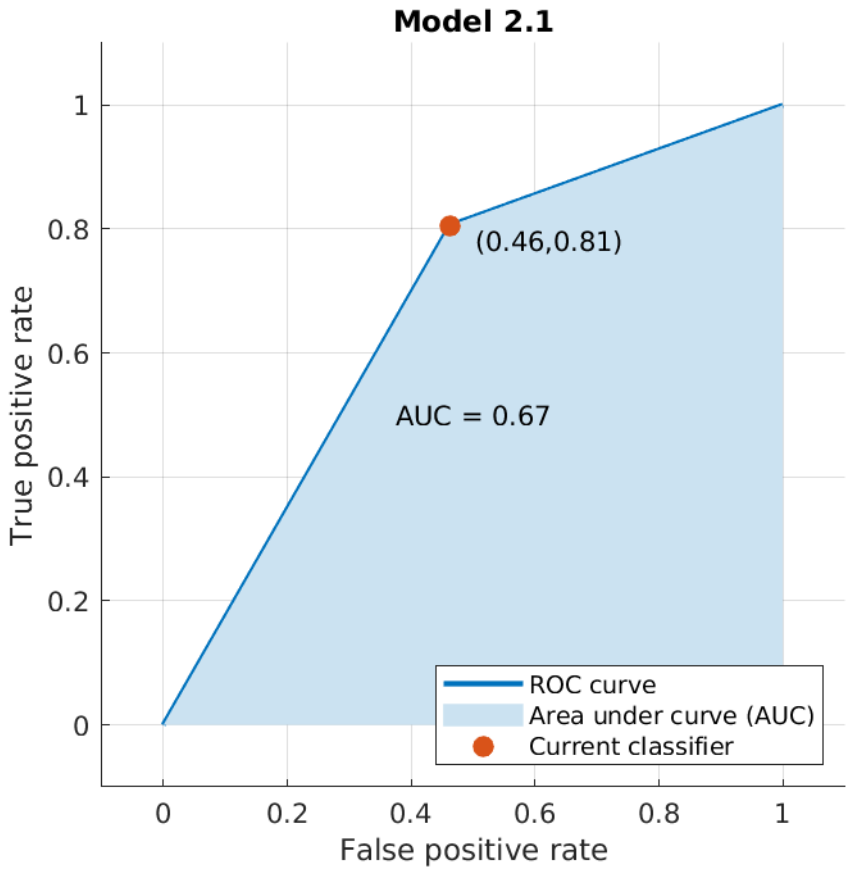


Το ποσοστό επιτυχίας του παραπάνω αλγορίθμου είναι 77,2%.

Ερώτημα 4.2:

Ο αλγόριθμος ταξινόμησης Κ-Κοντινότερου Γείτονα (KNearestNeighbor/KNN) εφαρμόζει την παρακάτω λογική για την ταξινόμηση των δειγμάτων. Όταν εκπαιδεύεται με κάποιο στοιχείο βρίσκει τα Κ κοντινότερα στοιχεία και θέτει ως στοιχείο εκπαίδευσης την κοινή κλάση ανάμεσα στα στοιχεία αυτά.
Το Κ στο KNN ουσιαστικά ρυθμίζει τους γείτονες τους οποίους ο ταξινομητής χωρίζει τα στοιχεία .
Τα αποτελέσματα για τις διάφορες τιμές του KNN τα βρίσκουμε εκτελώντας το αρχείο KNN.m και είναι τα εξής :





με ποσοστό επιτυχίας 71,2%.

Ανακεφαλαιώνοντας τα ποσοστά ακρίβειας ανά ταξινομητή παρατίθενται στον παρακάτω πίνακα:

Naive Bayes	5-fold cross-validation	Support Vector Machines	KNearestNeighbor
72,7 %	75,9%	77,2	71,2

Συνεπώς, κρίνεται καλύτερο με τα δοθέντα στοιχεία να χρησιμοποιήσουμε τον ταξινομητή

Support Vector Machines με Radial Basis Function kernel function