Classifiers

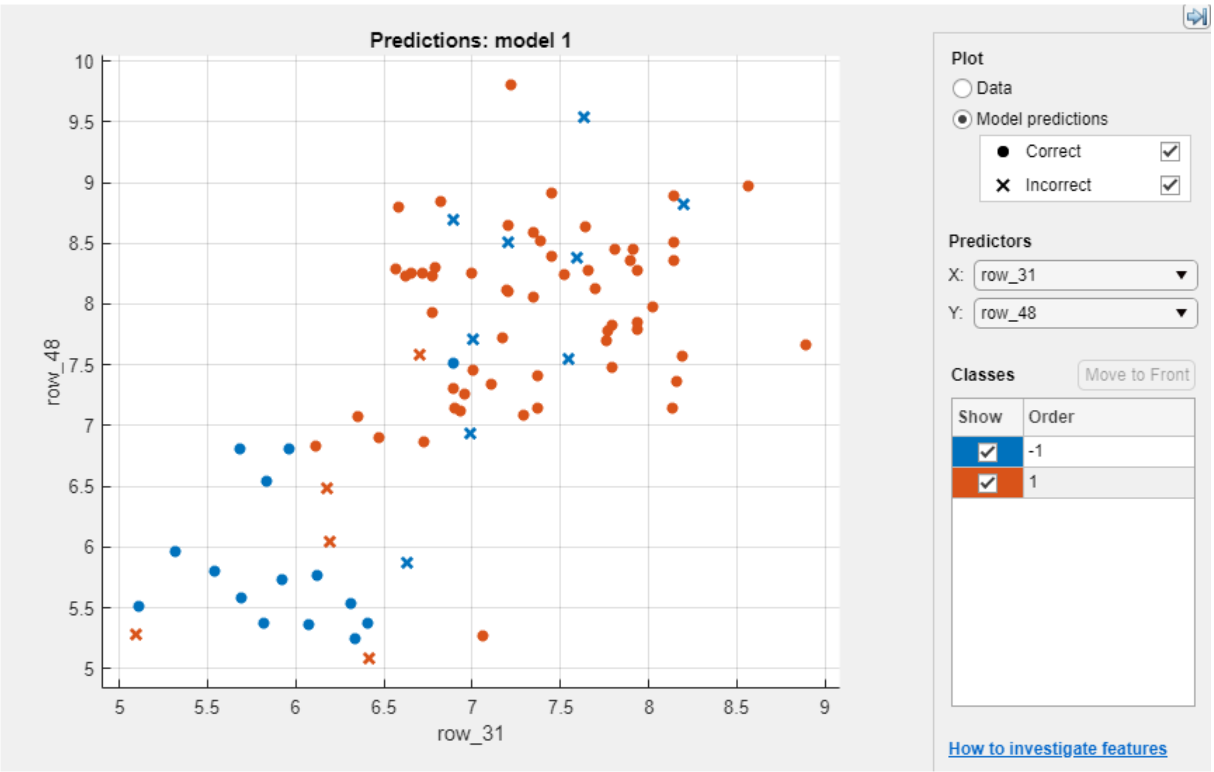
Report 1

Classification Quality Estamination

Author:
Piotr Pawełko
Piotr Wojsa

Laboratory date:
22.04.2022

1. Introduction



Plot

○ Data

● Model predictions

● Correct ☑

✕ Incorrect ☑

Predictors

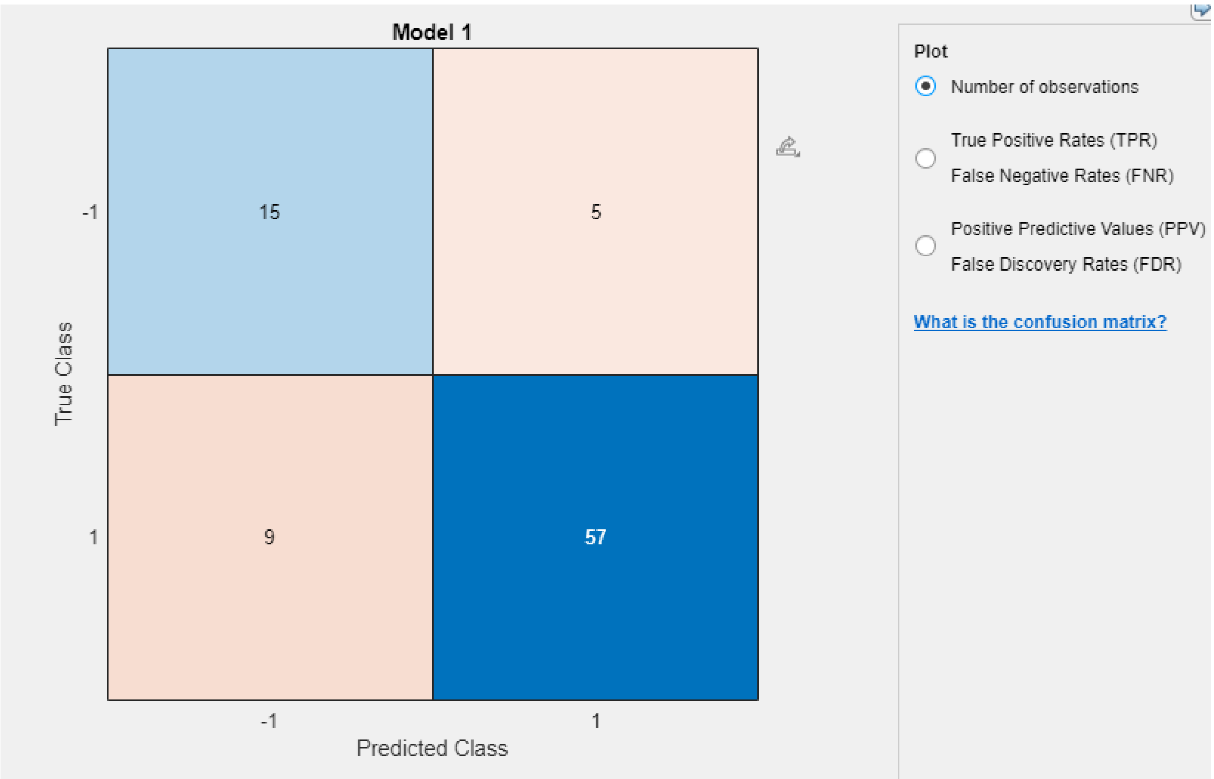
X: row_31 ▼

Y: row_48 ▼

Classes Move to Front

Show	Order
☑	-1
☑	1

How to investigate features



True Class \ Predicted Class	-1	1
-1	15	5
1	9	57

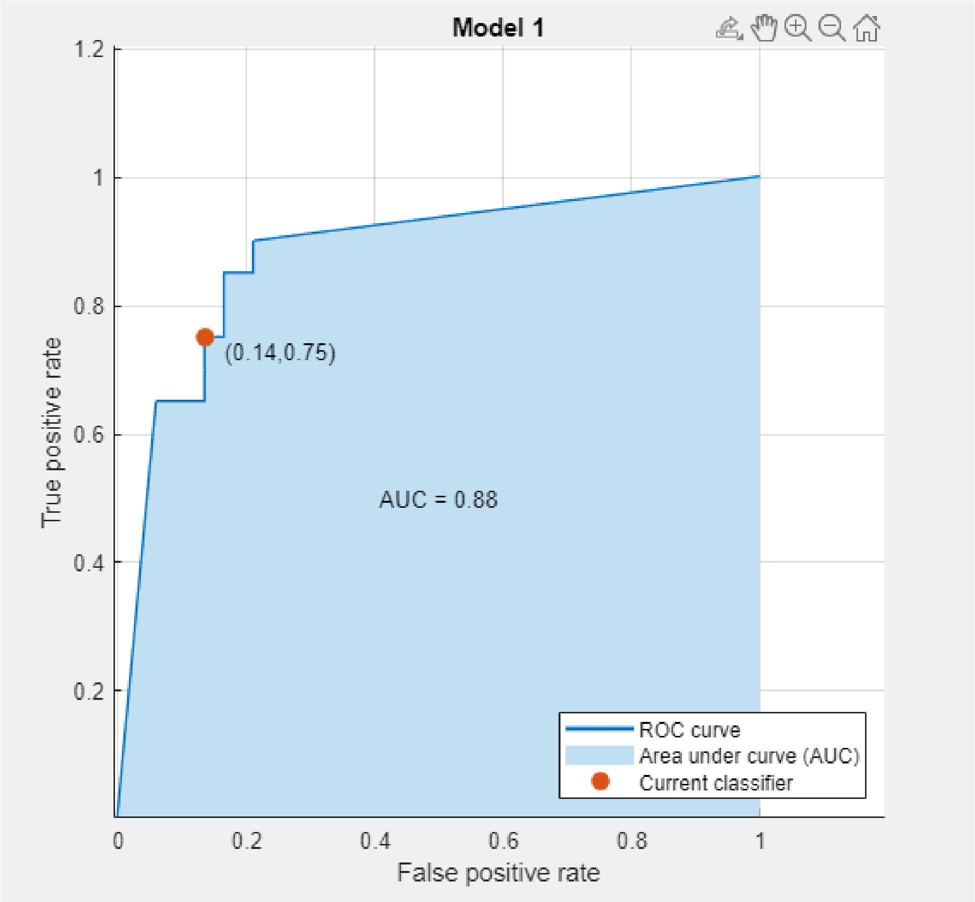
Plot

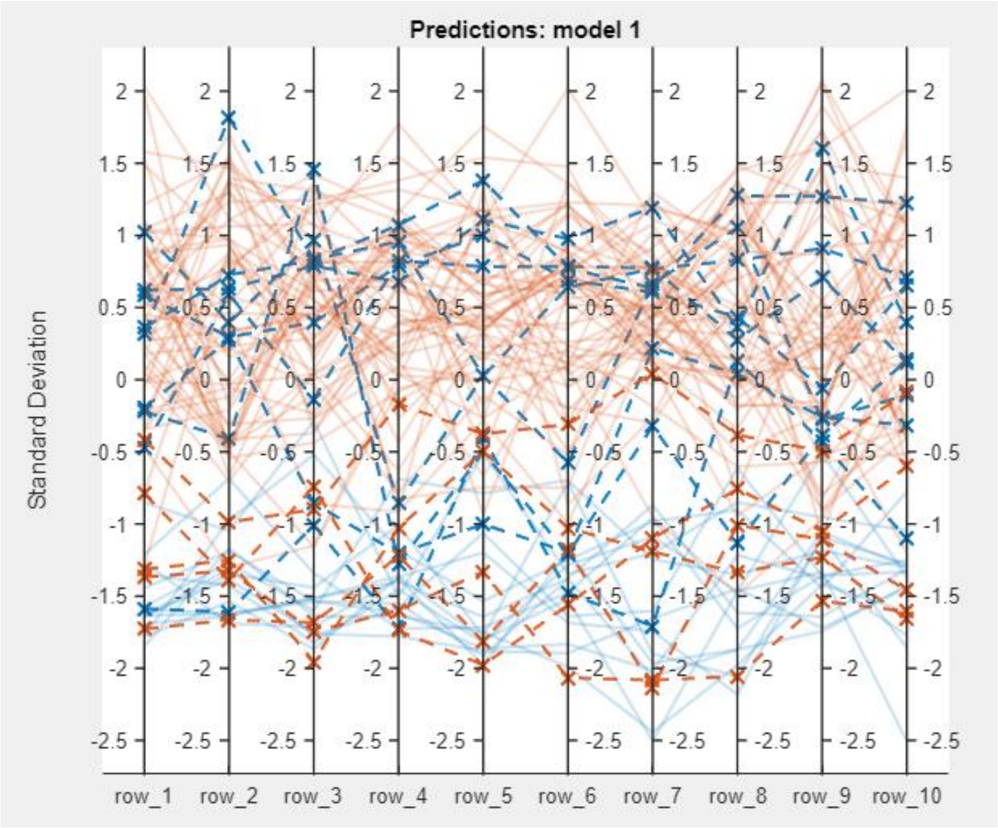
● Number of observations

○ True Positive Rates (TPR)
False Negative Rates (FNR)

○ Positive Predictive Values (PPV)
False Discovery Rates (FDR)

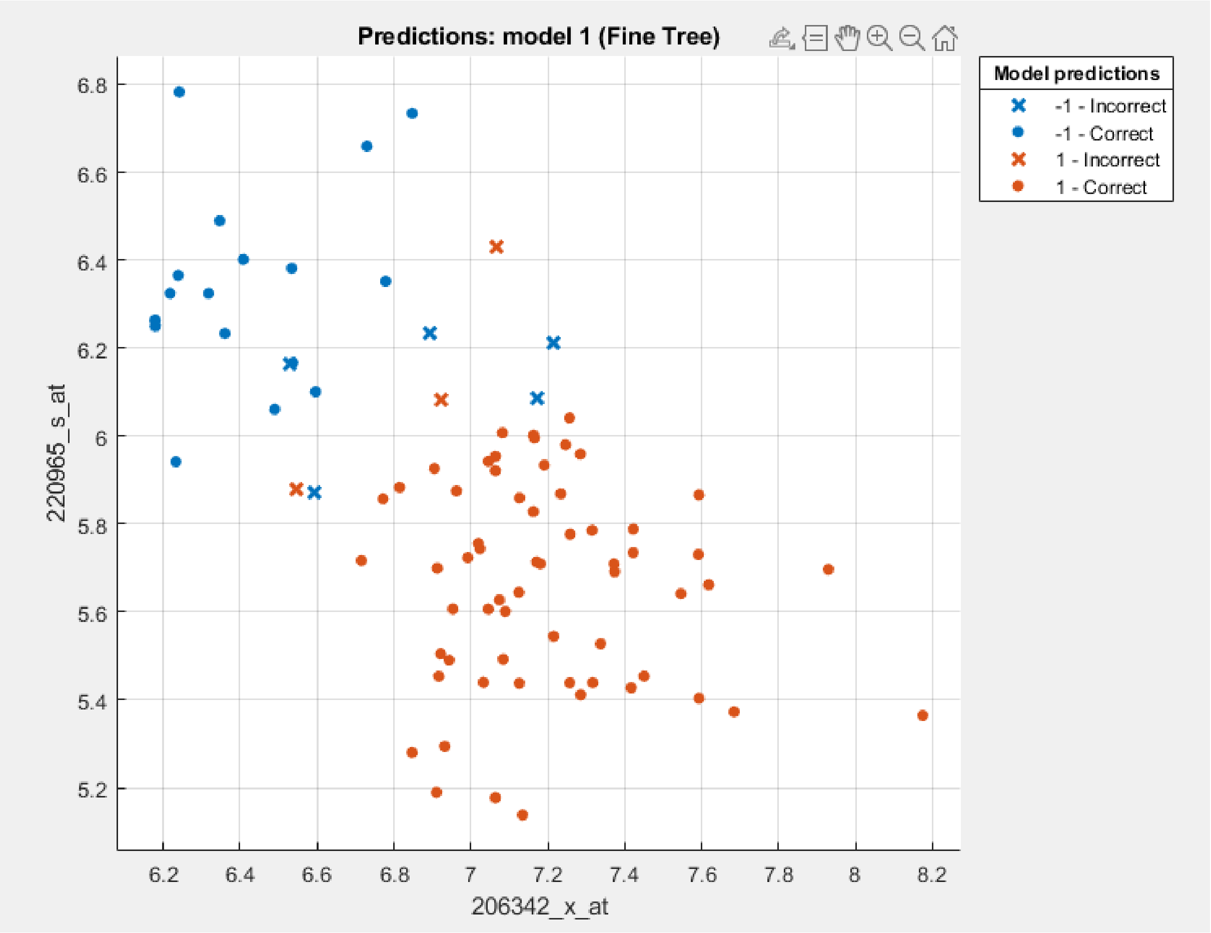
What is the confusion matrix?


Model 1
1.2
1
0.8
0.6
0.4
0.2
True positive rate
(0.14,0.75)
AUC = 0.88
ROC curve
Area under curve (AUC)
Current classifier
0
0.2
0.4
0.6
0.8
1
False positive rate


Predictions: model 1
Standard Deviation
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
row_1
row_2
row_3
row_4
row_5
row_6
row_7
row_8
row_9
row_10

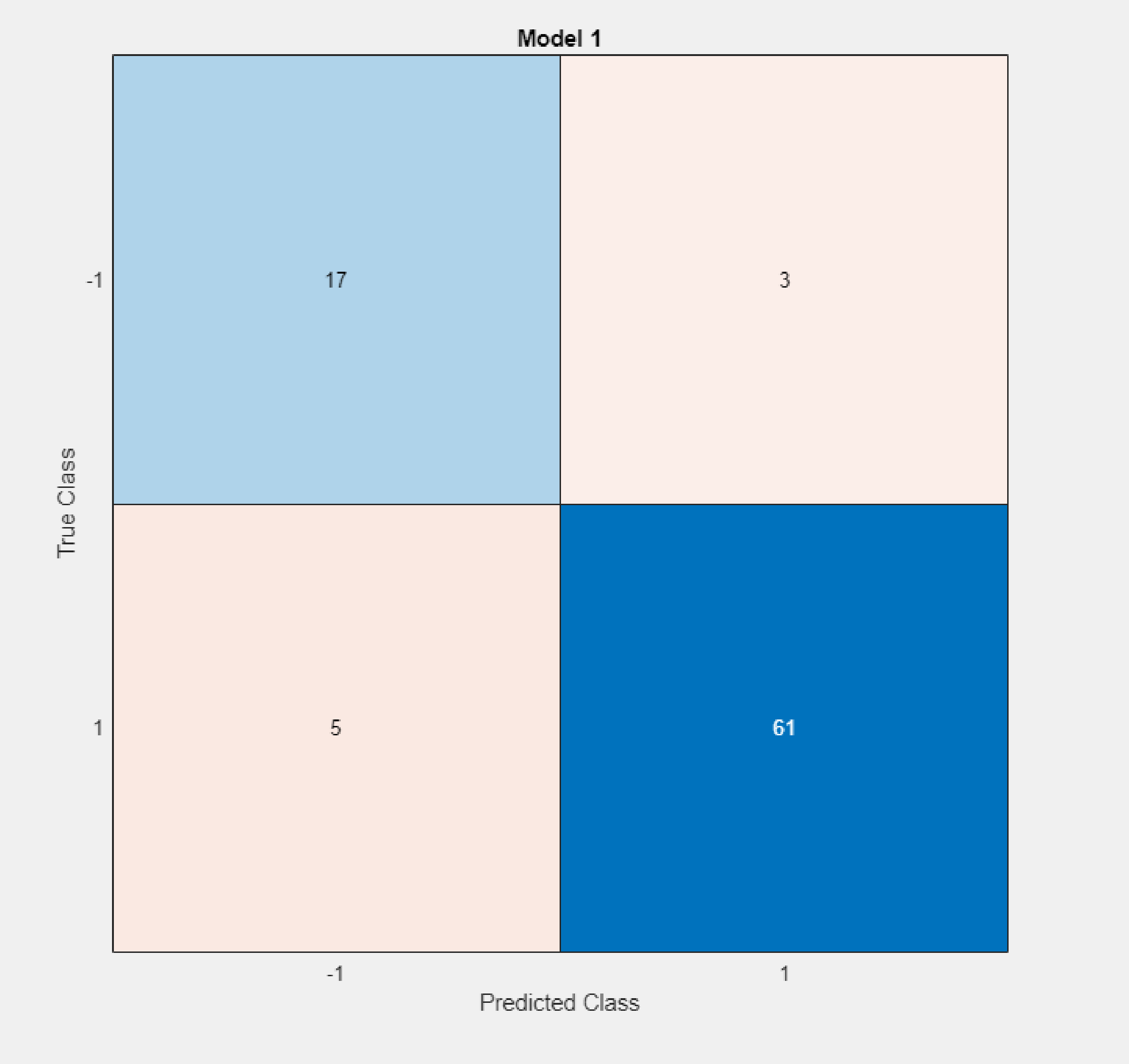
2. Task 1

I. Scatterplot



Data is not linear seperable.

II. Confusion matrix



TP – bottom right

TN – top left

FN – top right

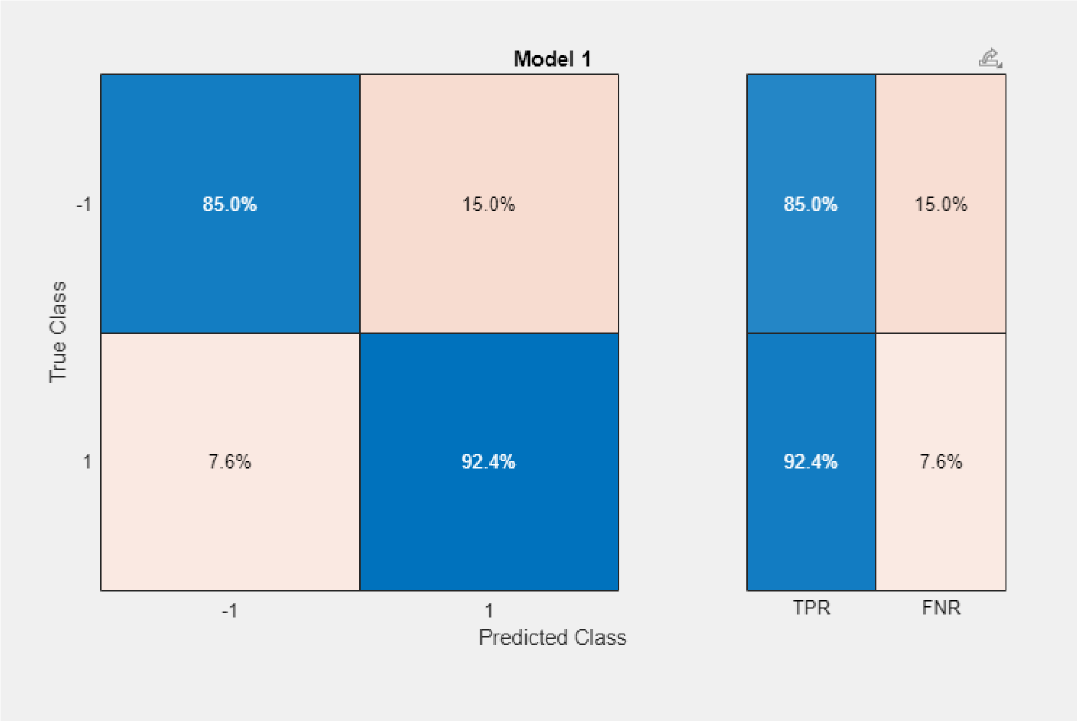
FP – bottom left

Accuracy	Error	Sensitivity	Specificity
0.907	0.093	0.9531	0.7727

III. TPR, FNR, PPV, FDR

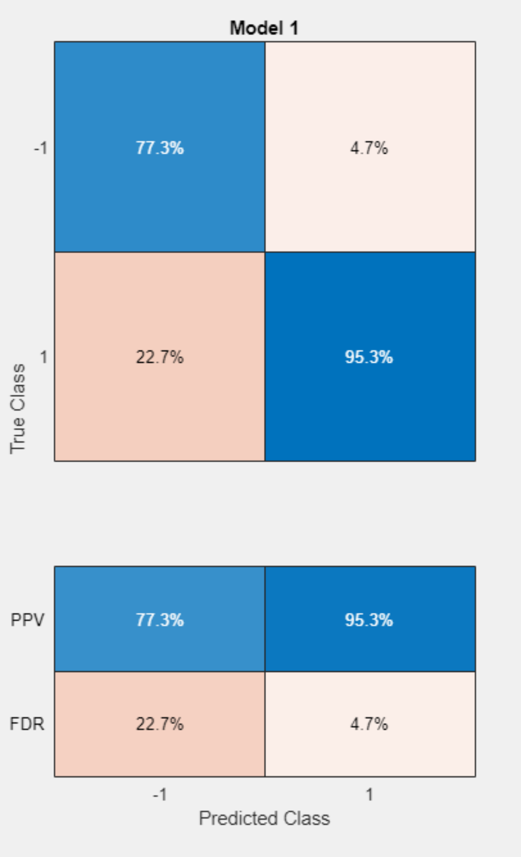
TPR(true positive rates) = TP/positive

FNR(false negative rates) = FN/negtive



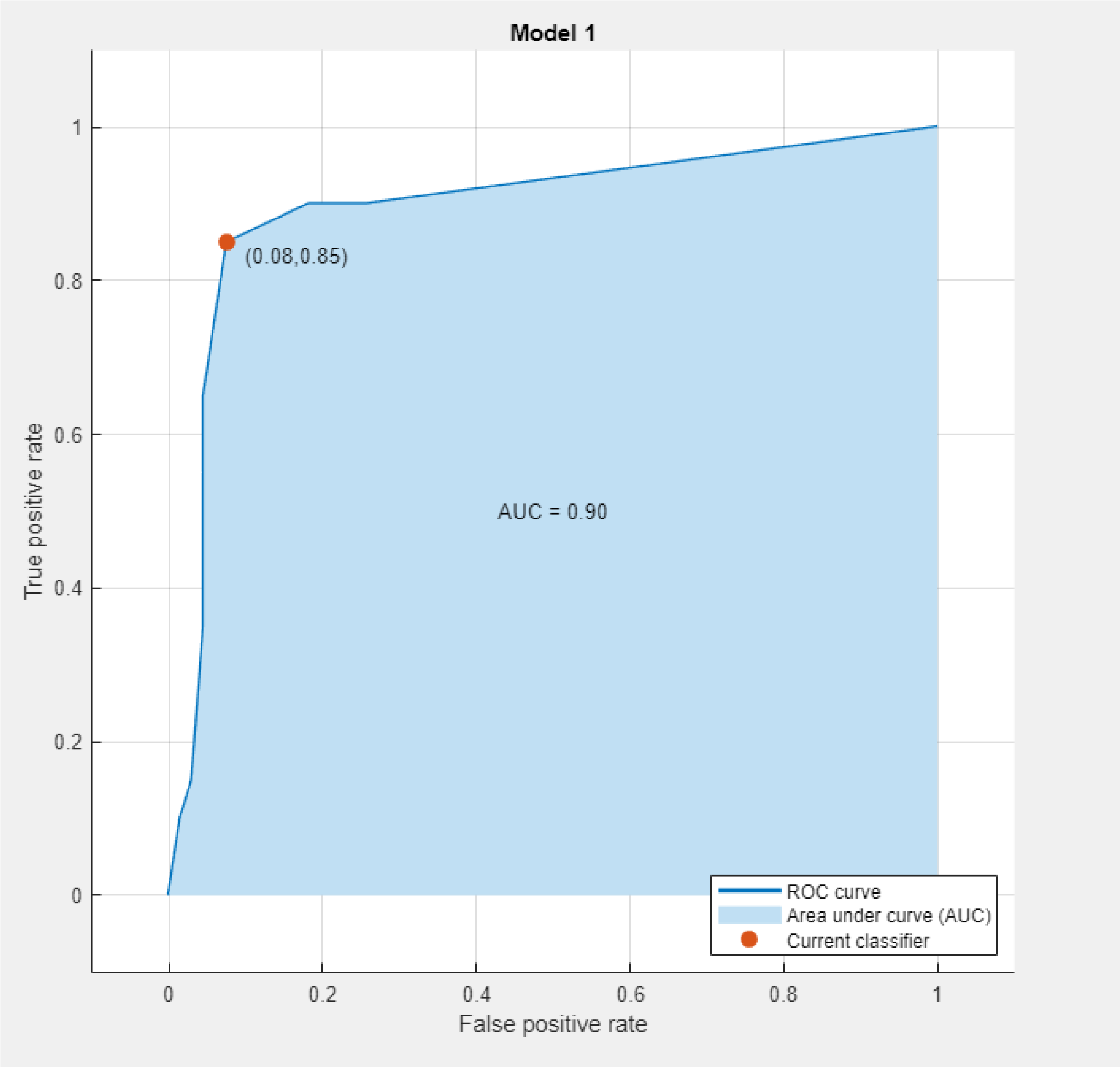
PPV(positive predictive value) = TP/(TN+FP)

FDR(false discovery rate) = FP/(FP+TP)



IV. ROC Curve

AUC = 0.9



AUC(Area under the curve) **ROC**(Recevier operating characheristic) – one of the most important evalutaion metrics for checking any classification model's performance. AUC - ROC curve is a performance measurment for the classification problems at various theshold settings. ROC is a probability curve and AUC represents the degree or measure of separibilty. It tells how much the model is capable of distingusishing betwee classes. [1] If AUC value is high there is better chance of right classification. ROC curve is plotted with TPR(true positive ratio) against the FPR(false positive ratio), TPR on y-axis and FPR on x-axis. The worst possible AUC value is 0.5 and it means that there is no discrimination, for 0.7 and higher value classification is acceptable.

[1] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

3. Task 2

Model is using 68th and 70th gene and Fine Tree training method.

	Accuracy	Error	Specificity	Sensitivity	AUC
5 Fold cross	0.9186	0.0814	0.8824	0.9275	0.91
10 Fold cross	0.9302	0.0698	0.85	0.9545	0.91
25% Holdout	1	0	1	1	1
50% Holdout	0.9535	0.0465	0.8333	1	0.96
75% Holdout	-	-	-	-	-
No validation	0.9767	0.0233	0.9091	1	0.99

For used data we got the best results for 25% Holdout Validation, the values here are the best possible, second best is no validation.

Overfitting - modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points. As a result, the model is useful in reference only to its initial data set, and not to any other data sets. Ways to prevent overfitting include cross-validation, in which the data being used for training the model is chopped into folds or partitions and the model is run for each fold. Then, the overall error estimate is averaged. [2]

4. Task 3

Model is using 68th and 70th gene.

	Accuracy	Error	Specificity	Sensitivity	AUC
SVM linear	0.9651	0.0349	0.9048	0.9846	0.99
SVM quadratic	0.9535	0.0465	0.9	0.9697	0.99
Fine KNN	0.9302	0.0698	0.8889	0.9412	0.88
Medium KNN	0.9651	0.0349	0.9474	0.9701	0.96
Simple Tree	0.9419	0.0581	0.8947	0.9552	0.92
Medium Tree	0.9419	0.0581	0.8947	0.9552	0.92

For this dataset the best result was got by SVM algorithms.

SVM(Support Vector Machine) – In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well

KNN(K-Nearest Neighbor) – tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of ‘K’ training data and class holds the highest probability will be selected. In the case of regression, the value is the mean of the ‘K’ selected training points.

Decision Tree – classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node.