WINE QUALITY PREDICTION USING ORANGE DATA MINING

COURSE NAME :MACHINE LEARNING
COURSE CODE :20CA2023

AGENDA

INTRODUCTION

OBJECTIVE

METHODOLOGY

RESULT

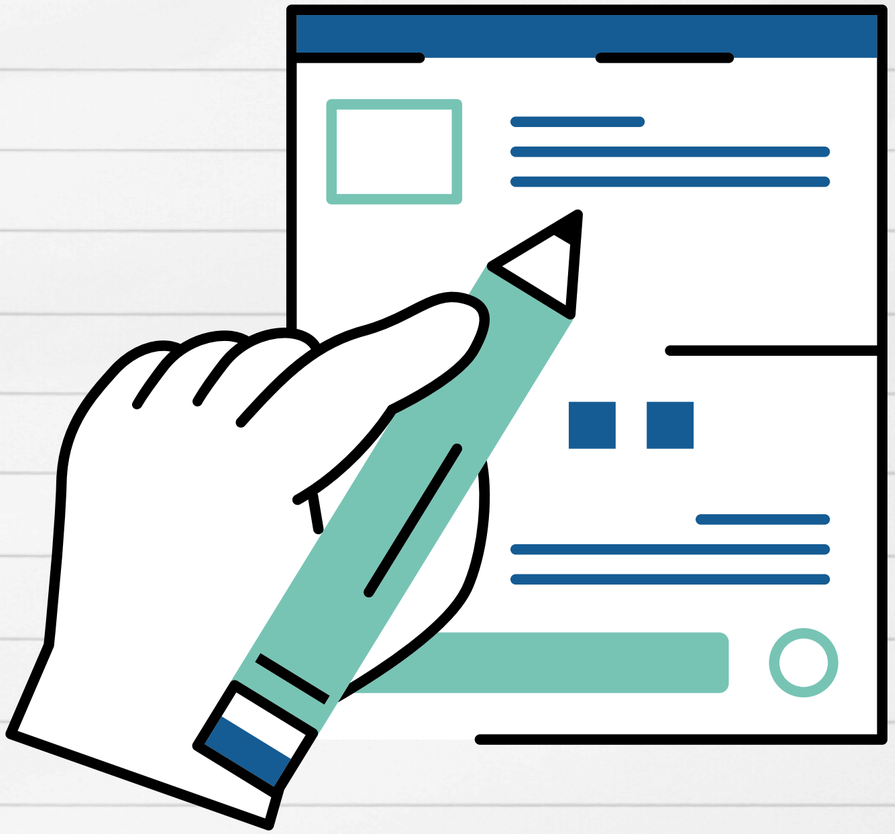
COMPARISON

CONCLUSION

INTRODUCTION:

TEAM MEMBERS

N. VIPUL RAJ -(URK22DS1025)
R. ADHITHAN -(URK22DS1011)
E. DIVITH -(URK22DS1012)



PROJECT TITLE: WINE QUALITY PREDICTION USING ORANGE DATA MINING

DATASET: THE PROJECT UTILIZES THE RED WINE QUALITY DATASET FROM THE UCI MACHINE LEARNING REPOSITORY, WHICH INCLUDES PHYSICOCHEMICAL PROPERTIES OF RED WINES, SUCH AS ACIDITY, SUGAR CONTENT, PH, AND ALCOHOL CONCENTRATION.

IMPORTANCE: PREDICTING WINE QUALITY IS CRITICAL FOR WINEMAKERS AND DISTRIBUTORS TO ENSURE CONSISTENT PRODUCT STANDARDS AND ENHANCE CUSTOMER SATISFACTION. AN ACCURATE PREDICTION CAN LEAD TO BETTER DECISIONS IN PRODUCTION AND MARKETING.

PROJECT TITLE: WINE QUALITY PREDICTION USING ORANGE DATA MINING

TOOL: THIS PROJECT EMPLOYS ORANGE DATA MINING, A VISUAL PROGRAMMING TOOL, TO EXPLORE AND PREDICT THE QUALITY OF RED WINES BASED ON THEIR PHYSICOCHEMICAL ATTRIBUTES. ORANGE'S SIMPLICITY ALLOWS FOR EASY DATA VISUALIZATION, MACHINE LEARNING, AND PREDICTIVE MODELING WITHOUT NEEDING ADVANCED CODING SKILLS.

This project explores how physicochemical properties of red wines, such as acidity, sugar content, and alcohol concentration, influence wine quality. Using Orange Data Mining, we built and tested machine learning models to predict wine quality.

OBJECTIVES

MAIN OBJECTIVE

THE MAIN OBJECTIVE IS TO ACCURATELY PREDICT THE QUALITY OF RED WINES BASED ON THEIR PHYSICOCHEMICAL CHARACTERISTICS.

SECONDARY OBJECTIVES

SECONDARY OBJECTIVES
TO INCLUDE COMPARING THE PERFORMANCE OF VARIOUS MACHINE LEARNING ALGORITHMS (E.G., RANDOM FOREST, GRADIENT BOOSTING, SVM, AND DECISION TREES) IN CLASSIFYING WINE QUALITY.

OBJECTIVES

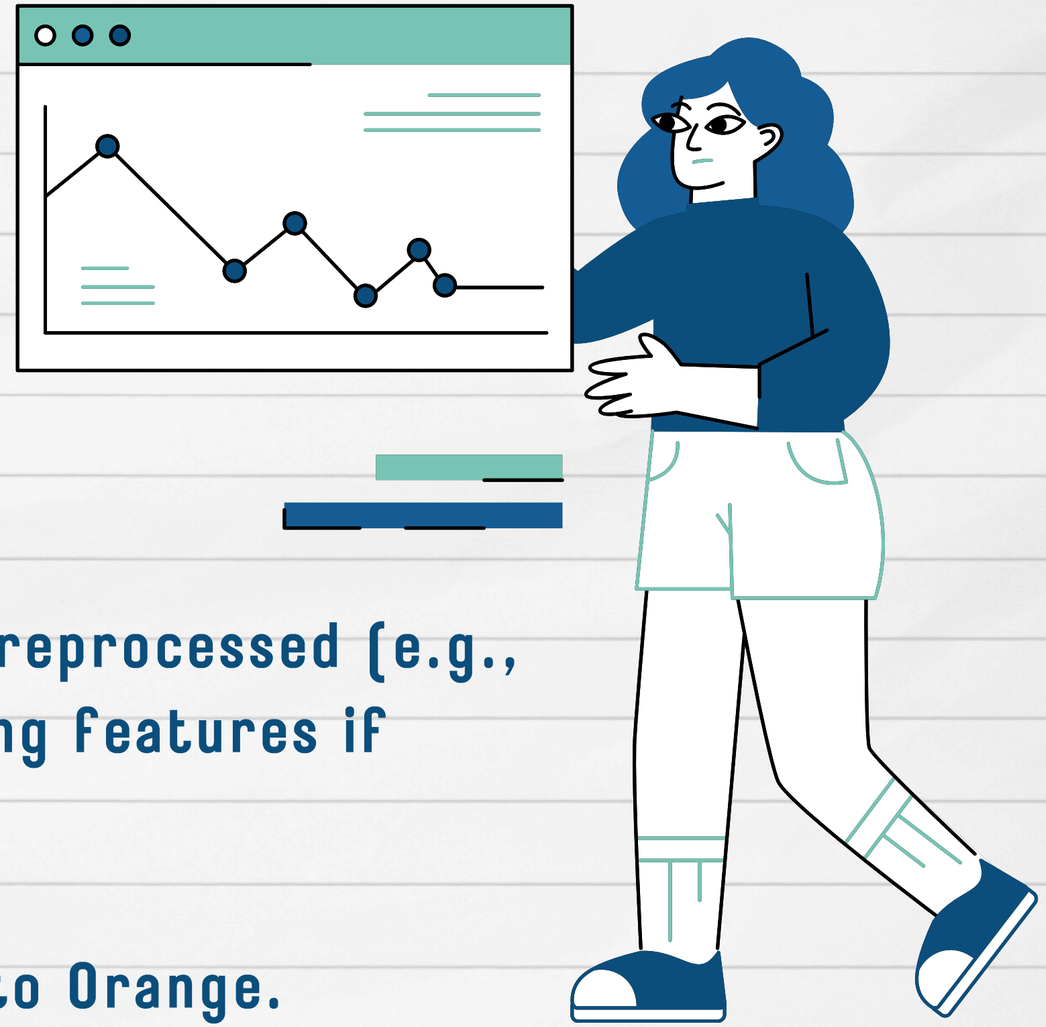
THE PRIMARY OBJECTIVE OF THIS STUDY IS TO PREDICT RED WINE QUALITY BASED ON PHYSICOCHEMICAL PROPERTIES SUCH AS PH, ALCOHOL, AND ACIDITY. WE AIM TO COMPARE DIFFERENT MACHINE LEARNING MODELS TO DETERMINE THE MOST ACCURATE AND INTERPRETABLE ONE.“

EMPHASIZE THE GOAL OF UNDERSTANDING THE IMPACT OF DIFFERENT FEATURES (E.G., ACIDITY, ALCOHOL) ON THE PREDICTION.

METHODOLOGY: PROCEDURES AND TOOLS

- DATA SOURCE: MENTION THE RED WINE QUALITY DATASET WHICH INCLUDES 11 FEATURES AND A QUALITY SCORE RANGING FROM 0 TO 10.
  - FIXED ACIDITY
  - VOLATILE ACIDITY
  - CITRIC ACID
  - RESIDUAL SUGAR
  - CHLORIDES
  - FREE SULFUR DIOXIDE
  - TOTAL SULFUR DIOXIDE
  - DENSITY
  - PH
  - SULPHATES
  - ALCOHOL
  - ACIDITY_RATIO
  - SULFUR_RATIO
  - QUALITY_CATEGORY

METHODOLOGY: PROCEDURES AND TOOLS



Data Preparation: Explain how the data was preprocessed (e.g., handling missing values, scaling, or normalizing features if necessary).

- Orange Workflow:
  - Data Loading: Load the dataset into Orange.
  - Visualization: Use Orange's visual tools to explore data distributions.

METHODOLOGY: PROCEDURES AND TOOLS

- Model Training: Apply machine learning models like Random Forest, Gradient Boosting, SVM, and Decision Trees.

Random Forest:

- Random Forest is an ensemble learning technique used for both classification and regression tasks. It creates multiple decision trees and merges them to get a more accurate and stable prediction.

Decision Trees:

- A Decision Tree is a tree-like structure where each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class label or target value.

METHODOLOGY: PROCEDURES AND TOOLS

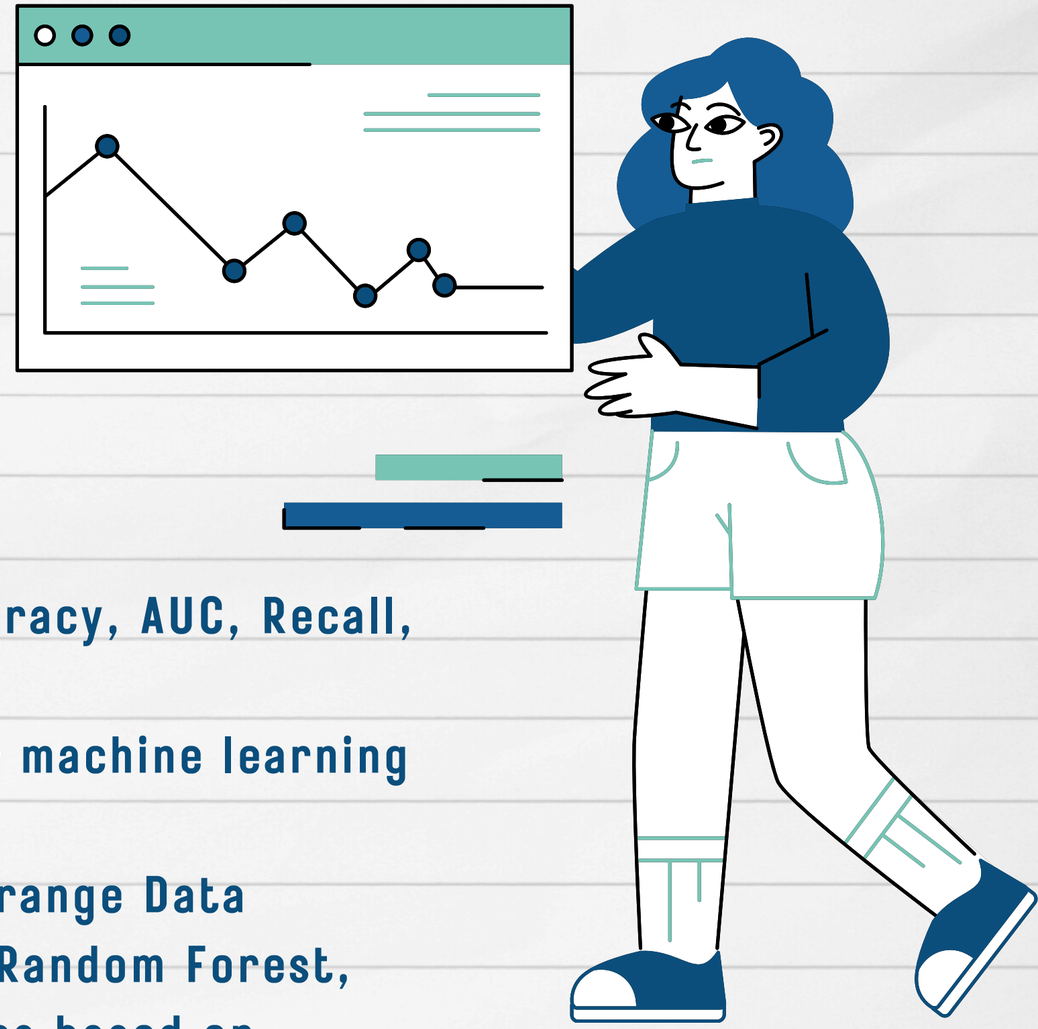
Gradient Boosting:

- Gradient Boosting is an ensemble technique that builds sequential trees, where each new tree corrects the errors made by the previous ones. It works by minimizing a loss function through gradient descent.

Support Vector Machines (SVM) :

- SVM is a supervised learning algorithm used for classification that works by finding the optimal hyperplane that separates different classes with the maximum margin. It is particularly useful in high-dimensional spaces

METHODOLOGY: PROCEDURES AND TOOLS



- Model Evaluation: Evaluate models using metrics like Accuracy, AUC, Recall, and MCC (Matthews Correlation Coefficient).
    - Why Orange?: Highlight the ease of use of Orange for machine learning workflows and rapid prototyping.

Example:“The dataset was preprocessed and analyzed using Orange Data Mining. We applied multiple machine learning algorithms like Random Forest, Gradient Boosting, and SVM, and compared their performances based on accuracy, AUC, and other evaluation metrics.“

RESULTS[CROSS VALIDATION] -1ST REVIEW

Gradient Boosting

AUC = 0.823,
Recall = 0.673,
MCC = 0.455

Random Forest

AUC = 0.838,
Recall = 0.710,
MCC = 0.511.

SVM

AUC = 0.693,
Recall = 0.529,
MCC = 0.231

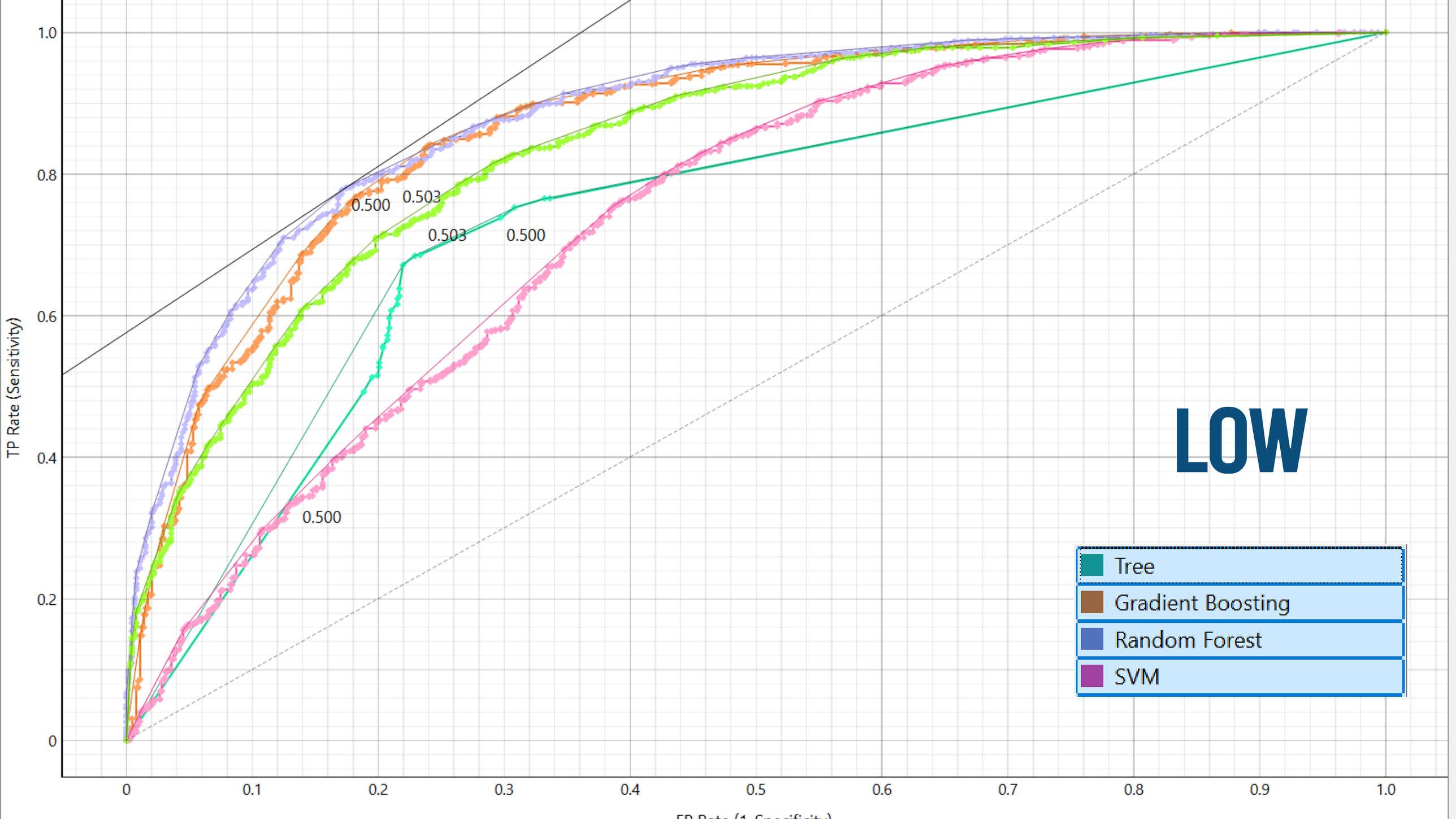
Decision Tree

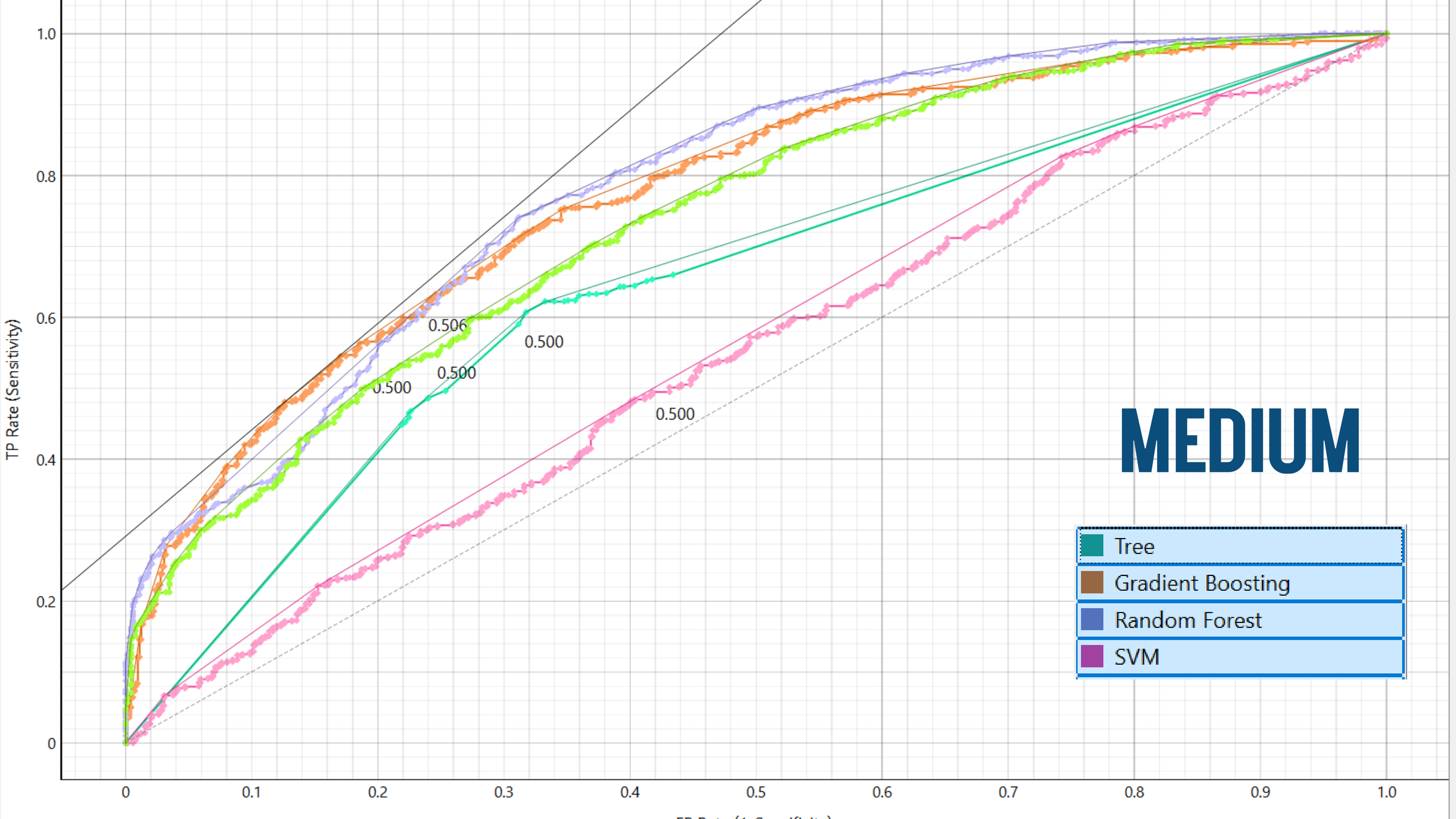
AUC = 0.681,
Recall = 0.587,
MCC = 0.324

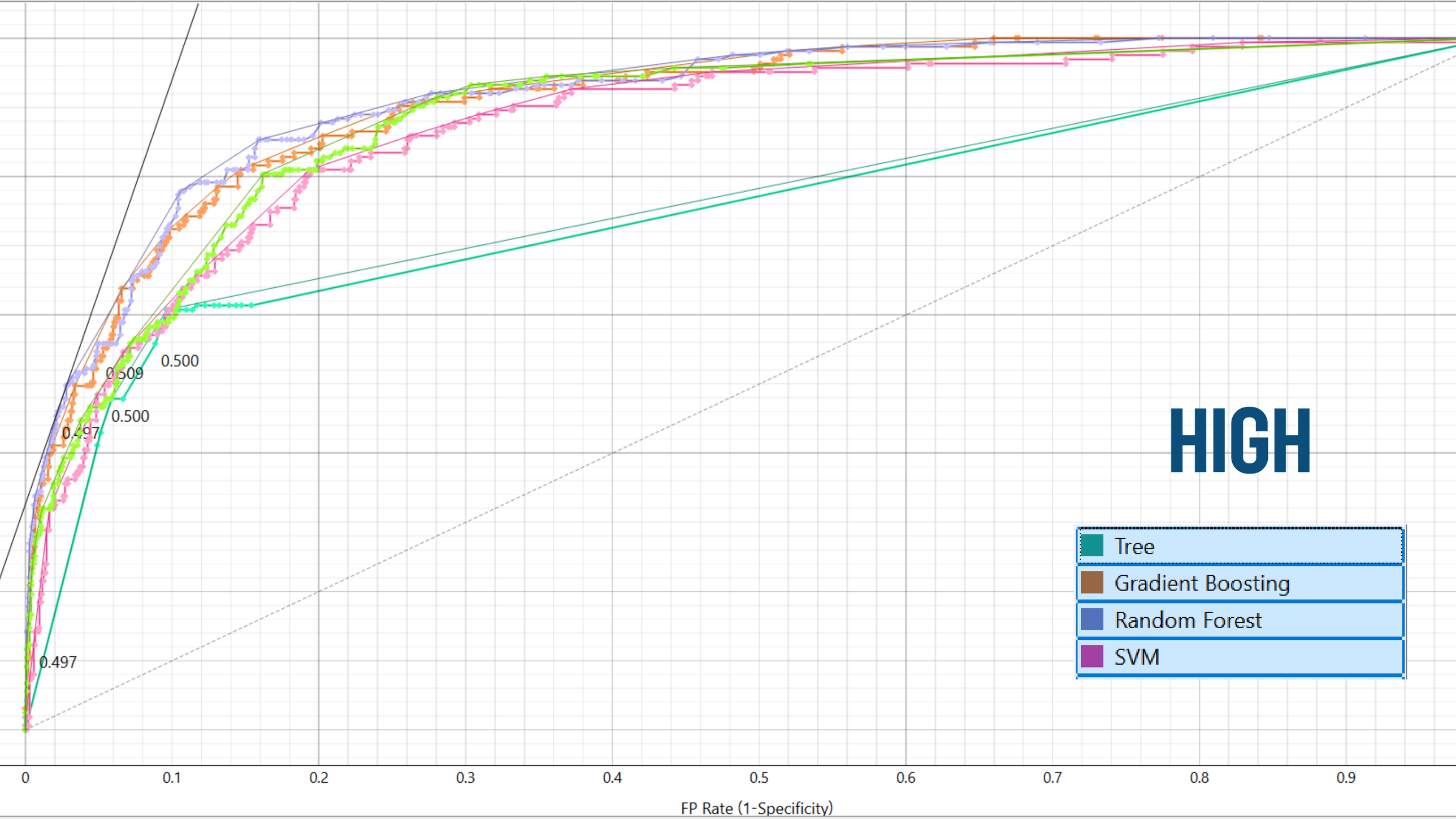
THE RANDOM FOREST MODEL ACHIEVED THE HIGHEST PERFORMANCE WITH AN AUC OF 0.838 AND AN MCC OF 0.511, DEMONSTRATING ITS ABILITY TO PREDICT WINE QUALITY MORE EFFECTIVELY THAN OTHER MODELS

RESULTS[CROSS VALIDATION]

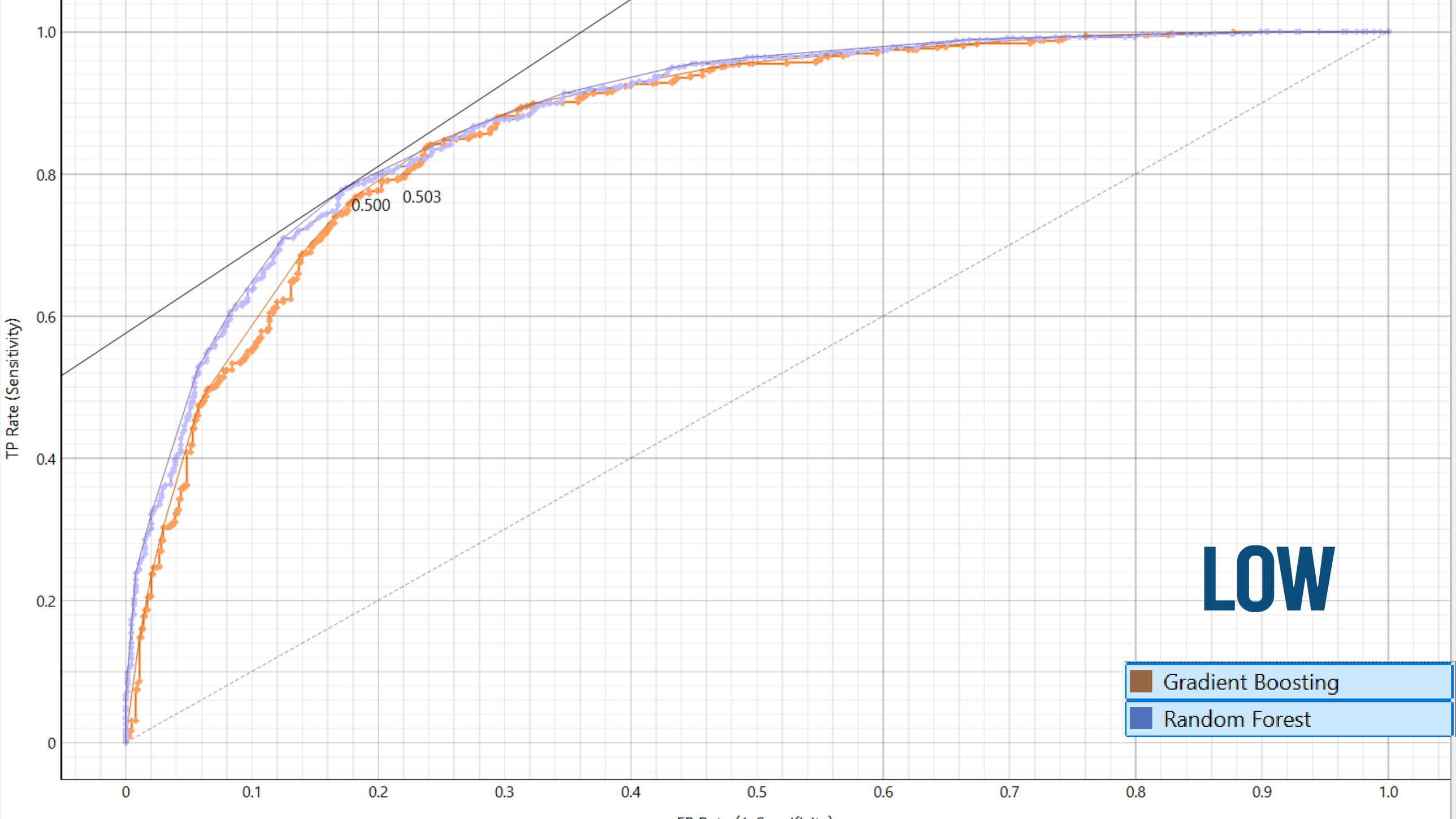
Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.702	0.615	0.612	0.613	0.615	0.369
Gradient Boosting	0.836	0.693	0.691	0.689	0.693	0.489
Random Forest	0.847	0.704	0.701	0.704	0.704	0.502
SVM	0.686	0.495	0.496	0.519	0.495	0.184

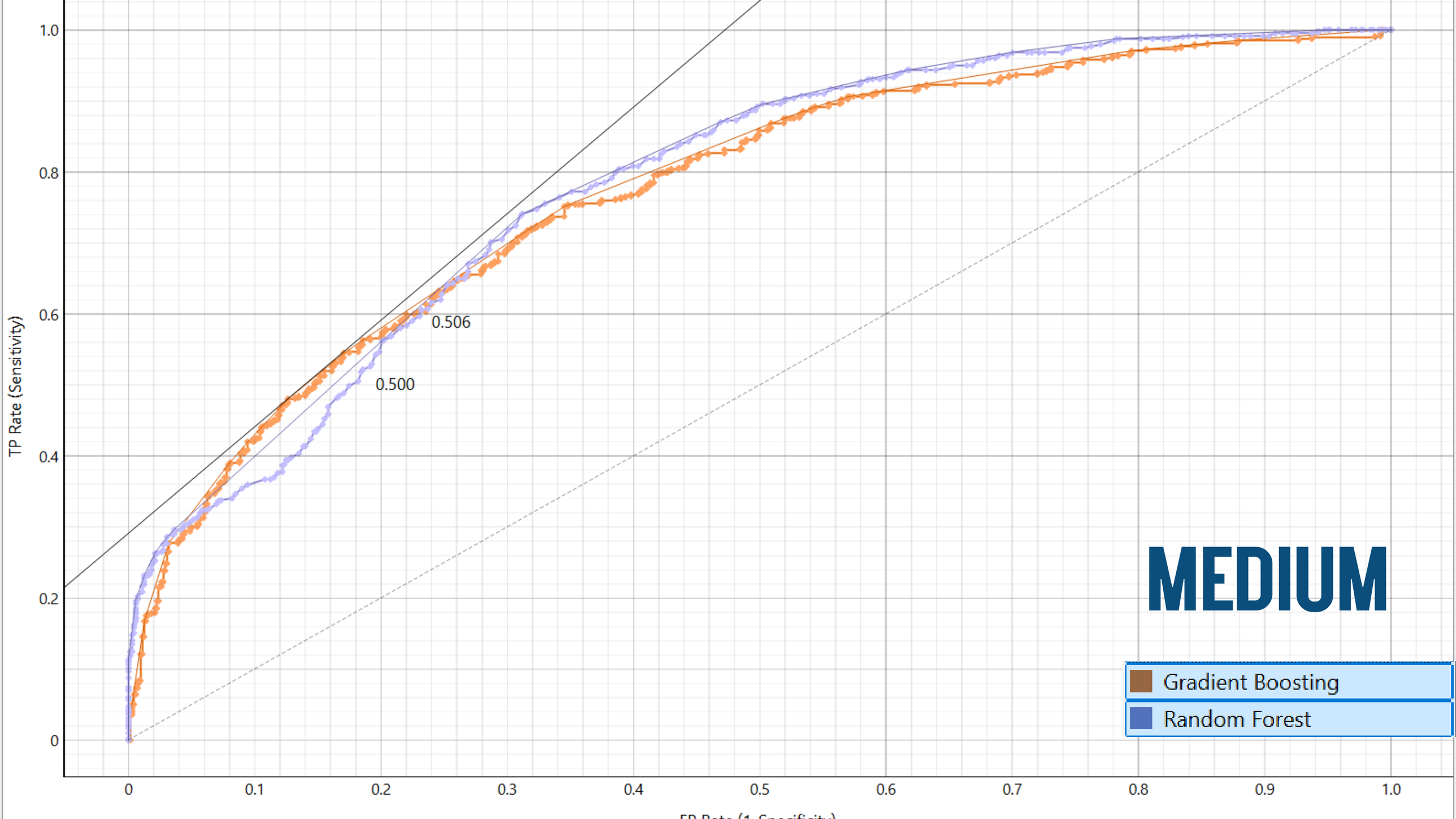

1.0
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.500
0.503
0.503
0.500
0.500
LOW
Tree
Gradient Boosting
Random Forest
SVM

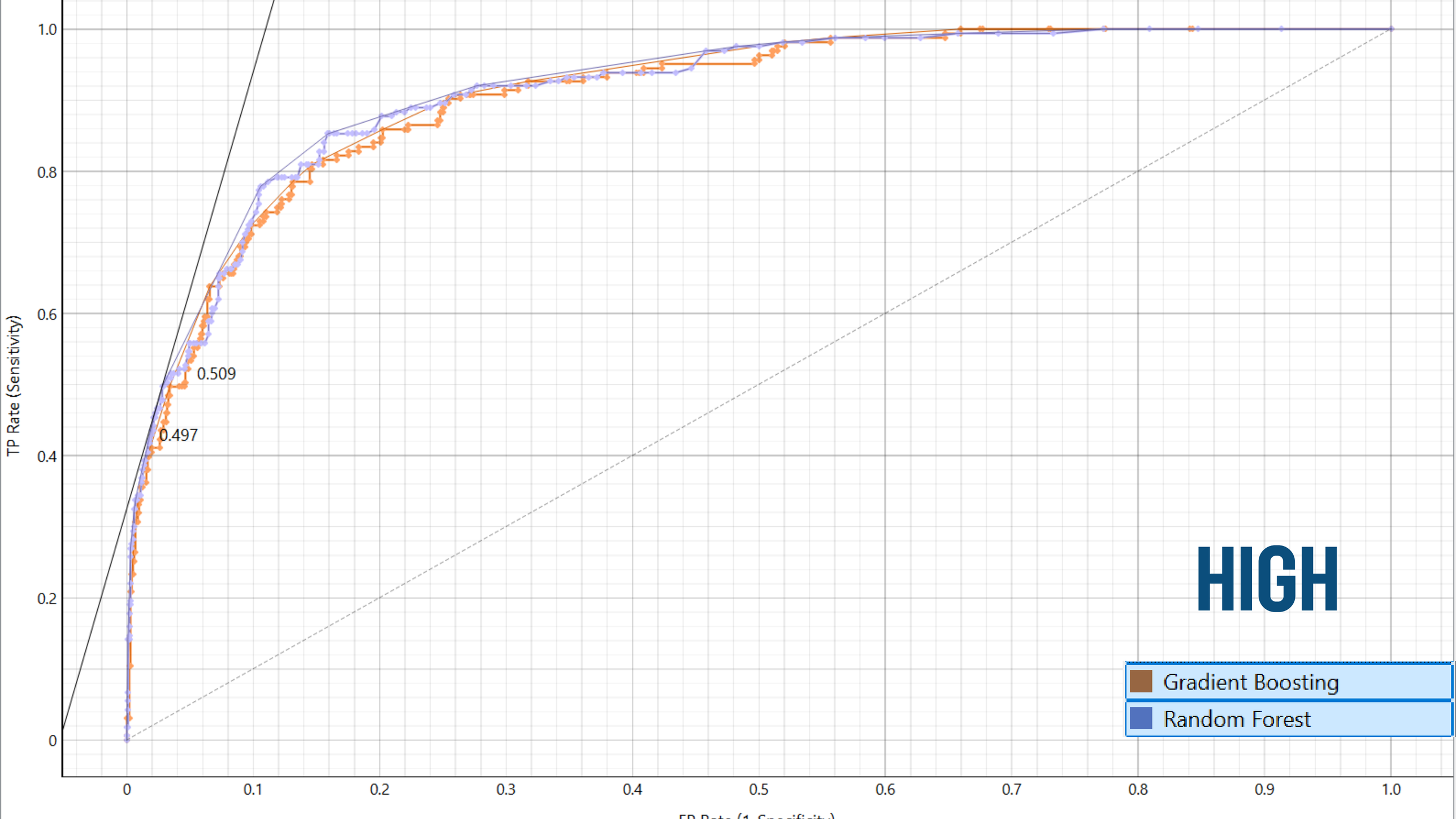

TP Rate (Sensitivity)
1.0
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.506
0.500
0.500
0.500
0.500
MEDIUM
Tree
Gradient Boosting
Random Forest
SVM


0.500
0.509
0.500
0.497
0.497
HIGH
Tree
Gradient Boosting
Random Forest
SVM
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
FP Rate (1-Specificity)

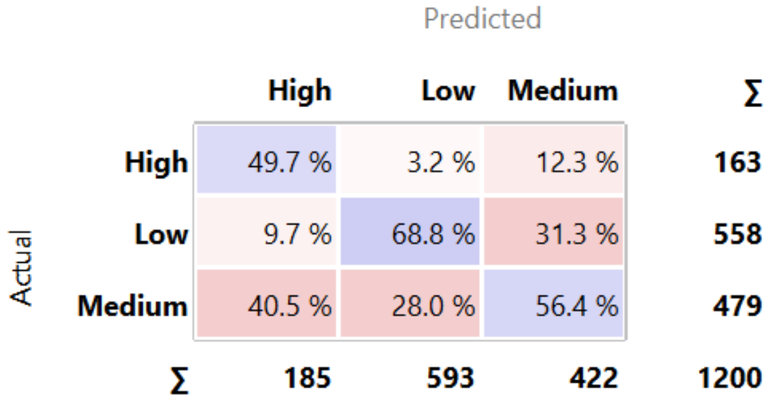
NOW COMPARING THE RANDOM FOREST AND GRADIENT BOOSTING


1.0
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.500
0.503
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
LOW
Gradient Boosting
Random Forest

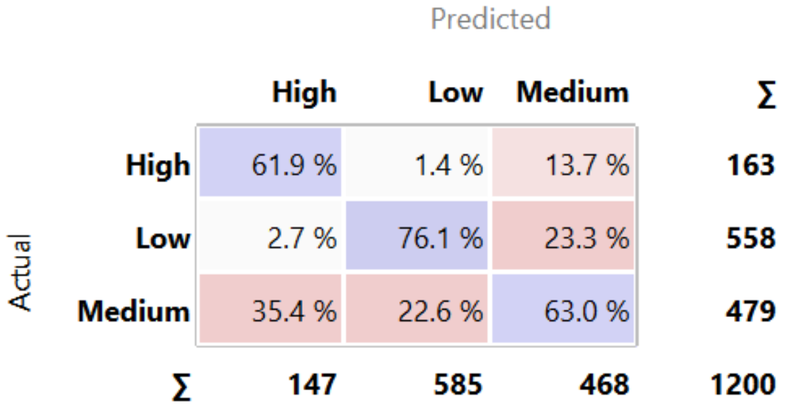

1.0
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.506
0.500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
MEDIUM
Gradient Boosting
Random Forest


TP Rate (Sensitivity)
1.0
0.8
0.6
0.4
0.2
0
0.509
0.497
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
HIGH
Gradient Boosting
Random Forest

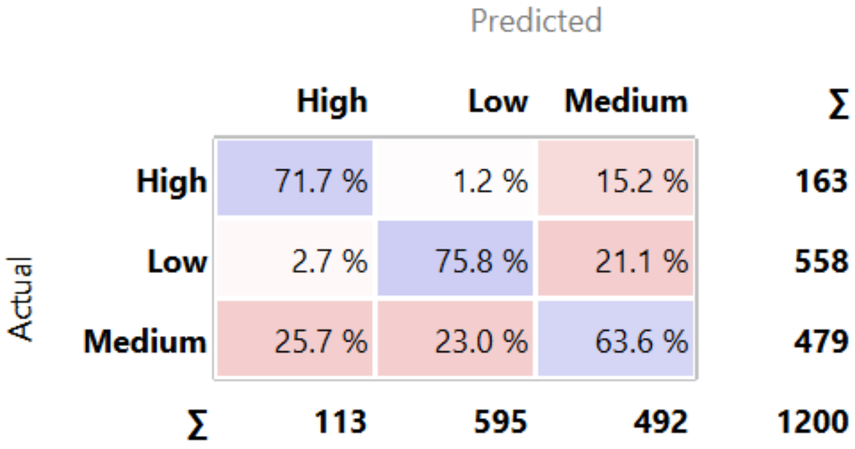
CONFUSION MATRIX ON CV



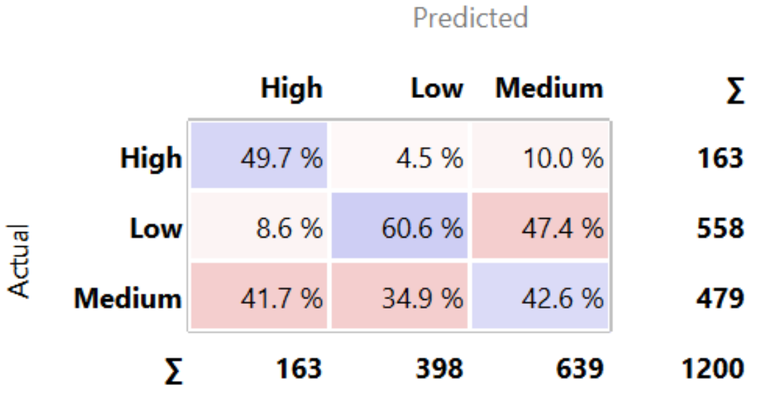
TREE



GRADIENT BOOSTING



RANDOM FOREST



SVM

RESULTS[CROSS VALIDATION]

Key Insights:

Random Forest performs the best in this scenario, with the highest AUC (0.847), accuracy (0.704), and MCC (0.502), making it the strongest model under stratified cross-validation.

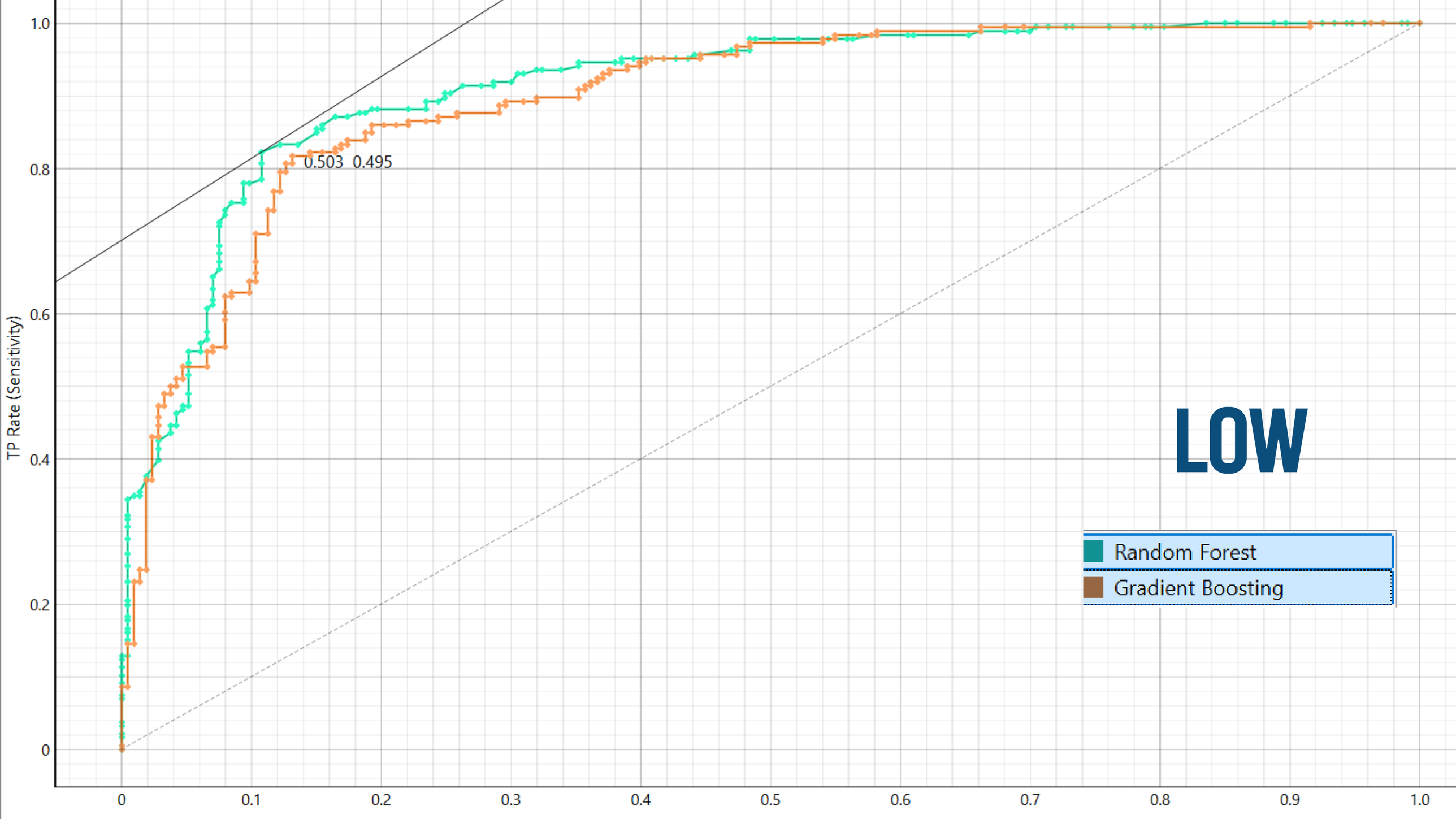
Gradient Boosting closely follows Random Forest, showing good performance across all metrics (AUC of 0.836 and CA of 0.693), but with slightly lower MCC (0.489).

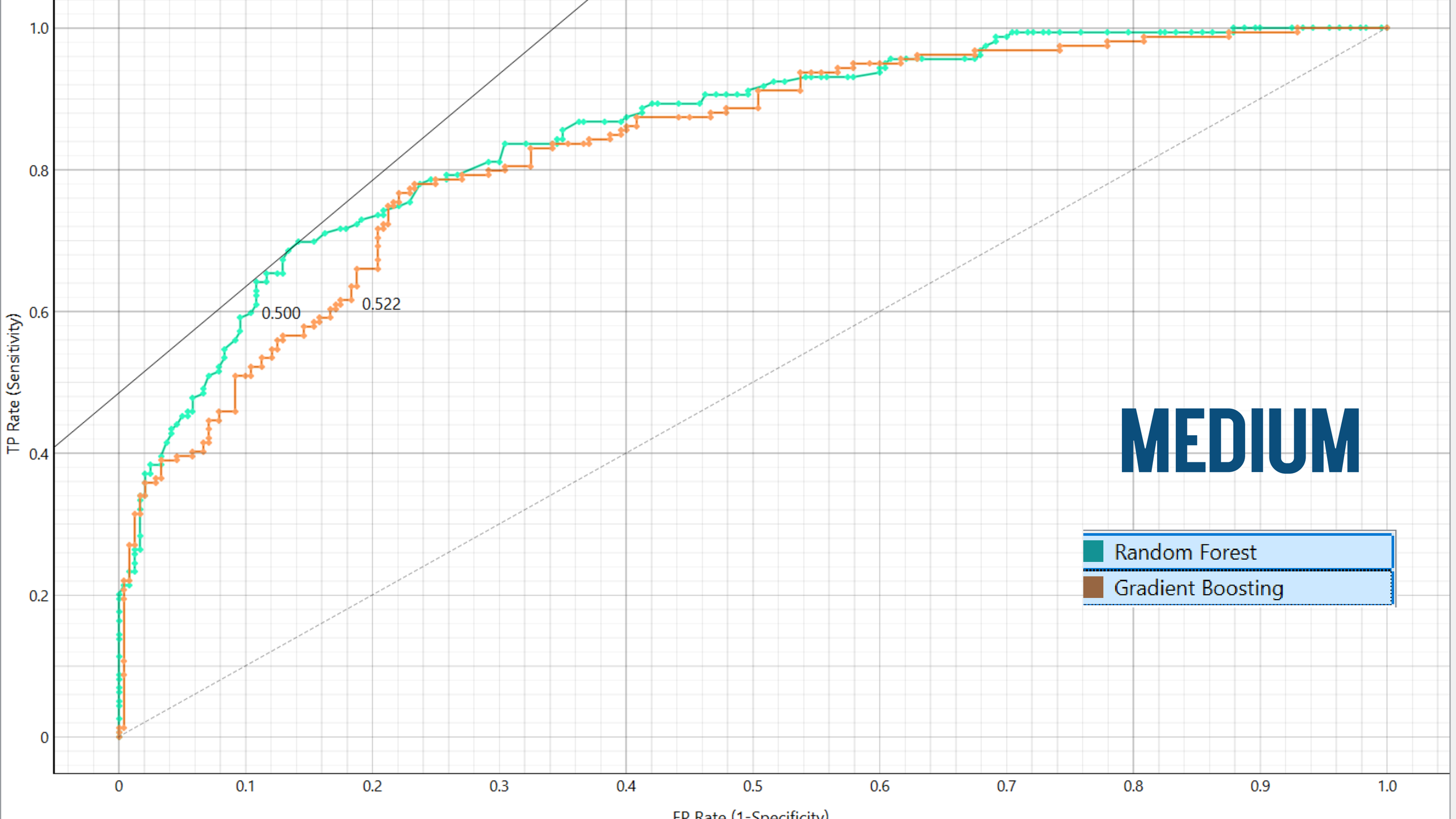
Decision Tree performs moderately well but shows lower metrics compared to the ensemble models.

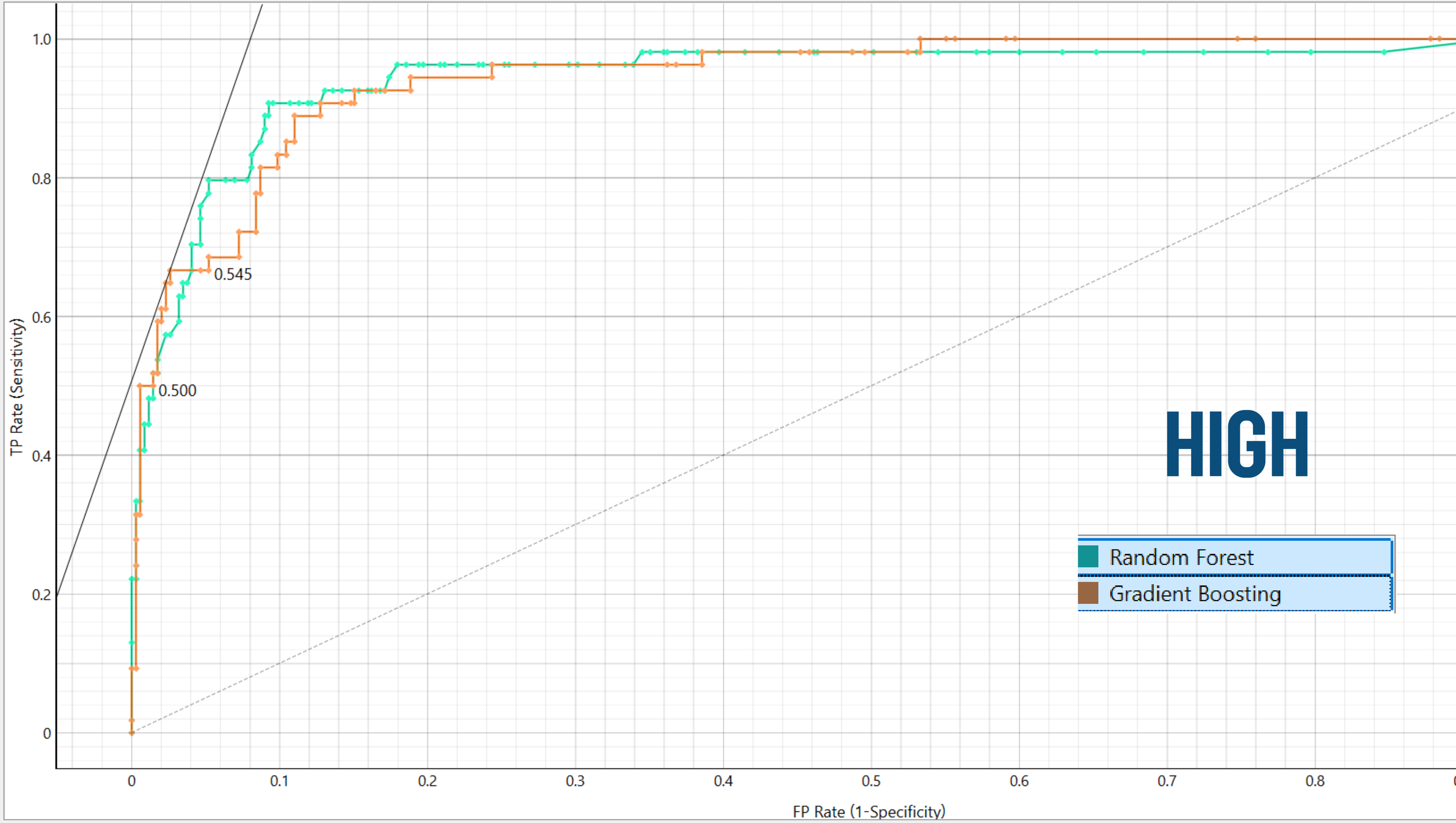
SVM struggles here, with notably low performance across all metrics (AUC 0.686, CA 0.495), indicating it‘s not suited for this dataset in a cross-validation scenario.

RESULTS[ON TEST DATA]

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.895	0.792	0.791	0.791	0.792	0.653
Gradient Boosting	0.880	0.744	0.744	0.744	0.744	0.579
SVM	0.664	0.489	0.488	0.516	0.489	0.161
Tree	0.734	0.659	0.656	0.656	0.659	0.436


1.0
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.503 0.495
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
LOW
Random Forest
Gradient Boosting


1.0
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.500
0.522
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
FP Rate (1-Specificity)
MEDIUM
Random Forest
Gradient Boosting


1.0
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.545
0.500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
FP Rate (1-Specificity)
HIGH
Random Forest
Gradient Boosting

CONFUSION MATRIX ON TEST DATA

Predicted

	High	Low	Medium	Σ
High	76.1 %	0.5 %	11.2 %	**54**
Low	2.2 %	83.9 %	14.4 %	**186**
Medium	21.7 %	15.5 %	74.4 %	**159**
Σ	**46**	**193**	**160**	**399**

RF

Actual

Predicted

	High	Low	Medium	Σ
High	63.8 %	0.5 %	10.7 %	**54**
Low	3.4 %	81.2 %	19.3 %	**186**
Medium	32.8 %	18.3 %	70.0 %	**159**
Σ	**58**	**191**	**150**	**399**

GB

RESULTS[ON TEST DATA]

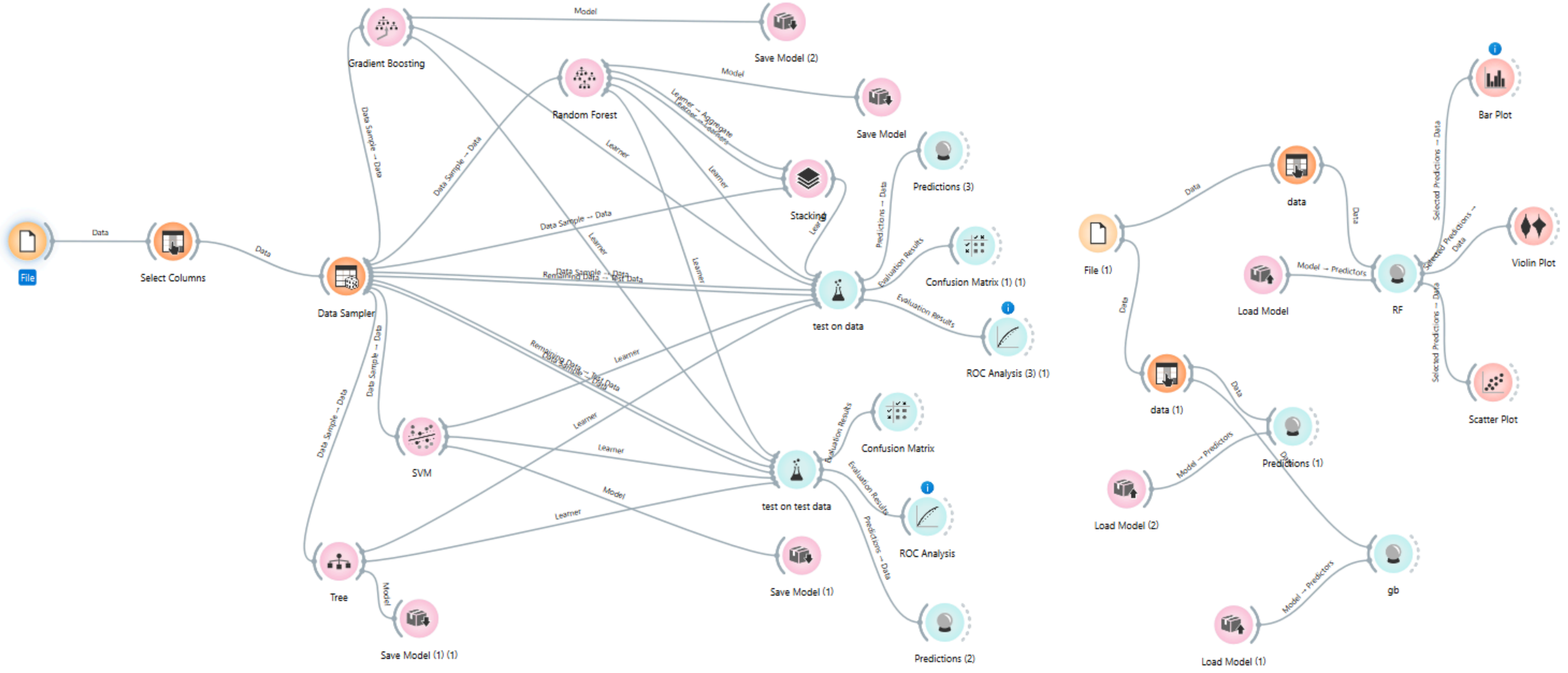
Key Insights:

Random Forest again stands out with the highest AUC (0.895), classification accuracy (0.792), and MCC (0.653), proving to be the most robust model in this setup.

Gradient Boosting also performs well, but its metrics are consistently lower than Random Forest (AUC of 0.880 and CA of 0.744).

Decision Tree shows reasonable performance but is significantly behind both ensemble methods in terms of accuracy and MCC.

SVM again performs poorly, showing the lowest metrics (AUC 0.664 and CA 0.489), similar to its performance in the cross-validation scenario.


File
Data
Select Columns
Data
Data Sampler
Gradient Boosting
Random Forest
Stacking
SVM
Tree
Model
Save Model (2)
Save Model
Save Model (1)
Save Model (1) (1)
Learner
Learner → Aggregate Learners
Data Sample → Data
Remaining Data → Test Data
Predictions → Data
Evaluation Results
Predictions (3)
Confusion Matrix (1) (1)
test on data
ROC Analysis (3) (1)
Confusion Matrix
test on test data
ROC Analysis
Predictions (2)
File (1)
data
data (1)
Load Model
Load Model (2)
Load Model (1)
Model → Predictors
RF
Predictions (1)
gb
Selected Predictions → Data
Bar Plot
Violin Plot
Scatter Plot

LOADING MODEL AND PREDICTIONS

Show probabilities for: Classes in data | Show classification errors | Restore Original Order

	Random Forest	error	quality_category	pH	alcohol	sulphates	density	otal sulfur dioxide	free sulfur dioxide	residual sugar	volatile acidity	chlorides
31	0.00 : 0.93 : 0.07 → Low	0.070	Low	3.35	10.1	0.54	0.99580	82.0	17.0	2.40	0.675	0.089
32	0.01 : 0.19 : 0.80 → Medium	0.197	Medium	3.46	10.6	0.57	0.99660	37.0	22.0	2.50	0.685	0.105
33	0.00 : 0.95 : 0.05 → Low	0.055	Low	3.17	9.8	0.66	0.99660	113.0	15.0	2.30	0.655	0.083
34	0.00 : 0.35 : 0.65 → Medium	0.348	Medium	3.45	9.4	0.52	0.99930	83.0	40.0	10.70	0.605	0.073
35	0.01 : 0.85 : 0.14 → Low	0.148	Low	3.38	9.2	0.55	0.99570	50.0	13.0	1.80	0.320	0.103
36	0.01 : 0.24 : 0.75 → Medium	0.245	Medium	3.40	9.6	0.55	0.99860	18.0	5.0	5.50	0.645	0.086
37	0.05 : 0.18 : 0.77 → Medium	0.227	Medium	3.42	10.8	0.60	0.99750	15.0	3.0	2.40	0.600	0.086
38	0.67 : 0.15 : 0.18 → High	0.330	High	3.23	9.7	0.73	0.99680	30.0	13.0	2.10	0.380	0.066
39	0.02 : 0.93 : 0.05 → Low	0.070	Low	3.50	9.8	0.48	0.99400	19.0	7.0	1.50	1.130	0.172
40	0.02 : 0.94 : 0.04 → Low	0.064	Low	3.33	10.5	0.83	0.99780	87.0	12.0	5.90	0.450	0.074
41	0.02 : 0.94 : 0.04 → Low	0.064	Low	3.33	10.5	0.83	0.99780	87.0	12.0	5.90	0.450	0.074
42	0.00 : 0.95 : 0.05 → Low	0.055	Low	3.26	9.3	0.51	0.99760	46.0	17.0	2.80	0.610	0.088
43	0.09 : 0.13 : 0.78 → Medium	0.221	Medium	3.21	10.5	0.90	0.99680	14.0	8.0	2.60	0.490	0.332
44	0.03 : 0.75 : 0.22 → Low	0.245	Low	3.30	10.3	1.20	0.99680	23.0	9.0	2.20	0.660	0.069
45	0.01 : 0.95 : 0.04 → Low	0.045	Low	3.48	9.5	0.52	0.99620	11.0	5.0	1.80	0.670	0.05
46	0.08 : 0.78 : 0.15 → Low	0.221	Low	3.90	13.1	0.56	0.99340	65.0	8.0	2.10	0.520	0.054
47	0.00 : 0.94 : 0.05 → Low	0.058	Low	3.25	9.2	0.73	0.99700	114.0	22.0	2.20	0.935	0.114
48	0.02 : 0.82 : 0.16 → Low	0.179	Low	3.25	9.5	0.58	0.99690	37.0	12.0	1.60	0.290	0.113
49	0.01 : 0.89 : 0.09 → Low	0.106	Low	3.34	9.2	0.56	0.99580	12.0	5.0	1.60	0.400	0.066
50	0.02 : 0.88 : 0.10 → Low	0.118	Low	3.32	9.2	0.58	0.99540	96.0	12.0	1.40	0.310	0.074

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.997	0.976	0.976	0.976	0.976	0.960

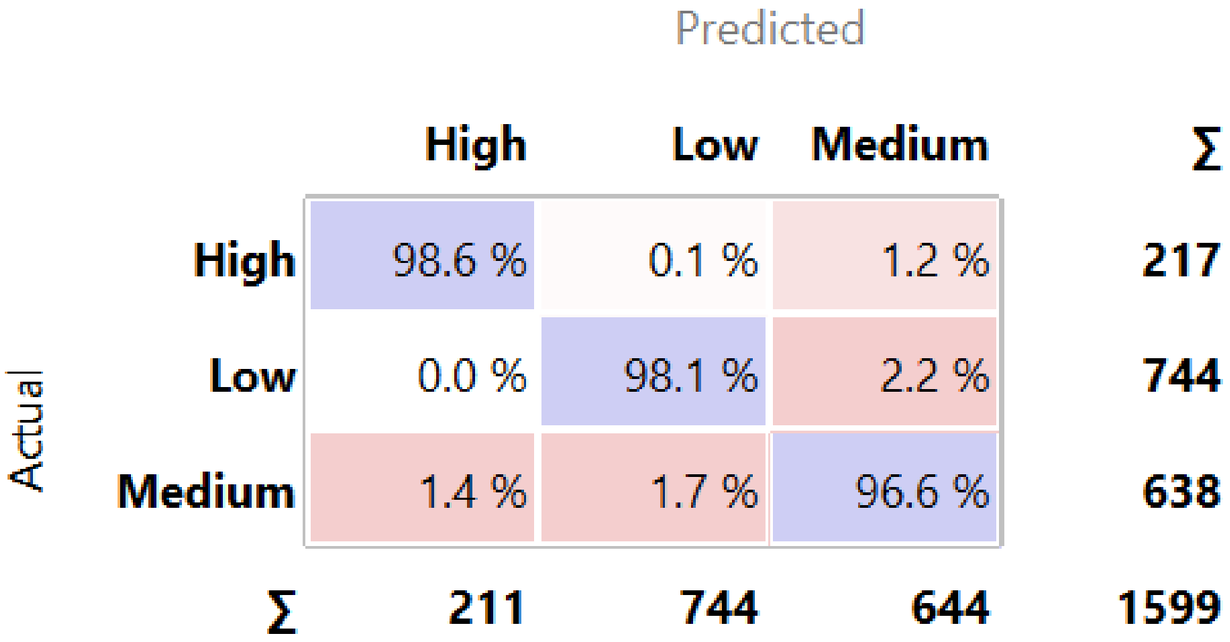
LOADING MODEL AND PREDICTIONS

Show probabilities for Classes known to the model ☑ Show classification errors Restore Original Order

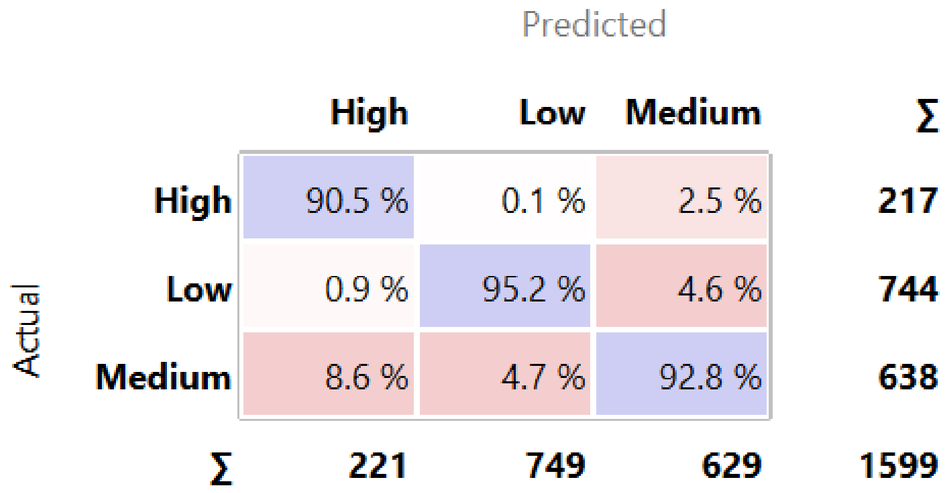
	Gradient Boosting	error	quality_category	pH	alcohol	sulphates	density	otal sulfur dioxide	free sulfur dioxide	residual sugar	volatile acidity	chlorides	
1	0.00 : 1.00 : 0.00 → Low	0.001	Low	3.51	9.4	0.56	0.99780	34.0	11.0	1.90	0.700	0.076	0
2	0.00 : 1.00 : 0.00 → Low	0.002	Low	3.20	9.8	0.68	0.99680	67.0	25.0	2.60	0.880	0.098	0
3	0.00 : 1.00 : 0.00 → Low	0.004	Low	3.26	9.8	0.65	0.99700	54.0	15.0	2.30	0.760	0.092	0
4	0.00 : 0.01 : 0.99 → Medium	0.007	Medium	3.16	9.8	0.58	0.99800	60.0	17.0	1.90	0.280	0.075	0
5	0.00 : 1.00 : 0.00 → Low	0.001	Low	3.51	9.4	0.56	0.99780	34.0	11.0	1.90	0.700	0.076	0
6	0.00 : 1.00 : 0.00 → Low	0.000	Low	3.51	9.4	0.56	0.99780	40.0	13.0	1.80	0.660	0.075	0
7	0.00 : 1.00 : 0.00 → Low	0.000	Low	3.30	9.4	0.46	0.99640	59.0	15.0	1.60	0.600	0.069	0
8	0.98 : 0.01 : 0.00 → High	0.016	High	3.39	10	0.47	0.99460	21.0	15.0	1.20	0.650	0.065	0
9	0.98 : 0.01 : 0.01 → High	0.019	High	3.36	9.5	0.57	0.99680	18.0	9.0	2.00	0.580	0.073	0
10	0.00 : 1.00 : 0.00 → Low	0.000	Low	3.35	10.5	0.80	0.99780	102.0	17.0	6.10	0.500	0.071	0
11	0.00 : 1.00 : 0.00 → Low	0.004	Low	3.28	9.2	0.54	0.99590	65.0	15.0	1.80	0.580	0.097	0
12	0.00 : 1.00 : 0.00 → Low	0.000	Low	3.35	10.5	0.80	0.99780	102.0	17.0	6.10	0.500	0.071	0
13	0.00 : 1.00 : 0.00 → Low	0.000	Low	3.58	9.9	0.52	0.99430	59.0	16.0	1.60	0.615	0.089	0
14	0.00 : 0.82 : 0.18 → Low	0.180	Low	3.26	9.1	1.56	0.99740	29.0	9.0	1.60	0.610	0.114	0
15	0.00 : 1.00 : 0.00 → Low	0.001	Low	3.16	9.2	0.88	0.99860	145.0	52.0	3.80	0.620	0.176	0
16	0.00 : 1.00 : 0.00 → Low	0.001	Low	3.17	9.2	0.93	0.99860	148.0	51.0	3.90	0.620	0.17	0
17	0.99 : 0.01 : 0.00 → High	0.010	High	3.30	10.5	0.75	0.99690	103.0	35.0	1.80	0.280	0.092	0
18	0.00 : 0.99 : 0.01 → Low	0.007	Low	3.11	9.3	1.28	0.99680	56.0	16.0	1.70	0.560	0.368	0
19	0.00 : 0.80 : 0.20 → Low	0.202	Low	3.38	9	0.50	0.99740	29.0	6.0	4.40	0.590	0.086	0
20	0.00 : 0.01 : 0.99 → Medium	0.007	Medium	3.04	9.2	1.08	0.99690	56.0	17.0	1.80	0.320	0.341	0

Model	AUC	CA	F1	Prec	Recall	MCC
Gradient Boosting	0.988	0.936	0.936	0.936	0.936	0.895

CONFUSION MATRIX



RANDOM FOREST



GRADIENT BOOSTING

COMPARISON AND CONCLUSION OF MODELS USING DIFFERENT SAMPLING TYPES

- Random Forest is the clear winner across both sampling techniques. It consistently provides the highest accuracy, AUC, and MCC, making it the most reliable model for your dataset.
- Gradient Boosting comes in second place, providing good performance across the board but falling slightly behind Random Forest in terms of accuracy and robustness.
- Decision Trees offer decent performance but are clearly outclassed by the ensemble methods. They might still be useful for their simplicity and interpretability.
- SVM consistently underperforms, making it unsuitable for this dataset and task.

Final Recommendation: Based on this comparison, Random Forest is the best choice for your wine quality prediction ,providing strong generalization across both cross-validation and test scenarios.

PROJECT REFERENCE

DATASET:RED_WIN

OUR PROJECT REPOSTROY

THANK YOU VERY MUCH!