Titanic Survival Dataset

Problem description

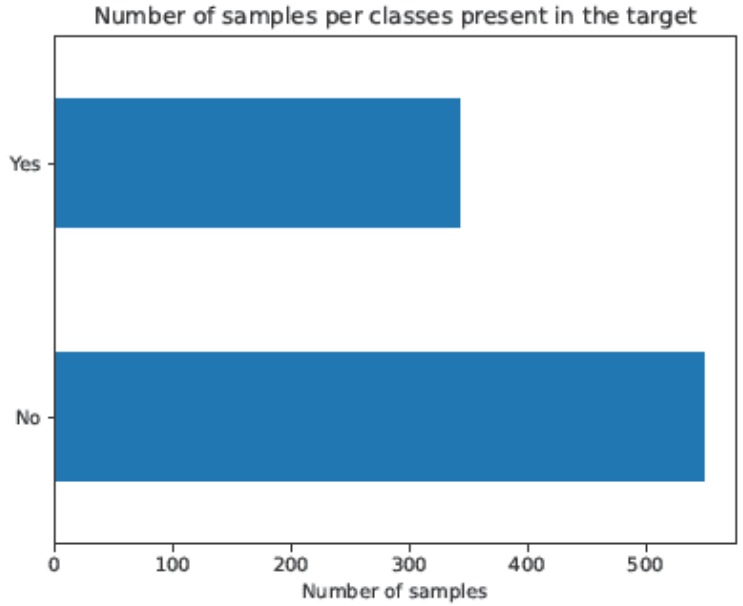
- Target
- Features

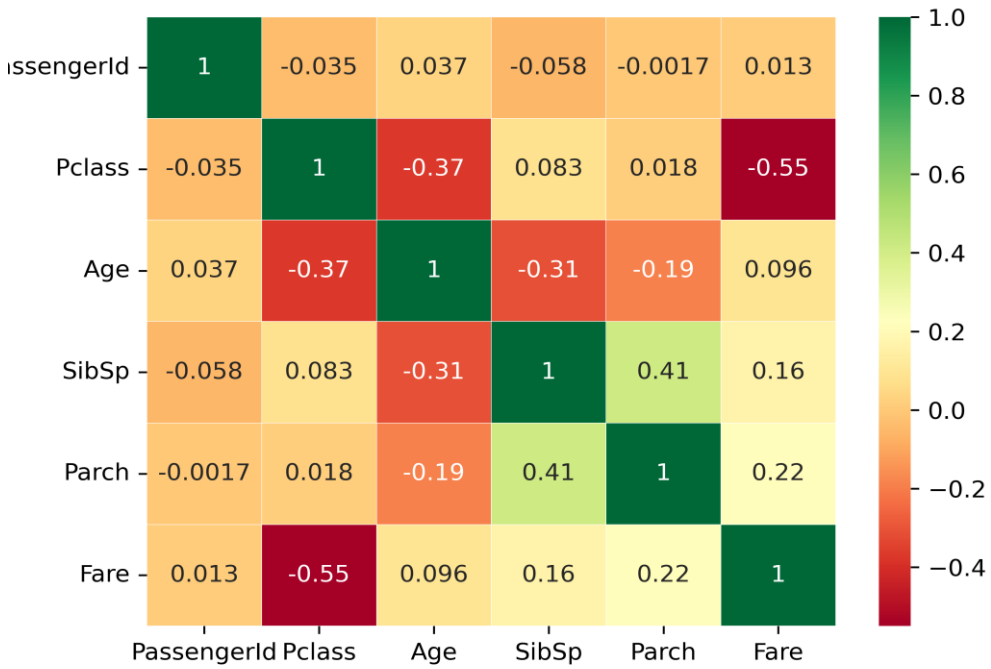
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	No	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	Yes	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	Yes	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	Yes	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	No	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

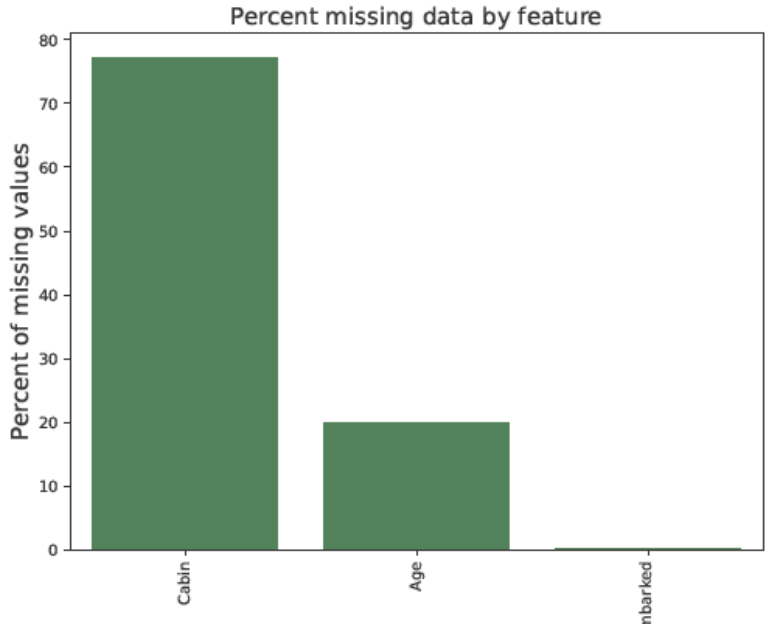
- **Sibsp** defines family relations (# of brother, sister, stepbrother, stepsister, …)
- **Parch** defines family relations (# of mother, father, daughter, son, …)
- **Embarked** is the port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
- **Pclass** is proxy for socio-economic status (1st = Upper, 2nd = Middle, 3rd = Lower)

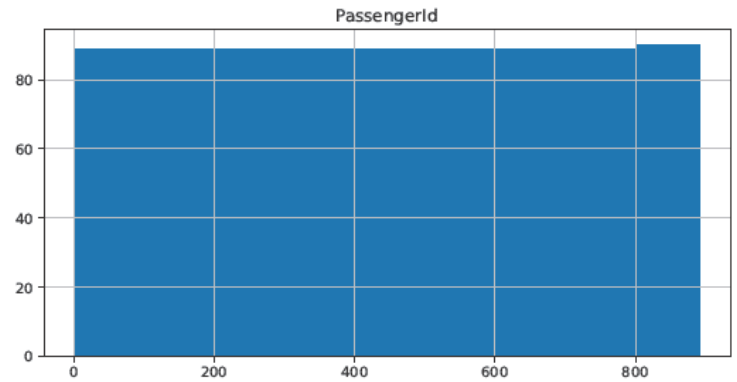
Data exploration

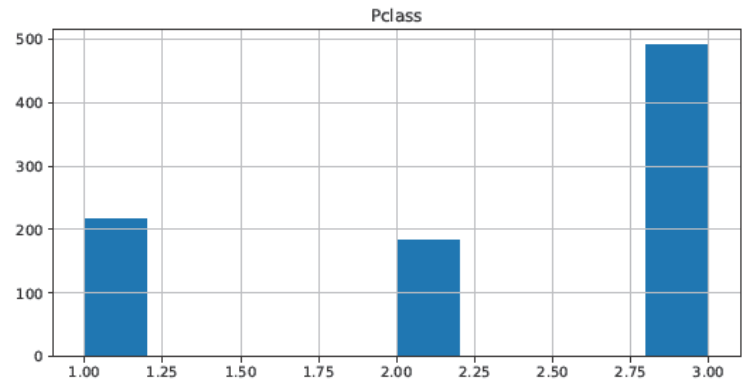
- Total # of examples: **891**
- Data type: **categorical** and **numerical**
- **Missing data**
- **Not balanced** (549 no, 342 yes)

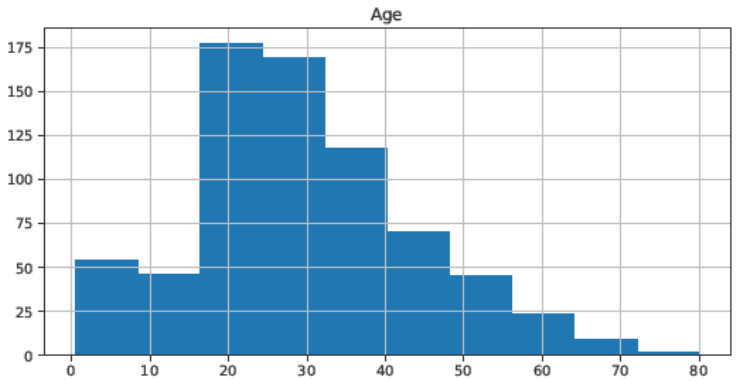


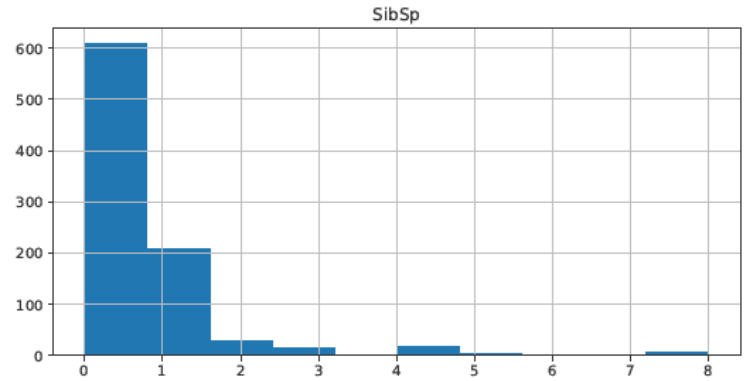


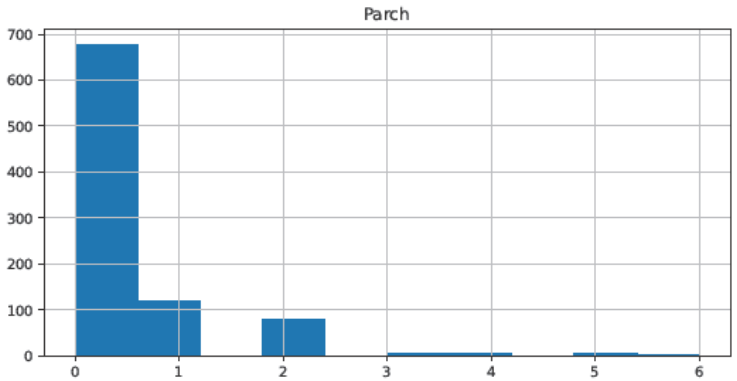


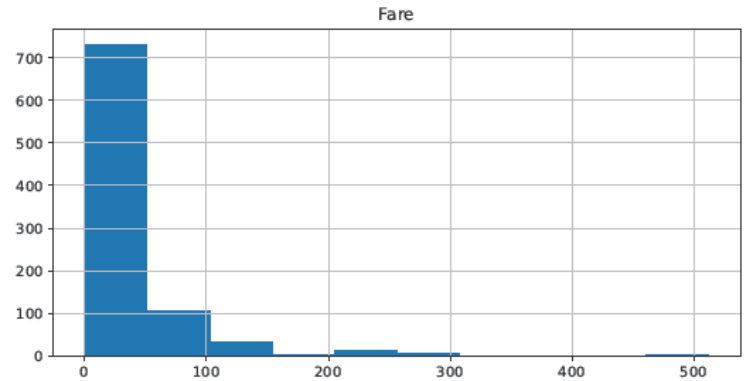

PassengerId
0
20
40
60
80
0
200
400
600
800

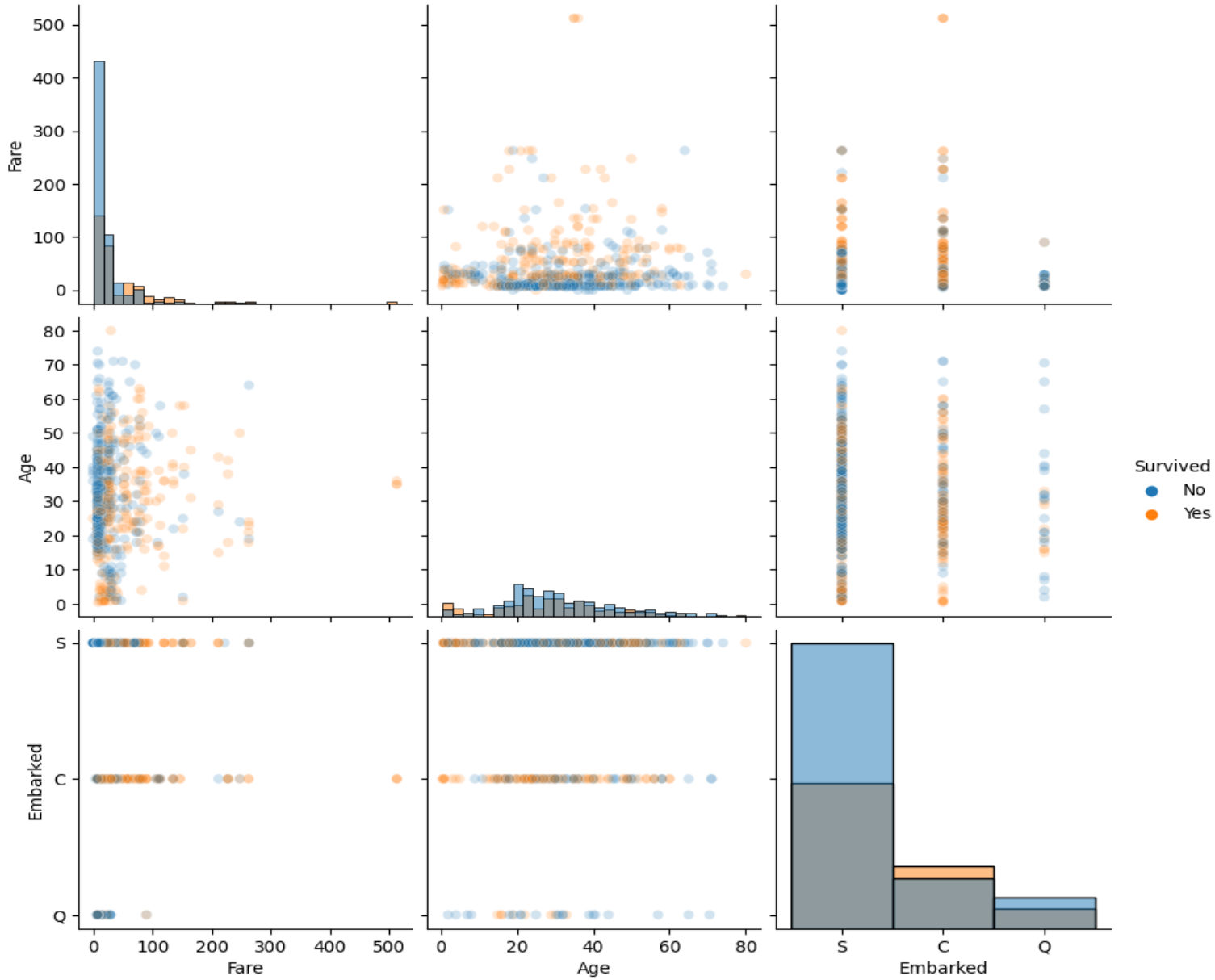

Pclass
0
100
200
300
400
500
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Age
0
25
50
75
100
125
150
175
0
10
20
30
40
50
60
70
80


SibSp
0
100
200
300
400
500
600
0
1
2
3
4
5
6
7
8


Parch
0
100
200
300
400
500
600
700
0
1
2
3
4
5
6

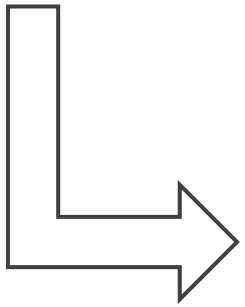

Fare
0
100
200
300
400
500
600
700
0
100
200
300
400
500


Fare
Age
Embarked
Fare
Age
Embarked
Survived
No
Yes

Features selection

- Too many Nan
- Correlated
- Unuseful

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	No	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	Yes	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	Yes	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	Yes	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	No	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S



	Survived	Pclass	Sex	SibSp	Parch	Fare	Embarked
0	No	3	male	1	0	7.2500	S
1	Yes	1	female	1	0	71.2833	C
2	Yes	3	female	0	0	7.9250	S
3	Yes	1	female	1	0	53.1000	S
4	No	3	male	0	0	8.0500	S

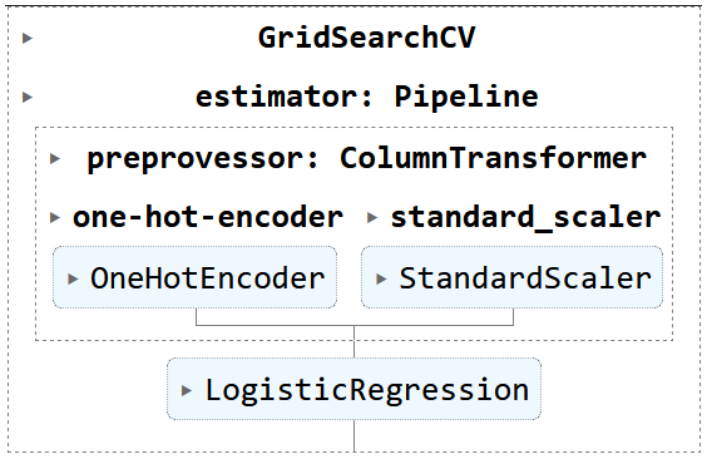
Machine Learning Models

ML Pipeline:
- Preprocessing (categorical/numerical)
- ML model

Hyperparameters Tuning:
- Grid-search

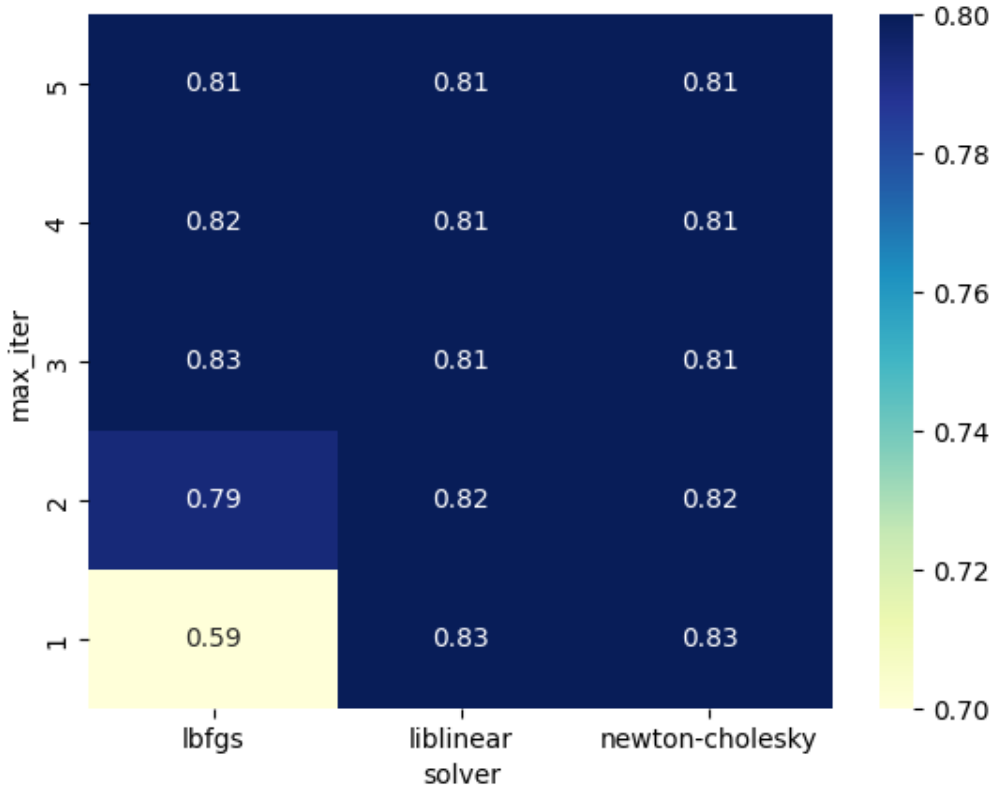
Overfitting check:
- Validation curve
- Training-error vs Test-error table

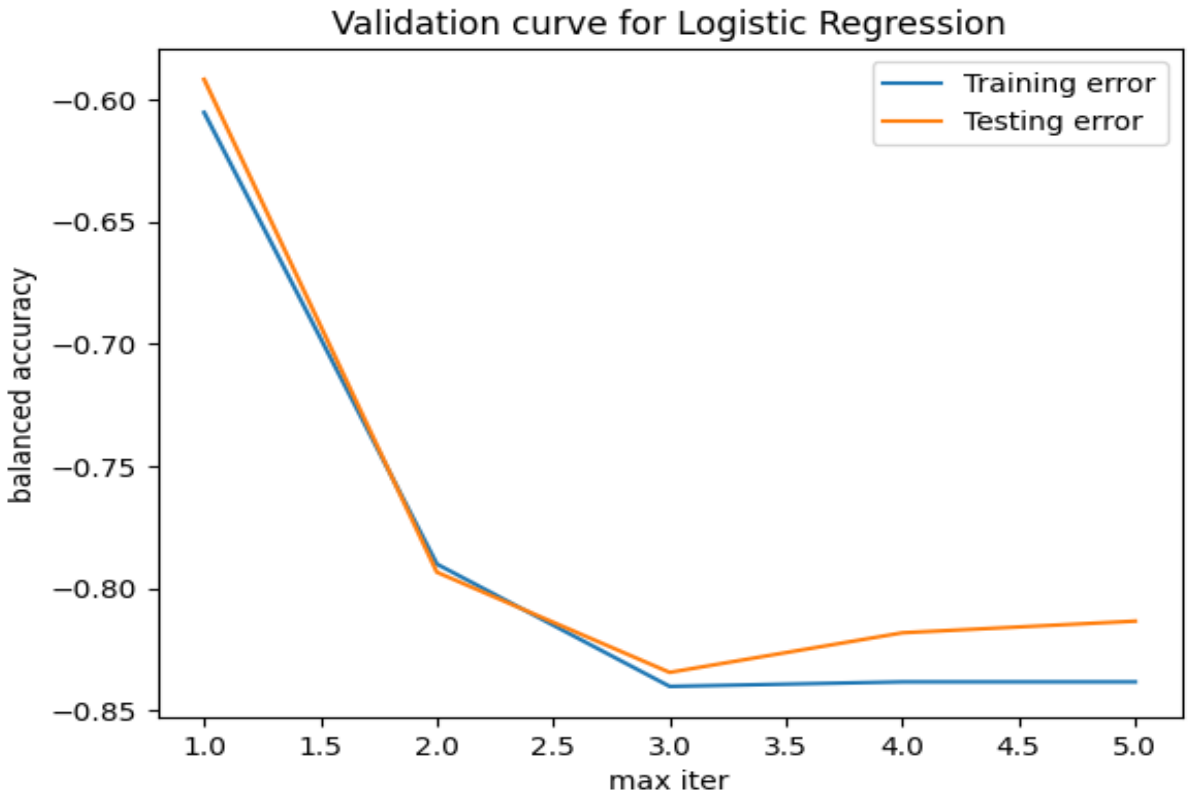
Performance Metrics:
- Balanced accuracy
- Model vs Dummy Classifier
- AUC (ROC-Curve)
- AP (PR-Curve)



ML Models:
- SVM
- Linear Discriminant Analysis
- Logistic Regression
- Gradient Decent
- Decision Tree
- AdaBoostClassifier
- Naive Bayes
- Random Forest
- KNN

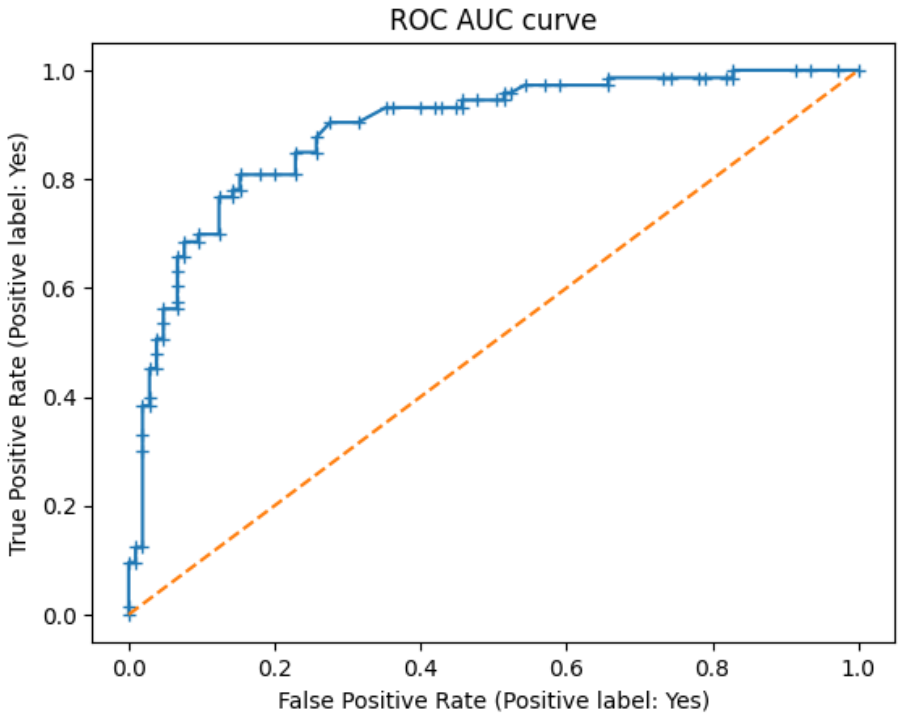
Logistic Regression: tuning and overfitting

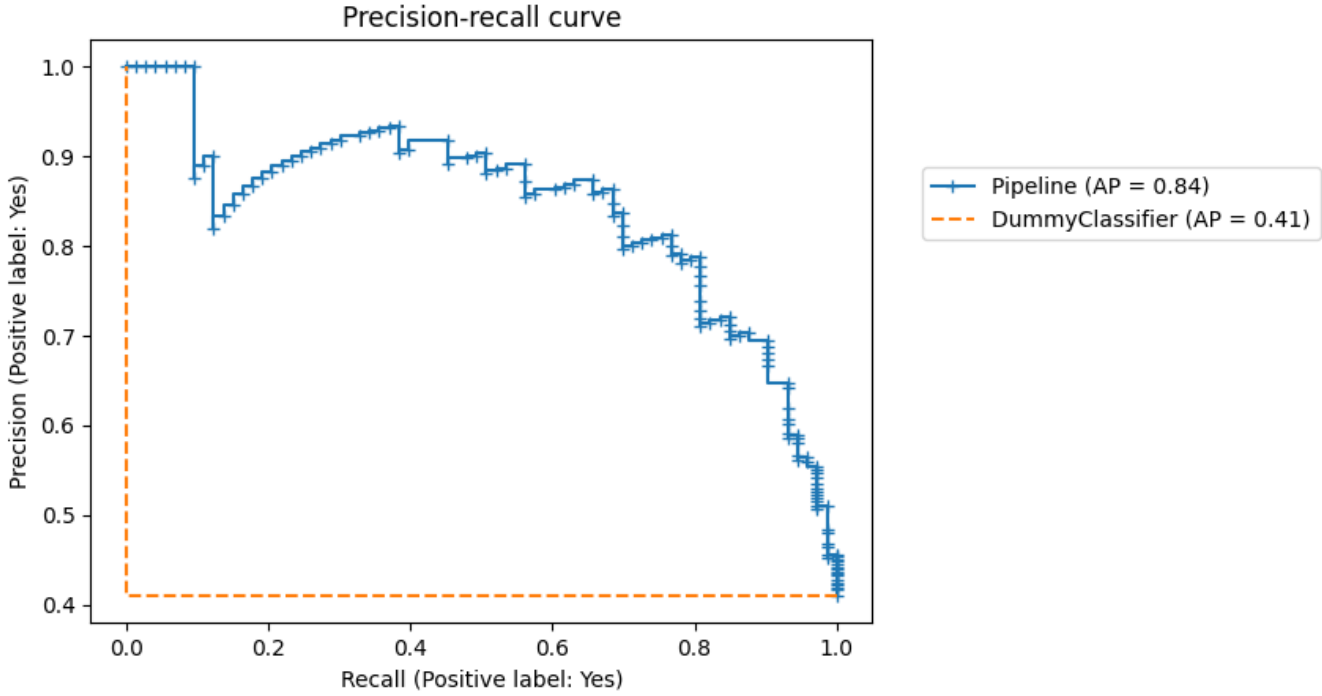




- **solver**: liblinear (dataset piccolo)
- **max_iter**: 3

Logistic Regression



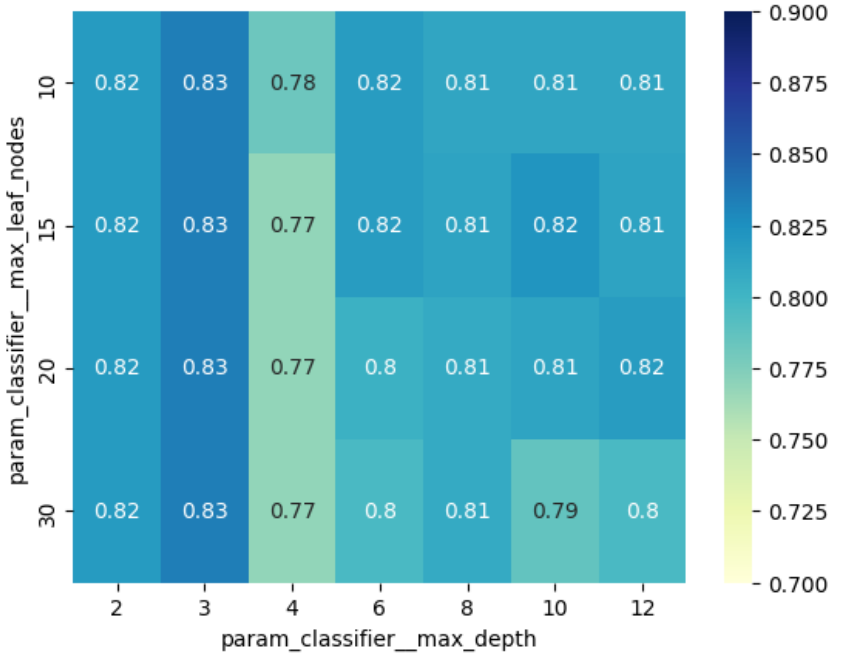


Balanced accuracy: 0.82

Decision Tree: tuning and overfitting

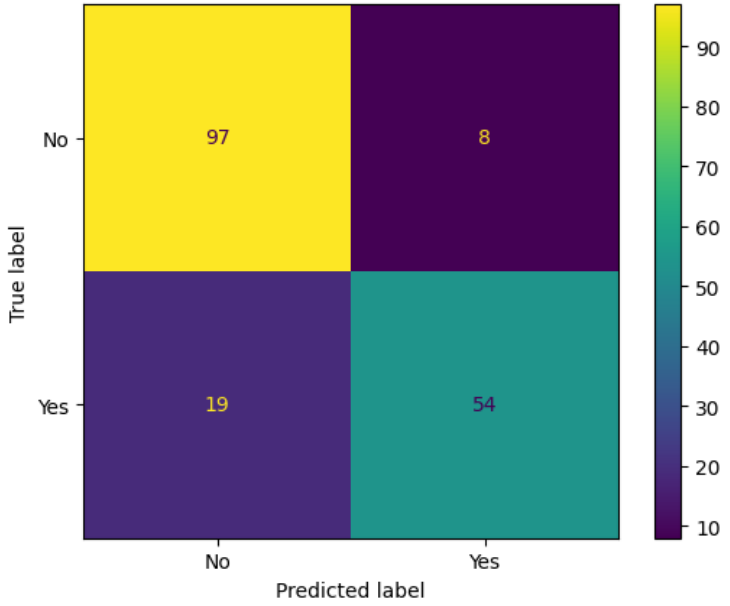
	param_classifier__max_leaf_nodes	param_classifier__max_depth	mean_test_score	rank_test_score	mean_train_score
5	15	3	0.833810	1	0.849293
7	30	3	0.833810	1	0.849293
6	20	3	0.833810	1	0.849293
4	10	3	0.833810	1	0.849293
18	20	8	0.824286	5	0.918407
25	15	12	0.822381	6	0.906367
3	30	2	0.818095	7	0.842356

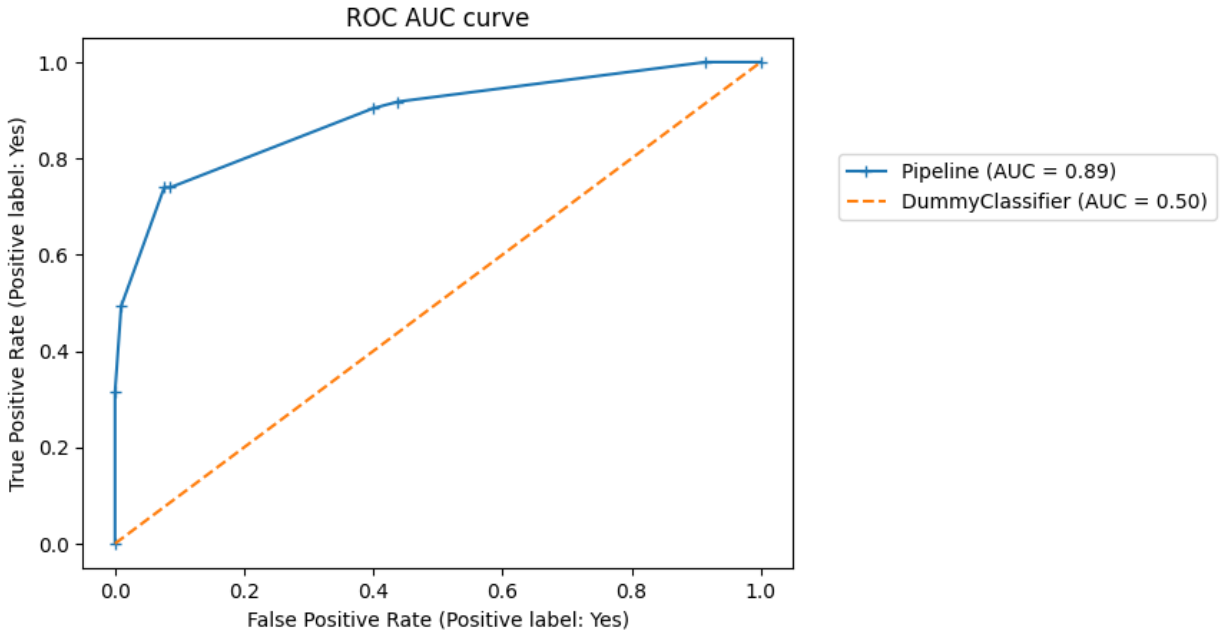
- **max_depth**: 3
- **max_leaf_nodes**: 10

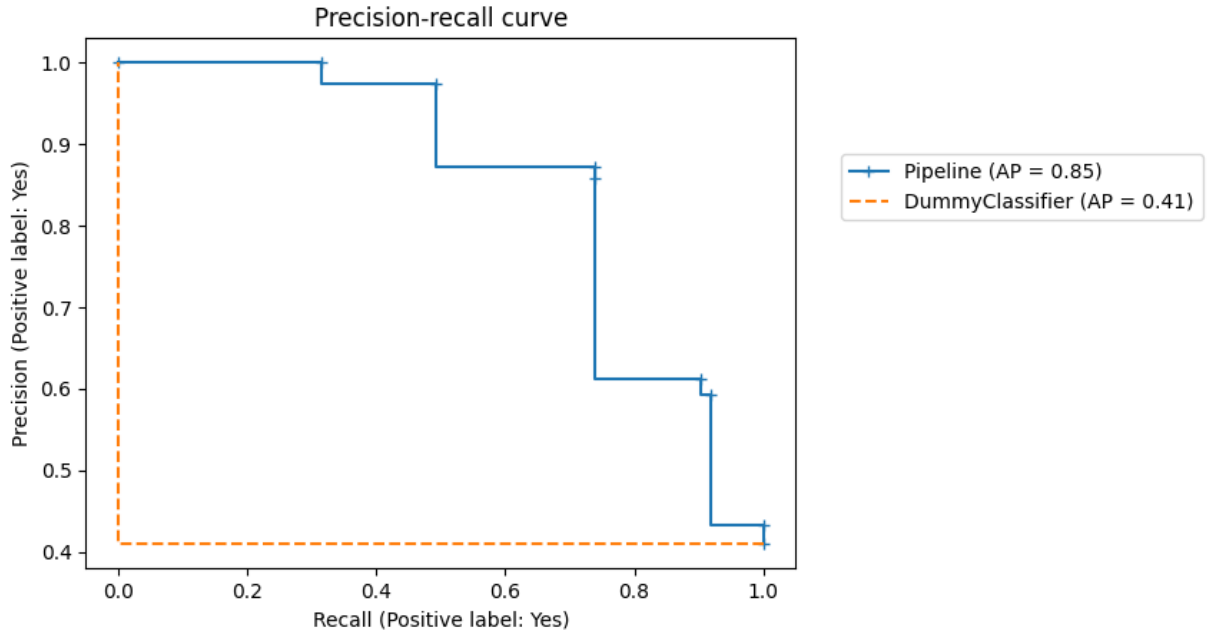


Decision Tree

Balanced accuracy: 0.83

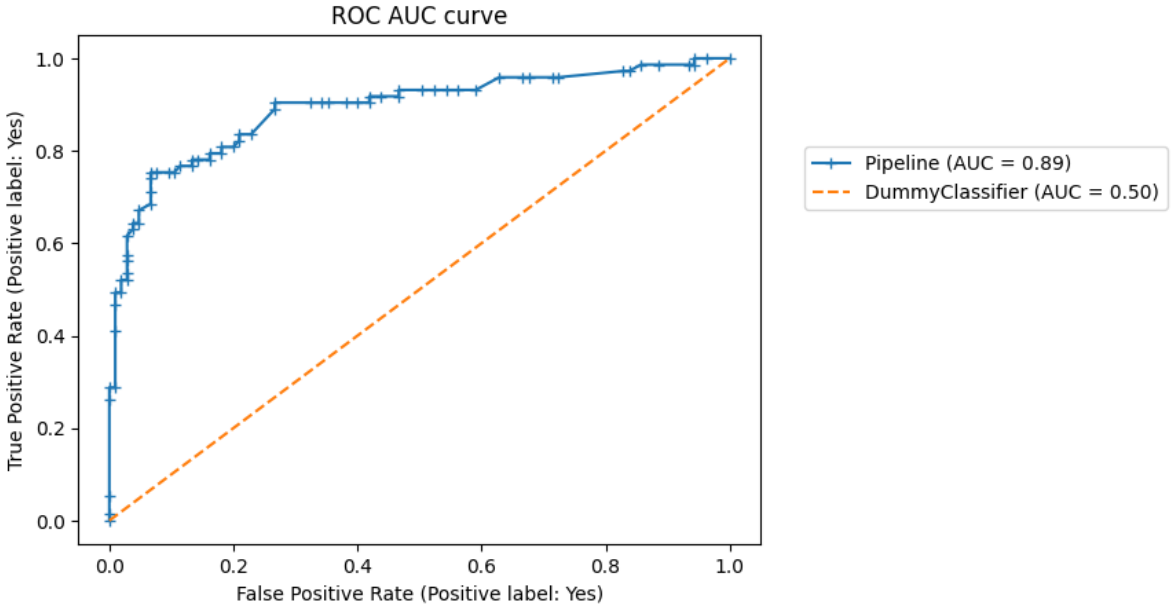


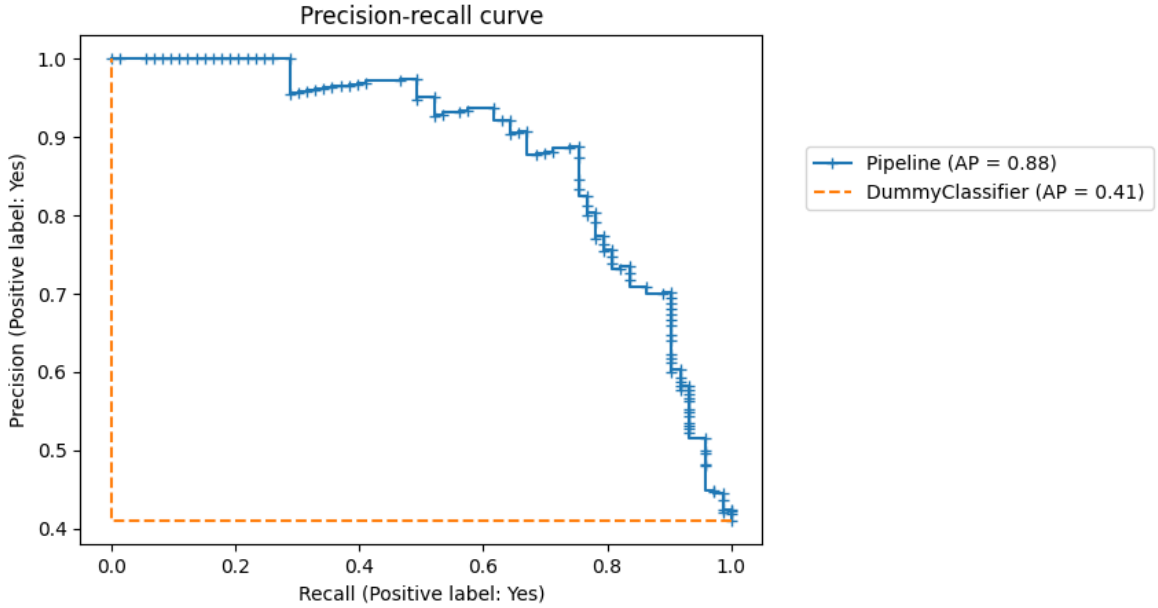




Random Forest

- Decision trees tend to overfitting -> **high variance and low bias**
- A possible solution to improve performance is **Bagging** -> **Random Forest**

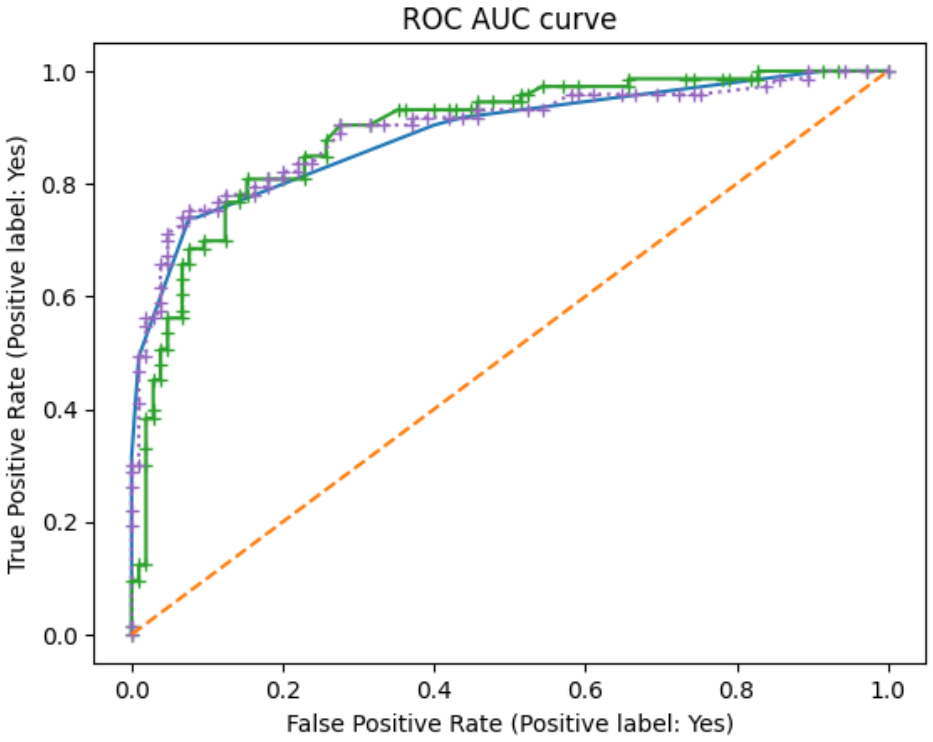


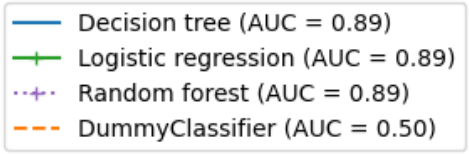


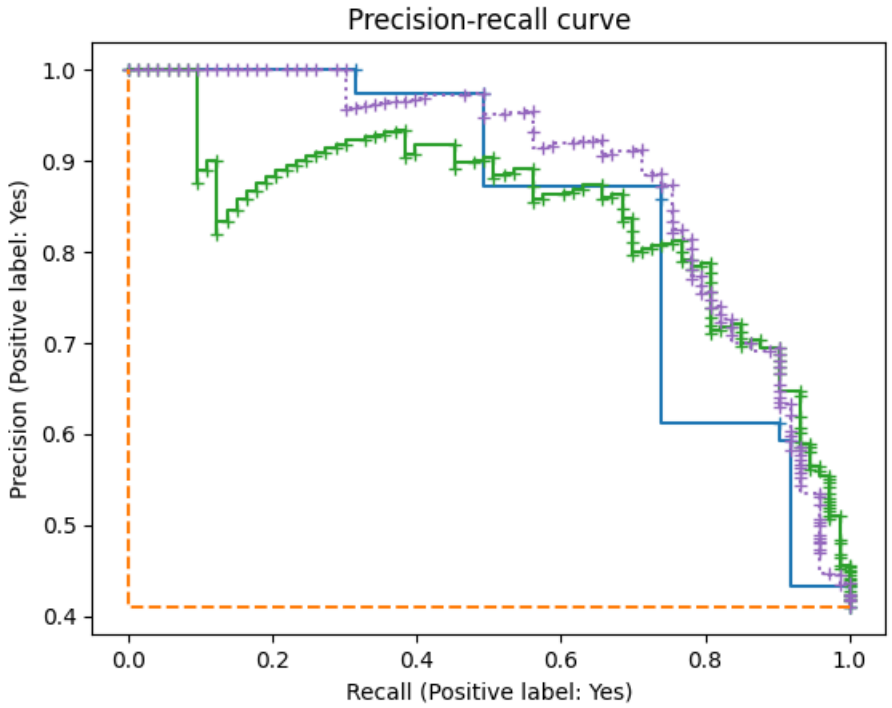
- **n_estimators**: 150
- **max_depth**: 3

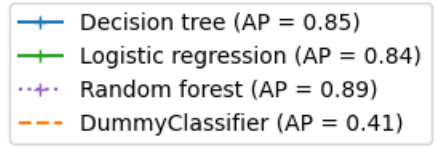
Balanced accuracy: 0.84

Model performance comparison









Conclusions

- The performance of the three models is **similar**
- The **Random Forest seems to perform better** (see AP values)
- Changing the set of features (adding ‘Age’ and ‘Cabin’) does not improve the performance
- In the future it woud be interesting to **evaluate other classifiers**

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	No	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	Yes	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	Yes	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	Yes	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	No	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	Yes	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C	C
3	4	Yes	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C	S
6	7	No	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E	S
10	11	Yes	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G	S
11	12	Yes	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C	S
...	...	...	...	...	...	...	...	...	...	...	...	...

Grazie per l’attenzione