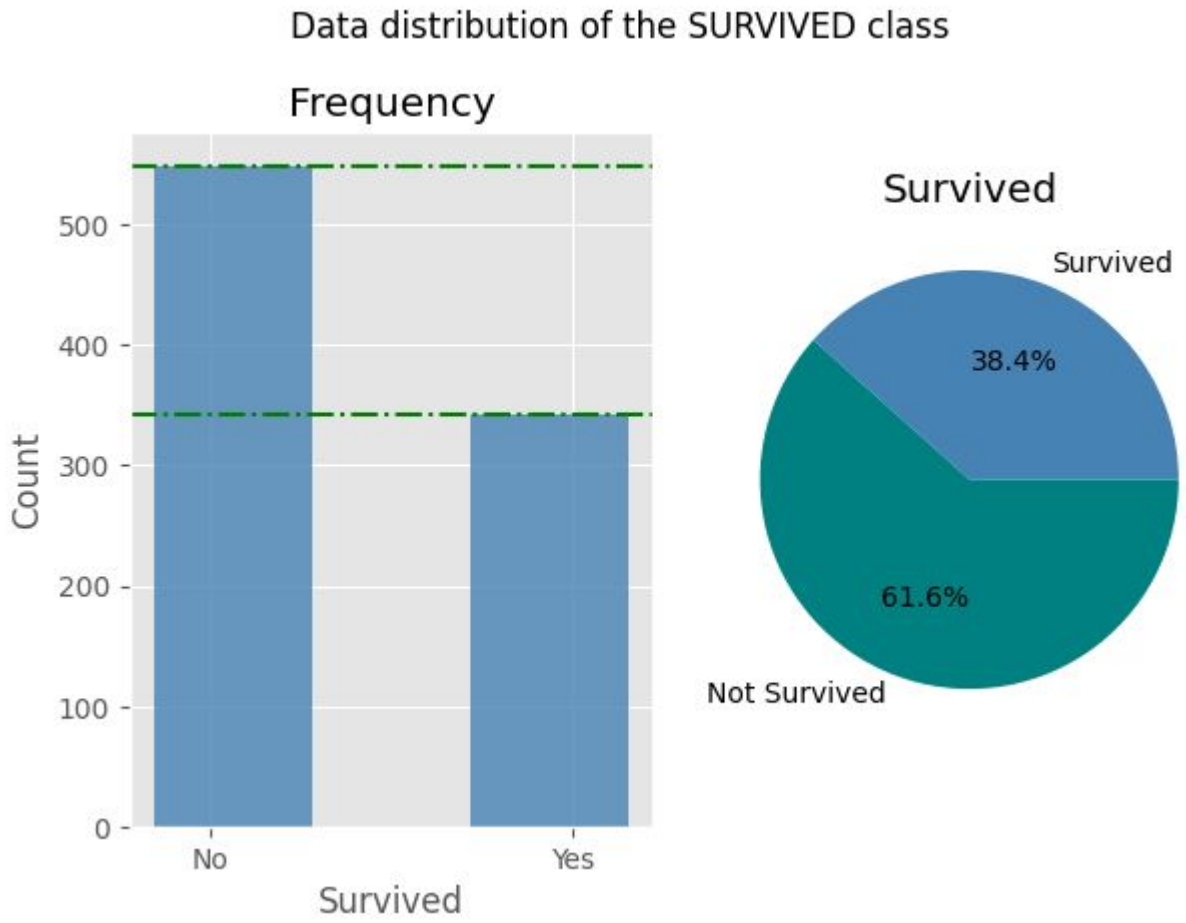
TITANIC: KAGGLE CHALLENGE

ADVANCED AI FOR DATA SCIENCE I

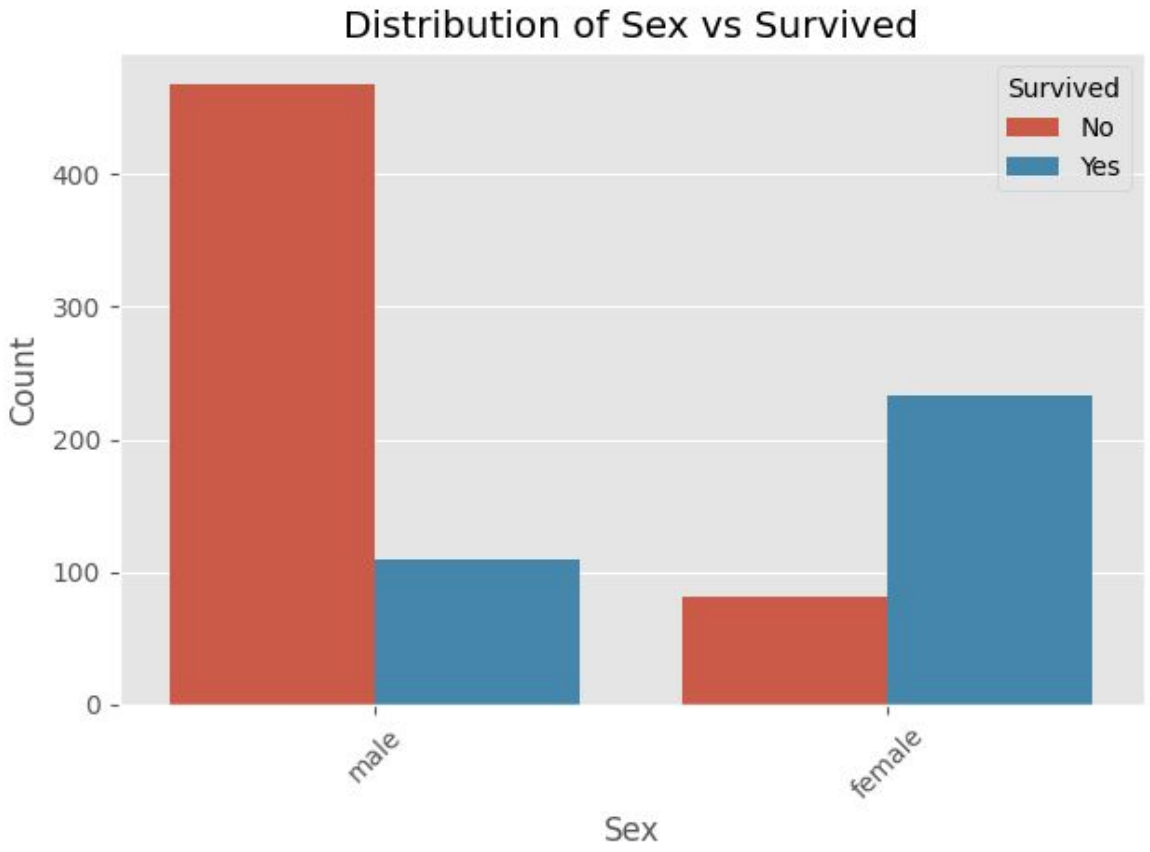
TEAM 7

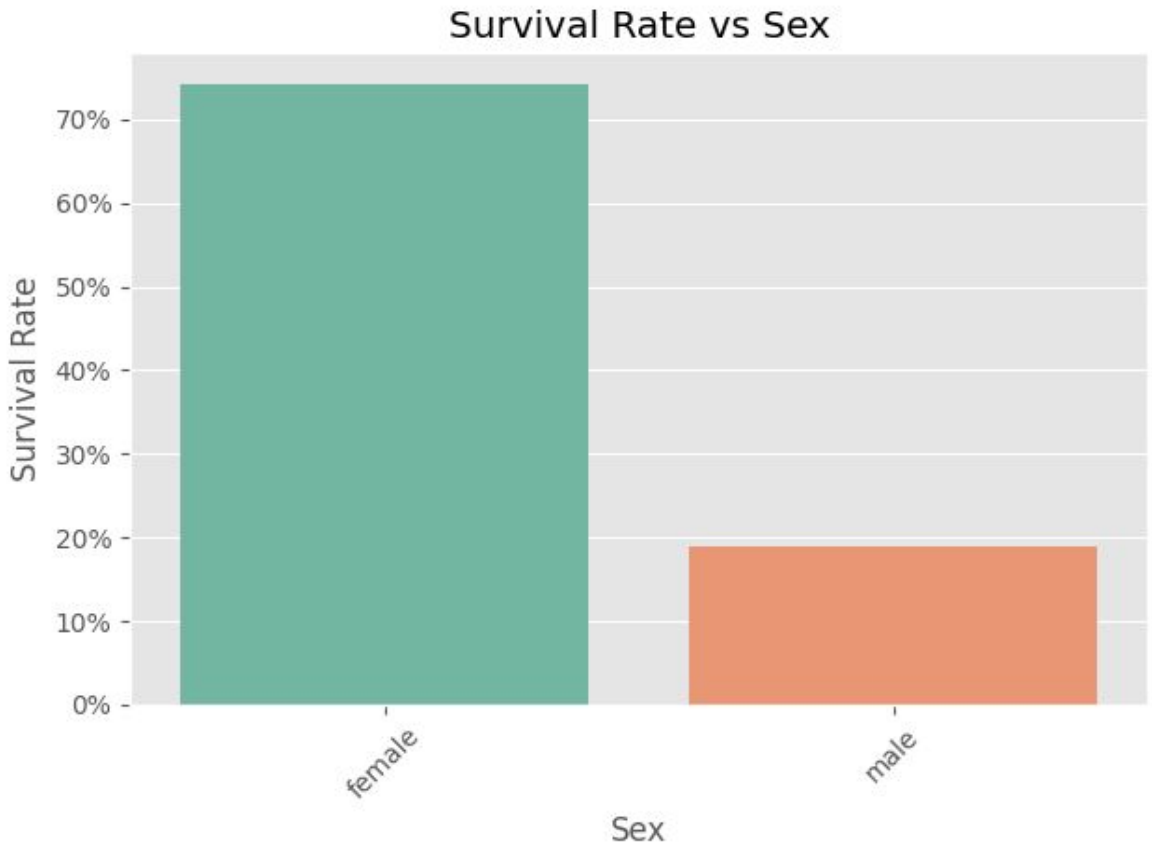
EXPLORATORY DATA ANALYSIS

TARGET DATA DISTRIBUTION

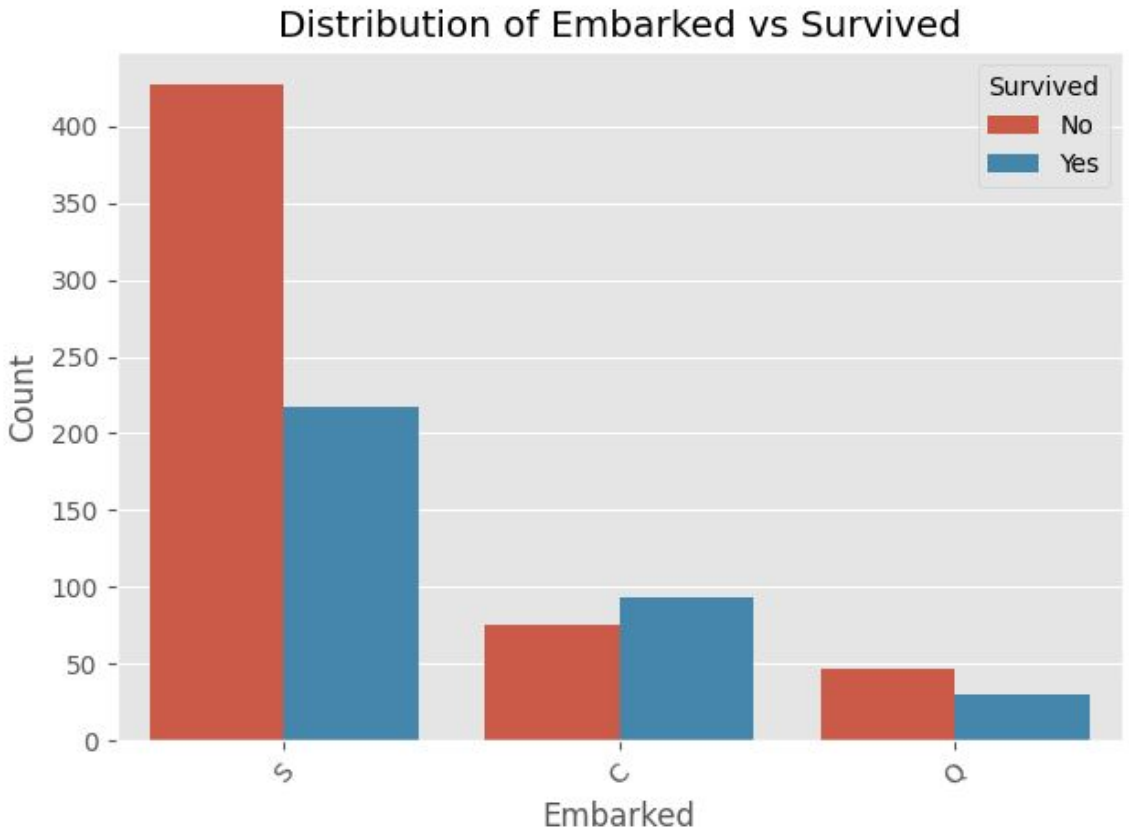


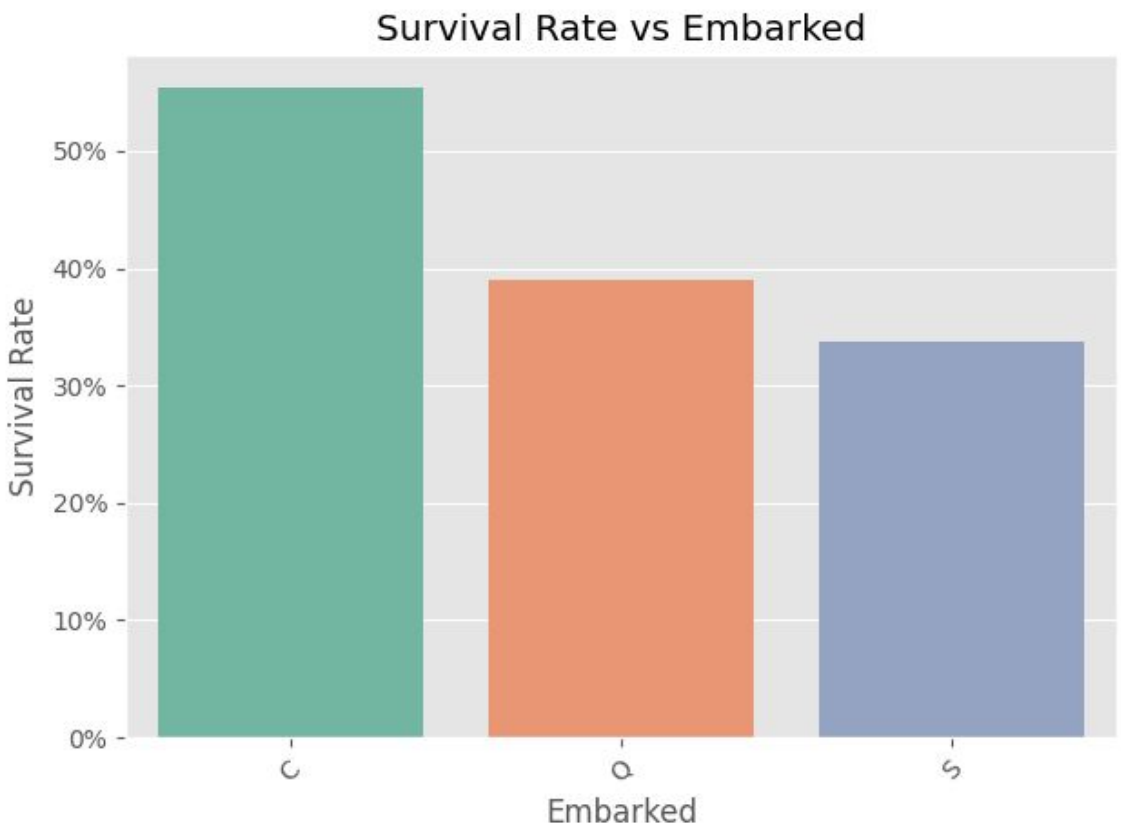
CATEGORICAL DATA DISTRIBUTION



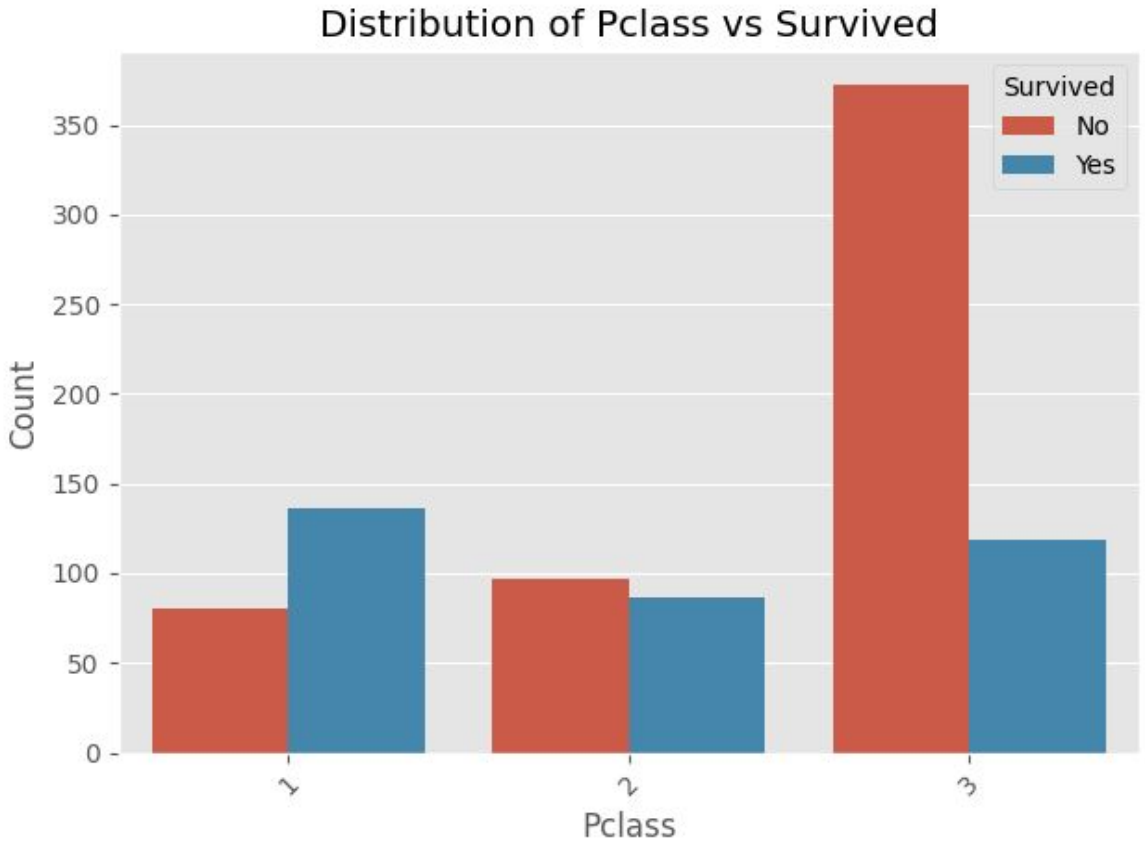


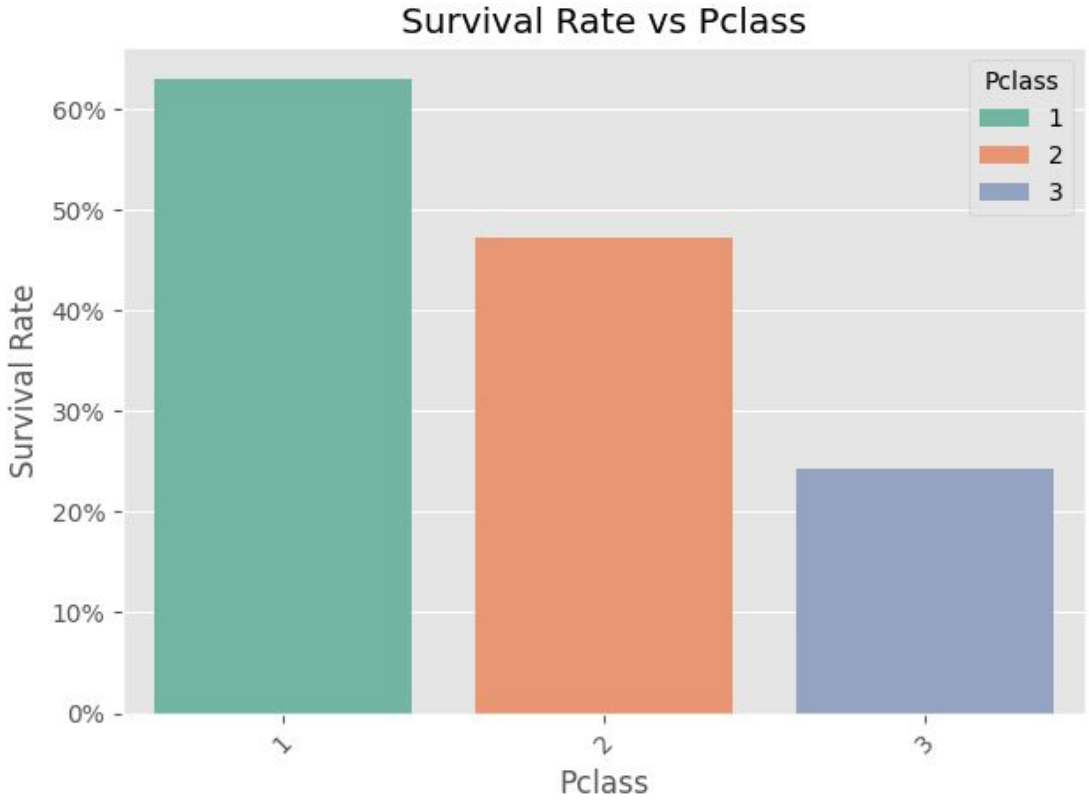
CATEGORICAL DATA DISTRIBUTION



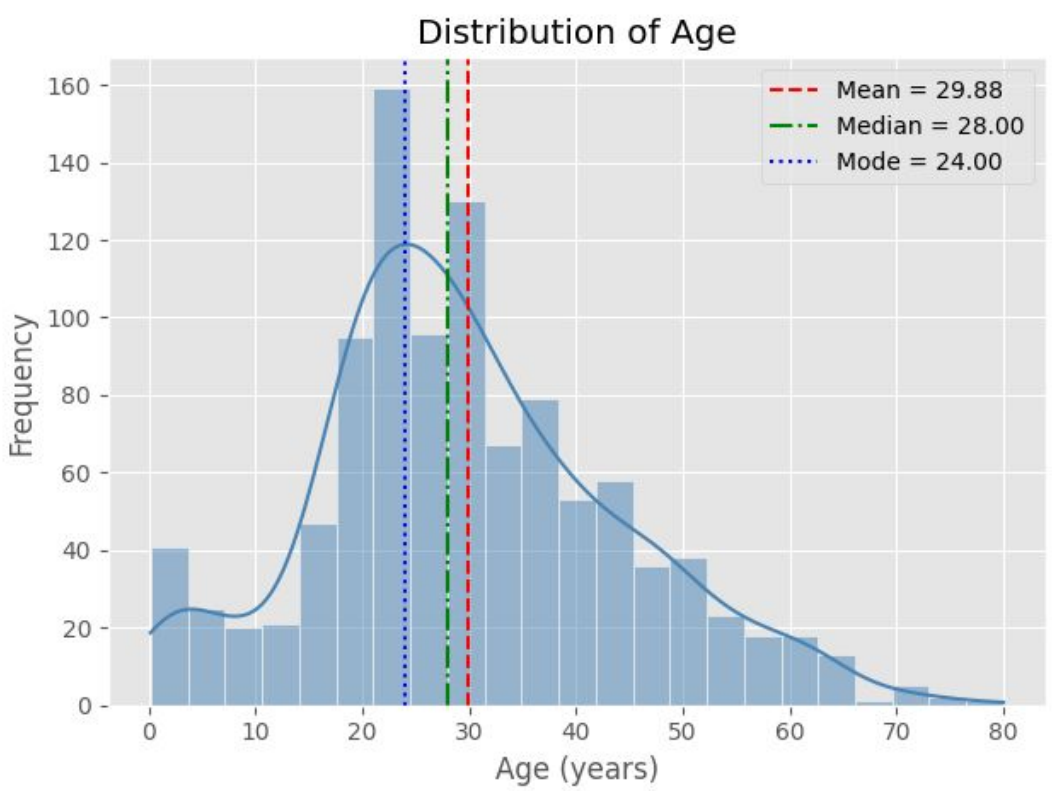


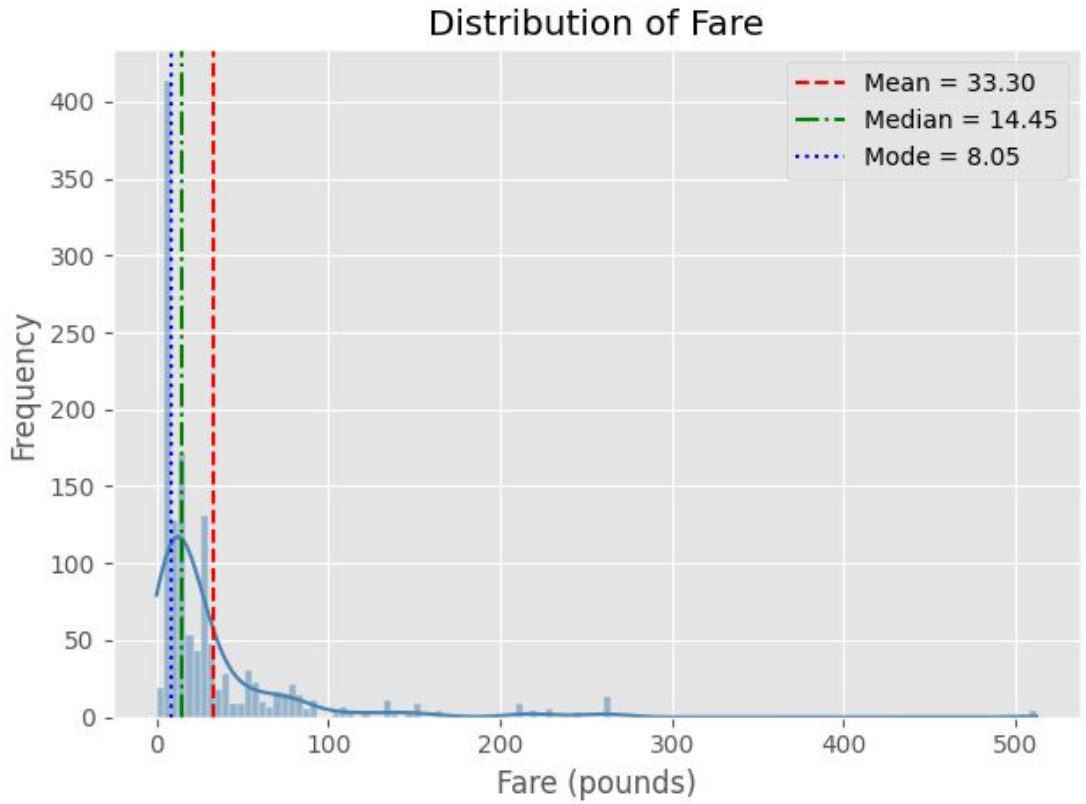
CATEGORICAL DATA DISTRIBUTION



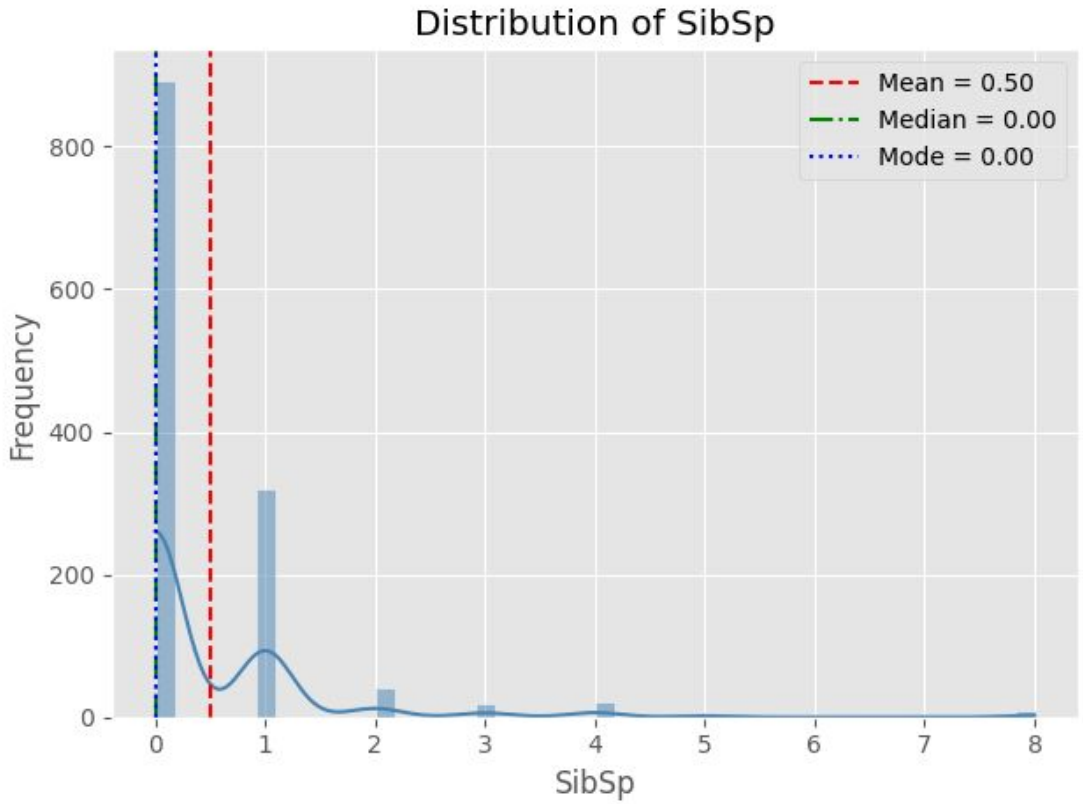


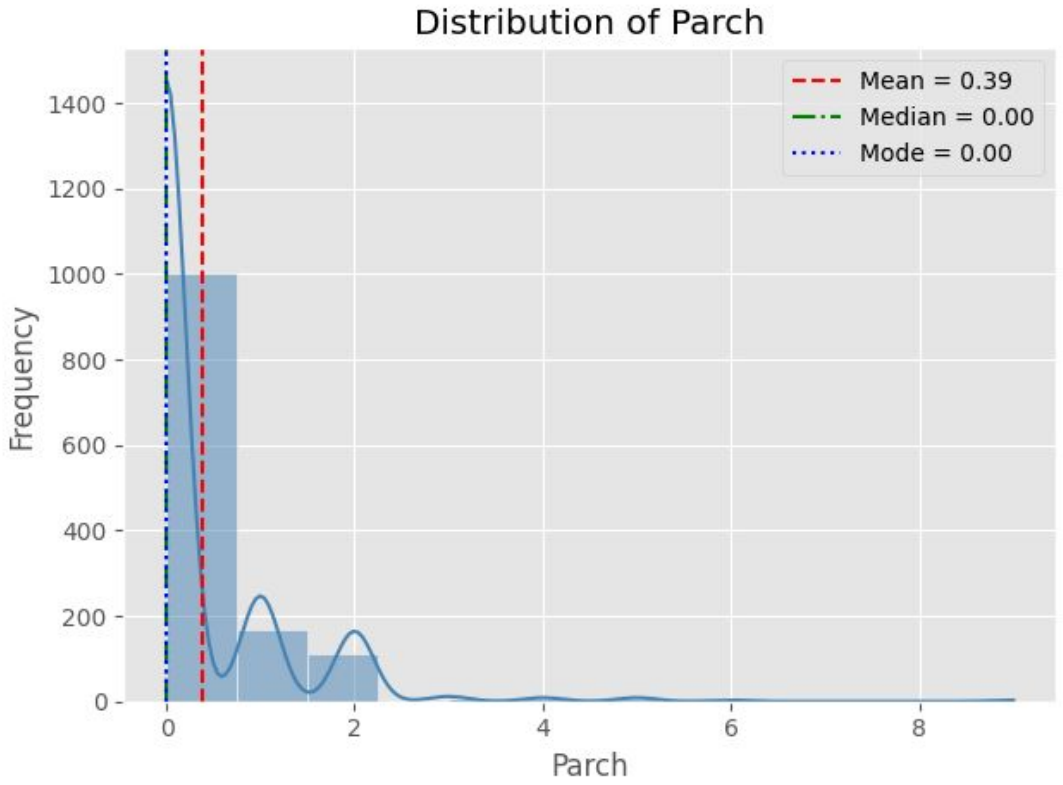
NUMERICAL DATA DISTRIBUTION



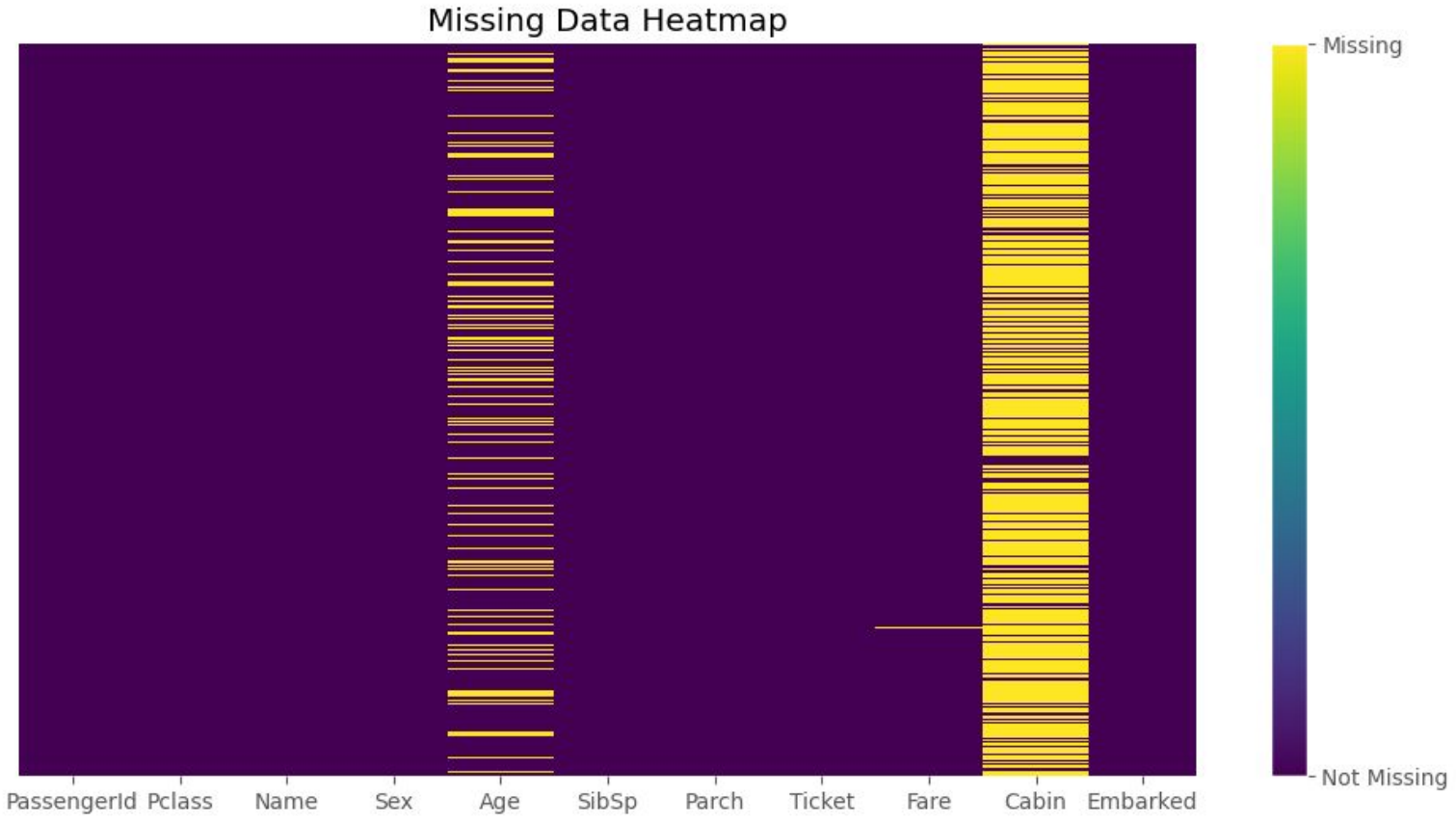


NUMERICAL DATA DISTRIBUTION





MISSING DATA



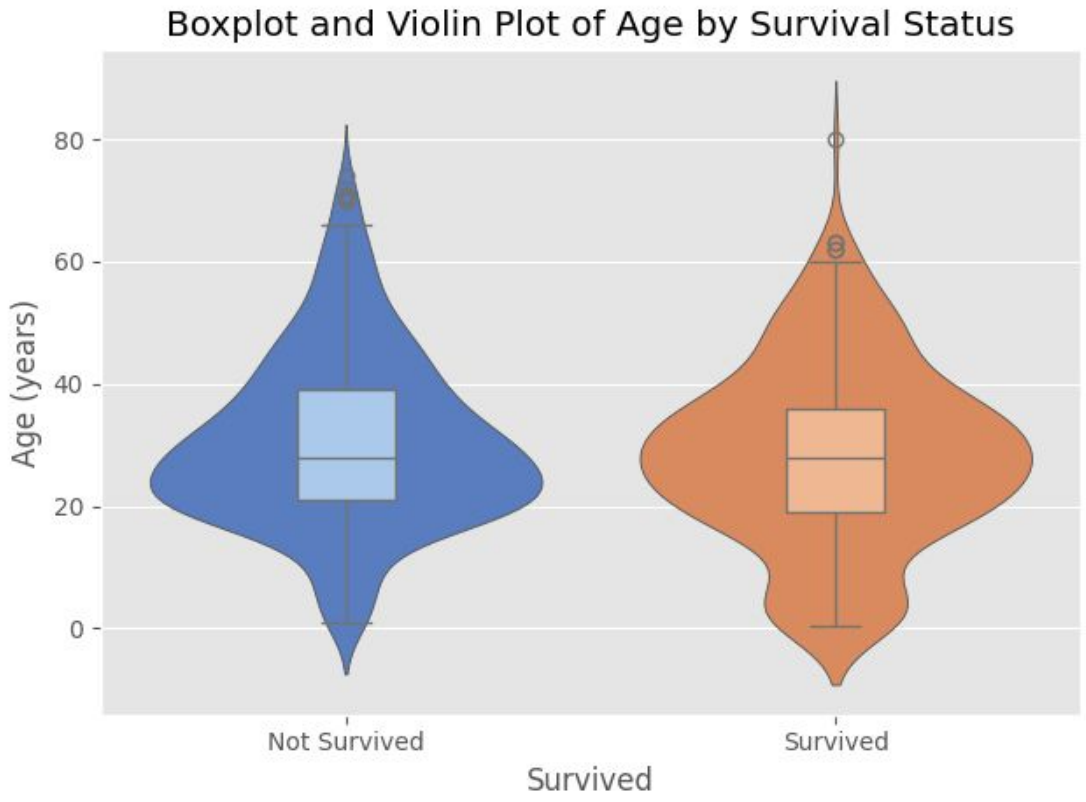
Cabin	1014
Age	263
Embarked	2
Fare	1

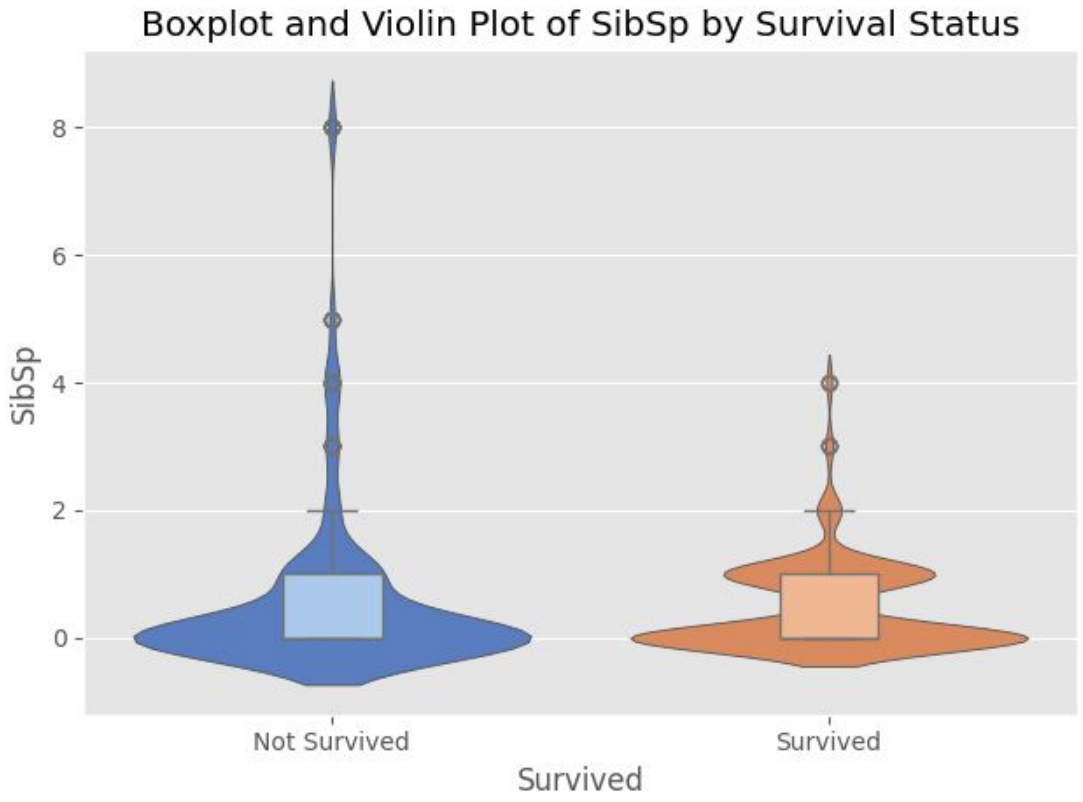
Median:
- Age
- Fare

Mode:
- Embarked

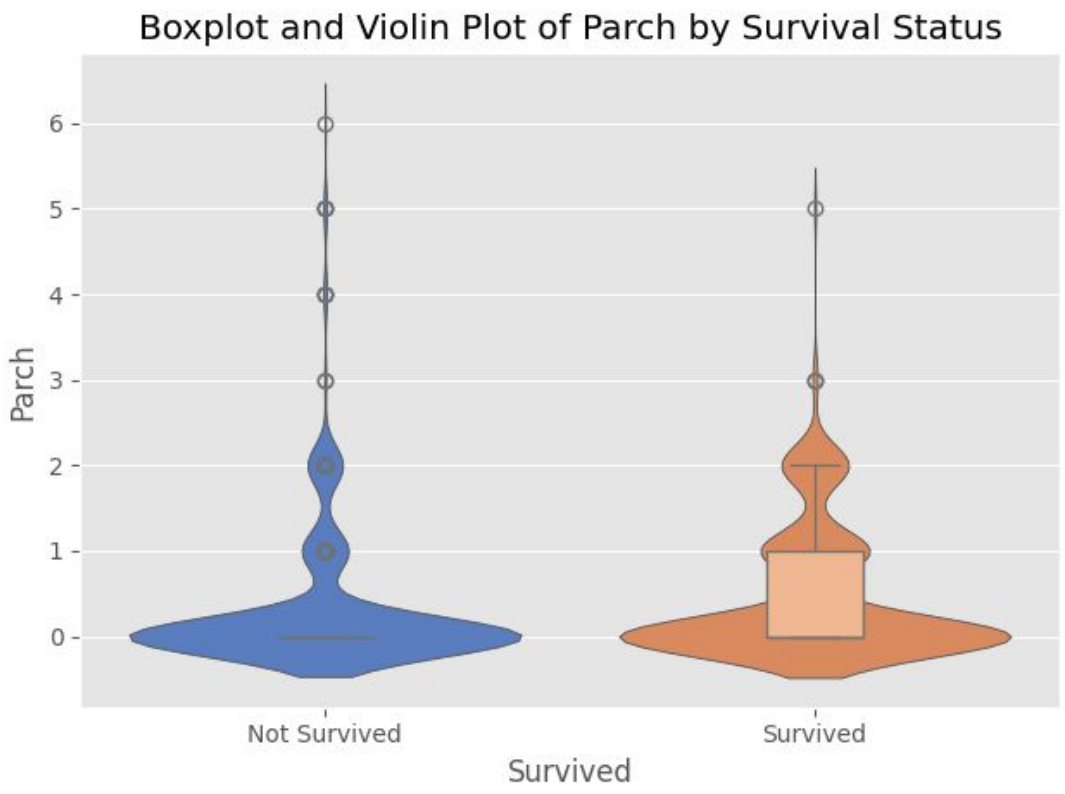
Remove:
- Cabin

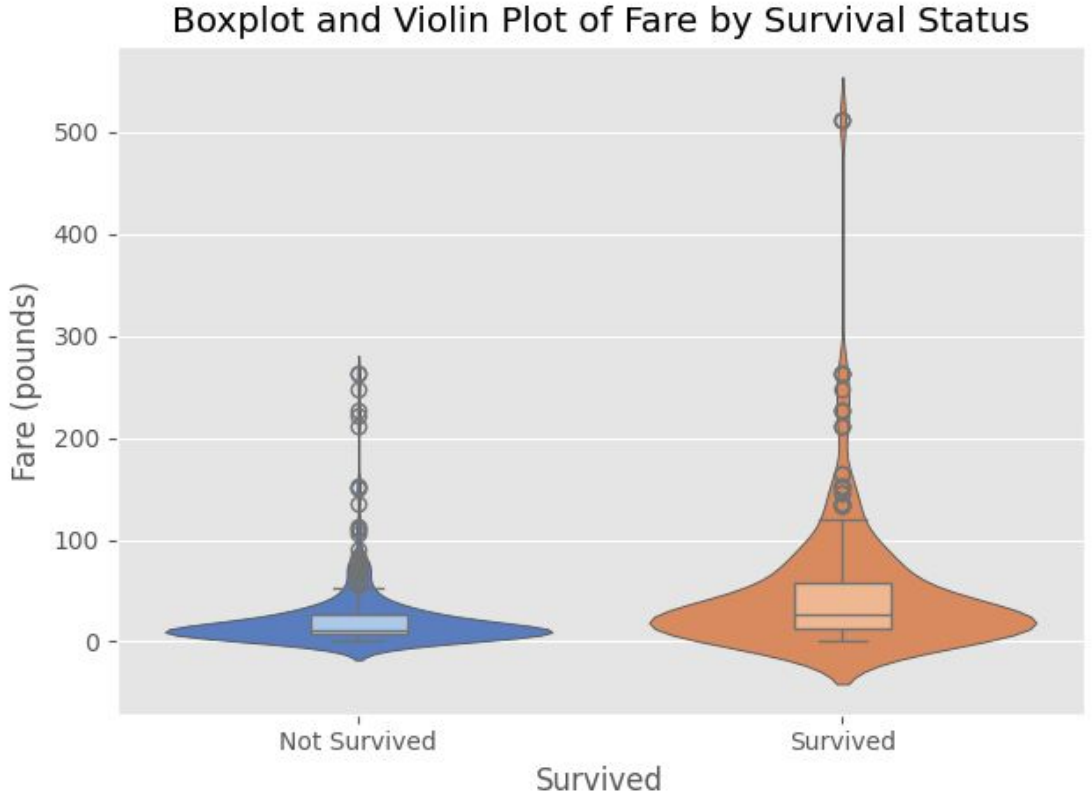
BOXPLOTS / VIOLIN PLOTS



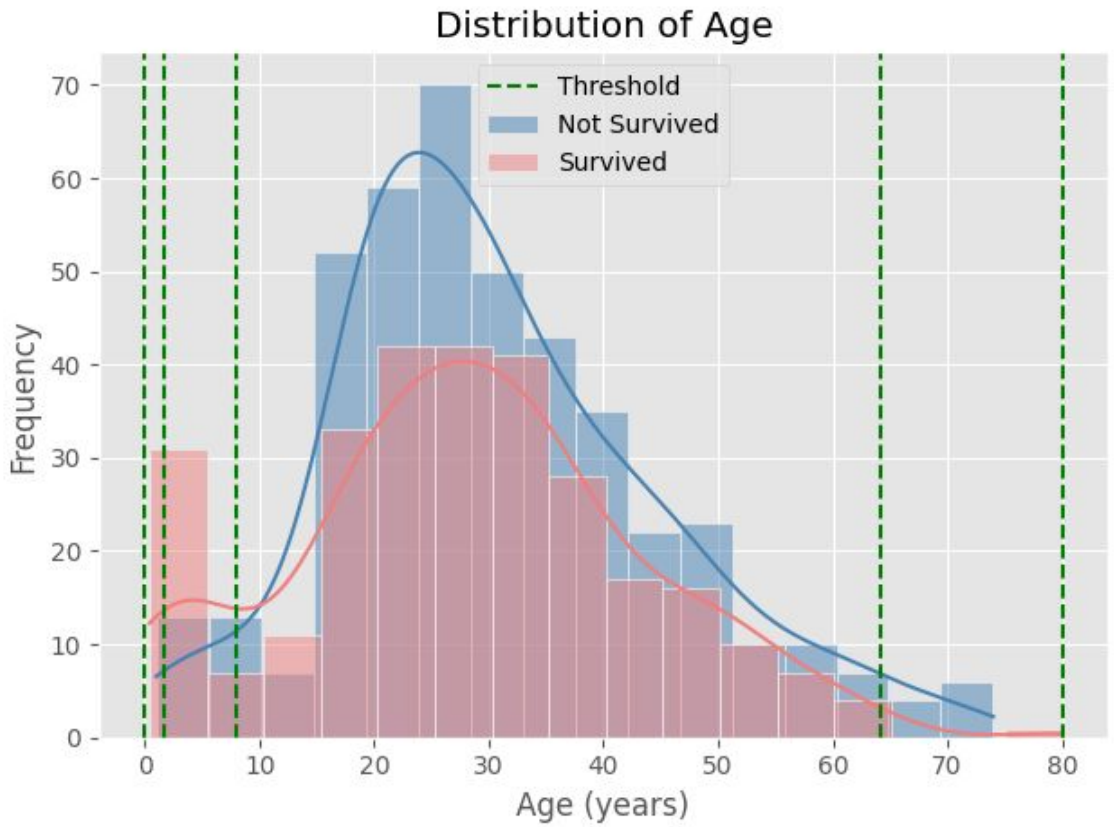


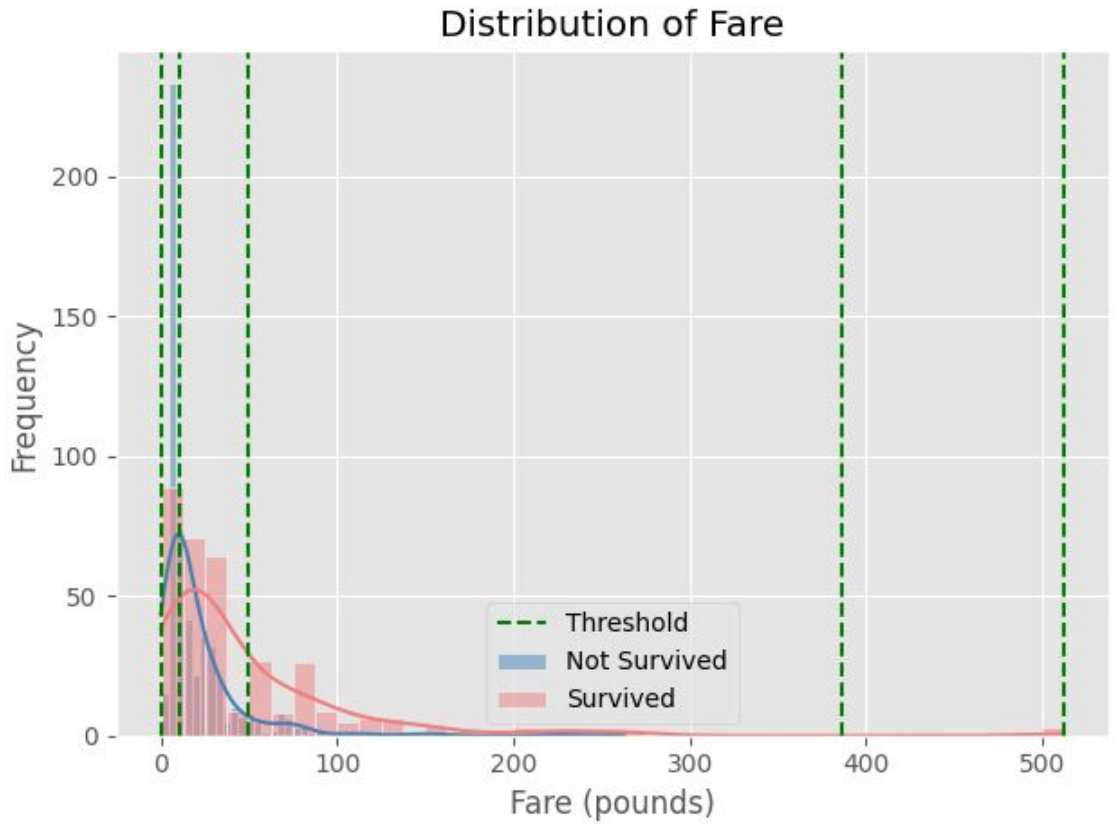
BOXPLOTS / VIOLIN PLOTS



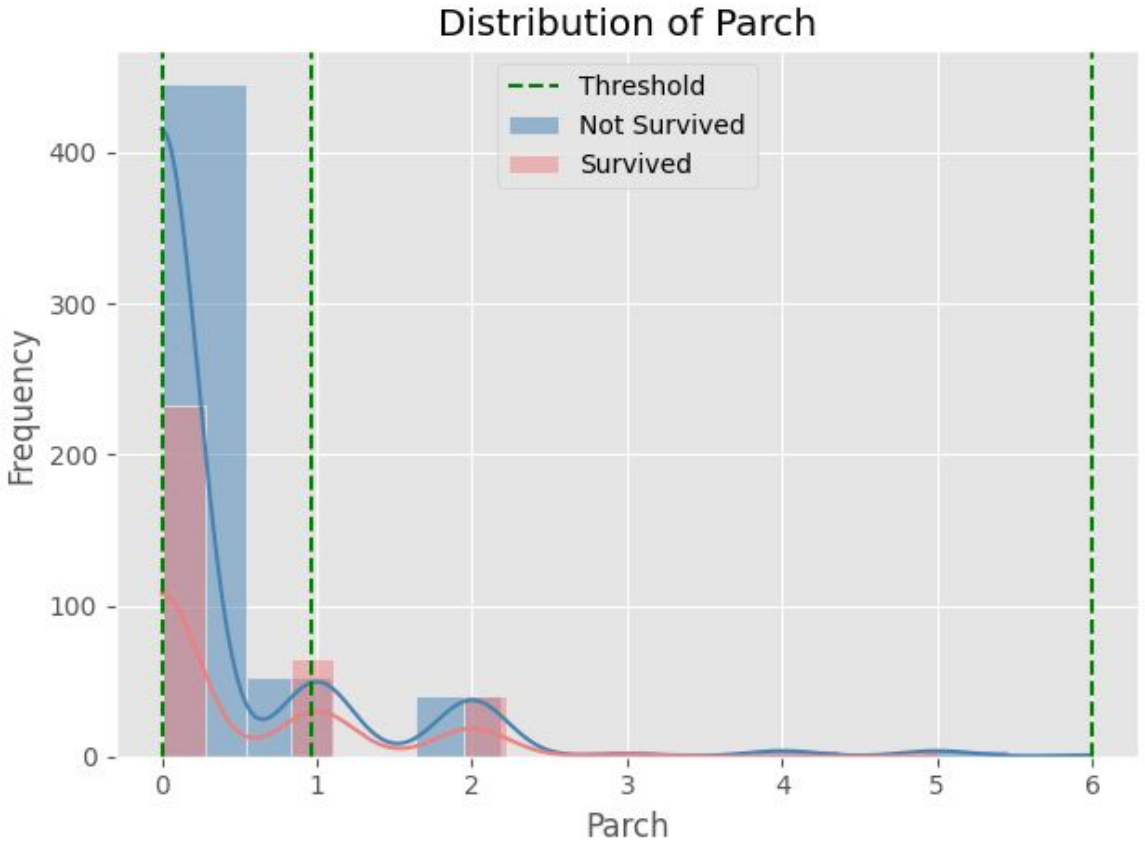


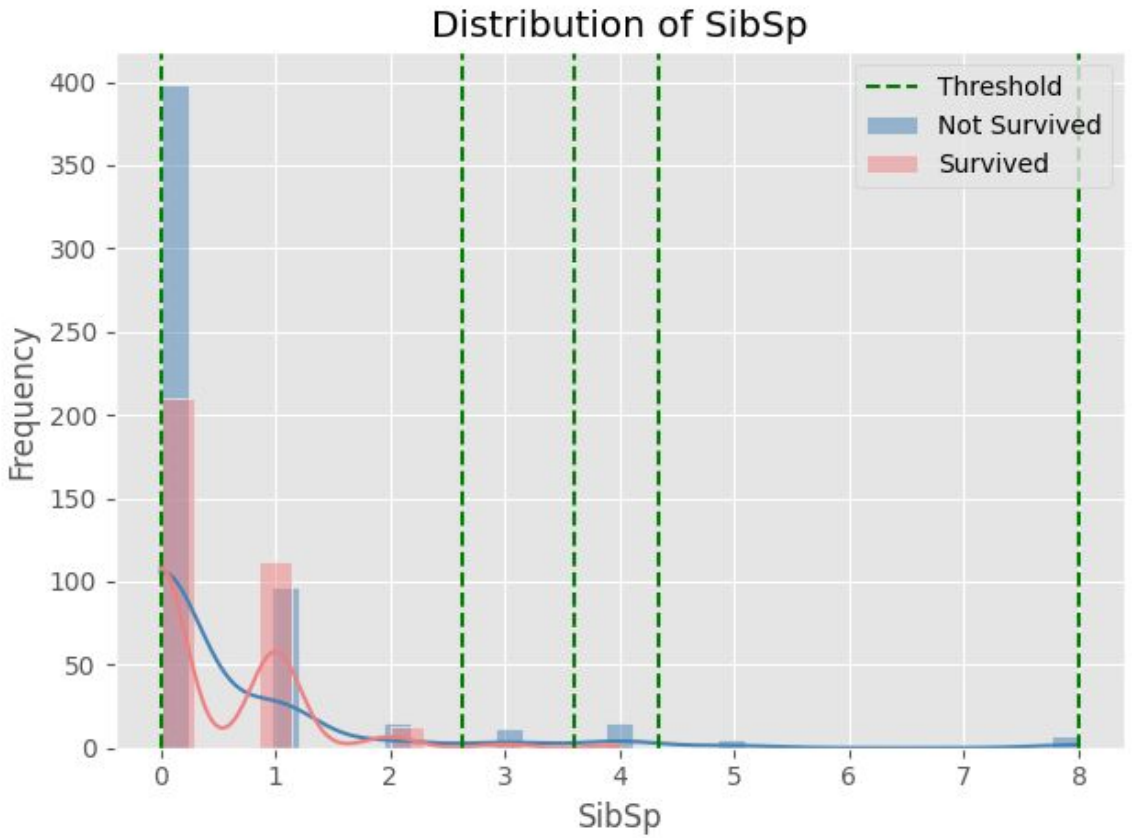
DISCRETIZATION



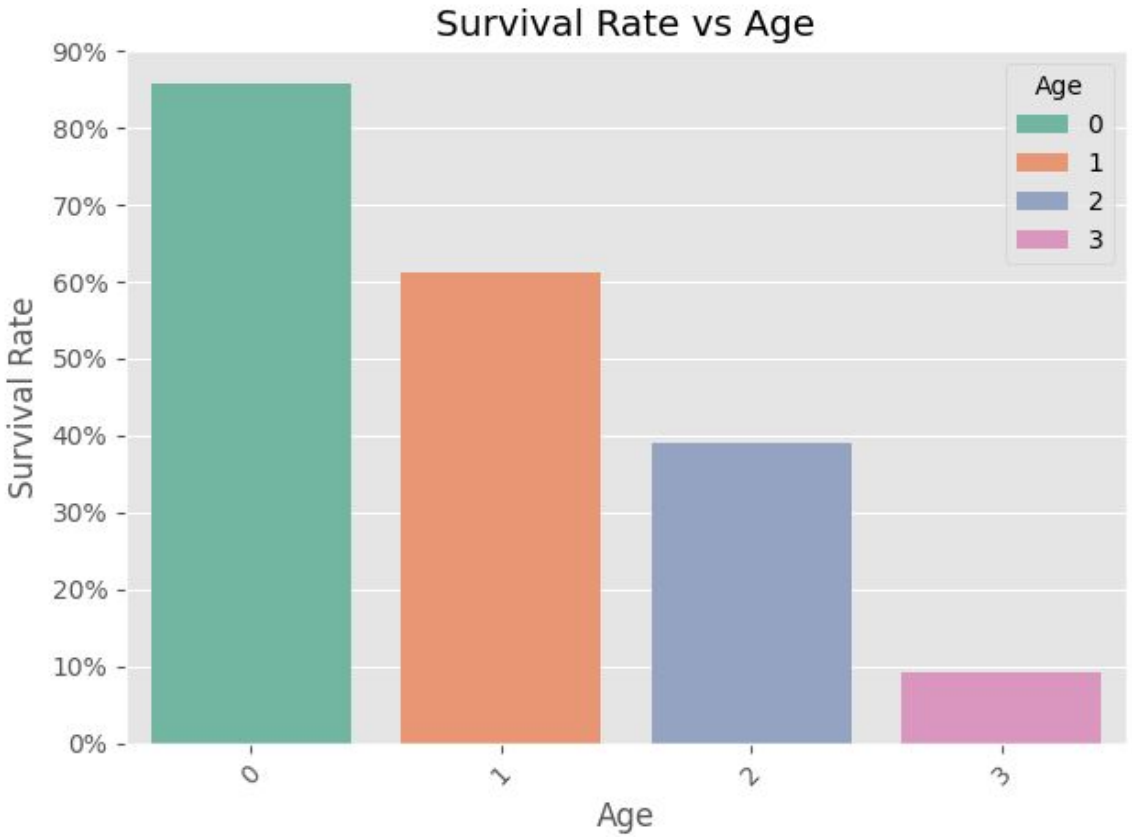


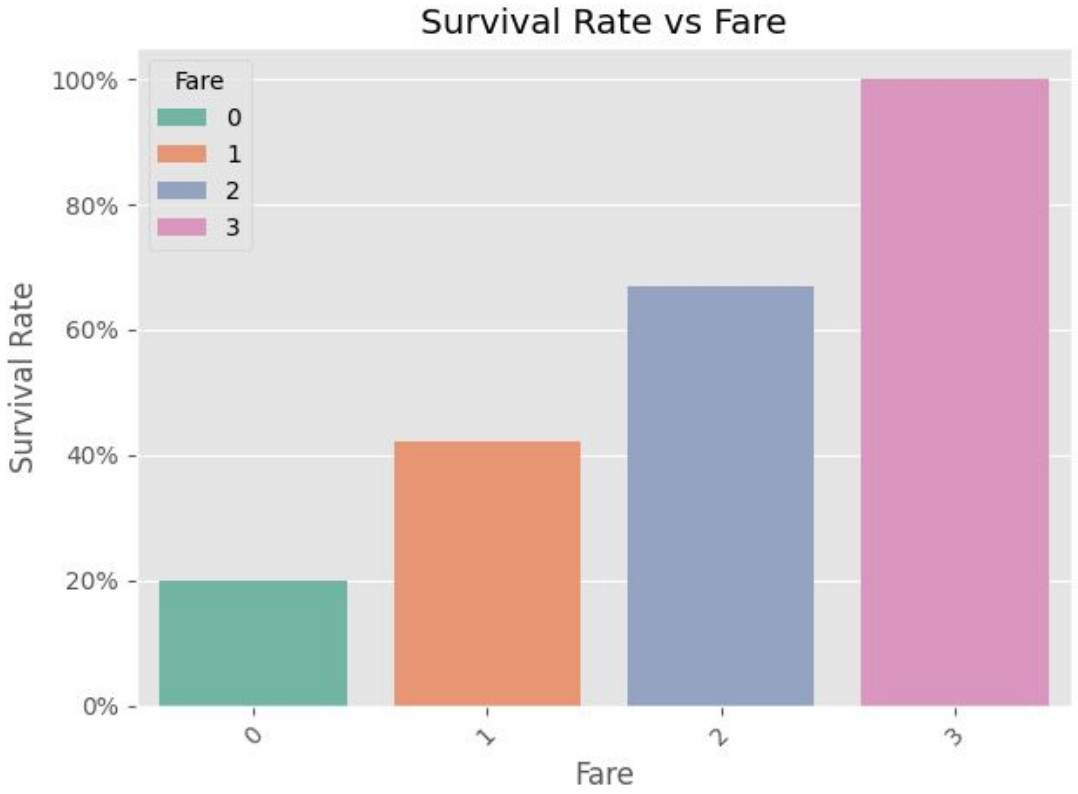
DISCRETIZATION



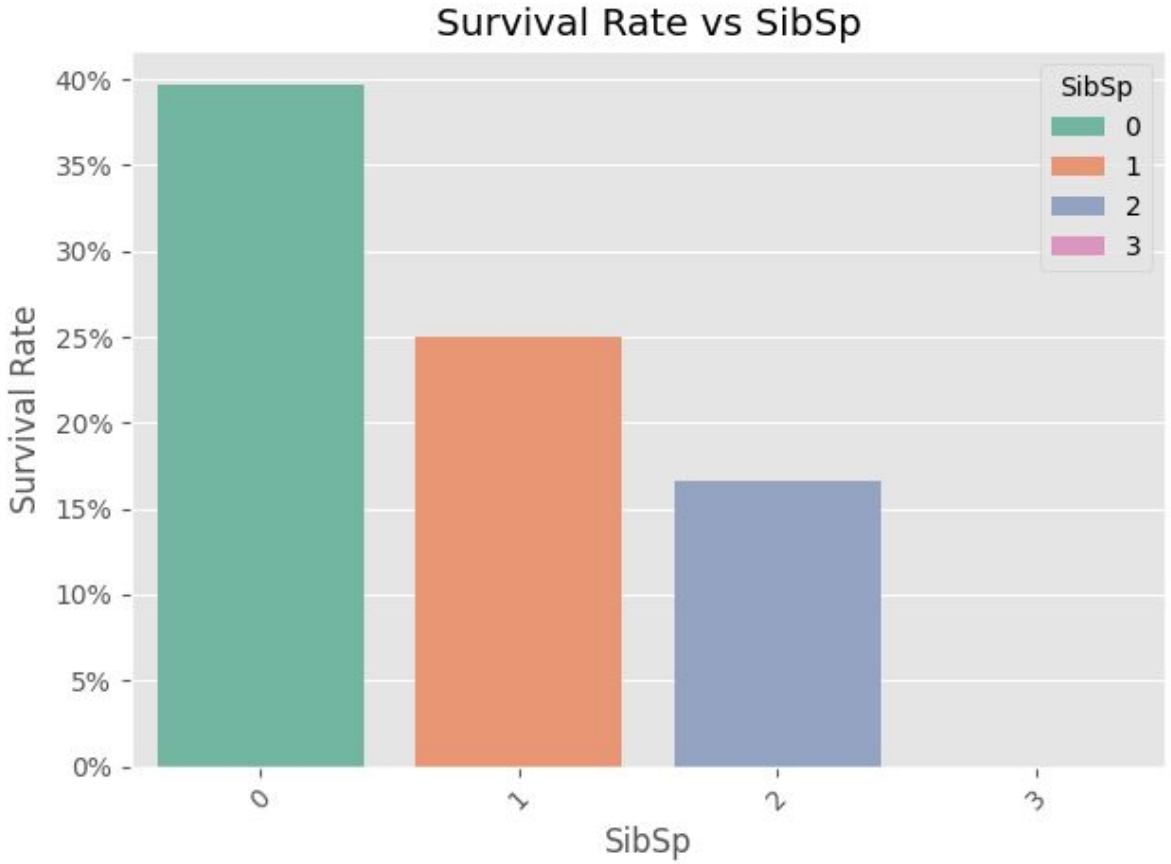


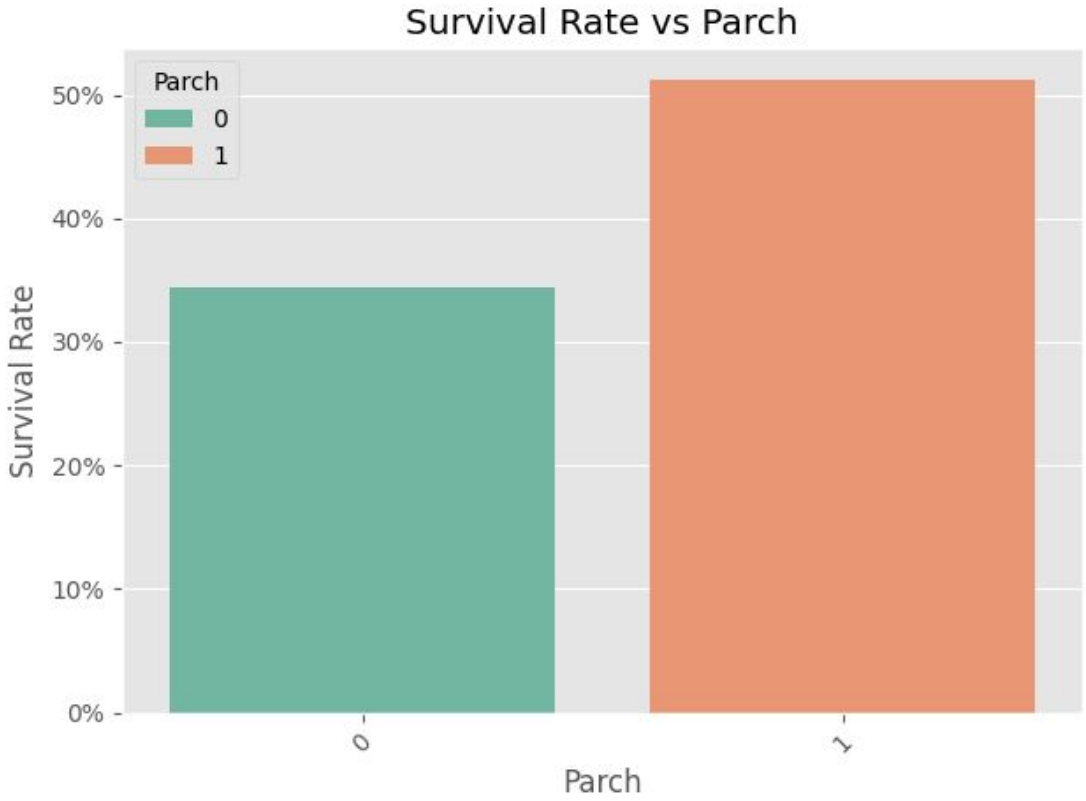
DISCRETIZATION RESULTS





DISCRETIZATION RESULTS



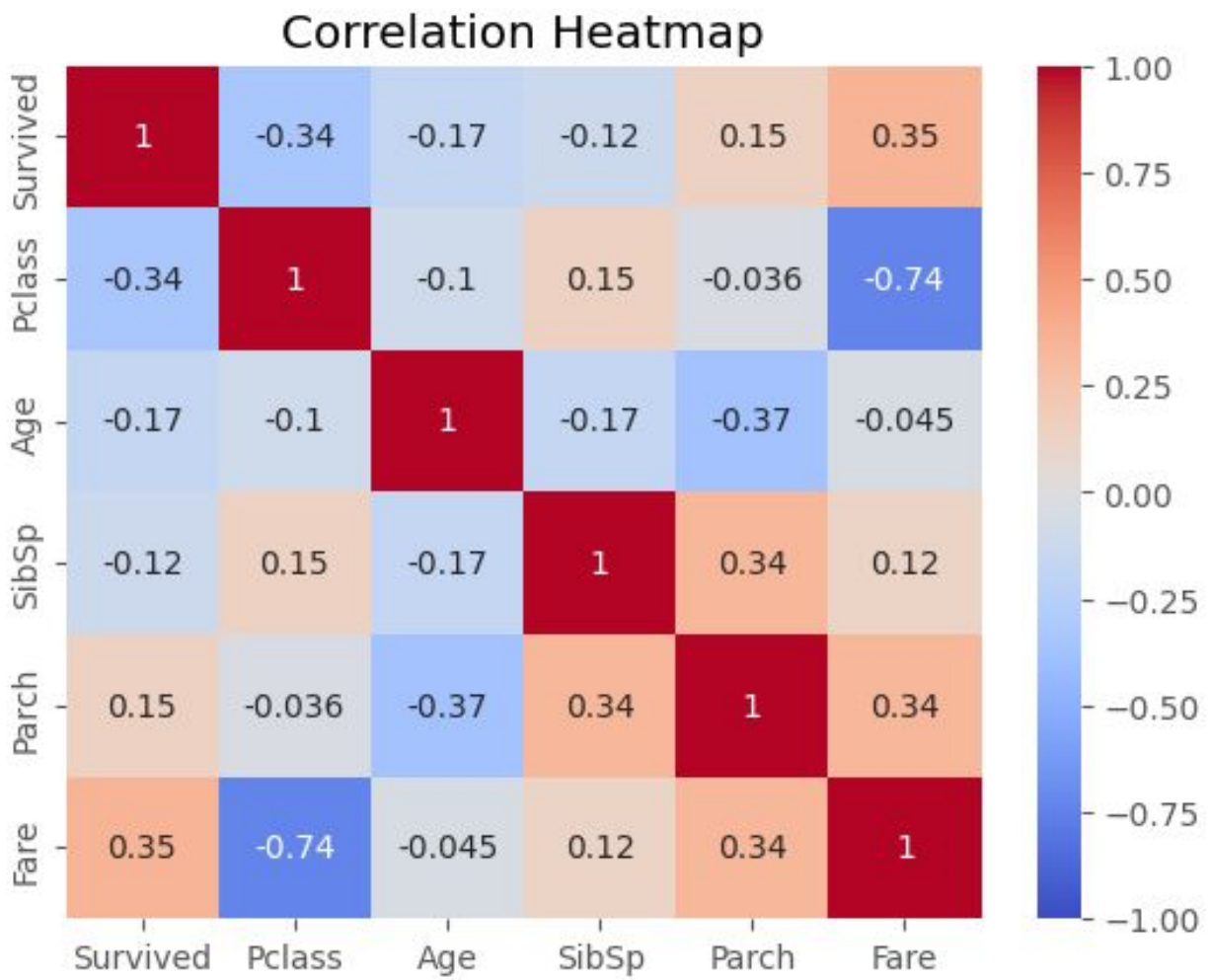


FEATURE MANIPULATION

NUMERICAL CORRELATION MATRIX

Delete:

- Parch
- Pclass



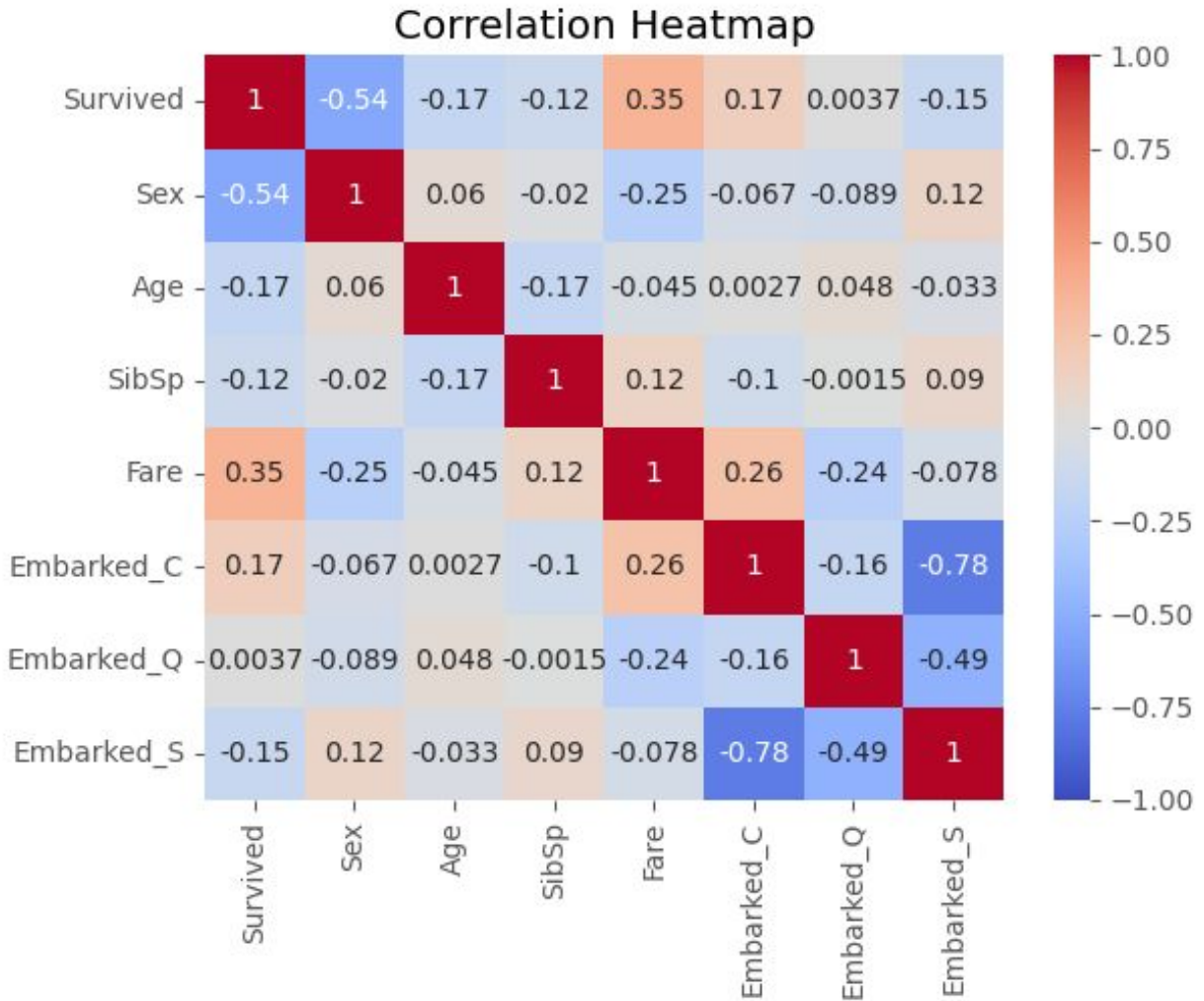
DATA ENCODING

1. **Delete:**
   - Ticket
   - Name

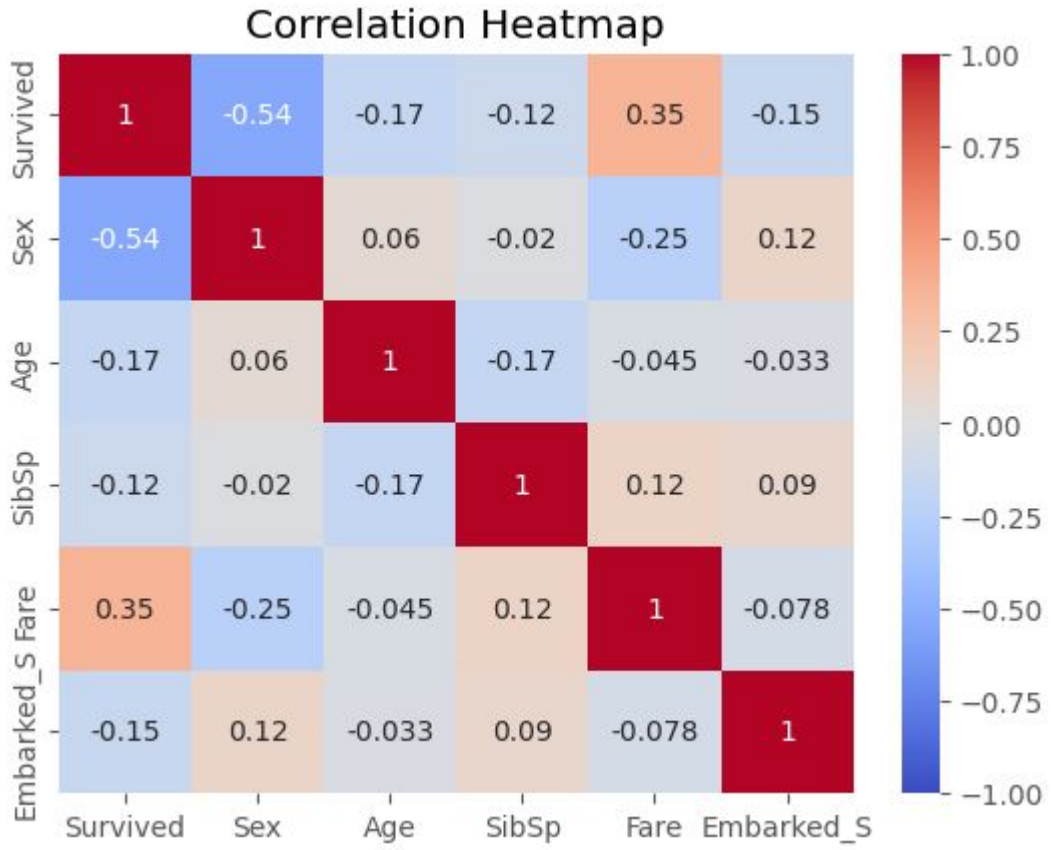
2. **One-Hot:**
   - Embarked

3. **Label:**
   - Sex

4. **Delete:**
   - Embarked_S
   - Embarked_Q

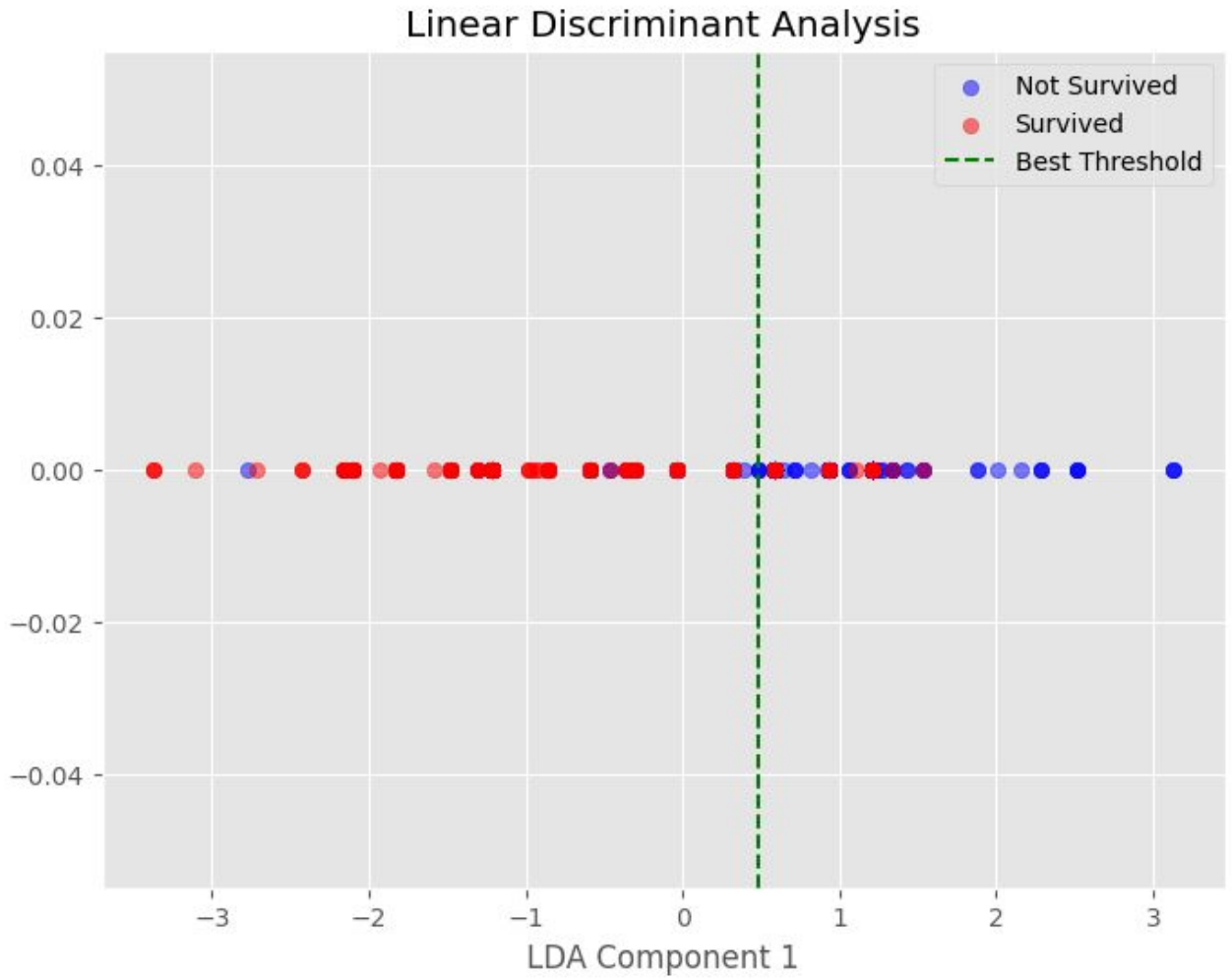


FINAL DATASET



Survived	Sex	Age	SibSp	Fare	Embarked_S
0	1	2	0	0	1
1	0	2	0	2	0
1	0	2	0	0	1
1	0	2	0	2	1
0	1	2	0	0	1
...	...	...	...	...	...
-1	1	2	0	0	1
-1	0	2	0	2	0
-1	1	2	0	0	1
-1	1	2	0	0	1
-1	1	2	0	1	0

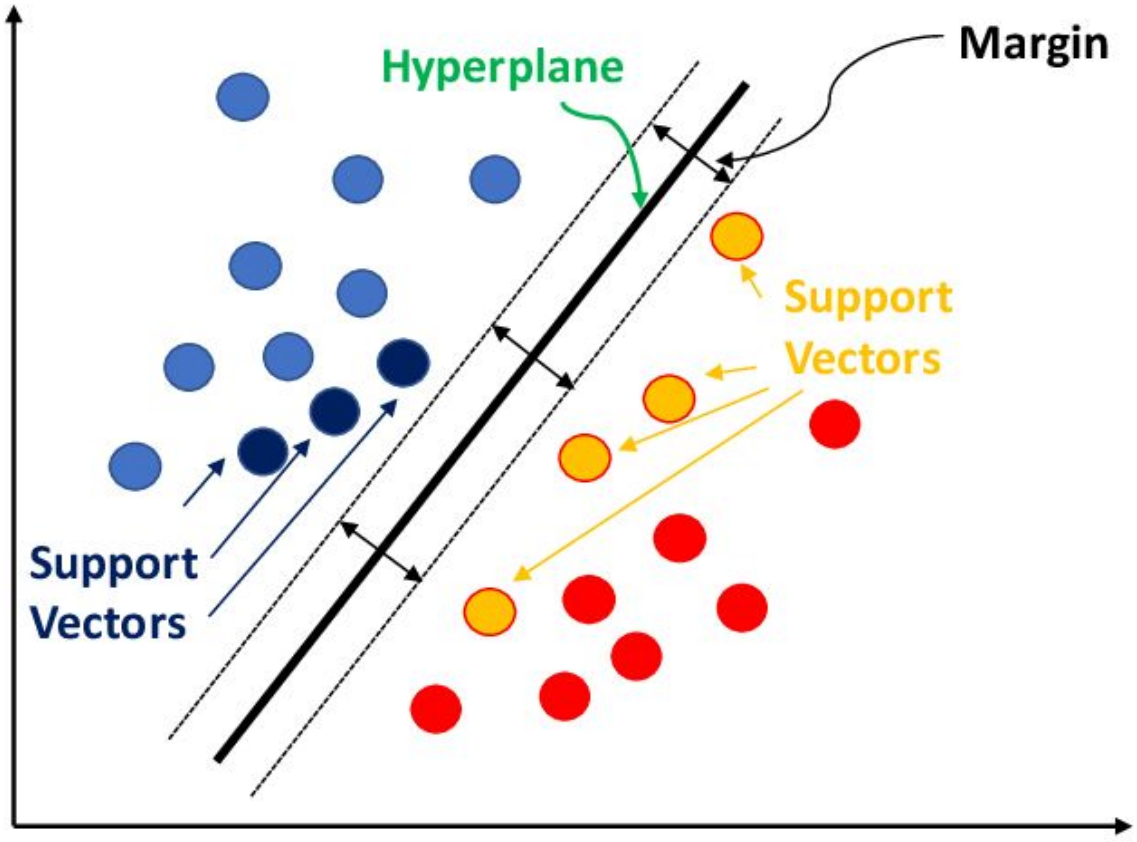
FINAL ANALYSIS



Mean F1 score: 0.7383
F1 score standard deviation: 0.032

MODEL SELECTION

SUPPORT VECTOR MACHINE

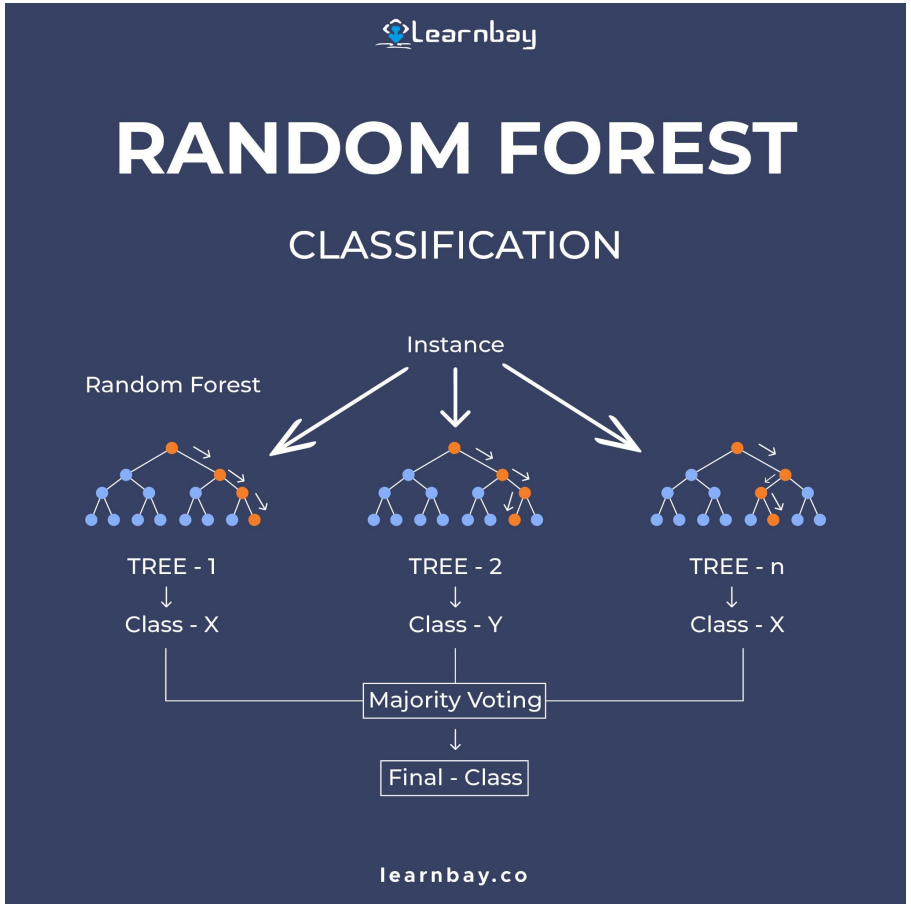


Reasons:

- linearly separable data
- Effective with non collinear data
- Works well with unbalanced classes

Technologies, D. (2022, 20 september). *What is a Support Vector Machine? - Datatron*. Datatron. https://datatron.com/what-is-a-support-vector-machine/

RANDOM FOREST

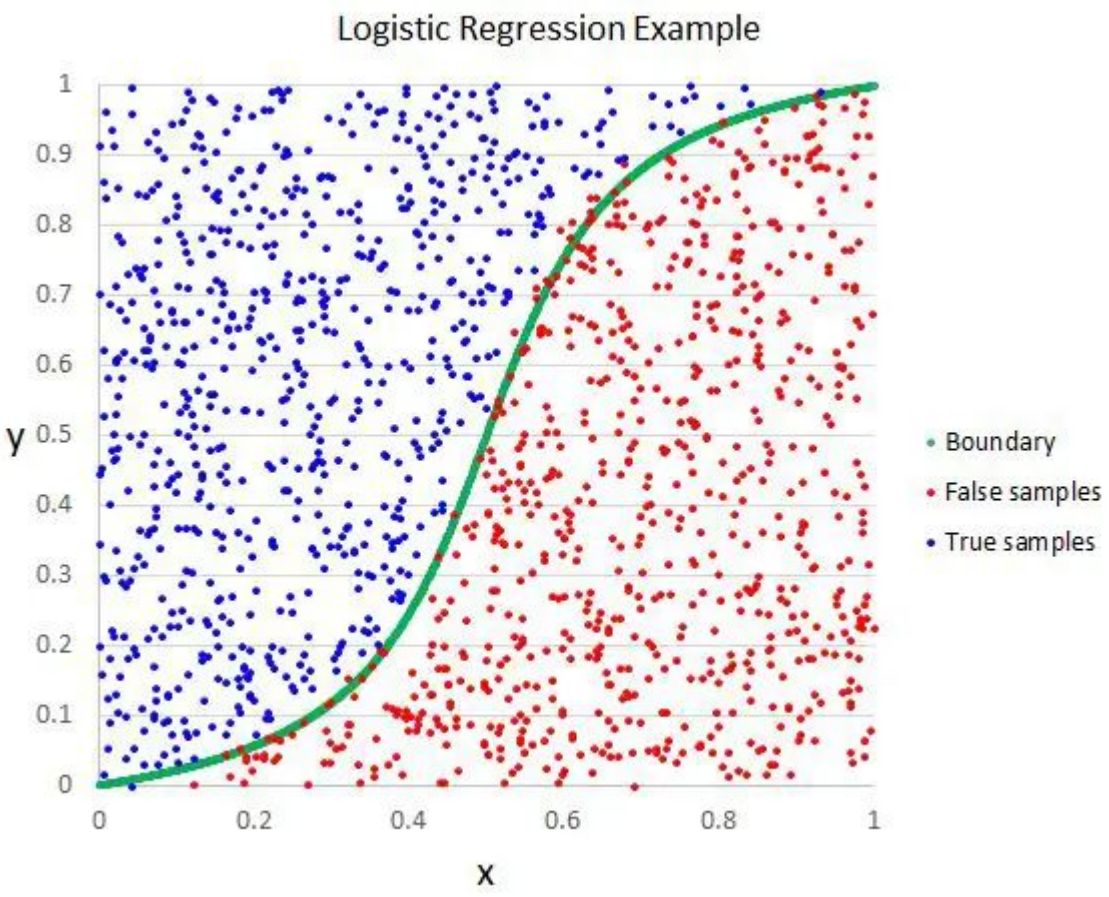


Reasons:

- Non-linear relationships
- Great for categorical and ordinal data
- Feature Importance

Cotten, S. (2022, 28 august). *Random Forest Classification*. Pinterest. https://pt.pinterest.com/pin/848787861035795075/

LOGISTIC REGRESSION



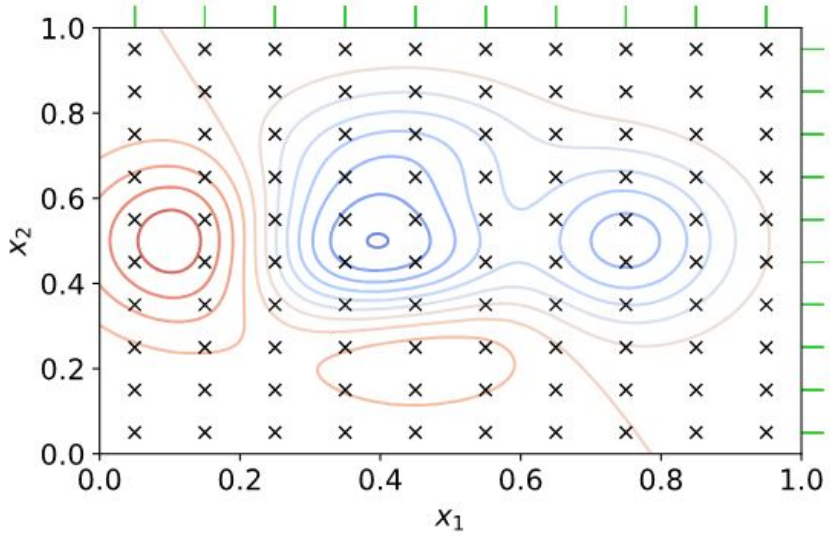
Reasons:

- Simplicity and probabilistic Interpretability
- Great for binary and discrete dataset
- No multicollinearity issues

Chakure, A. (2023, 22 diciembre). *What is logistic regression?* Built In. https://builtin.com/data-science/what-is-logistic-regression

MODEL TESTING

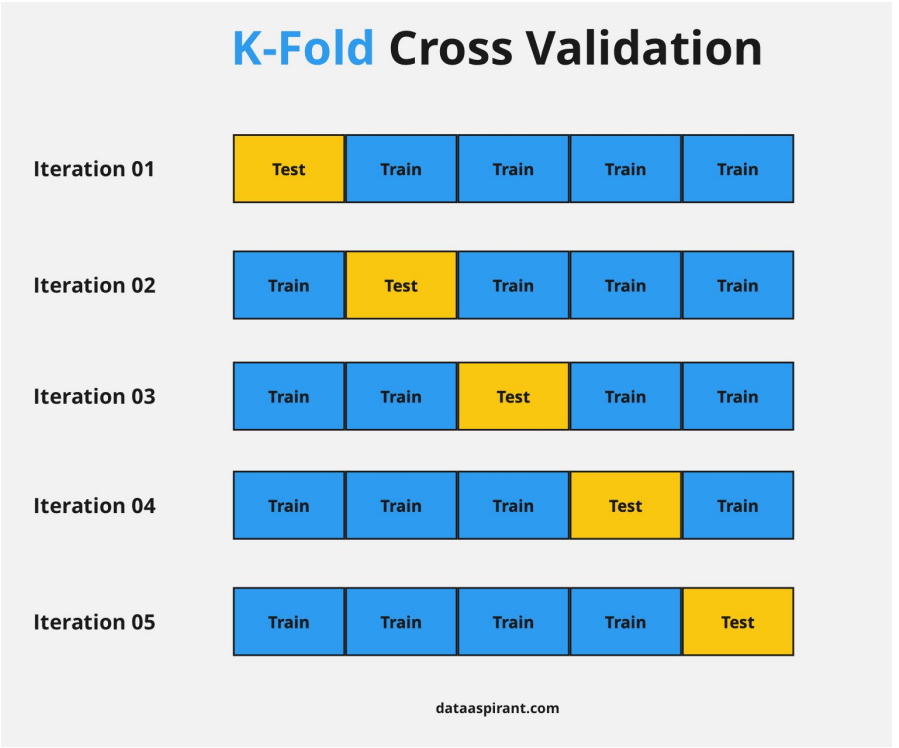
MODEL TESTING



Best parameters for SVM: {'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}

Best parameters for Random Forest: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200}

Best parameters for Logistic Regression: {'C': 1, 'penalty': 'l2', 'solver': 'saga'}



colaboradores de Wikipedia. (s. f.). *Archivo:Hyperparameter Optimization using Grid Search.svg - Wikipedia, la enciclopedia libre*. https://es.m.wikipedia.org/wiki/Archivo:Hyperparameter_Optimization_using_Grid_Search.svg

Arif, A., & Arif, A. (2023, 19 octubre). How Cross-Validation works in Machine Learning - Dataaspirant. *Dataaspirant - A Data Science Portal For Beginners*. https://dataaspirant.com/cross-validation/

GENERAL METRICS

Support Vector Machine

Metric	Mean	Std Dev
F1 Score	0.742523	0.056246
Accuracy	0.814826	0.031705
Recall	0.707486	0.050742
Specificity	0.880248	0.026977

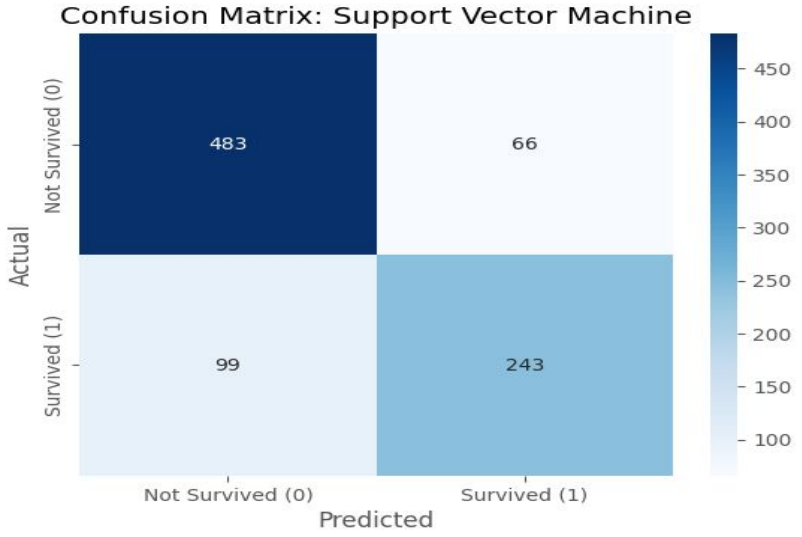
Random Forest

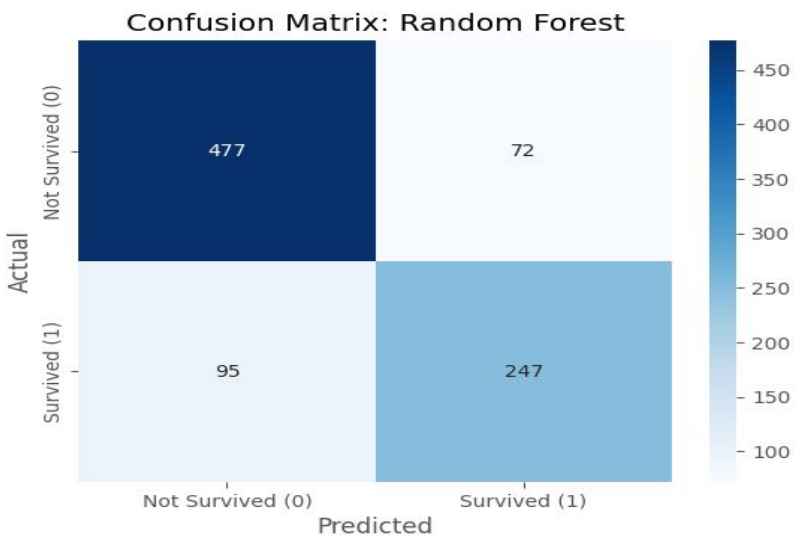
Metric	Mean	Std Dev
F1 Score	0.742810	0.046506
Accuracy	0.812573	0.020256
Recall	0.719153	0.051804
Specificity	0.869339	0.013334

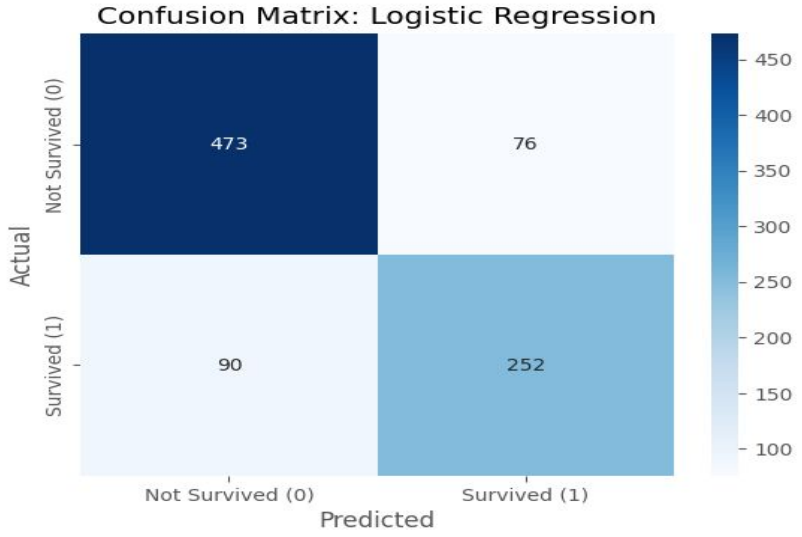
Logistic Regression

Metric	Mean	Std Dev
F1 Score	0.748503	0.042826
Accuracy	0.813690	0.019607
Recall	0.735077	0.046833
Specificity	0.862383	0.019189

CONFUSION MATRICES







BEST MODEL

LOGISTIC REGRESSION

Pros:

- Good mean F1 score (0.75)
- Good mean accuracy (0.81)
- Lower standard deviation for most metrics
- Higher mean recall
- Higher mean negative predictive value
- Lower mean miss rate
- Easy interpretation
- Fast training

Cons:

- Lower precision
- Higher fallout
- Lower specificity

Results:

- Kaggle Score: 0.7751
- ROC AUC = 0.87

PYSPARK COMPARISON

PySpark Training time:
3.85 seconds

SciKit Learn Training time: 0.03 seconds

Too small dataset

CONCLUSION

General objective:

- Predict Titanic survival of Kaggle dataset
- Recall ≥ 0.7
- Specificity ≥ 0.85

How to do it?:

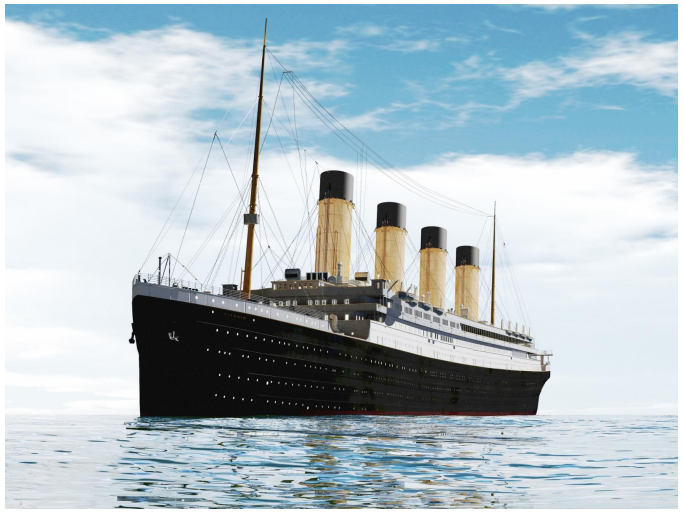
- Analyse the data
- Fill the missing data
- Select useful features
- Select correct algorithms
- Train the models
- Evaluate the model

Results:

- Transformed dataset
- Logistic regression model
- Mean F1 score: 0.74
- Mean Specificity: 0.86
- Mean Recall: 0.73
- Kaggle Score: 0.7751

How to improve?

- More exhaustive analysis
- Better feature engineering
- Better hyperparameter tuning



Q&A

THANK YOU

JESÚS ALEXANDER MEISTER CAREAGA A01656699

IKER S. BALI ELIZALDE A01656437

MICHELLE AGUIRRE MARTÍNEZ A01661592

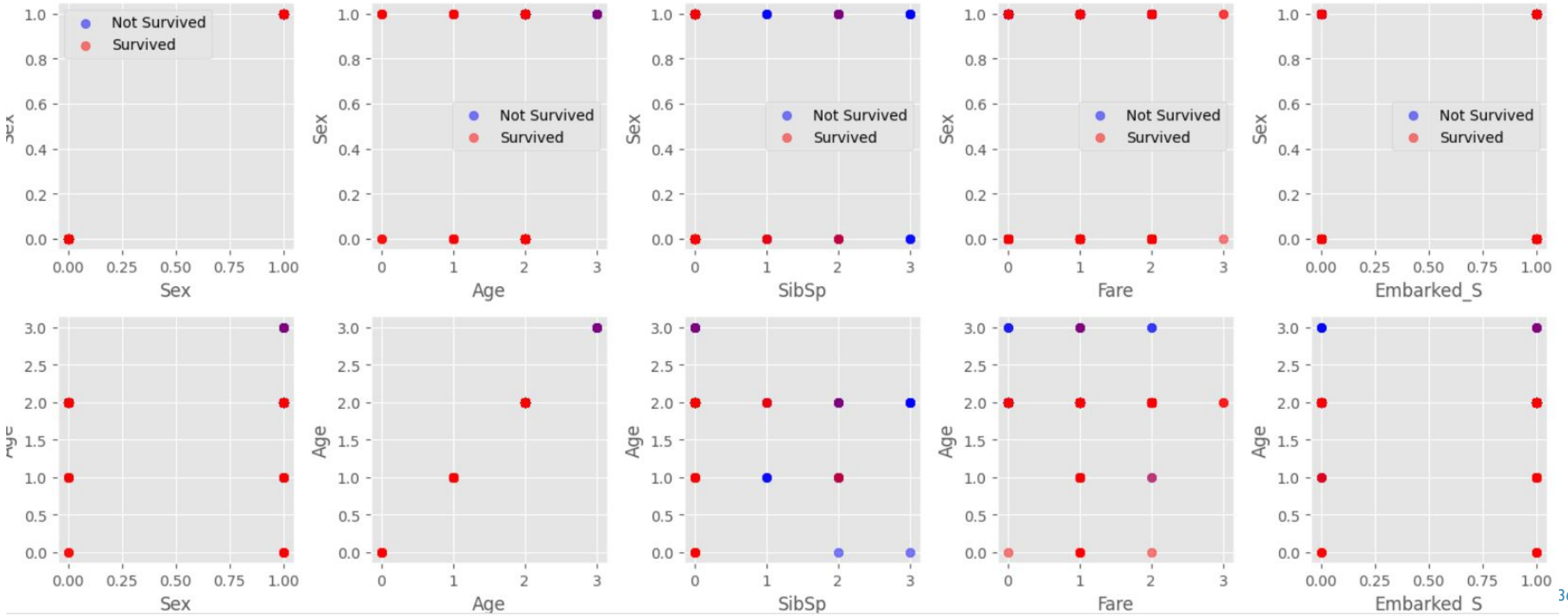
DIEGO SÁNCHEZ HERNÁNDEZ

A01783237

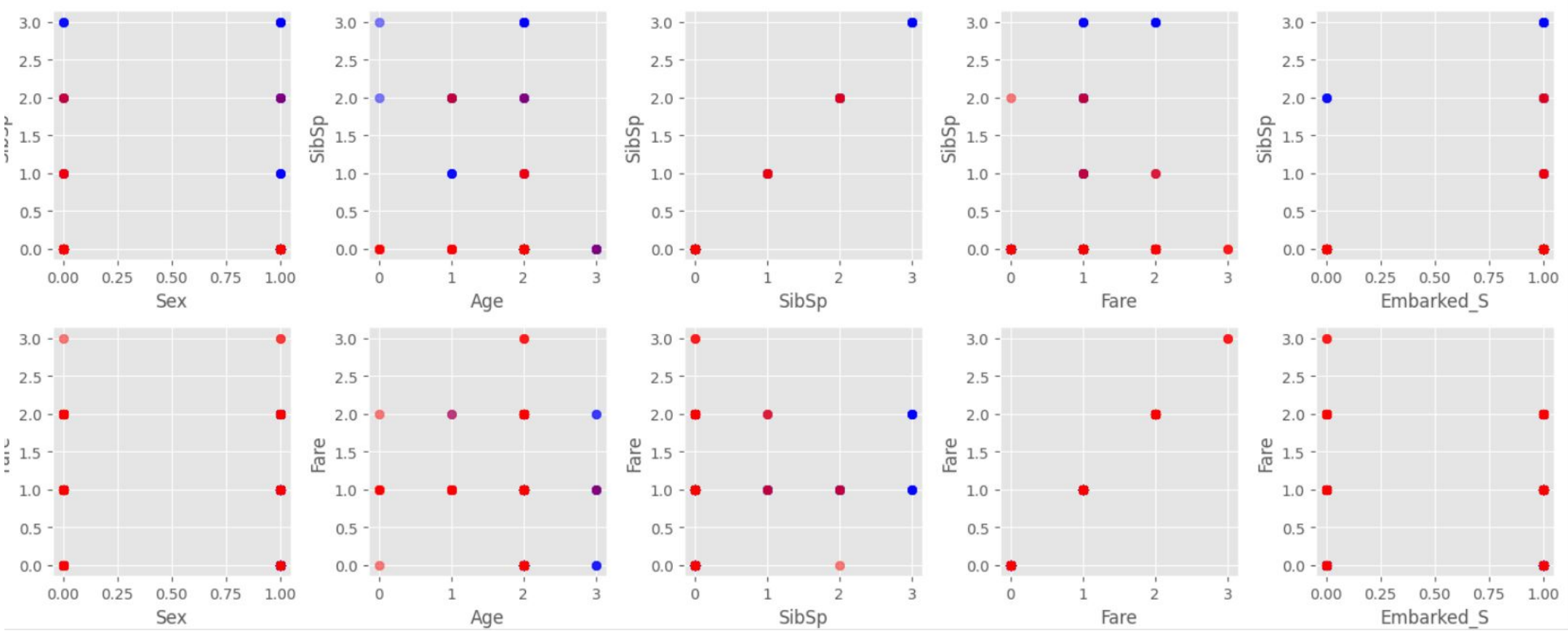
DUPLICATE DATA

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
289	290	1.0	3	Connolly, Miss. Kate	female	22.0	0	0	370373	7.7500	NaN	Q
897	898	NaN	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q
696	697	0.0	3	Kelly, Mr. James	male	44.0	0	0	363592	8.0500	NaN	S
891	892	NaN	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q

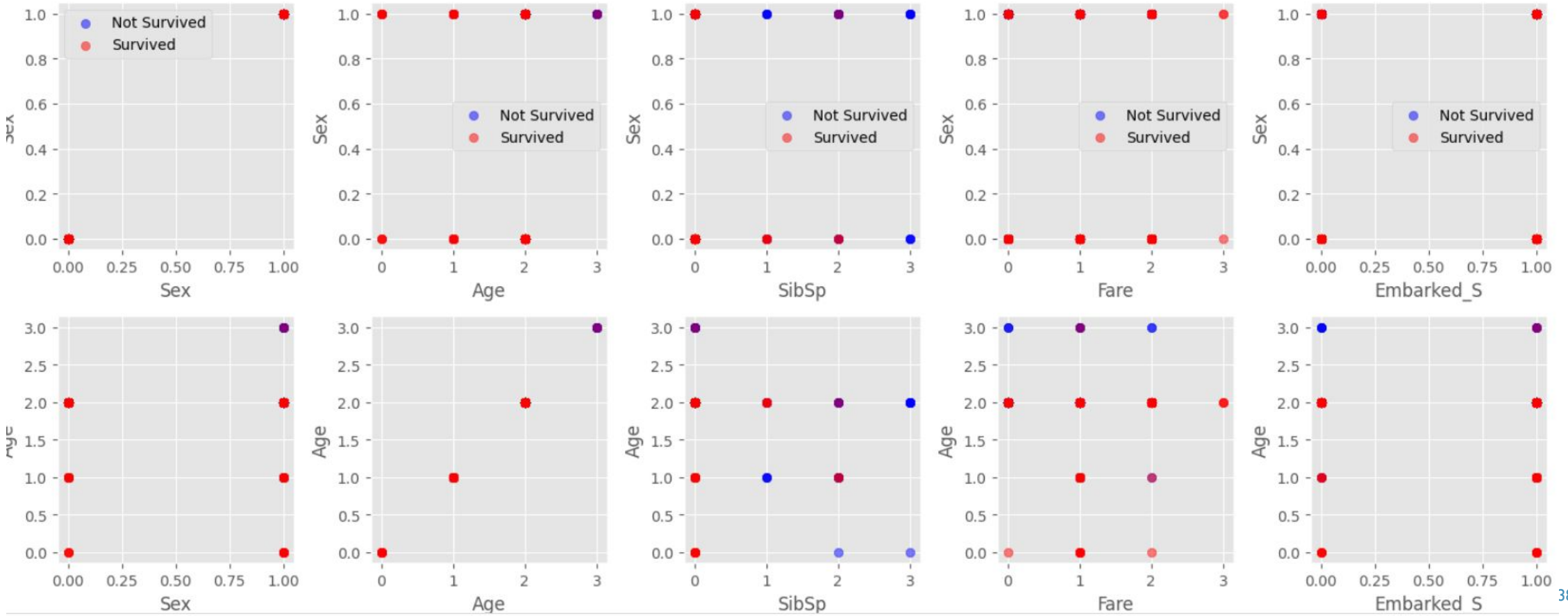
FINAL ANALYSIS



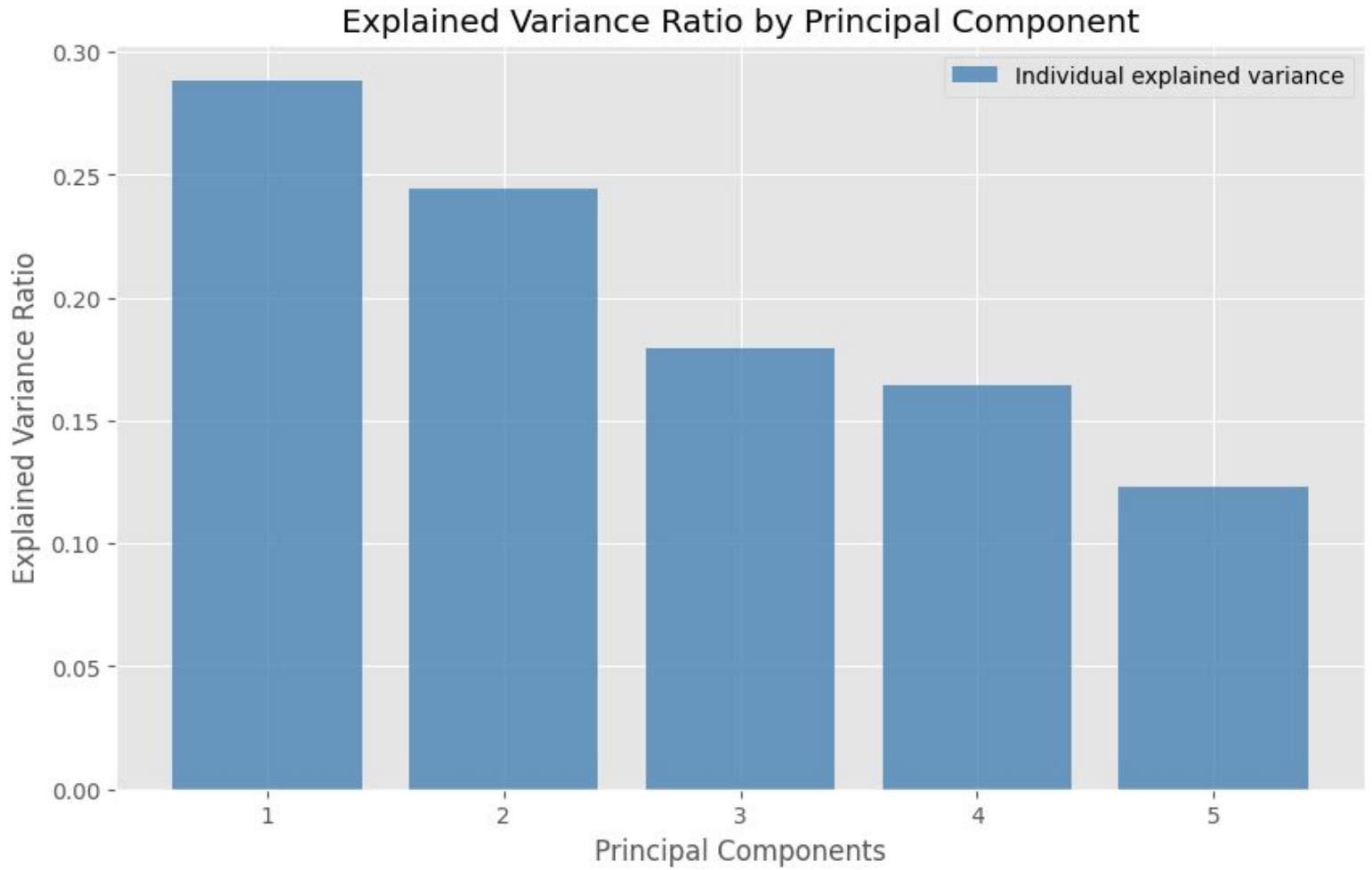
FINAL ANALYSIS



FINAL ANALYSIS



FINAL ANALYSIS



TRAINING WHILE LEAVING EMBARKED VARIABLES

Cross validation metrics for Support Vector Machine

Metric	Mean	Std Dev
F1 Score	0.738957	0.040573
Accuracy	0.810338	0.024610
Recall	0.703571	0.048641
Miss Rate	0.296429	0.048641
Specificity	0.876038	0.014349
Fall out	0.123962	0.014349
Precision	0.778481	0.030926
NPV	0.826777	0.023079

Cross validation metrics for Random Forest

Metric	Mean	Std Dev
F1 Score	0.739907	0.051683
Accuracy	0.819327	0.024783
Recall	0.664735	0.067257
Miss Rate	0.329551	0.077697
Specificity	0.913894	0.020353
Fall out	0.086106	0.020353
Precision	0.830201	0.024898
NPV	0.818928	0.033171

Cross validation metrics for Logistic Regression

Metric	Mean	Std Dev
F1 Score	0.743924	0.037784
Accuracy	0.811462	0.022509
Recall	0.718376	0.052111
Miss Rate	0.281624	0.052111
Specificity	0.868795	0.013628
Fall out	0.131205	0.013628
Precision	0.772491	0.026911
NPV	0.833086	0.025624

CLASSIFICATION REPORTS

Classification report for: Logistic Regression

	precision	recall	f1-score	support
0	0.823944	0.873134	0.847826	134.000000
1	0.790123	0.719101	0.752941	89.000000
accuracy	0.811659	0.811659	0.811659	0.811659
macro avg	0.807034	0.796118	0.800384	223.000000
weighted avg	0.810446	0.811659	0.809957	223.000000

Classification report for: Random Forest

	precision	recall	f1-score	support
0	0.808219	0.880597	0.842857	134.000000
1	0.792208	0.685393	0.734940	89.000000
accuracy	0.802691	0.802691	0.802691	0.802691
macro avg	0.800213	0.782995	0.788898	223.000000
weighted avg	0.801829	0.802691	0.799787	223.000000

Classification report for: Support Vector Machine

	precision	recall	f1-score	support
0	0.810811	0.895522	0.851064	134.000000
1	0.813333	0.685393	0.743902	89.000000
accuracy	0.811659	0.811659	0.811659	0.811659
macro avg	0.812072	0.790458	0.797483	223.000000
weighted avg	0.811818	0.811659	0.808295	223.000000

ALTERNATIVE MODEL TESTING

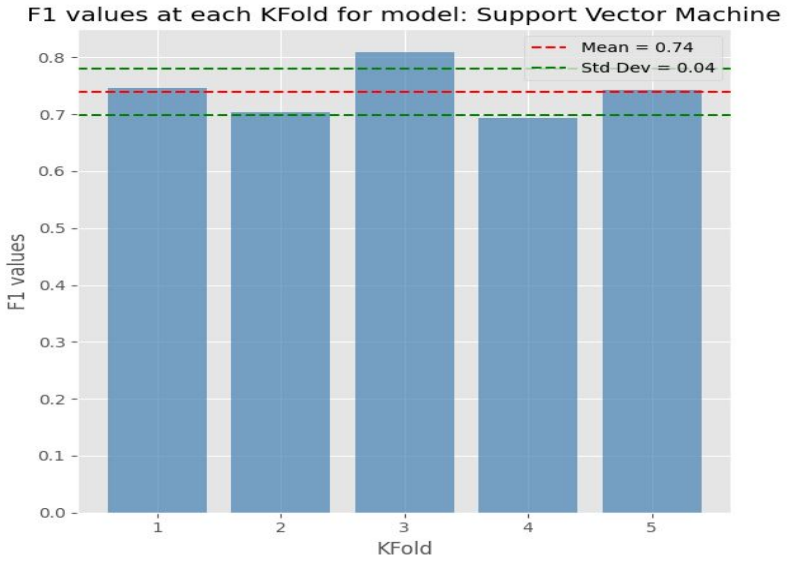
Cross validation metrics for K Nearest Neighbors

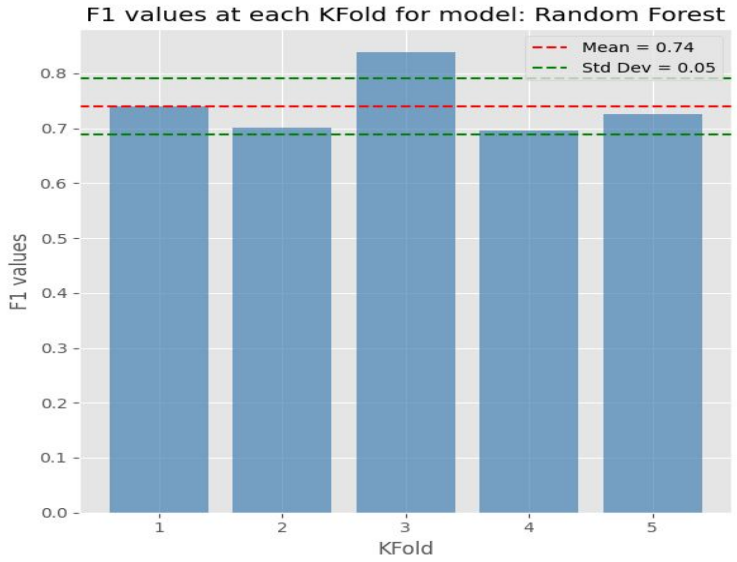
Metric	Mean	Std Dev
F1 Score	0.660918	0.094354
Accuracy	0.759802	0.049111
Recall	0.634352	0.159504
Specificity	0.834571	0.123142

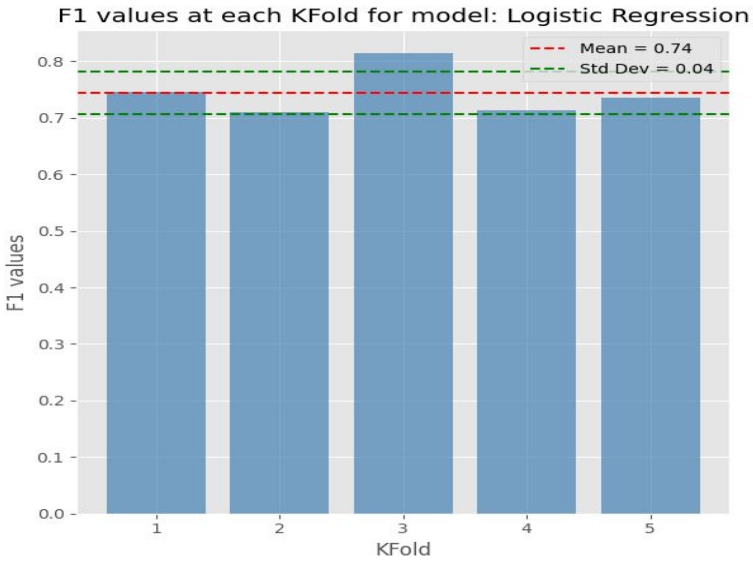
Cross validation metrics for Naive Bayes

Metric	Mean	Std Dev
F1 Score	0.682536	0.036157
Accuracy	0.786793	0.025011
Recall	0.598293	0.039931
Specificity	0.902960	0.028221

F1 VALUES


F1 values at each KFold for model: Support Vector Machine
Mean = 0.74
Std Dev = 0.04
F1 values
KFold


F1 values at each KFold for model: Random Forest
Mean = 0.74
Std Dev = 0.05
F1 values
KFold


F1 values at each KFold for model: Logistic Regression
Mean = 0.74
Std Dev = 0.04
F1 values
KFold

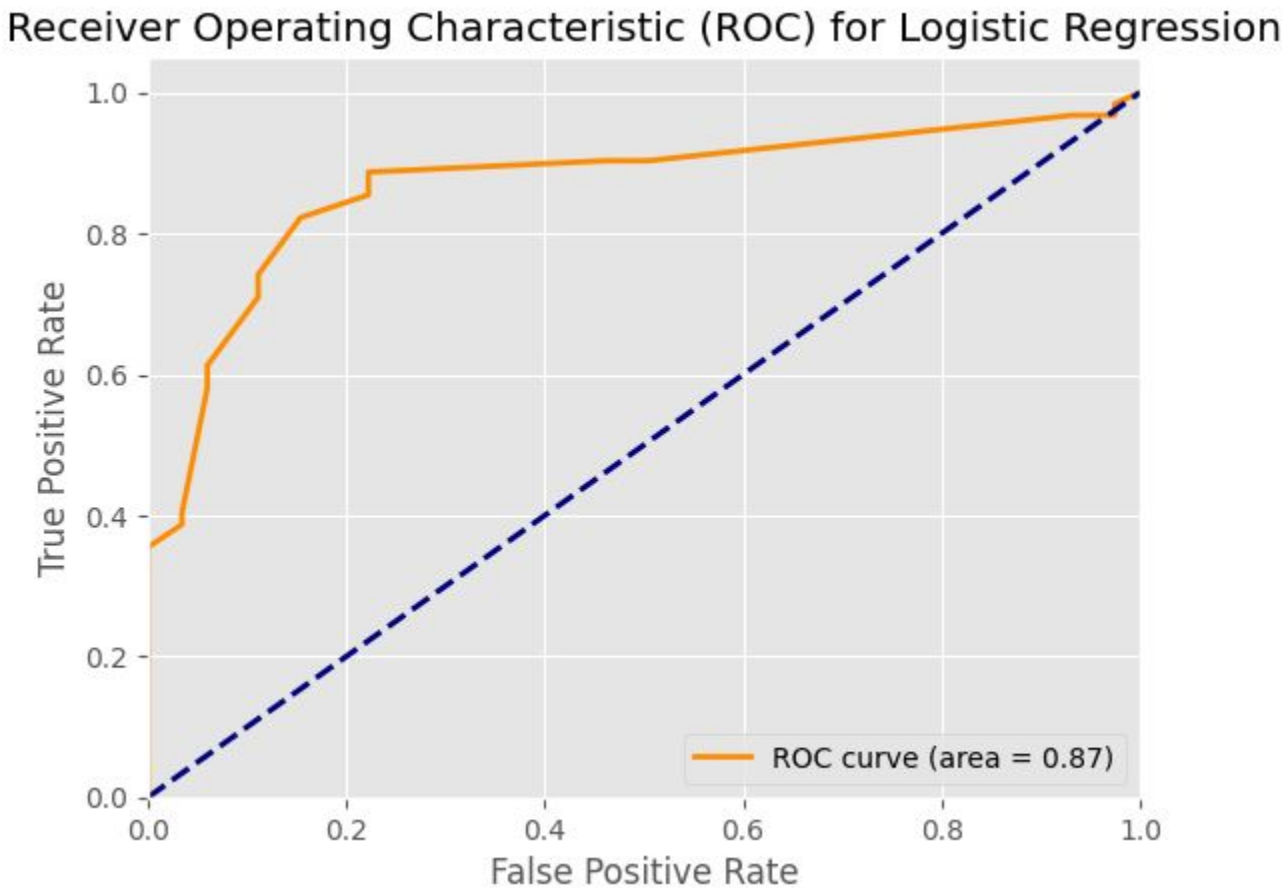
LOGISTIC REGRESSION

Classification report for: Logistic Regression

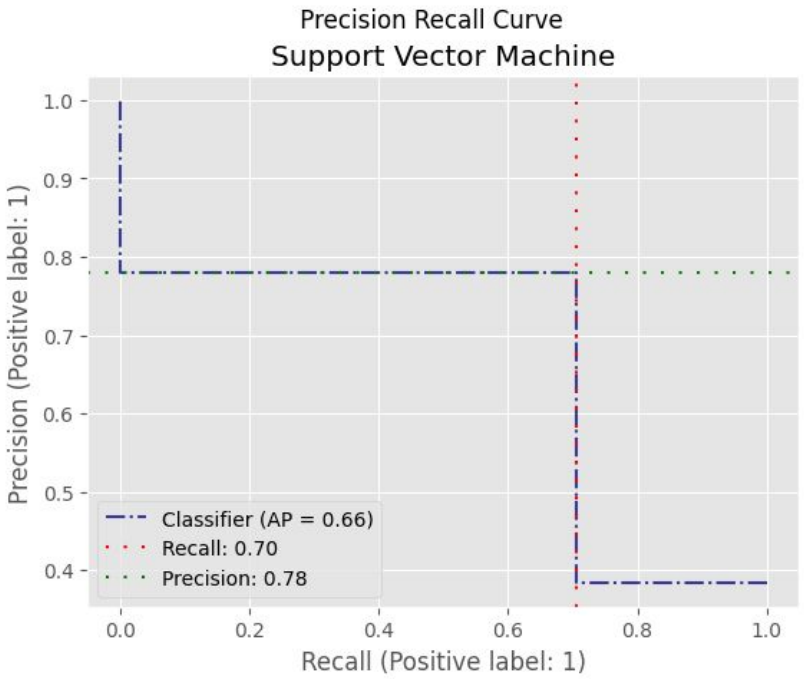
	precision	recall	f1-score	support
0	0.832461	0.868852	0.850267	549.000000
1	0.773585	0.719298	0.745455	342.000000
accuracy	0.811448	0.811448	0.811448	0.811448
macro avg	0.803023	0.794075	0.797861	891.000000
weighted avg	0.809862	0.811448	0.810036	891.000000

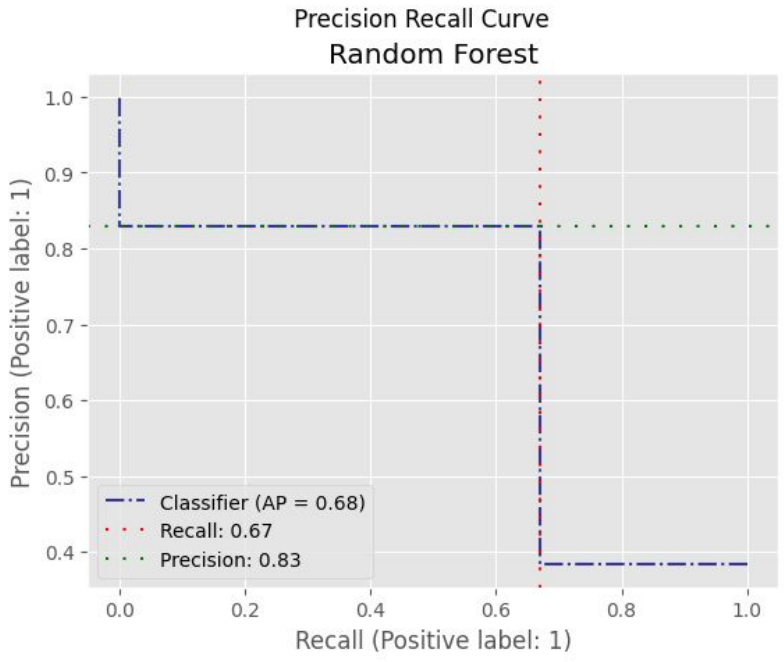
Feature	Coefficient
Intercept	3.145737
Sex	-2.414660
Age	-1.472008
SibSp	-1.390792
Fare	0.836028
Embarked_C	0.474054

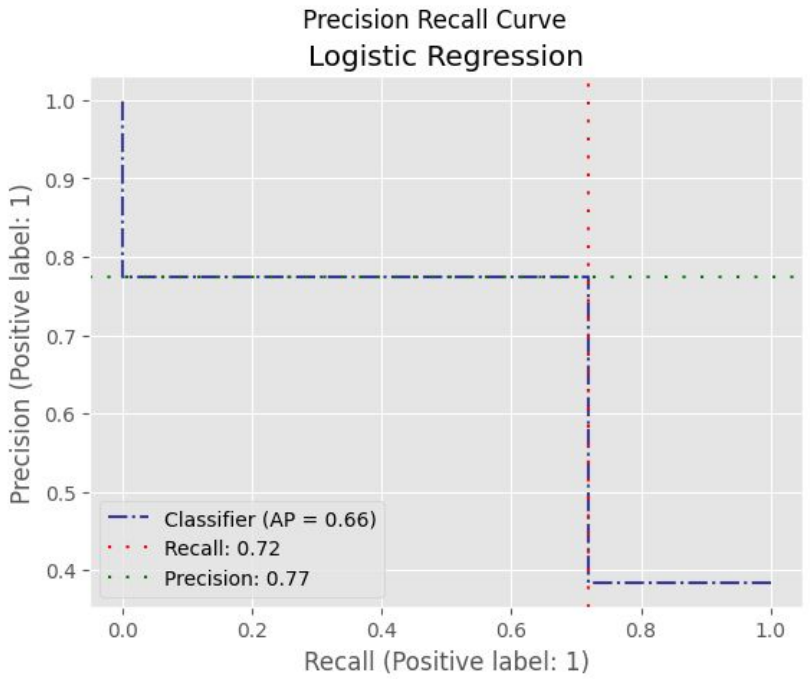
LOGISTIC REGRESSION



Precision Recall Curves







YOUDEN INDEX

Youden Index: *specificity + recall - 1*