ML Raport

AutoPrep

January 8, 2025

Abstract

This raport has been generated with AutoPrep.

Contents

1 Overview

1.1 System

System	Darwin
Machine	arm64
Processor	arm
Architecture	64bit
Python Version	3.10.5
Physical Cores	8
Logical Cores	8
CPU Frequency (MHz)	3204
Total RAM (GB)	16.00
Available RAM (GB)	3.35
Total Disk Space (GB)	228.27
Free Disk Space (GB)	9.16

Table 1: System overview.

1.2 Dataset

Number of samples	1047
Number of features	13
Number of numerical features	6
Number of categorical features	7

Table 2: Dataset Summary.

class	**number of observations**	**Percentage**
0	665	0.64
1	382	0.36

Table 3: Target class distribution.

classgit	**number of observations**	**Percentage**
pclass	0	0.00
name	0	0.00
sex	0	0.00
age	207	0.20
sibsp	0	0.00
parch	0	0.00
ticket	0	0.00
fare	1	0.00
cabin	813	0.78
embarked	1	0.00
boat	672	0.64
body	948	0.91
home___dest	453	0.43

Table 4: Missing values distribution.

class	type	dtype	space usage
pclass	numerical	uint8	9.4 kB
name	categorical	object	96.4 kB
sex	categorical	category	9.7 kB
age	numerical	float64	16.8 kB
sibsp	numerical	uint8	9.4 kB
parch	numerical	uint8	9.4 kB
ticket	categorical	object	75.1 kB
fare	numerical	float64	16.8 kB
cabin	categorical	object	42.1 kB
embarked	categorical	category	9.7 kB
boat	categorical	object	46.4 kB
body	numerical	float64	16.8 kB
home___dest	categorical	object	64.5 kB

Table 5: Features dtypes description.

index	count	mean	std	min	25%	50%	75%	max
pclass	1047.00	2.30	0.84	1.00	2.00	3.00	3.00	3.00
age	840.00	29.53	14.27	0.17	21.00	28.00	38.62	80.00
sibsp	1047.00	0.52	1.05	0.00	0.00	0.00	1.00	8.00
parch	1047.00	0.40	0.89	0.00	0.00	0.00	0.00	9.00
fare	1046.00	33.55	51.81	0.00	7.92	14.50	31.27	512.33
body	99.00	160.90	98.35	1.00	73.50	156.00	255.50	328.00

Table 6: Numerical features description.

index	count	unique	top	freq
name	1047	1046	Connolly, Miss. Kate	2
ticket	1047	773	CA. 2343	9
cabin	234	161	B57 B59 B63 B66	5
boat	375	25	13	34
home___dest	594	317	New York, NY	50

Table 7: Categorical features description.

2 Eda

2.1 Eda

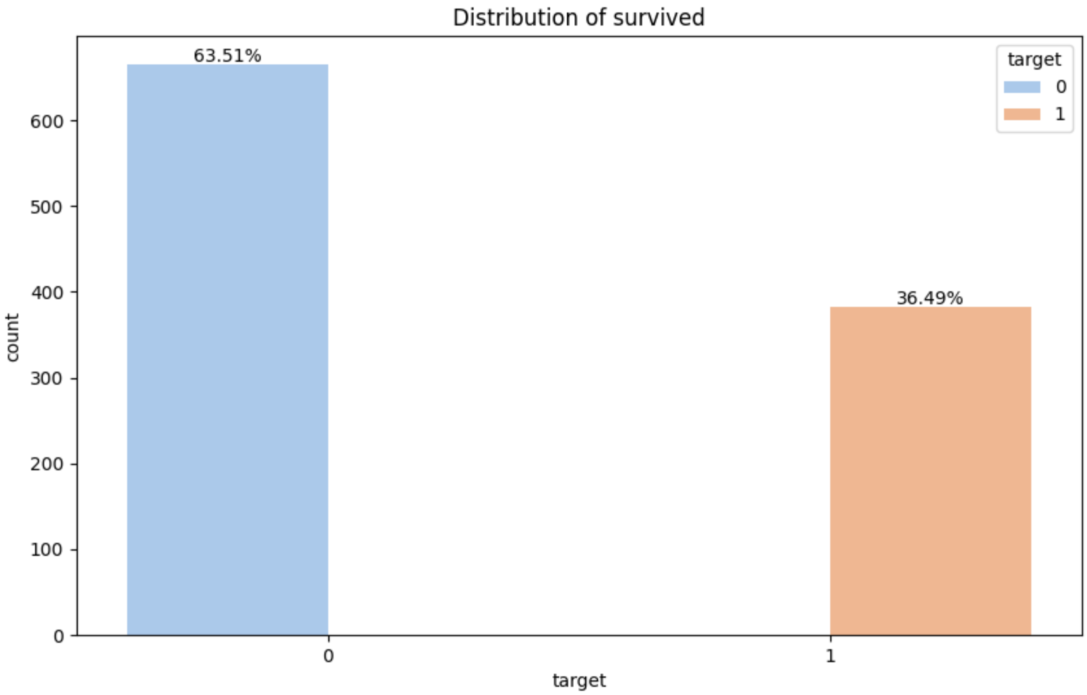

Figure 1: Target distribution.

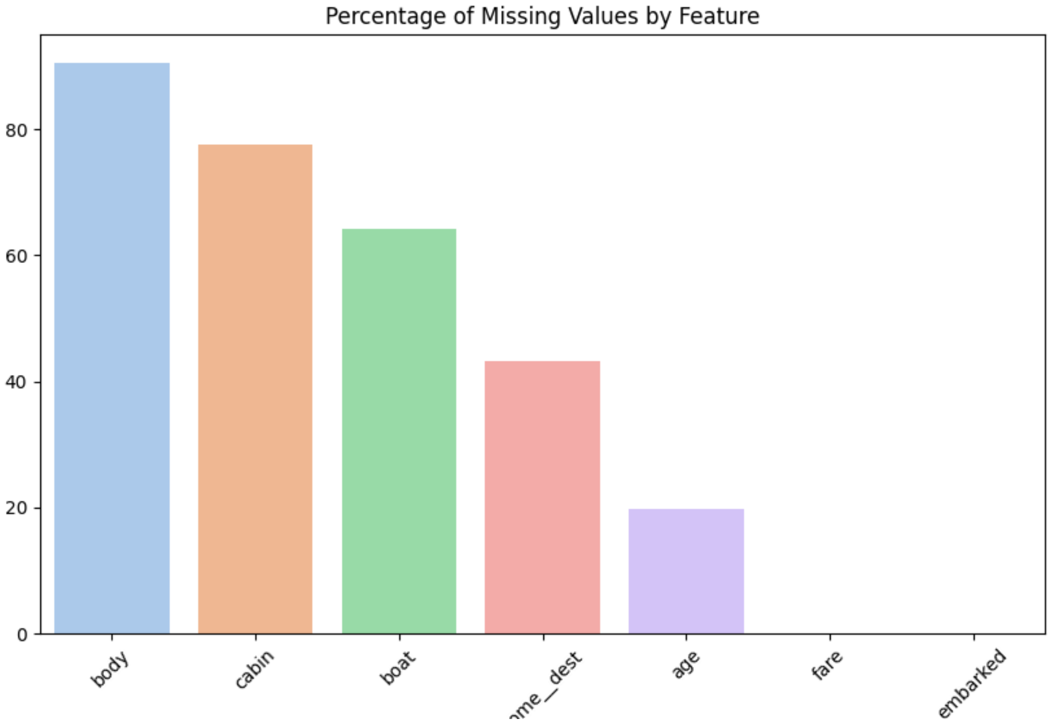

Figure 2: Missing values.

2.2 Categorical

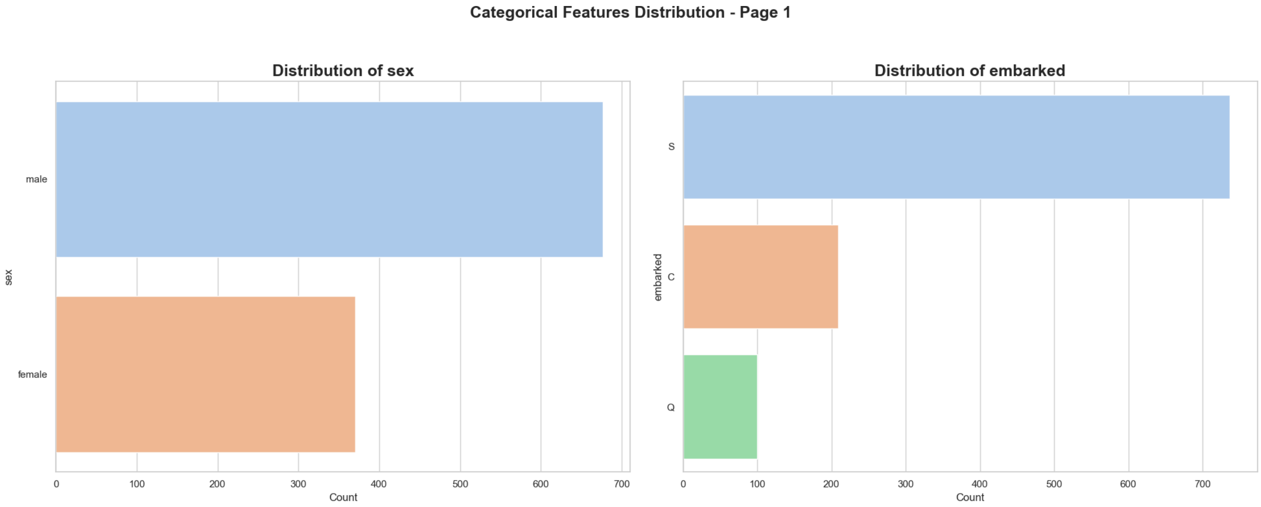

Figure 3: Categorical Features Distribution - Page 1

2.3 Numerical

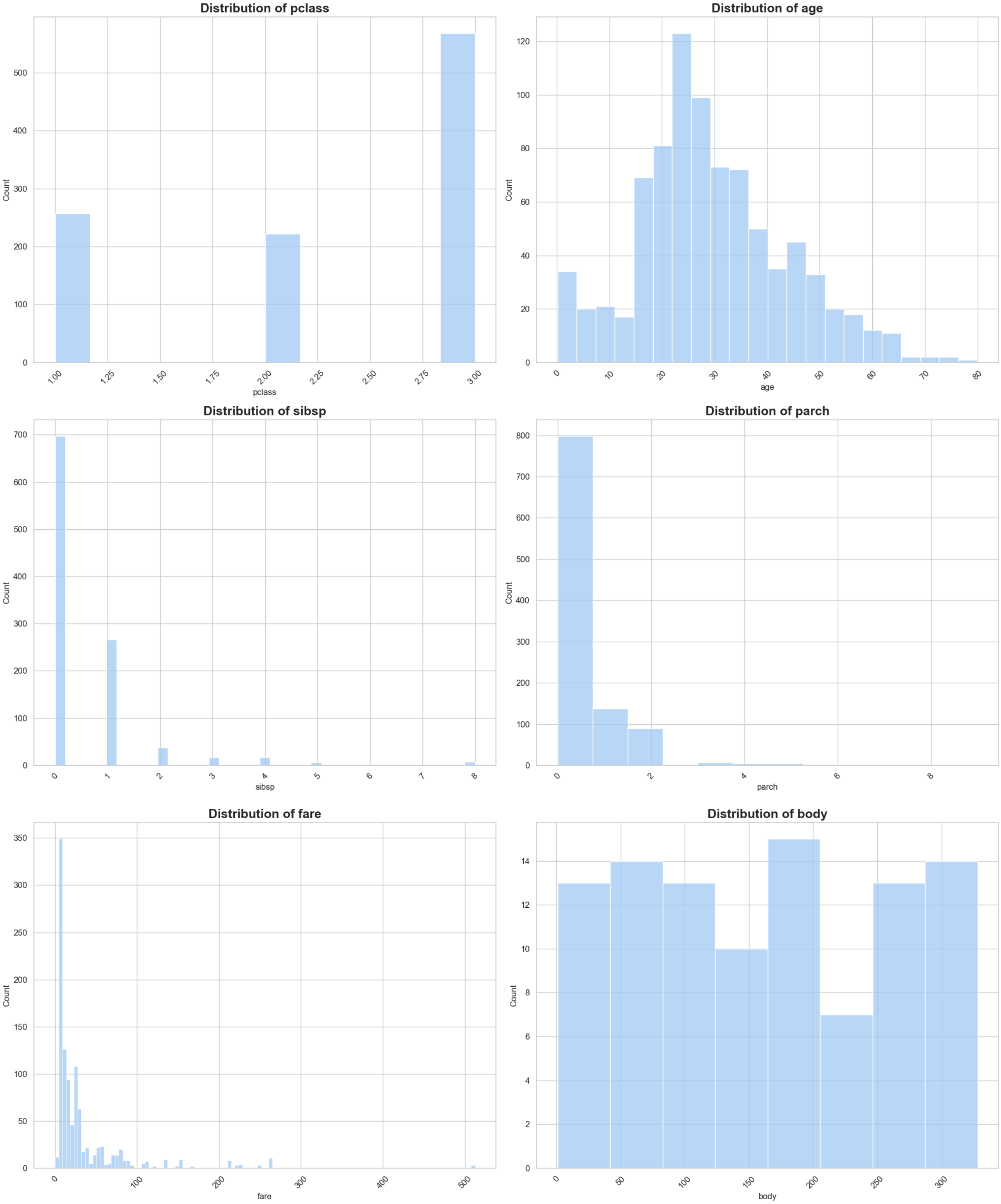

Figure 4: Numerical Features Distribution - Page 1

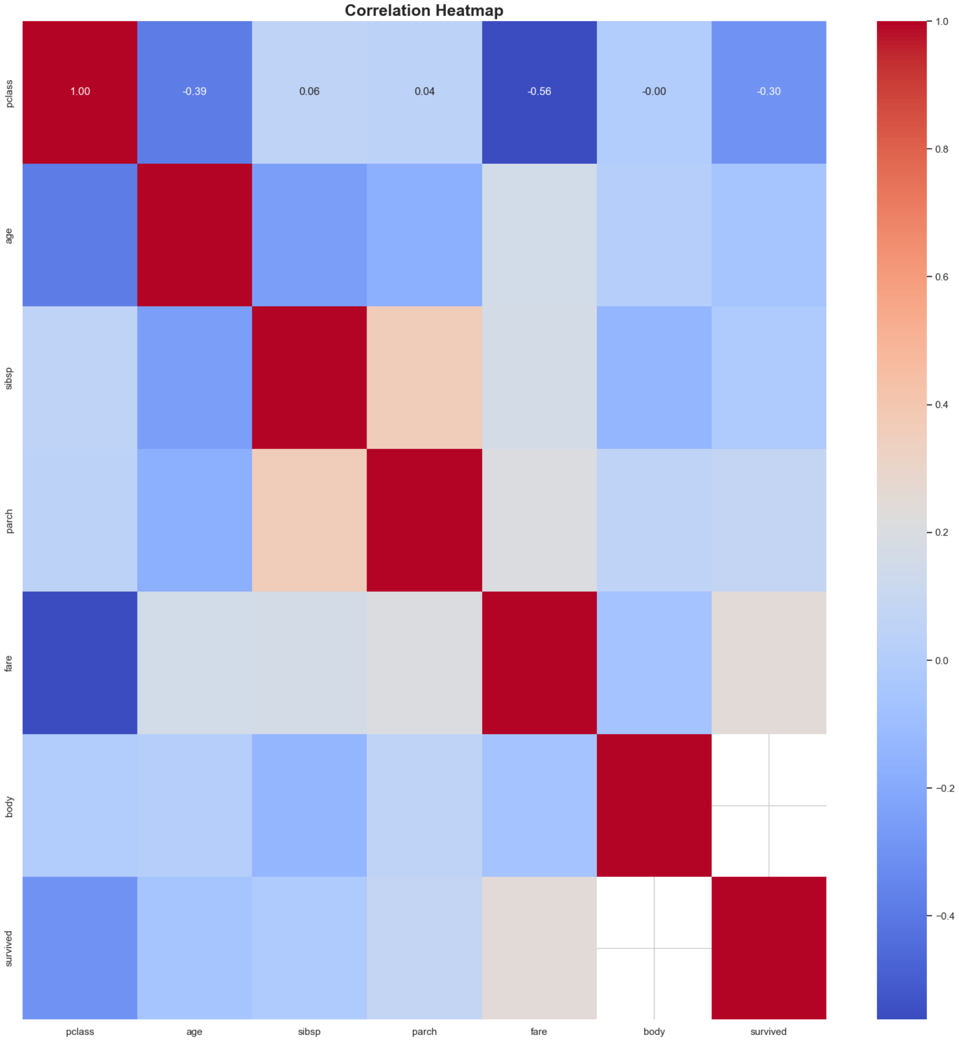

Figure 5: Correlation heatmap.

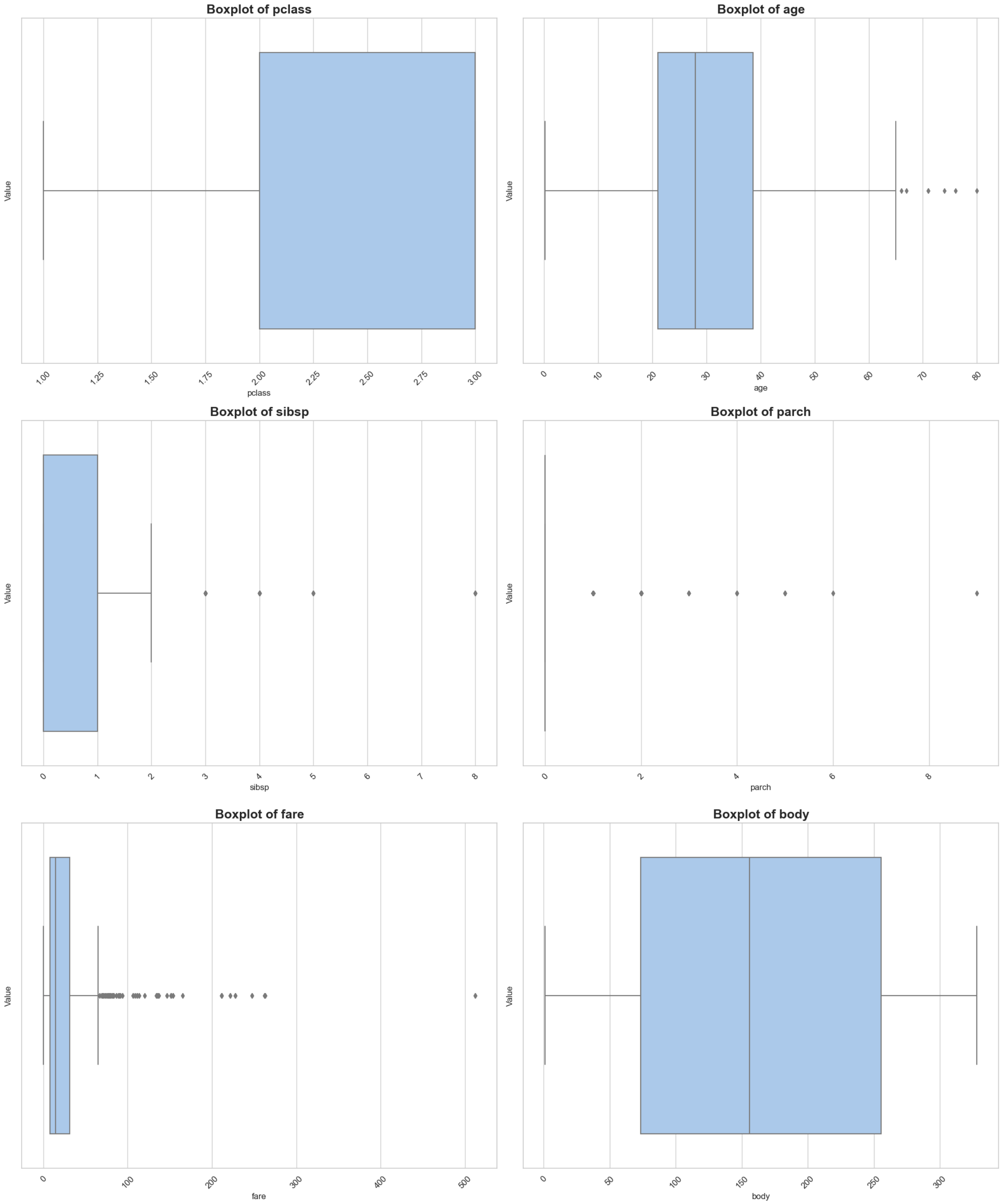

Figure 6: Boxplot page 1

3 Preprocessing

Category	Value
Unique created pipelines	2
All created pipelines (after exploading each step params)	6
All pipelines fit time	4 seconds
All pipelines score time	4 seconds
scores_count	6.00
scores_mean	0.76
scores_std	0.01
scores_min	0.75
scores_25%	0.75
scores_50%	0.76
scores_75%	0.77
scores_max	0.78
Scoring function	<class 'str'>
Scoring model	RandomForestClassifier

Table 8: Preprocessing pipelines runtime statistics.

index	steps
0	NAImputer, UniqueFilter, ColumnEncoder, ColumnScaler, CorrelationFilter
1	NAImputer, UniqueFilter, ColumnEncoder, ColumnScaler, VarianceFilter

Table 9: Pipelines steps overview.

score index	file name	score	fit duration	score duration
0	preprocessing_pipeline_0.joblib	0.78	a moment	a moment
1	preprocessing_pipeline_1.joblib	0.77	a moment	a moment
2	preprocessing_pipeline_2.joblib	0.77	a moment	a moment

Table 10: Best preprocessing pipelines.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "robust"}
4	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}

Table 11: 0th best pipeline overwiev on training set.

index	count	mean	std	min	25%	50%	75%	max
pclass	1047.00	0.00	1.00	-1.55	-0.36	0.84	0.84	0.84
name	1047.00	0.00	1.00	-1.73	-0.87	-0.00	0.87	1.73
age	1047.00	-0.00	1.00	-2.27	-0.57	-0.10	0.45	3.97
sibsp	1047.00	-0.00	1.00	-0.50	-0.50	-0.50	0.46	7.13
parch	1047.00	0.00	1.00	-0.44	-0.44	-0.44	-0.44	9.63
ticket	1047.00	-0.00	1.00	-1.68	-0.90	0.00	0.93	1.67
fare	1047.00	0.00	1.00	-0.65	-0.49	-0.37	-0.04	9.25
home___dest	1047.00	-0.00	1.00	-2.72	-0.18	0.23	0.30	2.01
sex_female	1047.00	0.00	1.00	-0.74	-0.74	-0.74	1.35	1.35
embarked_C	1047.00	-0.00	1.00	-0.50	-0.50	-0.50	-0.50	2.00
embarked_Q	1047.00	0.00	1.00	-0.32	-0.32	-0.32	-0.32	3.08
embarked_S	1047.00	-0.00	1.00	-1.55	-1.55	0.65	0.65	0.65

Table 12: 0th best pipeline output overview.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "standard"}
4	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}

Table 13: 1th best pipeline overwiev on training set.

index	count	mean	std	min	25%	50%	75%	max
pclass	1047.00	0.65	0.42	0.00	0.50	1.00	1.00	1.00
name	1047.00	0.50	0.29	0.00	0.25	0.50	0.75	1.00
age	1047.00	0.36	0.16	0.00	0.27	0.35	0.44	1.00
sibsp	1047.00	0.07	0.13	0.00	0.00	0.00	0.12	1.00
parch	1047.00	0.04	0.10	0.00	0.00	0.00	0.00	1.00
ticket	1047.00	0.50	0.30	0.00	0.23	0.50	0.78	1.00
fare	1047.00	0.07	0.10	0.00	0.02	0.03	0.06	1.00
home___dest	1047.00	0.58	0.21	0.00	0.54	0.62	0.64	1.00
sex_female	1047.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00
embarked_C	1047.00	0.20	0.40	0.00	0.00	0.00	0.00	1.00
embarked_Q	1047.00	0.10	0.29	0.00	0.00	0.00	0.00	1.00
embarked_S	1047.00	0.70	0.46	0.00	0.00	1.00	1.00	1.00

Table 14: 1th best pipeline output overview.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "minmax"}
4	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}

Table 15: 2th best pipeline overwiev on training set.

index	count	mean	std	min	25%	50%	75%	max
pclass	1047.00	-0.70	0.84	-2.00	-1.00	0.00	0.00	0.00
name	1047.00	0.00	0.58	-1.00	-0.50	0.00	0.50	1.00
age	1047.00	0.09	0.98	-2.14	-0.46	0.00	0.54	4.00
sibsp	1047.00	0.52	1.05	0.00	0.00	0.00	1.00	8.00
parch	1047.00	0.40	0.89	0.00	0.00	0.00	0.00	9.00
ticket	1047.00	-0.00	0.55	-0.92	-0.49	0.00	0.51	0.91
fare	1047.00	0.81	2.22	-0.62	-0.28	0.00	0.72	21.32
home__dest	1047.00	-0.48	2.06	-6.09	-0.86	0.00	0.14	3.66
sex_female	1047.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00
embarked_C	1047.00	0.20	0.40	0.00	0.00	0.00	0.00	1.00
embarked_Q	1047.00	0.10	0.29	0.00	0.00	0.00	0.00	1.00
embarked_S	1047.00	-0.30	0.46	-1.00	-1.00	0.00	0.00	0.00

Table 16: 2th best pipeline output overview.

4 Modeling

4.1 Overview

This part of the report presents the results of the modeling process. There were 6 classification models trained and 3 of them selected based on the ROC AUC score.
Models used in the modeling process

- XGBoost
- GaussianNaiveClassifier
- KNeighboursClassifier
- Logistic Regression
- SVC
- DecisionTreeClassifier

The table below presents the results of the modeling process on default parameters for each of the best piplelines. The models are sorted by the ROC AUC score in descending order.

Model	AUC Score	Pipeline
SVC	0.77	preprocessing_pipeline_0
XGBoost	0.76	preprocessing_pipeline_1
XGBoost	0.76	preprocessing_pipeline_2

Table 17: Results of the modeling process on default parameters

4.2 Hyperparameter tuning

This section presents the results of the hyperparameter tuning process for the best 3 models using RandomizedSearchCV.The following parameters grids were used for hyperparameter tuning:

Parameter	Values
C	[0.1, 1, 10, 100, 1000]
kernel	['linear', 'poly', 'rbf', 'sigmoid']
degree	[3, 4, 5]
gamma	['scale', 'auto']
random_state	[42]

Table 18: Parameter grid for SVC

Parameter	Values
max_depth	[3, 6, 7]
learning_rate	[0.01, 0.1, 0.3]
subsample	[0.5, 0.8, 1.0]
colsample_bytree	[0.8, 1.0]
objective	['binary:logistic', 'multi:softprob', 'reg:squarederror']

Table 19: Parameter grid for XGBoost

The table below presents the results of the hyperparameter tuning process for the best 3 models. The models are sorted by the ROC AUC score in descending order.

Model	Params	Pipeline	ROC AUC
XGBoost	{'subsample': 1.0, 'objective': 'binary:logistic', 'max_depth': 3, 'learning_rate': 0.1, 'colsample_bytree': 0.8}	preprocessing_pipeline_1	0.86
XGBoost	{'subsample': 1.0, 'objective': 'binary:logistic', 'max_depth': 3, 'learning_rate': 0.1, 'colsample_bytree': 0.8}	preprocessing_pipeline_2	0.86
SVC	{'random_state': 42, 'kernel': 'rbf', 'gamma': 'scale', 'degree': 3, 'C': 1}	preprocessing_pipeline_0	0.84

Table 20: Results of the hyperparameter tuning process on default parameters