ML Raport

AutoPrep

January 7, 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	3.11.10
Physical Cores	8
Logical Cores	8
CPU Frequency (MHz)	4056
Total RAM (GB)	16.00
Available RAM (GB)	5.56
Total Disk Space (GB)	460.43
Free Disk Space (GB)	247.70

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	int64	16.8 kB
name	categorical	object	96.4 kB
sex	categorical	category	9.7 kB
age	numerical	float64	16.8 kB
sibsp	numerical	int64	16.8 kB
parch	numerical	int64	16.8 kB
ticket	categorical	object	75.1 kB
fare	numerical	float64	16.8 kB
cabin	categorical	object	48.6 kB
embarked	categorical	category	9.7 kB
boat	categorical	object	51.8 kB
body	numerical	float64	16.8 kB
home___dest	categorical	object	68.2 kB

Table 5: Features description.

2 Eda

2.1 Eda

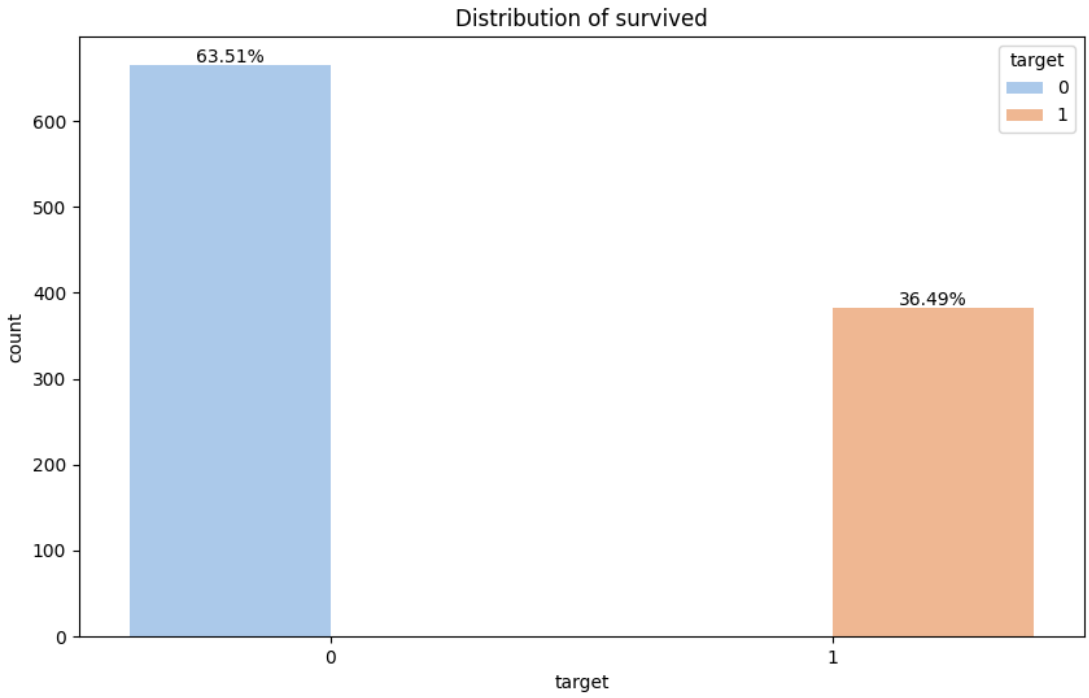

Figure 1: Target distribution.

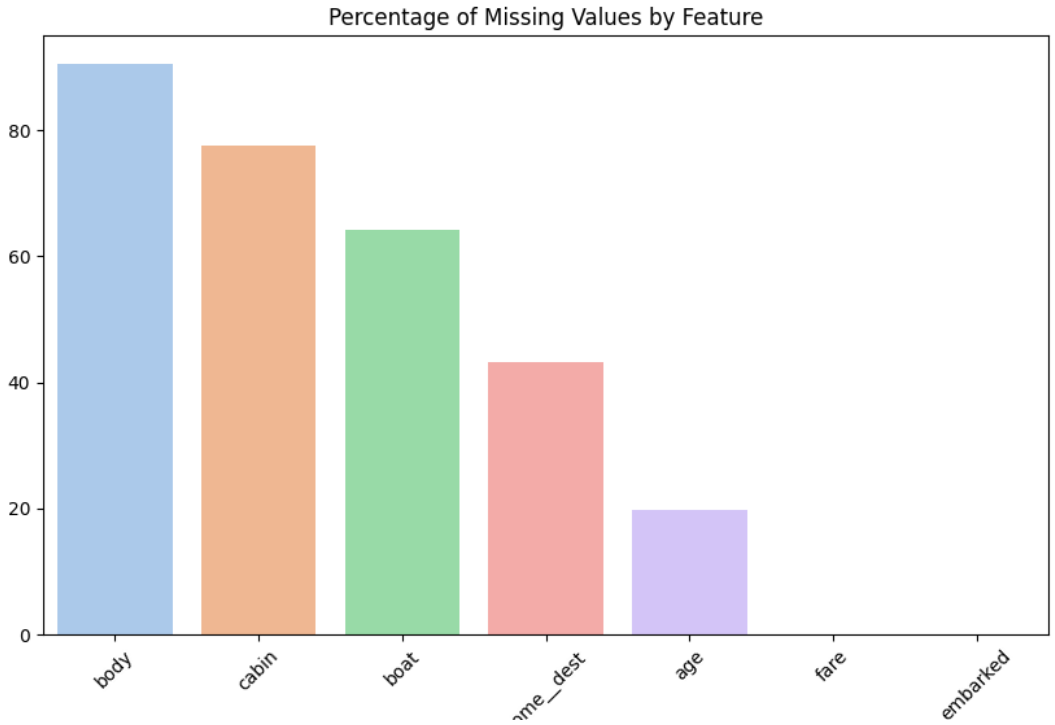

Figure 2: Missing values.

2.2 Categorical

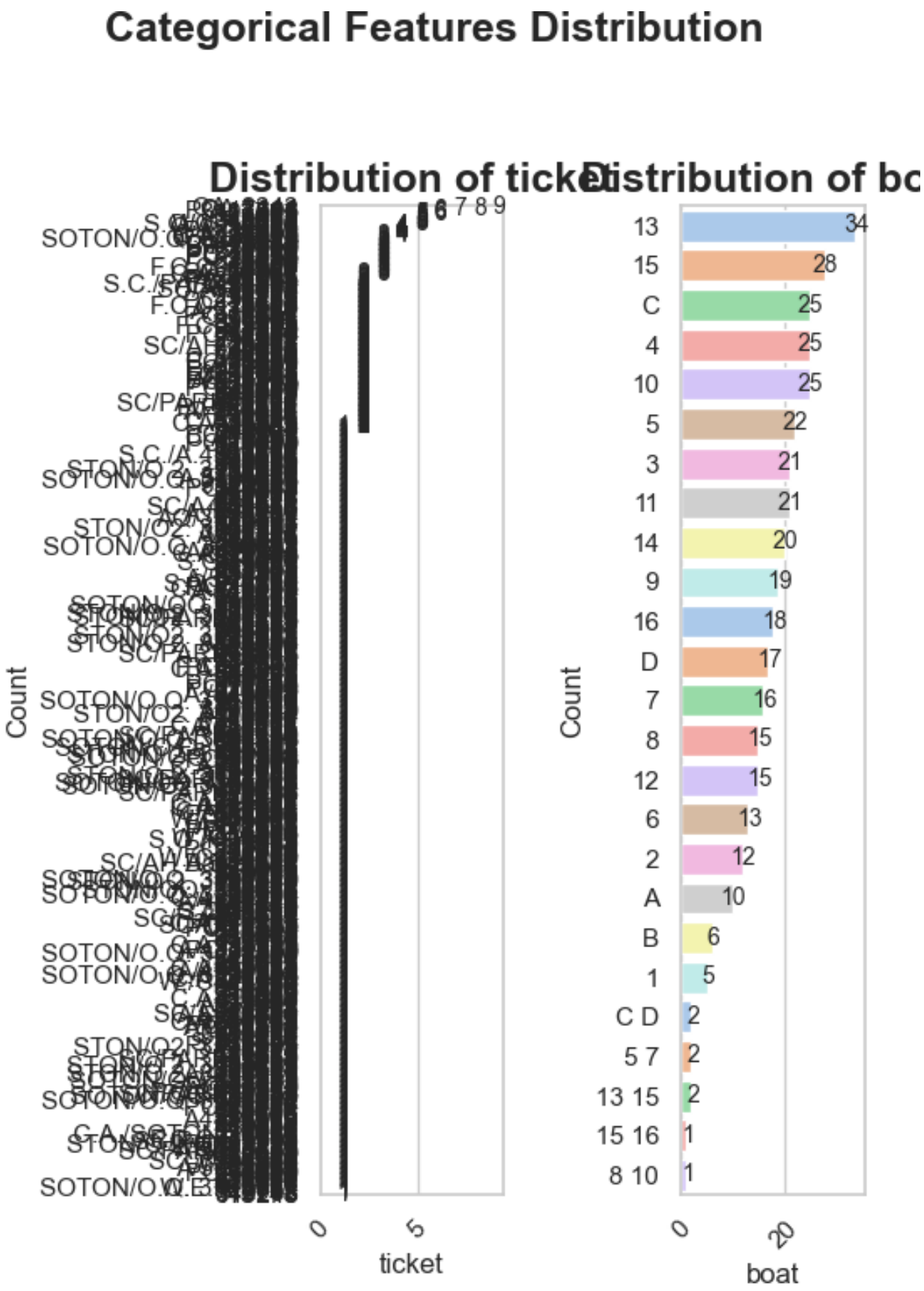

Figure 3: Categorical distribution.

2.3 Numerical

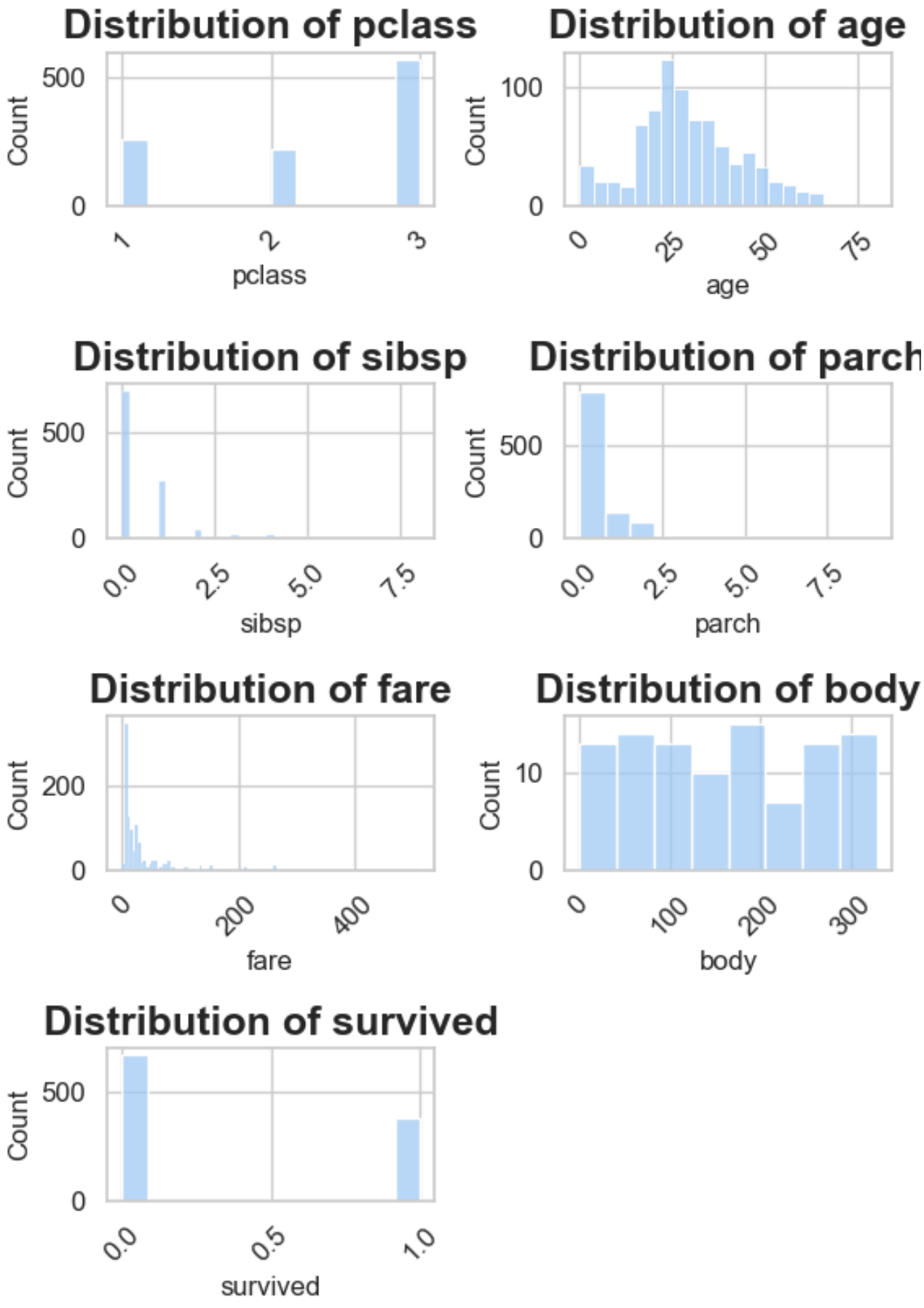

Figure 4: Numerical distribution.

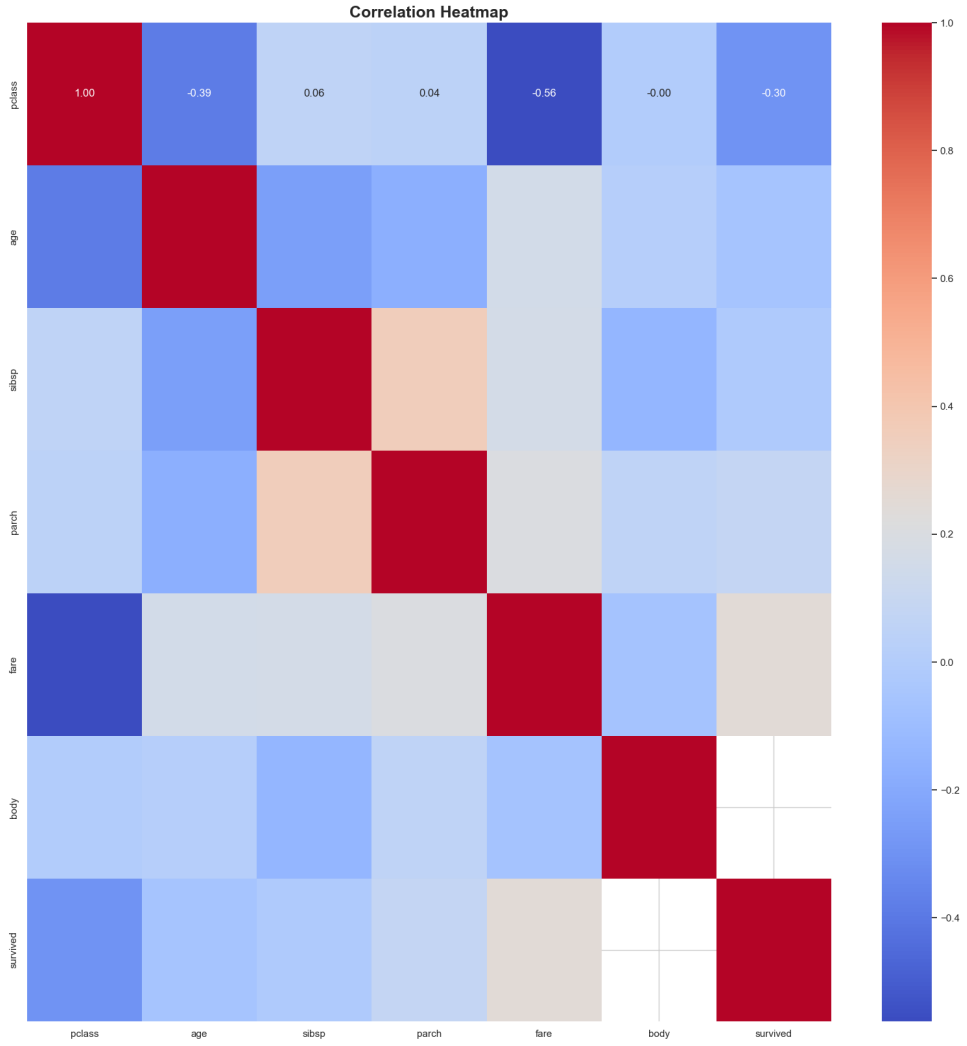

Figure 5: Correlation heatmap.

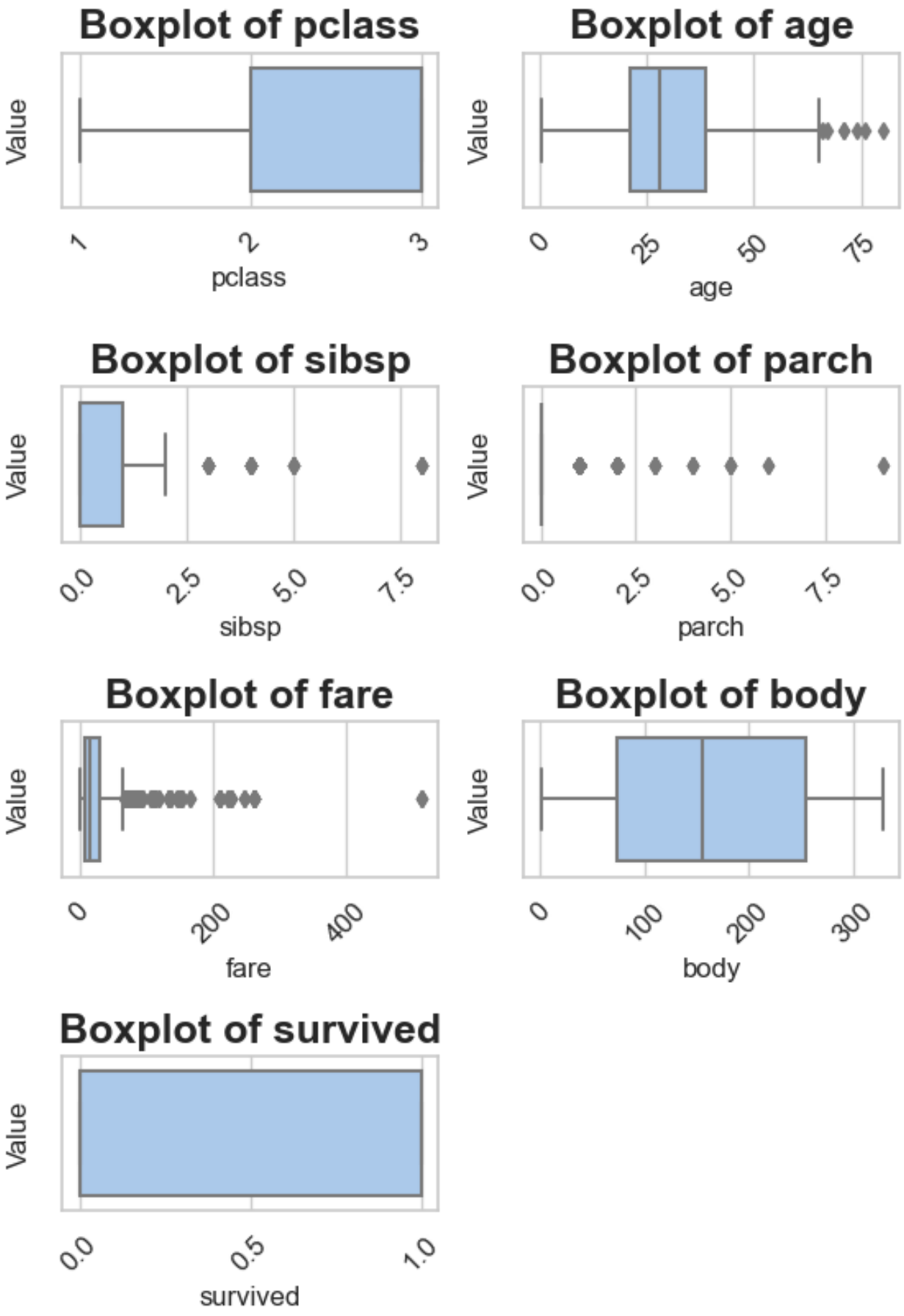

Figure 6: Boxplot.

3 Preprocessing

Category	Value
Unique created pipelines	2
All created pipelines (after exploading each step params)	6
All pipelines fit time	a second
All pipelines score time	2 seconds
scores_count	6.00
scores_mean	0.75
scores_std	0.00
scores_min	0.75
scores_25%	0.75
scores_50%	0.75
scores_75%	0.75
scores_max	0.75
Scoring function	<class 'str'>
Scoring model	RandomForestClassifier

Table 6: Preprocessing pipelines runtime statistics.

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

Table 7: Pipelines steps overview.

score index	file name	score	fit duration	score duration
0	preprocessing_pipeline_0.joblib	0.75	a moment	a moment
1	preprocessing_pipeline_1.joblib	0.75	a moment	a moment
2	preprocessing_pipeline_2.joblib	0.75	a moment	a moment

Table 8: 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	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "standard"}
3	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.	{}
4	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}

Table 9: 0th best pipeline overwiev.

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	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "minmax"}
3	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.	{}
4	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}

Table 10: 1th best pipeline overwiev.

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	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "robust"}
3	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.	{}
4	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}

Table 11: 2th best pipeline overwiev.