dibimbing

#Dibimbing #DigitalSkillFair30

Titanic
Machine Learning From Disaster

Bayu Indradinata

bayu indradinata

data analyst | data science enthusiast | bachelor of informatics engineering student

in bayu-indradinata

bayuindin

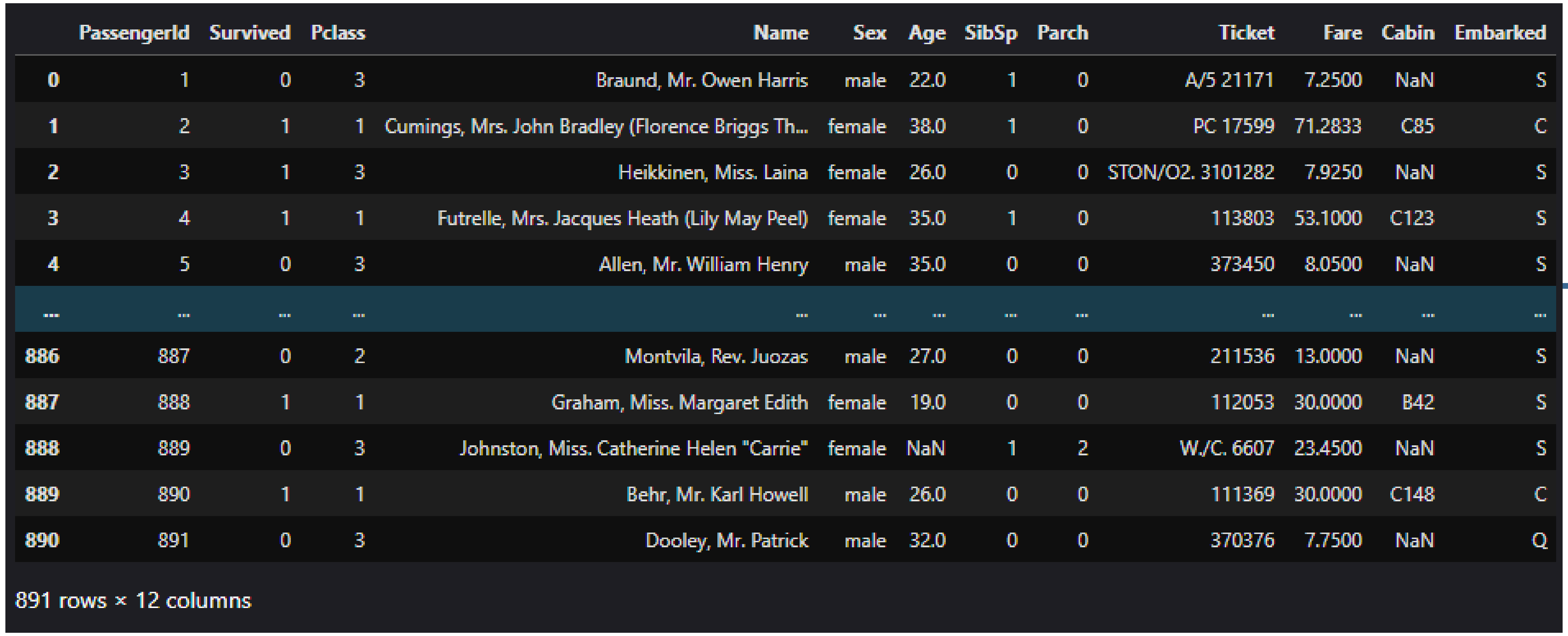
bayuindrad87@gmail.com

Deskripsi Dataset

kolom	deskripsi	key
PassengerId	identitas unik penumpang	
Survived	penumpang yang selamat atau tidak selamat	1 = survived, 0 = not survived
Pclass	kelas penumpang	1, 2, dan 3
Name	nama penumpang	
Sex	jenis kelamin penumpang	

Deskripsi Dataset

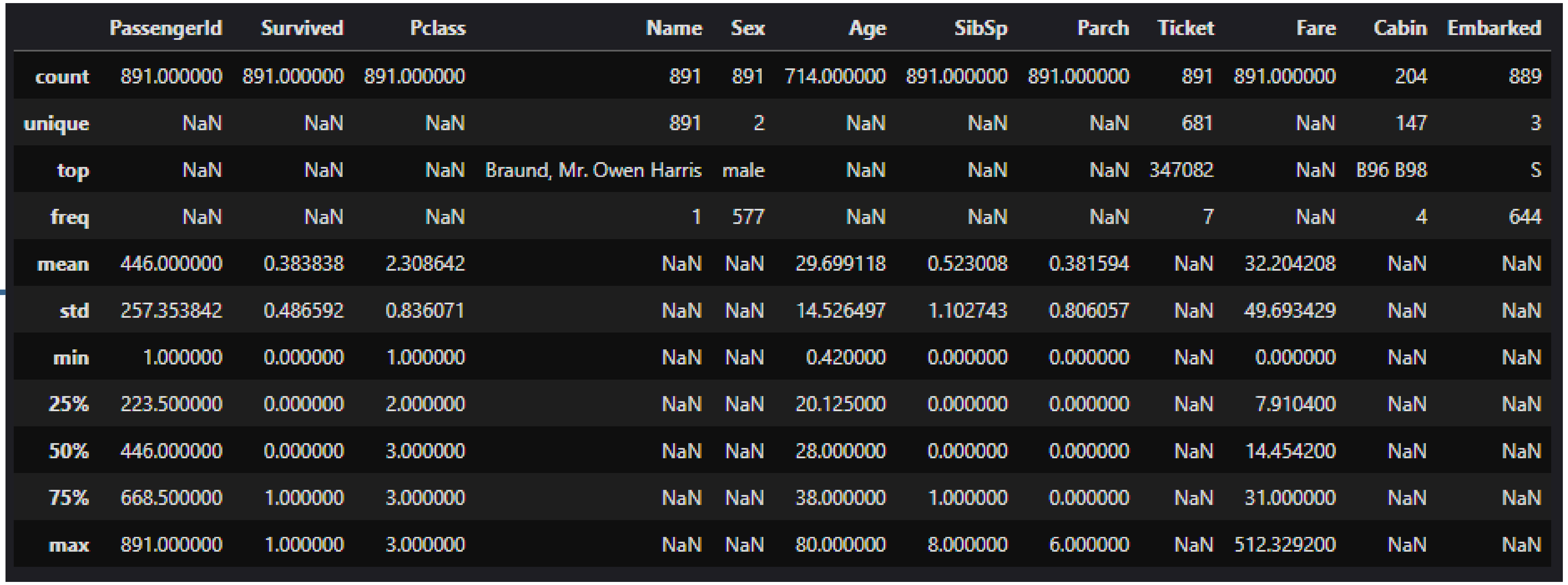
kolom	deskripsi	key
Age	usia	
SibSp	jumlah saudara kandung atau pasangan	
Parch	jumlah orang tua atau anak	
Ticket	nomor tiket	
Fare	tarif	
Cabin	nomor kabin	
Embarked	pelabuhan keberangkatan	C = Cherbourg, Q = Queenstown, S = Southampton



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Dataset

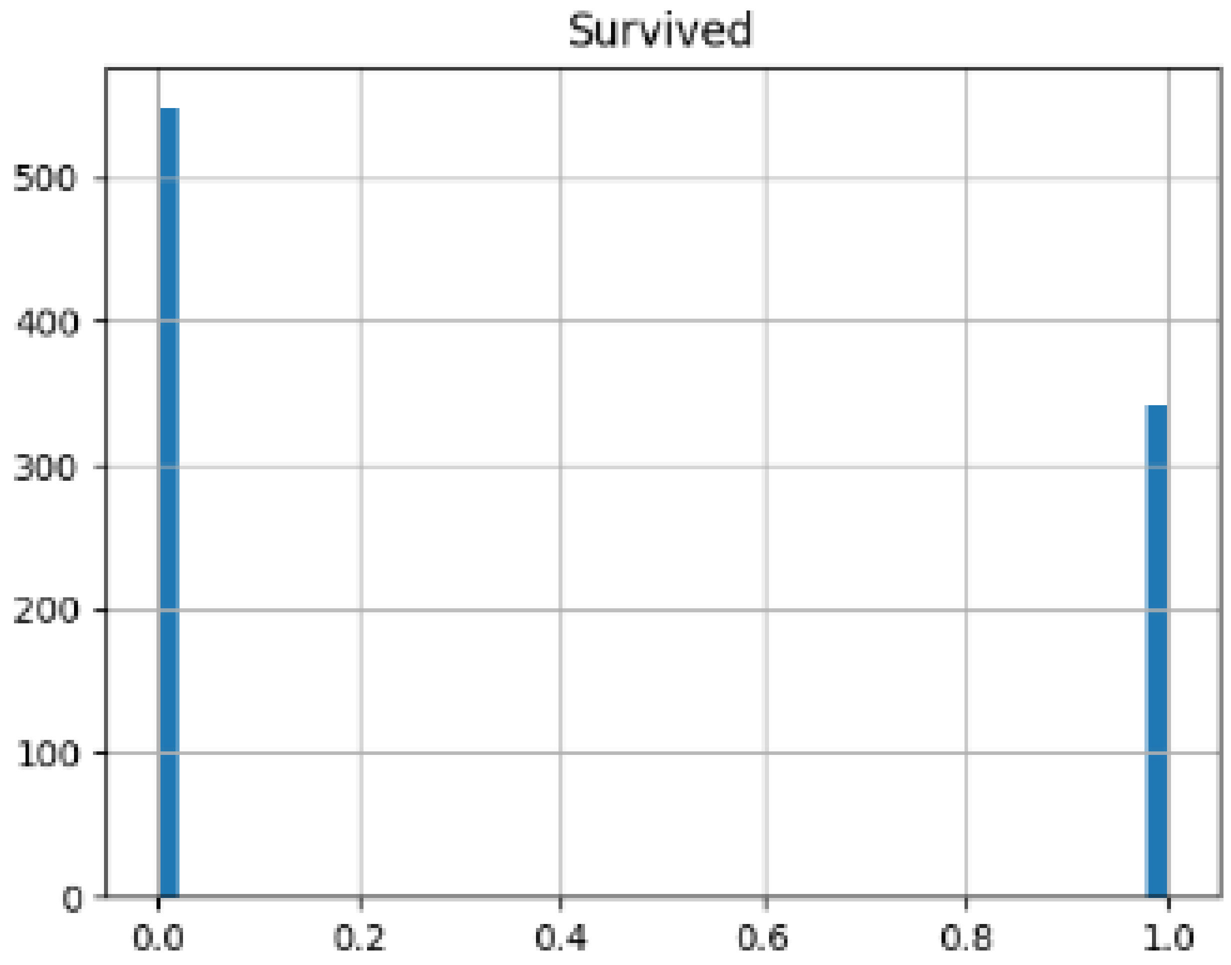


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
count	891.000000	891.000000	891.000000	891	891	714.000000	891.000000	891.000000	891	891.000000	204	889
unique	NaN	NaN	NaN	891	2	NaN	NaN	NaN	681	NaN	147	3
top	NaN	NaN	NaN	Braund, Mr. Owen Harris	male	NaN	NaN	NaN	347082	NaN	B96 B98	S
freq	NaN	NaN	NaN	1	577	NaN	NaN	NaN	7	NaN	4	644
mean	446.000000	0.383838	2.308642	NaN	NaN	29.699118	0.523008	0.381594	NaN	32.204208	NaN	NaN
std	257.353842	0.486592	0.836071	NaN	NaN	14.526497	1.102743	0.806057	NaN	49.693429	NaN	NaN
min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	223.500000	0.000000	2.000000	NaN	NaN	20.125000	0.000000	0.000000	NaN	7.910400	NaN	NaN
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	668.500000	1.000000	3.000000	NaN	NaN	38.000000	1.000000	0.000000	NaN	31.000000	NaN	NaN
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	6.000000	NaN	512.329200	NaN	NaN

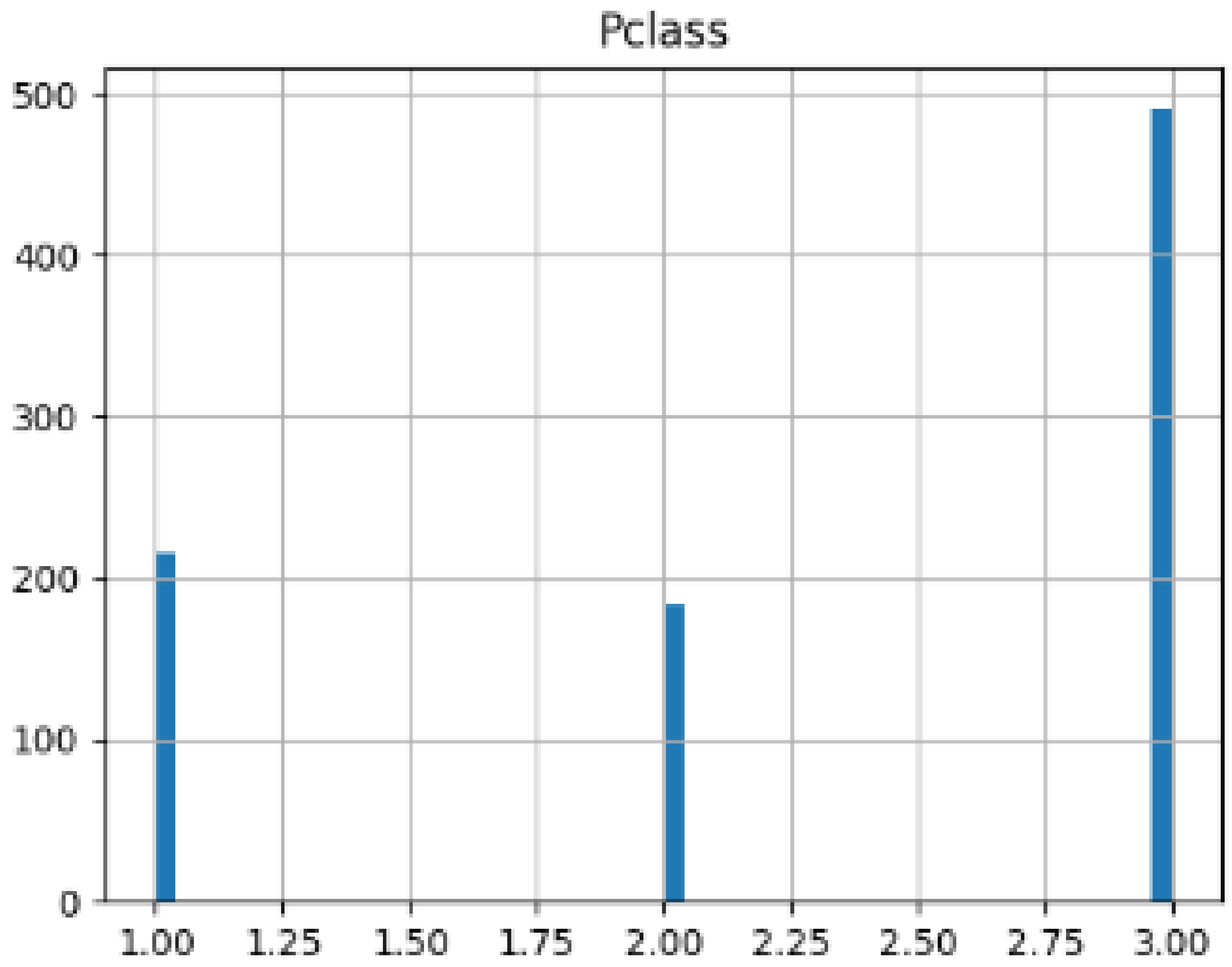
Deskripsi Dataset

ibimbing

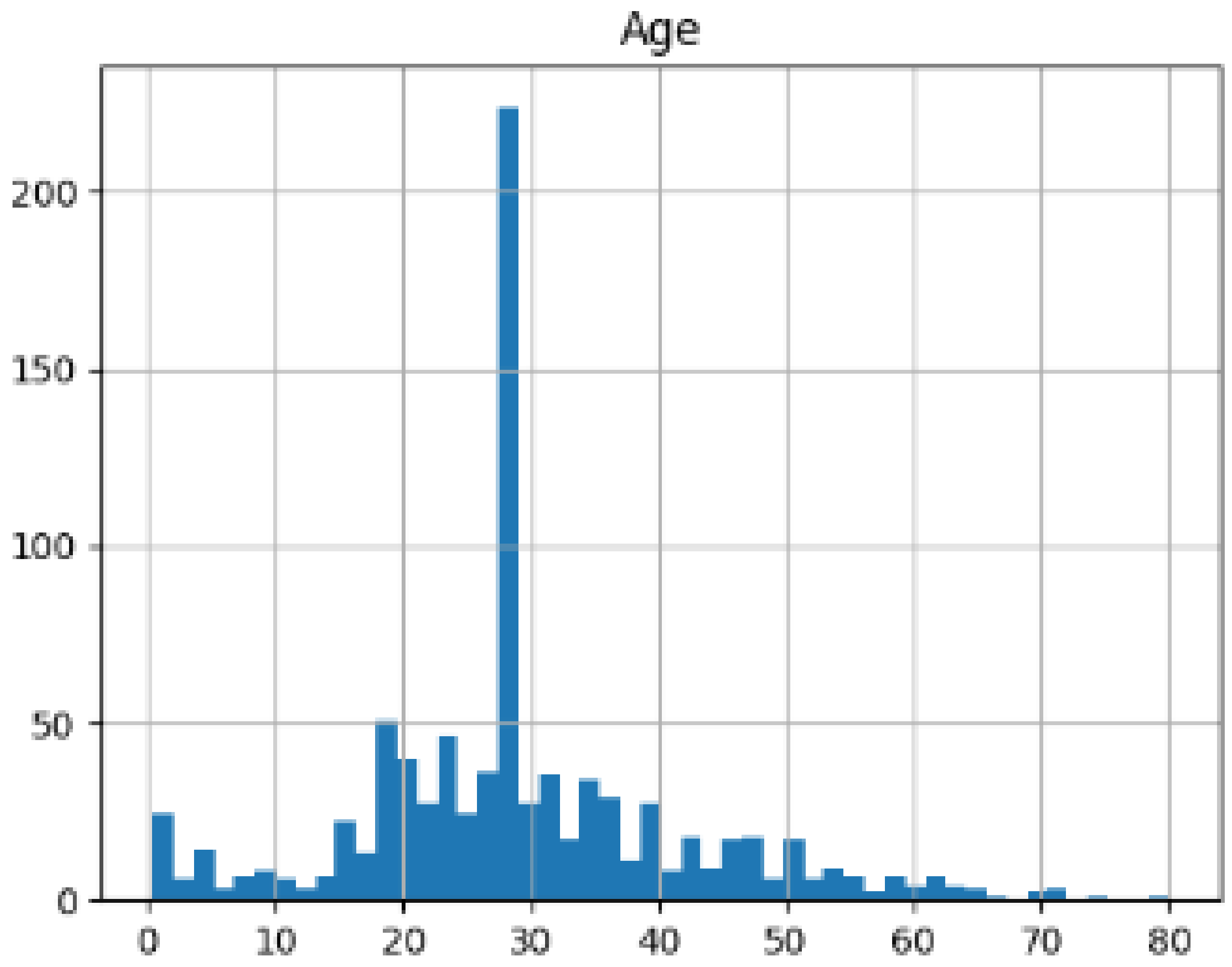
Distribusi Data



Paling banyak tidak selamat sekitar 500 keatas, sementara yang selamat sekitar 350an

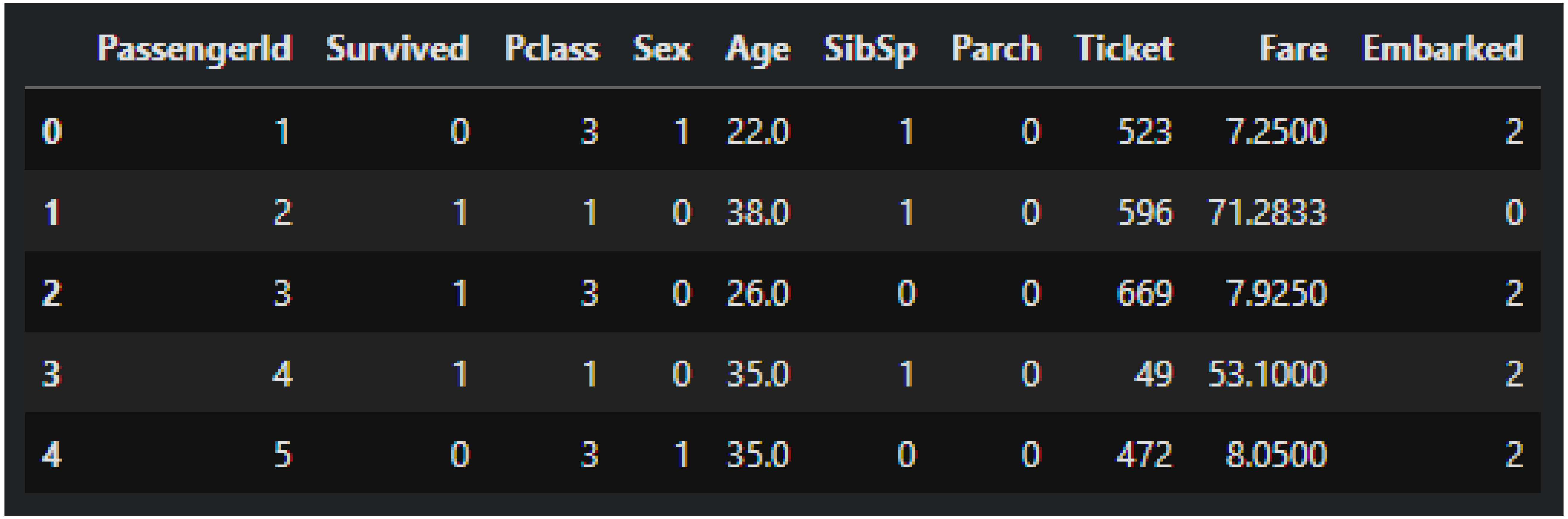


Penumpang kelas 3 paling banyak daripada kelas 1 dan 2



Usia penumpang paling banyak berada di usia mendekati 30

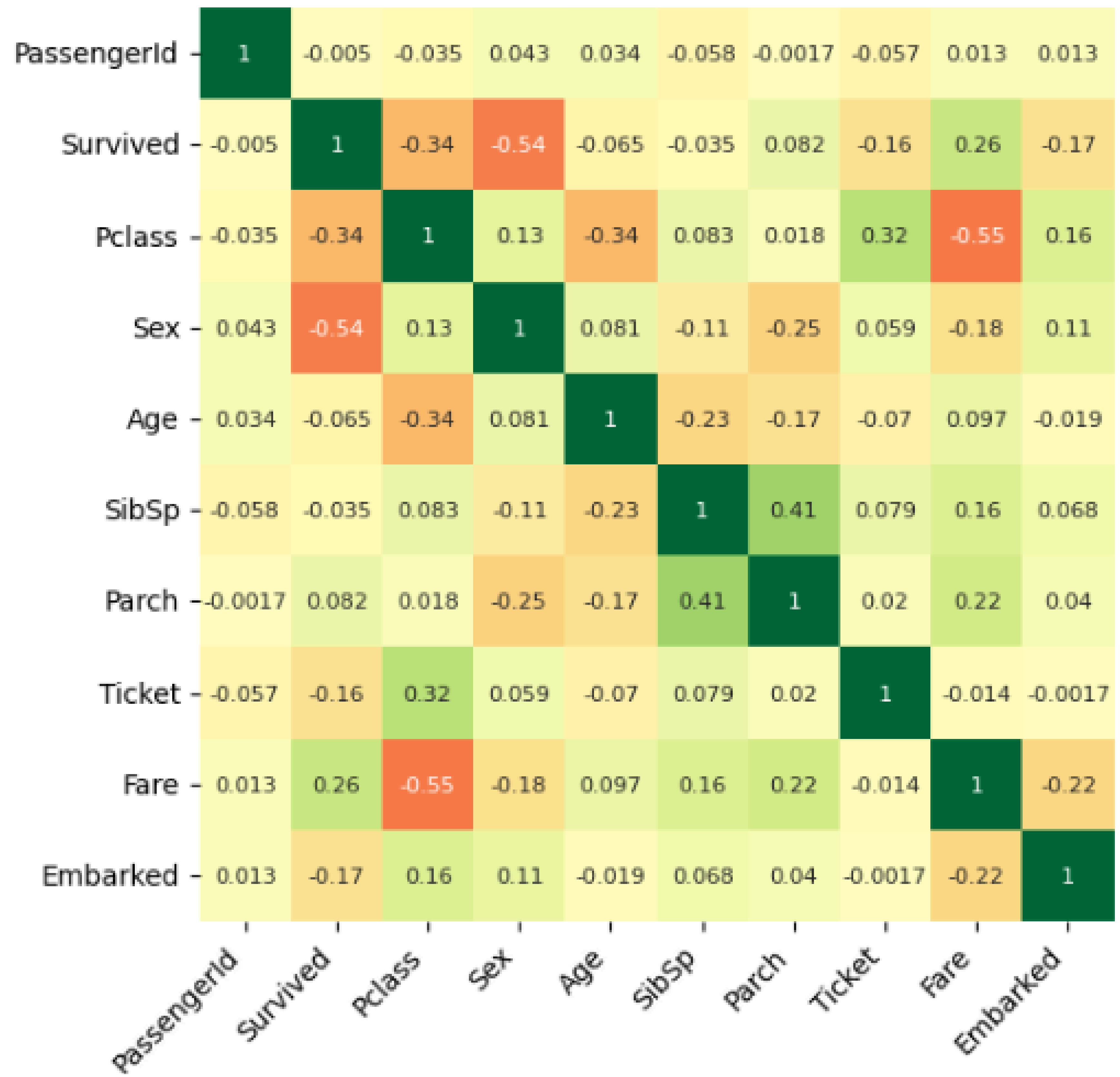
Feature Engineering



	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	1	22.0	1	0	523	7.2500	2
1	2	1	1	0	38.0	1	0	596	71.2833	0
2	3	1	3	0	26.0	0	0	669	7.9250	2
3	4	1	1	0	35.0	1	0	49	53.1000	2
4	5	0	3	1	35.0	0	0	472	8.0500	2

Feature engineering berfungsi untuk mengubah data kategorik menjadi numerik dengan modul LabelEncoder

Linear Correlation

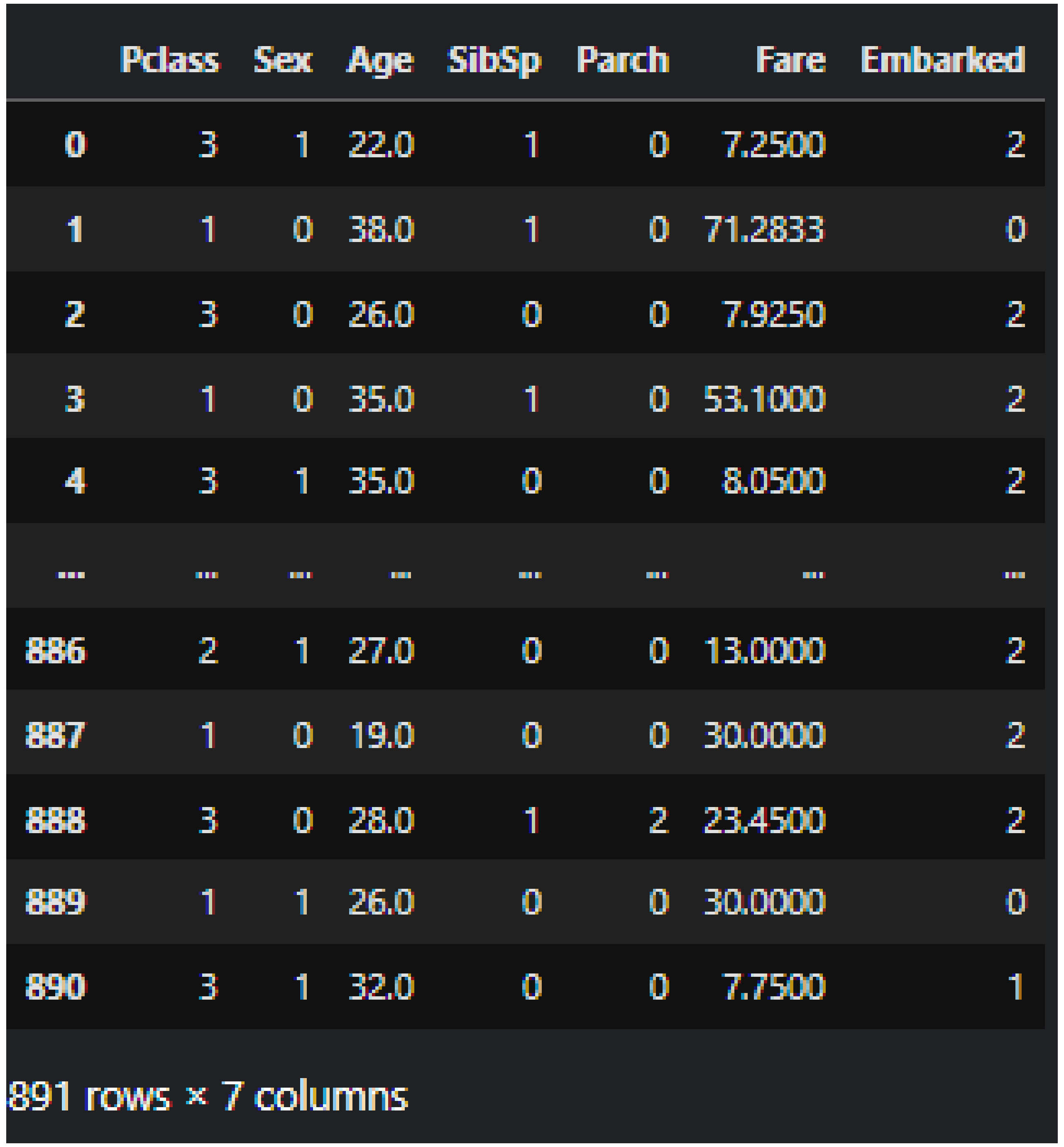


	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
PassengerId	1	-0.005	-0.035	0.043	0.034	-0.058	-0.0017	-0.057	0.013	0.013
Survived	-0.005	1	-0.34	-0.54	-0.065	-0.035	0.082	-0.16	0.26	-0.17
Pclass	-0.035	-0.34	1	0.13	-0.34	0.083	0.018	0.32	-0.55	0.16
Sex	0.043	-0.54	0.13	1	0.081	-0.11	-0.25	0.059	-0.18	0.11
Age	0.034	-0.065	-0.34	0.081	1	-0.23	-0.17	-0.07	0.097	-0.019
SibSp	-0.058	-0.035	0.083	-0.11	-0.23	1	0.41	0.079	0.16	0.068
Parch	-0.0017	0.082	0.018	-0.25	-0.17	0.41	1	0.02	0.22	0.04
Ticket	-0.057	-0.16	0.32	0.059	-0.07	0.079	0.02	1	-0.014	-0.0017
Fare	0.013	0.26	-0.55	-0.18	0.097	0.16	0.22	-0.014	1	-0.22
Embarked	0.013	-0.17	0.16	0.11	-0.019	0.068	0.04	-0.0017	-0.22	1

Grafik korelasi menunjukkan bahwa peluang hidup untuk selamat dipengaruhi oleh beberapa faktor, antara lain Parch dan Fare.

Sementara itu, Pclass dan Sex menjadi faktor yang memengaruhi ketidakselamatan penumpang

Linear Correlation

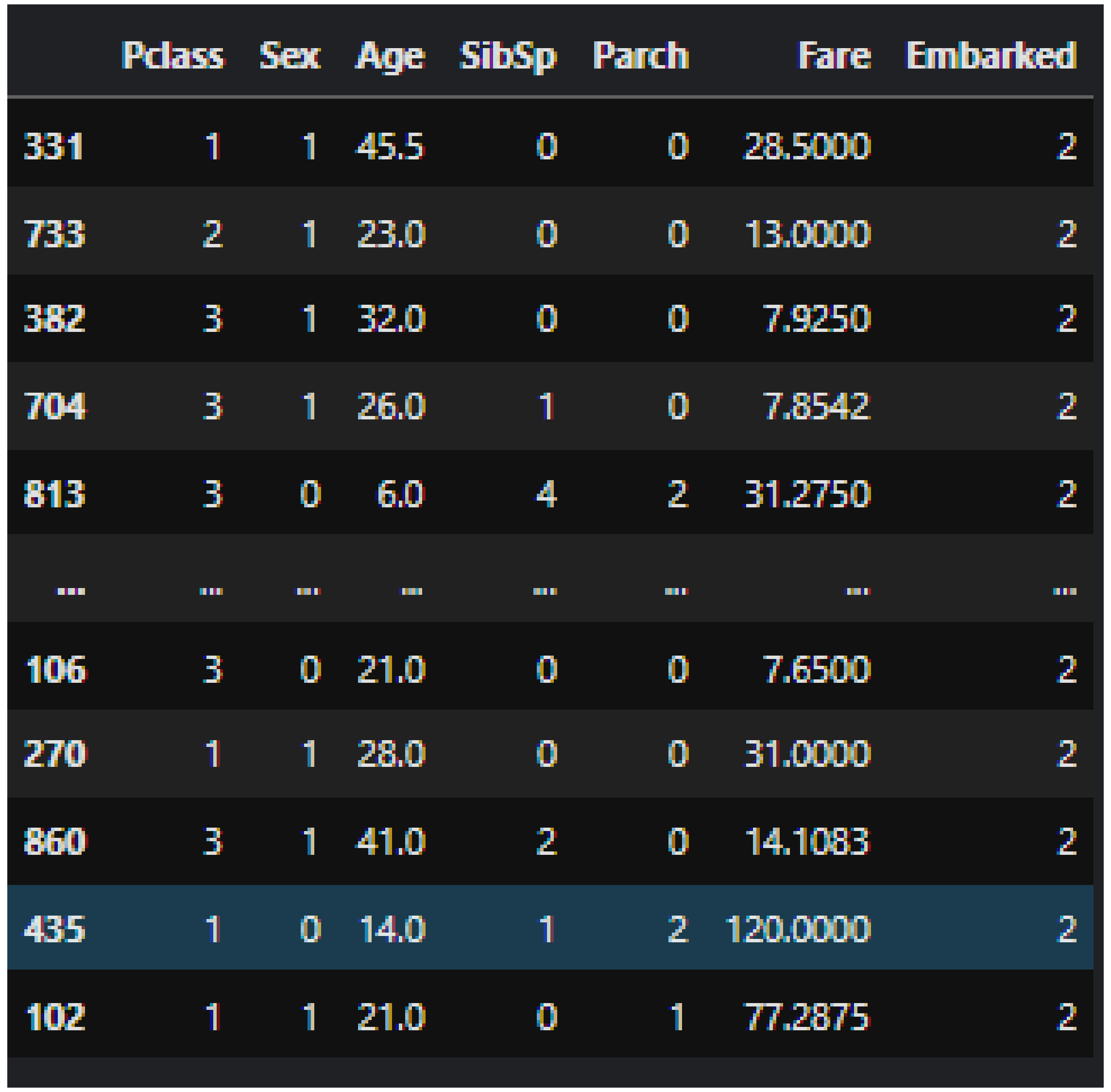


```
     Pclass  Sex  Age  SibSp  Parch     Fare  Embarked
0         3    1  22.0     1      0   7.2500         2
1         1    0  38.0     1      0  71.2833         0
2         3    0  26.0     0      0   7.9250         2
3         1    0  35.0     1      0  53.1000         2
4         3    1  35.0     0      0   8.0500         2
...     ...  ...   ...   ...    ...      ...       ...
886       2    1  27.0     0      0  13.0000         2
887       1    0  19.0     0      0  30.0000         2
888       3    0  28.0     1      2  23.4500         2
889       1    1  26.0     0      0  30.0000         0
890       3    1  32.0     0      0   7.7500         1

891 rows × 7 columns
```

seleksi fitur pada dataset ini digunakan untuk menyeleksi kolom-kolom yang fungsional

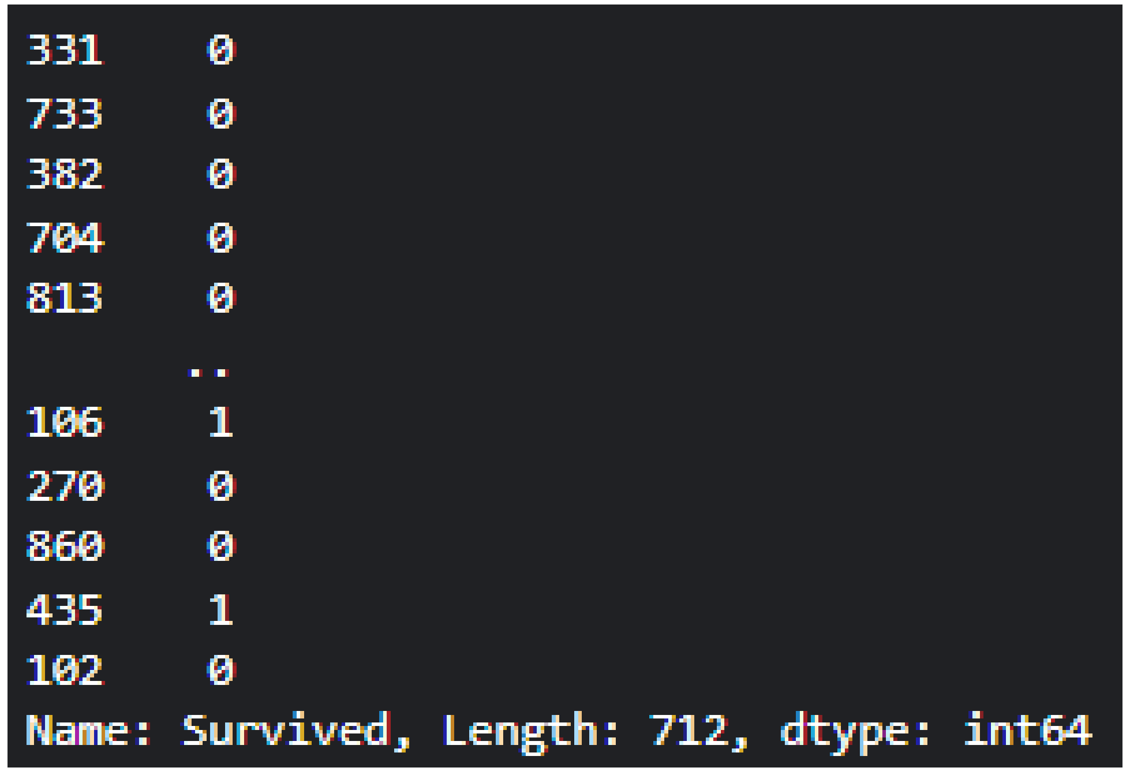
PassengerId	Survived
1	0
2	1
3	1
4	1
5	0
...	...
886	0
887	1
888	0
889	1
890	0



	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
331	1	1	45.5	0	0	28.5000	2
733	2	1	23.0	0	0	13.0000	2
382	3	1	32.0	0	0	7.9250	2
704	3	1	26.0	1	0	7.8542	2
813	3	0	6.0	4	2	31.2750	2
...	...	...	...	...	...	...	...
106	3	0	21.0	0	0	7.6500	2
270	1	1	28.0	0	0	31.0000	2
860	3	1	41.0	2	0	14.1083	2
435	1	0	14.0	1	2	120.0000	2
102	1	1	21.0	0	1	77.2875	2

Data Train (X_train)

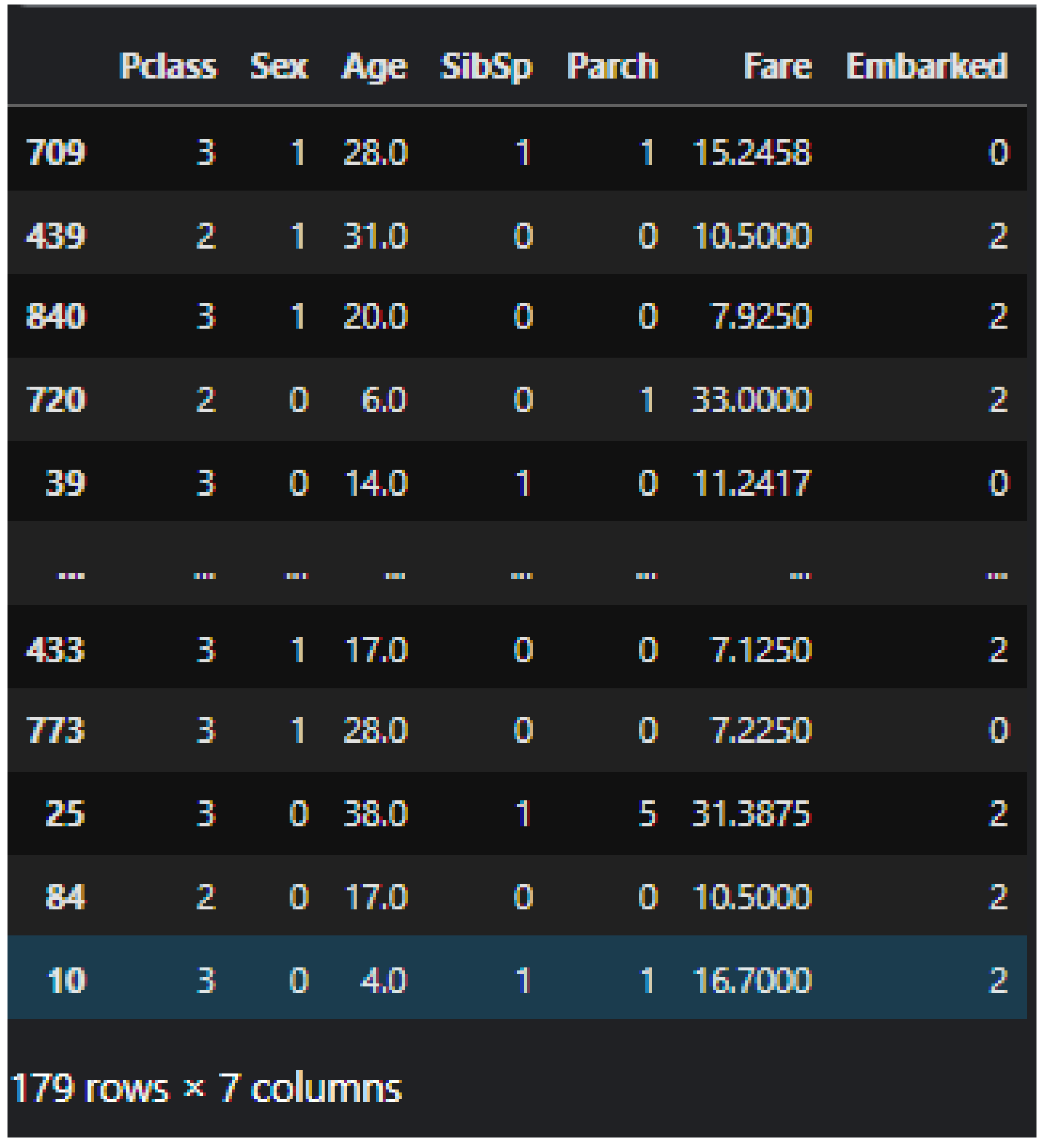
Modelling Random Forest



```
331    0
733    0
382    0
704    0
813    0
      ..
106    1
270    0
860    0
435    1
102    0
Name: Survived, Length: 712, dtype: int64
```

Data Train (y_train)

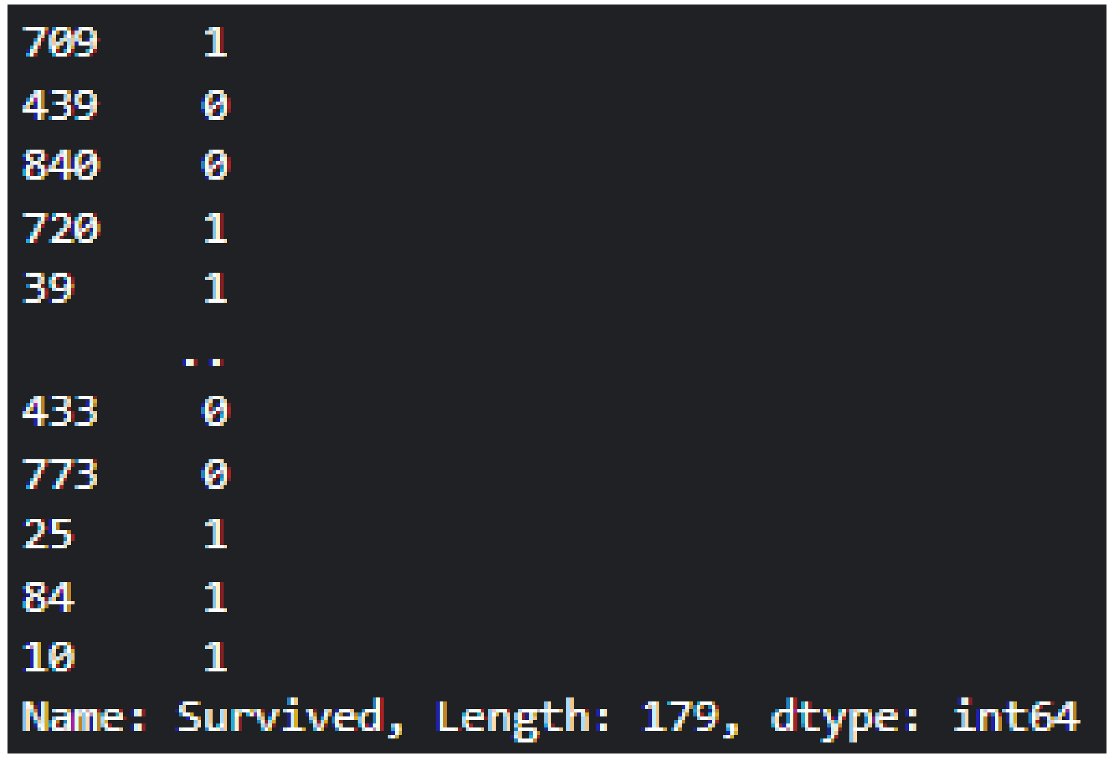
Modelling Logistic Regression



	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
709	3	1	28.0	1	1	15.2458	0
439	2	1	31.0	0	0	10.5000	2
840	3	1	20.0	0	0	7.9250	2
720	2	0	6.0	0	1	33.0000	2
39	3	0	14.0	1	0	11.2417	0
...	...	...	...	...	...	...	...
433	3	1	17.0	0	0	7.1250	2
773	3	1	28.0	0	0	7.2250	0
25	3	0	38.0	1	5	31.3875	2
84	2	0	17.0	0	0	10.5000	2
10	3	0	4.0	1	1	16.7000	2

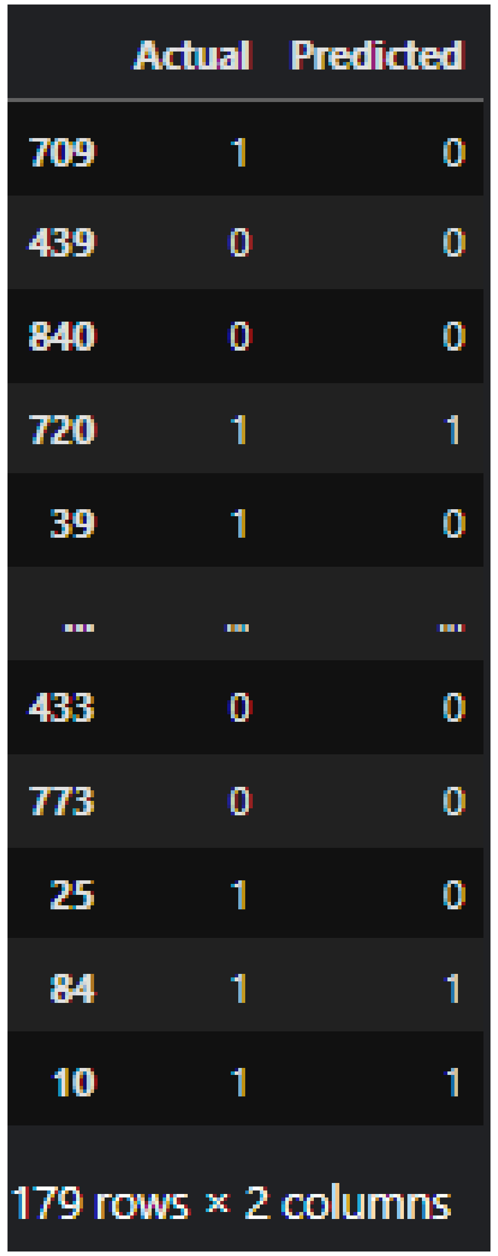
179 rows × 7 columns

Data Test (X_test)



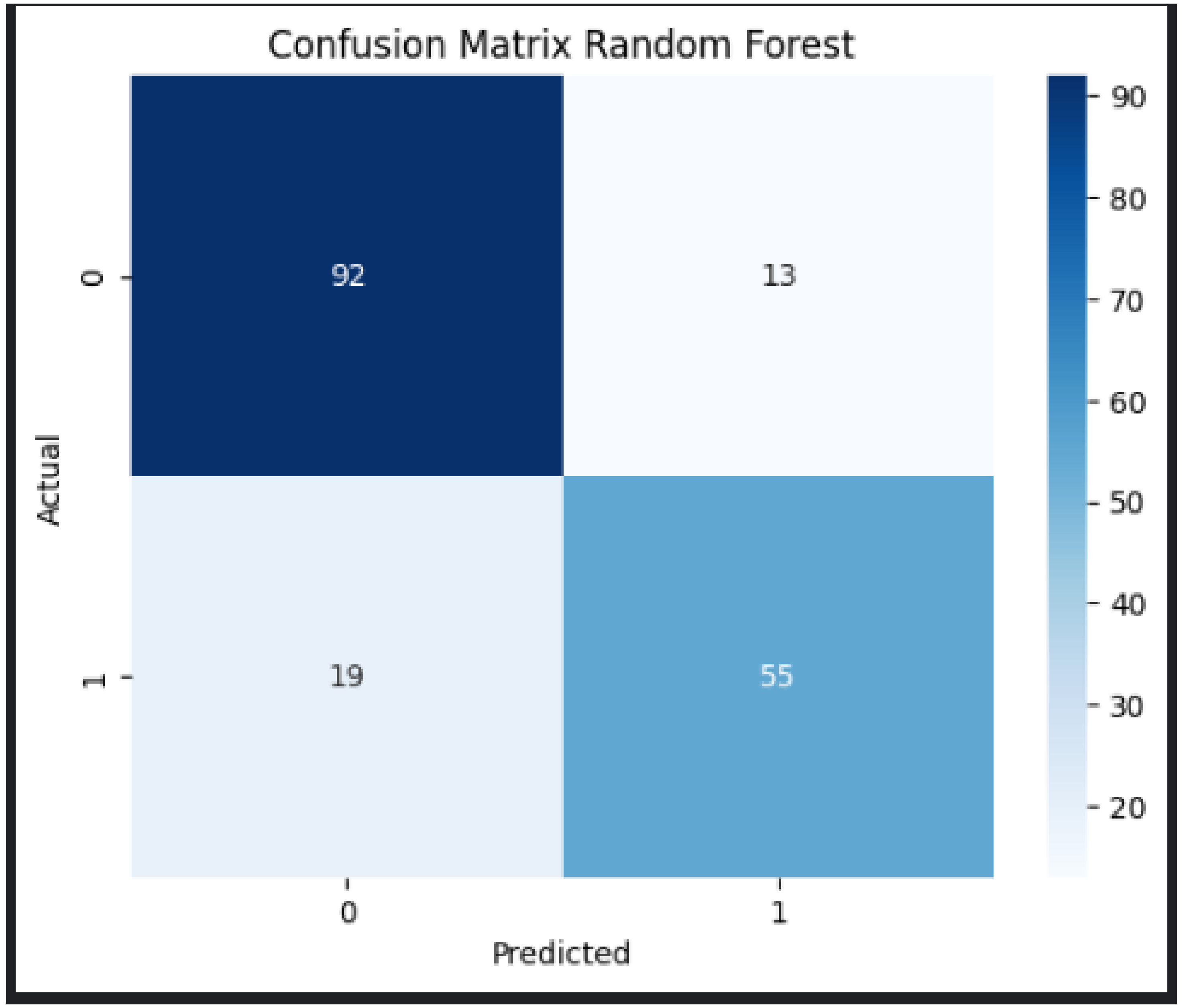
```
709     1
439     0
840     0
720     1
39      1
       ..
433     0
773     0
25      1
84      1
10      1
Name: Survived, Length: 179, dtype: int64
```

Data Train (y_test)

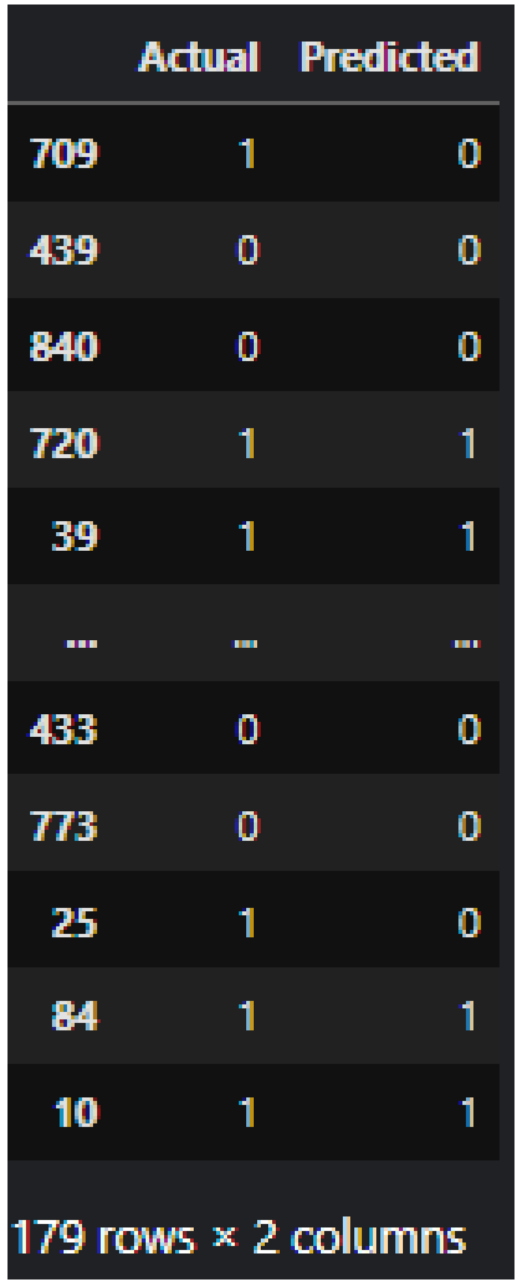


	Actual	Predicted
709	1	0
439	0	0
840	0	0
720	1	1
39	1	0
...	...	...
433	0	0
773	0	0
25	1	0
84	1	1
10	1	1

179 rows × 2 columns

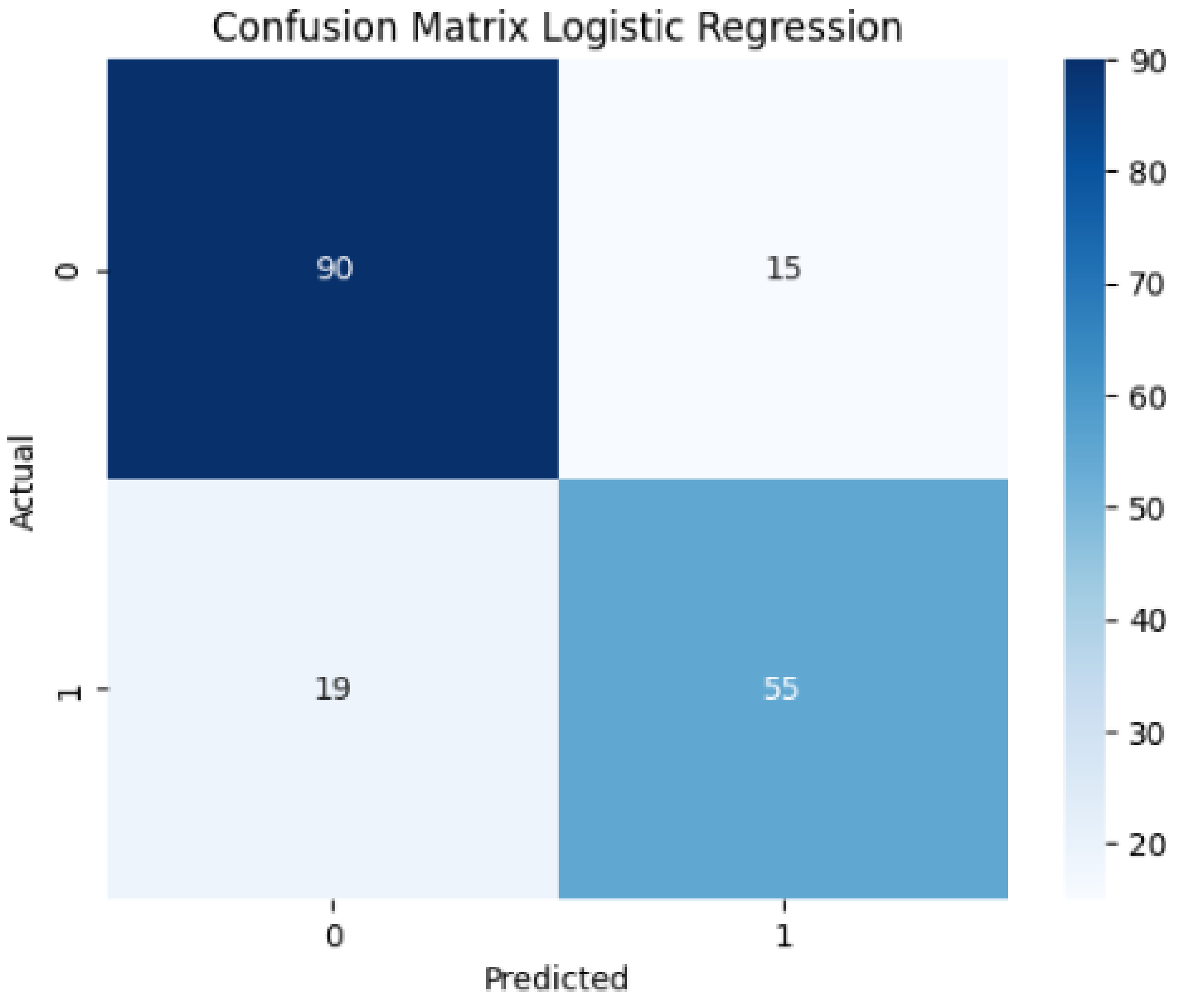


Evaluasi confusion matrix pada model Random Forest menunjukkan TP = 55, TN = 92, FP = 13, dan FN = 19



	Actual	Predicted
709	1	0
439	0	0
840	0	0
720	1	1
39	1	1
...	...	...
433	0	0
773	0	0
25	1	0
84	1	1
10	1	1

179 rows × 2 columns



Evaluasi confusion matrix pada model Logistic Regression menunjukkan TP = 55, TN = 92, FP = 13, dan FN = 19

Ukuran accuracy, precision dan recall digunakan untuk evaluasi performa machine learning

Accuracy	82%
Precision	81%
Recall	74%

Random Forest

Accuracy	81%
Precision	79%
Recall	74%

Logistic Regression

Model Random Forest memiliki accuracy dan precision yang setingkat lebih unggul dibanding Logistic Regression walaupun selisihnya tidak terlampau jauh