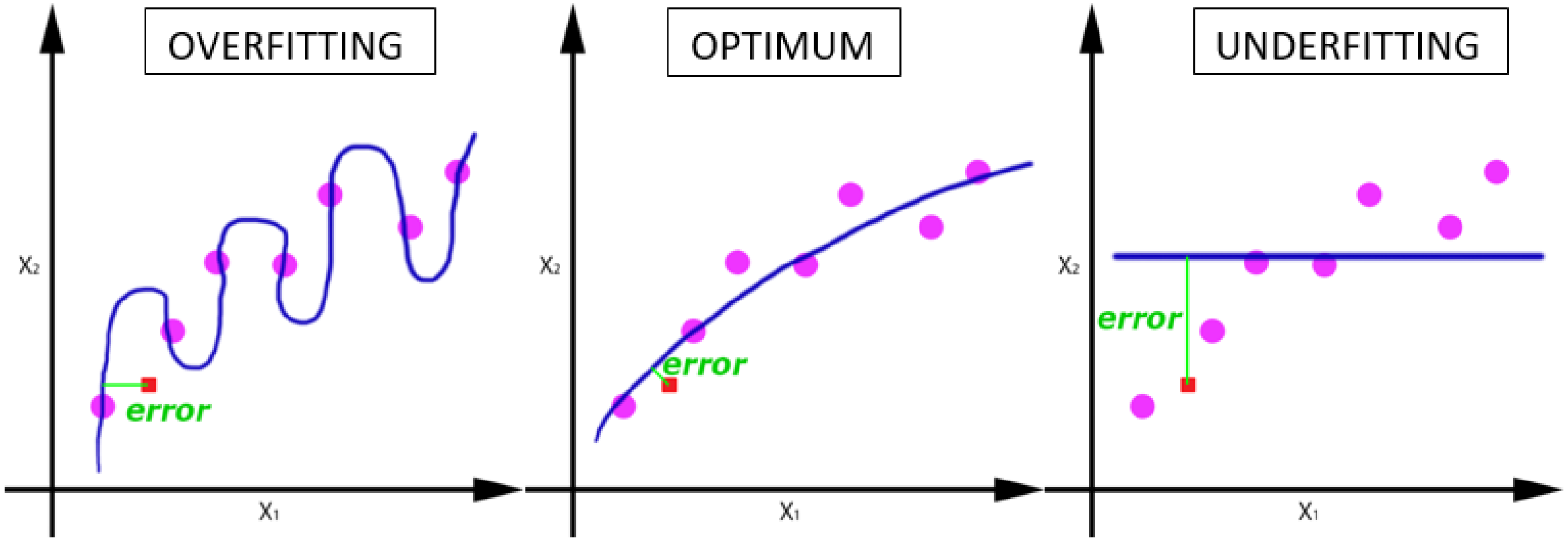
Wtorek 27.04

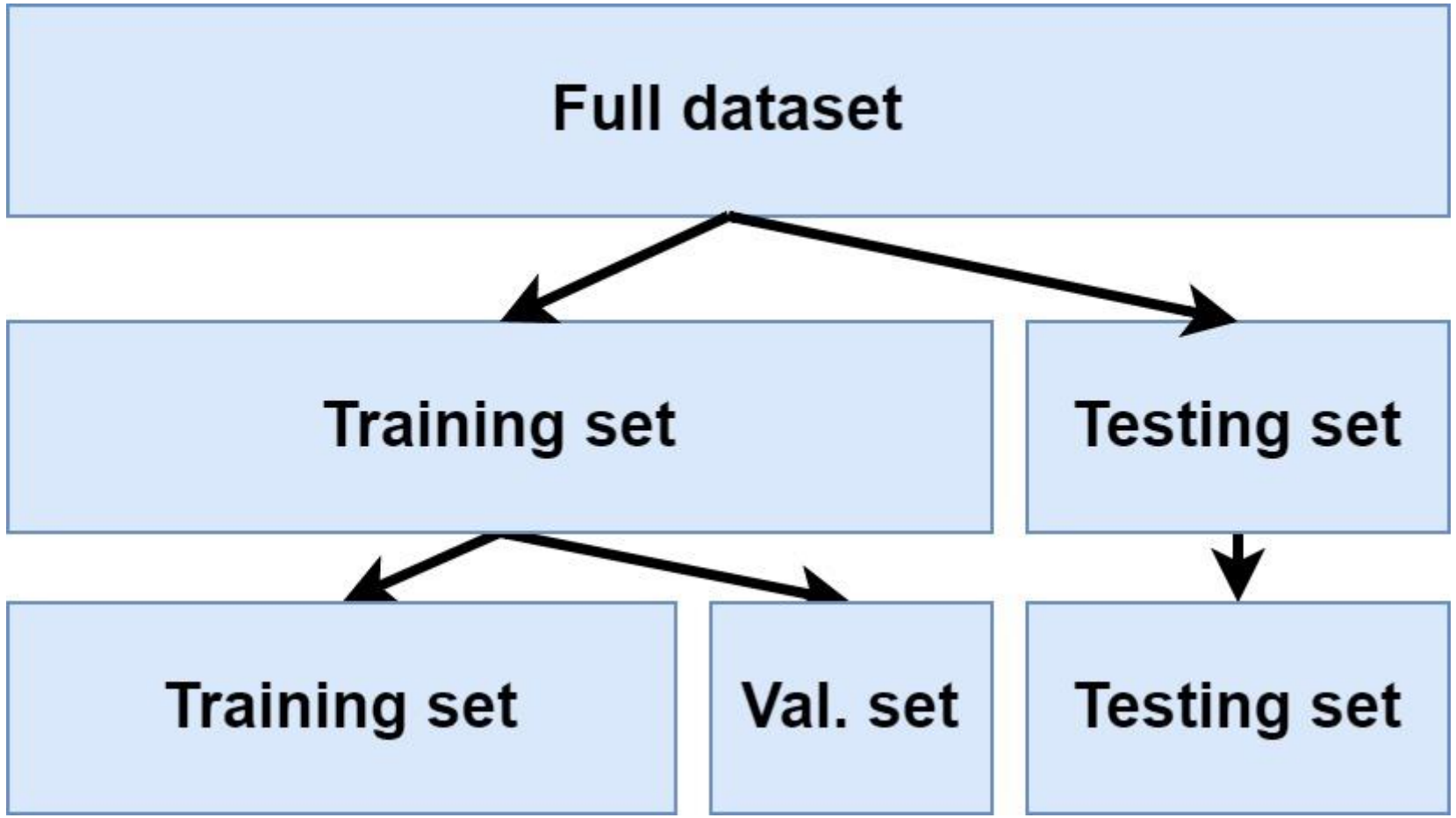
UMCS.ai

Klasyfikacja zbiorów – problem Titanica

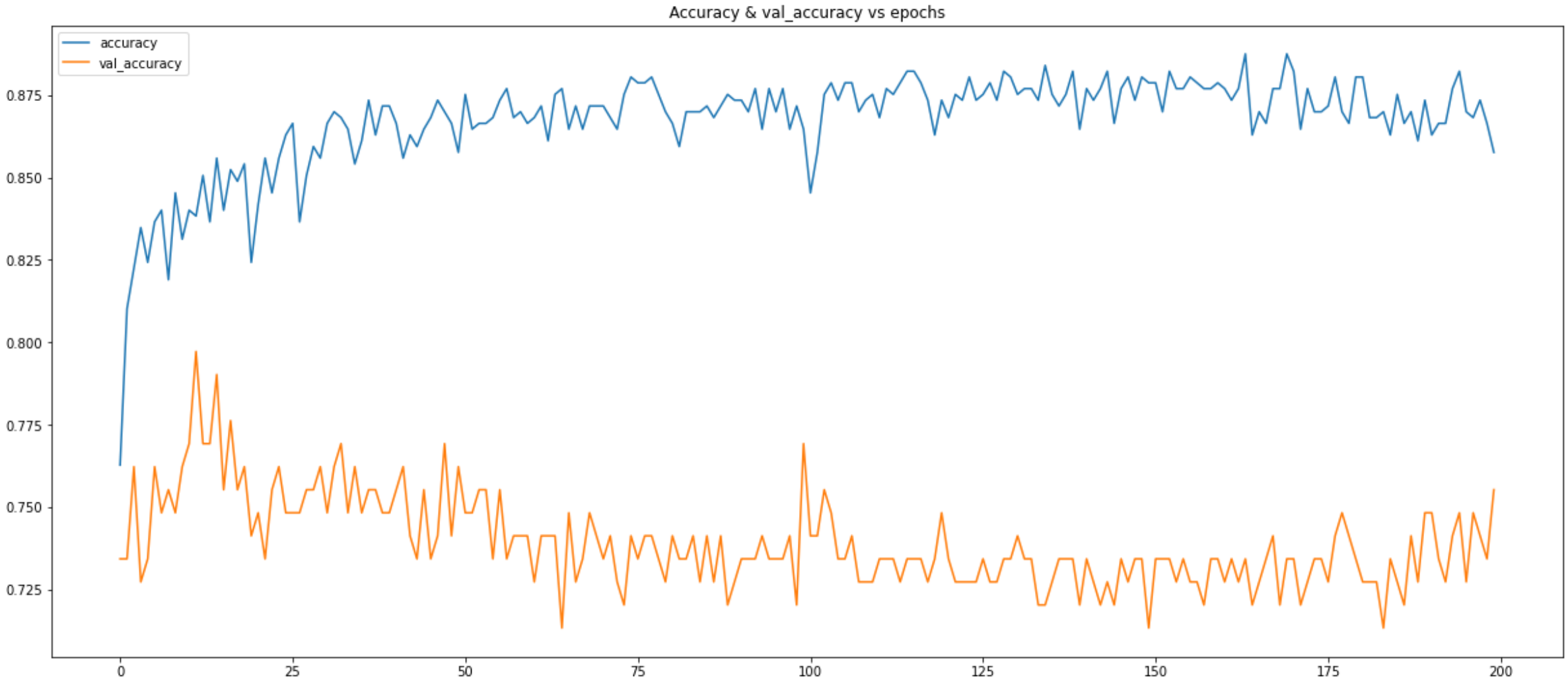
Overfitting



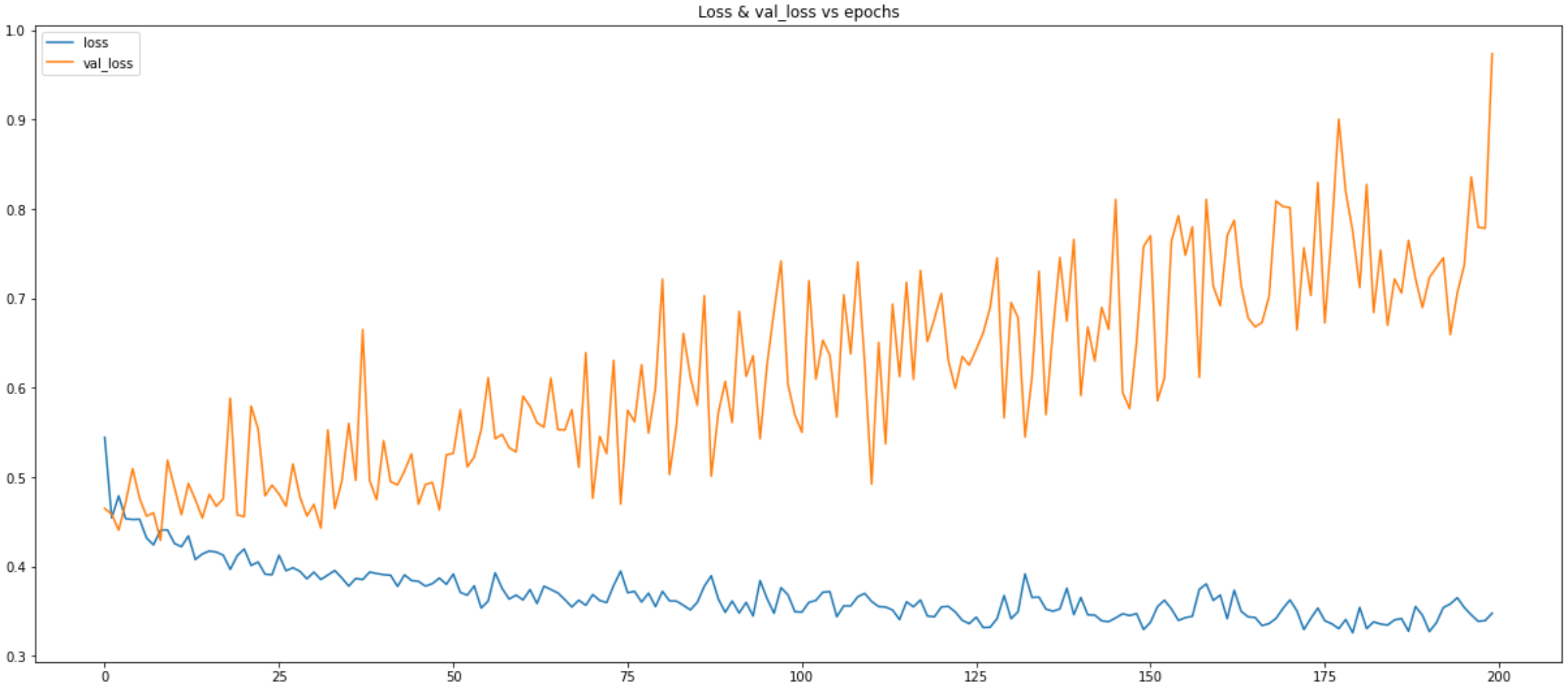
Overfitting – jak uniknąć?



Wartość accuracy na przestrzeni epok



Wartość loss na przestrzeni epok



Overfitting – jak uniknąć?

```
from tensorflow.keras.callbacks import EarlyStopping

history = model.fit(x=X_train, y=Y_train, epochs=200, batch_size=16,
                    validation_split=0.2, callbacks=[EarlyStopping(monitor='loss', patience=2)])
```

history.history[‚loss’] = wartości loss dla zbioru train
history.history[‚val_loss’] = wartości loss dla zbioru valid

Funkcje straty

Klasyfikacja binarna:
używamy aktywacji **sigmoid** oraz funkcji straty **binary crossentropy**

Klasyfikacja wieloklasowa:
używamy aktywacji **softmax** orax funkcji straty **categorical crossentropy**

Regresja:
MAE – mean absolute error
MSE – mean square error
MAPE – mean absolute percentage error

Można doczytać trochę więcej: https://neptune.ai/blog/keras-loss-functions

ZADANIE

- Przeanalizuj zbiór danych Titanic
- Wykonaj transformację kolumn do odpowiedniej postaci
- Zamodeluj dane (np. MLP – Keras, [KNN, SVM, Random Forest] – sklearn)

Kopalnia wiedzy:

- kaggle.com
- medium.com
- towardsdatascience.com

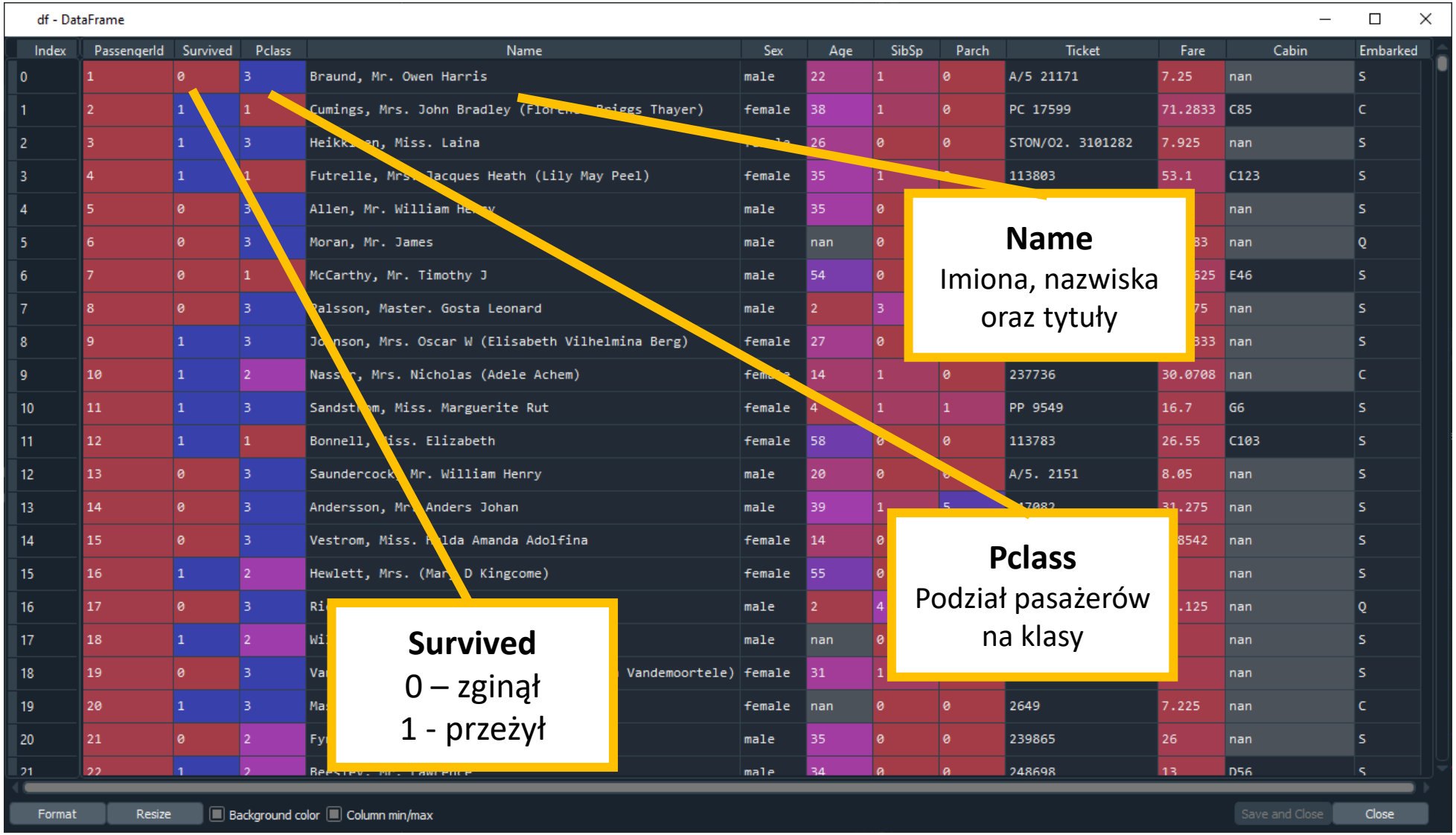
Analiza problemu

df - DataFrame

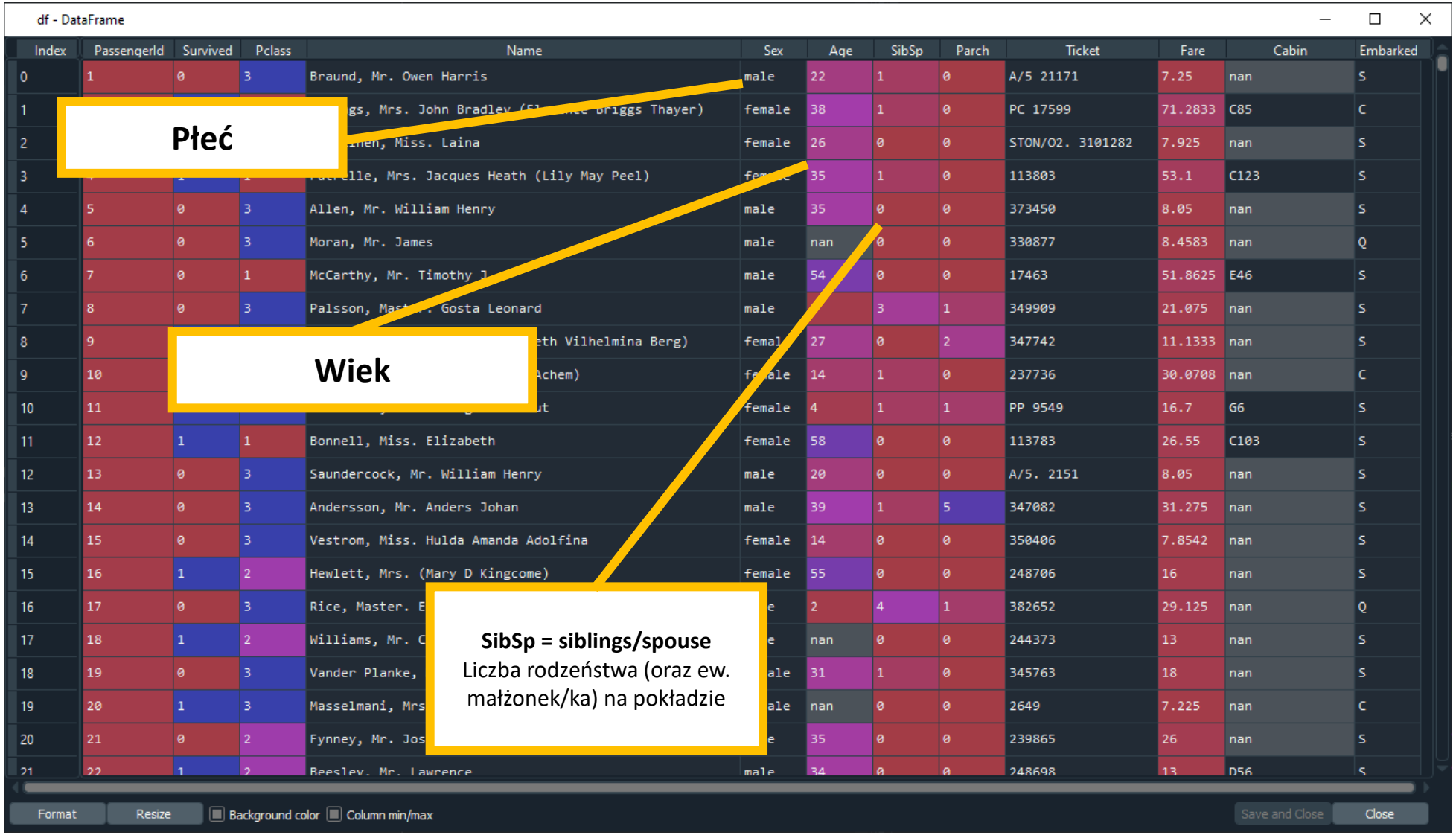
Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	nan	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	nan	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	nan	S
5	6	0	3	Moran, Mr. James	male	nan	0	0	330877	8.4583	nan	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	nan	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	nan	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	nan	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	nan	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275	nan	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	nan	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16	nan	S
16	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125	nan	Q
17	18	1	2	Williams, Mr. Charles Eugene	male	nan	0	0	244373	13	nan	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18	nan	S
19	20	1	3	Masselmani, Mrs. Fatima	female	nan	0	0	2649	7.225	nan	C
20	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26	nan	S
21	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S

Format Resize Background color Column min/max Save and Close Close

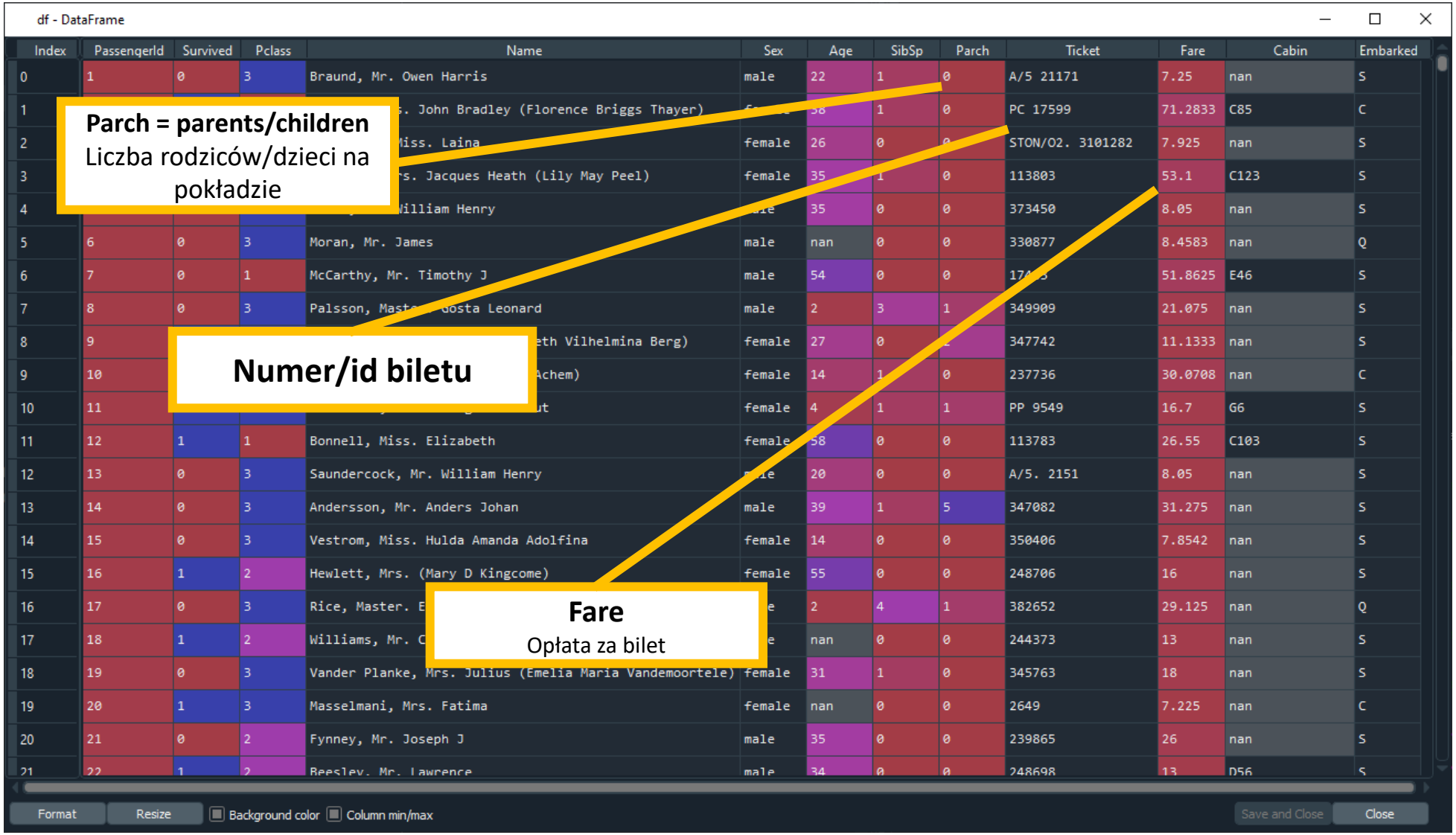
Analiza problemu



Analiza problemu

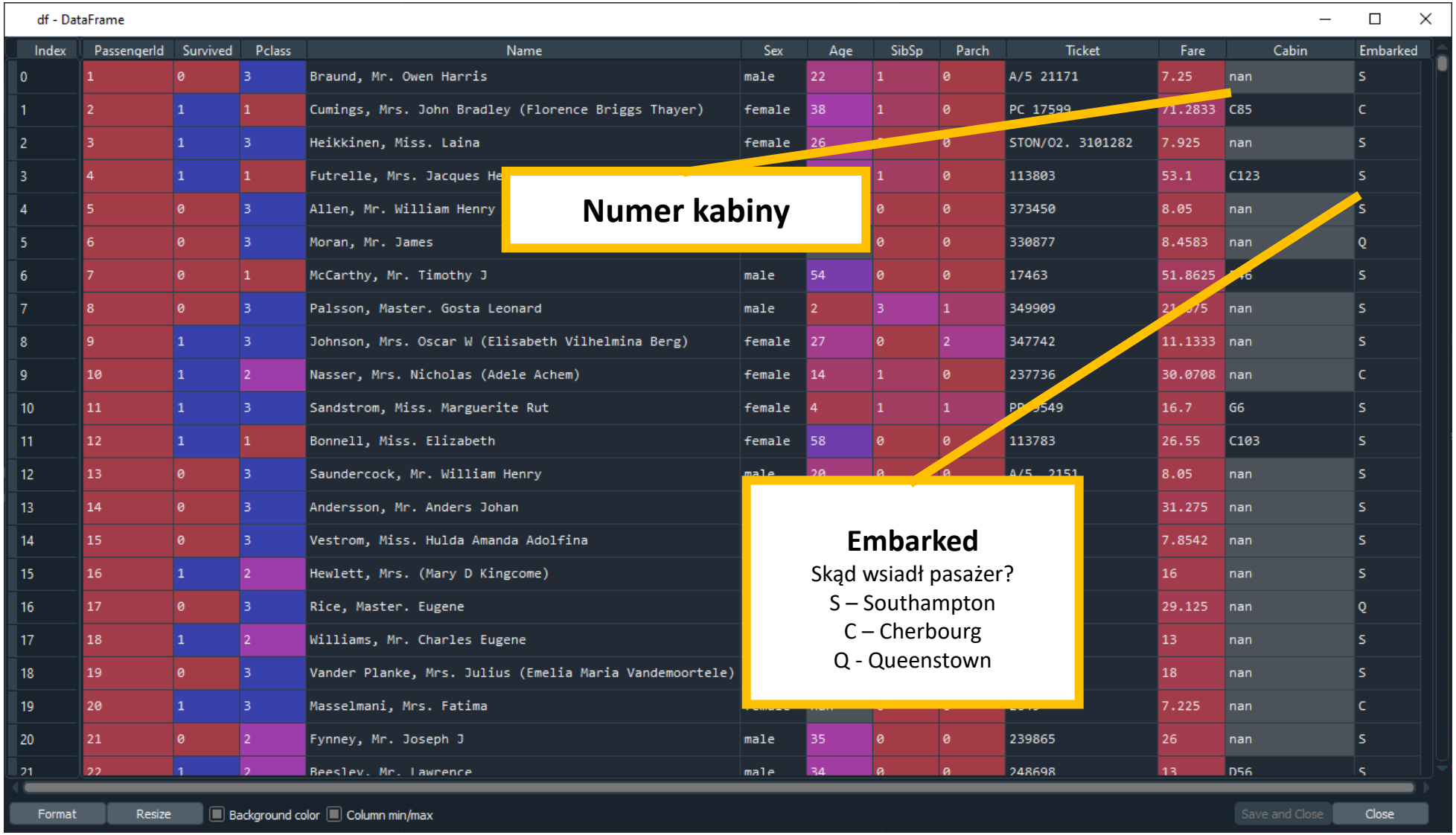


Analiza problemu



Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	nan	S
1				. John Bradley (Florence Briggs Thayer)		38	1	0	PC 17599	71.2833	C85	C
2				iss. Laina	female	26	0	0	STON/O2. 3101282	7.925	nan	S
3				s. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
4				illiam Henry		35	0	0	373450	8.05	nan	S
5	6	0	3	Moran, Mr. James	male	nan	0	0	330877	8.4583	nan	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	nan	S
8	9			eth Vilhelmina Berg)	female	27	0	2	347742	11.1333	nan	S
9	10			Achem)	female	14	1	0	237736	30.0708	nan	C
10	11			ut	female	4	1	1	PP 9549	16.7	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	nan	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275	nan	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	nan	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16	nan	S
16	17	0	3	Rice, Master. E	male	2	4	1	382652	29.125	nan	Q
17	18	1	2	Williams, Mr. C	male	nan	0	0	244373	13	nan	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18	nan	S
19	20	1	3	Masselmani, Mrs. Fatima	female	nan	0	0	2649	7.225	nan	C
20	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26	nan	S
21	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S

Analiza problemu



Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	nan	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	nan	S
3	4	1	1	Futrelle, Mrs. Jacques He			1	0	113803	53.1	C123	S
4	5	0	3	Allen, Mr. William Henry			0	0	373450	8.05	nan	S
5	6	0	3	Moran, Mr. James			0	0	330877	8.4583	nan	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	nan	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	nan	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	nan	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	nan	S
13	14	0	3	Andersson, Mr. Anders Johan						31.275	nan	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina						7.8542	nan	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)						16	nan	S
16	17	0	3	Rice, Master. Eugene						29.125	nan	Q
17	18	1	2	Williams, Mr. Charles Eugene						13	nan	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)						18	nan	S
19	20	1	3	Masselmani, Mrs. Fatima						7.225	nan	C
20	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26	nan	S
21	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S

Analiza danych

Read dataset

```
df = pd.read_csv('data.csv')
```

Analyse data

Print head / tail of dataset

```
print("Head:\n", df.head(n=5))
print("Tail:\n", df.tail(n=10))
```

Get column names

```
print("Column names: ", df.columns)
```

Describe dataframe

```
print(df.describe())
```

Check if dataframe has NaN values

```
print(df.isna())
print(df.isna().any())
```

Print feature correlations

```
print(df[['Sex', 'Survived']].groupby(['Sex'], as_index=False).mean())
print(df[['Pclass', 'Survived']].groupby(['Pclass'], as_index=False).mean())
print(df[['SibSp', 'Survived']].groupby(['SibSp'], as_index=False).mean())
```

Preprocessing danych

Drop column / row

```
df = df.drop(['PassengerId', 'Ticket'], axis=1)      # Column
df = df.drop(0, axis=0)                              # Row #0
```

Drop rows with any NaN value

```
df = df.dropna()
```

Fill NaN values

```
df['Age'] = df['Age'].fillna(df['Age'].dropna().median())      # Fill 'Age' with median
df['Age'] = df['Age'].fillna(method='ffill')                   # ffill = forward fill ; bfill = backward fill
```

Categorical feature normalization

```
df['Embarked'] = LabelEncoder().fit_transform(df['Embarked'])
```

Cast values to specified format

```
df['Age'] = df['Age'].astype('int')
```

New features (using list comprehension)

```
df['Age_child'] = [int(x<18) for x in df['Age’]]
df['Age_adult'] = [int(x>=18) for x in df['Age’]]
```

Scale values

```
from sklearn.preprocessing import MinMaxScaler
df[['Age', 'Fare']] = MinMaxScaler().fit_transform(df[['Age', 'Fare']])
```