

Julia Chmaj

Link do repozytorium: <https://github.com/Jul22/IO-projekt2>

SPRAWOZDANIE 2

Wybrana przeze mnie baza do analizy to „Telco Customer Churn”. Składa się z 7043 wierszy i 21 kolumn. Kolumny zawierają informacje o klientach firmy dostarczającej usługi internetowe i telefoniczne.

Celem projektu jest zbadanie klasyfikatorów, które najlepiej wskażą przewidywania dotyczące rezygnacji klientów z dostarczanych usług- kolumna „Churn” .

gender	Female	Male	Male	Male	Female	Female	Male
SeniorCitizen	0	0	0	0	0	0	0
Partner	Yes	No	No	No	No	No	No
Dependents	No	No	No	No	No	No	Yes
tenure	1	34	2	45	2	8	22
PhoneService	No	Yes	Yes	No	Yes	Yes	Yes
MultipleLines	No phone service	No	No	No phone service	No	Yes	Yes
InternetService	DSL	DSL	DSL	DSL	Fiber optic	Fiber optic	Fiber optic
OnlineSecurity	No	Yes	Yes	Yes	No	No	No
OnlineBackup	Yes	No	Yes	No	No	No	Yes
DeviceProtection	No	Yes	No	Yes	No	Yes	No
TechSupport	No	No	No	Yes	No	No	No
StreamingTV	No	No	No	No	No	Yes	Yes
StreamingMovies	No	No	No	No	No	Yes	No
Contract	Month-to-month	One year	Month-to-month	One year	Month-to-month	Month-to-month	Month-to-month
PaperlessBilling	Yes	No	Yes	No	Yes	Yes	Yes
PaymentMethod	Electronic check	Mailed check	Mailed check	Bank transfer (automatic)	Electronic check	Electronic check	Credit card (automatic)
MonthlyCharges	29.85	56.95	53.85	42.3	70.7	99.65	89.1
TotalCharges	29.85	1889.5	108.15	1840.75	151.65	820.5	1949.4
Churn	No	No	Yes	No	Yes	Yes	No

Postanowiłam usunąć kolumnę z ID klienta, ponieważ nic nie wносиła do klasyfikacji, oraz zamieniłam typ kolumny „SeniorCitizen” na kategorię, a „TotalCharges” na float.

TYP DANYCH

gender	object
SeniorCitizen	category
Partner	object
Dependents	object
tenure	int64
PhoneService	object
MultipleLines	object
InternetService	object
OnlineSecurity	object
OnlineBackup	object
DeviceProtection	object
TechSupport	object
StreamingTV	object
StreamingMovies	object
Contract	object
PaperlessBilling	object
PaymentMethod	object
MonthlyCharges	float64
TotalCharges	float64
Churn	object

PUSTE WARTOŚCI

```
df.isnull().sum()
```

gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	11
Churn	0

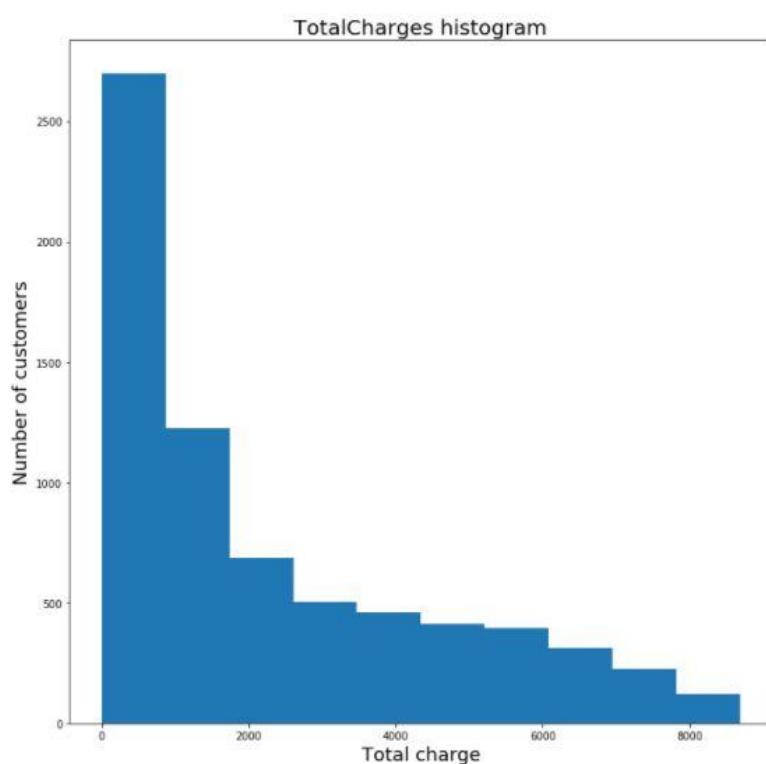
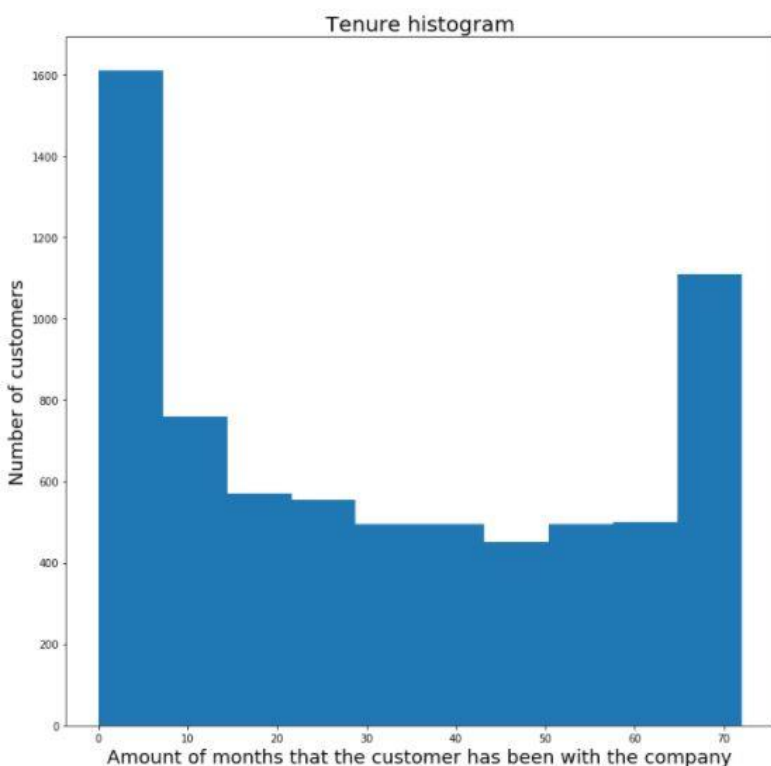
Tylko w jednej kolumnie znajduję braki danych i zastępuję je zerami.

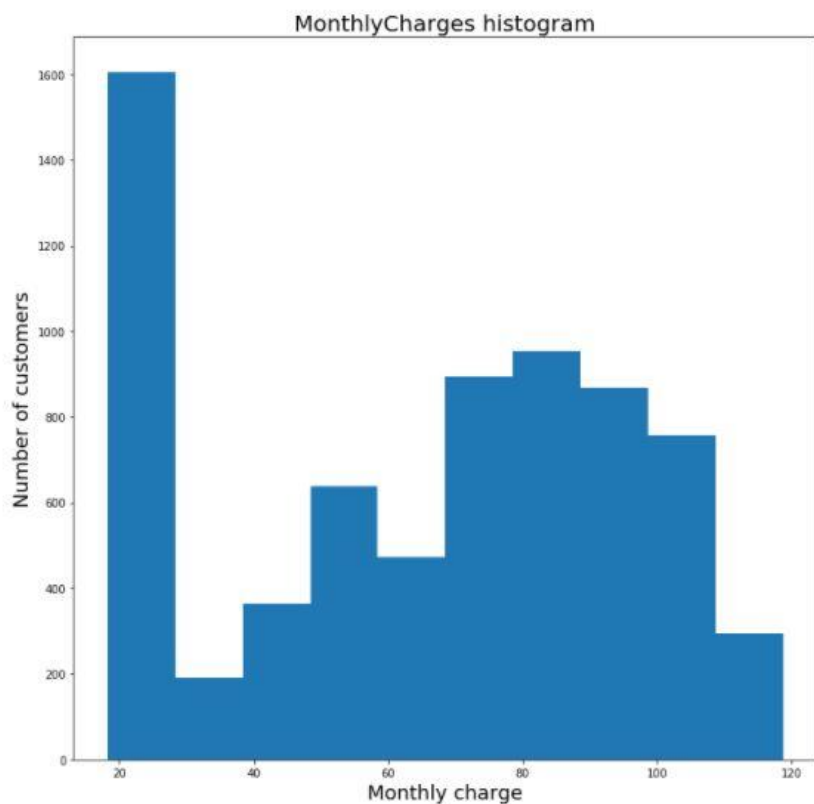
Aby wyłuskać podstawowe informacje o danych (średnia, częstość występowania odpowiedzi, max, min) podzieliłam dane na numeryczne i kategoryczne.

```
numeric_ds.describe()
```

	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000
mean	32.371149	64.761692	2279.734304
std	24.559481	30.090047	2266.794470
min	0.000000	18.250000	0.000000
25%	9.000000	35.500000	398.550000
50%	29.000000	70.350000	1394.550000
75%	55.000000	89.850000	3786.600000
max	72.000000	118.750000	8684.800000

Histogramy danych numerycznych:

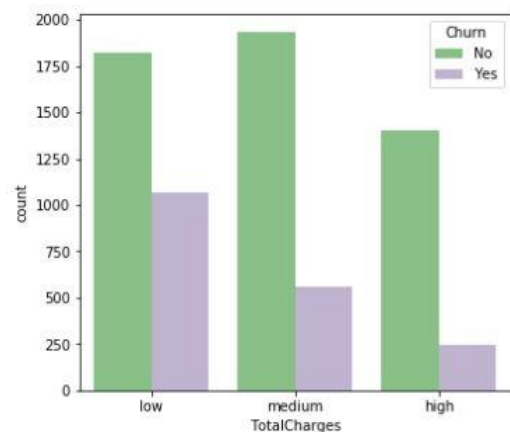
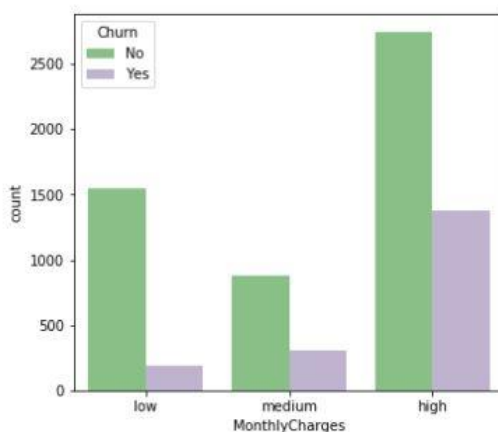
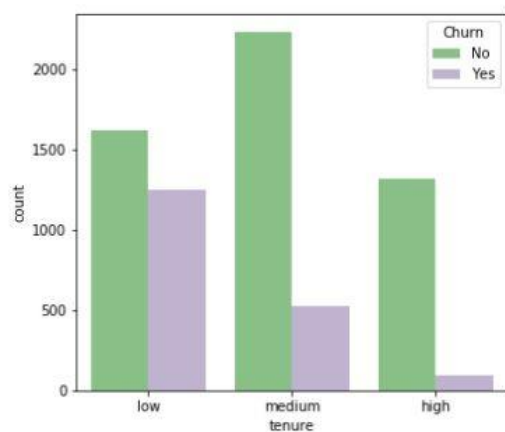




```
objects_ds.describe().T
```

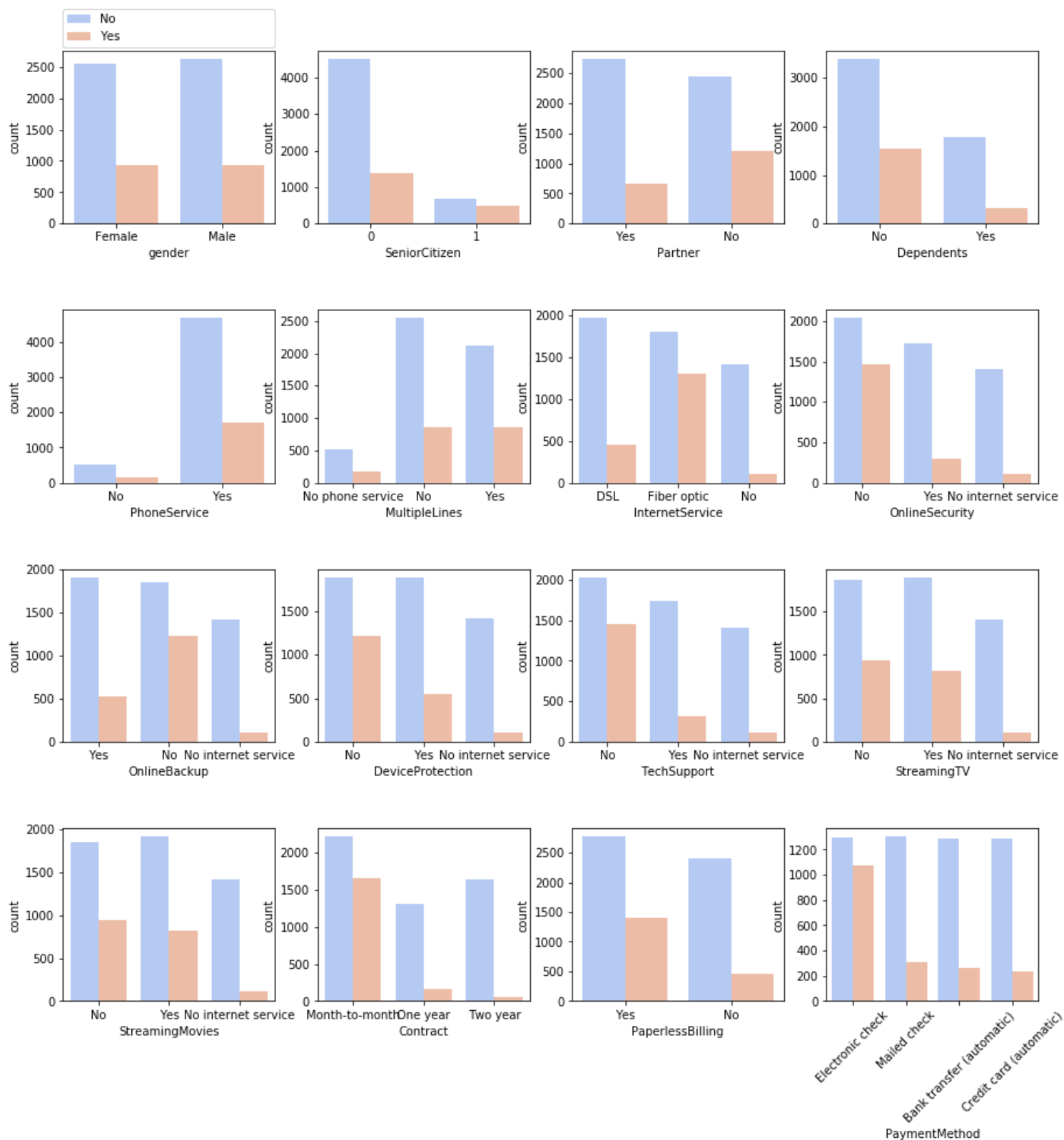
	count	unique	top	freq
gender	7043	2	Male	3555
SeniorCitizen	7043	2	0	5901
Partner	7043	2	No	3641
Dependents	7043	2	No	4933
PhoneService	7043	2	Yes	6361
MultipleLines	7043	3	No	3390
InternetService	7043	3	Fiber optic	3096
OnlineSecurity	7043	3	No	3498
OnlineBackup	7043	3	No	3088
DeviceProtection	7043	3	No	3095
TechSupport	7043	3	No	3473
StreamingTV	7043	3	No	2810
StreamingMovies	7043	3	No	2785
Contract	7043	3	Month-to-month	3875
PaperlessBilling	7043	2	Yes	4171
PaymentMethod	7043	4	Electronic check	2365
Churn	7043	2	No	5174

W celu lepszego zobrazowania wartości danych numerycznych podzieliłam je na przedziały low/medium/high.



Histogramy danych kategorycznych:

Wskazują ile osób zrezygnowało z usług lub zostało, w zależności od danej cechy.

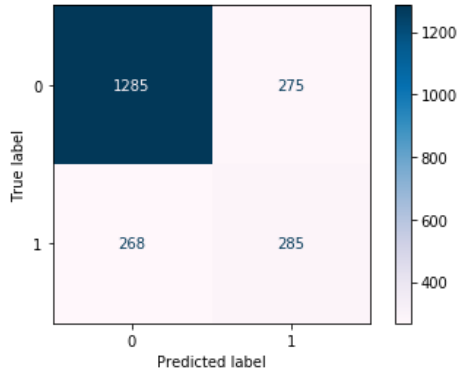
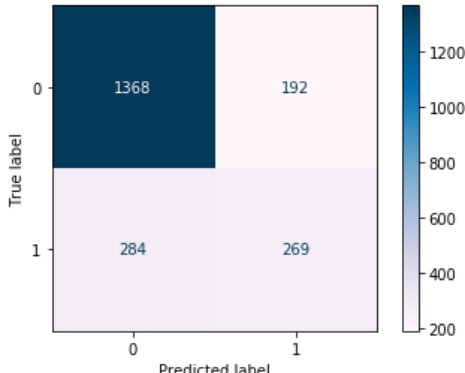


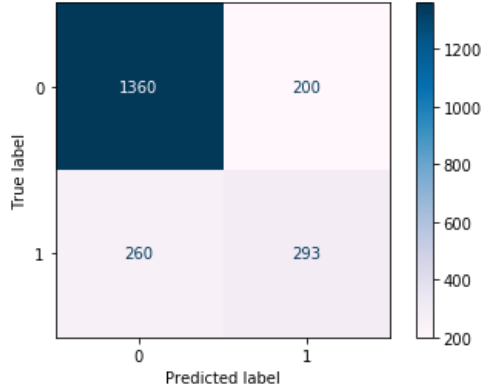
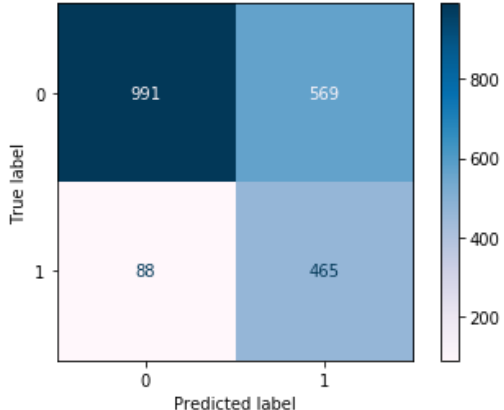
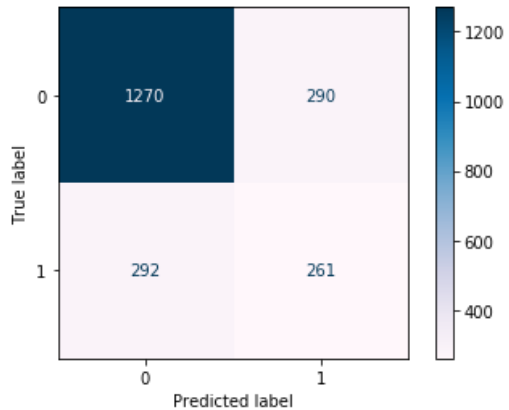
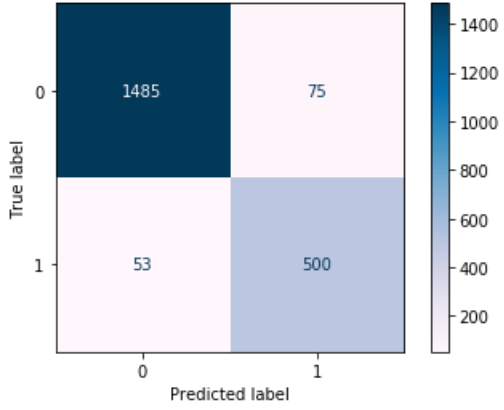
Baza danych do klasyfikacji składa się z danych numerycznych, które zostały podzielone na przedziały oraz danych katagorycznych rozdzielonych na kolumny, przez co wartości są 0-1.

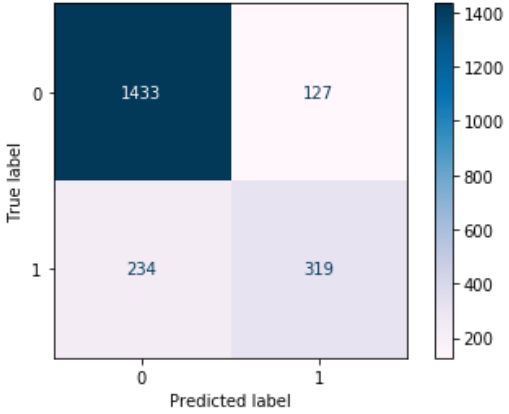
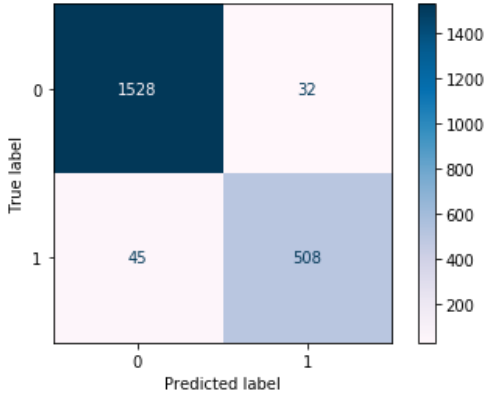
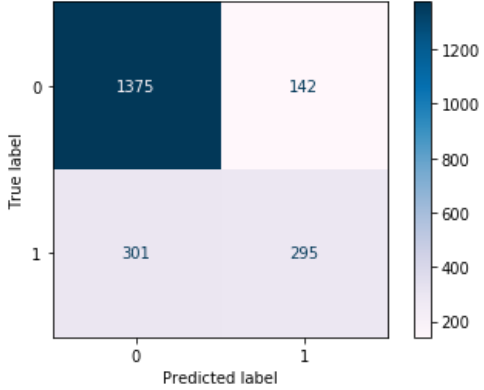
	0	1	2	3	4	5	6	7	8	9	...	7033	7034	7035	7036	7037	7038	7039	7040	7041	7042
tenure_high	0	0	0	0	0	0	0	0	0	1	...	0	1	0	0	1	0	1	0	0	1
tenure_low	1	0	1	0	1	1	0	1	0	0	...	0	0	1	1	0	0	0	1	1	0
tenure_medium	0	1	0	1	0	0	1	0	1	0	...	1	0	0	0	0	1	0	0	0	0
MonthlyCharges_high	0	0	0	0	1	1	1	0	1	0	...	1	1	1	1	0	1	1	0	1	1
MonthlyCharges_low	1	0	0	0	0	0	0	1	0	0	...	0	0	0	0	1	0	0	1	0	0
MonthlyCharges_medium	0	1	1	1	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
TotalCharges_high	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	1	0	0	1
TotalCharges_low	1	0	1	0	1	1	0	1	0	0	...	0	0	0	1	0	0	0	1	1	0
TotalCharges_medium	0	1	0	1	0	0	1	0	1	1	...	1	0	1	0	1	1	0	0	0	0
gender_Female	1	0	0	0	1	1	0	1	1	0	...	0	1	0	1	1	0	1	1	0	0
gender_Male	0	1	1	1	0	0	1	0	0	1	...	1	0	1	0	0	1	0	0	1	1
SeniorCitizen_0	1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1	1	1	1	0	1
SeniorCitizen_1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
Partner_No	0	1	1	1	1	1	1	1	0	1	...	1	1	1	1	1	0	0	0	0	1
Partner_Yes	1	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	1	1	1	1	0
Dependents_No	1	1	1	1	1	1	0	1	1	0	...	1	1	1	1	1	0	0	0	1	1
Dependents_Yes	0	0	0	0	0	0	1	0	0	1	...	0	0	0	0	0	1	1	1	0	0
PhoneService_No	1	0	0	1	0	0	0	1	0	0	...	0	0	0	1	0	0	0	1	0	0
PhoneService_Yes	0	1	1	0	1	1	1	0	1	1	...	1	1	1	0	1	1	1	0	1	1
MultipleLines_No	0	1	1	0	1	0	0	0	0	1	...	1	0	1	0	1	0	0	0	0	1
MultipleLines_No phone service	1	0	0	1	0	0	0	1	0	0	...	0	0	0	1	0	0	0	1	0	0
MultipleLines_Yes	0	0	0	0	0	1	1	0	1	0	...	0	1	0	0	0	1	1	0	1	0
InternetService_DSL	1	1	1	1	0	0	0	1	0	1	...	0	0	0	1	0	1	0	1	0	0
InternetService_Fiber optic	0	0	0	0	1	1	1	0	1	0	...	1	1	1	0	0	0	1	0	1	1
InternetService_No	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
OnlineSecurity_No	1	0	0	0	1	1	1	0	1	0	...	1	0	1	1	0	0	1	0	1	0

KLASYFIKATORY

- 1.kNN (dla k= 3,8,20)
2. Naive Bayes
- 3.Drzewo Decyzyjne
4. Multi-layer Perceptron
- 5.Gradient Boosting- wykorzystuje metodę spadku gradientu
6. Random Forest- złożony z drzew decyzyjnych
8. Support Vector Machines
7. Neural Network

<u>KLASYFIKATOR</u>	<u>ACCURACY</u>	<u>MACIERZ BŁĘDU</u>									
kNN, k=3	73.78%	 <table border="1" data-bbox="949 1093 1407 1460"> <thead> <tr> <th>True \ Pred</th> <th>0</th> <th>1</th> </tr> </thead> <tbody> <tr> <th>0</th> <td>1285</td> <td>275</td> </tr> <tr> <th>1</th> <td>268</td> <td>285</td> </tr> </tbody> </table>	True \ Pred	0	1	0	1285	275	1	268	285
True \ Pred	0	1									
0	1285	275									
1	268	285									
kNN, k=8	77.09%	 <table border="1" data-bbox="941 1505 1407 1877"> <thead> <tr> <th>True \ Pred</th> <th>0</th> <th>1</th> </tr> </thead> <tbody> <tr> <th>0</th> <td>1368</td> <td>192</td> </tr> <tr> <th>1</th> <td>284</td> <td>269</td> </tr> </tbody> </table>	True \ Pred	0	1	0	1368	192	1	284	269
True \ Pred	0	1									
0	1368	192									
1	284	269									

kNN k=20	79.32%	 <p>Confusion matrix for kNN k=20. The y-axis is 'True label' (0, 1) and the x-axis is 'Predicted label' (0, 1). The matrix shows 1360 true positives, 200 false positives, 260 false negatives, and 293 true negatives. A color bar on the right indicates counts from 200 to 1200.</p> <table border="1"><thead><tr><th>True label \ Predicted label</th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>1360</td><td>200</td></tr><tr><th>1</th><td>260</td><td>293</td></tr></tbody></table>	True label \ Predicted label	0	1	0	1360	200	1	260	293
True label \ Predicted label	0	1									
0	1360	200									
1	260	293									
Naive Bayes	70.61%	 <p>Confusion matrix for Naive Bayes. The y-axis is 'True label' (0, 1) and the x-axis is 'Predicted label' (0, 1). The matrix shows 991 true positives, 569 false positives, 88 false negatives, and 465 true negatives. A color bar on the right indicates counts from 200 to 800.</p> <table border="1"><thead><tr><th>True label \ Predicted label</th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>991</td><td>569</td></tr><tr><th>1</th><td>88</td><td>465</td></tr></tbody></table>	True label \ Predicted label	0	1	0	991	569	1	88	465
True label \ Predicted label	0	1									
0	991	569									
1	88	465									
Drzewo decyzyjne	73.69%	 <p>Confusion matrix for Drzewo decyzyjne. The y-axis is 'True label' (0, 1) and the x-axis is 'Predicted label' (0, 1). The matrix shows 1270 true positives, 290 false positives, 292 false negatives, and 261 true negatives. A color bar on the right indicates counts from 400 to 1200.</p> <table border="1"><thead><tr><th>True label \ Predicted label</th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>1270</td><td>290</td></tr><tr><th>1</th><td>292</td><td>261</td></tr></tbody></table>	True label \ Predicted label	0	1	0	1270	290	1	292	261
True label \ Predicted label	0	1									
0	1270	290									
1	292	261									
MLP	95.6%	 <p>Confusion matrix for MLP. The y-axis is 'True label' (0, 1) and the x-axis is 'Predicted label' (0, 1). The matrix shows 1485 true positives, 75 false positives, 53 false negatives, and 500 true negatives. A color bar on the right indicates counts from 200 to 1400.</p> <table border="1"><thead><tr><th>True label \ Predicted label</th><th>0</th><th>1</th></tr></thead><tbody><tr><th>0</th><td>1485</td><td>75</td></tr><tr><th>1</th><td>53</td><td>500</td></tr></tbody></table>	True label \ Predicted label	0	1	0	1485	75	1	53	500
True label \ Predicted label	0	1									
0	1485	75									
1	53	500									

Gradient Boosting	82.68%	 <p>Confusion matrix for Gradient Boosting:</p> <table><thead><tr><th></th><th>Predicted 0</th><th>Predicted 1</th></tr></thead><tbody><tr><th>True 0</th><td>1433</td><td>127</td></tr><tr><th>True 1</th><td>234</td><td>319</td></tr></tbody></table>		Predicted 0	Predicted 1	True 0	1433	127	True 1	234	319
	Predicted 0	Predicted 1									
True 0	1433	127									
True 1	234	319									
Random forest	96.59%	 <p>Confusion matrix for Random forest:</p> <table><thead><tr><th></th><th>Predicted 0</th><th>Predicted 1</th></tr></thead><tbody><tr><th>True 0</th><td>1528</td><td>32</td></tr><tr><th>True 1</th><td>45</td><td>508</td></tr></tbody></table>		Predicted 0	Predicted 1	True 0	1528	32	True 1	45	508
	Predicted 0	Predicted 1									
True 0	1528	32									
True 1	45	508									
SVM	79.03%	 <p>Confusion matrix for SVM:</p> <table><thead><tr><th></th><th>Predicted 0</th><th>Predicted 1</th></tr></thead><tbody><tr><th>True 0</th><td>1375</td><td>142</td></tr><tr><th>True 1</th><td>301</td><td>295</td></tr></tbody></table>		Predicted 0	Predicted 1	True 0	1375	142	True 1	301	295
	Predicted 0	Predicted 1									
True 0	1375	142									
True 1	301	295									

Sieć neuronowa

Sieć wykorzystuje funkcję aktywacji „selu” oraz „softmax”.

```
model = keras.Sequential()
model.add(keras.layers.Dense(52,activation="selu",kernel_initializer= "he_normal",input_dim = X_train.shape[1]))
model.add(keras.layers.BatchNormalization())
model.add(keras.layers.Dense(52,activation="selu",kernel_initializer= "he_normal"))
model.add(keras.layers.BatchNormalization())
model.add(keras.layers.Dense(20,activation="selu",kernel_initializer= "he_normal"))
model.add(keras.layers.BatchNormalization())

model.add(keras.layers.Dense(2,activation="softmax"))

model.compile(optimizer='adam',loss="sparse_categorical_crossentropy", metrics=["accuracy"])

model.summary()
```

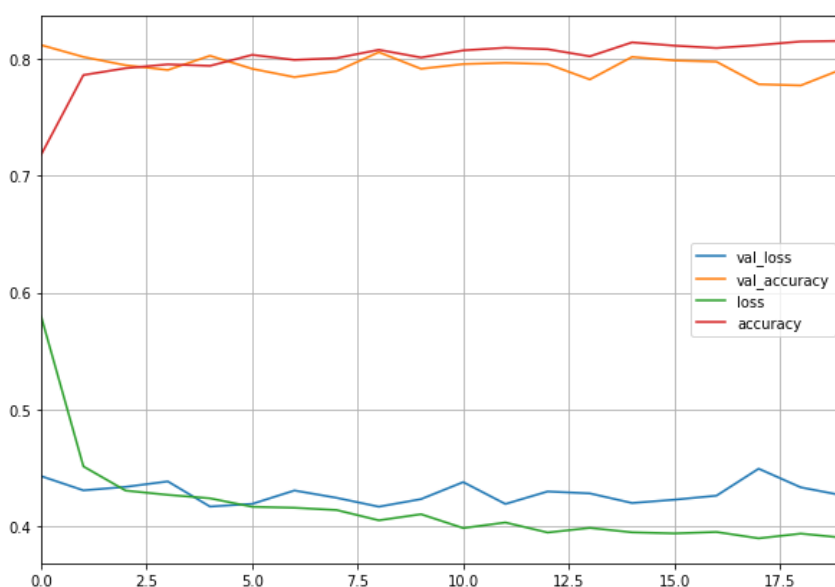
Model: "sequential_44"

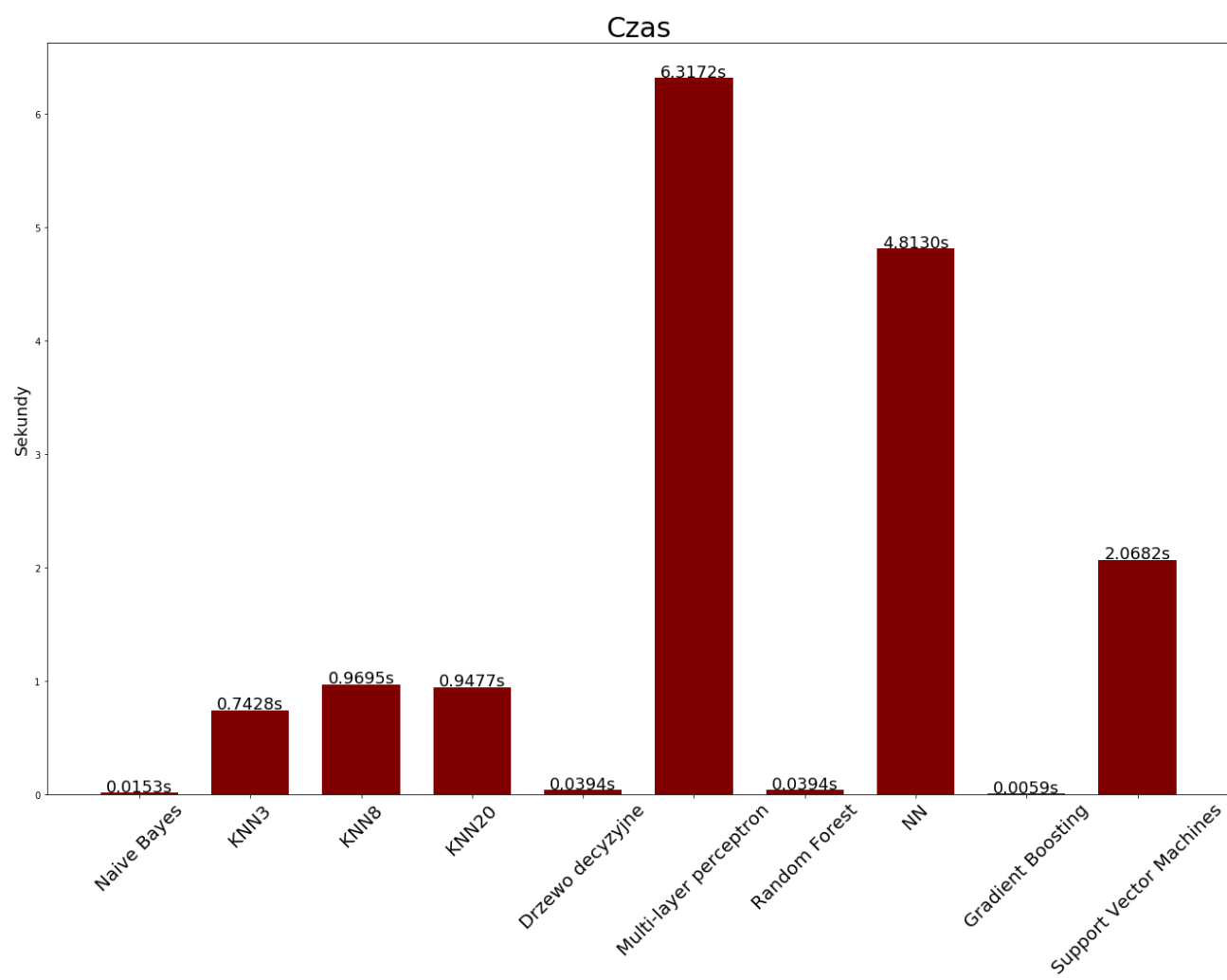
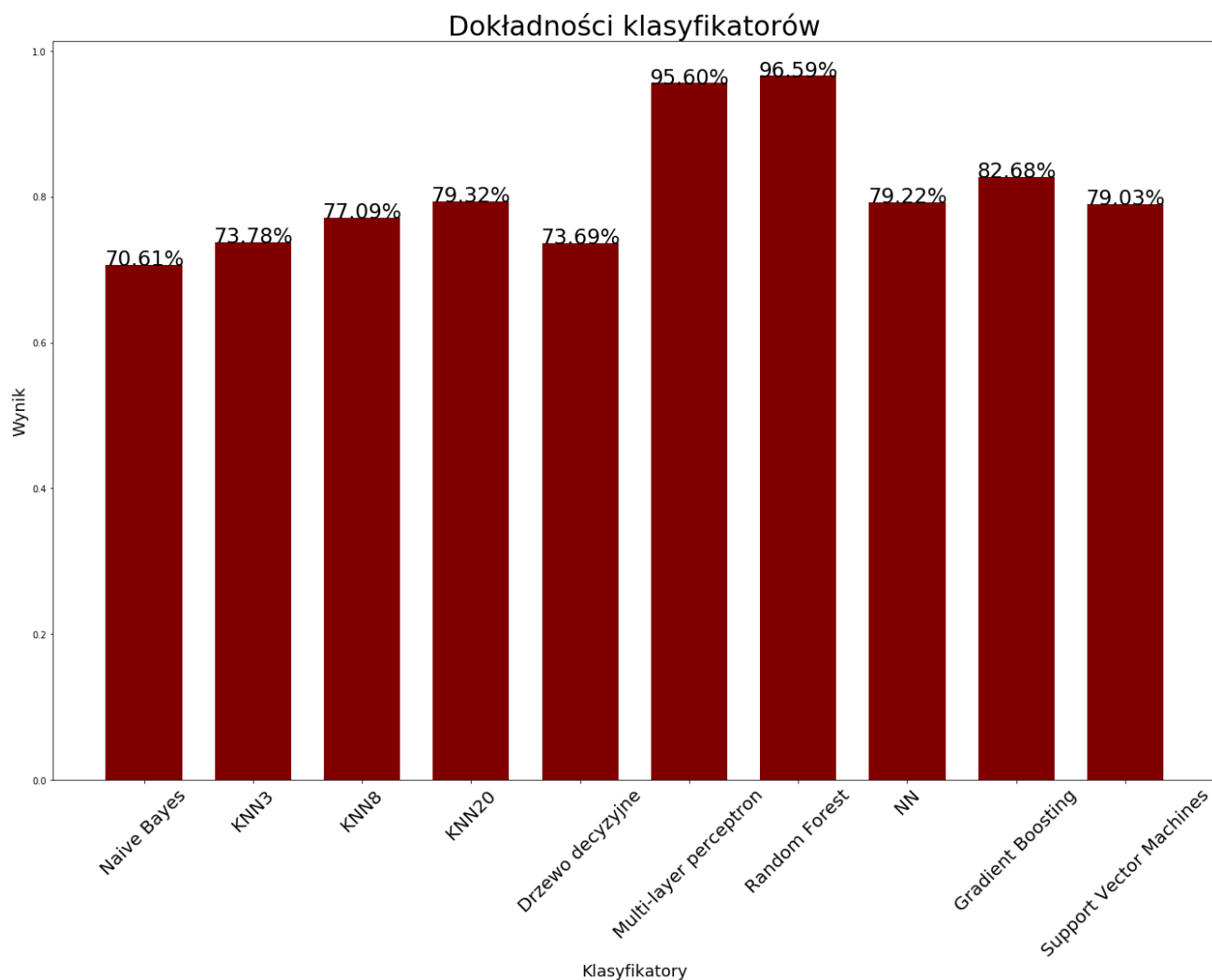
Layer (type)	Output Shape	Param #
dense_187 (Dense)	(None, 52)	2756
batch_normalization_27 (Batch Normalization)	(None, 52)	208
dense_188 (Dense)	(None, 52)	2756
batch_normalization_28 (Batch Normalization)	(None, 52)	208
dense_189 (Dense)	(None, 20)	1060
batch_normalization_29 (Batch Normalization)	(None, 20)	80
dense_190 (Dense)	(None, 2)	42
Total params: 7,110		
Trainable params: 6,862		
Non-trainable params: 248		

Accuracy: 79.22%

Macierz błędów:

	0	1
0	1421	131
1	382	179





Reguły asocjacyjne

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(PhoneService_Yes)	(SeniorCitizen_0)	0.903166	0.837853	0.755786	0.836818	0.998765	-0.000935	0.993657
1	(SeniorCitizen_0)	(PhoneService_Yes)	0.837853	0.903166	0.755786	0.902050	0.998765	-0.000935	0.988609
2	(Churn_No)	(SeniorCitizen_0)	0.734630	0.837853	0.640068	0.871279	1.039895	0.024556	1.259681
3	(SeniorCitizen_0)	(Churn_No)	0.837853	0.734630	0.640068	0.763938	1.039895	0.024556	1.124155
4	(PhoneService_Yes)	(Dependents_No)	0.903166	0.700412	0.632827	0.700676	1.000377	0.000239	1.000883
5	(Dependents_No)	(PhoneService_Yes)	0.700412	0.903166	0.632827	0.903507	1.000377	0.000239	1.003531
6	(PhoneService_Yes)	(Churn_No)	0.903166	0.734630	0.661934	0.732904	0.997650	-0.001559	0.993536
7	(Churn_No)	(PhoneService_Yes)	0.734630	0.903166	0.661934	0.901044	0.997650	-0.001559	0.978550