Porównanie algorytmów uczenia maszynowego w klasyfikacji jakości wody

Agata Bartczak

Maj 2024

1 Wstęp

Raport dotyczy porównania skuteczności różnych klasyfikatorów w przewidywaniu poziomu zdatności do spożycia wody na podstawie danych numerycznych. Przebadano następujące klasyfikatory: Logistic Regression, SVC (C-Support Vector Classification), Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, KNeighbours Classifier, Gradient Boosting Classifier, AdaBoost Classifier, Naive Bayes Classifier, MLP (sieci neuronowe). W tworzeniu kodu wykorzystano dwa poradniki: z kanału The Data Futrue lab [4] oraz z kanału Developer Ashish [5].

2 Charakterystyka danych i preprocessing

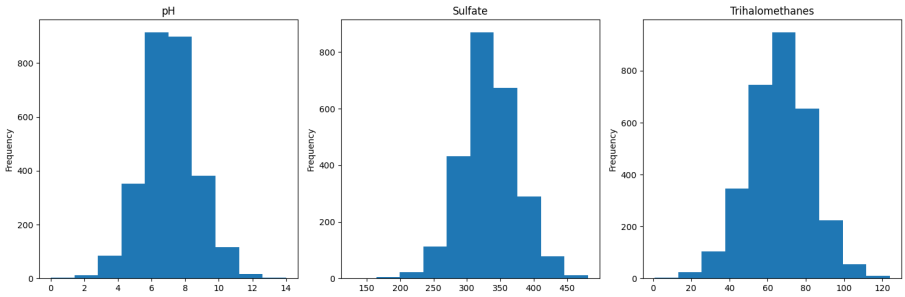
Dane wykorzystane w raporcie pochodzą z zestawu danych dostępnego na platformie Kaggle [1]. Zestaw zawiera 3276 wierszy z 10 kolumnami:

1. pH (ph): pH wody (od 0 do 14).
2. Twardość (Hardness): Zdolność wody do wytrącania mydła w mg/L.
3. Rozpuszczone substancje stałe (Solids): Całkowita zawartość substancji stałych w ppm.
4. Chloraminy (Chloramines): Ilość chloramin w ppm.
5. Siarczany (Sulfate): Ilość siarczanów rozpuszczonych w mg/L.
6. Przewodność (Conductivity): Przewodność elektryczna wody w $\mu S\,cm^{-1}$.
7. Węgiel organiczny (Organic_carbon): Ilość węgla organicznego w ppm.
8. Trihalometany (Trihalomethanes): Ilość trihalometanów w $\mu g\,L^{-1}$.
9. Zmętnienie (Turbidity): Miara właściwości emisyjnych światła wody w NTU (Nephelometric Turbidity Units).
10. Zdatność do spożycia (Potability): Wskaźnik czy woda jest bezpieczna do spożycia przez ludzi. 1 oznacza wodę zdatną do spożycia, 0 - niezdatną.

W dalszej części raportu będą używane oryginalne nazwy kolumn, które zostały podane w nawiasach.

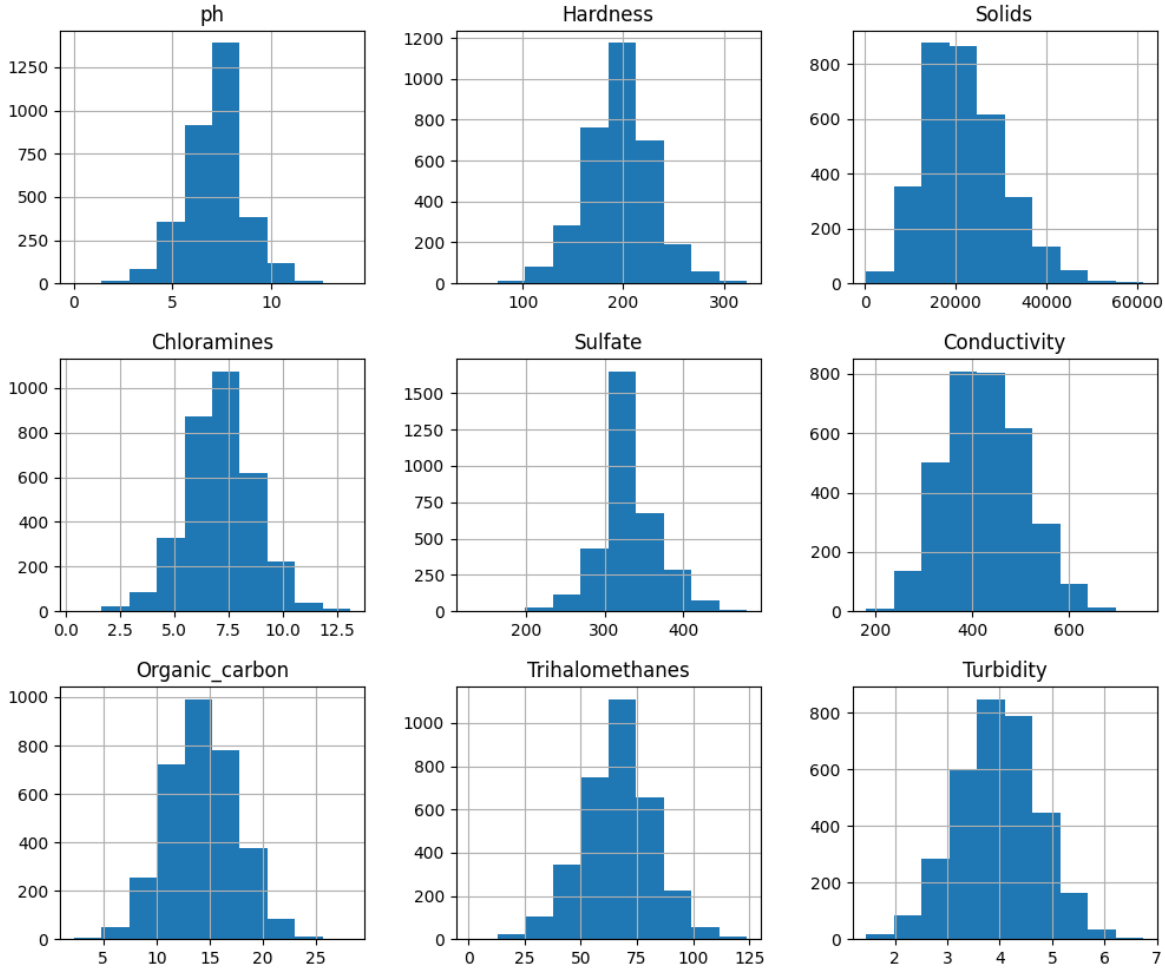
2.1 Uzupełnianie brakujących wartości

W oryginalnym zestawie danych brakowało wartości: w 491 wierszach ph, w 781 Sulfate, w 162 Trihalomethanes. Na poniższym wykresie przedstawiono histogramy dla tych kolumn przed ich uzupełnieniem:



Ze względu na rozkłady tych danych, zbliżonych kształtem do rozkładu normalnego, brakujące dane uzupełniono wartościami średnimi z danej kolumny.

Na poniższym wykresie przedstawiono histogramy dla wszystkich cech, po uzupełnieniu danych:

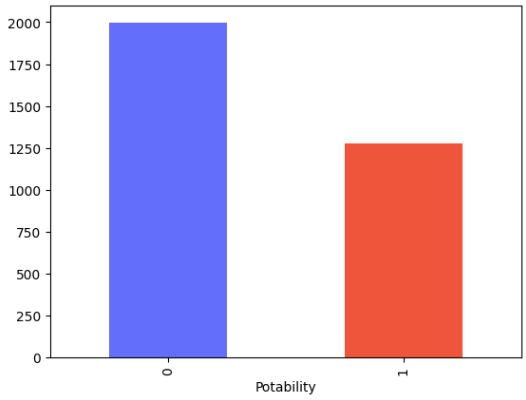


Z analizy wykresu wynika, że wszystkie cechy mają rozkłady zbliżone kształtem do rozkładu nor-

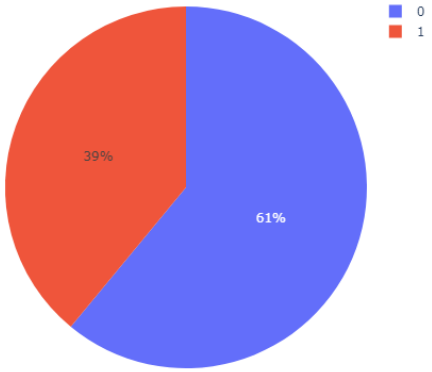
malnego.

2.2 Balansowanie próbek

W bazie danych są dwie klasy określające zdatność wody do spożycia: 0 - woda niezdatna z licznością próbek wynoszącą 1998, 1 - woda zdatna z 1278 próbkami). Ze względu na brak dużej przewagi w liczności którejś z klas zdecydowano o niebalansowaniu bazy danych. Poniżej graficznie przedstawiono liczność klas:



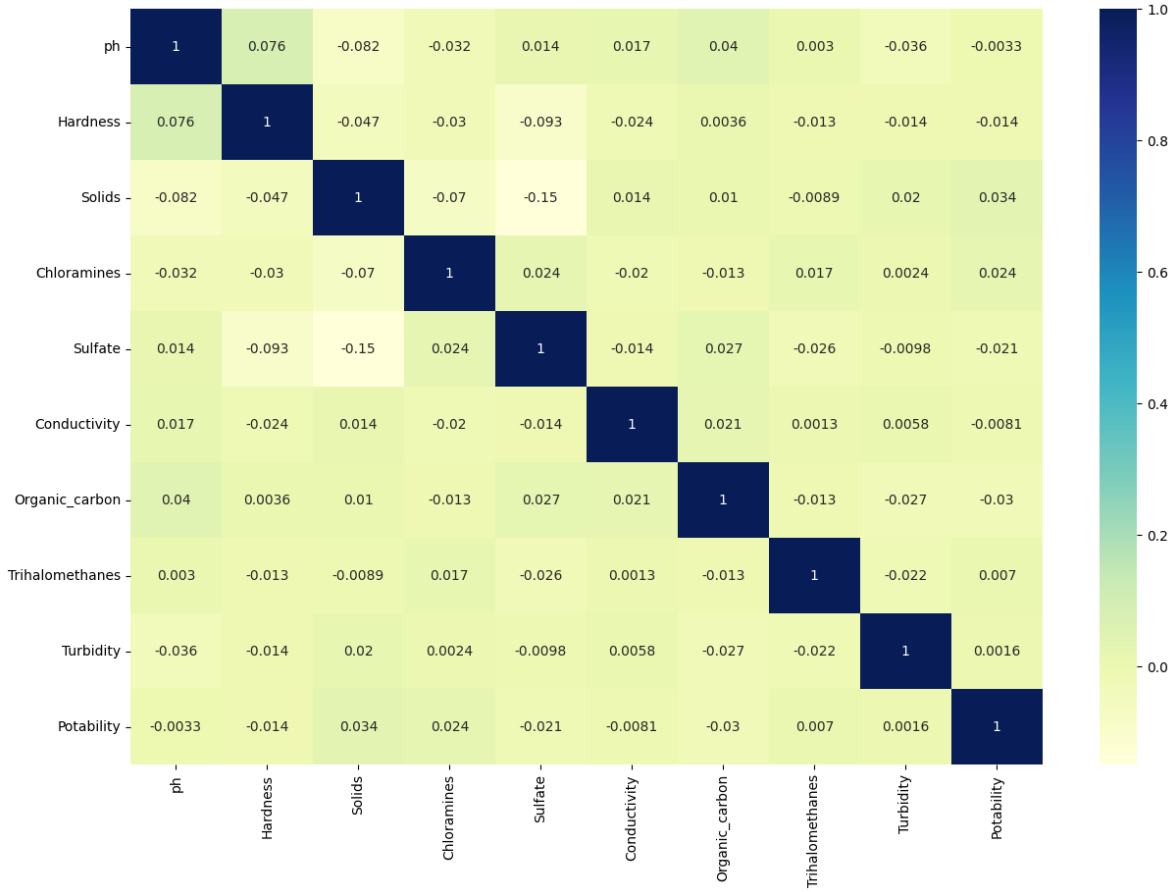
Rysunek 1: Liczność próbek



Rysunek 2: Rozkład procentowy próbek

2.3 Redukcja liczby cech

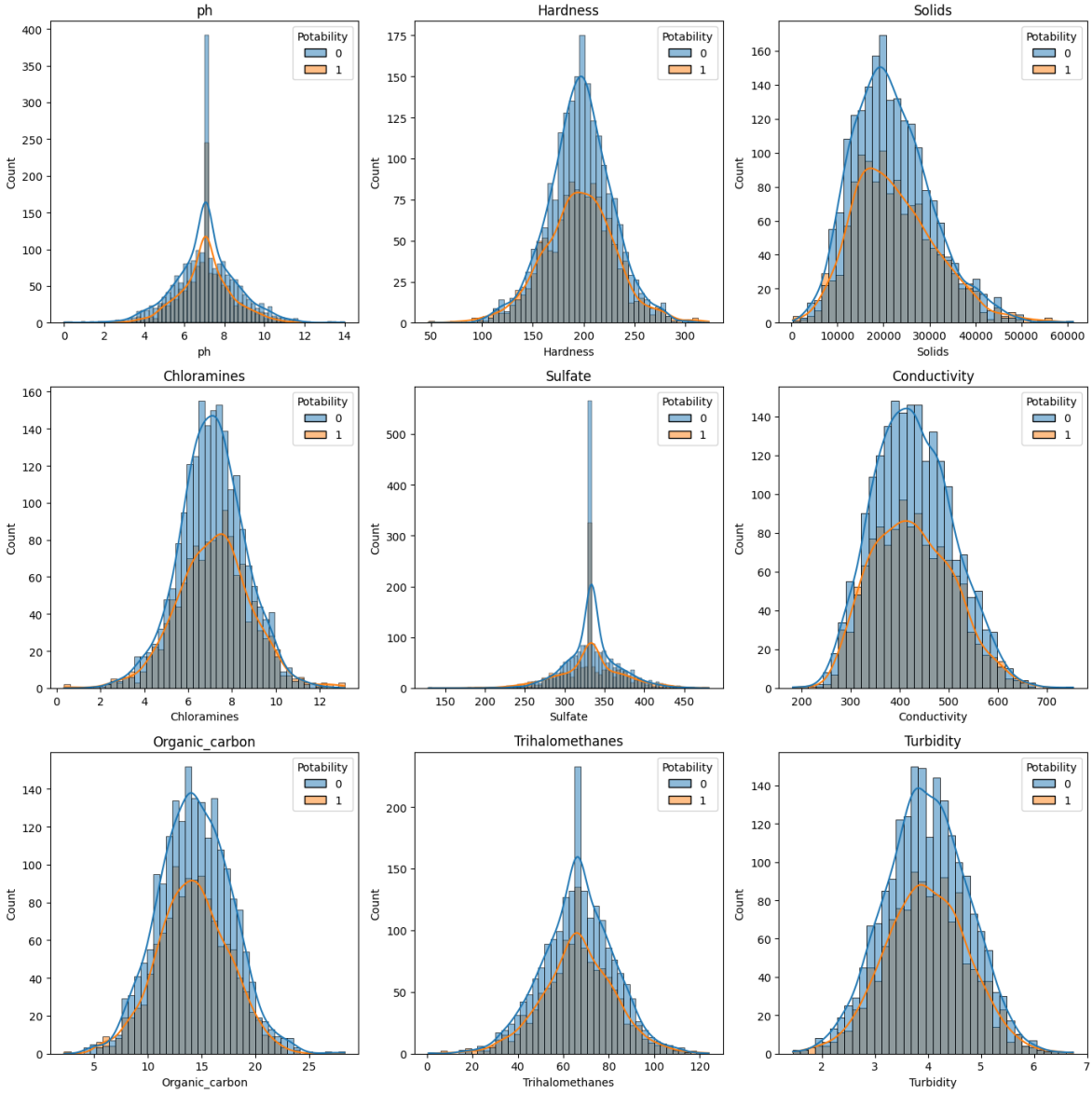
Na poniższym wykresie przedstawiono macierz korelacji dla kolumn w zestawie danych:



Korelacje pomiędzy poszczególnymi kolumnami nie różnią się znacząco od 0, więc niezredukowano liczby cechy.

2.4 Charakterystyka danych z podziałem na klasy

W celu wykrycia tego jak wartości poszczególnych cech rozkładają się dla wody zdatnej i niezdatnej do spożycia wykonano poniższe histogramy z podziałem na klasy:



Z analizy wykresu wynika, że histogramy dla każdej z cech nakładają się na siebie. Nie są przesunięte. Nie widać wyraźnych podziałów w rozkładach wartości.

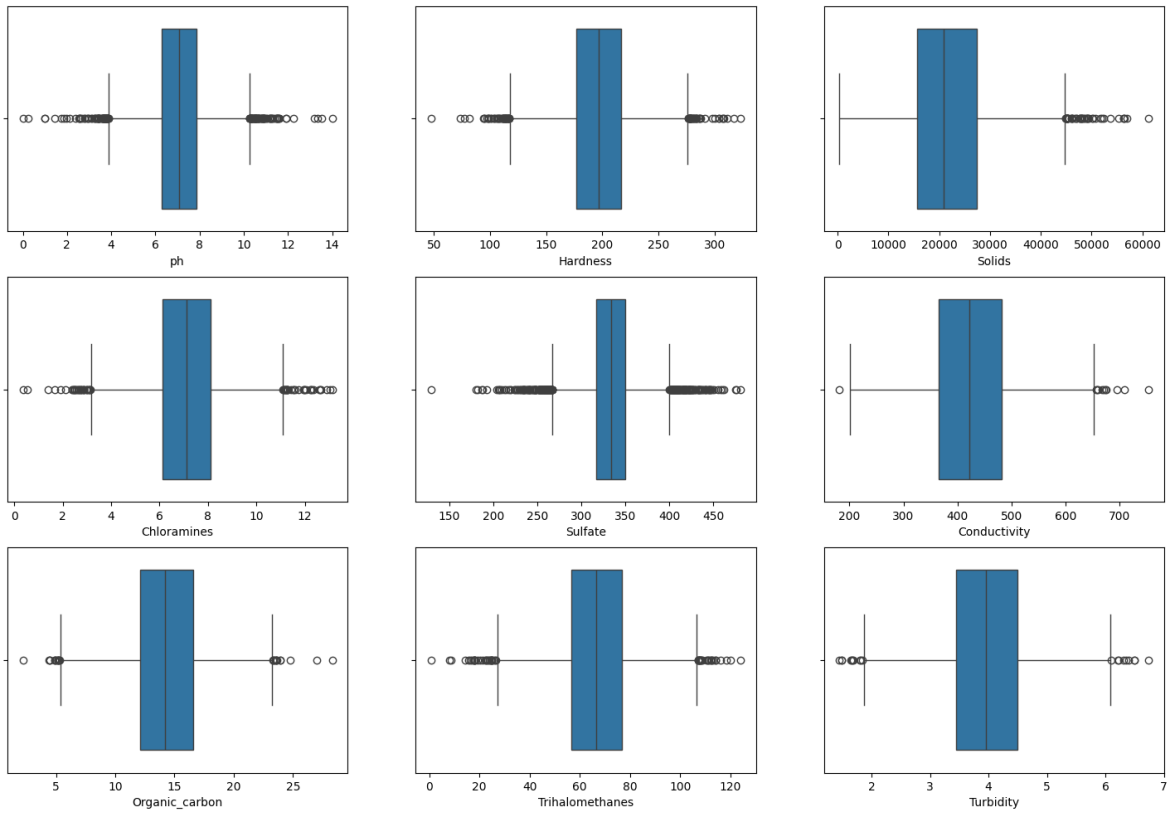
Dodatkowo obliczono wartości średnie dla każdej z cech z podziałem na wodę zdatną i niezdatną do spożycia:

Cecha	Średnia dla 0	Średnia dla 1	Względna różnica (%)
ph	7.0847	7.0748	0.1398
Hardness	196.7333	195.8007	0.4740
Solids	21777.4908	22383.9910	2.7850
Chloramines	7.0922	7.1693	1.0880
Sulfate	334.3717	332.8441	0.4569
Conductivity	426.7305	425.3838	0.3156
Organic_carbon	14.3643	14.1609	1.4163
Trihalomethanes	66.3085	66.5335	0.3393
Turbidity	3.9658	3.9683	0.0638

Różnice w średnich wartościach między klasami są niewielkie, co sugeruje, że pojedyncze cechy nie są wystarczającym kryterium do jednoznacznej klasyfikacji próbek.

2.5 Wartości odstające (outliers)

Na poniższym wykresie przedstawiono wartości odstające dla każdej z cech:



Również policzono liczbę wartości odstających i obliczono jaki procent wszystkich wierszy stanowią. Wyniki umieszczono w tabeli:

Kolumna	Liczba outliers	(%)
ph	142	4.33
Hardness	83	2.53
Solids	47	1.43
Chloramines	61	1.86
Sulfate	264	8.06
Conductivity	11	0.34
Organic_carbon	25	0.76
Trihalomethanes	54	1.65
Turbidity	19	0.58

Dla większości kolumn procent outliers jest niewielki. Waha się od 0.58% do 4%. Dla kolumny „Solids" procent outliers wyniósł 8%. Jednak biorąc pod uwagę charakterystykę tej cechy, wartości odstające (max = 481, min = 129) mieszczą się w normach mg/litr dla badanych wód (opis cechy Solids dla datasetu [1]). Stąd zdecydowano o pozostawieniu ich w zestawie danych.

3 Klasyfikacja

Przed klasyfikacją znormalizowano dane przy użyciu wzoru

$$z = (x - \mu)/\sigma$$

gdzie x to oryginalna wartość, μ to wartość średnia danej cechy, σ to odchylenie standardowe, a z to wartość po normalizacji.

3.1 Dobieranie parametrów

W celuznalezienia optymalnych parametrów dla różnych modeli klasyfikatorów, przeprowadzono testy na różnych zestawach parametrów. Do oceny dopsasowania modelu wykorzystano walidację krzyżową opartą na miarze dokładności (accuracy). Do walidacji krzyżowej wykorzystano 5 zestawów danych utworzonych za pomocą metody Stratified K-Fold, co pozwoliło na zachowanie oryginalnych proporcji klas w każdym zbiorze [2].

W poniższej tabeli przedstawiono średnie „accuracy" z walidacji krzyżowej dla omawianych klasyfikatorów z najlepszymi parametrami:

Model	Średnie accuracy (%)
Logistic Regression	61.02
Decision Tree	61.75
Random Forest	62.91
Extra Trees	64.07
SVC	65.08
K-Nearest Neighbors	63.28
AdaBoost	62.03
Gradient Boosting	62.18
Gaussian Naive Bayes	61.26

Oceniając „accuracy" najlepszy okazał się klasyfikator SVC.

3.2 Sieci neuronowe

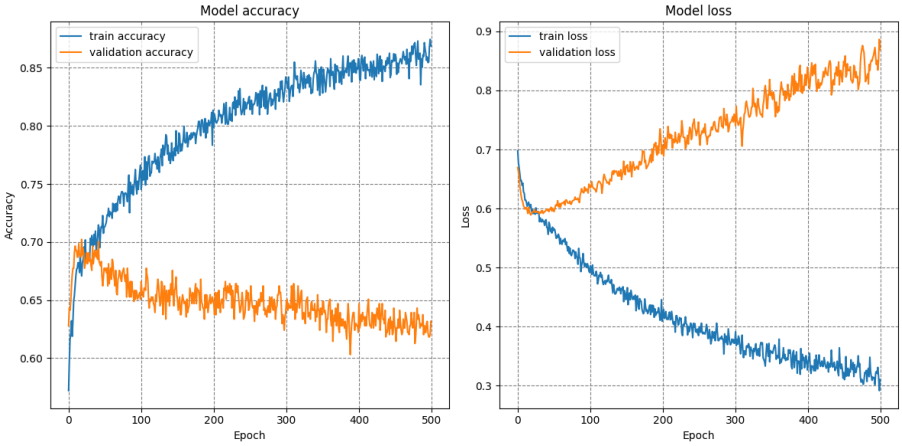
W celu wyszkolenia różnych sieci neuronowych podzielono dane na zbiór treningowy (80%) i zbiór testowy (20%). W procesie nauki zapisywano najlepszy model pod względem „accuracy" i do późniejszych

testów wykorzystywano najlepszą wersję danego modelu. Modele dobrano tak, aby sprawdzić jak różne czynniki (liczba warstw, liczba neuronów, obecność dropoutu, funkcja aktywacyjna [3], optymalizator) wpływają na dokładność modelu. Poniżej przedstawiono opis każdej z sieci oraz wykresy z learning curve oraz loss.

- **MLP1**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 256 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 128 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 64 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

  Optymalizator: Adam

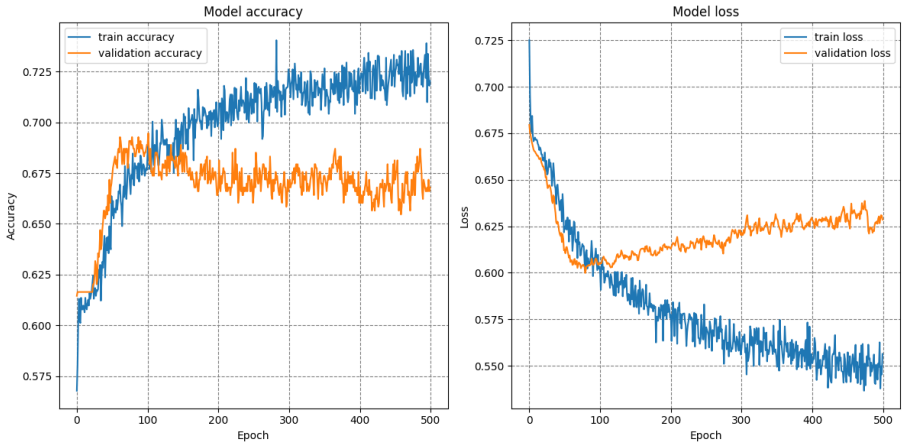
  Najlepsze validation accuracy: 70.23%



- **MLP2**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 64 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 16 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

  Optymalizator: Adam

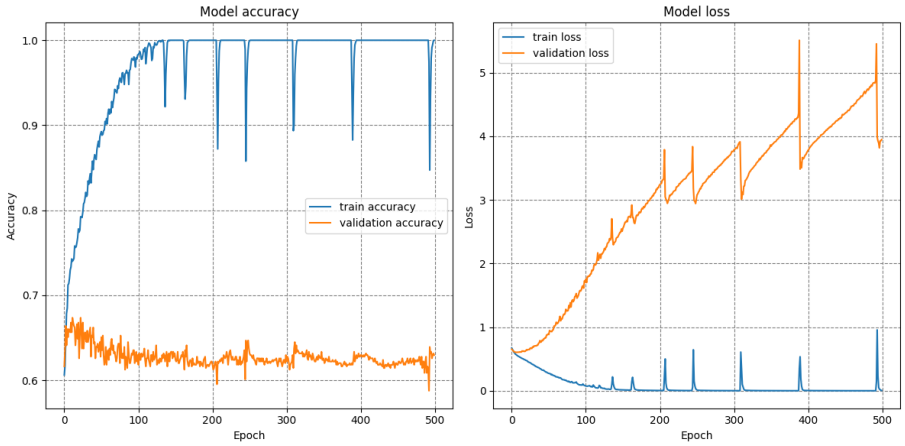
  Najlepsze validation accuracy: 69.47%



- **MLP3**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 64 neuronów z funkcją aktywacji ReLU.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji ReLU.
  - Warstwa ukryta: 16 neuronów z funkcją aktywacji ReLU.
  - Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

  Optymalizator: Adam

  Najlepsze validation accuracy: 67.37%

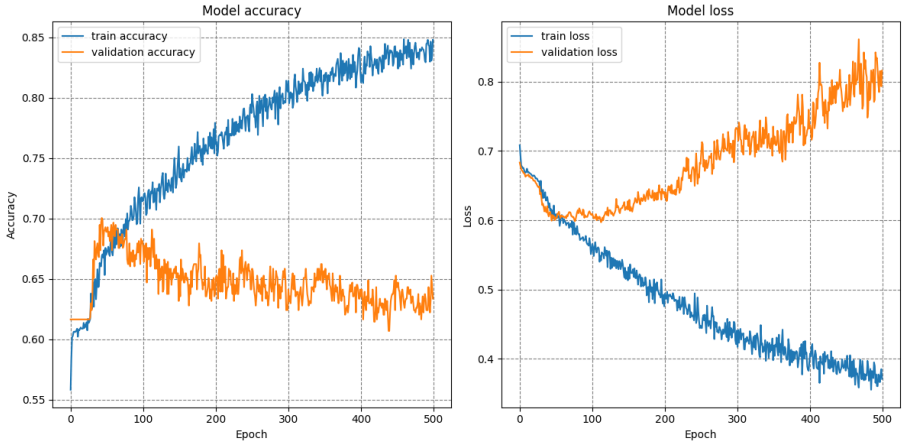


- **MLP4**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 256 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 128 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 64 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji ReLU.

- Warstwy Dropout z prawdopodobieństwem 0.5.
- Warstwa ukryta: 16 neuronów z funkcją aktywacji ReLU.
- Warstwy Dropout z prawdopodobieństwem 0.5.
- Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

Optymalizator: Adam

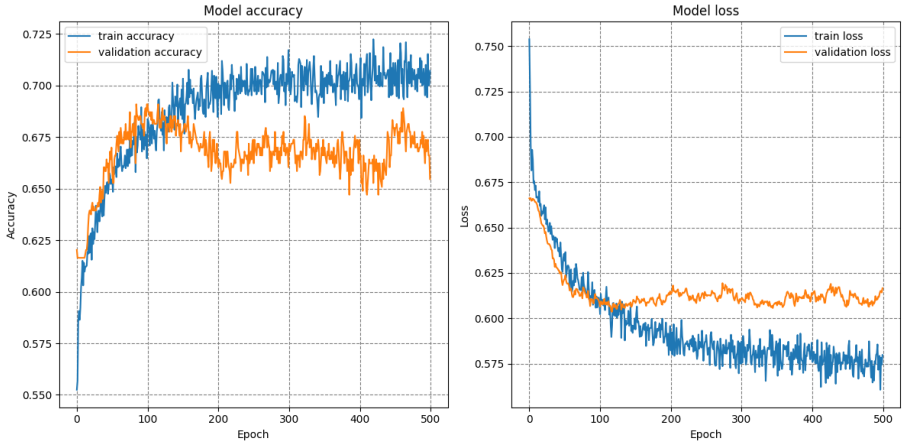
Najlepsze validation accuracy: 70.04%



- **MLP5**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 32 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

Optymaliztor: Adam

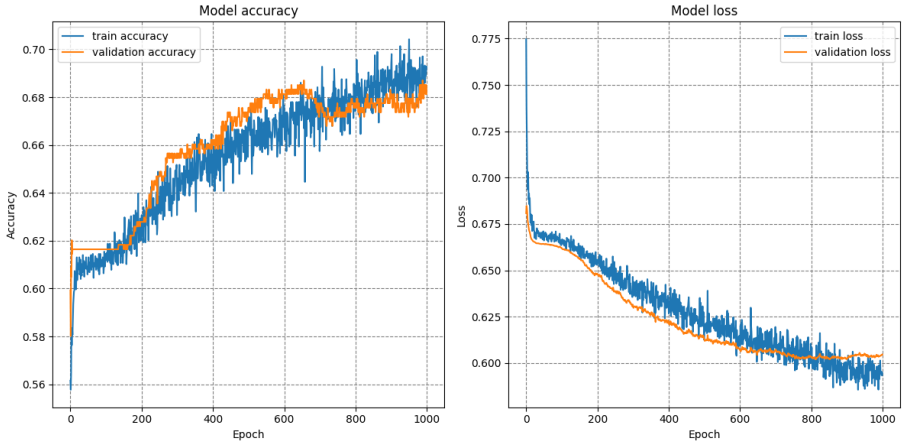
Najlepsze validation accuracy: 69.08%



- **MLP6**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 64 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 16 neuronów z funkcją aktywacji ReLU.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

  Optymalizator: Sigmoid

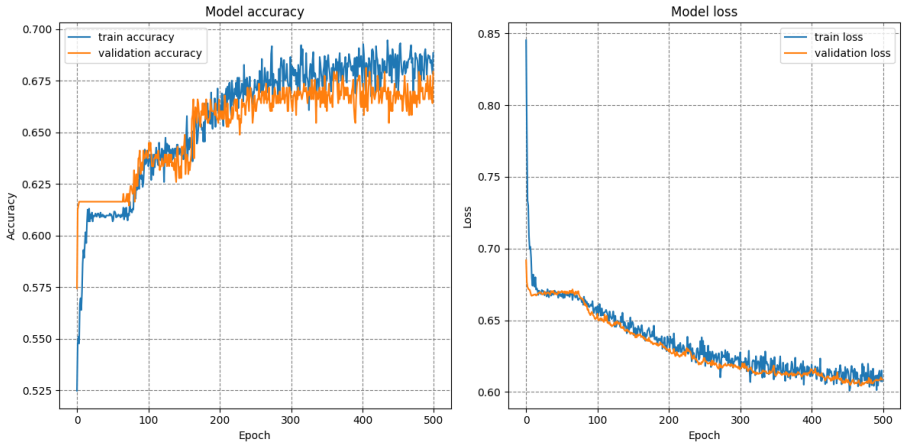
  Najlepsze validation accuracy: 68.70%



- **MLP7**: Sieć neuronowa złożona z warstw:
  - Warstwa wejściowa: 32 neuronów z funkcją aktywacji tanh.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji tanh.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa ukryta: 32 neuronów z funkcją aktywacji tanh.
  - Warstwy Dropout z prawdopodobieństwem 0.5.
  - Warstwa wyjściowa: 2 neurony z funkcją aktywacji softmax.

  Optymalizator: Adam

  Najlepsze validation accuracy: 68.13%



3.2.1 Wnioski dla sieci neuronowych

Z analizy learning curve oraz loss modeli wynika, że w przypadku większości z nich dłuższe trenowanie nie prowadziło do lepszych rezultatów. Zawarcie warstw Dropout pozytywnie wpływa na validation accuracy modeli. Zastosowanie optymalizatora sigmoid lub funkcji aktywacyjnej tanh dla warstw ukrytych prowadziło to wolniejszego, bardziej równomiernego procesu nauki. Większa liczba neuronów pozytywnie wpływała na validation accuracy. Najbardziej dokładnym modelem (oceniając validation accuracy) okazał się model „mlp1".

3.3 Porównanie wszystkich modeli

W celu oceny wydajności każdego z modeli przeprowadzono testowanie na zestawie danych testowych. Dla każdego modelu wygenerowano macierz błędów (Confusion Matrix), oraz obliczono accuracy, precision, F1-score i recall:

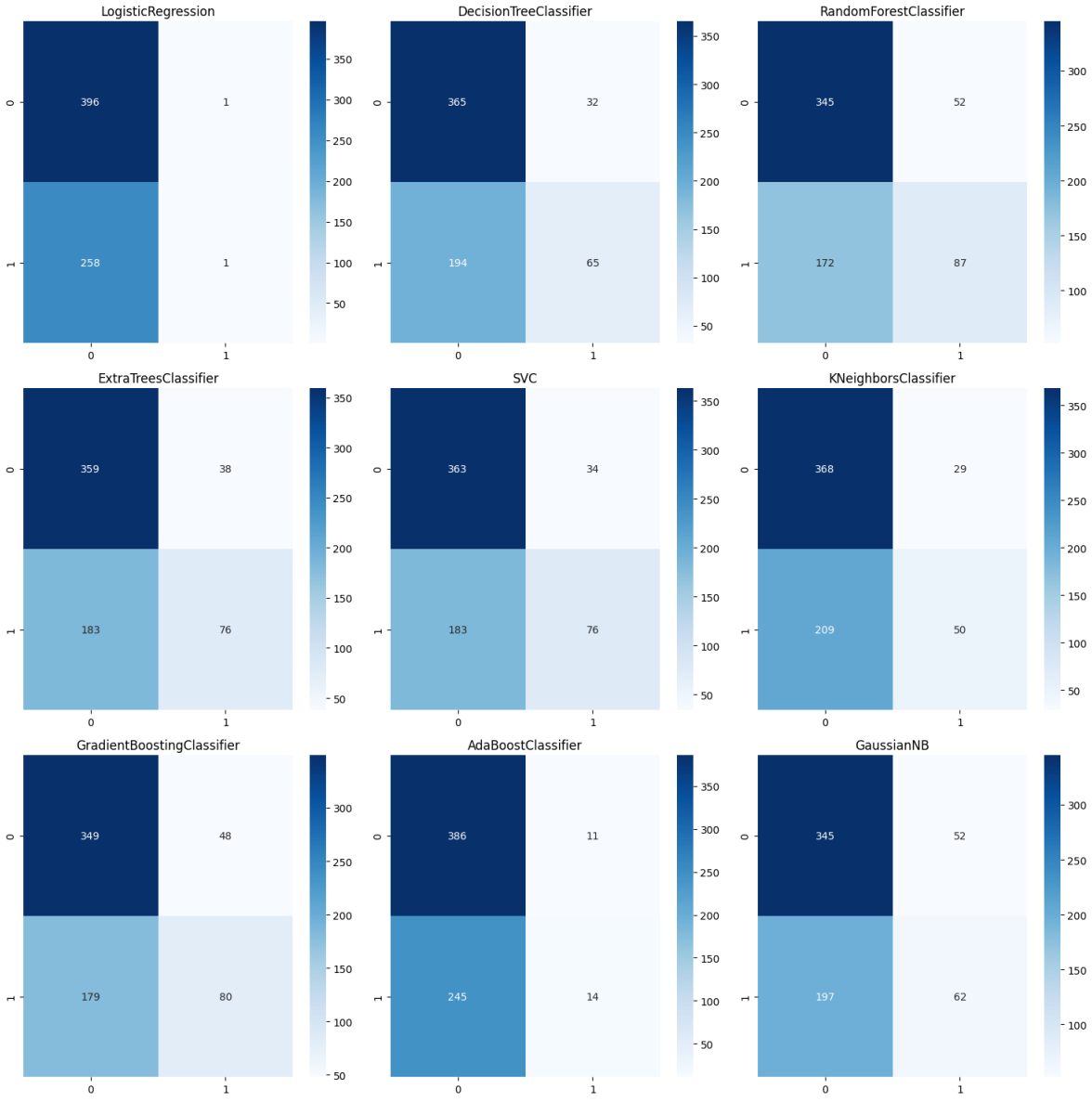
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

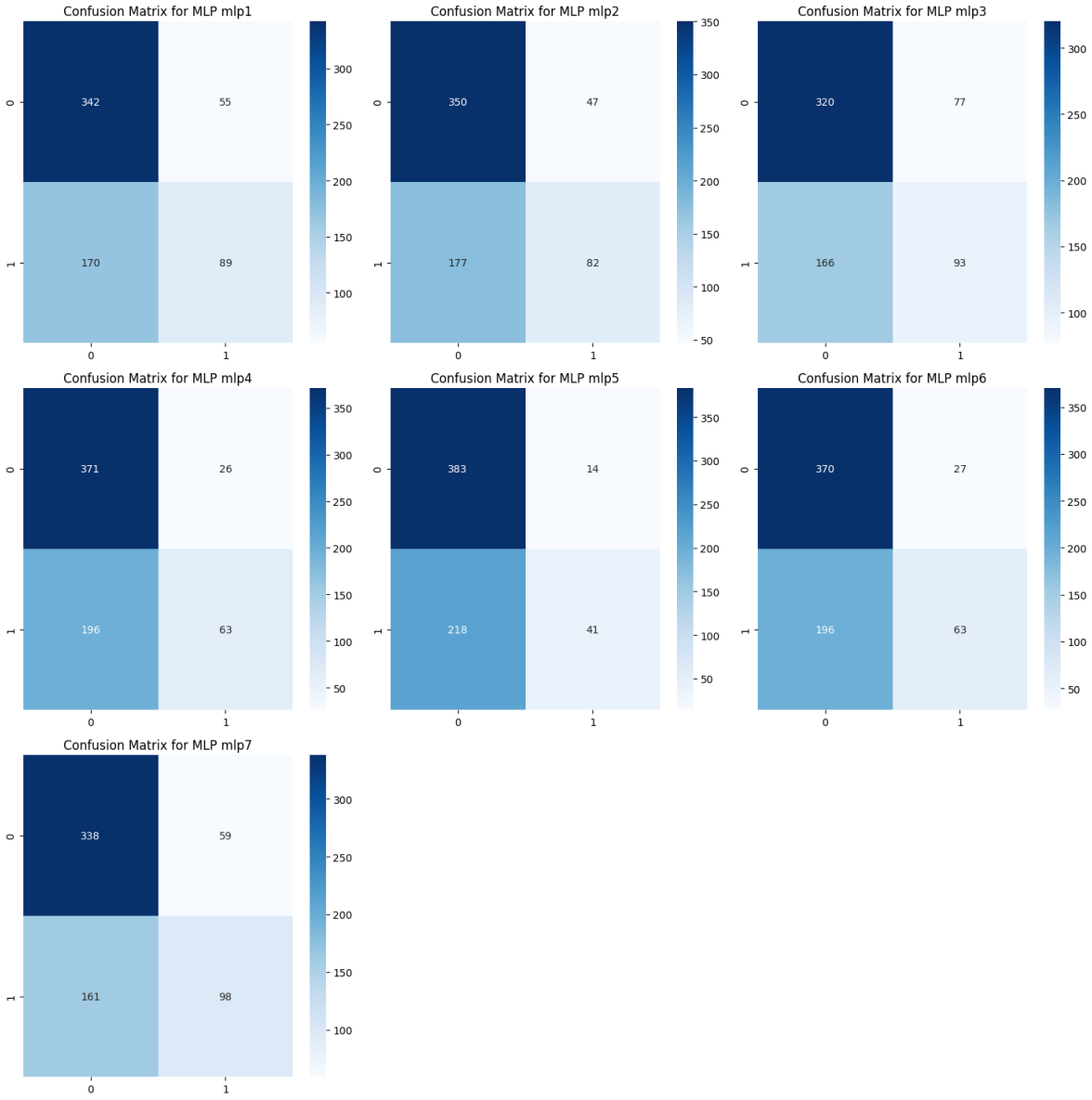
,

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Wyniki przedstawiono na poniższych wykresach i tabeli:


LogisticRegression
396
1
258
1
DecisionTreeClassifier
365
32
194
65
RandomForestClassifier
345
52
172
87
ExtraTreesClassifier
359
38
183
76
SVC
363
34
183
76
KNeighborsClassifier
368
29
209
50
GradientBoostingClassifier
349
48
179
80
AdaBoostClassifier
386
11
245
14
GaussianNB
345
52
197
62


Confusion Matrix for MLP mlp1
342
55
170
89
Confusion Matrix for MLP mlp2
350
47
177
82
Confusion Matrix for MLP mlp3
320
77
166
93
Confusion Matrix for MLP mlp4
371
26
196
63
Confusion Matrix for MLP mlp5
383
14
218
41
Confusion Matrix for MLP mlp6
370
27
196
63
Confusion Matrix for MLP mlp7
338
59
161
98

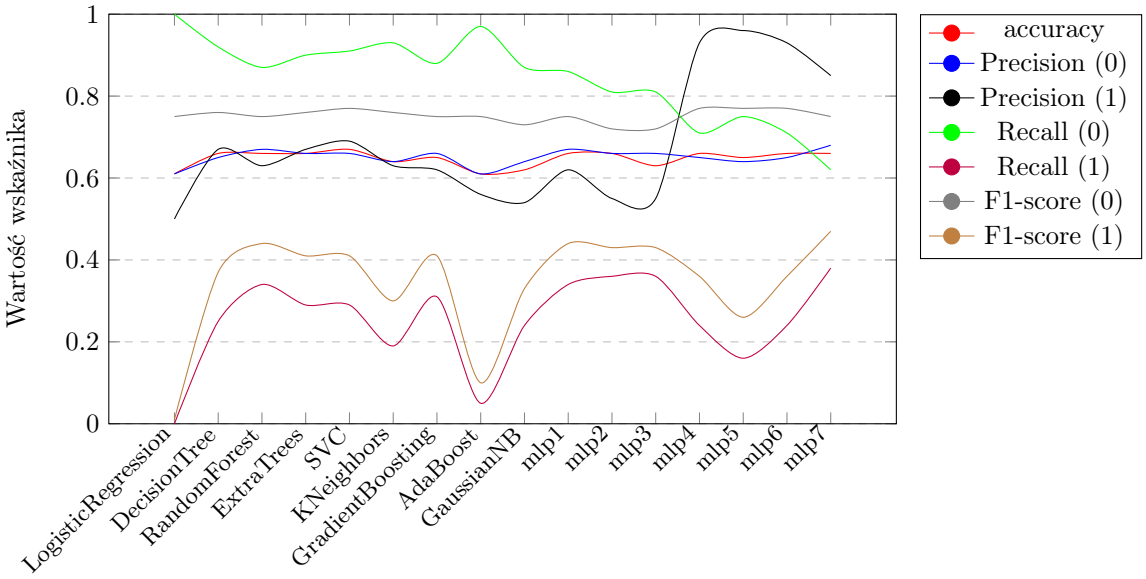


Tabela 1: Wyniki klasyfikacji dla różnych modeli

Model	Accuracy	Precision		Recall		F1-score		Suma wskaźników
		0	1	0	1	0	1	
LogisticRegression	0.61	0.61	0.50	1.00	0.00	0.75	0.01	3.48
DecisionTree	0.66	0.65	0.67	0.92	0.25	0.76	0.37	4.28
RandomForest	0.66	0.67	0.63	0.87	0.34	0.75	0.44	4.36
ExtraTrees	0.66	0.66	0.67	0.90	0.29	0.76	0.41	4.36
SVC	0.67	0.66	0.69	0.91	0.29	0.77	0.41	4.40
KNeighbors	0.64	0.64	0.63	0.93	0.19	0.76	0.30	4.09
GradientBoosting	0.65	0.66	0.62	0.88	0.31	0.75	0.41	4.28
AdaBoost	0.61	0.61	0.56	0.97	0.05	0.75	0.10	3.65
GaussianNB	0.62	0.64	0.54	0.87	0.24	0.73	0.33	3.97
mlp1	0.66	0.67	0.62	0.86	0.34	0.75	0.44	4.34
mlp2	0.66	0.66	0.55	0.81	0.36	0.72	0.43	4.19
mlp3	0.63	0.66	0.55	0.81	0.36	0.72	0.43	4.16
mlp4	0.66	0.65	0.93	0.71	0.24	0.77	0.36	4.22
mlp5	0.65	0.64	0.96	0.75	0.16	0.77	0.26	4.19
mlp6	0.66	0.65	0.93	0.71	0.24	0.77	0.36	4.32
mlp7	0.66	0.68	0.85	0.62	0.38	0.75	0.47	4.41

Z analizy macierzy błędów oraz wartości klasyfikacji wynika, że wszystkie modele radzą sobie gorzej w poprawnym klasyfikowaniu wody zdatnej do spożycia (oznaczonej przez 1). Wskaźniki F1-score oraz Recall są dużo niższe dla klasy „1”.

4 Wnioski

Wybrane modele nie mają odpowiednio wysokiej dokładności. Najwyższe accuracy dla testu na zbiorze testowym wyniosło 67% dla klasyfikatora SVC (C-Support Vector Classification), co nie jest wysokim wynikiem. Może to wynikać z wielkości zestawu danych oraz braku wyraźnego rozróżnienia rozkładu

danych dla każdej z cech. Rozkłady te są zbilżone do siebie kształtem. Modele mają dużo niższą skuteczność w klasyfikowaniu wody zdatnej do spożycia. Może to częściowo wynikać z balansu klas w zbiorze danych, w którym klasa „0" ma większą liczbę próbek od klasy „1". Biorąc pod uwagę sumę wszystkich wzkaźników najlepszymi modelami okazały się „mlp7" oraz SVC.

Literatura

[1] A. Kadiwal, *Water Quality*, https://www.kaggle.com/datasets/adityakadiwal/water-potability, [dostęp: 1 maja 2024].

[2] *sklearn.model_ selection.StratifiedKFold*, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html, [dostęp 15 maja 2024]

[3] Keras, *Layer activation functions*, https://keras.io/api/layers/activations/, [dostęp 15 maja 2024]

[4] The Data Future Lab, *Machine Learning Project | Water Quality Prediction* https://www.youtube.com/watch?v=yVFPf3_gvQI&t=1886s&ab_channel=TheDataFutureLab, [dostęp 1 maja 2024]

[5] Developer Ashish, *Kaggle Project- Water Quality Prediction using Machine Learning !! Drinking water potability* https://www.youtube.com/watch?v=MWLUtTlHxpw&t=4391s&ab_channel=DeveloperAshish, [dostęp 1 maja 2024]