

Optymalizacja hiperparametrów klasyfikatorów bazowych oraz heterogenicznego baggingu

T. Kołodziej

Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Wrocław, Polska
251004@student.pwr.edu.pl

M. Kubiak

Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Wrocław, Polska
249481@student.pwr.edu.pl

Streszczenie—Projekt bada skuteczność metod przeszukiwania GridSearch i MonteCarlo oraz optymalizację hiperparametrów klasyfikatorów bazowych: drzewa decyzyjnego, MLP, SVM i GNB. Celem jest porównanie skuteczności tych metod oraz dostosowanie hiperparametrów klasyfikatorów na różnych zbiorach danych. Zbadane zostanie również skuteczność heterogenicznego baggingu klasyfikatorów bazowych.

Index Terms—Bagging, Multi-Layer Perceptron(MLP), Support Vector Machine(SVM), Gauss Naive Bayes(GNB), Drzewo decyzyjne

I. WSTĘP

Jednym z zadań postawionych w dziedzinie uczenia maszynowego jest klasyfikacja obiektów. Polega ona na przypisaniu etykiety do przykładu z dziedziny danego problemu, wykorzystując algorytmy uczenia maszynowego. Zadanie przydzielania etykiet zostaje wykonane przez klasyfikatory, które po wstępnym wyuczeniu na testowym zbiorze danych, są w stanie samodzielnie określić klasę danych wejściowych. W zależności od zasady działania, cechy klasyfikatorów mogą zostać uznane za zalety lub też wady, uwzględniając szybkość działania, dokładność czy skłonność do przeuczenia.

Multi-Layer Perceptron(MLP) jest w stanie zapewnić wyższą sprawność w porównaniu z innymi klasyfikatorami, jednak charakteryzują go również długie czasy trenowania oraz możliwość utknięcia w minimach lokalnych [1]. Funkcją decyzyjną Support Vector Machine(SVM) jest hiperpłaszczyzna oddzielająca obserwacje należące do różnych klas, na podstawie wzorców informacji o tych obserwacjach, zwanych cechami. Płaszczyzna optymalna to taka, która separuje klasy z możliwie największym marginesem [2]. Gauss Naive Bayes(GNB) to probabilistyczny algorytm kwalifikacji, cieszący się dużą popularnością, dzięki prostocie, sprawności i solidności działania [3]. Drzewa decyzyjne klasyfikują dane na podstawie serii pytań, ułożonej w hierarchii, gdzie pytania posiadają określoną liczbę odpowiedzi, z których każda przekierowuje decyzję do kolejnego, podrzędnego pytania, tworząc tym samym strukturę drzewa [4].

W celu zniwelowania niekorzystnych cech klasyfikatorów, zostają one złożone w zespoły, składające się z wielu klasyfikatorów o mniej skomplikowanej strukturze, niż byłoby to wymagane w przypadku pojedynczego klasyfikatora. Zespoły

klasyfikatorów, takie jak Bagging lub Boosting, trenują klasyfikatory bazowe wykorzystując różne zestawy danych. W przypadku Baggingu, wynik zostaje wybrany na podstawie głosowania większościowego spośród wszystkich klasyfikatorów bazowych. Dzięki wytrenowaniu ich różnymi zestawami danych, zwiększona zostaje zdolność do generalizacji modelu, jednocześnie zwiększając jego odporność na przetrenowanie [5].

Celem projektu jest zbadanie skuteczności metod przeszukiwania Grid Search i Monte Carlo oraz optymalizacja hiperparametrów klasyfikatorów. W pierwszym eksperymencie przeprowadzimy przeszukiwanie Grid Search i Monte Carlo na 4 klasyfikatorach bazowych: Drzewo decyzyjne, MLP, SVM i GNB, w celu dostosowania hiperparametrów i sprawdzenia, które przeszukiwanie jest najskuteczniejsze na syntetycznym zbiorze danych. Drugi eksperyment ma na celu weryfikację skuteczności zoptymalizowanych klasyfikatorów na tle różnorodnych baz danych. W trzecim eksperymencie wykonany zostanie bagging klasyfikatorów bazowych. Hiperparametry zespołu klasyfikatorów również zostaną zoptymalizowane z wykorzystaniem Grid Search i Monte Carlo. Skuteczność klasyfikacji zespołu zostanie porównana z wynikami indywidualnych klasyfikatorów bazowych.

II. POWIĄZANE PRACE

Artykuł [1] opisuje klasyfikator MLP oraz łagodzeniu jego wad przez operowanie w zespole klasyfikatorów. SVM zostaje dokładnie opisany z podziałem na etapy analizy danych w [2], opierając się na przykładzie neuroobrazowania mózgu. [3] jest pracą poświęconą predykcji rezygnacji klientów, za pomocą algorytmu GNB, wprowadzając wykorzystanie równoległych obliczeń wartości średniej, odchylenia standardowego i prawdopodobieństwa, w celu zwiększenia szybkości działania algorytmu. W artykule [6], drzewa decyzyjne i SVM zostały przedstawione jako odpowiednie dla większości zadań klasyfikacji, w formie krótkiego przeglądu danych. Drzewa decyzyjne zostały również opisane w pracy [4], omawiając zasadę ich działania, rodzaje problemów, do których mogą one zostać zastosowane oraz podając ich zastosowania w dziedzinie biotechnologii. W ramach publikacji [7] zaprezentowano powstawanie, rozwój oraz zasady działania drzew decyzyjnych oraz metod Boosting i Bagging. W artykule [5],

Boosting przedstawiono jako prosty sposób na ograniczenie występowania problemu błędnej klasyfikacji i zmniejszenie wielkości zestawu uczącego, potrzebnego do osiągnięcia wyników porównywalnych do przypadku użycia pojedynczych klasyfikatorów. Artykuł [8] zagłębia się w temat heterogenicznych zespołów klasyfikatorów, badając wpływ różnorodności zestawów uczących i kompozycji bazowych klasyfikatorów w zespole, w celu osiągnięcia optymalnych wyników klasyfikacji.

III. ZAŁOŻENIA PROJEKTOWE

A. Research questions

- Która z metod optymalizacji hiperparametrów, jest skuteczniejsza?
- Jaki wpływ na jakość klasyfikacji ma dobór hiperparametrów klasyfikatora?
- W jakim stopniu bagging klasyfikatorów bazowych wpływa na skuteczność klasyfikacji?

B. Cele eksperymentu

- Porównanie metod GridSearch i MonteCarlo.
- Optymalizacja hiperparametrów klasyfikatorów bazowych dla zadania klasyfikacji.
- Optymalizacja hiperparametrów zespołu klasyfikatorów.
- Analiza otrzymanych wyników eksperymentów.

C. Opis zestawu danych

Badania zostaną przeprowadzone na zestawie dziesięciu, różnorodnych zestawów danych.

Tabela I
PARAMETRY ZESTAWÓW DANYCH

| Zestaw danych | Liczba instancji | Liczba cech | Liczba klas |
|---------------|------------------|-------------|-------------|
| australian | 690 | 14 | 2 |
| balance | 625 | 4 | 3 |
| breastcan | 683 | 9 | 2 |
| cryotherapy | 90 | 6 | 2 |
| diabetes | 768 | 8 | 2 |
| digit | 10992 | 17 | 10 |
| ecoli4 | 336 | 7 | 2 |
| german | 1000 | 24 | 2 |
| glass2 | 214 | 9 | 2 |
| heart | 270 | 13 | 2 |

IV. PLAN EKSPERYMENTU

A. Opis środowiska eksperymentalnego

Środowisko badawcze zrealizowane zostanie przy pomocy IDE PyCharm Community Edition 2022.3, będącego otwartym oprogramowaniem, umożliwiającym tworzenie i kompilowanie skryptów.

Oprogramowanie testowe stworzone zostanie w języku Python w wersji 3. Tworzenie i szkolenie modeli ML oraz analiza ich skuteczności i wyników zostanie zrealizowana dzięki bibliotece scikit-learn, zbudowanej na bibliotekach NumPy, SciPy i matplotlib.

B. Dokładny opis planu eksperymentu

W projekcie zbudowane zostaną cztery różne modele, wykorzystujące ten sam schemat działania, przyjmując na wejście bootstrap odpowiedniego rozmiaru i dokonując ewaluacji klas instancji. Dla modeli przeprowadzona zostanie walidacja krzyżowa, dzieląc zbiór na uczący i testowy. Po wytrenowaniu otrzymany zostanie współczynnik skuteczności klasyfikacji, na którego podstawie wyznaczone zostaną rankingi klasyfikatorów oraz średnia ranga każdego klasyfikatora, na tle wielu zbiorów danych. Następnie, wykorzystując uzyskane rangi, wykonane zostaną testy parowe klasyfikatorów, przy użyciu metody Wilcozona. Obliczona zostanie w-statystyka, mówiąca o tym, jak bardzo różnią się od siebie wyniki testów klasyfikatorów, oraz p-wartość, stanowiąca prawdopodobieństwo, że zaobserwowane różnice są przypadkowe. Im mniejsze p-value, tym mniejsze prawdopodobieństwo, że różnice są przypadkowe i tym mocniejsze dowody na istnienie rzeczywistej różnicy. Na podstawie uzyskanych statystyk wyznaczone zostaną macierze przewag i statystycznej istotności testów parowych. Porównując pozycje wyznaczonych macierzy, będziemy w stanie powiedzieć, czy uzyskane przewagi mogą zostać uznane jako istotne statystycznie.

Hiperparametry klasyfikatorów podlegające optymalizacji:

- GNB • 'var_smoothing': [1e-9, 1e-3, 1e-6, 1e-12]
- SVM • 'C': [0.1, 1, 10, 100],
• 'gamma': [0.01, 0.1, 1, 10],
- MLP • 'hidden_layer_sizes': [(10,), (20,), (30,)],
• 'alpha': [0.1, 0.01, 0.001],
- DT • 'max_depth': [2, 4, 6, 8],
• 'min_samples_split': [2, 4, 6, 8],
• 'min_samples_leaf': [1, 2, 4, 6]
- Bagging • 'voting': ['hard', 'soft'],
• 'weights': [[1, 1, 1, 1], [2, 1, 1, 1], [1, 2, 1, 1], [1, 1, 2, 1], [1, 1, 1, 2]]

W pierwszym etapie zostanie przeprowadzona optymalizacja hiperparametrów klasyfikatorów bazowych - SVM, GNB, MLP, Drzewa decyzyjne. Celem eksperymentu jest porównanie jakości klasyfikacji na syntetycznym zbiorze danych, po optymalizacji hiperparametrów metodami Grid Search i Monte Carlo.

W następnym badaniu przeprowadzony zostanie ten sam eksperyment, na większej puli zbiorów. Celem badania jest uzyskanie informacji o skuteczności klasyfikacji klasyfikatorów na tle różnorodnych zbiorów danych. Dzięki czemu zostanie uzyskana informacja, który z klasyfikatorów jest najodpowiedniejszy do generalnych problemów klasyfikacji.

W ostatnim etapie eksperymentu zostanie przeprowadzona integracja poszczególnych klasyfikatorów przez wykorzystanie baggingu. W takim klasyfikatorze złożonym, modele określają klasy obiektów osobno, po czym przeprowadzone zostaje głosowanie większościowe. Wynik głosowania spowoduje przypisanie pojedynczej klasy do instancji problemu. Wyniki klasyfikacji zespołu klasyfikatorów zostaną porównane z wynikami klasyfikatorów indywidualnych, pozwalając na ocenę wpływu baggingu na jakość klasyfikacji.

V. WYNIKI BADAŃ I ANALIZY STATYSTYCZNEJ

B. Eksperyment II

A. Eksperyment I

Tabela II
WYNIKI NA SYNTETYCZNYM ZBIORZE DANYCH

| Klasyfikator | GridSearch | MonteCarlo |
|--------------|------------|------------|
| <i>GNB</i> | 0,790 | 0,790 |
| <i>SVM</i> | 0,933 | 0,935 |
| <i>MLP</i> | 0,931 | 0,925 |
| <i>DT</i> | 0,782 | 0,790 |

Tabela III
RANGI KLASYFIKATORÓW NA SYNTETYCZNYM ZBIORZE DANYCH

| Klasyfikator | GridSearch | MonteCarlo |
|----------------|------------|------------|
| <i>GNB</i> | 1,500 | 1,500 |
| <i>SVM</i> | 1,000 | 2,000 |
| <i>MLP</i> | 2,000 | 1,000 |
| <i>DT</i> | 1,000 | 2,000 |
| Średnia | 1,375 | 1,625 |

Tabela IV
STATYSTYKI NA SYNTETYCZNYM ZBIORZE DANYCH

| Macierz w-statistic | | |
|---------------------|------------|------------|
| | GridSearch | MonteCarlo |
| <i>GridSearch</i> | 0,00 | -0,72 |
| <i>MonteCarlo</i> | 0,72 | 0,00 |

| Macierz p-value | | |
|-------------------|------------|------------|
| | GridSearch | MonteCarlo |
| <i>GridSearch</i> | 1,00 | 0,47 |
| <i>MonteCarlo</i> | 0,47 | 1,00 |

Tabela V
PRZEWAGI NA SYNTETYCZNYM ZBIORZE DANYCH

| Macierz przewag | | |
|-------------------|------------|------------|
| | GridSearch | MonteCarlo |
| <i>GridSearch</i> | 0 | 0 |
| <i>MonteCarlo</i> | 1 | 0 |

| Macierz statystycznej istotności (alpha=0,05) | | |
|---|------------|------------|
| | GridSearch | MonteCarlo |
| <i>GridSearch</i> | 0 | 0 |
| <i>MonteCarlo</i> | 0 | 0 |

Tabela VI
WYNIKI GRIDSEARCH NA 10 ZBIORACH DANYCH

| Zbiór danych | GNB | SVM | MLP | DT |
|--------------------|-------|-------|-------|-------|
| <i>australian</i> | 0,796 | 0,538 | 0,604 | 0,857 |
| <i>balance</i> | 0,642 | 0,903 | 0,933 | 0,574 |
| <i>breastcan</i> | 0,965 | 0,970 | 0,948 | 0,933 |
| <i>cryoteraphy</i> | 0,850 | 0,868 | 0,620 | 0,853 |
| <i>diabetes</i> | 0,705 | 0,500 | 0,574 | 0,701 |
| <i>digit</i> | 1,000 | 0,523 | 0,991 | 1,000 |
| <i>ecoli4</i> | 0,915 | 0,784 | 0,562 | 0,783 |
| <i>german</i> | 0,625 | 0,574 | 0,688 | 0,655 |
| <i>glass2</i> | 0,652 | 0,511 | 0,500 | 0,515 |
| <i>heart</i> | 0,833 | 0,555 | 0,695 | 0,754 |

Tabela VII
RANGI KLASYFIKATORÓW DLA GRIDSEARCH NA 10 ZBIORACH DANYCH

| Zbiór danych | GNB | SVM | MLP | DT |
|--------------------|------|------|------|------|
| <i>australian</i> | 3,00 | 1,00 | 2,00 | 4,00 |
| <i>balance</i> | 2,00 | 3,00 | 4,00 | 1,00 |
| <i>breastcan</i> | 3,00 | 4,00 | 2,00 | 1,00 |
| <i>cryoteraphy</i> | 2,00 | 4,00 | 1,00 | 3,00 |
| <i>diabetes</i> | 4,00 | 1,00 | 2,00 | 3,00 |
| <i>digit</i> | 3,50 | 1,00 | 2,00 | 3,50 |
| <i>ecoli4</i> | 4,00 | 3,00 | 1,00 | 2,00 |
| <i>german</i> | 2,00 | 1,00 | 4,00 | 3,00 |
| <i>glass2</i> | 4,00 | 2,00 | 1,00 | 3,00 |
| <i>heart</i> | 4,00 | 1,00 | 2,00 | 3,00 |
| Średnia | 3,15 | 2,10 | 2,10 | 2,65 |

Tabela VIII
STATYSTYKI GRIDSEARCH NA 10 ZBIORACH DANYCH

| Macierz w-statistic | | | | |
|---------------------|-------|------|-------|-------|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 0,00 | 1,85 | 2,00 | 1,13 |
| <i>SVM</i> | -1,85 | 0,00 | -0,23 | -0,94 |
| <i>MLP</i> | -2,00 | 0,23 | 0,00 | -1,17 |
| <i>DT</i> | -1,13 | 0,94 | 1,17 | 0,00 |

| Macierz p-value | | | | |
|-----------------|------|------|------|------|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 1,00 | 0,06 | 0,05 | 0,26 |
| <i>SVM</i> | 0,06 | 1,00 | 0,82 | 0,34 |
| <i>MLP</i> | 0,05 | 0,82 | 1,00 | 0,24 |
| <i>DT</i> | 0,26 | 0,34 | 0,24 | 1,00 |

Tabela IX
PRZEWAGI GRIDSEARCH NA 10 ZBIORACH DANYCH

| Macierz Przewag | | | | |
|-----------------|-----|-----|-----|----|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 0 | 1 | 1 | 1 |
| <i>SVM</i> | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 0 | 1 | 0 | 0 |
| <i>DT</i> | 0 | 1 | 1 | 0 |

| Macierz statystycznej istotności (alpha = 0,05) | | | | |
|---|-----|-----|-----|----|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 0 | 0 | 1 | 0 |
| <i>SVM</i> | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 1 | 0 | 0 | 0 |
| <i>DT</i> | 0 | 0 | 0 | 0 |

Tabela X
WYNIKI MONTECARLO NA 10 ZBIORACH DANYCH

| Zbiór danych | GNB | SVM | MLP | DT |
|--------------------|-------|-------|-------|-------|
| <i>australian</i> | 0,639 | 0,582 | 0,727 | 0,858 |
| <i>balance</i> | 0,642 | 0,786 | 0,906 | 0,576 |
| <i>breastcan</i> | 0,965 | 0,961 | 0,952 | 0,933 |
| <i>cryoteraphy</i> | 0,850 | 0,874 | 0,785 | 0,841 |
| <i>diabetes</i> | 0,712 | 0,500 | 0,603 | 0,700 |
| <i>digit</i> | 1,000 | 0,523 | 0,991 | 1,000 |
| <i>ecoli4</i> | 0,910 | 0,637 | 0,613 | 0,780 |
| <i>german</i> | 0,689 | 0,560 | 0,688 | 0,653 |
| <i>glass2</i> | 0,640 | 0,500 | 0,500 | 0,492 |
| <i>heart</i> | 0,837 | 0,555 | 0,662 | 0,759 |

Tabela XI
RANGI KLASYFIKATORÓW DLA MONTECARLO NA 10 ZBIORACH DANYCH

| Zbiór danych | GNB | SVM | MLP | DT |
|--------------------|------|------|------|------|
| <i>australian</i> | 3,00 | 1,00 | 3,00 | 4,00 |
| <i>balance</i> | 2,00 | 3,00 | 4,00 | 1,00 |
| <i>breastcan</i> | 4,00 | 3,00 | 2,00 | 1,00 |
| <i>cryoteraphy</i> | 3,00 | 4,00 | 1,00 | 2,00 |
| <i>diabetes</i> | 4,00 | 1,00 | 2,00 | 3,00 |
| <i>digit</i> | 3,50 | 1,00 | 2,00 | 3,50 |
| <i>ecoli4</i> | 4,00 | 2,00 | 1,00 | 3,00 |
| <i>german</i> | 4,00 | 1,00 | 3,00 | 2,00 |
| <i>glass2</i> | 4,00 | 2,50 | 2,50 | 1,00 |
| <i>heart</i> | 4,00 | 1,00 | 2,00 | 3,00 |
| Średnia | 3,45 | 1,95 | 2,25 | 2,35 |

Tabela XII
STATYSTYKI MONTECARLO NA 10 ZBIORACH DANYCH

| Macierz w-statistic | | | | |
|---------------------|-------|------|-------|-------|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 0,00 | 2,65 | 2,42 | 2,27 |
| <i>SVM</i> | -2,65 | 0,00 | -0,68 | -0,83 |
| <i>MLP</i> | -2,42 | 0,68 | 0,00 | -0,26 |
| <i>DT</i> | -2,27 | 0,83 | 0,26 | 0,00 |

| Macierz p-value | | | | |
|-----------------|------|------|------|------|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 1,00 | 0,01 | 0,02 | 0,02 |
| <i>SVM</i> | 0,01 | 1,00 | 0,50 | 0,41 |
| <i>MLP</i> | 0,02 | 0,50 | 1,00 | 0,79 |
| <i>DT</i> | 0,02 | 0,41 | 0,79 | 1,00 |

Tabela XIII
PRZEWAGI MONTECARLO NA 10 ZBIORACH DANYCH

| Macierz Przewag | | | | |
|-----------------|-----|-----|-----|----|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 0 | 1 | 1 | 1 |
| <i>SVM</i> | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 0 | 1 | 0 | 0 |
| <i>DT</i> | 0 | 1 | 1 | 0 |

| Macierz statystycznej istotności (alpha = 0,05) | | | | |
|---|-----|-----|-----|----|
| | GNB | SVM | MLP | DT |
| <i>GNB</i> | 0 | 1 | 1 | 1 |
| <i>SVM</i> | 1 | 0 | 0 | 0 |
| <i>MLP</i> | 1 | 0 | 0 | 0 |
| <i>DT</i> | 1 | 0 | 0 | 0 |

C. Eksperyment III

Tabela XIV
WYNIKI GRIDSEARCH BAGGING

| Zbiór danych | GNB | SVM | MLP | DT | Bagging |
|--------------------|-------|-------|-------|-------|---------|
| <i>australian</i> | 0,796 | 0,538 | 0,728 | 0,857 | 0,833 |
| <i>balance</i> | 0,642 | 0,903 | 0,934 | 0,581 | 0,748 |
| <i>breastcan</i> | 0,965 | 0,970 | 0,947 | 0,936 | 0,969 |
| <i>cryoteraphy</i> | 0,850 | 0,868 | 0,761 | 0,853 | 0,928 |
| <i>diabetes</i> | 0,705 | 0,500 | 0,583 | 0,700 | 0,693 |
| <i>digit</i> | 1,000 | 0,523 | 0,992 | 1,000 | 1,000 |
| <i>ecoli4</i> | 0,915 | 0,784 | 0,550 | 0,783 | 0,834 |
| <i>german</i> | 0,625 | 0,574 | 0,662 | 0,653 | 0,672 |
| <i>glass2</i> | 0,652 | 0,511 | 0,500 | 0,517 | 0,497 |
| <i>heart</i> | 0,833 | 0,555 | 0,600 | 0,750 | 0,836 |

Tabela XV
RANGI KLASYFIKATORÓW DLA GRIDSEARCH BAGGING

| Zbiór danych | GNB | SVM | MLP | DT | Bagging |
|--------------------|-----|-----|-----|-----|---------|
| <i>australian</i> | 3,0 | 1,0 | 2,0 | 5,0 | 4,0 |
| <i>balance</i> | 2,0 | 4,0 | 5,0 | 1,0 | 3,0 |
| <i>breastcan</i> | 3,0 | 5,0 | 2,0 | 1,0 | 4,0 |
| <i>cryoteraphy</i> | 2,0 | 4,0 | 1,0 | 3,0 | 5,0 |
| <i>diabetes</i> | 5,0 | 1,0 | 2,0 | 4,0 | 3,0 |
| <i>digit</i> | 4,0 | 1,0 | 2,0 | 4,0 | 4,0 |
| <i>ecoli4</i> | 5,0 | 3,0 | 1,0 | 2,0 | 4,0 |
| <i>german</i> | 2,0 | 1,0 | 4,0 | 3,0 | 5,0 |
| <i>glass2</i> | 5,0 | 3,0 | 2,0 | 4,0 | 1,0 |
| <i>heart</i> | 4,0 | 1,0 | 2,0 | 3,0 | 5,0 |
| Średnia | 3,5 | 2,4 | 2,3 | 3,0 | 3,8 |

Tabela XVI
STATYSTYKI GRIDSEARCH BAGGING

| Macierz w-statistic | | | | | |
|---------------------|-------|------|-------|-------|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 0,00 | 1,63 | 2,00 | 0,76 | -0,57 |
| <i>SVM</i> | -1,63 | 0,00 | -0,19 | -0,87 | -1,89 |
| <i>MLP</i> | -2,00 | 0,19 | 0,00 | -1,21 | -2,23 |
| <i>DT</i> | -0,76 | 0,87 | 1,21 | 0,00 | -1,40 |
| <i>Bagging</i> | 0,57 | 1,89 | 2,23 | 1,40 | 0,00 |

| Macierz p-value | | | | | |
|-----------------|------|------|------|------|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 1,00 | 0,10 | 0,05 | 0,45 | 0,57 |
| <i>SVM</i> | 0,10 | 1,00 | 0,85 | 0,38 | 0,06 |
| <i>MLP</i> | 0,05 | 0,85 | 1,00 | 0,23 | 0,03 |
| <i>DT</i> | 0,45 | 0,38 | 0,23 | 1,00 | 0,16 |
| <i>Bagging</i> | 0,57 | 0,06 | 0,03 | 0,16 | 1,00 |

Tabela XVII
PRZEWAGI GRIDSEARCH BAGGING

| Macierz Przewag | | | | | |
|-----------------|-----|-----|-----|----|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 0 | 1 | 1 | 1 | 0 |
| <i>SVM</i> | 0 | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 0 | 1 | 0 | 0 | 0 |
| <i>DT</i> | 0 | 1 | 1 | 0 | 0 |
| <i>Bagging</i> | 1 | 1 | 1 | 1 | 0 |

| Macierz statystycznej istotności (alpha = 0,05) | | | | | |
|---|-----|-----|-----|----|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 0 | 0 | 1 | 0 | 0 |
| <i>SVM</i> | 0 | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 1 | 0 | 0 | 0 | 1 |
| <i>DT</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Bagging</i> | 0 | 0 | 1 | 0 | 0 |

Tabela XVIII
WYNIKI MONTECARLO BAGGING

| Zbiór danych | GNB | SVM | MLP | DT | Bagging |
|--------------------|-------|-------|-------|-------|---------|
| <i>australian</i> | 0.639 | 0.582 | 0.727 | 0.858 | 0.847 |
| <i>balance</i> | 0.642 | 0.786 | 0.923 | 0.576 | 0.762 |
| <i>breastcan</i> | 0.965 | 0.961 | 0.947 | 0.933 | 0.971 |
| <i>cryoteraphy</i> | 0.850 | 0.874 | 0.888 | 0.841 | 0.866 |
| <i>diabetes</i> | 0.712 | 0.500 | 0.630 | 0.701 | 0.673 |
| <i>digit</i> | 1,000 | 0,523 | 0,991 | 1,000 | 0,996 |
| <i>ecoli4</i> | 0.910 | 0.637 | 0.625 | 0.781 | 0.759 |
| <i>german</i> | 0.689 | 0.560 | 0.703 | 0.647 | 0.677 |
| <i>glass2</i> | 0.640 | 0.500 | 0.500 | 0.492 | 0.495 |
| <i>heart</i> | 0.837 | 0.555 | 0.603 | 0.745 | 0.723 |

Tabela XIX
RANGI KLASYFIKATORÓW DLA MONTECARLO BAGGING

| Zbiór danych | GNB | SVM | MLP | DT | Bagging |
|--------------------|------|------|------|------|---------|
| <i>australian</i> | 2,00 | 1,00 | 3,00 | 5,00 | 4,00 |
| <i>balance</i> | 2,00 | 4,00 | 5,00 | 1,00 | 3,00 |
| <i>breastcan</i> | 4,00 | 3,00 | 2,00 | 1,00 | 5,00 |
| <i>cryoteraphy</i> | 2,00 | 4,00 | 5,00 | 1,00 | 3,00 |
| <i>diabetes</i> | 5,00 | 1,00 | 2,00 | 4,00 | 3,00 |
| <i>digit</i> | 4,50 | 1,00 | 2,00 | 4,50 | 3,00 |
| <i>ecoli4</i> | 5,00 | 2,00 | 1,00 | 4,00 | 3,00 |
| <i>german</i> | 4,00 | 1,00 | 5,00 | 2,00 | 3,00 |
| <i>glass2</i> | 5,00 | 3,50 | 3,50 | 1,00 | 2,00 |
| <i>heart</i> | 5,00 | 1,00 | 2,00 | 4,00 | 3,00 |
| Średnia | 3,85 | 2,15 | 3,05 | 2,75 | 3,20 |

Tabela XX
STATYSTYKI MONTECARLO BAGGING

| Macierz w-statistic | | | | | |
|---------------------|-------|------|-------|-------|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 0,00 | 2,61 | 1,06 | 1,74 | 1,17 |
| <i>SVM</i> | -2,61 | 0,00 | -1,40 | -0,94 | -1,51 |
| <i>MLP</i> | -1,06 | 1,40 | 0,00 | 0,64 | -0,53 |
| <i>DT</i> | -1,74 | 0,94 | -0,64 | 0,00 | -0,42 |
| <i>Bagging</i> | -1,17 | 1,51 | 0,53 | 0,42 | 0,00 |

| Macierz p-value | | | | | |
|-----------------|------|------|------|------|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 1,00 | 0,01 | 0,29 | 0,08 | 0,24 |
| <i>SVM</i> | 0,01 | 1,00 | 0,16 | 0,34 | 0,13 |
| <i>MLP</i> | 0,29 | 0,16 | 1,00 | 0,52 | 0,60 |
| <i>DT</i> | 0,08 | 0,34 | 0,52 | 1,00 | 0,68 |
| <i>Bagging</i> | 0,24 | 0,13 | 0,60 | 0,68 | 1,00 |

Tabela XXI
PRZEWAGI MONTECARLO BAGGING

| Macierz Przewag | | | | | |
|-----------------|-----|-----|-----|----|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 0 | 1 | 1 | 1 | 1 |
| <i>SVM</i> | 0 | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 0 | 1 | 0 | 1 | 0 |
| <i>DT</i> | 0 | 1 | 0 | 0 | 0 |
| <i>Bagging</i> | 0 | 1 | 1 | 1 | 0 |

| Macierz statystycznej istotności (alpha = 0,05) | | | | | |
|---|-----|-----|-----|----|---------|
| | GNB | SVM | MLP | DT | Bagging |
| <i>GNB</i> | 0 | 1 | 0 | 0 | 0 |
| <i>SVM</i> | 1 | 0 | 0 | 0 | 0 |
| <i>MLP</i> | 0 | 0 | 0 | 0 | 0 |
| <i>DT</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Bagging</i> | 0 | 0 | 0 | 0 | 0 |

D. Eksperyment IV - Klasyfikator SVM

Analizując wyniki klasyfikacji dla obu metod przeszukiwania, zawartych w tabelach VI, X, XIV i XVIII, warto zwrócić uwagę na relatywnie słabe wyniki SVM na tle pozostałych klasyfikatorów.

Wykazuje on bardzo słabą jakość klasyfikacji, często bliską klasyfikatorowi losowemu. W przypadku zbioru danych o nazwie "digits" pozostałe metody wyuczają się w stopniu

niemalże idealnym, natomiast SVM uzyskuje skuteczność w okolicy 50%.

Według pierwszych podejrzeń, mogło to wynikać z użycia bazowej wartości parametru "class_weight" klasyfikatora SVM.

Przeprowadzone zostały więc dodatkowe testy dla wszystkich zbiorów ze zmienioną wartością tego parametru na 'balanced', który oblicza wagę dla każdej klasy jako odwrotność liczby przykładów w danej klasie. Dzięki temu, model SVM jest bardziej odpowiedzialny za klasy, które są mniej reprezentowane w danym zbiorze.

Tabela XXII
WYNIKI DLA SVM (CLASS_WEIGHT='BALANCED')

| Zbiór danych | GridSearch | MonteCarlo |
|--------------|------------|------------|
| australian | 0,651 | 0,541 |
| balance | 0,932 | 0,896 |
| breastcan | 0,969 | 0,967 |
| cryoteraphy | 0,875 | 0,874 |
| diabetes | 0,500 | 0,500 |
| digit | 0,523 | 0,523 |
| ecoli4 | 0,781 | 0,704 |
| german | 0,560 | 0,560 |
| glass2 | 0,614 | 0,739 |
| heart | 0,558 | 0,555 |

Kolejnym parametrem podejrzanym o obniżanie skuteczności klasyfikatora był "decision_function_shape", decydujący o kształcie funkcji decyzyjnej.

Domyślne ustawienie parametru stanowi 'ovr'(all-vs-rest). W badaniu dodana została wartość 'ovo'(all-vs-all), stosowana jako wielo-klasowa strategia trenowania modelu.

Tabela XXIII
WYNIKI DLA SVM (DECISION_FUNCTION_SHAPE='OVO')

| Zbiór danych | GridSearch | MonteCarlo |
|--------------|------------|------------|
| australian | 0,538 | 0,630 |
| balance | 0,903 | 0,903 |
| breastcan | 0,970 | 0,960 |
| cryoteraphy | 0,868 | 0,874 |
| diabetes | 0,500 | 0,500 |
| digit | 0,523 | 0,523 |
| ecoli4 | 0,784 | 0,809 |
| german | 0,574 | 0,585 |
| glass2 | 0,511 | 0,524 |
| heart | 0,555 | 0,555 |

VI. WNIOSKI

A. Eksperyment I

Analizując wyniki klasyfikacji dla pojedynczego syntetycznego zbioru danych nie można zauważyć znaczących różnic w jakości klasyfikacji. Obie metody wykazują zadawalające działanie, skutecznie optymalizując hiperparametry klasyfikatorów. Według rang, metoda MonteCarlo uzyskuje nieznacznie lepsze wyniki, jednak na podstawie wyników testu parowego Wilcozona, w żadnym z testowanych zestawień nie możemy

odrzuć hipotezy zerowej. Oznacza to, że pomimo występujących różnic nie można jednoznacznie określić statystycznie lepszej metody.

B. Eksperyment II

Dla 10 zróżnicowanych zestawów danych, optymalizacja metodą GridSearch wykazała, że klasyfikator GNB klasyfikuje najlepiej ze wszystkich klasyfikatorów, jednakże wynik ten jest istotny statystycznie tylko w przypadku porównania z MLP.

W przypadku przeszukiwania MonteCarlo, klasyfikator GNB również okazał się najskuteczniejszy, jednakże tym razem wyniki testów parowych pozwalają stwierdzić, że jest on lepszy statystycznie od wszystkich pozostałych klasyfikatorów.

C. Eksperyment III

Analizując wyniki klasyfikatorów bazowych oraz heterogenicznego baggingu można zauważyć poprawienie skuteczności klasyfikacji w porównaniu z klasyfikatorami bazowymi. W przeszukiwaniu GridSearch, zespół klasyfikatorów wykazał się najlepszą skutecznością, a tuż za nim znalazł się klasyfikator GNB. Jedną istotną statystycznie przewagę ma on nad klasyfikatorem MLP.

W przypadku przeszukiwania MonteCarlo, wyniki okazały się inne. Metoda GNB wyprzedziła skutecznością zespół klasyfikatorów, który znalazł się na miejscu drugim. Pomimo widocznej przewagi baggingu nad trzema pozostałymi klasyfikatorami, nie można jednoznacznie stwierdzić, że jest on lepszy statystycznie.

D. Eksperyment IV

Wyniki wykazały, iż jakość klasyfikacji nie uległa poprawie przy optymalizacji wartości obu badanych parametrów, a nawet pogorszyła się w przypadku niektórych zbiorów. Takie rezultaty nie pozwoliły na wyjaśnienie niskiego współczynnika poprawnej klasyfikacji modelu SVM, co skłania do dalszej refleksji i głębszego zastanowienia się nad tym problemem.

LITERATURA

- [1] T. Windeatt, "Ensemble mlp classifier design," in *Computational Intelligence Paradigms*. Springer, 2008, pp. 133–147.
- [2] A. Mechelli and S. Viera, *Machine learning: methods and applications to brain disorders*. Academic Press, 2019.
- [3] D. T. Barus, R. Elfarizy, F. Masri, and P. Gunawan, "Parallel programming of churn prediction using gaussian naïve bayes," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2020, pp. 1–4.
- [4] C. Kingsford and S. L. Salzberg, "What are decision trees?" *Nature biotechnology*, vol. 26, no. 9, pp. 1011–1013, 2008.
- [5] T. Hothorn and B. Lausen, "Double-bagging: combining classifiers by bootstrap aggregation," *Pattern Recognition*, vol. 36, no. 6, pp. 1303–1309, 2003.
- [6] M. Somvanshi, P. Chavan, S. Tambade, and S. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *2016 international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2016, pp. 1–7.
- [7] C. D. Sutton, "Classification and regression trees, bagging, and boosting," *Handbook of statistics*, vol. 24, pp. 303–329, 2005.
- [8] M. P. Sesmero, J. A. Iglesias, E. Magán, A. Ledezma, and A. Sanchis, "Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles," *Applied Soft Computing*, vol. 111, p. 107689, 2021.