

WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH
POLITECHNIKA WARSZAWSKA

Analiza lasów losowych z baggingiem wykorzystującym
Borderline-SMOTE

PROJEKT INTERDYSCYPLINARNY
RAPORT

Michał Szewczak, Filip Szlingiert, Kornel Tłaczała

12 czerwca 2025

Spis treści

1	Wstęp	3
2	Metoda Borderline-Smote	3
2.1	Kroki Borderline-Smote	3
3	Oversampling w lasach losowych	3
4	Przeprowadzone Eksperymenty	3
5	Wyniki	4
5.1	Zbiór car eval 34 (niezbalansowanie: 12 : 1)	4
5.1.1	Wyniki	4
5.1.2	Wnioski	6
5.2	Zbiór Ecoli (niezbalansowanie: 8.6 : 1)	7
5.2.1	Wyniki	7
5.2.2	Wnioski	9
5.3	Zbiór Wine Quality (niezbalansowanie: 26:1)	10
5.3.1	Wyniki	10
5.3.2	Wnioski	12
6	Wnioski ogólne	13
7	Bibliografia	13

1 Wstęp

Często spotykanym w klasyfikacji problem jest problem niezbalansowania klas. Jest to sytuacja, w której jedna z klas jest istotnie mniej liczna od drugiej, co skutkuje faworyzacją predykcji klasy większościowej w przypadku wielu modeli klasyfikacyjnych. Jednak w wielu sytuacjach wykrycie klasy mniejszościowej może być znacznie istotniejsze z punktu widzenia użytkownika (np. oszustwa bankowe i pranie brudnych pieniędzy czy diagnostyka chorób). Jednym z częściej stosowanych sposobów radzenia sobie z tym problemem jest oversampling, czyli sztuczne dodanie do zbioru nowych elementów klasy mniejszościowej, tak by zbiór był bardziej zbalansowany. Nie zawsze oznacza to uzyskanie stosunku 1:1, gdyż w przypadku zbiorów z bardzo mocną dominacją klasy większościowej (np. 10:1) mogłoby to doprowadzić do zbytnej nadpredykcji klasy mniejszościowej. W ramach naszej pracy skupiliśmy się na metodzie oversamplingu znanej jako Borderline-SMOTE i jej użyciu w trenowaniu lasów losowych, a także porównaliśmy jej działanie z kilkoma innymi metodami.

2 Metoda Borderline-SMOTE

Metoda Borderline-SMOTE została wprowadzona jako udoskonalenie metody SMOTE, przykładające większą wagę do obserwacji bliskich granicy decyzyjnej. Zamiast tworzyć nowe próbki dla wszystkich przykładów klasy mniejszościowej (jak w SMOTE), metoda ta koncentruje się wyłącznie na przykładach leżących na granicy klas, które są bardziej narażone na błędną klasyfikację. Dzięki temu nowe dane syntetyczne w założeniu wzmacniają jedynie te punkty, które są kluczowe dla poprawy jakości klasyfikacji.

2.1 Kroki Borderline-SMOTE

Krok 1: Dla każdego elementu klasy mniejszościowej w zbiorze znajdujemy m najbliższych sąsiadów w całym zbiorze, a liczbę elementów z klasy większościowej wśród nich oznaczamy jako m' .

Krok 2: Jeżeli $m'=m$, to tą obserwację klasy mniejszościowej uznajemy za szum. Jeżeli $m' \leq m/2$, to obserwacja jest uznawana za "bezpieczną", a więc nie leżącą w pobliżu granicy. Jeżeli $m/2 < m' < m$ to wówczas uznajemy taką obserwację za "zagrożoną", czyli leżącą blisko granicy.

Krok 3: W ramach tego kroku generujemy sztuczne obserwacje do momentu uzyskania pożądanej liczby obserwacji klasy mniejszościowej. Dla generowania nowych obserwacji rozważamy tylko te z oryginalnych, które zostały zakwalifikowane jako "zagrożone". Następnie wybieramy n najbliższych sąsiadów spośród zagrożonych elementów klasy mniejszościowej, wyznaczamy sumę różnic między nimi i wybranym wcześniej punktem i dodajemy ją do tego punktu (pomnożoną przez pewien losowy współczynnik) otrzymując nową obserwację.

3 Oversampling w lasach losowych

Las losowy (ang. Random Forest) jest klasyfikatorem opartym na agregacji wyników składających się na niego drzew decyzyjnych. Wiele badań pokazuje, że jest to klasyfikator dość odporny zarówno na szum i outliery jak i na niezbalansowanie klas. Mimo tego badania nad poprawą klasyfikacji w takich zbiorach wskazują na możliwość uzyskania lepszych wyników poprzez użycie modyfikacji lasów, np. poprzez dodanie wag dla klas (klasa mniejszościowa ma wówczas większą wagę i kara za jej złe sklasyfikowanie jest większa), bądź poprzez techniki augmentacji danych. Augmentacja ta może być przeprowadzona zarówno na całym zbiorze, i wówczas z próbki bootstapowe dla poszczególnych drzew są losowane ze zbioru po augmentacji, bądź w procesie baggingu, tzn. oversampling następuje po wylosowaniu próbki dla każdego z drzew osobno.

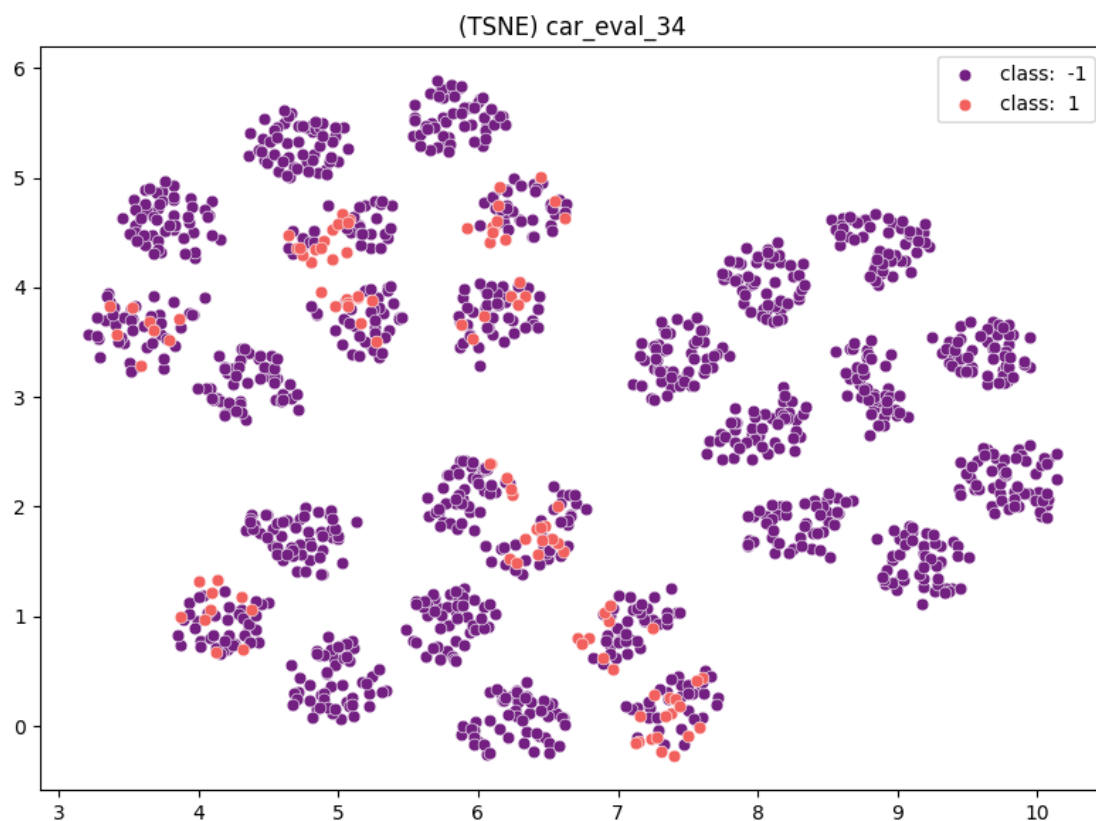
4 Przeprowadzone Eksperymenty

W ramach projektu badaliśmy wpływ doboru metody i momentu oversamplingu dla lasów losowych. W szczególności celem było sprawdzenie użyteczności metody borderline-SMOTE, zarówno w postaci augmentacji na całym zbiorze jak i przede wszystkim baggingu, poprzez ich implementację i porównanie z innymi metodami: losową augmentacją polegającą na wielokrotnym wyborze wylosowanych

obserwacji, Smote i ADASYN oraz baselinowym modelem bez augmentacji. metody te testowaliśmy na zbiorach udostępnianych w bibliotece ImbalancedLearn, charakteryzujących się różnymi stopniami niezbalansowania. Dla każdego ze zbioru model testowano wielokrotnie, początkowo z Sampling Rate 0.5, a następnie dla bardziej obiecujących zbiorów testowaliśmy również inne wartości tego parametru.

5 Wyniki

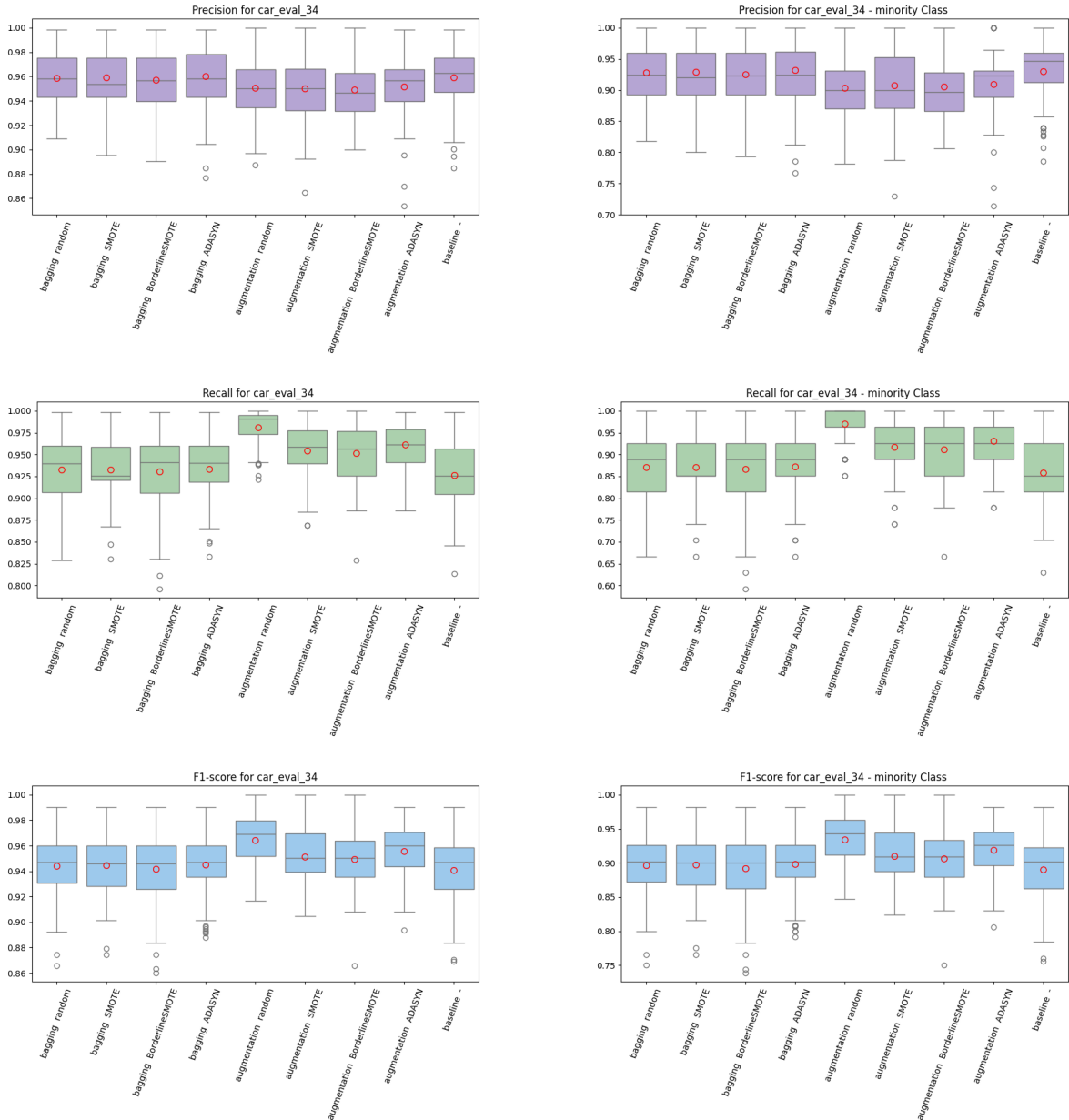
5.1 Zbiór car eval 34 (niezbalansowanie: 12 : 1)

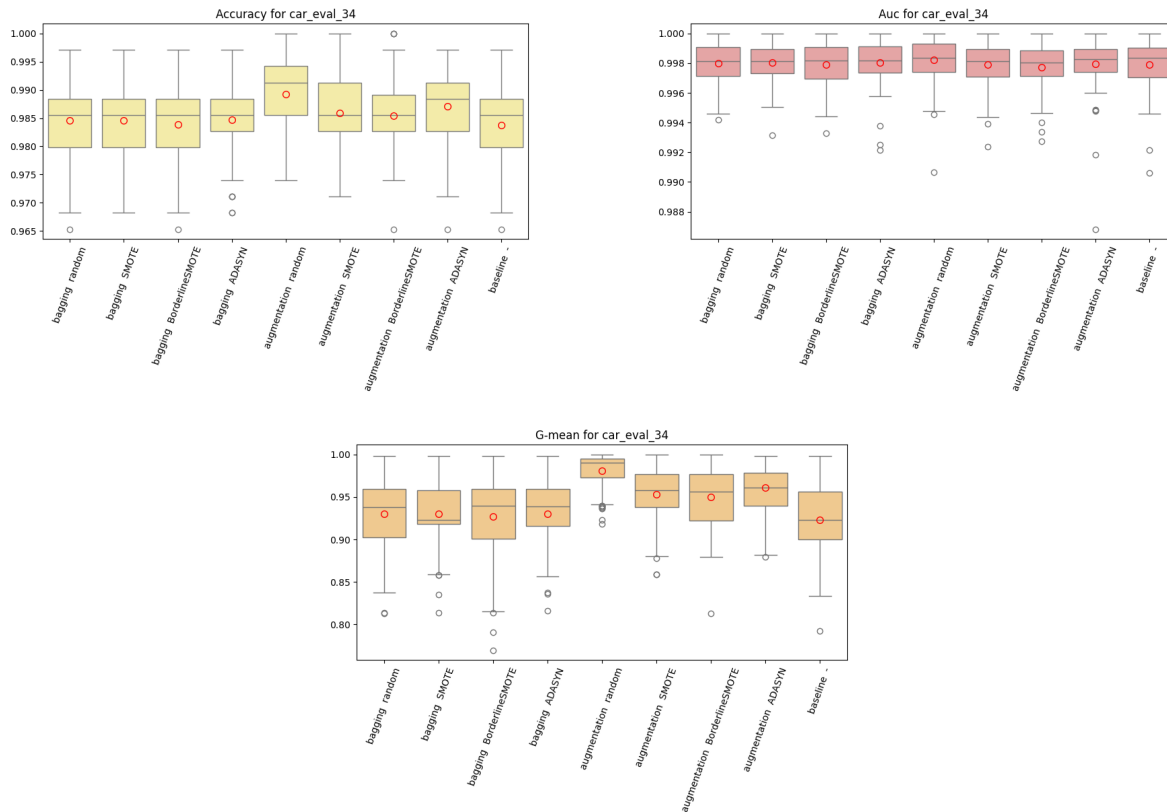


5.1.1 Wyniki

type	strategy	accuracy	auc	f1-score	g-mean	precision	recall
Minority class statistics							
bagging	random	-	-	0.8968	-	0.9281	0.8715
bagging	SMOTE	-	-	0.8973	-	0.9289	0.8711
bagging	BorderlineSMOTE	-	-	0.8925	-	0.9254	0.8670
bagging	ADASYN	-	-	0.8988	-	0.9317	0.8722
augmentation	random	-	-	0.9345	-	0.9034	0.9707
augmentation	SMOTE	-	-	0.9103	-	0.9070	0.9174
augmentation	BorderlineSMOTE	-	-	0.9066	-	0.9054	0.9111
augmentation	ADASYN	-	-	0.9188	-	0.9095	0.9315
baseline	-	-	-	0.8907	-	0.9302	0.8585
Average metrics							
bagging	random	0.9845	0.9980	0.9442	0.9299	0.9586	0.9328
bagging	SMOTE	0.9846	0.9980	0.9445	0.9299	0.9591	0.9326
bagging	BorderlineSMOTE	0.9839	0.9979	0.9419	0.9272	0.9571	0.9304
bagging	ADASYN	0.9848	0.9981	0.9453	0.9306	0.9605	0.9333
augmentation	random	0.9893	0.9982	0.9643	0.9805	0.9505	0.9808
augmentation	SMOTE	0.9859	0.9979	0.9513	0.9533	0.9500	0.9546
augmentation	BorderlineSMOTE	0.9854	0.9977	0.9494	0.9500	0.9489	0.9514
augmentation	ADASYN	0.9871	0.9980	0.9559	0.9608	0.9519	0.9617
baseline	-	0.9838	0.9979	0.9409	0.9230	0.9592	0.9264

Tabela 1: Porównanie średnich metryk dla zbioru Cars eval 34 przy Sampling Rate = 0.5.





5.1.2 Wnioski

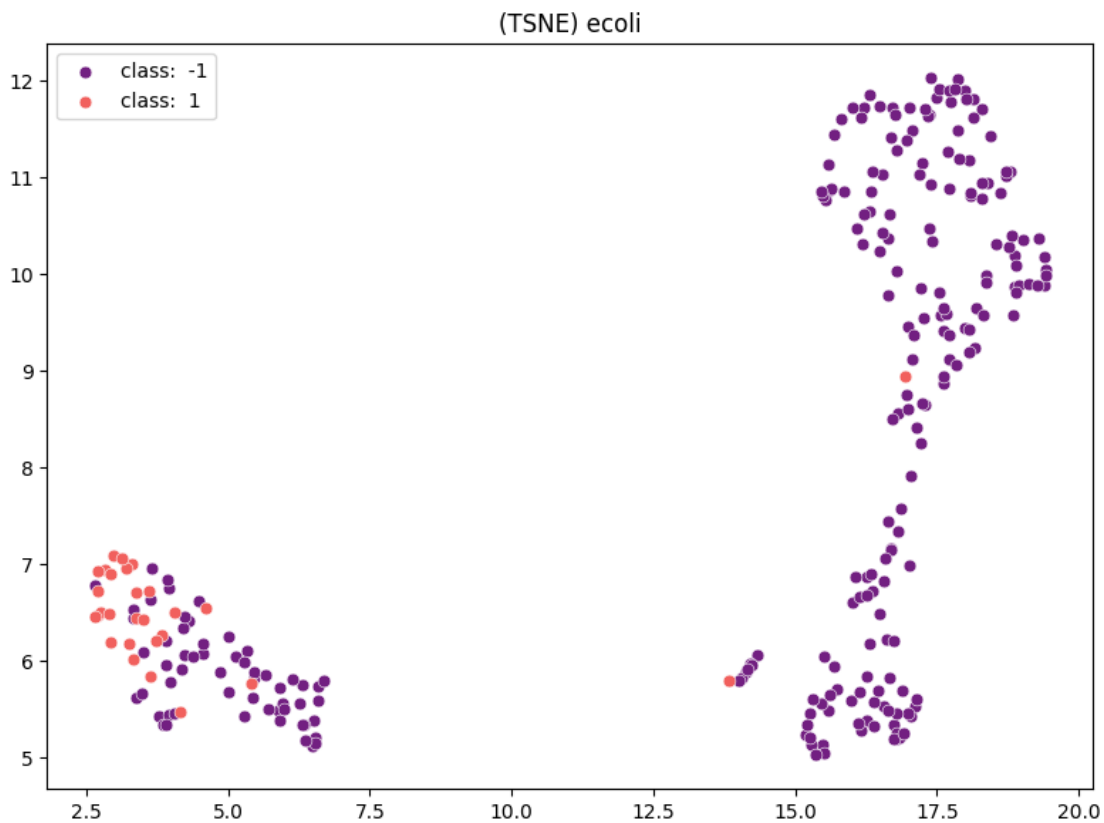
Warto zauważyć, że augmentacja danych generalnie wypadła lepiej niż bagging, jeśli chodzi o średnie metryki. W przypadku augmentacji, szczególnie przy strategii random, osiągnięto wyższe wartości accuracy, AUC, F1-score oraz recall w porównaniu do baggingu. Sugeruje to, że augmentacja może być bardziej skuteczna w poprawie jakości klasyfikacji dla tego zbioru danych.

Dla baggingu z BorderlineSMOTE, wyniki są nieco słabsze niż w przypadku innych metod oversamplingu, takich jak ADASYN czy klasyczny SMOTE. BorderlineSMOTE osiągnął F1-score na poziomie 0.9419 dla średnich metryk i 0.8925 dla klasy mniejszościowej, co jest wynikiem niższym niż w przypadku ADASYN czy nawet podstawowego SMOTE. Może to wynikać z faktu, że BorderlineSMOTE skupia się na przykładach znajdujących się na granicy decyzyjnej, co w niektórych przypadkach może prowadzić do mniej stabilnych predykcji.

Ciekawym spostrzeżeniem jest również to, że bagging charakteryzuje się wyższą precyzją niż augmentacja, ale niższym recall. W przypadku augmentacji mamy odwrotną sytuację – wyższy recall, ale nieco niższą precyzję. Może to oznaczać, że augmentacja lepiej radzi sobie z wykrywaniem przykładów klasy mniejszościowej, ale kosztem większej liczby fałszywych trafień.

Podsumowując, choć BorderlineSMOTE w baggingu dał rezultaty zbliżone do innych metod oversamplingu, to jednak nie wyróżnił się na ich tle. Augmentacja okazała się ogólnie lepszym rozwiązaniem, szczególnie w kontekście poprawy recall i F1-score.

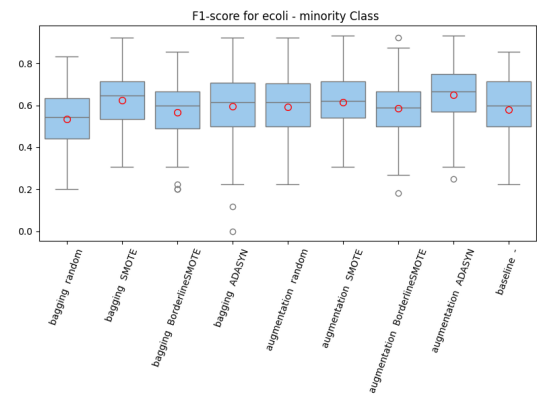
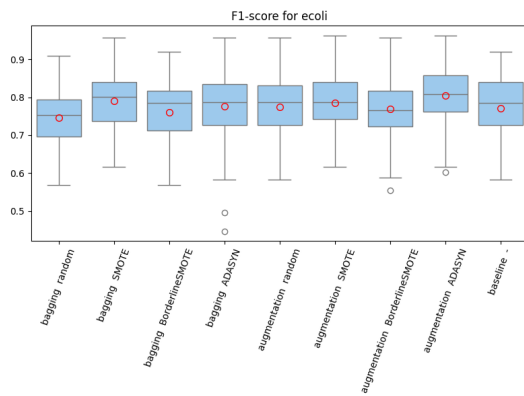
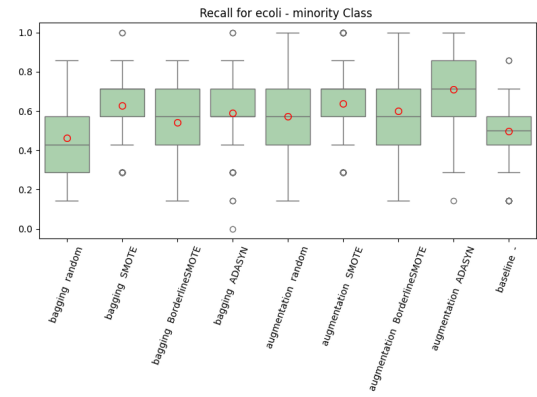
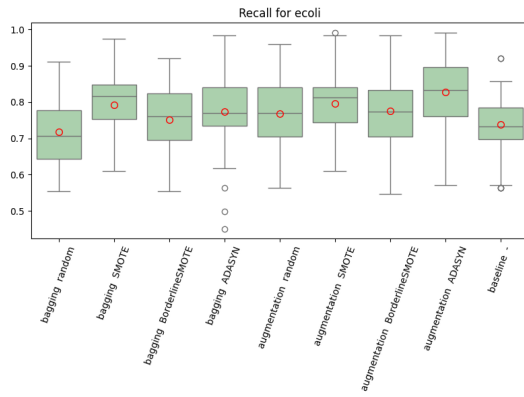
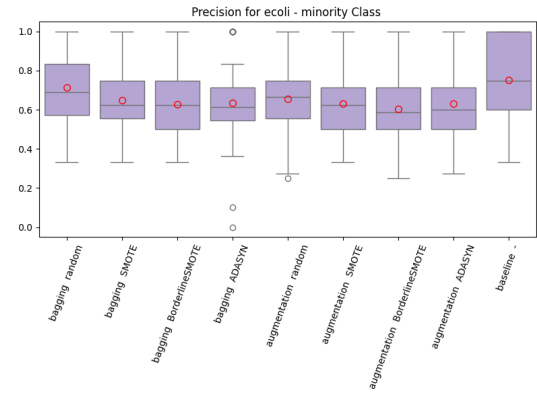
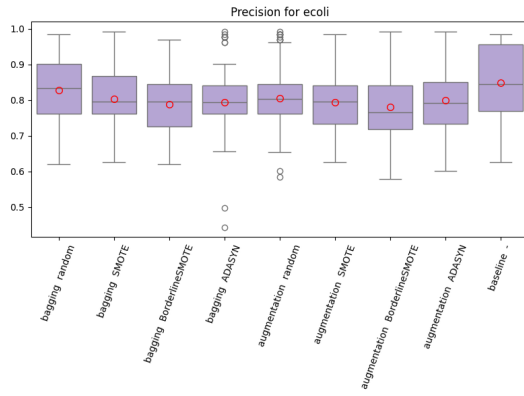
5.2 Zbiór Ecoli (niezbalansowanie: 8.6 : 1)

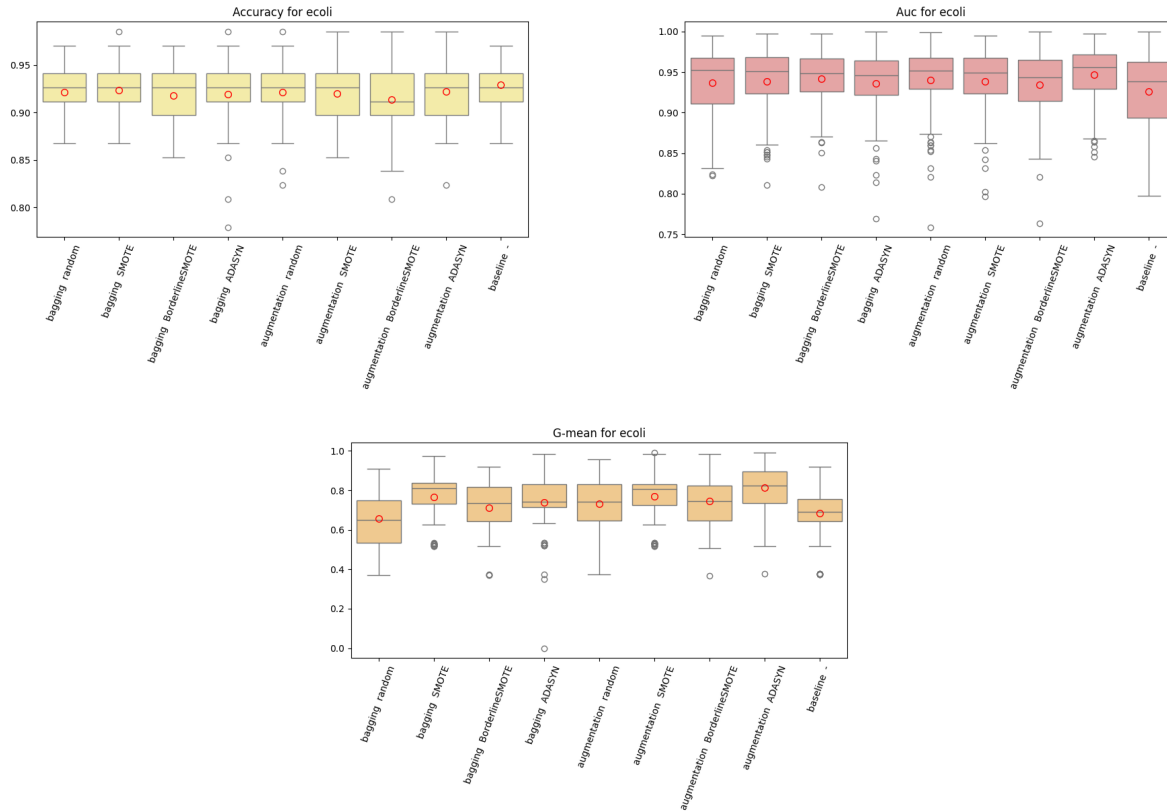


5.2.1 Wyniki

type	strategy	accuracy	auc	f1-score	g-mean	precision	recall
Minority class statistics							
bagging	random	-	-	0.5344	-	0.7151	0.4629
bagging	SMOTE	-	-	0.6239	-	0.6493	0.6271
bagging	BorderlineSMOTE	-	-	0.5670	-	0.6287	0.5429
bagging	ADASYN	-	-	0.5964	-	0.6340	0.5900
augmentation	random	-	-	0.5934	-	0.6566	0.5743
augmentation	SMOTE	-	-	0.6147	-	0.6298	0.6400
augmentation	BorderlineSMOTE	-	-	0.5864	-	0.6053	0.6014
augmentation	ADASYN	-	-	0.6517	-	0.6319	0.7100
baseline	-	-	-	0.5794	-	0.7516	0.4971
Average metrics							
bagging	random	0.9216	0.9369	0.7457	0.6572	0.8280	0.7186
bagging	SMOTE	0.9238	0.9382	0.7907	0.7664	0.8035	0.7925
bagging	BorderlineSMOTE	0.9178	0.9417	0.7607	0.7107	0.7887	0.7518
bagging	ADASYN	0.9193	0.9359	0.7757	0.7389	0.7937	0.7735
augmentation	random	0.9215	0.9403	0.7749	0.7321	0.8043	0.7678
augmentation	SMOTE	0.9203	0.9381	0.7850	0.7702	0.7944	0.7962
augmentation	BorderlineSMOTE	0.9134	0.9347	0.7689	0.7466	0.7799	0.7753
augmentation	ADASYN	0.9224	0.9466	0.8040	0.8127	0.7992	0.8284
baseline	-	0.9290	0.9262	0.7702	0.6856	0.8481	0.7378

Tabela 2: Porównanie średnich metryk dla zbioru Ecoli przy Sampling Rate = 0.5.





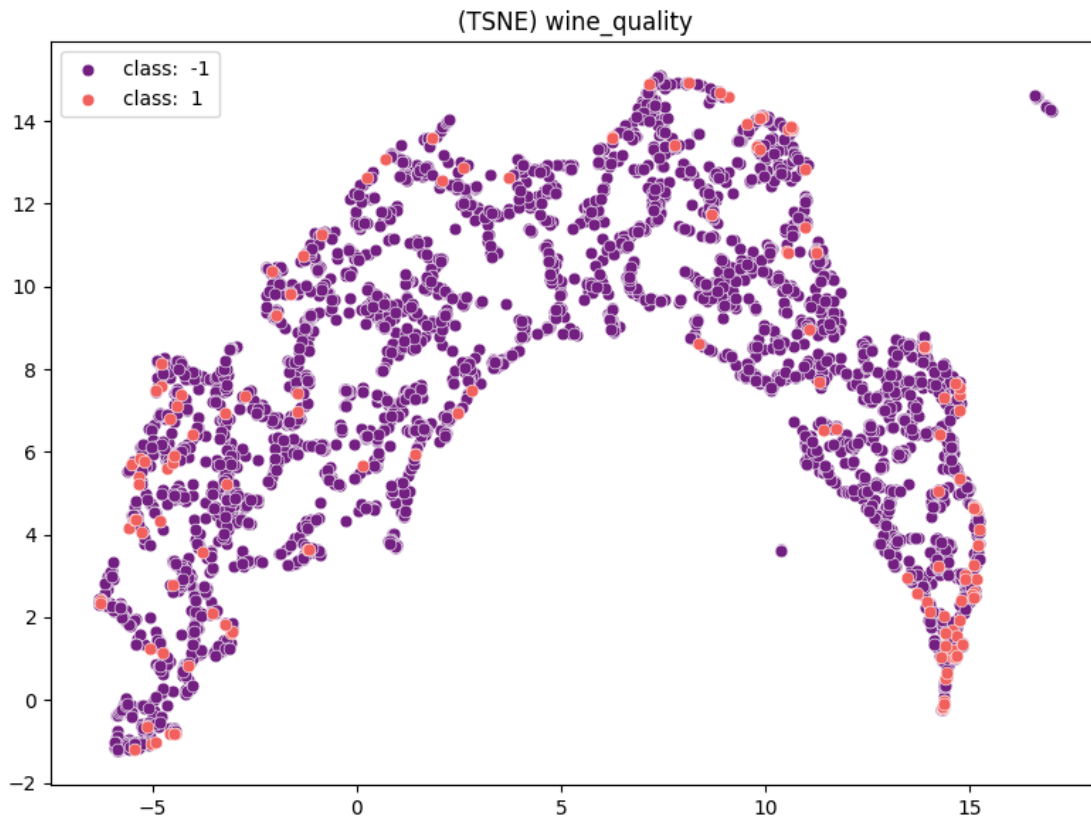
5.2.2 Wnioski

Wyniki pokazują, że w przypadku tego zbioru danych augmentacja przewyższa bagging pod względem skuteczności, szczególnie w kontekście klasy mniejszościowej. Najlepsze wyniki dla F1-score w tej klasie osiągnęła augmentacja z ADASYN, co jest znacząco lepsze niż jakkolwiek metoda baggingu. Widać też, że BorderlineSMOTE w baggingu wypadł słabiej niż klasyczny SMOTE czy ADASYN, co sugeruje, że w tym przypadku generowanie syntetycznych przykładów na granicy decyzyjnej nie przyniosło oczekiwanej poprawy.

Patrząc na średnie metryki dla całego zbioru, ponownie najlepsze wyniki osiągnęła augmentacja z ADASYN, co wskazuje na jej przewagę w poprawie czułości modelu. Bagging z kolei wykazał wyższą precyzję, ale niższy recall, co oznacza, że lepiej unika fałszywych trafień, ale gorzej wykrywa prawdziwe przypadki klasy mniejszościowej.

Dostosowanie Sampling Rate dla BorderlineSMOTE baggingu pozwoliło otrzymać nieznacznie gorsze wyniki większości metryk niż dla najlepszego rozwiązania - augmentacji ADASYN, natomiast istnieje podejrzenie, że inne strategie baggingowe są w stanie generować jeszcze lepsze wyniki po dostosowaniu Sampling Rate.

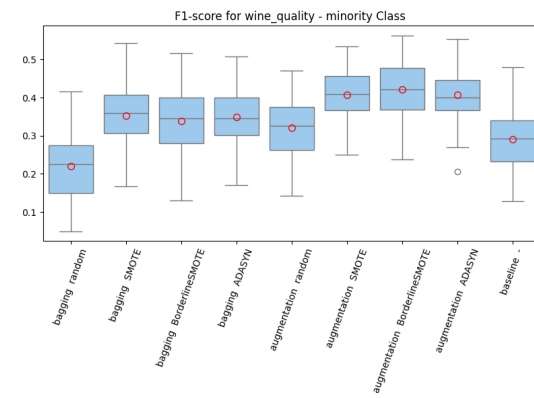
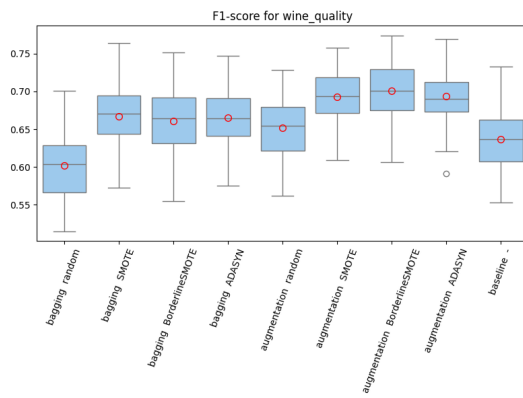
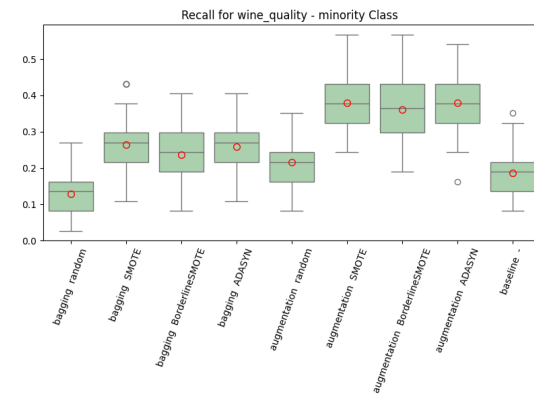
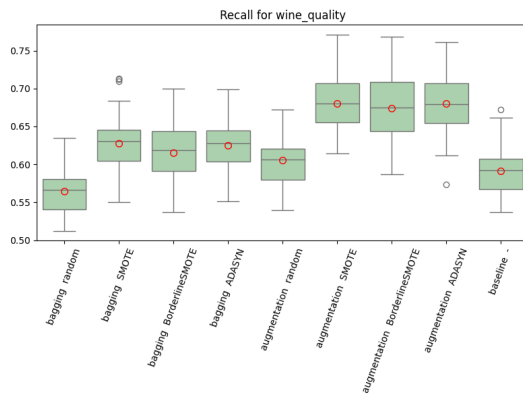
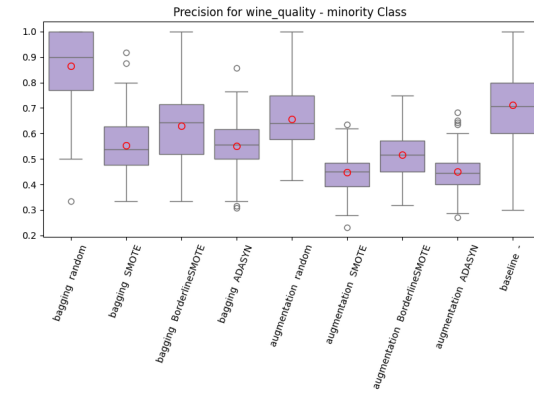
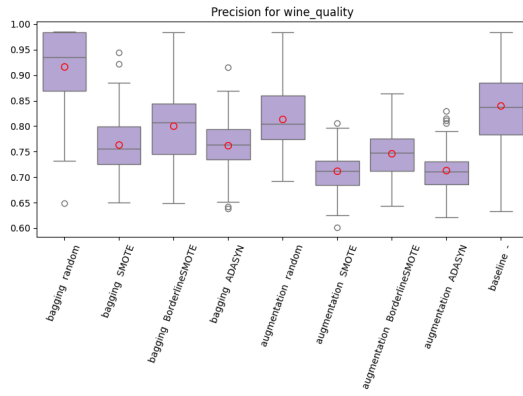
5.3 Zbiór Wine Quality (niezbalansowanie: 26:1)

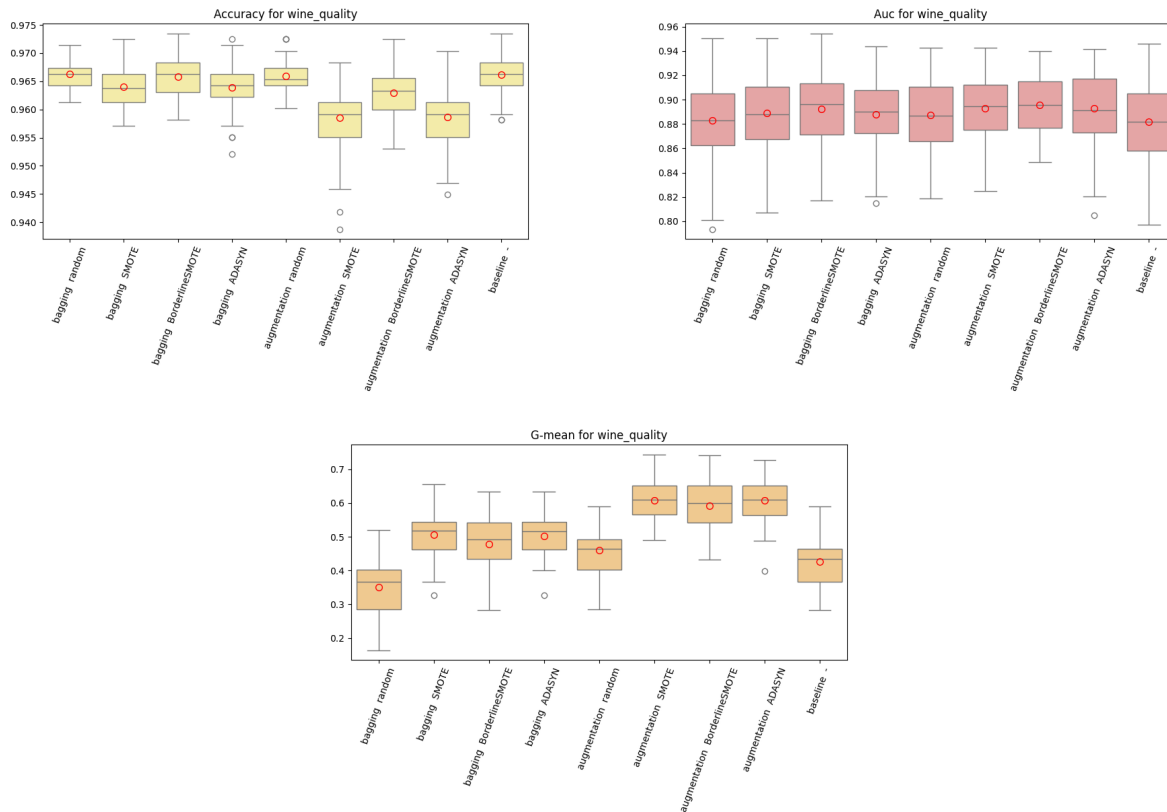


5.3.1 Wyniki

type	strategy	accuracy	auc	f1-score	g-mean	precision	recall
Minority class statistics (class 1)							
bagging	random	-	-	0.2208	-	0.8658	0.1292
bagging	SMOTE	-	-	0.3531	-	0.5542	0.2638
bagging	BorderlineSMOTE	-	-	0.3390	-	0.6299	0.2362
bagging	ADASYN	-	-	0.3487	-	0.5519	0.2584
augmentation	random	-	-	0.3211	-	0.6566	0.2165
augmentation	SMOTE	-	-	0.4078	-	0.4473	0.3795
augmentation	BorderlineSMOTE	-	-	0.4207	-	0.5174	0.3611
augmentation	ADASYN	-	-	0.4078	-	0.4500	0.3792
baseline	-	-	-	0.2902	-	0.7121	0.1862
Average metrics							
bagging	random	0.9663	0.8829	0.6018	0.3509	0.9164	0.5642
bagging	SMOTE	0.9640	0.8887	0.6673	0.5071	0.7630	0.6276
bagging	BorderlineSMOTE	0.9658	0.8925	0.6607	0.4792	0.8003	0.6153
bagging	ADASYN	0.9639	0.8880	0.6651	0.5026	0.7617	0.6250
augmentation	random	0.9659	0.8874	0.6518	0.4600	0.8133	0.6059
augmentation	SMOTE	0.9585	0.8926	0.6932	0.6078	0.7115	0.6803
augmentation	BorderlineSMOTE	0.9629	0.8954	0.7008	0.5927	0.7463	0.6738
augmentation	ADASYN	0.9587	0.8927	0.6932	0.6069	0.7129	0.6803
baseline	-	0.9661	0.8817	0.6364	0.4260	0.8405	0.5915

Tabela 3: Porównanie średnich metryk dla zbioru Wine Quality przy Sampling Rate = 0.5.





5.3.2 Wnioski

Wyniki pokazały różnice w skuteczności poszczególnych metod przetwarzania danych niezrównoważonych. W przypadku klasy mniejszościowej najwyższe wartości F1-score osiągnęły metody augmentacji, szczególnie BorderlineSMOTE oraz SMOTE i ADASYN, co świadczy o ich lepszej zdolności do balansowania precyzji i czułości w porównaniu z metodami baggingu. Warto zauważyć, że bagging z losowym oversamplingiem wypadł szczególnie słabo pod względem recall'u, co sugeruje, że ta strategia ma trudności z prawidłowym identyfikowaniem przykładów klasy mniejszościowej.

Jeśli chodzi o średnie metryki, augmentacja ponownie wykazała przewagę, osiągając wyższe wartości F1-score i G-mean w porównaniu do baggingu. Najlepsze wyniki w tej kategorii uzyskiwała augmentacja z BorderlineSMOTE, co potwierdza skuteczność tej metody w poprawie ogólnej jakości klasyfikacji. Bagging, mimo że osiągał wyższą precyzję, charakteryzował się znacząco niższym recall'em, co prowadziło do mniej zbalansowanych predykcji.

W kontekście implementacji BorderlineSMOTE w baggingu wyniki są nieco gorsze niż w przypadku augmentacji, co może wynikać z faktu, że generowanie syntetycznych przykładów na granicy decyzyjnej w połączeniu z bootstrapowaniem nie zawsze prowadzi do lepszej generalizacji. Warto zwrócić uwagę, że w augmentacji BorderlineSMOTE dał lepsze rezultaty niż w baggingu, co sugeruje, że sama metoda oversamplingu nie jest problemem, ale jej interakcja z baggingiem może wymagać dalszych optymalizacji.

Po dostosowaniu parametru Sampling Rate do wartości optymalnych, wyniki dla BorderlineSMOTE w baggingu ulegają znaczącej poprawie, osiągając precision na poziomie ok. 0.87 dla ogólnej klasyfikacji oraz ok. 0.7 dla klasy mniejszościowej, przy recallu wynoszącym odpowiednio ok. 0.62 i ok. 0.24. Te wartości wskazują, że po kalibracji metoda ta może stać się bardziej konkurencyjna, szczególnie w zakresie precyzji, która w tym przypadku przewyższa większość innych podejść. Jednakże, mimo tej poprawy, recall dla klasy mniejszościowej pozostaje stosunkowo niski, co sugeruje, że model nadal może mieć trudności z prawidłowym identyfikowaniem przykładów tej klasy w porównaniu do augmentacji, która w oryginalnych wynikach osiągała recall na poziomie 0.36–0.38.

W porównaniu do innych wariantów baggingu, takich jak SMOTE czy ADASYN, BorderlineSMOTE po dostrojeniu parametrów wypada lepiej pod względem precyzji, ale nadal ustępuje im w czułości. Można przypuszczać, że podobna optymalizacja Sampling Rate dla tych metod mogłaby przynieść jeszcze lepsze rezultaty, zwłaszcza że w oryginalnych wynikach SMOTE i ADASYN już osiągały wyższy recall dla klasy mniejszościowej. W szczególności ADASYN, który w podstawowej konfiguracji wypadł bardzo obiecująco, mógłby po dostosowaniu parametrów stać się liderem wśród metod baggingu.

6 Wnioski ogólne

Przeprowadzone eksperymenty na zbiorach danych o różnym stopniu niezbalansowania ujawniły pewne wzorce skuteczności analizowanych metod. W przypadku zbiorów o umiarkowanym niezbalansowaniu, zarówno augmentacja jak i bagging dawały zbliżone wyniki, przy czym augmentacja wykazywała nieznaczną przewagę w metrykach skupionych na klasie mniejszościowej, szczególnie w recallu. Jednak wraz ze wzrostem stopnia niezbalansowania, augmentacja zaczynała wyraźnie dominować, oferując lepszą równowagę między precyzją a czułością.

W kontekście potencjału rozwojowego, augmentacja wydaje się bardziej uniwersalnym rozwiązaniem, szczególnie w przypadkach skrajnego niezbalansowania danych. Jej zdolność do zachowania lepszej równowagi między metrykami, nawet przy wysokich współczynnikach niezbalansowania, czyni ją bardziej odporną na różne scenariusze danych. Bagging natomiast, choć w niektórych przypadkach osiąga wyższą precyzję, wykazuje większą wrażliwość na stopień niezbalansowania i wymaga dokładniejszego strojenia parametrów.

Porównując BorderlineSMOTE bagging z innymi podejściami, widać że metoda ta nie wyróżnia się wyraźnie na tle innych technik baggingu (jak SMOTE czy ADASYN), szczególnie w trudniejszych przypadkach. Jej główną zaletą jest możliwość osiągnięcia wysokiej precyzji po dostrojeniu parametrów, co może być wartościowe w specyficznych zastosowaniach. Jednak w ogólnym rozrachunku, w przypadku priorytetu dotyczącego wykrywalności klasy mniejszościowej, lepsze wyniki zapewniają techniki augmentacji, szczególnie z wykorzystaniem ADASYN lub klasycznego SMOTE.

7 Bibliografia

- [1] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*. In: *Advances in Intelligent Computing (Lecture Notes in Computer Science, vol. 3644)*, Springer, 2005, pp. 878–887.
- [2] Taghi M. Khoshgoftaar, Moiz Golawala, and Jason Van Hulse. *An Empirical Study of Learning from Imbalanced Data Using Random Forest*. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2007, pp. 310–317.
- [3] Chao Chen, Andy Liaw, and Leo Breiman. *Using Random Forest to Learn Imbalanced Data*. Technical Report, Department of Statistics, University of California, Berkeley, 2004.