Politechnika
Wrocławska

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Kierunek Teleinformatyka

Metody Sztucznej Inteligencji - Projekt

Implementacja algorytmu oversamplingu ADASYN

Autorzy:
Kałwa Weronika 263876
Postawa Sandra 263826
Zych Zuzanna 263882

Spis treści

1 Wstęp i przegląd literatury

1.1 Omówienie działania algorytmu oversamplingu ADASYN

Algorytm ADASYN (Adaptive Synthetic Sampling - Adaptacyjne Syntetyczne Próbkowanie) jest zaawansowaną techniką oversamplingu, mającą na celu zwiększenie dokładności klasyfikatorów w problemach związanych z niezbalansowanymi zbiorami danych. Główna idea algorytmu opiera się na generowaniu syntetycznych przykładów dla mniejszościowej klasy w taki sposób, aby równoważyć zbiór danych, skupiając się przede wszystkim na tych obszarach, gdzie granica decyzyjna między klasami jest niejasna. W porównaniu do innych metod, takich jak SMOTE, ADASYN dąży do bardziej zróżnicowanego i adaptacyjnego generowania próbek, poprawiając w ten sposób klasyfikację w trudniejszych obszarach [1].

1.2 Cel projektu

Celem projektu jest implementacja algorytmu ADASYN w kontekście balansowania zbioru danych w problemach klasyfikacji z niezrównoważonymi klasami. Projekt skupi się na zrozumieniu działania algorytmu, jego implementacji praktycznej oraz ocenie efektywności w porównaniu z innymi technikami oversamplingu.

1.3 Przegląd literatury

1.3.1 Omówienie istniejących metod

Oprócz ADASYN istnieją także inne metody oversampligu, takie jak:

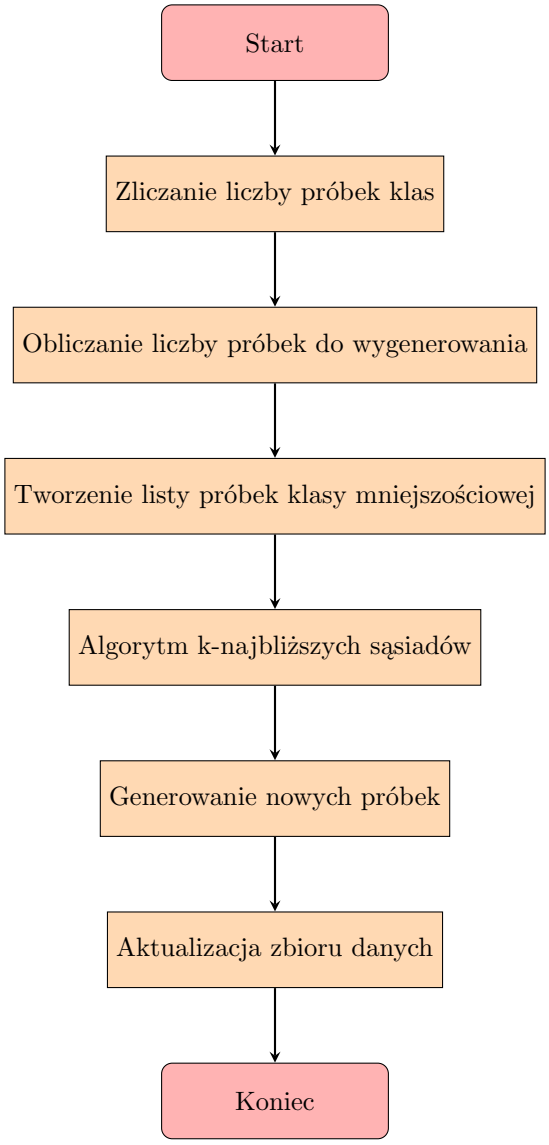
- SMOTE (Synthetic Minority Over-sampling Technique) [2]: Jest to technika statystyczna zwiększająca liczbę przypadków w zestawie danych w zrównoważony sposób. Składnik działa przez wygenerowanie nowych wystąpień z istniejących przypadków mniejszości, które są dostarczane jako dane wejściowe. Ta implementacja programu SMOTE nie zmienia liczby przypadków większościowych.
- BorderlineSMOTE [3]: Jest to modyfikacja metody SMOTE, ogranicza tworzenie nowych obiektów jedynie do granicy między przykładami z obydwu klas, co ma na celu zmniejszenie ryzyka przeuczenia.
- SVM-SMOTE [4]: Jest to metoda nadpróbkowania danych, która używa wektorów nośnych z SVM do generowania syntetycznych próbek klasy mniejszościowej przez interpolację między wektorami nośnymi a ich najbliższymi sąsiadami tej samej klasy, zwiększając liczbę danych na granicy decyzyjnej.
- BAGGING [5] polega na wykorzystaniu zbioru trenującego algorytmu klasyfikacji. Tworzony jest zbiór klasyfikatorów, z których każdy wykorzystuje algorytm i trenowany jest na zbiorze trenującym. Zbiór trenujący powstaje poprzez wylosowanie przykładów (ze zwracaniem) ze zbioru. Losowanie odbywa się zgodnie z rozkładem jednostajnym. Liczności zbioru. Każdy klasyfikator przydziela kategorię przykładowi. Ostateczna kategoria jest tą, która najczęściej była proponowana przez klasyfikatory
- NearMiss [6] to technika undersamplingu. Jej celem jest zrównoważenie rozkładu klas poprzez losowe eliminowanie przykładów z klasy większościowej. Gdy instancje dwóch różnych klas są bardzo blisko siebie, usuwamy instancje z klasy większościowej, aby zwiększyć odstępy między tymi klasami. Aby zapobiec problemowi utraty informacji w większości technik undersamplingu, szeroko stosuje się metody bliskich sąsiadów.
- K-Nearest Neighbors [7] tworzy wyimaginowaną granicę, aby sklasyfikować dane. Gdy do przewidywania dodawane są nowe punkty danych, algorytm dodaje ten punkt do najbliższego punktu granicy.

1.3.2 Metoda referencyjna

Głównym celem projektu jest zaimplementowanie algorytmu oversamplingu ADASYN za pomocą istniejących funkcji i implementacji. Do badania zostaną wykorzystane dane wygenerowane dane syntetyczne. W projekcie metodą, do której porównywane będą wyniki, będzie zaimportowany ADASYN, SMOTE oraz BorderlineSMOTE.

2 Metoda

Został utworzony własny estymator FutureAdasyn na podstawie algorytmu ADASYN, za pomocą metody nadpróbkowania klasy mniejszościowej w zbiorach danych z niezbalansowanymi klasami. Proces wygenerowania danych syntetycznych odbywa się poprzez wykorzystanie algorytmu k-najbliższych sąsiadów (k-NN) do generowania nowych próbek na podstawie istniejących próbek klasy mniejszościowej.



Rysunek 1: Schemat działania estymatora FutureAdasyn

3 Projekt eksperymentów:

Eksperymenty zakładają porównanie klasyfikacji zbiorów, dla wybranych metod. Początkowo opracowano własny algorytm, który został nazwany FutureADASYN, opierający się na algorytmie ADASYN. Następnie zarówno algorytm FutureAdasyn, jak i ADASYN został porównany z algorytmami SMOTE oraz BorderlineSMOTE. Na wszystkich algorytmach zostaną wykonane eksperymenty na danych syntetycznych oraz rzeczywistych.
Do wykonania eksperymentów zostaną użyte biblioteki: numpy, matplotlib, imblearn. Zostanie użyta technika walidacji krzyżowej do stabilizacji informacji o uzyskanych rezultatach.

Idea walidacji krzyżowej polega na podzieleniu dostępnych danych na kilka podzbiorów, zwanych częściami (ang. folds), a następnie przeprowadzeniu wielu iteracji treningu i walidacji modelu. W każdej iteracji jeden z podzbiorów zostaje wybrany jako zbiór walidacyjny, a pozostałe służą jako zbiór treningowy. Proces ten jest powtarzany wielokrotnie, aby każdy z podzbiorów został użyty przynajmniej raz jako zbiór walidacyjny. Ostatecznie, wyniki z poszczególnych iteracji są uśredniane, co daje ogólny wynik skuteczności modelu.

3.1 Metryki oceny wyników

Zostaną wykorzystane metryki oceny takie jak:

- Dokładność (Accuracy): mierzy procent poprawnie sklasyfikowanych próbek,
- Precyzja (Precision): określa stosunek poprawnie pozytywnie sklasyfikowanych przypadków do wszystkich pozytywnych przypadków,
- Czułość (Recall): określa stosunek poprawnie sklasyfikowanych pozytywnych przypadków do wszystkich rzeczywiście pozytywnych przypadków,
- F1-Score: średnia harmoniczna precyzji i czułości.

3.2 Testy statystyczne

Zostaną przeprowadzone testy statystyczne opierające się na teście t-Studenta, aby ocenić różnice między wynikami uzyskanymi dla różnych modeli lub różnych zestawów danych, oraz skuteczność różnych algorytmów.

3.3 Opis poszczególnych eksperymentów wraz z ich celami

Cel: Celem każdego eksperymentu jest zaobserwowanie działania algorytmów: ADASYN, FutureAdasyn, SMOTE, BorderlineSMOTE oraz porównania jego pracy w oparciu o algorytm ADASYN.
Opis: Wszystkie eksperymenty zostaną przeprowadzone na danych rzeczywistych oraz syntetycznych. A wyniki zostaną porównane za pomocą metryk: Accuracy, Precision, Recall, F1-Score.

3.4 Opis rzeczywistych zbiorów danych wraz z metodami ich pozyskania i/ lub sposobów generowania danych syntetycznych

3.4.1 Dane rzeczywiste:

Do testowania algorytmów została użyta rzeczywista baza danych *QSAR biodegradation* [8]. Jest to niezbalansowany zbiór danych zawierający wartości 41 atrybutów (deskryptorów molekularnych) zastosowanych do klasyfikacji 1055 substancji chemicznych na 2 klasy (gotowe i niełatwe do biodegradacji).

- Klasa 0: *gotowe do degradacji*, klasa mniejszościowa, 356 obiektów.
- Klasa 1: *niełatwe do degradacji*, klasa większościowa, 699 obiektów.

3.4.2 Dane syntetyczne:

Dane syntetyczne zostały wygenerowane za pomocą funkcji *make_classification* biblioteki *scikit-learn*. Zbiór niezbalansowany o wadze 0.1 dla klasy 0 oraz 0.9 dla klasy 1. Przykład kodu:

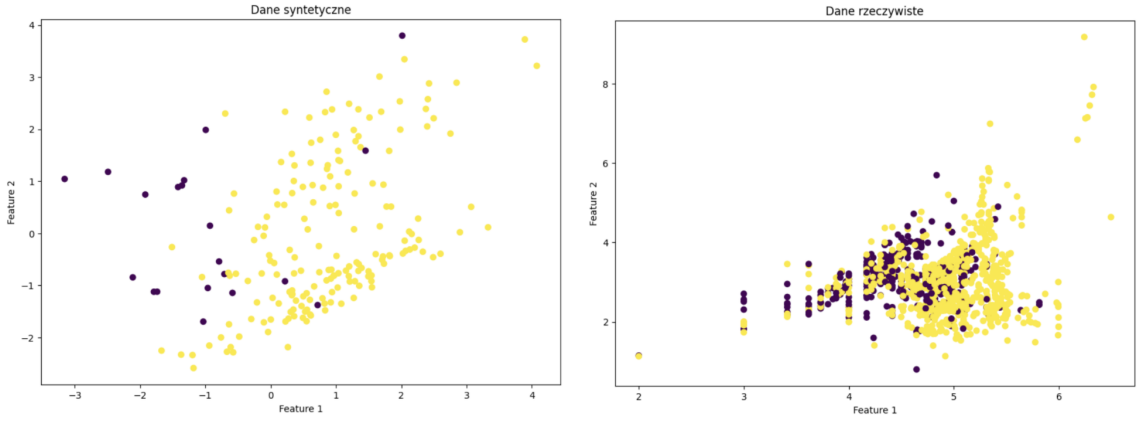
```
X, y = make_classification(n_samples=200,
                           n_features=2,
                           n_informative=2,
                           n_repeated=0,
                           n_redundant=0,
                           random_state=2137,
                           weights=[0.1, 0.9])
```

4 Wyniki eksperymentów

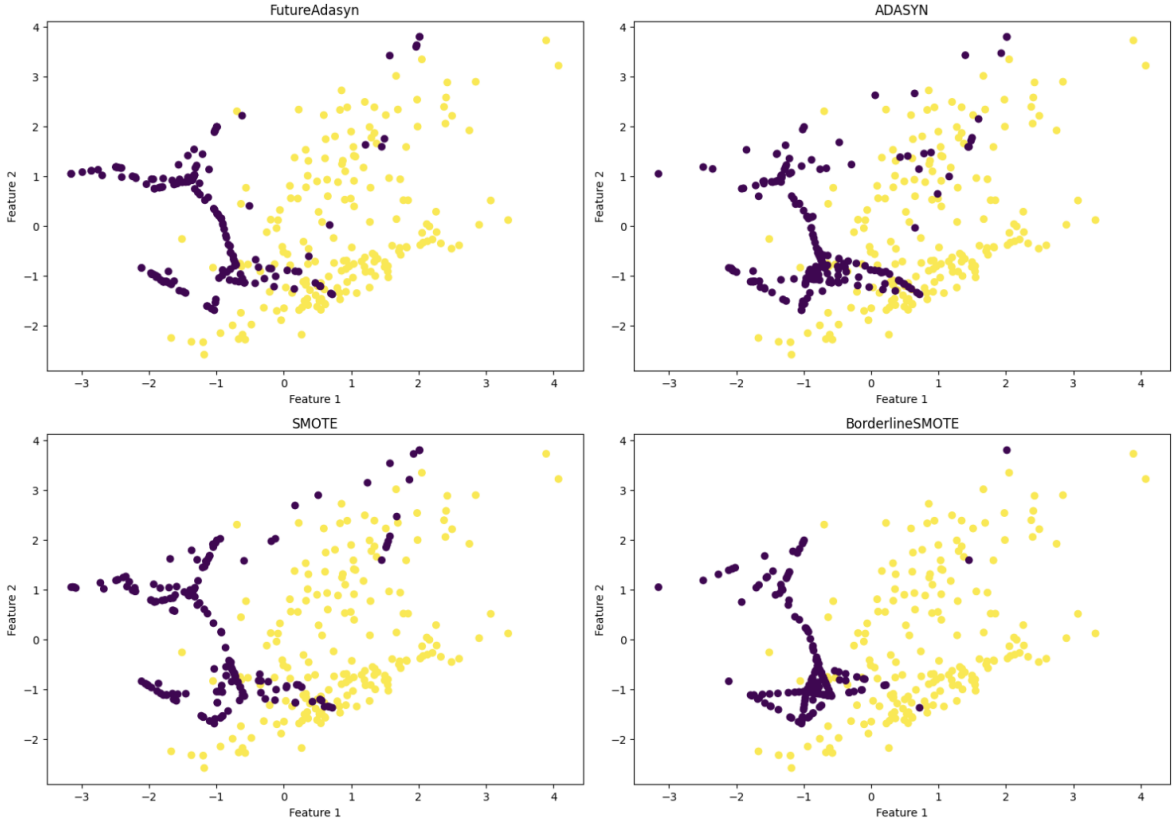
Na rysunku 2 zostały przedstawione dane syntetyczne oraz rzeczywiste przed zastosowaniem algorytmów oversamplingu. Można zauważyć, iż w obu przypadkach przeważa kolor żółty, co oznacza, że liczność klasy żółtej można nazwać klasą większościową ponieważ jest ona kilka razy większy niż klasa fioletowa, którą można natomiast nazwać klasą mniejszościową.

Na kolejnym rysunku (rys.3) widoczna jest zmiana w liczebności klasy mniejszościowej czyli fioletowej po zastosowaniu metod oversamplingu w przypadku danych syntetycznych.

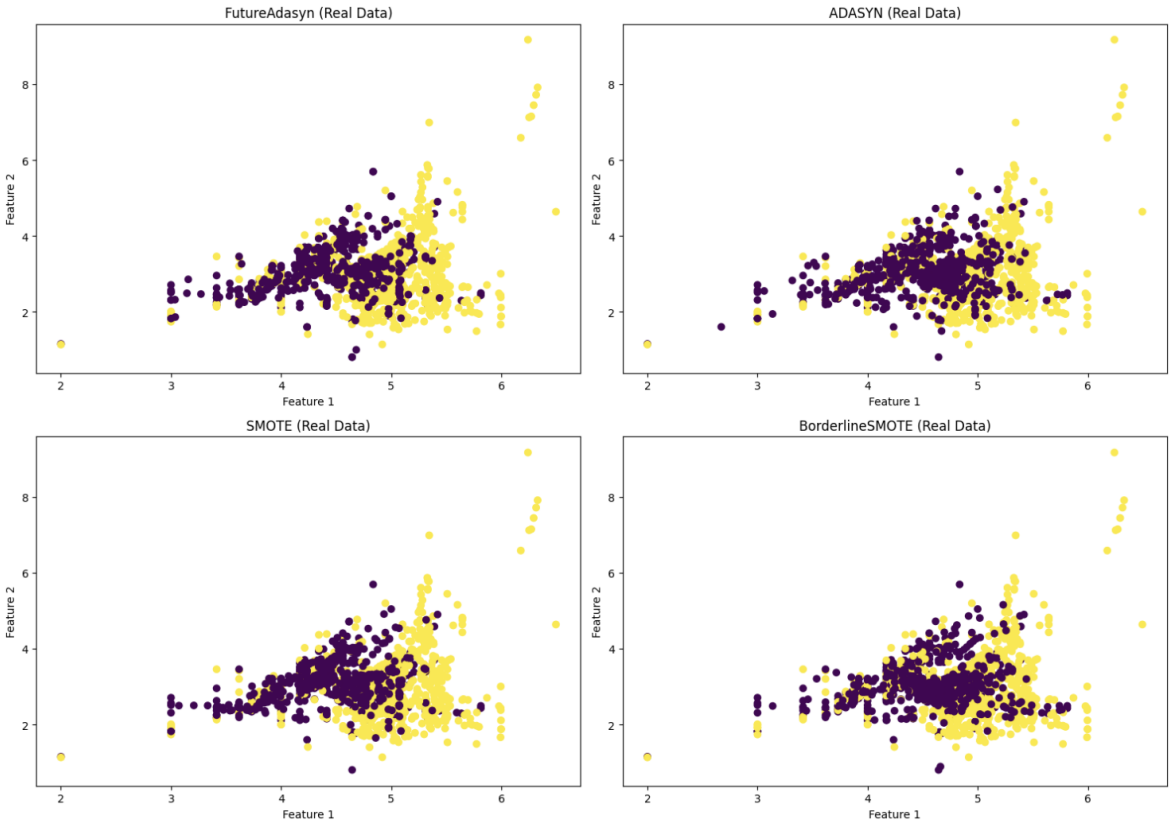
Podobna zmiana jest zauważalna w przypadku danych rzeczywistych (rys.4), ponieważ tutaj również klasa mniejszościowa, czyli fioletowa zwiększyła znacznie liczebność.



Rysunek 2: Dane syntetyczne i rzeczywiste przed oversamplingiem



Rysunek 3: Dane syntetyczne po oversamplingu



Rysunek 4: Dane rzeczywiste po oversamplingu

Tabela 1: Dane syntetyczne

Metryki	Dokladnosc	Precyzja	F1-score	Recall
Adasyn	0.8616 +/- 0.0689	0.882 +/- 0.097	0.8604 +/- 0.0651	0.8444 +/- 0.0572
Future Adasyn	0.8806 +/- 0.0347	0.8808 +/- 0.042	0.8803 +/- 0.0365	0.8833 +/- 0.0643
BorderlineSMOTE	0.9472 +/- 0.0297	0.9673 +/- 0.0434	0.9463 +/- 0.0297	0.9278 +/- 0.0377
SMOTE	0.8833 +/- 0.0379	0.8825 +/- 0.0241	0.8826 +/- 0.0399	0.8833 +/- 0.0567

Tabela 2: Dane rzeczywiste

Metryki	Dokladnosc	Precyzja	F1-score	Recall
Adasyn	0.8125 +/- 0.0324	0.9172 +/- 0.0392	0.7811 +/- 0.0404	0.681 +/- 0.0453
Future Adasyn	0.8355 +/- 0.0241	0.9092 +/- 0.0347	0.8183 +/- 0.0351	0.748 +/- 0.0633
BorderlineSMOTE	0.8226 +/- 0.0253	0.9524 +/- 0.0228	0.7919 +/- 0.0349	0.6795 +/- 0.0509
SMOTE	0.839 +/- 0.0305	0.9039 +/- 0.0346	0.8238 +/- 0.0422	0.7609 +/- 0.0695

Analizując dane z tabel zauważalne jest, iż najwyższą wartość dokładności w przypadku danych syntetycznych osiągnął BorderlineSMOTE, natomiast w przypadku danych rzeczywistych najwyższą wartość dokładności osiągnął Future Adasyn.

Tabela 3: Wyniki t-statistic oraz p-value dla $accuracy_{real}$ z wykorzystaniem danych rzeczywistych

Method	t-statistic	p-value
Future Adasyn	-1.1352	0.2892
BorderlineSMOTE	-0.4883	0.6384
SMOTE	-1.192	0.2674

Tabela 4: Wyniki t-statistic oraz p-value dla $precision_{real}$ z wykorzystaniem danych rzeczywistych

Method	t-statistic	p-value
Future Adasyn	0.3084	0.7656
BorderlineSMOTE	-1.5507	0.1596
SMOTE	0.5098	0.6239

Tabela 5: Wyniki t-statistic oraz p-value dla $f1_{real}$ z wykorzystaniem danych rzeczywistych

Method	t-statistic	p-value
Future Adasyn	-1.3929	0.2012
BorderlineSMOTE	-0.4068	0.6948
SMOTE	-1.4607	0.1822

Tabela 6: Wyniki t-statistic oraz p-value dla $recall_{real}$ z wykorzystaniem danych rzeczywistych

Method	t-statistic	p-value
Future Adasyn	-1.7232	0.1231
BorderlineSMOTE	0.0437	0.9662
SMOTE	-1.9265	0.0902

Tabela 7: Wyniki t-statistic oraz p-value dla $accuracy_{synthetic}$ z wykorzystaniem danych syntetycznych

Method	t-statistic	p-value
Future Adasyn	-0.4914	0.6363
BorderlineSMOTE	-2.2834	0.0518
SMOTE	-0.5528	0.5955

Tabela 8: Wyniki t-statistic oraz p-value dla $precision_{synthetic}$ z wykorzystaniem danych syntetycznych

Method	t-statistic	p-value
Future Adasyn	0.0211	0.9837
BorderlineSMOTE	-1.6058	0.147
SMOTE	-0.0117	0.9909

Tabela 9: Wyniki t-statistic oraz p-value dla $f1_{synthetic}$ z wykorzystaniem danych syntetycznych

Method	t-statistic	p-value
Future Adasyn	-0.5319	0.6092
BorderlineSMOTE	-2.3995	0.0432
SMOTE	-0.5813	0.577

Tabela 10: Wyniki t-statistic oraz p-value dla $recall_{synthetic}$ z wykorzystaniem danych syntetycznych

Method	t-statistic	p-value
Future Adasyn	-0.9037	0.3926
BorderlineSMOTE	-2.4333	0.041
SMOTE	-0.9661	0.3623

W tabelach 3-10 przedstawiono wyniki obliczonych wartości p-value, które uzyskano za pomocą wykonanego testu t-Studenta. Test odbył się po kolei dla każdej z metod, obliczał wartość p-value między dwiema metodami. Jedną niezmienną metodą do której zostały porównane był Adasyn. Dzięki wykonaniu tych testów jesteśmy w stanie określić czy odrzucić hipotezę zerową. Jeżeli p-value jest poniżej poziomu istotności, to przyjmujemy na korzyść hipotezę alternatywną.

Hipoteza zerowa zakłada, że średnie dwóch populacji są równe, co oznacza brak istotnych różnic między nimi. Z kolei hipoteza alternatywna sugeruje, że średnie tych dwóch populacji są na tyle różne, że można uznać te różnice za istotne statystycznie.

5 Wnioski

Spoglądając na ryskunki 2 oraz 3 możemy stwierdzić, iż zbiory danych rzeczywistych w porównianiu do syntetycznych znacznie się różnią od siebie. Jednakże patrząc oddzilenie na dane syntetyczne czy dane rzeczywiste to 4 przypadki wyglądają dość porównywalnie, posiadają podobne rozmieszczenie jak i podobna ilość próbek.

Analizując wyniki otrzymane w tabelkach 1 i 2 możemy zauważyć, że BorderlineSMOTE dominuje na danych syntetycznych, osiągając najwyższe wyniki we wszystkich metrykach. SMOTE wykazuje się najbardziej zrównoważonymi wynikami na danych rzeczywistych, co czyni go najbardziej wszechstronnym modelem. Future Adasyn prezentuje dobre, stabilne wyniki na obu zestawach danych, oferując kompromis między precyzją a Recall. Adasyn osiąga najniższe wartości Recall na danych rzeczywistych, co sugeruje, że może nie być najlepszym wyborem dla bardziej złożonych danych.

W przypadku danych rzeczywistych żadna z metod nie wykazuje statystycznie istotnych różnic w żadnej z analizowanych metryk, co może świadczyć, że ich wyniki są porównywalne do referencyjnej metody. Natomiast w przypadku danych syntetycznych dotyczących dokładności BorderlineSMOTE jest bliski granicy istotności z p-value = 0.0518, w F-1 score BorderSMOTE wykazuje statystycznie istotną różnicę p-value = 0.043 oraz w Recall wykazuje statystycznie istotną różnicę p-value = 0.041. Wszystkie te wyniki mogą wskazywaćna to, że może być znacząco lepszy od referencyjnej metody. Natomiast pozostałe metody takie jak Future Adasyn i SMOTE nie wykazują statystycznie istotnych różnic w żadnej z metryk, co sugeruje, że ich wyniki są porównywalne do referencyjnej metody.

Podsumowując możemy dojść do wniosku, iż w przypadku danych rzeczywistych żadna z metod nie pokazała przewagi nad innymi, jednakże w przypadku danych syntetycznych już widzimy ta różmnicę, ponieważ metoda BorderlineSMOTE jest szczególnie skuteczna na tle innych metod.

6 Bibliografia

Literatura

[1] https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa [maj 2024]

[2] https://learn.microsoft.com/pl-pl/azure/machine-learning/component-reference/smote?view=azureml-api-2

[3] https://pb.edu.pl/oficyna-wydawnicza/wp-content/uploads/sites/4/2021/12/Modelowanie-i-optymalizacja-1.pdf [maj 2024]

[4] https://www.blog.trainindata.com/oversampling-techniques-for-imbalanced-data/ [maj 2024]

[5] https://repo.pw.edu.pl/docstore/download/WEiTI-ab4d82b3-a859-462f-a20b-2823dec969b1/pandrusz_Metauczenie+a+mo%C5%BCliwo%C5%9B%C4%87+poprawy+skuteczno%C5%9Bci+klasyfikacji.pdf [marzec 2024]

[6] https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/ [marzec 2024]

[7] https://www.geeksforgeeks.org/regression-using-k-nearest-neighbors-in-r-programming/?ref=lbp [maj 2024]

[8] https://archive.ics.uci.edu/dataset/254/qsar+biodegradation [kwiecień 2024]