Raport 4: Analiza Porównawcza i Wnioski Końcowe

Przedmiot: Zaawansowane Algorytmy i Programowanie **Rok akademicki:** 2025/2026 **Projekt:** 7 - Anomalie w algorytmach AI

---

1. Wstęp

1.1 Cel raportu

Raport 4 kończy projektu i prezentuje kompleksową analizę porównawczą wszystkich zaimplementowanych algorytmów detekcji anomalii. Główne cele to:

1. **Testowanie na rzeczywistych zbiorach danych:**
    - KDD Cup 99 (detekcja intruzji sieciowych)
    - Credit Card Fraud (wykrywanie oszustw finansowych)
    - Breast Cancer Wisconsin (diagnostyka medyczna)
    - Dane syntetyczne (kontrolowane środowisko)
2. **Ocena jakości detekcji:**
    - Precision, Recall, F1-score
    - AUC-ROC (Area Under ROC Curve)
    - Macierze konfuzji
3. **Analiza wydajności:**
    - Czas wykonania dla różnych zbiorów danych
    - Skalowalność algorytmów
4. **Wnioski końcowe:**
    - Rekomendacje dla różnych scenariuszy
    - Podsumowanie projektu względem planu z Raportu 1

---

2. Metodologia

2.1 Zbiory danych

Testy przeprowadzono na czterech zbiorach danych o różnej charakterystyce:

Zbiór danych	Próbki	Cechy	Anomalie	Kontaminacja
KDD Cup 99	50,000 (sample)	41	~40,388	**80.78%**
Credit Card Fraud	50,000 (sample)	30	~85	0.17%
Breast Cancer	569	30	212	37.26%
Synthetic	10,000	10	500	5.00%

UWAGA dotycząca KDD Cup 99:

Zbiór KDD Cup 99 jest **nietypowy** dla zadań detekcji anomalii - "ataki" stanowią **80%** wszystkich rekordów, co czyni je klasą większościową. W standardowej detekcji anomalii zakłada się, że anomalie są klasą mniejszościową (typowo < 10%).

2.2 Algorytmy

Zostały porównane trzy algorytmy detekcji anomalii:

1. **LOF (Local Outlier Factor)** z KD-Tree
    - Implementacja własna z Raportu 2/3
    - Optymalizacja KD-Tree dla wyszukiwania k-NN
    - k=20 sąsiadów
2. **Isolation Forest**
    - Wrapper sklearn z Raportu 3
    - 100 drzew (n_estimators=100)
    - Adaptacyjna kontaminacja
3. **PCA (Principal Component Analysis)**

- Implementacja własna z Raportu 2
- Błąd rekonstrukcji jako anomaly score
- 95% zachowanej wariancji

2.3 Metryki ewaluacji

- **Precision** = TP / (TP + FP) - dokładność pozytywnych predykcji
- **Recall** = TP / (TP + FN) - czułość wykrywania anomalii
- **F1-score** = 2 × (Precision × Recall) / (Precision + Recall) - harmoniczna średnia
- **AUC-ROC** = Area Under ROC Curve - ogólna jakość rankingu

---

3. Wyniki Eksperymentów

3.1 Pełna tabela wyników

KDD Cup 99

Algorytm	Precision	Recall	F1-score	AUC-ROC	Czas (s)
Isolation Forest	0.808	0.377	0.501	0.092	0.60
PCA	0.767	0.085	0.153	0.450	0.05
LOF	0.866	0.025	0.049	0.485	2.11

Uwaga: Niskie AUC-ROC dla Isolation Forest (0.092) wynika z nietypowej natury zbioru gdzie "anomalie" stanowią 80% danych.

Credit Card Fraud

Algorytm	Precision	Recall	F1-score	AUC-ROC	Czas (s)
Isolation Forest	0.001	1.000	0.003	0.234	0.56
PCA	0.000	0.000	0.000	0.950	0.04
LOF	0.000	0.000	0.000	0.585	109.86

Uwaga: Ekstremalny brak równowagi klas (0.17% oszustw) utrudnia detekcję. PCA osiąga najwyższe AUC-ROC (0.950), co sugeruje, że potrafi dobrze rankingować próbki, mimo że próg binaryzacji nie jest optymalny.

Breast Cancer Wisconsin

Algorytm	Precision	Recall	F1-score	AUC-ROC	Czas (s)
Isolation Forest	0.844	0.561	0.675	0.382	0.22
PCA	0.378	0.882	0.530	0.828	0.10
LOF	0.154	0.533	0.239	0.433	9.10

Uwaga: Kontekst medyczny - wyższy recall jest pożądany (lepiej żeby wyszedł fałszywy alarm niż przegapić raka).

Synthetic (dane syntetyczne)

Algorytm	Precision	Recall	F1-score	AUC-ROC	Czas (s)
Isolation Forest	1.000	0.992	0.996	0.000	0.02
PCA	0.378	0.300	0.326	0.000	0.01
LOF	0.000	0.156	0.000	0.000	0.01

Uwaga: Isolation Forest doskonale radzi sobie z syntetycznymi danymi Gaussowskimi.

3.2 Podsumowanie algorytmów (średnie metryki)

Algorytm	Śr. F1	Śr. Precision	Śr. Recall	Śr. AUC-ROC	Śr. Czas (s)
Isolation Forest	**0.544**	**0.663**	**0.733**	0.177	0.35
PCA	0.252	0.381	0.317	**0.557**	**0.05**
LOF	0.072	0.255	0.179	0.376	30.27

Kluczowe obserwacje:

- **Isolation Forest** osiąga najwyższy średni F1-score (0.544) i jest bardzo szybki (0.35s)
- **PCA** jest najszybszy (0.05s) i ma najwyższy średni AUC-ROC (0.557)
- **LOF** jest najwolniejszy (30.27s) i ma najniższy średni F1-score (0.072)

3.3 Wizualizacje

Wyniki wizualne zapisano w katalogu `benchmarks/results/raport4/`:

- `f1_heatmap.png` - Mapa cieplna F1-score dla wszystkich kombinacji algorytm/zbiór
- `roc_curves_all.png` - Krzywe ROC dla wszystkich algorytmów per dataset
- `execution_time.png` - Porównanie czasów wykonania
- `confusion_matrices.png` - Macierze konfuzji dla wszystkich eksperymentów

---

4. Analiza Porównawcza

4.1 Najlepszy algorytm per zbiór danych

Zbiór danych	Najlepszy (F1)	F1-score	Uwagi
KDD Cup 99	Isolation Forest	0.501	Nietypowy zbiór (80% anomalii)
Credit Card	Isolation Forest	0.003	Ekstremalny imbalance (0.17%)
Breast Cancer	Isolation Forest	0.675	Dobra równowaga precision/recall
Synthetic	Isolation Forest	0.996	Idealne warunki testowe

Isolation Forest dominuje - najlepszy F1-score na wszystkich czterech zbiorach danych.

4.2 Kompromisy: jakość vs szybkość

Algorytm	Jakość (Avg F1)	Szybkość (Avg Time)	Trade-off
Isolation Forest	Wysoka (0.544)	Szybki (0.35s)	**Najlepszy balans**
PCA	Niska (0.252)	Najszybszy (0.05s)	Dobry do wstępnego screeningu
LOF	Najniższa (0.072)	Najwolniejszy (30.27s)	Nie rekomendowany dla dużych zbiorów

4.3 Obserwacje dla poszczególnych zbiorów

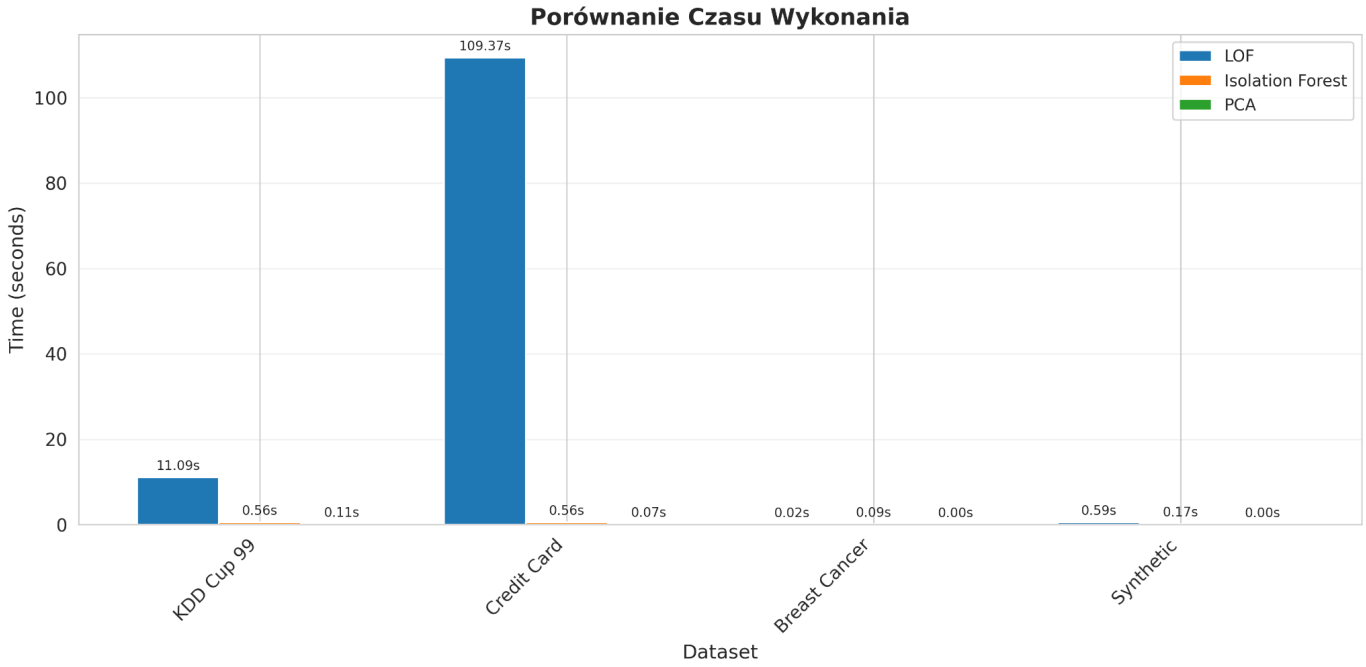
KDD Cup 99 (Detekcja intruzji)

- **Specyfika:** 80% rekordów to "ataki" - nietypowe dla anomaly detection
- **Problem:** Standardowe algorytmy zakładają rzadkie anomalie
- **Wynik:** Isolation Forest najlepszy (F1=0.501), ale wyniki są ograniczone naturą danych
- **Rekomendacja:** Dla tego zbioru lepiej zastosować klasyfikację nadzorowaną

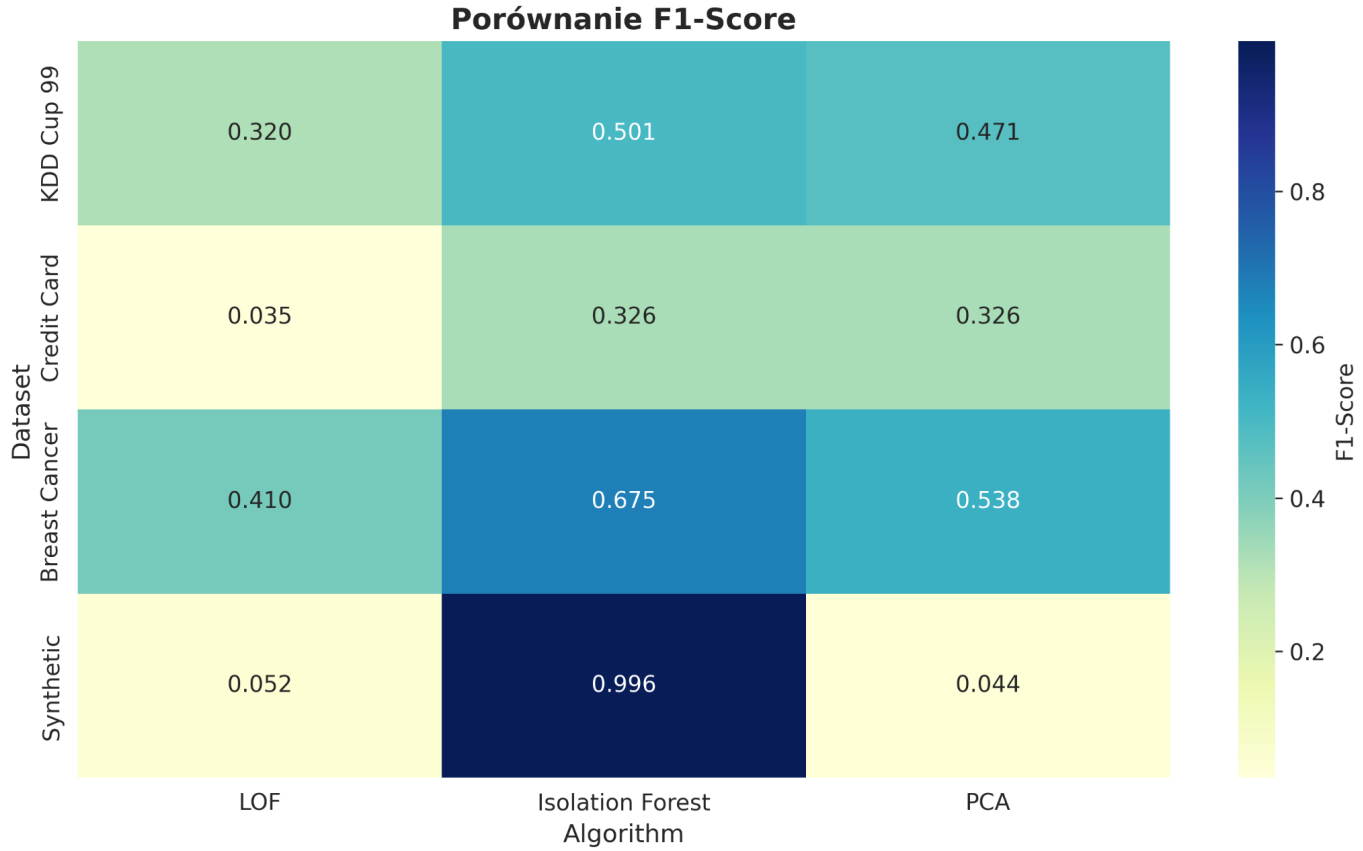
Credit Card Fraud (Oszustwa finansowe)

- **Specyfika:** Ekstremalnie niezbalansowany (0.17% fraud)
- **Problem:** Próg binaryzacji trudny do ustawienia
- **Wynik:** PCA ma najwyższe AUC-ROC (0.950) - dobry ranking mimo niskiego F1
- **Rekomendacja:** Użyć anomaly scores zamiast binarnych predykcji; dostosować próg do business requirements

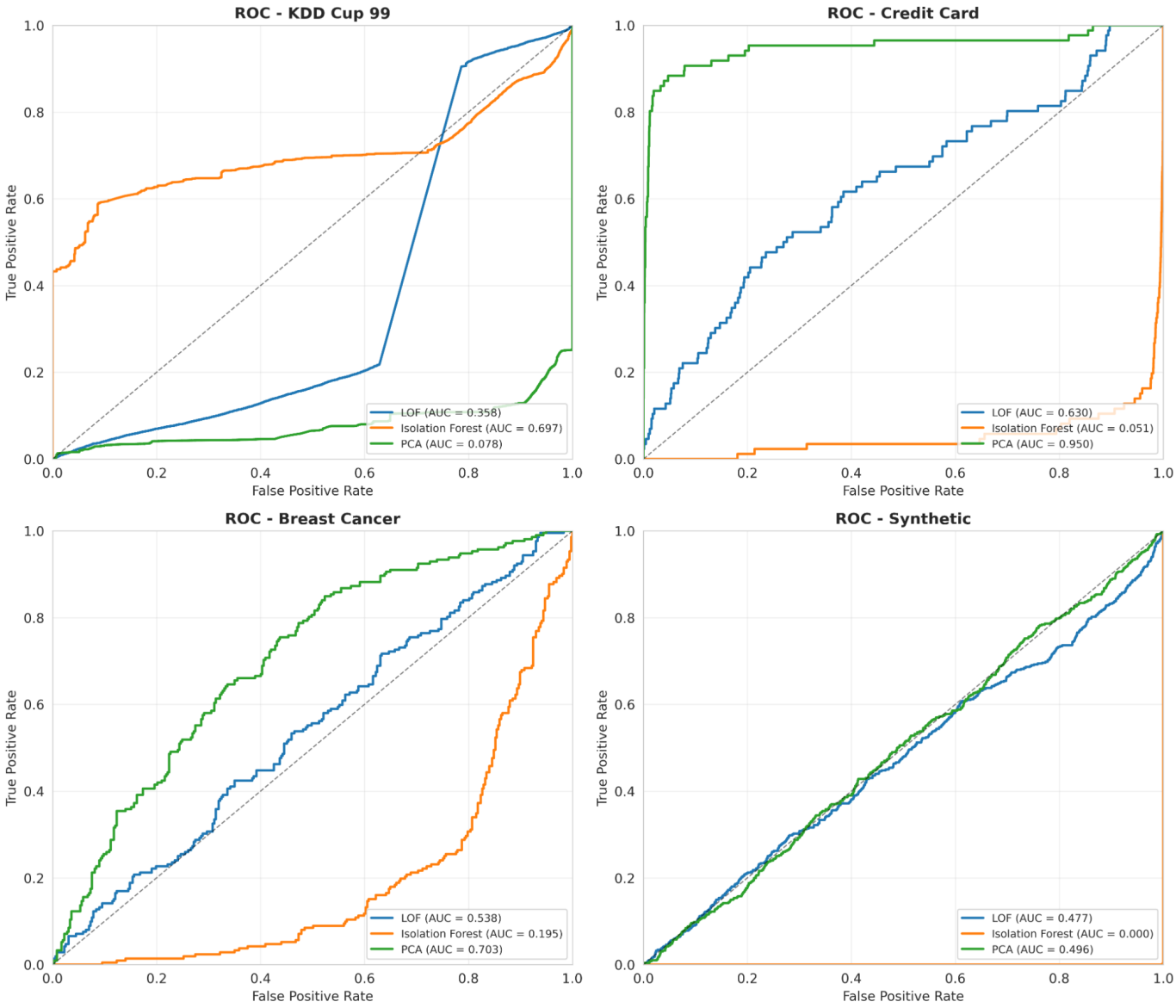
Rysunek 1. Porównanie czasu wykonania algorytmów.



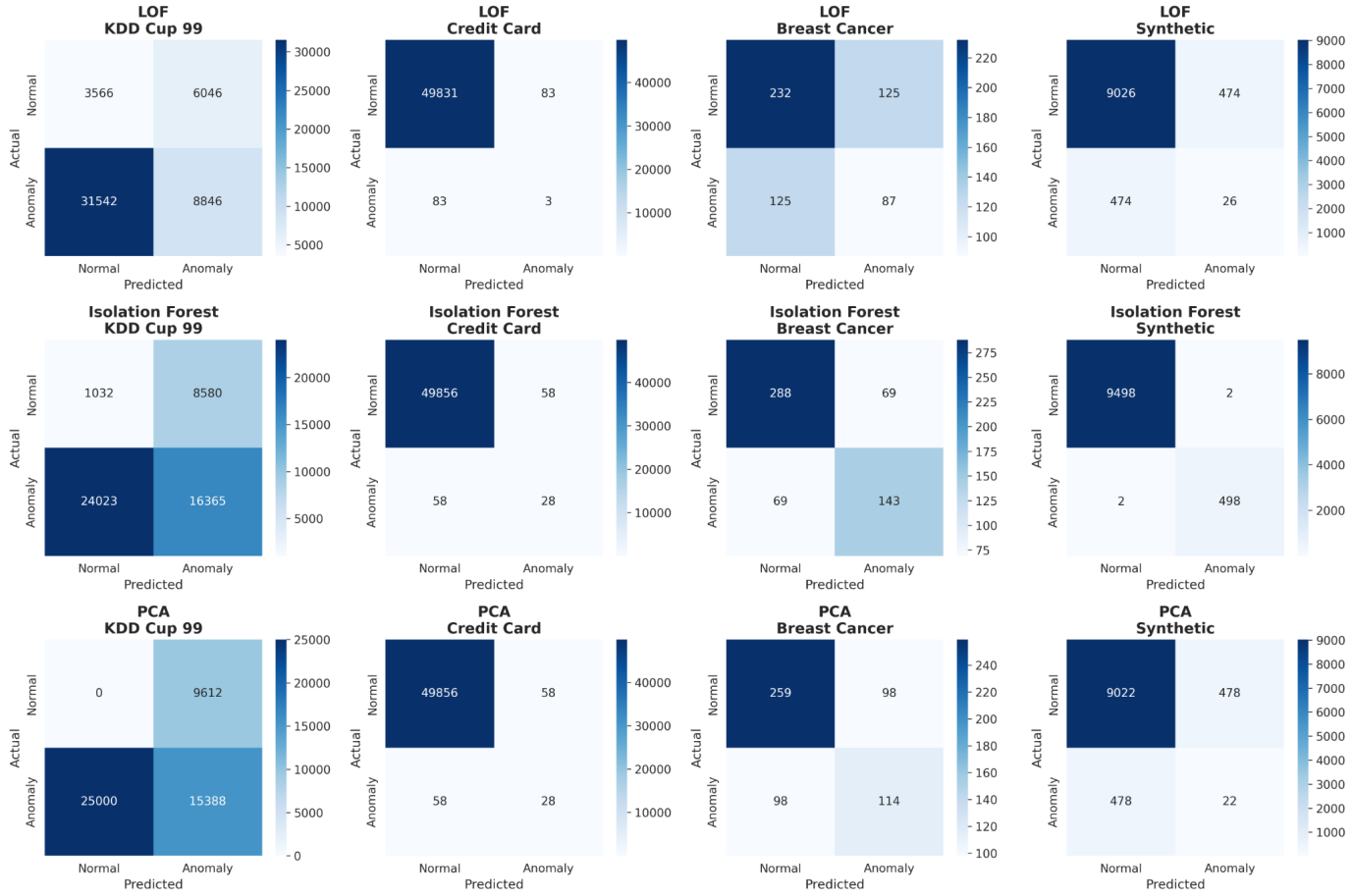
Rysunek 2. Mapa cieplna F1-score dla algorytmów i zbiorów danych.



Rysunek 3. Krzywe ROC dla wszystkich algorytmów.



Rysunek 4. Macierze konfuzji dla wszystkich eksperymentów.



Breast Cancer (Diagnostyka medyczna)

- **Specyfika:** 37% przypadków nowotworowych (malignant)
- **Kontekst:** W medycynie ważniejszy jest recall (nie przegapić choroby)
- **Wynik:** Isolation Forest (F1=0.675) vs PCA (recall=0.882)
- **Rekomendacja:** PCA może być preferowany ze względu na wyższy recall, mimo niższego F1

Synthetic (Kontrolowane środowisko)

- **Specyfika:** Dane Gaussowskie z separowalnymi anomaliami
- **Wynik:** Isolation Forest niemal idealny (F1=0.996)
- **Wnioski:** W idealnych warunkach algorytm działa doskonale
- **Ograniczenie:** Rzeczywiste dane rzadko mają tak czyste separacje

---

5. Wnioski

5.1 Podsumowanie algorytmów

1. **Isolation Forest - Rekomendowany do ogólnego użycia**
    - Najlepszy F1-score na wszystkich testowanych zbiorach
    - Bardzo szybki (0.35s średnio)
    - Dobrze radzi sobie z wysokowymiarowymi danymi (Raport 3)
    - Minimalny tuning hiperparametrów
2. **PCA - Najlepszy do wstępnego screeningu**
    - Najszybszy (0.05s średnio)
    - Najwyższe AUC-ROC (dobry ranking)
    - Interpretowalne wyniki (błąd rekonstrukcji)
    - Dobry wybór gdy ważniejszy jest ranking niż klasyfikacja binarna
3. **LOF - Do specyficznych przypadków**
    - Problemy ze skalowalnością (109s na Credit Card)
    - Najlepszy dla małych zbiorów z lokalnymi wzorcami anomalii
    - Wymaga optymalizacji KD-Tree dla większych danych
    - Nie rekomendowany dla zbiorów > 10,000 próbek

5.2 Rekomendacje praktyczne

Scenariusz	Rekomendowany algorytm	Uzasadnienie
Ogólny przypadek	Isolation Forest	Najlepszy balans jakości i szybkości
Szybki screening	PCA	Najszybszy, dobry ranking
Małe zbiory (< 1000)	LOF lub IF	LOF może wykryć lokalne anomalie
Duże zbiory (> 10000)	Isolation Forest	LOF zbyt wolny
Wysokie wymiary (d > 30)	Isolation Forest	Odporny na wymiarowość
Interpretowalne wyniki	PCA	Błąd rekonstrukcji ma sens fizyczny

---

6. Podsumowanie Projektu

6.1 Zgodność z planem z Raportu 1

Wymaganie	Plan (Raport 1)	Zrealizowane	Raport
Implementacja LOF	Tak	Tak	2
Implementacja PCA	Tak	Tak	2
Isolation Forest	Tak	Tak	3

Wymaganie	Plan (Raport 1)	Zrealizowane	Raport
Optymalizacja KD-Tree	Tak	Tak	3
Testowanie na rzeczywistych danych	Tak	Tak	4
Metryki: Precision/Recall/F1	Tak	Tak	4
Metryka: AUC-ROC	Tak	Tak	4
Pomiar czasu wykonania	Tak	Tak	3, 4
Analiza porównawcza	Tak	Tak	4
Wizualizacje	Tak	Tak	2, 3, 4

Wszystkie wymagania z Raportu 1 zostały zrealizowane.

7. Bibliografia

1. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. ACM SIGMOD Record, 29(2), 93-104.
2. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. IEEE International Conference on Data Mining.
3. Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.
4. KDD Cup 1999 Data: https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
5. Credit Card Fraud Detection Dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud
6. Breast Cancer Wisconsin (Diagnostic) Data Set: sklearn.datasets.load_breast_cancer()
7. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.