

Ja robie dla klasyfikacji, można podobne dodac do klasteryzacji

Klasyfikacja

✅ Co dokładnie masz zrobić – krok po kroku:

◆ Krok 1: Dane

- Pobierz dane – najlepiej UCI Heart Disease Dataset
→ Zawiera: wiek, ciśnienie, cholesterol, EKG itd.

◆ Krok 2: Preprocessing

- Uzupełnij brakujące dane (np. medianą)
- Przeskaluj dane (standaryzacja)
- Zakoduj zmienne katagoryczne (np. one-hot encoding)

◆ Krok 3: Zbuduj kilka modeli klasyfikacyjnych

Zrób porównanie co najmniej 3 metod:

Model	Bias ↓	Wariancja ↓
Boosting (XGBoost)	✅	❌ może być wysoka
Random Forest	❌	✅
Drzewo decyzyjne	❌	❌

Opcjonalnie: dodaj np. SVM, sieć neuronową, Rotation Forest (jak w PDF-ie).

◆ Krok 4: Trening i walidacja

- Użyj np. k-fold cross-validation (k=5)
- Zbieraj metryki:



◆ Krok 5: Optymalizacja i analiza

- Porównaj modele:
 - Który ma niski bias? (czyli wysoką trafność na zbiorze testowym)
 - Który ma niską wariancję? (czyli niskie różnice między foldami)
- Wskaż kompromis: np. „Boosting osiąga najlepsze F1, ale RF ma niższą wariancję”.

◆ Krok 6: Wyniki i wnioski

- Zrób tabelkę porównawczą
- Zrób wykresy (np. słupkowe metryk, krzywe ROC)
- Zinterpretuj: które metody najlepiej nadają się do diagnozy chorób serca i dlaczego

1. Accuracy (dokładność)

To po prostu:

ile % wszystkich przykładów zostało poprawnie zaklasyfikowanych?

Wzór:


ini

 Kopiuuj

 Edytuj

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- TP – True Positives (model dobrze przewidział chorych)
- TN – True Negatives (model dobrze przewidział zdrowych)
- FP – False Positives (model błędnie uznał zdrowych za chorych)
- FN – False Negatives (model nie wykrył choroby)

 **Uwaga:** jeśli dane są niezbalansowane (np. 90% zdrowych, 10% chorych), accuracy może być mylące.

2. F1-score

To średnia harmoniczna precyzji i czułości (recall).

F1-score mówi:

jak dobrze model radzi sobie z wykrywaniem chorych pacjentów, uwzględniając zarówno błędy fałszywie dodatnie, jak i fałszywie ujemne.

Wzór:


ini

 Kopiuuj

 Edytuj

$$F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

- **Precision** = $TP / (TP + FP)$ – z tych, których uznał za chorych, ile naprawdę było chorych?
- **Recall** = $TP / (TP + FN)$ – z prawdziwie chorych, ile znalazł?

 **F1-score jest lepszy od Accuracy, gdy masz niezbalansowane klasy.**

3. AUC-ROC

ROC (Receiver Operating Characteristic) to krzywa pokazująca zależność między:

- True Positive Rate (czułość)
- False Positive Rate

AUC (Area Under Curve) to:

obszar pod krzywą ROC – im bliżej 1, tym lepiej.

- AUC bliskie 0.5 = losowy klasyfikator
- AUC bliskie 1 = super klasyfikator

🚀 Świetna metryka, bo jest niezależna od progu decyzyjnego!

4. Odchylenie standardowe (wariancja wyników)

Jeśli robisz **k-fold cross-validation**, to trenujesz model kilka razy (np. 5 razy na różnych podziałach danych).

Dla każdej iteracji masz np. inną wartość accuracy. Wtedy:

Odchylenie standardowe pokazuje, jak bardzo wyniki „skaczą” między foldami.

- Niskie SD → model stabilny (niskie ryzyko przeuczenia)
- Wysokie SD → model niestabilny (duża wariancja)

Do tego dodałem jeszcze Naive Bayes (nb)

Wyniki heart_1 (gdzie jest 5000 rekordów):

```
=== Wyniki dla zbioru heart_1.csv ===
rf best params: {'clf__max_depth': None, 'clf__n_estimators': 50}
svm best params: {'clf__C': 10, 'clf__gamma': 'scale'}
ada best params: {'clf__learning_rate': 1.5, 'clf__n_estimators': 200}

=== Metryki na zbiorze testowym (po tuning) ===

Model: rf
Accuracy: 0.994
F1-score: 0.990
AUC-ROC: 0.998

Model: svm
Accuracy: 0.982
F1-score: 0.971
AUC-ROC: 0.988

Model: ada
Accuracy: 0.959
F1-score: 0.934
AUC-ROC: 0.991

=== Cross-validation (Accuracy ± SD) po tuningu ===
rf: 0.992 ± 0.004
svm: 0.979 ± 0.007
ada: 0.960 ± 0.010
```

Wnioski:

Random Forest (RF)

- Wyniki: Accuracy 0.994, F1 0.990, AUC 0.998
- Bias/Wariancja: Niski bias (bardzo dobrze dopasowuje się do danych), niska wariancja (mało różnic między CV wynikami).
- Wniosek: Bardzo silny model, dobrze radzi sobie z danymi, minimalne przetrenowanie (overfitting).

SVM

- **Wyniki:** Accuracy 0.982, F1 0.971, AUC 0.988
- **Bias/Wariancja:** Średni bias, niska wariancja (wyniki CV są stabilne).
- **Wniosek:** Model dobrze dopasowany, lekko bardziej restrykcyjny niż RF, ale nadal bardzo skuteczny.

AdaBoost

- **Wyniki:** Accuracy 0.959, F1 0.934, AUC 0.991
- **Bias/Wariancja:** Większy bias niż RF i SVM, umiarkowana wariancja (nieco większa rozpiętość wyników CV).
- **Wniosek:** Mniej elastyczny model, który jednak dobrze generalizuje, nieco bardziej konserwatywny.

Wyniki heart_2 (gdzie jest tylko 300 rekordow):

```
=== Wyniki dla zbioru heart_2.csv ===
rf best params: {'clf__max_depth': 5, 'clf__n_estimators': 100}
svm best params: {'clf__C': 1, 'clf__gamma': 'scale'}
ada best params: {'clf__learning_rate': 0.5, 'clf__n_estimators': 50}

=== Metryki na zbiorze testowym (po tuning) ===

Model: rf
Accuracy: 0.813
F1-score: 0.682
AUC-ROC: 0.890

Model: svm
Accuracy: 0.773
F1-score: 0.638
AUC-ROC: 0.863

Model: ada
Accuracy: 0.840
F1-score: 0.684
AUC-ROC: 0.902

=== Cross-validation (Accuracy  $\pm$  SD) po tuningu ===
rf: 0.843  $\pm$  0.063
svm: 0.783  $\pm$  0.089
ada: 0.826  $\pm$  0.067
```

Wnioski:

Random Forest

- **Wyniki:** Accuracy 0.813, F1 0.682, AUC 0.890
- **Bias/Wariancja:** Większy bias niż na heart_1 (max_depth=5 ogranicza złożoność), umiarkowana wariancja (większa od heart_1).
- **Wniosek:** Model bardziej uproszczony, może niedopasowywać się do trudniejszych danych (wyższy bias), ale stabilniejszy.

SVM

- **Wyniki:** Accuracy 0.773, F1 0.638, AUC 0.863
- **Bias/Wariancja:** Wyższy bias (C=1, umiarkowana regularyzacja), stosunkowo duża wariancja (niestabilne wyniki CV).
- **Wniosek:** Model mniej skuteczny, prawdopodobnie zbyt prosty dla tego zbioru lub dane są bardziej złożone.

AdaBoost

- **Wyniki:** Accuracy 0.840, F1 0.684, AUC 0.902
- **Bias/Wariancja:** Najmniejszy bias spośród trzech, ale większa wariancja (rozpiętość wyników CV).
- **Wniosek:** AdaBoost radzi sobie najlepiej na trudniejszym zbiorze, bo adaptacyjnie poprawia słabe klasyfikatory, ale jest bardziej podatny na zmienność.