

## Autorzy

Mateusz Ostaszewski 325203  
Michał Sadowski 325221

## Zadanie

Połączenie lasu losowego z SVM w zadaniu klasyfikacji. Postępujemy tak jak przy tworzeniu lasu losowego, tylko co drugi klasyfikator w lesie to SVM. Jeden z klasyfikatorów (SVM lub drzewo ID3) może pochodzić z istniejącej implementacji.

## Interpretacja i doprecyzowanie treści zadania

Celem zadania jest stworzenie hybrydowego modelu klasyfikatora, który łączy drzewa ID3 i maszyny wektorów nośnych (SVM).

## Algorytmy

1. Drzewo decyzyjne (ID3): Zaimplementujemy algorytm ID3 do budowy drzew decyzyjnych, który wybiera podział w węźle na podstawie maksymalizacji zysku informacyjnego (information gain).
2. SVM (Support Vector Machine): Będziemy korzystać z dostępnej implementacji SVM z biblioteki scikit-learn.

## Integracja w modelu hybrydowym:

- Dla każdego klasyfikatora generujemy losowy podzbiór danych treningowych.
- Co drugi klasyfikator jest zastępowany SVM. Wynik końcowy jest określany na podstawie głosowania większościowego.

## Metryki

•

$$\text{Dokładność} = \frac{TP + TN}{TP + TN + FP + FN}$$

•

$$F1 = 2 \cdot \frac{\text{Precyzja} \cdot \text{Czułość}}{\text{Precyzja} + \text{Czułość}}$$

•

$$\text{Precyzja} = \frac{TP}{TP + FP}$$

- 

$$\text{Czułość} = \frac{TP}{TP + FN}$$

- Macierz pomyłek

## **Eksperymenty**

### **Zbiory danych**

#### **1. Iris**

- Liczba przykładów: 150
- Liczba klas: 3 - sprowadzone do 2 (dla SVM)
- Liczba cech: 4 (długość/szerokość kielicha i płatek)

#### **2. Wine Quality**

- Liczba przykładów: 4898 (wino białe), 1599 (wino czerwone)
- Liczba klas: od 0 do 10 – oceny jakości wina (przyjmujemy od 0 do 5 za wina słabe, a powyżej za wina dobre)
- Liczba cech: 11 (chemiczne właściwości)

#### **3. Telecom Churn**

- Liczba przykładów: 3150
- Liczba klas: 2 (churn lub nie churn)
- Liczba cech: 21 (informacje o klientach, użycie usług).

OPISAĆ CO ZOSTAŁO ZMIENIONE !!!!!!!!!!!!!

**Walidacja modeli zostanie przeprowadzona za pomocą walidacji krzyżowej z podziałem na 5 podzbiorów.**

## **Eksperymenty numeryczne**

### **Wpływ parametrów modelu na skuteczność**

TODO

### **Skuteczność hybrydowego modelu**

Przeprowadziliśmy eksperyment, w którym porównaliśmy nasz hybrydowy las (nazywany dalej HybridRandomForest) z bazowym lasem losowym oraz modelem SVM z biblioteki scikit-learn (RandomForest). Modele z scikit-learn zostały stworzone z domyślnymi parametrami, natomiast nasz hybrydowy las został stworzony z hiperparametrami wyznaczonymi w poprzednim eksperymencie. Warto zaznaczyć, że trenowaliśmy oraz walidowaliśmy modele przy użyciu walidacji krzyżowej o stopniu 5, co oznacza, że wyniki są uśrednione.

## Wyniki dla Telecom Churn

Run Name	Cr	Duration	accuracy	f1	precision	recall	classifier_class	dataset_name
Group: load_ch... 3			0.9408465...	0.7875456...	0.8586482...	0.7380471...	-	-
popular-cow-932	i	2.1min	0.96380952...	0.88434997...	0.88802033...	0.88080808...	HybridRand...	load_churn
sassy-pug-481	i	1.6s	0.95555555...	0.85213775...	0.89179222...	0.81616161...	RandomFor...	load_churn
receptive-croc-997	i	0.9s	0.90317460...	0.62614910...	0.79613229...	0.51717171...	SVC	load_churn

Figure 1: Telecom Churn

## Porównanie metryk F1, Precyzja, Czulość i Dokladność

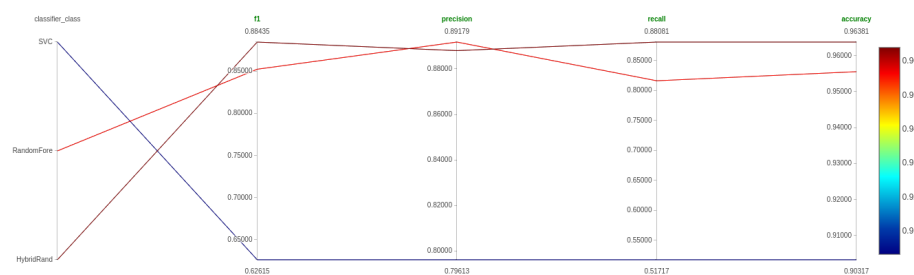


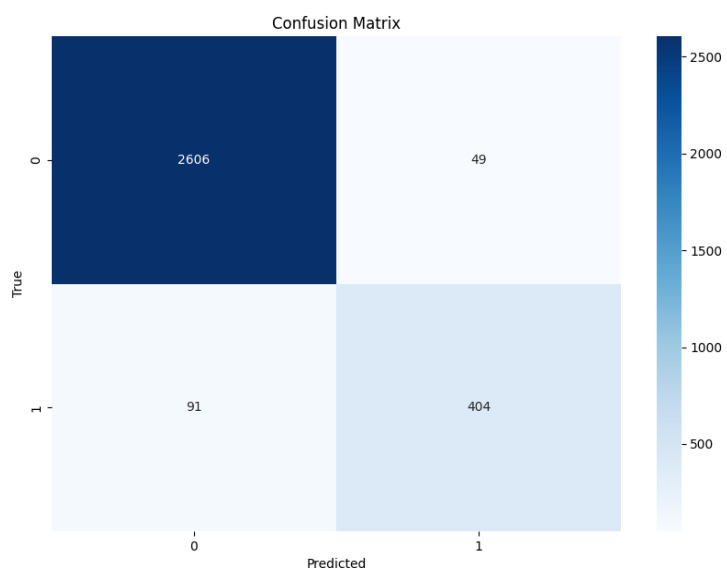
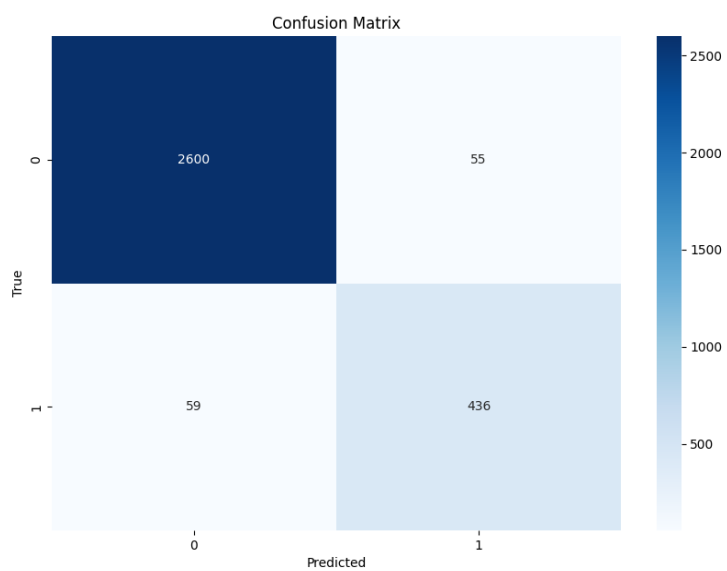
Figure 2: Telecom Churn

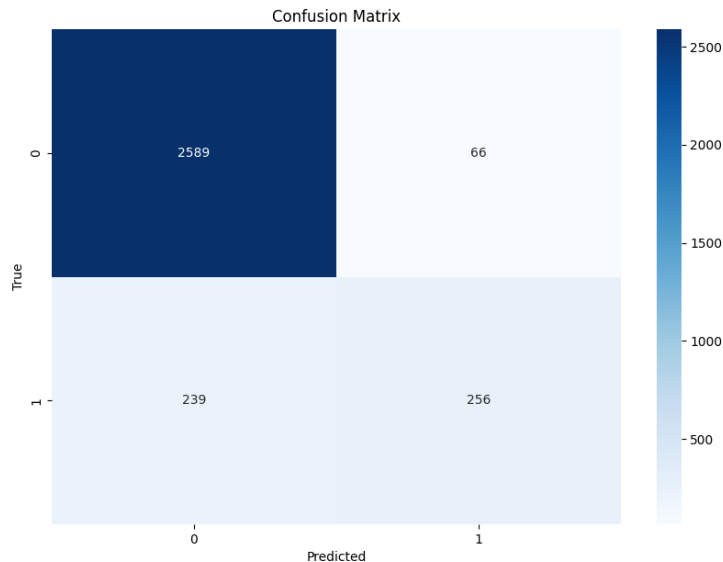
Wykres przedstawia wyniki dla czterech metryk dla trzech modeli.

### Obserwacje

- **SVC:**
  - F1: ~0,63 (niskie).
  - Precyzja: ~0,79.
  - Czulość: ~0,52 (bardzo niskie).
  - Dokladność: ~0,90.
  - Pomimo przyzwoitej dokladności, niska czulość sugerują, że model ma problem z poprawnym rozpoznaniem klasy 1.
- **RandomForestClassifier:**
  - F1: ~0,85 (dobre).
  - Precyzja i Czulość: ~0,89 i ~0,82 (dobry balans).
  - Dokladność: ~0,96.
  - Wyniki wskazują na stabilny model o dobrej wydajności.
- **HybridRandomForest:**
  - F1: ~0,90 (najwyższe).
  - Precyzja i Czulość: ~0,89 i ~0,88 (bardzo dobry balans).
  - Dokladność: ~0,96.
  - Model osiąga najlepsze wyniki we wszystkich metrykach (oprócz minimalnie wyższej Precyzji dla RandomForestClassifier), co czyni go liderem na tym zbiorze.

## Macierze pomyłek





Wizualizacja macierzy pomyłek potwierdza wcześniejsze wnioski. Liderem pozostaje HybridRandomForest, który popełnia bardzo mało błędów.

## Wnioski

W kontekście analizy churn (utrata klientów), **Czułość** jest szczególnie istotną metryką, ponieważ pozwala wykrywać jak najwięcej przypadków pozytywnych (np. klientów, którzy mogą odejść). HybridRandomForest osiągnął najlepszy wynik, co czyni go szczególnie użytecznym w tym zastosowaniu.

## Wyniki dla Wine Quality

Run Name	Created	Duration	accuracy	f1 $\bar{F}_1$	precision	recall	classifier_class	dataset_name
Group: load_iris 3	-		1 (average)	1 (average)	1 (average)	1 (average)	-	-
<a href="#">luxuriant-auk-762</a>	2 minutes ago	0.6s	1	1	1	1	SVC	load_iris
<a href="#">resilient-hare-879</a>	22 hours ago	0.8s	1	1	1	1	RandomFor...	load_iris
<a href="#">languid-yak-687</a>	1 day ago	1.7s	1	1	1	1	HybridRand...	load_iris

Figure 3: Wine Quality

## Porównanie metryk F1, Precyzja, Czułość i Dokładność

### Obserwacje

- SVC:
  - **F1**: ~0,82 .
  - **Precyzja**: ~0,76.

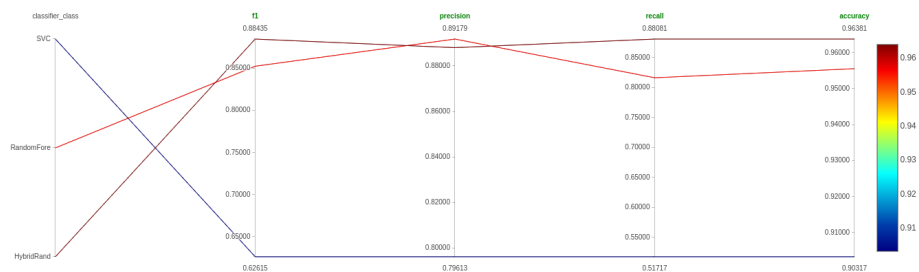
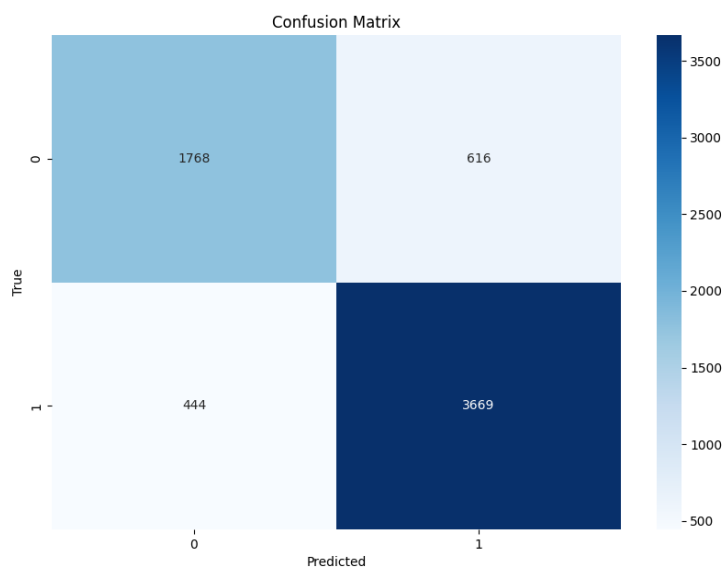
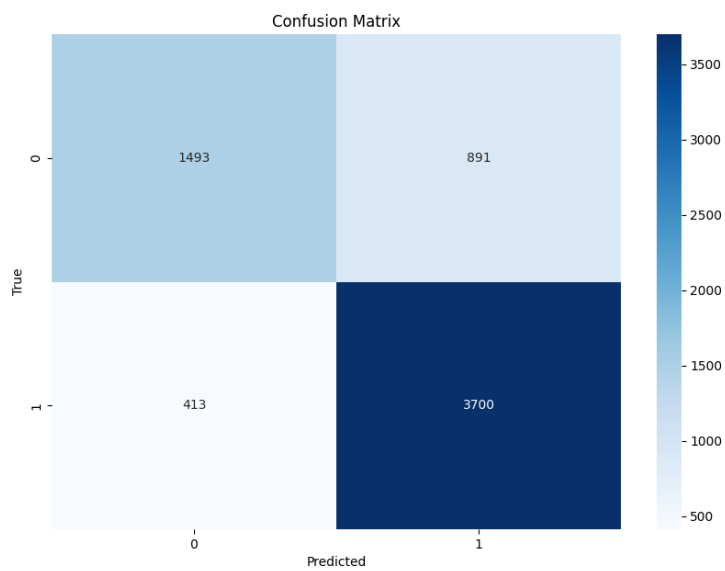
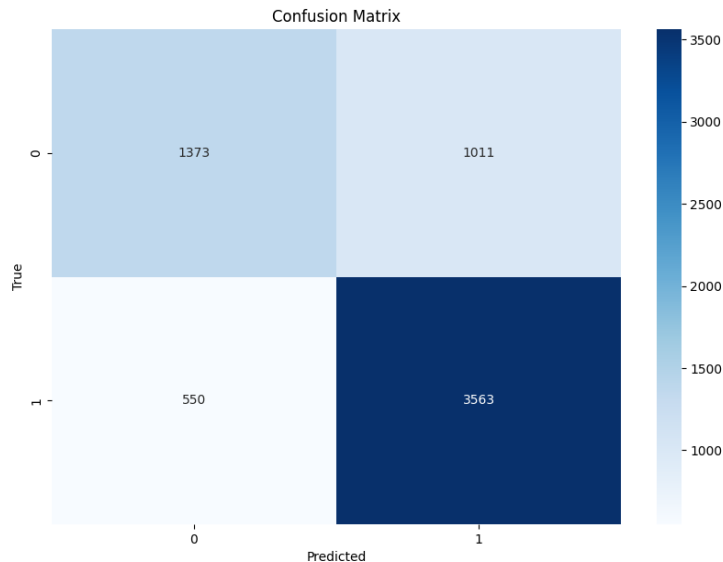


Figure 4: Wykres przedstawia wyniki dla czterech metryk dla trzech modeli.

- **Czułość:** ~0,87.
- **Dokładność:** ~0,78.
- Model charakteryzuje się wysokim poziomem czułości i niskimi innymi metrykami, co może sugerować, że model zazwyczaj przewidywa klasę 1.
- **RandomForestClassifier:**
  - **F1:** ~0,87.
  - **Precyzja:** ~0,86.
  - **Czułość:** ~0,89.
  - **Dokładność:** ~0,84.
  - Model prezentuje bardzo dobre wyniki we wszystkich metrykach, zwłaszcza dokładność.
- **HybridRandomForest:**
  - **F1:** ~0,85.
  - **Precyzja:** ~0,80.
  - **Czułość:** ~0,90.
  - **Accuracy:** ~0,80.

Macierze pomyłek





Wizualizacja macierzy pomyłek potwierdza wcześniejsze wnioski. Bazowy RandomForestClassifier zdecydowanie lepiej radzi sobie z klasą, której jest mniej.

### Wnioski

RandomForestClassifier okazał się liderem na zbiorze Wine Quality. HybridRandomForest charakteryzuje się sensownymi wynikami, lecz ma tendencję do przewidywania klasy z większą ilością próbek.

### Wyniki dla Iris

Run Name	Created	Duration	accuracy	f1	precision	recall	classifier_class	dataset_name
Group: load_iris 3	-	1 (average)	1 (average)	1 (average)	1 (average)	1 (average)	-	-
<a href="#">luxuriant-auk-762</a>	2 minutes ago	0.6s	1	1	1	1	SVC	load_iris
<a href="#">resilient-hare-879</a>	22 hours ago	0.8s	1	1	1	1	RandomFor...	load_iris
<a href="#">languid-yak-687</a>	1 day ago	1.7s	1	1	1	1	HybridRand...	load_iris

Figure 5: Iris

Wszystkie modele uzyskały perfekcyjne wyniki na tym zbiorze danych. Zbiór ten okazał się “zbyt prosty” dla wszystkich porównywanych modeli.

**Wnioski z eksperymentu** Podsumowując, HybridRandomForest okazał się najbardziej efektywnym modelem w analizie churn, natomiast RandomForestClassifier był liderem na zbiorze Wine Quality. SVM nie sprawdził się dobrze w



żadnym z zadań. Istotną uwagę jest to że nasz autorski HybridRandomForest jest wielokrotnie wolniejszy od modeli z scikit-learn.

## **Wnioski**

TODO