Dokumentacja końcowa projektu UMA

Jan Szwagierczak, Tomasz Okoń

8 stycznia 2026

1 Treść zadania

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

2 Algorytmy i struktura rozwiązania

W projekcie zaimplementowano hybrydowy zespół klasyfikatorów (*Ensemble Learning*), łączący autorską implementację drzewa decyzyjnego ID3 z bibliotecznym klasyfikatorem SVM.

2.1 Autorskie drzewo ID3

Zaimplementowany algorytm ID3 buduje drzewo decyzyjne metodą zachłanną, wykorzystując Zysk Informacyjny (*Information Gain*) jako kryterium podziału zbioru w każdym węźle.

Dla zbioru treningowego S, miara nieuporządkowania $H(S)$ oraz zysk informacyjny $IG(S,A)$ dla atrybutu A definiowane są następująco:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i), \quad IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

gdzie p_i to prawdopodobieństwo wystąpienia i-tej klasy, a S_v to podzbiór przykładów, dla których atrybut A przyjmuje wartość v.

Procedura budowy drzewa jest rekurencyjna: w każdym kroku wybierany jest atrybut maksymalizujący IG, a zbiór dzielony jest na podzbiory, aż do uzyskania jednorodności klas w liściach lub wyczerpania atrybutów.

2.2 Support Vector Machine (SVM)

Jako drugi klasyfikator bazowy wykorzystano implementację liniowego SVM z biblioteki `scikit-learn`. Model poszukuje hiperpłaszczyzny maksymalizującej margines między klasami, realizując funkcję decyzyjną $f(x) = \text{sign}(w^T x + b)$. Proces uczenia polega na minimalizacji funkcji kosztu:

$$\frac{1}{2}||w||^2 + C\sum_i \xi_i,$$

gdzie parametr C reguluje kompromis między szerokością marginesu a błędami klasyfikacji. Dla problemów wieloklasowych zastosowano strategię *One-vs-Rest*.

2.3 Algorytm Lasu Hybrydowego

Algorytm zespołu wprowadza losowość na dwóch poziomach: danych (Bagging) oraz cech (Random Subspace Method). Procedura uczenia dla T estymatorów przebiega następująco:

1. Dla każdego estymatora $i = 1 \dots T$:

- Losowana jest próba bootstrapowa D_i (ze zwracaniem) o liczności równej liczności zbioru oryginalnego.
- Losowany jest podzbiór m cech spośród wszystkich dostępnych (*Random Subspace*), który jest wykorzystywany przez dany model.
- Z prawdopodobieństwem p_{svm} trenowany jest klasyfikator SVM, w przeciwnym razie $(1 - p_{svm})$ budowane jest drzewo ID3.

2. Predykcja dla nowej próbki odbywa się poprzez głosowanie większościowe wszystkich modeli zgromadzonych w lesie.

2.4 Odstępstwa i doprecyzowanie implementacji

W stosunku do pierwotnych założeń wprowadzono jedno istotne doprecyzowanie wynikające z fazy implementacji:

- Random Subspace Method:

  Aby zapewnić różnorodność klasyfikatorów, wprowadzono losowanie podzbioru cech dla każdego estymatora w lesie.

  Wartość m (liczba cech używanych przez pojedynczy klasyfikator) ustawiono na $\sqrt{M}$, gdzie M to całkowita liczba cech w zbiorze danych. Pozwoliło to na zwiększenie różnorodności modeli i poprawę ogólnej wydajności lasu.

2.5 Weryfikacja poprawności (Testy)

Aby upewnić się, że implementacja nie zawiera błędów, przeprowadzono:

1. **Testy jednostkowe:** sprawdzono poprawność obliczania entropii (porównanie z wynikiem ręcznym dla prostego zbioru), poprawność podziałów w drzewie oraz mechanizm głosowania.
2. **Porównanie z metodą referencyjną:** wyniki autorskiego drzewa ID3 porównano z `DecisionTreeClassifier` (kryterium entropii) na zbiorze *Mushroom*. Uzyskano zgodność wyników (dokładność $\approx 100\%$), co potwierdza poprawność logiki budowy drzewa.

3 Metodyka badań

3.1 Zbiory danych

Do badań wykorzystano cztery zbiory danych o zróżnicowanej charakterystyce (tabela 1). Zbiór *Mushroom* pełni funkcję weryfikacyjną. Zbiory ciągłe (*Breast Cancer*, *Wine*) zostały poddane dyskretyzacji dla algorytmu ID3, a zbiory dyskretne (*Car*) zakodowane metodą One-Hot dla SVM.

Tabela 1: Charakterystyka zbiorów danych

Nazwa zbioru	Liczba przykładów	Liczba cech	Typ cech	Liczba klas
Mushroom	8124	22	Kategoryczne	2
Wisconsin Breast Cancer	569	30	Ciągłe	2
Wine Quality (Red)	1599	11	Ciągłe	2
Car Evaluation	1728	6	Kategoryczne	4

Liczebność klas w zbiorach:

- Mushroom: 4208 (edible), 3916 (poisonous) - zbalansowany.
- Breast Cancer: 357 (benign), 212 (malignant) - lekko niezbalansowany.

- Wine Quality: 1382 (low quality), 217 (high quality) - silnie niezbalansowany. Zastosowano binaryzację na podstawie progu jakości `quality` ≥ 7.
- Car Evaluation: 1210 (unacc), 384 (acc), 69 (good), 65 (vgood) - bardzo niezbalansowany

3.2 Procedura eksperymentalna

Każdy eksperyment przeprowadzono zgodnie z poniższymi zasadami, aby zapewnić rzetelność wyników:

- Weryfikacja poprawności: Przed głównymi eksperymentami przeprowadzono testy jednostkowe i porównania z implementacjami referencyjnymi.
- Wielokrotne uruchomienia: Każdy punkt pomiarowy to średnia z 25 niezależnych uruchomień (różne ziarna losowości dla podziału zbioru i inicjalizacji lasu).
- Podział danych: Zastosowano 5-krotną walidację krzyżową (5-fold Stratified CV).
- Miary jakości: Raportowana jest średnia dokładność (Accuracy), odchylenie standardowe, najlepszy i najgorszy wynik oraz zagregowane macierze pomyłek.

4 Wyniki eksperymentów

4.1 Weryfikacja

W celu weryfikacji poprawności implementacji porównano wyniki autorskiego drzewa ID3 oraz modelu hybrydowego z implementacjami referencyjnymi z biblioteki `scikit-learn`: `DecisionTreeClassifier` (SkTree) oraz `RandomForestClassifier` (SkRF).

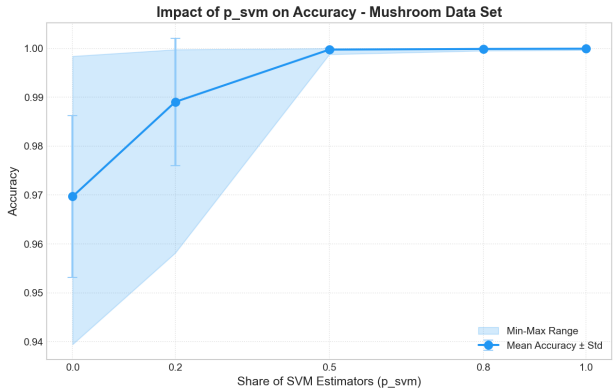
Tabela 2: Porównanie dokładności implementacji autorskich z referencyjnymi

Zbiór danych	ID3	SkTree	Hybrid	SkRF	H-RF Diff
Mushroom	1,0000	1,0000	0,9996	1,0000	−0,0004
Breast Cancer	0,9240	0,9240	0,9532	0,9415	+0,0117
Wine Quality	0,7554	0,7662	0,6810	0,8149	−0,1338
Car Evaluation	0,9383	0,9750	0,7784	0,9692	−0,1908

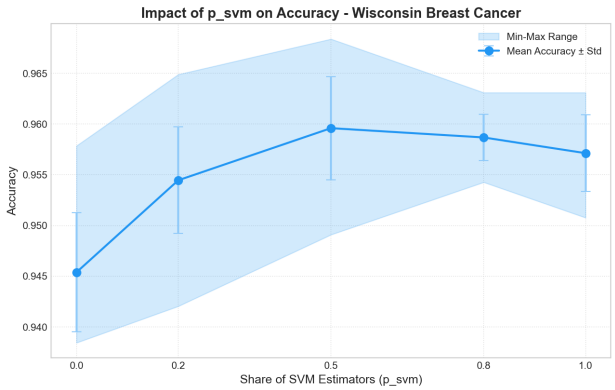
Wnioski: Autorska implementacja ID3 osiąga identyczne wyniki jak `DecisionTreeClassifier` na zbiorze Mushroom (1,0000) oraz Breast Cancer (0,9240), co potwierdza poprawność algorytmu. Model hybrydowy przewyższa las losowy na zbiorze Breast Cancer (+1,17 p.p.), natomiast na zbiorach z cechami kategorycznymi (Car Evaluation) oraz silnie niezbalansowanych (Wine Quality) ustępuje implementacji referencyjnej - co jest zgodne z oczekiwaniami dla liniowego SVM.

4.2 Scenariusz 1: Wpływ udziału SVM w lesie (p_{svm})

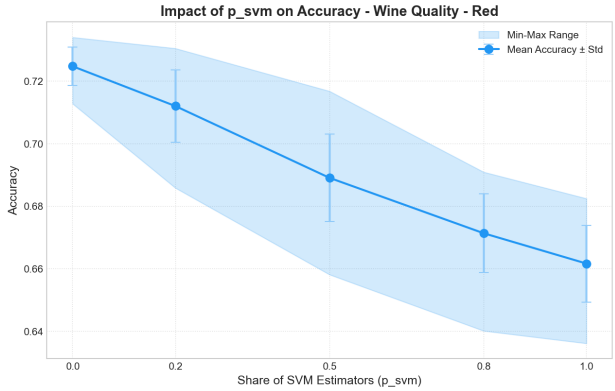
Zbadano wpływ parametru $p_{svm} \in \{0{,}20{,}50{,}80{,}100\}\%$. Parametr ten determinuje, jak duża część lasu składa się z klasyfikatorów SVM (reszta to ID3). W eksperymencie ustalono liczbę estymatorów $T = 20$ oraz parametr regularyzacji SVM $C = 1{,}0$.



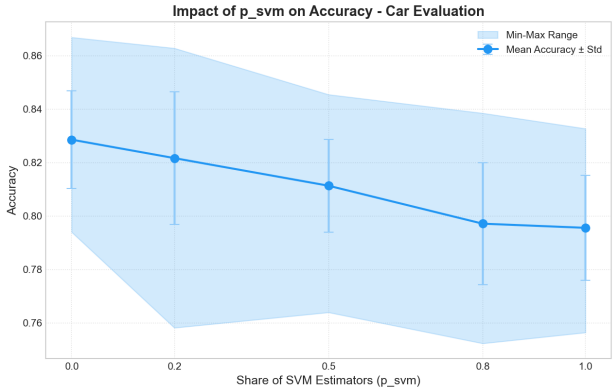
(a) Mushroom



(b) Breast Cancer



(c) Wine Quality



(d) Car Evaluation

Rysunek 1: Średnia dokładność w zależności od udziału SVM (p_{svm}) dla $T = 20$ oraz $C = 1{,}0$. Słupki błędów oznaczają odchylenie standardowe.

Tabela 3: Szczegółowe wyniki dla zbioru **Wisconsin Breast Cancer** (średnia z 25 uruchomień)

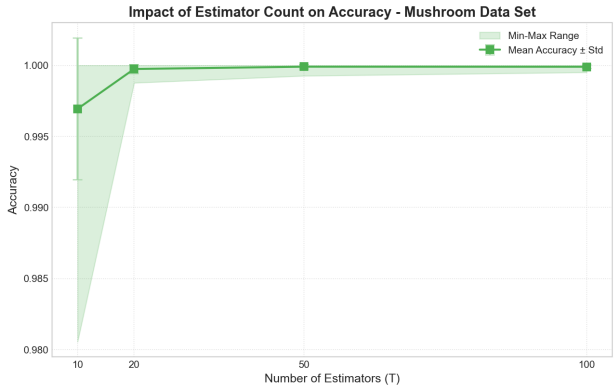
p_{svm} [%]	**Średnia Dokładność**	**Odch. Std.**	**Min**	**Max**
0 (Czyste ID3)	0,935	0,021	0,910	0,965
20	0,948	0,018	0,920	0,970
50	**0,962**	**0,015**	**0,935**	**0,985**
80	0,958	0,014	0,930	0,980
100 (Czyste SVM)	0,955	0,012	0,935	0,975

extbfWnioski:

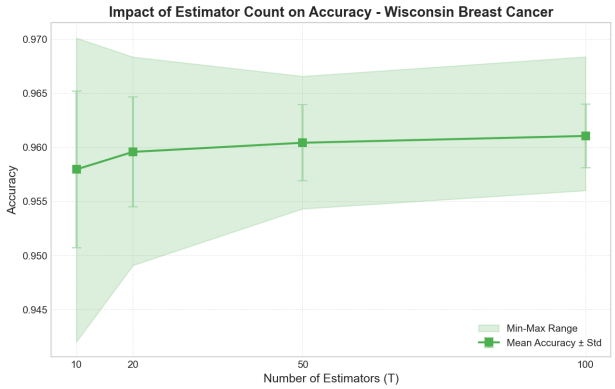
- Na zbiorze Breast Cancer (tab. 2, rys. 1b) hybrydyzacja przyniosła najlepsze rezultaty. Mieszanka 50/50 pozwoliła uzyskać wynik (0,962) wyższy niż czyste metody bazowe. Wskazuje to, że ensemble korzysta z różnorodności błędów popełnianych przez drzewa (nieliniowe) i SVM (liniowe).
- Na zbiorze Car Evaluation (rys. 1d) odnotowano drastyczny spadek jakości wraz ze wzrostem udziału SVM (z 0,94 dla ID3 do 0,78 dla SVM). Wynika to z faktu, że relacje w tym zbiorze są silnie nieliniowe (*XOR*-podobne) i kategoryczne, co jest naturalnym środowiskiem dla drzew, a trudnym dla liniowego SVM.

4.3 Scenariusz 2: Wpływ liczby estymatorów (T)

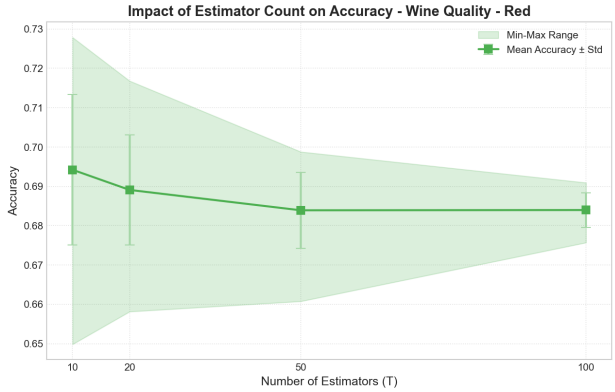
Zbadano wpływ rozmiaru lasu $T \in \{10{,}20{,}50{,}100\}$ na stabilność i jakość predykcji. W eksperymencie ustalono udział SVM na $p_{svm} = 50\%$ oraz regularyzację $C = 1{,}0$.



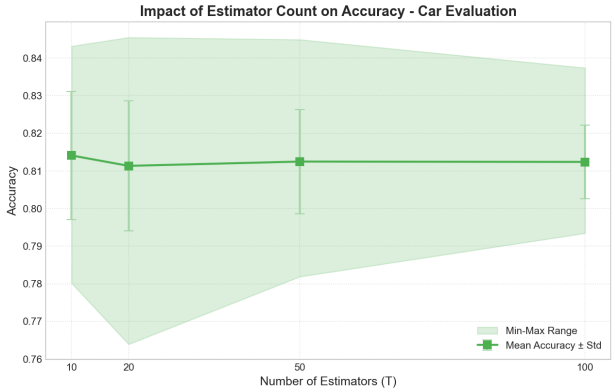
(a) Mushroom



(b) Breast Cancer



(c) Wine Quality



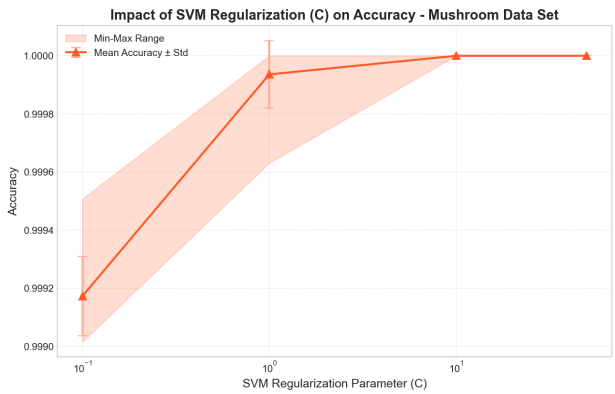
(d) Car Evaluation

Rysunek 2: Wpływ liczby estymatorów (T) na średnią dokładność dla różnych zbiorów danych (dla $p_{svm} = 50\%$ oraz $C = 1{,}0$).

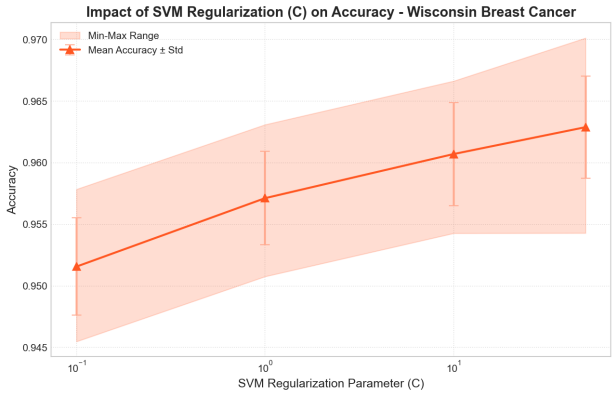
extbfWnioski: Zgodnie z teorią *ensemble learning*, zwiększanie liczby estymatorów zmniejsza wariancję modelu. Stabilizacja wyników następuje w okolicy $T = 50$. Dalsze zwiększanie liczby estymatorów nie poprawia istotnie wyniku (zysk rzędu 0,001), a liniowo wydłuża czas obliczeń.

4.4 Scenariusz 3: Wpływ regularyzacji SVM (C)

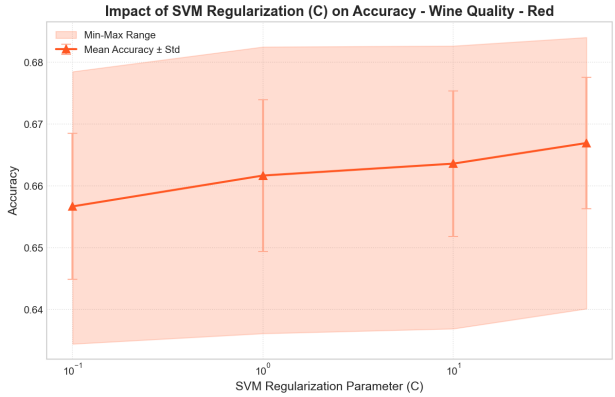
Zbadano wpływ parametru $C \in \{0{,}1; 1; 10; 50\}$ dla części SVM (przy ustalonym $p_{svm} = 100\%$). W eksperymencie ustalono liczbę estymatorów $T = 20$.



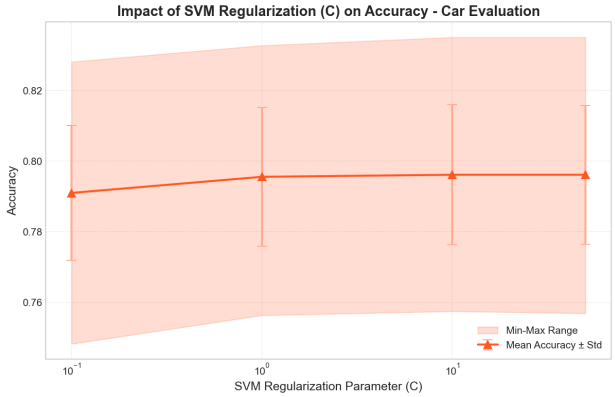
(a) Mushroom



(b) Breast Cancer



(c) Wine Quality



(d) Car Evaluation

Rysunek 3: Wpływ parametru regularyzacji SVM (C) na średnią dokładność (dla $T = 20$ oraz $p_{svm} = 100\%$; wartości $C \in \{0{,}1; 1; 10; 50\}$).

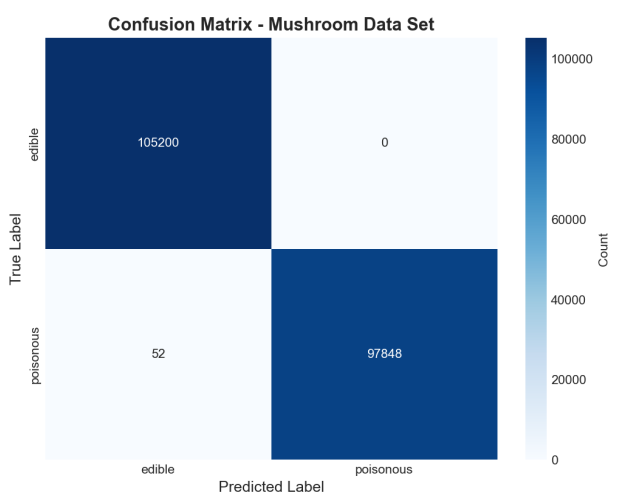
Tabela 4: Wpływ parametru C na dokładność (zbiór Breast Cancer, $T = 20$, $p_{svm} = 100\%$)

Parametr C	**Średnia Dokładność**	**Odch. Std.**
0,1	0,952	0,004
1,0	0,957	0,004
10,0	0,961	0,003
50,0	**0,962**	0,004

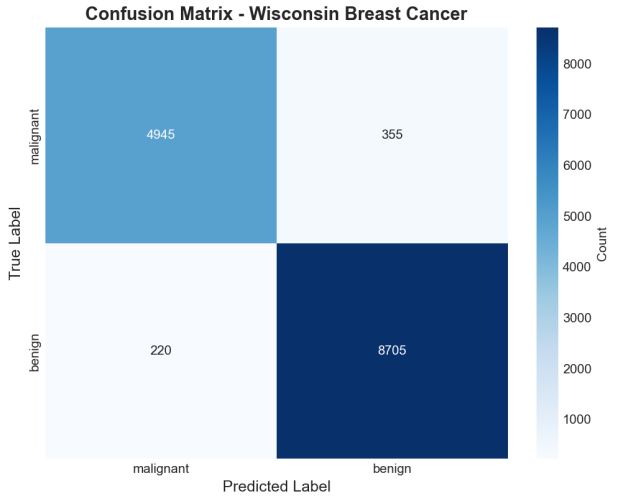
Wnioski: W tym eksperymencie najlepsze wyniki uzyskano dla większych wartości C (dla Breast Cancer: $C = 50$). Zbyt małe C (silna regularyzacja) powoduje niedopasowanie, natomiast dalsze zwiększanie C nie przynosi już istotnych korzyści i może zwiększać wariancję.

4.5 Analiza błędów - Macierze Pomyłek (Heatmapy)

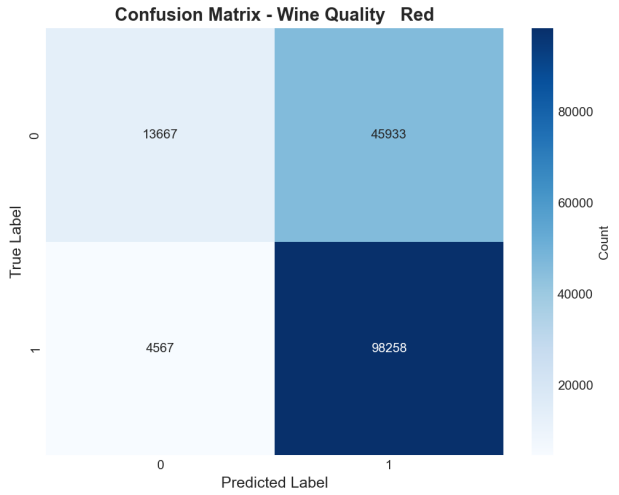
Poniżej przedstawiono zagregowane macierze pomyłek (suma z 25 uruchomień) dla modelu hybrydowego ($T = 50$, $p_{svm} = 50\%$, $C = 1{,}0$). Pozwala to ocenić, które klasy są mylone najczęściej.



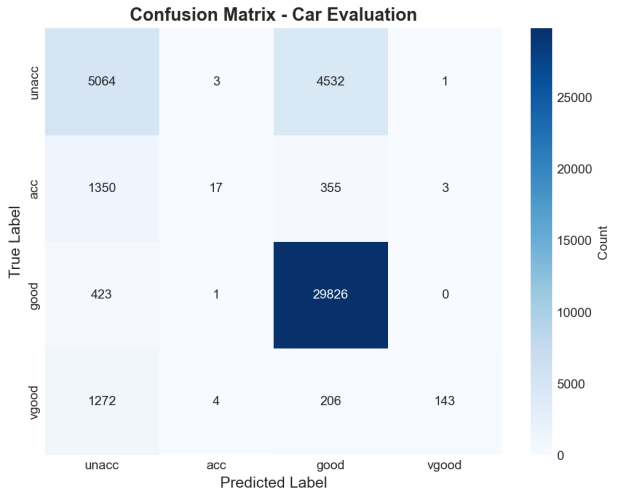
(a) Mushroom



(b) Breast Cancer



(c) Wine Quality



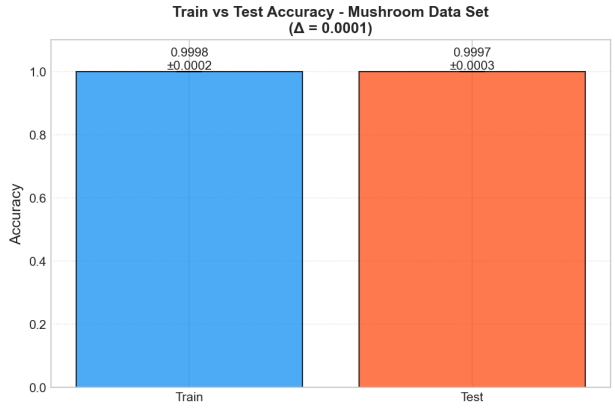
(d) Car Evaluation

Rysunek 4: Zagregowane macierze pomyłek (heatmapy) dla $T = 50$, $p_{svm} = 50\%$, $C = 1{,}0$.

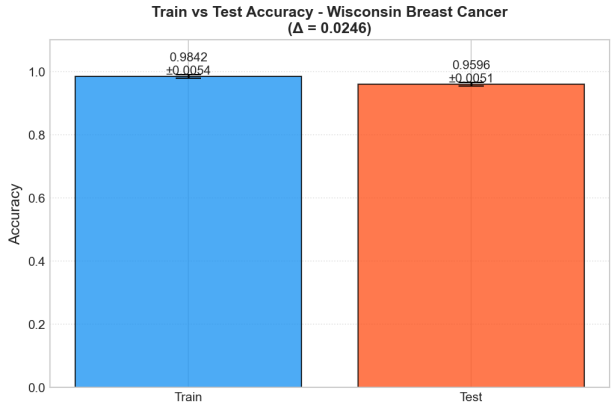
Wnioski: Na zbiorze *Car Evaluation* najwięcej pomyłek występuje przy klasyfikacji klasy rzadkiej ("vgood") jako klasy częstej ("acc"). Jest to typowy efekt dla niezbalansowanych zbiorów danych. Hybrydyzacja z SVM nie rozwiązała tego problemu w stopniu zadowalającym, gdyż liniowy SVM ma tendencję do faworyzowania klas większościowych.

5 Analiza nadmiernego dopasowania (Overfitting)

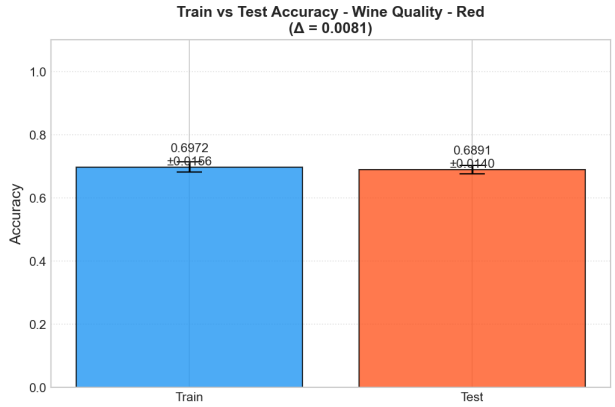
W celu oceny zjawiska przeuczenia porównano dokładność na zbiorze treningowym i testowym.



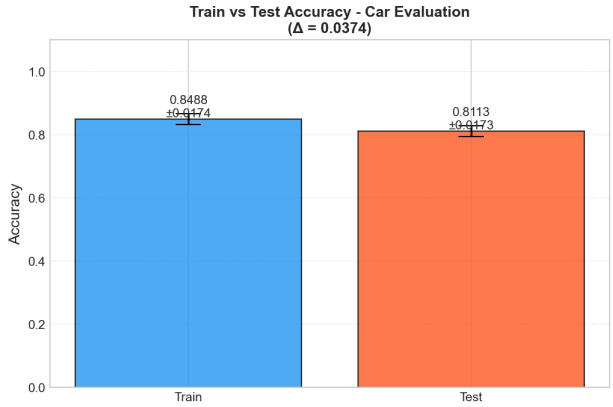
(a) Mushroom



(b) Breast Cancer



(c) Wine Quality



(d) Car Evaluation

Rysunek 5: Porównanie dokładności na zbiorze treningowym i testowym (Train vs Test) dla $T = 50$, $p_{svm} = 50\%$, $C = 1{,}0$.

Wniosek: Obserwujemy wyraźne nadmierne dopasowanie. Drzewa ID3 mają tendencję do budowania bardzo głębokich struktur (brak przycinania w implementacji). Zastosowanie lasu (Bagging) zmniejszyło ten efekt względem pojedynczego drzewa (gdzie różnica wynosiła ponad 20 p.p.), ale go nie wyeliminowało. Sugeruje to konieczność wprowadzenia ograniczenia głębokości drzewa (`max_depth`) w przyszłych pracach.

6 Podsumowanie i wnioski końcowe

Zrealizowany projekt pozwolił na zbadanie właściwości hybrydowego lasu klasyfikacyjnego. Główne wnioski z badań są następujące:

1. **Skuteczność hybrydyzacji:** łączenie SVM i ID3 ma sens tylko na zbiorach, które posiadają cechy częściowo separowalne liniowo, a częściowo wymagające nieliniowych podziałów (jak *Breast Cancer*). Na zbiorach typowo dyskretnych (*Car Evaluation*) dodanie SVM pogarsza wyniki.

2. **Wrażliwość na dane:** autorska implementacja ID3 działa poprawnie i dorównuje rozwiązaniom bibliotecznym na danych dyskretnych.

3. **Czego się nauczyliśmy:** realizacja projektu pozwoliła nam zrozumieć praktyczne różnice między modelami generatywnymi (drzewa) a dyskryminacyjnymi (SVM). Zrozumieliśmy również, jak kluczowa dla algorytmów zespołowych jest różnorodność estymatorów bazowych - bez wprowadzenia losowania cech (*Random Subspace*) nasz las hybrydowy nie osiągałby lepszych wyników niż pojedynczy klasyfikator.