

Dokumentacja końcowa projektu UMA

Jan Szwagierczak, Tomasz Okoń

9 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, 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 z największym zyskiem informacyjnym, a zbiór dzielony jest na podzbiory, aż do uzyskania takich samych 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:

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

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

2.3 Algorytm Lasu

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 (na danych ciągłych/One-Hot), w przeciwnym razie $(1 - p_{svm})$ budowane jest drzewo ID3 (na danych dyskretnych).
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 następujące doprecyzowania:

- **Random Subspace Method:** Aby zapewnić różnorodność klasyfikatorów, wprowadzono losowanie podzbioru cech dla każdego estymatora. Wartość m (liczba cech używanych przez pojedynczy klasyfikator) domyślnie ustawiono na \sqrt{M} (gdzie M to liczba cech), z dynamicznym dostosowaniem dla zbiorów o bardzo małej liczbie atrybutów.
- **Obsługa danych heterogenicznych:** Las obsługuje jednocześnie dwie reprezentacje danych: dyskretną (dla ID3) oraz ciągłą/zakodowaną One-Hot (dla SVM), przekazując do konkretnego estymatora odpowiedni format danych wejściowych.

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` 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 (metoda kubelkowa), natomiast zbiory dyskretne (*Mushroom*, *Car Evaluation*) zostały zakodowane metodą One-Hot Encoding na potrzeby klasyfikatora 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

Szczegółowy rozkład klas w zbiorach:

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

- **Wine Quality (Red):** 744 (low quality), 855 (high quality) - zbiór lekko niezbalansowany. Zastosowano binaryzację etykiet, przyjmując próg jakości $\text{quality} \geq 6$ jako klasę pozytywną.
- **Car Evaluation:** 1210 (unacc), 384 (acc), 69 (good), 65 (vgood) - zbiór silnie niezbalansowany.

3.2 Procedura eksperymentalna

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

- **Weryfikacja poprawności:** Przed przystąpieniem do głównych eksperymentów wykonano testy jednostkowe oraz porównano wyniki autorskiej implementacji z metodami referencyjnymi (biblioteka `scikit-learn`).
- **Wielokrotne uruchomienia:** Każdy pomiar stanowi średnią z 25 niezależnych uruchomień algorytmu. W każdej iteracji stosowano inne ziarno losowości (`random_state`).
- **Podział danych:** Zastosowano 5-krotną walidację krzyżową, co w połączeniu z 25 powtórzeniami daje łącznie 125 prób trenowania dla każdej konfiguracji parametrów.
- **Miary jakości:** Raportowana jest średnia dokładność (Accuracy), odchylenie standardowe, wynik minimalny i maksymalny oraz zagregowane macierze pomyłek (sumowane ze wszystkich uruchomień).

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).

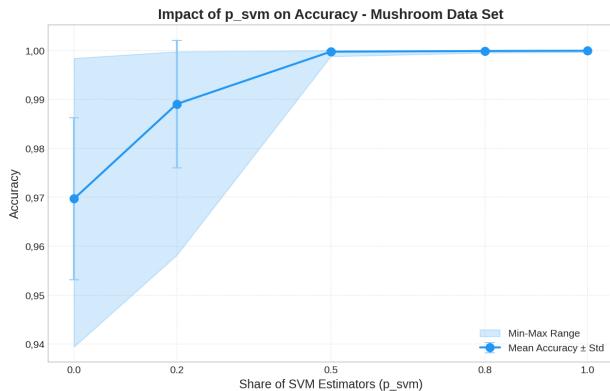
Tabela 2: Porównanie dokładności implementacji autorskich z referencyjnymi

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	0,7458	0,7542	0,7250	0,7812	-0,0563
Car Evaluation	0,9383	0,9750	0,7784	0,9692	-0,1908

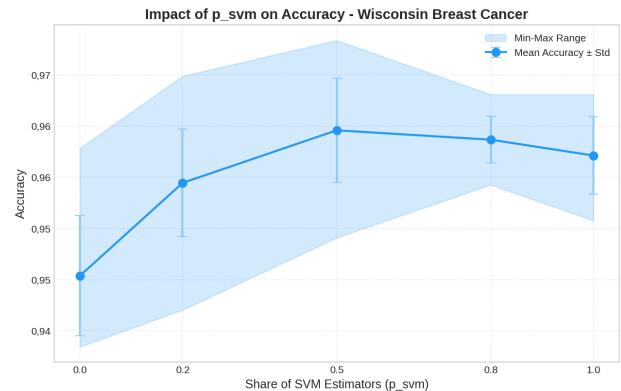
Wnioski: Autorska implementacja ID3 osiąga identyczne wyniki jak `DecisionTreeClassifier` na zbiorze Mushroom (1,0000) oraz Breast Cancer (0,9240), co potwierdza poprawność algorytmu. Model hybrydowy przewyższa las losowy na zbiorze Breast Cancer (+1,17 p.p.), natomiast na zbiorach z cechami kategorycznymi (Car Evaluation) oraz wieloklasowych radzi sobie gorzej niż metoda referencyjna - co jest zgodne z oczekiwaniemi dla liniowego SVM. Na zbiorze Wine Quality model hybrydowy (0,725) ustępuje referencyjnemu lasowi losowemu (0,781) o 5,6 p.p.

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

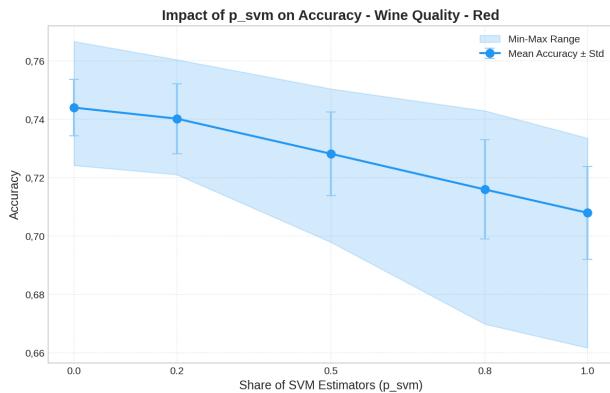
Zbadano wpływ parametru $p_{svm} \in \{0,20,50,80,100\}\%$. Parametr ten określa, jak duża część lasu składa się z klasyfikatorów SVM (reszta to ID3). W eksperymencie ustalone liczbę estymatorów $T = 20$ oraz parametr regularizacji SVM $C = 1,0$.



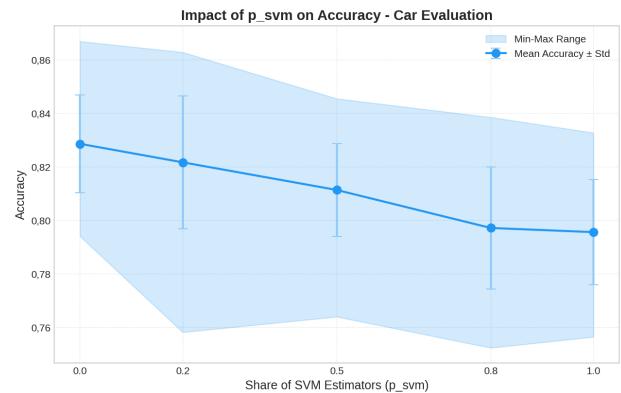
(a) Mushroom



(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)	0,945	0,006	0,938	0,958
20	0,954	0,005	0,942	0,965
50	0,960	0,005	0,949	0,968
80	0,959	0,002	0,954	0,963
100 (Czyste SVM)	0,957	0,004	0,951	0,963

Tabela 4: Szczegółowe wyniki dla zbioru **Wine Quality** (średnia z 25 uruchomień)

p_{svm} [%]	Średnia Dokładność	Odch. Std.	Min	Max
0 (Czyste ID3)	0,744	0,010	0,724	0,767
20	0,740	0,012	0,721	0,760
50	0,728	0,014	0,698	0,750
80	0,716	0,017	0,670	0,743
100 (Czyste SVM)	0,708	0,016	0,662	0,734

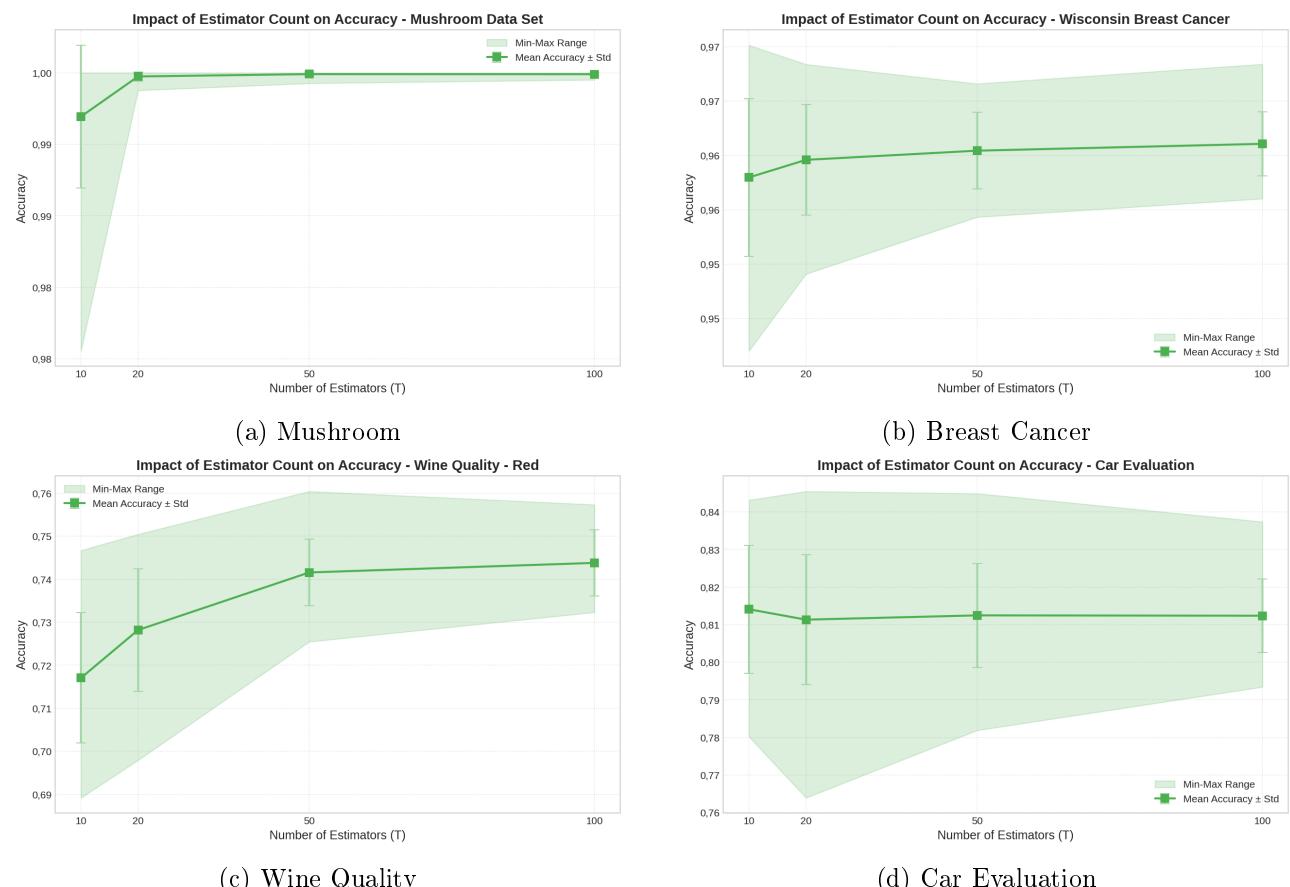
Wnioski:

- Na zbiorze Mushroom (rys. 1a) obserwujemy odwrotny trend niż na pozostałych zbiorach - zwiększenie udziału SVM poprawia wyniki. Dla czystego ID3 ($p_{svm} = 0\%$) dokładność wynosi ok. 0,97, podczas gdy dla czystego SVM ($p_{svm} = 100\%$) osiągamy wartość zbieżną do 1,00. Zbiór ten jest liniowo separowalny, co stanowi idealne środowisko dla SVM.

- Na zbiorze Breast Cancer (tab. 3, rys. 1b) hybrydyzacja przyniosła najlepsze rezultaty. Optimum uzyskano dla $p_{svm} = 50\%$ (dokładność 0,960), co jest wynikiem wyższym niż dla czystego ID3 (0,945) oraz czystego SVM (0,957). Sugeruje to, że model hybrydowy skutecznie wykorzystuje różniące się własności zastosowanych klasyfikatorów: drzewo umożliwia modelowanie nieliniowych granic decyzyjnych, a SVM zapewnia liniową separację w przestrzeni cech.
- Na zbiorze Wine Quality (tab. 4, rys. 1c) obserwujemy tendencję spadkową - najlepsze wyniki uzyskano dla czystego ID3 ($p_{svm} = 0\%$, dokładność 0,744), a zwiększenie udziału SVM stopniowo pogarszało jakość klasyfikacji (spadek do 0,708 dla $p_{svm} = 100\%$). Sugeruje to, że relacje w tym zbiorze są nieliniowe i lepiej modelowane przez drzewa decyzyjne niż przez liniowy SVM.
- Na zbiorze Car Evaluation (rys. 1d) odnotowano podobny trend jak dla Wine Quality - spadek jakości wraz ze wzrostem udziału SVM (z 0,829 dla ID3 do 0,796 dla SVM). Relacje w tym zbiorze są silnie nieliniowe i kategoryczne, co faworyzuje drzewa decyzyjne.

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 ustalonono udział SVM na $p_{svm} = 50\%$ oraz regularyzację $C = 1,0$.



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

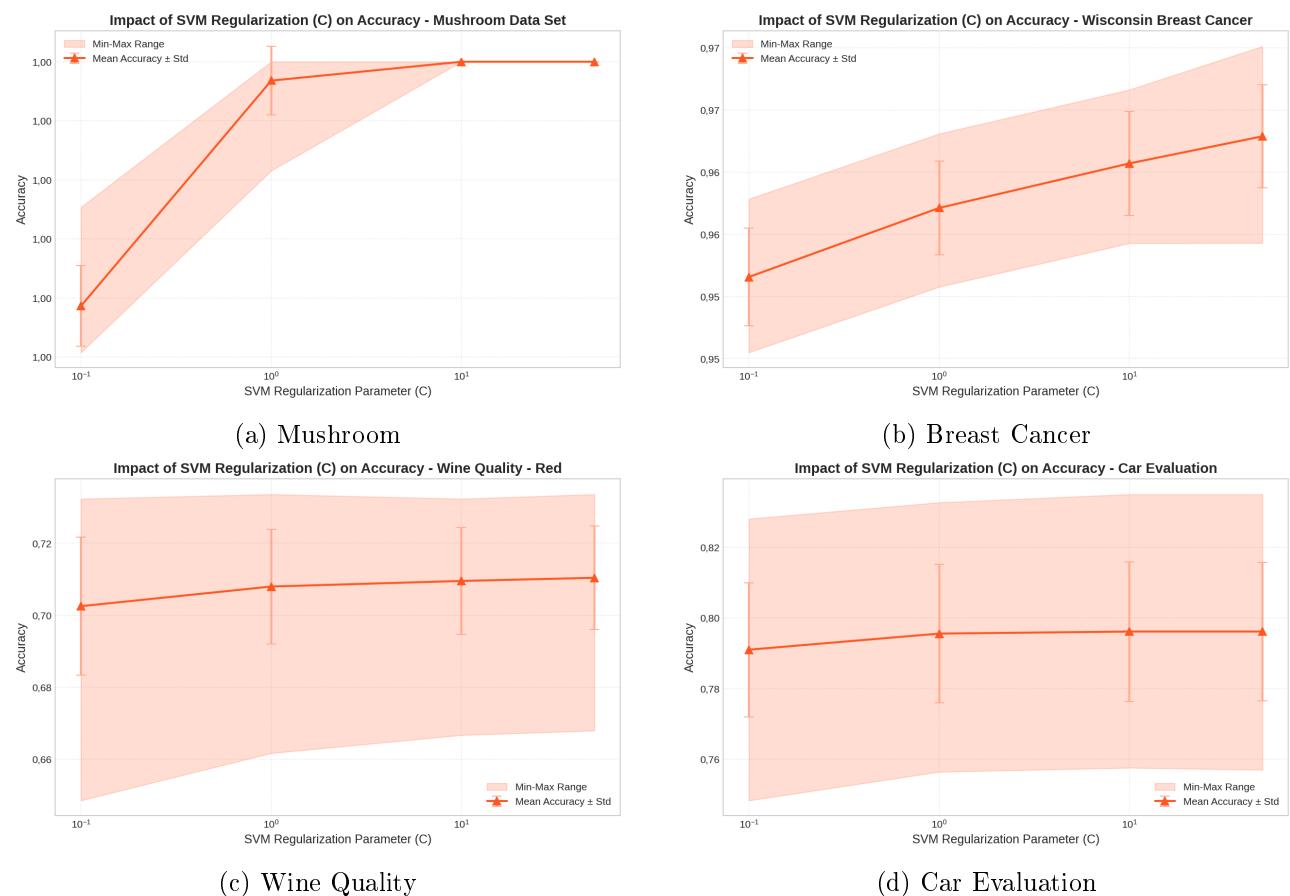
Wnioski:

- Na zbiorze Mushroom (rys. 2a) obserwujemy szybką zbieżność do optimum - już dla $T = 10$ osiągamy dokładność 0,997, a dla $T = 20$ wynosi ona 0,9997. Dalsze zwiększenie liczby estymatorów nie przynosi istotnej poprawy (0,9999 dla $T = 50$ i $T = 100$).

- Na zbiorze Breast Cancer (rys. 2b) zwiększenie T powoduje stopniowy wzrost dokładności (z 0,958 dla $T = 10$ do 0,961 dla $T = 100$) oraz zmniejszenie odchylenia standardowego (z 0,007 do 0,003), co świadczy o większej stabilności modelu.
- Na zbiorze Wine Quality (rys. 2c) zwiększenie liczby estymatorów poprawia dokładność (z 0,717 dla $T = 10$ do 0,744 dla $T = 100$) oraz zmniejsza wariancję (odchylenie standardowe spada z 0,015 do 0,008). Na zbiorze Car Evaluation (rys. 2d) dokładność stabilizuje się na poziomie 0,812, a odchylenie standardowe spada z 0,017 do 0,010.
- Stabilizacja wyników następuje w okolicy $T = 50$. Dalsze zwiększenie liczby estymatorów nie poprawia istotnie wyniku, 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 ustalone liczbę estymatorów $T = 20$.



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

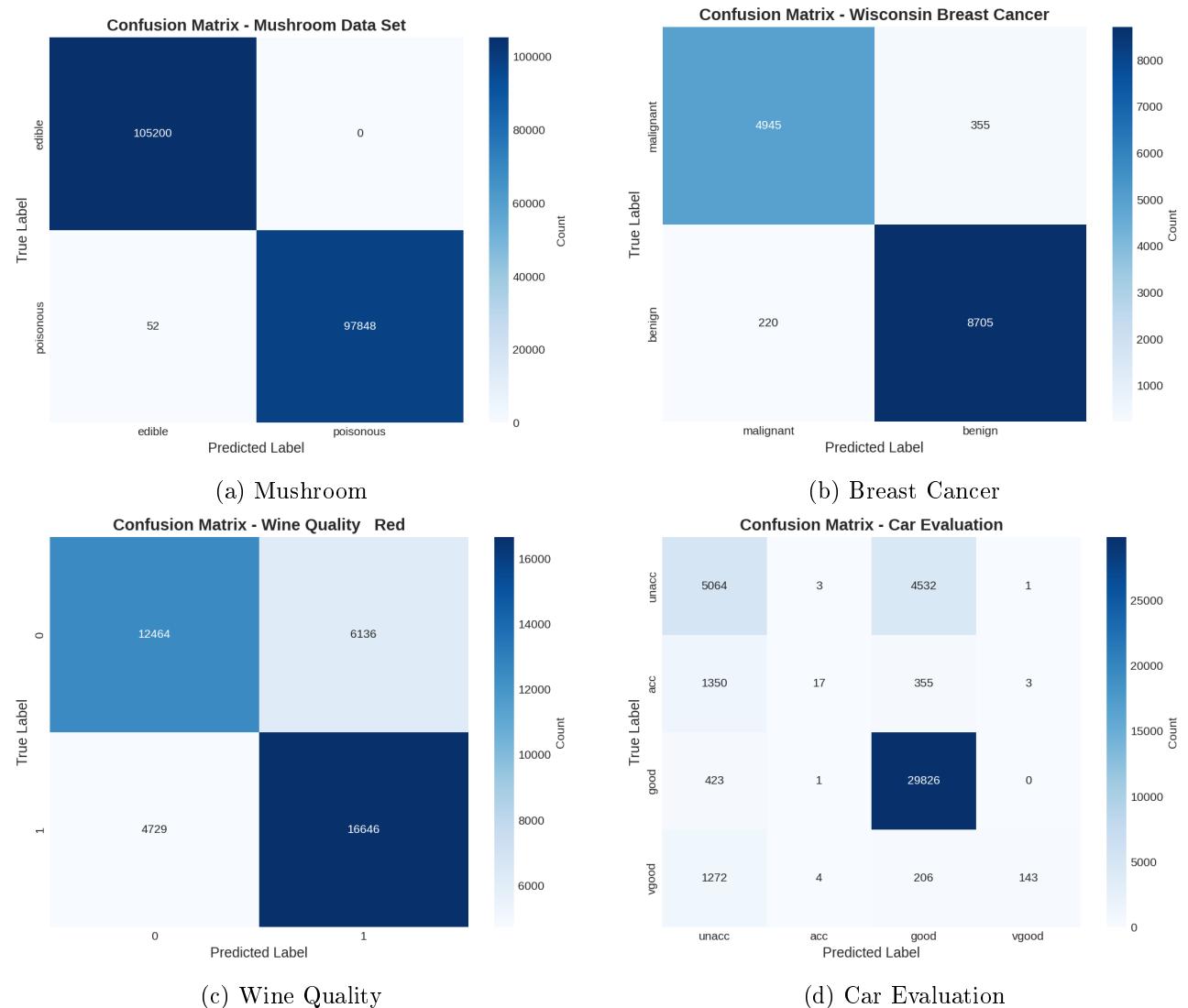
Wnioski:

- Na zbiorze Mushroom (rys. 3a) obserwujemy wyraźny wpływ parametru C - dla $C = 0,1$ dokładność wynosi 0,999, natomiast dla $C \geq 10$ osiągamy perfekcyjną klasyfikację (1,000). Zbiór ten jest liniowo separowalny, więc większe C pozwala na dokładniejsze dopasowanie hiperpłaszczyzny.
- Na zbiorze Breast Cancer (rys. 3b) zwiększenie C poprawia wyniki - od 0,952 dla $C = 0,1$ do 0,963 dla $C = 50$. Wzrost jest stopniowy i stabilny, co sugeruje, że dane są dobrze separowalne liniowo, a słaba regularyzacja (wysokie C) nie prowadzi do przeuczenia.

- Na zbiorach Wine Quality (rys. 3c) oraz Car Evaluation (rys. 3d) wpływ parametru C jest minimalny - dokładność zmienia się odpowiednio w zakresie 0,703-0,710 oraz 0,791-0,796. Wynika to z faktu, że liniowy SVM nie jest w stanie dobrze zamodelować nieliniowych zależności w tych zbiorach, niezależnie od wartości regularyzacji.
- Ogólnie, zbyt małe C (silna regularyzacja) powoduje niedopasowanie modelu, natomiast dla zbiorów nieliniowych zwiększenie C nie przynosi istotnych korzyści.

4.5 Analiza błędów - Macierze Pomyłek (Heatmapy)

Poniżej przedstawiono zagregowane macierze pomyłek (suma z 25 uruchomień) dla modelu hybrydowego ($T = 20$, $p_{svm} = 50\%$, $C = 1,0$). Pozwala to ocenić, które klasy są mylone najczęściej.



Rysunek 4: Zagregowane macierze pomyłek (heatmapy) dla $T = 20$, $p_{svm} = 50\%$, $C = 1,0$.

Wnioski:

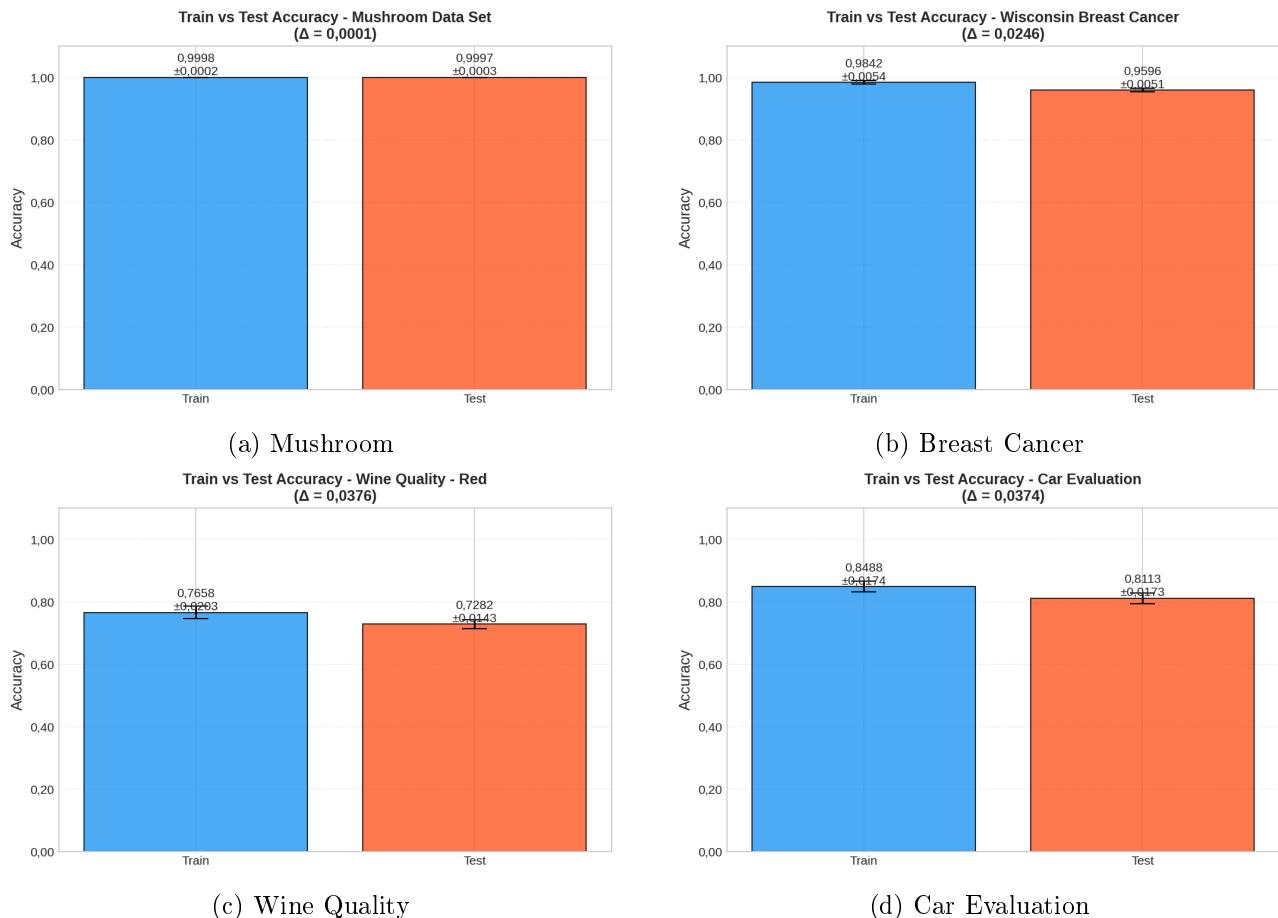
- Na zbiorze Mushroom (rys. 4a) obserwujemy niemal perfekcyjną klasyfikację - tylko 52 błędy na ponad 200 000 predykcji (zagregowanych z 25 uruchomień). Błędy są jednostronne tylko False Negative, co oznacza, że model sporadycznie klasyfikuje grzyby trujące jako jadalne, ale nigdy odwrotnie. Wynika to najprawdopodobniej z samej charakterystyki zbioru.
- Na zbiorze Breast Cancer (rys. 4b) model popełnia więcej błędów typu FP (355 przypadków benign sklasyfikowanych jako malignant) niż FN (220 przypadków malignant sklasyfikowanych

jako benign). W kontekście diagnostyki medycznej taki rozkład błędów jest akceptowalny - lepiej fałszywie zaalarmować niż przeoczyć chorobę.

- Na zbiorze Wine Quality (rys. 4c) rozkład błędów jest bardziej zrównoważony. Model popełnia 6 136 błędów typu FP (low quality sklasyfikowane jako high quality) oraz 4 729 błędów typu FN (high quality sklasyfikowane jako low quality). Większa liczba FP może wynikać z tendencji modelu do klasyfikowania granicznych przypadków jako wyższą jakość.
- Na zbiorze Car Evaluation (rys. 4d) najczęściej mylone są klasy "unacc" i "good" (4532 błędów) oraz "vgood" sklasyfikowana jako "unacc" (1272 błędów). Klasy rzadkie ("acc", "vgood") są bardzo słabo rozpoznawane - tylko 17 z 1725 przypadków "acc" i 143 z 1625 przypadków "vgood" zostało poprawnie sklasyfikowanych. Potwierdza to, że hybrydyzacja z liniowym SVM, bez wsparcia znacznym preprocessingiem danych, nie radzi sobie z niezbalansowanymi, wieloklasowymi zbiorami.

4.6 Analiza nadmiernego dopasowania (Overfitting)

W celu oceny zjawiska przeuczenia porównano dokładność na zbiorze treningowym i testowym dla modelu hybrydowego ($T = 20$, $p_{svm} = 50\%$, $C = 1,0$).



Rysunek 5: Porównanie dokładności na zbiorze treningowym i testowym (Train vs Test) dla $T = 20$, $p_{svm} = 50\%$, $C = 1,0$.

Wnioski:

- Na zbiorze Mushroom (rys. 5a) nie obserwujemy przeuczenia - dokładność treningowa (0,9998) i testowa (0,9997) są niemal identyczne (różnica $\Delta = 0,008$ p.p.). Zbiór jest łatwy do klasyfikacji i model dobrze generalizuje.

- Na zbiorze Breast Cancer (rys. 5b) występuje umiarkowane przeuczenie - dokładność treningowa wynosi 0,984, a testowa 0,960 (różnica $\Delta = 2,46$ p.p.). Jest to akceptowalny poziom, wskazujący na dobrą generalizację modelu.
- Na zbiorze Wine Quality (rys. 5c) występuje umiarkowane przeuczenie - dokładność treningowa wynosi 0,766, a testowa 0,728 (różnica $\Delta = 3,76$ p.p.). Wynik jest porównywalny z innymi zbiorami i wskazuje na typowe zachowanie modelu.
- Na zbiorze Car Evaluation (rys. 5d) obserwujemy również umiarkowane przeuczenie- dokładność treningowa wynosi 0,849, a testowa 0,811 (różnica $\Delta = 3,74$ p.p.).
- Na testowanych zbiorach model hybrydowy wykazuje umiarkowane przeuczenie. Różnice między dokładnością treningową a testową mieszczą się w zakresie 2,5-3,8 p.p., co jest akceptowalne i wskazuje na dobrą zdolność generalizacji.

5 Podsumowanie i wnioski końcowe

Zrealizowany projekt pozwolił na zbadanie właściwości hybrydowego lasu klasyfikacyjnego łączącego drzewa ID3 z klasyfikatorami SVM. Główne wnioski z badań są następujące:

1. Czego się nauczyliśmy:

- Jak dobór klasyfikatora bazowego zależy od charakteru danych: SVM (liniowy) preferuje zbiory (w przybliżeniu) liniowo separowalne, natomiast ID3 lepiej radzi sobie z nielinijnymi zależnościami oraz danymi kategorycznymi.
- Jak istotna jest różnorodność estymatorów w zespole: zastosowanie *Random Subspace Method* (domyślnie $m = \sqrt{M}$) okazało się kluczowe, aby modele w lesie podejmowały odmienne decyzje i żeby głosowanie miało sens.
- Jak interpretować wyniki eksperymentów: analiza wariancji (odchylenia standardowego) oraz macierzy pomyłek pozwala lepiej ocenić zachowanie, efektywność modelu niż sama dokładność.