Dokumentacja projektu - Las z SVM

Antoni Grajek, Natalia Pieczko

22 maja 2025

1 Temat projektu

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

2 Specyfikacja tematu

Projekt zakłada stworzenie hybrydowego klasyfikatora zespołowego, który łączy las losowy z klasyfikatorami SVM jako część modeli bazowych. Finalny klasyfikator składać się będzie z m modeli bazowych, z czego $p\%$ to SVM, a pozostałe to drzewa decyzyjne - każdy model trenowany na osobnych, bootstrapowych próbkach danych. Końcowa predykcja będzie uzyskana przez głosowanie.

Drzewa decyzyjne (ID3) tworzone będą na podstawie próbek bootstrapowych (ze zwracaniem). Dla każdego z węzłów wybierany jest atrybut oraz próg, które pozwalają maksymalizować zysk informacji. Drzewo rozbudowywane jest, aż do spełnienia kryterium stopu.

Klasyfikator SVM szukający maksymalnego marginesu rozdzielającego klasy z miękkim marginesem dopuszczając błędy w kontrolowany sposób oraz w nieliniowej wersji z wykorzystaniem funkcji jądrowych (jądro rbf).

3 Opis algorytmów

Opis algorytmu ID3

Algorytm ID3 (Iterative Dichotomiser 3) buduje drzewo decyzyjne, wybierając na każdym kroku atrybut, który maksymalizuje zysk informacyjny, co oznacza zmniejszenie entropii danych. Entropia $H(S)$ dla zbioru danych S jest obliczana wzorem:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

gdzie c to liczba klas, a p_i to proporcja instancji w S należących do klasy i. Zysk informacyjny $IG(S, A)$ dla atrybutu A to:

$$IG(S, A) = H(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} H(S_v)$$

gdzie S_v to podzbiór S, w którym atrybut A ma wartość v. Algorytm ID3 wybiera atrybut o najwyższym zysku informacyjnym, który maksymalizuje redukcję entropii zbioru.

Algorytm zaczyna od całego zbioru danych i rekurencyjnie dzieli go na podstawie atrybutu o najwyższym zysku informacyjnym, aż podzbiory będą czyste (wszystkie instancje tej samej klasy) lub nie będzie więcej atrybutów do podziału.

Przykładowe obliczenia:

Załóżmy, że mamy zbiór danych S z dwoma atrybutami A i B, klasyfikującymi na dwie klasy: 1 (4 przykłady) i 2 (6 przykładów). Obliczamy entropię $H(S)$, entropię dla podzbiorów S_A i S_B, a następnie zysk informacyjny $IG(S, A)$.

Obliczenia dla entropii całego zbioru S oraz zysku informacyjnego dla atrybutu A są następujące:

$$H(S) = -\left(0,4 \log_2 0,4 + 0,6 \log_2 0,6\right) \approx 0,971$$

$$IG(S, A) = H(S) - (0,5 \cdot 0,971 + 0,5 \cdot 0,722) = 0,1245$$

Zysk informacyjny $IG(S, A) = 0,1245$ oznacza, jak bardzo podział na podstawie atrybutu A zmniejsza niepewność w zbiorze S.

Algorithm 1 Algorytm ID3

Require: Y: zbiór klas, D: zbiór atrybutów wejściowych, $U \neq \emptyset$: zbiór par uczących

1: **if** $\forall\{x_i, y_i\} \in U \quad y_i == y$ **then**
2: **return** Liść zawierających klasę y
3: **end if**
4: **if** $|D| == 0$ **then**
5: **return** Liść zawierający najczęstszą klasę w U
6: **end if**
7: $d \leftarrow \arg\max_{d \in D} InfGain(d, U)$
8: $U_j \leftarrow \{\{x_i, y_i\} \in U : x_i[d] = d_j\}$, gdzie d_j – j-ta wartość atrybutu d
9: **return** drzewo z korzeniem d oraz krawędziami $d_j, j = 1, 2, \ldots$, prowadzącymi do drzew: ID3$(Y, D - \{d\}, U_1)$, ID3$(Y, D - \{d\}, U_2)$, $\ldots$

Opis algorytmu SVM

Support Vector Machine (SVM) to algorytm uczenia nadzorowanego, głównie do klasyfikacji, który znajduje optymalną płaszczyznę oddzielającą punkty danych różnych klas z maksymalnym marginesem. Dla danych liniowo separowalnych problem optymalizacji to:

$$\min_{\mathbf{w},b,i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$$

przy ograniczeniu:

$$y_i(\mathbf{w^T} \cdot \mathbf{x}_i - b) \geqslant 1 - \xi_i \quad \xi_i \geqslant 0$$

gdzie $\mathbf{w}$ to wektor wag, b to bias, $\mathbf{x}_i$ to wektory wejściowe, a $y_i \in \{-1, 1\}$ to etykiety klas. Margines jest maksymalizowany przez minimalizację $\|\mathbf{w}\|$, a punkty na granicach marginesu to wektory wsparcia. W przypadkach, gdy dane nie są liniowo separowalne — to znaczy, gdy nie istnieje hiperpłaszczyzna, która mogłaby skutecznie oddzielić próbki należące do różnych klas — klasyczny SVM w przestrzeni oryginalnych cech okazuje się niewystarczający. Przykładem może być zbiór danych, w którym próbki jednej klasy rozmieszczone są koncentrycznie wewnątrz okręgu, a próbki drugiej klasy otaczają je na zewnątrz. W takim przypadku granica decyzyjna musi mieć charakter nieliniowy, aby skutecznie oddzielić klasy. Rozwiązaniem tego problemu jest zastosowanie funkcji jądra, które umożliwiają niejawne odwzorowanie danych do przestrzeni o wyższej liczbie wymiarów, gdzie możliwe jest liniowe rozdzielenie próbek.

W tym projekcie używany będzie **radial basis function (RBF)** jako funkcja jądra, co umożliwia efektywne rozdzielenie klas w przypadkach nieliniowo separowalnych. Kernel RBF jest popularny w zadaniach klasyfikacyjnych, ponieważ pozwala na mapowanie danych do wyższej przestrzeni wymiarowej, gdzie stają się one separowalne liniowo.

Równanie dla funkcji jądra RBF jest następujące:

$$kernel(x, x') = \exp(-\gamma\|x - x'\|^2)$$

gdzie γ to parametr regulujący szerokość jądra.

Algorithm 2 Algorytm SVM (ogólny)

Require: U: zbiór danych uczących $\{(x_i, y_i)\}$, gdzie x_i to wektor cech, $y_i \in \{+1, -1\}$ to etykieta klasy, $i = 1, 2, \ldots, n$
Require: C: parametr regularyzacji (kontroluje kompromis między marginesem a błędami klasyfikacji)
Require: *kernel*: funkcja jądra (np. liniowa, wielomianowa, RBF), opcjonalna

Krok 1: Przygotowanie danych
if dane nie są liniowo separowalne **then**
    Przekształć dane x_i za pomocą funkcji jądra $kernel(x_i, x_j)$ ▷ Np. kernel RBF: $e^{-\gamma||x_i - x_j||^2}$
end if
Krok 2: Znajdź hiperpłaszczyznę
Zainicjalizuj wagi w (wektor normalny do hiperpłaszczyzny) i przesunięcie b ▷ Początkowo $w = 0$, $b = 0$
Znajdź w i b, które maksymalizują margines, spełniając warunki:

for każdy punkt $(x_i, y_i) \in U$ **do**
    $y_i(w \cdot x_i - b) \geqslant 1 - \xi_i$ ▷ Warunek marginesu, ξ_i to błąd dla punktu i
end for
Minimalizuj $||w||^2 + C \sum \xi_i$ ▷ Maksymalizacja marginesu z regularyzacją
Krok 3: Wybierz wektory nośne
Wektory nośne to punkty x_i, dla których $y_i(w \cdot x_i + b) = 1$
Krok 4: Klasyfikacja nowych danych
function KlasyfikujNowyPunkt(x_{new})
    Oblicz $f(x_{new}) = w \cdot x_{new} + b$
    if $f(x_{new}) \geqslant 0$ **then**
        return $+1$
    else
        return -1
    end if
end function
return model SVM: $\{w, b, wektorynone\}$

Algorytm Random Forest

Random Forest to algorytm uczenia maszynowego, który bazuje na podejściu *ensemble learning*, czyli łączeniu wyników wielu modeli w celu uzyskania lepszej dokładności i stabilności. Główną ideą Random Forest jest budowa wielu drzew decyzyjnych oraz w naszym przypadku SVM'ów i głosowanie, aby uzyskać końcową prognozę.

Proces działania algorytmu Random Forest można podzielić na kilka kluczowych etapów:

1. **Generowanie wielu drzew decyzyjnych/SVM:** Random Forest tworzy wiele klasyfikatorów, z których każdy jest trenowany na losowo wybranym podzbiorze danych. Zamiast używać całego zbioru treningowego, dla każdego klasyfikatora jest losowy podzbiór danych z powtórzeniami (tzw. *bootstrap sampling*).

2. **Losowy wybór cech dla podziału:** W procesie uczenia, zamiast rozważać wszystkie dostępne cechy, algorytm losowo wybiera tylko podzbiór cech do rozważenia w każdym węźle. Dzięki temu każdy klasyfikator jest niezależny i różni się od pozostałych, co poprawia zdolność generalizacji modelu.

3. **Głosowanie:** Po wygenerowaniu końcowa prognoza modelu jest uzyskiwana na podstawie głosowania wyników uzyskanych. W przypadku klasyfikacji, najczęściej stosuje się głosowanie większościowe, gdzie klasa, którą przewiduje najwięcej drzew, staje się wynikiem finalnym.

Przykład: Modele przewidują przynależność do klasy 1 lub 0, dany zestaw atrybutów został przyporządkowany przez pojedyncze modele znajdujące się w lesie po kolei jako klasa: $[1, 0, 0, 0, 1, 0, 1, 0]$. Random Forrest poprzez głosowanie większościowe zwraca klasę 0, która była najczęstszą klasyfikacją danego zestawu atrybutów w obrębie lasu.

Algorithm 3 Algorytm Random Forest z SVM i Drzewami Decyzyjnymi

Require: U: zbiór danych uczących $\{(x_i, y_i)\}$, gdzie x_i to wektor cech, y_i to etykieta klasy, n to liczba klasyfikatorów, D: zbiór cech

Określ proporcję klasyfikatorów SVM w lesie (np. 50% SVM, 50% drzewa)
for $i = 1$ to n **do** ▷ Generowanie klasyfikatorów
    Losuj próbkę bootstrapową U_i z U
    if i należy do grupy SVM **then**
        Zbuduj klasyfikator SVM na próbie U_i z wybranym jądrem (np. RBF)
    else
        Zbuduj drzewo decyzyjne na próbie U_i
    end if
end for
Głosowanie:
for każdy nowy przykład x_{new} **do**
    Wyniki głosowania ← pusta lista
    for każdy klasyfikator C_i w lesie **do**
        Uzyskaj przewidywanie z klasyfikatora C_i na przykładzie x_{new}
        Dodaj przewidywanie do listy **Wyniki głosowania**
    end for
    Wykonaj głosowanie większościowe na wynikach z listy **Wyniki głosowania**
    Przypisz wynik głosowania jako wynik klasyfikacji dla x_{new}
end for
return Wynik klasyfikacji

4 Plan eksperymentów

Eksperymenty mają na celu zbadanie wpływu dodania klasyfikatorów SVM do lasu losowego na jakość klasyfikacji, wrażliwość na parametry, stabilność wyników oraz czas działania.

4.1 Porównanie jakości klasyfikacji dla SVM/ID3

Dla ustalonej liczby modeli bazowych (np. $m = 50$), zbadane będą różne proporcje między SVM, a ID3. Sprawdzenie, jaka proporcja klasyfikatorów SVM daje najlepsze rezultaty.

SVM/ID3	Liczba drzew	Liczba SVM
0%	50	0
25%	12	38
50%	25	25
75%	38	12
100%	0	50

Tabela 1: Stosunek liczby SVM do liczby ID3

4.2 Wpływ liczby modeli bazowych

Dla wybranej proporcji ID3/SVM, przetestowane zostały różne wartości m - liczby modeli bazowych: m = 1, 10, 25, 50, 75, 100. Badanie ma na celu sprawdzenie jak liczba klasyfikatorów wpływa na jakość i stabilność predykcji.

4.3 Wpływ parametrów SVM

Dla klasyfikatorów SVM z miękkim marginesem zbadany został wpływ wartości parametru C (kosztu naruszenia ograniczeń klasyfikacyjnych) (1, 5, 10) na jakość wyników.

4.4 Stabilność wyników

Każda konfiguracja uruchomiona została 25 razy z różnymi losowymi ziarnami, aby móc oszacować średnią, odchylenie standardowe, a także minimum i maksimum dokładności. Tym sposobem zobrazowane zostało, czy model jest stabilny, a jego predykcje nie są przypadkowe.

4.5 Porównanie z klasycznymi modelami

Wyniki modelu hybrydowego porównane zostały z klasycznym lasem losowym. Sprawdzone zostało, czy model hybrydowy radzi sobie lepiej niż klasyczne podejścia oraz w jakich warunkach.

5 Miary jakości

Skuteczność klasyfikatorów w tym zadaniu oceniana będzie na podstawie:

- **Accuracy** (Dokładność) - określa jaki procent przykładów (pozytywnych i negatywnych) został sklasyfikowany poprawnie,
- **Confusion matrix** (Macierz pomyłek) - tabela przedstawiająca liczbę poprawnych oraz błędnych klasyfikacji dla każdej z klas. Pokazuje TP, FP, TN i FN,
- **Precision** (Precyzja) - pokazuje jaki procent przykładów, które zostały zakwalifikowane jako pozytywne należy faktycznie do tej klasy,
- **Recall** (Czułość) - pokazuje jaki procent przykładów rzeczywiście należących do danej klasy został wykryty,
- **F1-score** (Miara-F1) - średnia harmoniczna precyzji i czułości, która okazuje się szczególnie przydatna przy nierównomiernym rozkładzie klas,

- **Średnia, odchylenie standardowe oraz najlepszy i najgorszy wynik** - zostały obliczone na podstawie eksperymentów uruchomionych 25 razy w różnych konfiguracjach.

6 Wybór zbiorów danych

Na potrzeby testów zdecydowaliśmy się użyć czterech zbiorów wymienionych poniżej. Są łatwo dostępne na Kaggle, popularne, reprezentatywne i zróżnicowane pomiędzy sobą, więc na wstępie zakładamy, że będą odpowiednie dla naszego klasyfikatora. Zbiór *Mushrooms* jest dość prosty, więc posłużył nam do weryfikacji poprawności naszego rozwiązania.

Zbiór danych	Liczba klas	Liczba przykładów	Liczba cech
Mushrooms	2	8124	22
Wine Quality	6	1143	11
Breast Cancer Wisconsin	2	569	30
Crop Recommendation	22	2200	7

Tabela 2: Wybrane zbiory danych

Odnośniki do zbiorów danych na Kaggle:

- Mushroom Classification
- Wine Quality Dataset
- Breast Cancer Wisconsin Dataset
- Crop Recommendation Dataset

7 Wyniki eksperymentów

7.1 Zbiór Mushrooms

Mushrooms W przypadku zbioru Mushroom wszystkie konfiguracje modelu – niezależnie od liczby klasyfikatorów bazowych czy proporcji ID3/SVM – osiągały wyniki ekstremalnie bliskie lub równe 100% dokładności. Jest to rezultat zgodny z oczekiwaniami, ponieważ zbiór ten został uwzględniony głównie w celach orientacyjnych, aby przetestować poprawność działania implementacji. Służył jako weryfikacja, czy nasz model hybrydowy działa zgodnie z założeniami i czy jego predykcje są spójne z wynikami uzyskiwanymi przez gotowe implementacje biblioteczne, takie jak klasyczny las losowy dostępny w `sklearn` lub nasz autorski las wykorzystujący biblioteczne drzewa decyzyjne z `sklearn`.

7.2 Zbiór Wine Quality

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	51,60	19,78	22,18	24,87	2,79	44,98	56,33
10	0,25	51,90	19,95	22,30	24,84	2,69	46,29	56,33
10	0,50	54,90	21,52	23,58	26,54	3,30	50,22	61,14
10	0,75	60,05	28,53	28,25	36,05	3,36	54,15	67,25
10	1,00	62,67	33,18	32,65	35,84	3,47	56,33	71,18
25	0,00	51,35	19,66	22,07	24,83	2,78	44,98	56,33
25	0,25	51,90	19,93	22,31	25,00	2,88	44,98	56,77
25	0,50	53,76	20,95	23,16	26,72	3,69	46,72	61,57
25	0,75	61,99	28,97	28,80	37,16	2,42	56,77	65,94
25	1,00	64,59	35,11	34,18	38,71	3,30	59,39	75,11
50	0,00	51,44	19,70	22,11	24,88	2,78	44,98	56,33
50	0,25	51,84	19,90	22,29	24,98	2,89	44,54	57,21
50	0,50	53,99	21,05	23,24	27,14	3,33	46,29	59,39
50	0,75	62,93	30,05	29,64	38,21	2,34	58,52	69,43
50	1,00	65,36	35,90	35,00	39,68	3,02	60,26	71,62
75	0,00	51,53	19,73	22,15	24,90	2,76	44,98	55,90
75	0,25	51,91	19,94	22,32	25,00	2,91	44,98	56,77
75	0,50	53,38	20,72	22,96	26,27	3,44	44,98	60,70
75	0,75	63,20	30,10	29,74	37,46	2,57	58,95	69,43
75	1,00	65,34	36,51	35,55	40,19	2,96	60,70	73,80
100	0,00	51,37	19,67	22,08	24,82	2,76	44,98	55,90
100	0,25	52,10	20,02	22,41	25,08	3,06	44,98	57,64
100	0,50	53,34	20,68	22,95	25,93	3,49	46,29	60,26
100	0,75	63,48	30,41	29,98	37,89	3,09	58,52	69,87
100	1,00	65,82	36,06	35,12	39,57	2,96	59,39	74,24

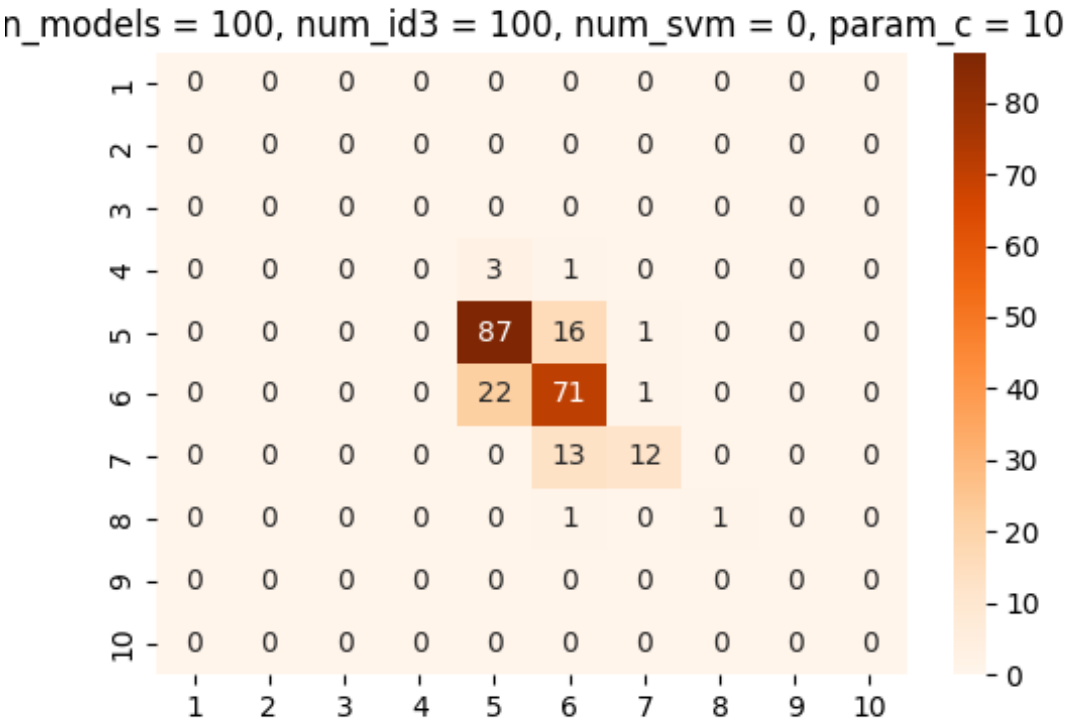
Tabela 3: Wyniki dla zbioru danych Wine Quality, C = 1

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	54,01	21,12	23,12	24,51	3,23	44,54	58,95
10	0,25	54,72	21,39	23,43	24,72	3,20	45,85	58,52
10	0,50	57,47	22,54	24,56	26,87	3,25	48,91	63,32
10	0,75	60,49	28,38	28,24	35,61	3,16	55,46	66,81
10	1,00	62,67	33,29	32,67	35,54	2,70	58,52	69,43
25	0,00	54,04	21,13	23,13	24,56	3,25	46,72	59,39
25	0,25	54,93	21,49	23,51	24,78	3,50	48,03	60,26
25	0,50	56,80	22,32	24,34	25,33	3,52	49,78	64,63
25	0,75	63,02	29,94	29,57	38,64	3,09	56,33	70,74
25	1,00	64,44	35,63	34,65	39,50	3,40	58,95	73,36
50	0,00	54,34	21,27	23,27	24,66	3,41	46,72	60,70
50	0,25	55,07	21,54	23,57	24,87	3,42	48,03	61,57
50	0,50	57,19	22,37	24,44	25,50	3,22	49,78	65,07
50	0,75	63,39	30,05	29,75	37,17	3,36	57,64	70,74
50	1,00	65,24	36,54	35,59	40,43	3,13	59,83	71,62
75	0,00	54,41	21,30	23,31	24,73	3,51	45,41	60,70
75	0,25	55,32	21,64	23,67	24,97	3,22	48,03	60,26
75	0,50	57,54	22,50	24,59	25,65	3,24	50,22	64,63
75	0,75	63,20	30,45	29,95	38,33	3,16	58,52	72,49
75	1,00	65,54	36,03	35,15	39,61	3,29	59,39	73,80
100	0,00	54,25	21,24	23,25	24,64	3,37	46,29	60,26
100	0,25	55,21	21,60	23,63	24,96	3,13	46,72	59,83
100	0,50	57,29	22,44	24,50	25,23	3,15	50,66	63,76
100	0,75	63,63	30,26	29,92	37,68	3,11	57,21	69,43
100	1,00	65,52	36,53	35,53	40,56	2,84	60,70	72,49

Tabela 4: Wyniki dla zbioru danych Wine Quality, C = 5

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	56,40	21,98	24,03	24,92	2,85	47,16	60,70
10	0,25	56,70	22,09	24,16	25,02	2,99	49,78	62,88
10	0,50	58,18	22,85	24,85	26,18	3,34	51,53	65,07
10	0,75	60,84	28,70	28,51	35,37	2,87	55,90	65,94
10	1,00	62,39	32,43	32,06	34,93	3,66	56,77	70,74
25	0,00	56,16	21,92	23,96	24,91	2,90	48,91	60,70
25	0,25	56,86	22,18	24,26	25,13	2,64	50,66	62,88
25	0,50	58,01	22,73	24,76	26,83	2,97	51,53	65,07
25	0,75	62,72	29,40	29,21	36,89	2,90	55,90	68,12
25	1,00	64,42	34,79	34,00	38,25	2,81	59,83	71,18
50	0,00	56,21	21,94	23,97	24,90	2,77	48,47	61,14
50	0,25	56,86	22,18	24,22	25,10	2,88	49,78	62,88
50	0,50	58,39	22,95	24,93	27,03	2,83	51,53	64,19
50	0,75	63,28	30,24	29,86	37,05	3,07	56,33	69,87
50	1,00	65,33	35,92	34,97	39,99	3,30	60,70	72,93
75	0,00	56,38	22,01	24,04	24,96	2,83	47,60	61,57
75	0,25	57,07	22,26	24,30	25,16	2,86	49,34	62,01
75	0,50	58,29	22,92	24,90	27,68	3,09	50,66	65,07
75	0,75	63,58	30,45	30,03	37,72	3,04	58,52	72,93
75	1,00	65,61	36,39	35,57	39,50	3,59	60,26	75,98
100	0,00	56,47	22,04	24,07	25,00	2,76	48,03	61,14
100	0,25	57,24	22,33	24,39	25,24	2,73	49,78	62,01
100	0,50	58,29	22,94	24,90	27,70	3,30	49,78	65,50
100	0,75	63,39	30,24	29,84	37,88	3,07	56,33	70,74
100	**1,00**	**65,76**	**36,89**	**35,78**	**41,30**	**3,57**	**61,14**	**74,67**

Tabela 5: Wyniki dla zbioru danych Wine Quality, C = 10



Rysunek 1: Macierz pomyłek dla najlepszego klasyfikatora

Liczba modeli	Las losowy z biblioteki	Badany las losowy ID3
1	54,18	53,21
10	64,07	62,40
25	65,43	64,43
50	65,89	65,33
75	66,44	65,61
100	66,25	65,76

Tabela 6: Porównanie dokładności naszej implementacji lasu z ID3 z implementacją biblioteczną lasu losowego sklearn na zbiorze Wine Quality (średnia z 25 uruchomień)

Wine Quality W przypadku zbioru Wine Quality zaobserwowano, że najlepsze wyniki uzyskiwały zespoły z dominacją ID3. Modele czysto oparte na ID3 (ID3/SVM = 1.0) osiągały średnią dokładność 65,76%. Zwiększanie udziału SVM prowadziło do pogorszenia wyników – nawet przy wysokich wartościach parametru C, co sugeruje, że SVM nie był dobrze dopasowany do charakterystyki tego zbioru (duża liczba klas, relatywnie mało przykładów).

Możliwą przyczyną jest fakt, że dane były wyłącznie numeryczne, a jednocześnie trudne do separacji liniowej, co stanowi wyzwanie dla SVM. Drzewa decyzyjne lepiej radziły sobie w tym kontekście. Nie zauważono oznak nadmiernego dopasowania – wyniki były stabilne między próbami.

Wniosek: metoda z dominującym ID3 dobrze sprawdza się przy zbiorach numerycznych o wysokiej liczbie klas.

7.3 Zbiór Breast Cancer Wisconsin

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	92,56	91,69	90,57	93,64	2,47	86,84	97,37
10	0,25	93,02	92,21	91,11	94,05	2,37	86,84	97,37
10	0,50	93,40	92,66	91,56	94,44	2,28	88,60	97,37
10	0,75	95,09	94,60	93,88	95,65	1,65	92,11	99,12
10	1,00	95,82	95,47	95,35	95,64	1,81	90,35	99,12
25	0,00	92,67	91,82	90,75	93,63	2,37	87,72	97,37
25	0,25	93,26	92,51	91,46	94,21	2,34	88,60	97,37
25	0,50	93,54	92,83	91,81	94,45	2,04	89,47	97,37
25	0,75	95,65	95,25	94,76	95,93	1,69	92,11	99,12
25	1,00	96,11	95,78	95,66	95,99	1,51	92,11	99,12
50	0,00	92,81	92,00	90,98	93,69	2,38	88,60	97,37
50	0,25	93,19	92,44	91,43	94,07	2,17	89,47	97,37
50	0,50	93,37	92,65	91,64	94,21	2,22	89,47	97,37
50	0,75	95,89	95,52	95,02	96,19	1,48	93,86	99,12
50	1,00	96,25	95,93	95,79	96,14	1,23	93,86	99,12
75	0,00	92,77	91,96	90,94	93,64	2,38	88,60	97,37
75	0,25	93,16	92,41	91,42	93,99	2,15	89,47	97,37
75	0,50	93,47	92,75	91,71	94,40	2,15	89,47	97,37
75	0,75	95,96	95,60	95,12	96,24	1,49	92,98	99,12
75	1,00	96,04	95,71	95,57	95,94	1,55	92,98	99,12
100	0,00	92,77	91,96	90,95	93,63	2,33	88,60	97,37
100	0,25	93,23	92,50	91,51	94,04	2,16	89,47	97,37
100	0,50	93,54	92,83	91,78	94,48	2,22	88,60	97,37
100	0,75	96,18	95,82	95,32	96,47	1,40	93,86	99,12
100	1,00	96,46	96,16	95,97	96,41	1,50	92,98	100,00

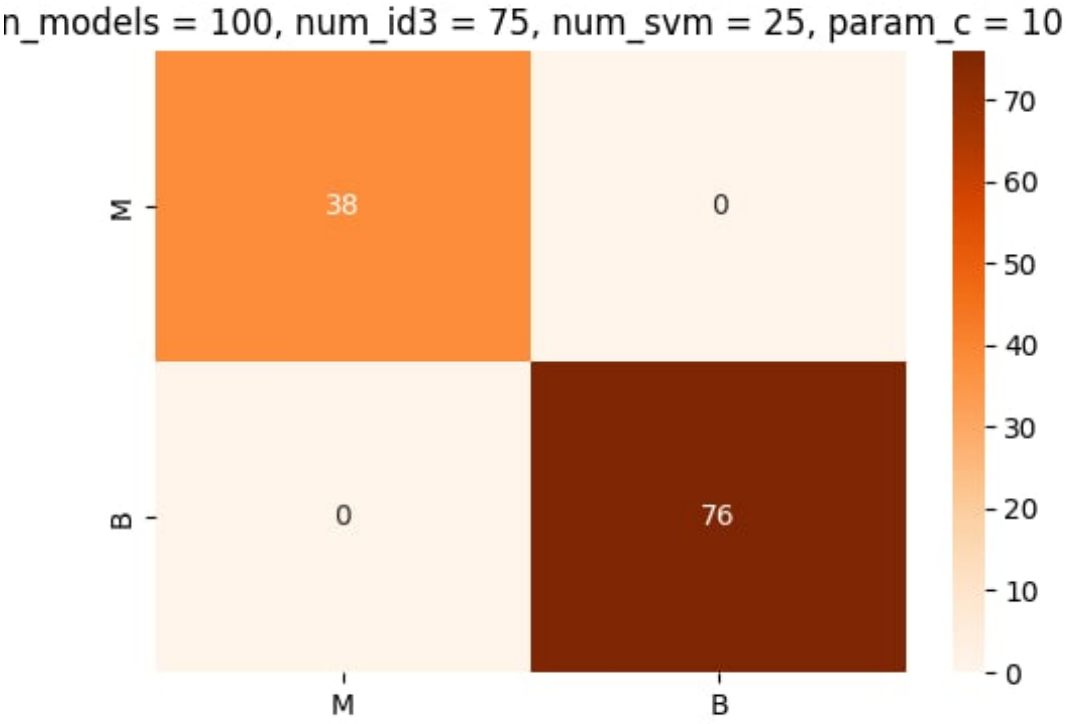
Tabela 7: Wyniki dla zbioru danych Breast Cancer Wisconsin, C = 1

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	92,49	91,68	90,77	93,11	2,25	88,60	96,49
10	0,25	92,88	92,12	91,25	93,46	2,25	88,60	96,49
10	0,50	93,89	93,24	92,33	94,64	2,23	89,47	98,25
10	0,75	95,61	95,19	94,60	96,00	1,63	92,11	98,25
10	1,00	96,04	95,68	95,35	96,12	1,68	91,23	99,12
25	0,00	92,70	91,94	91,09	93,25	2,27	88,60	96,49
25	0,25	93,12	92,40	91,58	93,66	2,21	88,60	97,37
25	0,50	93,65	93,00	92,21	94,17	2,21	89,47	97,37
25	0,75	95,82	95,45	95,01	96,03	1,61	92,11	99,12
25	1,00	95,96	95,62	95,44	95,90	1,74	92,11	99,12
50	0,00	92,81	92,05	91,21	93,36	2,13	88,60	96,49
50	0,25	93,12	92,40	91,56	93,69	2,27	88,60	97,37
50	0,50	93,75	93,11	92,28	94,35	2,20	89,47	98,25
50	0,75	96,11	95,74	95,29	96,34	1,43	93,86	99,12
50	1,00	96,35	96,04	95,89	96,27	1,21	93,86	99,12
75	0,00	92,81	92,05	91,18	93,38	2,25	88,60	97,37
75	0,25	93,16	92,44	91,62	93,70	2,23	88,60	97,37
75	0,50	93,54	92,88	92,09	94,04	2,19	89,47	97,37
75	0,75	96,46	96,15	95,80	96,61	1,48	94,74	100,00
75	1,00	96,42	96,13	95,98	96,35	1,49	93,86	99,12
100	0,00	92,74	91,98	91,15	93,25	2,23	88,60	97,37
100	0,25	93,16	92,45	91,63	93,69	2,19	88,60	97,37
100	0,50	93,75	93,11	92,32	94,31	2,18	89,47	98,25
100	0,75	96,32	95,99	95,54	96,57	1,51	92,11	99,12
100	1,00	96,39	96,08	95,89	96,34	1,22	93,86	99,12

Tabela 8: Wyniki dla zbioru danych Breast Cancer Wisconsin, C = 5

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	92,56	91,78	90,95	93,06	2,37	88,60	97,37
10	0,25	92,84	92,10	91,28	93,34	2,24	88,60	96,49
10	0,50	94,18	93,57	92,78	94,76	2,31	88,60	98,25
10	0,75	95,89	95,50	94,86	96,37	1,52	92,98	98,25
10	1,00	95,33	94,93	94,73	95,28	1,88	92,11	98,25
25	0,00	92,67	91,90	91,11	93,14	2,20	88,60	97,37
25	0,25	93,40	92,75	92,04	93,81	2,23	89,47	97,37
25	0,50	94,00	93,43	92,80	94,32	2,32	90,35	97,37
25	0,75	96,21	95,89	95,57	96,28	1,50	93,86	99,12
25	1,00	96,07	95,73	95,58	95,95	1,45	92,98	99,12
50	0,00	92,63	91,86	91,03	93,13	2,27	88,60	97,37
50	0,25	93,37	92,71	91,98	93,78	2,28	89,47	97,37
50	0,50	94,14	93,56	92,86	94,58	2,17	90,35	97,37
50	0,75	96,42	96,11	95,68	96,64	1,33	93,86	99,12
50	1,00	96,25	95,93	95,71	96,23	1,46	92,98	99,12
75	0,00	92,77	92,01	91,18	93,29	2,24	89,47	97,37
75	0,25	93,47	92,82	92,10	93,89	2,15	89,47	97,37
75	0,50	94,00	93,42	92,81	94,31	2,14	90,35	97,37
75	0,75	96,42	96,10	95,71	96,61	1,21	93,86	99,12
75	1,00	96,39	96,09	95,95	96,28	1,34	92,98	99,12
100	0,00	92,70	91,93	91,09	93,24	2,32	88,60	97,37
100	0,25	93,40	92,74	91,99	93,85	2,21	89,47	97,37
100	0,50	94,14	93,58	92,95	94,48	2,09	90,35	97,37
100	**0,75**	**96,56**	**96,26**	**95,87**	**96,75**	**1,38**	**93,86**	**100,00**
100	1,00	96,35	96,04	95,81	96,33	1,54	92,98	99,12

Tabela 9: Wyniki dla zbioru danych Breast Cancer Wisconsin, C = 10



Rysunek 2: Macierz pomyłek dla najlepszego klasyfikatora

Liczba modeli	Las losowy z biblioteki	Badany las losowy z ID3
1	91,76	92,88
10	95,47	95,34
25	96,32	96,07
50	96,32	96,25
75	96,11	96,39
100	96,15	96,35

Tabela 10: Porównanie dokładności naszej implementacji lasu z ID3 z implementacją biblioteczną lasu losowego sklearn na zbiorze Breast Cancer Wisconsin (średnia z 25 uruchomień)

Breast Cancer Wisconsin W przypadku zbioru Breast Cancer Wisconsin, wszystkie konfiguracje zespołu osiągały bardzo wysokie wyniki, co sugeruje, że dane te są łatwe do klasyfikacji niezależnie od zastosowanego modelu. Średnia dokładność dla wielu konfiguracji przekraczała 95%, a najlepsze wyniki oscylowały wokół 97–98%.

Dane w tym zbiorze są w pełni numeryczne, ale wyraźnie separowalne – większość przypadków można przypisać do jednej z dwóch klas (łagodny/złośliwy) przy pomocy prostych reguł. W takich warunkach zarówno ID3, jak i SVM radzą sobie bardzo dobrze.

Warto zauważyć, że czyste modele ID3 osiągały dobre wyniki, ale najlepsze rezultaty uzyskano przy proporcji ID3/SVM = 0,75 oraz najwyższej wartości parametru C. Dodanie SVM pozwalało na lepsze uchwycenie granic decyzyjnych w przypadku mniej jednoznacznych przypadków.

Nie zaobserwowano oznak overfittingu – wyniki były stabilne w wielu próbach.

Wniosek: Algorytm bardzo dobrze nadaje się do zbiorów z dobrze separowalnymi klasami oraz większą liczbą cech numerycznych. W takich przypadkach najlepsze wyniki osiąga się przy umiarkowanym połączeniu SVM i ID3. Wydaje się, że metoda jest odporna na drobne zmiany parametrów i dobrze generalizuje.

Dodatkowo warto zauważyć, że dla lasów z wieloma klasyfikatorami podmienienie drzew decyzyjnych bibliotecznymi SVM poprawiło czas treningu przy jednoczesnym zachowaniu wysokiego poziomu dokładności.

7.4 Zbiór Crop Recommendation

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	97,64	97,61	97,66	97,97	0,69	96,14	98,86
10	0,25	97,76	97,74	97,79	98,09	0,72	96,36	99,09
10	0,50	97,72	97,68	97,73	98,13	0,83	95,68	99,09
10	0,75	97,55	97,51	97,61	97,99	1,15	94,32	98,86
10	1,00	95,56	95,46	95,64	96,36	1,15	92,27	97,73
25	0,00	97,67	97,64	97,70	97,99	0,69	96,36	99,09
25	0,25	97,85	97,82	97,87	98,15	0,69	96,59	99,32
25	0,50	98,12	98,11	98,15	98,40	0,71	96,82	99,32
25	0,75	98,11	98,06	98,13	98,41	0,66	96,82	99,32
25	1,00	95,99	95,93	96,12	96,64	1,02	94,32	97,95
50	0,00	97,65	97,63	97,68	97,99	0,72	96,14	99,09
50	0,25	97,79	97,77	97,82	98,11	0,70	96,59	99,09
50	0,50	98,17	98,15	98,19	98,46	0,67	97,05	99,09
50	0,75	98,12	98,10	98,18	98,43	0,83	95,91	99,55
50	1,00	96,20	96,09	96,25	96,90	1,06	93,86	97,73
75	0,00	97,65	97,62	97,67	97,98	0,76	96,14	99,09
75	0,25	97,76	97,74	97,79	98,08	0,72	96,36	99,09
75	0,50	98,23	98,22	98,27	98,50	0,67	97,05	99,32
75	0,75	98,22	98,19	98,24	98,50	0,73	97,05	99,55
75	1,00	96,27	96,19	96,33	96,97	0,80	94,32	97,95

Tabela 11: Wyniki dla zbioru danych Crop Recommendation, C = 1

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	98,25	98,24	98,30	98,46	0,49	97,05	98,86
10	0,25	98,34	98,35	98,39	98,54	0,53	97,27	99,09
10	0,50	98,25	98,25	98,31	98,53	0,86	96,14	99,55
10	0,75	97,28	97,21	97,30	97,77	0,97	95,45	98,86
10	1,00	95,37	95,25	95,43	96,17	1,22	92,05	97,27
25	0,00	98,29	98,28	98,33	98,47	0,51	97,05	99,09
25	0,25	98,41	98,41	98,45	98,59	0,42	97,50	99,09
25	0,50	98,65	98,64	98,66	98,81	0,43	97,95	99,32
25	0,75	98,14	98,09	98,17	98,37	0,66	96,82	99,77
25	1,00	95,95	95,84	96,03	96,62	1,20	93,86	98,18
50	0,00	98,28	98,28	98,33	98,47	0,50	97,05	99,09
50	0,25	98,39	98,39	98,43	98,57	0,45	97,50	99,09
50	0,50	98,68	98,68	98,69	98,83	0,36	97,95	99,55
50	0,75	98,22	98,17	98,23	98,50	0,70	96,82	99,32
50	1,00	96,41	96,36	96,53	97,09	1,01	93,18	99,09
75	0,00	98,30	98,30	98,34	98,47	0,47	97,27	99,32
75	0,25	98,39	98,39	98,42	98,55	0,47	97,27	99,32
75	0,50	98,64	98,64	98,66	98,77	0,38	97,95	99,32
75	0,75	98,57	98,56	98,61	98,77	0,60	97,50	99,55
75	1,00	96,27	96,18	96,30	97,05	0,92	94,09	97,95

Tabela 12: Wyniki dla zbioru danych Crop Recommendation, C = 5

L. modeli	ID3/SVM	Accuracy	F1 Score	Recall	Precision	Std	Min	Max
10	0,00	98,35	98,35	98,40	98,51	0,49	97,27	99,09
10	0,25	98,46	98,47	98,51	98,62	0,50	97,27	99,09
10	0,50	98,45	98,45	98,49	98,63	0,46	97,27	99,55
10	0,75	97,41	97,34	97,42	97,86	1,07	95,00	99,32
10	1,00	95,39	95,32	95,49	96,15	1,27	92,05	97,73
25	0,00	98,34	98,35	98,40	98,47	0,50	97,27	99,09
25	0,25	98,43	98,43	98,48	98,56	0,51	97,27	99,32
25	0,50	98,63	98,63	98,66	98,76	0,37	97,73	99,32
25	0,75	98,16	98,14	98,21	98,40	0,85	96,36	99,55
25	1,00	96,00	95,92	96,09	96,75	1,08	92,95	97,50
50	0,00	98,29	98,30	98,35	98,42	0,49	97,27	99,09
50	0,25	98,54	98,54	98,58	98,65	0,45	97,50	99,55
50	0,50	98,69	98,70	98,73	98,79	0,32	97,95	99,32
50	0,75	98,29	98,28	98,36	98,51	0,67	97,05	99,32
50	1,00	96,06	95,96	96,12	96,81	1,02	93,86	97,50
75	0,00	98,35	98,36	98,41	98,47	0,47	97,27	99,09
75	0,25	98,53	98,54	98,58	98,63	0,35	97,95	99,32
75	**0,50**	**98,77**	**98,78**	**98,81**	**98,89**	**0,31**	**98,18**	**99,32**
75	0,75	98,32	98,28	98,35	98,57	0,77	97,05	99,55
75	1,00	96,13	96,06	96,19	96,90	1,01	93,41	97,73

Tabela 13: Wyniki dla zbioru danych Crop Recommendation, C = 10

Liczba modeli	Las losowy z biblioteki	Badany las losowy z ID3
1	91,99	89,46
10	98,12	95,39
25	98,23	96,00
50	98,21	96,06
75	98,27	96,13

Tabela 14: Porównanie dokładności naszej implementacji lasu z ID3 z implementacją biblioteczną lasu losowego sklearn na zbiorze Crop Recomendations (średnia z 25 uruchomień)

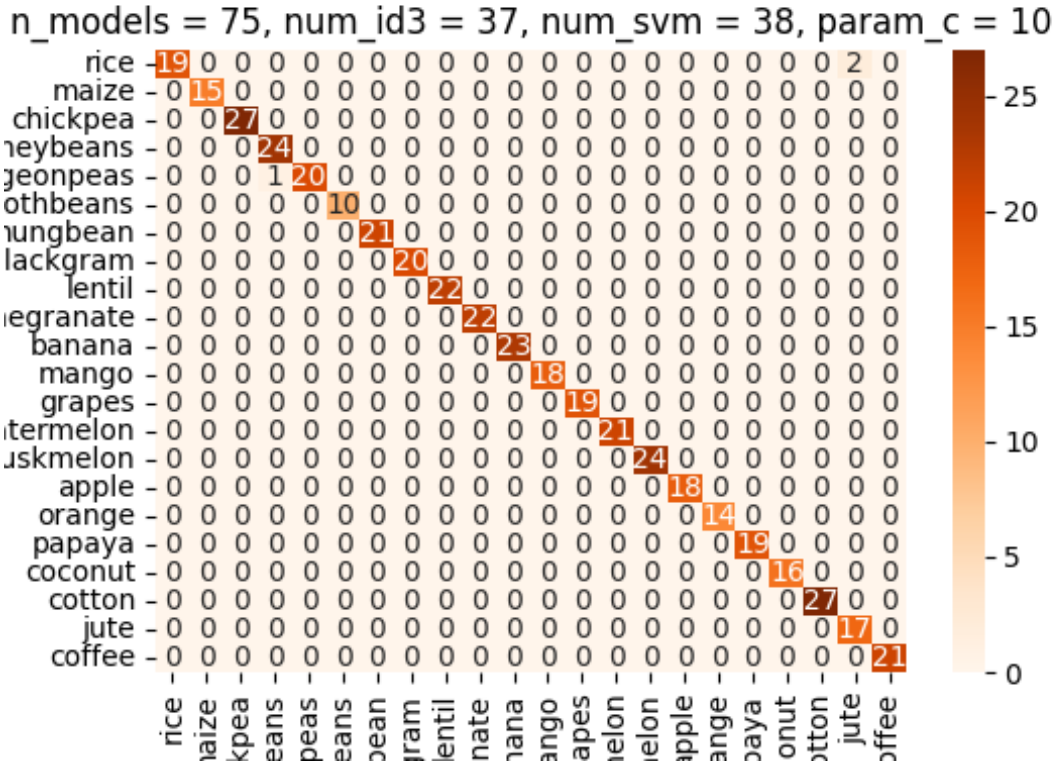
Crop Recommendation Zbiór danych Crop Recommendation okazał się najłatwiejszy do klasyfikacji spośród wszystkich testowanych (z wyłączeniem Mushrooms). Wszystkie konfiguracje osiągały bardzo wysoką dokładność, przekraczającą 95%, a najlepszy wynik – 98,77% – uzyskano dla proporcji ID3/SVM = 0,5, przy 75 modelach bazowych i $C = 10$.

Warto podkreślić, że zbiór zawiera tylko 7 cech numerycznych i aż 22 klasy, jednak ich rozdzielność jest bardzo wyraźna , co znacząco ułatwia zadanie klasyfikacyjne. Modele oparte wyłącznie na ID3 dawały gorsze wyniki – ich dokładność była niższa zarówno w ujęciu średnim, jak i maksymalnym.

To sugeruje, że dla bardziej złożonych zadań wieloklasowych, wykorzystanie SVM jako uzupełnienia ID3 może poprawiać skuteczność klasyfikatora, zwłaszcza przy odpowiednio dobranym parametrze C.

W przypadku tego zbioru nie zostały zbadane predykcje dla 100 modeli bazowych. Złożoność obliczeniowa na tym etapie mocno wzrosła, a wyniki nie poprawiały się znacząco, więc zdecydowaliśmy się zakończyć eksperymenty.

Wniosek: Metoda najlepiej sprawdza się w konfiguracji hybrydowej, co pokazuje, że łączenie różnych typów klasyfikatorów pozwala lepiej uchwycić zależności w danych. Wysoka skuteczność nawet przy 22 klasach pokazuje, że podejście to może być z powodzeniem stosowane w zadaniach rekomendacyjnych opartych na danych numerycznych. Należy jednak uważać na zbyt duży udział ID3, który może pogorszyć uogólnianie modelu.



Rysunek 3: Macierz pomyłek dla najlepszego klasyfikatora

8 Wnioski

Zgodnie z teorią, **drzewa decyzyjne ID3** są szybkie, łatwe do interpretacji i dobrze działają na danych kategorycznych lub takich, gdzie zależności między cechami a klasami są jasno zdefiniowane. Jednak ich największą wadą jest **tendencja do przeuczenia** (ang. *overfitting*), zwłaszcza przy zbyt głębokich drzewach i niewielkiej liczbie danych. Z kolei **maszyny wektorów nośnych (SVM)** są bardziej odporne na przeuczenie i dobrze radzą sobie z problemami nieliniowymi (przy odpowiednim kernelu i wartości parametru C), ale są wrażliwe na liczbę klas oraz wymagają odpowiedniego doboru hiperparametrów.

Nasze obserwacje pokazują, że teoria pokrywa się z praktyką:

- **Dla dużych zbiorów o dużej liczbie parametrów** (np. *Crop Recommendation*) wszystkie konfiguracje osiągały bardzo dobre wyniki. Szczególnie skuteczne okazały się konfiguracje hybrydowe ID3+SVM. W przypadku zbioru *Crop Recommendation*, najlepszy wynik **98,77%** uzyskano przy proporcji ID3/SVM = 0,5, 75 modelach bazowych oraz $C = 10$.
- **W zbiorze *Wine Quality*** lepsze wyniki osiągały modele oparte wyłącznie na ID3 – najlepsza konfiguracja dała średnią dokładność **65,76%** dla 100 klasyfikatorów. Sugeruje to, że dla małych zbiorów wieloklasowych prostsze modele mogą generalizować lepiej niż bardziej złożone SVM-y.
- **W zbiorze *Breast Cancer Wisconsin*** najlepsze rezultaty (ok. **97–98%** dokładności) uzyskano przy przy zastostwaniu klasycznego lasu losowego lub dodaniu niewielkich liczb klasyfikatorów SVM.

Nie odnotowano wyraźnych oznak przeuczenia – wyniki były spójne w wielu próbach, a odchylenia stndardowe zazwyczaj nie przekraczały kilku punktów procentowych. Wydaje się, że losowość (np. wybór próbek) oraz zastosowanie modelu lasu losowego skutecznie ograniczały ryzyko nadmiernego dopasowania.

Czego się nauczyliśmy:

- **Nie istnieje jeden najlepszy model** — skuteczność zależy od charakterystyki danych: liczby klas, rodzaju cech (numeryczne vs. kategoryczne), separowalności czy liczby przykładów.
- **Parametr C w SVM** nie miał tak jednoznacznego wpływu, jak można by oczekiwać. W niektórych zbiorach (np. Crop) jego większe wartości poprawiały jakość generalizacji, ale w innych (np. Wine Quality) efekt był minimalny lub negatywny.

Wniosek ogólny: Klasyfikacja za pomocą zespołów ID3 i SVM jest skuteczna i uniwersalna, jednak **najlepsze rezultaty uzyskuje się poprzez dostosowanie proporcji modeli do charakterystyki danych**. Dzięki temu projektowi nauczyliśmy się nie tylko praktycznie stosować metody zespołowe, ale również interpretować wyniki z uwzględnieniem kontekstu danych oraz unikać nadinterpretacji niewielkich różnic.