Metody Klasyfikacji

Eksploracja danych - Lista nr 3

Ksawery Józefowski, 277513

2025-05-25

Spis treści

1 Zadanie nr 1

1.1 Wprowadzenie

W tym zadaniu przeprowadzimy klasyfikację gatunków irysów z wykorzystaniem modelu regresji liniowej. Wykorzystamy zbiór danych `iris`, który zawiera informacje o czterech cechach morfologicznych trzech gatunków irysów.

1.2 Analizowane dane

Tabela 1: Statystyki opisowe zbioru Iris

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
	1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
	Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
	Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	NA
	3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	NA
	Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	NA

Zbiór danych `iris` zawiera 150 obserwacji, po 50 dla każdego z trzech gatunków: setosa, versicolor i virginica. Każda obserwacja opisana jest czterema cechami: długość płatka (`Petal.Length` - PL), szerokość płatka (`Petal.Width` - PW), długość działki kielicha (`Sepal.Length` - SL) i szerokość działki kielicha (`Sepal.Width` - SW).

1.3 Podział danych na zbiór uczący i testowy

Tabela 2: Rozkład klas w zbiorze uczącym

Var1	Freq
setosa	34
versicolor	29
virginica	37

Tabela 3: Rozkład klas w zbiorze testowym

Var1	Freq
setosa	16
versicolor	21
virginica	13

Zbiór danych dzielimy na `uczący` i `testowy`. W tym przypadku zastosowaliśmy losowy podział w proporcji 2/3 do 1/3, co jest standardowym podejściem. Ustawienie ziarna generatora liczb pseudolosowych `(set.seed(123))` zapewnia, że podział będzie reprodukowalny.

1.4 Konstrukcja klasyfikatora i wyznaczenie prognoz

Tabela 4: Podsumowanie modelu regresji liniowej

	Estimate	Std. Error	t value	p value
(Intercept)	1.293	0.245	5.275	0.000
Petal.Length	0.239	0.070	3.416	0.001
Petal.Width	0.652	0.117	5.590	0.000
Sepal.Length	-0.151	0.071	-2.143	0.035
Sepal.Width	-0.026	0.071	-0.366	0.715

W tym punkcie skonstruowaliśmy klasyfikator wykorzystujący model regresji liniowej do predykcji gatunków irysów.Warto zauważyć, że choć regresja liniowa jest tradycyjnie metodą przewidywania wartości ciągłych, w tym przypadku zastosowaliśmy ją do problemu klasyfikacji poprzez odpowiednie przekształcenie zmiennej docelowej i zaokrąglenie wyników. Jest to uproszczone podejście, które może działać dobrze dla liniowo separowalnych klas, co w przypadku zbioru iris częściowo ma miejsce (zwłaszcza dla gatunku setosa).

1.5 Ocena jakości modelu

Tabela 5: Macierz pomyłek - zbiór uczący

1	2	3
34	0	0
0	28	1
0	1	36

```
## [1] "Błąd klasyfikacji - zbiór uczący: 0.02"
```

Tabela 6: Macierz pomyłek - zbiór testowy

1	2	3
16	0	0
0	19	2
0	0	13

```
## [1] "Błąd klasyfikacji - zbiór testowy: 0.04"
```

Model regresji liniowej zbudowany na podstawie oryginalnych czterech cech irysów wykazał bardzo dobrą skuteczność klasyfikacyjną. Błąd klasyfikacji na zbiorze uczącym wyniósł zaledwie 2%, natomiast na zbiorze testowym – 4%. Wysoka dokładność predykcji sugeruje, że dane są w dużej mierze liniowo separowalne, a przyjęte założenie dotyczące użycia regresji liniowej jako klasyfikatora było w tym przypadku uzasadnione.

Analiza macierzy pomyłek potwierdza, że klasyfikator doskonale radzi sobie z rozpoznawaniem gatunku `setosa`, który został sklasyfikowany bezbłędnie we wszystkich przypadkach, zarówno w zbiorze uczącym, jak i testowym. Błędy klasyfikacji pojawiły się wyłącznie w rozróżnianiu gatunków `versicolor` i `virginica`.

1.6 Budowa modelu liniowego dla rozszerzonej przestrzeni cech

Tabela 7: Model regresji z cechami wielomianowymi

	Estimate	Std. Error	t value	p value
(Intercept)	1.944	2.066	0.941	0.349
Petal.Length	-0.387	0.741	-0.523	0.603
Petal.Width	1.865	1.280	1.456	0.149
Sepal.Length	0.498	0.894	0.557	0.579
Sepal.Width	-1.354	0.697	-1.941	0.056
PL2	0.081	0.149	0.544	0.588
PW2	0.415	0.383	1.082	0.282
SL2	-0.280	0.156	-1.789	0.077
SW2	-0.239	0.159	-1.503	0.137
PL_PW	-0.305	0.431	-0.706	0.482
PL_SW	-0.143	0.279	-0.513	0.610
PL_SL	0.156	0.266	0.586	0.559
PW_SL	0.053	0.347	0.153	0.879
PW_SW	-0.492	0.406	-1.214	0.228
SL_SW	0.639	0.302	2.116	0.037

Tabela 8: Macierz pomyłek - zbiór uczący (model z cechami wielomianowymi)

1	2	3
34	0	0
0	29	0
0	0	37

```
## [1] "Błąd klasyfikacji - zbiór uczący (model z cechami wielomianowymi): 0"
```

Tabela 9: Macierz pomyłek - zbiór testowy (model z cechami wielomianowymi))

1	2	3
16	0	0
0	20	1
0	0	13

```
## [1] "Błąd klasyfikacji - zbiór testowy (model z cechami wielomianowymi): 0.02"
```

W celu poprawy jakości klasyfikacji, model regresji został rozszerzony o składniki wielomianowe drugiego stopnia, obejmujące zarówno kwadraty cech (`PL2`, `PW2`, `SL2`, `SW2`), jak i ich iloczyny parami (`PL·PW`, `PL·SW`, `PL·SL`, `PW·SL`, `PW·SW`, `SL·SW`). Celem tej modyfikacji było uchwycenie nieliniowych zależności między zmiennymi oraz umożliwienie lepszego rozdzielenia klas w bardziej złożonej przestrzeni cech.

Nowy model wykazał jeszcze lepszą skuteczność klasyfikacji niż wersja liniowa. Na zbiorze uczącym uzyskano zerowy błąd klasyfikacji (0%), co oznacza, że wszystkie obserwacje zostały sklasyfikowane poprawnie. Również w zbiorze testowym wynik uległ poprawie — błąd spadł z 4% do 2%, co oznacza tylko jedną błędną klasyfikację spośród 50 obserwacji. Gatunek `setosa` nadal został sklasyfikowany bezbłędnie, natomiast niewielka pomyłka dotyczyła rozróżnienia pomiędzy `versicolor` a `virginica`.

Warto zauważyć, że pomimo dodania wielu dodatkowych składników, tylko jeden z nich — iloczyn `SL·SW` — okazał się statystycznie istotny na poziomie 5% (p-value = 0.037), co sugeruje, że nie wszystkie cechy wielomianowe wnoszą istotną informację. Mimo to, nawet jeśli pojedyncze składniki nie są istotne statystycznie, to łącznie mogą one poprawiać zdolność modelu do odwzorowania granic decyzyjnych.

Podsumowując, rozszerzenie modelu o składniki wielomianowe przyniosło zauważalną poprawę skuteczności klasyfikacji, zwłaszcza poprzez eliminację błędów w zbiorze uczącym oraz ograniczenie błędów w zbiorze testowym. Uzyskane wyniki wskazują, że uwzględnienie nieliniowych zależności między cechami może być wartościowe w przypadku problemów klasyfikacyjnych, nawet przy stosunkowo prostych metodach jak regresja liniowa.

2 Zadanie nr 2

2.1 Wprowadzenie

Celem niniejszej analizy jest porównanie wybranych algorytmów klasyfikacyjnych pod kątem ich skuteczności w rozróżnianiu klas na podstawie danych chemicznych win z zestawu `Wine`, dostępnego w pakiecie `HDclassif`. W analizie zostaną uwzględnione trzy algorytmy klasyfikacyjne: metoda k-najbliższych sąsiadów (k-NN), drzewa decyzyjne oraz klasyfikator Bayesa. Jak i dodatkowo użyte zostaną wersje tych algorytmów z zastosowaniem PCA i algorytm regresji. Analiza będzie obejmować wstępne zapoznanie się z danymi, ocenę dokładności poszczególnych modeli, a także weryfikację, które cechy mają największe znaczenie dla poprawnej klasyfikacji.

2.2 Wybór i zapoznanie się z danymi

Charakterystyka zbioru danych:

- Liczba przypadków (obserwacji): 178
- Liczba zmiennych (cech): 13
- Liczba klas: 3

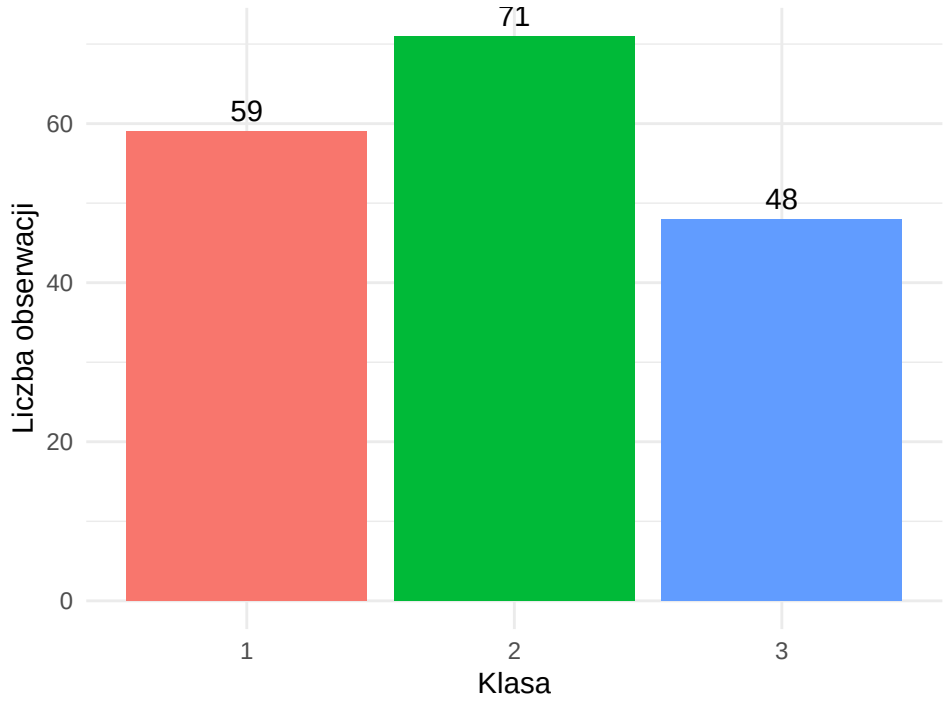
Zmienna `class` zawiera etykiety klas i została odpowiednio przekonwertowana na typ `factor`.

W zbiorze nie występują brakujące dane (`NA`).

Wszystkie zmienne mają typ `numeric`, z wyjątkiem `class`, która została skonwertowana do `factor`.

2.3 Wstępna analiza danych

2.3.1 Rozkład klas



Rysunek 1: Rozkłady klas

Tabela 10: Rozkład klas - liczebność i proporcje

	Liczebność	Proporcja
1	59	0.331
2	71	0.399
3	48	0.270

W zbiorze danych `Wine` występują 3 klasy (oznaczone jako `1`, `2` i `3`). Rozkład liczebności klas nie jest zupełnie równomierny – najliczniejsza klasa to klasa 2, która stanowi około 39.9% całego zbioru. Przypisując wszystkie obiekty do tej klasy, uzyskalibyśmy błąd klasyfikacji równy około 0.6. Oznacza to, że taki naiwny klasyfikator pomyliłby się w ok. 60% przypadków.

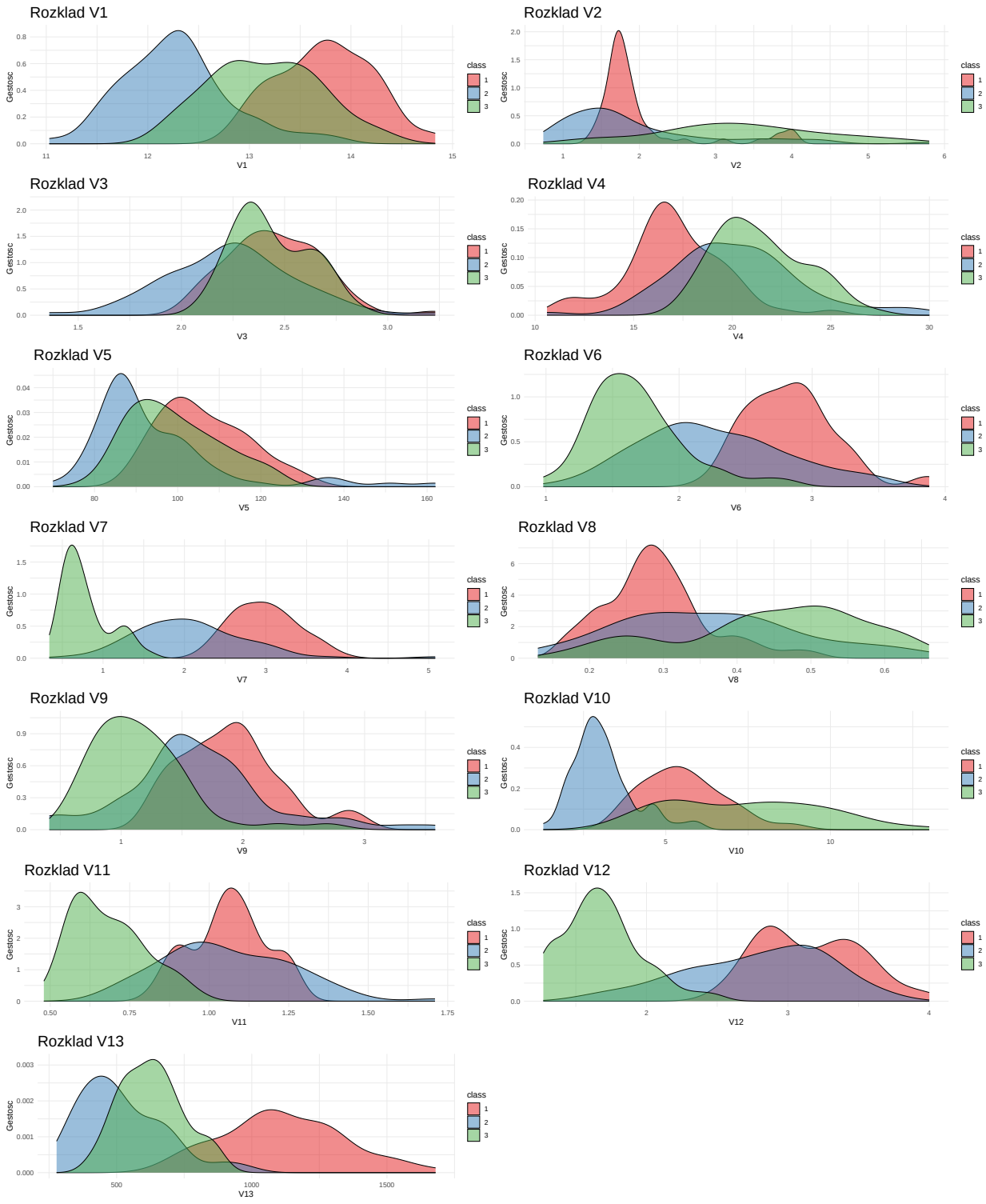
2.3.2 Analiza zmienności

Tabela 11: Wariancje cech chemicznych

	Wariancja
V13	99166.717
V5	203.989
V4	11.153
V10	5.374
V2	1.248
V7	0.998
V1	0.659
V12	0.504
V6	0.392
V9	0.328
V3	0.075
V11	0.052
V8	0.015

Zmienność cech w zbiorze `Wine` jest bardzo zróżnicowana – wariancje poszczególnych zmiennych różnią się bardzo. Oznacza to, że przed zastosowaniem niektórych algorytmów, niezbędna będzie standaryzacja danych, aby uniknąć dominacji cech o większej skali.

2.3.3 Zdolności dyskryminacyjne



Rysunek 2: Rozkłady cech V1–V13 według klasy

Z rysunku 2 można wywnioskować, że najlepiej dyskryminującymi cechami są - `V7`, `V13`, `V12`. Cechy które również dość dobrze dyskryminują klasy to - `V10`, `V1` i `V11`. Najgorzej dyskryminuje cecha `V3`.

2.4 Ocena dokładności klasyfikacji i porównanie metod

2.4.1 Po standaryzacji i bez PCA

Tabela 12: Macierz pomyłek K-NN

1	2	3
20	0	0
1	25	0
0	0	13

```
## [1] "Błąd klasyfikacji K-NN: 0.017"
```

Tabela 13: Macierz pomyłek Drzewo

1	2	3
19	0	1
3	22	1
0	0	13

```
## [1] "Błąd klasyfikacji Drzewo: 0.085"
```

Tabela 14: Macierz pomyłek Bayes

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Błąd klasyfikacji Bayes: 0"
```

Tabela 15: Macierz pomyłek Regresja

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Błąd klasyfikacji Regresja: 0"
```

2.4.2 Po standaryzacji i PCA

Tabela 16: Macierz pomyłek K-NN po PCA

1	2	3
20	0	0
1	25	0
0	0	13

```
## [1] "Błąd klasyfikacji K-NN po PCA: 0.017"
```

Tabela 17: Macierz pomyłek Drzewa po PCA

1	2	3
18	2	0
0	26	0
0	0	13

```
## [1] "Błąd klasyfikacji Drzewa po PCA: 0.034"
```

Tabela 18: Macierz pomyłek Bayes po PC

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Błąd klasyfikacji Bayes po PCA: 0"
```

Tabela 19: Macierz pomyłek Regresji po PCA

1	2	3
20	0	0
0	26	0
0	0	13

```
## [1] "Błąd klasyfikacji Regresji po PCA: 0"
```

2.4.3 Interpretacja

W przypadku klasyfikacji bez redukcji wymiarowości, najwyższą dokładność osiągnęły modele `Bayesa` i `Regresji`, które poprawnie sklasyfikowały wszystkie obserwacje w zbiorze testowym (błąd `0%`). Nieco słabsze wyniki uzyskał klasyfikator `K-NN` z błędem `1.7%`, gdzie jedna obserwacja klasy 2 została błędnie przypisana do klasy 1. Najmniej dokładne okazało się `drzewo decyzyjne`, które popełniło `8.5%` błędów, głównie poprzez mylenie obserwacji między klasami 1 i 2 oraz 1 i 3.

Po zastosowaniu `PCA` zaobserwowano ogólną poprawę skuteczności modeli. Najbardziej widoczna zmiana dotyczyła `drzewa decyzyjnego`, którego błąd zmniejszył się ponad dwukrotnie (z `8.5%` do `3.4%`). W tym przypadku model nadal mylił obserwacje między klasami 1 i 2, ale w mniejszym stopniu. Klasyfikatory `Bayesa` i `Regresji` zachowały perfekcyjną skuteczność (`0%` błędów), podobnie jak w przypadku analizy bez `PCA.` Algorytm `K-NN` utrzymał błąd na poziomie `1.7%`, co wskazuje, że redukcja wymiarowości nie wpłynęła znacząco na jego działanie.

Po wynikach można zauważyć, że modele generalizują dobrze. Nie występuje znacząca różnica w dokładności, co sugeruje brak nadmiernego dopasowania. Wyjątkiem jest `drzewo decyzyjne`, które przed zastosowaniem `PCA` wykazywało większą tendencję do błędów, co mogło wynikać z jego naturalnej skłonności do przeuczenia. Po redukcji wymiarowości jego skuteczność znacząco się poprawiła, co potwierdza, że `PCA` może być szczególnie korzystne dla tego typu modeli.

2.4.4 Wersja zaawansowana

Tabela 20: Błędy klasyfikacji (cross-validation 10-fold)

	CV_bez_PCA	CV_z_PCA
KNN	0.059	0.067
Tree	0.134	0.092
Bayes	0.042	0.042
Regresja	0.042	0.042

Porównując wyniki klasyfikatorów uzyskane metodą jednorazowego podziału danych ze skutecznością ocenioną za pomocą walidacji krzyżowej, widać, że pojedynczy podział przeszacowuje możliwości modeli. Klasyfikatory `Bayesa` i `regresji`, które wcześniej osiągały `100%` trafności, w `cross-validation` uzyskały błąd na poziomie `4.2%`. `K-NN`, wcześniej z błędem `1.7%`, osiągnął `5.9%`, a `drzewo` z `8.5%` wzrosło do `13.4%`. Po zastosowaniu `PCA`, największą poprawę odnotowano dla `drzewa decyzyjnego` (spadek błędu do `9.2%`), co potwierdza, że redukcja wymiarowości pomaga mu unikać przeuczenia. Pozostałe algorytmy nie zyskały znacząco. `Bayes` i `regresja` utrzymały poziom `4.2%`, a `k-NN` zanotował niewielkie *pogorszenie* (`6.7%`). Wnioski wskazują, że walidacja krzyżowa daje bardziej realistyczny obraz skuteczności modeli, a `PCA` najbardziej wspiera modele podatne na przeuczenie, jak drzewa decyzyjne.

2.5 Ocena na podstawie różnych parametrów

Tabela 21: Porównanie błędów klasyfikacji dla różnych konfiguracji

Metoda	Wszystkie cechy	Top 3 cechy	PCA (2 składowe)	Najlepszy parametr
K-NN	0.042	0.067	0.067	k = 5
Drzewo	0.127	0.203	0.084	depth = 2
Bayes	0.033	0.092	0.084	-
Regresja	0.034	0.092	0.067	-

Najlepsze wyniki dla metody `K-NN` osiągnięto wykorzystującej wszystkie dostępne cechy z parametrem k=5, gdzie błąd klasyfikacji wyniósł zaledwie `4.2%`. W przypadku zastosowania jedynie trzech najlepszych cech (`V7`, `V12`, `V13`) błąd wzrósł do `6.7%`, a dla dwóch głównych składowych `PCA` utrzymał się na poziomie `6.7%`, co wskazuje, że pełny zestaw cech zapewnia najwyższą dokładność tej metodzie.

Dla `drzew decyzyjnych` obserwujemy odmienny trend, podczas gdy wykorzystanie wszystkich cech dało błąd `12.7%` (przy optymalnej głębokości równej 2), redukcja wymiarowości okazała się korzystna. W przypadku `PCA` błąd spadł do `8.4%`, co sugeruje, że `drzewa` zyskują na uproszczeniu przestrzeni cech. Co ciekawe, zastosowanie jedynie trzech wybranych cech pogorszyło wyniki (błąd `20.3%`), co może wskazywać na utratę istotnych informacji dyskryminacyjnych.

Metoda `Bayesa` i `regresji` wykazały podobne wzorce, najlepsze wyniki osiągnięto przy pełnym zestawie cech (odpowiednio `3.3%` i `3.4%` błędu). W obu przypadkach redukcja cech prowadziła do pogorszenia skuteczności, przy czym `PCA` okazało się lepszym rozwiązaniem niż wybór trzech pojedynczych cech (błędy `8.4%` vs `9.2%` dla `Bayesa` i `6.7%` vs `9.2%` dla `regresji`).

2.6 Wnioski

Najlepsze wyniki klasyfikacji uzyskano przy wykorzystaniu pełnego zestawu cech predykcyjnych dla większości metod. `K-NN` osiągał optymalną skuteczność przy odpowiednio dobranym parametrze k, `regresja` i `Bayes` również najlepiej działały bez redukcji liczby cech. Dla `drzewa decyzyjnego` konieczne było ograniczenie głębokości i zastosowanie `PCA`, co znacząco poprawiało jakość klasyfikacji.

Spośród analizowanych metod, `regresja` i naiwny `Bayes` okazały się najskuteczniejsze i najbardziej stabilne. `K-NN` dawał dobre wyniki, jednak był bardziej wrażliwy na wybór parametru. `Drzewo decyzyjne` charakteryzowało się największą zmiennością, bez redukcji liczby cech wykazywało tendencję do przeuczenia.

Wybór schematu oceny miał istotny wpływ na wnioski, ponieważ ocena na podstawie pojedynczego podziału danych często przeszacowywała skuteczność modeli. Walidacja krzyżowa ujawniała rzeczywistą jakość metod, szczególnie tych podatnych na przeuczenie.