Raport 3

Eksploracja danych

Olga Foriasz 277529, Szymon Smoła 282252

2025-05-28

Spis treści

1 Klasyfikacja na bazie modelu regresji liniowej

1.1 Analizowane dane

W pierwszej części raportu będziemy korzystać z ramki danych **iris** składającej się z: * trzech klas dzielących nasze obserwacje (Setosa, Versicolor, Virginica) * czterech zmiennych objaśniających (Petal Length, Petal Width, Sepal Length, Sepal Width), które dla uproszczenia zapiszemy jako PL, PW, SL, SW * stu pięćdziesięciu rekordów bez brakujących wartości

1.2 Podział danych na zbiór uczący i testowy, konstrukcja klasyfikatora i wyznaczenie prognoz

Przeprowadzono losowy podział danych na **zbiór uczący** i **zbiór testowy** w proporcji **2:1**

```
set.seed(123)
n <- nrow(iris)
```

```r
learning.set.index <- sample(1:n,2/3*n)

# Zbiór uczący
learning.set <- iris[learning.set.index,]

# Zbiór testowy
test.set <- iris[-learning.set.index,]
```

Korzystając z funkcji **lm()** otrzymnao trzy modele regresji liniowej z danych uczących. Na ich podstawie wyznaczono prognozowane etykietki klas dla przypadków ze zbioru uczącego i testowego. Za ich pomocą otrzymano macierze błędu dla każdego ze zbiorów.

Tabela 1: Macierz pomyłek - zbiór uczący

	Rzeczywiste klasy		
	setosa	**versicolor**	**virginica**
setosa	34	0	0
versicolor	0	14	4
virginica	0	15	33

Tabela 2: Macierz pomyłek - zbiór testowy

	Rzeczywiste klasy		
	setosa	**versicolor**	**virginica**
setosa	16	0	0
versicolor	0	14	1
virginica	0	7	12

Na podstawie wyznaczonych macierzy błędu widzimy, że klasyfikacja obiektów klasy **setosa** przebiega bez problemu, a obiekty klas **versicolor** oraz **virginica** często są klasyfikowane niepoprawnie. Widoczny pomiędzy tymi klasami jest efekt maskowania. Dokładność modelu regresji liniowej dla zbioru uczącego wynosi 0.81, a dla zbioru testowego 0.84.

1.3 Budowa modelu liniowego dla rozszerzonej przestrzeni cech

Powtórzono konstrukcję modelu i ocenę jego dokładności, tym razem budując model regresji po uzupełnieniu wyjściowych cech o składniki wielomianowe stopnia 2 (tzn.: PL^2, PW^2, SL^2, SW^2, PL · PW, PL · SW, PL · SL, PW · SL, PW · SW, SL · SW). Użyto tego samego ziarna generacji losowej w celu rozważenia poprawy nowego modelu.

Tabela 3: Macierz pomyłek - zbiór uczący

	Rzeczywiste klasy		
	setosa	**versicolor**	**virginica**
setosa	34	0	0
versicolor	0	29	1
virginica	0	0	36

Tabela 4: Macierz pomyłek - zbiór testowy

	Rzeczywiste klasy		
	setosa	**versicolor**	**virginica**
setosa	16	0	0
versicolor	0	20	0
virginica	0	1	13

Nowy model jest zauważalnie dokładniejszy od poprzedniego. W zbiorze uczącym popełnił on tylko jeden błąd klasyfikując obiekt typu **virginica** jako typ **versicolor**. W zbiorze testowym podobnie zaobserwowano tylko jeden błąd (tym razem przypisano obiekt typu **versicolor** do typu **virginica**). Dokładność poprawionego modelu regresji liniowej dla **zbioru uczącego** wynosi 0.99, a dla **zbioru testowego** 0.98.

1.4 Wnioski

Prostota zbioru danych **iris** pozwala na dobre działanie modelu regresji liniowej lecz widoczny jest w nim efekt maskowania klas. Model liniowy dla rozszerzonej przestrzeni cech zwiększa dokładność wyników klasyfikacji.

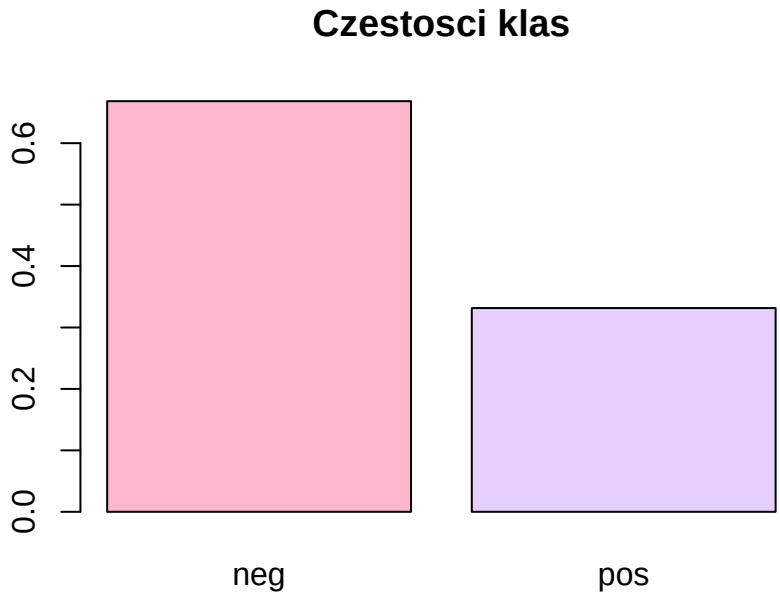
2 Porównanie metod klasyfikacji

2.1 Wybór i zapoznanie się z danymi

Nasze dane składają się z 768 obserwacji i 9 cech. Mamy jedną zmienną cukrzyca, która dzieli pacjentów na dwie grupy - chorych i zdrowych(neg/pos). Wszystkie zmienne są prawidłowych typów. W danych brakujące wartości oznaczone są jako NA. W nastepujących kolummach jest ich: glukoza - 5, cisnienie - 35, triceps - 227, insulina - 374, BMI - 11.

2.2 Analiza

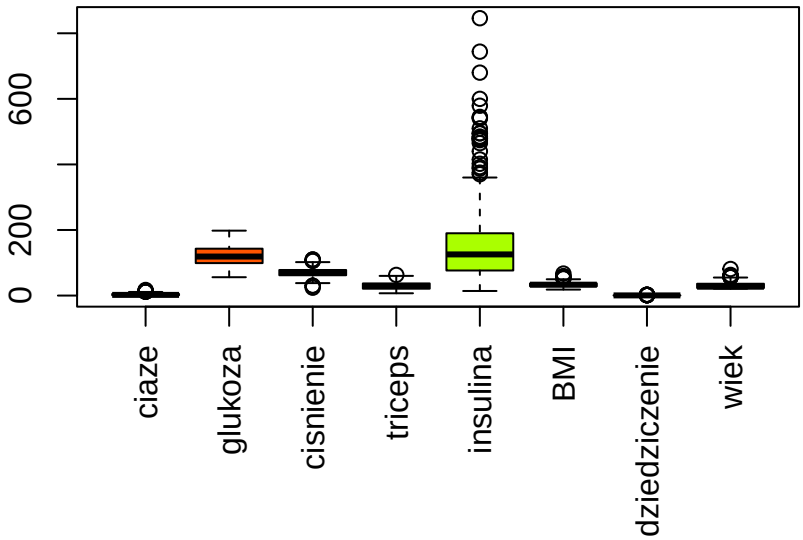
2.2.1 Wstępna analiza danych



Rysunek 1: Rozkład klas

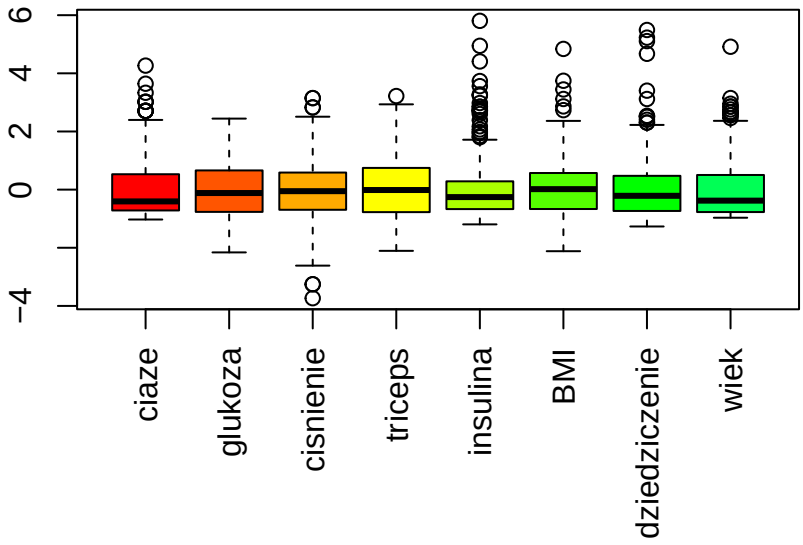
Widzimy na Rysunku 1, że rozkład klas nie jest symetryczny. Istnieje widoczna, duża dysproporcja pomiędzy klasami. Prawie dwukrotnie więcej osób należy do grupy neg (tzn. osób zdrowych). Przypisując wszystkie obiekty do jednej, najczęściej występującej klasy otrzymalibyśmy błąd 33.16%.

Na poniższym rysunku, przedstawimy jak rozkładają się zmienne. Ma to na celu sprawdzenie, czy standarzyacja w niektórych przypadkach może być konieczna.



Rysunek 2: Rozkład przed standaryzacją

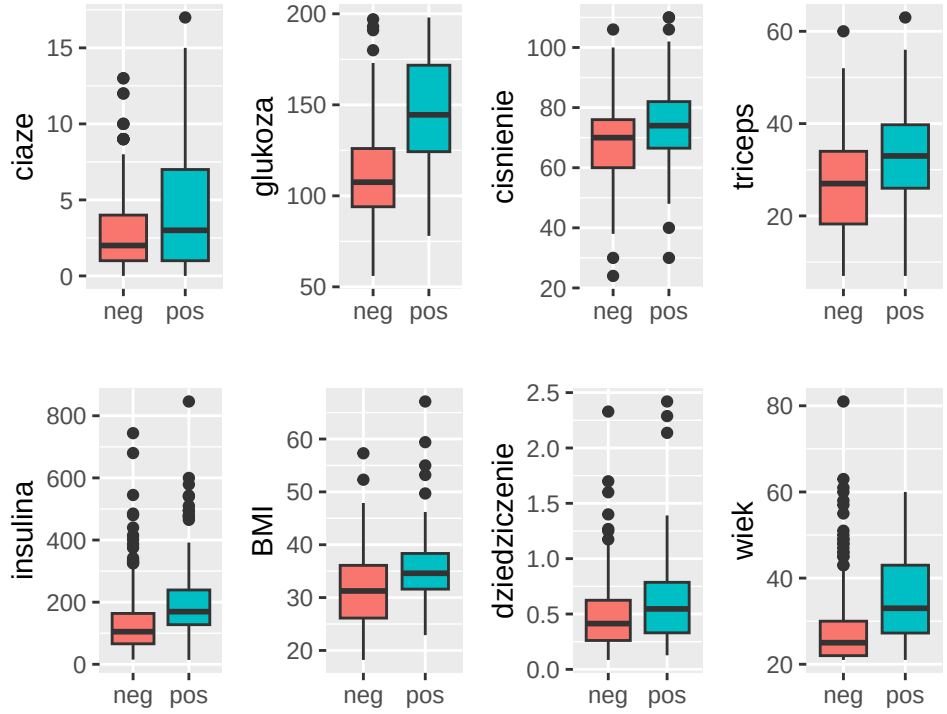
Widoczne na rysunku 2 dane cechują się istotną zmiennością poszczególnych cech. Może to oznaczać potrzebę standaryzacji, zatem zastosujemy ją dla naszych danych, dam gdzie będzie to potrzebne.



Rysunek 3: Rozkład po standaryzacji

Rysunek 3 przedstawia dane po standaryzacji.

2.2.2 Badanie zdolności dykryminacyjnej zmiennych



Rysunek 4: Zdolności dyskryminacyjne

Z Rysunku 4 możemy odczytać, które zmienne mają najlepsze zdolności dyskryminacyjne. Do tych zmiennych należą glukoza, insulina oraz wiek - co jest zgodne z oczekiwaniami. Te zmienne są najbardziej powiązane ze zmienna cukrzyca, dlatego wyniki, które otrzymaliśmy są takie, jakich się spodziewano.

2.2.3 Metoda k-najbliższych sąsiadów

Do tej metody, dane zostały zestandaryzowane, ponieważ opiera się ona na odelgłościach, gdzie zmienne o większych zmiennościach dominowałyby. Aby uniknąć takiej sytuacji zastosowana została standaryzacja.

2.2.3.1 Ocena dokładności klasyfikacji - liczba sąsiadów

Tabela 5: Procentowa macierz pomyłek - zbiór testowy (k=2)

	neg	pos
neg	79	42
pos	21	57

Tabela 6: Procentowa macierz pomyłek - zbiór uczący (k=2)

	neg	pos
neg	90	24
pos	10	76

```
## Błąd klasyfikacji na zbiorze uczącym: 0.149
```

```
## Błąd klasyfikacji na zbiorze testowym: 0.275
```

Tabela 7: Procentowa macierz pomyłek - zbiór testowy (k=5)

	neg	pos
neg	84	50
pos	16	50

Tabela 8: Procentowa macierz pomyłek - zbiór uczący (k=5)

	neg	pos
neg	94	26
pos	6	74

```
## Błąd klasyfikacji na zbiorze uczącym: 0.13
```

```
## Błąd klasyfikacji na zbiorze testowym: 0.267
```

Tabela 9: Procentowa macierz pomyłek - zbiór testowy (k=7)

	neg	pos
neg	79	51
pos	21	49

Tabela 10: Procentowa macierz pomyłek - zbiór uczący (k=7)

	neg	pos
neg	94	39

pos	6	61

```
## Błąd klasyfikacji na zbiorze uczącym: 0.161
```

```
## Błąd klasyfikacji na zbiorze testowym: 0.328
```

Korzystając z Tabel o numerach 5-10, które kolejno przedstawiają macierz pomyłek dla 2,5 oraz 7 najbliższych sąsiadów oraz wyliczonych błędach klasyfikacyjnych z rozdzieleniem na zbiór testowy oraz uczący, możemy zaobserwować, że najlepszy wynik - najmniejszy błąd na zbiorze uczącym wyszedł dla k=5 - 13%. Natomiast przy liczbie sąsiadów równej 7 wynosił więcej (16,1%) niż przy równej 2 (14,9%). Zbiór testowy także miał najmniejszy błąd klasyfikacyjny przy k=5, wynoszący 26,7%, a także analogicznie największy dla k=7 - aż 32,8%

2.2.3.2 Ocena dokładności klasyfikacji - wybrany podzbiór zmiennych Następnie porównamy dokładność klasyfikacji, wybierając tylko niektóre zmienne, przy ustalonej liczbie sąsiadów - 5. :

Zbiór składający się ze zmiennych: glukoza, wiek, BMI oraz dziedziczenie

Tabela 11: Procentowa macierz pomyłek - zbiór testowy (k=5)

	neg	pos
neg	81	45
pos	19	55

Tabela 12: Procentowa macierz pomyłek - zbiór uczący (k=5)

	neg	pos
neg	89	26
pos	11	74

```
## Błąd klasyfikacji na zbiorze uczącym: 0.157
```

```
## Błąd klasyfikacji na zbiorze testowym: 0.267
```

Zbiór składający się ze zmiennych: glukoza oraz wiek

Tabela 13: Procentowa macierz pomyłek - zbiór testowy (k=5)

	neg	pos

neg	85	48
pos	15	52

Tabela 14: Procentowa macierz pomyłek - zbiór uczący (k=5)

	neg	pos
neg	89	33
pos	11	67

```
## Błąd klasyfikacji na zbiorze uczącym: 0.184
```

```
## Błąd klasyfikacji na zbiorze testowym: 0.252
```

Możemy zatem zobaczyć, że ograniczenie się jedynie do dwóch zmiennych mających największe zdolności dyskryminujące, wcale nie poprawia wyników błędów klasyfikacyjnych. Dla zbioru 4 zmiennych błąd dla zbioru testowego i uczącego wynosił odpowiednio 26,7% oraz 15,7%, podczas gdy dla zbioru tylko dwóch zmiennych - glukozy oraz wieku - także odpowiednio 25,2% oraz 18,4%

2.2.3.3 Wnioski dla metody k-NN: Dane w użytej metodzie zostały podzielone na zbiór uczący i testowy w proporcjach 1:2, a jako miary oceny użyto macierzy pomyłek oraz błędów klasyfikacji. Na podstawie przeprowadzonych eksperymentów z metodą k-najbliższych sąsiadów (k-NN) można stwierdzić, że skuteczność klasyfikacji zależy zarówno od liczby sąsiadów (k), jak i od doboru zmiennych.

Wyniki pokazały, że wielkość błędu zależy od prawidłowego doboru liczby sąsiadów. W naszych danych najlepszym parametrem k z liczb 2, 5 i 7, okazała się wartość k = 5. Wtedy osiągnięto najmniejszy błąd dla zbioru testowego - 26,7%, jak i uczącego - 13%.

Dodatkowo sprawdzono skuteczność klasyfikacji przy użyciu tylko czterech najbardziej istotnych zmiennych (glukoza, wiek, BMI, dziedziczenie) oraz następnie jedynie dwóch (glukoza, wiek). Wyniki były najlepsze dla zbioru uczącego, gdy pod uwagę wzięto cały zbiór (13%), natomiast błąd dla zbioru testowego wyszedł najmniejszy dla zbioru dwóch zmiennych - 19,1%.

2.2.4 Naiwny Klasyfikator Bayesowski

Najpierw wyliczyone zostają błędy klasyfikacyjne oraz macierze pomłek dla całego zbioru danych:

```
## Błąd klasyfikacji - zbiór uczący: 0.234
```

```
## Błąd klasyfikacji - zbiór testowy: 0.214
```

Tabela 15: Procentowa macierz pomyłek - zbiór uczący

	neg	pos
neg	82	33
pos	18	67

Tabela 16: Procentowa macierz pomyłek - zbiór testowy

	neg	pos
neg	88	42
pos	12	57

Następnie dla zbioru składającego się ze zmiennych: glukoza, wiek, BMI, dziedziczenie:

```
## Błąd klasyfikacji - zbiór uczący: 0.215
```

```
## Błąd klasyfikacji - zbiór testowy: 0.198
```

Tabela 17: Procentowa macierz pomyłek - zbiór uczący

	neg	pos
neg	86	36
pos	14	64

Tabela 18: Procentowa macierz pomyłek - zbiór testowy

	neg	pos
neg	88	38
pos	12	62

A także dla zbioru złożonego jedynie ze zmiennych glukoza oraz wiek:

```
## Błąd klasyfikacji - zbiór uczący: 0.218
```

```
## Błąd klasyfikacji - zbiór testowy: 0.214
```

Tabela 19: Procentowa macierz pomyłek - zbiór uczący

	neg	pos
neg	86	36
pos	14	64

Tabela 20: Procentowa macierz pomyłek - zbiór testowy

	neg	pos
neg	88	38
pos	12	62

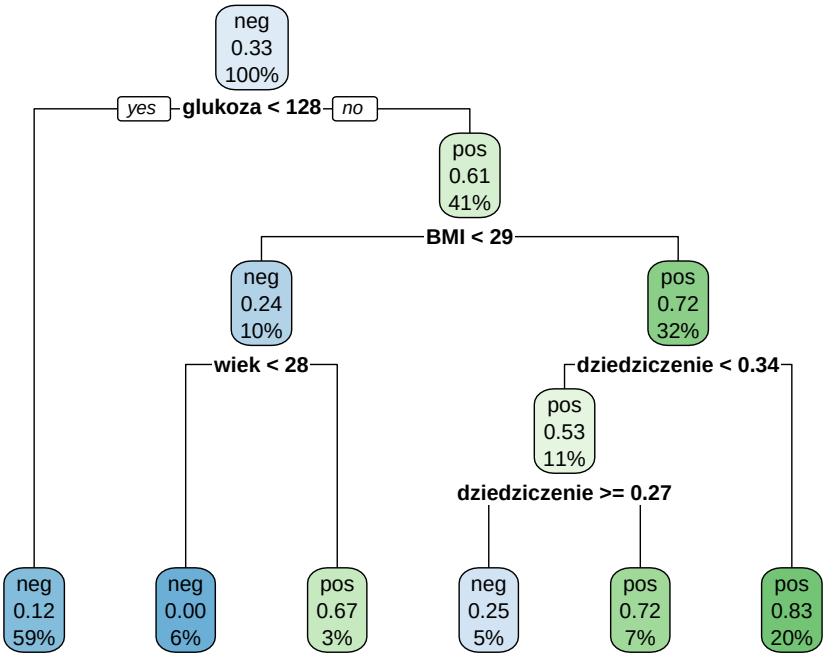
2.2.4.1 Wnioski dla Naiwnego Klasyfikatora Bayesowskiego: Na podstawie przeprowadzonych testów można stwierdzić, że najlepsze wyniki klasyfikacji osiąga model wykorzystujący 4 zmienne – błąd klasyfikacji na zbiorze testowym wyniósł tylko 19,8%. Także ten model miał najniższy błąd na zbiorze uczącym, wyniósł on 21,5%.

Najprostszy model, który używa tylko dwóch zmiennych (glukozy i insuliny), osiągnął błąd testowy 21,4% i błąd uczący 21,8%. Choć jego wyniki były stosunkowo stabilne, to dokładność była niższa niż w przypadku modelu z zestawem 4 zmiennych mających najlepsze zdolności dyskryminacyjne.

Natomiast model używający wszystkich zmiennych miał największy błąd dla zbioru uczącego - aż 23,4%, natomiast dla zbioru testowego wynosił tyle samo co dla zbioru dwóch zmiennych - 21,4%.

2.2.5 Drzewa klasyfikacyjne

2.2.5.1 Drzewa dla różnych zbiorów Najpierw rozważmy drzewo klasyfikacyjne dla wszystkich zmiennych:



Rysunek 5: Drzewo klasyfikacyjne - wszystkie zmienne

Tabela 21: Macierz pomyłek (zbiór uczący)

	neg	pos
neg	159	22
pos	17	63

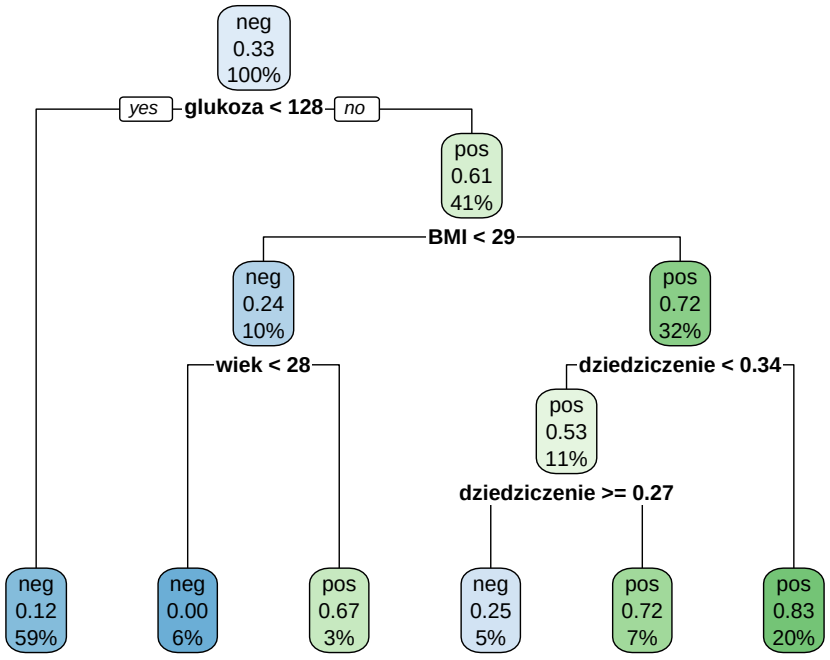
Tabela 22: Macierz pomyłek (zbiór testowy)

	neg	pos
neg	72	20
pos	14	25

```
## Błąd klasyfikacji - zbiór uczący: 0.149
```

```
## Błąd klasyfikacji - zbiór testowy: 0.26
```

Oraz drzewo dla wybranych 4 zmiennych o najlepszych zdolnościach dyskryminacyjnych: glukoza, wiek, BMI oraz dziedziczenie:



Rysunek 6: Drzewo klasyfikacyjne - wybrane zmienne

Tabela 23: Macierz pomyłek (zbiór uczący)

	neg	pos

neg	159	22
pos	17	63

Tabela 24: Macierz pomyłek (zbiór testowy)

	neg	pos
neg	72	20
pos	14	25

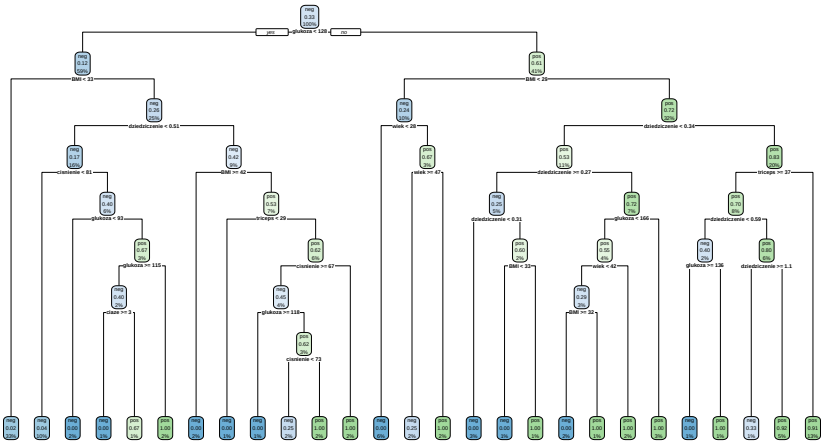
```
## Błąd klasyfikacji - zbiór uczący: 0.149
```

```
## Błąd klasyfikacji - zbiór testowy: 0.26
```

Na podstawie wyników dwóch pierwszych drzew klasyfikacyjnych można zauważyć, że oba modele osiągnęły podobną skuteczność.

Pierwsze drzewo, zbudowane na podstawie wszystkich zmiennych miało niski błąd na zbiorze uczącym - 15%, ale wyższy na zbiorze testowym - 26%. Oznacza to, że dobrze dopasowało się do dostarczonych danych, ale trochę gorzej radziło sobie z nowymi danymi. W analizowanym przypadku wyniki dla zbioru składającego się z wybranych zmiennych są analogiczne do tych, dla całego zbioru.

2.2.5.2 Drzewa klasyfikacyjne - zmiany parametrów Rozważmy najpierw drzewa klasyfikacyjne dla całego zbioru danych z parametrami tymi, co poprzednio:



Rysunek 7: Drzewo klasyfikacyjne- cp=.01

Tabela 25: Macierz pomyłek (zbiór uczący)

	neg	pos
neg	171	6

pos	5	79

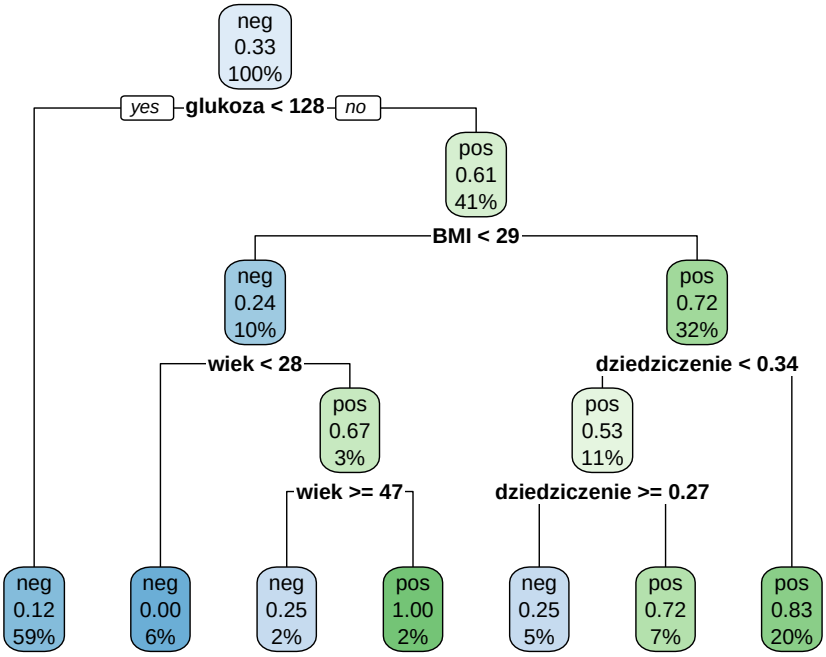
Tabela 26: Macierz pomyłek (zbiór testowy)

	neg	pos
neg	70	28
pos	16	17

```
## Błąd klasyfikacji - zbiór uczący: 0.042
```

```
## Błąd klasyfikacji - zbiór testowy: 0.336
```

Następnie spójrzmy na drzewa ze zmienionymi parametrami (m.in. cp, maxdepth oraz minsplit):



Rysunek 8: Drzewo klasyfikacyjne - cp=.02

Tabela 27: Macierz pomyłek (zbiór uczący)

	neg	pos
neg	162	23
pos	14	62

Tabela 28: Macierz pomyłek (zbiór testowy)

	neg	pos
neg	74	22
pos	12	23

```
## Błąd klasyfikacji - zbiór uczący: 0.142
```

```
## Błąd klasyfikacji - zbiór testowy: 0.26
```

Na podstawie powyższych drzew klasyfikacyjnych można zauważyć, że oba modele osiągnęły podobną skuteczność. Po zmianie parametrów takich jak cp, minsplit oraz maxdepth widać, że przy dokładniejszym i bardziej rozbudowanym drzewie, znacznie maleje błąd klasyfikacji dla zbioru uczącego. W rozbudowanym drzewie wynoisł raptem 4,2%, podczas gdy dla mniej dokładnego aż 14,2%.

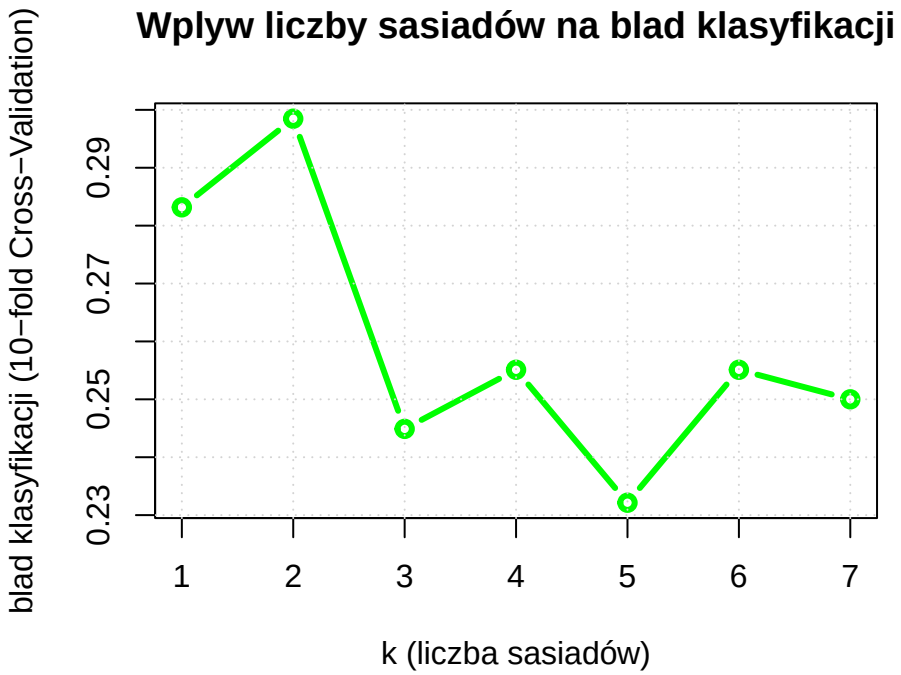
Jendak błąd dla zbioru testowego przy bardziej rozbudowanym drzewie wzrósł z 26% do 33,6%.

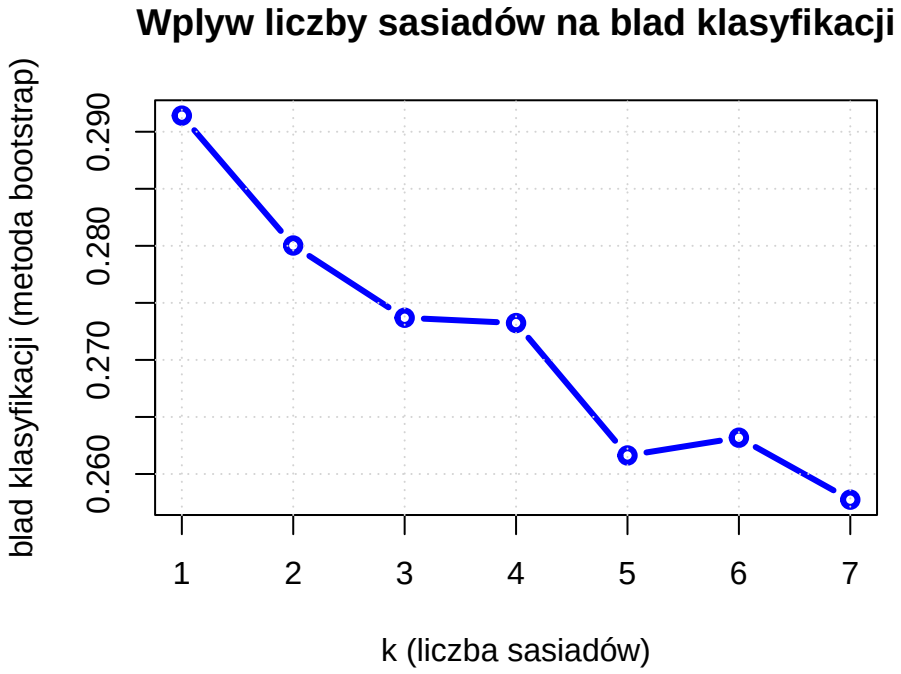
2.2.5.3 Wnioski dla drzew klasyfikacyjnych: Na podstawie analizy dwóch pierwszych drzew (rysuki 5, 6) klasyfikacyjnych można stwierdzić, że oba modele osiągnęły podobną skuteczność – błąd klasyfikacji na zbiorze testowym w obu przypadkach wynosił 26%. Pierwsze drzewo oparto na wszystkich zmiennych i uzyskano niski błąd na zbiorze uczącym - 14,9%. Drugie drzewo zbudowano jedynie na czterech cechach o najlepszych zdolnościach dyskryminacyjnych. Nie poprawiło to jednak błądu uczącego, który wynosił także 14,9%. Pokazuje to, że uproszczenie modelu nie pogarsza jego jakości.

Dodatkowo, zmiana parametrów drzewa (takich jak cp, minsplit i maxdepth) pozwoliła uzyskać lepsze dopasowanie do danych uczących – im bardziej rozbudowane drzewo, tym mniejszy błąd dla danych uczących. Jednak wpływ na dane testowe był znacząco gorszy – błąd zmienił się z 26% na 33,6%.

2.3 Porównanie metod

2.3.1 Metoda kNN





Zaawansowane metody sprawdzania, takie jak 10-krotna walidacja krzyżowa i metoda bootstrap, dały podobne wyniki jak podział danych na część uczącą i testową. W walidacji krzyżowej, najlepsze wyniki dla algorytmu k-NN osiągnięto przy liczbie sąsiadów równej 5. W przypadku metody bootstrap różnica w błędzie między k = 5 a k = 7 była bardzo mała, więc oba te ustawienia dawały podobne rezultaty.

2.3.2 Metoda - Naiwny Klasyfikator Bayesowski

```
## Błąd klasyfikacji Naive Bayes (10-fold CV): 0.227
```

```
## Błąd klasyfikacji Naive Bayes (Bootstrap): 0.235
```

W przypadku Naiwnego Klasyfikatora Bayesowskiego najlepszy wynik również uzyskano przy zastosowaniu najprostszego podejścia – błąd klasyfikacji wynosił wtedy 21,4%. Przy wykorzystaniu bardziej zaawansowanych metod oceny dokładności, takich jak 10-fold Cross-Validation i bootstrap, błędy były nieco wyższe i wynosiły odpowiednio 22,7% oraz 23,5%. Pokazuje to, że choć metody zaawansowane dają bardziej wiarygodną ocenę, w tym przypadku najprostszy schemat podziału danych przyniósł najlepszy rezultat.

2.3.3 Metoda - drzewa klasyfikacyjne

```
## Błąd klasyfikacji - Drzewo decyzyjne (10-fold CV): 0.245
```

```
## Błąd klasyfikacji - Drzewo decyzyjne (Bootstrap): 0.255
```

Obie zaawansowane metody dały błędy klasyfikacyjne na podobnym poziomie – i były one niższe niż przy prostszym podejściu. Przy jednokrotnym podziale zbioru, nawet przy dobrze dobranych parametrach, błąd wynosił 26%. Dzięki zastosowaniu bardziej zaawansowanych metod oceny, błąd klasyfikacji spadł: w przypadku 10-fold Cross-Validation do 24,5%, a przy metodzie bootstrap do 25,5%. Pokazuje to, że bardziej wiarygodne metody oceny pozwalają na uzyskanie dokładniejszych szacunków skuteczności modelu.

2.4 Wnioski końcowe

- Dla metody k-NN:

Korzystając z zaawansowanych metod oceny dokładności, zauważono, że wraz ze wzrostem liczby sąsiadów (k) błąd klasyfikacji się zmniejsza. Przy podziale danych tylko raz na zbiór uczący i testowy, najlepsze wyniki uzyskano, gdy użyto dwóch zmiennych o najwyższej zdolności rozróżniania klas. Natomiast w przypadku zaawansowanych metod, takich jak walidacja krzyżowa, minimalny błąd klasyfikacji był niższy niż przy jednokrotnym podziale danych. Mimo to, zależność między liczbą sąsiadów a błędem klasyfikacji pozostała podobna

- Dla Naiwnego Klasyfikatora Bayesowskiego:

Przy jednorazowym podziale danych na zbiór uczący i testowy, błąd klasyfikacji wyniósł 21,4%. Dla bardziej zaawansowanych metod wynik był nieco gorszy – w metodzie 10-fold Cross-Validation błąd wyniósł 22,7%, a przy użyciu metody bootstrap 23,5%. Oznacza to, że w tym przypadku najlepszy rezultat uzyskano przy najprostszym podejściu. Natomiast analizując różne zestawy cech, najlepsze wyniki osiągnięto, gdy do klasyfikacji użyto czterech zmiennych: glukozy, wieku, BMI oraz dziedziczenie.

- Dla drzew klasyfikacyjnych:

W przypadku drzew klasyfikacyjnych, najprostsze podejście – czyli jednorazowy podział na zbiór uczący i testowy – dało najsłabsze wyniki. Nawet po zmianie parametrów, najniższy błąd

klasyfikacji wyniósł 26%. Zastosowanie zaawansowanych metod przyniosło lepsze rezultaty: przy 10-fold Cross-Validation błąd spadł do 24,5%, a przy metodzie bootstrap wyniósł 25,5%

Spośród analizowanych metod, najlepsze wyniki dał algorytm k-NN – szczególnie przy liczbie sąsiadów k = 5. Najniższy błąd klasyfikacji (21,4%) uzyskano przy prostym podziale danych. Nieco gorsze, ale porównywalne wyniki dawały metody 10-fold CV i bootstrap. Na drugim miejscu pod względem skuteczności był klasyfikator Naive Bayes, z błędem rzędu 22–23%. Najsłabiej wypadły drzewa decyzyjne – ich błąd klasyfikacji był najwyższy, nawet po optymalizacji parametrów, i wynosił od 24,5% do 26%.

W porównaniu trzech metod klasyfikacyjnych najlepsze wyniki uzyskano dla Naiwnego Klasyfikatora Bayesowskiego – błąd klasyfikacji na zbiorze testowym wynosił 21,4%, a przy metodach zaawansowanych (10-fold CV i bootstrap) odpowiednio 22,7% i 23,5%. Najlepsze rezultaty osiągnięto, gdy użyto czterech zmiennych: glukozy, wieku, BMI i dziedziczenia.

Metoda k-NN dała dobre wyniki przy liczbie sąsiadów równej 5 – błąd testowy wynosił wtedy 26,7%. Zarówno CV, jak i bootstrap potwierdziły, że k = 5 to optymalna wartość.

Drzewa decyzyjne wypadły najsłabiej – błąd na zbiorze testowym wynosił co najmniej 26%, a przy bardziej rozbudowanym modelu nawet 33,6%.

W przypadku analizowanych danych metoda Naiwnego Bayesa daje najlepsze wyniki, jednak różnice w otrzymywanych błędach są niewielkie. Natomiast drzewa klasyfikacyjne oraz metoda k-Najbliższych Sąsiadów dają bardzo zbliżone, niemal identyczne błędy klasyfikacyjne dla zbiorów testowych przy jednokrotnym podziale danych.

Wybór schematu oceny dokładności miał wpływ na uzyskane wyniki, jednak nie zmienił ogólnych wniosków dotyczących skuteczności porównywanych metod klasyfikacyjnych. Niezależnie od zastosowanego podejścia – czy był to jednokrotny podział danych, 10-fold Cross-Validation czy metoda bootstrap – najlepsze wyniki uzyskano dla Naiwnego Klasyfikatora Bayesowskiego. Metoda k-Najbliższych Sąsiadów dawała nieco gorsze rezultaty, a najniżej oceniono drzewa klasyfikacyjne. Choć błędy klasyfikacji różniły się nieznacznie w zależności od wybranej metody oceny (np. były nieco wyższe w przypadku cross-validation i bootstrapu niż przy prostym podziale), kolejność skuteczności metod pozostała taka sama. Oznacza to, że wybór schematu oceny miał wpływ na dokładne wartości błędów, ale nie wpłynął istotnie na końcowe wnioski dotyczące porównania metod.