Prognozowanie przeżycia u pacjentów z niewydolnością serca

Klasyfikacja binarna

Agnieszka Olechnowicz (211023), Patrycja Pobuta (211028),

Patrycja Król (230077), Radosław Polak (229500)

Spis treści

1. Streszczenie

Praca jest poświęcona predykcji przeżycia pacjentów z niewydolnością serca na podstawie dodatkowych charakterystyk, które mogą mieć wpływ na ich śmiertelność w tym zakresie. Do predykcji zbudowano cztery klasyfikatory binarne – Naiwny Klasyfikator Bayesa, Drzewo Decyzyjne, klasyfikator na podstawie Regresji Logistycznej oraz model hybrydowy, skonstruowany na podstawie pozostałych trzech. Otrzymane klasyfikatory poddano ocenie poprzez zastosowanie ich na danych testowych oraz nowo wygenerowanych. Ich jakość zmierzono przy pomocy macierzy trafień oraz mierników Precyzji, NPV, Specyficzności, Wrażliwości, Dokładności oraz F1-score. Ponadto wyniki predykcji przedstawiono także na krzywej ROC. Dokonano również porównania między uzyskanymi klasyfikatorami zestawiając ich krzywe ROC na jednym wykresie.

2. Słowa kluczowe

Niewydolność serca, choroby, zawał serca, metoda drzew klasyfikacyjnych, metoda Naiwnego Klasyfikatora Bayesa, model hybrydowy, klasyfikacja binarna, regresja logistyczna, prognozowanie przeżycia

3. Wprowadzenie

Niewydolność serca to stan, w którym serce nie jest w stanie zapewnić wystarczającego przepływu krwi zgodnie z zapotrzebowaniem organizmu. Najczęstszymi przyczynami takiego stanu jest zawał serca, nadciśnienie, choroba niedokrwienna serca, choroby zastawkowe, kardiomiopatie czy nadmierne spożycie alkoholu[1]. W 2022 roku problem ten dotknął około 64 miliony ludzi na świecie, przede wszystkim wśród osób starszych, które są szczególnie narażone na ten stan.

4. Cel i zakres badania

Celem badania jest prognoza przeżycia pacjentów z niewydolnością serca na podstawie dodatkowych charakterystyk, które mogą mieć wpływ na rozwój choroby. Wśród nich znajdują się informacje o chorobach współistniejących (lub ich braku), wyniki badań medycznych (np. poziom enzymu CPK czy frakcja wyrzutowa EF) oraz dane dotyczące wieku i płci osób, a także tego, czy palą papierosy. Do wykonania prognozy zbudowane zostały cztery różne klasyfikatory binarne – Naiwny Klasyfikator Bayesa, Drzewo Decyzyjne, klasyfikator na podstawie regresji logistycznej oraz model hybrydowy oparty na trzech pozostałych, a następnie sprawdzona została ich jakość na danych testowych oraz nowo wygenerowanych.

Dane wykorzystane w badaniu to zbiór 5000 obserwacji dotyczących pacjentów z niewydolnością serca, opisanych przez 13 cech klinicznych, pochodzące z platformy Kaggle[2]. Zbiór ten został sztucznie wygenerowany na podstawie oryginalnego zbioru zawierającego rekordy 299 pacjentów zebrane w Faisalabad Institute of Cardiology i Allied Hospital w

Faisalabad (Punjab, Pakistan) od kwietnia do grudnia 2015 roku z badania Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Ali Raza, M. (2017).[3]

5. Przegląd literatury

Oryginalny zbiór, na podstawie którego wygenerowane zostały użyte dane, jest szeroko używany w badaniach w kontekście klasyfikacji. Poniżej przytoczono kilka z nich.

Już Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M. i Ali Raza, M. (2017)[3] w artykule przedstawiającym zebrane dane stworzyli model śmiertelności przy użyciu regresji Coxa. Wykorzystali także takie narzędzia jak wykres Kaplana Meiera, reszty Martingale'a, nonogramy, krzywą ROC czy bootstrapping. Stwierdzili, że wiek, dysfunkcja nerek (stężenie kreatyniny w surowicy wyższe niż prawidłowe 1,5), wysokie ciśnienie krwi, niższe wartości frakcji wyrzutowej i anemia są istotnymi czynnikami ryzyka śmiertelności wśród pacjentów z niewydolnością serca.

Chicco, D. i Jurman, G. (2020)[4] w swojej pracy skupili się na budowie różnych klasyfikatorów pozwalających przewidzieć śmiertelność pacjentów oraz wyróżnieniu najważniejszych czynników na to wpływających. Wykorzystali do tego regresję liniową, trzy metody oparte na drzewach (Random Forest, One Rule i Drzewo Decyzyjne), Artificial Neural Network, SVM liniowy i z gaussowskim rozkładem, metodę k-najbliższych sąsiadów, Naive Bayes oraz Gradient Boosting. Na podstawie przeprowadzonych badań doszli do wniosku, że kreatynina w surowicy i frakcja wyrzutowa są wystarczające do przewidzenia przeżycia pacjentów z niewydolnością serca, a nawet, że wzięcie pod uwagę tylko tych dwóch czynników może prowadzić do dokładniejszych przewidywań niż wykorzystanie w całości oryginalnych cech zbioru danych.

Zaman, S. M. M., Qureshi W. M., Raihan M. M. S., Shams A. B. & Sultana S. (2021)[5] ze względu na niezbilansowaną liczebność klas zastosowali Synthetic Minority Oversampling Technique (SMOTE). Do samej predykcji zastosowano dwie metody uczenia nienadzorowanego (metoda k-średnich i rozmyta metoda k-średnich) oraz trzy klasyfikatory ucznia nadzorowanego (Random Forest, XGBoost and Drzewo Decyzyjne). Wyniki wykazały lepszą wydajność nadzorowanych algorytmów Uczenia Maszynowego w porównaniu z modelami nienadzorowanymi. Ponadto, aby poprawić dokładność predykcji zaproponowany został zespołowy algorytm uczenia maszynowego na podstawie wcześniej stworzonych, który uzyskał wzrost wydajności do 99,98% w zakresie dokładności, precyzji, recall i F1-score z AUC wynoszącym 0,99.

6. Przedstawienie zmiennych

Oznaczenie oraz opis zmiennych dostępnych w wykorzystywanym zbiorze danych przedstawiony został w poniższej tabeli:

	Zmienna	Opis
X1	Wiek	wiek (w latach)
X2	Anemia	czy pacjent ma anemię (0 - nie, 1 - tak)
X3	Kinaza_fosfokreatynowa	poziom enzymu CPK (kinaza fosfokreatynowa) we krwi (mcg/l)
X4	Cukrzyca	czy pacjent ma cukrzycę (0 - nie, 1 - tak)
X5	Frakcja_wyrzutowa	stosunek objętości wyrzutowej serca (SV) do objętości końcoworozkurczowej komory serca (EDV) (procent)
X6	Wysokie_ciśnienie_krwi	czy pacjent ma nadciśnienie (0 - nie, 1 - tak)
X7	Płytki_krwi	płytki krwi we krwi (kilopłytki/ml)
X8	Kreatynina_w_surowicy	poziom kreatyniny w surowicy we krwi (mg/dl)
X9	Stężenie_sodu_w_surowicy	poziom sodu w surowicy we krwi (mEq/L)
X10	Płeć	płeć (0 - kobieta, 1 - mężczyzna)
X11	Palenie	czy pacjent pali czy nie (0 - nie, 1 - tak)
X12	Czas	okres obserwacji (liczba dni)
Y	Przypadek_śmiertelny	czy pacjent zmarł w okresie obserwacji (0 - nie, 1 - tak)

Tabela 1. Opis wykorzystywanych zmiennych

7. Wstępna analiza danych

7.1. Statystyki opisowe

Na początek zostały obliczone statystyki opisowe dla wszystkich zmiennych. Poniżej zamieszczona tabela przedstawia, jak kształtowały się wartości średnie, odchylenie standardowe, mediana, wartość minimalna i maksymalna, zakres, skośność, kurtoza, błąd standardowy oraz kwartyl 1 i 3.

	mean	sd	median	min	max	range	skew	kurtosis	se	Q0.25	Q0.75
X1	60.29	11.70	60.0	40.0	95.0	55.0	0.45	-0.10	0.17	50.0	68.0
X2	0.47	0.50	0.0	0.0	1.0	1.0	0.10	-1.99	0.01	0.0	1.0
X3	586.76	976.73	248.0	23.0	7861.0	7838.0	4.40	24.40	13.81	121.0	582.0
X4	0.44	0.50	0.0	0.0	1.0	1.0	0.24	-1.94	0.01	0.0	1.0
X5	37.73	11.51	38.0	14.0	80.0	66.0	0.49	-0.07	0.16	30.0	45.0
X6	0.36	0.48	0.0	0.0	1.0	1.0	0.56	-1.68	0.01	0.0	1.0
X7	265075.40	97999.76	263358.0	25100.0	850000.0	824900.0	1.16	5.17	1385.93	215000.0	310000.0
X8	1.37	1.01	1.1	0.5	9.4	8.9	4.62	27.67	0.01	0.9	1.4
X9	136.81	4.46	137.0	113.0	148.0	35.0	-1.00	3.73	0.06	134.0	140.0
X10	0.65	0.48	1.0	0.0	1.0	1.0	-0.61	-1.63	0.01	0.0	1.0
X11	0.31	0.46	0.0	0.0	1.0	1.0	0.81	-1.34	0.01	0.0	1.0
X12	130.68	77.33	113.0	4.0	285.0	281.0	0.11	-1.23	1.09	74.0	201.0
Y	0.31	0.46	0.0	0.0	1.0	1.0	0.80	-1.35	0.01	0.0	1.0

Tabela 2. Statystyki opisowe dla analizowanych danych.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	Y
wsp_zm	19.402	105.269	166.462	112.964	30.515	131.969	36.971	73.753	3.263	74.098	148.581	59.173	147.96

Tabela 3. Współczynniki zmienności dla poszczególnych zmiennych.

Ze względu na niską wartość współczynnika zmienności dla zmiennej X9, która jest znacznie mniejsza od 10, w dalszych analizach w kontekście klasyfikacji zmienna ta nie będzie brana pod uwagę. Natomiast zmienna ta została jeszcze uwzględniona na macierzy korelacji między

zmiennymi znajdującej się poniżej oraz w następnym podrozdziale, który dotyczy wizualizacji danych.

X1 - wiek

Średni wiek badanych wynosi 60,29 lat, a jego rozkład charakteryzuje się odchyleniem standardowym 11,70. Mediana wieku to 60 lat, z wartościami wahającymi się od 40 do 95 lat (zakres 55 lat). Rozkład wieku jest lekko asymetryczny dodatnio (skośność 0,45) i ma kurtozę bliską normalnej rozkładowi (-0,10). Wartości 25. i 75. percentyla wynoszą odpowiednio 50 lat oraz 68 lat.

X2 - Anemia

Anemia jest zmienną binarną, gdzie wartość 1 wskazuje na obecność anemii, a 0 na jej brak. Średnia wynosi 0,47, co oznacza, że około 47% badanych ma anemię. Odchylenie standardowe wynosi 0,50, a mediana to 0. Zakres wartości to 0–1, a rozkład jest minimalnie asymetryczny dodatnio (skośność 0,10) i spłaszczony (kurtoza -1,99).

X3 – Kinaza fosfokreatynowa

Średnie stężenie kinazy fosfokreatynowej wynosi 586,76 U/L, jednak zmienna ta charakteryzuje się dużą zmiennością (odchylenie standardowe 976,73 U/L) i skrajnymi wartościami od 23 do 7861 U/L (zakres 7838 U/L). Mediana wynosi 248 U/L, a rozkład jest silnie asymetryczny dodatnio (skośność 4,40) i wysoce szpiczasty (kurtoza 24,40). Wartości kwartylowe to 121 U/L oraz 582 U/L.

X4 – Cukrzyca

Cukrzyca jest zmienną binarną, gdzie wartość 1 wskazuje na obecność cukrzycy, a 0 na jej brak. Średnia wynosi 0,44, co oznacza, że około 44% badanych ma cukrzycę. Odchylenie standardowe to 0,50, a mediana wynosi 0. Rozkład jest lekko asymetryczny ujemnie (skośność -0,24) i spłaszczony (kurtoza -1,94).

X5 – Frakcja wyrzutowa

Średnia wartość frakcji wyrzutowej wynosi 37,73%, z odchyleniem standardowym 11,51%. Mediana to 38%, a wartości mieszczą się w przedziale od 14% do 66% (zakres 52%). Rozkład jest lekko asymetryczny dodatnio (skośność 0,49) i zbliżony do normalnego (kurtoza -0,07). Wartości kwartylowe wynoszą 30% oraz 45%.

X6 – Wysokie ciśnienie krwi

Wysokie ciśnienie krwi to zmienna binarna, gdzie 1 oznacza obecność nadciśnienia, a 0 brak. Średnia wynosi 0,36, co oznacza, że około 36% badanych ma nadciśnienie. Odchylenie standardowe wynosi 0,48, a mediana to 0. Rozkład jest minimalnie asymetryczny dodatnio (skośność 0,56) i spłaszczony (kurtoza -1,68).

X7 - Płytki krwi

Średnia liczba płytek krwi wynosi 265075,40 jednostek, z odchyleniem standardowym 97999,76 jednostek. Mediana to 263358 jednostek, a wartości mieszczą się w przedziale od 25100 do 850000 jednostek (zakres 824900 jednostek). Rozkład jest umiarkowanie asymetryczny dodatnio (skośność 1,16) i bardziej szpiczasty (kurtoza 5,17). Wartości kwartylowe to 215000 jednostek oraz 310000 jednostek.

X8- Kreatynina w surowicy

Średnia wartość kreatyniny w surowicy wynosi 1,37 mg/dL, z odchyleniem standardowym 1,01 mg/dL. Mediana to 1,1 mg/dL, a wartości wahają się od 0,5 do 9,4 mg/dL (zakres 8,9 mg/dL). Rozkład jest silnie asymetryczny dodatnio (skośność 4,62) i wyjątkowo szpiczasty (kurtoza 27,67). Wartości kwartylowe to 0,9 mg/dL oraz 1,4 mg/dL.

X9 - Stężenie sodu w surowicy

Średnie stężenie sodu w surowicy wynosi 136,81 mmol/L, z odchyleniem standardowym 41,64 mmol/L. Mediana to 137 mmol/L, a wartości wahają się od 35 do 148 mmol/L (zakres 113 mmol/L). Rozkład jest silnie asymetryczny ujemnie (skośność -3,73) i wysoce szpiczasty (kurtoza 10,06). Wartości kwartylowe wynoszą 134 mmol/L oraz 140 mmol/L.

X10 - Płeć

Płeć jest zmienną binarną, gdzie 1 oznacza mężczyzn, a 0 kobiety. Średnia wynosi 0,65, co oznacza, że 65% badanych to mężczyźni, a 35% to kobiety. Odchylenie standardowe wynosi 0,46, a mediana to 1. Rozkład jest umiarkowanie asymetryczny ujemnie (skośność -0,62) i spłaszczony (kurtoza -1,63).

X11 - Palenie

Palenie to zmienna binarna, gdzie 1 oznacza osoby palące, a 0 osoby niepalące. Średnia wynosi 0,36, co oznacza, że 36% badanych pali papierosy. Odchylenie standardowe wynosi 0,48, a mediana to 0. Rozkład jest silnie asymetryczny ujemnie (skośność -1,34) i bliski normalnemu (kurtoza 0,01).

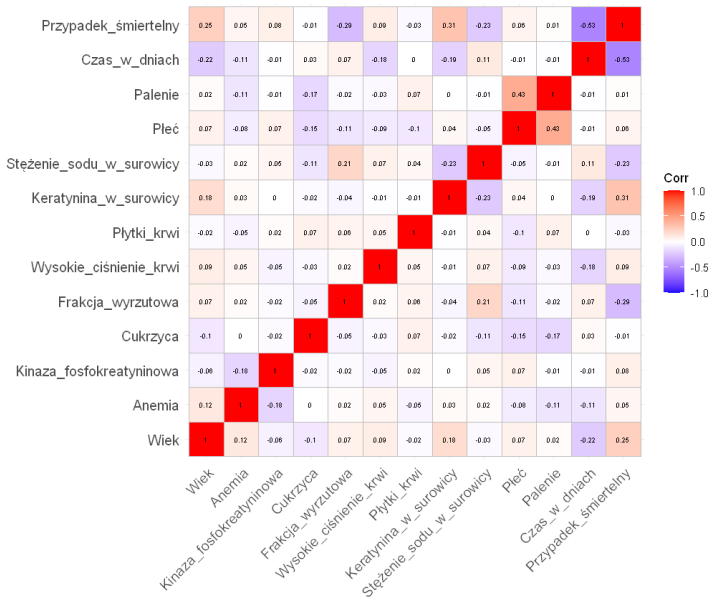
X12 - Czas

Czas obserwacji wynosi średnio 130,68 jednostek, z odchyleniem standardowym 77,33 jednostek. Mediana to 113 jednostek, a wartości wahają się od 0 do 285 jednostek (zakres 281 jednostek). Rozkład jest prawie symetryczny (skośność -0,11) i lekko spłaszczony (kurtoza -1,23). Wartości kwartylowe wynoszą 74 jednostki oraz 201 jednostek.

Y – Przypadek śmiertelny

Średnia wynosi 0,31, co oznacza, że około 31% badanych stanowiły przypadki śmiertelne. Odchylenie standardowe wynosi 0,46, a mediana to 0. Wartości mieszczą się w zakresie od 0

do 1. Rozkład tej zmiennej jest lekko asymetryczny ujemnie (skośność -0,80) i spłaszczony (kurtoza -1,35). Wartości pierwszego i trzeciego kwartyla to odpowiednio 0 i 1.

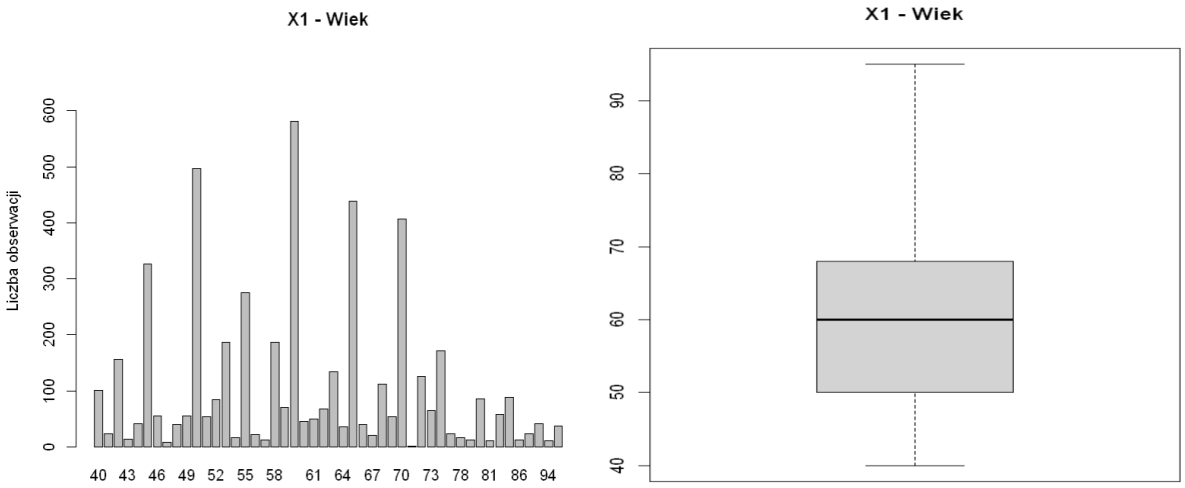


Wykres 1. Macierz korelacji między zmiennymi

Widać, że wśród analizowanych zmiennych nie znajdują się takie, które byłyby silnie ze sobą skorelowane.

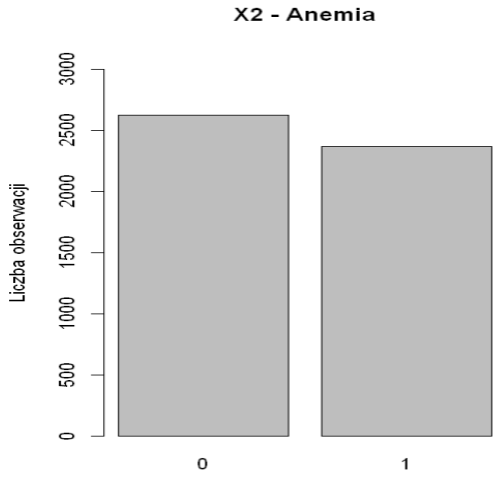
7.2. Wizualizacja danych

Na poniższych wykresach przedstawiono rozkłady cech, natomiast dla zmiennych ciągłych stworzono również boxploty.



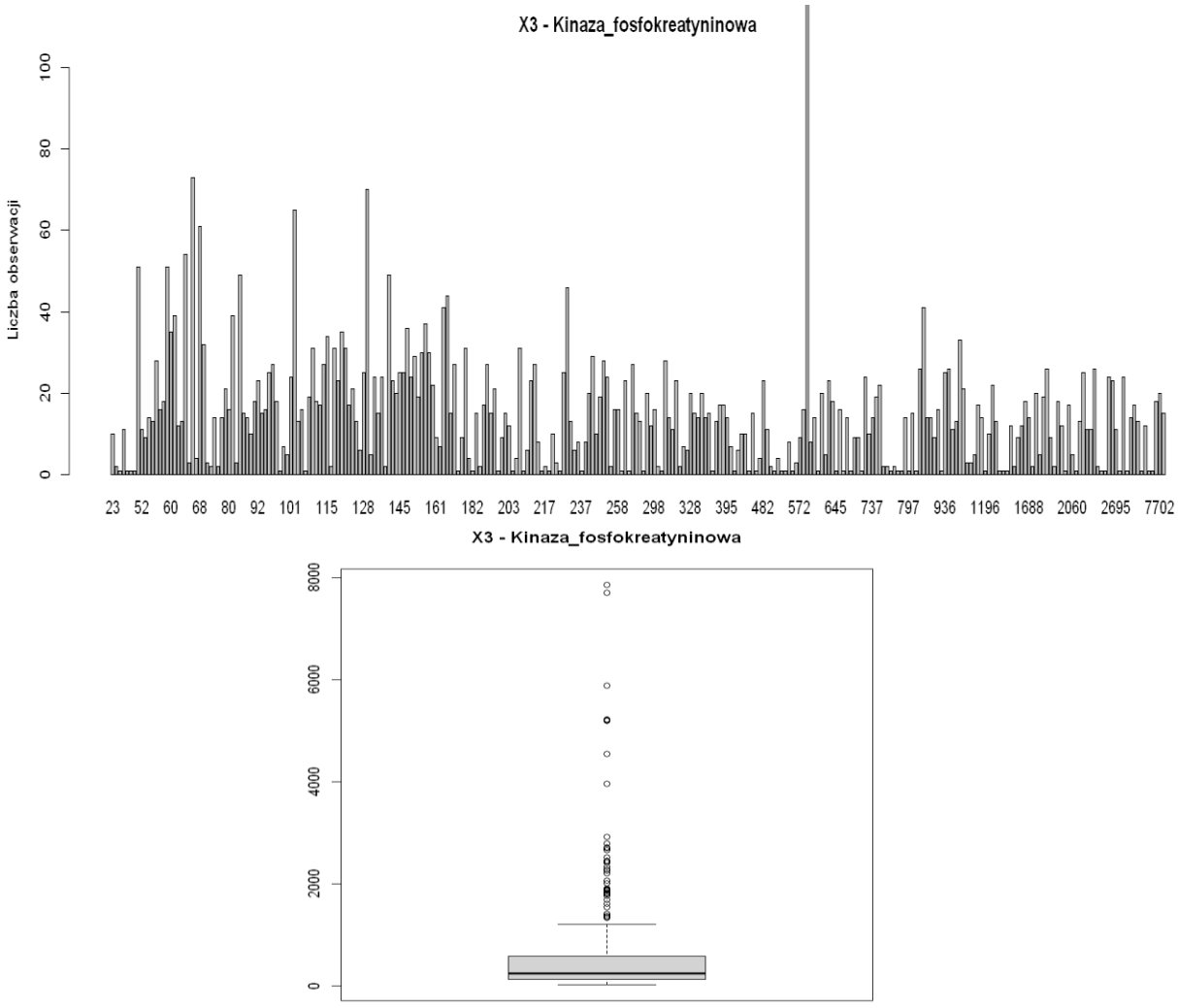
Wykres 2. Wykres rozkładu i wykres pudełkowy dla zmiennej X1.

Histogram pokazuje, że wiek badanych jest rozproszony z kilkoma szczytami, szczególnie w przedziale 58-64 lat. Wykres pudełkowy wskazuje, że mediana wynosi około 60 lat, a większość obserwacji mieści się w zakresie 50-10 lat bez wyraźnych wartości odstających.



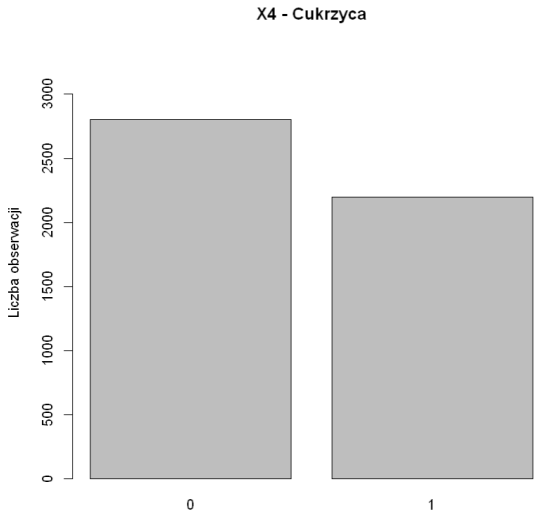
Wykres 3. Wykres rozkładu dla zmiennej X2.

Rozkład zmiennej X2 wskazuje, że wartość 0 (brak anemii) występuje częściej z liczbą obserwacji około 2500, a wartość 1 (występowanie anemii) jest rzadsza, z liczbą obserwacji około 2300. Grupa bez anemii jest nieznacznie większa, ale rozkład jest dość zrównoważony.



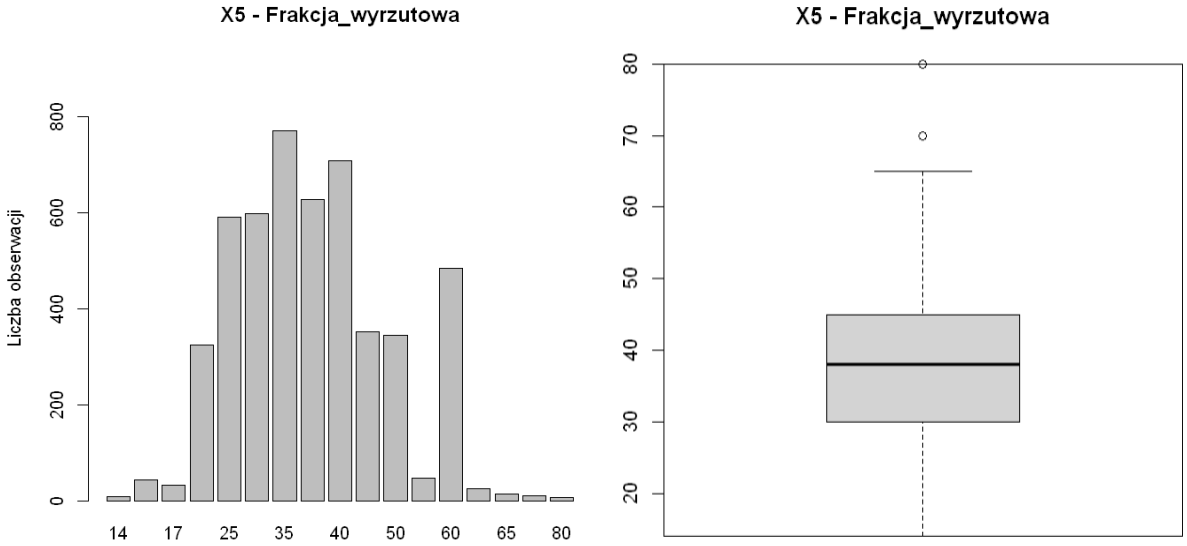
Wykres 4. Wykres rozkładu i wykres pudełkowy dla zmiennej X3.

Większość wartości X3 mieści się poniżej 400, ale występują liczne wartości odstające. Jeden ekstremalny o wartości około 7702, może silnie zaburzać rozkład. Wykres pudełkowy potwierdza skrajnie wysokie wartości odstające.



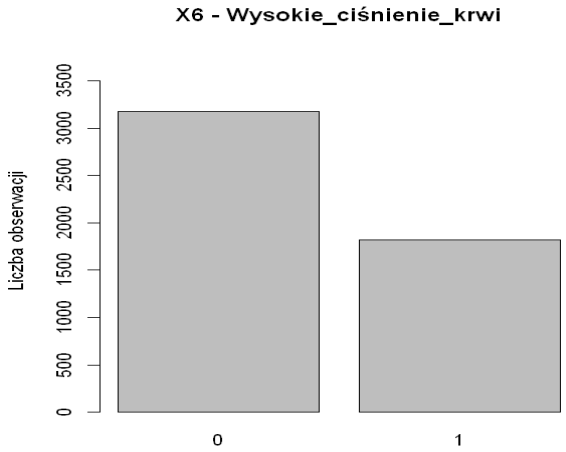
Wykres 5. Wykres rozkładu dla zmiennej X4.

Rozkład zmiennej X4 pokazuje, że wartość 0 (brak cukrzycy) występuje częściej, z liczbą obserwacji około 2800, natomiast wartość 1 (występowanie cukrzycy) jest mniej liczna, z liczbą obserwacji około 2200. Grupa osób bez cukrzycy przeważa, ale różnica nie jest znacząca, co sugeruje dość zrównoważony rozkład.



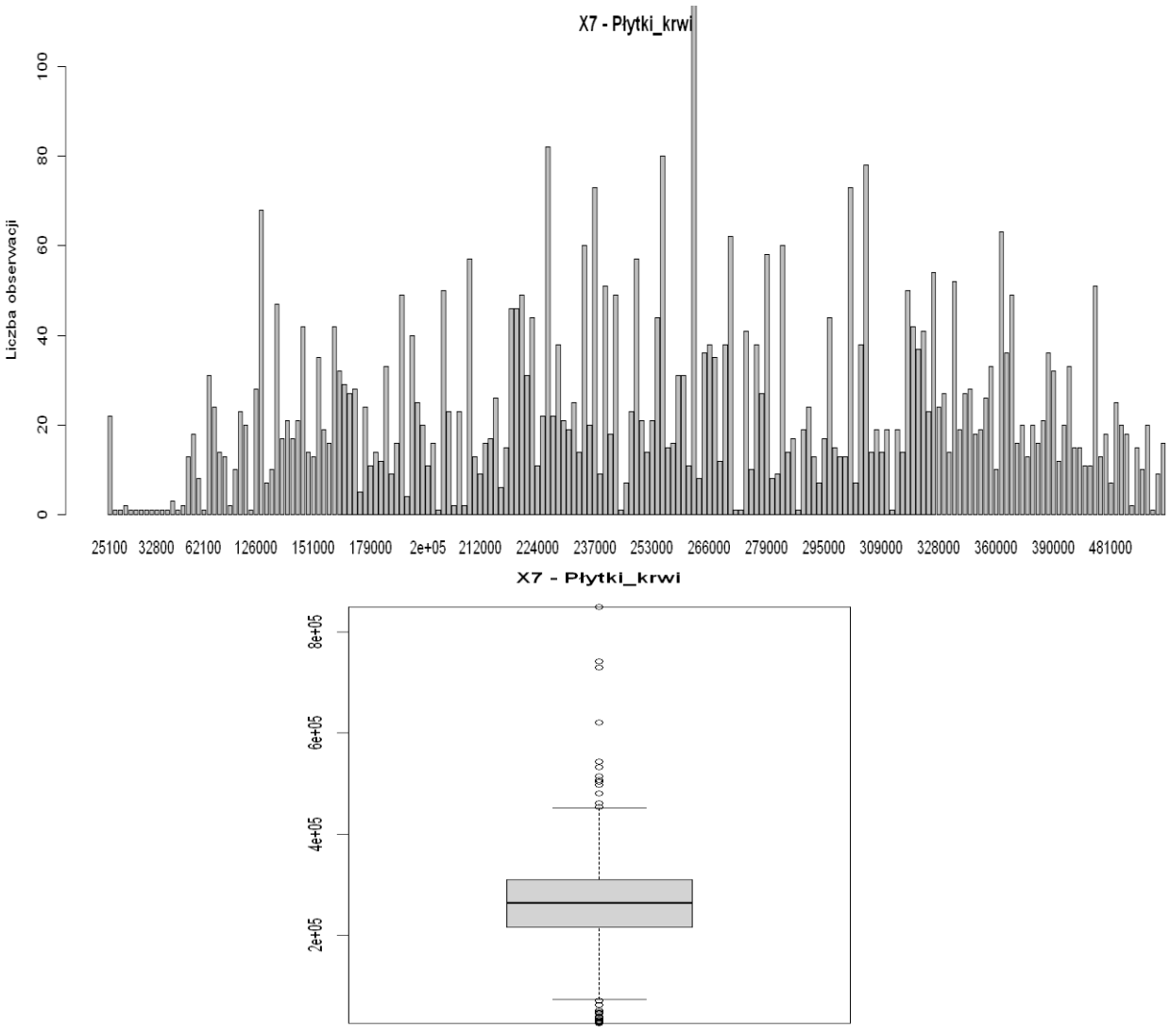
Wykres 6. Wykres rozkładu i wykres pudełkowy dla zmiennej X5.

Rozkład zmiennej X5 jest symetryczny z większością wartości w przedziale 30-50. Wykres pudełkowy pokazuje, że mediana wynosi około 40, a kilka wartości odstających przekracza 60. Większość obserwacji mieści się między 30 a 50.



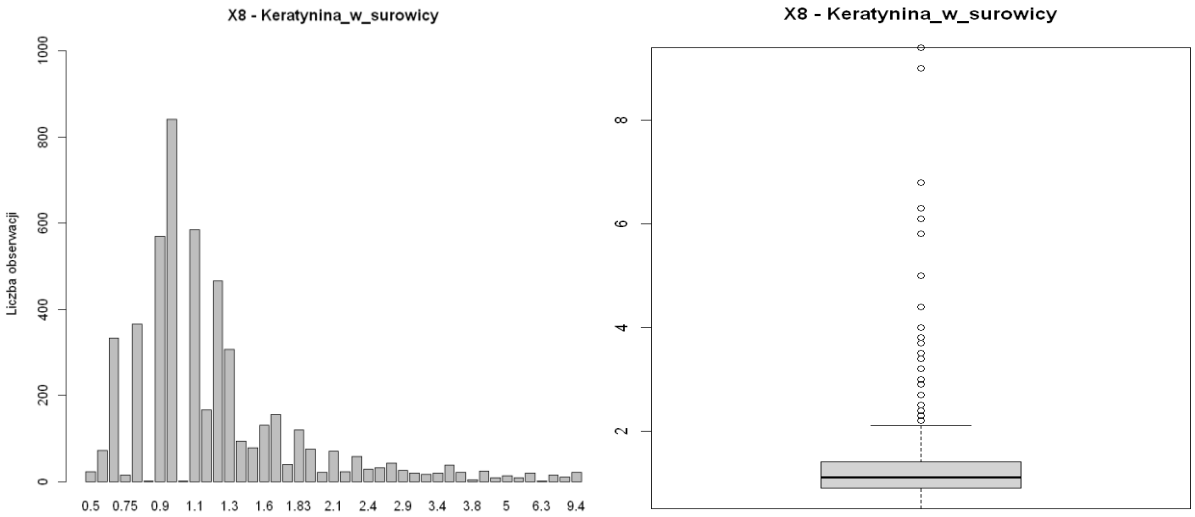
Wykres 7. Wykres rozkładu dla zmiennej X6.

Rozkład zmiennej X6 pokazuje, że wartość 0 (brak wysokiego ciśnienia) występuje częściej z liczbą obserwacji około 3300, a wartość 1 (wysokie ciśnienia) ma mniej obserwacji, około 1700. Osoby bez wysokiego ciśnienia krwi stanowią większość badanej grupy.



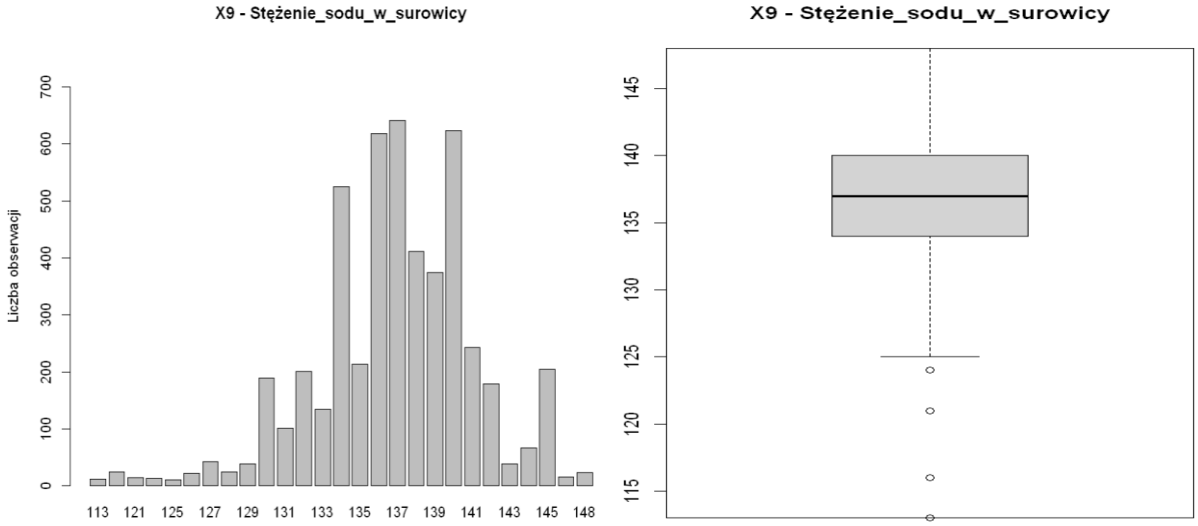
Wykres 8. Wykres rozkładu i wykres pudełkowy dla zmiennej X7.

Rozkład zmiennej X7 jest asymetryczny z wartościami skoncentrowanymi głownie między 150000 a 300000. Występują liczne wartości odstające powyżej 400000, co potwierdza wykres pudełkowy. Mediana wynosi około 250000.



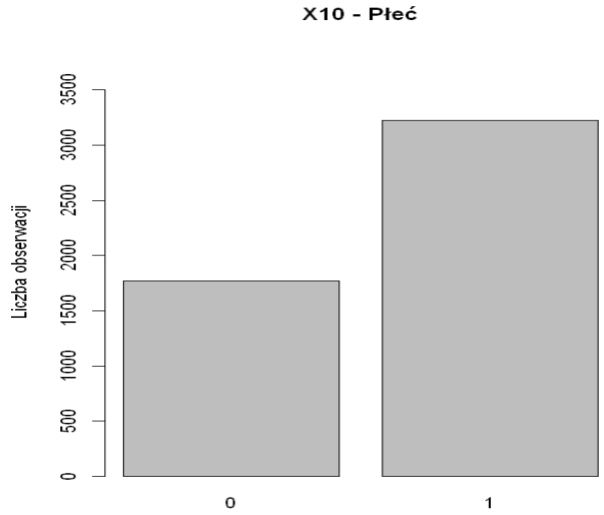
Wykres 9. Wykres rozkładu i wykres pudełkowy dla zmiennej X8.

Rozkład zmiennej X8 jest silnie prawostronnie asymetryczny. Większość wartości mieści się w przedziale 0,7 – 1,4, z wyraźnym pikiem około 1,1. Wykres pudełkowy pokazuje liczne wartości odstające powyżej 2, co świadczy o obecności ekstremalnych obserwacji.



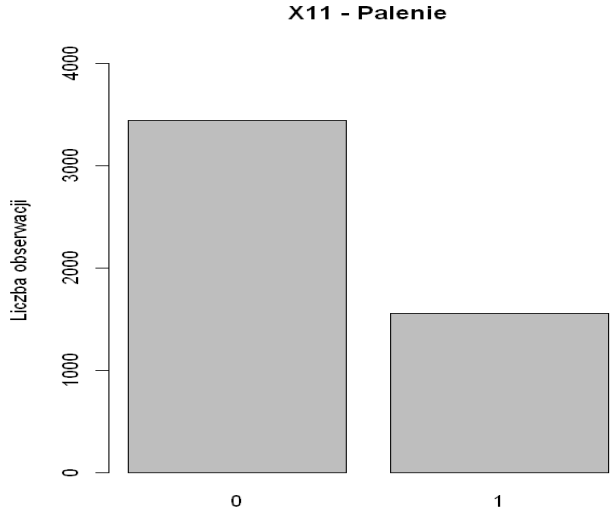
Wykres 10. Wykres rozkładu i wykres pudełkowy dla zmiennej X9.

Rozkład zmiennej X9 jest lekko asymetryczny z większością wartości w przedziale 134-140. Wykres pudełkowy pokazuje, że mediana wynosi około 137, a kilka wartości odstających znajduje się poniżej 125. Rozkład jest względnie skupiony, bez ekstremalnie odbiegających od normy wartości.



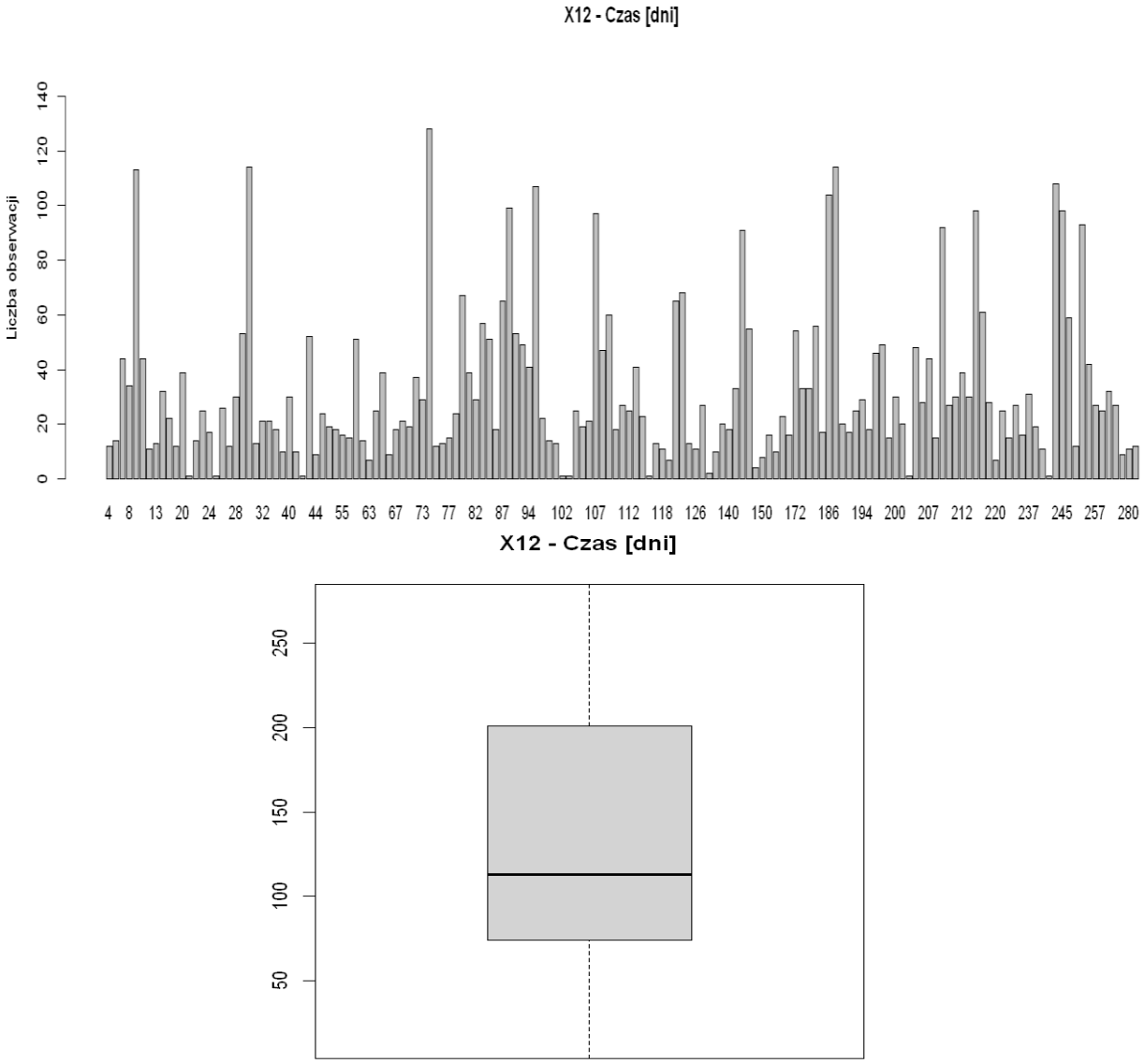
Wykres 11. Wykres rozkładu dla zmiennej X10.

Rozkład zmiennej X10 pokazuje, że wartość 1 (mężczyzna) występuje częściej od wartości 0 (kobieta). Grupa mężczyzn przeważa nad grupą kobiet w analizowanej próbie.



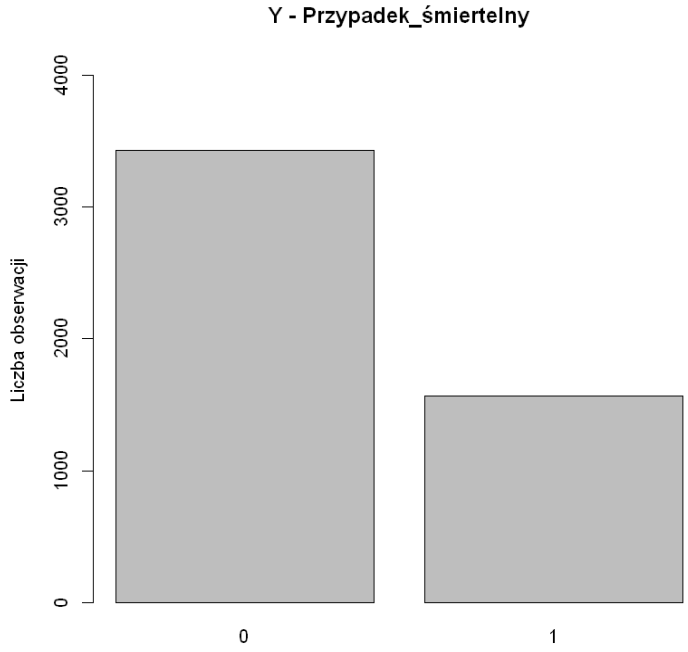
Wykres 12. Wykres rozkładu dla zmiennej X11.

Rozkład zmiennej X12 pokazuje, że wartość 1 (palacz) występuje znacznie rzadziej od wartości 0 (osoba niepaląca). Grupa niepalących zdecydowanie przeważa nad palącymi w analizowanej próbie.



Wykres 13. Wykres rozkładu i wykres pudełkowy dla zmiennej X12.

Rozkład zmiennej X12 jest rozproszony z licznymi wartościami w całym zakresie od 4 do 280 dni. Brak wyraźnych skupień, chociaż niektóre wartości są częstsze. Wykres pudełkowy pokazuje, że mediana wynosi około 100 dni, a większość wartości mieści się między 50 a 200 dni, bez widocznych wartości odstających.



Wykres 14. Wykres rozkładu dla zmiennej Y.

Rozkład zmiennej Y pokazuje, że wartość 0 występuje częściej niż wartość 1, zatem grupa osób, które przeżyły jest znacząco większa od grupy przypadków śmiertelnych.

7.3. Braki danych

Zbiór danych wykorzystany do analizy był kompletny i nie zawierał żadnych braków danych.

7.4. Obserwacje odstające

Obserwacje odstające zostały zastąpione wartościami wąsów boxplotów w zależności od tego czy wartość odstawała poniżej dolnego lub powyżej górnego wąsa.

7.5. Transformacja danych

Na danych została dokonana transformacja związana z wyżej wspomnianymi obserwacjami odstającymi oraz normalizacja min-max.

8. Opis metod wykorzystanych w pracy

8.1 Metoda Naiwnego Klasyfikatora Bayesa

Naiwny Klasyfikator Bayesa (ang. *Naive Bayes Classifier*) to prosta, ale potężna metoda klasyfikacji oparta na twierdzeniu Bayesa. Jest często stosowana w zadaniach takich jak analiza tekstu (np. filtrowanie spamu, klasyfikacja sentymentów) oraz w innych problemach, gdzie istotna jest szybkość i efektywność. Nazwa "naiwny" wynika z przyjęcia założeń upraszczających, które często nie są spełnione w rzeczywistych danych.

Stosujemy wzór Bayesa, którego postać ogólną można przedstawić następująco w postaci prawdopodobieństwa warunkowego:

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

natomiast dla naszego przypadku klasyfikacji równanie to można przekształcić do poniższej formy:

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)}$$

gdzie: C – klasa, x – zestaw zmiennych objaśniających, P(C) - prawdopodobieństwo przyporządkowania do klasy C bez informacji o warunku, który ma spełniać x (prawdopodobieństwo a priori), P(x|C) - prawdopodobieństwo poprawnego przyporządkowania do klasy C dla zestawu x, P(x) – prawdopodobieństwo dla x (predyktora) bez informacji o klasie, P(C|x) oznacza prawdopodobieństwo przyporządkowania do klasy C przy spełnieniu danego warunku przez zestaw zmiennych x.

Następnie powyższe równanie można przekształcić do postaci, zakładając, że x = (x_1, …, x_n), gdzie x_1, … , x_n to kolejne cechy w postaci wektora x:

$$P(C|x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n|C)P(C)}{P(x_1, \dots, x_n)}$$

a skoro zakładamy, że zmienne są między sobą niezależne to można to przekształcić:

$$P(C|x_1, \dots, x_n) = \frac{P(x_1|C)P(x_2|C) \dots P(x_n|C)P(C)}{P(x_1)P(x_2) \dots P(x_n)} = \frac{P(C)\prod_{i=1}^{n} P(x_i|C)}{P(x_1)P(x_2) \dots P(x_n)}$$

gdzie mianownik dla określonych danych wejściowych jest stały, więc możemy pozbyć się tego czynnika otrzymując zależność:

$$P(C|x_1, \dots, x_n) \propto P(C)\prod_{i=1}^{n} P(x_i|C)$$

Ostatecznie chcemy stworzyć model klasyfikatora. W tym celu znajdujemy prawdopodobieństwo danego zestawu danych wejściowych dla wszystkich możliwych wartości klas C i wybieramy dane wynik wyjściowy o maksymalnym prawdopodobieństwie. Matematycznie wyraża się to następująco:

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

gdzie: k – k-ta klasa, K – liczba wszystkich możliwych klas

Kluczowe dla naiwnego klasyfikatora Bayesa jest (naiwne) założenie, że atrybuty są parami niezależne. Było to sprawdzane poprzez macierz korelacji dla analizowanych zmiennych i na jej podstawie stwierdzono, że nie występuje silna korelacja między żadną parą zmiennych.

8.2. Metoda Klasyfikacyjnego Drzewa Decyzyjnego

Metoda drzew klasyfikacyjnych jest techniką analizy danych, która pozwala przypisywać przypadki lub obiekty do klas zmiennej zależnej (jakościowej) na podstawie wartości jednej lub więcej zmiennych objaśniających. Stanowi ona połączenie elementów statystyki i hierarchicznej analizy skupień, umożliwiając podział danych na logiczne grupy według określonych reguł decyzyjnych.

Drzewo klasyfikacyjne składa się z kilku podstawowych elementów:

- Korzeń (root): Punkt początkowy drzewa, reprezentujący cały zbiór danych.
- Węzły (nodes): Punkty, w których podejmowane są decyzje na podstawie warunków określonych dla zmiennych objaśniających.
- Gałęzie (branches): Ścieżki prowadzące z jednego węzła do kolejnego, wskazujące możliwe wyniki decyzji.
- Liście (leaves): Końcowe węzły drzewa, w których odczytuje się przynależność obserwacji do konkretnej klasy.

Proces klasyfikacji polega na przejściu przez drzewo od korzenia do jednego z liści, gdzie zostaje przypisana klasa odpowiadająca analizowanej obserwacji.

Wykorzystaną miarą niejednorodności węzła jest Indeks Giniego. Indeks różnorodności Giniego jest używany przez algorytm CART (classification and regression tree, czyli drzewa klasyfikacyjne i regresyjne) dla drzew klasyfikacyjnych. Wskaźnik ten mierzy jak często wybrany losowo element zestawu danych byłby niepoprawnie oznaczony, gdyby został oznaczony losowo i niezależnie zgodnie z rozkładem etykiet klas w badanym zestawie danych. Osiąga wartość minimalną 0 gdy wszystkie przypadki w węźle lądują w pojedynczej kategorii docelowej.Dla zestawu elementów z K klasami i względnymi częstościami p_i, i $\in \{1, 2, , \ldots K\}$ prawdopodobieństwo wybrania elementu z etykietą i wynosi p_i, a prawdopodobieństwo błędnej kategoryzacji tego elementu wynosi $\sum_{k \neq i} p_k = 1 - p_i$. Indeks Giniego jest obliczany przez zsumowanie iloczynów parami tych prawdopodobieństw dla każdej etykiety klasy:

$$I_G(p) = 1 - \sum_{i=1}^{K} p_i^2$$

8.3. Klasyfikator na podstawie regresji logistycznej

Do klasyfikacji binarnej zastosowana może być także regresja logistyczna. Hosmer i Lemeshow (2000) podają jej postać jako $P(Y = 1|\mathrm{x}) = \frac{e^{g(x)}}{1+e^{g(x)}}$, gdzie g(**x**) oznacza logit, $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$, a x_i to kolejne zmienne objaśniające, $i = 1,2, \ldots, p$. Na podstawie modelu dla konkretnych wartości cech uzyskuje się prawdopodobieństwo, że Y przyjmie wartość 1. Klasyfikacji dokonuje się na podstawie zestawienia z wartością progową, która najczęściej przyjmuje wartość $p^* = 0{,}5$. Jeśli uzyskane prawdopodobieństwo jest większe od p*, przyjmuje się, że Y^ = 1, w przeciwnym przypadku Y^ = 0. Do oszacowania modelu regresji logistycznej można użyć metody największej wiarygodności.

8.4. Model hybrydowy

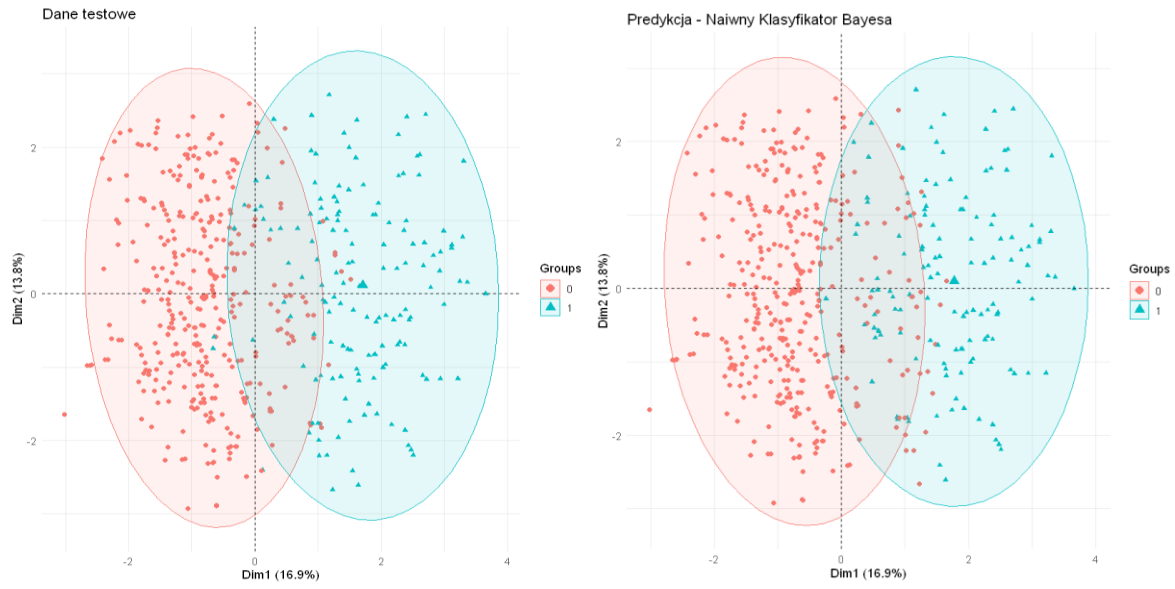
Model hybrydowy jest połączniem wyników trzech algorytmów klasyfikacyjnych: Naiwnego Klasyfikatora Bayesa, Drzewa Decyzyjnego i Regresji Logistycznej. Celem tej metody jest uzyskanie lepszych wyników predykcji niż w przypadku stosowania każdego z tych modeli indywidualnie. Model hybrydowy wykorzystuje średnią ważoną wyników trzech modeli bazowych w oparciu o ich efektywność na danych testowych, przy czym wagi są ustalane na podstawie wskaźnika F1-score, co pozwala uwzględnić zarówno precyzję, jak i wrażliwość modeli. Na podstawie średniej ważonej prawdopodobieństw ustalamy ostateczną predykcję klasy dla każdego rekordu.

9. Wyniki przeprowadzonych badań

W celu dokonania klasyfikacji zbiór danych został podzielony na zbiór treningowy (uczący) oraz zbiór testowy. Podział dokonano w proporcji 0,8 tzn. 80% danych zostało wykorzystane w procesie tworzenia klasyfikatorów, a pozostałe 20% danych służy do przetestowania działania tych klasyfikatorów i porównania uzyskanych rezultatów z wartościami rzeczywistymi. Zbiór danych składa się z 5000 rekordów, czyli 4000 z nich stanowiły zbiór uczący.

9.1. Metoda Naiwnego Klasyfikatora Bayesa

Wyniki dla danych testowych:



Wykres 15. Wizualizacja PCA predykcji Naiwnym Klasyfikatorem Bayesa dla danych testowych

Macierz trafień i statystyki predykcji dla danych testowych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	614	66
1	91	229

Tabela 4. Macierz trafień dla danych testowych – Naiwny Klasyfikator Bayesa

Model poprawnie sklasyfikował 614 przypadków negatywnych (TN) i 229 przypadków pozytywnych (TP).

Popełnił 91 błędów fałszywie negatywnych (FN) oraz 66 błędów fałszywie pozytywnych (FP).

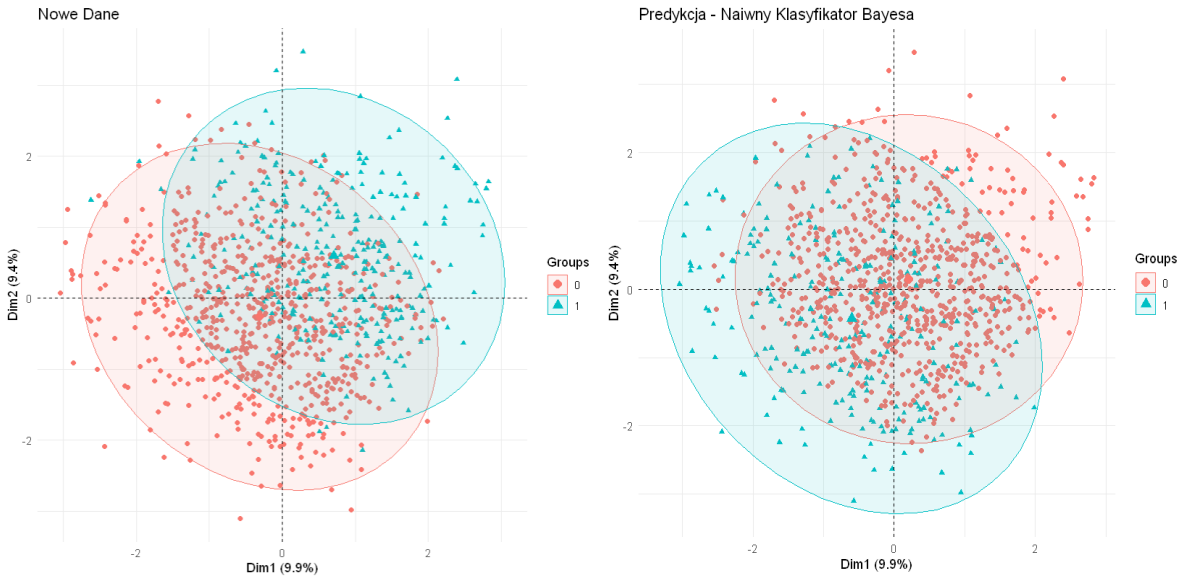
Precyzja	0,871
NPV	0,776
Specyficzność	0,716
Wrażliwość	0,903
Dokładność	0,843
F1-score	0,887

Tabela 5. Wskaźniki jakości dla danych testowych – Naiwny Klasyfikator Bayesa

Precyzja: Wynosi 0,871, co oznacza, że 87,1% wszystkich przewidywań pozytywnych było trafnych.
Wartość predykcyjna negatywna (NPV): Wynosi 0,776, co oznacza, że 77,6% wszystkich przewidywań negatywnych było prawidłowych.
Specyficzność: Wynosi 0,716, co oznacza, że model poprawnie rozpoznał 71,6% rzeczywistych przypadków negatywnych.
Wrażliwość: Wynosi 0,903, co wskazuje, że model wykrył 90,3% wszystkich rzeczywistych przypadków pozytywnych.
Dokładność: Wynosi 0,843, co oznacza, że model poprawnie sklasyfikował 84,3% wszystkich przypadków.
F1-score: Wynosi 0,887, co wskazuje na dobre wyważenie pomiędzy wrażliwością a precyzją.

Podsumowanie dla danych testowych
Model Naive Bayes w danych testowych wykazuje wysoką skuteczność, szczególnie w klasyfikowaniu pozytywnych przypadków (wysoka czułość) i utrzymuje przyzwoitą precyzję. Dobra dokładność oraz F1-score wskazują na solidność predykcji.

Wyniki dla nowych danych:



Wykres 16. Wizualizacja PCA predykcji Naiwnym Klasyfikatorem Bayesa dla nowych danych

Macierz trafień i statystyki predykcji dla nowych danych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	502	192
1	242	64

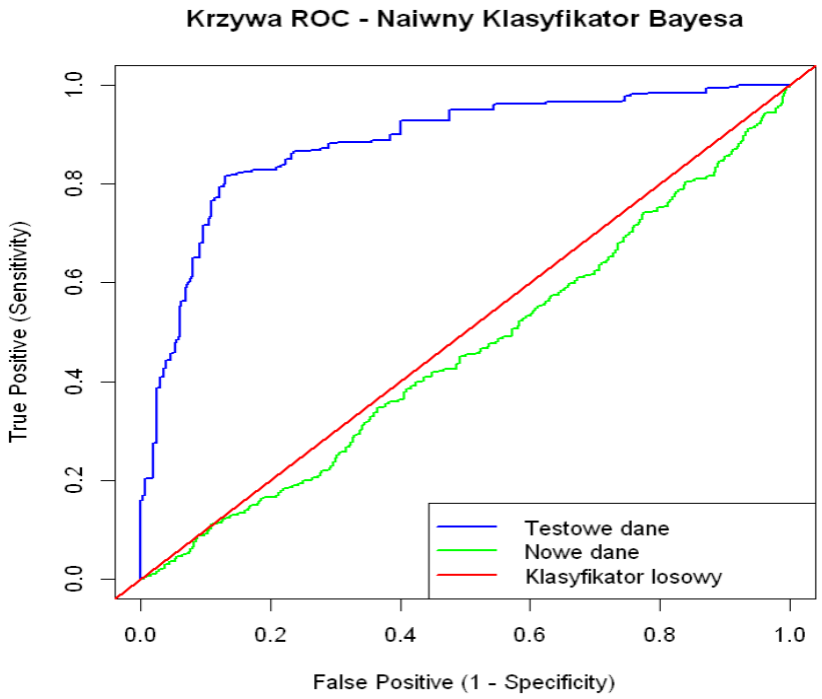
Tabela 6. Macierz trafień dla nowych danych – Naiwny Klasyfikator Bayesa

Model poprawnie sklasyfikował 502 przypadki pozytywne (TP) i 64 przypadki negatywne (TN).
Popełnił 242 błędy fałszywie pozytywne (FP) oraz 192 błędy fałszywie negatywne (FN).

Precyzja	0,675
NPV	0,25
Specyficzność	0,209
Wrażliwość	0,723
Dokładność	0,566
F1-score	0,698

Tabela 7. Wskaźniki jakości dla nowych danych– Naiwny Klasyfikator Bayesa

Precyzja: Wynosi 0,675, co oznacza, że 67,5% wszystkich przewidywań pozytywnych było trafnych.
Wartość predykcyjna negatywna (NPV): Wynosi jedynie 0,25, co oznacza, że tylko 25% przewidywań negatywnych było poprawnych.
Specyficzność: Wynosi 0,209, co oznacza, że model poprawnie rozpoznał tylko 20,9% rzeczywistych przypadków negatywnych.
Wrażliwość: Wynosi 0,723, co wskazuje, że model wykrył 72,3% rzeczywistych przypadków pozytywnych.
Dokładność: Wynosi 0,566, co wskazuje, że model poprawnie sklasyfikował jedynie 56,6% wszystkich przypadków.
F1-score: Wynosi 0,698, co wskazuje na akceptowalne, choć znacznie słabsze wyważenie między wrażliwością a precyzją w porównaniu do danych testowych.



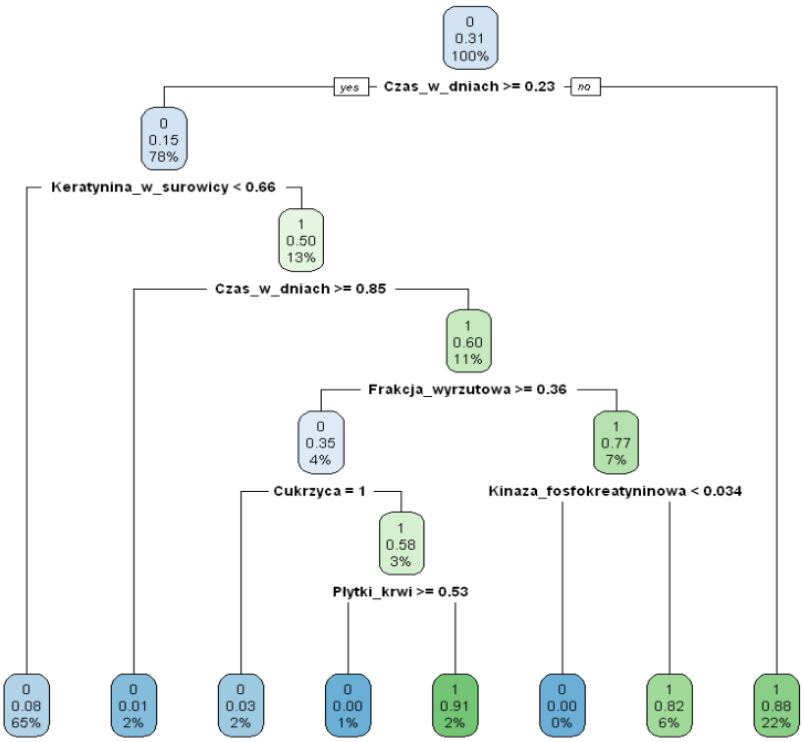
Wykres 17. Krzywa ROC – Naiwny Klasyfikator Bayesa

Krzywa ROC pokazuje jakość klasyfikatora w różnych progach decyzyjnych:

Krzywa dla danych testowych (niebieska) znajduje się znacznie powyżej linii klasyfikatora losowego i krzywej dla nowych danych (zielona), co potwierdza lepszą wydajność modelu w danych testowych.
Obszar pod krzywą (AUC) dla danych testowych jest większy, co potwierdza wyższą zdolność modelu do odróżniania przypadków pozytywnych od negatywnych w porównaniu z nowymi danymi.

9.2. Metoda Klasyfikacyjnego Drzewa Decyzyjnego

Wizualizacja powstałego klasyfikacyjnego drzewa decyzyjnego:



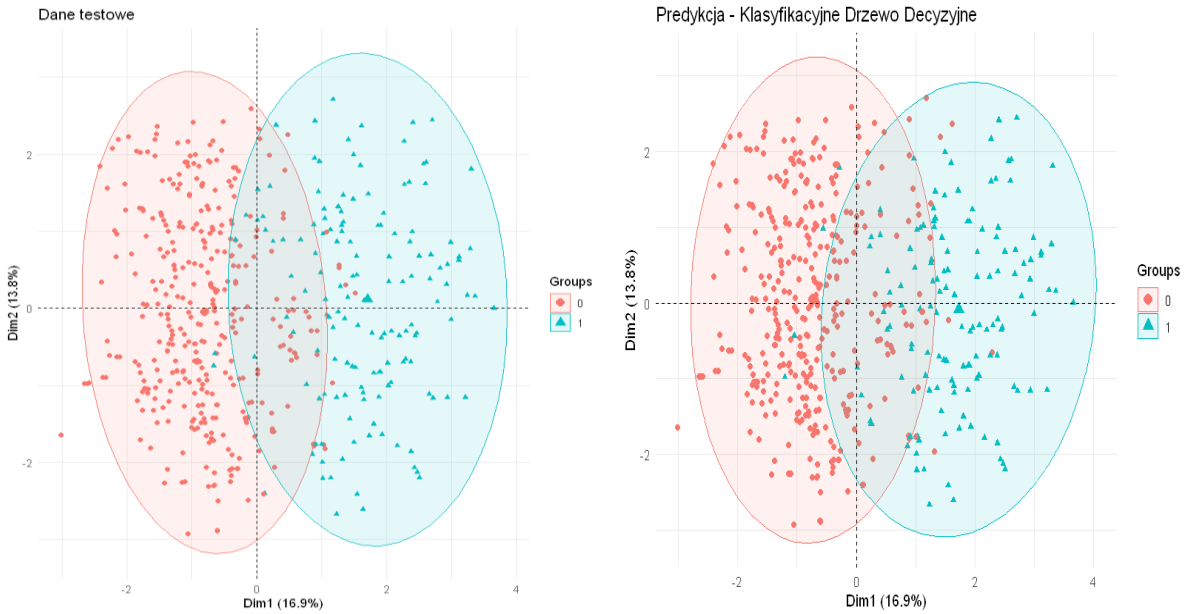
Wykres 18. Struktura powstałego drzewa klasyfikacyjnego

Wykres przedstawia strukturę drzewa decyzyjnego stworzonego w celu klasyfikacji danych. Każdy węzeł drzewa reprezentuje test dla konkretnej zmiennej, a gałęzie wychodzące z węzła odpowiadają różnym wartościom tej zmiennej. Liście drzewa (zielone prostokąty) wskazują na przewidywaną klasę.

Analiza drzewa:

- Pierwszy podział następuje ze względu na zmienną "Czas_w_dniach". Jeśli wartość jest mniejsza od 0,23, model przewiduje klasę "1" z prawdopodobieństwem 100%.
- Dla wartości "Czas_w_dniach" większych od 0,23, drzewo rozgałęzia się na podstawie zmiennej "Keratynina_w_surowicy".
- Dalsze podziały zależą od wartości kolejnych zmiennych, takich jak "Frakcja_wyrzutowa", "Wiek", "Płytki_krwi" czy "Cukrzyca".

Wyniki dla danych testowych:



Wykres 19. Wizualizacja PCA predykcji Klasyfikacyjnym Drzewem Decyzyjnym dla danych testowych

Macierz trafień i statystyki dla danych testowych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	635	45
1	67	253

Tabela 8. Macierz trafień dla danych testowych – Klasyfikacyjne Drzewo Decyzyjne

Macierz trafień pokazuje liczbę poprawnych i błędnych klasyfikacji dla danych testowych.

635 przypadków należących do klasy "0" zostało poprawnie sklasyfikowanych.
253 przypadków należących do klasy "1" zostało poprawnie sklasyfikowanych.
67 przypadków należących do klasy "1" zostało błędnie sklasyfikowanych jako klasa "0".
45 przypadków należących do klasy "0" zostało błędnie sklasyfikowanych jako klasa "1".

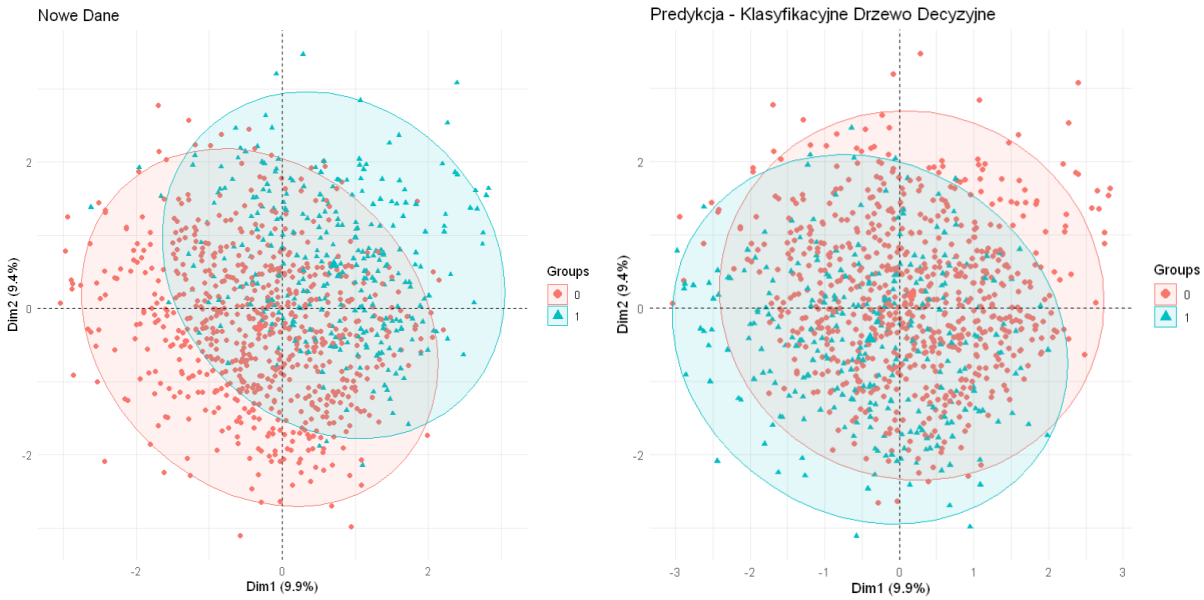
Precyzja	0,905
NPV	0,849
Specyficzność	0,791
Wrażliwość	0,934
Dokładność	0,888
F1-score	0,919

Tabela 9. Wskaźniki jakości dla danych testowych – Klasyfikacyjne Drzewo Decyzyjne

- **Precyzja:** Wynosi 0,905, co oznacza, że 90,5% wszystkich przewidywań pozytywnych było trafnych.
- **Wartość predykcyjna negatywna (NPV):** Wynosi jedynie 0,849, co oznacza, że tylko 84,9% przewidywań negatywnych było poprawnych.

- **Specyficzność:** Wynosi 0,791, co oznacza, że model poprawnie rozpoznał tylko 79,1% rzeczywistych przypadków negatywnych.
- **Wrażliwość:** Wynosi 0,934, co wskazuje, że model wykrył 93,4% rzeczywistych przypadków pozytywnych.
- **Dokładność:** Wynosi 0,888, co wskazuje, że model poprawnie sklasyfikował jedynie 88,8% wszystkich przypadków.
- **F1-score:** Wynosi 0,919, co wskazuje na dobre wyważenie pomiędzy wrażliwością a precyzją.

Wyniki dla nowych danych:



Wykres 20. Wizualizacja PCA predykcji Klasyfikacyjnym Drzewem Decyzyjnym dla nowych danych

Macierz trafień i statystyki predykcji dla nowych danych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	484	210
1	225	81

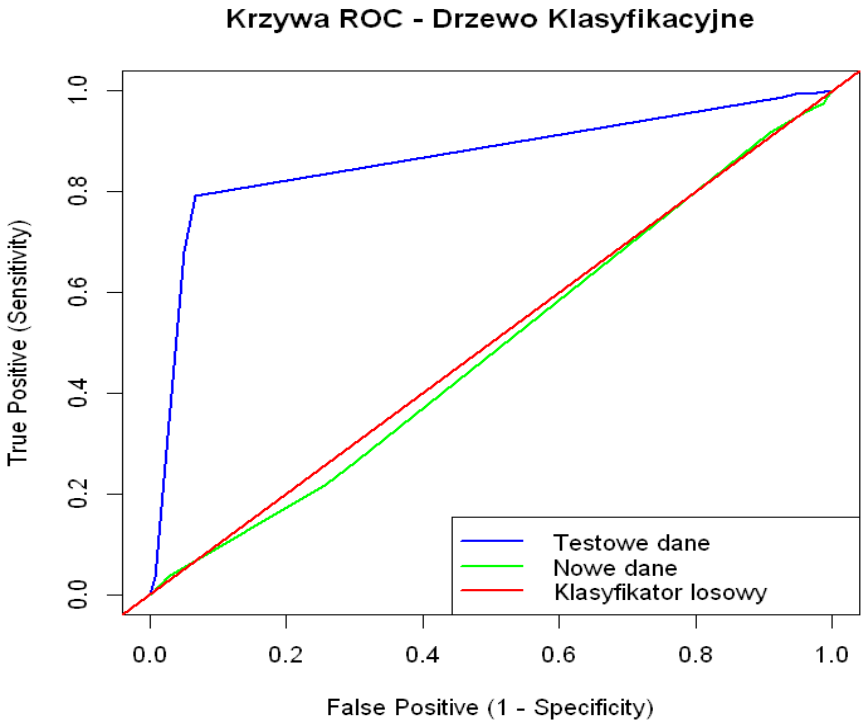
Tabela 10. Macierz trafień dla nowych danych – Klasyfikacyjne Drzewo Decyzyjne

484 razy model poprawnie przewidział wartość 0.
81 razy model poprawnie przewidział wartość 1.
225 razy model błędnie przewidział wartość 0 (powinna być 1).
210 razy model błędnie przewidział wartość 1 (powinno być 0).

Precyzja	0,683
NPV	0,278
Specyficzność	0,265
Wrażliwość	0,697
Dokładność	0,565
F1-score	0,69

Tabela 11. Wskaźniki jakości dla nowych danych – Klasyfikacyjne Drzewo Decyzyjne

- **Precyzja:** Wynosi 0,683, co oznacza, że 68,3% wszystkich przewidywań pozytywnych było trafnych.
- **Wartość predykcyjna negatywna (NPV):** Wynosi jedynie 0,278, co oznacza, że tylko 27,8% przewidywań negatywnych było poprawnych.
- **Specyficzność:** Wynosi 0,265, co oznacza, że model poprawnie rozpoznał tylko 26,5% rzeczywistych przypadków negatywnych.
- **Wrażliwość:** Wynosi 0,697, co wskazuje, że model wykrył 69,7% rzeczywistych przypadków pozytywnych.
- **Dokładność:** Wynosi 0,565, co wskazuje, że model poprawnie sklasyfikował jedynie 56,5% wszystkich przypadków.
- **F1-score:** Wynosi 0,69, co wskazuje na akceptowalne, ale znacznie słabsze wyważenie między wrażliwością a precyzją w porównaniu do danych testowych.



Wykres 21. Krzywa ROC – Klasyfikacyjne Drzewo Decyzyjne

Niebieska linia reprezentuje wydajność klasyfikatora na danych testowych.
Zielona linia reprezentuje wydajność klasyfikatora na nowych danych.
Czerwona linia reprezentuje losowy klasyfikator, który losowo przypisuje etykiety klas.

Im bliżej krzywa ROC znajduje się w lewym górnym rogu wykresu, tym lepsza jest wydajność klasyfikatora. Widzimy, że klasyfikator drzewa decyzyjnego działa dobrze na danych testowych, natomiast dla nowych danych ten klasyfikator jest zbliżony do klasyfikatora losowego. Stąd dla takiego zestawu nowych danych wyniki predykcji są znacznie gorsze.

9.3. Klasyfikator na podstawie regresji logistycznej

Oszacowania regresji logistycznej uzyskane w programie R prezentują się następująco:

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.17502    0.24494   0.715   0.4749
Wiek                      2.20112    0.24432   9.009  < 2e-16 ***
Anemia                    0.10182    0.10108   1.007   0.3138
Kinaza_fosfokreatyninowa  0.96131    0.15862   6.060 1.36e-09 ***
Cukrzyca                 -0.06290    0.09845  -0.639   0.5229
Frakcja_wyrzutowa        -3.52580    0.23627 -14.923  < 2e-16 ***
Wysokie_ciśnienie_krwi    0.44356    0.10132   4.378 1.20e-05 ***
Płytki_krwi              -0.47835    0.22359  -2.139   0.0324 *
Keratynina_w_surowicy     3.23466    0.19117  16.921  < 2e-16 ***
Płeć                      0.10041    0.11551   0.869   0.3847
Palenie                   0.13420    0.11648   1.152   0.2493
Czas_w_dniach            -5.83504    0.23699 -24.622  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

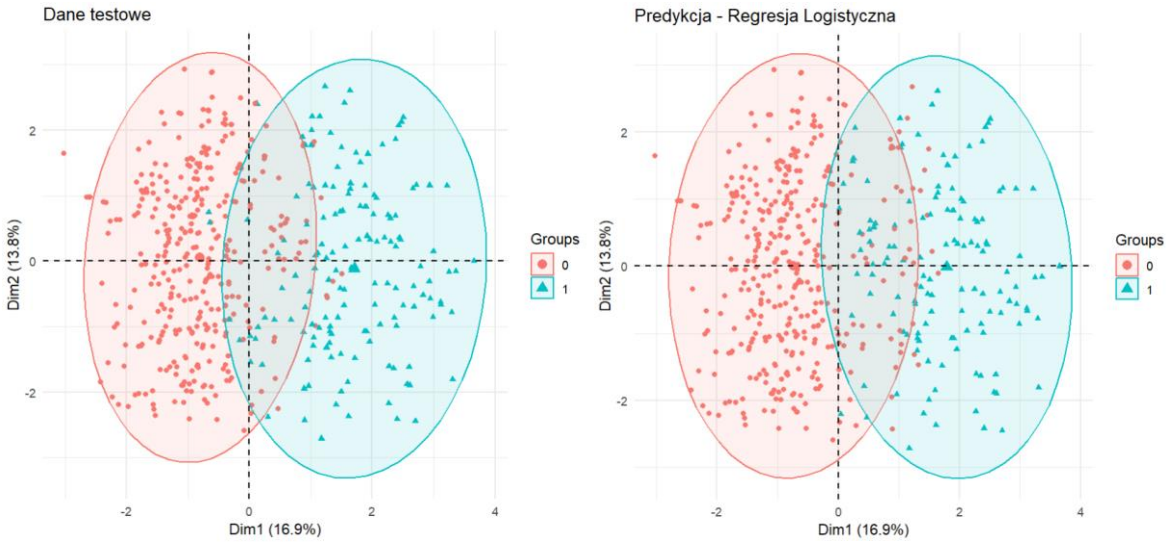
Tabela 12. Oszacowania współczynników, błędy standardowe, statystyka z oraz p-value dla zmiennych w regresji logistycznej

Obecność gwiazdek przy zmiennych wskazuje na ich istotność statystyczną oraz jej poziom (znaczenie poszczególnych symboli powyżej przy “Signif. codes"), dlatego od razu można zauważyć, że część ze zmiennych (oraz stała) okazała się nieistotna statystycznie. Takimi zmiennymi są Anemia, Cukrzyca, Płeć oraz Palenie, co może sugerować, że według tego modelu, nie wpływają istotnie na śmiertelność związaną z niewydolnością serca, jednak ze względu, że wszystkie pozostałe klasyfikatory opierały się na całym zestawie zmiennych, nie zostały one usunięte, aby umożliwić porównanie klasyfikatorów ze sobą.

Znaki przy oszacowaniach współczynników wskazują kierunek wpływu danej zmiennej na prawdopodobieństwo, że pacjent umrze, np. dodatni współczynnik przy zmiennej wiek wskazuje, że wraz z wzrostem wieku to prawdopodobieństwo było większe, natomiast ujemny współczynnik przy czasie sugeruje, że im dłuższy był okres obserwacji, tym mniejsze prawdopodobieństwo śmierci (inaczej, im dłużej pacjent żył podczas okresu obserwacji, tym większe prawdopodobieństwo, że przeżyje).

Na podstawie uzyskanego modelu dokonano predykcji wartości zmiennej objaśnianej dla danych testowych oraz nowo wygenerowanych, stosując wartość progową $p^* = 0{,}5$. Wyniki tych działań przedstawione zostały poniżej.

Wyniki dla danych testowych:



Wykres 22. Wizualizacja PCA predykcji Regresją Logistyczną dla danych testowych

Macierz trafień i statystyki dla danych testowych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	615	65
1	94	226

Tabela 13. Macierz trafień dla danych testowych – Regresja Logistyczna

Macierz trafień pokazuje liczbę poprawnych i błędnych klasyfikacji dla danych testowych.

615 przypadków należących do klasy "0" zostało poprawnie sklasyfikowanych.
226 przypadków należących do klasy "1" zostało poprawnie sklasyfikowanych.
94 przypadków należących do klasy "1" zostało błędnie sklasyfikowanych jako klasa "0".
65 przypadków należących do klasy "0" zostało błędnie sklasyfikowanych jako klasa "1".

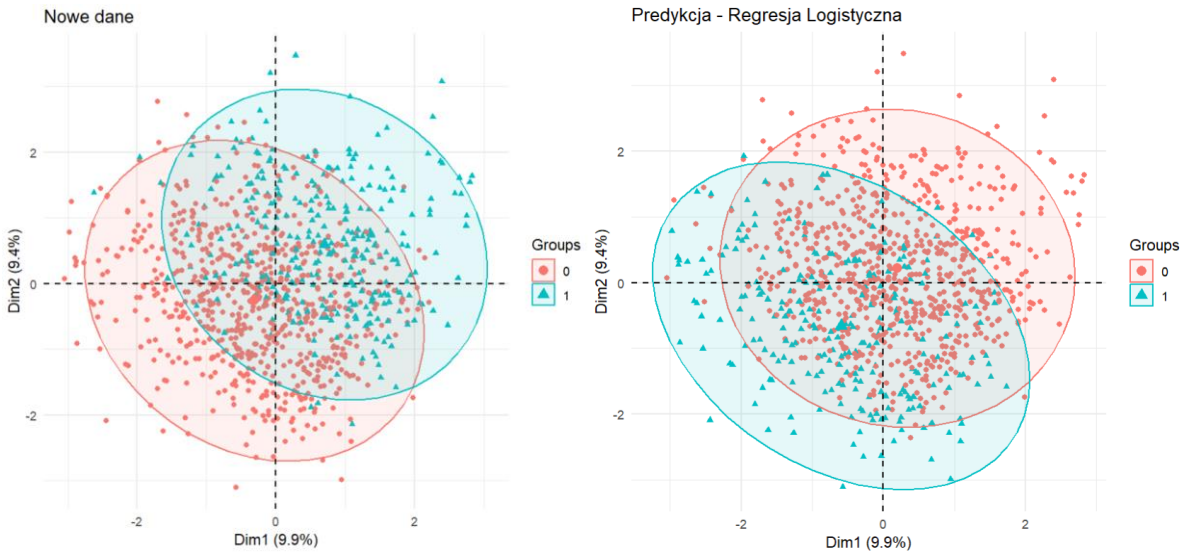
Precyzja	0,867
NPV	0,777
Specyficzność	0,706
Wrażliwość	0,904
Dokładność	0,841
F1-score	0,886

Tabela 14. Wskaźniki jakości dla danych testowych – Regresja Logistyczna

- **Precyzja:** Wynosi 0,867, co oznacza, że 86,7% wszystkich przewidywań pozytywnych było trafnych.
- **Wartość predykcyjna negatywna (NPV):** Wynosi 0,777, co oznacza, że tylko 77,7% przewidywań negatywnych było poprawnych.
- **Specyficzność:** Wynosi 0,706, co oznacza, że model poprawnie rozpoznał tylko 70,6% rzeczywistych przypadków negatywnych.

- **Wrażliwość:** Wynosi 0,904, co wskazuje, że model wykrył 90,4% rzeczywistych przypadków pozytywnych.
- **Dokładność:** Wynosi 0,841, co wskazuje, że model poprawnie sklasyfikował jedynie 84,1% wszystkich przypadków.
- **F1-score:** Wynosi 0,886, co wskazuje na całkiem dobre wyważenie pomiędzy wrażliwością a precyzją.

Wyniki dla nowych danych:



Wykres 23. Wizualizacja PCA predykcji Regresją Logistyczną dla nowych danych

Macierz trafień i statystyki dla nowych danych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	508	186
1	241	65

Tabela 15. Macierz trafień dla nowych danych – Regresja Logistyczna

Macierz trafień pokazuje liczbę poprawnych i błędnych klasyfikacji dla danych testowych.

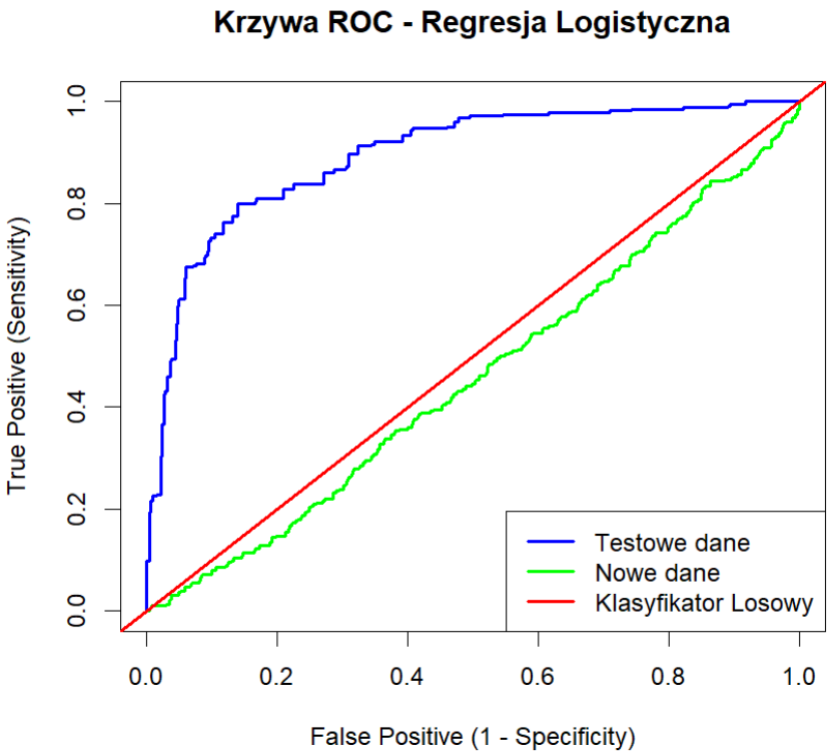
508 przypadków należących do klasy "0" zostało poprawnie sklasyfikowanych.
Tylko 65 przypadków należących do klasy "1" zostało poprawnie sklasyfikowanych.
241 przypadków należących do klasy "1" zostało błędnie sklasyfikowanych jako klasa "0".
186 przypadków należących do klasy "0" zostało błędnie sklasyfikowanych jako klasa "1".

Precyzja	0,678
NPV	0,259
Specyficzność	0,212
Wrażliwość	0,732
Dokładność	0,573
F1-score	0,704

Tabela 16. Wskaźniki jakości dla nowych danych – Regresja Logistyczna

- **Precyzja:** Wynosi 0,678, co oznacza, że tylko 67,8% wszystkich przewidywań pozytywnych było trafnych.
- **Wartość predykcyjna negatywna (NPV):** Wynosi zaledwie 0,259, co oznacza, że tylko 25,9% przewidywań negatywnych było poprawnych.
- **Specyficzność:** Wynosi tylko 0,212, co oznacza, że model poprawnie rozpoznał jedynie 21,2% rzeczywistych przypadków negatywnych.
- **Wrażliwość:** Wynosi 0,732, co wskazuje, że model wykrył 73,2% rzeczywistych przypadków pozytywnych.
- **Dokładność:** Wynosi 0,573, co wskazuje, że model poprawnie sklasyfikował jedynie 57,3% wszystkich przypadków.
- **F1-score:** Wynosi 0,704, co wskazuje na dość dobre wyważenie pomiędzy wrażliwością a precyzją.

Jakość klasyfikatora przedstawiono także na wykresie krzywej ROC.

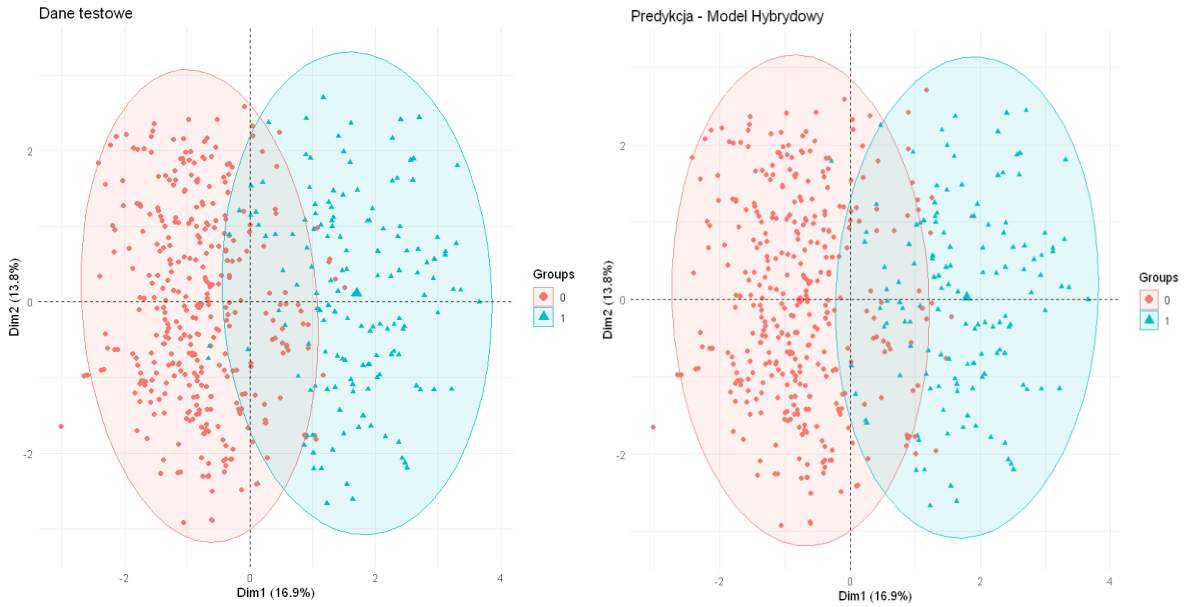


Wykres 24. Krzywa ROC – Regresja Logistyczna

Wykres ten jest zbliżony do tych poprzednio zaprezentowanych metod klasyfikacji. Tak jak dla Naiwnego Klasyfikatora Bayesa oraz Drzewa Decyzyjnego klasyfikator oparty o regresję logistyczną dość dobrze radzi sobie z klasyfikacją na danych testowych, natomiast dla nowo wygenerowanych danych radzi sobie dużo gorzej, a uzyskane tak wyniki są bliskie temu, jakby zupełnie losowo przypisać obserwacjom klasy.

9.4. Model hybrydowy

Wyniki dla danych testowych:



Wykres 25. Wizualizacja PCA predykcji modelem hybrydowym dla danych testowych

Macierz trafień i statystyki predykcji dla danych testowych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	633	47
1	70	250

Tabela 17. Macierz trafień dla danych testowych – Model Hybrydowy

633 razy model poprawnie przewidział wartość 0.
250 razy model poprawnie przewidział wartość 1.
70 razy model błędnie przewidział wartość 0 (powinna być 1).
47 razy model błędnie przewidział wartość 1 (powinno być 0).

Precyzja	0,9
NPV	0,842
Specyficzność	0,781
Wrażliwość	0,931
Dokładność	0,883
F1-score	0,915

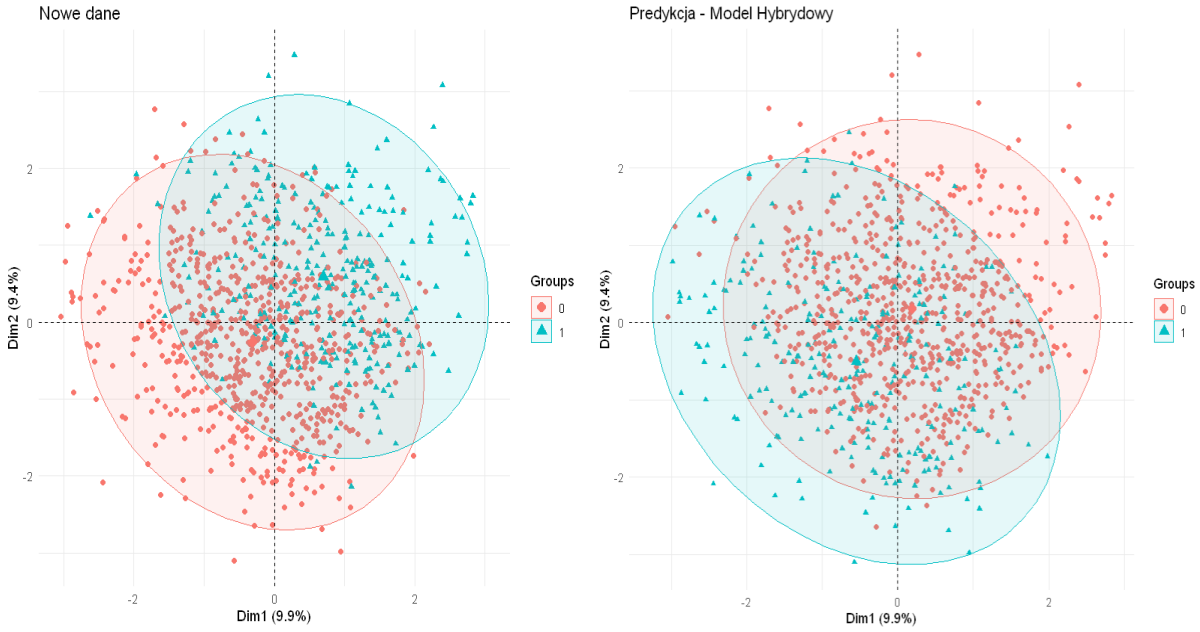
Tabela 18. Wskaźniki jakości dla danych testowych – Model Hybrydowy

- **Precyzja:** Wynosi 0,9, co oznacza, że 90% wszystkich przewidywań pozytywnych było trafnych.
- **Wartość predykcyjna negatywna (NPV):** Wynosi jedynie 0,842, co oznacza, że 84,2% przewidywań negatywnych było poprawnych.

- **Specyficzność:** Wynosi 0,781, co oznacza, że model poprawnie rozpoznał 78,1% rzeczywistych przypadków negatywnych.
- **Wrażliwość:** Wynosi 0,931, co wskazuje, że model wykrył 93,1% rzeczywistych przypadków pozytywnych.
- **Dokładność:** Wynosi 0,883, co wskazuje, że model poprawnie sklasyfikował 88,3% wszystkich przypadków.
- **F1-score:** Wynosi 0,915, co wskazuje na dobre wyważenie pomiędzy wrażliwością a precyzją.

Podsumowanie dla danych testowych:
Model hybrydowy uzyskał bardzo dobre wyniki dla danych testowych, charakteryzując się wysoką precyzją, wrażliwością i ogólną dokładnością. Dzięki dobrze wyważonemu F1-score (0,915) może on być szczególnie skuteczny w klasyfikacji przypadków pozytywnych, co czyni go wartościowym narzędziem do analizy.

Wyniki dla nowych danych:



Wykres 26. Wizualizacja PCA predykcji modelem hybrydowym dla nowych danych

Macierz trafień i statystyki predykcji dla nowych danych:

	Wartości przewidziane	
Wartości rzeczywiste	0	1
0	508	186
1	232	74

Tabela 19. Macierz trafień dla nowych danych – Model Hybrydowy

508 razy model poprawnie przewidział wartość 0.
74 razy model poprawnie przewidział wartość 1.
232 razy model błędnie przewidział wartość 0 (powinna być 1).
186 razy model błędnie przewidział wartość 1 (powinno być 0).

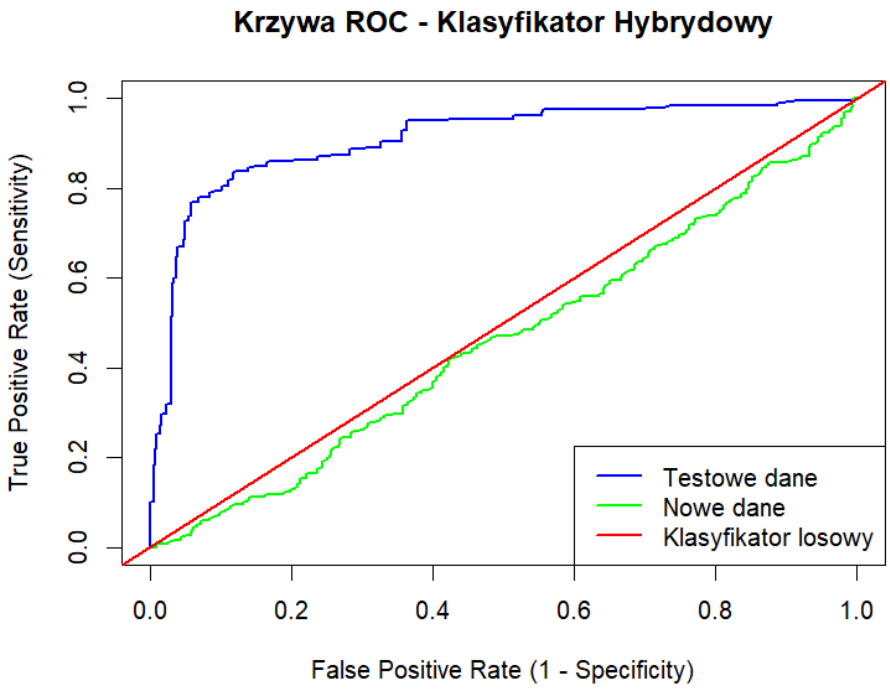
Precyzja	0,686
NPV	0,285
Specyficzność	0,242
Wrażliwość	0,732
Dokładność	0,582
F1-score	0,709

Tabela 20. Wskaźniki jakości dla nowych danych – Model Hybrydowy

- **Precyzja:** Wynosi 0,686, co oznacza, że 68,6% wszystkich przewidywań pozytywnych było trafnych.
- **Wartość predykcyjna negatywna (NPV):** Wynosi 0,285, co oznacza, że tylko 28,5% przewidywań negatywnych było poprawnych.
- **Specyficzność:** Wynosi 0,242, co oznacza, że model poprawnie rozpoznał tylko 24,2% rzeczywistych przypadków negatywnych.
- **Wrażliwość:** Wynosi 0,732, co wskazuje, że model wykrył 73,2% rzeczywistych przypadków pozytywnych.
- **Dokładność:** Wynosi 0,582, co wskazuje, że model poprawnie sklasyfikował jedynie 58,2% wszystkich przypadków.
- **F1-score:** Wynosi 0,709, co wskazuje na w miarę dobre wyważenie pomiędzy wrażliwością a precyzją.

Podsumowanie dla nowych danych

Model hybrydowy dla nowych danych dobrze radzi sobie z identyfikacją przypadków klasy 1, co potwierdza jego wysoka wrażliwość i F1-score. Jednak ma trudności z poprawnym rozpoznaniem klasy 0. W obecnej formie model może być bardziej przydatny w sytuacjach, gdzie ważniejsze jest minimalizowanie błędów dla klasy pozytywnej, kosztem błędów w klasie negatywnej.



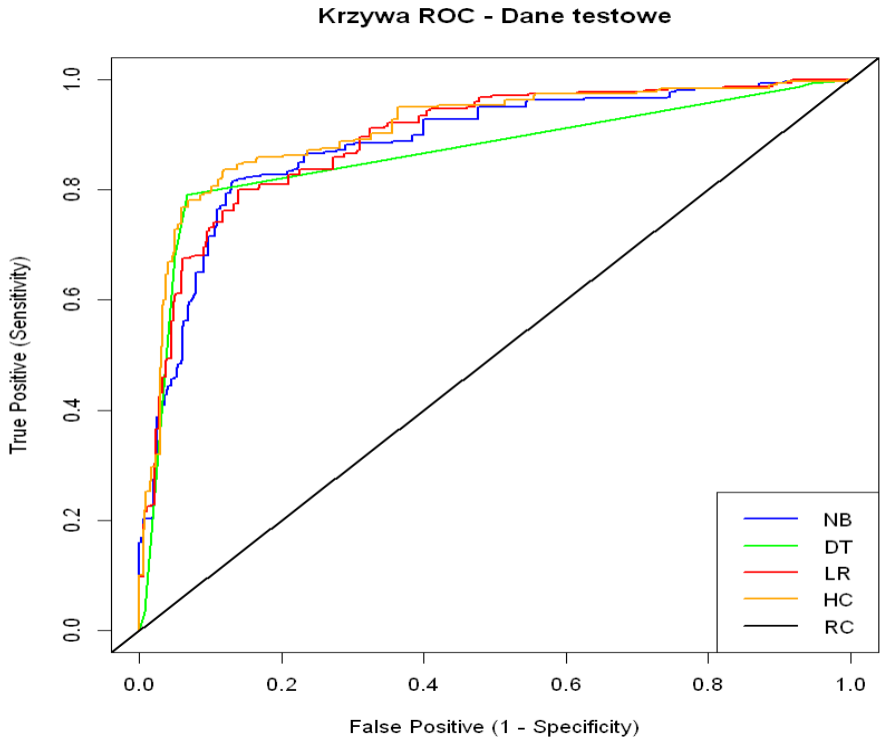
Wykres 27. Krzywa ROC – Model Hybrydowy

Im bliżej krzywa ROC znajduje się lewym górnym rogu wykresu, tym lepsza jest wydajność klasyfikatora. Klasyfikator losowy modelu hybrydowego działa dobrze na danych testowych, jednak dla nowych danych, zbliżony jest do klasyfikatora losowego. Dlatego, dla takiego zestawu nowych danych wyniki predykcji są znacznie gorsze.

Podsumowanie

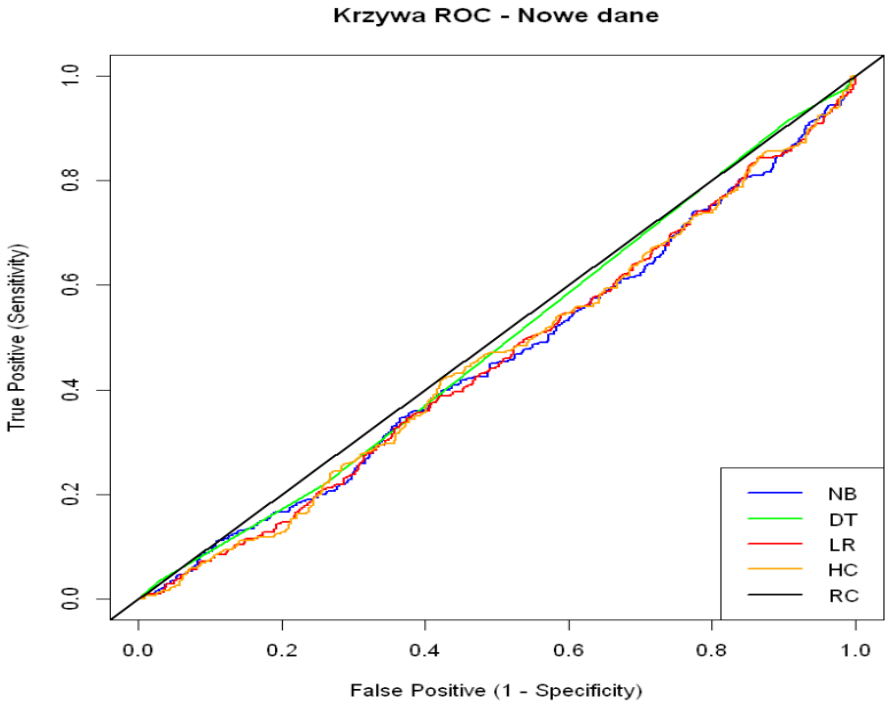
Model hybrydowy osiągnął bardzo dobre wyniki na danych testowych, co potwierdza jego zdolność do łączenia zalet poszczególnych modeli bazowych: Naiwnego Klasyfikatora Bayesa, Drzewa Decyzyjnego i Regresji Logistycznej. Jednak analiza wyników na nowych danych nie jest tak dobra. Model ma problem z poprawnym rozpoznawaniem klasy negatywnej, co prowadzi do obniżenia wartości specyficzności i NPV na nowych danych. Natomiast jego wrażliwość pozostaje względnie dobra, co oznacza, że nadal skutecznie wykrywa przypadki klasy pozytywnej.

10. Podsumowanie i wnioski



Wykres 28. Porównanie działania klasyfikatorów na danych testowych

Wyniki osiągnięte dla danych testowych są dość dobre i zbliżone do siebie dla wszystkich analizowanych metod klasyfikacji. Jednak nieco lepiej wykazały się: Klasyfikacyjne Drzewo Decyzyjne (linia zielona) oraz Model Hybrydowy (linia pomarańczowa) w porównaniu do pozostałych metod.



Wykres 29. Porównanie działania klasyfikatorów na nowych danych

Z kolei dla nowych danych wyniki dla poszczególnych klasyfikatorów są raczej słabe i zbliżone do klasyfikatora losowego, a nawet nieco gorsze od niego. Wynika to z faktu, że wszystkie modele dla nowych danych miały problem z poprawną predykcją klasy negatywnej. Ponadto tak słabe wyniki dla nowych danych wynikają także z tego, że nowy zestaw danych był generowany całkowicie losowo. Przez to mogły występować przypadki, gdzie wylosowany zestaw wartości zmiennych wejściowych mógł bardziej odpowiadać danej klasie, ale mimo to wartość wyjściowa była ustawiana jako klasa przeciwna. W takiej sytuacji klasyfikator mógł nawet dobrze i sensownie dokonywać predykcji, ale porównując ten wynik z etykietą nadaną przy generowaniu danych mogły występować rozbieżności skutkujące pogorszeniem się jakości ostatecznych rezultatów.

11. Bibliografia

[1] McMurray, J. & Pfeffer, M. (2005). Heart failure. *The Lancet.* 365 (9474), s. 1877–1889

[2] https://www.kaggle.com/datasets/aadarshvelu/heart-failure-prediction-clinical-records/data

[3] Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Ali Raza, M. (2017). Survival analysis of heart failure patients: A case study

[4] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making, 20.*

[5] Zaman, S. M. M., Qureshi W. M., Raihan M. M. S., Shams A. B. & Sultana S. (2021). Survival Prediction of Heart Failure Patients using Stacked Ensemble Machine Learning Algorithm, *2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, Dhaka, Bangladesh, pp. 117-120,

[6] Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression. Second Edition. *Wiley Series in Probability and Statistics*

[7] Pełka M. Uniwersytet Ekonomiczny we Wrocławiu, ZASTOSOWANIE DRZEW KLASYFIKACYJNYCH DLA DANYCH SYMBOLICZNYCH W OCENIE PREFERENCJI KONSUMENTÓW

[8] Kazienko, P., Lughofer, E. & Trawiński, B. (2013). Hybrid and Ensemble Methods in Machine Learning. *Journal of Universal Computer Science*, vol. 19, no. 4 (2013), 457-461