Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Uczenie Maszynowe

Dokumentacja końcowa

Piotr Patek, Jan Potaszyński

Warszawa, 2024

Spis treści

1. Ogólne założenia

W ramach projektu z przedmiotu Uczenie Maszynowe przydzielone zostało następujące zadanie:

„Las losowy z naiwnym klasyfikatorem bayesowskim (NBC) w zadaniu klasyfikacji. Postępujemy tak jak przy tworzeniu lasu losowego, tylko co drugi klasyfikator w lesie to NBC. Jeden z klasyfikatorów (NBC lub drzewo ID3) może pochodzić z istniejącej implementacji. Przed rozpoczęciem realizacji projektu proszę zapoznać się z zawartością: `https://staff.elka.pw.edu.pl/~rbiedrzy/UMA/index.html`."

Na podstawie treści zadania autorzy zdecydowali się zaimplementować zarówno klasyfikator w postaci drzewa ID3, jak i naiwny klasyfikator bayesowski. Na podstawie stworzonych modeli lokalnych tworzony będzie model w postaci lasu losowego, gdzie co drugie drzewo decyzyjne zostanie zastąpione naiwnym klasyfikatorem bayesowskim oraz las utworzony z samych drzew ID3.

2. Implementacja algorytmów drzew ID3, Naiwnego Klasyfikatora Bayesowskiego i Lasu Losowego

2.1. Założenia matematyczne

2.1.1. Drzewo decyzyjne ID3

Drzewo decyzyjne ID3 to klasyczne drzewo decyzyjne tworzone za pomocą algorytmu o nazwie ID3. Algorytm ten polega na takim rozgałęzianiu drzewa, by sumaryczna entropia $H(S)$ definiowana następującym wzorem:

$$H(S) = -\sum_{x \in X} p(x) \cdot \log_2 p(x) \tag{2.1}$$

gdzie:

— S - Dane, dla których liczona jest entropia,
— X - Zbiór klas w zbiorze S,
— $p(x)$ - Stosunek liczby elementów w zbiorze S z przypisaną klasą $x \in X$ do liczby wszystkich elementów w zbiorze S.

była z każdym następnym rozgałęzieniem możliwie najbardziej zmniejszana. Innaczej, żeby przyrost wiedzy $G(S, S_a, S_b)$ definiowany wzorem:

$$G(S, S_a, S_b) = H(S) - (H(S_a) + H(S_b)) \tag{2.2}$$

gdzie:

— S - Dane, dla których liczona jest entropia,
— S_a, S_b - Dychotomiczny podział danych S.

był jak największy.

Najprostszą metodą podziału danych jest porównanie jej do pewnej wybranej wartości. Na przykład:

$$\begin{gathered} s \in S \\ \lambda_s \geqslant value \implies s \in S_a \\ \lambda_s < value \implies s \in S_b \end{gathered} \tag{2.3}$$

gdzie λ_s jest wartością atrybutu, na podstawie którego dokonujemy podziału, a $value$ wartością wybraną przez algorytm, na podstawie której zachodzi podział.

Rozgałęzienia są dokonywane aż nie zajdzie jeden z tzw. warunków stopu. W tym przypadku jeden z następujących trzech:

— Gdy rozpatrywany zbiór S jest 1 elementowy (jest to szczególny przypadek następnego warunku),
— Gdy wszystkie elementy rozpatrywanego zbioru S należą do tej samej klasy $x \in X$,
— Gdy osiągana jest maksymalna głębokość drzewa (ustalana arbitralnie przez trenującego)

W takim wypadku tworzony jest liść, czyli ostateczna klasyfikacja $x \in X$ (w przypadku, gdy zachodzi trzeci warunek liść przyjmuje klasę częściej występującą w zbiorze)

Algorytm ID3 tworzenia drzewa decyzyjnego w pseudokodzie wygląda następująco:

Algorithm 1 Algorytm ID3

Definicja Podziel(k, λ) - dokonaj podziału zbioru S ze względu na wartość λ parametru k
Definicja Entropia(Podziel(k, λ)) - oblicz entropię tak dokonanego podziału
Definicja Warunek STOP:
- Jeśli Wszystkie elementy S są tej samej klasy
- Jeśli maksymalna głębokość drzewa została osiągnięta

Stworzenie pierwszego węzła dla zbioru S
Realizacja algorytmu:
if Warunek STOP **then**
    Zamiana tego węzła na liść, klasyfikujący wg najliczniejszej klasy w zbiorze S
end if
k_{best} = Dowolny atrybut k
for Każdy rodzaj atrybutu k **do**
    λ_{best} = Dowolna wartość λ atrybutu k
    for Wartość λ atrybutu k dla każdego z elementów z S **do**
        if Entropia(Podziel(k,λ)) < Entropia(Podziel(k,λ_{best})) **then**
            $\lambda_{\text{best}} = \lambda$
        end if
    end for
    if Entropia(Podziel(k,λ_{best})) < Entropia(Podziel(k_{best},λ_{best})) **then**
        $k_{\text{best}} = k$
    end if
end for
Dodanie k_{best} i λ_{best} jako parametrów decyzyjnych tego węzła
(S_a, S_b) = Podziel(k_{best}, λ_{best})
Stworzenie węzłów dla zbiorów S_a oraz S_b i dodanie ich jako węzłów potomnych tego węzła.
Realizacja algorytmu dla węzłów potomnych

Przykładowe obliczenia entropii dla podziału 3 elementowego zbioru S

Załóżmy, że posiadamy 3 elementowy zbiór S zawierający próbki o jednym atrybucie ciągłym. Każda próbka ma przydzieloną jedną z dwóch klas 0 lub 1. Naszym celem jest taki podział pierwotnego zbioru S na podzbiory, aby zmniejszać entropię zbioru po podziale i maksymalizować przyrost wiedzy (ang. information gain).

Wartość atrybutu	1	2	3
Klasyfikacja	0	0	1

Tab. 2.1: Wartości atrybutu i klasy przypisane próbkom z przykładowego zbioru S

Entropia dla niepodzielonego zbioru:

$$H(S) = -\frac{1}{3} \cdot \log_2 \frac{1}{3} - \frac{2}{3} \cdot \log_2 \frac{2}{3} \approx 0{,}918 \tag{2.4}$$

Entropia dla podziału wg kryterium (Atrybut $\geqslant 2$):

$$H(S_{21}) + H(S_{22}) = -1 \cdot \log_2 1 - \frac{1}{2} \cdot \log_2 \frac{1}{2} - \frac{1}{2} \cdot \log_2 \frac{1}{2} = 1 \tag{2.5}$$

Entropia dla podziału wg kryterium (Atrybut $\geqslant 3$):

$$H(S_{31}) + H(S_{32}) = -1 \cdot \log_2 1 - 1 \cdot \log_2 1 = 0 \tag{2.6}$$

Przyrost wiedzy dla podziału wg kryterium (Atrybut $\geqslant 2$):

$$G(S, S_{21}, S_{22}) = H(S) - (H(S_{21}) + H(S_{22})) = -0{,}082 \tag{2.7}$$

Przyrost wiedzy dla podziału wg kryterium (Atrybut $\geqslant 3$):

$$G(S, S_{31}, S_{32}) = H(S) - (H(S_{31}) + H(S_{32})) = 1 \tag{2.8}$$

Z obliczeń wynika, że podział według kryterium (Atrybut $\geqslant 3$) jest podziałem preferowanym przez algorytm ID3.

Tak stworzone drzewo użyte może być do klasyfikacji obiektu w następujący sposób. Obiekt zaczyna od pierwszego wierzchołka (zwanego korzeniem). Obliczana jest wartość funkcji entropii dla zbioru, na podstawie której wybierany jest najlepszy atrybut wg którego powinien nastąpić następny podział i jego próg. Następne wierzchołki są tworzone tak długo jak nie zostanie spełnione kryterium stopu. Wtedy wierzchołek zostaje liściem, któremu przypisana jest konkretna klasa w klasyfikacji.

2.1.2. Naiwny klasyfikator bayesowski

Naiwny klasyfikator bayesowski to metoda klasyfikacji oparta na powszechnie znanym w statystyce twierdzeniu Bayesa.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.9}$$

gdzie A, B to zdarzenia losowe.

Na potrzeby uczenia maszynowego, zmienne losowe A i B zastąpimy poprzez zmienne y i $x_1, \cdots, x_n$ oznaczające odpowiednio klasę przyporządkowaną danej próbce i wektor atrybutów danej próbki.

Po wspomnianych modyfikacjach wzór wygląda następująco:

$$P(y|x_1, \cdots, x_n) = \frac{P(x_1, \cdots, x_n|y)P(y)}{P(x_1, \cdots, x_n)} \tag{2.10}$$

W tej chwili możemy zastosować główne założenie algorytmu naiwnego klasyfikatora bayesowskiego, od którego pochodzi jego nazwa, a mianowicie założenie (naiwne i niepotwierdzone), że predyktory modelu są niezależne względem siebie.

$$P(x_1, \cdots, x_n|y) = \prod_{j=1}^{n} P(x_j|y) \tag{2.11}$$

W praktyce prawdopodobieństwo warunkowe, że atrybut x_i przyjmie daną wartość pod warunkiem, że próbka należy do danej klasy y estymuje się na podstawie zbioru trenującego $\mathbb{T}$ w następujący sposób:

$$P(x_i|y) = \frac{|\mathbb{T}_{y,x_i}|}{|\mathbb{T}_y|} \tag{2.12}$$

Prawdopodobieństwa *a priori* możemy obliczyć w bardzo prosty sposób na podstawie naszego zbioru trenującego, zliczając liczbę wystąpień danej klasy w tym zbiorze, a następnie licząc jej stosunek do wielkości całego zbioru trenującego.

$$P(y) = \frac{|\mathbb{T}_y|}{|\mathbb{T}|} \tag{2.13}$$

Ostatecznie wzór na predykcję klasy dla próbki $\mathbf{x}$ o określonych atrybutach wygląda następująco:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.14}$$

gdzie:

— n - liczba atrybutów dla przykładów,
— x_i - i-ty atrybut przykładu,
— y - dowolna klasa ze zbioru danych,
— $P(y)$ - prawdopodobieństwo *a priori* dla danej klasy,
— $P(x_i|y)$ - prawdopodobieństwo warunkowe wystąpienia danej klasy dla danej wartości atrybutu,
— $\hat{y}$ - predykcja modelu.

Dodatkowo należy jeszcze wspomnieć, że naiwny klasyfikator bayesowski w podstawowej formie działa na atrybutach dyskretnych. W tym celu, przed etapem trenowania modelu należy dokonać dyskretyzacji wszystkich ciągłych argumentów.

2.1.3. Zmodyfikowany las losowy

Zmodyfikowany las losowy polega na stworzeniu wielu różnych klasyfikatorów (w tym przypadku drzew losowych oraz naiwnych klasyfikatorów bayesowskich, w stosunku 50/50)(ilość klasyfikatorów może być arbitralnie wybrana przez trenującego). Klasyfikatory tworzone są za pomocą tzw. próby bootstrap z liczebnością próby równą liczebności populacji. Innymi słowy z n-elementowego zbioru danych wybieramy n-elementową próbę za pomocą losowania ze zwracaniem z naszego zbioru danych. Dzięki temu klasyfikatory będą się od siebie różniły. Predykcja odbywa się przez głosowanie, czyli zliczana jest ilość poszczególnych klasyfikacji dokonanych przez każdy z klasyfikatorów w lesie i wybierana jest klasyfikacja częściej występująca.

2.2. Implementacja algorytmów

2.2.1. Ogólny opis implementacji

W ramach projektu zaimplementowano algorytmy drzewa decyzyjnego ID3, naiwnego klasyfikatora bayesowskiego i algorytm zmodyfikowanego lasu losowego. Kod został napisany zgodnie z przedstawionymi założeniami matematycznymi w języku Python. Algorytmy zostały zaimplementowane możliwie uniwersalnie, tj. mogą zostać uruchomione na dowolnym zbiorze o atrybutach oraz klasyfikacjach wyrażanych liczbowo. Same programy zostały zgrupowane w jeden pakiet pythonowy **uma24z-nbc-random-forest** przez co ich wykorzystanie jest bardzo proste i intuicyjne. Dodatkowo, zgodnie z założeniami przedstawionymi w dokumentacji wstępnej, kod został pokryty w części testami jednostkowymi.

W implementacji algorytmu ID3 możliwa jest modyfikacja głębokości maksymalnej drzewa, natomiast w lesie losowym możliwy jest wybór liczby modeli składających się na las oraz ułamek, mówiący o tym jaką część z tych modeli stanowią drzewa ID3 (pozostałe modele są naiwnymi klasyfikatorami Bayesa). Drzewa wchodzące w skład lasu losowego są lekko zmodyfikowane, tzn. w każdym węźle przed podziałem wybierane jest losowo bez zwracania $\sqrt{n}$ atrybutów, spośród których algorytm wybierze ten dla którego podział drzewa przyniesie największy przyrost informacji i najbardziej zmniejszy entropię w węzłach potomnych. Ponadto wszystkie drzewa w lesie losowym nie mają ograniczenia głębokości (tj. ograniczenie jest na tyle duże, że jest nieosiągalne w żadnym ze sprawdzanych przypadków - jest równe 1e10). Te dwie zmiany razem z zastosowaniem próby bootstrapowej przy tworzeniu zbioru treningowego na modelu lokalnego ma na celu maksymalne zwiększenie różnorodności drzew, a więc głównego mechanizmu, który sprawia, że las losowy osiąga na wielu zbiorach danych bardzo dobre rezultaty.

2.2.2. Napotkane problemy i rozwiązania

Tak jak wspomniano algorytmy zostały zaimplementowane w języku Python, z wykorzystaniem popularnych bibliotek takich jak **numpy** czy **scikit-learn**. Niestety mimo tego, przez charakterystykę Pythona, który jest językiem interpretowanym, wydajność algorytmów nie jest oszałamiająca. W tym celu w algorytmie lasu losowego dodano możliwość trenowania w trybie wielowątkowym, z możliwością podania liczby wykorzystywanych wątków. Poprawiło to wydajność programu, ale docelowo i tak należałoby się zastanowić w przyszłości nad implementacją algorytmów w języku kompilowanym takim jak C++ czy Rust i wyprowadzeniem ich interfejsu do Pythona. Byłby to dobry kompromis pomiędzy szybkością działania programu i wygodą korzystania z niego.

2.2.3. Zależności programu i instalacja

Pakiet **uma24z-nbc-random-forest** może być zainstalowany z wykorzystaniem narzędzia **pip** po jego wywołaniu na głównym folderze **uma24z-nbc-random-forest** znajdującym się w skompresowanym archiwum lub na repozytorium dostępnym na GitHubie.

Link do GitHuba `https://github.com/ViisteK528/UMA`

3. Eksperymenty

Dla każdego zbioru wykonano po 4 eksperymenty, nie wliczając wstępnej analizy danych (jak sprawdzanie rozkładu danych itp.)

— Naiwny klasyfikator bayesowski
— Drzewo ID3 z ograniczeniem głębokści
— Las losowy z 50 drzewami ID3
— Las losowy ID3, NBC w ilości kolejno 25,25

Każda powtórzenie eksperymentu przeprowadzane było według poniższej procedury:

1. Podzielenie zbioru na trenujący (90%) oraz testujący (10%) z wykorzystaniem funkcji *train_test_split* z pakietu *sklearn* w sposób losowy,
2. Podział zbioru treningowego na 5 części z wykorzystanie walidacji krzyżowej do trenowania modelu (4 części wykorzystywane były do trenowania, natomiast 5 część wykorzystywana była do walidacji modelu)
3. Wybór modelu, który uzyskał najwyższą dokładność na zbiorze walidującym w i-tej iteracji,
4. Sprawdzenie modelu na zbiorze testującym

Eksperymenty z wykorzystaniem lasu losowego, zmodyfikowanego lasu losowego i drzew decyzyjnych ID3 powtarzane były 25 razy, natomiast w przypadku naiwnego klasyfikatora bayesowskiego zdecydowano się na 50-krotne powtórzenie powyższej sekwencji. Następnie rezultaty były zapisywane w celu analizy skuteczności i charakterystyk algorytmów.

Wyniki zostały uśrednione i przedstawione w formie macierzy pomyłek, krzywej charakterystyki operacyjnej odbiornika (ROC - Receiver Operating Characteristics), średniej precyzji i średniej skuteczności z jej maksimum, minimum oraz odchyleniem standardowym.

W przypadku Drzewa ID3 z ograniczeniem głębokości wykonywany jest iteracyjny test sprawdzający skuteczność różnych głębokości drzewa. Do eksperymentu wybierana jest jedna głębokość. Sam eksperyment porównujący różne głębokości też jest udokumentowany w formie wykresu dokładności od głębokości, zarówno dla zbioru treningowego jak i testującego. Przyjęto, że sprawdzane wartości będą mieścić się w zakresie $[0; 10]$.

Liczba uczonych modeli w ramach jednego eksperymentu jest równa od 125 do 250. W każdej iteracji spośród 5 modeli wybierany był ten, który uzyskiwał najwyższą dokładność na zbiorze walidacyjnym. Z tego powodu najdłuższe eksperymenty trwały nawet ok. 3 godzin.

Zbiór Mushroom ze względu na swoją trywialność był wykorzystywany tylko na etapie testów implementacji modelów, a ostatecznie na potrzeby eksperymentów zastąpiono go zbiorem Credit Score. Ponadto niekonkluzywność zbioru Healthcare (którą zobaczyć będzie można w wynikach eksperymentów) wymusiła znalezienie jeszcze jednego zbioru. Tym zbiorem jest zbiór CDC Diabetes Health Indicators z repozytorium UCI ML.

Przed wykonaniem eksperymentów dane uległy wstępnej obróbce, która polegała na wyselekcjonowaniu użytecznych atrybutów, kubełkowaniu atrybutów o charakterze ciągłym, zamianie atrybutów o charakterze nominalnym na kodowanie za pomocą gorącej jedynki oraz zamianę formatu klasyfikacji ze *string* na *int* w zakresie $\{0, 1\}$ lub $\{0, 1, 2\}$.

3.1. Zbiór Wine

3.1.1. O danych

Zbiór Wine składa się z 11 atrybutów będących chemicznymi właściwościami danego wina. Domyślna klasyfikacja jest oceną w skali 1-10 jakości wina. W zbiorze wine korzystamy ze wszystkich dostępnych atrybutów. Dane układają się zgodnie z rozkładem Gaussa. Jedyną zmianą w stosunku do oryginalnych danych jest dyskretyzacja klasyfikacji ze skali $1-10$ do skali $\{0,1\}$ gdzie 1 przypisywane jest dla win z oceną większą niż 5, natomiast reszta win klasyfikowana jest jako 0. Zbiór ten jako jedyny nie został przez nas idealnie zbalansowany - Stosunek win o klasyfikacji 1 do win o klasyfikacji 0 jest równy 7:5. Poniżej przedstawiono tabelę opisującą zbiór, na którym przeprowadzono eksperymenty

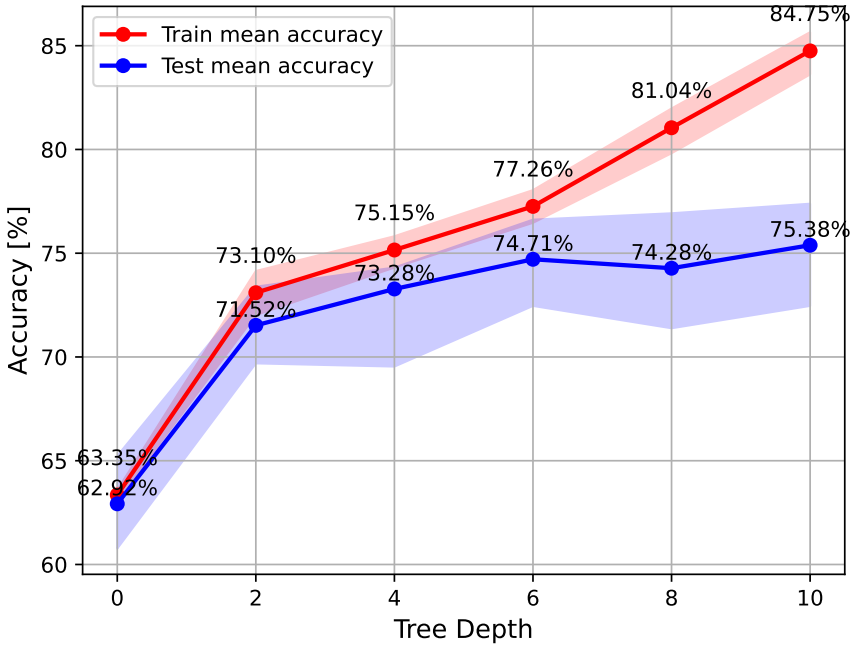
Tab. 3.1: Podstawowe informacje o zbiorze

Liczba próbek	**Liczba klas**	**Liczba atrybutów**
6497	2	12

Tab. 3.2: Atrybuty próbek ze zbioru Wine Quality

Polska nazwa atrybutu	**Angielska nazwa atrybutu**	**Typ atrybutu**
kwasowość stała	fixed acidity	ciągły
kwasowość ulotna	volatile acidity	ciągły
zawartość kwasu cytrynowego	citric acid	ciągły
zawartość cukru resztkowego	residual sugar	ciągły
zawartość chlorków	chlorides	ciągły
wolnego dwutlenku siarki SO_2	free sulfur dioxide	ciągły
ilość całkowitego dwutlenku siarki SO_2	total sulfur dioxide	ciągły
gęstość	density	ciągły
pH	pH	ciągły
zawartość siarczanów	sulphates	ciągły
zawartość alkoholu	alcohol	ciągły
kolor	color	binarny

3.1.2. Wybór maksymalnej głębokości dla drzewa ID3



Rys. 3.1: Wykres skuteczności drzewa ID3 od maksymalnej głębokości dla zbioru Wine

Na podstawie wykresu przedstawiającego zależność skuteczności na zbiorze trenującym i testującym od maksymalnej głębokości drzewa podjęto decyzję o wykorzystaniu w eksperymentach z ID3 maksymalnej głębokości równej 6. Dla większych wartości skuteczność na zbiorze testującym nie zwiększała się znacznie przy znacznym wzroście skuteczności na zbiorze treningowym, co jest klasycznym przypadkiem overfittingu.

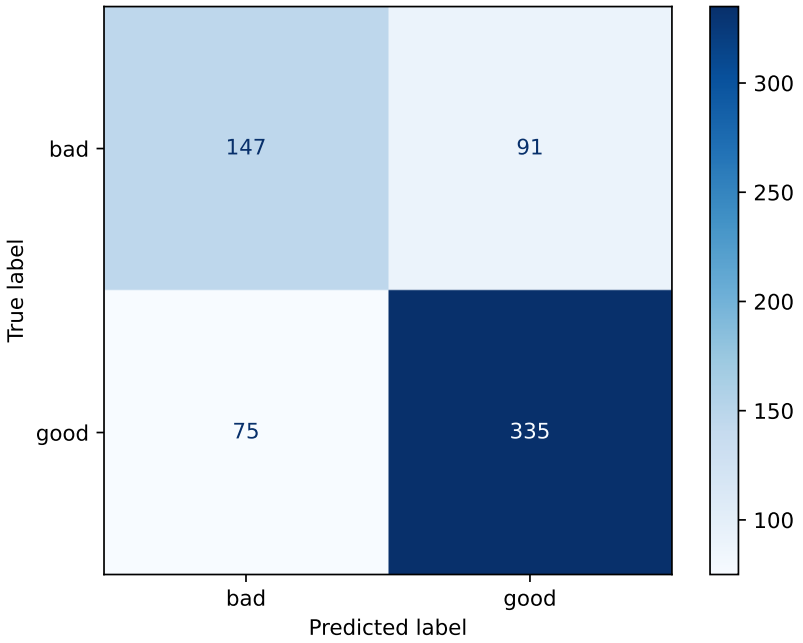
3.1.3. ID3 z ograniczeniem

Tab. 3.3: Procentowa dokładność klasyfikacji; drzewo ID3; 25 powtórzeń

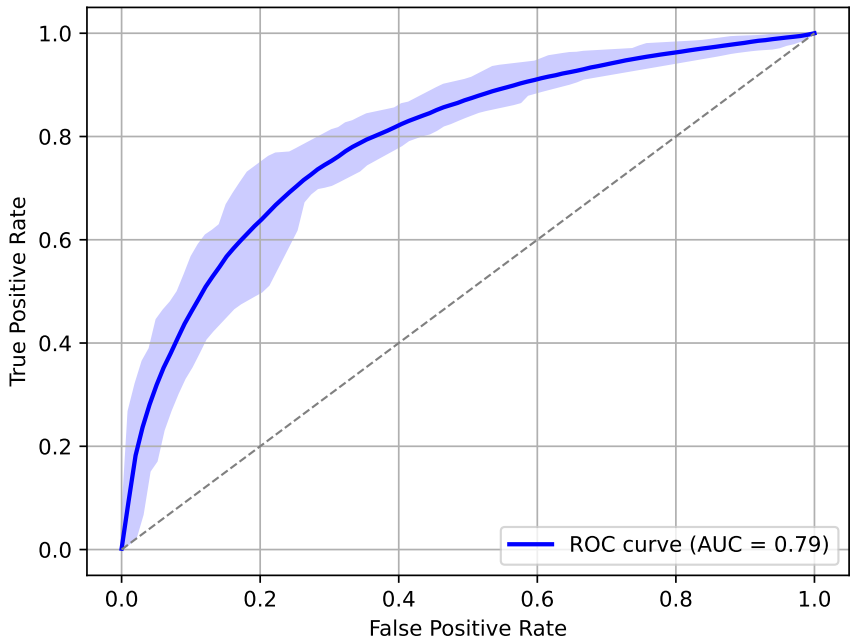
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	77,51	74,28
Odchylenie standardowe	0,46	1,39
Maksimum	78,18	77,23
Minimum	76,30	71,54

Tab. 3.4: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Wine (drzewo ID3, 25 powtórzeń)

Parametr	Wartość
Precyzja	0,787
True Positive Rate	0,817
False Positive Rate	0,382



Rys. 3.2: Uśredniona macierz pomyłek; drzewo ID3; 25 powtórzeń



Rys. 3.3: Uśredniona ROC; drzewo ID3; 25 powtórzeń

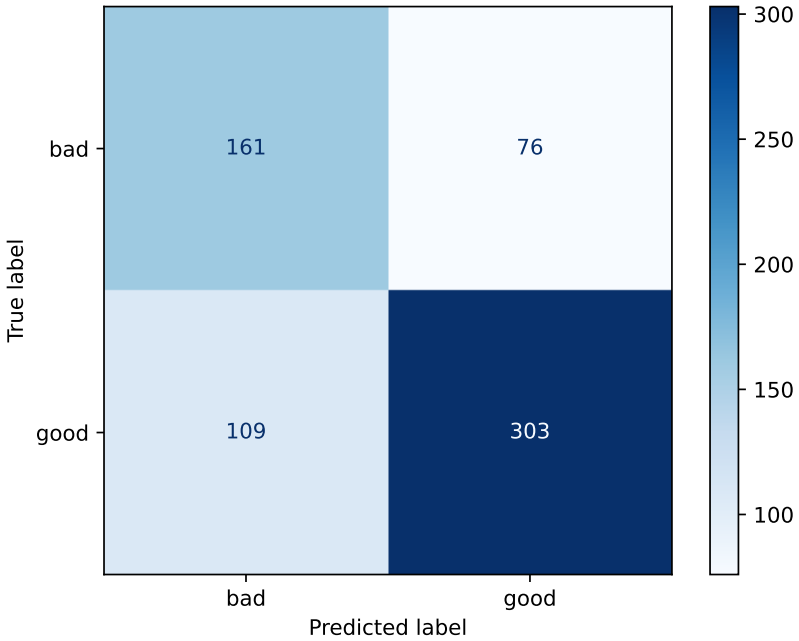
3.1.4. Naiwny klasyfikator Bayesa

Tab. 3.5: Procentowa dokładność klasyfikacji; Naiwny klasyfikator Bayesa; 50 powtórzeń

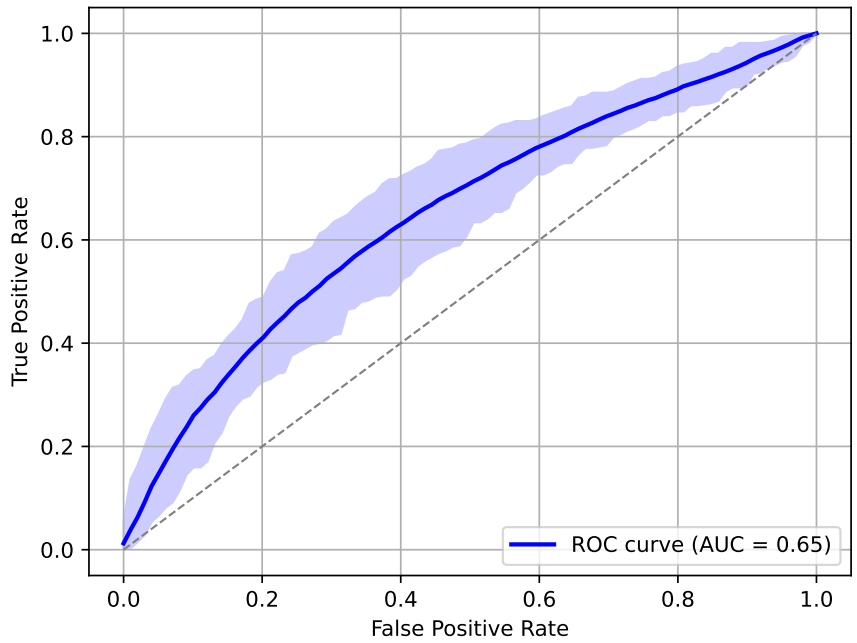
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	74,69	71,44
Odchylenie standardowe	0,33	1,75
Maksimum	75,73	75,38
Minimum	74,14	68,15

Tab. 3.6: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Wine (Naiwny klasyfikator Bayesa, 50 powtórzeń)

Parametr	Wartość
Precyzja	0,80
True Positive Rate	0,74
False Positive Rate	0,32



Rys. 3.4: Uśredniona macierz pomyłek; Naiwny klasyfikator Bayesa; 50 powtórzeń



Rys. 3.5: Uśredniona ROC; Naiwny klasyfikator Bayesa; 50 powtórzeń

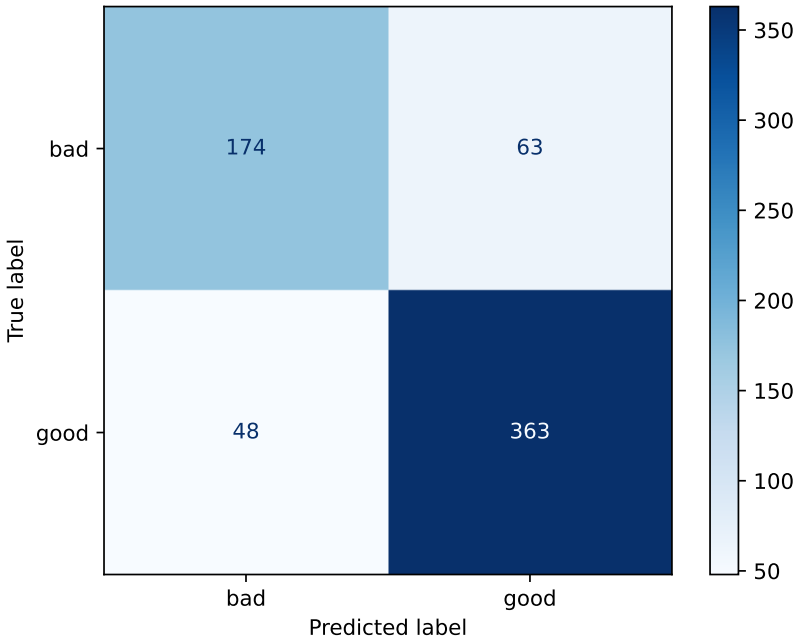
3.1.5. Klasyczny las losowy

Tab. 3.7: Procentowa dokładność klasyfikacji; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

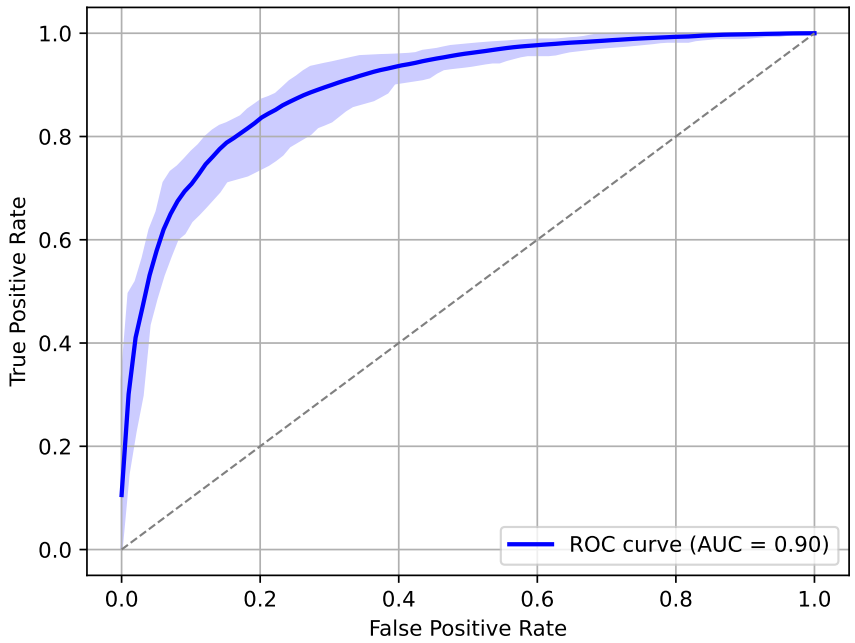
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	96,71	82,80
Odchylenie standardowe	0,13	1,25
Maksimum	97,06	85,08
Minimum	96,51	79,08

Tab. 3.8: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Wine (klasyczny las losowy, 50 drzew ID3, 25 powtórzeń)

Parametr	Wartość
Precyzja	0,85
True Positive Rate	0,88
False Positive Rate	0,27



Rys. 3.6: Uśredniona macierz pomyłek; klasyczny las losowy (50 drzew ID3); 25 powtórzeń



Rys. 3.7: Uśredniona ROC; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

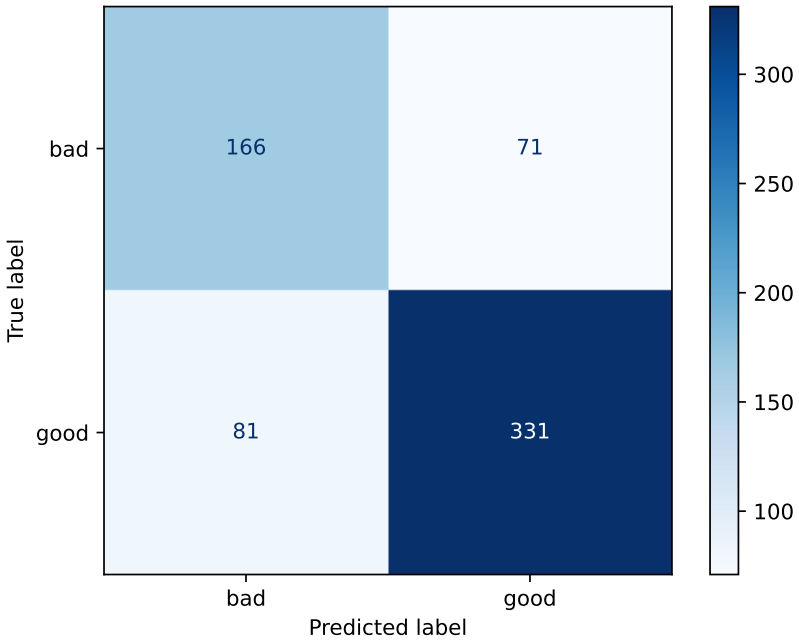
3.1.6. Zmodyfikowany las losowy ID3/NBC

Tab. 3.9: Procentowa dokładność klasyfikacji; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

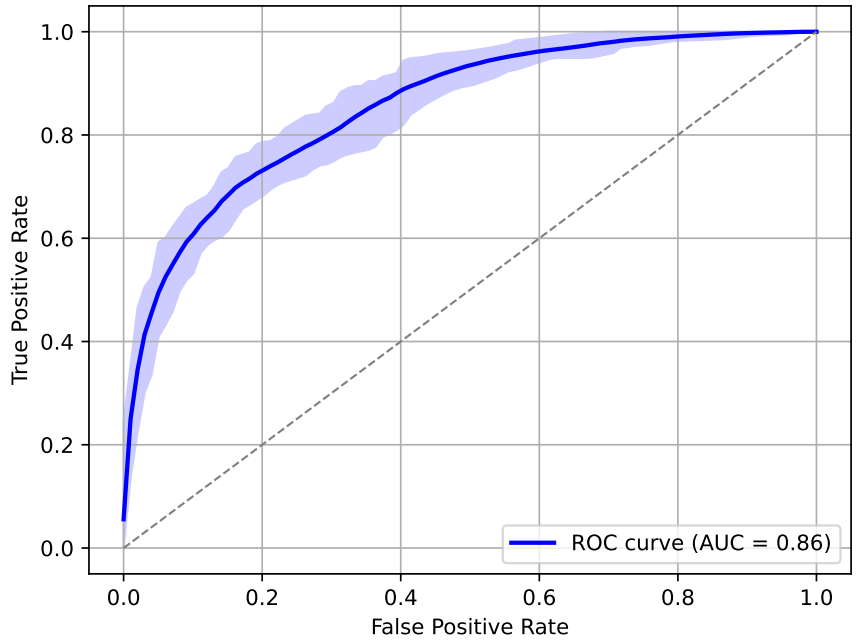
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	83,85	76,53
Odchylenie standardowe	0,35	1,84
Maksimum	84,64	80,46
Minimum	83,34	72,00

Tab. 3.10: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Wine (zmodyfikowany las losowy, 25 drzew ID3 i 25 NBC, 25 powtórzeń)

Parametr	Wartość
Precyzja	0,82
True Positive Rate	0,80
False Positive Rate	0,30



Rys. 3.8: Uśredniona macierz pomyłek; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń



Rys. 3.9: Uśredniona ROC; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

3.1.7. Podsumowanie

Tab. 3.11: Zestawienie dokładności klasyfikacji na zbiorze Wine dla różnych modeli (na zbiorze testującym)

Model	Średnia	Maksimum	Minimum	Odchylenie standardowe
Drzewo decyzyjne ID3	74,28	77,23	71,54	1,39
Naiwny klasyfikator Bayesa	71,44	75,38	68,15	1,75
Las losowy (50 drzew ID3)	82,80	85,08	79,08	1,25
Las losowy (25 ID3 / 25 NBC)	76,53	80,46	72,00	1,84

Z eksperymentów przeprowadzonych na zbiorze Wine można wyciągnąć całkiem interesujące wnioski. Z danych przedstawionych w tabeli 3.44 widać, że najlepszą skutecznością cechuje się standardowy las losowy. Jednocześnie, co widać w tabeli 3.7 występuje w nim największe nadmierne dopasowanie, które objawia się znacząco lepszą skutecznością modelu na zbiorze treningowym w porównaniu do zbioru testowego. Zmodyfikowany las losowy oferuje natomiast zauważalnie gorszą skuteczność klasyfikacji i największe odchylenie standardowe, ale co ciekawe zachowuje wysoki wynik współczynnika prawdziwych pozytywnych (w porównaniu do klasycznego lasu) i posiada większą odporność na nadmierne dopasowanie.

Analiza krzywych ROC potwierdza wnioski wyciągnięte z wstępnej analizy danych tabelarycznych. Najlepiej z klasyfikowaniem dobrych win radzi sobie klasyczny las losowy, co widać na rys. 3.9. Najlepszy i najgorszy model zaznaczone na wykresie nieznacznie odbiegają od ukazanej średniej. Najgorszym modelem jest natomiast naiwny klasyfikator bayesowski, który radzi sobie lepiej od klasyfikatora losowego, ale jego średnia charakterystyka jest mocno spłaszczona. Zauważalna jest także znacząca różnica pomiędzy modelem najlepszym i najgorszym.

W kontekście stosunku jakości modelu do jego złożoności obliczeniowej najlepiej wypada klasyczne drzewo decyzyjne, które oferuje oczywiście gorszą skuteczność i większy błąd, ale jednocześnie wymaga znacznie krótszego czasu trenowania.

Na podstawie wyników można dostrzec dwa "kierunki" powodujące zwiększenie skuteczności klasyfikacji:

1. Zmiana modeli pojedynczych na multimodele - Widać, skuteczność modeli NBC oraz ID3 jest mniejsza od skuteczności obu lasów. Pojedyncze modele ze względu na dane, na których zostały wytrenowane mogą nie być wrażliwe na niektóre zależności. Jeżeli jednak odpowiednio dużo innych modeli te zależności wykrywa, to klasyfikacja może być poprawna mimo. W ten sposób niedoskonałości pojedynczych modeli są eliminowane.
2. Zamiana NBC na drzewa - Na przykładzie lasu widzimy, że las z mieszanką NBC i ID3 osiąga gorsze wyniki od lasu z samymi ID3. Analogicznie pojedyncze drzewa ID3 radzą sobie lepiej niż pojedyncze NBC. Pojawienie się takiego "kierunku" na przykładzie tego zbioru może świadczyć o tym, że atrybuty każdej w sposób zależny od siebie wpływają na wartość klasyfikacji. Na takie zależności NBC jest "ślepy".

3.2. Zbiór Healthcare

3.2.1. O danych

Dane Healthcare zawierają domyślnie 15 różnych parametrów. Pominięto atrybuty takie jak Imię, Identyfikatory, Nazwy Szpitala, Imię Doktora ze względu na zbyt dużą ilość różnych instancji w porównaniu do rozmiaru zbioru treningowego oraz nieistotność danych przy klasyfikacji. Pozostałe po selekcji (9 atrybutów) dane Healthcare rozkładają się bardzo równomiernie. Atrybuty ciągłe, wbrew oczekiwaniom nie układają się według rozkładu Gaussa, a przybierają rozkład zbliżony bardziej do jednostajnego. Atrybuty o charakterze nominalnym zostały zakodowane z pomocą gorącej jedynki. Po użyciu gorącej jedynki dane mają 29 atrybutów. Klasyfikacja tego zbioru jest trójwartościowa i jest przewidywanym stanem pacjenta: abnormal, normal lub inconclusive. Zbiór jest idealnie zbalansowany. Poniżej przedstawiono tabelę opisującą zbiór, na którym przeprowadzono eksperymenty

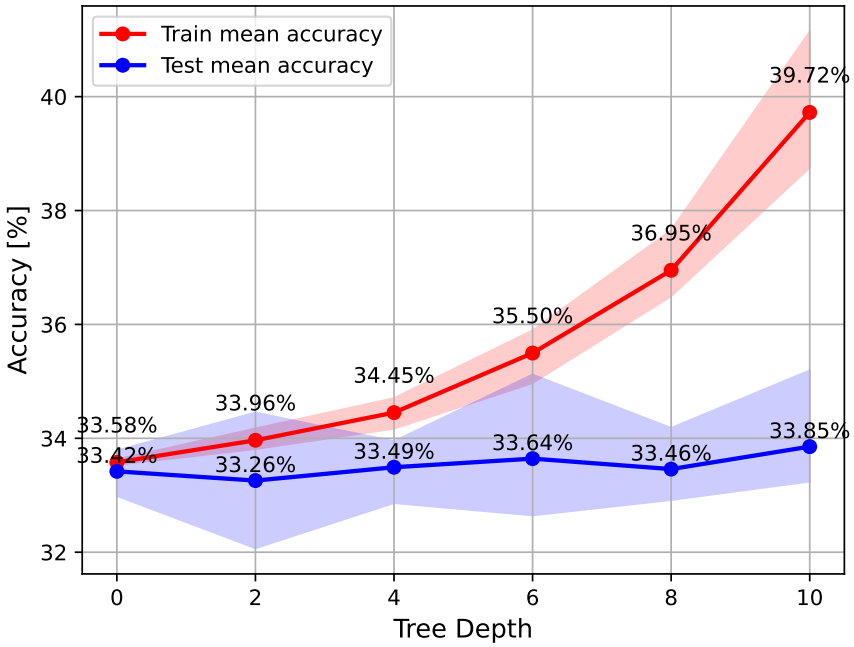
Tab. 3.12: Podstawowe informacje o zbiorze

Liczba próbek	**Liczba klas**	**Liczba atrybutów**
55000	3	9

Tab. 3.13: Atrybuty próbek ze zbioru Healthcare

Polska nazwa atrybutu	**Angielska nazwa atrybutu**	**Typ atrybutu**
wiek	age	dyskretny
płeć	gender	binarny
grupa krwi	blood type	nominalny (gorąca jedynka)
dolegliwość	medical condition	nominalny (gorąca jedynka)
ubezpieczyciel	insurance provider	nominalny (gorąca jedynka)
kwota rachunku	billing amount	ciągły
rodzaj przyjęcia	admission type	dyskretny
lek	medicatiom	nominalny (gorąca jedynka)
czas w szpitalu	time in hospital	dyskretny

3.2.2. Wybór ograniczenia do ID3



Rys. 3.10: Wykres skuteczności drzewa ID3 od maksymalnej głębokości dla zbioru Healthcare

W przypadku zbioru Healthcare zauważalne jest, że mimo zwiększania głębokości drzewa decyzyjnego jego skuteczność na danych testowych nie wzrastała i utrzymywała się na poziomie klasyfikatora losowego. Jedynym efektem takich działań było stopniowe dopasowywanie się klasyfikatora do zbioru trenującego. Zdecydowano więc w eksperymentach użyć drzewa decyzyjne z ograniczeniem głębokości do wartości 6.

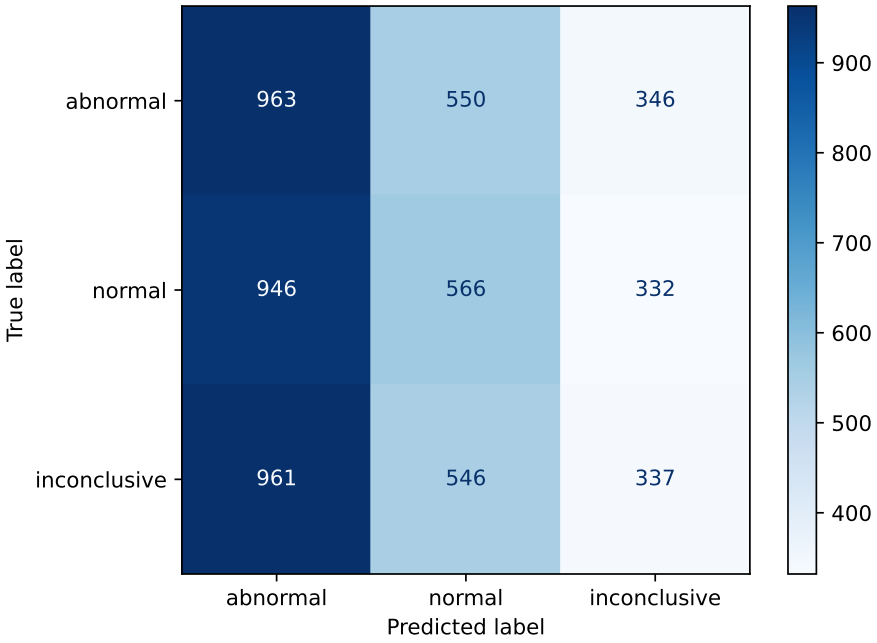
3.2.3. ID3 z ograniczeniem

Tab. 3.14: Procentowa dokładność klasyfikacji; drzewo ID3; 25 powtórzeń

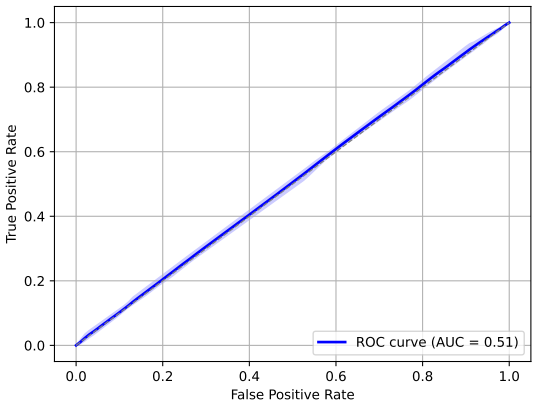
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	35,38	33,63
Odchylenie standardowe	0,31	0,79
Maksimum	35,86	34,65
Minimum	34,89	31,60

Tab. 3.15: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla poszczególnych klas (drzewo ID3, 25 powtórzeń)

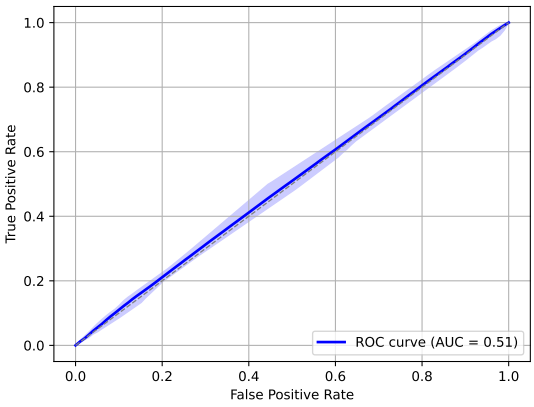
Klasa	Precyzja	T-P Rate	F-P Rate
Abnormal	0,34	0,52	0,52
Normal	0,34	0,31	0,30
Inconclusive	0,33	0,18	0,18



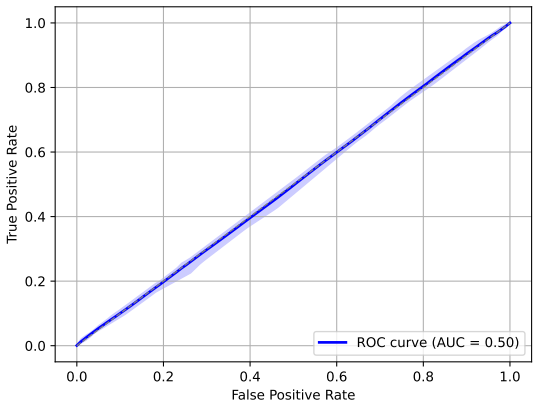
Rys. 3.11: Uśredniona macierz pomyłek; drzewo ID3; 25 powtórzeń



(a) Klasyfikacja Abnormal



(b) Klasyfikacja Normal



(c) Klasyfikacja Inconclusive

Rys. 3.12: Uśredniona ROC; drzewo ID3; 25 powtórzeń

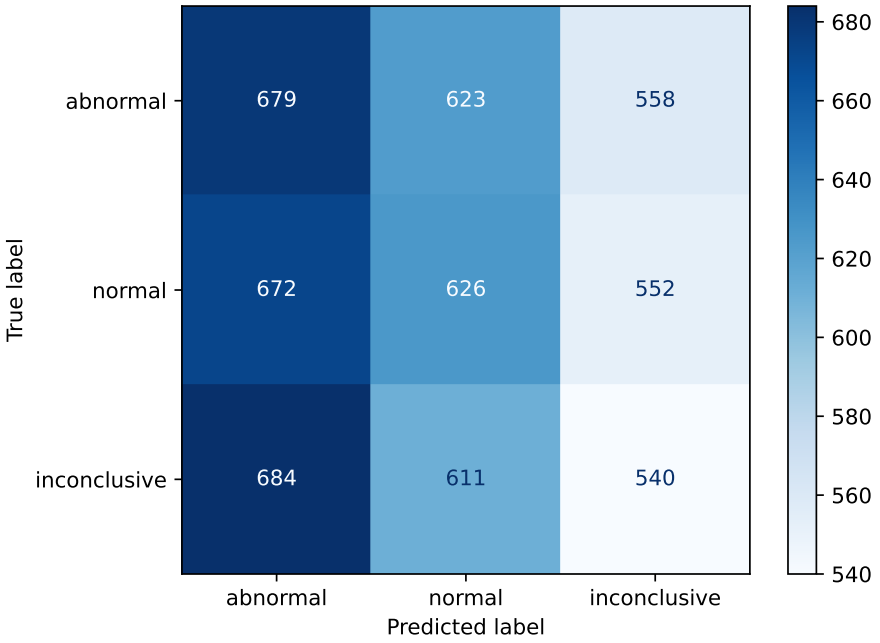
3.2.4. Naiwny klasyfikator Bayesa

Tab. 3.16: Procentowa dokładność klasyfikacji; Naiwny klasyfikator Bayesa; 25 powtórzeń

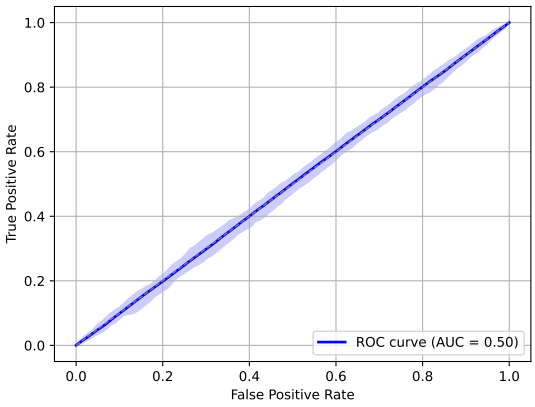
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	35,97	33,27
Odchylenie standardowe	0,14	0,57
Maksimum	36,35	34,67
Minimum	35,51	32,20

Tab. 3.17: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla poszczególnych klas (Naiwny klasyfikator Bayesa, 25 powtórzeń)

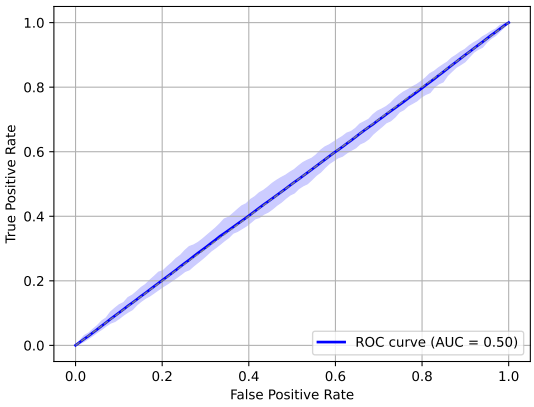
Klasa	Precyzja	T-P Rate	F-P Rate
Abnormal	0,33	0,37	0,37
Normal	0,34	0,34	0,33
Inconclusive	0,33	0,29	0,30



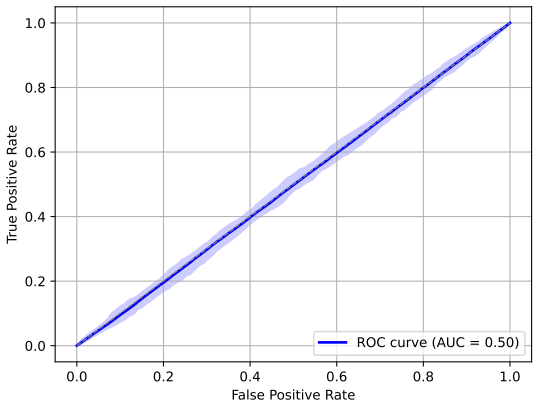
Rys. 3.13: Uśredniona macierz pomyłek; Naiwny klasyfikator Bayesa; 25 powtórzeń



(a) Klasyfikacja Abnormal



(b) Klasyfikacja Normal



(c) Klasyfikacja Inconclusive

Rys. 3.14: Uśredniona ROC; Naiwny klasyfikator Bayesa; 25 powtórzeń

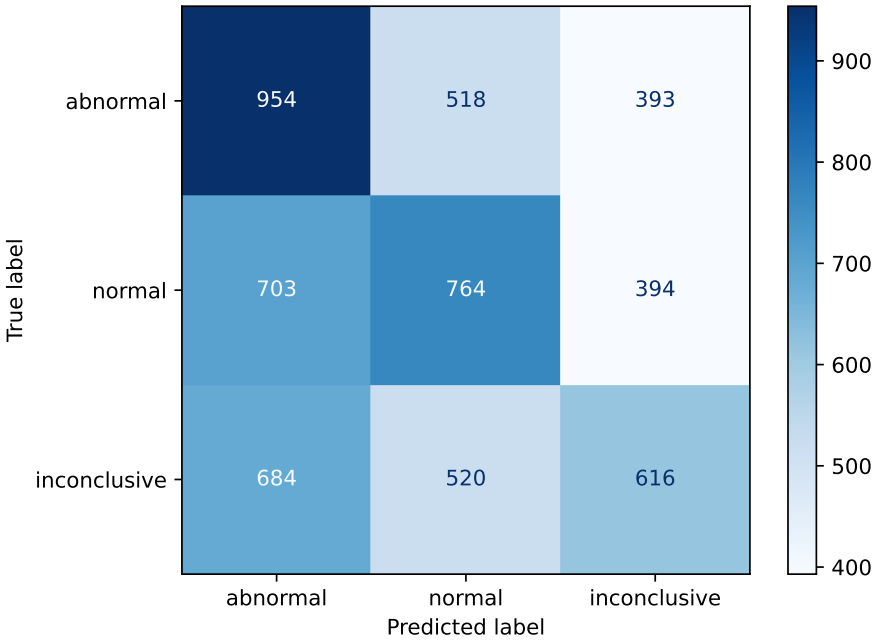
3.2.5. Klasyczny las losowy

Tab. 3.18: Procentowa dokładność klasyfikacji; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

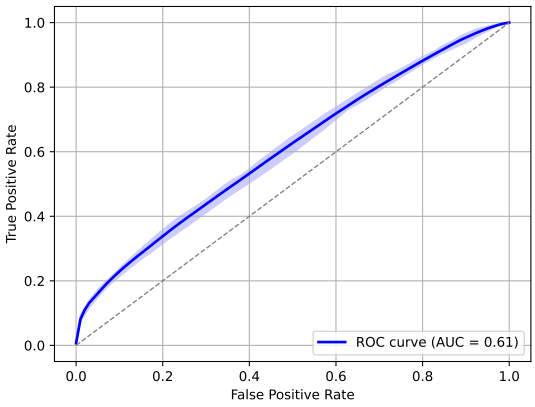
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	87,93	42,08
Odchylenie standardowe	0,15	0,58
Maksimum	88,16	42,76
Minimum	87,64	40,65

Tab. 3.19: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla poszczególnych klas (klasyczny las losowy z 50 drzew ID3, 25 powtórzeń)

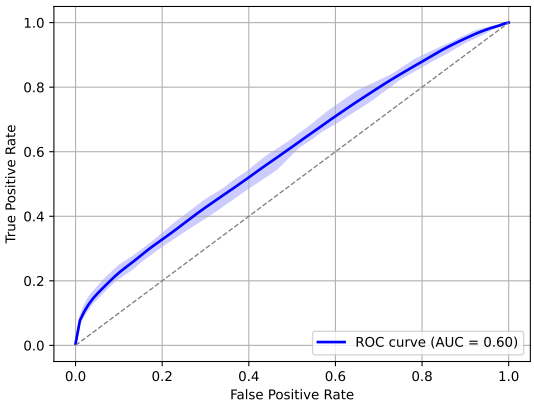
Klasa	Precyzja	T-P Rate	F-P Rate
Abnormal	0,41	0,51	0,38
Normal	0,42	0,41	0,28
Inconclusive	0,44	0,34	0,21



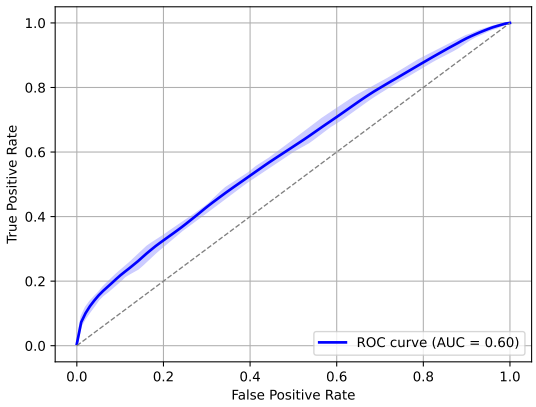
Rys. 3.15: Uśredniona macierz pomyłek; klasyczny las losowy (50 drzew ID3); 25 powtórzeń



(a) Klasyfikacja Abnormal



(b) Klasyfikacja Normal



(c) Klasyfikacja Inconclusive

Rys. 3.16: Uśredniona ROC; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

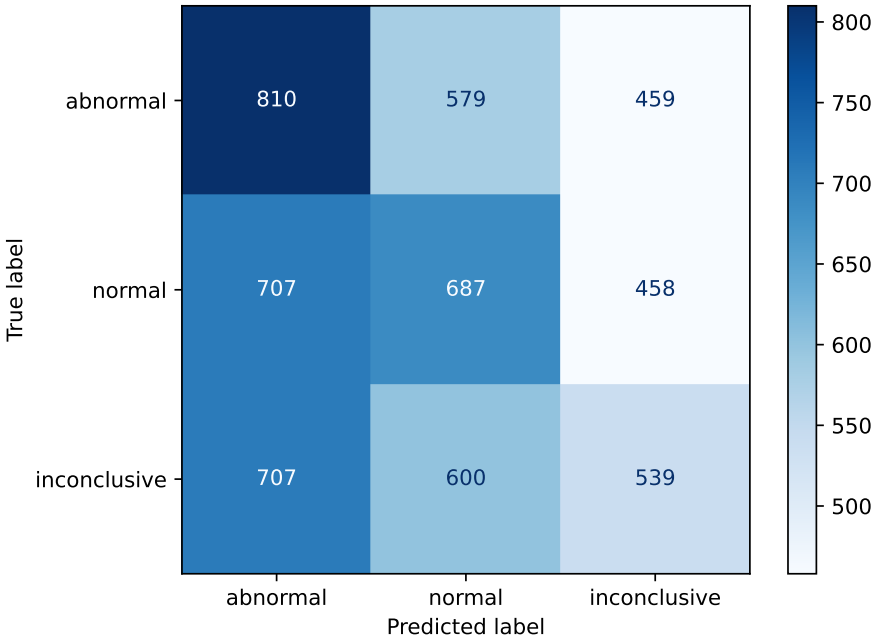
3.2.6. Zmodyfikowany las losowy ID3/NBC

Tab. 3.20: Procentowa dokładność klasyfikacji; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

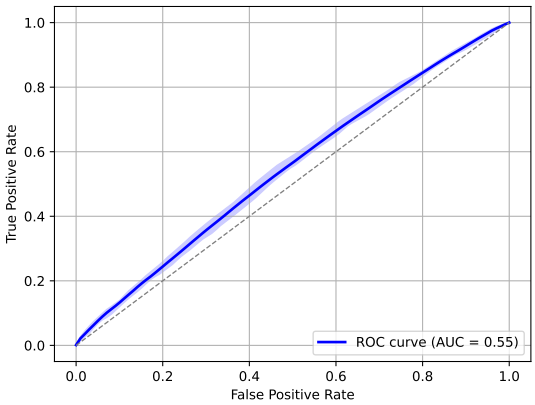
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	60,25	36,71
Odchylenie standardowe	1,38	0,66
Maksimum	62,47	37,68
Minimum	58,23	35,71

Tab. 3.21: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla poszczególnych klas (las losowy z 25 drzew ID3 i 25 NBC, 25 powtórzeń)

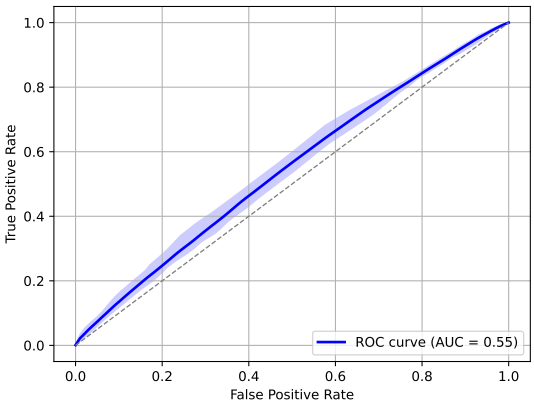
Klasa	Precyzja	T-P Rate	F-P Rate
Abnormal	0,36	0,44	0,38
Normal	0,37	0,37	0,32
Inconclusive	0,37	0,29	0,25



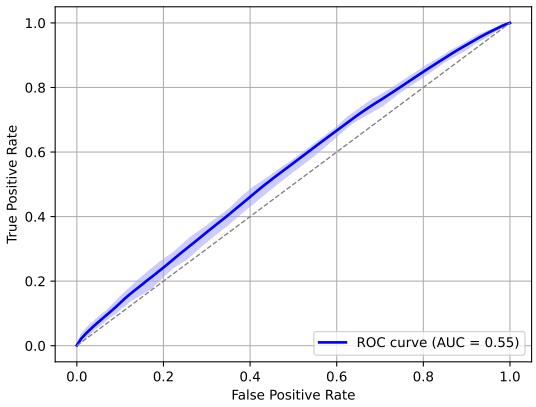
Rys. 3.17: Uśredniona macierz pomyłek; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń



(a) Klasyfikacja Abnormal



(b) Klasyfikacja Normal



(c) Klasyfikacja Inconclusive

Rys. 3.18: Uśredniona ROC; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

3.2.7. Podsumowanie

Tab. 3.22: Zestawienie dokładności klasyfikacji na zbiorze Healthcare dla różnych modeli (na zbiorze testującym)

Model	Średnia	Maksimum	Minimum	Odchylenie standardowe
Drzewo decyzyjne ID3	33,63	34,65	31,60	0,79
Naiwny klasyfikator Bayesa	33,27	34,67	32,20	0,57
Las losowy (50 drzew ID3)	42,08	42,76	40,65	0,58
Las losowy (25 ID3 / 25 NBC)	36,71	37,68	35,71	0,66

Wyniki eksperymentów na zbiorze Healthcare różnią się znacząco od wyników poprzedniego eksperymentów. Pierwszy wniosek przychodzący do głowy jest następujący - Zbiór Healthcare okazał się być mało konkluzywny. W przypadku NBC oraz drzewa ID3 skuteczność klasyfikacji była porównywalna z losową klasyfikacją (około 33%), natomiast najlepsze klasyfikatory (w tym przypadku las ID3) osiągnął średnią skuteczność na poziomie 44,03% co jest też bardzo słabym wynikiem. Stwierdzenie o niekonkluzywności (albo raczej małej konkluzywności) danych oparte jest nie tylko na słabych wynikach otrzymanych w powyżej przedstawionych eksperymentach i niewiarygodnych rozkładach parametrów ciągłych, ale również na podstawie dodatkowych eksperymentów przeprowadzonych z wykorzystaniem klasyfikatorów z pakietu sklearn (była to odpowiedź na podejrzenie błędnej implementacji klasyfikatorów, jednak osiągnięcie przez klasyfikatory z sklearn, uznawane za poprawnie zaimplementowane, podobnej skuteczności pomogło rozwiać wątpliwości) oraz na wynikach projektów wykonanych na tym samym zbiorze przez użytkowników Kaggle (fakt, że dużej części ML-owego społeczeństwa wychodziły podobnie słabe wyniki zwiększył przekonanie o poprawności zaimplementowanych klasyfikatorów).

Na zbiorze tym zaobserwowano wyraźnie zjawisko overfittingu. W szczególności na przykładzie klasycznego lasu losowego ID3 widać wyraźnie zjawisko przetrenowania - na zbiorze treningowym skuteczność sięga 87,93%, natomiast na zbiorze testowym 42,08%. Gdyby eksperyment był przeprowadzany za każdym razem z takim samym podziałem danych na zbiór testowy i trenujący, można by było postulować o pewnych wzorach zależności między atrybutami, które występują w zbiorze trenującym i nie występują, bądź są niedoreprezentowane w zbiorze testującym. Ponieważ jednak podział na dane treningowe i testujące za każdym razem wykonywany jest w sposób losowy, inny od siebie, takie twierdzenie jest nieuprawnione.

3.3. Zbiór Credit Score

3.3.1. O danych

Dane Credit Score zawierają 27 atrybutów.Wybrano rozkładają się bardzo zgodnie z rozkładem Gaussa z niewielkimi odstępami na końcach zakresów. Pominięto atrybuty takie jak Imię, Identyfikatory, Numery karty/Ubezpieczenia ze względu na zbyt dużą ilość różnych instancji w porównaniu do rozmiaru zbioru treningowego oraz nieistotność. Ponadto wykreślono rzędy, w których występowały braki atrybutów lub błędnie wprowadzone dane. Wartości ciągłe zamieniono na dyskretne zgodnie z percentylowym rozkładem atrybutu w taki sposób, żeby każdy pierwotnie ciągły atrybut mógł przyjąć jedną z 50 wartości. Klasyfikacja tego zbioru jest trójwartościowa i jest oceną zdolności kredytowej: poor, standard, good. Zbiór jest idealnie zbalansowany. Poniżej przedstawiono tabelę opisującą zbiór, na którym przeprowadzono eksperymenty

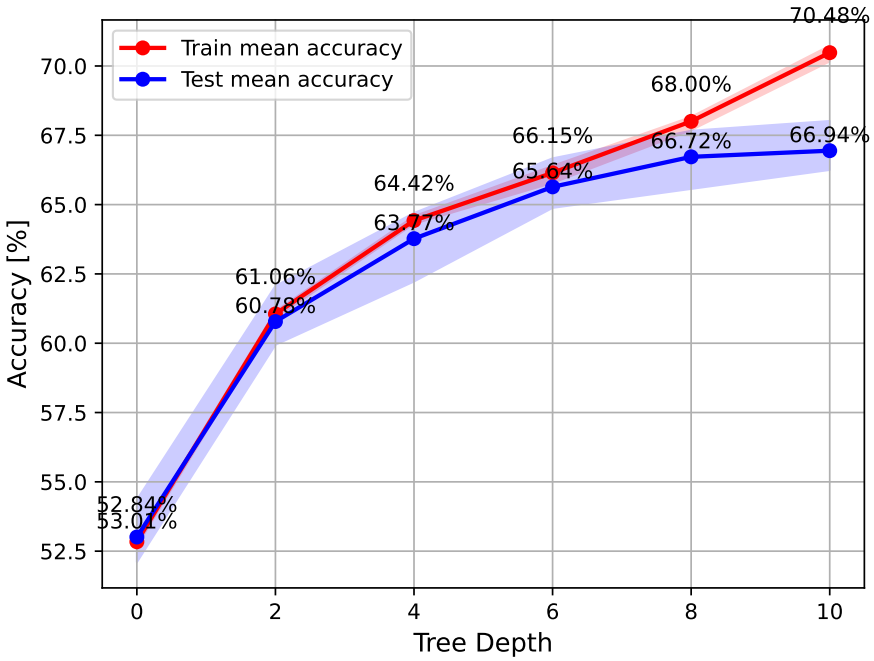
Tab. 3.23: Podstawowe informacje o zbiorze

Liczba próbek	**Liczba klas**	**Liczba atrybutów**
24645	3	14

Tab. 3.24: Atrybuty próbek ze zbioru Credit Score

Polska nazwa atrybutu	**Angielska nazwa atrybutu**	**Typ atrybutu**
wiek	age	dyskretny
roczny przychód	annual income	ciągły (percentyle)
miesięczna wypłata	monthly inhand salary	ciągły (percentyle)
stopa procentowa	interest rate	ciągły (percentyle)
liczba pożyczek	number of loans	dyskretny
opóźnienie płatności	delay from due day (payment)	dyskretny
liczba spóźnionych płatności	number of delayed payments	dyskretny
liczba zapytań kredytowych	number of credit inquiries	dyskretny
dług	outstanding debt	ciągły (percentyle)
wskaźnik wykorzystania kredytu	credit utilization ratio	ciągły (percentyle)
wiek korzystania z usług kredytowych	credit history age	ciągły (percentyle)
całkowite miesięczne EMI	total EMI per month	ciągły (percentyle)
wartości miesięcznych inwestycji	amount invested monthly	ciągły (percentyle)
miesięczny bilans finansowy	monthly balance	ciągły (percentyle)

3.3.2. Wybór ograniczenia do ID3



Rys. 3.19: Wykres skuteczności drzewa ID3 od maksymalnej głębokości dla zbioru Credit Score

Na podstawie wykresu przedstawiającego zależność skuteczności na zbiorze trenującym i testującym od maksymalnej głębokości drzewa podjęto decyzję o wykorzystaniu w eksperymentach z ID3 maksymalnej głębokości równej 8. Parametr ten wybrano ze względu na wzrost skuteczności modelu na zbiorze trenującym, przy braku poprawy modelu na zbiorze testującym dla głębokości 10.

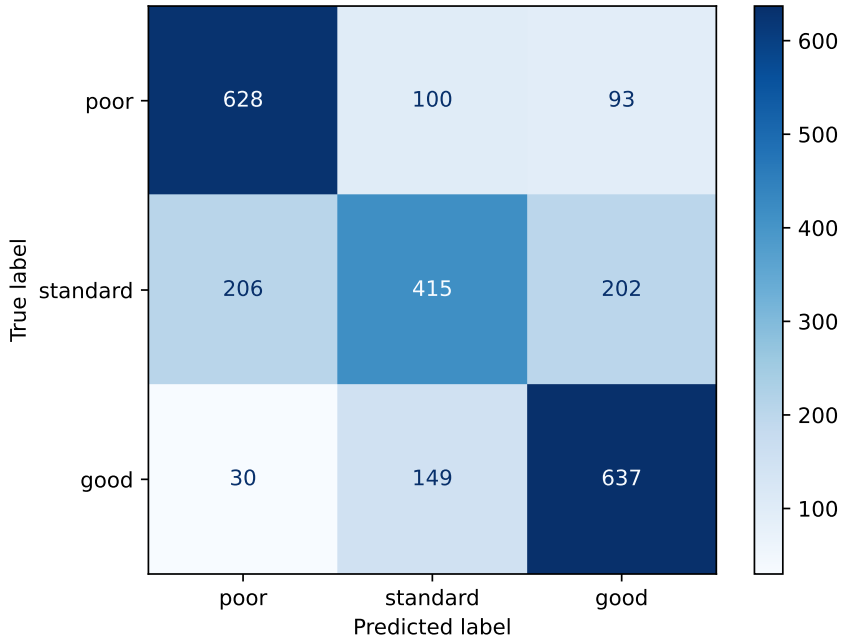
3.3.3. ID3 z ograniczeniem

Tab. 3.25: Procentowa dokładność klasyfikacji; drzewo ID3; 25 powtórzeń

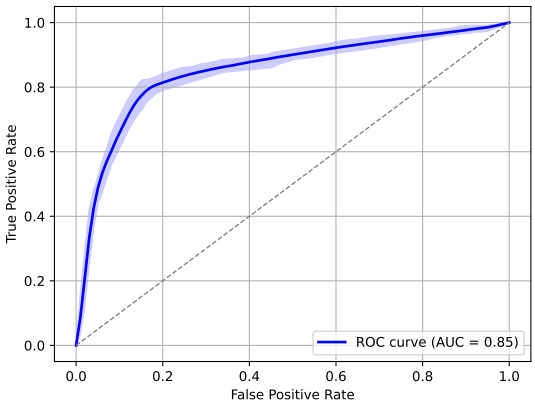
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	70,57	68,22
Odchylenie standardowe	0,22	1,04
Maksimum	70,96	71,44
Minimum	70,16	66,65

Tab. 3.26: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Credit Score (drzewo ID3, 25 powtórzeń)

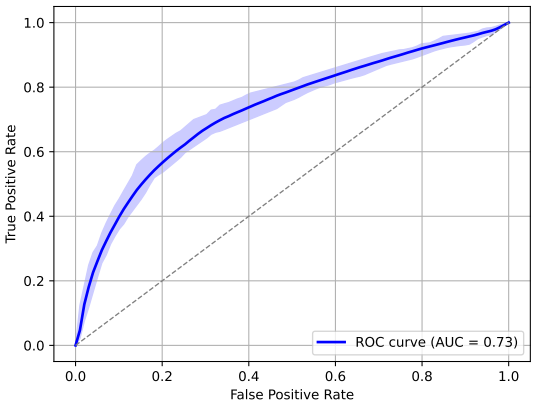
Klasa	Precyzja	T-P Rate	F-P Rate
Poor	0,73	0,76	0,15
Standard	0,63	0,50	0,15
Good	0,68	0,78	0,18



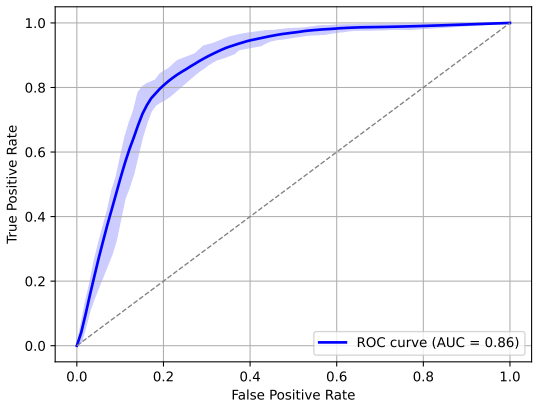
Rys. 3.20: Uśredniona macierz pomyłek; drzewo ID3; 25 powtórzeń



(a) Klasyfikacja Poor



(b) Klasyfikacja Standard



(c) Klasyfikacja Good

Rys. 3.21: Uśredniona ROC; drzewo ID3; 25 powtórzeń

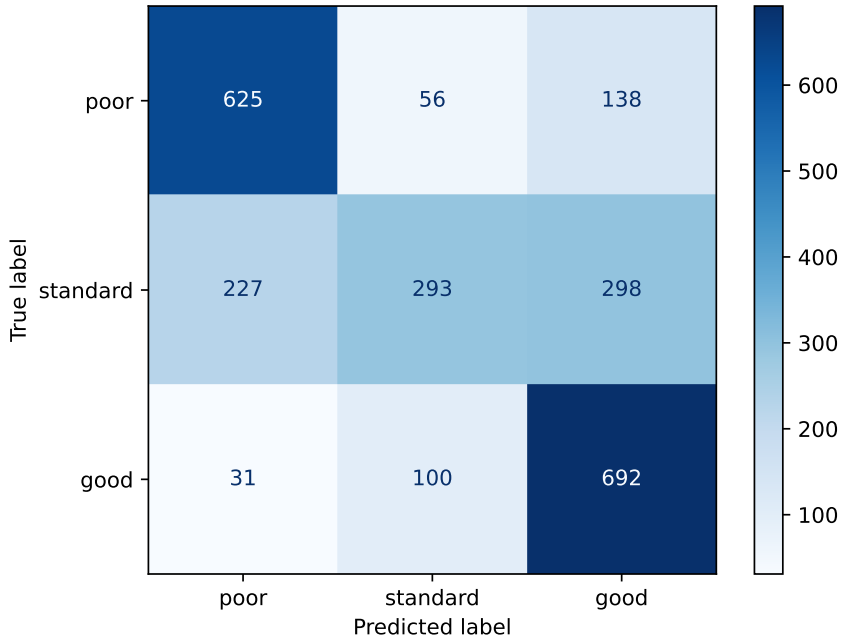
3.3.4. Naiwny klasyfikator Bayesa

Tab. 3.27: Procentowa dokładność klasyfikacji; Naiwny klasyfikator Bayesa; 25 powtórzeń

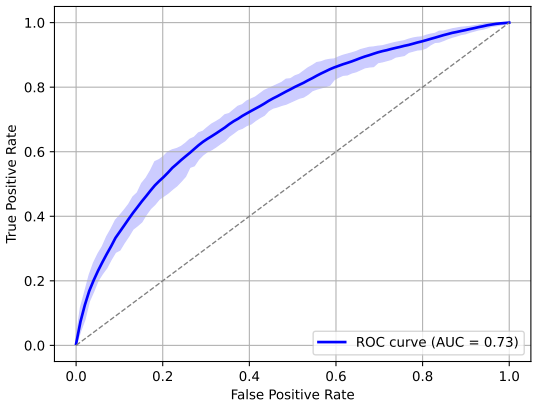
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	65,49	65,37
Odchylenie standardowe	0,12	1,00
Maksimum	65,74	67,38
Minimum	65,28	62,76

Tab. 3.28: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Credit Score (Naiwny klasyfikator Bayesa, 50 powtórzeń)

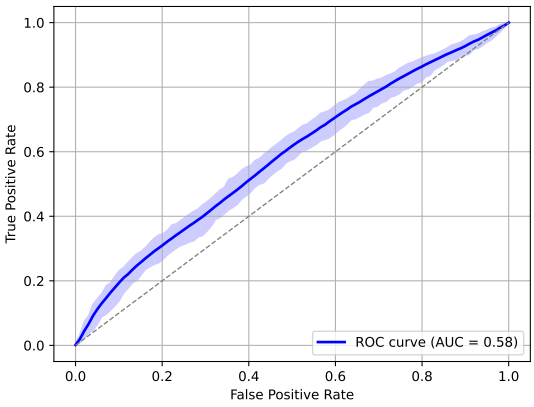
Klasa	Precyzja	T-P Rate	F-P Rate
Poor	0,71	0,76	0,16
Standard	0,65	0,36	0,10
Good	0,61	0,84	0,27



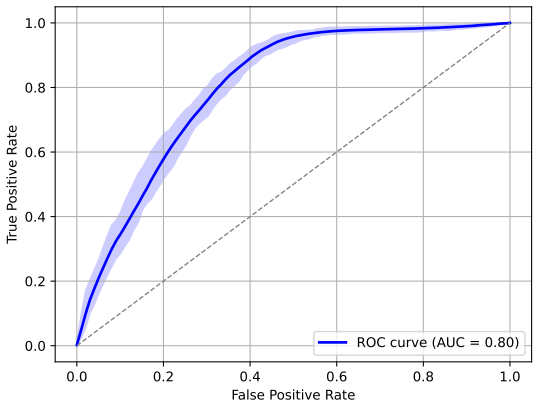
Rys. 3.22: Uśredniona macierz pomyłek; Naiwny klasyfikator Bayesa; 25 powtórzeń



(a) Klasyfikacja Poor



(b) Klasyfikacja Standard



(c) Klasyfikacja Good

Rys. 3.23: Uśredniona ROC; Naiwny klasyfikator Bayesa; 25 powtórzeń

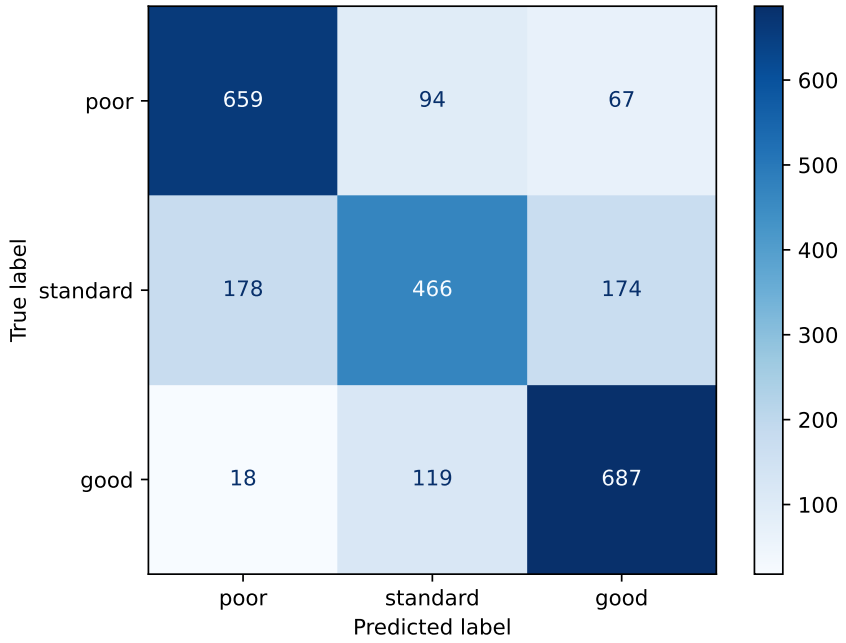
3.3.5. Klasyczny las losowy

Tab. 3.29: Procentowa dokładność klasyfikacji; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

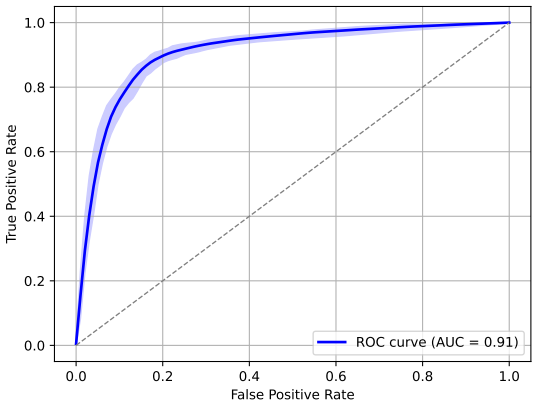
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	94,80	73,54
Odchylenie standardowe	0,09	0,65
Maksimum	95,02	74,85
Minimum	94,68	72,41

Tab. 3.30: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Credit Score (klasyczny las losowy, 50 drzew ID3, 25 powtórzeń)

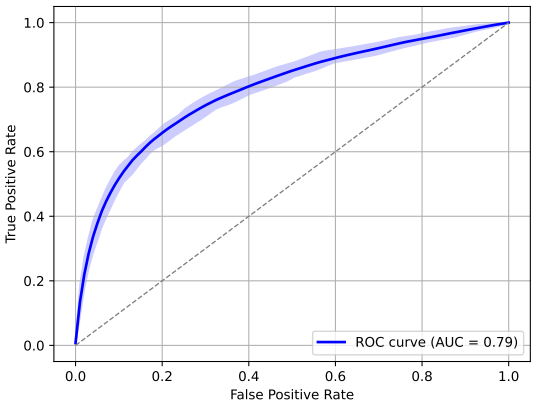
Klasa	Precyzja	T-P Rate	F-P Rate
Poor	0,77	0,80	0,12
Standard	0,69	0,57	0,13
Good	0,74	0,83	0,15



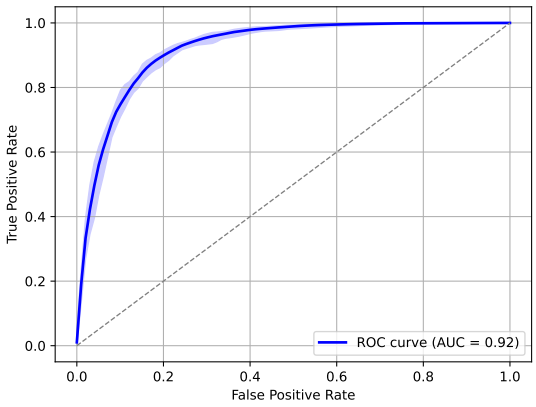
Rys. 3.24: Uśredniona macierz pomyłek; klasyczny las losowy (50 drzew ID3); 25 powtórzeń



(a) Klasyfikacja Poor



(b) Klasyfikacja Standard



(c) Klasyfikacja Good

Rys. 3.25: Uśredniona ROC; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

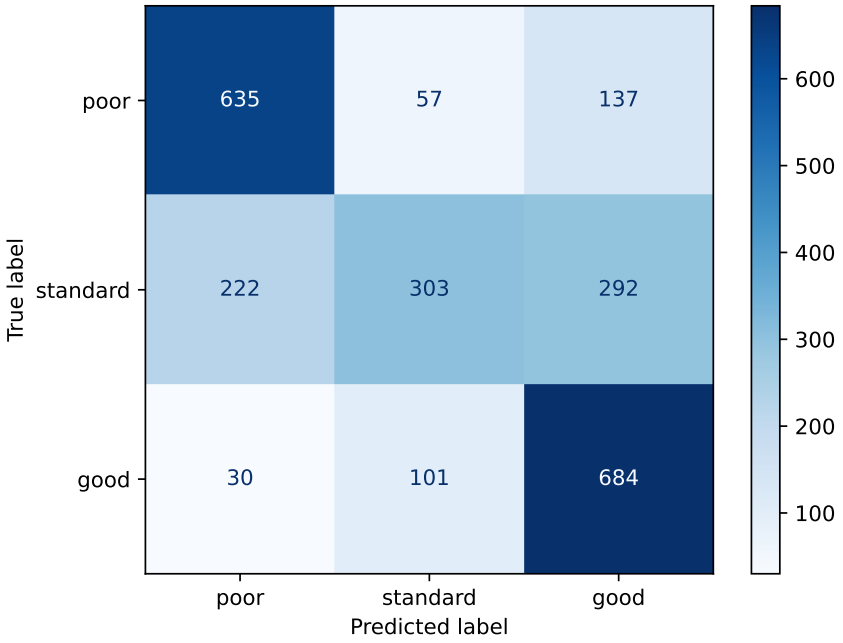
3.3.6. Zmodyfikowany las losowy ID3/NBC

Tab. 3.31: Procentowa dokładność klasyfikacji; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

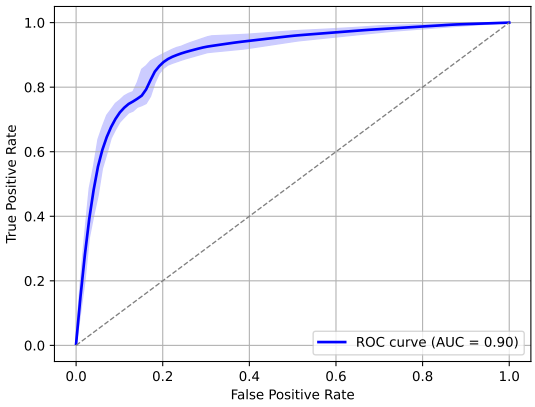
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	67,35	65,86
Odchylenie standardowe	0,20	1,08
Maksimum	67,69	68,48
Minimum	66,97	64,10

Tab. 3.32: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Credit Score (zmodyfikowany las losowy, 25 drzew ID3 i 25 NBC, 25 powtórzeń)

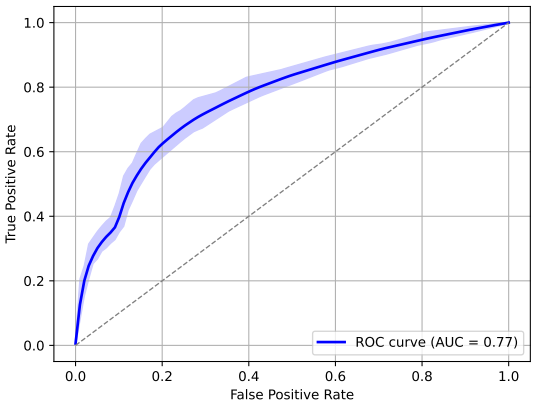
Klasa	Precyzja	T-P Rate	F-P Rate
Poor	0,72	0,77	0,15
Standard	0,66	0,37	0,10
Good	0,61	0,84	0,26



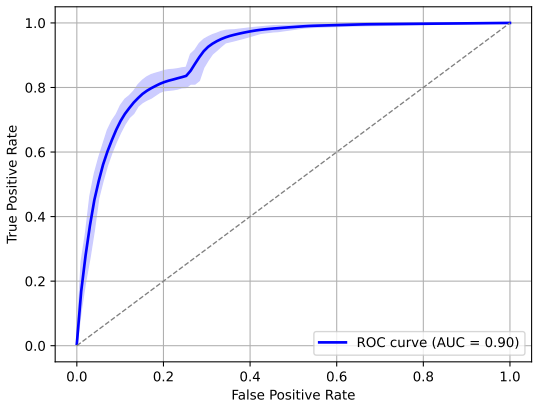
Rys. 3.26: Uśredniona macierz pomyłek; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń



(a) Klasyfikacja Poor



(b) Klasyfikacja Standard



(c) Klasyfikacja Good

Rys. 3.27: Uśredniona ROC; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

3.3.7. Podsumowanie

Tab. 3.33: Zestawienie dokładności klasyfikacji na zbiorze Credit Score dla różnych modeli (na zbiorze testującym)

Model	Średnia	Maksimum	Minimum	Odchylenie standardowe
Drzewo decyzyjne ID3	68,22	71,44	66,65	1,04
Naiwny klasyfikator Bayesa	65,37	67,38	62,76	1,00
Las losowy (50 drzew ID3)	73,54	74,85	72,41	0,65
Las losowy (25 ID3 / 25 NBC)	65,86	68,48	64,10	1,08

Wyniki eksperymentów na zbiorze Credit Score pokazują podobne zależności między modelami co wyniki eksperymentów na zbiorze Wine. Najbardziej podatny na przetrenowanie jest klasyczny las losowy ID3. Może to być jednak nie do końca zgodne z teorią i ogólną własnością multimodeli, które to powinny przetrenowanie minimalizować. Jednak biorąc pod uwagę, że w tym projekcie porównywany jest las drzew bez ograniczenia głębokości z pojedynczymi lasami o podanym ograniczeniu głębokości właściwość ta może faktycznie być odwrócona, gdyż drzewa bez ograniczenia głębokości w lesie trenowane są na próbach bootstrapowych, ale jednak podobnych do siebie, co sprawia, że drzewa te, są do siebie bardzo podobne. Zmniejszenie głębokości drzew w lesie klasycznym mogłoby zredukować zjawisko overfittingu. Natomiast las zmodyfikowany kosztem obniżenia skuteczności jest uodporniony na przetrenowanie. Dzieje się tak dzięki obecności Naiwnych klasyfikatorów Bayesa.

Na podstawie krzywych ROC oraz macierzy pomyłek, bardzo zauważalny jest fakt, że klasa standard jest o wiele rzadziej wybierana przez klasyfikatory od pozostałych klas. W szczególności dzieje się tak w przypadku Naiwnego klasyfikatora Bayesa lub zmodyfikowanego lasu, który to z kolei zawiera Naiwne klasyfikatory Bayesa. Można wnioskować, że instancje klasy standard są do siebie mniej podobne od instancji pozostałych klas, tj. ciężko wyróżnić charakterystyczne dla tej klasy właściwości/zależności.

W zbiorze Wine wybrano drzewo ID3 jako klasyfikator będący kompromisem między złożonością obliczeniową, a skutecznością. W tym przypadku można by zrobić podobnie, jednak koszt związany z obniżeniem skuteczności będzie o wiele większy, co może spowodować pozostanie przy wyborze klasycznego lasu losowego ID3, będącego najskuteczniejszym modelem.

Z wartości odchyleń standardowych oraz minimów i maksimów w skutecznościach oraz przebiegach krzywych ROC na pewno wynika duża powtarzalność eksperymentów i osiągnięcie przez modele dużej skuteczności, bliskiej granicy maksymalnej skuteczności jaką może osiągnąć model o danej architekturze.

Podobnie jak w przypadku zbioru Wine można dostrzec dwa "kierunki" powodujące wzrost skuteczności klasyfikacji. Są to:

1. Zmiana modeli pojedynczych na multimodele - Skuteczność zmodyfikowanego lasu ID3/NBC jest większa od skuteczności klasyfikatora NBC oraz Skuteczność klasycznego lasu ID3 jest większa od skuteczności pojedynczego drzewa ID3. "Kierunek" ten jest w przypadku zbioru Credit Score na tyle prominentny, że pomimo występowania również drugiego "kierunku ID3-NBC" skuteczność drzewa ID3 jest mniejsza od skuteczności zmodyfikowanego lasu losowego NBC/ID3
2. Zamiana NBC na drzewa - Skuteczność ID3 jest większa od skuteczności NBC, natomiast skuteczność klasycznego lasu ID3 jest większa od skuteczności zmodyfikowanego lasu ID3/NBC.

3.4. Zbiór Diabetes

3.4.1. O danych

Dane Diabetes zawierają 21 atrybutów będących informacjami o stylu życia danego człowieka. Nawet dane nie będące w rzeczywistości danymi liczbowymi (lecz o charakterze porządkowym) są w tym zbiorze reprezentowane przez liczby całkowite. Dane są binarne lub całkowite. Korzystamy ze wszystkich dostępnych atrybutów. Jedynym problemem w zbiorze było niezbalansowanie, więc usunięto losowo część z przykładów z klasyfikacją 0 aby liczba klasyfikacji 0 i 1 była równa. Poniżej przedstawiono tabelę opisującą zbiór, na którym przeprowadzono eksperymenty

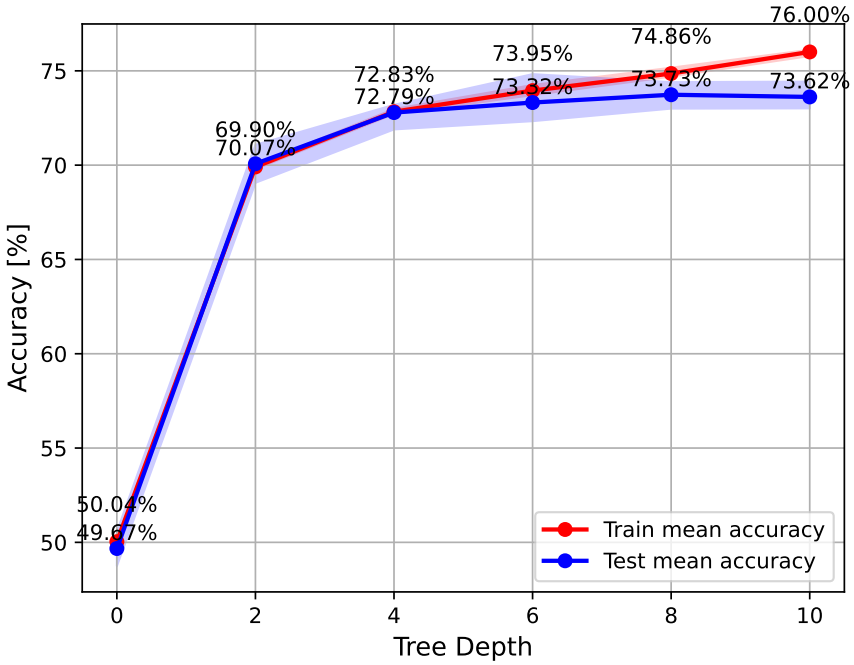
Tab. 3.34: Podstawowe informacje o zbiorze

Liczba próbek	**Liczba klas**	**Liczba atrybutów**
70692	2	21

Tab. 3.35: Atrybuty próbek ze zbioru Diabetes

Polska nazwa atrybutu	**Angielska nazwa atrybutu**	**Typ atrybutu**
wysokie ciśnienie krwi	high blood pressure	binarny
wysoki cholesterol	high cholesterol	binarny
badanie cholesterolu lat	cholesterol check	binarny
bmi	bmi	dyskretny
palacz	smoker	binarny
udar	stroke	binarny
choroba wieńcowa	coronary heart disease	binarny
aktywność fizyczna	physical activity	binarny
owoce	fruits	binarny
warzywa	veggies	binarny
ciężki alkoholizm	heavy drinker	binarny
ubezpieczenie zdrowotne	healthcare coverage	binarny
brak wizyty u doktora	no doctor visit	binarny
ogólne zdrowie	general health	dyskretny
zdrowie psychiczne	mental health	dyskretny
zdrowie fizyczne	physical health	dyskretny
problemy z chodzeniem	walking difficulties	binarny
płeć	sex	binarny
wiek	age	dyskretny
wykształcenie	education	dyskretny
przychód	income	dyskretny

3.4.2. Wybór ograniczenia do ID3



Rys. 3.28: Wykres skuteczności od maksymalnej głębokości

Na podstawie wykresu przedstawiającego zależność skuteczności na zbiorze trenującym i testującym od maksymalnej głębokości drzewa podjęto decyzję o wykorzystaniu w eksperymentach z ID3 maksymalnej głębokości równej 8. Dla większych wartości skuteczność na zbiorze testującym minimalnie malała co może świadczyć o niewielkim overfittingu.

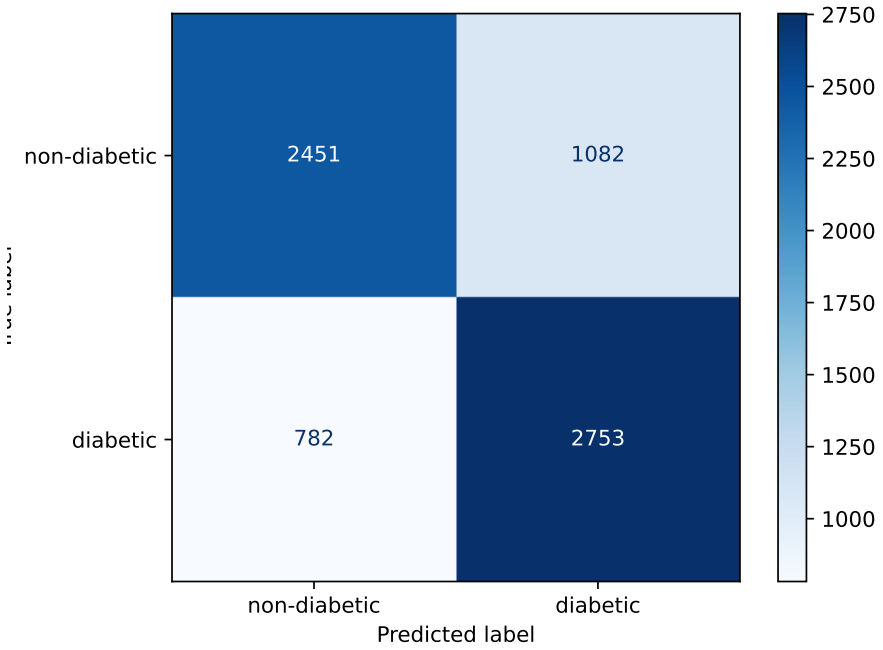
3.4.3. ID3 z ograniczeniem

Tab. 3.36: Procentowa dokładność klasyfikacji; drzewo ID3; 25 powtórzeń

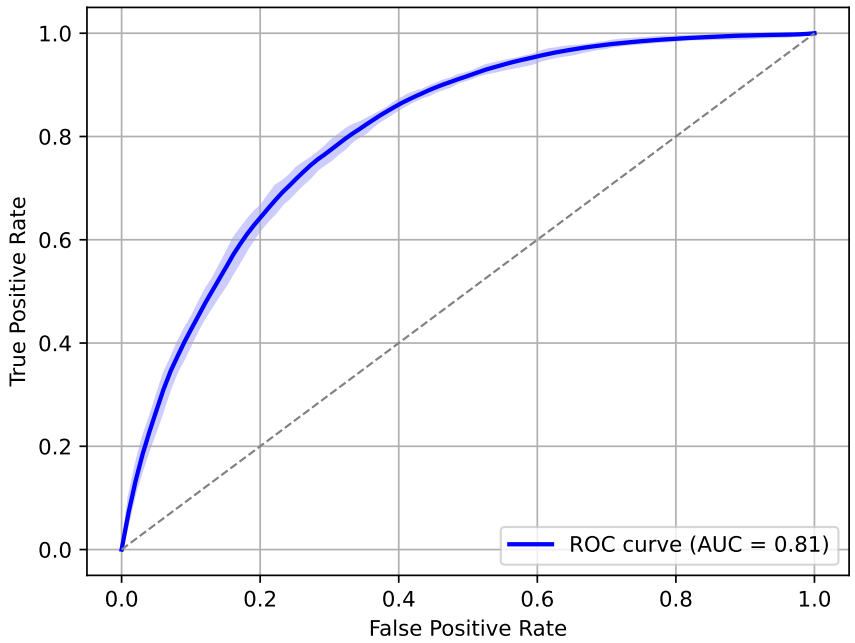
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	74,92	73,62
Odchylenie standardowe	0,09	0,41
Maksimum	75,10	74,37
Minimum	74,75	72,87

Tab. 3.37: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Diabetes (drzewo ID3, 25 powtórzeń)

Parametr	Wartość
Precyzja	0,72
True Positive Rate	0,78
False Positive Rate	0,31



Rys. 3.29: Uśredniona macierz pomyłek; drzewo ID3; 25 powtórzeń



Rys. 3.30: Uśredniona ROC; drzewo ID3; 25 powtórzeń

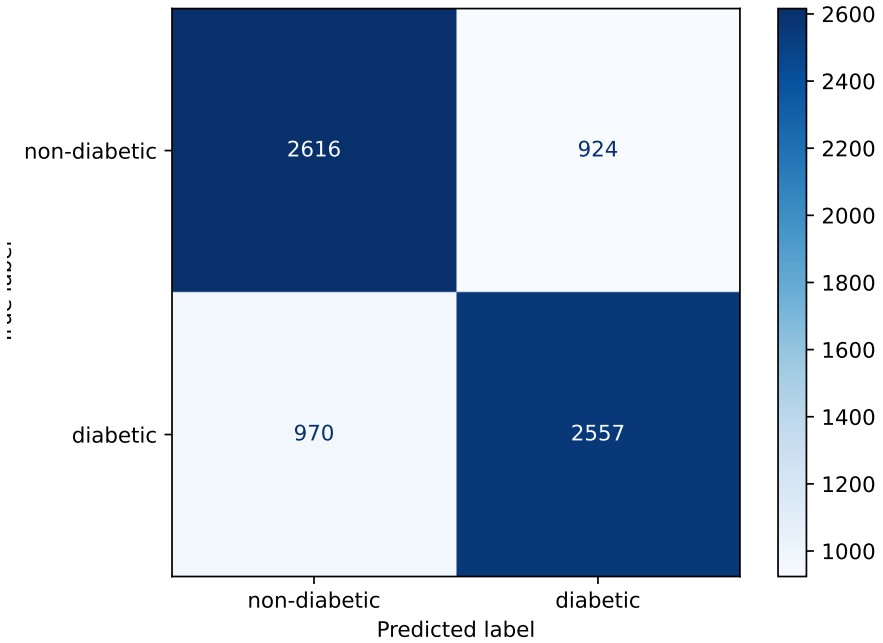
3.4.4. Naiwny klasyfikator Bayesa

Tab. 3.38: Procentowa dokładność klasyfikacji; Naiwny klasyfikator Bayesa; 50 powtórzeń

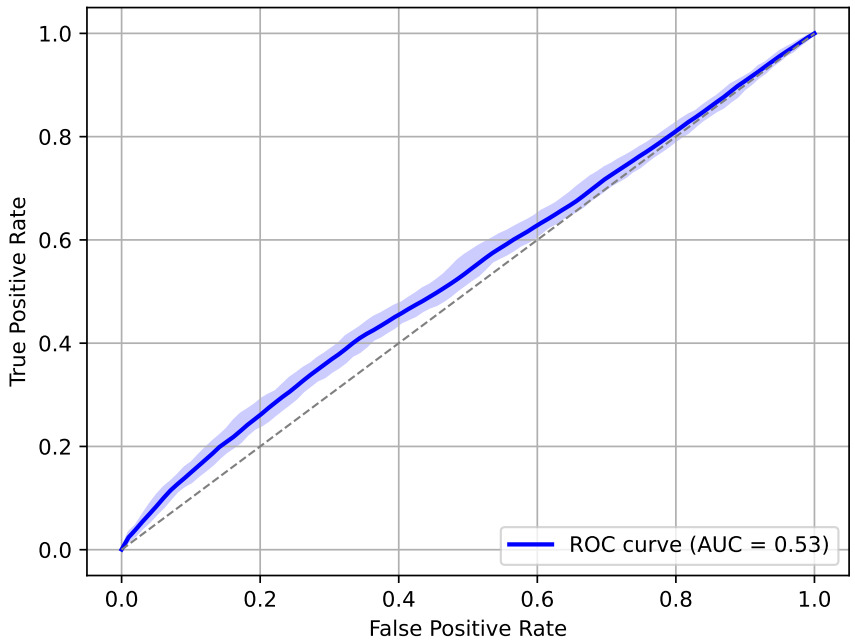
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	73,15	73,20
Odchylenie standardowe	0,06	0,47
Maksimum	73,25	74,37
Minimum	73,00	72,31

Tab. 3.39: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Diabetes (Naiwny klasyfikator Bayesa, 50 powtórzeń)

Parametr	Wartość
Precyzja	0,735
True Positive Rate	0,725
False Positive Rate	0,261



Rys. 3.31: Uśredniona macierz pomyłek; Naiwny klasyfikator Bayesa; 50 powtórzeń



Rys. 3.32: Uśredniona ROC; Naiwny klasyfikator Bayesa; 50 powtórzeń

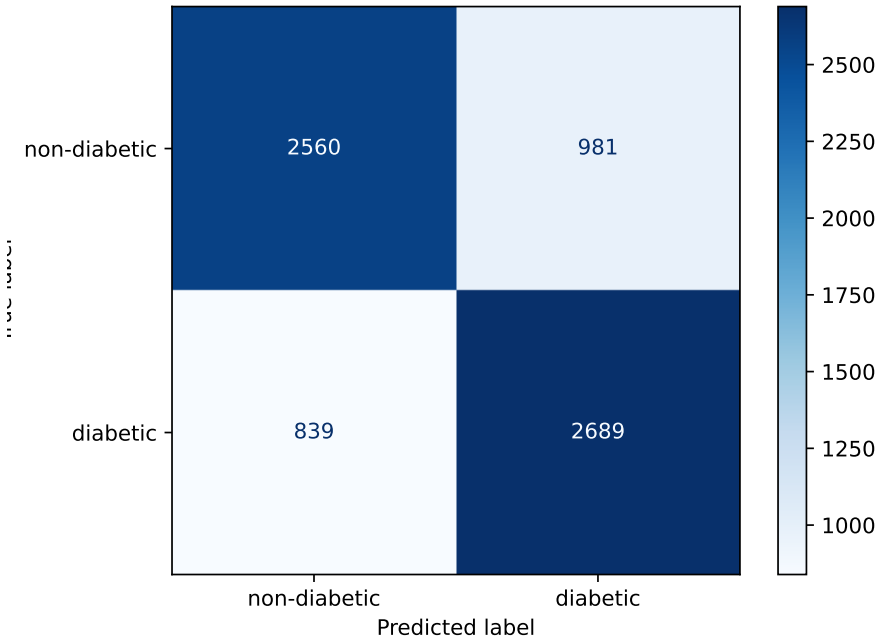
3.4.5. Klasyczny las losowy

Tab. 3.40: Procentowa dokładność klasyfikacji; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

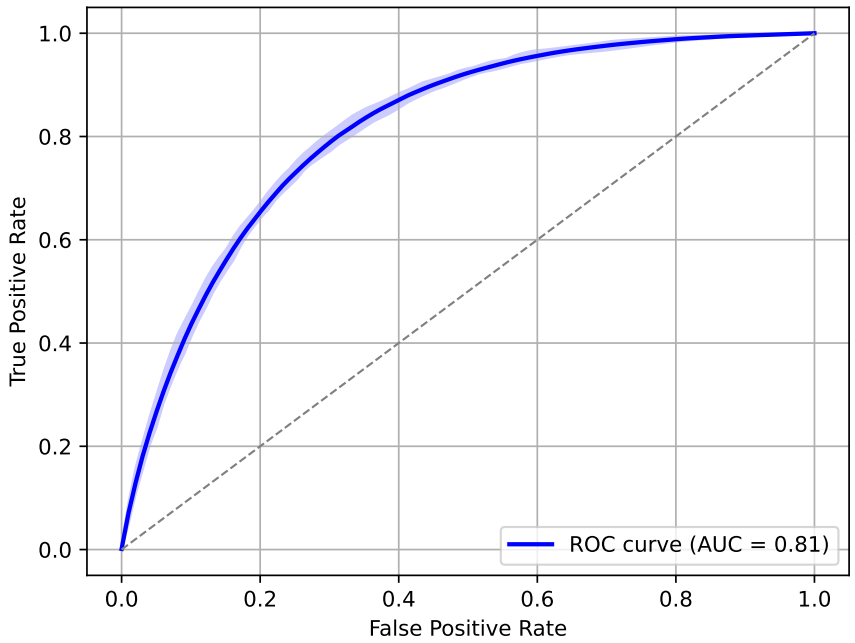
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	90,51	74,26
Odchylenie standardowe	0,12	0,46
Maksimum	90,77	75,30
Minimum	90,28	73,59

Tab. 3.41: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Diabetes (klasyczny las losowy, 50 drzew ID3, 25 powtórzeń)

Parametr	Wartość
Precyzja	0,73
True Positive Rate	0,76
False Positive Rate	0,28



Rys. 3.33: Uśredniona macierz pomyłek; klasyczny las losowy (50 drzew ID3); 25 powtórzeń



Rys. 3.34: Uśredniona ROC; klasyczny las losowy (50 drzew ID3); 25 powtórzeń

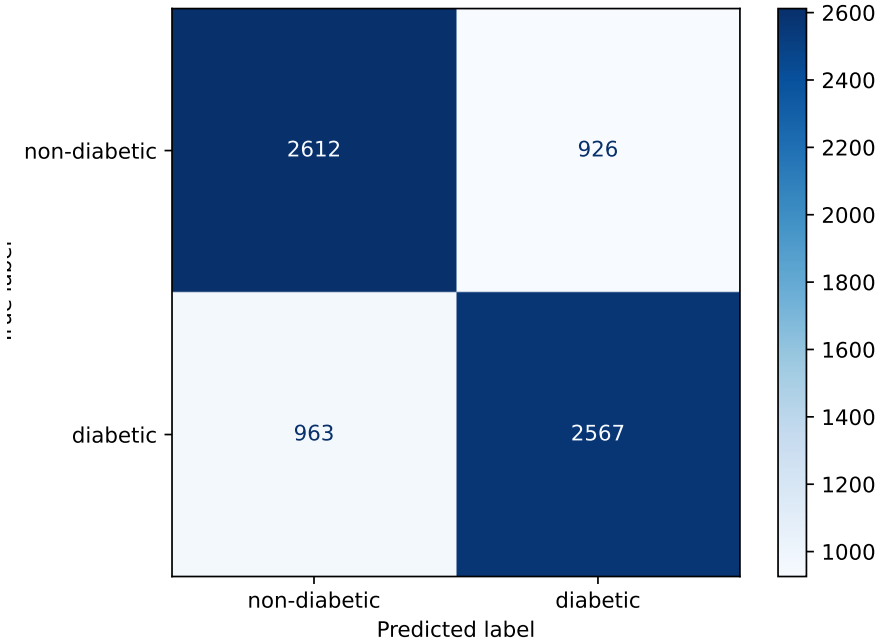
3.4.6. Zmodyfikowany las losowy ID3/NBC

Tab. 3.42: Procentowa dokładność klasyfikacji; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

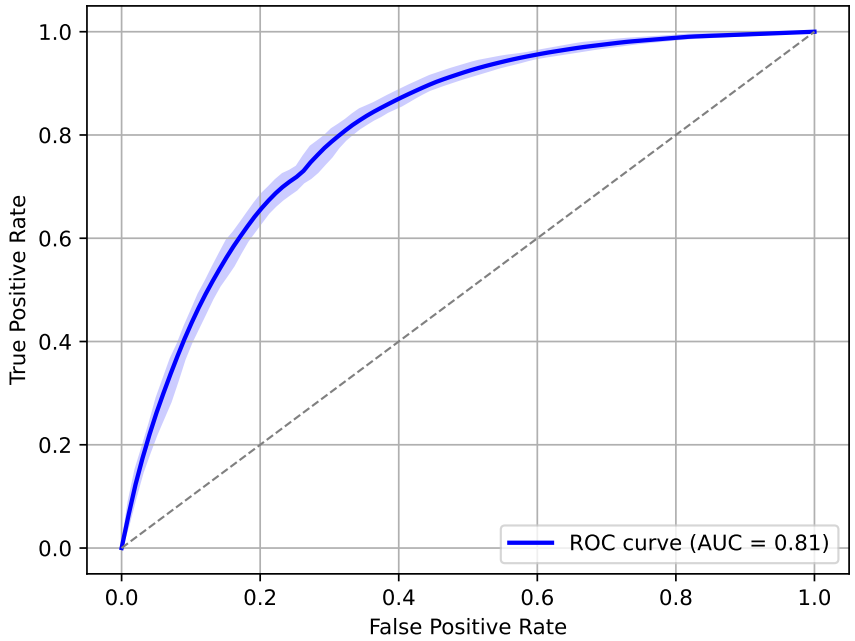
Skuteczność na zbiorze	Trenującym	Testującym
Średnia	73,67	73,27
Odchylenie standardowe	0,12	0,55
Maksimum	74,05	74,27
Minimum	73,49	72,15

Tab. 3.43: Precyzja, współczynnik prawdziwych pozytywnych i fałszywych pozytywnych dla zbioru danych Diabetes (zmodyfikowany las losowy, 25 drzew ID3 i 25 NBC, 25 powtórzeń)

Parametr	Wartość
Precyzja	0,74
True Positive Rate	0,73
False Positive Rate	0,26



Rys. 3.35: Uśredniona macierz pomyłek; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń



Rys. 3.36: Uśredniona ROC; zmodyfikowany las losowy (25 drzew ID3 i 25 NBC); 25 powtórzeń

3.4.7. Podsumowanie

Tab. 3.44: Zestawienie dokładności klasyfikacji na zbiorze Diabetes dla różnych modeli (na zbiorze testującym)

Model	Średnia	Maksimum	Minimum	Odchylenie standardowe
Drzewo decyzyjne ID3	73,62	74,37	72,87	0,41
Naiwny klasyfikator Bayesa	73,20	74,37	72,31	0,47
Las losowy (50 drzew ID3)	74,26	75,30	73,59	0,46
Las losowy (25 ID3 / 25 NBC)	73,27	74,27	72,15	0,55

Wyniki eksperymentów na zbiorze Diabetes prezentują się w odmienny sposób od wyników poprzednich eksperymentów. Przede wszystkim różnice w skuteczności między modelami są bardzo, bardzo niewielkie. Świadczyć to może o pewnej charakterystyce zbioru - wartości atrybutów są od siebie bardzo mało zależne. Jest to z resztą zgodne z przewidywaniami, które można by snuć na podstawie faktycznej interpretacji atrybutów (np. palenie, albo wskaźnik BMI bez względu na inne atrybuty i w każdym kontekście będzie pozytywnie skorelowane z występowaniem cukrzycy).

Z danych widać, że podobnie jak w poprzednich eksperymentach klasyczny las losowy za każdym razem podlega zjawisku overfittingu. Rozwiązanie tego problemu może być analogiczne jak w przypadku Credit Score - zmniejszenie maksymalnej głębokości drzew w lesie.

Tak samo, jak we wszystkich poprzednich modelach największą skuteczność osiąga klasyczny las losowy. Jednak podczas szukania kompromisu pomiędzy złożonością obliczeniową, a skutecznością klasyfikacji można pokusić się o wybór nawet najgorszego z modeli - Naiwnego klasyfikatora Bayesa, ponieważ jego skuteczność w sposób bliski nieznaczącemu odbiega od najlepszego wyniku (tym bardziej od pozostałych).

Odchylenia standardowe oraz minima i maksima w skutecznościach oraz przebiegach krzywych ROC świadczą o dużej powtarzalności eksperymentów i osiągnięciu przez modele granicznie dużej skuteczności dla danej architektury modelu.

Mimo bardzo zbliżonych do siebie skuteczności dla różnych modeli na tym zbiorze danych dalej można dostrzec dwa "kierunki" powodujące, w tym przypadku subtelny (ale jednak), wzrost skuteczności. Są to oczywiście:

1. Zmiana modeli pojedynczych na multimodele - Skuteczność zmodyfikowanego lasu ID3/NBC jest większa od skuteczności klasyfikatora NBC oraz Skuteczność klasycznego lasu ID3 jest większa od skuteczności pojedynczego drzewa ID3.
2. Zamiana NBC na drzewa - Skuteczność ID3 jest większa od skuteczności NBC, natomiast skuteczność klasycznego lasu ID3 jest większa od skuteczności zmodyfikowanego lasu ID3/NBC. W tym przypadku, to ten właśnie kierunek okazuje się być bardziej znaczący, co widać po różnicy skuteczności między drzewem ID3, a zmodyfikowanym lasem losowym ID3/NBC.

3.5. Podsumowanie

3.5.1. Skuteczności Modeli na Zbiorach

Tab. 3.45: Porównanie procentowej dokładności klasyfikacji różnych modeli na badanych zbiorach

	Wine	Diabetes	Credit Score	Healthcare
ID3	74,28	73,62	68,22	33,63
NBC	71,44	73,20	65,37	33,27
Las ID3	82,80	74,26	73,54	42,08
Las ID3/Bayes	76,53	73,27	65,86	36,71

3.5.2. Poprawianie skuteczności

We wszystkich wykonanych eksperymentach widoczna jest omawiana już wyżej kilkukrotnie "kierunkowa zależność". Polega ona na zauważeniu dwóch kierunków zmian modelu powiązanych z poprawą skuteczności. Te kierunki to:

1. Zmiana modeli pojedynczych na multimodele - Skuteczność zmodyfikowanego lasu ID3/NBC jest zawsze większa od skuteczności klasyfikatora NBC, natomiast skuteczność klasycznego lasu ID3 jest zawsze większa od skuteczności pojedynczego drzewa ID3.
2. Zamiana NBC na drzewa - Skuteczność drzewa ID3 jest zawsze większa od skuteczności NBC, natomiast skuteczność klasycznego lasu ID3 jest zawsze większa od skuteczności zmodyfikowanego lasu ID3/NBC.

Ze względu na te 2 kierunki las ID3 zawsze okazywał się najskuteczniejszym modelem, natomiast pojedynczy naiwny klasyfikator Bayesa zawsze okazywał się najmniej skuteczny. O tym czy większą skuteczność osiągało pojedyncze drzewo ID3, czy zmodyfikowany las losowy ID3/NBC decydowała istotność w danym zbiorze poszczególnego kierunku.

3.5.3. Overfitting

Modelem najbardziej podatnym na przetrenowanie okazał się klasyczny las losowy ID3. Wynika to jednak najprawdopodobniej z faktu, że w zaimplementowanym w ramach tego projektu lesie losowym wszystkie drzewa nie mają ograniczenia głębokości, natomiast pojedyncze drzewa ID3 mają ustawione ograniczenie głębokości od 4 do 10 (w zależności od zbioru danych - dla każdego głębokości zostawały dobierane indywidualnie). Zgodnie z teorią przewidywać można, że zwiększanie ilości podmodeli w multimodelu powinno zmniejszać overfitting, tak samo zmniejszanie maksymalnej głębokości drzewa powinno zmniejszać overfitting.

Z doświadczeń wynika również, że naiwny klasyfikator Bayesa jest modelem mało podatnym na overfitting.

3.5.4. Złożoność obliczeniowa

Modele testowane w ramach eksperymentów można by ułożyć zgodnie z ich złożonością obliczeniową w sposób rosnący w następujący sposób

1. Nawiny klasyfikator Bayesa
2. Drzewo ID3
3. Zmodyfikowany las losowy ID3/NBC
4. Klasyczny las losowy ID3

Jednak ze złożonością obliczeniową nie zawsze wiąże się skuteczność. Dla niektórych zbiorów większą skuteczność od Zmodyfikowanego lasu losowego ID3/NBC osiągały pojedyncze Drzewa ID3.