Spis treści

1. Cel i zakres projektu

Celem projektu jest przeprowadzenie analizy danych dotyczących czynników wiążących się z występowaniem zapalenia wątroby typu C (HCV) na różnych etapach choroby, w tym zwłóknień wątroby oraz marskości wątroby.

W zakres projektu wchodzą opis wykorzystywanego zbioru danych, analiza statystyczna danych, analiza danych metodami eksploracji takimi jak badanie ważności atrybutów, generowanie reguł klasyfikujących, indukcja drzew decyzyjnych, modelowanie sztucznych sieci neuronowych oraz tworzenie modelu regresji logistycznej. Analizy danych przeprowadzono przy użyciu oprogramowania Statistica oraz WEKA.

2. Opis i analiza statystyczna danych

W niniejszym projekcie przeprowadzono analizę danych pozyskanych z repozytorium UCI Machine Learning Repository. Zbiór danych dostępny jest do pobrania pod adresem: https://archive.ics.uci.edu/ml/datasets/HCV+data. Zbiór zawiera informacje na temat wieku, płci i wybranych parametrów biochemicznych krwi osób zdrowych, podejrzanych o zakażenie HCV oraz chorych na HCV o różnych stopniach zaawansowania choroby. Najwcześniejszy etap dotyczy osób określonych jako chore na HVC. Kolejne stadium choroby to zwłóknienia wątroby, natomiast najpoważniejszy etap to marskość wątroby [8].

Zbiór zawiera 615 instancji oraz 14 atrybutów, które wymieniono poniżej:

1) ID pacjenta,
2) klasa:
    - 0 - dawca krwi (pacjent zdrowy),
    - 0s - dawca krwi podejrzany o zakażenie HCV,
    - 1 - hepatitis (pacjent chory na zapalenie wątroby typu C),
    - 2 - fibrosis (pacjent ze zwłóknieniami wątroby),
    - 3 - cirrhosis (pacjent z marskością wątroby);
3) wiek,
4) płeć (m, f),
5) ALB - poziom albuminy,
6) ALP - poziom fosfatazy alkalicznej,
7) ALT - poziom aminotransferazy alaninowej,
8) AST - poziom aminotransferazy asparaginowej,
9) BIL - poziom bilirubiny,
10) CHE - poziom cholinoesterazy,
11) CHOL - poziom cholesterolu,
12) CREA - poziom kreatyniny,
13) GGT - poziom gammaglutamylotransferazy,
14) PROT - poziom białka całkowitego.

Instancje są podzielone na pięć klas, których liczebność zestawiono w tabeli 1.

Tabela 1. Liczebność poszczególnych klas

Numer klasy	Opis klasy	Liczebność klasy
0	Dawca krwi	533
0s	Dawca krwi podejrzany o zakażenie HCV	7
1	Pacjent chory na HCV	24
2	Pacjent ze zwłóknieniami wątroby	21
3	Pacjent z marskością wątroby	30

Najliczniejszą klasą była klasa osób zdrowych (dawców krwi). W ich przypadku wszystkie parametry krwi powinny przyjmować wartości określone w przyjętych normach. Pacjenci chorzy zostali przypisani do trzech kategorii o podobnej liczebności (21-30 osób) w zależności od rozpoznania etapu choroby. Dodatkowo, wyróżniono także niewielką klasę 0s, do której należeli dawcy krwi z podejrzeniem zakażenia HCV.

Poza klasą, charakter kategoryczny miał także atrybut płeć, natomiast pozostałe przyjmowały wartości liczb rzeczywistych (charakter numeryczny).

Atrybut ID został pominięty w analizach, ponieważ przypadki w zbiorze danych były posortowane według klasy. Wynika z tego, że identyfikator zostałby uznany za bardzo istotny czynnik decydujący o diagnozie pacjenta, jednak byłby to błędny wniosek, gdyż w rzeczywistości ID jest nadawane pacjentom kolejno, bez względu na ich przynależność do poszczególnych klas. Identyfikator nie ma wpływu na stan zdrowia pacjenta, więc nie może być użyte do klasyfikowania przypadków.

W badaniu wzięło udział 377 mężczyzn i 238 kobiet. Wiek pacjentów zawierał się w przedziale od 19 do 77 lat (średni wiek wynosił 47,4 ± 10,1 lat). Atrybuty o numerach 5-14 odnoszą się do wyników badań ilościowych krwi. W niektórych instancjach wystąpiły brakujące dane dotyczące wartości co najmniej jednego parametru krwi.

Atrybuty o charakterze numerycznym poddano analizie statystycznej. W poniższej tabeli zestawiono wartości podstawowych parametrów statystycznych tych atrybutów.

Tabela 2. Parametry statystyczne atrybutów numerycznych

Atrybut	Średnia	Mediana	Moda	Minimum	Maksimum	Współczynnik zmienności	Wariancja	Ochylenie standardowe	Skośność	Kurtoza
WIEK	47,41	47,0	46,0	19,0	77,0	21,21	101,1	10,06	0,3	-0,4
ALB	41,62	42,0	39,0	14,9	82,2	13,89	33,4	5,78	-0,2	6,0
ALP	68,28	66,2	Wiele	11,3	416,6	38,12	677,5	26,03	4,7	55,0
ALT	28,45	23,0	16,6	0,9	325,3	89,52	648,7	25,47	5,5	47,1
AST	34,79	25,9	Wiele	10,6	324,0	95,13	1095,0	33,09	4,9	30,8
BIL	11,40	7,3	6,0	0,8	254,0	172,62	387,0	19,67	8,4	83,2
CHE	8,20	8,3	7,5	1,4	16,4	26,91	4,9	2,21	-0,1	1,3
CHOL	5,37	5,3	Wiele	1,4	9,7	21,10	1,3	1,13	0,4	0,7
CREA	81,29	77,0	74,0	8,0	1079,1	61,21	2475,7	49,76	15,2	280,1
GGT	39,53	23,3	Wiele	4,5	650,9	138,27	2987,8	54,66	5,6	43,7
PROT	72,04	72,2	71,9	44,8	90,0	7,50	29,2	5,40	-1,0	3,5

Jak widać w tabeli 2, wyniki są bardzo zróżnicowane. W niektórych przypadkach wartości maksymalne parametrów są oddalone od średniej o kilkaset jednostek. Wskazuje to na wystąpienie pacjentów o bardzo odstających wynikach, co może być związane z zapaleniem wątroby. Rozkłady atrybutów są w większości leptokurtyczne (wartość kurtozy jest dodatnia), czyli wykresy są bardziej wysmukłe od rozkładu normalnego. Rozkłady ALB, CHE i PROT są słabo lewostronnie skośne, a pozostałe są prawostronnie skośne. Szczególnie wysoką skośnością charakteryzują się rozkłady atrybutów CREA oraz BIL. Współczynnik zmienności przyjmuje bardzo wysokie wartości w przypadku poziomu bilirubiny oraz GGT. Silną zmienność wykazują także AST, ALT oraz CREA. Pozostałe atrybuty charakteryzują się słabą i przeciętną zmiennością.

W celu zbadania normalności rozkładów atrybutów numerycznych (wiek oraz parametry krwi) wykonano test Shapiro-Wilka. Test wykonano grupami (w klasach). Założono poziom istotności $\alpha = 0{,}05$. Hipotezy brzmiały:

H0: Rozkład atrybutu X jest rozkładem normalnym.

H1: Rozkład atrybutu X nie jest rozkładem normalnym.

Jeśli uzyskane $p < \alpha$, odrzucono hipotezę zerową i przyjęto, że rozkład atrybutu nie jest rozkładem normalnym. Jeśli p okazało się $> \alpha$, rozkład atrybutu określono jako rozkład normalny. Zestawienie uzyskanych wniosków przedstawiono w poniższej tabeli. Litera N oznacza, że rozkład atrybutu w określonej klasie jest rozkładem normalnym, natomiast x, że rozkład odbiega od rozkładu normalnego.

Tabela 3. Zestawienie wyników testu Shapiro-Wilka dla atrybutów w klasach

Atrybut	Normalność rozkładu atrybutu w klasach				
	0	**0s**	**1**	**2**	**3**
Wiek	x	N	N	N	N
ALB	x	x	N	N	N
ALP	x	N	x	N	x
ALT	x	x	x	x	x
AST	x	N	x	x	x
BIL	x	N	x	x	x
CHE	x	N	N	N	x
CHOL	x	N	x	N	N
CREA	x	N	x	N	x
GGT	x	N	x	N	x
PROT	x	x	N	N	N

Jak widać powyżej, w klasie 0 żaden z atrybutów nie charakteryzował się rozkładem normalnym, natomiast w klasach 0s i 2 prawie wszystkie atrybuty go posiadały. W przypadku, gdy wykonywano test Shapiro-Wilka dla wszystkich atrybutów globalnie (bez uwzględniania podziału na klasy) uzyskano takie same wyniki, jak w klasie 0, czyli rozkład żadnego z atrybutów nie był rozkładem normalnym.

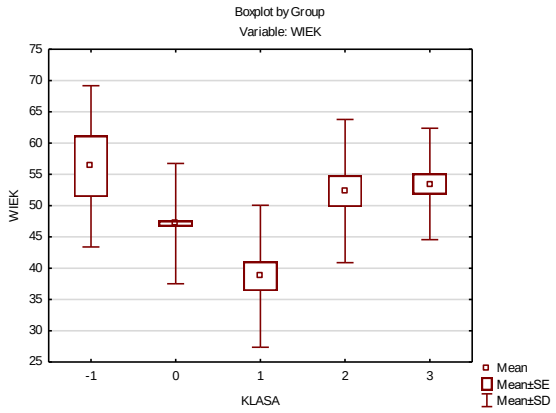
Aby określić, które z atrybutów mogą być istotne w stawianiu diagnozy, przeprowadzono test umożliwiający porównywanie grup - test Kruskala-Wallisa. Wybrano tę metodę, ponieważ nie wymaga ona rozkładu normalnego atrybutów. Porównano wszystkie atrybuty numeryczne w klasach i uzyskano informacje o istotnych statystycznie różnicach wartości atrybutów pomiędzy klasami. W poniższej tabeli zaznaczono znakiem X klasy pomiędzy którymi wystąpiły te różnice.

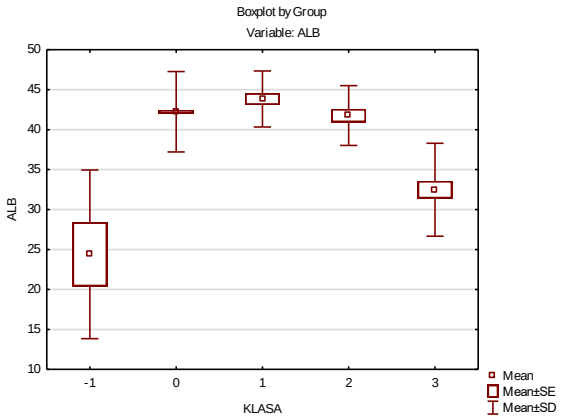
Tabela 4. Zestawienie istotnych statystycznie różnic wartości atrybutów pomiędzy klasami

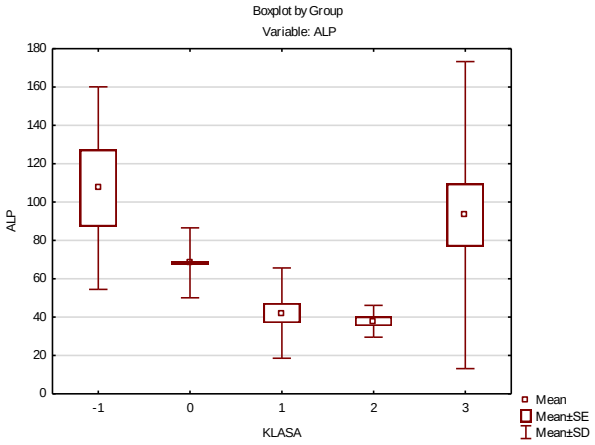
Atrybut	Istotne statystycznie różnice wystąpiły pomiędzy klasami:									
	0 i 1	0 i 2	0 i 3	0 i 0s	1 i 2	1 i 3	1 i 0s	2 i 3	2 i 0s	3 i 0s
Wiek	X		X		X	X	X			
ALB			X	X		X	X	X		
ALP	X	X				X	X	X	X	
ALT			X					X		X
AST	X	X	X							
BIL	X	X	X				X		X	X
CHE			X			X		X		
CHOL		X	X			X				
CREA										
GGT	X	X	X	X						
PROT		X		X			X	X		X

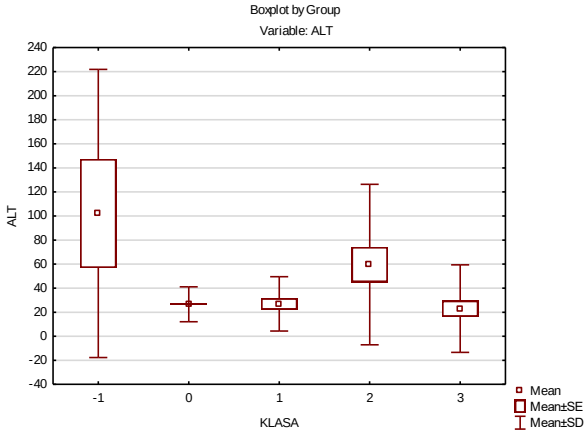
Jak widać w tabeli 4, większość atrybutów przyjmuje istotnie różne wartości w zależności od klasy. Dzięki temu można określić, który z parametrów może być istotny w klasyfikacji przypadków. Atrybuty AST oraz GGT przyjmują znacząco inne wartości w przypadku osób zdrowych niż w przypadku osób chorych. Podobnie jest z poziomem bilirubiny (BIL), jednak opierając się jedynie na tym parametrze nie można jednoznacznie stwierdzić, czy dana instancja związana jest z klasą 0 czy 0s. Z kolei parametry ALT i CHE różnią się przyjmowaną wartością w przypadku osób chorych na marskość wątroby (klasa 3) niż innych pacjentów. Atrybut CREA natomiast nie wykazuje znaczących statystycznie różnic wartości w poszczególnych klasach.

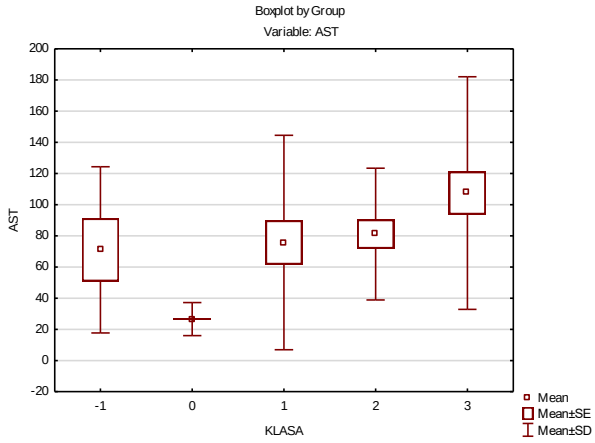
W celu uwidocznienia różnic wartości poszczególnych atrybutów pomiędzy klasami wygenerowano wykresy ramka-wąsy. Każdy wykres przedstawia średnią wartość atrybutu w poszczególnych klasach wraz z zakresem błędu standardowego (ramka) oraz odchylenia standardowego (wąsy). Ze względu na brak możliwości nadania w programie Statistica klasie 0s jej kodu (ma on format tekstowy, a wymagany jest format liczbowy), na wykresach klasa ta jest reprezentowana kodem -1.

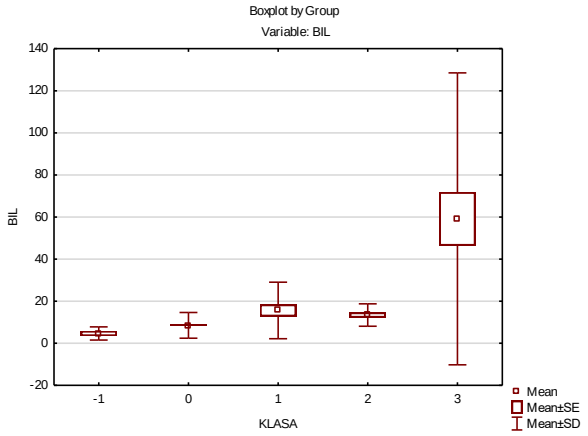

Boxplot by Group
Variable: WIEK
WIEK
KLASA
Mean
Mean±SE
Mean±SD

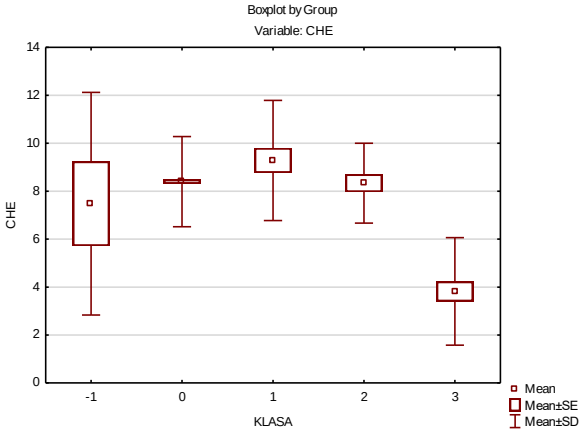

Boxplot by Group
Variable: ALB
ALB
KLASA
Mean
Mean±SE
Mean±SD

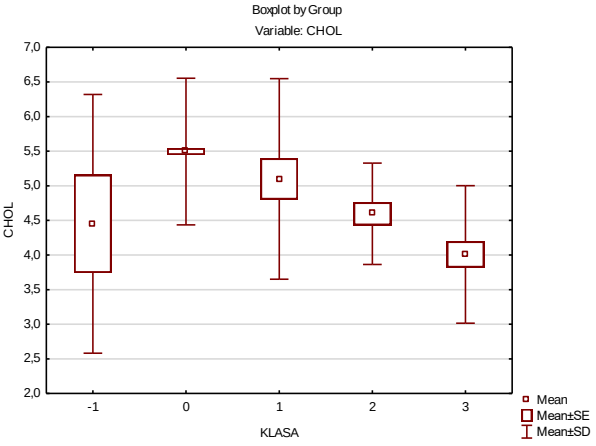

Boxplot by Group
Variable: ALP
ALP
KLASA
Mean
Mean±SE
Mean±SD

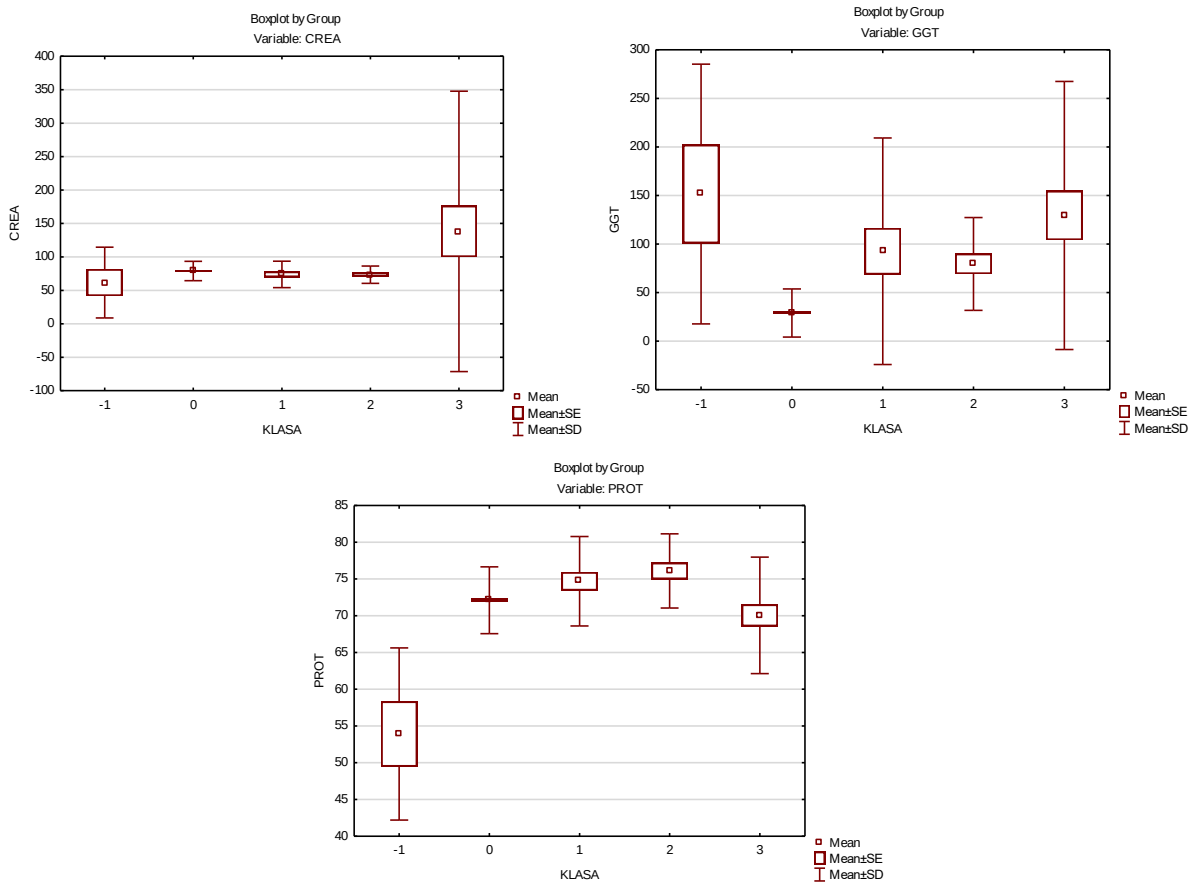

Boxplot by Group
Variable: ALT
ALT
KLASA
Mean
Mean±SE
Mean±SD


Boxplot by Group
Variable: AST
AST
KLASA
Mean
Mean±SE
Mean±SD


Boxplot by Group
Variable: BIL
BIL
KLASA
Mean
Mean±SE
Mean±SD


Boxplot by Group
Variable: CHE
CHE
KLASA
Mean
Mean±SE
Mean±SD


Boxplot by Group
Variable: CHOL
CHOL
KLASA
Mean
Mean±SE
Mean±SD



Rys. 1. Wykresy ramka-wąsy atrybutów w klasach

Przedstawione wykresy w sposób bardziej obrazowy przedstawiają podobieństwa i różnice pomiędzy wartościami atrybutów w klasach. Można zauważyć, że bardziej zaawansowane stadia choroby (klasy 2 i 3) wystąpiły u osób starszych niż samo zakażenie bez zwłóknień i marskości (klasa 1). Natomiast wartość parametru ALB maleje wraz z postępem choroby. Średnia wartość poziomu cholesterolu jest wyższa u osób zdrowych niż chorych. Z kolei parametry GGT i AST przyjmują znacznie niższe wartości u osób zdrowych niż w innych grupach. W przypadku klasy osób podejrzanych o zakażenie HCV (klasa 0s, na wykresach jako -1) zakresy wartości większości atrybutów są duże, co wskazuje na niejednorodność tej grupy.

3. Analiza przy użyciu metod eksploracji danych

3.1. Wybór istotnych atrybutów

Pierwszym etapem analizy było określenie istotności atrybutów poprzez wykrycie ich powiązań z atrybutem klasyfikującym. W tym celu wykorzystano algorytmy CfsSubsetEval, GainRatioAttributeEval oraz InfoGainAttributeEval wraz z metodami wyszukiwania BestFirst oraz Ranker.

Ewaluator CfSubsetEval ocenia wartość podzbioru atrybutów poprzez rozważenie indywidualnej zdolności predykcyjnej każdej z cech wraz ze stopniem redundancji

(nadmiaru) pomiędzy nimi. Preferowane są podzbiory atrybutów, które są silnie skorelowane z klasą, a jednocześnie mają niską korelację z innymi atrybutami [3]. Algorytm GainRatioAttributeEval dokonuje oceny istotności atrybutów poprzez pomiar ich współczynnika wzmocnienia w odniesieniu do klasy, natomiast test InfoGainAttributeEval poprzez pomiar przyrostu informacji w odniesieniu do klasy [7].

Metoda BestFirst przeszukuje przestrzeń podzbiorów atrybutów przez lokalne przeszukiwanie zachłanne (greedy hill climbing) i wyświetla najlepsze rozwiązania problemu. Z kolei Ranker sortuje atrybuty według indywidualnych ocen, a także usuwa atrybuty, które znajdują się poniżej pewnej określonej wartości oceny.

Poniżej zaprezentowano wyniki oceny ważności analizowanych atrybutów przy użyciu trzech opisanych powyżej algorytmów.

a)

```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 80
        Merit of best subset found:    0.522

Attribute Subset Evaluator (supervised, Class (nominal): 1 Class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,4,5,6,7,8,9,11,12,13 : 10
                     Age
                     ALB
                     ALP
                     ALT
                     AST
                     BIL
                     CHE
                     CREA
                     GGT
                     PROT
```

b)

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 Class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.7236   9 CHE
 0.4968   6 ALT
 0.381    4 ALB
 0.299    7 AST
 0.1759   5 ALP
 0.1687   8 BIL
 0.1641  12 GGT
 0.1565  13 PROT
 0.1441  11 CREA
 0.1378  10 CHOL
 0.0732   2 Age
 0.0104   3 Sex

Selected attributes: 9,6,4,7,5,8,12,13,11,10,2,3 : 12
```

c)

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 Class):
        Information Gain Ranking Filter

Ranked attributes:
 0.3433   7 AST
 0.1867   6 ALT
 0.166    8 BIL
 0.1604   4 ALB
 0.1497   9 CHE
 0.144    5 ALP
 0.1418  12 GGT
 0.1195  13 PROT
 0.0807  10 CHOL
 0.0792  11 CREA
 0.0776   2 Age
 0.01     3 Sex

Selected attributes: 7,6,8,4,9,5,12,13,10,11,2,3 : 12
```

Rys. 2. Wynik badania ważności atrybutów a) CfsSubsetEval + BestFirst, b) GainRatioAttributeEval + Ranker, c) InfoGainAttributeEval + Ranker

Z przeprowadzonych analiz atrybutów wynika, że największe powiązanie z atrybutem klasyfikującym wykazują parametry ALT, AST, CHE, ALB, BIL oraz ALP. Pozostałe parametry krwi oraz płeć i wiek pacjenta były zdecydowanie mniej powiązane z klasą. Podczas analizy danych z użyciem algorytmu CfsSubsetEval wraz z metodą wyszukiwania

BestFirst, atrybuty płeć i poziom cholesterolu (CHOL) zostały uznane jako nieistotne i pominięte.

Na podstawie uzyskanych wyników podjęto decyzję o prowadzeniu dalszych analiz na podstawie jedenastu atrybutów. Nie uwzględniano płci pacjenta, ponieważ atrybut ten wykazał zdecydowanie mniejsze powiązanie z klasą niż inne atrybuty.

3.2. Generowanie reguł klasyfikujących

Przed rozpoczęciem wyznaczania reguł asocjacyjnych konieczne było wykonanie dyskretyzacji. Ciągłe atrybuty przekształcono w dyskretne za pomocą filtra *Discretize* dostępnego w zakładce *Preprocess* programu WEKA. Wybrano podział na 10 przedziałów o równej szerokości.

Reguły klasyfikujące wyznaczano przy użyciu algorytmu Apriori poprzez zaznaczenie w jego ustawieniach opcji *True* parametru *car*. Dzięki temu pominięto wyznaczanie wszystkich reguł asocjacyjnych i wybieranie spośród nich reguł klasyfikujących, czyli tych których następnikiem jest występowanie jednej z klas.

Założono minimalną ufność na poziomie 0,95 oraz wsparcie w przedziale od 0,8 do 1. W wyniku działania algorytmu uzyskano 4 przedstawione poniżej reguły.

1. AST='(-inf-41.94]' CREA='(-inf-115.11]' 520 ==> Category=0 504 conf:(0.97)

2. AST='(-inf-41.94]' BIL='(-inf-26.12]' CREA='(-inf-115.11]' 511 ==> Category=0 495 conf:(0.97)

3. AST='(-inf-41.94]' 526 ==> Category=0 506 conf:(0.96)

4. AST='(-inf-41.94]' BIL='(-inf-26.12]' 517 ==> Category=0 497 conf:(0.96)

Wyznaczone reguły klasyfikacyjne dotyczyły wartości parametrów krwi wskazujących na przynależność instancji do klasy 0, czyli osób zdrowych. We wszystkich regułach pojawia się wartość parametru AST. Wynika z tego, że osoby zdrowe miały niski poziom aminotransferazy asparaginowej. Zaobserwowano także wpływ niskich wartości kreatyniny i bilirubiny na przynależność instancji do klasy 0. Uzyskane reguły nie odnoszą się jednak do występowania innych klas. Z tego powodu zmniejszono wartość minimalnego wsparcia oraz ufności, jednak nawet przy bardzo niskich wartościach tych parametrów (na poziomie 0,01) żadna z wygenerowanych reguł nie wskazywała na zależności z innymi klasami. Również zmiana liczby przedziałów podczas dyskretyzacji nie pozwoliła na otrzymanie reguł innych niż wskazujące na występowanie klasy 0.

Zakładając ufność na poziomie 0,95 i minimalne wsparcie 0,5 uzyskano 34 reguły klasyfikujące. Wśród nich wystąpiły trzy reguły jednopoziomowe:

ALP='(51.83-92.36]' 399 ==> Category=0 386 conf:(0.97)

AST='(-inf-41.94]' 526 ==> Category=0 506 conf:(0.96)

GGT='(-inf-69.14]' 532 ==> Category=0 491 conf:(0.92)

Większość pozostałych reguł zawierała jako jeden z warunków ten sam poprzednik, co jedna z powyższych reguł, więc wymagały spełnienia tego samego warunku

oraz dodatkowych. Poza nimi wystąpiły jeszcze dwie reguły, które nie są związane z atrybutami ALP, AST i GGT:

ALT='(-inf-33.34]' BIL='(-inf-26.12]' 439 ==> Category=0 404 conf:(0.92)

BIL='(-inf-26.12]' CREA='(-inf-115.11]' 579 ==> Category=0 522 conf:(0.9)

Uzyskane reguły są regułami mocnymi o wsparciu na poziomie co najmniej 63% (386 z 615 instancji potwierdza regułę) i minimalnej ufności wynoszącej 0,9. Na ich podstawie można stwierdzić, że osobę można określić jako zdrową, jeśli jej parametry krwi spełniają jeden z następujących warunków:

- ALP w zakresie od 51,83 do 92,36,
- AST o wartości niższej lub równej 41,94,
- GGT o wartości niższej lub równej 69,14,
- ALT o wartości niższej lub równej 33,34 i BIL o wartości niższej lub równej 26,12,
- BIL o wartości niższej lub równej 26,12 i CREA o wartości niższej lub równej 115,11.

Aby uzyskać reguły dotyczące innych klas zmodyfikowano zbiór danych. Usunięto z niego instancje należące do klas 0 i 0s, dzięki czemu można było wygenerować reguły pozwalające na rozróżnianie poszczególnych poziomów zaawansowania choroby. Ze względu na znaczne zmniejszenie się zbioru danych niemożliwe było otrzymanie tak mocnych reguł, jak w przypadku klasy 0. Przy minimalnym zaufaniu na poziomie 0,7 i minimalnym wsparciu 0,1 (8 instancji z 75 potwierdzało regułę) wygenerowano 23 reguły. Poprzez odrzucenie reguł zawierających powtarzające się fragmenty poprzedników, zbiór reguł zredukowano do czterech dotyczących klasy 3 i dwóch dotyczących klasy 1.

CHE='(-inf-2.919]' 14 ==> Category=3 14 conf:(1)

ALP='(51.83-92.36]' 11 ==> Category=3 8 conf:(0.73)

BIL='(29.9-54.8]' 10 ==> Category=3 8 conf:(0.8)

ALT='(-inf-26.61]' CHOL='(3.078-3.902]' 10 ==> Category=3 9 conf:(0.9)

ALP='(-inf-51.83]' AST='(-inf-47.43]' 11 ==> Category=1 8 conf:(0.73)

ALP='(-inf-51.83]' CHE='(8.915-10.414]' 11 ==> Category=1 8 conf:(0.73)

Powyższe reguły wskazują, że jeśli u pacjenta występuje niski poziom cholinoesterazy (CHE) lub podwyższony poziom fosfatazy alkalicznej (ALP) albo bilirubiny (BIL), to cierpi on prawdopodobnie na marskość wątroby. Podobnie jest jeśli poziom ALT u pacjenta jest niski przy jednoczesnym poziomie cholesterolu w zakresie 3,078-3,902. Jeśli jednak podwyższony poziom ALP występuje wraz z niskim poziomem AST lub z CHE w zakresie 9,915-10,414, pacjent może być zakażony HCV jednak bez zaawansowanych uszkodzeń wątroby. Należy jednak pamiętać, że reguły te zostały wygenerowane na podstawie danych jedynie chorych pacjentów, więc nie można na ich podstawie jednoznacznie określić, czy pacjent jest zdrowy, czy chory. Można jedynie wyciągać wnioski dotyczące osób chorych i określać stadium choroby. W celu uzyskania reguł pozwalających na klasyfikowanie

pacjentów do wszystkich pięciu klas występujących w oryginalnym zbiorze danych należy dodać do niego więcej instancji należących do klas 0s, 1, 2 oraz 3 i prowadzić analizy na wszystkich danych.

3.3. Indukcja drzew decyzyjnych

W celu stworzenia klasyfikatora, który umożliwiałby wygenerowanie schematu procesu klasyfikacji instancji, analizowane dane poddano działaniu czterech algorytmów indukujących drzewa decyzyjne, które opisano poniżej.

Drzewo J48 jest generowane przy użyciu algorytmu C4.52, który dzieli pierwotny zestaw danych względem każdej ze zmiennych. W ten sposób powstaje tyle wariantów podziału, ile w zestawie jest zmiennych objaśniających. Dla każdego podziału liczona jest wartość metryki information gain, która zdefiniowana jest jako przyrost entropii w każdym z podzbiorów. Zmienna o najwyższym współczynniku information gain staje się pierwszym węzłem drzewa. Następnie dla wszystkich podzbiorów powtarza się tę operację do wyczerpania wszystkich instancji [11]. Algorytm RandomTree do konstruowania drzewa, które uwzględnia K losowo wybranych atrybutów w każdym węźle. Nie wykonuje żadnego cięcia [9]. REPTree - algorytm budujący drzewo decyzyjne na podstawie zdobytych informacji, a następnie przycina je opierając się na metodzie redukcji błędów (z dopasowaniem do tyłu). Tylko raz sortuje wartości dla atrybutów numerycznych. Brakujące wartości rozwiązuje się poprzez podzielenie odpowiednich instancji na części (np. jak w C4. 5) [6]. Klasyfikator RandomForest to metoda uczenia maszynowego używana m.in. w klasyfikacji i regresji, która polega na konstruowaniu wielu drzew decyzyjnych w czasie uczenia i generowaniu klasy, która jest dominantą klas (klasyfikacja) lub przewidywaną średnią (regresja) poszczególnych drzew [4]. Algorytm RandomForest jest odporny na braki danych i wartości odstające, a także na przeuczenie [1].

Wybrane klasyfikatory były testowane przy użyciu walidacji krzyżowej (*Cross-validation, 10 folds*). Uzyskane wyniki klasyfikacji przedstawiono poniżej.

Tabela 5. Zestawienie parametrów oceny klasyfikacji

Parametr oceny klasyfikacji	J48	RandomForest	RandomTree	REPTree
Skuteczność klasyfikacji [%]	91,06	92,52	90,24	90,24
Średni błąd bezwzględny	0,04	0,04	0,04	0,05
Średni błąd kwadratowy	0,18	0,14	0,20	0,18
Statystyka Kappa	0,60	0,65	0,58	0,56
TP Rate	0,911	0,925	0,902	0,902
FP Rate	0,235	0,245	0,214	0,256
Czułość	0,911	0,925	0,902	0,902
Precyzja	0,897	0,906	0,896	0,882
Pole pod krzywą ROC	0,854	0,988	0,876	0,883

Jak widać w tabeli 5, wszystkie klasyfikatory charakteryzowały się bardzo wysoką skutecznością klasyfikacji - powyżej 90%. Najlepsze wyniki uzyskano stosując algorytm RandomForest. W jego przypadku pole pod krzywą ROC wynosiło prawie 0,99, co oznacza, że uzyskana krzywa ROC jest praktycznie identyczna z krzywą klasyfikacji idealnej. Drzewa RandomTree i REPTree uzyskały taki sam procent poprawnie przypisanych instancji. Wystąpiły pomiędzy nimi niewielkie różnice w wartościach innych parametrów oceny klasyfikacji, jednak nie można jednoznacznie określić, która z nich była dokładniejsza. Algorytm J48 natomiast, pomimo wyższej liczby poprawnych przypisań niż drzewa RandomTree i REPTree charakteryzował się niższą wartością pola pod krzywą ROC, co świadczy o bardziej losowym klasyfikowaniu instancji.

Aby określić, w przypadku których klas pojawiło się najwięcej błędów przypisań instancji, przeanalizowano macierze pomyłek (Rys. 3).

a)

a	b	c	d	e	<-- classified as
524	1	2	6	0	a = 0
3	0	0	1	3	b = 0s
10	0	8	3	3	c = 1
7	0	5	7	2	d = 2
2	1	1	5	21	e = 3

b)

a	b	c	d	e	<-- classified as
531	1	0	1	0	a = 0
5	0	0	1	1	b = 0s
11	0	7	4	2	c = 1
5	0	6	8	2	d = 2
2	0	0	5	23	e = 3

c)

a	b	c	d	e	<-- classified as
520	2	1	9	1	a = 0
3	1	1	0	2	b = 0s
9	1	6	6	2	c = 1
4	0	8	7	2	d = 2
4	0	3	2	21	e = 3

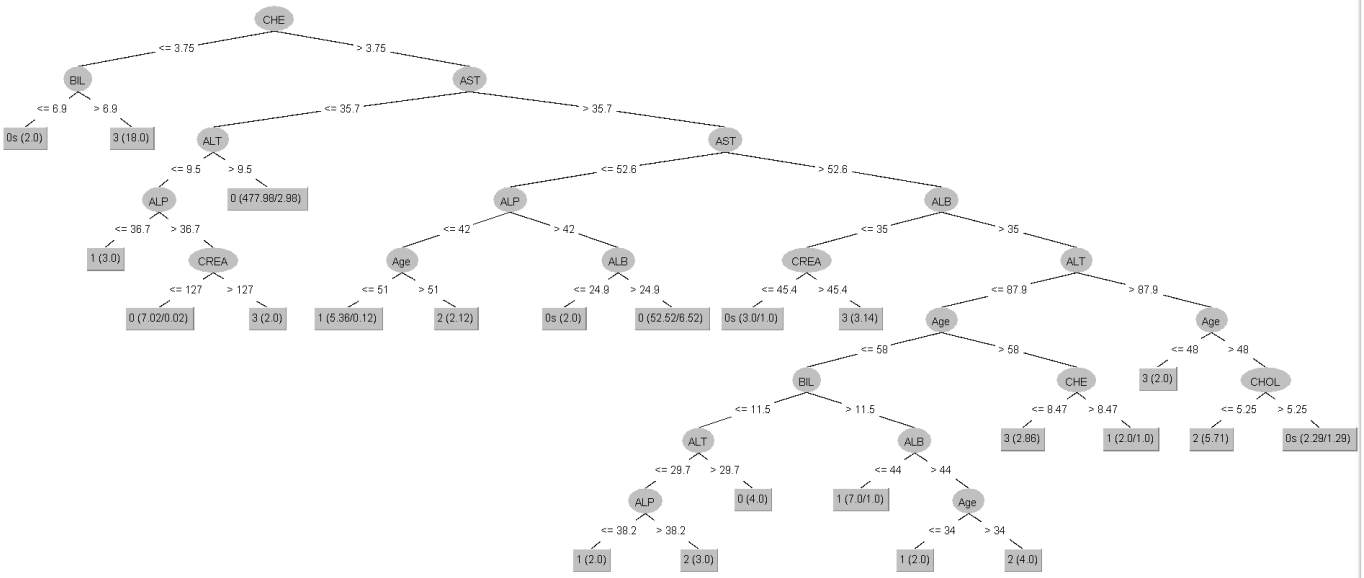
d)

a	b	c	d	e	<-- classified as
524	1	2	3	3	a = 0
4	0	0	2	1	b = 0s
11	0	3	6	4	c = 1
6	1	2	5	7	d = 2
3	0	1	3	23	e = 3

Rys. 3. Macierze pomyłek klasyfikacji przy użyciu algorytmu a) J48, b) RandomForest, c) RandomTree, d) REPTree

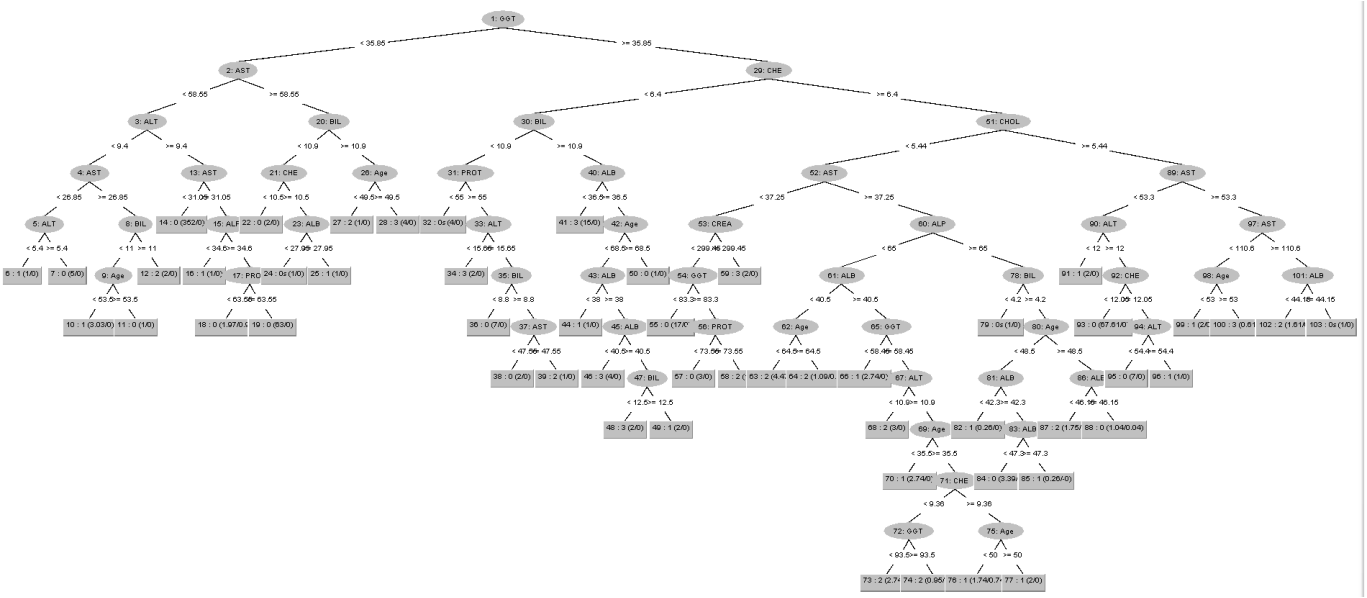
Z macierzy pomyłek wynika, że największe trudności pojawiły się z prawidłowym przypisaniem instancji należących do klasy 0s. Przy użyciu algorytmu RandomTree prawidłowo zaklasyfikowano tylko jeden przypadek należący do tej grupy, natomiast inne drzewa decyzyjne nie wykazały ani jednego przypisania True Positive. Wynika to prawdopodobnie z faktu, że grupa ta nie jest dokładnie zdefiniowana, ponieważ obejmuje osoby zdrowe podejrzane o zakażenie HCV wśród których mogą być zarówno osoby zdrowe jak i chore. Jednak, jeśli nawet wystąpiłyby charakterystyczne dla tej grupy pacjentów reguły, klasa ta jest mało liczna, więc nie ma wystarczająco dużo instancji, na przykładzie których klasyfikator mógłby nauczyć się poprawnego działania. Również pomiędzy klasami 1 i 2 (zakażenie HCV o małym zaawansowaniu i zwłóknienie wątroby) występowało wiele wzajemnych błędnych przypisań. Wśród klas obejmujących osoby chore, największą czułość i precyzję zaobserwowano w klasie 3, co świadczy prawdopodobnie o występowaniu łatwiej wykrywalnych różnic pomiędzy nią a innymi grupami. Różnice te wynikają zapewne ze znacznie zmienionej fizjologii wątroby związanej z jej bardzo dużym wyniszczeniem.

Algorytmy J48, RandomTree i REPTree umożliwiły wizualizację uzyskanych modeli w postaci drzew decyzyjnych, które przedstawiono na rysunkach 4-6.



Rys. 4. Uzyskane drzewo decyzyjne - J48

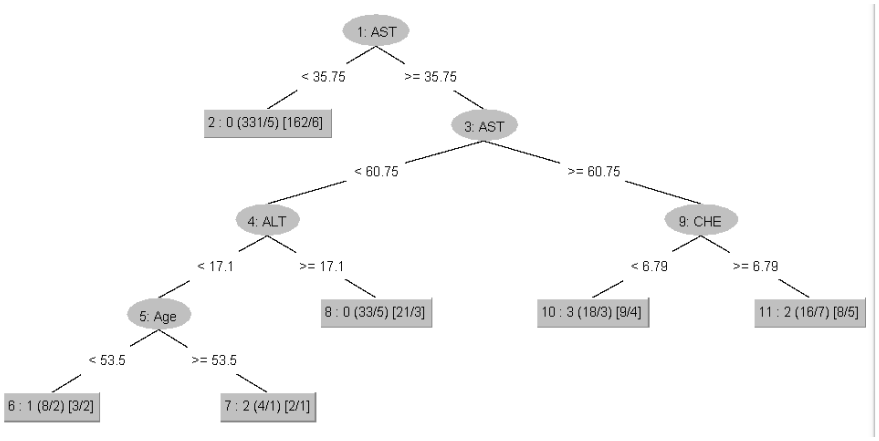
Schemat jest dość rozbudowany i bierze pod uwagę większość atrybutów. We wnioskowaniu nie zostały uwzględnione parametry GGT oraz PROT. Przypisania do poszczególnych klas mogą być dokonywane na kilka sposobów w zależności od wybranej ścieżki związanej z wartościami atrybutów.



Rys. 5. Uzyskane drzewo decyzyjne - RandomTree

Drzewo RandomTree jest bardzo rozbudowane i znajdują się na nim wszystkie analizowane atrybuty. Możliwe jest uzyskanie takiego samego wyniku klasyfikacji różnymi ścieżkami, przy czym na niektórych z nich wartość tego samego parametru jest oceniania

kilkukrotnie, np. aby zaklasyfikować instancję do klasy 0 (liść nr 7 na schemacie) należy rozpatrzyć wartość parametrów GGT, AST, ALT, a następnie ponownie AST i ALT.



Rys. 6. Uzyskane drzewo decyzyjne - REPTree

Schemat REPTree nie jest skomplikowany, co wiąże się z algorytmem jego działania i przycinaniem drzewa. Drzewo nie przedstawia ścieżki do klasy 0s, a pozostałe klasy pojawiają się na maksymalnie dwóch liściach. Podczas wnioskowania pod uwagę brane są zaledwie cztery atrybuty. Najkrótsza ścieżka ocenia wartość jedynie parametru AST. Jeśli jego wartość jest mniejsza niż 35,75 instancja przypisywana jest do klasy osób zdrowych. Ten sam parametr był brany pod uwagę w jednej z wygenerowanych reguł klasyfikujących. Jego wartość graniczna wynosiła w niej 41,94, jednak widać, że w obu algorytmach wygenerowano podobne warunki.

3.4. Modelowanie sztucznych sieci neuronowych

Klasyfikację danych przeprowadzono także przy użyciu sztucznych sieci neuronowych, czyli systemów przeznaczonych do przetwarzania informacji, których budowa i zasada działania są w pewnym stopniu wzorowane na działaniu fragmentów biologicznego systemu nerwowego. Zaletami sztucznych sieci neuronowych są możliwość komputerowego rozwiązywania praktycznych problemów bez ich wcześniejszej matematycznej formalizacji, brak konieczności odwoływania się do teoretycznych założeń na temat rozwiązywanego problemu oraz zdolność uczenia się na podstawie przykładów i możliwość automatycznego uogólniania zdobytej wiedzy (generalizacja). W strukturze sieci wyróżnia się warstwę wejściową, warstwę wyjściową oraz warstwy ukryte. Liczba warstw ukrytych waha się od zera do dwóch. Sieci o większej liczbie warstw ukrytych są inteligentniejsze, ale także trudniejsze do uczenia [10].

Do klasyfikacji użyto dwóch algorytmów: MultilayerPerceptron i RBFNetwork.

Sieć typu MLP (MultilayerPerceptron) jest rodzajem jednokierunkowej sieci neuronowej składającą się z warstwy wejściowej, jednej lub dwóch warstw ukrytych złożonych z neuronów sigmoidalnych oraz warstwy wyjściowej złożonej z neuronów sigmoidalnych lub liniowych. Uczenie perceptronu wielowarstwowego realizowane jest najczęściej przy użyciu metody wstecznej propagacji błędów.

Sieć typu RBF (RBFNetwork) to jednokierunkowa sieć neuronowa, w której wykorzystywana jest technika radialnych funkcji bazowych (RBF – Radial Basis Functions) i stosowane są neurony radialne. Sieć radialna w typowym kształcie składa się (patrz rysunek) z warstwy wejściowej, warstwy ukrytej złożonej z dużej liczby neuronów radialnych i warstwy wyjściowej, wypracowującej odpowiedź sieci. Neurony radialne służą do rozpoznawania powtarzalnych i charakterystycznych cech grup (skupisk) danych wejściowych. Konkretny neuron radialny ulega pobudzeniu, gdy sieć radialna konfrontowana jest z przypadkiem podobnym do tego, który nauczył się on wcześniej rozpoznawać jako reprezentanta pewnej grupy. W warstwie wyjściowej sieci radialnej najczęściej występuje jeden neuron liniowy.

Klasyfikatory testowano przy użyciu walidacji krzyżowej (*Cross-validation, folds 10*). Zbudowano kilka sieci neuronowych każdego typu przy różnych wartościach parametrów. W przypadku sieci MLP zmieniane były konfiguracja warstw ukrytych, współczynnik uczenia oraz współczynnik momentum, natomiast sieć RBF modyfikowano poprzez ustalanie wartości minimalnego odchylenia standardowego i liczby klastrów. Ze względu na bardzo długi czas budowania klasyfikatora, jego testowanie odbyło się przy wyłączonej opcji wyświetlania schematu sieci (*GUI: False*).

W przypadku klasyfikatora MultilayerPerceptron największą skuteczność klasyfikacji wśród sieci o jednej warstwie ukrytej uzyskano stosując konfigurację t warstw ukrytych, współczynnik uczenia na poziomie 0,3 oraz współczynnik momentum o wartości 0,2. Testowano także sieci o większej liczbie warstw ukrytych. Największą skuteczność uzyskano stosując sieć trzywarstwową o liczbie węzłów 10, 11 i 12. Ograniczona wydajność komputera nie pozwoliła na tworzenie i testowanie bardziej rozbudowanych sieci. Z kolei sieć RBF uzyskała najlepsze wyniki przy minimalnym odchyleniu standardowym wynoszącym 1 i 20 klastrach. Możliwe, że przy większej liczbie klastrów, skuteczność klasyfikacji byłaby wyższa, jednak niemożliwe było przetestowanie tego typu sieci ze względu na bardzo długi czas budowania modelów. Parametry oceny tych klasyfikatorów zestawiono w tabeli 6.

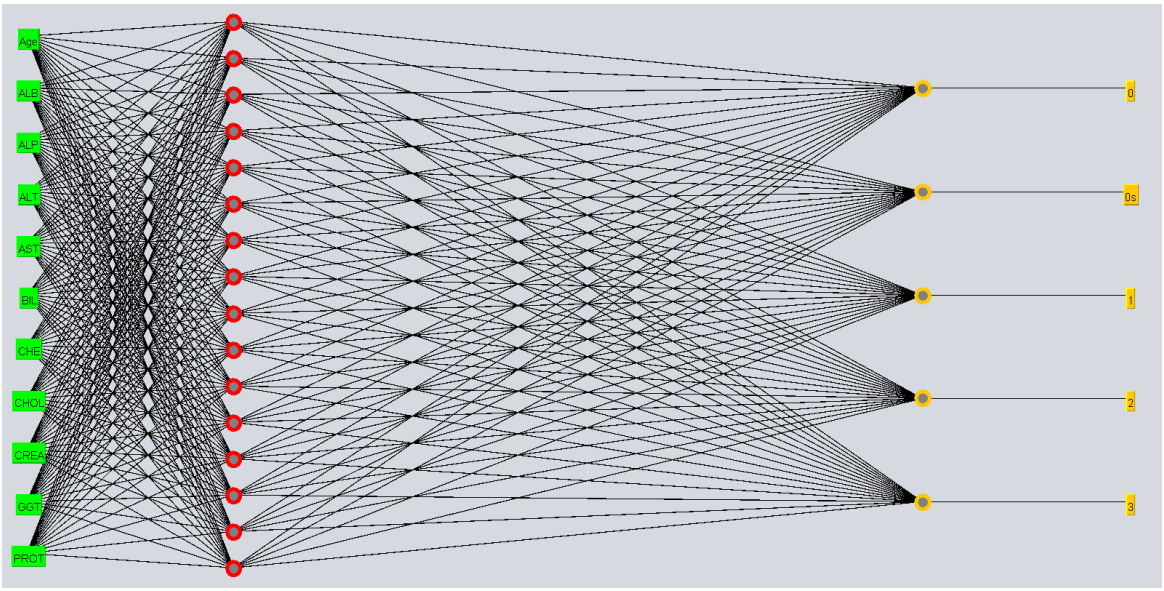
Tabela 6. Zestawienie parametrów oceny klasyfikacji

Parametr oceny klasyfikacji	Sieć MLP warstwy ukryte t	Sieć MLP warstwy ukryte 10,11,12	Sieć RBF
Skuteczność klasyfikacji [%]	93,66	93,82	89,27
Średni błąd bezwzględny	0,03	0,04	0,06
Średni błąd kwadratowy	0,14	0,14	0,18
Statystyka Kappa	0,73	0,73	0,49
TP Rate	0,937	0,938	0,893
FP Rate	0,150	0,129	0,403
Czułość	0,937	0,938	0,893
Precyzja	0,932	0,927	0,875
Pole pod krzywą ROC	0,956	0,954	0,914

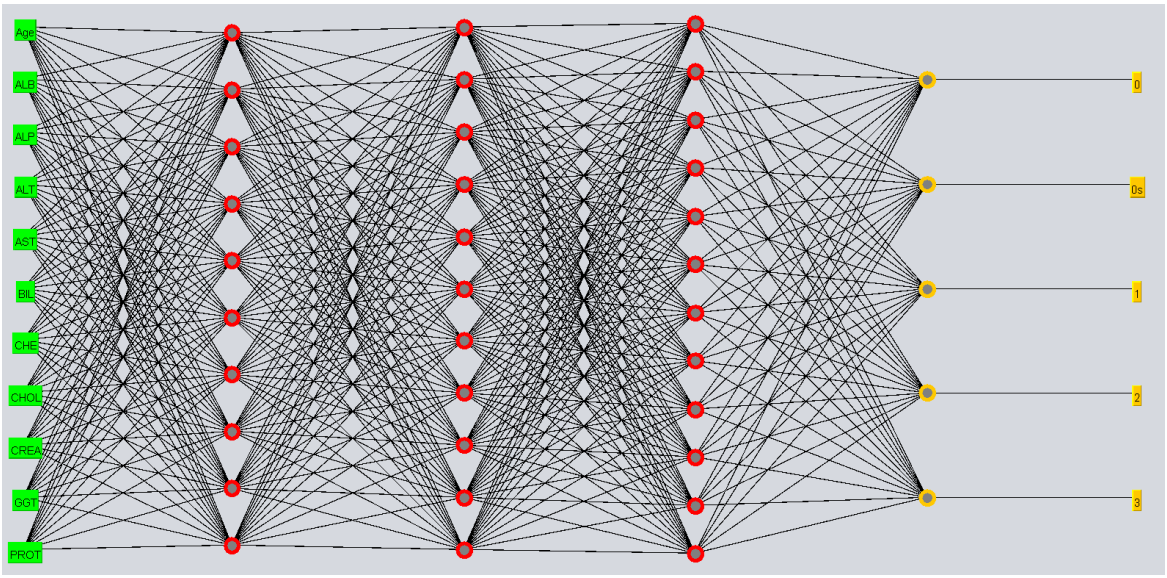
Z podsumowania parametrów oceny klasyfikacji wynika, że największą skuteczność uzyskano stosując bardziej rozbudowaną sieć typu MLP. Jednak sieć o jednej warstwie ukrytej otrzymała bardzo podobne wyniki skuteczności, a jednocześnie wykazała większą precyzję i mniejszy średni błąd bezwzględny. Sieć typu RBF uzyskała gorsze wyniki wszystkich parametrów niż sieć MLP. Jej skuteczność wynosiła niecałe 90%.

Poniżej (Rys. 7) przedstawiono schematy obu zastosowanych sieci MLP, na których widoczne są różnice pomiędzy ich strukturą.

a)



b)



Rys. 7. Uzyskana sieć typu MLP a) warstwy ukryte t, b) warstwy ukryte 10,11,12

Sieć typu t (Rys. 7a) posiada jedną warstwę ukrytą, która składa się z liczby węzłów równej sumie liczby atrybutów i klas. W analizowanym przypadku węzłów jest 16. Z kolei liczba warstw ukrytych i węzłów drugiej z zastosowanych sieci została zdefiniowana ręcznie. Stworzono trzy warstwy ukryte, z których każda miała inną liczbę węzłów. Pierwsza składała się z 10 węzłów, druga z 11, natomiast trzecia z 12 węzłów (Rys. 7b). Jak widać, sieć

trójwarstwowa posiada zdecydowanie więcej połączeń, co sprawia, że jest bardziej rozbudowana.

W celu porównania poprawności przypisania instancji należących do poszczególnych klas, przedstawiono macierze pomyłek analizowanych sieci MLP. Sieć RBF pominięto, ponieważ uzyskała widocznie gorsze wyniki.

a)

a	b	c	d	e	<-- classified as
528	1	2	1	1	a = 0
3	4	0	0	0	b = 0s
6	0	11	4	3	c = 1
4	1	5	10	1	d = 2
1	0	0	6	23	e = 3

b)

a	b	c	d	e	<-- classified as
531	1	0	1	0	a = 0
4	0	1	2	0	b = 0s
7	0	10	4	3	c = 1
1	0	4	12	4	d = 2
0	0	1	5	24	e = 3

Rys. 8. Macierze pomyłek klasyfikacji przy użyciu sieci neuronowej typu MLP a) z jedną warstwą ukrytą, b) z trzema warstwami ukrytymi

Jak widać na podstawie macierzy pomyłek, sieć z trzema warstwami ukrytymi, mimo, że dała lepsze rezultaty w ogólnym podsumowaniu klasyfikacji, gorzej poradziła sobie z poprawnym przypisywaniem instancji należących do klas 0s oraz 1 niż sieć z jedną warstwą ukrytą. Żaden z przypadków z grupy pacjentów z podejrzeniem HCV (klasa 0s) nie został sklasyfikowany prawidłowo przez tę sieć, większość tych instancji została przypisana do klasy osób zdrowych. Występujące błędy klasyfikacji są analogiczne do błędów popełnianych przez algorytmy drzew decyzyjnych. Instancje klasy 0s są często określane jako należące do klasy 0, podobnie jak przypadki z klasy 1. Z kolei klasa 2 jest mylnie uznawana za klasę 1, a klasa 3 za klasę 2. Najpoprawniej klasyfikowana jest bardzo liczna klasa osób zdrowych, co pozwala stwierdzić, że skuteczność sieci neuronowych poprawiłaby się, gdyby zbiór danych dotyczących osób chorych został zwiększony i algorytmy mogłyby uczyć się na większej liczbie instancji.

3.5. Model regresji logistycznej

W celu stworzenia modelu matematycznego, który opisywałby prawdopodobieństwo, że pacjent jest osobą zdrową biorąc pod uwagę jego wiek i parametry biochemiczne krwi, zastosowano regresję logistyczną. Zbiór danych jest liczny, co wspiera zastosowanie tej metody.

Model logistyczny określa prawdopodobieństwo warunkowe, że zmienna Y przyjmuje wartość równą 1 dla zmiennych niezależnych $x_1, x_2, \dots, x_k$.

$$P(Y = 1 \,|x_1, x_2, \dots, x_k) = P(X) = \frac{e^{a_0 + \sum_{i=1}^{k} a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^{k} a_i x_i}}$$

gdzie:

Y - zmienna dychotomiczna przyjmująca wartość Y = 1 dla zdarzeń pożądanych (w analizowanym przypadku - osoba zdrowa) lub Y = 0 dla zdarzeń niepożądanych (w analizowanym przypadku - osoba chora),

$a_i, i = 0, \dots, k$ - współczynniki regresji,

$x_1, x_2, \dots, x_k$ - zmienne niezależne.

Dodatkowo oblicza się także iloraz szans, który określa stosunek szans wystąpienia danego zdarzenia w jednej grupie do szansy jego wystąpienia w innej grupie.

$$S(A) = \frac{P(A)}{1 - P(A)}$$

Jak widać, model regresji logistycznej wymaga jednak występowania jedynie dwóch klas, dlatego przed jego zbudowaniem zmodyfikowano zbiór danych. Klasy 1, 2 oraz 3 połączono w jedną oznaczającą osobę chorą. Klasa 0 pozostała bez zmian i dotyczyła osób zdrowych. Z kolei przypadki należące do klasy 0s zostały usunięte ze zbioru danych, gdyż nie można było jednoznacznie określić, czy są to osoby zdrowe czy chore. Klasa ta była mało liczna, więc wykluczenie tych instancji nie zaburzyło zbioru danych. Ostatecznie zastosowany zbiór zawierał 608 instancji.

Wykonano analizę regresji logistycznej danych przy zastosowaniu ich automatycznego podziału na zbiory uczący i testujący różnymi metodami (*Percentage split* oraz *Cross-validation*), a także przy różnych wartościach parametrów tych metod. Najlepszy klasyfikator uzyskano stosując *Percentage split = 65%*. W wyniku podziału zbiór uczący składał się z 395 instancji, a zbiór testujący z 213 instancji. Wyniki działania algorytmu przedstawiono na kolejnych rysunkach.

```
Correctly Classified Instances         208               97.6526 %
Incorrectly Classified Instances         5                2.3474 %
Kappa statistic                          0.8759
Mean absolute error                      0.047
Root mean squared error                  0.1514
Relative absolute error                 21.7897 %
Root relative squared error             47.0155 %
Total Number of Instances              213

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,200    0,974      1,000    0,987      0,883    0,946     0,987     zdrowy
                 0,800    0,000    1,000      0,800    0,889      0,883    0,946     0,906     chory
Weighted Avg.    0,977    0,177    0,977      0,977    0,975      0,883    0,946     0,977

=== Confusion Matrix ===

   a   b   <-- classified as
 188   0 |   a = zdrowy
   5  20 |   b = chory
```

Rys. 9. Podsumowanie działania klasyfikatora

Dokładność zbudowanego klasyfikatora jest bardzo wysoka. Prawidłowo zostało przyporządkowanych aż 97,6% instancji ze zbioru testującego. Błędnie sklasyfikowane zostało 5 z 25 instancji, które w rzeczywistości należały do klasy osób chorych. Oznacza to, że klasyfikator charakteryzuje się wysoką precyzją, ale niższą czułością w przypadku

wykrywania instancji należących do klasy "chory". Wynika z tego, że algorytm może błędnie uznać diagnozowany przypadek za osobę zdrową, pomimo, że będzie ona chora. Uzyskane wartości parametrów oceny klasyfikatora takie jak pole pod krzywą ROC, czy współczynnik F-Measure są wysokie, co wskazuje na dobrą dokładność klasyfikacji.

Obliczone zostały także współczynniki regresji logistycznej oraz iloraz szans, których wartości przedstawiono poniżej.

a)

Coefficients...

Variable	Class zdrowy
Age	0.0088
ALB	0.2011
ALP	0.0713
ALT	0.0163
AST	-0.0918
BIL	-0.0752
CHE	-0.1276
CHOL	0.9492
CREA	-0.0235
GGT	-0.0351
PROT	-0.2149
Intercept	8.3981

b)

Odds Ratios...

Variable	Class zdrowy
Age	1.0089
ALB	1.2227
ALP	1.0739
ALT	1.0164
AST	0.9123
BIL	0.9275
CHE	0.8802
CHOL	2.5836
CREA	0.9768
GGT	0.9655
PROT	0.8066

Rys. 10. Współczynniki regresji (a) oraz iloraz szans (b)

Na podstawie uzyskanych współczynników można zapisać funkcję logitową jako:

$$g(x) = 8{,}3981 + 0{,}0088 \cdot wiek + 0{,}2011 \cdot ALB + 0{,}0713 \cdot ALP + 0{,}0163 \cdot ALT - 0{,}0918 \cdot AST - 0{,}0752 \cdot BIL - 0{,}1276 \cdot CHE + 0{,}9492 \cdot CHOL - 0{,}0235 \cdot CREA - 0{,}0351 \cdot GGT - 0{,}2149 \cdot PROT$$

Uzyskany model regresji logistycznej można przedstawić równaniem:

$$P(Y = 1|X) = \frac{e^{8{,}3981+0{,}0088 \cdot wiek+0{,}2011 \cdot ALB+0{,}0713 \cdot ALP+0{,}0163 \cdot ALT-0{,}0918 \cdot AST-0{,}0752 \cdot BIL-0{,}1276 \cdot CHE+0{,}9492 \cdot CHOL-0{,}0235 \cdot CREA-0{,}0351 \cdot GGT-0{,}2149 \cdot PROT}}{1 + e^{8{,}3981+0{,}0088 \cdot wiek+0{,}2011 \cdot ALB+0{,}0713 \cdot ALP+0{,}0163 \cdot ALT-0{,}0918 \cdot AST-0{,}0752 \cdot BIL-0{,}1276 \cdot CHE+0{,}9492 \cdot CHOL-0{,}0235 \cdot CREA-0{,}0351 \cdot GGT-0{,}2149 \cdot PROT}}$$

Powyższe równanie, po podstawieniu wartości atrybutów, pozwala określić prawdopodobieństwo, że dana instancja należy do klasy osób zdrowych.

Model regresji logistycznej zbudowano także przy użyciu modułu Stepwise Model Builder w programie Statistica. W tym przypadku funkcja logitowa ma postać:

$$g(x) = 2{,}7515 + 0{,}0287 \cdot ALB + 0{,}0200 \cdot ALT - 0{,}0983 \cdot AST - 0{,}0580 \cdot BIL - 0{,}1379 \cdot CHE + 0{,}9028 \cdot CHOL - 0{,}0037 \cdot CREA - 0{,}0241 \cdot GGT$$

Jak widać, uzyskana funkcja nie uwzględnia wieku, a także parametrów ALP oraz PROT, ponieważ zostały one odrzucone przez algorytmu podczas budowania modelu. Większość pozostałych atrybutów otrzymało podobnej wartości współczynniki, jak w przypadku modelu zbudowanego w programie WEKA. Jedynie kreatynina (CREA) została uznana za mniej istotny czynnik.

Ostatecznie model można przedstawić jako:

$$P(Y = 1|X) == \frac{e^{2,7515+0,0287 \cdot ALB+0,0200 \cdot ALT-0,0983 \cdot AST-0,0580 \cdot BIL-0,1379 \cdot CHE+0,9028 \cdot CHOL-0,0037 \cdot CREA-0,0241 \cdot GGT}}{1+e^{2,7515+0,0287 \cdot ALB+0,0200 \cdot ALT-0,0983 \cdot AST-0,0580 \cdot BIL-0,1379 \cdot CHE+0,9028 \cdot CHOL-0,0037 \cdot CREA-0,0241 \cdot GGT}}$$

Model charakteryzuje się współczynnikiem determinancji na poziomie R^2 = 0,7293, co oznacza, że w 73% wyjaśnia dlaczego osoba jest zdrowa. Jest to dość dobry wynik.

4. Wnioski

Zastosowanie metod eksploracji danych w niniejszym projekcie umożliwiło wygenerowanie z analizowanego zbioru reguł i modeli klasyfikacyjnych dotyczących rozpoznawania zakażenia HCV oraz stopnia jego zaawansowania (zakażenie bez zaawansowanych objawów, zwłóknienia wątroby oraz marskość wątroby).

Zbadanie ważności atrybutów różnymi testami ważności stosując różne metody wyszukiwania pozwoliło na ograniczenie ich liczby i uwzględnianie jedynie wieku i wybranych parametrów biochemicznych krwi pacjenta jako istotnych czynników. Przy użyciu algorytmu Apriori wygenerowano reguły klasyfikacyjne dotyczące przypisywania pacjenta do klasy osób zdrowych oraz, po zmodyfikowaniu zbioru danych, przyporządkowywania instancji z grupy osób chorych do odpowiedniego stadium choroby.

Drzewa decyzyjne są czytelnym i intuicyjnym do odczytania sposobem wizualizacji klasyfikacji danych. Oprogramowanie WEKA pozwala na tworzenie ich w łatwy i szybki sposób przy użyciu wielu rodzajów klasyfikatorów. Zaletą tworzenia klasyfikacji w tym programie jest również automatyczne wyświetlanie wielu parametrów oceniających jej dokładność, w tym macierzy pomyłek, co bardzo ułatwia sprawdzenie, czy stworzona klasyfikacja jest zadowalająca.

Klasyfikatory różnią się między sobą sposobem działania, co sprawia, że mają różną dokładność przy klasyfikowaniu tych samych danych. W przypadku analizowanych danych najlepszym algorytmem indukującym drzewa decyzyjne okazał się RandomForest (skuteczność klasyfikacji na poziomie 92,5%). Z kolei w przypadku sztucznych sieci neuronowych lepsze wyniki uzyskano stosując sieć typu MLP niż RBF.

Sieć MLP umożliwia ustalenie konfiguracji warstw ukrytych. Sieć o trzech warstwach z kilkunastoma węzłami tworzącymi każdą z nich dała podobne rezultaty, co sieć jednowarstwowa z większą liczbą węzłów. Tworzenie bardziej rozbudowanych sieci wymaga bardzo wysokiej wydajności komputera.

Modele logistyczne pozwoliły matematycznie opisać prawdopodobieństwo, że dana osoba jest osobą zdrową. Stosując różne algorytmy (w programach WEKA i Statistica) na tych samych danych uzyskano różniące się między sobą modele. Charakteryzowały się dobrą skutecznością (model w WEKA 97,65% poprawnie przypisanych instancji, model w Statistica R^2=0,73). Czułość modelu mogłaby być wyższa - pojawiały się problemy z klasyfikacją instancji należących do klasy osób chorych. Analogicznie można byłoby zbudować modele dla wystąpienia choroby. Należy jednak pamiętać, że regresja logistyczna

ogranicza się do takich zbiorów danych, w których instancje mogą być przypisywane jedynie do dwóch klas.

Problemem okazała się niewielka liczność klas innych niż klasa 0 (osoby zdrowe). Z tego powodu klasyfikatory popełniały wiele błędów w przypisywaniu instancji należących do klas osób chorych i podejrzanych o zakażenie (0s, 1, 2, 3), a także do generowania reguł związanych jedynie z klasą osób zdrowych. Zwiększenie zbioru danych dotyczących mniej licznych grup pozwoliłoby klasyfikatorom na dokładniejsze uczenie, co poprawiłoby skuteczność klasyfikacji.

Przed rozpoczęciem analizy należy zapoznać się ze zbiorem danych. Niektóre atrybuty mogą zakłócać proces klasyfikacji. Identyfikator instancji może być nadawany losowo i powodować próbę wyszukiwania nieistniejących w rzeczywistości reguł. Jednak może być także ustalany zgodnie z przynależnością przypadku do klasy. W takiej sytuacji uwzględnienie go w analizach spowoduje uznanie go za bardzo ważny atrybut i opieranie klasyfikacji na jego wartości, pomimo, że nie ma on nic wspólnego z badanym zjawiskiem, np. chorobą.

Analiza statystyczna umożliwia zapoznanie się z danymi. Zastosowanie testów takich jak test Kruskala-Wallisa lub ANOVA, czy analiza korelacji pozwalają na wykrycie różnic pomiędzy grupami lub zależności między zmiennymi. Dzięki temu możliwy jest wybór atrybutów istotnych z punktu widzenia dalszych analiz, podobnie jak przy wykonywaniu badania ważności atrybutów przy użyciu algorytmów w programie WEKA. Analiza statystyczna umożliwia także wizualizację danych.

Metody eksploracji danych mogą być użyteczne w medycynie do wyszukiwania parametrów diagnostycznych. Na podstawie bazy danych algorytmy pozwalają określić podobieństwa pomiędzy instancjami należącymi do tej samej klasy oraz różnice pomiędzy przypadkami z różnych klas. Na tej podstawie możliwe jest nauczenie klasyfikatora wykrywania wartości wskazujących na chorobę i diagnozowania. Klasyfikatory te muszą charakteryzować się wysoką skutecznością, a także czułością, żeby nie doprowadzić do zbagatelizowania choroby i niewprowadzenia leczenia.

5. Literatura

[1] Demski T.: Od pojedynczych drzew do losowego lasu. StatSoft Polska sp. z o.o. Dostęp online [15.01.2021]: statsoft.pl

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Dostęp online [15.01.2021]: https://archive.ics.uci.edu/ml/datasets

[3] Frank E., Hall M.A., Pal C.J., Witten I.H.: The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016

[4] Ho, The random subspace method for constructing decision forests, „IEEE Transactions on Pattern Analysis and Machine Intelligence", 20 (8), 1998, s. 832 - 844,

[5] Hoffmann G et al. Using machine learning techniques to generate laboratory diagnostic pathways - a case study. J Lab Precis Med 2018; 3: 58-67

[6] Krawczyńska-Piechna A.: Predicting the length of a post-accident absence in construction with decision trees and their ensembles. Archives of Civil Engineering. 3(2020), s. 365-376

[7] Kurs Data Mining WEKA. Dostęp online [15.01.2021]: https://weka.sourceforge.io/

[8] Lichtinghagen R et al. J Hepatol 2013; 59: 236-42

[9] Materiały dotyczące oprogramowania WEKA. Dostęp online [15.01.2021]: https://weka.sourceforge.io/

[10] Tadeusiewicz R., Szaleniec M.: Leksykon sieci neuronowych. Wydawnictwo Fundacji "Projekt Nauka". Wrocław 2015

[11] Zalewska M., Zalewski W.: Zastosowanie metody drzewa decyzyjnego w analizie problemów makroekenomicznych. Economics and Management, 4/2012, s. 58-69