POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki

Praca dyplomowa magisterska

ANALIZA RZECZYWISTEGO PROBLEMU MEDYCZNEGO Z WYKORZYSTANIEM REGUŁOWYCH METOD WSPOMAGANIA DECYZJI

Kornelia Staszewska, 136803

Promotor
dr hab. inż. Miłosz Kadziński, prof. PP

POZNAŃ 2022

Tutaj będzie karta pracy dyplomowej;
oryginał wstawiamy do wersji dla archiwum PP, w pozostałych kopiach wstawiamy ksero.

Spis treści

Streszczenie

Celem pracy jest zastosowanie wybranych regułowych i drzewiastych metod wspomagania decyzji do dwóch rzeczywistych zbiorów danych opisujących pary chcące skorzystać z metody in vitro. Ze względu na powszechność problemu niepłodności wzrasta popularność stosowania technik wspomaganego rozrodu. Dla usprawnienia fazy klasyfikacji pacjentów do odpowiedniej kategorii można wykorzystać metody wspomagania decyzji. W literaturze opisane jest zastosowanie metody Electre Tri C do tego problemu. Analiza przedstawiona w niniejszej pracy obejmuje zarówno dobór metod, ustalenie ich najkorzystniejszych parametryzacji w toku eksperymentów walidacyjnych, jak i omówienie otrzymanych rezultatów. Przedstawiono opis użytych algorytmów analizy decyzji, wykorzystujących reguły lub drzewa decyzyjne, wraz z przykładami ilustrującymi ich działanie. Zostały one przygotowane na podstawie jednego z rozpatrywanych zbiorów. Dla metod regułowych dane wejściowe stanowią przybliżenia unii klas przygotowane dzięki zastosowaniu DRSA i VC-DRSA. Algorytmy DOMLEM i DOMApriori posłużyły do indukcji odpowiednio minimalnych i satysfakcjonujących zbiorów reguł, zaś do klasyfikacji wariantów wykorzystane zostały dwa podejścia, proste i zaawansowane. Modele drzewiaste zbudowano przy wykorzystaniu CART i losowych lasów. W pracy zawarta jest także interpretacja uzyskanych reguł i drzew oraz ocena skuteczności działania poszczególnych modeli.

Abstract

The main goal of this thesis is to apply rule- and tree-based methods to two real datasets that refer to infertile couples who want to try in vitro. Infertility is a common issue, which is why assisted reproduction techniques (ART) are growing in popularity. In vitro states aa the most frequently used ones. To improve its first stage, which includes the classification of couples into predefined categories, decision support methods can be applied. In the literature, we can find an example in which the Electre Tri C method was investigated. This thesis emphasizes the selection of appropriate methods, their parametrization by validation experiments, evaluation of models, and the interpretation of results. It also includes a description of the applied methods with detailed examples based on one of the considered datasets. Regarding rule-based approaches, an approximation of unions of classes was computed based on the DRSA and VC-DRSA approaches. To induce minimal and satisfactory sets of decision rules, DOMLEM and DOMApriori algorithms were applied. Object classification was also investigated in two variants, simple and advanced. In the decision trees section, CART and Random Forest were presented and applied.

Rozdział 1

Wstęp

Niepłodność definiowana jest jako brak możliwości uzyskania ciąży przez okres 12 miesięcy, mimo regularnych starań, za które przyjmuje się stosunki płciowe odbywane od 2 do 4 razy w tygodniu i rezygnację ze stosowania jakichkolwiek metod antykoncepcji. Została ona uznana przez Światową Organizację Zdrowia za chorobę cywilizacyjną XXI wieku. Według statystyk dotyka ona jedną na dziesięć par w wieku rozrodczym na całym świecie. W Europie problem obserwowany jest w jeszcze szerszej skali – boryka się z nim 18% par [5]. Wymienione elementy bezpośrednio wpływają na popularność stosowania tak zwanych technik wspomaganego rozrodu. Jedną z najpopularniejszych jest in vitro, polegające na połączeniu komórki jajowej i plemnika poza organizmem kobiety – w warunkach laboratoryjnych, a następnie przeniesieniu zarodka do jamy macicy. Kluczowym etapem przygotowawczym do zastosowania in vitro jest sformułowanie poprawnej oceny pary.

W tradycyjnym podejściu lekarz zobowiązany był do ręcznej analizy informacji o pacjentach i wyciągnięcia odpowiednich wniosków pozwalających na ich właściwą klasyfikację na podstawie wiedzy eksperckiej. W literaturze opisane zostało podejście proponujące usprawnienie procesu, wykorzystujące metodę Electre Tri C [7] do wstępnej klasyfikacji par. W artykule zaprezentowane zostało zastosowanie konkretnej metody wspomagania decyzji do problemu leczenia niepłodności metodą in vitro. Co więcej, dla innych problemów medycznych zastosowanie podejść z dziedziny wspomagania decyzji jest popularnym rozwiązaniem stosowanym na przykład do analizy problemu leczenia schorzeń nowotworowych (patrz np. [3], [11] oraz [18]).

Istotnym aspektem takiej analizy jest określenie jej celu. W przypadku opisanym w pracy jest nim z jednej strony maksymalizacja prawdopodobieństwa osiągnięcia ciąży przez parę, a z drugiej – minimalizowanie prawdopodobieństwa, że będzie to ciąża mnoga. Zaproponowane modele miałyby docelowo sugerować odpowiednią ocenę dla danej pary, przypisaną przy wzięciu pod uwagę nie tylko opisu pacjentów, lecz także danych historycznych, zgromadzonych przy ocenie par, które dotychczas skorzystały z pomocy kliniki. Następnie, rekomendacja systemu mogłaby być analizowana przez specjalistę leczenia niepłodności. Z tym ostatnim związany jest wymóg, aby sposób podejmowania decyzji przez model był przejrzysty i zrozumiały dla człowieka. Taki warunek spełniają modele regułowe oraz drzewa decyzyjne czy ich zespoły.

1.1 Cel i zakres pracy

Celem pracy jest wykonanie analizy rzeczywistego problemu medycznego, którym jest leczenie niepłodności metodą in vitro, przy wykorzystaniu odpowiednio dobranych metod wspomagania decyzji. Dane wejściowe wykorzystane w pracy stanowią dwa zbiory pozyskane w klinice leczenia niepłodności w Lizbonie. Warto zaznaczyć, że atrybuty obserwowane w obu zbiorach nie są

takie same, jednak widoczne są powiązania semantyczne między nimi. Ponadto obserwacje zgromadzone w obu zbiorach są mało liczne, dlatego w pracy zdecydowano się przeprowadzić analizę przy wykorzystaniu kilku metod, które można podzielić na dwie grupy. Pierwszą stanowią metody wielokryterialnego sortowania, zaś drugą drzewa decyzyjne. Motywacją dla wykorzystania tych podejść jest ich wyjaśnialność oraz łatwość interpretacji wykorzystywanych przez nie modeli preferencji. Co istotne, w pracy użyto dwóch grup metod również ze względu na chęć porównania skuteczności modeli regułowych i drzewiastych. W zakres wykonanych badań wchodzą:

- analiza statystyczna zbiorów danych;
- dobór i wykorzystanie metod wielokryterialnego sortowania i drzew decyzyjnych;
- przeprowadzenie eksperymentów walidacyjnych i wybór najlepszych parametryzacji dla poszczególnych algorytmów;
- interpretacja otrzymanych wyników;
- porównanie rezultatów otrzymanych różnymi metodami zarówno dla tych samych zbiorów, jak i ogólnie dla rozpatrywanego problemu.

1.2 Struktura pracy

Praca składa się z sześciu rozdziałów, z których pierwszy przedstawia rozważany problem medyczny, motywację dla jego wyboru oraz zakres przeprowadzonych eksperymentów.

Rozdział 2 poświęcony jest analizie zbiorów danych, uwzględniającej interpretację poszczególnych atrybutów, statystyki i porównanie struktur obserwacji. Zawiera on także uwagi, które mogą wpływać na wyniki dalszych eksperymentów.

W Rozdziale 3 przedstawione są metody wielokryterialnego sortowania oraz drzewa CART i losowe lasy wykorzystane w pracy. Zawarte jest też uzasadnienie motywacji stojącej za ich wyborem. Działanie algorytmów jest zilustrowane przykładami pochodzącymi z pierwszego z analizowanych zbiorów danych.

Rozdział 4 stanowi raport wyników, otrzymanych metodami opisanymi w rozdziale 3. Zrealizowana jest w nim także próba porównania rezultatów, a przedstawione wnioski dotyczą zarówno skuteczności działania dla poszczególnych zbiorów danych, jak i ogólnie dla analizowanego problemu medycznego w odniesieniu do obu zbiorów.

Podsumowanie pracy zawarte jest w Rozdziale 5. Stanowi ono syntezę wniosków sformułowanych w pracy, wskazuje możliwe kierunki dalszych badań nad zagadnieniem i komentuje zaobserwowane trudności czy ograniczenia.

Rozdział 2

Analiza zbiorów danych

Wykorzystane w pracy dwa rzeczywiste zbiory, zawierające dane dotyczące leczenia niepłodności, zostały przygotowane na podstawie informacji z prywatnej kliniki płodności w Lizbonie (CEME-ARE). Decydentami, przypisującymi odpowiednie kategorie poszczególnym parom rozumianym w problemie jako warianty, byli embriolodzy z wspomnianej placówki medycznej.

Dla pierwszego zbioru przypisanie poszczególnych kategorii odbywało się przy uwzględnieniu ocen wariantów na siedmiu kryteriach [7]:

- age – wiek kobiety, istotny dla oceny pary ze względu na fakt, iż młodsze kobiety mają większe szanse na pozytywny wynik aplikacji metod wspomaganego rozrodu;
- infertility – całkowita liczba lat trwania niepłodności; im jest ona większa tym mniejsza szansa na sukces zastosowania leczenia;
- oocytes – wskaźnik pozyskanych oocytów, ustalany na podstawie wartości różnicy liczby pozyskanych oocytów i wartości idealnej równej 12; na kryterium może zostać przypisana jedna z 7 możliwych wartości:
  - 1 – jeśli różnica ma wartość 0,
  - 2 – jeśli różnica przyjmuje wartość 1 lub -1,
  - 3 – jeśli różnica wynosi -2 lub jest większa bądź równa 2,
  - 4 – jeśli różnica ma wartość - 3 lub -4,
  - 5 – jeśli różnica to -5 lub -6,
  - 6 – jeśli różnica ma wartość -7 lub -8,
  - 7 – jeśli różnica przyjmuje wartość mniejszą bądź równą -9.

  Im przypisana wartość jest mniejsza, tym jest ona korzystniejsza dla pary.
- woman_eval – ogólna ocena zdrowia kobiety dokonana przez ginekologa lub położnika; jest złożonym kryterium, na którym kobieta może otrzymać jedną z wartości całkowitoliczbowych od 1 do 7, przy czym preferowane są wyższe wartości;
- sperm – ocena jakości spermy, przyjmująca:
  - wartość 1 dla kriokonserwowanych plemników z biopsji jąder,
  - wartość 2 dla kriokonserwowanej spermy z ejakulacji,
  - wartość 3 dla spermy z ejakulacji.

Preferowane są większe wartości. Kryterium stanowi jedyny męski czynnik w procesie decyzyjnym.

- morpho_quality – suma z wartości oceny morfologicznej czterech embrionów niezależnie w skali całkowitoliczbowej od 1 do 5, w której preferowane są większe wartości;
- develop_quality – średnia wartość oceny rozwoju czterech embrionów w skali całkowitoliczbowej od 1 do 5, w której preferowane są większe wartości.

Kryteria woman_eval, morpho_quality i develop_quality są bardzo złożone. Przygotowanie ocen na nich wymaga zaangażowania w cały proces ginekologów, położników i embriologów. Bierze pod uwagę również ich doświadczenie.W Tabeli 2.1 zostało przedstawione skrócone podsumowanie informacji o wyżej opisanych kryteriach.

Tablica 2.1: Skrócona interpretacja kryteriów dla zbioru pierwszego

Nazwa kryterium	Opis kryterium	Typ wartości	Kierunek preferencji
age	wiek kobiety	liczba całkowita	koszt
infertility	liczba lat trwania niepłodności	liczba całkowita	koszt
oocytes	efektywna liczba oocytów	liczba całkowita z przedziału od 1 do 7	koszt
woman_eval	ogólna ocena płodności kobiety	liczba całkowita z przedziału od 1 do 7	zysk
sperm	pochodzenie spermy	liczba całkowita z przedziału od 1 do 3	zysk
morpho_quality	morfologiczna jakość czterech najlepszych embrionów	liczba całkowita z zakresu od 4 do 20	zysk
develop_quality	ocena rozwoju czterech najlepszych embrionów	liczba całkowita z zakresu od 1 do 5	zysk

Jeśli chodzi o klasy decyzyjne przypisywane wariantom to istnieją cztery możliwości:

- wartość 1 dla transferu czterech embrionów,
- wartość 2 dla transferu trzech embrionów,
- wartość 3 dla transferu dwóch embrionów,
- wartość 4 dla transferu jednego embrionu.

Preferowane są wyższe wartości – im para otrzymuje kategorię o wyższym numerze tym większa szansa na powodzenie zastosowania metod wspomaganego rozrodu oraz tym mniejsza szansa na wystąpienie niepożądanych efektów, za które uznawane są na przykład ciąże mnogie.

Drugi wykorzystany w pracy zbiór przypisanie poszczególnych klas decyzyjnych również uzależniał od ocen wariantów na siedmiu kryteriach [4]:

- age – wiek kobiety w momencie pozyskiwania oocytów, mniejsze wartości są preferowane;
- pregnancy – informacja o tym, czy parze udało się wcześniej osiągnąć ciążę – jeśli tak przypisana zostaje wartość 1, w przeciwnym wypadku – 0, korzystniejsza jest ocena 1;
- attempts – kryterium, na którego wartość wpływają dwa czynniki:
  - całkowita liczba pozyskanych oocytów,
  - liczba zakończonych porażką prób transferu embrionów.

  Jeśli liczba pozyskanych oocytów wynosi 1 lub 2 oraz nieudanych transferów embrionów odbyło się nie więcej niż 3 to ocena wynosi 1. Dla takich samych wartości liczby pozyskanych oocytów i więcej niż trzech nieudanych transferów wartość ustalana jest na 2. Gdy pozyskano powyżej 2 oocytów kryterium przyjmuje wartość 3 dla odpowiednio maksymalnie 3 zakończonych porażką prób transferu embrionów oraz wartość 4 w przeciwnym przypadku.
- endometrium – opisuje chłonność endometrium, wartości zależą od grubości błony mierzonej miedzy 10 a 14 dniem cyklu, ocena 1 odpowiada grubości między 8 a 12 milimetrów, zaś 2 wszystkim pozostałym osiąganym wartościom pomiaru, preferowana jest mniejsza wartość;
- sperm – informacja o jakości spermy uwzględniająca dwa czynniki:
  - pochodzenie spermy: pozyskana z jąder lub otrzymana w wyniku ejakulacji,
  - koncentrację plemników w preparacie.

  Dla spermy pozyskanej z jąder ocena na kryterium to 5, w przeciwnym wypadku przypisywane są oceny od 1 do 4, zgodnie ze schematem:
  - 1 – koncentracja powyżej 15 milionów,
  - 2 – koncentracja między 5 a 15 milionów,
  - 3 – koncentracja między 1 a 5 milionów,
  - 4 – koncentracja poniżej 1 miliona.

  Preferowane są niższe wartości.
- frozen_embryos – opisuje liczbę zamrożonych embrionów, przyjmuje wartości całkowite między 1 a 10 - korzystniejsze są te większe;
- cleavage_stage – informuje o stopniu podziału embrionu i stanowi element jego morfologicznej ewaluacji, na ocenę składają się trzy podkryteria:
  - dzień transferu,
  - liczba blastomerów,
  - stopień fragmentacji.

  W Tabeli 2.2 opisano przyjętą skalę przy czym preferowane są dla niej mniejsze wartości.

Tablica 2.2: Skala na kryterium cleavage_stage w zbiorze drugim

Wartość oceny	Warunek przypisania
1	dzień 3 i 8 blastomerów lub dzień 2 i 4 blastomery, fragmentacja 0%
2	dzień 3 i 8 blastomerów lub dzień 2 i 4 blastomery, fragmentacja poniżej 10%
3	dzień 3 i 8 blastomerów lub dzień 2 i 4 blastomery, fragmentacja między 10 a 20%
4	dzień 3 i od 6 do 12 blastomerów z wyjątkiem 8, fragmentacja 0%
5	dzień 3 i od 6 do 12 blastomerów z wyjątkiem 8, fragmentacja poniżej 10%
6	dzień 3 i od 6 do 12 blastomerów z wyjątkiem 8, fragmentacja między 10 a 20%
7	dzień 3 i mniej niż 6 blastomerów lub dzień 2 i od 2 do 4 blastomerów poza 4, fragmentacja 0%
8	dzień 3 i mniej niż 6 blastomerów lub dzień 2 i od 2 do 4 blastomerów poza 4, fragmentacja poniżej 10%
9	dzień 3 i mniej niż 6 blastomerów lub dzień 2 i od 2 do 4 blastomerów poza 4, fragmentacja między 10 a 20%

Opisane wyżej kryteria można podzielić na dwie grupy zgodnie ze schematem, ocena pary (age, pregnancy, attempts, endometrium, sperm) oraz czynniki embriologiczne (frozen_embryos, cleavage_stage). W Tabeli 2.3 zostało przedstawione skrócone podsumowanie informacji o wyżej opisanych kryteriach.

Tablica 2.3: Skrócona interpretacja kryteriów dla zbioru drugiego

Nazwa kryterium	Opis kryterium	Typ wartości	Kierunek preferencji
age	wiek kobiety	liczba całkowita	koszt
pregnancy	czy parze udało się dotychczas otrzymać ciążę	liczba całkowita 0 lub 1	zysk
attempts	złożenie informacji o dotychczasowej liczbie nieudanych prób transferu embrionów i całkowitej liczbie pozyskanych oocytów	liczba całkowita od 1 do 4	koszt
endometrium	ocena chłonności endometrium	liczba całkowita 1 lub 2	koszt
sperm	ocena jakości spermy	liczba całkowita od 1 do 5	koszt
frozen_embryos	liczba zamrożonych embrionów	liczba całkowita od 1 do 10	zysk
cleavage_stage	morfologiczna ocena embrionu	liczba całkowita od 1 do 9	koszt

Każdy z wariantów może zostać przypisany do jednej z czterech kategorii:

- wartość 1 dla podwójnego transferu blastocysty,
- wartość 2 dla podwójnego transferu embrionu w fazie podziału,
- wartość 3 dla pojedynczego transferu blastocysty,
- wartość 4 dla pojedynczego transferu embrionu w fazie podziału.

Preferowane klasy decyzyjne oznaczane są większymi numerami. Im wartość jest mniejsza, tym wzrasta prawdopodobieństwo wystąpienia ciąży mnogiej.

2.1 Pierwszy zbiór danych

Obserwacje zebrane w zbiorze danych zostały przedstawione w Tabeli 2.4. W zbiorze zanotowanych zostało 51 obserwacji. Na przykład wariant a42 reprezentuje parę, w której kobieta ma 26 lat, okres niepłodności trwa już 5 lat, efektywna liczba oocytów przyjmuje wartość 3 co oznacza, że pozyskano 10 lub 14 bądź więcej oocytów, ogólne zdrowie kobiety ocenione jest na 2, nasienie stanowią krikonserwowane plemniki z biopsji jąder, ocena morfologiczna dla embrionów była dość wysoka, natomiast ocena rozwoju embrionów przyjęła najwyższą możliwą wartość. Dla tej pary proponowane jest zaklasyfikowanie do grupy 3, czyli transferu dwóch embrionów. Wariant a51 opisuje parę, w której kobieta jest w wieku 40 lat, okres niepłodności trwa od 4 lat, efektywna liczba oocytów przyjmuje wartość 5 (pozyskano 7 lub 6 oocytów), ogólnie zdrowie kobiety ocenione zostało na 3 w siedmiopunktowej skali, sperma pochodzi z ejakulacji, pod względem morfologicznym embriony zostały ocenione średnio na 3 punkty, zaś ich rozwój na 4. Sugerowane postępowanie w tym przypadku to transfer trzech embrionów, czyli klasa 2.

Tablica 2.4: Pierwszy zbiór danych

Object	age	infertility	oocytes	woman_eval	sperm	morpho_quality	develop_quality	class
a1	36	1	6	4	3	16	3	2
a2	28	1	2	5	2	20	5	3
a3	38	1	6	1	3	14	4	3
a4	28	6	2	6	3	8	3	3
a5	44	3	2	3	1	11	3	2
a6	42	4	4	3	3	14	5	2
a7	30	1	6	4	2	7	3	3
a8	33	2	4	4	3	17	5	3
a9	36	3	5	2	3	9	2	3
a10	27	2	2	4	3	16	5	3
a11	39	3	4	3	3	5	1	2
a12	36	1	4	3	3	14	2	3
a13	37	2	4	5	3	15	3	3
a14	27	2	3	3	3	18	4	3
a15	30	3	3	2	3	16	5	3
a16	33	4	5	4	3	16	3	3
a17	38	6	3	4	3	13	4	2

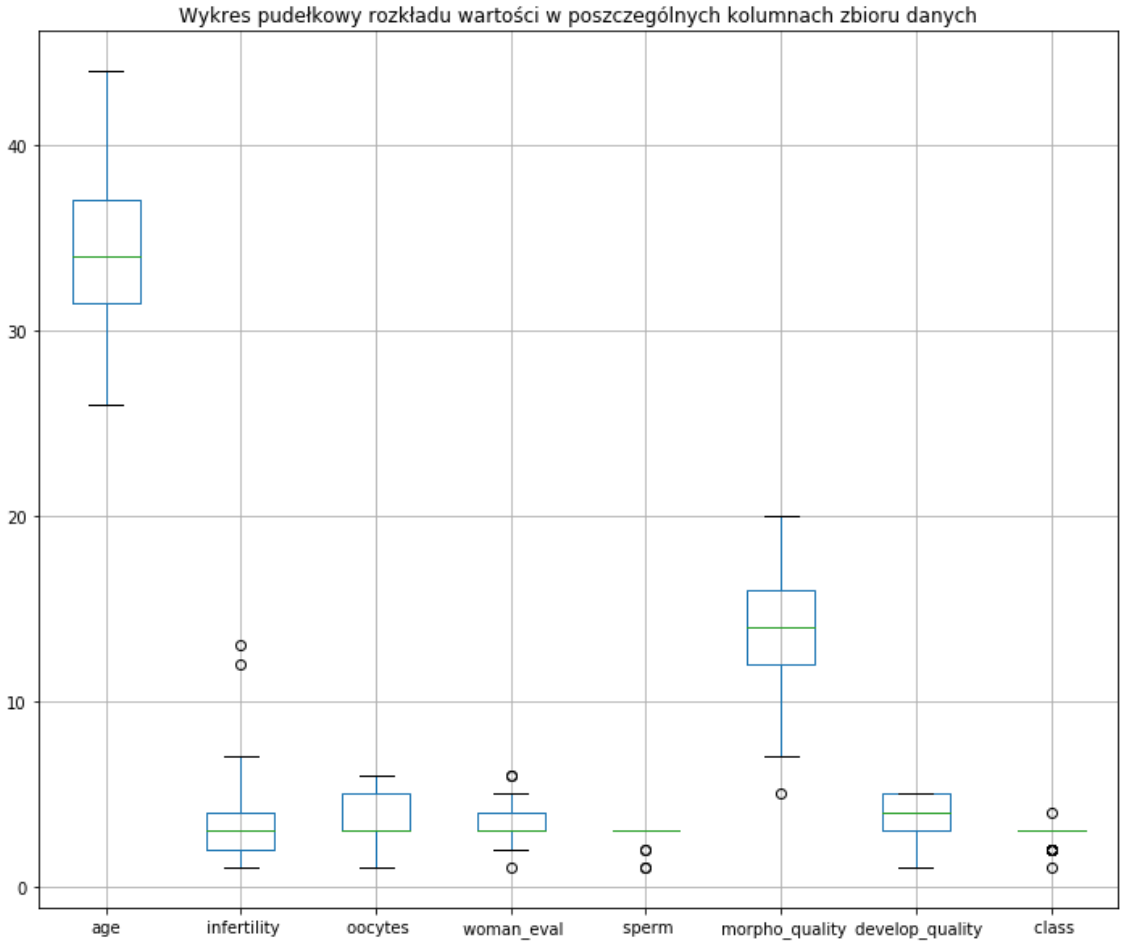
Kontynuacja na następnej stronie

Object	age	infertility	oocytes	woman_eval	sperm	morpho_quality	develop_quality	class
a18	30	2	3	4	3	17	5	3
a19	32	3	1	3	3	14	4	3
a20	30	3	3	3	3	20	5	3
a21	33	12	5	3	3	14	4	3
a22	37	2	6	2	3	14	1	3
a23	40	3	1	2	3	12	3	1
a24	34	3	1	3	3	10	3	3
a25	28	2	3	3	3	12	4	3
a26	33	3	5	3	3	12	2	3
a27	37	6	1	4	3	17	5	3
a28	34	3	3	3	3	15	4	3
a30	33	5	5	4	3	9	4	2
a31	34	13	1	2	3	15	5	3
a32	33	5	4	3	3	14	5	3
a33	39	3	5	2	3	8	2	2
a34	37	3	5	4	3	16	4	3
a35	34	3	3	2	3	14	5	3
a36	38	4	4	5	3	13	5	2
a37	33	3	4	5	3	16	5	3
a38	34	2	4	3	3	16	4	3
a39	35	2	3	5	3	19	5	3
a40	34	3	2	4	3	13	5	3
a41	38	7	6	4	3	10	1	3
a42	26	5	3	2	1	15	5	3
a43	34	2	3	3	3	10	3	3
a44	28	3	4	3	3	14	4	3
a45	30	1	3	6	3	10	1	3
a46	31	2	3	2	3	16	5	3
a47	33	2	3	4	3	15	4	3
a48	41	2	3	3	3	19	5	3
a49	40	2	5	2	3	12	2	3
a50	32	3	1	3	3	16	4	3
a51	40	4	5	3	3	12	4	2

Na Rysunku 2.1 zostały przedstawione, w formie wykresów pudełkowych, rozkłady wartości poszczególnych atrybutów w zbiorze. Analizując te dane, można zauważyć, że największa różnorodność możliwych ocen obserwowana jest na kryteriach *age* oraz *morpho_quality*, natomiast najmniejsza na kryterium *sperm*. Zaobserwowane zależności mogą skutkować mniejszym wykorzystaniem atrybutu *sperm* do budowy poszczególnych modeli. Warto też zauważyć, że w zbiorze obserwacje z poszczególnych klas mają następujące liczności:

- klasa 1: liczność 1,
- klasa 2: liczność 9,
- klasa 3: liczność 40,

- klasa 4: liczność 1.



Rysunek 2.1. Rozkład wartości atrybutów w pierwszym zbiorze danych

W Tabeli 2.5 zebrano dane statystyczne dla zbioru danych. Z ich analizy można wyciągnąć wniosek, iż przeciętna para chcąca skorzystać z metody in vitro ma następującą charakterystykę:

- kobieta ma 34 lata,
- niepłodność trwa 3 lata,
- liczba pozyskanych oocytów wynosi 10 lub 14 bądź więcej,
- kobieta jest oceniona na 3 punkty w siedmiostopniowej skali, na której ocenę przypisuje ginekolog lub położnik,
- sperma pochodzi z ejakulacji,
- morfologiczny rozwój embrionów jest oceniany na średnio 3 lub 4 punkty,
- rozwój embrionów jest oceniany na 4 z pięciu możliwych punktów,
- proponowany jest transfer dwóch embrionów.

Tablica 2.5: Statystyki pierwszego zbioru danych

	Liczba unikatowych wartości	Średnia	Odchylenie standardowe	Mediana
age	17	34.18	4.24	34.0
infertility	9	3.37	2.37	3.0
oocytes	6	3.51	1.46	3.0
woman_eval	6	3.31	1.10	3.0
sperm	3	2.88	0.43	3.0
morpho_quality	15	13.75	3.32	14.0
develop_quality	5	3.76	1.27	4.0
class	4	2.80	0.49	3.0

Z interpretacji powyżej przedstawionych informacji wynika, iż zbiór danych zawiera dość mało obserwacji, a jego analiza może sprawić problemy ze względu na silne niezbalansowanie obserwacji z poszczególnych klas. Najlepsza i najgorsza kategoria reprezentowane są tylko przez pojedyncze obserwacje, co będzie skutkowało niewielką skutecznością modeli dla obserwacji potencjalnie należących do tych klas. Dodatkowo wartości atrybutu sperm są bardzo zbliżone do siebie, a aż 47 obserwacji jest ze względu na niego nierozróżnialnych, co potencjalnie wpływać może na jego niewielką rolę przy budowaniu modeli. Co istotne, problem tego typu nie powinien dotyczyć pozostałych sześciu atrybutów warunkowych.

2.2 Drugi zbiór danych

Obserwacje zebrane w drugim zbiorze danych zostały przedstawione w Tabeli 2.6. Zawiera on 25 obiektów. Na przykład wariant a24 reprezentuje parę, w której kobieta ma 29 lat, nie udało się dotychczas osiągnąć ciąży, liczba pozyskanych oocytów to 1 lub 2 i wykonano maksymalnie 3 nieudane transfery embrionów, grubość endometrium wynosi między 8 a 12 mm, sperma ma koncentrację powyżej 15 milionów plemników, zamrożono 1 embrion oraz w trzecim dniu mamy od 6 do 12 blastomerów, z wyjątkiem 8 a fragmentacja nie przekracza 10%. Dla tej pary proponowany jest podwójny transfer embrionu w fazie podziału, czyli klasa 2. Dla wariantu a25 wiek kobiety określony jest na 24 lata, udało się osiągnąć ciążę, liczba pozyskanych oocytów to 1 lub 2 i wykonano maksymalnie 3 nieudane transfery embrionów, endometrium ma grubość inną niż zawarta w przedziale 8–12 mm, sperma ma koncentrację między 1 a 5 milionów plemników, zamrożono 10 embrionów i w dniu trzecim mamy 8 blastomerów lub w drugim – 4 a fragmentacja nie sięga 10%. Tej parze sugeruje się podwójny transfer blastocysty – klasę 1.

Tablica 2.6: Drugi zbiór danych

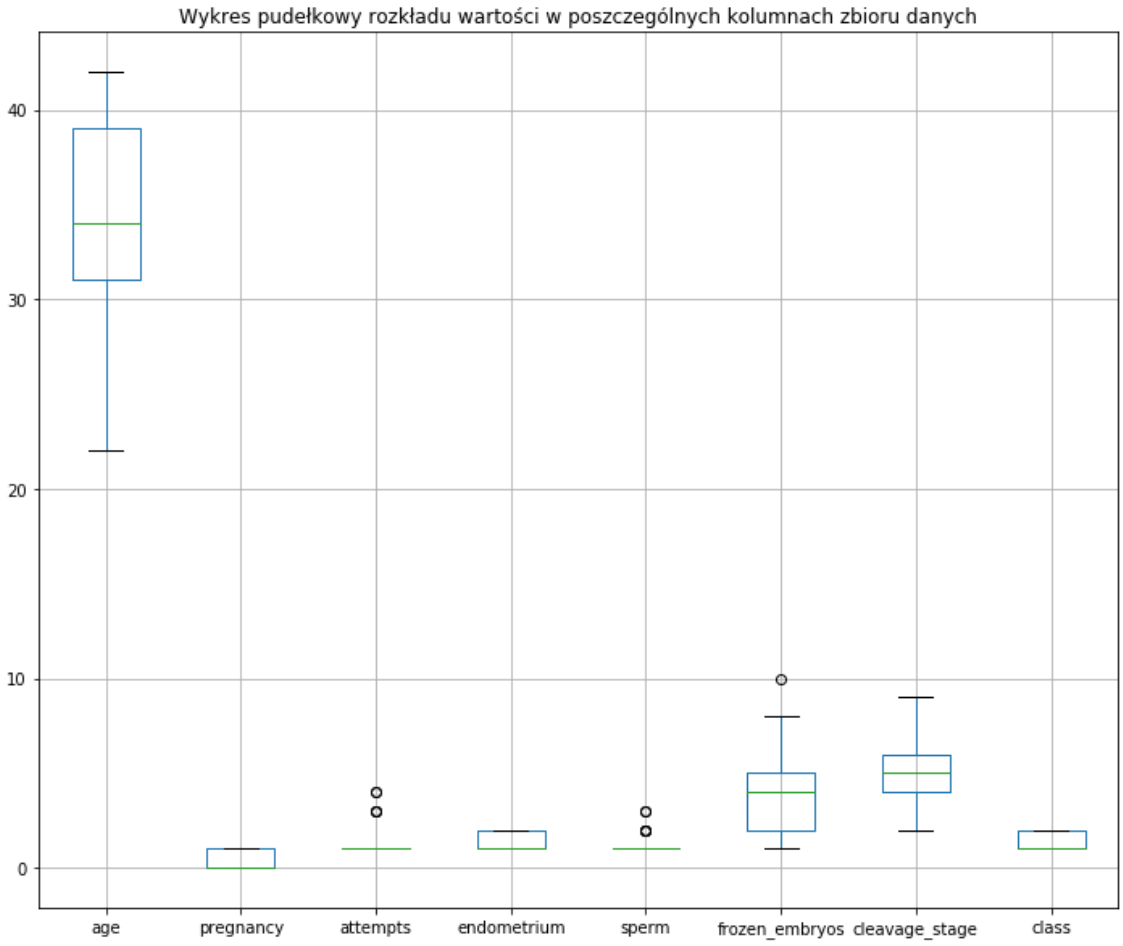
Object	age	pregnancy	attempts	endom.	sperm	frozen_embr.	cleavage_stage	class
a1	23	0	3	1	1	4	5	1
a2	39	0	1	2	1	1	5	2
a3	42	0	3	1	1	2	8	1
a4	37	1	1	1	2	2	5	2
a5	34	1	1	1	2	2	5	2
a6	32	0	1	2	1	8	8	2
a7	27	0	4	2	1	5	5	1
a8	34	1	1	1	1	6	2	1

Kontynuacja na następnej stronie

Object	age	pregnancy	attempts	endom.	sperm	frozen_embr.	cleavage_stage	class
a9	31	0	1	1	1	4	8	2
a10	36	0	1	2	1	2	2	1
a11	22	0	1	1	1	2	9	2
a12	31	1	1	1	2	3	5	2
a13	33	1	3	1	2	5	2	1
a14	42	1	1	2	1	4	4	1
a15	40	0	1	1	1	3	5	1
a16	32	1	1	2	1	8	2	2
a17	36	0	1	2	1	7	2	1
a18	34	1	1	1	1	6	5	1
a19	39	1	1	2	1	4	5	1
a20	40	1	1	1	1	5	6	1
a21	32	0	1	1	1	4	5	1
a22	37	0	1	1	3	4	8	1
a23	41	1	4	2	1	3	9	2
a24	29	0	1	1	1	1	5	2
a25	24	1	1	2	3	10	2	1

Na Rysunku 2.2 zostały przedstawione rozkłady wartości poszczególnych atrybutów. Z analizy wykresów pudełkowych można wywnioskować, że największa różnorodność obserwowana jest na kryteriach *age* i *frozen_embryos*, zaś najmniejsza – *attempts* i *sperm*. Jeśli chodzi o liczności obserwacji z poszczególnych klas w zbiorze danych to:

- dla klasy 1: liczność 15,
- dla klasy 2: liczność 10,
- dla klasy 3: liczność 0,
- dla klasy 4: liczność 0.



Rysunek 2.2. Rozkład wartości atrybutów w drugim zbiorze danych

W Tabeli 2.7 zebrano dane statystyczne dla zbioru danych. Przeciętna para chcąca skorzystać z metody in vitro jest opisana następującymi ocenami:

- kobieta jest w wieku 34 lat,
- dotychczas nie udało się osiągnąć ciąży,
- liczba pozyskanych oocytów to 1 lub 2 i wykonano maksymalnie 3 nieudane transfery embrionów,
- endometrium ma grubość między 8 a 12 mm,
- sperma ma koncentrację powyżej 15 milionów plemników,
- zamrożono 4 embriony,
- w trzecim dniu mamy od 6 do 12 blastomerów, z wyjątkiem 8 a fragmentacja nie przekracza 10%,
- sugerowany jest podwójny transfer blastocysty – klasa 1.

Tablica 2.7: Statystyki drugiego zbioru danych

	Liczba unikatowych wartości	Średnia	Odchylenie standardowe	Mediana
age	15	33.88	5.75	34.0
pregnancy	2	0.48	0.51	0.0
attempts	3	1.48	1.00	1.0
endometrium	2	1.40	0.50	1.0
sperm	3	1.32	0.63	1.0
frozen_embryos	9	4.20	2.31	4.0
cleavage_stage	6	5.08	2.27	5.0
class	2	1.40	0.50	1.0

W ramach interpretacji powyżej zamieszczonych informacji o zbiorze danych należy zwrócić uwagę na niewielką liczbę obserwacji. Stosunek obserwacji z poszczególnych klas ma się jak 2:3, co potencjalnie może czynić ten zbiór nieco łatwiejszym do wykorzystania niż poprzedni, jednak należy też zauważyć, że żadna z obserwacji nie należy do klasy 3 ani 4. Skutkiem tego będzie fakt, iż zbudowany model żadnej obserwacji nie będzie przypisywał do tych klas. Problemem mogą okazać się atrybuty *attempts* i *sperm*, gdyż w dominującej liczbie obserwacji przyjmują one dokładnie tę samą wartość, co, dla wykorzystywanych w dalszej części pracy algorytmów, może czynić je niezbyt użytecznymi i nie wpływającymi na końcową decyzję.

2.3 Obserwacje końcowe

Mediana wieku dla obu zbiorów danych przyjmuje tę samą wartość. W związku z tym można stwierdzić, iż charakterystyka wiekowa obu grup obserwacji jest podobna, a różnicą jest podejście do gromadzonych informacji, choć tu również widoczne są podobieństwa – pojawiają się kryteria odnoszące się do analogicznych informacji, jednak przypisywane oceny mają różne skale lub są agregacją innych danych szczegółowych. Zasadniczą różnicą, na którą należy zwrócić uwagę, jest liczba klas decyzyjnych wykorzystanych w obu problemach — dla pierwszego zbioru są 4, zaś dla drugiego 2. Potencjalnie będzie ona wpływała na dokładność klasyfikacji, która, dodatkowo ze względu na korzystniejszy bilans obserwacji z poszczególnych klas, może okazać się wyższa dla zbioru drugiego.

Rozdział 3

Podstawy teoretyczne

3.1 Metody wielokryterialnego wspomagania decyzji

Metody wielokryterialnego wspomagania decyzji stosowane są do problemów, w których przypisanie wariantu do odpowiedniej klasy decyzyjnej zależne jest od wartości jego ocen uzyskanych na zbiorze kryteriów o ustalonych kierunkach preferencji określanych jako jeden z możliwych dwóch typów:

- koszt – preferowane są niskie wartości ocen,
- zysk – preferowane są wysokie wartości ocen.

W tym kontekście sortowanie narzuca wymóg, aby także klasy, do których przypisywane są warianty, miały określony kierunek preferencji. Taka definicja idealnie pasuje do zagadnienia badanego w pracy, stąd motywacja do wykorzystania metod wielokryterialnego sortowania. Wśród nich wybrano Podejście Zbiorów Przybliżonych oparte na Relacji Dominacji (DRSA) oraz jego rozszerzenie – Podejście Zbiorów Przybliżonych oparte na Relacji Dominacji ze Zmienną Spójnością (VC-DRSA) [9]. Taki wybór podyktowany był faktem, że bazowanie na teorii zbiorów przybliżonych jest użyteczne dla problemów, w których może pojawiać się, wynikająca z granularności ocen na kryteriach, niespójność w danych wejściowych. W badanym problemie jest ona obserwowalna. Dodatkowo, wykorzystanie podstawowego wariantu metody i jej rozszerzenia pozwoli na wyciągnięcie pełniejszych wniosków w ramach analizy. Wybór modelu ze zmienną spójnością powodowany jest jego prostą interpretacją, jak również chęcią przygotowania modelu, który – ze względu na niewielką liczbę obserwacji w zbiorach danych – nie będzie nadmiernie dopasowany do przykładów wykorzystanych w fazie jego budowy. Dzięki temu będzie on potencjalnie miał większą zdolność uogólniania wiedzy i wykorzystania jej do klasyfikacji nowych obserwacji.

Aplikacja metod wielokryterialnego sortowania pociąga za sobą możliwość użycia różnych algorytmów indukcji reguł decyzyjnych [17]. W pracy przetestowane zostaną dwie procedury, z których pierwsza pozwala na wygenerowania minimalnego zbioru reguł decyzyjnych, zaś druga – satysfakcjonującego. Wejściem dla obu algorytmów będą wyniki otrzymane po zastosowaniu do oryginalnych danych metod DRSA i VC-DRSA.

W celu przetestowania skuteczności poszczególnych modeli zostaną wykorzystane również dwa odmienne podejścia do klasyfikacji obiektów przy wykorzystaniu reguł decyzyjnych. Jedno z nich to standardowe, bardzo proste podejście, zaś drugie stanowi przykład bardziej wyrafinowanej metody [2].

3.1.1 Podejście Zbiorów Przybliżonych oparte na Relacji Dominacji

Podstawowym pojęciem, kluczowym do zrozumienia działania metody DRSA, jest dominacja. Mówimy, że wariant A dominuje wariant B, jeśli na każdym kryterium przyjmuje on oceny takie same bądź lepsze niż B. Warto tu zauważyć, że dla kryterium typu zysk lepszą oceną jest ta o wartości większej, natomiast dla kryterium typu koszt – o wartości mniejszej. Oznaczając zbiór kryteriów przez F oraz ocenę wariantu A na kryterium $j \in F$ przez $g_j(A)$, relacja dominacji (D_P) zachodzi, gdy dla $\forall_{j \in F}$ zachodzi:

- dla j będącego kryterium typu zysk: $g_j(A) \geq g_j(B)$,
- dla j będącego kryterium typu koszt: $g_j(A) \leq g_j(B)$.

O dominacji możemy mówić w sensie silnym, wówczas wymagane jest, aby przynajmniej na jednym kryterium ocena wariantu dominującego była lepsza niż zdominowanego oraz, w sensie słabym, gdzie dla zajścia relacji wystarczającym jest równość wartości ocen obu porównywanych wariantów na wszystkich kryteriach.

W podejściu zbiorów przybliżonych opartym na relacji dominacji korzystamy z unii klas, które są grupami klas decyzyjnych. Wyróżniamy:

- unie klas w górę $Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s$ dla $t = 2, ..., m$ – grupują klasy co najmniej tak dobre jak dana klasa,
- unie klas w dół $Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s$ dla $t = 1, ..., m-1$ – grupują klasy co najwyżej tak dobre jak dana klasa.

Przydatne jest także pojęcie stożka dominacji definiowane w dwóch wersjach:

- pozytywny stożek dominacji $D_P^+ = \{y \in U : yD_Px\}$, gdzie U jest zbiorem wszystkich wariantów – zbiór wariantów dominujących dany wariant,
- negatywny stożek dominacji $D_P^- = \{y \in U : xD_Py\}$, gdzie U jest zbiorem wszystkich wariantów – zbiór wariantów zdominowanych przez dany wariant.

Główna idea metody polega na budowaniu przybliżeń unii klas. Wyróżniamy:

- przybliżenie unii klas w górę, które powstaje poprzez analizę pozytywnych stożków dominacji:
  - dolne przybliżenie unii klas w górę $\underline{P}(Cl_t^{\geq}) = \{x \in U : D_P^+(x) \subseteq Cl_t^{\geq}\}$,
  - górne przybliżenie unii klas w górę $\overline{P}(Cl_t^{\geq}) = \bigcup_{x \in Cl_t^{\geq}} D_P^+(x)$,
  - brzeg unii klas w górę $Bn_P(Cl_t^{\geq}) = \overline{P}^{(}Cl_t^{\geq}) - \underline{P}^{(}Cl_t^{\geq})$,
- przybliżenie unii klas w dół, które powstaje poprzez analizę negatywnych stożków dominacji:
  - dolne przybliżenie unii klas w dół $\underline{P}(Cl_t^{\leq}) = \{x \in U : D_P^-(x) \subseteq Cl_t^{\leq}\}$,
  - górne przybliżenie unii klas w dół $\overline{P}(Cl_t^{\leq}) = \bigcup_{x \in Cl_t^{\leq}} D_P^-(x)$,
  - brzeg unii klas w dół $Bn_P(Cl_t^{\leq}) = \overline{P}^{(}Cl_t^{\leq}) - \underline{P}^{(}Cl_t^{\leq})$.

Otrzymane przybliżenia służą następnie jako dane wejściowe dla algorytmów indukcji reguł decyzyjnych, które opisane zostaną w dalszej części pracy.

Przykład z pierwszego zbioru danych

Aby zilustrować działanie metody przedstawiony zostanie przykład jej zastosowania dla podzbioru, w którym za obserwacje służące do budowy modelu wzięto wszystkie z wyjątkiem a3. Powstały wówczas unie klas w dół zgromadzone w Tabeli 3.1 oraz unie klas w górę – Tabela 3.2. Przykładowo, do unii *co najwyżej kasa 2* należą wszystkie obiekty, które w kolumnie *class* z Tabeli 2.4 mają przypisaną wartość 1 lub 2, zaś do unii *co najmniej kasa 2* przypisane są obiekty z klas 2, 3 i 4.

Tablica 3.1: Unie klas w dół dla metody DRSA i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najwyżej klasa 1	[a23]
co najwyżej klasa 2	[a1, a5, a6, a11, a17, a23, a30, a33, a36, a51]
co najwyżej klasa 3	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]

Tablica 3.2: Unie klas w górę dla metody DRSA i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najmniej klasa 2	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]
co najmniej klasa 3	[a2, a4, a7, a8, a9, a10, a12, a13, a14, a15, a16, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a31, a32, a34, a35, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50]
co najmniej klasa 4	[a28]

Pozytywne i negatywne stożki dominacji dla wariantów zostały przedstawione odpowiednio w Tabelach 3.3 i 3.4. Na przykład, dla obiektu a1 nie istnieje inny wariant mający na wszystkich kryteriach oceny przynajmniej tak samo dobre jak a1, dlatego w jego pozytywnym stożku dominacji ujęty jest tylko on sam. Natomiast w jego negatywnym stożku dominacji, oprócz niego samego, są także obiekty a22 i a41, które, zgodnie z definicją relacji dominacji, mają oceny takie same lub gorsze na całym zbiorze kryteriów.

Tablica 3.3: Pozytywne stożki dominacji dla metody DRSA i zbioru pierwszego bez wariantu a3

Obiekt	Obiekty dominujące dany obiekt
a1	[a1]
a2	[a2]
a4	[a4]
a5	[a2, a5, a10, a19, a40, a50]
a6	[a6, a8, a10, a18, a20, a37, a39, a48]
a7	[a2, a7]
a8	[a8, a18]
a9	[a8, a9, a10, a12, a14, a15, a18, a19, a20, a24, a25, a26, a28, a35, a37, a38, a39, a40, a43, a44, a46, a47, a50]
a10	[a10]
a11	[a8, a10, a11, a12, a13, a14, a18, a19, a20, a24, a25, a37, a38, a39, a40, a43, a44, a45, a47, a50]
a12	[a12]
a13	[a13, a39]
a14	[a14]
a15	[a10, a15, a18, a20]
a16	[a8, a10, a16, a18, a37]
a17	[a10, a17, a18, a27, a39, a40, a47]
a18	[a18]
a19	[a19, a50]
a20	[a20]
a21	[a8, a10, a14, a18, a19, a20, a21, a32, a37, a44, a47, a50]
a22	[a1, a8, a10, a12, a13, a14, a18, a22, a38, a39, a46, a47]
a23	[a19, a23, a50]
a24	[a19, a24, a50]
a25	[a10, a14, a25]
a26	[a8, a10, a14, a18, a19, a20, a25, a26, a37, a44, a47, a50]
a27	[a27]
a28	[a10, a15, a18, a20, a28, a35, a40, a46]
a29	[a10, a14, a18, a20, a29, a39, a47, a50]
a30	[a8, a10, a18, a30, a37, a47]
a31	[a31]
a32	[a8, a10, a18, a20, a32, a37]
a33	[a8, a9, a10, a12, a13, a14, a15, a18, a19, a20, a24, a25, a26, a28, a33, a34, a35, a37, a38, a39, a40, a43, a44, a46, a47, a50]
a34	[a8, a10, a18, a34, a37, a39]
a35	[a10, a15, a18, a20, a35, a46]
a36	[a36, a37, a39]
a37	[a37]
a38	[a8, a10, a14, a18, a38]
a39	[a39]
a40	[a10, a40]
a41	[a1, a8, a10, a13, a16, a17, a18, a27, a34, a36, a37, a39, a40, a41, a45, a47]
a42	[a42]
a43	[a10, a14, a18, a25, a43, a47]
a44	[a10, a14, a44]
a45	[a45]
a46	[a10, a18, a46]
a47	[a10, a18, a47]
a48	[a39, a48]
a49	[a8, a10, a12, a13, a14, a18, a25, a38, a39, a46, a47, a49]
a50	[a50]
a51	[a8, a10, a14, a18, a19, a20, a25, a34, a36, a37, a38, a39, a40, a44, a47, a50, a51]

Tablica 3.4: Negatywne stożki dominacji dla metody DRSA i zbioru pierwszego bez wariantu a3

Obiekt	Obiekty zdominowane przez dany obiekt
a1	[a1, a22, a41]
a2	[a2, a5, a7]
a4	[a4]
a5	[a5]
a6	[a6]
a7	[a7]
a8	[a6, a8, a9, a11, a16, a21, a22, a26, a30, a32, a33, a34, a38, a41, a49, a51]
a9	[a9, a33]
a10	[a5, a6, a9, a10, a11, a15, a16, a17, a21, a22, a25, a26, a28, a29, a30, a32, a33, a34, a35, a38, a40, a41, a43, a44, a46, a47, a49, a51]
a11	[a11]
a12	[a9, a11, a12, a22, a33, a49]
a13	[a11, a13, a22, a33, a41, a49]
a14	[a9, a11, a14, a21, a22, a25, a26, a29, a33, a38, a43, a44, a49, a51]
a15	[a9, a15, a28, a33, a35]
a16	[a16, a41]
a17	[a17, a41]
a18	[a6, a8, a9, a11, a15, a16, a17, a18, a21, a22, a26, a28, a29, a30, a32, a33, a34, a35, a38, a41, a43, a46, a47, a49, a51]
a19	[a5, a9, a11, a19, a21, a23, a24, a26, a33, a51]
a20	[a6, a9, a11, a15, a20, a21, a26, a28, a29, a32, a33, a35, a51]
a21	[a21]
a22	[a22]
a23	[a23]
a24	[a9, a11, a24, a33]
a25	[a9, a11, a25, a26, a33, a43, a49, a51]
a26	[a9, a26, a33]
a27	[a17, a27, a41]
a28	[a9, a28, a33]
a29	[a29]
a30	[a30]
a31	[a31]
a32	[a21, a32]
a33	[a33]
a34	[a33, a34, a41, a51]
a35	[a9, a28, a33, a35]
a36	[a36, a41, a51]
a37	[a6, a9, a11, a16, a21, a26, a30, a32, a33, a34, a36, a37, a41, a51]
a38	[a9, a11, a22, a33, a38, a49, a51]
a39	[a6, a9, a11, a13, a17, a22, a29, a33, a34, a36, a39, a41, a48, a49, a51]
a40	[a5, a9, a11, a17, a28, a33, a40, a41, a51]
a41	[a41]
a42	[a42]
a43	[a9, a11, a33, a43]
a44	[a9, a11, a21, a26, a33, a44, a51]
a45	[a11, a41, a45]
a46	[a9, a22, a28, a33, a35, a46, a49]
a47	[a9, a11, a17, a21, a22, a26, a29, a30, a33, a41, a43, a47, a49, a51]
a48	[a6, a48]
a49	[a49]
a50	[a5, a9, a11, a19, a21, a23, a24, a26, a29, a33, a50, a51]
a51	[a51]

Otrzymane poprzez analizę stożków dominacji przybliżenia poszczególnych unii klas zostały przedstawione w Tabelach 3.5, 3.6, 3.7, 3.8, 3.9 oraz 3.10. Dla przykładu, w dolnym przybliżeniu unii *co najmniej klasa 4* nie mamy żadnych obiektów, ponieważ nie mamy wariantów, które są zdominowane tylko przez warianty z unii *co najmniej klasa 4*, potencjalny kandydat (mający przypisaną klasę 4 jako jedyny w zbiorze) – a28 – jest zdominowany przez a9, a28 i a33, odpowiednio z klas 3, 4 i 2. Natomiast w górnym przybliżeniu unii *co najmniej klasa 4* pojawia się już kilka obiektów: a10, a15, a18, a20, a28, a35, a40, a46 – są to te obiekty, które znajdują się w pozytywnym stożku dominacji wariantu a28. Jeśli chodzi o brzeg klasy 4 to pokrywa się on z jej górnym przybliżeniem, ponieważ w dolnym przybliżeniu nie mamy żadnych wariantów. Przygotowane przybliżenia to:

- dolne przybliżenie unii klas w górę przedstawione w Tablicy 3.5,

Tablica 3.5: Dolne przybliżenie unii klas w górę dla metody DRSA i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najmniej klasa 2	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]
co najmniej klasa 3	[a2, a4, a7, a8, a9, a10, a12, a13, a14, a15, a16, a18, a19, a20, a21, a24, a25, a26, a27, a28, a29, a31, a32, a34, a35, a37, a38, a39, a40, a42, a43, a44, a45, a46, a47, a48, a49, a50]
co najmniej klasa 4	[]

- górne przybliżenie unii klas w górę zostało przedstawione w Tablicy 3.6,

Tablica 3.6: Górne przybliżenie unii klas w górę dla metody DRSA i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najmniej klasa 2	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]
co najmniej klasa 3	[a1, a2, a4, a7, a8, a9, a10, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a31, a32, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50]
co najmniej klasa 4	[a10, a15, a18, a20, a28, a35, a40, a46]

- brzeg unii klas w górę został przedstawiony w Tablicy 3.7,

Tablica 3.7: Brzeg unii klas w górę dla metody DRSA i zbioru pierwszego bez wariantu a3

Klasa	Obiekty
2	[]
3	[a1, a17, a22, a36, a41]
4	[a10, a15, a18, a20, a28, a35, a40, a46]

- dolne przybliżenie unii klas w dół zostało przedstawione w Tablicy 3.8,

Tablica 3.8: Dolne przybliżenie unii klas w dół dla metody DRSA i zbioru pierwszego bez wariantu a3

,

Unia	Obiekty
co najwyżej klasa 1	[a23]
co najwyżej klasa 2	[a5, a6, a11, a23, a30, a33, a51]
co najwyżej klasa 3	[a1, a2, a4, a5, a6, a7, a8, a9, a11, a12, a13, a14, a16, a17, a19, a21, a22, a23, a24, a25, a26, a27, a29, a30, a31, a32, a33, a34, a36, a37, a38, a39, a41, a42, a43, a44, a45, a47, a48, a49, a50, a51]

- górne przybliżenie unii klas w dół zostało przedstawione w Tablicy 3.9,

Tablica 3.9: Górne przybliżenie unii klas w dół dla metody DRSA i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najwyżej klasa 1	[a23]
co najwyżej klasa 2	[a1, a5, a6, a11, a17, a22, a23, a30, a33, a36, a41, a51]
co najwyżej klasa 3	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]

- brzeg unii klas w dół został przedstawiony w Tablicy 3.10

Tablica 3.10: Brzeg unii klas w dół dla metody DRSA i zbioru pierwszego bez wariantu a3

Klasa	Obiekty
1	[]
2	[a1, a17, a22, a36, a41]
3	[a10, a15, a18, a20, a28, a35, a40, a46]

3.1.2 Podejście Zbiorów Przybliżonych oparte na Relacji Dominacji ze Zmienną Spójnością

Podejście Zbiorów Przybliżonych oparte na Relacji Dominacji ze Zmienną Spójnością jest wariantem metody DRSA, w którym, w stosunku do oryginalnej metody, w inny sposób wyznaczane są przybliżenia unii klas. Korzystamy tu ze stopnia spójności – l, który przyjmuje wartości z przedziału $(0, 1>$. Określa on, w jakim stopniu dopuszczane są niespójności w budowie poszczególnych przybliżeń, co wpływa na możliwość zbudowania modelu bardziej odpornego niż klasyczny na szum w danych wejściowych. Poszczególne przybliżenia w metodzie VC-DRSA wyznaczane są w oparciu o pozytywne i negatywne stożki dominacji, jednak tu ich definicje są następujące:

- przybliżenie unii klas w górę, które powstaje poprzez analizę pozytywnych stożków dominacji przy uwzględnieniu stopnia spójności l:
  - dolne przybliżenie unii klas w górę $\underline{P}^l(Cl_t^{\geq}) = \{x \in Cl_t^{\geq} : \frac{|D_P^+(x) \cap Cl_t^{\geq}|}{|D_P^+(x)|} \geq l\}$,
  - górne przybliżenie unii klas w górę $\overline{P}^l(Cl_t^{\geq}) = U - \underline{P}^l(Cl_{t-1}^{\leq})$,
  - brzeg unii klas w górę $Bn_P(Cl_t^{\geq}) = \overline{P}^l(Cl_t^{\geq}) - \underline{P}^l(Cl_t^{\geq})$;

- przybliżenie unii klas w dół, które powstaje poprzez analizę negatywnych stożków dominacji przy uwzględnieniu stopnia spójności l:
  - dolne przybliżenie unii klas w dół $\underline{P}^l(Cl_t^{\leq}) = \{x \in Cl_t^{\leq} : \frac{|D_P^-(x) \cap Cl_t^{\leq}|}{|D_P^-(x)|} \geq l\}$,
  - górne przybliżenie unii klas w dół $\overline{P}(Cl_t^{\leq}) = U - \underline{P}^l(Cl_{t+1}^{\leq})$,
  - brzeg unii klas w dół $Bn_P(Cl_t^{\leq}) = \overline{P}^l(Cl_t^{\leq}) - \underline{P}^l(Cl_t^{\leq})$.

W podanych wyżej definicjach U jest zbiorem wszystkich wariantów.

Przykład z pierwszego zbioru danych

Aby zilustrować działanie metody zostanie przedstawiony przykład jej zastosowania dla zbioru bez uwzględnienia wariantu a3, który będzie obiektem testowym wykorzystanym w dalszej części pracy (analogicznie jak dla DRSA). Przyjmijmy $l = 0.8$.

Zgodnie z opisem metody, zarówno unie klas jak i stożki dominacji pozostaną dokładnie takie jak dla opisu przykładu dotyczącego DRSA. Różnice pojawią się dla przybliżeń unii klas. Otrzymane przybliżenia przedstawione są w Tabelach 3.11, 3.12, 3.13, 3.14, 3.15 oraz 3.16. Dla przykładu, w dolnym przybliżeniu unii *co najmniej klasa 4* nie mamy żadnych obiektów. Potencjalny kandydat (mający przypisaną klasę 4 jako jedyny w zbiorze) – a28 – jest zdominowany przez a9, a28 i a33, odpowiednio z klas 3, 4 i 2, czyli dla niego:

$$\frac{|D_P^-(x) \cap Cl_t^{\leq}|}{|D_P^-(x)|} = \frac{|\{a9, a28, a33\} \cap \{a28\}|}{|\{a9, a28, a33\}|} = \frac{1}{3} \leq 0.8 = l$$

Natomiast w górnym przybliżeniu unii *co najmniej klasa 4* pojawiają się już obiekty: a28, a35 – są to te obiekty, które nie występują w dolnym przybliżeniu unii *co najwyżej klasa 3*. Jeśli chodzi o brzeg klasy 4 to pokrywa się on z jej górnym przybliżeniem, ponieważ w dolnym przybliżeniu nie mamy żadnych wariantów. Przygotowane z wykorzystaniem metody VC-DRSA przybliżenia unii klas to:

- dolne przybliżenie unii klas w górę przedstawione w Tablicy 3.11,

Tablica 3.11: Dolne przybliżenie unii klas w górę dla metody VC-DRSA, $l = 0.8$ i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najmniej klasa 2	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]
co najmniej klasa 3	[a2, a4, a7, a8, a9, a10, a12, a13, a14, a15, a16, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a31, a32, a34, a35, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50]
co najmniej klasa 4	[]

- górne przybliżenie unii klas w górę przedstawione w Tablicy 3.12,

Tablica 3.12: Górne przybliżenie unii klas w górę dla metody VC-DRSA, $l = 0.8$ i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najmniej klasa 2	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]
co najmniej klasa 3	[a1, a2, a4, a7, a8, a9, a10, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a31, a32, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50]
co najmniej klasa 4	[a28, a35]

- brzeg unii klas w górę przedstawiony w Tablicy 3.13,

Tablica 3.13: Brzeg unii klas w górę dla metody VC-DRSA, $l = 0.8$ i zbioru pierwszego bez wariantu a3

Klasa	Obiekty
2	[]
3	[a1, a17, a36]
4	[a28, a35]

- dolne przybliżenie unii klas w dół przedstawione w Tablicy 3.14,

Tablica 3.14: Dolne przybliżenie unii klas w dół dla metody VC-DRSA, $l = 0.8$ i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najwyżej klasa 1	[a23]
co najwyżej klasa 2	[a5, a6, a11, a23, a30, a33, a51]
co najwyżej klasa 3	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a29, a30, a31, a32, a33, a34, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]

- górne przybliżenie unii klas w dół przedstawione w Tablicy 3.15,

Tablica 3.15: Górne przybliżenie unii klas w dół dla metody VC-DRSA, $l = 0.8$ i zbioru pierwszego bez wariantu a3

Unia	Obiekty
co najwyżej klasa 1	[a23]
co najwyżej klasa 2	[a1, a5, a6, a11, a17, a23, a30, a33, a36, a51]
co najwyżej klasa 3	[a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51]

- brzeg unii klas w dół przedstawiony w Tablicy 3.16.

TABLICA 3.16: Brzeg unii klas w dół dla metody VC-DRSA, $l = 0.8$ i zbioru pierwszego bez wariantu a3

Klasa	Obiekty
1	[]
2	[a1, a17, a36]
3	[a28, a35]

3.1.3 Algorytmy indukcji reguł decyzyjnych

Z wyznaczonych przybliżeń unii klas możemy indukować zbiory reguł decyzyjnych. Wyróżniamy pięć następujących typów reguł [17]:

- typu 1 – indukowane z dolnych przybliżeń unii klas w górę,
- typu 2 – indukowane z górnych przybliżeń unii klas w górę,
- typu 3 – indukowane z dolnych przybliżeń unii klas w dół,
- typu 4 – indukowane z górnych przybliżeń unii klas w dół,
- typu 5 – indukowane z brzegów brzegów.

Reguły typu 1 i 3 określamy mianem pewnych, typu 2 i 4 – przybliżonych, a reguły typu 5 to reguły możliwe. W pracy zostaną wykorzystane tylko reguły pewne.

Algorytmy indukcji reguł mogą wyznaczać minimalne, satysfakcjonujące lub pełne zbiory. O minimalności mówimy, gdy każdy obiekt jest pokrywany przez jak najmniejszą liczbę reguł – wówczas reguły pokrywające obiekty już opisane innymi formułami są pomijane w takich zbiorach. Pojęcie satysfakcjonującego zbioru reguł określa taką grupę reguł, w której każda reguła spełnia postawione wymagania dotyczące na przykład minimalnego wsparcia – będącego liczbą obiektów pokrytych przez regułę – czy długości części warunkowej reguły (na przykład interesują nas tylko reguły zawierające w części warunkowej koniunkcję maksymalnie trzech wyrażeń). Natomiast pełny zbiór reguł nie narzuca na wyrażenia żadnego z opisanych warunków.

Algorytm DOMLEM

Algorytm DOMLEM pozwala na wyznaczenie minimalnego zbioru reguł decyzyjnych. Jego pseudokod w wersji DRSA, dla indukcji reguł typu 1 i przy założeniu, że kryteria są typu zysk, przedstawiono w Algorytmie 1 [17][s.95-96].

Algorytm DOMLEM uruchamiamy niezależnie w celu otrzymania reguł różnych typów – modyfikacji w stosunku do przedstawionej w Algorytmie 1 wersji ulegają, poza przybliżeniami na wejściu, także znaki w warunkach elementarnych. Kolejność iteracji zakłada rozpoczynanie budowania zbioru reguł od najsilniejszych unii klas, czyli tych, w skład których wchodzą obiekty z najmniejszej liczby klas (na przykład dla reguł typu 3 rozpoczynamy od unii co najwyżej klasa 1). Budowanie reguł polega na rozpatrywaniu wszystkich kandydatów na warunki elementarne i wyborze najlepszego z nich. Ocena realizowana może być przy wykorzystaniu różnych miar. W pracy są to maksymalizowane odpowiednio:

- stosunek liczby przykładów pokrywanych przez regułę i należących do aktualnie rozważanego przybliżenia do ogólnej liczby przykładów pokrywanych przez regułę,

- liczba przykładów pokrywanych przez regułę i należących do aktualnie rozważanego.

Jeśli dla porównywanych reguł pierwsza wartość jest taka sama – porównujemy drugą. Gdy tu również zachodzi równość za lepszą uznawana jest pierwsza z reguł.

Warunek akceptacji reguły zakłada sprawdzenie, czy pokryte są wyłącznie obiekty z danego przybliżenia oraz czy reguła ma niepustą część warunkową.

Algorytm DOMLEM może służyć także indukcji reguł przy wykorzystaniu metody VC-DRSA, wówczas modyfikacji ulega warunek akceptacji reguły – sprawdzamy czy dokładność reguły (stosunek liczby pokrytych przykładów z danego przybliżenia i liczby wszystkich pokrytych przykładów) jest większa bądź równa niż przyjęta wartość stopnia spójności.

Algorithm 1 Psudokod algorytmu DOMLEM dla DRSA

procedure DOMLEM(L_{upp})
  $R_{\geq} \leftarrow \varnothing$
  for każdego przybliżenia $K \in L_{upp}$ **do**
    $E \leftarrow FIND\_RULES(K, F)$ ▷ F - zbiór kryteriów
    for każdej reguły r inE **do**
      if r jest minimalna **then**
        $R_{\geq} \leftarrow R_{\geq} \cup r$
      end if
    end for
  end for
end procedure

procedure FIND_RULES(K, F)
  $G \leftarrow K$ ▷ zbiór obiektów nie pokryty dotychczas poprzez elementy z **P**
  P $\leftarrow \varnothing$ ▷ lokalne pokrycie zbioru k
  while $G \neq \varnothing$ **do**
    $P \leftarrow \varnothing$ ▷ kandydat na część warunkową reguły
    $S \leftarrow G$ ▷ zbiór obiektów pokrywany przez P
    while $P = \varnothing$ or not $([P] \subseteq K)$ **do**
      $w \leftarrow \varnothing$ ▷ najlepszy kandydat na warunek elementarny
      for każdego kryterium $a_i \in F$ **do**
        $Cond \leftarrow \{(f(a_i, x) \geq r_{ai}) : \exists x \in S$ taki, że $(f(a_i, x) = r_{ai})\}$ ▷ dla każdego przykładu pozytywnego z S rozważ utworzenie warunku elementarnego
        for każdy warunek elementarny $new\_p \in Cond$ **do**
          if $evaluate(P \cup new\_p, P \cup w)$ **then** ▷ sprawdź, czy warunek new_p jest lepszy niż w
            $w \leftarrow new\_p$
            $w\_eval \leftarrow new\_eval\_p$
          end if
        end for
      end for
      $P \leftarrow P \cup \{w\}$ ▷ dołącz najlepszy warunek do koniunkcji P
      $S \leftarrow S \cap [w]$ ▷ ogranicz zbiór obiektów
    end while
    for każdy warunek elementarny $w \in P$ **do**
      if $[P - \{w\}] \subseteq K$ **then**
        $P \leftarrow P - \{w\}$
      end if
      P $\leftarrow$ **P** $\cup \{P\}$
      $G \leftarrow K - \cup_{P \in \mathbf{P}}[P]$
    end for
  end while
  utwórz reguły **E** na podstawie lokalnego pokrycia **P**
end procedure

Przykład z pierwszego zbioru danych

Kontynuując rozpoczęty przykład dla DRSA, z wykorzystaniem DOMLEM otrzymujemy reguły zebrane w Tabeli 3.17. Na przykład, *Reguła 1* pokrywa tylko obiekt a23 – jedyny w przybliżeniu unii *co najwyżej klasa 1.*

Tablica 3.17: Reguły dla algorytmu DOMLEM, metody DRSA i pierwszego zbioru danych bez wariantu a3

Oznaczenie	Reguła
Reguła 1	$(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \wedge (infertility \geq 3.0) \implies (class \leq 1)$
Reguła 2	$(age \geq 39.0) \wedge (infertility \geq 3.0) \implies (class \leq 2)$
Reguła 3	$(morpho\_quality \leq 9.0) \wedge (age \geq 33.0) \wedge (infertility \geq 5.0) \implies (class \leq 2)$
Reguła 4	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 5	$(infertility \geq 4.0) \implies (class \leq 3)$
Reguła 6	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 7	$(age \geq 35.0) \implies (class \leq 3)$
Reguła 8	$(sperm \leq 2.0) \implies (class \leq 3)$
Reguła 9	$(age \leq 36.0) \wedge (oocytes \leq 4.0) \implies (class \geq 3)$
Reguła 10	$(morpho\_quality \geq 15.0) \wedge (oocytes \leq 5.0) \implies (class \geq 3)$
Reguła 11	$(age \leq 36.0) \wedge (infertility \leq 3.0) \wedge (oocytes \leq 5.0) \implies (class \geq 3)$
Reguła 12	$(infertility \leq 2.0) \wedge (oocytes \leq 5.0) \implies (class \geq 3)$
Reguła 13	$(age \leq 34.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 14	$(age \leq 32.0) \implies (class \geq 3)$
Reguła 15	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 16	$(woman\_eval \geq 3.0) \implies (class \geq 2)$
Reguła 17	$(infertility \leq 2.0) \implies (class \geq 2)$

Zgodnie z algorytmem w pierwszej kolejności zbudowano kandydatów na część warunkową reguły zebranych w Tabeli 3.18. Analizując oceny w niej ujęte, najlepszym kandydatem jest ten o numerze 1, jednak pokrywa on, oprócz obiektu a23, który powinien być pokryty, także pięć innych obiektów. W związku z tym reguła musi zostać rozszerzona, a zbiór kandydatów stanowią warunki ujęte w Tablicy 3.19. Na ich podstawie najlepszym kandydatem jest więc 3, czyli otrzymujemy:

$$(age \geq 40.0) \wedge (woman\_eval \leq 2.0)$$

Jednak nadal pokryte są obiekty nienależące do przybliżenia unii *co najwyżej klasa 1*, więc w kolejnym kroku znów rozszerzamy regułę, rozpatrując kandydatów przedstawionych w Tabeli 3.20. Tym razem najlepszy kandydat to 1. Identyczne oceny ma 5., jednak w tej sytuacji bierzemy pierwszego. Ostatecznie, dodanie tego warunku do dotychczas zbudowanej koniunkcji powoduje

otrzymanie reguły, którą można zaakceptować, ponieważ pokrywa tylko obiekty z rozpatrywanego przybliżenia. Ponadto jest to jedyna reguła dla tego przybliżenia unii i jest ona minimalna.

Tablica 3.18: Kandydaci na część warunkową reguły dla algorytmu DOMLEM, metody DRSA i pierwszego zbioru danych bez wariantu a3

Numer	Kandydat	Ocena kandydata na kryterium 1	Ocena kandydata na kryterium 2
1	$(age \geq 40.0)$	$\frac{1}{6}$	1
2	$(infertility \geq 3.0)$	$\frac{1}{31}$	1
3	$(oocytes \geq 1.0)$	$\frac{1}{51}$	1
4	$(woman\_eval \leq 2.0)$	$\frac{1}{12}$	1
5	$(sperm \leq 3.0)$	$\frac{1}{51}$	1
6	$(morpho\_quality \leq 12.0)$	$\frac{1}{16}$	1
7	$(develop\_quality \leq 3.0)$	$\frac{1}{18}$	1

Tablica 3.19: Kandydaci na rozszerzenie części warunkowej reguły dla algorytmu DOMLEM, metody DRSA i pierwszego zbioru danych bez wariantu a3

Numer	Kandydat	Ocena kandydata na kryterium 1	Ocena kandydata na kryterium 2
1	$(infertility \geq 3.0)$	$\frac{1}{4}$	1
2	$(oocytes \geq 1.0)$	$\frac{1}{6}$	1
3	$(woman\_eval \leq 2.0)$	$\frac{1}{2}$	1
4	$(sperm \leq 3.0)$	$\frac{1}{5}$	1
5	$(morpho\_quality \leq 12.0)$	$\frac{1}{4}$	1
6	$(develop\_quality \leq 3.0)$	$\frac{1}{3}$	1

Tablica 3.20: Kandydaci na kolejne rozszerzenie części warunkowej reguły dla algorytmu DOMLEM, metody DRSA i pierwszego zbioru danych bez wariantu a3

Numer	Kandydat	Ocena kandydata na kryterium 1	Ocena kandydata na kryterium 2
1	$(infertility \geq 3.0)$	$\frac{1}{1}$	1
2	$(oocytes \geq 1.0)$	$\frac{1}{2}$	1
3	$(sperm \leq 3.0)$	$\frac{1}{2}$	1
4	$(morpho\_quality \leq 12.0)$	$\frac{1}{2}$	1
5	$(develop\_quality \leq 3.0)$	$\frac{1}{1}$	1

W ramach kontynuacji przykładu dla VC-DRSA, otrzymujemy reguły ujęte w Tabeli 3.21. *Reguła 1* powstaje tu analogicznie jak dla przykładu z DRSA, również składa się z trzech warunków, gdyż po dodaniu pierwszego dokładność reguły to $\frac{1}{6}$ i jest ona mniejsza niż próg spójności ustawiony na wartość 0.8. Po dodaniu drugiego warunku sytuacja ma się analogicznie:
dokładność = $\frac{1}{2} \leq$ l. Dopiero po dodaniu trzeciego warunku możemy zaakceptować regułę.

Tablica 3.21: Reguły dla algorytmu DOMLEM, metody VC-DRSA, $l = 0.8$ i pierwszego zbioru danych bez wariantu a3

Oznaczenie	Reguła
Reguła 1	$(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \wedge (infertility \geq 3.0) \implies (class \leq 1)$
Reguła 2	$(age \geq 39.0) \wedge (infertility \geq 3.0) \implies (class \leq 2)$
Reguła 3	$(morpho\_quality \leq 9.0) \wedge (age \geq 33.0) \wedge (infertility \geq 5.0) \implies (class \leq 2)$
Reguła 4	$(age \geq 27.0) \implies (class \leq 3)$
Reguła 5	$(sperm \leq 2.0) \implies (class \leq 3)$
Reguła 6	$(age \leq 38.0) \implies (class \geq 3)$
Reguła 7	$(infertility \leq 2.0) \implies (class \geq 3)$
Reguła 8	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 9	$(woman\_eval \geq 3.0) \implies (class \geq 2)$
Reguła 10	$(infertility \leq 2.0) \implies (class \geq 2)$

Algorytm DOMApriori

Algorytm DOMApriori pozwala na wyznaczenie satysfakcjonującego zbioru reguł. Pseudokod algorytmu w wersji dla reguł typu pierwszego, przy założeniu, że wszystkie kryteria są typu zysk przedstawiono w Algorytmie 2 [17][p.100-101]. Zaprezentowano pseudokod dla wersji DRSA wraz z komentarzem pozwalającym na wykorzystanie go dla VC-DRSA.

Algorytm 2 uruchamiany jest kolejno dla poszczególnych przybliżeń unii poczynając od najsilniejszych. W wersji dla indukcji reguł innych typów lub pracy na kryteriach typu koszt odpowiednim aktualizacjom ulegają znaki w kandydatach na części warunkowe reguł. Pierwszy krok metody polega na zbudowaniu wszystkich kandydatów na potencjalne reguły i odfiltrowaniu tych, którzy spełniają narzucony przez użytkownika wymóg minimalnego wsparcia. Następnie, iteracyjnie powtarzana jest procedura budowy nowych kandydatów i zapamiętywania jedynie tych, dla których spełniony jest wymóg minimalnego wsparcia. Do poprawnego działania algorytmu wymagane jest utrzymywanie liczników przykładów pozytywnych pokrytych przez zbiór warunków (c.positive_support) i negatywnych (c.negative_support). Te pierwsze stanowią warianty pasujące do części warunkowej reguły i należące do aktualnie rozpatrywanego przybliżenia unii, zaś te drugie określają obiekty również spełniające część warunkową, jednak nie należące do $\underline{P}(Cl_s^{\geq})$. Ostatnim krokiem jest przeprowadzenie dwuetapowego testu minimalności reguł:

- wewnątrz danego przybliżenia unii,
- przeciwko silniejszym przybliżeniom unii.

Przedstawiony pseudokod korzysta z dwóch funkcji pomocniczych, których schematy zostały umieszczone w Algorytmie 3 i 4. Funkcje te służą odpowiednio zbudowaniu inicjalnej rodziny zaakceptowanych silnych zbiorów warunków (L_1) (patrz Algorytm 3) i tworzeniu rodzin kandydujących zbiorów warunków (C_k) (patrz Algorytm 4).

Algorithm 2 Psudokod algorytmu DOMApriori - część główna

procedure DOMAPRIORI($\underline{P}(Cl_s^{\geq})$)
  $C_1 \leftarrow \varnothing$
  $L_1 \leftarrow createconditions(\underline{P}(Cl_s^{\geq}))$
  $k \leftarrow 2$
  while $L_{k-1} \neq \varnothing$ and $(k \leq maxlength)$ **do** ▷ powtarzaj dopóki można znaleźć silne zbiory warunków
    $C_k \leftarrow apriori2\_gen(L_{k-1})$ ▷ znajdź nowych kandydatów
    for wszystkich $x \in U$ **do**
      $Updatesupport(C_k, x)$
    end for
    $L_k \leftarrow \{c \in C_k : c.positive\_support \geq minsupport\}$
    $k \leftarrow k + 1$
  end while ▷ generuj reguły
  for $allL_k$ **do** ▷ rozważaj k od 1 do $maxlength(L_k)$
    for each itemset $c \in L_k$ **do**
      if $c.negative\_support > 0$ **then** ▷ sprawdź warunek akceptacji reguły dla DRSA
      ▷ lub $c.positive\_support$ $(c.positive\_support + c.negative\_support) < l$ dla VC-DRSA
        usuń c z L_k
      end if
    end for
    for $k = 2$ to $maxlength(L_k)$ **do** ▷ test minimalności wewnątrz przybliżenia unii
      for każdy $c \in L_k$ **do**
        for $j = k$ downto 2 **do**
          for każdy $c1 \in L_j$ **do**
            if c1 posiada przynajmniej tak ogólną część warunkową jak c and dokładność koniunkcji c $\geq$ dokładność koniunkcji c1 **then**
              usuń c z L_k
            end if
          end for
        end for
      end for
    end for
  end for
  utwórz reguły $R_{\geq}$ na podstawie $\bigcup_k L_k$
end procedure

Algorithm 3 Psudokod algorytmu DOMApriori – funkcja CREATECONDITIONS

function CREATECONDITIONS($\underline{P}(Cl_s^{\geq})$) ▷ stwórz L_1
  for każdego $x \in \underline{P}(Cl_s^{\geq})$ **do**
    for każdego kryterium $a_i \in C$ **do**
      $r_{ai} \leftarrow f(a_i, x)$
      if warunek $(f(a_i, x) \geq r_{ai}) \notin C_1$ **then**
        dodaj $(f(a_i, x) \geq r_{ai})$ do C_1
      end if
    end for
  end for
  for każdego $x \in U$ **do**
    $Updatesupport(C_1, x)$ ▷ uaktualnij liczniki c.positive_support i c.negative_support dla wszystkich $c \in C_1$, które pokrywają x
  end for
  $L_1 \leftarrow \{c \in C_1 : c.positive\_support \geq minsupport\}$ ▷ odrzuć za słabe warunki
end function

Algorithm 4 Psudokod algorytmu DOMApriori – funkcja APRIORI_2GEN

function APRIORI2_GEN(L_{k-1})
  insert into C_k
  select $p.item_1, p.item_2, ..., p.item_{k-2}, p.item_{k-1}, q.item_{k-1}$ ▷ wykorzystaj porządek leksykograficzny warunków przy połączeniu dwóch silnych zbiorów $L_{k-1}p, L_{k-1}q$
  from $L_{k-1}p, L_{k-1}q$
  where
  $p.negative\_support > 0$ ▷ nie rozszerzaj pewnych reguł
  $q.negative\_support > 0$
  $p.item_1 = q.item_1$
  $p.item_2 = q.item_2, ..., p.item_{k-2} = q.item_{k-2}$
  $p.item_{k-1} < q.item_{k-1}$ ▷ zbiory $L_{k-1}p, L_{k-1}q$ mogą się różnić tylko na ostatniej pozycji
  $p.item_{k-1}, q.item_{k-1}$ nie dotyczą tego samego kryterium
  for wszystkich (k-1) podzbiorów $s \in C_k$ **do** ▷ podzbiory silnych warunków muszą być silne
    if $s \notin L_{k-1}$ **then**
      usuń s z C_k
    end if
  end for
end function

Przykład z pierwszego zbioru danych

Kontynuując przykład dla metody DRSA i korzystając z algorytmu DOMApriori otrzymujemy zbiór reguł przedstawiony w Tablicy 3.22. Są to reguły wyindukowane z dolnych przybliżeń unii klas w dół (pełny zbiór liczy 69 reguł dlatego zaprezentowano jego część) przy minimalnym wsparciu równym 1 i maksymalnej długości reguły ustalonej na wartość 3.

Na przykład indukując reguły dla przybliżenia unii *co najwyżej klasa 1*, do którego należy tylko wariant a23, otrzymujemy w pierwszym kroku rodzinę silnych zbiorów przedstawioną w Tabeli 3.23.

W kolejnym etapie budujemy zbiór C_2, ponieważ L_1 nie jest pusty oraz k=2, czyli spełniony jest warunek $k \leq maxlength$. Powstają wówczas kandydaci ujęci w Tabeli 3.24. Ponadto dla każdego $c \in C_k : c.positive\_support \geq minsupport$, więc $C_2 = L_2$. Następnie, ponieważ L_2 jest pusty oraz $k = 3 \geq minsupport$, generujemy zbiór C_3 przedstawiony w Tabeli 3.25.

Otrzymujemy wniosek analogiczny do poprzedniego, czyli $C_3 = L_3$ i ponieważ k=4 kończymy fazę generacji reguł. Następnie sprawdzamy, dla których reguł spełniony jest warunek $c.negative\_support > 0$ i usuwamy te reguły, wówczas otrzymujemy tylko regułę:

$$(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (woman\_eval \leq 2.0) \implies class \leq 1$$

Kończymy działanie algorytmu dla tego przybliżenia, gdyż reguła jest minimalna.

TABLICA 3.22: Reguły indukowane z dolnych przybliżeń unii klas w dół dla algorytmu DOMApriori przy $min\_support = 1$, $max\_length = 3$ i metody DRSA dla pierwszego zbioru danych bez wariantu a3

Oznaczenie	Reguła
Reguła 1	$(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (woman\_eval \leq 2.0) \implies (class \leq 1)$
Reguła 2	$(age \geq 42.0) \implies (class \leq 2)$
Reguła 3	$(morpho\_quality \leq 5.0) \implies (class \leq 2)$
Reguła 4	$(woman\_eval \leq 3.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 5	$(sperm \leq 1.0) \wedge (morpho\_quality \leq 14.0) \implies (class \leq 2)$
Reguła 6	$(sperm \leq 1.0) \wedge (develop\_quality \leq 4.0) \implies (class \leq 2)$
Reguła 7	$(age \geq 39.0) \wedge (infertility \geq 3.0) \implies (class \leq 2)$
Reguła 8	$(age \geq 39.0) \wedge (morpho\_quality \leq 11.0) \implies (class \leq 2)$
Reguła 9	$(age \geq 39.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 10	$(age \geq 33.0) \wedge (sperm \leq 1.0) \implies (class \leq 2)$
Reguła 11	$(age \geq 33.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 12	$(morpho\_quality \leq 9.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 13	$(morpho\_quality \leq 8.0) \wedge (develop\_quality \leq 2.0) \implies (class \leq 2)$
Reguła 14	$(infertility \geq 3.0) \wedge (woman\_eval \leq 3.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 15	$(infertility \geq 3.0) \wedge (oocytes \geq 4.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 16	$(infertility \geq 3.0) \wedge (woman\_eval \leq 4.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 17	$(woman\_eval \leq 3.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 18	$(infertility \geq 4.0) \wedge (woman\_eval \leq 3.0) \wedge (morpho\_quality \leq 12.0) \implies (class \leq 2)$
Reguła 19	$(infertility \geq 4.0) \wedge (oocytes \geq 4.0) \wedge (morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 20	$(infertility \geq 4.0) \wedge (woman\_eval \leq 4.0) \wedge (morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 21	$(age \geq 33.0) \wedge (infertility \geq 4.0) \wedge (morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 22	$(sperm \leq 2.0) \implies (class \leq 3)$
Reguła 23	$(infertility \geq 4.0) \implies (class \leq 3)$
Reguła 24	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 25	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 26	$(morpho\_quality \leq 12.0) \implies (class \leq 3)$
Reguła 27	$(age \geq 35.0) \implies (class \leq 3)$

Tablica 3.23: Rodzina silnych zbiorów warunków L_1 dla algorytmu DOMApriori przy $min\_support = 1$, $max\_length = 3$ i metody DRSA dla pierwszego zbioru danych bez wariantu a3

Numer	Silny zbiór
1	$\{(age \geq 40.0)\}$
2	$\{(infertility \geq 3.0)\}$
3	$\{(oocytes \geq 1.0)\}$
4	$\{(woman\_eval \leq 2.0)\}$
5	$\{(sperm \leq 3.0)\}$
6	$\{(morpho\_quality \leq 12.0)\}$
7	$\{(develop\_quality \leq 3.0)\}$

Tablica 3.24: Rodzina zbiorów kandydujących warunków C_2 dla algorytmu DOMApriori przy $min\_support = 1$, $max\_length = 3$ i metody DRSA dla pierwszego zbioru danych bez wariantu a3

Numer	Kandydat
1	$\{(age \geq 40.0) \wedge (infertility \geq 3.0)\}$
2	$\{(age \geq 40.0) \wedge (oocytes \geq 1.0)\}$
3	$\{(age \geq 40.0) \wedge (woman\_eval \leq 2.0)\}$
4	$\{(age \geq 40.0) \wedge (sperm \leq 3.0)\}$
5	$\{(age \geq 40.0) \wedge (morpho\_quality \leq 12.0)\}$
6	$\{(age \geq 40.0) \wedge (develop\_quality \leq 3.0)\}$
7	$\{(infertility \geq 3.0) \wedge (oocytes \geq 1.0)\}$
8	$\{(infertility \geq 3.0) \wedge (woman\_eval \leq 2.0)\}$
9	$\{(infertility \geq 3.0) \wedge (sperm \leq 3.0)\}$
10	$\{(infertility \geq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
11	$\{(infertility \geq 3.0) \wedge (develop\_quality \leq 3.0)\}$
12	$\{(oocytes \geq 1.0) \wedge (woman\_eval \leq 2.0)\}$
13	$\{(oocytes \geq 1.0) \wedge (sperm \leq 3.0)\}$
14	$\{(oocytes \geq 1.0) \wedge (morpho\_quality \leq 12.0)\}$
15	$\{(oocytes \geq 1.0) \wedge (develop\_quality \leq 3.0)\}$
16	$\{(woman\_eval \leq 2.0) \wedge (sperm \leq 3.0)\}$
17	$\{(woman\_eval \leq 2.0) \wedge (morpho\_quality \leq 12.0)\}$
18	$\{(woman\_eval \leq 2.0) \wedge (develop\_quality \leq 3.0)\}$
19	$\{(sperm \leq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
20	$\{(sperm \leq 3.0) \wedge (develop\_quality \leq 3.0)\}$
21	$\{(morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 3.0)\}$

TABLICA 3.25: Rodzina zbiorów kandydujących warunków C_3 dla algorytmu DOMApriori przy $min\_support = 1$, $max\_length = 3$ i metody DRSA dla pierwszego zbioru danych bez wariantu a3

Numer	Kandydat
1	$\{(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (oocytes \geq 1.0)\}$
2	$\{(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (woman\_eval \leq 2.0)\}$
3	$\{(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (sperm \leq 3.0)\}$
4	$\{(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
5	$\{(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (develop\_quality \leq 3.0)\}$
6	$\{(age \geq 40.0) \wedge (oocytes \geq 1.0) \wedge (woman\_eval \leq 2.0)\}$
7	$\{(age \geq 40.0) \wedge (oocytes \geq 1.0) \wedge (sperm \leq 3.0)\}$
8	$\{(age \geq 40.0) \wedge (oocytes \geq 1.0) \wedge (morpho\_quality \leq 12.0)\}$
9	$\{(age \geq 40.0) \wedge (oocytes \geq 1.0) \wedge (develop\_quality \leq 3.0)\}$
10	$\{(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \wedge (sperm \leq 3.0)\}$
11	$\{(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \wedge (morpho\_quality \leq 12.0)\}$
12	$\{(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \wedge (develop\_quality \leq 3.0)\}$
13	$\{(age \geq 40.0) \wedge (sperm \leq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
14	$\{(age \geq 40.0) \wedge (sperm \leq 3.0) \wedge (develop\_quality \leq 3.0)\}$
15	$\{(age \geq 40.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 3.0)\}$
16	$\{(infertility \geq 3.0) \wedge (oocytes \geq 1.0) \wedge (woman\_eval \leq 2.0)\}$
17	$\{(infertility \geq 3.0) \wedge (oocytes \geq 1.0) \wedge (sperm \leq 3.0)\}$
18	$\{(infertility \geq 3.0) \wedge (oocytes \geq 1.0) \wedge (morpho\_quality \leq 12.0)\}$
19	$\{(infertility \geq 3.0) \wedge (oocytes \geq 1.0) \wedge (develop\_quality \leq 3.0)\}$
20	$\{(infertility \geq 3.0) \wedge (woman\_eval \leq 2.0) \wedge (sperm \leq 3.0)\}$
21	$\{(infertility \geq 3.0) \wedge (woman\_eval \leq 2.0) \wedge (morpho\_quality \leq 12.0)\}$
22	$\{(infertility \geq 3.0) \wedge (woman\_eval \leq 2.0) \wedge (develop\_quality \leq 3.0)\}$
23	$\{(infertility \geq 3.0) \wedge (sperm \leq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
24	$\{(infertility \geq 3.0) \wedge (sperm \leq 3.0) \wedge (develop\_quality \leq 3.0)\}$
25	$\{(infertility \geq 3.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 3.0)\}$
26	$\{(oocytes \geq 1.0) \wedge (woman\_eval \leq 2.0) \wedge (sperm \leq 3.0)\}$
27	$\{(oocytes \geq 1.0) \wedge (woman\_eval \leq 2.0) \wedge (morpho\_quality \leq 12.0)\}$
28	$\{(oocytes \geq 1.0) \wedge (woman\_eval \leq 2.0) \wedge (develop\_quality \leq 3.0)\}$
29	$\{(oocytes \geq 1.0) \wedge (sperm \leq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
30	$\{(oocytes \geq 1.0) \wedge (sperm \leq 3.0) \wedge (develop\_quality \leq 3.0)\}$
31	$\{(oocytes \geq 1.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 3.0)\}$
32	$\{(woman\_eval \leq 2.0) \wedge (sperm \leq 3.0) \wedge (morpho\_quality \leq 12.0)\}$
33	$\{(woman\_eval \leq 2.0) \wedge (sperm \leq 3.0) \wedge (develop\_quality \leq 3.0)\}$
34	$\{(woman\_eval \leq 2.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 3.0)\}$
35	$\{(sperm \leq 3.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 3.0)\}$

Rozwijając przykład dla metody VC-DRSA i korzystając z algorytmu DOMApriori otrzymujemy zbiór reguł ujęty w Tablicy 3.26, wyindukowanych z dolnych przybliżeń unii klas w dół przy minimalnym wsparciu równym 1 i maksymalnej długości reguły ustalonej na wartość 3.

Na przykład indukując reguły dla przybliżenia unii *co najwyżej klasa 1*, do którego należy tylko wariant a23, otrzymujemy rezultaty analogiczne jak dla DRSA, jedynie w ostatnim kroku odrzucamy reguły, dla których zachodzi:

$$c.positive\_support/(c.positive\_support + c.negative\_support) < l$$

Ostatecznie otrzymujemy dla tego przybliżenia taki sam wynik jak przy metodzie DRSA.

TABLICA 3.26: Reguły indukowane z dolnych przybliżeń unii klas w dół dla algorytmu DOMApriori przy $min\_support = 1$, $max\_length = 3$ i metody VC-DRSA, $l = 0.8$ dla pierwszego zbioru danych bez wariantu a3

Oznaczenie	Reguła
Reguła 1	$(age \geq 40.0) \wedge (infertility \geq 3.0) \wedge (woman\_eval \leq 2.0) \implies (class \leq 1)$
Reguła 2	$(age \geq 42.0) \implies (class \leq 2)$
Reguła 3	$(morpho\_quality \leq 5.0) \implies (class \leq 2)$
Reguła 4	$(woman\_eval \leq 3.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 5	$(sperm \leq 1.0) \wedge (morpho\_quality \leq 14.0) \implies (class \leq 2)$
Reguła 6	$(sperm \leq 1.0) \wedge (develop\_quality \leq 4.0) \implies (class \leq 2)$
Reguła 7	$(age \geq 39.0) \wedge (infertility \geq 3.0) \implies (class \leq 2)$
Reguła 8	$(age \geq 39.0) \wedge (morpho\_quality \leq 11.0) \implies (class \leq 2)$
Reguła 9	$(age \geq 39.0) \wedge (oocytes \geq 4.0) \implies (class \leq 2)$
Reguła 10	$(age \geq 39.0) \wedge (morpho\_quality \leq 14.0) \implies (class \leq 2)$
Reguła 11	$(age \geq 39.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 12	$(age \geq 39.0) \wedge (develop\_quality \leq 4.0) \implies (class \leq 2)$
Reguła 13	$(age \geq 33.0) \wedge (sperm \leq 1.0) \implies (class \leq 2)$
Reguła 14	$(age \geq 33.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 15	$(morpho\_quality \leq 9.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 16	$(morpho\_quality \leq 8.0) \wedge (develop\_quality \leq 2.0) \implies (class \leq 2)$
Reguła 17	$(infertility \geq 3.0) \wedge (woman\_eval \leq 3.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 18	$(infertility \geq 3.0) \wedge (oocytes \geq 4.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 19	$(infertility \geq 3.0) \wedge (woman\_eval \leq 4.0) \wedge (morpho\_quality \leq 8.0) \implies (class \leq 2)$
Reguła 20	$(woman\_eval \leq 3.0) \wedge (morpho\_quality \leq 12.0) \wedge (develop\_quality \leq 1.0) \implies (class \leq 2)$
Reguła 21	$(infertility \geq 4.0) \wedge (woman\_eval \leq 3.0) \wedge (morpho\_quality \leq 12.0) \implies (class \leq 2)$
Reguła 22	$(infertility \geq 4.0) \wedge (oocytes \geq 4.0) \wedge (morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 23	$(infertility \geq 4.0) \wedge (woman\_eval \leq 4.0) \wedge (morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 24	$(age \geq 33.0) \wedge (infertility \geq 4.0) \wedge (morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 25	$(infertility \geq 1.0) \implies (class \leq 3)$
Reguła 26	$(sperm \leq 3.0) \implies (class \leq 3)$
Reguła 27	$(sperm \leq 2.0) \implies (class \leq 3)$
Reguła 28	$(morpho\_quality \leq 20.0) \implies (class \leq 3)$
Reguła 29	$(develop\_quality \leq 5.0) \implies (class \leq 3)$
Reguła 30	$(woman\_eval \leq 6.0) \implies (class \leq 3)$
Reguła 31	$(infertility \geq 4.0) \implies (class \leq 3)$
Reguła 32	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 33	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 34	$(oocytes \geq 1.0) \implies (class \leq 3)$
Reguła 35	$(morpho\_quality \leq 12.0) \implies (class \leq 3)$
Reguła 36	$(age \geq 35.0) \implies (class \leq 3)$
Reguła 37	$(age \geq 26.0) \implies (class \leq 3)$

3.1.4 Algorytmy klasyfikacji

Na podstawie wyindukowanych reguł możemy klasyfikować nowe przykłady, czemu służy wykorzystanie algorytmów klasyfikacji. W pracy wykorzystano dwa z nich. Pierwszy jest prostą metodą [8], zaś drugi wykorzystuje bardziej zaawansowane podejście [2].

Standardowy algorytm

Pierwszym krokiem metody jest znalezienie zbioru reguł R o części warunkowej dopasowującej się do obiektu, który chcemy zaklasyfikować. Następnie postępowanie zależy od mocy wyznaczonego zbioru. Możliwe są trzy sytuacje:

- $|R| = 0$ (w zbiorze nie ma żadnych elementów)
  Wynikiem klasyfikacji jest przypisanie obiektu do każdej z rozważanych klas decyzyjnych.
- $|R| = 1$ (w zbiorze znajduje się dokładnie jedna reguła)
  Wynik klasyfikacji zależy od typu dopasowanej reguły.
  - reguły typu 1 lub 2 (co najmniej klasa)
    Obiektowi zostaje przypisana najgorsza z klas, na które wskazuje reguła (na przykład co najmniej klasa 2 - wynik to klasa 2).
  - reguły typu 3 lub 4 (co najwyżej klasa)
    Obiektowi zostaje przypisana najlepsza z klas, na które wskazuje reguła (na przykład co najwyżej klasa 2 - wynik to klasa 2).
- $|R| > 1$ (liczba reguł w zbiorze jest większa niż 1)
  Jeśli każda z reguł w zbiorze wskazuje tę samą klasę stosujemy zasadę z poprzedniego punktu, w przeciwnym wypadku wyznaczamy przecięcie unii klas biorąc:
  - dla unii co najmniej klasa: najgorszą z klas (na przykład co najmniej klasa 3 i co najmniej klasa 2 - wynik co najmniej klasa 2),
  - dla unii co najwyżej klasa: najlepszą z klas (na przykład co najwyżej klasa 3 i co najwyżej klasa 2 - wynik co najwyżej klasa 3).

  Gdy wszystkie reguły wskazywały tylko na unie tego samego typu – odpowiednio co najmniej lub co najwyżej – to powyższy wynik jest ostateczny. W przeciwnym wypadku wyznaczamy przecięcie powyższych odpowiedzi, czyli zbiór klas ujętych przez każdą z dwóch odpowiedzi (dla przykładu: co najmniej klasa 2 i co najwyżej klasa 3 – otrzymujemy klasa 2 i 3). Jeśli przecięcie jest zbiorem pustym obiekt może zostać sklasyfikowany do każdej lub do żadnej klasy.

Przykład z pierwszego zbioru danych

Klasyfikowany obiekt to usunięty z oryginalnego zbioru pierwszego wariant a3. Przyjmuje on następujące oceny na poszczególnych kryteriach zebrane w Tabeli 3.27.

Tablica 3.27: Testowy wariant a3 z pierwszego zbioru danych

Object	age	infertility	oocytes	woman_eval	sperm	morpho_quality	develop_quality	class
a3	38	1	6	1	3	14	4	3

Kontynuacja przykładu dla DRSA i DOMLEM

Do przykładu a3 z pierwszego zbioru danych dopasowano reguły ujęte w Tabeli 3.28. Zgodnie z danymi w niej zawartymi dopasowaniu uległo 5 reguł opisujących odpowiednio przybliżenie unii *najwyżej klasa 3* oraz *co najmniej klasa 2*, dlatego przykład a3 zostanie sklasyfikowany do klas 2 lub 3 zgodnie z opisem postępowania dla $|R| = 1$, czyli w propozycji algorytmu zawiera się oryginalna decyzja dla tego wariantu.

Tablica 3.28: Reguły dopasowane do wariantu a3 z pierwszego zbioru danych dla algorytmu DOMLEM i metody DRSA

Oznaczenie	Reguła
Reguła 4	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 6	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 7	$(age \geq 35.0) \implies (class \leq 3)$
Reguła 15	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 17	$(infertility \leq 2.0) \implies (class \geq 2)$

Kontynuacja przykładu dla DRSA i DOMApriori

W Tabeli 3.29 zebrano reguły, których część warunkową spełnia przykład a3 z pierwszego zbioru danych. Dopasowaniu uległo w tym przypadku $|R| = 9$ reguł. Opisują one przybliżenia unii:

- co najwyżej klasa 3,
- co najmniej klasa 3,
- co najmniej klasa 2.

Biorąc pod uwagę wypisane przybliżenia szukamy przecięcia dla zbiorów *co najwyżej klasa 3* i *co najmniej klasa 2*. Otrzymujemy wynik, że a3 należy do klasy 2 lub 3, zatem w propozycji algorytmu zawiera się oryginalna decyzja dla tego wariantu.

Tablica 3.29: Reguły dopasowane do wariantu a3 z pierwszego zbioru danych dla algorytmu DOMApriori, $min\_support = 1$, $max\_length = 3$ i metody DRSA

Oznaczenie	Reguła
Reguła 24	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 25	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 27	$(age \geq 35.0) \implies (class \leq 3)$
Reguła 37	$(infertility \leq 3.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 64	$(age \leq 41.0) \wedge (morpho\_quality \geq 14.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 66	$(infertility \leq 2.0) \implies (class \geq 2)$
Reguła 67	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 68	$(develop\_quality \geq 4.0) \implies (class \geq 2)$
Reguła 69	$(morpho\_quality \geq 13.0) \implies (class \geq 2)$

Kontynuacja przykładu dla VC-DRSA i DOMLEM

Do przykładu a3 z pierwszego zbioru danych dopasowano reguły ujęte w Tabeli 3.30. Zbiór dopasowanych reguł liczy w tym przypadku 5 elementów, a podjęta decyzja to przypisanie klasy 2 lub 3 analogicznie jak w poprzednim przykładzie.

Tablica 3.30: Reguły dopasowane do wariantu a3 z pierwszego zbioru danych dla algorytmu DOMLEM i metody VC-DRSA, $l = 0.8$

Oznaczenie	Reguła
Reguła 4	$(age \geq 27.0) \implies (class \leq 3)$
Reguła 6	$(age \leq 38.0) \implies (class \geq 3)$
Reguła 7	$(infertility \leq 2.0) \implies (class \geq 3)$
Reguła 8	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 10	$(infertility \leq 2.0) \implies (class \geq 2)$

Kontynuacja przykładu dla VC-DRSA i DOMApriori

W Tabeli 3.31 zebrano reguły, których część warunkową spełnia przykład a3 z pierwszego zbioru danych. Dla tego przykładu otrzymujemy $|R| = 108$.
Dopasowane reguły odpowiadają następującym przybliżeniom unii klas:

- co najwyżej klasa 2,
- co najwyżej klasa 3,
- co najmniej klasa 3,
- co najmniej klasa 2.

Ostatecznie poszukujemy przecięcia dla przybliżeń unii *co najwyżej klasa 3* i *co najmniej klasa 2*. Wynikiem jest przypisanie obiektowi a3 klasy 2 lub 3, czyli w propozycji algorytmu zawiera się oryginalna decyzja dla tego wariantu.

Tablica 3.31: Reguły dopasowane do wariantu a3 z pierwszego zbioru danych dla algorytmu DOMApriori , $min\_support = 1$, $max\_length = 3$ i metody VC-DRSA, $l = 0.8$

Oznaczenie	Reguła
Reguła 5	$(sperm \leq 1.0) \wedge (morpho\_quality \leq 14.0) \implies (class \leq 2)$
Reguła 6	$(sperm \leq 1.0) \wedge (develop\_quality \leq 4.0) \implies (class \leq 2)$
Reguła 13	$(age \geq 33.0) \wedge (sperm \leq 1.0) \implies (class \leq 2)$
Reguła 25	$(infertility \geq 1.0) \implies (class \leq 3)$
Reguła 26	$(sperm \leq 3.0) \implies (class \leq 3)$
Reguła 28	$(morpho\_quality \leq 20.0) \implies (class \leq 3)$
Reguła 30	$(woman\_eval \leq 6.0) \implies (class \leq 3)$
Reguła 32	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 33	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 34	$(oocytes \geq 1.0) \implies (class \leq 3)$
Reguła 36	$(age \geq 35.0) \implies (class \leq 3)$
Reguła 37	$(age \geq 26.0) \implies (class \leq 3)$
Reguła 39	$(sperm \geq 2.0) \implies (class \geq 3)$
Reguła 42	$(develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 43	$(oocytes \leq 6.0) \implies (class \geq 3)$

Oznaczenie	Reguła
Reguła 44	$(morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 45	$(infertility \leq 2.0) \implies (class \geq 3)$
Reguła 48	$(infertility \leq 3.0) \implies (class \geq 3)$
Reguła 51	$(morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 52	$(develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 53	$(morpho\_quality \geq 14.0) \implies (class \geq 3)$
Reguła 57	$(develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 59	$(develop\_quality \geq 1.0) \implies (class \geq 3)$
Reguła 60	$(morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 61	$(morpho\_quality \geq 12.0) \implies (class \geq 3)$
Reguła 62	$(morpho\_quality \geq 13.0) \implies (class \geq 3)$
Reguła 63	$(infertility \leq 13.0) \implies (class \geq 3)$
Reguła 65	$(age \leq 38.0) \implies (class \geq 3)$
Reguła 66	$(sperm \geq 1.0) \implies (class \geq 3)$
Reguła 67	$(age \leq 41.0) \implies (class \geq 3)$
Reguła 72	$(sperm \geq 2.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 73	$(sperm \geq 2.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 74	$(sperm \geq 2.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 75	$(sperm \geq 2.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 76	$(sperm \geq 2.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 85	$(infertility \leq 3.0) \wedge (sperm \geq 2.0) \implies (class \geq 3)$
Reguła 86	$(infertility \leq 3.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 88	$(infertility \leq 3.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 91	$(infertility \leq 3.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 92	$(infertility \leq 3.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 94	$(infertility \leq 3.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 95	$(infertility \leq 3.0) \wedge (morpho\_quality \geq 12.0) \implies (class \geq 3)$
Reguła 96	$(infertility \leq 3.0) \wedge (morpho\_quality \geq 13.0) \implies (class \geq 3)$
Reguła 111	$(morpho\_quality \geq 14.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 128	$(infertility \leq 4.0) \wedge (sperm \geq 2.0) \implies (class \geq 3)$
Reguła 129	$(infertility \leq 4.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 131	$(infertility \leq 4.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 133	$(infertility \leq 4.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 134	$(infertility \leq 4.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 136	$(infertility \leq 4.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 138	$(morpho\_quality \geq 10.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 139	$(morpho\_quality \geq 13.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 143	$(infertility \leq 5.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 144	$(infertility \leq 5.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 145	$(infertility \leq 5.0) \wedge (morpho\_quality \geq 12.0) \implies (class \geq 3)$
Reguła 146	$(infertility \leq 5.0) \wedge (morpho\_quality \geq 13.0) \implies (class \geq 3)$
Reguła 151	$(age \leq 38.0) \wedge (infertility \leq 3.0) \implies (class \geq 3)$
Reguła 154	$(age \leq 38.0) \wedge (infertility \leq 4.0) \implies (class \geq 3)$
Reguła 155	$(age \leq 38.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 156	$(age \leq 38.0) \wedge (infertility \leq 5.0) \implies (class \geq 3)$

Oznaczenie	Reguła
Reguła 157	$(age \leq 41.0) \wedge (develop\_quality \geq 5.0) \implies (class \geq 3)$
Reguła 158	$(age \leq 41.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 159	$(age \leq 41.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 161	$(age \leq 41.0) \wedge (infertility \leq 3.0) \implies (class \geq 3)$
Reguła 163	$(age \leq 41.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 164	$(age \leq 41.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 166	$(age \leq 41.0) \wedge (morpho\_quality \geq 14.0) \implies (class \geq 3)$
Reguła 168	$(age \leq 41.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 169	$(age \leq 41.0) \wedge (infertility \leq 4.0) \implies (class \geq 3)$
Reguła 170	$(age \leq 41.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 171	$(age \leq 41.0) \wedge (morpho\_quality \geq 13.0) \implies (class \geq 3)$
Reguła 180	$(infertility \leq 3.0) \wedge (sperm \geq 2.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 181	$(infertility \leq 3.0) \wedge (sperm \geq 2.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 182	$(infertility \leq 3.0) \wedge (sperm \geq 2.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 183	$(infertility \leq 3.0) \wedge (sperm \geq 2.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 218	$(infertility \leq 4.0) \wedge (sperm \geq 2.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 219	$(infertility \leq 4.0) \wedge (sperm \geq 2.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 220	$(infertility \leq 4.0) \wedge (sperm \geq 2.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 221	$(infertility \leq 4.0) \wedge (sperm \geq 2.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 233	$(infertility \leq 5.0) \wedge (sperm \geq 2.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 234	$(infertility \leq 5.0) \wedge (sperm \geq 2.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 256	$(infertility \leq 5.0) \wedge (morpho\_quality \geq 10.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 257	$(infertility \leq 5.0) \wedge (morpho\_quality \geq 13.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 261	$(age \leq 38.0) \wedge (infertility \leq 4.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 262	$(age \leq 38.0) \wedge (morpho\_quality \geq 10.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 266	$(age \leq 38.0) \wedge (infertility \leq 5.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 267	$(age \leq 38.0) \wedge (infertility \leq 5.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 273	$(age \leq 41.0) \wedge (infertility \leq 3.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 274	$(age \leq 41.0) \wedge (infertility \leq 3.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 277	$(age \leq 41.0) \wedge (infertility \leq 3.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 278	$(age \leq 41.0) \wedge (infertility \leq 3.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 290	$(age \leq 41.0) \wedge (morpho\_quality \geq 14.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 292	$(age \leq 41.0) \wedge (infertility \leq 4.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 293	$(age \leq 41.0) \wedge (infertility \leq 4.0) \wedge (morpho\_quality \geq 7.0) \implies (class \geq 3)$
Reguła 295	$(age \leq 41.0) \wedge (infertility \leq 4.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 296	$(age \leq 41.0) \wedge (infertility \leq 4.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 298	$(age \leq 41.0) \wedge (infertility \leq 4.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 299	$(age \leq 41.0) \wedge (morpho\_quality \geq 10.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 300	$(age \leq 41.0) \wedge (morpho\_quality \geq 13.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 301	$(age \leq 41.0) \wedge (infertility \leq 5.0) \wedge (develop\_quality \geq 3.0) \implies (class \geq 3)$
Reguła 304	$(age \leq 41.0) \wedge (infertility \leq 5.0) \wedge (morpho\_quality \geq 9.0) \implies (class \geq 3)$
Reguła 305	$(age \leq 41.0) \wedge (infertility \leq 5.0) \wedge (develop\_quality \geq 2.0) \implies (class \geq 3)$
Reguła 308	$(age \leq 41.0) \wedge (infertility \leq 5.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 309	$(age \leq 41.0) \wedge (infertility \leq 5.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 310	$(age \leq 41.0) \wedge (infertility \leq 5.0) \wedge (morpho\_quality \geq 13.0) \implies (class \geq 3)$

Oznaczenie	Reguła
Reguła 311	$(age \leq 44.0) \implies (class \geq 2)$
Reguła 313	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 314	$(morpho\_quality \geq 5.0) \implies (class \geq 2)$

Zaawansowany algorytm

Pierwszą fazą algorytmu jest znalezienie zbioru reguł (Cov_z), których część warunkowa dopasowuje się do klasyfikowanego przykładu (z). Następnie, w zależności od mocy zbioru Cov_z, postępujemy zgodnie ze schematem:

- $|Cov_z| = 0$
  Dokładnie tak jak dla poprzedniego algorytmu, wynikiem klasyfikacji jest przypisanie obiektu do każdej z rozważanych klas.

- $|Cov_z| = 1$
  Dla dopasowanej reguły ρ i każdej z możliwych klas decyzyjnych obliczamy współczynnik $Score_\rho(Cl_t, z) = \frac{|Cond_\rho \cap Cl_t|^2}{|Cond_\rho||Cl_t|}$, w którym:
  - $Cond_\rho$ oznacza zbiór obiektów dopasowujących się do warunkowej części reguły,
  - Cl_t określa zbiór obiektów należących do klasy t.

  Obiektowi zostaje przypisana klasa, dla której wartość współczynnika jest największa.

- $|Cov_z| > 1$
  Dla dopasowanych reguł i każdej możliwej klasy decyzyjnej wyznaczamy wartość współczynnika $Score_{Cov_z}(Cl_t, z) = Score^+_{Cov_z}(Cl_t, z) - Score^-_{Cov_z}(Cl_t, z)$, na którą składają się:
  - $Score^+_{Cov_z}(Cl_t, z) = \frac{|(Cond_{\rho_1} \cap Cl_t) \cup \ldots \cup (Cond_{\rho_k} \cap Cl_t)|^2}{|Cond_{\rho_1 \cup \ldots \cup Cond_{\rho_k}}||Cl_t|}$, który jest obliczany na postawie reguł (Cov^+_z) sugerujących przypisanie z do klasy takiej, że: $Cl_t \subseteq Cl^{\geq}_s$ i $Cl_t \subseteq Cl^{\leq}_q$,
  - $Score^-_{Cov_z}(Cl_t, z) = \frac{|(Cond_{\rho_{k+1}} \cap Cl^{\geq}_{\rho_{k+1}}) \cup \ldots \cup (Cond_{\rho_l} \cap Cl^{\geq}_{\rho_l}) \cup (Cond_{\rho_{l+1}} \cap Cl^{\leq}_{\rho_{l+1}}) \cup \ldots \cup (Cond_{\rho_h} \cap Cl^{\leq}_{\rho_h})|^2}{|Cond_{\rho_{k+1}} \cup \ldots \cup Cond_{\rho_l} \cup Cond_{\rho_{l+1}} \cup \ldots \cup Cond_{\rho_h}||Cl^{\geq}_{\rho_{k+1}} \cup \ldots \cup Cl^{\geq}_{\rho_l} \cup Cl^{\leq}_{\rho_{l+1}} \cup \ldots \cup Cl^{\leq}_{\rho_h}|}$, który jest obliczany na postawie reguł (Cov^-_z) sugerujących przypisanie z do klasy takiej, że: $Cl_t \cap Cl^{\geq}_s = \varnothing$ i $Cl_t \cap Cl^{\leq}_q = \varnothing$.

  Ostateczny wynik przypisania do klasy wybierany jest na podstawie maksymalnej wartości współczynnika $Score_{Cov_z}(Cl_t, z)$

Przykład ze pierwszego zbioru danych

Klasyfikowany obiekt to, jak w przykładzie dla prostego algorytmu, usunięty z oryginalnego zbioru pierwszego wariant a3. Wartości osiągane przez ten wariant na poszczególnych kryteriach zostały przedstawione w Tabeli 3.27.

Kontynuacja przykładu dla DRSA i DOMLEM

Korzystając z danych w Tabeli 3.28 do wariantu dopasować można 5 reguł, dlatego algorytm będzie realizowany w ostatniej z przedstawionych wersji.

Poszczególne współczynniki zostały wyznaczone w następujący sposób:

- dla klasy 1:
  - $Cov^{+}_{a3} = \{Reg4, Reg6, Reg7\}$,
    Jedyny obiekt należący do klasy 1 to a23.

Tablica 3.32: Elementy pomocnicze do wyznaczenia $Score^{+}_{Cov_{a3}}(Cl_1, a3)$

Reguła	$Cond_\rho$	$Cond_\rho \cap Cl_1$
Reguła 4	{a1, a4, a5, a7, a9, a11, a12, a13, a14, a16, a17, a19, a21, a22, a23, a24, a25, a26, a29, a30, a33, a34, a38, a41, a43, a44, a45, a47, a49, a50, a51}	$\{a23\}$
Reguła 6	{a1, a6, a7, a8, a9, a11, a12, a13, a16, a21, a22, a26, a30, a32, a33, a34, a36, a37, a38, a41, a44, a49, a51}	$\varnothing$
Reguła 7	{a1, a5, a6, a9, a11, a12, a13, a17, a22, a23, a27, a29, a33, a34, a36, a39, a41, a48, a49, a51}	$\{a23\}$

  - Na podstawie danych w Tabeli 3.32 wyznaczamy

$$Score^{+}_{Cov_{a3}}(Cl_1, a3) = \frac{|\{a23 \cup \varnothing \cup a23\}|^2}{39 \star 1} = \frac{1}{39}$$

  - $Cov^{-}_{a3} = \{Reg15, Reg17\}$,
    Do przybliżenia unii *co najmniej klasa 2* należą elementy: {a1, a2, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28, a29, a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a48, a49, a50, a51}

Tablica 3.33: Elementy pomocnicze do wyznaczenia $Score^{-}_{Cov_{a3}}(Cl_1, a3)$

Reguła	$Cond_\rho$	$Cond_\rho \cap Cl^{\geq}_2$
Reguła 15	{a1, a2, a4, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28', 'a29', a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a50}	{a1, a2, a4, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a20, a21, a22, a24, a25, a26, a27, a28', 'a29', a30, a31, a32, a33, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a44, a45, a46, a47, a50}
Reguła 17	{a1, a2, a7, a8, a10, a12, a13, a14, a18, a22, a25, a38, a39, a43, a45, a46, a47, a48, a49}	{a1, a2, a7, a8, a10, a12, a13, a14, a18, a22, a25, a38, a39, a43, a45, a46, a47, a48, a49}

  - Na podstawie danych w Tabeli 3.33 wyznaczamy:

$$Score^{-}_{Cov_{a3}}(Cl_1, a3) = \frac{|46|^2}{46 \star 49} = \frac{46}{49}$$

  - $Score_{Cov_{a3}}(Cl_1, a3) = \frac{1}{39} - \frac{46}{49} \approx -0.9131$;

- dla klasy 2:
  - $Cov^{+}_{a3} = \{Reg4, Reg6, Reg7, Reg15, Reg17\}$,
  - $Score^{+}_{Cov_{a3}}(Cl_2, a3) \approx 0.18$,
  - $Cov^{-}_{a3} = \{\varnothing\}$,
  - $Score^{-}_{Cov_{a3}}(Cl_2, a3) = 0$,
  - $Score_{Cov_{a3}}(Cl_2, a3) \approx 0.18$;
- dla klasy 3:
  - $Cov^{+}_{a3} = \{Reg4, Reg6, Reg7, Reg15, Reg17\}$,
  - $Score^{+}_{Cov_{a3}}(Cl_3, a3) \approx 0.78$,
  - $Cov^{-}_{a3} = \{\varnothing\}$,
  - $Score^{-}_{Cov_{a3}}(Cl_3, a3) = 0$,
  - $Score_{Cov_{a3}}(Cl_3, a3) \approx 0.78$;
- dla klasy 4:
  - $Cov^{+}_{a3} = \{Reg15, Reg17\}$,
  - $Score^{+}_{Cov_{a3}}(Cl_4, a3) \approx 0.0217$,
  - $Cov^{-}_{a3} = \{Reg4, Reg6, Reg7\}$,
  - $Score^{-}_{Cov_{a3}}(Cl_4, a3) \approx 0.7959$,
  - $Score_{Cov_{a3}}(Cl_4, a3) \approx -0.7742$.

Ostateczny wynik klasyfikacji to przypisanie obiektowi a3 z pierwszego zbioru danych klasy 3, co jest zgodne z oryginalną decyzją.

Kontynuacja przykładu dla DRSA i DOMApriori
Zgodnie z Tabelą 3.29 do wariantu a3 pasuje 9 reguł. Otrzymujemy:

- dla klasy 1:
  - $Score^{+}_{Cov_{a3}}(Cl_1, a3) \approx 0.0256$,
  - $Score^{-}_{Cov_{a3}}(Cl_1, a3) \approx 0.9796$,
  - $Score_{Cov_{a3}}(Cl_1, a3) \approx -0.954$;
- dla klasy 2:
  - $Score^{+}_{Cov_{a3}}(Cl_2, a3) \approx 0.18$,
  - $Score^{-}_{Cov_{a3}}(Cl_2, a3) \approx 0.675$,
  - $Score_{Cov_{a3}}(Cl_2, a3) \approx -0.495$;

- dla klasy 3:
  - $Score^{+}_{Cov_{a3}}(Cl_3, a3) \approx 0.78$,
  - $Score^{-}_{Cov_{a3}}(Cl_3, a3) = 0$,
  - $Score_{Cov_{a3}}(Cl_3, a3) \approx 0.78$;
- dla klasy 4:
  - $Score^{+}_{Cov_{a3}}(Cl_4, a3) \approx 0.0208$,
  - $Score^{-}_{Cov_{a3}}(Cl_4, a3) \approx 0.7959$,
  - $Score_{Cov_{a3}}(Cl_4, a3) \approx -0.7751$.

Ostatecznie wariantowi a3 z pierwszego zbioru danych zostaje przypisana poprawna klasa 3.

Kontynuacja przykładu dla VC-DRSA i DOMLEM

Dla danych w Tabeli 3.30 do wariantu a3 z pierwszego zbioru danych pasuje 5 reguł. Otrzymujemy:

- dla klasy 1:
  - $Score^{+}_{Cov_{a3}}(Cl_1, a3) \approx 0.0204$,
  - $Score^{-}_{Cov_{a3}}(Cl_1, a3) \approx 0.9388$,
  - $Score_{Cov_{a3}}(Cl_1, a3) \approx -0.9184$;
- dla klasy 2:
  - $Score^{+}_{Cov_{a3}}(Cl_2, a3) \approx 0.18$,
  - $Score^{-}_{Cov_{a3}}(Cl_2, a3) \approx -.9091$,
  - $Score_{Cov_{a3}}(Cl_2, a3) \approx -0.7291$;
- dla klasy 3:
  - $Score^{+}_{Cov_{a3}}(Cl_3, a3) \approx 0.78$,
  - $Score^{-}_{Cov_{a3}}(Cl_3, a3) = 0$,
  - $Score_{Cov_{a3}}(Cl_3, a3) \approx 0.78$;
- dla klasy 4:
  - $Score^{+}_{Cov_{a3}}(Cl_4, a3) \approx 0.0217$,
  - $Score^{-}_{Cov_{a3}}(Cl_4, a3) \approx 0.9596$,
  - $Score_{Cov_{a3}}(Cl_4, a3) \approx -0.9379$.

Algorytm przypisuje wariantowi a3 klasę zgodną z prawdziwą, czyli 3.

Kontynuacja przykładu dla VC-DRSA i DOMApriori

Zgodnie z Tabelą 3.31 do wariantu a3 z pierwszego zbioru danych pasuje 108 reguł. Otrzymujemy:

- dla klasy 1:
  - $Score^{+}_{Cov_{a3}}(Cl_1, a3) \approx 0.02$,
  - $Score^{-}_{Cov_{a3}}(Cl_1, a3) \approx 0.98$,
  - $Score_{Cov_{a3}}(Cl_1, a3) \approx -0.96$;

- dla klasy 2:
  - $Score^{+}_{Cov_{a3}}(Cl_2, a3) \approx 0.18$,
  - $Score^{-}_{Cov_{a3}}(Cl_2, a3) \approx 0.8$,
  - $Score_{Cov_{a3}}(Cl_2, a3) \approx -0.62$;
- dla klasy 3:
  - $Score^{+}_{Cov_{a3}}(Cl_3, a3) \approx 0.78$,
  - $Score^{-}_{Cov_{a3}}(Cl_3, a3) = 0$,
  - $Score_{Cov_{a3}}(Cl_3, a3) \approx 0.78$;
- dla klasy 4:
  - $Score^{+}_{Cov_{a3}}(Cl_4, a3) \approx 0.02$,
  - $Score^{-}_{Cov_{a3}}(Cl_4, a3) \approx 0.98$,
  - $Score_{Cov_{a3}}(Cl_4, a3) \approx -0.96$.

Podobnie jak w poprzednich przykładach algorytm klasyfikacji przypisał obiektowi a3 z pierwszego zbioru danych poprawną klasę 3.

3.2 Drzewa decyzyjne

Drzewo decyzyjne to struktura, w której wierzchołki reprezentują możliwe do podjęcia alternatywne decyzje. Na ich podstawie ustalany jest odpowiedni wybór końcowy, zaznaczony w liściu. Początkowy wierzchołek drzewa decyzyjnego to korzeń. Przejście od niego do liścia wyznacza – jeden z możliwych – kompletny scenariusz podjęcia decyzji w procesie opisanym drzewem [13]. Algorytmy tworzące drzewa decyzyjne są szeroko stosowane do celów wspomagania decyzji. W literaturze opisanych jest wiele różnych metod budowania drzew decyzyjnych, rozważane są także podejścia, w których tworzone są zespoły drzew.

Najpopularniejsze algorytmy służące do budowy drzew decyzyjnych to ID3, CART, C4.5. każdy z nich opiera się na rekurencyjnym podziale przestrzeni atrybutów. Różnice wynikają ze stosowania różnych kryteriów podziału oraz założeń na temat danych wejściowych [10]. W pracy wykorzystane zostaną Classification And Regression Trees (CART) [19] dla problemu klasyfikacji, ze względu na fakt, iż pozwalają na budowanie w wierzchołkach warunków monotonicznych oraz dobrze radzą sobie z obserwacjami odstającymi.

Drugi wykorzystany w pracy model to Random Forest [1]. Jest to złożony model, w którym podstawowymi komponentami są drzewa decyzyjne. Wykorzystanie algorytmu powinno potencjalnie wpłynąć na uzyskanie wyższej trafności klasyfikacji, przy zachowaniu interpretowalności i prostoty modelu. Podstawową zaletą tego podejścia jest wykorzystanie kilku klasyfikatorów, które pracują niezależnie, a następnie ich odpowiedzi składają się na końcową decyzję.

3.2.1 Drzewa klasyfikacyjne i regresyjne

Algorytm CART [20] stosowany do problemu klasyfikacji w celu zbudowania drzewa decyzyjnego dokonuje podziału danych treningowych na grupy, uwzględniając ich jednorodność. W celu wybrania odpowiedniego kryterium podziału stosowane są w nim miary takie jak indeks Gini czy

entropia. Każdy podział powoduje powstanie nowego wierzchołka w modelu. Jest on opisany warunkiem podziału postaci:

$$wartosc\_na\_kryterium \leq wartosc\_podzialu.$$

Generacja drzewa jest procesem rekurencyjnym – w każdym wierzchołku, począwszy od korzenia, uruchamiany jest mechanizm, pozwalający na wprowadzenie nowego podziału. Algorytm kończy swoje działanie, gdy spełnione są warunki stopu, wśród których uwzględniane są kryteria takie jak:

- w zbiorze, który ma zostać podzielony, wszystkie obiekty mają przypisaną tą samą klasę decyzyjną,
- liczba elementów w zbiorze do podziału nie spełnia predefiniowanego warunku (na przykład podziałowi ulegają tylko zbiory o liczności większej niż 3),
- głębokość dotychczas zbudowanego drzewa osiąga maksymalną założoną wartość,
- w zbiorze do podziału występuje mniej niż założona liczba różnych klas, do których przypisane są przykłady,
- podział zbioru w żaden możliwy sposób nie powoduje osiągnięcia większej niż założony próg wartości miary oceniającej poszczególne podziały.

Algorytm podzielony jest na dwa etapy:

- budowa drzewa polega na wykonywaniu rekurencyjnych podziałów, dopóki nie są spełnione warunki stopu, jej pseudokod jest przedstawiony w Algorytmie 5.

Algorithm 5 Procedura budowy drzewa CART

function GROWING_TREE(zbiór treningowy X, zbiór atrybutów A, wyjściowa zmienna y)
  zainicjuj drzewa T korzeniem
  if jeśli wszystkie kryteria stopu są spełnione **then**
    drzewo T ma jeden wierzchołek z najczęściej występującą klasą w X jako etykietą
  else
    znajdź najlepsze kryterium $a \in A$ z wykorzystaniem funkcji oceny podziału
    otaguj wierzchołek przy wykorzystaniu zbudowanego warunku: $x_{i,a} \leq v$
    for $\{x_{i,a} \leq v, x_{i,a} > v\}$ **do**
      $X_i \leftarrow \{x_i \in X$, które spełniają rozważany w iteracji warunek$\}$
      $A_i \leftarrow A - a$
      $growing\_tree(X_i, A_i, y)$
      połącz nowy wierzchołek z wierzchołkiem otagowanym przez $x_{i,a} \leq v$
    end for
  end if
  return pruningtree(X, A, y)
end function

- procedura przycinania zbudowanego drzewa, w celu otrzymania optymalnego modelu, jej pseudokod jest przedstawiony w Algorytmie 6.

Algorithm 6 Procedura przycinania drzewa CART

function PRUNINGTREE(zbiór treningowy X, zbiór atrybutów A, wyjściowa zmienna y)
  $T_1 \leftarrow T(0)$
  $a_1 \leftarrow 0$
  $k \leftarrow 1$
  while T_k ma przynajmniej jeden wierzchołek **do**
    for wszystkich wierzchołków nieterminalnych $t \in T_k$ **do**
      $g_k(T) = \frac{R(T)-R(T_{k,t})}{L(T_{k,t})-1}$
    end for
    $a_{k+1} = min_t g_k(t)$
    if $g_k(t) = a_{k+1}$ **then**
      przytnij drzewo w t aby otrzymać T_{k+1}
    end if
    $k \leftarrow k + 1$
  end while
  return przycięte drzewo decyzyjne
end function

Do przycinania wykorzystana jest miara minimalnego kosztu-złożoności. Każdemu poddrzewu przypisujemy koszt:

$$C_a(T) = R(T) + aL(T)$$

gdzie $R(T)$ to ułamek błędnie sklasyfikowanych przykładów w drzewie T, $L(T)$ to liczba liści w drzewie T, zaś $a \geq 0$ to stała określająca jak bardzo karane są złożone drzewa.

Jak już wcześniej zostało zaznaczone, do wyboru najlepszego podziału może być wykorzystany indeks Gini, który w wierzchołku t jest zdefiniowany jako:

$$Gini(t) = 1 - \sum_{k=0}^{K} \left(\frac{n(k|t)}{n(t)} \right)^2$$

gdzie: k oznacza klasę, $n(k|t)$ określa liczbę przykładów w wierzchołku t, które mają przypisaną klasę k, zaś $n(t)$ liczbę wszystkich przykładów w wierzchołku. Ostatecznie, do wybrania najlepszego podziału stosowana jest miara:

$$Gini(t)_{split} = \frac{n(t_L)}{n(t)} Gini(t_L) + \frac{n(t_R)}{n(t)} Gini(t_R)$$

gdzie t_L i t_R to odpowiednio lewe i prawe dziecko wierzchołka t. Najlepszy podział to taki, dla którego opisana miara osiąga najmniejszą wartość.

Druga z miar to entropia. Do oceny stosowana jest następująca jej postać:

$$H(t)_{split} = -\frac{n(t_L)}{n(t)} \sum_{k=0}^{K} \left(\frac{n(k|t_L)}{n(t_L)} log(\frac{n(k|t)}{n(t)}) \right) - \frac{n(t_R)}{n(t)} \sum_{k=0}^{K} \left(\frac{n(k|t_R)}{n(t)} log(\frac{n(k|t)}{n(t_L)}) \right)$$

która również podlega minimalizacji.

Przykład z pierwszego zbioru danych

Rozpatrzmy dokładnie taki sam zbór treningowy jak dla poprzedniego podrozdziału, czyli taki, w którym do treningu użyte są wszystkie warianty z pierwszego zbioru danych z wyjątkiem a3, który posłuży do testów. Przy wykorzystaniu modelu DecisionTreeClassifier [14] z biblioteki sklearn [16] i korzystaniu z kryterium wyboru najlepszego podziału indeks Gini otrzymano drzewo decyzyjne przedstawione na Rysunku 3.1. Zaznaczone są na nim warunki podziału, najlepsza wartość indeksu Gini, liczba przykładów oraz wartości na wybranym kryterium w danym wierzchołku.

Głębokość zbudowanego drzewa wynosi 6. Istnieje 12 możliwych ścieżek, z których 7 kończy się w liściach sugerujących przypisanie klasy 3, 3 – klasy 2 i po jednej klas odpowiednio 4 i 1. Do podziału wykorzystane są wszystkie kryteria z wyjątkiem tego, które opisuje jakość i pochodzenie spermy. W korzeniu obserwacje różnicowane są w oparciu o wiek kobiety.

Klasyfikacja testowego przykładu a3 z pierwszego zbioru danych przebiega ścieżką, w której począwszy od korzenia wędrujemy do lewego wierzchołka ($oocytes \leq 4.5$), następnie do prawego ($woman\_eval \leq 3.5$) i do lewego liścia, w którym przykładowi przypisana zostaje zgodna z oryginalną klasa 3.

Rysunek 3.1. Drzewo decyzyjne CART dla pierwszego zbioru danych bez wariantu a3, indeks Gini jako miara podziału

3.2.2 Losowy las

Algorytm budowy losowego lasu [12] został przedstawiony w formie pseudokodu w Algorytmie 7. Działanie algorytmu opiera się na budowie z góry określonej liczby drzew, w taki sposób, że zbiorem treningowym dla każdego z nich jest losowa próbka danych z oryginalnego zbioru treningowego. Sam proces budowy drzewa w stosunku do tradycyjnego algorytmu ulega nieznacznej modyfikacji – zamiast szukać globalnie najlepszego podziału ustalamy losowy, mały zbiór kryteriów i dopiero z niego wybieramy najlepszy podział, co znacząco ogranicza czas poszukiwania kryterium podziału.

Algorithm 7 Pseudokod algorytmu losowego lasu

```
function RANDOMFOREST(zbiór treningowy X, zbiór atrybutów A, liczba składowych drzew B)
    H ← 0
    for i ∈ 1, .., B do
        X^(i) ← niezależna próbka przykładów z X
        h_i ← randomized_tree_learn(X^(i),A)
        H ← H ∪ {h_i}
    end for
    return H
end function

function RANDOMIZED_TREE_LEARN(zbiór treningowy X, zbiór atrybutów A)
    for każdego wierzchołka do
        a ← mały podbiór zbioru A
        dokonaj podziału na podstawie najlepszej, zgodnie z kryterium podziału, cechy w a
    end for
    return zbudowane drzewo
end function
```

Jeśli chodzi o dokonywanie klasyfikacji dla nowych przykładów, to proces jest realizowany w następujących krokach:

1. dla każdego drzewa znajdź rekomendowaną przez nie klasę dla przykładu,
2. sprawdź, która z decyzji jest rekomendowana przez największą liczbę drzew,
3. przypisz powyższą jako wynik klasyfikacji.

Przykład z pierwszego zbioru danych

Analogicznie jak dla przykładu dla CART dane uczące stanowią oryginalny zbiór z wyłączeniem wariantu a3, który posłuży do testów. Wizualizacja przykładu została zrealizowana z wykorzystaniem modelu RandomForestClassifier [15] również z biblioteki sklearn [16]. Wybrana została miara indeks Gini oraz ograniczenie do budowy zespołu składającego się z trzech drzew. Na Rysunkach 3.2, 3.3 oraz 3.4 zaprezentowane zostały kolejne estymatory wchodzące w skład modelu.

Pierwsze drzewo (patrz Rysunek 3.2) ma głębokość 4 i pozwala na klasyfikację przykładów do klas 1 (w jednym liściu), 2 (w dwóch liściach) oraz 3 (w czterech liściach). Podział w korzeniu wykonywany jest na podstawie kryterium $infertility$, zaś pozostałe wykorzystane kryteria to *age*, *woman_eval* i *oocytes*. W fazie budowy drzewa pracowano na 29 przykładach ze zbioru danych.

Drugi estymator (patrz Rysunek 3.3) ma głębokość 5 i zawiera 8 liści, z których 5 sugeruje przypisanie obiektom klasy 3, 2 – klasy 2 i jeden klasy 4. Wiek kobiety jest kryterium podziału dokonywanego w korzeniu. Pozostałe użyte w tym drzewie kryteria to *develop_quality*, *woman_eval*, *oocytes*, *morpho_quality*, *sperm* oraz $infertility$, czyli wykorzystane zostały wszystkie dostępne atrybuty. Estymator został zbudowany na podstawie próbki 34 przykładów ze zbioru danych.

Ostatnie, trzecie, drzewo (patrz Rysunek 3.4) ma głębokość równą 5. Pozwala na klasyfikację przykładów do wszystkich możliwych klas, odpowiednio:

- 4 liście sugerują przypisanie klasy 3,
- 3 liście sugerują przypisanie klasy 2,
- 1 liść sugeruje przypisanie klasy 4,
- 1 liść sugeruje przypisanie klasy 1.

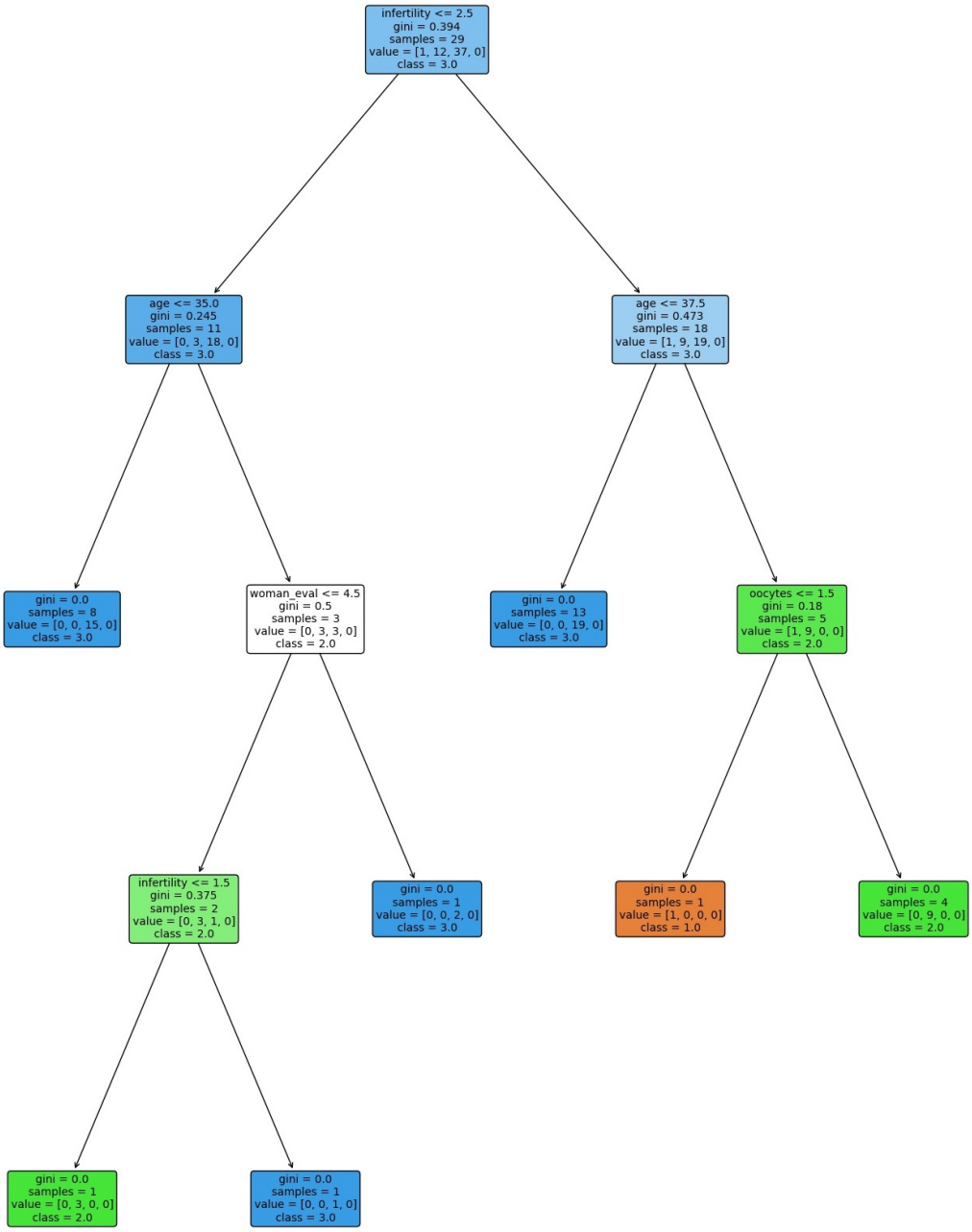
Ten estymator dokonuje podziału w korzeniu na podstawie wartości na kryterium *morpho_quality*, pozostałe użyte kryteria to: $infertility$, *sperm*, *oocytes*, *woman_eval* oraz *age*. Faza budowy drzewa zrealizowana została z wykorzystaniem 29 przykładów ze zbioru danych.

Patrząc na cały zespół, poszczególne estymatory wykorzystują różne kryteria podziału, co więcej – tylko drugi wykorzystał każdy możliwy atrybut. Powstałe drzewa mają zbliżone głębokości. Każde z nich może sugerować przypisanie do innych zestawów klas, tylko ostatnie uwzględnia pełny wachlarz możliwości. Pierwsze drzewo jest najbardziej zbalansowanym klasyfikatorem. Drugie i trzecie mają bardziej rozbudowane lewe poddrzewa (patrząc od korzenia).

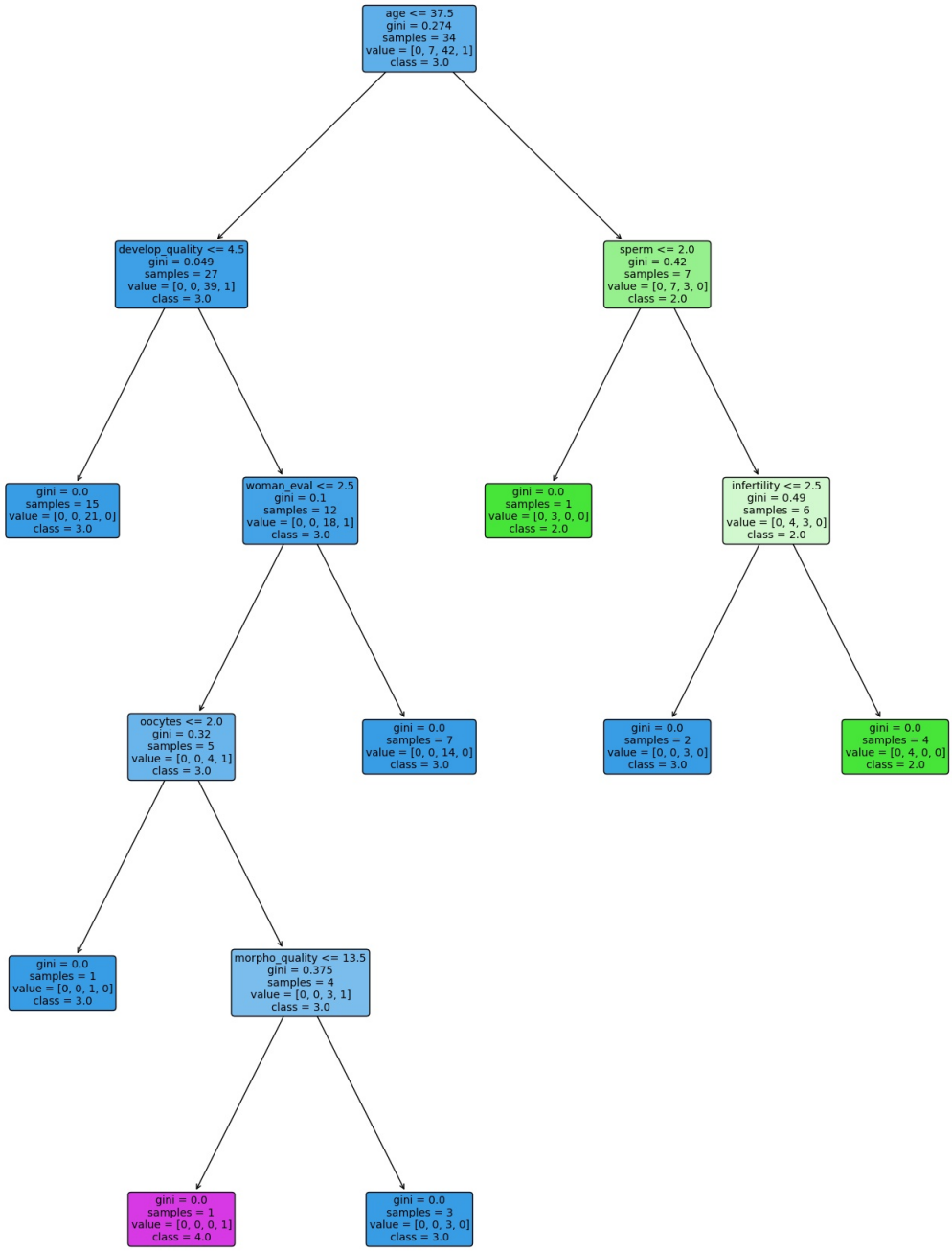
Dla testowanego przykładu kolejne estymatory (patrz Rysunek 3.2, 3.3 oraz 3.4) przydzielają następujące klasy:

- estymator 1: klasa 2
  Otrzymana w wyniku przejścia do lewego wierzchołka
  $$age \leq 35.0$$
  następnie prawego
  $$woman\_eval \leq 4.5$$
  i lewego
  $$infertility \leq 1.5$$
  oraz na koniec do lewego liścia.
- estymator 2: klasa 3
  Jest wynikiem przejścia do prawego wierzchołka
  $$sperm \leq 2.0$$
  następnie prawego
  $$infertility \leq 2.5$$
  i lewego liścia.
- estymator 3: klasa 3
  Otrzymana na ścieżce do prawego wierzchołka
  $$woman\_eval \leq 3.5$$
  i następnie prawego liścia.

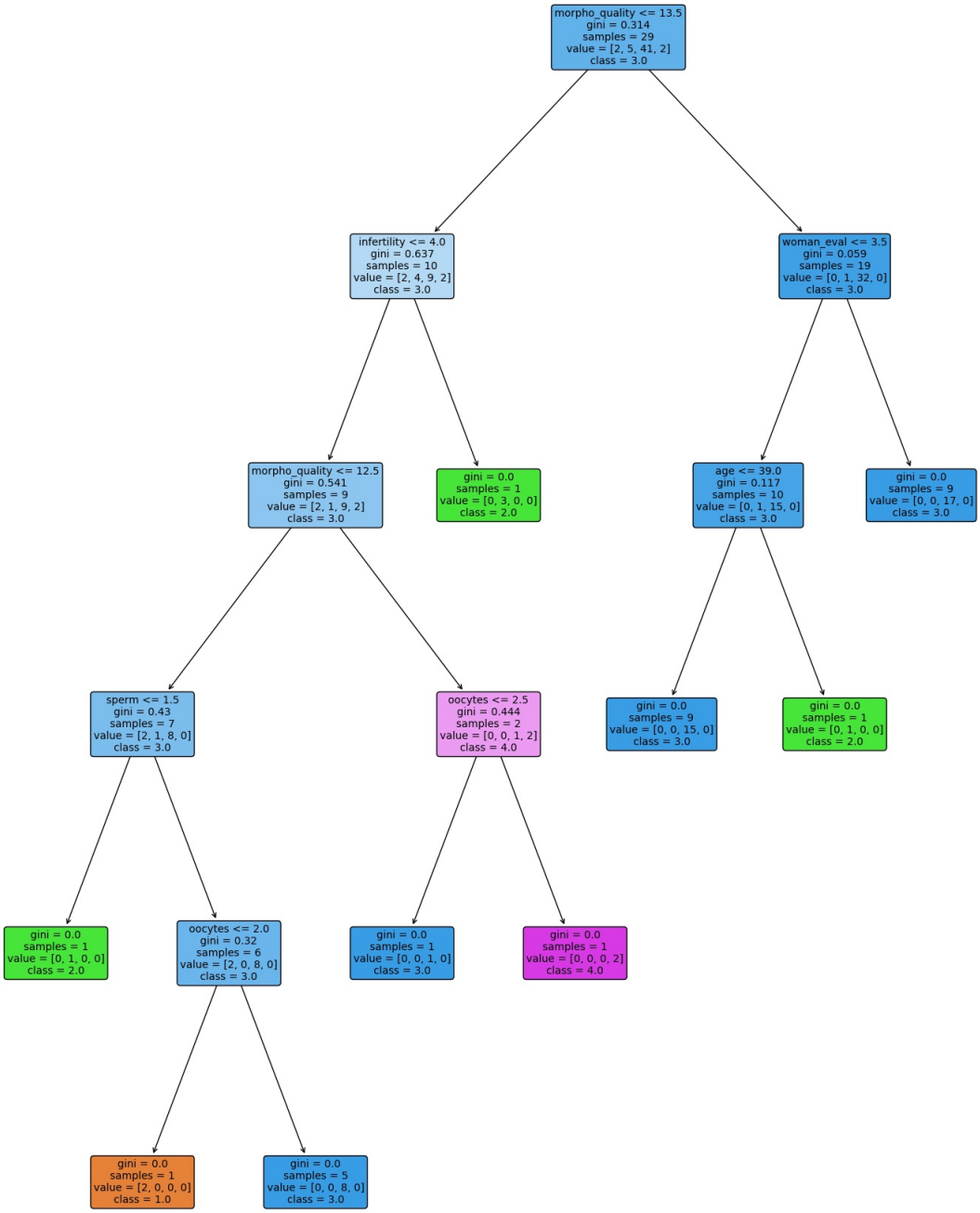
Ostatecznie predykcja klasy 3 została przypisana przez 2 z 3 drzew składowych, zatem wariant a3 ma przypisaną poprawną klasę 3.



Rysunek 3.2. Losowy las - pierwsze drzewo dla zbioru pierwszego bez wariantu a3, indeks Gini jako miara podziału, 3 estymatory



Rysunek 3.3. Losowy las - drugie drzewo dla zbioru pierwszego bez wariantu a3, indeks Gini jako miara podziału, 3 estymatory



Rysunek 3.4. Losowy las - trzecie drzewo dla zbioru pierwszego bez wariantu a3, indeks Gini jako miara podziału, 3 estymatory

Rozdział 4

Rezultaty

Opisane w poprzednim rozdziale metody zostały wykorzystane do przygotowania kilku modeli. Modele drzew CART i losowy las były implementowane przy wykorzystaniu biblioteki sklearn [16], zaś pozostałe metody zaimplementowano samodzielnie przy wykorzystaniu języka Python. Repozytorium kodów źródłowych oraz pełne wyniki umieszczone są w załącznikach do pracy. Przetestowano następujące scenariusze eksperymentów:

- wykorzystanie metody DRSA, algorytmu DOMLEM i prostego algorytmu klasyfikacji,
- wykorzystanie metody VC-DRSA, algorytmu DOMLEM i prostego algorytmu klasyfikacji,
- wykorzystanie metody DRSA, algorytmu DOMLEM i zaawansowanego algorytmu klasyfikacji,
- wykorzystanie metody VC-DRSA, algorytmu DOMLEM i zaawansowanego algorytmu klasyfikacji,
- wykorzystanie metody DRSA, algorytmu DOMApriori i prostego algorytmu klasyfikacji,
- wykorzystanie metody VC-DRSA, algorytmu DOMApriori i prostego algorytmu klasyfikacji,
- wykorzystanie metody DRSA, algorytmu DOMApriori i zaawansowanego algorytmu klasyfikacji,
- wykorzystanie metody VC-DRSA, algorytmu DOMApriori i zaawansowanego algorytmu klasyfikacji,
- wykorzystanie drzew CART,
- wykorzystanie losowego lasu.

Przy wykorzystaniu metody DRSA nie było potrzeby doboru wartości parametrów. Dla VC-DRSA należało dobrać wartość poziomu spójności – testowane były liczby z przedziału od 0.05 do 0.95 włącznie, z krokiem co 0.05. Wartość 1 nie była testowana ze względu na fakt, iż dla poziomu spójności równego 1 metoda VC-DRSA i DRSA otrzymują zawsze te same wyniki. Dodatkowo, testowanie odbywało się z krokiem co 0.05 dlatego, że zmiany w otrzymywanych wynikach dla tak małych zbiorów nie są znaczące dla małych zmian wartości parametru. W przypadku algorytmu DOMApriori koniecznym było ustalenie minimalnego wsparcia reguły i maksymalnej jej długości. Ostatecznie, wymagano wsparcia większego bądź równego 1 ze względu na fakt, iż dla większych wartości minimalnego wsparcia nie były pokrywane wszystkie obiekty. Maksymalną

długość reguły ustalono na 3, gdyż dłuższe reguły pokrywały już tylko obiekty dotychczas pokryte oraz znacząco wpływały na wydłużenie czasu działania algorytmu. Dla DOMLEM oraz algorytmów klasyfikacji nie ustalano parametrów.

Dla drzew sprawdzano entropię i indeks Gini dla szukania najlepszego podziału oraz uwzględnianie bądź nie wagi klas w algorytmie – tę możliwość stwarza implementacja algorytmu CART z biblioteki sklearn. Natomiast dla losowego lasu, oprócz elementów wymienionych dla drzew, dobierano także liczbę estymatorów w zespole, wartość parametru a określającego karę dla złożonych struktur w procedurze przycinania zbudowanych drzew oraz czy budując poszczególne estymatory dobierać niezależne próbki przykładów ze zbioru danych czy pracować na pełnym zbiorze.

Dla oceny poszczególnych modeli przeprowadzono walidacje krzyżową. Posłużono się metodą Leave-one-out [6], która polega na zbudowaniu tylu modeli ile obserwacji w zbiorze n, z których każdy uczony jest na $n-1$ wariantach, a testowany na jednym, niewykorzystanym do budowy modelu, przykładzie.

W Podrozdziałach 4.1 oraz 4.2 przestawione są ogólnie otrzymane wyniki wraz z dokładnym opisem rezultatów osiągniętych przez najlepsze zbudowane modele.

4.1 Pierwszy zbiór danych

W Tabeli 4.1 zebrano wartości trafności oraz miary F1-score dla najlepszych wersji, budowanych na pierwszym zbiorze danych, modeli DRSA i VC-DRSA. Przy wykorzystaniu algorytmu DOMApriori minimalne wsparcie zostało ustalone na 1 a maksymalna długość reguły na 3.

Na podstawie Tabeli 4.1 modele DRSA osiągały trafność klasyfikacji z przedziału $0.59-0.76$. Wartość 0.76 osiągnięto dla modelu korzystającego z algorytmu DOMApriori i zaawansowanej metody klasyfikacji. Oznacza ona, że średnia trafność dla 51 zbudowanych modeli – przy wykorzystaniu walidacji metodą Leave-one-out – plasuje się na poziomie nieco wyższym niż 75%, około 3 na 4 modele dokonywały poprawnej klasyfikacji testowego wariantu. Można też zauważyć, że niewielki wpływ na wartość tej miary ma wybór algorytmu indukcji reguł decyzyjnych, zaś ważną role odgrywa metoda klasyfikacji – dla zaawansowanego algorytmu rezultaty są istotnie lepsze (prosty algorytm klasyfikacji: $0.59-0.63$, zaawansowany: $0.75-0.76$, zatem trafność wyższa o około 15%).

Dla VC-DRSA osiągnięte wartości miary trafności były na zbliżonym poziomie ($0.75-0.78$) – najsłabiej wypadła wersja z algorytmem DOMLEM i prostą metodą klasyfikacji. Eksperymenty walidacyjne pokazały, że takie modele klasyfikowały poprawnie średnio 75% przykładów, czyli, w odniesieniu do Leave-one-out, co czwarty zbudowany model przypisywał niepoprawną klasę testowemu przykładowi. Warto zauważyć, że w tym przypadku wybrany algorytm klasyfikacji nie miał takiego znaczenia jak dla DRSA. Istotna była natomiast metoda indukcji reguł decyzyjnych – dla algorytmu DOMLEM wysoka trafność była osiągana kosztem bardzo niskiej spójności – nie jest to pożądane. Dopuszczenie dużej niespójności skutkuje powstaniem modelu zbyt uniwersalnego, za mało dopasowanego do problemu.

W każdym z ujętych w Tabeli 4.1 modeli F1-score osiąga wartość 0, co podkreśla sygnalizowany wcześniej problem zbioru danych - występuje w nim zdecydowanie za mało przykładów dla dwóch z czterech klas decyzyjnych.

Tablica 4.1: Wartości trafności i F1-score dla pierwszego zbioru danych - modele DRSA

Model	Algorytm ind. reguł	Algorytm klasyf.	Poziom spójności	Trafność	F1-score
DRSA	DOMLEM	prosty	-	0.59	0.0
VC-DRSA	DOMLEM	prosty	0.45	0.75	0.0
DRSA	DOMLEM	zaawansowany	-	0.75	0.0
VC-DRSA	DOMLEM	zaawansowany	0.25	0.78	0.0
DRSA	DOMApriori	prosty	-	0.63	0.0
VC-DRSA	DOMApriori	prosty	0.9	0.78	0.0
DRSA	DOMApriori	zaawansowany	-	0.76	0.0
VC-DRSA	DOMApriori	zaawansowany	0.9	0.78	0.0

Tabela 4.2 zawiera wartości trafności i miary F1-score dla modeli drzew i lasów. Przedstawione są także parametry poszczególnych wersji algorytmów. Dla wagi klas subbalans określanie wag klas przygotowywane jest na podstawie przykładów z próbki dla danego estymatora.

Wśród drzew CART najlepszy wynik osiągnięto nie stosując ważenia klas i korzystając z entropii jako miary do oceny podziału. Dla tej wersji 84% wszystkich zbudowanych modeli poprawnie zaklasyfikowało testowy przykład. Ogółnie dla drzew CART trafność klasyfikacji mieści się w przedziale $0.67 - 0.84$.

Najlepsze losowe lasy to te, które wykorzystywały entropię jako miarę podziału, nie wykorzystywały ważenia klas oraz składały się odpowiednio z 3 estymatorów. Osiągnęły one 86% trafności klasyfikacji, co oznacza że spośród 51 modeli walidacyjnych 44 poprawnie zaklasyfikowały testowy przykład. Rozważane były także zespoły o większej liczbie drzew, jednak nie sprawdzały się one lepiej niż te zaraportowane w Tabeli 4.2. Jeśli chodzi o ważenie klas dla tego zbioru najlepiej sprawdza się pominięcie go. Natomiast lepszą miarą oceny podziału okazała się w tym przypadku entropia – średnio o 2% wyższe wyniki niż dla indeksu Gini. Ogólnie trafność klasyfikacji dla losowych lasów mieściła się w przedziale $0.73 - 0.86$.

Miara F1-score osiąga, zarówno dla drzew CART jak i losowych lasów, zbliżone wartości do trafności klasyfikacji.

Tablica 4.2: Wartości trafności i F1-score dla pierwszego zbioru danych - modele drzew

Model	Waga klas	Miara oceny podziału	Liczba estymatorów	Trafność	F1-score
drzewa CART	niestosowana	indeks Gini	-	0.82	0.82
drzewa CART	niestosowana	entropia	-	0.84	0.84
drzewa CART	zbalansowana	indeks Gini	-	0.73	0.73
drzewa CART	zbalansowana	entropia	-	0.67	0.67
losowy las	niestosowana	indeks Gini	2	0.76	0.76
losowy las	niestosowana	entropia	2	0.78	0.78
losowy las	zbalansowana	indeks Gini	2	0.73	0.73
losowy las	zbalansowana	entropia	2	0.73	0.73
losowy las	subbalans	indeks Gini	2	0.73	0.73
losowy las	subbalans	entropia	2	0.73	0.73
losowy las	niestosowana	indeks Gini	3	0.78	0.78
losowy las	niestosowana	entropia	3	0.86	0.86
losowy las	zbalansowana	indeks Gini	3	0.76	0.76
losowy las	zbalansowana	entropia	3	0.82	0.82
losowy las	subbalans	indeks Gini	3	0.76	0.76
losowy las	subbalans	entropia	3	0.76	0.76

Porównując modele DRSA i drzewiaste lepiej sprawdziły się te drugie, jednak różnice nie są tu znaczące, jeśli chodzi o wartości trafności. Dla miary F1-score zdecydowanie lepsze rezultaty osiągane są przez drzewa i lasy – o około 70%. Mimo problematycznego rozkładu obserwacji z poszczególnych klas, przygotowane modele osiągają dość wysoką wartość obu raportowanych miar.

Porównując DRSA i VC-DRSA można stwierdzić, iż wprowadzenie obniżonego poziomu spójności pozytywnie wpływa na skuteczność działania modelu, natomiast drzewa CART i losowy las dla tego zbioru danych sprawdzają się podobnie.

W Tabeli 4.3 zebrano dane na temat reguł wyindukowanych przez rozważane wersje metody DRSA - wartości długości, wsparcia, pokrycia i siły to dane uśrednione po zbiorze reguł dla poszczególnych modeli.

Analizując dane w Tabeli 4.3 można zauważyć, że reguły indukowane przy wykorzystaniu algorytmu DOMLEM są średnio krótsze od tych, które zbudowane zostały z wykorzystaniem DOMA-

priori, co jest naturalną konsekwencją tego, jakie typy zbiorów reguł powstają przy wykorzystaniu tych metod. Ponadto dla DOMLEM powstałe reguły cechuje wyższe średnie wsparcie, pokrycie i siła – co również jest następstwem indukcji minimalnego zbioru reguł w tym przypadku. Największą różnicę możemy zaobserwować w liczbie powstałych reguł – dla DOMApriori jest ich kilkakrotnie więcej.

Tablica 4.3: Statystyki reguł dla pierwszego zbioru danych

Model	Długość	Wsparcie	Pokrycie	Siła	Liczba reguł
DRSA, DOMLEM, prosty	1.65	18.41	0.53	0.36	17
VC-DRSA, DOMLEM, prosty, 0.45	1.1	24.1	0.87	0.47	10
DRSA, DOMLEM, zaawansowany	1.65	18.41	0.53	0.36	17
VC-DRSA, DOMLEM, zaawansowany, 0.25	1.18	22.27	0.81	0.44	11
DRSA, DOMApriori, prosty	2.07	12.6	0.37	0.25	73
VC-DRSA, DOMApriori, prosty, 0.9	2.25	23.57	0.59	0.46	186
DRSA, DOMApriori, zaawansowany	2.07	12.6	0.37	0.25	73
VC-DRSA, DOMApriori, zaawansowany, 0.9	2.45	22.31	0.57	0.44	176

Warto teraz dokładniej przyjrzeć się kilku najlepszym przygotowanym modelom. Dla metody DRSA wybrany został wariant dla algorytmu DOMLEM i złożonej metody klasyfikacji. Osiąga on trafność na poziomie 75%. Jest to wynik bliski najlepszemu osiągniętemu dla tej metody. Taki wybór podyktowany jest zdecydowanie mniejszą liczbą reguł niż dla wersji korzystającej z DOMApriori. W Tabeli 4.4 przedstawiono reguły decyzyjne dla tego modelu. Nie pojawiły się wśród nich reguły opisujące obiekty należące do klasy *co najmniej 4*. Wystąpiła jedna reguła opisująca warianty, dla których przypisana klasa to co najwyżej 1. Warunki często budowane są w oparciu o atrybut wieku kobiety, natomiast tylko jedna reguła wykorzystuje atrybut pochodzenie nasienia. Kluczowymi atrybutami są niepłodność i oceny morfologiczna oraz rozwoju dla embrionów. Wśród wyindukowanych reguł pojawiło się 10 reguł składających się z jednego warunku. Warto też zauważyć, że najlepiej opisanymi uniami klas są: *co najmniej klasa 3* oraz *co najwyżej klasa 3*, co wynika z niezbalansowania obserwacji z poszczególnych klas w zbiorze - najwięcej obiektów ma przypisaną klasę 3. Rozpatrując przykładową regułę:

$$(age \leq 34.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3),$$

która kobietom mającym co najwyżej 34 lata i embrionom uzyskującym z punktu widzenia morfologicznego ocenę co najmniej 10 przypisuje klasę 3 lub 4, można zauważyć, że im młodsza jest pacjentka i im wyższa wartość oceny budowy embrionów tym przypisana zostaje lepsza klasa. Jest to zgodne z intuicją specjalistów.

Tablica 4.4: Reguły wyindukowane dla metody DRSA przy wykorzystaniu algorytmu DOMLEM dla pierwszego zbioru danych

Oznaczenie	Reguła
Reguła 1	$(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \wedge (infertility \geq 3.0) \implies (class \leq 1)$
Reguła 2	$(age \geq 39.0) \wedge (infertility \geq 3.0) \implies (class \leq 2)$
Reguła 3	$(morpho\_quality \leq 9.0) \wedge (age \geq 33.0) \wedge (infertility \geq 5.0) \implies (class \leq 2)$
Reguła 4	$(develop\_quality \leq 4.0) \implies (class \leq 3)$
Reguła 5	$(infertility \geq 4.0) \implies (class \leq 3)$
Reguła 6	$(oocytes \geq 4.0) \implies (class \leq 3)$
Reguła 7	$(age \geq 35.0) \implies (class \leq 3)$
Reguła 8	$(sperm \leq 2.0) \implies (class \leq 3)$
Reguła 9	$(age \leq 34.0) \wedge (morpho\_quality \geq 10.0) \implies (class \geq 3)$
Reguła 10	$(morpho\_quality \geq 15.0) \wedge (oocytes \leq 5.0) \implies (class \geq 3)$
Reguła 11	$(age \leq 36.0) \wedge (infertility \leq 3.0) \wedge (oocytes \leq 5.0) \implies (class \geq 3)$
Reguła 12	$(age \leq 32.0) \implies (class \geq 3)$
Reguła 13	$(infertility \leq 1.0) \wedge (develop\_quality \geq 4.0) \implies (class \geq 3)$
Reguła 14	$(infertility \leq 2.0) \wedge (oocytes \leq 5.0) \implies (class \geq 3)$
Reguła 15	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 16	$(woman\_eval \geq 3.0) \implies (class \geq 2)$
Reguła 17	$(infertility \leq 2.0) \implies (class \geq 2)$

Spośród modeli VC-DRSA do dalszej analizy, mimo niskiego poziomu spójności ($l = 0.25$), weźmy wersję dla DOMLEM i zaawansowanego algorytmu klasyfikacji. Taki wybór podyktowany jest zdecydowanie mniejszą liczbą reguł, niż dla pozostałych, równie dobrych jeśli chodzi o wartość trafności, wariantów. Zbiór reguł dla tej wersji, ujęty w Tabeli 4.5, podobnie jak poprzedni, nie zawiera reguł opisujących przypisanie do klasy *co najmniej 4*, a dominują w nim reguły dotyczące przybliżeń klasy 2 i 3, co odzwierciedla liczbę obiektów z poszczególnych klas w danych wejściowych. Warunki w regułach najczęściej dotyczą wieku kobiety – im jest ona młodsza tym proponowane są przypisania do lepszych klas. Dla 9 reguł pojawia się pojedynczy warunek, tylko 2 składają się z koniunkcji dwóch warunków, co związane jest z obniżonym poziomem spójności. Analizując przykładową regułę:

$$(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \implies (class \leq 1),$$

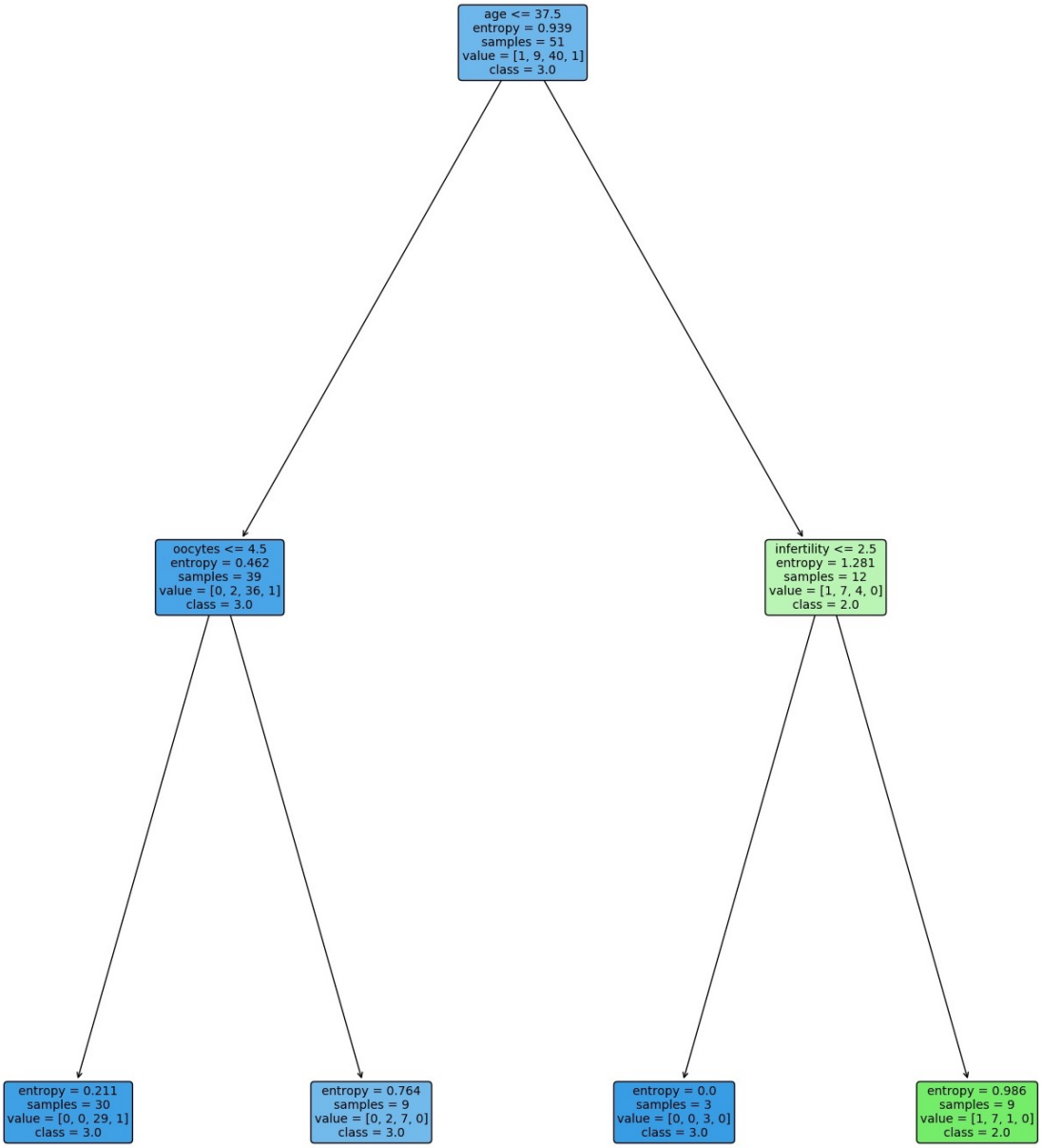
która kobietom w wieku 40 lat lub starszym i uzyskującym w ocenie specjalisty notę nie większą niż 2 przypisuje klasę 1, możemy zauważyć, że pacjentki starsze, słabo ocenione przez ginekologów czy położników klasyfikowane są do najsłabszej z rozważanych klas. Jest to zgodne z intuicyjnym rozważaniem. Warto też zauważyć, że w zbiorze reguł najmniej warunków dotyczy oceny embrionu - jest to różnica w stosunku do poprzedniego przykładu. Prawdopodobnie wynika ona z niskiego poziomu spójności, skutkującego także małą liczbą reguł w minimalnym zbiorze.

TABLICA 4.5: Reguły wyindukowane dla metody VC-DRSA, $l = 0.25$ przy wykorzystaniu algorytmu DOMLEM dla pierwszego zbioru danych

Oznaczenie	Reguła
Reguła 1	$(age \geq 40.0) \wedge (woman\_eval \leq 2.0) \implies (class \leq 1)$
Reguła 2	$(age \geq 38.0) \implies (class \leq 2)$
Reguła 3	$(oocytes \geq 6.0) \wedge (age \geq 36.0) \implies (class \leq 2)$
Reguła 4	$(morpho\_quality \leq 9.0) \implies (class \leq 2)$
Reguła 5	$(age \geq 27.0) \implies (class \leq 3)$
Reguła 6	$(sperm \leq 2.0) \implies (class \leq 3)$
Reguła 7	$(age \leq 38.0) \implies (class \geq 3)$
Reguła 8	$(infertility \leq 2.0) \implies (class \geq 3)$
Reguła 9	$(age \leq 39.0) \implies (class \geq 2)$
Reguła 10	$(woman\_eval \geq 3.0) \implies (class \geq 2)$
Reguła 11	$(infertility \leq 2.0) \implies (class \geq 2)$

Najlepszym zbudowanym drzewem CART jest dla tego zbioru wariant niewykorzystujący wagi klas i stosujący entropię jako miarę oceny podziału (Rysunek 4.1).

W drzewie (Rysunek 4.1) są 4 liście, z których 3 sugerują przypisanie obiektom klasy 3. Pozostały liść dotyczy klasy 2. Pierwszy podział dokonany jest za względu na wiek kobiety, co podkreśla istotność tego kryterium w procesie decyzyjnym. Kolejne dwa alternatywne podziały różnicują warianty ze względu na wartości na kryteriach oceniających odpowiednio liczbę oocytów i niepłodność kobiety. Warto też zauważyć, że drzewo mogłoby zostać uproszczone, gdyż kobietom mającym mniej niż 38 lat, niezależnie od kolejnej decyzji, zostanie przypisana klasa 3. Zaletą zbudowanego klasyfikatora jest niewątpliwie jego niska złożoność - jego głębokość wynosi 2.



Rysunek 4.1. Najlepsze zbudowane drzewo CART, entropia jako miara oceny podziału, dla pierwszego zbioru danych

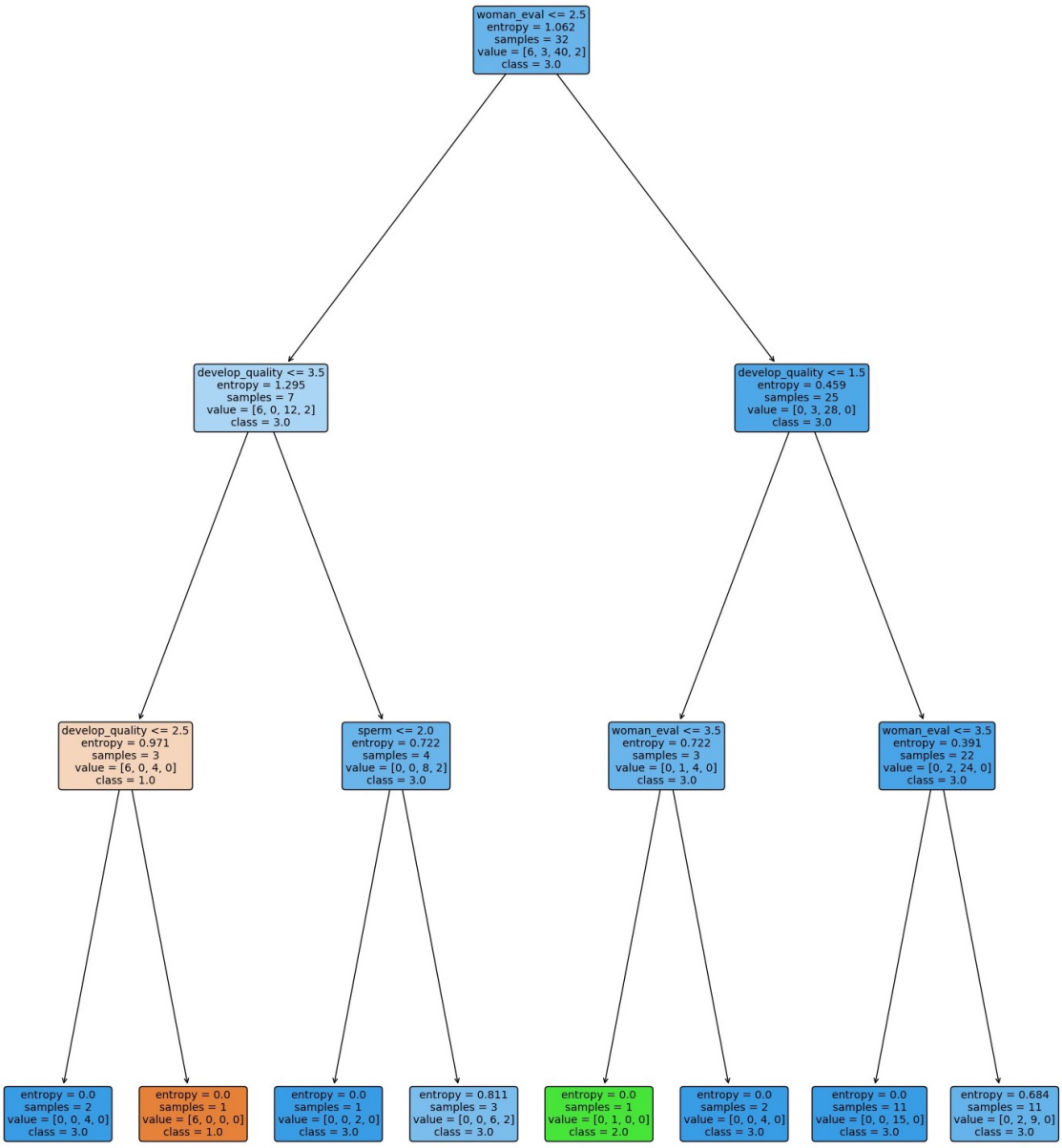
Wśród modeli lasów opisany zostanie wariant niestosujący wag klas, składający się z 3 drzew i aplikujący entropię do oceny podziału. Na Rysunkach 4.2, 4.3 oraz 4.4. zostały przedstawione drzewa wchodzące w skład zespołu.

Pierwszy z estymatorów (Rysunek 4.2) pracuje na 32 obiektach zbioru danych. Występuje w nim 8 liści, z których 6 sugeruje przypisanie obiektom klasy 3 oraz po jednym odpowiednio klas 1 i 2. Podziały dokonywane są w oparciu o atrybuty dotyczące ogólnej oceny kobiety przez specjalistę, pochodzenia spermy i oceny rozwoju embrionów. Jest to ciekawa losowa próbka kryteriów, gdyż dla wcześniej zaprezentowanych modeli nie były one istotne. Dodatkowo, zaproponowane drzewo możnaby uprościć: w wierzchołkach $sperm \leq 2$ oraz $woman\_eval \leq 3.5$ zarówno lewy jak i prawy liść przypisuje obiektom klasę 3. Zredukowałoby to także liczbę wykorzystanych kryteriów o atrybut podziału dotyczący pochodzenia spermy. Powstałe drzewo ma głębokość 3, jest więc bardziej złożone niż najlepszy wariant dla drzew CART (Rysunek 4.1).

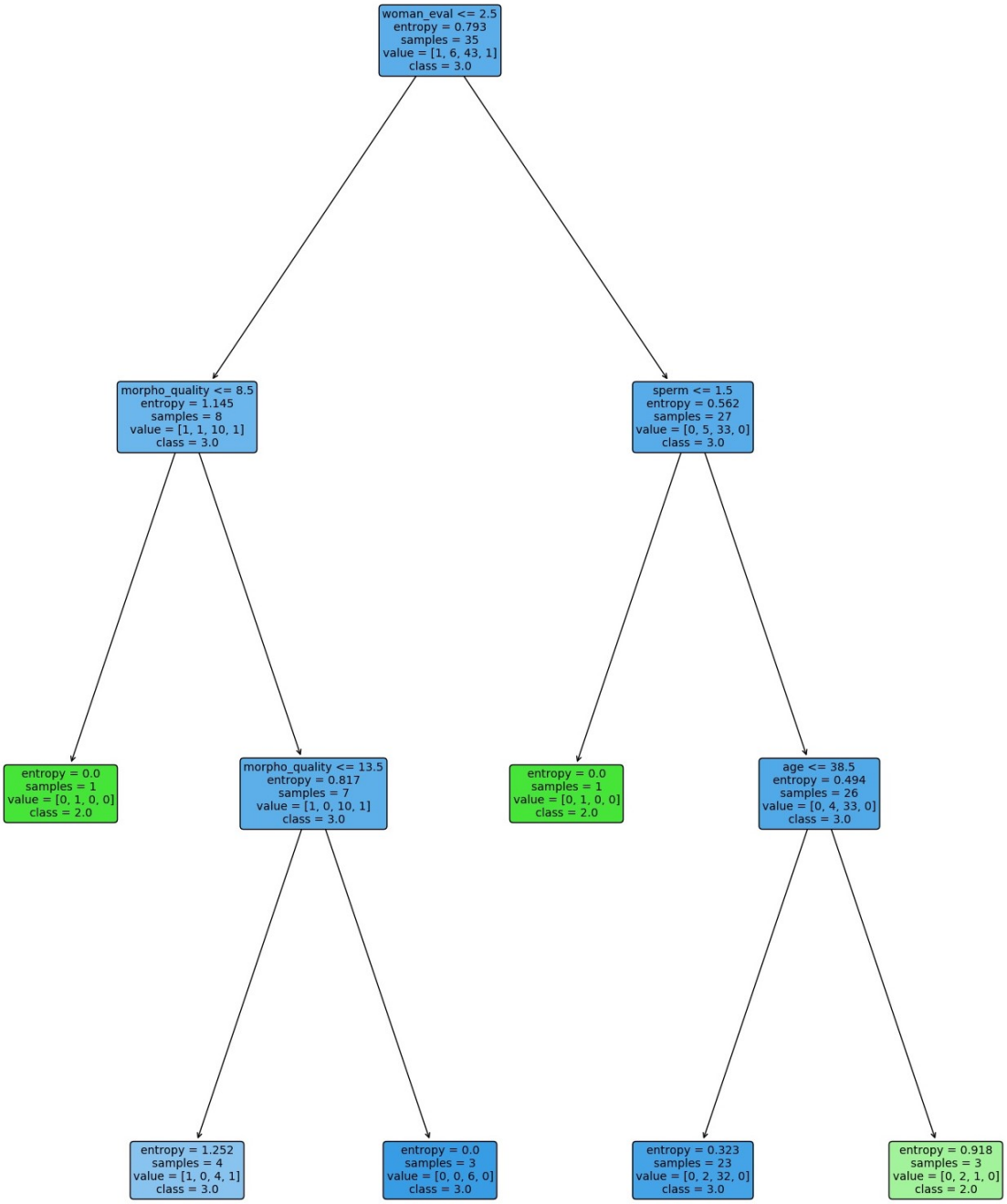
Drugi estymator (Rysunek 4.3) zawiera 6 liści, 3 z nich sugerują przypisanie klasy 3, a pozostałe – klasy 2. Drzewo zbudowane zostało na podstawie 35 przykładów ze zbioru. Wykorzystane kryteria opisują ocenę kobiety przez specjalistę, pochodzenie spermy, morfologiczną ocenę embrionów i wiek kobiety. Najistotniejszy podział dokonany jest, jak dla poprzedniego estymatora (Rysunek 4.2.), na podstawie oceny kobiety przez specjalistę. Zbudowane drzewo również można uprościć: w wierzchołku $morpho\_quality \leq 13.5$ zarówno w lewym jak i prawym liściu obiektom przypisywana jest klasa 3.

Ostatni, trzeci, estymator (Rysunek 4.4) pracuje na 34 obiektach. Składa się z 5 liści, z których 3 sugerują przypisanie klasy 3, a pozostałe – klasy 2. Użyte kryteria opisują pochodzenie spermy, morfologiczną i rozwojową ocenę embrionów oraz niepłodność pary. Zaproponowane drzewo mogłoby zostać uproszczone - w wierzchołku $develop\_quality \leq 3.5$ oba liście przypisują obiektom klasę 3. Głębokość drzewa jest taka sama jak w przypadku pozostałych estymatorów w zespole – wynosi 3.

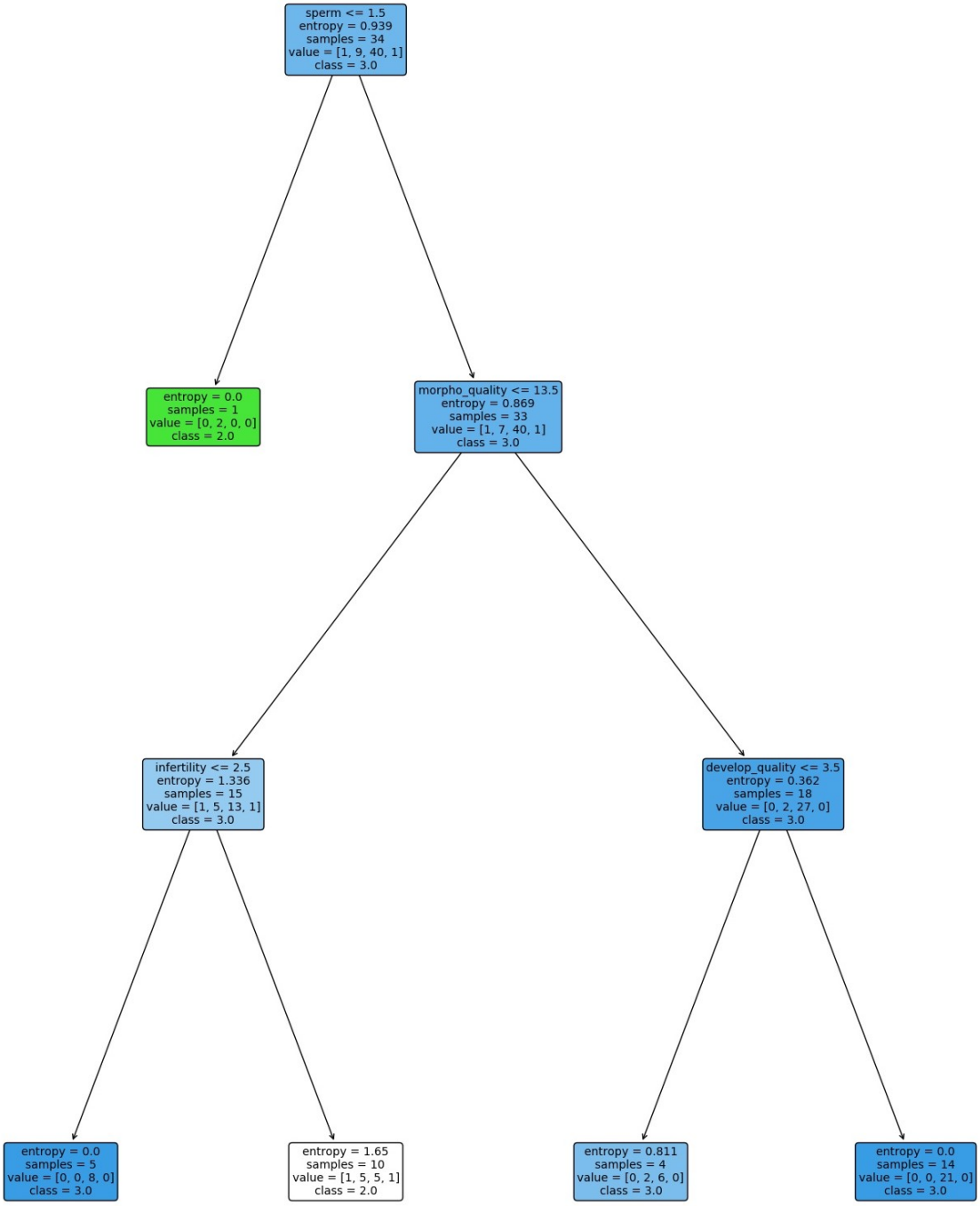
Cały zespół (Rysunki 4.2, 4.3 oraz 4.4) stanowi dość rozbudowaną i ciekawą propozycję klasyfikatora uwzględniającego wiele kryteriów. Pozwala na klasyfikowanie obiektów do trzech spośród 4 możliwych klas decyzyjnych.



Rysunek 4.2. Najlepszy zbudowany losowy las dla pierwszego zbioru danych, stosujący entropię jako miarę oceny podziału – pierwszy estymator



Rysunek 4.3. Najlepszy zbudowany losowy las dla pierwszego zbioru danych, stosujący entropię jako miarę oceny podziału – drugi estymator



Rysunek 4.4. Najlepszy zbudowany losowy las dla pierwszego zbioru danych, stosujący entropię jako miarę oceny podziału – trzeci estymator

4.2 Drugi zbiór danych

W Tabeli 4.6 zebrano wartości trafności oraz miary F1-score dla najlepszych wersji, budowanych na drugim zbiorze danych, modeli DRSA i VC-DRSA. Przy wykorzystaniu algorytmu DOMApriori, analogicznie jak dla poprzedniego zbioru, minimalne wsparcie zostało ustalone na 1 a maksymalna długość reguły na 3.

Wśród modeli DRSA (Tabela 4.6) najlepiej sprawdził się ten, w którym wykorzystano algorytm DOMApriori i zaawansowaną metodę klasyfikacji. Uzyskał 40-procentową trafność w eksperymentach walidacyjnych, jest ona jednak dość niska. Ogólnie wartości trafności dla metody DRSA mieściły się dla tego zbioru w przedziale $0.36 - 0.4$. Biorąc pod uwagę F1-score, najlepiej sprawdził się algorytm DOMLEM i prosta metoda klasyfikacji. Uzyskał wartość 0.48.

Dla VC-DRSA (Tabela 4.6) najwyższy wynik osiągnięty został dla zaawansowanego algorytmu klasyfikacji – 0.4. Wybór algorytmu indukcji reguł decyzyjnych nie wpływał na wartość trafności w tym przypadku, natomiast dla DOMLEM zanotowane zostały większe wartości F1-score (0.48 i 0.5). Trafność, podobnie jak dla DRSA jest niska, mieści się w przedziale $0.36 - 0.4$. Patrząc na F1-score, najlepszy model VC-DRSA korzystał z algorytmu DOMLEM i zaawansowanej metody klasyfikacji.

Porównując wyniki dla VC-DRSA i DRSA są one, zarówno jeśli chodzi o trafność jak i F1-score na zbliżonym poziomie. Dla tego zbioru obie metody uzyskały dość słabe rezultaty.

Tablica 4.6: Wartości trafności i F1-score dla drugiego zbioru danych - modele DRSA

Model	Algorytm ind. reguł	Algorytm klasyf.	Poziom spójności	Trafność	F1-score
DRSA	DOMLEM	prosty	-	0.36	0.48
VC-DRSA	DOMLEM	prosty	0.9	0.36	0.48
DRSA	DOMLEM	zaawansowany	-	0.36	0.45
VC-DRSA	DOMLEM	zaawansowany	0.8	0.4	0.5
DRSA	DOMApriori	prosty	-	0.36	0.45
VC-DRSA	DOMApriori	prosty	0.9	0.36	0.45
DRSA	DOMApriori	zaawansowany	-	0.4	0.46
VC-DRSA	DOMApriori	zaawansowany	0.9	0.4	0.46

W Tabeli 4.7 zawarte są wartości trafności i miary F1-score dla modeli drzew i lasów. Przedstawione są także parametry poszczególnych wersji algorytmów. Dla wagi klas subbalans określanie wag klas przygotowywane jest na podstawie przykładów z próbki dla danego estymatora.

Jeśli chodzi o drzewa CART, nie jest tu obserwowalna różnica w wartościach trafności i F1-score raportowanych w Tabeli 4.7. Co więcej, wartości dla obu miar są takie same i wynoszą 0.56.

Dla losowych lasów (Tabela 4.7) lepsze rezultaty osiągane są przez modele składające się z większej liczby estymatorów. Mieszczą się one w przedziale $0.68 - 0.76$. Ponadto dla lasów składających się z dwóch drzew lepszą miarą oceny podziału okazała się entropia (zysk około 4%), natomiast dla tych z trzech – indeks Gini (zysk około 4%). Dla większej niż 3 liczby estymatorów w zespole

nie był obserwowany wzrost wartości żadnej z raportowanych miar stąd dane te nie są zawarte w Tabeli 4.7.

Losowe lasy sprawdziły się znacząco lepiej niż drzewa CART. może być to związane z faktem wykorzystania w zespole kilku drzew, pracujących na różnych próbkach danych i kryteriach.

Tablica 4.7: Wartości trafności i F1-score dla drugiego zbioru danych - modele drzew

Model	Waga klas	Miara oceny podziału	Liczba estymatorów	Trafność	F1-score
drzewa CART	niestosowana	indeks Gini	-	0.56	0.56
drzewa CART	niestosowana	entropia	-	0.56	0.56
drzewa CART	zbalansowana	indeks Gini	-	0.56	0.56
drzewa CART	zbalansowana	entropia	-	0.56	0.56
losowy las	niestosowana	indeks Gini	2	0.68	0.68
losowy las	niestosowana	entropia	2	0.76	0.76
losowy las	zbalansowana	indeks Gini	2	0.64	0.64
losowy las	zbalansowana	entropia	2	0.64	0.64
losowy las	subbalans	indeks Gini	2	0.68	0.68
losowy las	subbalans	entropia	2	0.76	0.76
losowy las	niestosowana	indeks Gini	3	0.76	0.76
losowy las	niestosowana	entropia	3	0.76	0.76
losowy las	zbalansowana	indeks Gini	3	0.72	0.72
losowy las	zbalansowana	entropia	3	0.68	0.68
losowy las	subbalans	indeks Gini	3	0.76	0.76
losowy las	subbalans	entropia	3	0.68	0.68

Porównując wyniki dla metod DRSA i drzew zdecydowanie lepiej sprawdziły się te drugie – tu uzyskane rezultaty są dość wysokie i sięgają nawet 76% zarówno dla trafności jak i F1-score.

Tabela 4.8 zawiera dane na temat reguł wyindukowanych przez poszczególne warianty metody DRSA – wartości długości, wsparcia, pokrycia i siły to dane uśrednione po zbiorze reguł dla poszczególnych modeli.

Z danych w Tabeli 4.8 można wyciągnąć wniosek, że indukowane reguły miały średnio dwa warunki. Ponadto im mniejsza jest wartość średniej długości reguł, tym większe jest średnie wsparcie, pokrycie oraz siła. Wykorzystanie algorytmu DOMApriori skutkowało powstaniem większych zbiorów reguł, co bezpośrednio wynika z faktu, że indukuje on satysfakcjonujący, a nie minimalny – jak DOMLEM, zbiór reguł. Poszczególne modele budują reguły wspierane przez średnio dwa obiekty. Pokrywa to od 25 do 45% wariantów z poszczególnych przybliżeń. Ze względu na mniejszą liczbę reguł faworyzowane może być podejście DOMLEM, wpływa to na czytelność i interpretowalność modelu.

Tablica 4.8: Statystyki reguł dla drugiego zbioru danych

Model	Długość	Wsparcie	Pokrycie	Siła	Liczba reguł
DRSA, DOMLEM, prosty	1.86	2.29	0.32	0.09	7
VC-DRSA, DOMLEM, prosty, 0.45	1.86	2.29	0.32	0.09	7
DRSA, DOMLEM, zaawansowany	1.86	2.29	0.32	0.09	7
VC-DRSA, DOMLEM, zaawansowany, 0.25	1.83	3.5	0.45	0.14	6
DRSA, DOMApriori, prosty	2.21	1.74	0.25	0.07	19
VC-DRSA, DOMApriori, prosty, 0.9	2.21	1.74	0.25	0.07	19
DRSA, DOMApriori, zaawansowany	2.21	1.74	0.25	0.07	19
VC-DRSA, DOMApriori, zaawansowany, 0.9	2.21	1.74	0.25	0.07	19

Przyglądając się dokładniej kilku najlepszym przygotowanym modelom i rozpoczynając od metody DRSA, wybrana została wersja korzystająca z algorytmu DOMApriori oraz zaawansowanego algorytmu klasyfikacji. Co istotne, mimo generacji satysfakcjonującego, a nie minimalnego zbioru reguł dysponujemy 19 regułami, co bezpośrednio wynika z wielkości zbioru danych. W Tabeli 4.9 zostały przedstawione przygotowane reguły.

Spośród reguł przedstawionych w Tabeli 4.9 9 stanowią konstrukcje o dwuelementowej części warunkowej, 7 o trójelementowej, zaś pozostałe 3 to jednoelementowe wersje. W aż 11 regułach wykorzystany jest warunek na kryterium wieku kobiety, co świadczy o istotności tego elementu. Zbiór reguł pokrywa tylko dwa typy przybliżeń unii klas: *co najwyżej klasa 1* i *co najmniej klasa 2*, gdyż w przykładach w zbiorze występują tylko te dwie z czterech możliwych klas decyzyjnych. Weźmy przykładową regułę:

$$(age \leq 31.0) \wedge (attempts \leq 1.0) \wedge (sperm \leq 2.0) \implies (class \geq 2)$$

która kobietom w wieku 31 lat lub poniżej, od których pozyskano 1 lub 2 oocyty i liczba nieudanych transferów embrionów nie przekracza 3 oraz przy wykorzystaniu spermy o koncentracji plemników w preparacie powyżej 5 milionów przypisuje klasę 2 lub wyższą. Dla takiego sformułowania do lep-

szych klas przypisywane są pary młodsze, mające niewiele nieudanych prób transferu embrionów i dysponujące spermą o wysokiej koncentracji plemników, co jest zgodne z intuicją.

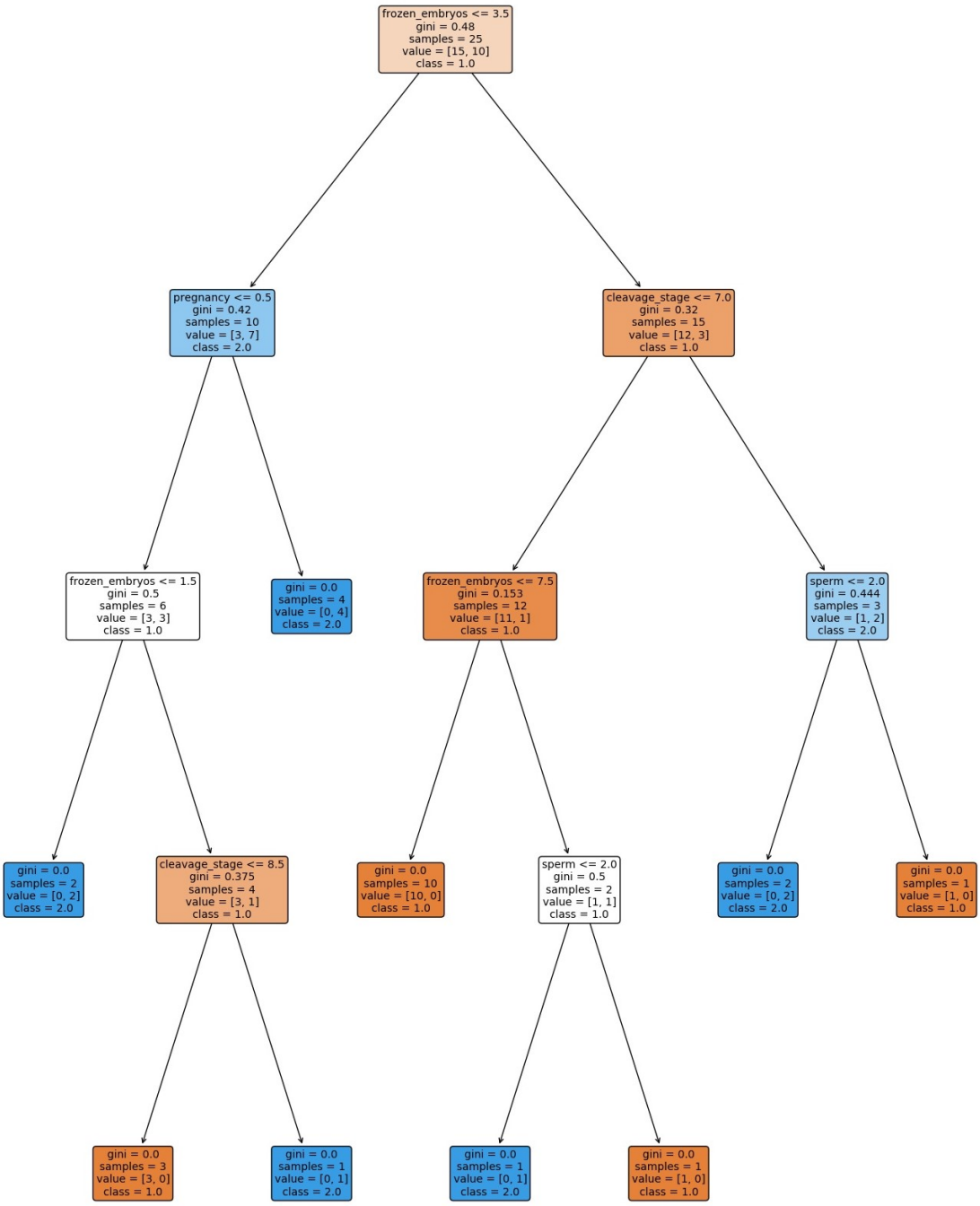
Tablica 4.9: Reguły wyindukowane dla metody DRSA przy wykorzystaniu algorytmu DOMLEM dla drugiego zbioru danych

Oznaczenie	Reguła
Reguła 1	$(age \geq 42.0) \implies (class \leq 1)$
Reguła 2	$(sperm \geq 3.0) \implies (class \leq 1)$
Reguła 3	$(pregnancy \leq 0.0) \wedge (attempts \geq 3.0) \implies (class \leq 1)$
Reguła 4	$(pregnancy \leq 0.0) \wedge (sperm \geq 2.0) \implies (class \leq 1)$
Reguła 5	$(attempts \geq 3.0) \wedge (frozen\_embryos \leq 2.0) \implies (class \leq 1)$
Reguła 6	$(attempts \geq 3.0) \wedge (sperm \geq 2.0) \implies (class \leq 1)$
Reguła 7	$(endometrium \geq 2.0) \wedge (sperm \geq 2.0) \implies (class \leq 1)$
Reguła 8	$(sperm \geq 2.0) \wedge (cleavage\_stage \geq 8.0) \implies (class \leq 1)$
Reguła 9	$(age \geq 40.0) \wedge (pregnancy \leq 0.0) \implies (class \leq 1)$
Reguła 10	$(age \geq 40.0) \wedge (frozen\_embryos \leq 2.0) \implies (class \leq 1)$
Reguła 11	$(age \geq 23.0) \wedge (frozen\_embryos \leq 2.0) \wedge (cleavage\_stage \geq 8.0) \implies (class \leq 1)$
Reguła 12	$(age \geq 33.0) \wedge (pregnancy \leq 0.0) \wedge (cleavage\_stage \geq 8.0) \implies (class \leq 1)$
Reguła 13	$(age \leq 22.0) \implies (class \geq 2)$
Reguła 14	$(sperm \leq 2.0) \wedge (frozen\_embryos \geq 8.0) \implies (class \geq 2)$
Reguła 15	$(age \leq 32.0) \wedge (pregnancy \geq 1.0) \wedge (endometrium \leq 1.0) \implies (class \geq 2)$
Reguła 16	$(age \leq 32.0) \wedge (pregnancy \geq 1.0) \wedge (sperm \leq 2.0) \implies (class \geq 2)$
Reguła 17	$(age \leq 32.0) \wedge (sperm \leq 2.0) \wedge (cleavage\_stage \leq 2.0) \implies (class \geq 2)$
Reguła 18	$(age \leq 31.0) \wedge (attempts \leq 1.0) \wedge (endometrium \leq 1.0) \implies (class \geq 2)$
Reguła 19	$(age \leq 31.0) \wedge (attempts \leq 1.0) \wedge (sperm \leq 2.0) \implies (class \geq 2)$

Wśród modeli VC-DRSA najlepiej sprawdziła się wersja korzystająca a algorytmu DOMApriori i zaawansowanej metody klasyfikacji. Jednak, z uwagi na wysoki poziom spójności ($l = 0.9$), wersja ta nie różni się od poprzednio opisanej i nie wymaga dodatkowego komentarza.

Dla drzew CART do opisu wybrana została wersja niestosująca ważenia klas oraz wykorzystująca indeks Gini jako miarę do oceny podziału. Zbudowane w tym przypadku drzewo zostało przedstawione na Rysunku 4.5.

Drzewo (Rysunek 4.5) składa się z 9 liści, z których 5 sugeruje przypisanie obiektom klasy 2, zaś pozostałe 4 – klasy 1. Główny podział wykonywany jest na podstawie kryterium mówiącego o liczbie zamrożonych embrionów. Pozostałe wykorzystane kryteria dotyczą informacji czy parze udało się dotychczas osiągnąć ciążę, oceny morfologicznej rozwoju embrionu oraz jakości i pochodzenia spermy. Zbudowane drzewo ma głębokość 4, ma więc dość złożoną strukturę.



Rysunek 4.5. Najlepsze zbudowane drzewo CART dla zbioru drugiego, indeks Gini jako miara oceny podziału

Spośród losowych lasów opisany zostanie wariant składający się z 3 estymatorów, nie stosujący wag klas oraz wykorzystujący entropię jako miarę oceny podziału. Poszczególne estymatory należące do zespołu zostały przedstawione na Rysunkach 4.6, 4.7 oraz 4.8.

Pierwszy estymator (Rysunek 4.6) zawiera 9 liści, z których 5 sugeruje przypisanie obiektowi klasy 1, a pozostałe 4 – klasy 2. Do zbudowania drzewa wykorzystanych zostało 16 wariantów ze zbioru. Użyte kryteria podziału dotyczą oceny liczby pozyskanych oocytów i nieudanych transferów embrionów (*attempts*), jakości i pochodzenia spermy, wieku kobiety, liczby embrionów, ich oceny morfologicznej oraz informacji o tym czy parze udało się dotychczas osiągnąć ciążę. Głębokość drzewa wynosi 5, zatem jest to dość złożony estymator.

Kolejne drzewo (Rysunek 4.7) również zbudowane zostało na podstawie 16 obiektów i zawiera 9 liści. Tym razem jednak 5 z nich sugeruje przypisanie klasy 2, a pozostałe 4 – klasy 1. Do podziału wykorzystanych jest 5 różnych kryteriów:

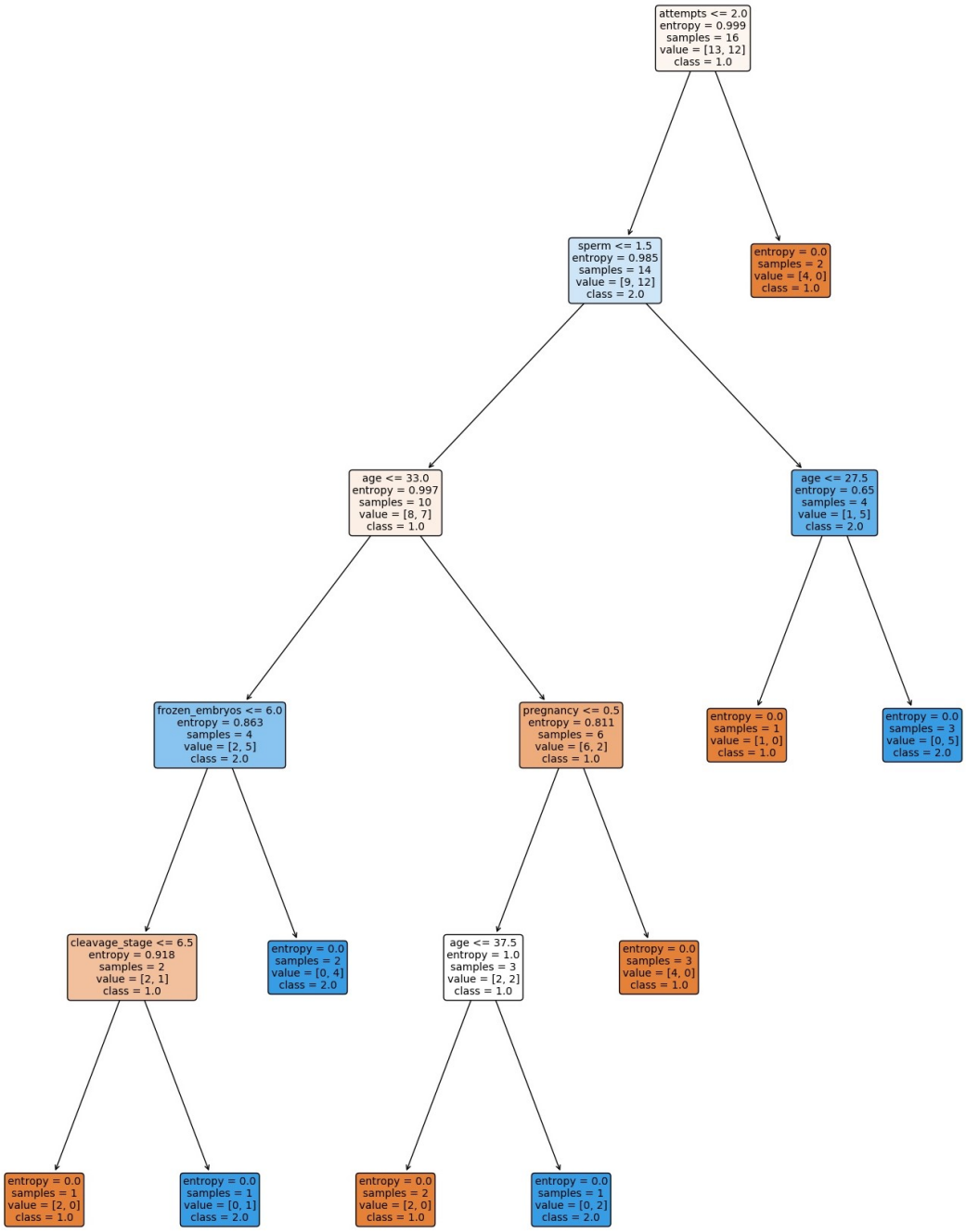
- liczba zamrożonych embrionów,
- wiek kobiety,
- ocena endometrium,
- czy parze udało się dotychczas osiągnąć ciążę,
- ocena morfologiczna stopnia rozwoju embrionów.

Trzeci estymator (Rysunek 4.8) ma głębokość 7, jest więc najbardziej złożonym drzewem w całym zespole. Zbudowany został na podstawie 14 przykładów i zawiera 8 liści, czyli mniej niż w pozostałych dwóch estymatorach w losowym lesie. Przypisanie klasy 1 jest sugerowane przez 4 liście, natomiast klasy 2 przez pozostałe 4. Wykorzystano kryteria podziału dotyczące następujących atrybutów:

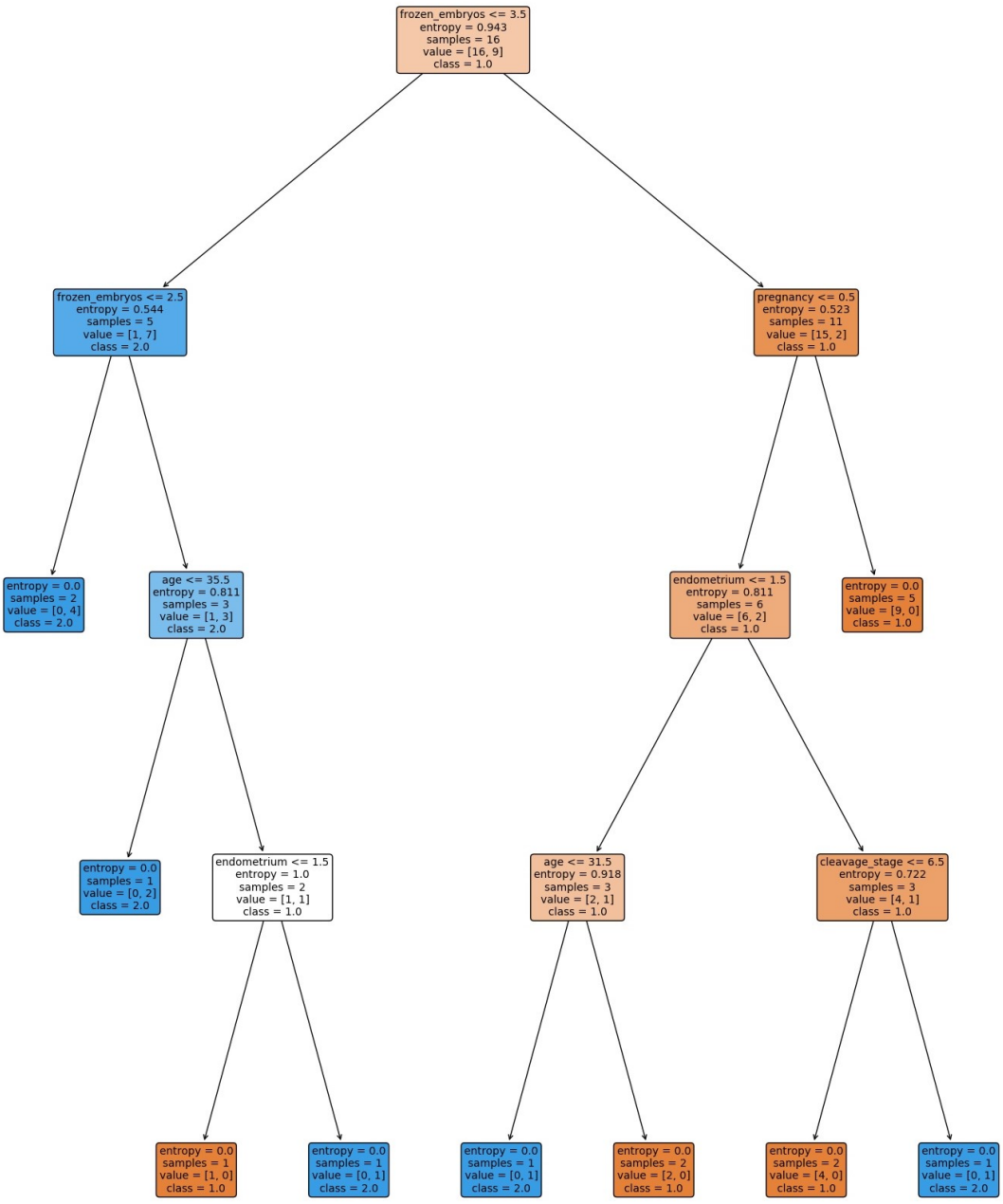
- ocena liczy pozyskanych oocytów i nieudanych prób transferu embrionów,
- wiek kobiety,
- liczba zamrożonych embrionów,
- ocena pochodzenia i jakości spermy,
- ocena morfologiczna stopnia rozwoju embrionów.

Podział ze względu na wiek kobiety wystąpił w aż 3 spośród 7 wierzchołków w drzewie.

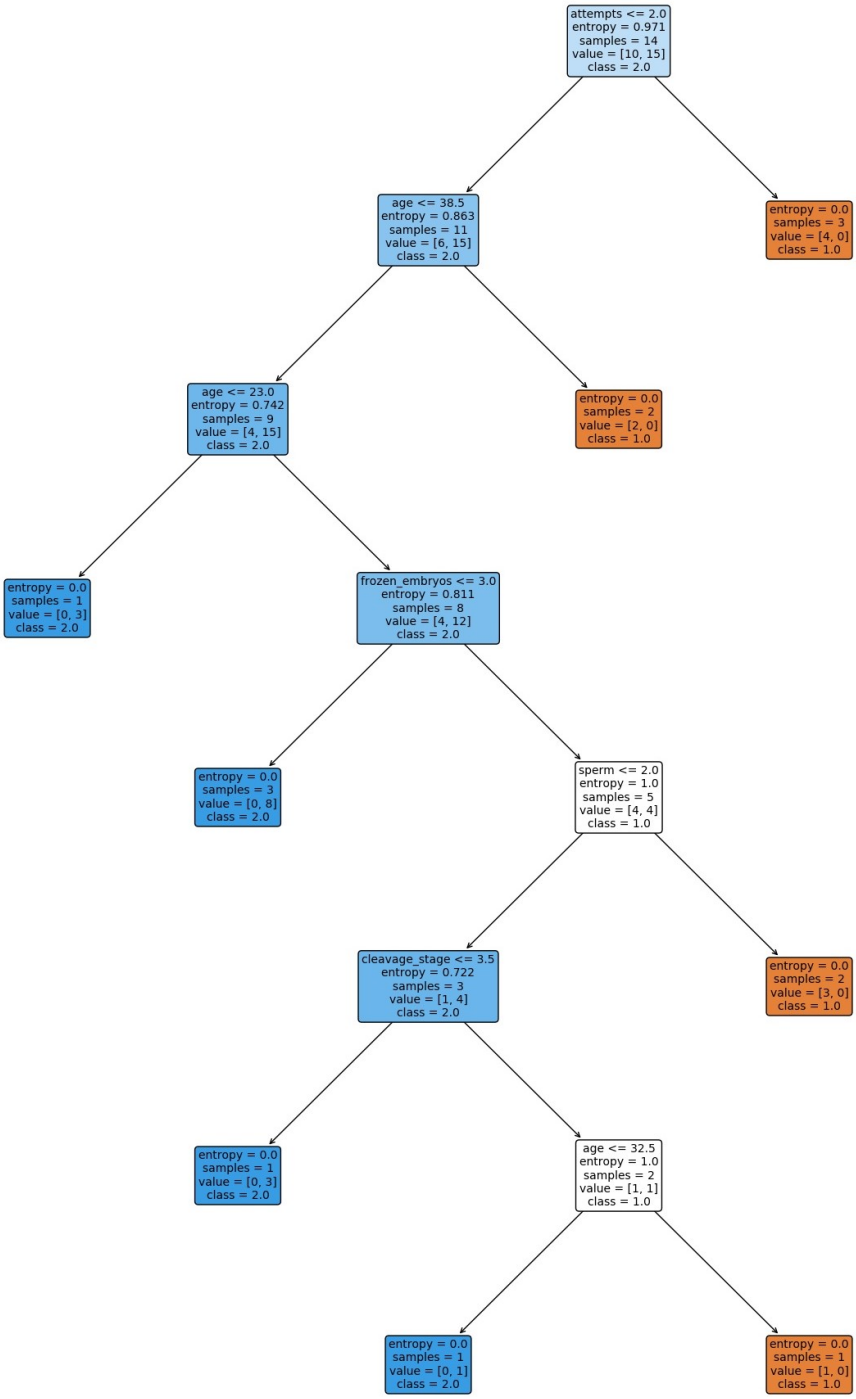
Losowy las przedstawiony na Rysunkach 4.6, 4.7 i 4.8 jest dość złożonym klasyfikatorem. Biorąc pod uwagę 3 estymatory w zespole, zostały wykorzystane wszystkie kryteria ze zbioru danych.



Rysunek 4.6. Najlepszy zbudowany losowy las dla drugiego zbioru danych wykorzystujący entropię jako miarę oceny podziału – pierwszy estymator



Rysunek 4.7. Najlepszy zbudowany losowy las dla drugiego zbioru danych wykorzystujący entropię jako miarę oceny podziału – drugi estymator



Rysunek 4.8. Najlepszy zbudowany losowy las dla drugiego zbioru danych wykorzystujący entropię jako miarę oceny podziału – trzeci estymator

4.3 Porównanie wyników

Dla obu analizowanych zbiorów udało się osiągnąć zadowalające rezultaty, gdyż w ocenie należy uwzględnić zarówno niewielką liczbę przykładów uczących jak i silne niezbalansowanie obserwacji z poszczególnych klas.

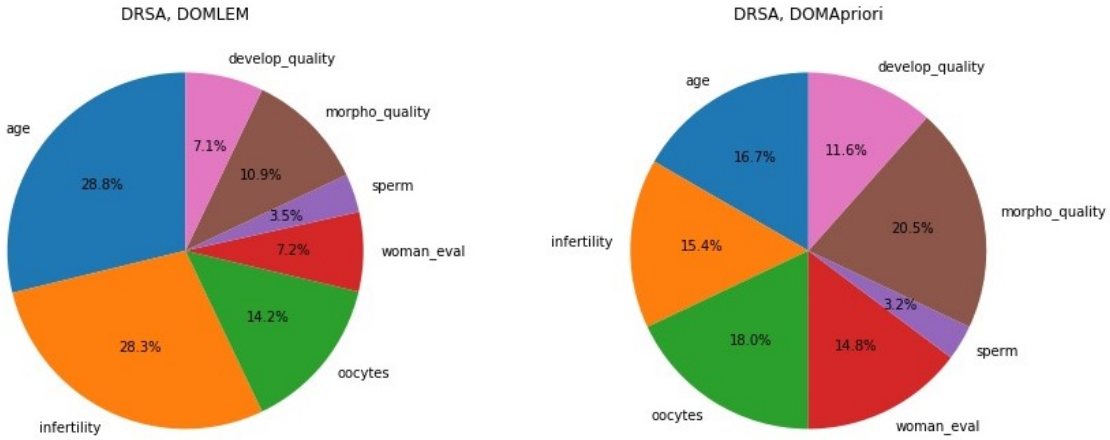
Metody DRSA i VC-DRSA lepiej sprawdziły się dla zbioru pierwszego niż drugiego. Może być to związane z faktem, że było w nim więcej obserwacji. Daje to szansę zbudowania reguł pod-

partych większą liczbą przykładów. Podobną zależność obserwujemy dla drzew CART i losowych lasów, jednak w tym przypadku różnice nie są aż tak duże.

Dla obu zbiorów lepsze rezultaty udało się osiągnąć, korzystając z podejść bazujących na drzewach decyzyjnych. Przyczyną wystąpienia tej zależności może być fakt, iż dostarczone dane były dość silnie zaszumione, co sprawia trudność podejściom regułowym. Metoda VC-DRSA, która pozwala na obniżenie stopnia spójności, a tym samym na przeciwdziałanie problemom wynikającym z szumu w obserwacjach, działała porównywalnie lub lepiej niż podstawowa wersja DRSA. Dla zbioru pierwszego w dwóch spośród czterech raportowanych najlepszych modeli VC-DRSA sugerowane było ustalenie bardzo niskiego stopnia spójności (odpowiednio 0.25 i 0.45). Nie jest to jednak pożądana parametryzacja, gdyż dopuszcza zbyt duże niespójności w zbiorze danych, a tym samym nie pozwala na przygotowanie właściwie dopasowanego modelu.

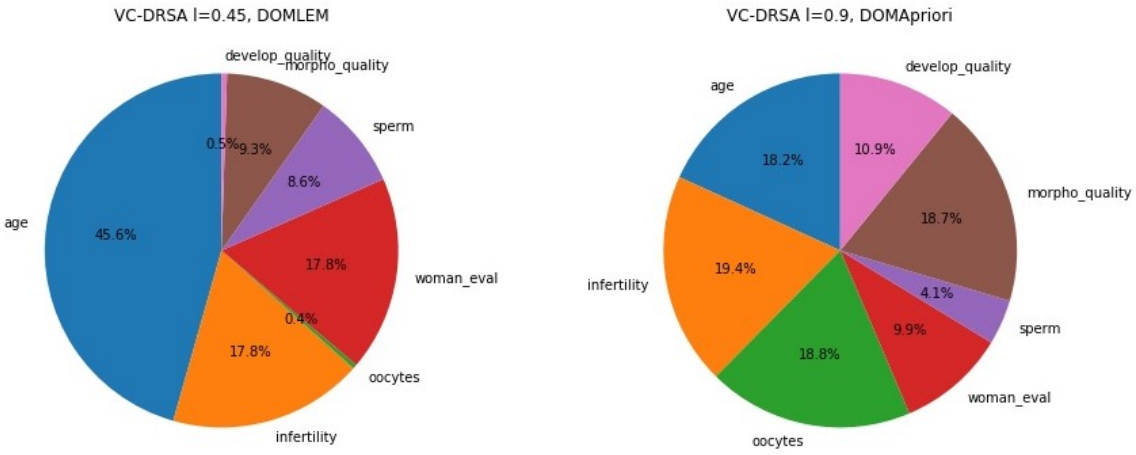
Warto zauważyć, że dla obu zbiorów bardzo istotnym, często wykorzystywanym przez modele atrybutem był wiek kobiety. Atrybut odnoszący się do jakości czy pochodzenia spermy w obu przypadkach był stosunkowo rzadko wykorzystywany, co może być związane z faktem, że jego wartości były bardzo mało zróżnicowane w obu zbiorach. Co więcej, taka dysproporcja miedzy częstością wykorzystania atrybutu wieku kobiety i jakości spermy wskazuje na poprawne działanie algorytmów – potrafiły one właściwie wybrać najistotniejsze atrybuty i wykorzystać je do różnicowania przypisania wariantów do poszczególnych klas decyzyjnych.

Przyjrzyjmy się jeszcze dokładniej kwestii wykorzystania atrybutów przez poszczególne modele dla obu zbiorów. Na Rysunku 4.9 przedstawiono częstość występowania atrybutów w regułach dla modeli DRSA budowanych na zbiorze pierwszym. Z ich analizy wynika, że kryteria *age*, *infertility* oraz *oocytes* były najczęściej używane przez model korzystający z algorytmu DOMLEM. Stanowiły 71.3% wszystkich wykorzystanych atrybutów. Najrzadziej wykorzystane kryterium to *sperm*, mające 3.5% udziału. Dla modelu DRSA indukującego reguły przy wykorzystaniu DOMApriori najczęściej wykorzystywane atrybuty to *morpho_quality*, *oocytes* i *age*, stanowiące 55.2% wszystkich wykorzystanych kryteriów. Warto jednak zauważyć, że w tej wersji wszystkie atrybuty, z wyjątkiem *sperm* były podobnie istotne (wartości między 11.6 a 20.5%), co wynika z faktu indukowania satysfakcjonującego zbioru reguł wspierających minimum jeden obiekt i mających długość nieprzekraczającą 3.



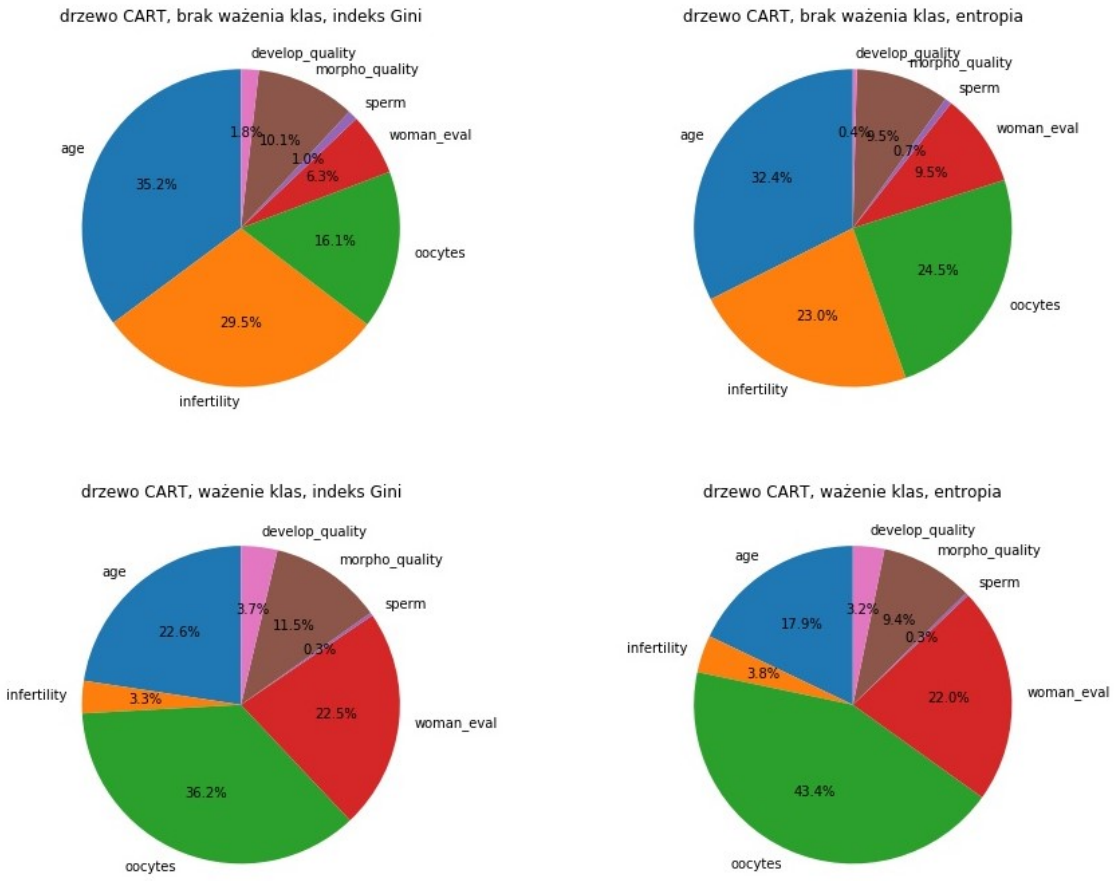
Rysunek 4.9. Wykorzystanie poszczególnych atrybutów dla DRSA – pierwszy zbiór

Na Rysunku 4.10 przedstawiono częstość występowania atrybutów w regułach dla modeli VC-DRSA budowanych na pierwszym zbiorze danych. Ich analiza pozwala stwierdzić, że dla algorytmu DOMLEM najistotniejszymi kryteriami były *age*, *infertility* oraz *woman_eval*. Stanowiły one 81.2% warunków we wszystkich regułach. Najmniej ważne okazały się kryteria *oocytes* i *develop_quality*, których udział nie sięgał nawet 1%. Dla VC-DRSA i algorytmu DOMApriori wszystkie atrybuty, z wyjątkiem *sperm* miały podobny udział w warunkach reguł.



Rysunek 4.10. Wykorzystanie poszczególnych atrybutów dla VC-DRSA – pierwszy zbiór

Na Rysunku 4.11 przedstawiono w formie wykresów kołowych częstość występowania poszczególnych atrybutów dla modeli drzew CART budowanych na podstawie pierwszego zbioru danych. Z analizy tych danych wynika, że w wersjach niestosujących ważenia klas najczęściej warunki podziału oparte były na kryteriach *age*, *infertility* oraz *ooctes* (odpowiednio 80.8% dla indeksu Gini i 79.9% dla entropii), zaś najrzadziej na atrybutach *sperm* i *develop_quality*. Dla drzew stosujących ważenie klas najistotniejsze okazały się kryteria *oocytes*, *woman_eval* i *age* (odpowiednio 81.3% dla indeksu Gini i 83.3% dla entropii), zaś najmniej istotne: *sperm*, *develop_quality* i *infertility*.

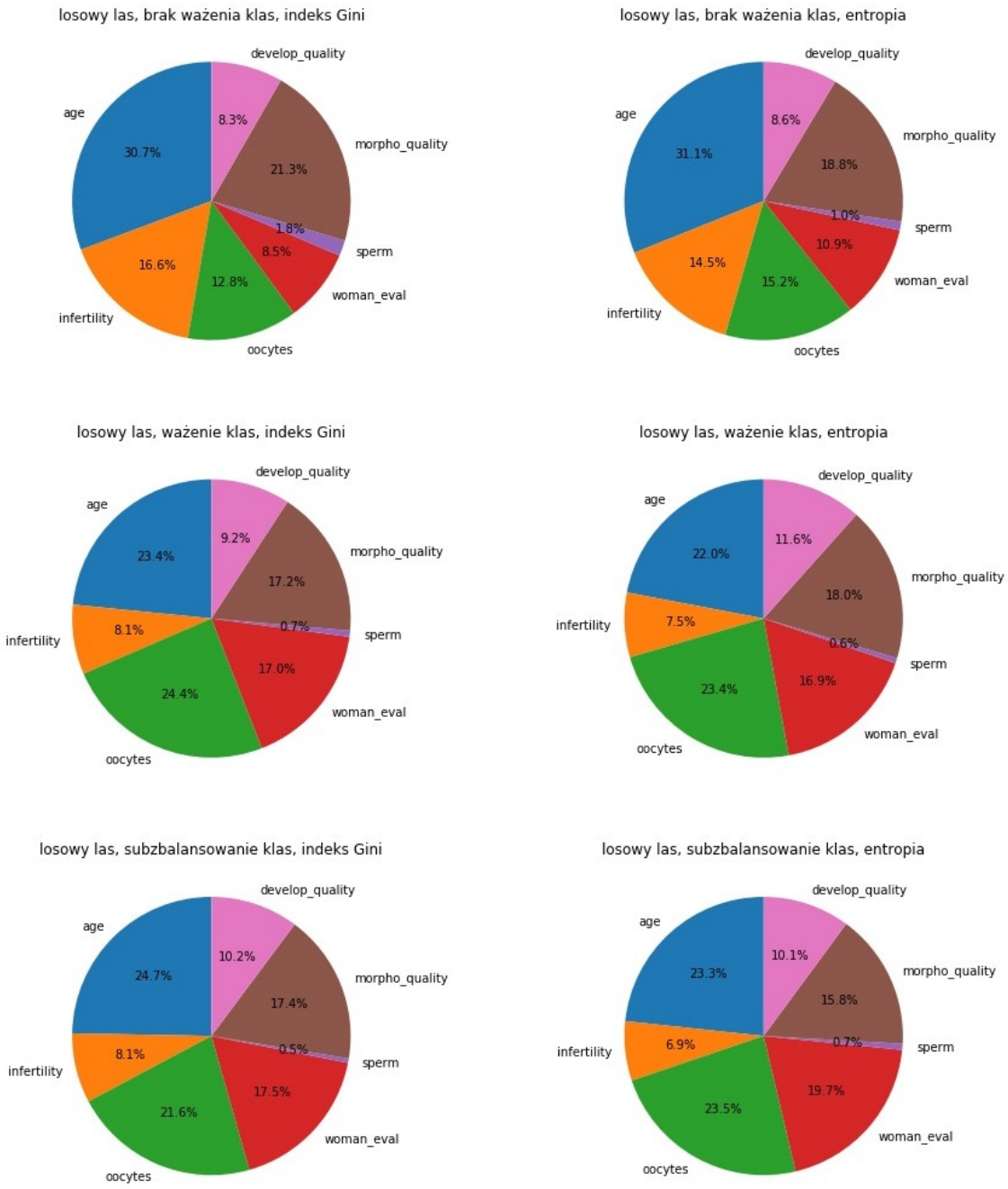


Rysunek 4.11. Wykorzystanie poszczególnych atrybutów dla drzew CART – pierwszy zbiór

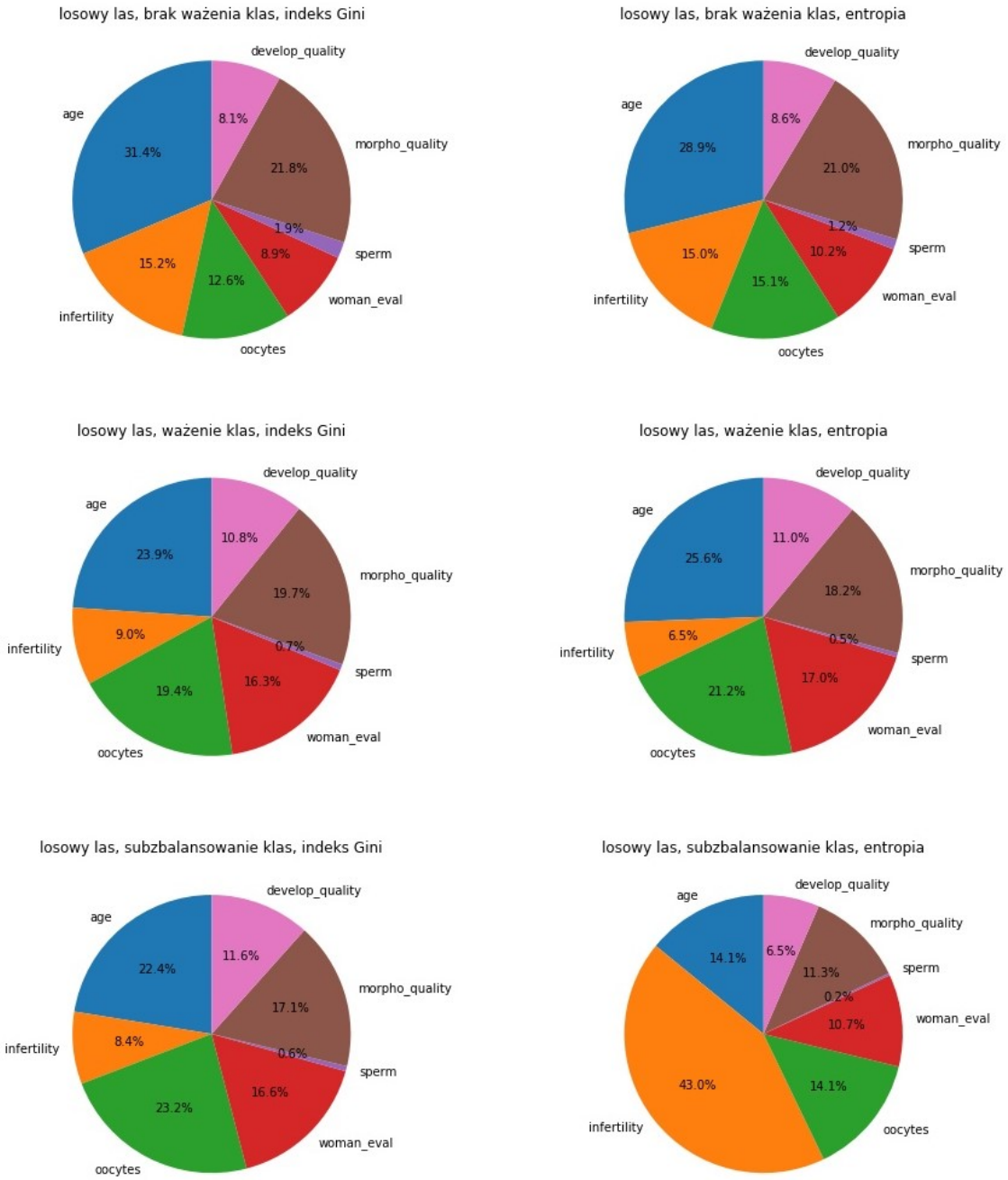
Na Rysunkach 4.12 oraz 4.13 zaprezentowano w formie wykresów kołowych częstość występowania poszczególnych atrybutów dla modeli losowych lasów składających się z odpowiednio 2 i 3 estymatorów oraz zbudowanych w oparciu o pierwszy zbiór danych.

Z analizy informacji zawartych na Rysunku 4.12 wynika, że dla niestosowania ważenia klas najistotniejsze kryteria to *age* i *morpho_quality*, zaś najmniej istotny jest atrybut *sperm*. Przy zastosowaniu wagi klas lub jej subzbalansowanej wersji najczęściej wykorzystywano kryteria *age* i *oocytes* a najrzadziej – *sperm*.

Analiza danych przedstawionych na Rysunku 4.13 pozawala stwierdzić, że najmniej istotnym kryterium, w każdej wersji, było *sperm*. Najważniejsze kryteria to, w zależności od wersji, *age*, *ooctes* i $infertility$, przy czym ostatnie dwa wymienione miały szczególne znaczenie dla subzbalansowania klas odpowiednio dla indeksu Gini i entropii jako miary podziału.

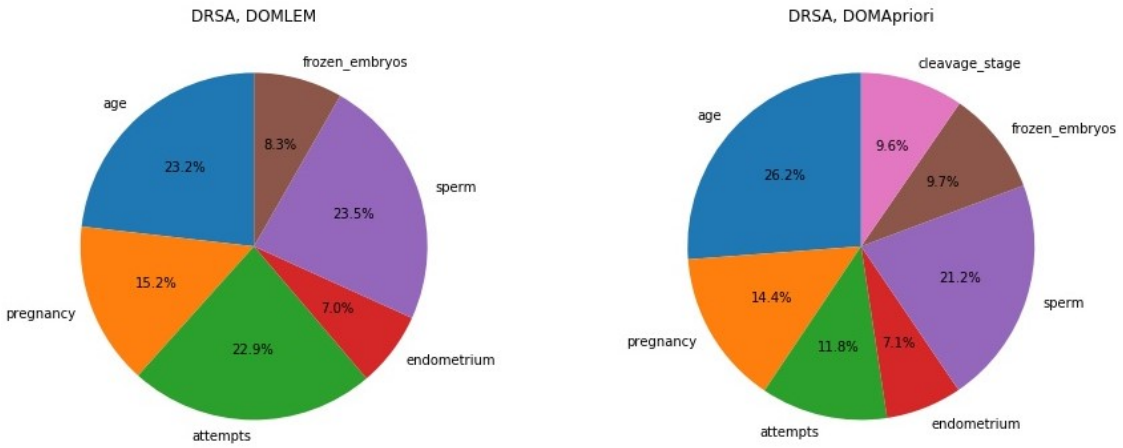


Rysunek 4.12. Wykorzystanie poszczególnych atrybutów dla losowych lasów o 2 estymatorach – pierwszy zbiór

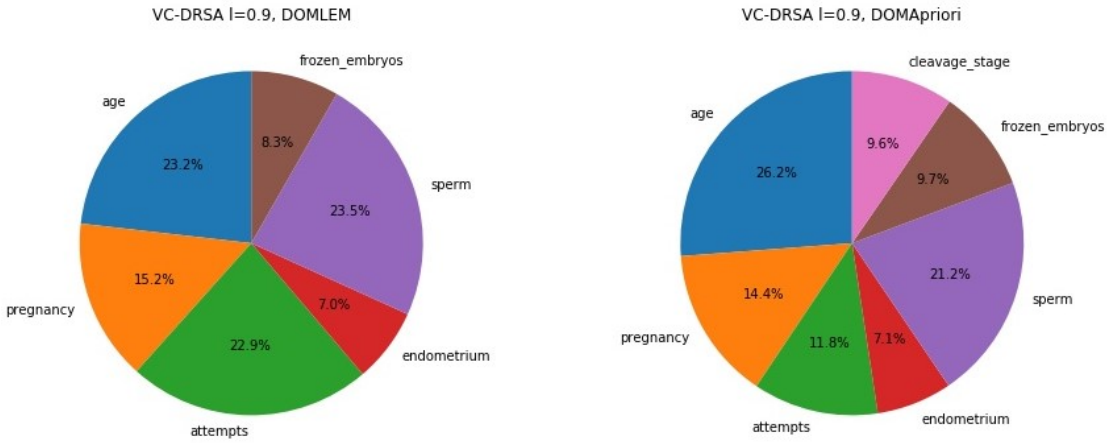


Rysunek 4.13. Wykorzystanie poszczególnych atrybutów dla losowych lasów o 3 estymatorach – pierwszy zbiór

Dla drugiego zbioru danych na Rysunkach 4.14 i 4.15 zaprezentowano w formie wykresów kołowych częstości występowania poszczególnych atrybutów w regułach zbudowanych przy wykorzystaniu metod DRSA i VC-DRSA. Najistotniejsze kryteria to w tym przypadku *sperm*, *age*, *attempts* i *pregnancy*, które stanowiły odpowiednio dla DOMLEM i DOMApriori 84.8% oraz 73.6% wszystkich atrybutów wykorzystanych w warunkach reguł. Najmniej popularne atrybuty to *endometrium* i *cleavage_stage*, z których drugi nie był użyty w ani jednej regule wyindukowanej algorytmem DOMLEM. Warto zauważyć, że dla DRSA i VC-DRSA istotności atrybutów są takie same.

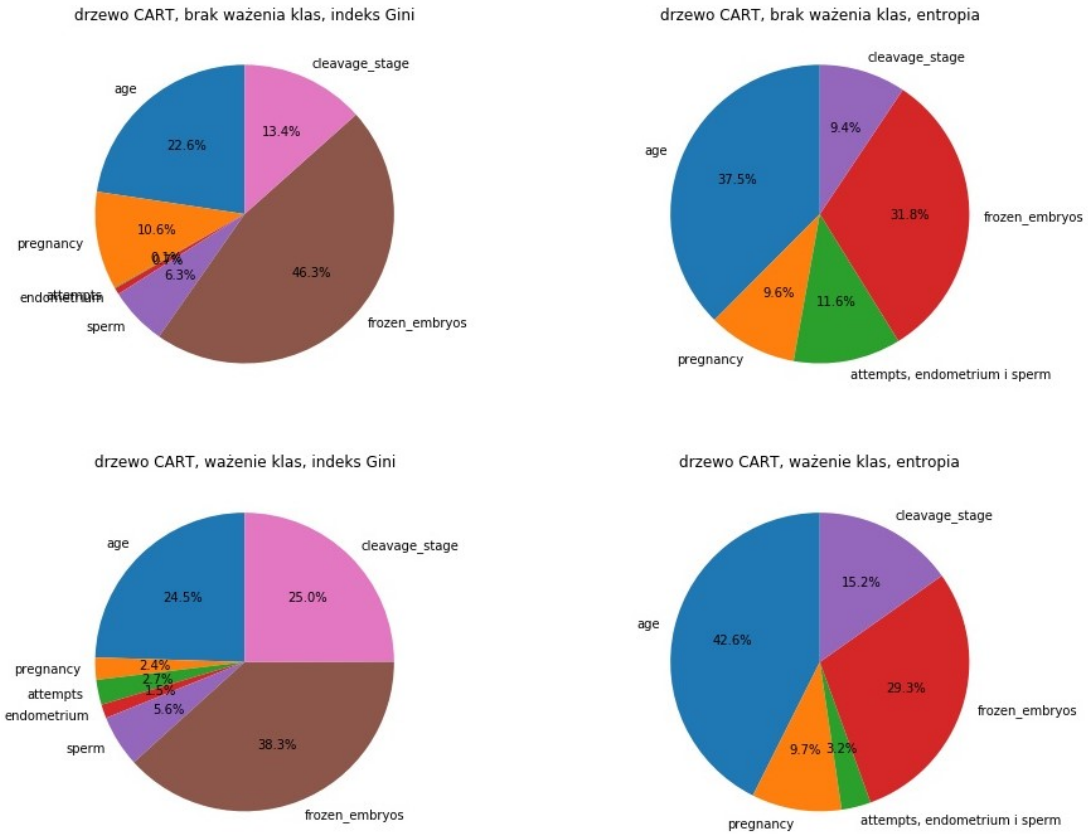


Rysunek 4.14. Wykorzystanie poszczególnych atrybutów dla DRSA – drugi zbiór



Rysunek 4.15. Wykorzystanie poszczególnych atrybutów dla VC-DRSA – drugi zbiór

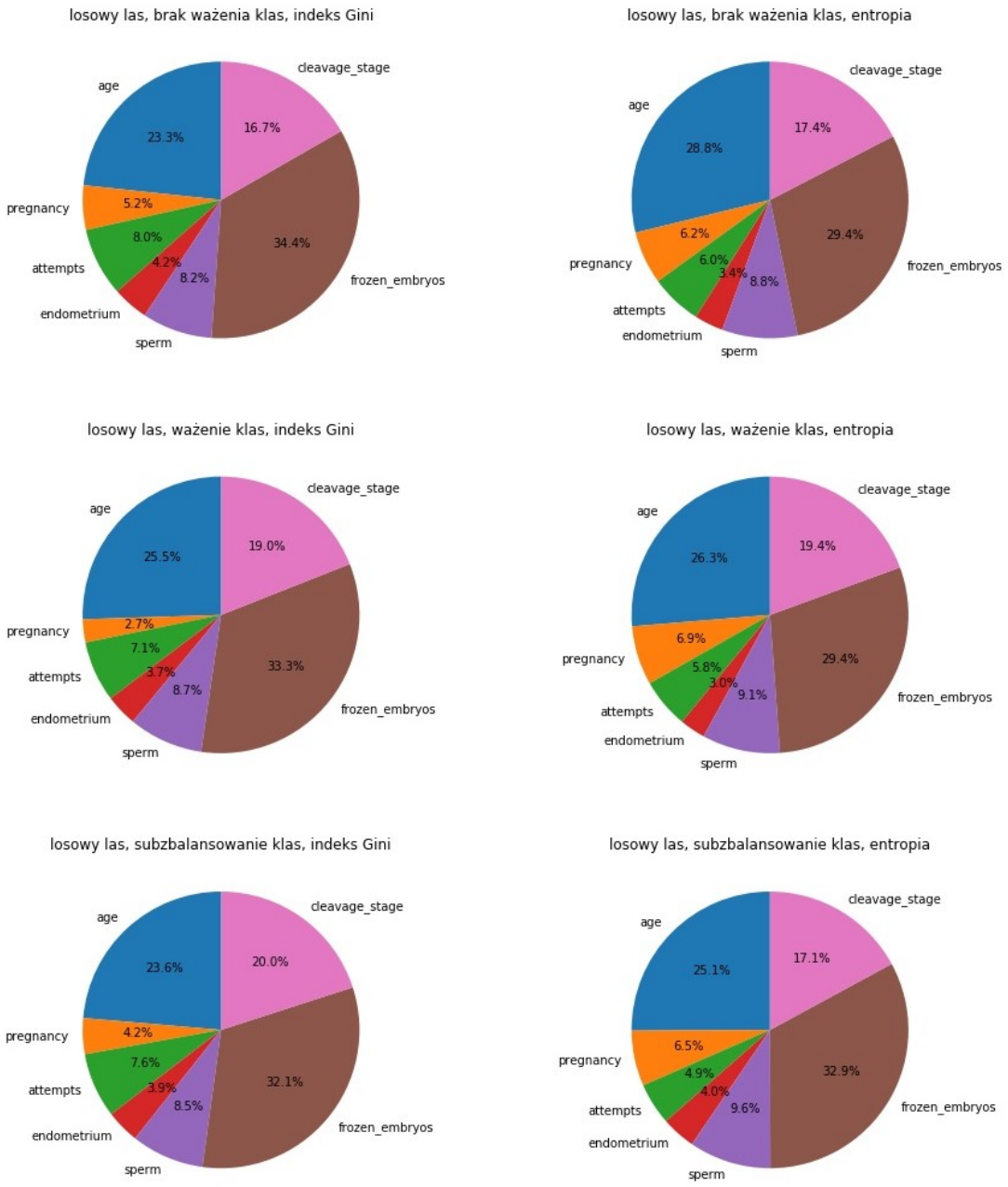
Na Rysunku 4.16 przedstawiono częstości wykorzystania poszczególnych atrybutów dla drzew CART budowanych w oparciu o dane zawarte w drugim zbiorze. Niezależnie od wersji najistotniejszymi kryteriami są *forzen_embryos*, *age* i *cleavage_stage*. Sumarycznie stanowią od 82.3% do 96.3% wszystkich ustalonych kryteriów podziału wykorzystanych w tych modelach. Pozostałe kryteria, w szczególności *endometrium*, *attempts* i *sperm*, są zdecydowanie mniej istotne dla drzew CART budowanych dla zbioru drugiego.



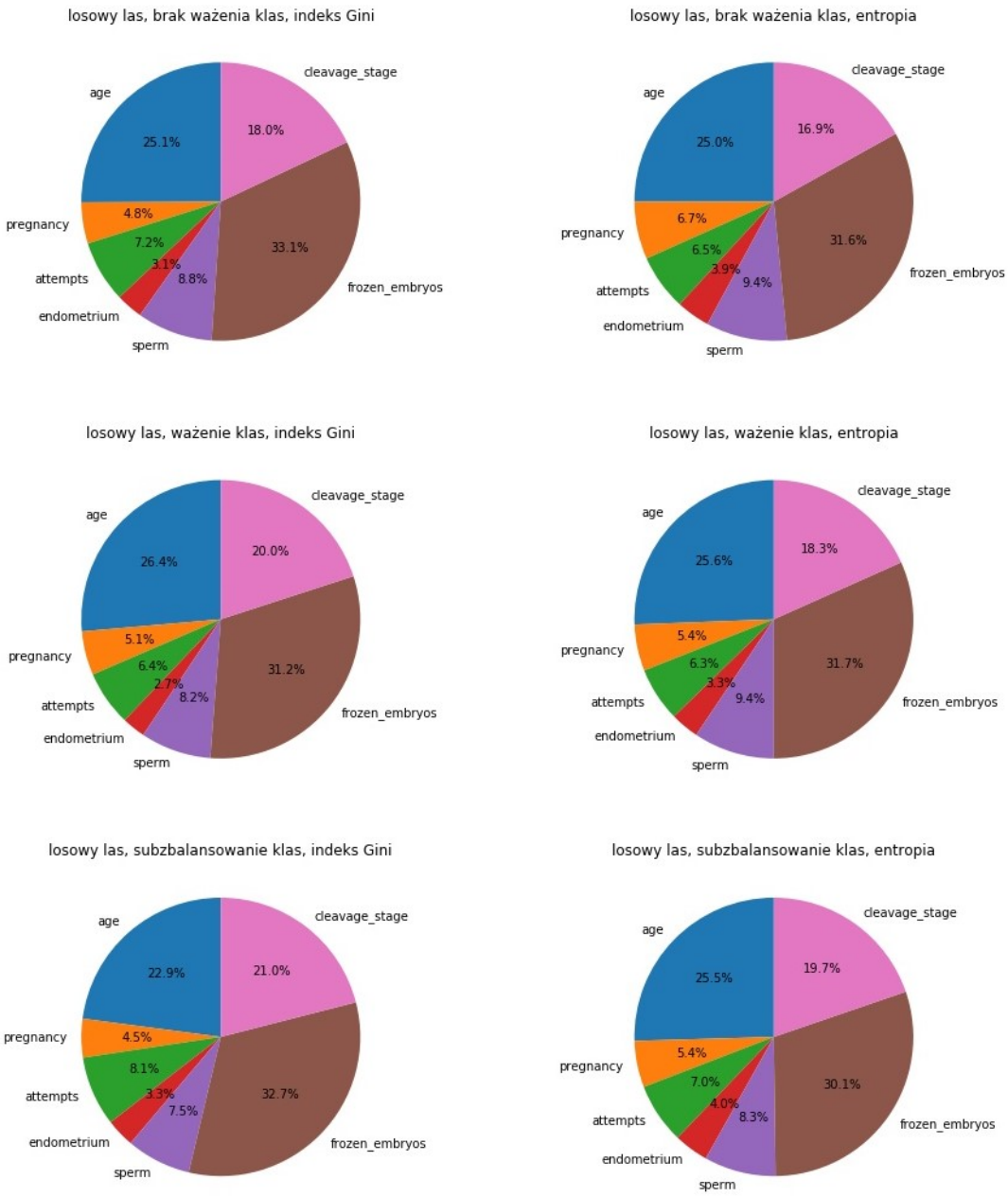
Rysunek 4.16. Wykorzystanie poszczególnych atrybutów dla drzew CART – drugi zbiór

Rysunki 4.17 i 4.18 przedstawiają w formie wykresów kołowych częstości wykorzystania poszczególnych atrybutów dla losowych lasów o odpowiednio 2 i 3 estymatorach, budowanych na drugim zbiorze danych.

Z analizy danych przedstawionych na Rysunku 4.17 wynika, iż najistotniejszymi atrybutami były *frozen_embryos*, *age* oraz *cleavage_stage*. Najmniejszą rolę odgrywały *pregnancy*, *attempts* i *endometrium*. Dla losowych lasów składających się z trzech estymatorów (patrz Rysunek 4.18) obserwujemy analogiczną ważność poszczególnych atrybutów.



Rysunek 4.17. Wykorzystanie poszczególnych atrybutów dla losowych lasów o 2 estymatorach – drugi zbiór



Rysunek 4.18. Wykorzystanie poszczególnych atrybutów dla losowych lasów o 3 estymatorach – drugi zbiór

Patrząc całościowo na częstości wykorzystania poszczególnych atrybutów dla zbioru pierwszego (Rysunki 4.9 – 4.13) i drugiego (Rysunki 4.14 – 4.18), najistotniejszym atrybutem w obu przypadkach okazał się wiek kobiety. W wielu przypadkach małe znaczenie ma atrybut opisujący spermę. Dla obu zbiorów popularne są atrybuty opisujące embriony. W kilku modelach niewielką rolę odgrywa opis płodności pary, definiowany jako liczba lat niepłodności w przypadku zbioru pierwszego oraz podsumowanie nieudanych transferów embrionów dla zbioru drugiego.

Rozdział 5

Podsumowanie

W pracy magisterskiej została przeprowadzona analiza dwóch zbiorów danych dotyczących leczenia niepłodności metodą in vitro. Zrealizowano etap opisu zbiorów i ich statystyk. Przedstawione zostały wszystkie wykorzystane w pracy metody i algorytmy wraz z przykładami ich działania pochodzącymi z pierwszego zbioru danych. Przygotowano także raport wyników dla najlepszych przetestowanych i poddanych walidacji modeli dla obu zbiorów danych.

W wyniku eksperymentów udało się potwierdzić ważność atrybutu opisującego wiek kobiety i mniejszą rolę tego, który dotyczył jakości i pochodzenia spermy. Zbudowane modele miały różną skuteczność działania, jednak dla obu zbiorów danych lepiej sprawdziły się modele drzewiaste, w szczególności losowe lasy. Podejścia regułowe pozwoliły na osiągnięcie lepszych rezultatów dla zbioru pierwszego niż drugiego. W dalszym ciągu były to wyniki słabsze, nieznacznie, niż dla modeli drzew decyzyjnych.

Warto jednak zaznaczyć, że wszystkie najlepsze modele były interpretowalne i zrozumiałe dla człowieka. Ponadto zbudowane reguły i drzewa cechowało dobre dopasowanie do problemu, rozumiane jako zgodność modeli z obserwacjami wiedzy dziedzinowej. Za przykład można tu podać przypisywanie lepszych kategorii parom, w których kobiety były młodsze.

Otrzymane modele mogłyby osiągnąć prawdopodobnie większą trafność dla zbiorów danych o większej liczbie obserwacji. Co więcej, niebagatelny wpływ na jakość klasyfikacji ma także rozkład obserwacji z poszczególnych klas – dla analizowanych zborów występowało silnie niezbalansowanie. Ponadto drugi zbiór danych nie zawierał obserwacji z klas 3 i 4, co skutkowało budową modeli pracujących tylko na dwóch z czterech klas decyzyjnych, a więc takich, które nie miały zdolności poprawnej klasyfikacji obiektów, którym sugerowane byłoby przypisanie klasy innej niż 1 lub 2.

Ciekawą propozycją rozszerzenia badań przeprowadzonych w ramach pracy byłoby wykorzystanie do problemu innych zespołów drzew. Klasyfikatory złożone najlepiej sprawdziły się dla zastosowania do leczenia niepłodności metodą in vitro, stąd inne ich warianty mogłyby przynieść wzrost trafności modeli.

Innym postępowaniem, które mogłoby pozytywnie wpłynąć na skuteczność działania modeli jest wykorzystanie metod radzenia sobie z niezbalansowaniem obserwacji z poszczególnych klas takich jak dodanie nowych przykładów z klas najmniej licznych czy usunięcie odpowiedniej liczby obserwacji z klas najbardziej licznych. Jednak tego typu postępowanie dla zbioru drugiego nadal nie rozwiązałoby problemu braku obserwacji z klas oznaczonych symbolami 3 i 4.

Literatura

[1] Leo Breiman. Random forests. *Machine Learning*, 2001.

[2] Jerzy Błaszczyński, Salvatore Greco, and Roman Słowiński. Multi-criteria classification – a new scheme for application of dominance-based decision rules. *European Journal of Operational Research*, 181(3):1030–1044, 2007.

[3] Kimbroe Carter, Nathan Ritchey, Frank Castro, Leonard Caccamo, Edward Kessler, and Barbara Erickson. Analysis of three decision-making methods: A breast cancer patient as a model. *Medical decision making : an international journal of the Society for Medical Decision Making*, 19:49–57, 01 1999.

[4] Maria Manuela Silva Sá Couto. The application of decision making techniques in infertile couples. *None*, 2017.

[5] dr n.med. Anna Bednarska-Czerwińska. Niepłodność - choroba cywilizacyjna. czy można jej zapobiec? `https://www.gyncentrum.pl/blog/pl/nieplodnosc-choroba-cywilizacyjna-czy-mozna-jej-zapobiec`.

[6] André Elisseeff and Massimiliano Pontil. Leave-one-out error and stability of learning algorithms with applications stability of randomized learning algorithms source. *International Journal of Systems Science - IJSySc*, 6, 01 2002.

[7] JR Figueira, J Almeida-Dias, S Matias, B Roy, MJ Carvalho, and CE Plancha. Electre tri-c, a multiple criteria decision aiding sorting model applied to assisted reproduction. *International journal of medical informatics*, 80(4):262—273, April 2011.

[8] Salvatore Greco, Benedetto Matarazzo, and Slowinski Roman. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. *European Journal of Operational Research*, 138:247–259, 02 2002.

[9] Salvatore Greco, Benedetto Matarazzo, Slowinski Roman, and J. Stefanowski. Variable consistency model of dominance-based rough sets approach. In *International Conference on Rough Sets and Current Trends in Computing*, volume 2005, pages 170–181, 12 2001.

[10] Badr Hssina, Abdelkarim MERBOUHA, Hanane Ezzikouri, and Mohammed Erritali. A comparative study of decision tree id3 and c4.5. *(IJACSA) International Journal of Advanced Computer Science and Applications*, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 07 2014.

[11] Paul Mangiameli, David West, and Rohit Rampal. Model selection for medical diagnosis decision support systems. *Decision Support Systems*, 36:247–259, 01 2004.

[12] matthewb. Random forests. `https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf`.

[13] Lior Rokach and Oded Maimon. *Decision Trees*, volume 6, pages 165–192. None, 01 2005.

[14] scikit. Decisiontreeclassifier. `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier`.

[15] scikit. Randomforestclassifier. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=randomforestclassifier#sklearn.ensemble.RandomForestClassifier`.

[16] scikit. sklearn. `https://scikit-learn.org/stable/`.

[17] Jerzy Stefanowski. *Algorytmy indukcji reuł decyzyjnych w odkrywaniu wiedzy*. Wydawnictwo Politechniki Poznańskiej, 2 2001.

[18] David West, Paul Mangiameli, Rohit Rampal, and Vivian West. Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*, 162:532–551, 04 2005.

[19] Yisehac Yohannes and John Hoddinott. Classification and regression trees: An introduction. *None*, 01 1999.

[20] Nick Z. Zacharis. Classification and regression trees (cart) for predictive modeling in blended learning. *International Journal of Intelligent Systems and Applications*, 10:1–9, 2018.

Instytut Informatyki, Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

Skład przy użyciu systemu LaTeX na platformie Overleaf.