Drzewo C 4.5

Maciej Urbaniak 200842

April 9, 2018

1 Wstęp

Drzewo J48 jest generowane za pomocą algorytmu C4.5, który dzieli dane używając kolejnych atrybutów. Do każdego podziału przeliczane jest "information gain" a największa z nich staje się pierwszym węzłem drzewa. Operacja ta powtarzana jest dla wszystkich podzbiorów. Przy czym każda linia wychodząca od wybranego węzła jest określana jedną z możliwych wartości atrybutu wyznaczonego dla niego w kroku poprzednim. Wzrost "information gain" jest tym większy im bardziej jednorodny jest rozkład prawdopodobieństwa wybranego atrybutu.

2 Parametry

2.1 Glass

Z powodu niewielkiej ilości danych dla klas 3, 5 i 6 przy losowaniu zbioru uczącego często występuje sytuacja w której jest ich zbyt mało lub wcale, by prawidłowo wytrenować klasyfikator.

Table 1: Podsumowanie zbioru glass

Correctly Classified Instances	140 65.4206%
Incorrectly Classified Instances	74 34.5794%
Kappa statistic	0.5375
Mean absolute error	0.1401
Root mean squared error	0.3118
Relative absolute error	56.8135%
Root relative squared error	88.9542%
Total Number of Instances	214

Statystyka Kappa określa zależność między zaproponowanym przydziałem obiektów do klasy ("random guessing") a otrzymanymi wynikami.

Table 2: Podsumowanie zbioru glass

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,686	0,174	0,658	0,686	0,671	0,507	0,782	0,589	1
0,579	0,159	0,667	0,579	0,620	0,435	0,711	0,562	2
0,294	0,046	0,357	0,294	0,323	0,272	0,772	0,240	3
0,769	0,025	0,667	0,769	0,714	0,696	0,870	0,733	5
1,000	0,039	0,529	1,000	0,692	0,713	0,979	0,537	6
0,828	0,027	0,828	0,828	0,828	0,801	0,914	0,701	7
0,654	0,124	0,655	0,654	0,650	0,523	0,787	0,573	Średnia

TP Rate wskazuje, jaki procent obserwacji z wybranej klasy jest poprawnie zaklasyfikowany (true positive).

FP Rate to jaka część obiektów nienależących do danej klasy została do niej błędnie zaklasyfikowana (false positive).

Precision to miara precyzji przyporządkowania obiektu do właściwej klasy.

Natomiast Recall oznacza poprawne pokrycie wybranej klasy.

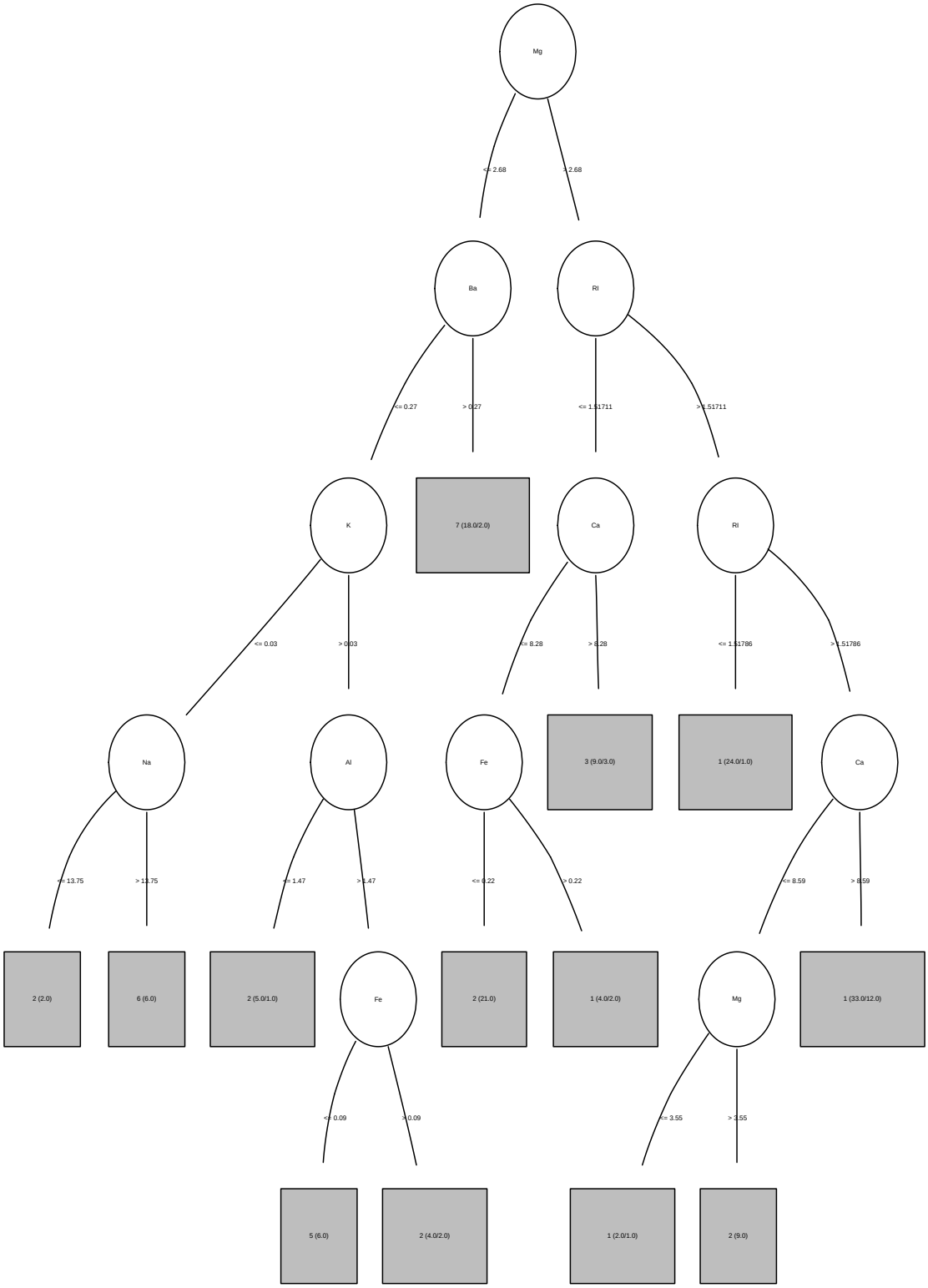

Figure 1: Glass C 4.5 Tree

2.2 Wine

Dzięki podzieleniu obiektów na zbilansowane pod względem ilości klasy wyniki klasyfikacji są dokładne.

Table 3: Podsumowanie zbioru wine

Correctly Classified Instances	170 95.5056%
Incorrectly Classified Instances	8 4.4944%
Kappa statistic	0.9317
Mean absolute error	0.039
Root mean squared error	0.174
Relative absolute error	8.8804%
Root relative squared error	37.1292%
Total Number of Instances	178

Table 4: Podsumowanie zbioru wine

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,983	0,017	0,967	0,983	0,975	0,962	0,994	0,986	1
0,944	0,037	0,944	0,944	0,944	0,906	0,956	0,937	2
0,938	0,015	0,957	0,938	0,947	0,928	0,966	0,922	3
0,955	0,025	0,955	0,955	0,955	0,931	0,971	0,949	Średnia

Table 5: Wpływ parametru R

Parametr	R=TRUE	R=FALSE
F1-score	0,864	0,927

Parametr R odpowiada za przycinanie drzewa i choć brak redukcji rozmiaru wpływa pozytywnie na wyniki, w szczególności dla danych uczących, to może znacznie zaniżać ogólność drzewa co zwiększa ryzyko overfittingu.

Table 6: Wpływ parametru C

Parametr C	0.1	0.2	0.3	0.4	0.5
F1-score	0,916	0,932	0,904	0,910	0,915

Parametr C (confidence factor) określa częstość przycinania drzewa. Mniejsze wartości to więcej przycinania.

Table 7: Wpływ parametru M

Parametr M	2	5	10	20	30
F1-score	0,921	0,916	0,921	0,847	0,845

Parametr M określa minimalną liczbę instancji na liść.

Table 8: Wpływ parametru N

Parametr N	2	4	5	6	10
F1-score	0,833	0,830	0,834	0,836	0,840

Podział danych na zbiory przy czym jeden z nich używany jest do przycinania a reszta do rozrostu drzewa.

Table 9: Wpływ rozmiaru kroswalidacji

Parametr K	2	3	4	5	6	7	8	9	10	11	12
F1-score	0,876	0,865	0,875	0,904	0,921	0,898	0,916	0,927	0,938	0,916	0,899

Porównanie najlepszych wyników:
Bayes Stratified 10fold:.....0.9616
C 4.5 Tree:.....0,938

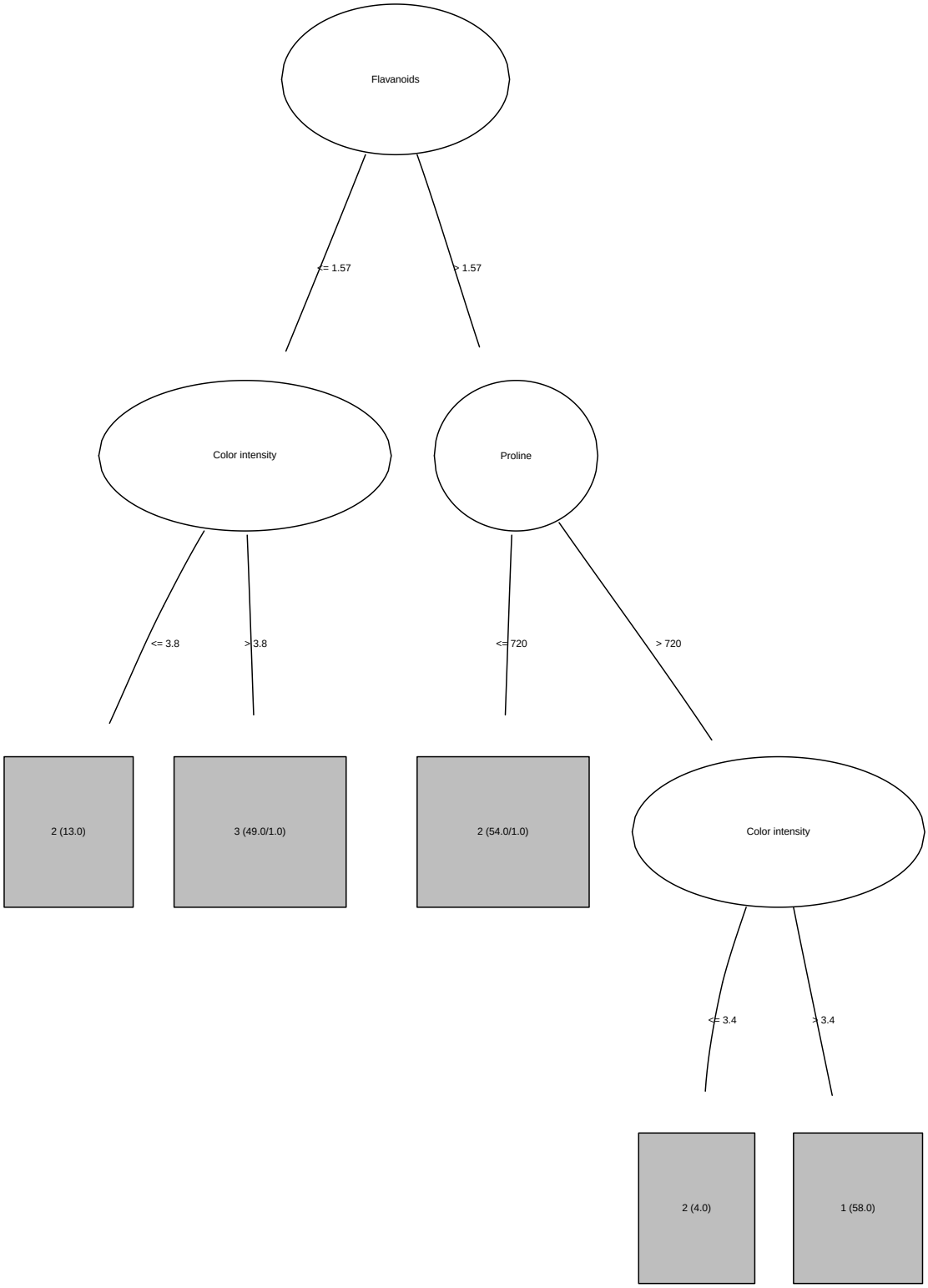

Figure 2: Parametr R = FALSE

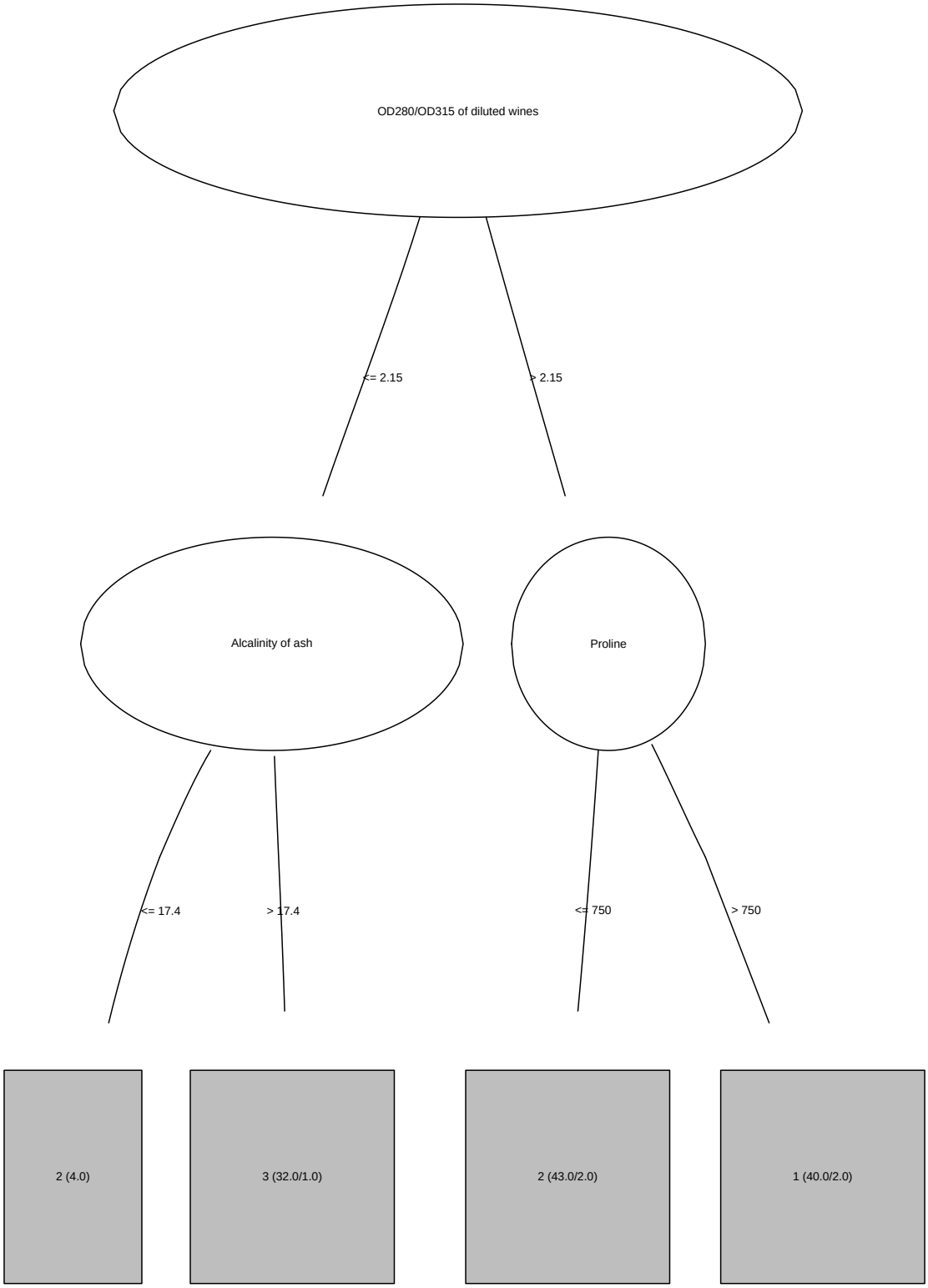

Figure 3: Parametr R = TRUE

2.3 Diabetes

Zbiór poniższych danych zawiera wartości zerowe dla przykładowo BMI czy ciśnienia krwi co wskazuje na niepełność tego zbioru informacji. Mimo tego drzewo C 4.5 jest w stanie poprawnie kwalifikować większość danych.

Table 10: Podsumowanie zbioru diabetes

Correctly Classified Instances	554 72.1354%
Incorrectly Classified Instances	214 27.8646%
Kappa statistic	0.352
Mean absolute error	0.3571
Root mean squared error	0.4462
Relative absolute error	78.5766%
Root relative squared error	93.6164%
Total Number of Instances	768

Table 11: Podsumowanie zbioru diabetes

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,848	0,515	0,754	0,848	0,798	0,358	0,721	0,789	0
0,485	0,152	0,631	0,485	0,549	0,358	0,721	0,561	1
0,721	0,388	0,711	0,721	0,711	0,358	0,721	0,709	Średnia

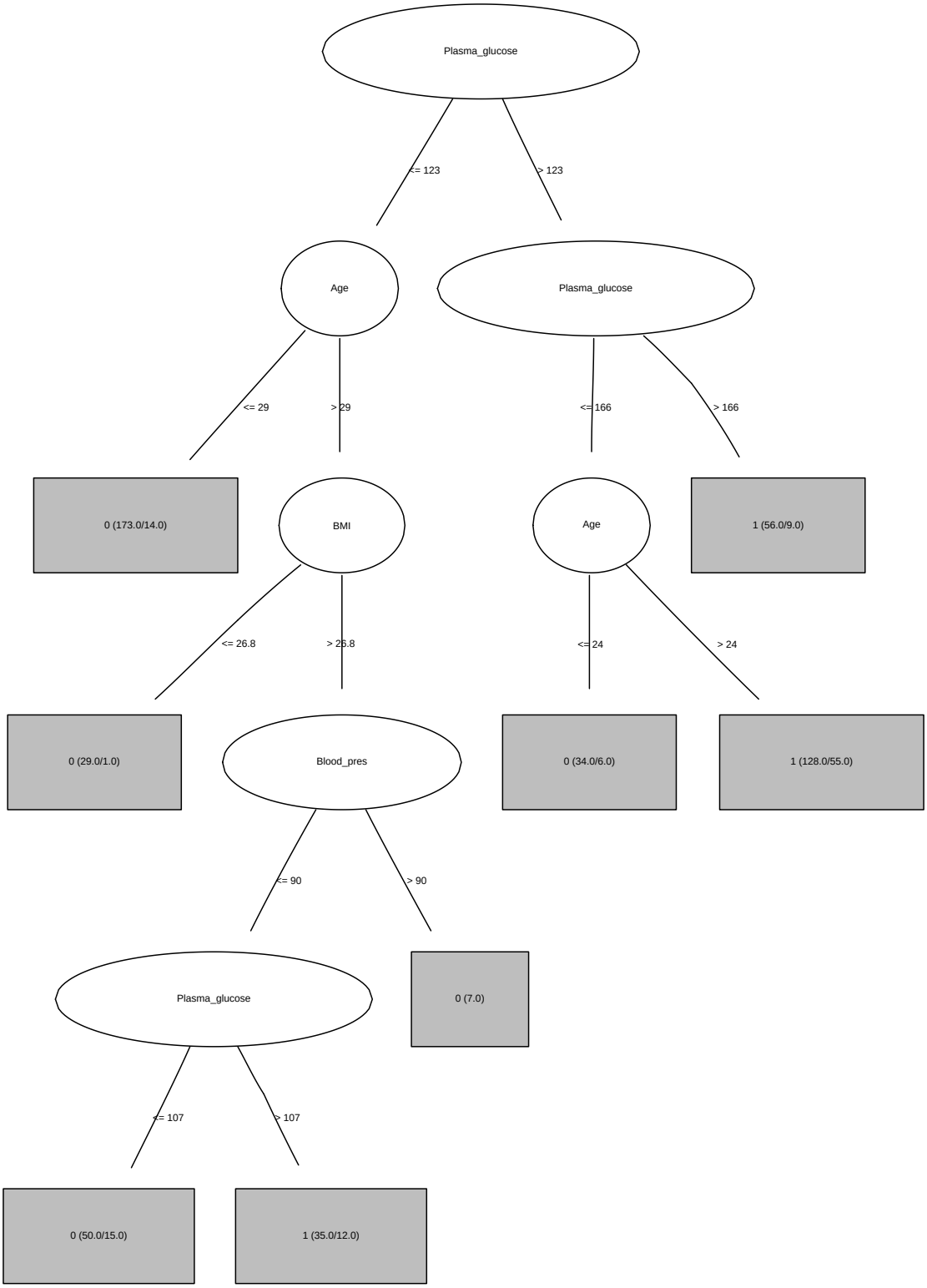

Figure 4: Diabetes C 4.5 Tree

3 Wnioski

Podobnie jak w przypadku klasyfikatora Bayes'owskiego najgorsze wyniki występowały dla zbioru glass z powodu nierównego podziału obiektów oraz ich małej liczby dla niektórych klas. Ale mimo tego wynik F1-score był znacznie lepszy dla drzewa C 4.5 (0.650) niż dla klasyfikatora Bayes'a(0.4462).