Projekt - Drzewa Decyzyjne I

Sebastian Michoń 136770, Marcin Zatorski 136834
grupe L5

1 Metoda ogólna - dane dyskretne

1. Wpierw dataset jest wczytany do formatu dataframe biblioteki pandas. Usuwane są nazwiska pasażerów i ich nry id.

2. Wiek pasażera jest transformowany zgodnie z wytycznymi (to dotyczy tylko obliczeń dla 1. części projektu).

3. Drzewo jest tworzone jako tablica podzbiorów dataframe'a: na początku, w korzeniu drzewa (w 0. elemencie listy) dany jest cały dataframe; jeśli możliwy jest split w danym wierzchołku (i gain ratio wynikający z tego splita jest większy od 0), to dataframe w wierzchołku jest dzielony według kolumny, split na której daje najwyższy gain ratio. Każdy powstający podzbiór dataframe'a jest dodawany na koniec tablicy wierzchołków. Razem z nimi dla każdego wierzchołka uzyskuję informacje o nrze ojca i splicie, w wyniku którego powstał (który jest wygodny do późniejszego rysowania drzewa).

4. Entropia, entropia warunkowa, gain ratio, information gain i intristic information są liczone jawnie ze wzoru - w ogólności złożoność obliczeniowa kalkulowania kolejnego splita to $O(n \sum_{i=0}^{m} x_i)$, gdzie n to liczba wierszy dataseta w danym wierzchołku, m to liczba atrybutów, a x_i to liczba różnych wartości w i-tym atrybucie dla dataframe'a w tym wierzchołku.

5. W momencie, w którym nie da się uzyskać lepszego od 0 gain ratio w danym wierzchołku (może zajść, gdy żadna pojedyncza kolumna nie pozwala lepiej sklasyfikować zbioru danych niż sam wierzchołek), nie jest dokonywany żaden split (co nie oznacza, że nie byłoby zasadnym go dokonać - natomiast algorytm podziału jest zachłanny, więc działa w taki a nie inny sposób). Jeśli nie da się dokonać żadnego kolejnego splita konstrukcja drzewa się kończy. Przykład dataseta, w którym 2 splity prowadziłyby do wzrostu współczynnika informacji, choć żaden nie zostanie wykonany: (Y to atrybut, którego wartość chcemy poznać w zależności od c1, c2)

c1	c2	Y
A	B	1
B	A	1
A	A	0
B	B	0

6. W ostatniej fazie rysowane jest drzewo - w osobnym pliku otwierającym się na poziomie jupytera, jako że wygodniej jest przeglądać duży obrazek właśnie w takiej formie niż jako standardowy obrazek wewnątrz jupytera; wierzchołek opisany jest w formacie "1:x / 0:y", który oznacza, że w wierzchołku jest x obserwacji mających w polu "Survived" wartość 1 i y obserwacji mających w polu "Survived" wartość 0. Na krawędzi pokazywany jest split który do powstania tego wierzchołka doprowadził.

2 Metoda dla danych ciągłych

1. Dla danych ciągłych split na atrybucie C dzieli zbiór obserwacji na 2 części: Obserwacje z wartością x atrybutu C: $x \leq f$ i obserwacje z wartością x atrybutu C: $x > f$ dla f nazywanego dalej miejscem splita.

2. Miejsce splita jest wybierane jako takie miejsce, podział w którym maksymalizuje gain ratio. Proba znalezienia gain ratio opisanym powyżej (dla danych dyskretnych) algorytmem działałaby w złożoności $O(n^2)$ dla n będącego liczbą wierszy dataseta, co jest niestaysfakcjonujące.

3. Można także posortować dataset po atrybucie C, przesuwać potencjalne f do przodu, modyfikując dynamicznie listy wartości do obliczania warunkowej entropii i intristic info w czasie stałym i kalkulować je także w czasie stałym - złożoność zatem wyniesie $O(nlog(n))$

3 Działanie algorytmu - Logi (wersja na 5.0)

W kolejnych wierszach: wszystkie gain ratio dla atrybutów, w # wybrany atrybut do podziału, w przypadku atrybutu ciągłego punkt splita dający daną wartość gain ratio

```
New vertex is processed: this vertex contains
- 40 survived observations
- 60 deceased observations
The splits that led to the advent of this vertex were:
[]
column: Pclass, gain ratio: 0.05960489889898, attribute discrete
column: Sex, gain ratio: 0.40323636523376316, attribute discrete
column: Age, gain ratio: 0.16497395662937295, split value: 1
column: SibSp, gain ratio: 0.02511018959580578, attribute discrete
column: Parch, gain ratio: 0.01464001497707953, attribute discrete
############### Chosen attribute: Sex, value of gain: 0.40323636523376316

New vertex is processed: this vertex contains
- 7 survived observations
- 53 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male']
column: Pclass, gain ratio: 0.07729306659385578, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.4355418320733391, split value: 1
column: SibSp, gain ratio: 0.020306536706473522, attribute discrete
column: Parch, gain ratio: 0.03512786197394762, attribute discrete
############### Chosen attribute: Age, value of gain: 0.4355418320733391

New vertex is processed: this vertex contains
- 33 survived observations
- 7 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female']
column: Pclass, gain ratio: 0.12003370114124165, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.08066119202255055, split value: 40
```

```
column: SibSp, gain ratio: 0.16649361093302303, attribute discrete
column: Parch, gain ratio: 0.013322955978332338, attribute discrete
############### Chosen attribute: SibSp, value of gain: 0.16649361093302303

New vertex is processed: this vertex contains
- 1 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age <= 1']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 6 survived observations
- 53 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1']
column: Pclass, gain ratio: 0.07134448543370803, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.05346216522701795, split value: 40
column: SibSp, gain ratio: 0.016521875902366145, attribute discrete
column: Parch, gain ratio: 0.0438446848069315, attribute discrete
############### Chosen attribute: Pclass, value of gain: 0.07134448543370803

New vertex is processed: this vertex contains
- 11 survived observations
- 4 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1']
column: Pclass, gain ratio: 0.2997723162709534, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.17110871057416346, split value: 14
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.11576241483250045, attribute discrete
############### Chosen attribute: Pclass, value of gain: 0.2997723162709534

New vertex is processed: this vertex contains
- 19 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 0']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
```

```
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 2 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 3']
column: Pclass, gain ratio: 0.274017542121281, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 1.0, split value: 8
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.579380164285695, attribute discrete
############### Chosen attribute: Age, value of gain: 1.0

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 2']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 1 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 4']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 5']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe
```

```
New vertex is processed: this vertex contains
- 1 survived observations
- 37 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.05569120277229924, split value: 32
column: SibSp, gain ratio: 0.03969457544077502, attribute discrete
column: Parch, gain ratio: 0.07483597867418908, attribute discrete
############### Chosen attribute: Parch, value of gain: 0.07483597867418908

New vertex is processed: this vertex contains
- 4 survived observations
- 9 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 1']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.4686651882055394, split value: 40
column: SibSp, gain ratio: 0.222057265152258, attribute discrete
column: Parch, gain ratio: 0.059246976190095675, attribute discrete
############### Chosen attribute: Age, value of gain: 0.4686651882055394

New vertex is processed: this vertex contains
- 1 survived observations
- 7 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 2']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 1.0, split value: 12
column: SibSp, gain ratio: 0.04755786045881497, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### Chosen attribute: Age, value of gain: 1.0

New vertex is processed: this vertex contains
- 4 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1', 'Pclass = 1']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 5 survived observations
```

```
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1', 'Pclass = 2']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 2 survived observations
- 4 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1', 'Pclass = 3']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.4871971762327021, split value: 4
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.7336804366512111, attribute discrete
############### Chosen attribute: Parch, value of gain: 0.7336804366512111

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 3', 'Age <= 8']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 2 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 3', 'Age > 8']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 29 deceased observations
The splits that led to the advent of this vertex were:
```

```
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 0']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 1 survived observations
- 3 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 1']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 1.0, split value: 7
column: SibSp, gain ratio: 0.5408520829727552, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### Chosen attribute: Age, value of gain: 1.0

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 5']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 3 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 2']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 3']
column: Pclass, gain ratio: 0.0, attribute discrete
```

```
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 4 survived observations
- 2 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 1', 'Age <= 40']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0.16794857383343836, split value: 13
column: SibSp, gain ratio: 0.7336804366512111, attribute discrete
column: Parch, gain ratio: 0.3015619649304495, attribute discrete
############### Chosen attribute: SibSp, value of gain: 0.7336804366512111

New vertex is processed: this vertex contains
- 0 survived observations
- 7 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 1', 'Age > 40']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 1 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 2', 'Age <= 12']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 7 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 2', 'Age > 12']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
```

```
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 1 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1', 'Pclass = 3', 'Parch = 1']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 4 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1', 'Pclass = 3', 'Parch = 0']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 1 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = female', 'SibSp = 1', 'Pclass = 3', 'Parch = 5']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
############### No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 3 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 1', 'Age <= 7']

column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
```

```
column: Parch, gain ratio: 0.0, attribute discrete
################ No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 1 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 3', 'Parch = 1', 'Age > 7']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
################ No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 4 survived observations
- 0 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 1', 'Age <= 40', 'SibSp = 0']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
################ No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 1', 'Age <= 40', 'SibSp = 3']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
################ No chosen attribute, end of processing for this subset of dataframe

New vertex is processed: this vertex contains
- 0 survived observations
- 1 deceased observations
The splits that led to the advent of this vertex were:
['Sex = male', 'Age > 1', 'Pclass = 1', 'Age <= 40', 'SibSp = 1']
column: Pclass, gain ratio: 0.0, attribute discrete
column: Sex, gain ratio: 0.0, attribute discrete
column: Age, gain ratio: 0, split value: 0
column: SibSp, gain ratio: 0.0, attribute discrete
column: Parch, gain ratio: 0.0, attribute discrete
################ No chosen attribute, end of processing for this subset of dataframe
```

4 Działanie algorytmu - Generowane drzewa

4.1 Dyskretny Age

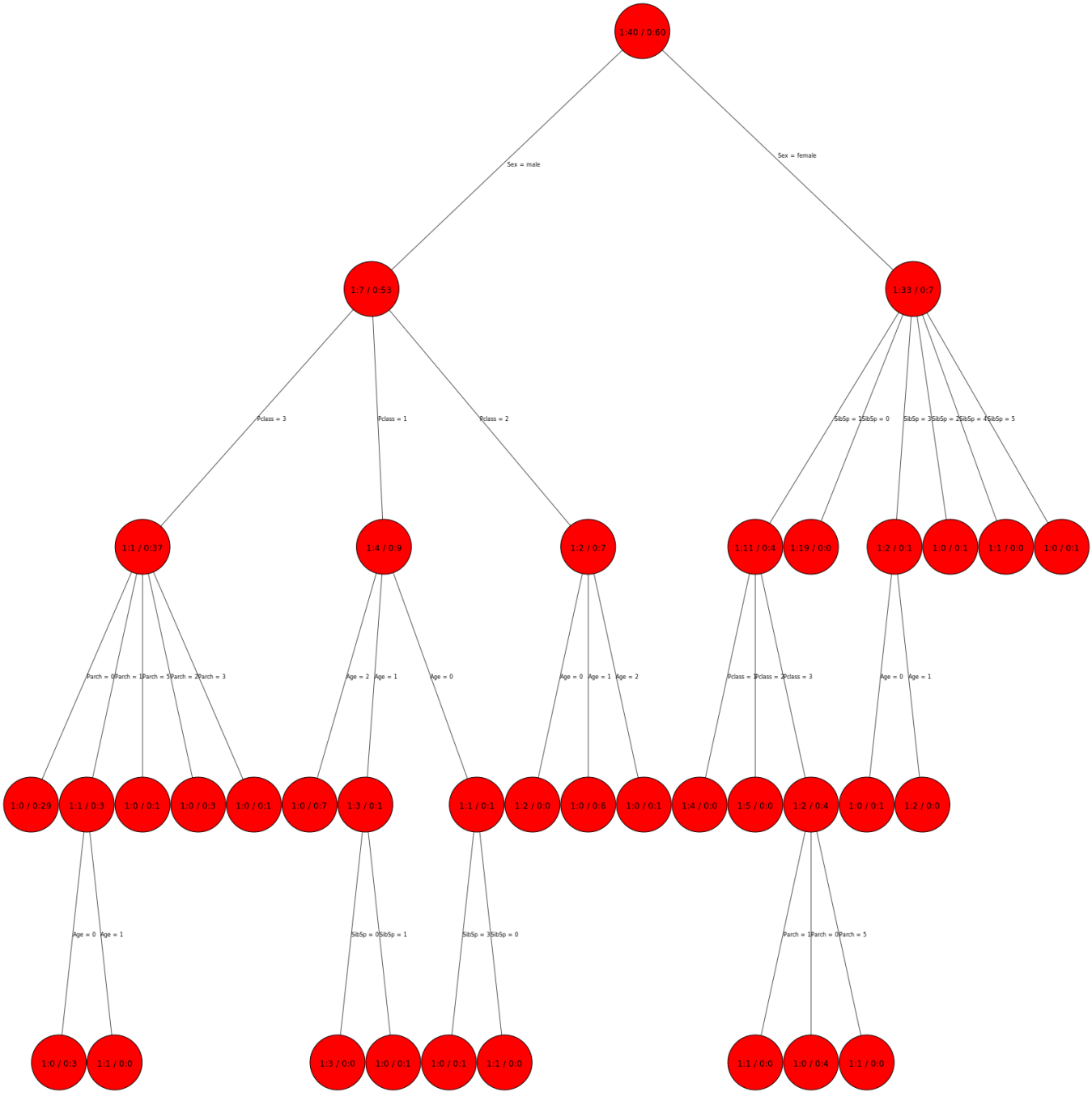

Figure 1: Age jest atrybutem dyskretnym

4.2 Ciągły Age

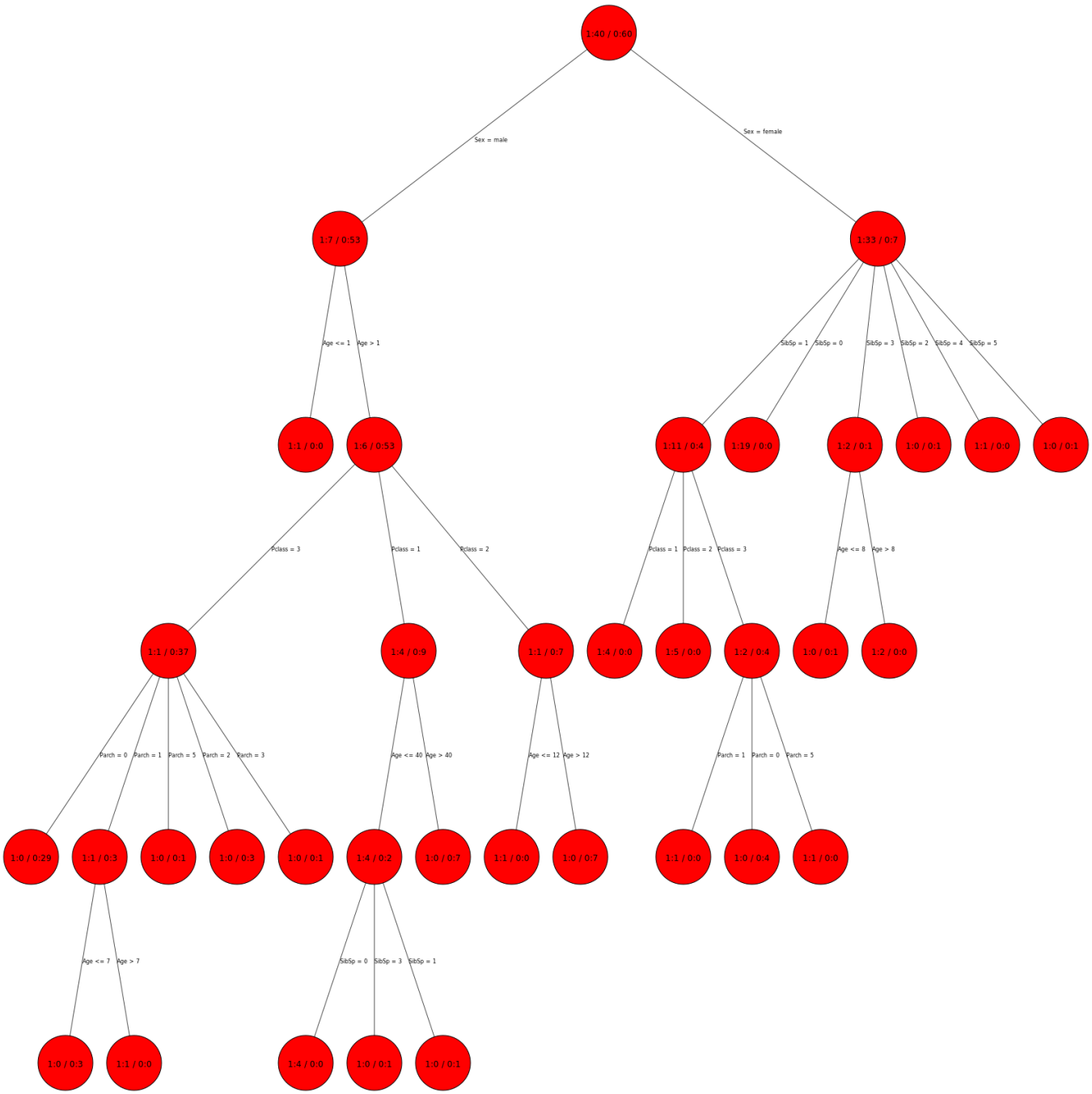

Figure 2: Age jest atrybutem ciągłym