Универзитет у Нишу

Електронски факултет Ниш

Интелигентни системи

Трећи домаћи задатак – Стабла одлучивања

Студент:

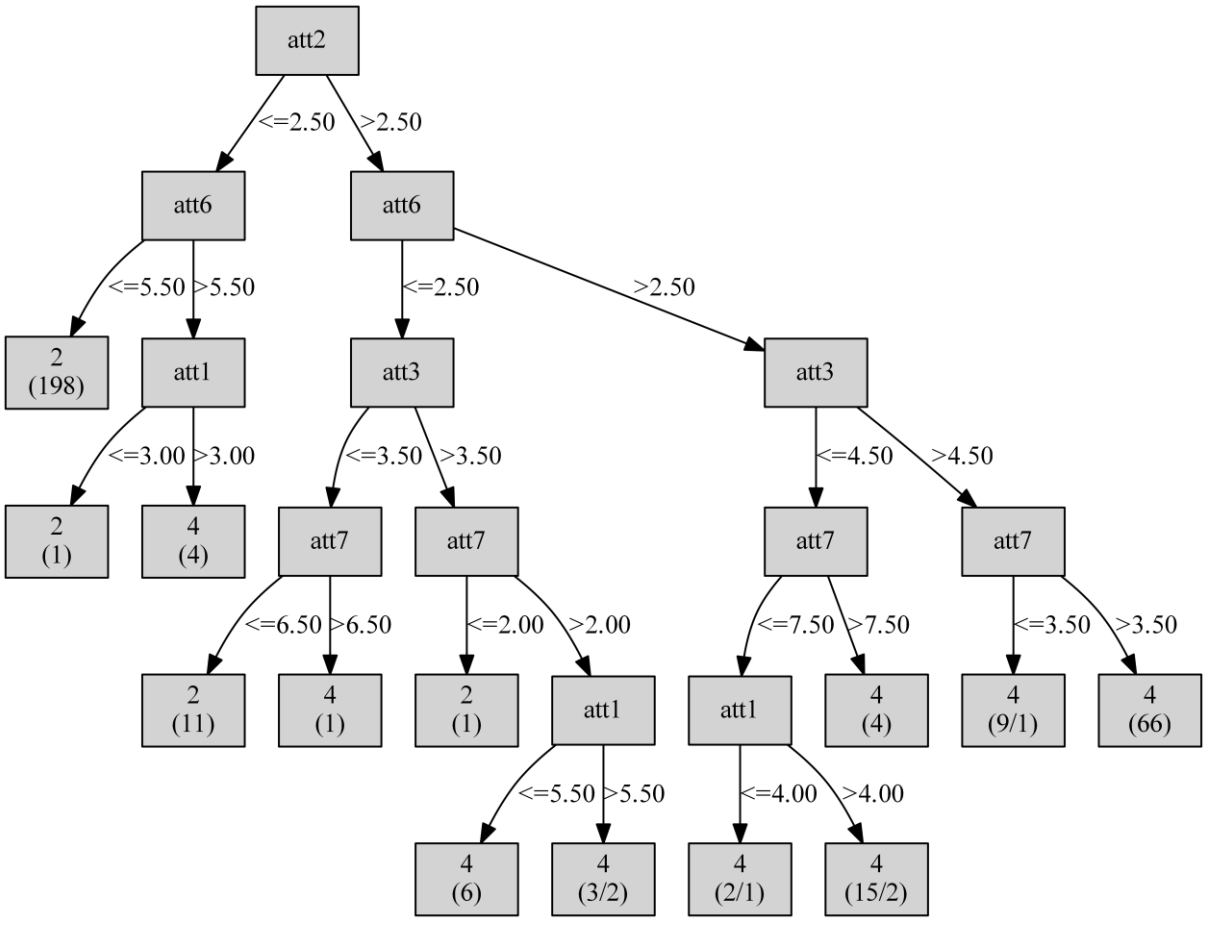
Димитрије Јовић, 928

Ниш, јануар 2020. год.

С А Д Р Ж А Ј

1. ФОРМУЛАЦИЈА ПРОБЛЕМА

На часу биологије, свако од ученика добио је задатак да изврши неку класификацију болести. Наши јунаци, Јанко и Миладин, договорили су се да овај задатак одраде у тиму. Као задатак, добили су да изврше класификацију тумора, односно да на основу параметара одреде да ли је тумор бенигни или малигни. У почетку су покушавали да тај проблем реше ручно, односно да уоче неки критеријум по којем би извршили класификацију, без асистенције рачунара. Међутим, пошто је скуп података био велики, да би се ручно извршила класификација, односно време потребно за извршење класификације би било велико, решили су да класификацију изврше уз помоћ рачунара. Миладин, који је добар програмер, знао је да постоје одређени алгоритми који тај посао могу да олакшају. Да би утврдили, који параметри утичу највише на процес класификације, решили су да подигну стабло одлуке. Међутим, били су у дилеми који алгоритам одабрати. Узимајући у обзир типове атрибута који се користе за класификацију, као и специфичности алгоритама, односно време извршења и прецизност класификације, одлучили су се да примене *CART* алгоритам. Такође, извршили су и обрада података. Захваљујући тим алгоритмима, процес само класификације је брзо завршен, а и проценат тачности класификације је био изузетно велики, што је резултирало томе да је су они одабрани да представе школу на недељи науке са пројектом класификације тумора.



2. ИМПЛЕМЕНТАЦИЈА СТАБЛА ОДЛУЧИВАЊА

Стабло има много аналогија у стварном животу, а испада да је утицало на широко подручје машинског учења, покривајући и класификацију и регресију. У анализи одлука, стабло одлуке може се користити за визуелно и експлицитно представљање одлука и начина одлучивања. Иако је често коришћен алат у анализи података за израду стратегије за постизање одређеног циља, он се такође широко користи у машинском учењу.

Приликом генерисања стабла одлуке, потребно је специфицирати неке параметре, као што су максимална дубина стабла, број листова у стаблу, критеријум на основу којег ће се вршити генерисање стабла одлуке, такође потребно је навести који је класни атрибут. Треба напоменути да су се у оквиру овог домаћег задатка користила подразумевана подешавања приликом генерисања стабла одлуке и за *CART* и за *ID3* алгоритам. Разлика између ова два алгоритма је у критеријуму на основу којег се креира стабло одлуке. Наиме, код *ID3* алгоритма као критерујум гранања се користи информацијски добит датог атрибута, док се код *CART* алгоритма користи *gini index*. Такође треба напоменути, да *ID3* алгоритам ради само уколико су атрибути номинални, док *CART* ради и са номиналним и са нумеричким типовима атрибута. Алгоритам *C4.5* није узет у разматрање јер је идентичан *CART* алгоритму, једина разлика је у индексу који се користи за гранање, у његовом случају је то *gain index*.

Што се тиче скупа података који је коришћен, скуп је преузет са http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 и бави се класификацијом тумора, односно да ли је тумор бенигни или малигни. Подаци су нумеричког типа, где се скуп састоји од 8 атрибута који утичи на процес класификације, једног класног атрибута, и атрибута који представља јединствени број пацијента. Такође, извршено је и препроцесирање података, односно програмски су уклоњени атрибути који не утичу на процес класификације, недефинисане вредности, односно вредности код којих је није унешена вредност, атрибута су замењене неком подразумеваном вредношћу. На слици испод дат је изглед скупа података.

	id	att1	att2	att3	att4	att5	att6	att7	att8	att9	class
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1018561	2	1	2	1	2	1	3	1	1	2
10	1033078	2	1	1	1	2	1	1	1	5	2
11	1033078	4	2	1	1	2	1	2	1	1	2
12	1035283	1	1	1	1	1	1	3	1	1	2
13	1036172	2	1	1	1	2	1	2	1	1	2
14	1041801	5	3	3	3	2	3	4	4	1	4
15	1043999	1	1	1	1	2	3	3	1	1	2
16	1044572	8	7	5	10	7	9	5	5	4	4
17	1047630	7	4	6	4	6	1	4	3	1	4
18	1048672	4	1	1	1	2	1	2	1	1	2
19	1049815	4	1	1	1	2	1	3	1	1	2
20	1050670	10	7	7	6	4	10	4	1	2	4
21	1050718	6	1	1	1	2	1	3	1	1	2
22	1054590	7	3	2	10	5	10	5	4	4	4
23	1054593	10	5	5	3	6	7	7	10	1	4
24	1056784	3	1	1	1	2	1	2	1	1	2
25	1057013	8	4	5	1	2	?	7	3	1	4

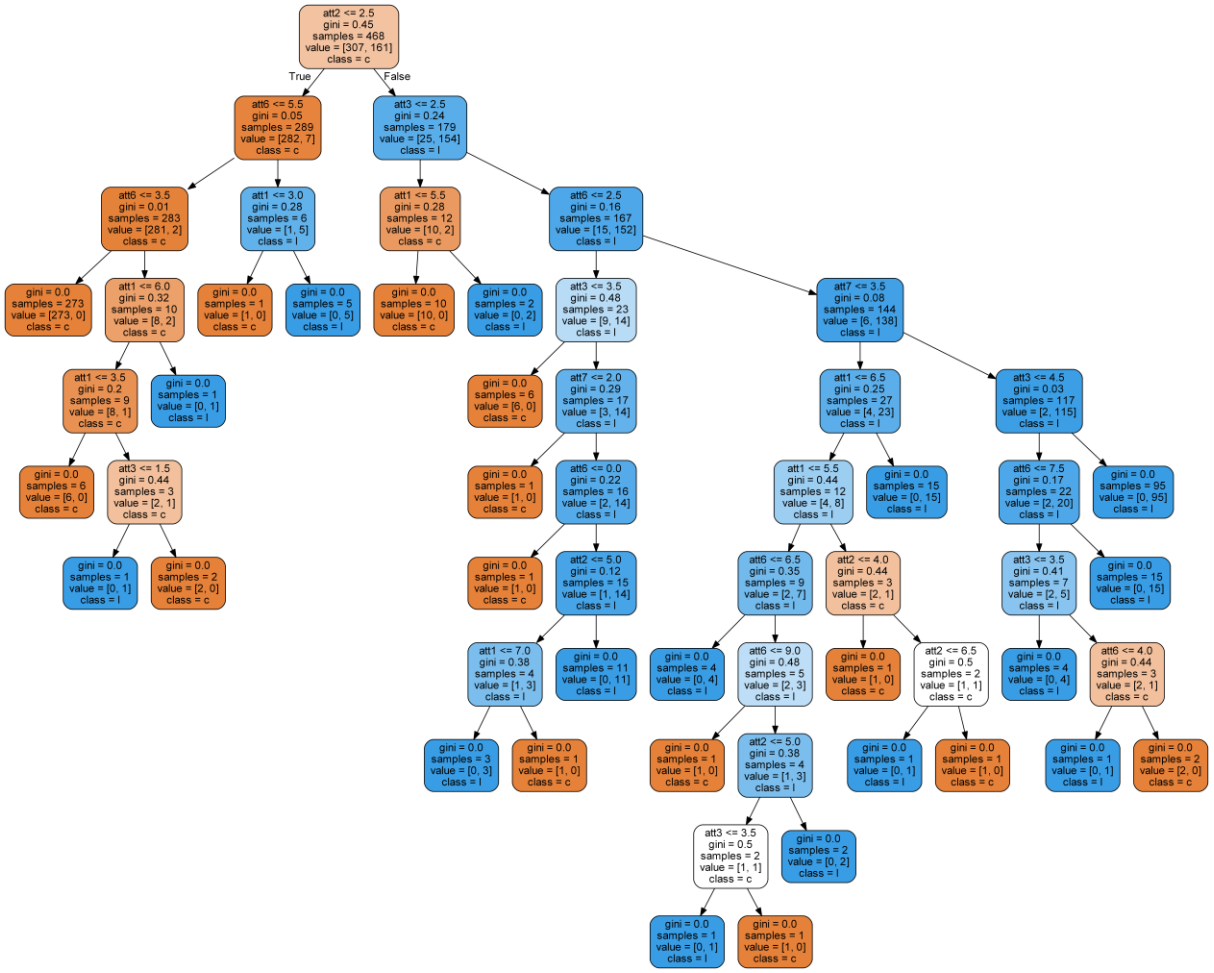
Слика 1: Скуп података

На слици испод дат је приказ прецизности и времена извршења алгоритма. Узимајући у обзир да је у оквиру библиотеке која садржи *ID3* алгоритам имплементирана и дискретизација атрибута, тај алгоритам је знатно спорији од *CART* алгоритма, јер је скуп података нумеричког типа, док им се прецизност класификације незнатно разликује.

```
Elapsed time CART:  0.0009975433349609375 seconds
Accuracy of CART: 0.9393939393939394
Elapsed time ID3:  0.014926433563232422 seconds
Accuracy of ID3: 0.9523809523809523
```

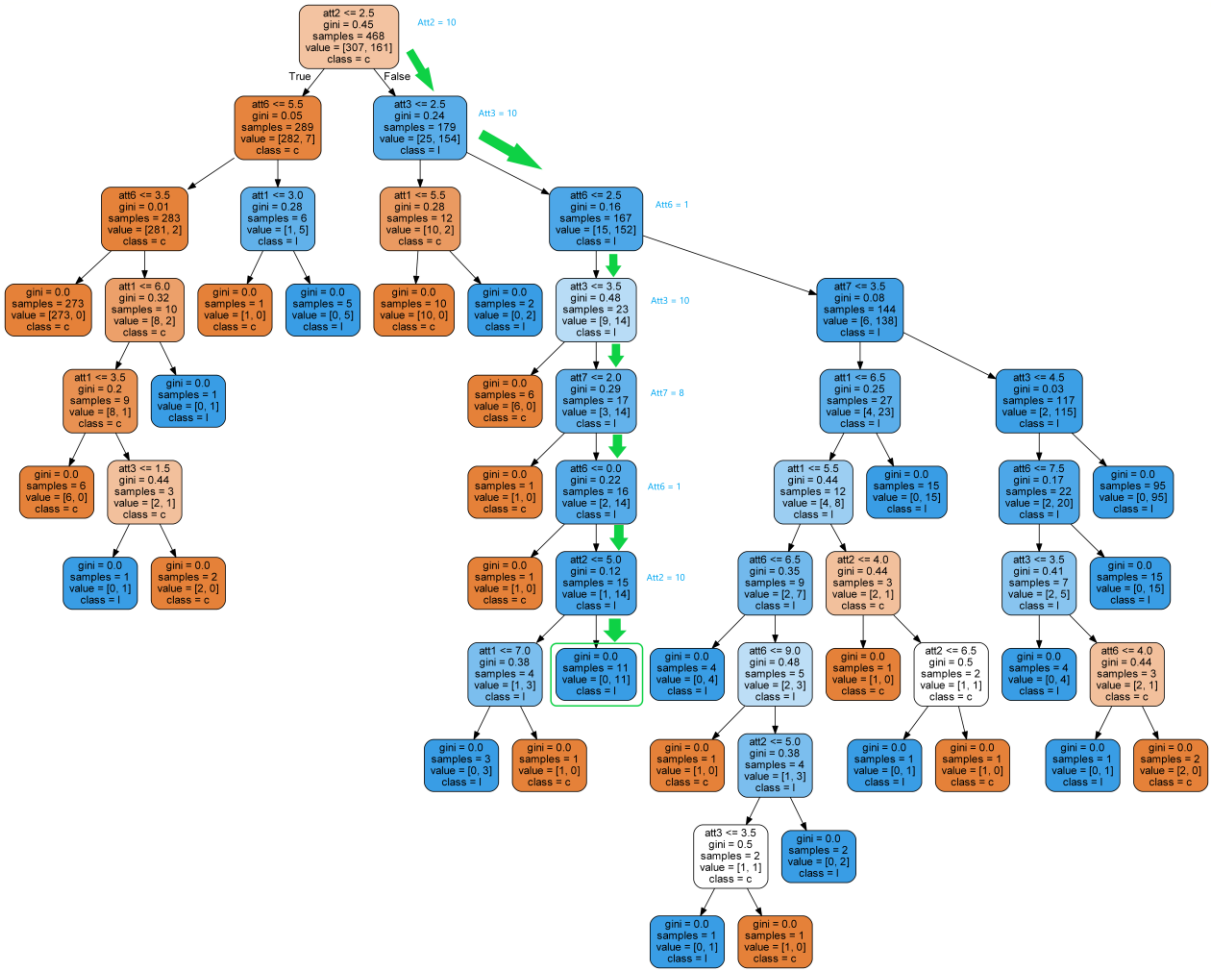
Слика 2: Перформансе алгоритама

На основу скупа података, специфичности и перформанси алгоритама, одлучено је да се за класификацију користи *CART* алгоритам. На слици испод је приказано стабло одлуке, након чега ће бити описан начин класификације.



Слика 3: Стабло генерисано *CART* алгоритмом

За податак **1103608,10,10,10,4,8,1,8,10,1,?** где није познато којој класи припада. На основу стабла одлучивања могуће је одредити припадност класи. На слици испод приказан је процес одлучивања. Након одлучивања утврђено је да податак припада класи малигних тумора.



Слика 4: Процес одлучивања

3. РЕФЕРЕНЦЕ

[1] https://scikit-learn.org/stable/modules/tree.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[3] https://pypi.org/project/decision-tree-id3/

[4] https://scikit-learn.org/stable/modules/feature_selection.html#removing-features-with-low-variance

[5] https://pandas.pydata.org

[6] https://pandas.pydata.org/pandas-docs/stable/reference/frame.html

[7] https://numpy.org/devdocs/