Центар за пословно одлучивање
Факултет организационих наука
Универзитет у Београду

ID3 алгоритам

ID3[1] алгоритам представља алгоритам за изградњу *дрва одлучивања*[2] на основу података. Први пут га је описао Рос Кинлан (Ross Quinlan) са Сиднејског универзитета, 1975. године у књизи *Machine Learning,* [1].

Дрво одлучивања представља графички приказ начина на који доносилац одлуке одлучује. Традиционално се израђује ручно од стране експерта из области, који из искуства у одлучивању генерише дрво које се може искористити за доношење одлука у будућности.

ID3 алгоритам покушава да замени експерта у прављењу дрва одлучивања, тако што ће дрво генерисати из података којима су описани ранији случајеви одлучивања у сличним ситуацијама. То значи да је за рад ID3 алгоритма неопходно постојање записа о историји одлучивања (случајева у прошлости), најчешће у табеларном облику:

дан	**врем. прилике**	**температура**	**влажност**	**ветар**	**играти?**
1	сунчано	вруће	висока	слаб	не
2	сунчано	вруће	висока	јак	не
3	облачно	вруће	висока	слаб	да
4	кишовито	топло	висока	слаб	да
5	кишовито	хладно	нормална	слаб	да
6	кишовито	хладно	нормална	јак	не
7	облачно	хладно	нормална	јак	да
8	сунчано	топло	висока	слаб	не
9	сунчано	хладно	нормална	слаб	да
10	кишовито	топло	нормална	слаб	да
11	сунчано	топло	нормална	јак	да
12	облачно	топло	висока	јак	да
13	облачно	вруће	нормална	слаб	да
14	кишовито	топло	висока	јак	не

Горе наведеном табелом се описују случајеви из прошлости, када се на основу временских прилика одлучивало да ли играти неку игру која се игра на отвореном простору. Атрибути (колоне) појединачно описују компоненте временских прилика сваког случаја (реда), док атрибут *играти* представља одлуку у том случају и назива се *атрибут одлуке*[3].

Основни проблем је: *какву одлуку донети када се у будућности догоде некакве временске прилике?*

[1] акроним од: Iterative Dichotomiser 3 („Tree“)
[2] у преводу често се назива и Стабло одлучивања (eng. Decision tree)
[3] неретко се јављају и синоними за овај атрибут, на пример: *output, target, label*, итд.

Пошто овакве историје случајева могу бити доста велике, претраживање целе базе случајева за случајем који највише одговара новонасталом не долази у обзир. Са друге стране, дрво одлучивања које одговара подацима о случајевима могло би знатно олакшати доношење одлуке. Зато се поставља следећи проблем: *како генерисати дрво одлучивања на основу података које ће бити што је могуће лакше за доношење одлука?* Јасно је да је такво дрво што је могуће „мање“, тј. садржи минималан број чворова и грана. Другачије речено, тражи се дрво које у просеку има што мањи број корака којима се долази до одлуке[4]. ID3 алгоритам управо даје одговор на ово питање.

Алгоритам ради итеративно, тако да ће до решења долазити у више корака.
У првом кораку треба одредити атрибут којим ће се дрво гранати *у корену*. За овај проблем се уводи концепт ентропије, позајмљен из Информационе теорије.

Ентропија представља меру неуређености система, тј. у нашем случају, неизвесности о томе коју одлуку треба донети.
Један од начина да се ентропија измери је и путем следеће формуле коју је дефинисао Клод Шенон (C. Shannon)[5]:

$$H(S) = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i)$$

где $H(S)$ представља ентропију скупа случајева S, p_i вероватноћу да ће бити донешена одлука i, а n број различитих одлука које могу бити донешене.
Вероватноће се могу рачунати преко формуле:

$$p_i = \frac{|C_i|}{|S|}$$

где је $|C_i|$ број случајева са одлуком i, а $|S|$ укупан број случајева.

У нашем случају, почетна ентропија износи:

$$H(S) = -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right] = 0,940$$

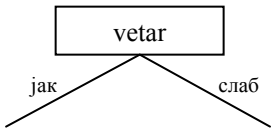
напомена: следећа математичка релација може бити од користи када се логаритми рачунају на ручним калкулаторима:

$$\log_2 X = \frac{\log_{10} X}{\log_{10} 2} = \frac{\log_{10} X}{0,301}$$

[4] ово представља примену познатог Окамовог принципа (Occam's razor), који преферира једноставније теорије и објашњења у односу на сложене.
[5] по овој формули ентропија представља просечан број битова потребних да би се идентификовало стање система (одлука)

Вратимо се на питање избора атрибута за гранање дрва одлучивања.
Увођењем атрибута X у корену дрво се грана на више грана, тј. почетни скуп случајева се дели на више подскупова (у општем случају, дрво се грана на онолико грана колико има вредности атрибута X).

На пример, почетно стабло из примера можемо преко атрибута *ветар* гранати на две гране:



Ентропија у систему после увођења атрибута X у корену за гранање дрвета рачуна се по следећој формули:

$$H(X,S)=\sum_{i=1}^{n}\left(\frac{|S_i|}{|S|}\cdot H(S_i)\right)$$

где су: $|S_i|$ број елемената са особином X_i , $|S|$ укупан број случајева у скупу S, а $H(S_i)$ ентропија подскупа случајева са особином X_i

У нашем примеру, гранајући дрво са атрибутом *ветар* систем ће имати следећу количину ентропије:

$$H(vetar,S)=\frac{|S_{jak}|}{|S|}\cdot H(S_{jak})+\frac{|S_{slab}|}{|S|}\cdot H(S_{slab})=\frac{6}{14}\cdot H(S_{jak})+\frac{8}{14}\cdot H(S_{slab})=0,892$$

пошто је:

$$H(S_{jak})=-\left[0,5\cdot\log_2(0,5)+0,5\cdot\log_2(0,5)\right]=1$$

$$H(S_{slab})=-\left[0,75\cdot\log_2(0,75)+0,25\cdot\log_2(0,25)\right]=0,811$$

Примећујемо да се ентропија смањила у односу на почетну ентропију у систему, што је и очекивано, с обзиром да смо открившри вредност атрибута X дошли до одређене информације о систему. Количина информације коју смо добили откривајући вредност атрибута може се и измерити следећом формулом:

$$I(X,S)=H(S)-H(X,S)$$

где се $I(X,S)$ назива **информациона добит** за одлуку од атрибута x

У нашем примеру имамо да је

$$I(vetar,S)=0,940-0,892=0,048$$

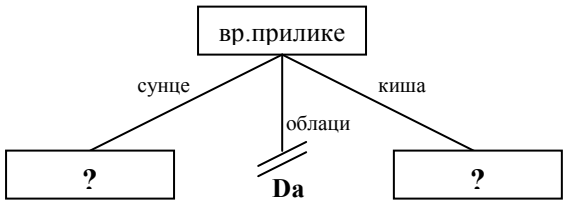
Дакле, информациона добит нам показује колико се ентропија система смањује ако познајемо вредност атрибута X.

Идеја ID3 алгоритма је да за гранање дрва одлучивања одабере онај атрибут који носи највише информација о одлуци, тј. чија је **информациона добит највећа**.

У нашем примеру, рачунајући информационе добити за преостале атрибуте добијамо:

$$I\left(vr.prilike, S\right) = 0,246$$
$$I\left(vlažnost, S\right) = 0,136$$
$$I\left(temperatura, S\right) = 0,030$$
$$I\left(vetar, S\right) = 0,048$$

Дакле, по ID3 алгоритму, бирамо *вр.прилике* за атрибут којим ћемо гранати дрво одлучивања:



Даље, итеративно примењујемо исти поступак за сваку грану, са циљем да гранамо дрво до нивоа када је одлука сигурна.

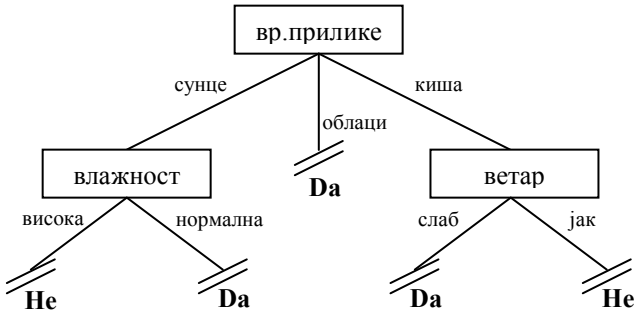
Ако погледамо наш пример, примећујемо да је у подскупу облачних дана одлука увек била да се игра. Ентропија у том подсистему је једнака нули и ту даље гранање није потребно. На том делу дрва означавамо да је одлука да се игра.
На преосталим гранама ентропија је већа од нуле, па ћемо ту увести неки од преосталих атрибута да би даље гранали дрво. Поступак, значи, понављамо посебно за подскуп сунчаних и за подскуп кишовитих дана.

У општем случају, поступак се понавља док не дођемо до сигурне одлуке, или нам понестане атрибута за гранање, ако су сви атрибути искоришћени.

На подскупу случајева сунчаних дана (лева грана стабла), информационе добити преосталих атрибута су:

$$I\left(vlažnost, S_{sunce}\right) = 0,971$$
$$I\left(temperatura, S_{sunce}\right) = 0,571$$
$$I\left(vetar, S_{sunce}\right) = 0,079$$

Дакле, узимамо *влажност* да поново гранамо дрво на грани сунчаних дана.
Са друге стране, за кишовите дане, истим поступком, бирамо атрибут *ветар*.
После спровођења целог алгоритма, коначно дрво има изглед:



Овим је поступак примене ID3 алгоритма завршен.
Оно што се може приметити је да атрибут *температура* није коришћен у дрвету зато што, по ID3 алгоритму, не носи довољно информација о одлуци. Даље, до одлуке се долази у највише два корака.

Посебна погодност дрва одлучивања јесте што се лако може претворити у низ „ако–онда“ правила. У нашем примеру имали бисмо неке од следећих правила:

IF (*вр.прилике=сунчано*) AND (*влажност=висока*) THEN *играти=не*
IF (*вр.прилике=облачно*) THEN *играти=да*
итд.

Треба још напоменути да је ID3 алгоритам у ствари хеуристика. То значи да њиме не морамо доћи до најбољег (најмањег) стабла, већ да се добија стабло које је „довољно добро“. Ипак, примена овог алгоритма у индустрији показује да је у пракси доста ефективан. Неке од досадашњих примена су: медицинска дијагноза, процена ризика код одобравања кредита, откривање грешака у опреми, Веб претраживање, итд.

ID3 алгоритам има и својих мана. Неки од тих недостатака су отклоњени у нешто млађим алгоритмима који се базирају на ID3 алгоритму. „C4.5“ алгоритам истог аутора уноси нека побољшања која се односе на рад са недостајућим вредностима, рад са атрибутима који имају контунуалне вредности, скраћивање дрва, израда правила, итд.

За детаљније упознавање са ID3 алгоритмом погледати наведене референце.

Пример „преживљавање“

Проблем: На основу описа плода из природе треба одлучити да ли је сигурно хранити се истим. [6]

Табела случајева има следећи изглед:

воће	кожа	боја	величина	месо	закључак
1	длакаво	браон	велико	тврдо	сигурно
2	длакаво	зелена	велико	тврдо	сигурно
3	глатко	црвена	велико	меко	опасно
4	длакаво	зелена	велико	меко	сигурно
5	длакаво	црвена	мало	тврдо	сигурно
6	глатко	црвена	мало	тврдо	сигурно
7	глатко	браон	мало	тврдо	сигурно
8	длакаво	зелена	мало	меко	опасно
9	глатко	зелена	мало	тврдо	опасно
10	длакаво	црвена	велико	тврдо	сигурно
11	глатко	браон	велико	меко	сигурно
12	глатко	зелена	мало	меко	опасно
13	длакаво	црвена	мало	меко	сигурно
14	глатко	црвена	велико	тврдо	опасно
15	глатко	црвена	мало	тврдо	сигурно
16	длакаво	зелена	мало	тврдо	опасно

Треба генерисати дрво одлучивања применом ID3 алгоритма.

У првом кораку одређујемо атрибут за гранање у корену. Одређујемо прво почетну ентропију:

$$H(S)=\frac{|S_{sigurno}|}{|S|}\cdot\log_2\frac{|S_{sigurno}|}{|S|}+\frac{|S_{opasno}|}{|S|}\cdot\log_2\frac{|S_{opasno}|}{|S|}=\frac{10}{16}\cdot\log_2\frac{10}{16}+\frac{6}{16}\cdot\log_2\frac{6}{16}=0,954$$

а затим и ентропије после увођења сваког атрибута појединачно:

$$H(koža,S)=\frac{|S_{dlakavo}|}{|S|}\cdot H(S_{dlakavo})+\frac{|S_{glatko}|}{|S|}\cdot H(S_{glatko})=\frac{8}{16}\cdot H(S_{dlakavo})+\frac{8}{16}\cdot H(S_{glatko})$$

$$H(S_{dlakavo})=-\left[\frac{6}{8}\cdot\log_2\frac{6}{8}+\frac{2}{8}\cdot\log_2\frac{2}{8}\right]=0,811$$

$$H(S_{glatko})=-\left[\frac{4}{8}\cdot\log_2\frac{4}{8}+\frac{4}{8}\cdot\log_2\frac{4}{8}\right]=1$$

$$H(koža,S)=0,906$$

[6] пример преузет из [6]

$$H(veličina, S) = \frac{|S_{veliko}|}{|S|} \cdot H(S_{veliko}) + \frac{|S_{malo}|}{|S|} \cdot H(S_{malo}) = \frac{7}{16} \cdot H(S_{veliko}) + \frac{9}{16} \cdot H(S_{malo})$$

$$H(S_{veliko}) = -\left[\frac{5}{7} \cdot \log_2 \frac{5}{7} + \frac{2}{7} \cdot \log_2 \frac{2}{7}\right] = 0,863$$

$$H(S_{malo}) = -\left[\frac{5}{9} \cdot \log_2 \frac{5}{9} + \frac{4}{9} \cdot \log_2 \frac{4}{9}\right] = 0,991$$

$$H(veličina, S) = 0,935$$

$$H(meso, S) = \frac{|S_{tvrdo}|}{|S|} \cdot H(S_{tvrdo}) + \frac{|S_{meko}|}{|S|} \cdot H(S_{meko}) = \frac{10}{16} \cdot H(S_{tvrdo}) + \frac{6}{16} \cdot H(S_{meko})$$

$$H(S_{tvrdo}) = -\left[\frac{7}{10} \cdot \log_2 \frac{7}{10} + \frac{3}{10} \cdot \log_2 \frac{3}{10}\right] = 0,881$$

$$H(S_{meko}) = -\left[\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6}\right] = 1$$

$$H(meso, S) = 0,926$$

$$H(boja, S) = \frac{3}{16} \cdot H(S_{braon}) + \frac{6}{16} \cdot H(S_{zeleno}) + \frac{7}{16} \cdot H(S_{crveno})$$

$$H(S_{braon}) = -\left[\frac{3}{3} \cdot \log_2 \frac{3}{3} + \frac{0}{3} \cdot \log_2 \frac{0}{3}\right] = 0$$

$$H(S_{zeleno}) = -\left[\frac{2}{6} \cdot \log_2 \frac{2}{6} + \frac{4}{6} \cdot \log_2 \frac{4}{6}\right] = 0,915$$

$$H(S_{crveno}) = -\left[\frac{5}{7} \cdot \log_2 \frac{5}{7} + \frac{2}{7} \cdot \log_2 \frac{2}{7}\right] = 0,863$$

$$H(boja, S) = 0,721$$

Даље рачунамо информационе добити, које ћемо користити за избор атрибута за гранање:

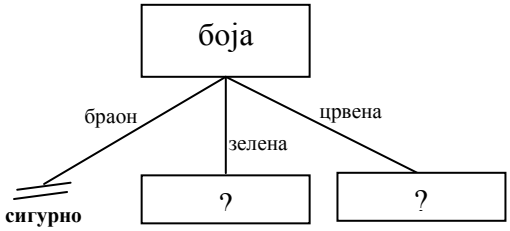
$$I(koža, S) = H(S) - H(koža, S) = 0,954 - 0,906 = 0,048$$

$$I(veličina, S) = H(S) - H(veličina, S) = 0,954 - 0,935 = 0,019$$

$$I(meso, S) = H(S) - H(meso, S) = 0,954 - 0,926 = 0,028$$

$$I(boja, S) = H(S) - H(boja, S) = 0,954 - 0,721 = 0,233$$

Дакле, бирамо атрибут *боја* за гранање. Видимо да је за браон боју ентропија спуштена на нулу, што значи да даље гранање ове гране није потребно. Преостале две гране треба још гранати, што ћемо чинити у наставку. Дрво засад има изглед:



Настављамо са гранањем подскупа случајева (воћки) које имају зелену боју. Одређујемо ентропије, а затим и информационе добити за све преостале атрибуте:

$$H\left(koža, S_{zeleno}\right) = \frac{\left|S_{dlakavo,zeleno}\right|}{\left|S_{zeleno}\right|} \cdot H\left(S_{dlakavo,zeleno}\right) + \frac{\left|S_{glatko,zeleno}\right|}{\left|S_{zeleno}\right|} \cdot H\left(S_{glatko,zeleno}\right) =$$

$$= \frac{4}{6} \cdot H\left(S_{dlakavo,zeleno}\right) + \frac{2}{6} \cdot H\left(S_{glatko,zeleno}\right)$$

$$H\left(S_{dlakavo,zeleno}\right) = -\left[\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4}\right] = 1$$

$$H\left(S_{glatko,zeleno}\right) = -\left[\frac{0}{2} \cdot \log_2 \frac{0}{2} + \frac{2}{2} \cdot \log_2 \frac{2}{2}\right] = 0$$

$$H\left(koža, S_{zeleno}\right) = 0,667$$

$$H\left(veličina, S_{zeleno}\right) = \frac{2}{6} \cdot H\left(S_{veliko,zeleno}\right) + \frac{4}{6} \cdot H\left(S_{malo,zeleno}\right)$$

$$H\left(S_{veliko,zeleno}\right) = -\left[\frac{2}{2} \cdot \log_2 \frac{2}{2} + \frac{0}{2} \cdot \log_2 \frac{0}{2}\right] = 0$$

$$H\left(S_{malo,zeleno}\right) = -\left[\frac{0}{4} \cdot \log_2 \frac{0}{4} + \frac{4}{4} \cdot \log_2 \frac{4}{4}\right] = 0$$

$$H\left(veličina, S_{zeleno}\right) = 0$$

$$H\left(meso, S_{zeleno}\right) = \frac{3}{6} \cdot H\left(S_{tvrdo,zeleno}\right) + \frac{3}{6} \cdot H\left(S_{meko,zeleno}\right)$$
$$H\left(S_{tvrdo,zeleno}\right) = -\left[\frac{1}{3} \cdot \log_2 \frac{2}{3} + \frac{2}{3} \cdot \log_2 \frac{2}{3}\right] = 0,915$$
$$H\left(S_{meko,zeleno}\right) = -\left[\frac{1}{3} \cdot \log_2 \frac{1}{3} + \frac{2}{3} \cdot \log_2 \frac{2}{3}\right] = 0,915$$
$$H\left(meso, S_{zeleno}\right) = 0,915$$

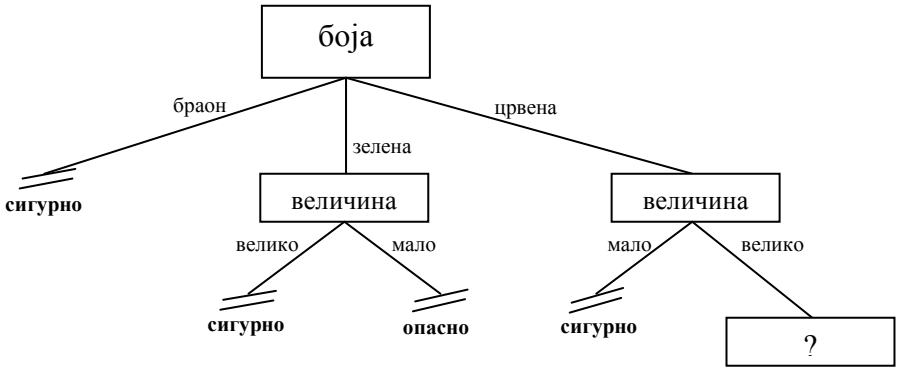
$$I\left(koža, S_{zeleno}\right) = H\left(S_{zeleno}\right) - H\left(koža, S_{zeleno}\right) = 0,915 - 0,667 = 0,248$$
$$I\left(veličina, S_{zeleno}\right) = H\left(S_{zeleno}\right) - H\left(veličina, S_{zeleno}\right) = 0,915 - 0 = 0,915$$
$$I\left(meso, S_{zeleno}\right) = H\left(S_{zeleno}\right) - H\left(meso, S_{zeleno}\right) = 0,915 - 0,915 = 0$$

Дакле, грану са зеленом бојом даље гранамо са атрибутом *величина*. Видимо да су у том случају велике воћке сигурне, док су мале опасне.

За подскуп црвених воћки добијамо следеће информационе добити:

$$I\left(koža, S_{crveno}\right) = H\left(S_{crveno}\right) - H\left(koža, S_{crveno}\right) = 0,863 - 0,571 = 0,292$$
$$I\left(veličina, S_{crveno}\right) = H\left(S_{crveno}\right) - H\left(veličina, S_{crveno}\right) = 0,863 - 0,392 = 0,471$$
$$I\left(meso, S_{crveno}\right) = H\left(S_{crveno}\right) - H\left(meso, S_{crveno}\right) = 0,863 - 0,801 = 0,062$$

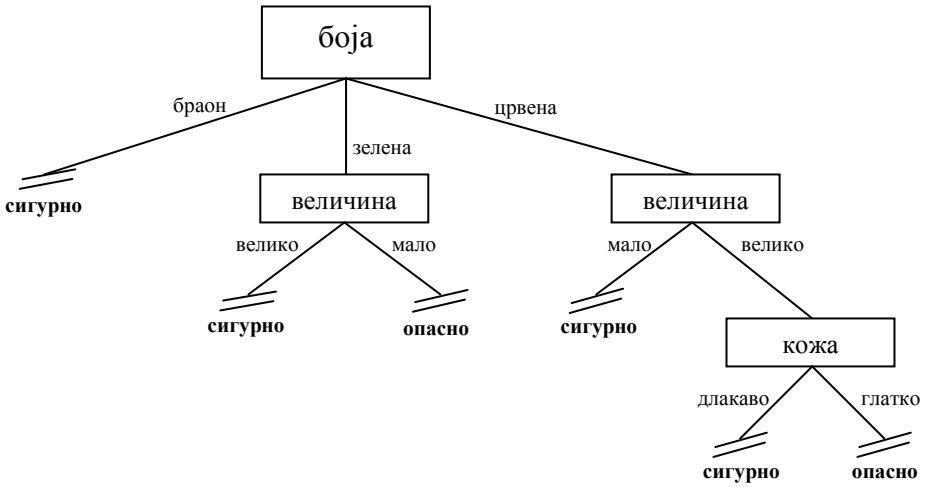
Дакле, грану са црвеном бојом даље гранамо такође са атрибутом *величина*. У овом случају за мале воћке можемо рећи да су сигурне, док за велике не можемо са сигурношћу рећи. Пошто су преостала још два атрибута (кожа и месо), настављамо гранање за црвене велике воћке. Дрво у овом тренутку изгледа на следећи начин:



Информационе добити за атрибуте на подскупу црвених великих воћки рачунамо на исти начин и добијамо:

$$I\left(koža, S_{crveno,veliko}\right) = H\left(S_{crveno,veliko}\right) - H\left(koža, S_{crveno,veliko}\right) = 0,915 - 0 = 0,915$$
$$I\left(meso, S_{crveno,veliko}\right) = H\left(S_{crveno,veliko}\right) - H\left(meso, S_{crveno,veliko}\right) = 0,915 - 0,667 = 0,248$$

Коначно дрво одлучивања има изглед:



Овим је задатак завршен.

Напомена: подаци за овај пример су фиктивни и знање добијено применом ID3 алгоритма над њима није за употребу.

Референце:

[1] Quinlan, J. R. (1986). Induction of decision trees, *Machine Learning* 1(1): 81–106.

[2] Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal,* 379–423 and 623–656.
http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html

[3] Course materials: CSE5230 Tutorial: The ID3 Decision Tree Algorithm
Monash University, Faculty of Information Technology, Australia
http://www.csse.monash.edu.au/courseware/cse5230/assets/tutorials/decisiontreesTute.pdf

[4] Public material: *An Implementation of ID3 --- Decision Tree Learning Algorithm*
Wei Peng, Juhua Chen and Haiping Zhou
University of New South Wales, School of Computer Science & Engineering, Sydney, Australia
http://wwwpeople.arch.usyd.edu.au/~wpeng/DecisionTree2.pdf

[5] Course materials: *CIS 587: Introduction to Artificial Intelligence*
Computer and informational sciences, Temple University
Dr. Giorgio P. Ingargiola
http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html

[6] Web material: *Data Mining*, Peter Ross
Faculty of Engineering and Computing, Napier University
http://www.dcs.napier.ac.uk/~peter/vldb/dm/node11.html