KLASIFIKIMI I STATUSIT EKONOMIK ME ANË TË PEMËVE VENDIM-MARRËSE, NJË RAST STUDIMI I FAMILJEVE SHQIPTARE

***QENDRAJ D., XHAFAJ E., MUKLI L.**

Universiteti "Aleksandër Moisiu" Durrës, Fakulteti i Teknologjisë së Informacionit, Departamenti i Matematikës

e-mail: daniela_qendraj@hotmail.com

Përmbledhje

Ndërtimi i pemëve vendim-marrëse nga disa të dhëna është një disiplinë e largët në kohë. Pema vendim-marrëse zakonisht është përdorur për klasifikim sepse ajo ka një strukturë hierarkike të thjeshtë për përdoruesin, duke kuptuar kështu edhe vendim-marrësin. Në këtë artikull ne do të klasifikojmë gjendjen ekonomike të familjeve shqiptare duke përdorur algoritmin e pemës vendim-marrëse që është ID3 algoritëm, dhe më pas të aplikojmë të dhënat edhe në paketën Weka. Kemi gjetur që si rrënjë e pemës vendim-marrëse shërben numri i personave për familje, i cili është po ashtu edhe më i rëndësishmi atribut ndër të tjerët. Kemi aplikuar Wekën për të ndërtuar pemën binare për këto raste dhe kemi parë se pema binare është më e vogël se pema vendim-marrëse.

Abstract

The construction of the decision trees from data is a longstanding discipline. The decision tree has been generally used for classification, because it is the simple hierarchical structure for the user to understand the decision maker. In this paper we will classify the economic state of the Albanian Family, by using the decision tree algorithm which is ID3 Algorithm, and then applying the data even in the Weka software. We have found that the root of the decision tree is the number of the persons for a family, which is also the most important from the others attribute. We have run with Weka the binary tree for these cases, and we have seen that the binary tree is smaller than the simple decision tree.

Fjalëkyçe. Pemë vendim-marrëse, ID3 algoritëm, paketa Weka.

Hyrje

Vendim-marrja është përzgjedhja e alternativës më të mirë për të arritur në një rezultat optimal në kushtet e një situate sado pak të ndërlikuar. Në këto kushte vendim-marrësi, qoftë ky një person apo një organ drejtues shqyrton të gjitha alternativat nga këndvështrime apo kritere të ndryshme, të cilat përcaktojnë vlerën e alternativave. Në këto rrethana shtrohet problemi i marrjes së një vendimi optimal, ku përmendim metodat e vendim-marrjes shumëkriterëshe (MCDM→ vendim-marrja shumë përmasore).

Problemat vendim-marrëse janë zgjidhur gjërësisht me anë të modeleve të programimit linear, në të cilat numri i zgjidhjeve të mundshme është zakonisht i

madh. Nga këto modele është shumë e vështirë të evidentohen zgjidhjet e lejueshme, pra ato alternativa që i duhen vendim-marrjes. Në metodat e MCDM edhe pse numri i alternativave është i vogël, ato janë të disponueshme që në momentin fillestar të shtrimit të problemit. Në praktikë kriteret e vlerësimit të alternativave mund të jenë me natyra cilësore ose sasiore. Kur kriteret janë cilësore ato nuk mund të shprehen numerikisht, pasi kriteret nuk kanë njësi matëse të përbashkëta. Ndaj në këto kushte problemet e vendim-marrjes nuk mund të trajtohen me modelet e njohura lineare.

Një nga modelet e MCDM është ndërtimi i pemës vendim-marrëse, si një proçes renditje dhe klasifikimi njëkohësisht, duke pasur në krye vendim-marrësin (Quinlan 1986). ID3 është i pari algoritëm vendim-marrës që është ndërtuar nga Ross Quinlan, ky algoritëm ndërton një pemë vendim-marrëse nga një bashkësi fikse të dhënash, zakonisht diskrete. Gjethja e pemës përmban emrin e klasës së atributit, nqse kulmi nuk është gjethe atëherë është një kulm vendimi, i cili shërben si një atribut test për degë të reja të pemës.

Ross Quinlan ka punuar me këto lloj pemësh vendimi, të cilat duken të thjeshta dhe teknikisht të lehta për t'u përdorur. Teoria e tij përdor vetitë e kulmeve të pemës së vendimit, në mënyrë rekursive lart-poshtë, krahasime dhe gjykime të bazuara në vlerat e ndryshme të atributeve. Si input janë bashkësia e të dhënave ndërsa si rezultat është pema që i përngjason një diagrame të orientuar, ku secila gjethe përfaqëson një klasë vendimi. Gjethja përfaqëson vendimin që i përket një klase të dhënash, pasibëhet verifikimi i çdo testi nga rrënja në gjethe.

Në vijim bëhet përshkrimi i të dhënave, i modelit të përdorur për të ndërtuar pemën vendim-marrëse, si dhe pemën binare ekuivalente me të.

Të dhënat dhe rasti i familjeve shqiptare

Të dhënat janë marrë nga Anketa e Matjes së nivelit të jetesës për vitin 2012. Në studim janë marrë në konsideratë 6671 familje që përbëjnë dhe njësinë e vrojtimit të cilat janë përfaqësuese për të 4 rajonet: qëndror, bregdetar, malor, dhe Tiranën. Lidhur me përcaktimin e konceptit se "kush është i varfër" në Shqipëri: Një individ quhet i varfër nëse niveli i tij i shpenzimeve për frymë bie nën nivelin minimal që nevojitet për të plotësuar nevojat bazë për ushqime dhe artikuj joushqimor të këtij individi. Ky nivel minimal konsumi quhet "nivel i varfërisë" dhe është një kufi që përfaqëson pikën e ndarjes midis të varfërve dhe atyre që nuk janë të varfër. Kufiri absolut i varfërisë përkufizohet në lidhje me një prag të paracaktuar ose me përballimin e standardeve minimale të lidhura me një shportë konsumi (artikuj ushqimorë jo ushqimorë).

Më poshtë, në tabelën nr 1, pasqyrohen të dhënat që kemi zgjedhur rastësisht. Meqenëse të dhënat kanë atribute cilësore që karakterizohen nga dy ndryshore

secili, atëherë ne mund të konvertojmë këto cilësi në ndryshore binare me vlera 0,1.

Përllogaritja e kufirit absolut të varfërisë e mbështetetur mbi këtë shportë konsumi të një grup zgjedhjeje popullsie të marrë nga vëzhgimi dhe është konvertuar në masë monetare. Kjo ka çuar në zgjedhjen e një kufiri varfërie

Tabela 1. Atributet dhe statusi ekonomik i familjeve

Madhësia e familjes	Zona	Kryefamiljari	Statusi ekonomik
7	rural	mashkull	I varfër
6	rural	femer	I varfër
8	rural	mashkull	I varfër
12	urban	mashkull	I varfër
6	rural	mashkull	I varfër
6	rural	mashkull	I varfër
8	rural	mashkull	I varfër
11	urban	mashkull	I varfër
3	urban	femer	Jo i varfër
4	rural	mashkull	I varfër
6	rural	mashkull	I varfër
7	rural	mashkull	I varfër
5	rural	mashkull	I varfër
4	urban	mashkull	Jo i varfër
10	urban	mashkull	I varfër
5	rural	mashkull	I varfër
5	urban	mashkull	I varfër
6	rural	mashkull	I varfër
5	rural	mashkull	I varfër
2	rural	mashkull	Jo i varfër

5	rural	mashkull	I varfër
2	rural	mashkull	Jo i varfër
3	urban	mashkull	Jo i varfër
3	rural	mashkull	Jo i varfër
3	urban	mashkull	Jo i varfër
2	urban	mashkull	Jo i varfër
4	urban	mashkull	Jo i varfër
2	urban	mashkull	Jo i varfër
3	rural	mashkull	Jo i varfër
2	urban	mashkull	Jo i varfër

absolut prej 4891 lekë për frymë në muaj. Kufiri i varfërisë ekuivalent me çmimet e vitit 2012 është 5442 lekë për frymë në muaj (INSTAT 2012).

Nga zgjedhje të rastësishme në bashkësinë e 6671 familjeve kemi marrë 30 prej tyre për ndërtimin e një pjese të pemës së vendimit duke aplikuar algoritmin ID3 si dhe duke i kthyer të dhënat e tabelës në elemente binare, për të ndërtuar gjithashtu pemën e vendimit binare. Kemi zgjedhur të ndërtojmë ketë pemë binare për të gjitha këto vendime, pasi edhe mënyra për të gjetur rrugën optimale në pemë është edhe më e thjeshtë, sepse nga secili kulm dalin vetëm dy alternativa ose asnjë. Në rast se nuk del asnjë atëherë ky është një kulm vendimi final. Pema binare si ekuivalente e pemës vendim-marrëse ka numër më të vogël gjethesh të cilat janë ato kulme vendimi nga ku nuk dalin më alternativa.

Algoritmi ID3, entropia dhe pseudokode

Algoritmi ID3 është implementuar për t'u përdorur në ndryshore diskrete dhe të vazhduara. Algoritmi përdor entropinë e Shannon (Claude Shannon 1948) dhe një veti statistike që quhet *Përfitim Informacioni* (IG). Në termodinamikë entropia mat rregullsinë ose jorregullsinë e një sistemi, ndërsa në teorinë e informacionit entropia mat sa i sigurtë ose i pa sigurtëështë vlera e një variabli rasti.

Entropia tregon se vlerat më të vogla implikojnë më pak pasiguri, vlerat e mëdha implikojnë më shumë pasiguri. Nqse kemi një bashkësi S me n atribute, atëherë entropia përkufizohet si:

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

ku me p_i është raporti i vlerave që ka S në klasën i referuar atributit vendim.

Përfitimi i informacionit IG mat ndryshimet efektive të entropisë, pasi është bërë një vendim i bazuar në vlerat e një atributi. Në kontekstin e ndërtimit të një peme vendimi, ne jemi të interesuar që të dimë se sa shumë informacion duhet të trajtohet për atributin e vendimit, ai që do të jetë kryesori në radhë, pasi njohim vlerën e atributit A.

Formula e përfitimit të informacionit është :

$$Gain\,(S,A) = Entropy\,(S) - \sum_{j=1}^{n} [p_j \cdot Entropy\,(p_j)]$$

ku p_j është raporti i të gjithë vlerave që ka atributi A ndaj bashkësisë S.

Gain (S,A) llogaritet për të renditur atributet dhe ndërtimin e pemës vendim-marrëse, ku në secilin kulm lokalizohet atributi që ka vlerën më të madhe të përfitimit të informacionit (*Information Gain),* krahasuar me atributet e tjera që nuk merren në konsideratë përgjatë rrugës që nga rrënja. Kështu çdo atribut klasifikohet hap pas hapi në çdo nënndarje të re. Kjo proçedurë i ngjan shumë thirrjes së famshme“ përça dhe sundo” (Zhang, 2011).

Në bazë të këtij informacioni ne do të zgjedhim atributin që ka vlerën më të lartë, duke e quajtur atë *rrënj*ë të pemës sonë. Pasi të kemi gjetur rrënjën, ky atribut lëviz nga bashkësia dhe për nivelin tjetër të dhënat do të ndahen sipas vlerave të këtij atributi. Më poshtë jepet entropia e statusit ekonomik si entropi e nje sistemi të tërë:

$$Entropy\,(S) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i = -\frac{18}{30}\log_2\frac{18}{30} - \frac{12}{30}\log_2\frac{12}{30} = 0.968$$

Tani llogaritim entropinë për atributin→ madhësi e familjes:

$$Entropi\,(S_{n\geq 4.5}) = -\frac{17}{17}\log_2\frac{17}{17} = 0$$

$$Entropi\,(S_{n<4.5}) = -\frac{12}{13}\log_2\frac{12}{13} - \frac{1}{13}\log_2\frac{1}{13} = 0.39$$

$$Gain\,(S, Madh.fam.) = Entropi\,(S) - \frac{17}{30}Entropi\,(S_{n\geq 4.5}) - \frac{13}{30}Entropi\,(S_{n<4.5})$$

$$= 0.968 - 0.566 \cdot 0 - 0.433 \cdot 0.39 = 0.799$$

Tani llogaritim entropinë për atributin→Zona:

$$Entropi\,(S_{rurale}) = -\frac{14}{18}\log_2\frac{14}{18} - \frac{4}{18}\log_2\frac{4}{18} = 0.75$$

$$Entropi\,(S_{urbane}) = -\frac{4}{12}\log_2\frac{4}{12} - \frac{8}{12}\log_2\frac{8}{12} = 0.91$$

$$Gain\ (S, Zona) = Entropi\,(S) - \frac{18}{30}Entropi\ (S_{rurale}) - \frac{12}{30}Entropi\,(S_{urbane})$$

$$= 0.968 - 0.6 \cdot 0.75 - 0.4 \cdot 0.91 = 0.154$$

Tani llogaritim entropinë për atributin→ Kryefamiljari:

$$Entropi\,(S_{mashkull}) = -\frac{17}{28}\log_2\frac{17}{28} - \frac{11}{28}\log_2\frac{11}{28} = 0.96$$

$$Entropi\,(S_{femer}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Gain\ (S, Kryef.) = Entropi\,(S) - \frac{28}{30}Entropi\ (S_{mashkull}) - \frac{2}{30}Entropi\,(S_{femer})$$

$$= 0.968 - 0.933 \cdot 0.96 - 0.066 \cdot 1 = 0.138$$

Siç shihet atributi me vlerë më të madhe është Madhësia e familjes, e cila do të jetë edhe rrënja e pemës sonë, atributi tjetër vjen Kryefamiljari e më pas atributi Zona. Kemi përcaktuar kështu vetëm rrënjën. Në vazhdim për çdo nënndarje në pemë aplikohet algoritmi nga e para, e kështu me radhë.

ID3 Algoritëm, (Quinlan 1986)

- S është një bashkësi tiparesh, gjithashtu quhet edhe hapësirë e tipareve.
- C është bashkësia e klasave.
- $c: S \to C$ është funksioni klasifikim ideal për S.
- $D = \{(s_1, c(s)), \ldots\ldots\ldots (s_n, c(s_n))\} \subseteq SxC$ është bashkësia e shembujve;

ID3 (D, Atributi, Etiketimi)

- Krijo një kulm t në pemë

Në qoftë se gjithë shembujt e D janëpozitivë, atëherë shëno të vetmin kulm me "+".

Në qoftë se gjithë shembujt e D janë negative, atëherë shëno të vetmin kulm me "-".

Etiketo "t" me vlerën më të madhe nga vlerat në D

- Nqse atributi është bosh, atëherë rikthehu në kulmin t.

Le të jetë A^* atributi më i mirë i klasifikuar nga bashkësia e shembujve

Shënoj t atributin vendim të A^*

Për çdo vlerë të mundëshme "a" në A^* bëj:

Shto një degë të re peme poshtë t, korresponduese të A^* ="a"

Përndryshe,

Shkruaj Pema ID3 (D_a, Atribute/ A^* , Etiketë)

- Rikthehu në t

Aplikimi i paketës Weka

Weka është një paketë e re për klasifikim, e cila u zhvillua për herë të parë në Universitetin Waikato në Zelandën e Re (1999). WEKA përfaqëson Waikato Environment for Knowledge Analysis, ku paraqitet me anë të një zogu të llojit Meka, i cili gjendet vetëm në Zelandën e Re. Weka është shkruar në një gjuhë programimi objekti të orientuar në Java, e përshtatshme për të gjithë llojet e kompjuterave. Pikërisht algoritmi ynë ID3 implementohet në Weka me anë të përzgjedhjes tek Trees-J48 Pruned (Hall, 2009). Në figurën 1 ne kemi paraqitur klasifikimin e të gjithë atributeve cilësore të krahasuar ndërmjet tyre.

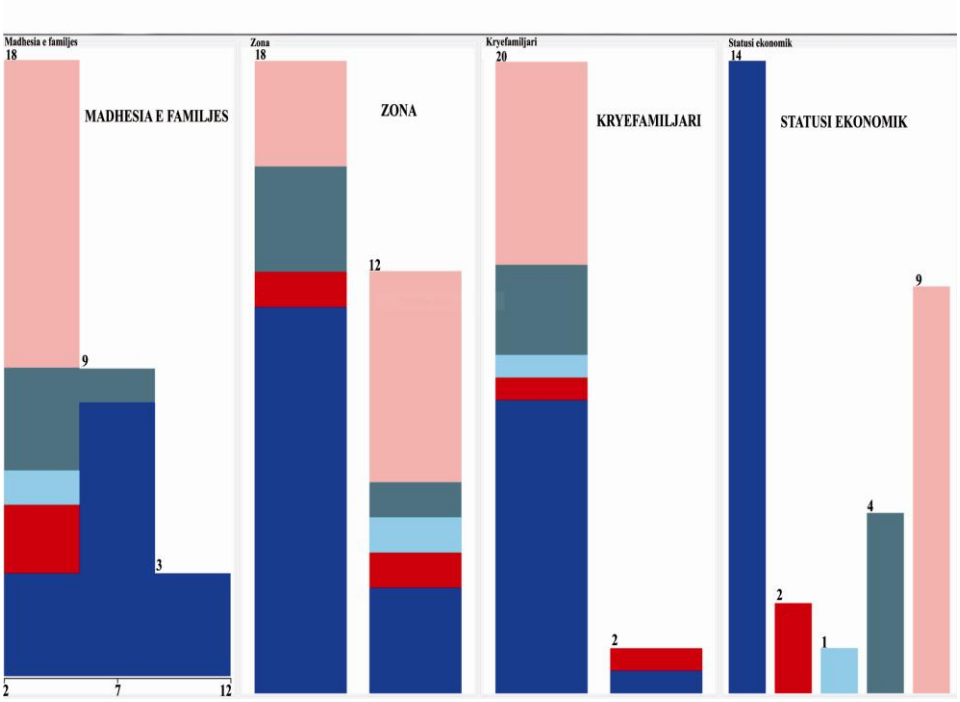

Figura 1. Klasifikimi i atributeve me anë të Weka

Për të ndërtuar pemën në Weka duhen ndjekur këto hapa:

Tabela jonë u ndërtua në Excel, në modelin CSV Comma Delimited. Kemi aplikuar opsionet: Classify, Choose J 48 Prunned tree, Use training data set, Run, Visualize tree (Nelson, 2010).

Pema jonë vendim-marrëse ka klasifikuar më së miri gjithë të dhënat tona. Rrënja e saj përbëhet nga atributi me entropinë më të madhe që është "madhësia e familjes", më pas në nivel të dytë, pasi rillogaritet entropia e nivelit pasardhës, ajo mbetet sërish me vlerë më të madhe se atributet e tjera.

Algoritmi ka klasifikuar duke marrë dy degë me opsione më të mëdha se 4.5 dhe më të vogël se 4.5 në lidhje me rrënjën, më pas sipas atributit fitues të radhës vendosen edhe variablat cilësorë të secilit.

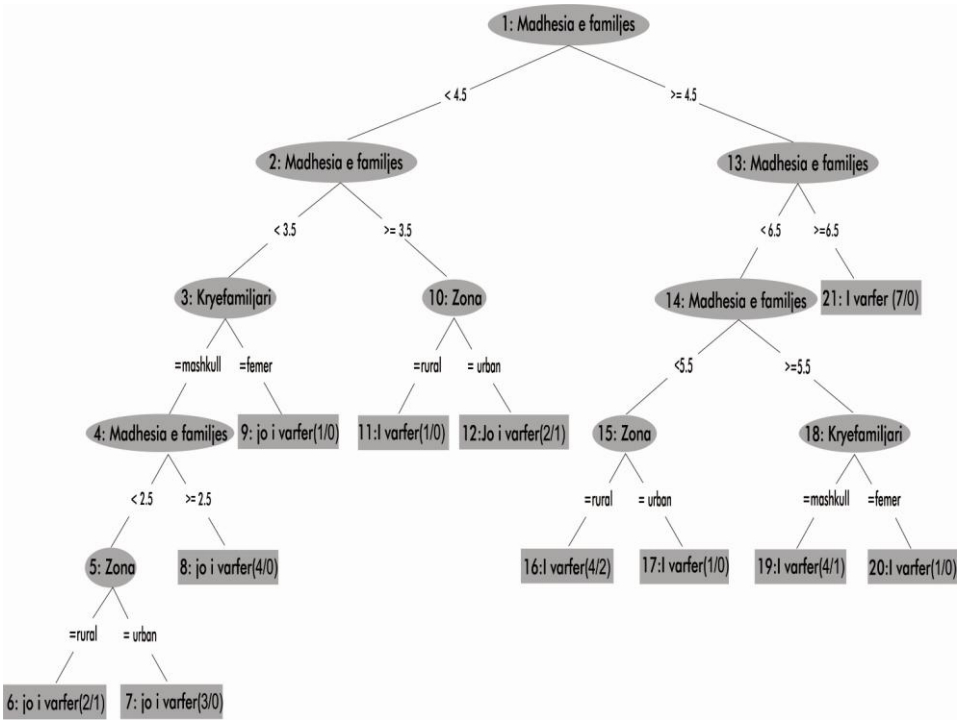

Figura2. Paraqitja e pemës vendim-marrëse me anë të Weka

Ne shohim se ana e majtë e pemës binare ka klasifikuar "jo të varfërit", përveç një gjethe që përmban të varfër, ndërsa ana e djathtë ka klasifikuar "të varfërit", gjithçka në një hierarki të tabelës fillestare (Qendraj & Xhafaj 2015).

Për të ndërtuar pemën ekuivalente binare bëjmë këto përcaktime:

Madhësiae familjes	Nr $\geq 4.5 \rightarrow 0$	Nr $< 4.5 \rightarrow 1$
Zona	Rurale $\rightarrow 1$	Urbane $\rightarrow 0$
Kryefamijlari	Mashkull $\rightarrow 0$	Femër $\rightarrow 1$
Statusi ekonomik	I varfër $\rightarrow 1$	Jo i varfër$\rightarrow 0$

Tabela 2. Kalimi në variabla binare

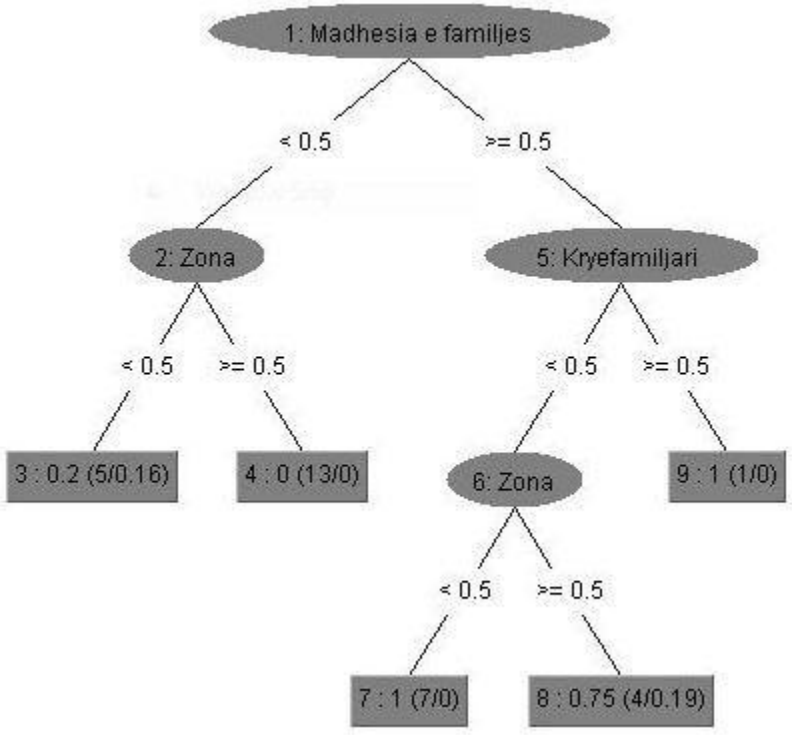

Figura3. Paraqitja e klasifikimit binarë

Edhe në pemën binare atributi "madhësi e familjes"është sërish rrënja e pemës. Këtu atributet kanë vlera 0;1, kështu që algoritmi do klasifikojë me atribute me variabla numerike jo më cilësore si më lart. Madhësia e pemës është më e vogël, por sërish në anën e majtë janë klasifikuar "jo të varfër", p.sh madhësia e

familjes$\rightarrow \leq 0.5 \rightarrow$ zona $\rightarrow \geq 0.5 \rightarrow 0(jo\ të\ varfër)$,13 të tillë.

Konkluzione

Siç shihet nga paraqitja e pemëve, ne kemi një informacion më të plotë klasifikimi kur kemi pemën vendim-marrëse të aplikuar me atribute cilësore dhe sasiore, ndërsa tek pema binare kemi më pak informacion klasifikimi. Pema binare është e përshtatëshme kur kemi një sasi shumë të madhe të dhënash dhe kur kemi atribute cilësore që marrin vetëm dy vlera. Nëse tek të dhënat, atributet janë cilësore me 3 ose më shumë tipare atëherë pema binare nuk është e përshtatshme.

Literatura

Quinlan J. R. (1986): Induction of decision trees, Machine Learning ,vol 1; 81-100

Zhang Q. (2011): Application of ID3 Algorithm in Exercise Prescription, Proccedings of the International Conference on Electric and Electronics; 669-675

Hall M. (2009): The Weka Data Mining Software,Sigkdd Explorations,vol 11;no 1,23-56

Soni S., Pillai J. (2008): An expert case-based system using decision tree Induction for Weight Management Conseling to Obese Children,International Journal of Computer Science and Applications, vol 1; no 2, 301-310

Nelson T. (2010): Java Universal Network/ Graph Framework [Online]

Qendraj D., Xhafaj E. (2015): Evaluating risk factors of being obese by using ID3 algorithm in Weka software,European Scientific Journal,vol 11, no 24. 261-267