

Adatbányászat projectmunka dokumentáció

Lakókocsi biztosítási adatok

A FELADAT

Referencia cikk:

Lakókocsi biztosítási adatok

Problémakör:

Ezt az adathalmazt a Coil 2000 adatbányászati versenyen használták föl a versenyzők. Egy biztosító társaság ügyfeleinek adatait tartalmazza. Az érdeklődés tárgya, hogy vajon az érdeklődő köt-e lakókocsi biztosítást. Lehetséges ügyfelenként 86 adatot tartunk nyilván: 43 szociodemográfiai összetevő származik az ügyfél irányító számából, 43 pedig az ügyfél egyéb biztosítási kötvény tulajdonjogából képződik.

A motivációt az ügyfelekkel való levelezés adta. A leveleket csak azoknak az ügyfeleknek akarták elküldeni, akik nagy valószínűséggel meg fogják vásárolni a biztosítási kötvényt. Ennek következtében a mélyebb probléma az ügyfeleknek azon részhalmazának kiválasztása, akik biztosítási kötvény vásárlásának valószínűsége egy valószínűségi küszöb felett van. (Akikről tudjuk, hogy már kötöttek biztosítást, azok természetesen ki lesznek véve ebből a részhalmazból.) A küszöbszám függ a levelezés költségétől és a biztosítás megkötéséből származó bevételből. Tehát a feladat nem annak megjóslása, hogy ki rendelkezik biztosítási kötvénnyel, hanem a tesztadatbázisból annak a 800 személynek a kiválasztása, akinek a legnagyobb a valószínűsége a biztosítás megkötésére. Minden egyes 800 -as halmazból a biztosítási kötvényesek tényleges száma adja a megoldás jóságát.

A feladat:

Adatbányászati feladat annak megjóslása, hogy

- **mely ügyfelek a potenciális érdeklődők a lakókocsi biztosításra**
- **adott személy fog-e kötni lakókocsi biztosítást**

Először is némi próbálgatás útján fel kell deríteni és ki kell analizálni az adatokat.

Az adatok ábrázolás egyfajta betekintést ad az adathalmaz bizonyos sajátosságaiba. Azután fel kell készíteni az adatokat az adatbányászatra. Fontos kiválasztani a megfelelő jellemzőket és új tulajdonságokat konstruálni a meglévőkből, végül a nyertes versenyzőjéhez közeli megoldást találni.

A feladat célja egy üzleti alapú leírás megadása, mely tisztázza hogy miért kötnek lakókocsi biztosítást az ügyfelek, és miben különböznek azok akik biztosítást kötnek azoktól akik nem kötnek. A megoldás felhasználhatja a regressziós egyenletet, a döntési fát, a neutrális hálózatokat, nyelvi leíró formákat, grafikus reprezentációt, vagy egyéb módszereket.

Kihívás, nehézségek:

A jellemző kiválasztás és kinyerés nagyon fontos a feladat szempontjából. Ezt csak aztután lehet helyesen megtenni, hogy tekintélyes mennyiségű idő lett fordítva az adathalmaz megismerésére.

Az adatok ingadozóak: a tanuló adathalmazból az ügyfeleknek csak az 5 – 6% -a köti meg a biztosítást. Nincs hiányos vagy zajos adat.

Az adatbázis:

Az 1.7 MB méretű adatbázis 9823 rekordból áll:

- 5823 ügyfél rekord a tanuló adatbázisban, tartalmazza azt az információt is, hogy az ügyfél kötött-e biztosítást
- 4000 rekord a teszt adatbázisban. Ezekről az ügyfelekről, csak a verseny szervezői tudják, hogy kötöttek-e biztosítást.

86 attribútum:

- 43 szociodemográfiai összetevő származik az ügyfél irányító számából
- 43 az ügyfél egyéb biztosítási kötvény tulajdonjogából képződik

Részletes leírás az adatokról:

Az adatokat a Sentient Machine Research holland adatbányászattal foglalkozó cég biztosította, valós adatokat tartalmaz.

Lényeges file -ok:

- ticdata.txt: tanuló és validáló adatbázis, 5823 ügyfél adatait tartalmazza. Mindegyik rekord 86 jellemzőből áll. 1-43 között szociodemográfiai adatot, 44-86 között a termék tulajdonlásából származó adatot tartalmaz. A 86. jellemző a CARAVAN: mobil otthon biztosítások száma, a célérték.
- ticeval.txt: teszt adatbázis, 4000 ügyféladatait tartalmazza. Formátuma és felépítése megegyezik a ticdata.txt file -ével, csak a célérték hiányzik belőle. A versenyzőknek javasolt, hogy csak a célértéket adják meg a megoldásban.
- dictionary.txt: a jellemzők és értékeik jelentése az adatszótárban található.
- targets.txt: a jelenleg biztosítással rendelkező ügyfelek.

Minden adathalmaz tabulátorral elválasztott értékeket tartalmaz.

A MEGOLDÁS

Eszközök, szoftverek:

Weka, Wekawrapper, OpenOffice Calc

Jellemzőkinyerés, reprezentáció:

A cél az, hogy a dimenziószám ~20 körüli értékre csökkenjen.

1. Előfeldolgozás: dimenzió redukció

Az adatszótár alapján egyes jellemzők elhagyhatóak.

○ Adattisztítás

Az egymást kizáró logikai értékek közül csak az egyiket tartom meg.

~~14 MFGEKIND Household without children~~

15 MFWEKIND Household with children

A jövedelem eloszlását tükrözi az átlagos jövedelem is:

~~37 MINKM30 Income < 30.000~~

~~38 MINK3045 Income 30-45.000~~

~~39 MINK4575 Income 45-75.000~~

~~40 MINK7512 Income 75-122.000~~

~~41 MINK123M Income >123.000~~

42 MINKGEM Average income

○ Jellemzők összevonása

Vallások összevonása vallásos kategóriába

~~6 MGODRK Roman catholic see L3~~

~~7 MGODPR Protestant ...~~

~~8 MGODOV Other religion~~

~~8 MGODOV Religion~~

~~9 MGODGE No religion~~

6 MGODRE Religion

Kapcsolatok összevonása kapcsolatban és egyedülálló kategóriák felállítása

~~10 MRELGE Married~~

~~11 MRELSA Living together~~

~~12 MRELOV Other relation~~

~~13 MFALLEEN Singles~~

10 MRELRL Relation

Képzettség összevonása egy, számértékkel jellemzett kategóriába

~~16 MOPLHOOG High level education~~

~~17 MOPLMIDD Medium level education~~

~~18 MOPLLAAG Lower level education~~

16 MOPLLEVE Level of education

Szociális helyzet összevonása egy, számértékkel jellemzett kategóriába

~~25 MSKA Social class A~~

~~26 MSKB1 Social class B1~~

~~27 MSKB2 Social class B2~~

~~28 MSKC Social class C-~~
~~29 MSKD Social class D-~~
25 MSKCL Social class

2. Jellemzőszelekció

A meglévő biztosítás jobban tükrözi a biztosítás kötésére való hajlandóságot, mint a megkötött biztosítások száma (ahol nem általános és nem autó biztosításról van szó).

44 PWAPART Contribution private third party insurance see L4

~~45 PWABEDR Contribution third party insurance (firms) ...~~

~~46 PWALAND Contribution third party insurance (agriculture)~~

47 PPERSAUT Contribution car policies

48 PBESAUT Contribution delivery van policies

49 PMOTSCO Contribution motorcycle/scooter policies

50 PVRAAUT Contribution lorry policies

51 PAANHANG Contribution trailer policies

52 PTRACTOR Contribution tractor policies

53 PWERKT Contribution agricultural machines policies

54 PBROM Contribution moped policies

55 PLEVEN Contribution life insurances

56 PPERSONG Contribution private accident insurance policies

57 PGEZONG Contribution family accidents insurance policies

58 PWAOREG Contribution disability insurance policies

59 PBRAND Contribution fire policies

60 PZEILPL Contribution surfboard policies

61 PPLEZIER Contribution boat policies

62 PFIETS Contribution bicycle policies

63 PINBOED Contribution property insurance policies

64 PBYSTAND Contribution social security insurance policies

~~66 AWABEDR Number of third party insurance (firms) ...~~

~~67 AWALAND Number of third party insurance (agriculture)~~

~~69 ABESAUT Number of delivery van policies~~

~~70 AMOTSCO Number of motorcycle/scooter policies~~

~~71 AVRAAUT Number of lorry policies~~

~~72 AAANHANG Number of trailer policies~~

~~73 ATRACTOR Number of tractor policies~~

~~74 AWERKT Number of agricultural machines policies~~

~~75 ABROM Number of moped policies~~

~~76 ALEVEN Number of life insurances~~

~~77 APERSONG Number of private accident insurance policies~~

~~78 AGEZONG Number of family accidents insurance policies~~

~~79 AWAOREG Number of disability insurance policies~~

~~80 ABRAND Number of fire policies~~

~~81 AZEILPL Number of surfboard policies~~

~~82 APLEZIER Number of boat policies~~

~~83 AFIETS Number of bicycle policies~~

~~84 AINBOED Number of property insurance policies~~

~~85 ABYSTAND Number of social security insurance policies~~

3. Adatok transzformálása

A vallásos oszlopot a következő összevonással és transzformációval helyettesítettem:

~~6 MGODRK Roman catholic see L3~~

~~7 MGODPR Protestant ...~~

~~8 MGODOV Other religion~~

~~9 MGODGE No religion~~

6 MGODRE Religion: if (#6+#7+#8>#9) TRUE else FALSE

A kapcsolatban oszlopot a következő összevonással és transzformációval helyettesítettem:

~~10 MRELGE Married~~

~~11 MRELSA Living together~~

~~12 MRELOV Other relation~~

~~13 MFALLEEN Singles~~

10 MRELRL Relation: if (#10>#13 OR #11>#13 OR #12>#13) TRUE else FALSE

A háztulajdon gyerekekkel oszlopot a következő transzformációval helyettesítettem:

~~14 MFGEKIND Household without children~~

15 MFWEKIND Household with children : if (#15>#14) TRUE else FALSE

A képzettségre 1-3 értékkel a legmagasabb végzettséget jellemzem:

~~16 MOPLHOOG High level education~~

~~17 MOPLMIDD Medium level education~~

~~18 MOPLLAAG Lower level education~~

16 MOPLLEVE Level of education:

switch max(#16,#17,#18) case#16: 1; case #17: 2; case #18: 3;

~~25 MSKA Social class A~~

~~26 MSKB1 Social class B1~~

~~27 MSKB2 Social class B2~~

~~28 MSKC Social class C~~

~~29 MSKD Social class D~~

25 MSKCL Social class

switch max(#25,#26,#27,#28,#29) case#25: A; case #26: B1; case #27: B2; case #28: C; case #29: D;

A következő mezők értékei 0..9 egész számok. A mezők logikai értéket tartalmaznak, ezért logikai értékre konvertáltam, ahol 0: False, 1..9: TRUE

30 MHHUUR Rented house

31 MHKOOP Home owners

32 MAUT1 1 car

33 MAUT2 2 cars

34 MAUT0 No car

35 MZFONDS National Health Service

36 MZPART Private health insurance

A célfüggvény 0-1 értékeket vehet föl. A lakókocsi biztosítást kötött attribútumot jobban kifejezi az igen/nem érték. Ezért a 0: yes, 1: no értékekre cseréltem.

Diszkretizálás: nincsenek valós értékű adatok.

Adattisztítás: nincsenek zajos vagy hiányzó adatok.

Számosságcsökkentés: lehetőség lenne heurisztika alapján az adatok csökkentésére.

Pl.: mindenki rendelkezik házzal, és a 3 -nál több házzal rendelkezők nem kötöttek lakókocsi biztosítást, ezért a 3 -nál több házzal rendelkezők adatai törölhetőek lennének, de a viszonylag kevés rekordszám miatt nem szeretnék törölni.

Normalizálás: az adatok többsége 0..9 egész értékeket vehet föl, ezért nem szükséges.

A dimenzió számot sikerült 86 -ról 46 -ra redukálni.

A Weka ezt az adatbázist valós időben már fel tudja dolgozni (< 500 sec):

java -Xmx300m -jar /opt/weka/weka.jar

A memória növelésével közepesen erős laptopon (IBM T42) futott.

Adatbázis konvertálás

Az adatbázis TXT file -ban, szöveges formátumban van megadva. A jellemző kinyerést és az ARFF formátumba (attribute-relation file format) konvertálást az OpenOffice programmal végeztem.

1. OpenOffice Calc táblázatkezelővel Text Import, az oszlopok Tab -bal szeparált értékek
2. A fentiekben leírt törölt attribútumokra oszlopok törlése
3. A tanuló adatbázisban nem, csak a teszt adatbázisban a fentiekben leírtak alapján a feladat szempontjából irreleváns rekordok törlése szűrők segítségével
4. Mentés másként, file típus: CSV TEXT kiterjesztés: .arff
5. OpenOffice Writer szövegszerkesztővel a táblázat elejét sorok beszúrásával kiegészítettem a Weka által definiált fejléccel és megjegyzésekkel

@relation caravan

```
@attribute MOSTYPE numeric
@attribute MAANTHUI numeric
@attribute MGEMOMV numeric
@attribute MGEMLEEF numeric
@attribute MOSHOOFD numeric
@attribute MGODRE {TRUE,FALSE}
@attribute MRELRL {TRUE,FALSE}
@attribute MFW EKIND {TRUE,FALSE}
@attribute MOPLLEVE {1,2,3}
@attribute MBERHOOG numeric
@attribute MBERZELF numeric
@attribute MBERBOER numeric
@attribute MBERMIDD numeric
@attribute MBERARBG numeric
@attribute MBERARBO numeric
@attribute MSKCL {A,B1,B2,C,D}
@attribute MHHUUR {TRUE,FALSE}
@attribute MHKOOP {TRUE,FALSE}
@attribute MAUT1 {TRUE,FALSE}
@attribute MAUT2 {TRUE,FALSE}
```

@attribute MAUTO {TRUE,FALSE}
@attribute MZFONDS {TRUE,FALSE}
@attribute MZPART {TRUE,FALSE}
@attribute MINKGEM numeric
@attribute MKOOPKLA numeric
@attribute PWAPART numeric
@attribute PPERSAUT numeric
@attribute PBESAUT numeric
@attribute PMOTSCO numeric
@attribute PVRAAUT numeric
@attribute PAANHANG numeric
@attribute PTRACTOR numeric
@attribute PWERKT numeric
@attribute PBROM numeric
@attribute PLEVEN numeric
@attribute PPERSONG numeric
@attribute PGEZONG numeric
@attribute PWAOREG numeric
@attribute PBRAND numeric
@attribute PZEILPL numeric
@attribute PPLEZIER numeric
@attribute PFIETS numeric
@attribute PINBOED numeric
@attribute PBYSTAND numeric
@attribute AWAPART numeric
@attribute APERSAUT numeric
@attribute CARAVAN {yes,no}

@data

BINÁRIS OSZTÁLYOZÁS

A feladat egy bináris osztályozási problémára vezethető vissza.

Input, tanulóhalaz (train): adott egy tanító halmaz, ami kétféle jellemzővektorokból áll: pozitív és negatív (jellemzi az objektumot, ill., nem jellemzi azt).

Output egy teszt halmazra: pontok (scores), amik megadják, hogy az adott tesztvektor mennyire jellemzi az objektumot, ill., mennyire tartozik a pozitív objektumok közé.

Rokon terület a regresszióval. A regressziónál az általános feladat:

Input vektor: $X = (X_1, X_2, \dots, X_p)$

Output: Y (valós)

Y -t jósoljuk egy $f(X)$ függvény segítségével úgy, hogy a várható hiba $E(L(Y, f(X)))$ minimális legyen.

Négyzetes hiba: $L(Y, f(X)) = (Y - f(X))^2$

Y : 1, 0 bináris értékű (1: az osztályhoz tartozik, 0: nem tartozik az osztályhoz) – pontosabban yes/no.

A hibafüggvény, amit minimalizálni szeretnénk:

- Neuronhálónál marad a Least Squares Error
- MCE: Minimum Classification Error: itt a tévesztések számát minimalizáljuk.

Bináris (diszkriminatív) osztályozás: a tanító halmaz tartalmaz egyaránt pozitív és negatív példákat is, amelyeket tanítás során felhasználunk.

Pl. neuronháló, döntési fa, support vector machine, vector quantisation, ...

Sűrűség alapú modellek:

Gaussian Mixture Model: A pozitív adatokat Gauss elszlások súlyozott összegével modellezzük, és ezáltal kapunk egy valószínűségi modellt. Ezt használhatjuk fel rangsorolásra a tesztnél.

Normális eloszlás: 1-GMM. Könnyen számolható.

Parzen becslő: minden tanítópontba egy-egy egyforma (de megadható) szórású többváltozós Gauss sűrűségfüggvényt illeszt. Az inputot ezen sűrűségfüggvények összegével modellezi. (A szóráson v. szélességen sok múlhat. Egy normalizálás után már

Behatároló módszerek:

K-center, vagy K-medoid módszerek: klaszerezésnél.

NNDD (Nearest Neighbour Data Description): legközelebbi szomszéd alapú adatjellemzés

SVDD (Support Vector Data Description):

A tanítóadatokat egy magasabb dimenziós ún. kernel térben „karikázza be”.

„Reconstruction” metódusok:

Az input adatokból előállítanak olyan - jóval kevesebb számú - jellemzővektort, melyek segítségével megadják az adatok legvalószínűbb elhelyezkedését.

Pl.

- K-means: az adatokat a középvektorokkal jellemezzük.
- SOM: szintén. Csak a referencia vektorok a „súlyvektorok”
- LVQ: Learn Vector Quantization: ez már felügyelt tanulás (bináris, v. több osztályos)

Elválasztó hiperfelület meghatározása:

Megadunk egy megbízhatósági küszöbértéket (pl. 95%, 99%), és keressük azt a score-küszöböt, hogy a pozitív tanulóminták 95%-a (99%-a, stb.) az osztályhoz tartozó, a maradék outlier legyen.

Ilyenek pl. Gauss modellek, Parzen, k-means,...

Kiértékelés:

Rangsorolásnál: ROC és AUC

Osztályozásnál: Accuracy, Precision, F-measure, ...

WEKA

A tanuló adatbázison elért eredmény alapján az osztályozási algoritmusok rangsorolva:

Alkalmazott megoldási mód	Korrekt osztályozás
Random Committee, Random Tree	99.3130%
IBk instance-based classifier using 1 nearest neighbour for classification	99.3130%
KStar Beta	99.3130%
RandomTree	99.3130%
Random forest of 10 trees, each constructed while considering 6 random features.	99.2614%
IB1 classifier	99.2614%
Multilayer Perceptron: neuronháló	96.9083%
PART	96.5132 %
NBTree	94.4349%
Bagging	94.4177%
Decorate	94.2460%
Ripple DOWn Rule Learner(Ridor) rules	94.1773 %
Locally weighted learning	94.1086%
Classification Via Regression	94.1086%
OrdinalClass Classifier	94.1086%
Logistic Regression with ridge parameter	94.1086%
Radial basis function network	94.1086%
AdaBoostM1: Base classifiers and their weights	94.1086%
BinarySMO	94.1086%
Attribute Selected Classifier	94.1086%
MultiBoostAB: Base classifiers and their weights	94.1086%
MultiClass Classifier	94.1086%
Filtered Classifier	94.1086%
Simple Logistic	94.1086%
Cross-validated Parameter selection	94.1086%
Granding	94.1086%

Stacking	94.1086%
Decision Stump	94.1086%
Vote combines the probability distributions of these base learners	94.1086%
StackingC	94.1086%
MultiScheme selection using error on training data	94.1086%
J48 pruned tree	94.1086%
Logistic model tree	94.1086%
REPTree	94.1086%
Single conjunctive rule learner	94.1086%
ZeroR	94.1086%
Decision Table	94.1086%
OneR classifier	94.1086 %
JRIP rules	94.1086 %
RacedIncrementalLogitBoost: Best committee on validation data	94.0914%
Alternating decision tree	94.0570%
VotedPerceptron	94.0398%
LogitBoost	94.0227%
Bayes Network Classifier	88.5435%
Naive Bayes Classifier	85.2284%
Threshold Selector	84.5929%

A 95% fölötti algoritmusokat választom ki további optimalizálásra.

A tanuló adatbázison elért eredményt egy betanult algoritmus adta, ami nem biztos, hogy a teszt adatbázison is jól fog szerepelni. Ezért a tanuló adatbázist kétfelé szedem. Az 5823 rekordból 4000 -et meghagyok a tanuló adatbázisban, 1832 -at pedig optimalizálás céljából a **validációs adatbázisba** különítek el, így a tanuló és a validációs adatbázison elért eredményű algoritmust használom majd a teszt adatbázisra.

Alkalmazott megoldási mód	Tanuló db	Validációs db
Random forest of 10 trees, each constructed while considering 6 random features.	99.5000%	91.1636 %
Random Committee, Random Tree	99.5500%	90.9440%
Multilayer Perceptron: neuronháló	97.300%	90.8342 %
PART	97.0000%	90.7245%
KStar Beta	99.5250%	89.4621%
RandomTree	99.5500%	89.0779%

IB1 classifier	99.5500%	88.6937%
IBk instance-based classifier using 1 nearest neighbour for classification	99.5500%	88.5840%

A validációs adatokon elért értékek alapján a **Random forest** algoritmust választom a teszt adatok feldolgozásához. A Random forest a tesztadatokat **92.8500%** -ban helyesen osztályozta.

Futásának eredménye és kimenete:

=== Run information ===

```

Scheme:      weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1
Relation:    caravan
Instances:   4000
Attributes:  47
...
Test mode:   user supplied test set: 4000 instances

```

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 6 random features.
Out of bag error: 0.139

Time taken to build model: 7.48 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	3714	92.85	%
Incorrectly Classified Instances	286	7.15	%
Kappa statistic	0.058		
Mean absolute error	0.1035		
Root mean squared error	0.2502		
Relative absolute error	95.9015	%	
Root relative squared error	104.702	%	
Total Number of Instances	4000		

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.053	0.015	0.188	0.053	0.083	yes
0.985	0.947	0.941	0.985	0.963	no

=== Confusion Matrix ===

```

  a    b  <-- classified as
13 230 |    a = yes
56 3701 |    b = no

```

A validációs adatokon elért értékek alapján a **PART** algoritmust választom szabályosságok, törvényszerűségek feltárásához. A PART a tesztadatokat **90.7245%** -ban helyesen osztályozta.

A PART algoritmus lakókocsi biztosítást köt szabályai logikai formulával leírva:

MRELRL = TRUE AND MBERBOER <= 0 AND MZFONDS = TRUE AND MSKCL = D AND MHKOOP = TRUE AND PBROM <= 0 AND MKOOPKLA > 2 AND PPERSAUT > 0: yes (5.0/1.0)

MRELRL = TRUE AND MHKOOP = FALSE AND MGEMOMV > 2 AND PLEVEN <= 1: yes (10.0)

MRELRL = TRUE AND MAUT2 = TRUE AND MZFONDS = FALSE AND PPERSAUT > 0: yes (4.0)

MRELRL = TRUE AND MAUT2 = FALSE AND PBYSTAND > 0: yes (5.0/1.0)

MRELRL = TRUE AND MAUT2 = TRUE AND MZPART = FALSE AND MBERMIDD <= 4 AND MOSHOOFD <= 8 AND PWAPART <= 1 AND PPERSAUT > 0 AND APERSAUT <= 1 AND MBERMIDD <= 3: yes (3.0/1.0)

MRELRL = TRUE AND MHHUUR = FALSE AND MZPART = TRUE AND MINKGEM <= 6 AND MGEMOMV > 3: yes (6.0)

MRELRL = TRUE AND PBROM > 0 AND APERSAUT <= 1 AND PMOTSCO <= 0 AND MOPLLEVE = 3 AND MOSHOOFD <= 6: yes (5.0)

MRELRL = TRUE AND MSKCL = B1: yes (7.0)

MRELRL = TRUE AND MAUT2 = TRUE AND MZPART = TRUE AND MSKCL = C AND MGODRE = FALSE AND MAUT0 = TRUE: yes (3.0)

MRELRL = TRUE AND MSKCL = C AND PPERSAUT <= 0 AND MAUT0 = FALSE AND MAUT2 = TRUE AND MBERARBG > 1: yes (6.0/1.0)

MRELRL = TRUE AND MZPART = FALSE AND MHHUUR = FALSE: yes (4.0)

MRELRL = TRUE AND MSKCL = B2 AND MBERMIDD <= 2 AND MZPART = TRUE: yes (17.0/4.0)

MRELRL = TRUE AND MSKCL = B2 AND PMOTSCO > 0: yes (4.0/1.0)

MRELRL = TRUE AND MSKCL = D: yes (4.0/1.0)

MRELRL = TRUE AND MHHUUR = FALSE AND PPLEZIER <= 1 AND MSKCL = A AND PPERSAUT > 0: yes (5.0/1.0)

MRELRL = TRUE AND MAUT2 = FALSE AND PBRAND > 1 AND MZPART = TRUE AND MGEMOMV <= 3: yes (13.0)

MRELRL = TRUE AND MHHUUR = FALSE AND PBROM <= 0 AND MOSHOOFD > 2: yes (10.0/2.0)

MRELRL = TRUE AND MHHUUR = TRUE AND MZPART = TRUE AND MAUT2 = TRUE AND MGODRE = FALSE AND APERSAUT <= 1 AND MKOOPKLA <= 6: yes (6.0/2.0)

MHHUUR = TRUE AND MRELRL = TRUE AND MZPART = FALSE: yes (6.0/1.0)

MRELRL = TRUE AND MHHUUR = TRUE AND MAUT2 = TRUE AND AWAPART <= 0 AND APERSAUT > 1 AND MOSHOOFD > 1: yes (5.0)

MRELRL = TRUE AND MHHUUR = TRUE AND MAUT2 = TRUE AND AWAPART <= 0 AND MGODRE = TRUE AND MBERARBG <= 2 AND MOSHOOFD <= 5 AND MOSHOOFD <= 2 AND APERSAUT <= 1 AND MBERZELF <= 2 AND PFIETS <= 0 AND MFWEKIND = FALSE AND PBRAND <= 1 AND MINKGEM > 5: yes (5.0/2.0)

MHHUUR = TRUE AND MRELRL = TRUE AND MAUT2 = FALSE: yes (4.0/1.0)

MRELRL = TRUE AND MHHUUR = FALSE: no (4.0)

MRELRL = TRUE AND MAANTHUI > 1 AND APERSAUT <= 1 AND MINKGEM > 4 AND PLEVEN <= 3: yes (3.0)

MRELRL = TRUE AND PBYSTAND > 2 AND MOPLLEVE = 3: yes (3.0/1.0)

MRELRL = TRUE AND MGODRE = TRUE AND APERSAUT <= 1 AND MBERMIDD > 4: yes (14.0/2.0)

MRELRL = TRUE AND MGODRE = TRUE AND MBERMIDD <= 4 AND APERSAUT > 1 AND MGEMOMV > 2: yes (5.0)

MRELRL = TRUE AND MBERMIDD <= 4 AND MGODRE = TRUE AND PGEZONG <= 0 AND PINBOED <= 0 AND APERSAUT <= 1 AND MSKCL = A AND MKOOPKLA > 7 AND PBRAND <= 4: yes (9.0/1.0)

MGODRE = TRUE AND MRELRL = TRUE AND APERSAUT <= 1 AND MSKCL = A: yes (10.0/4.0)

MOPLLEVE = 3 AND PINBOED <= 0 AND MGODRE = TRUE AND MRELRL = TRUE AND MAUT0 = TRUE AND MKOOPKLA <= 5 AND MBERZELF > 0: yes (6.0)

MBERBOER <= 0: yes (7.0)

Az Occam-féle borotva elvet alkalmazva, miszerint a legegyszerűbb magyarázat a legnagyobb valószínűséggel a helyes – erre a következtetésre jutottam:

Vallásos emberek, egészség- vagy életbiztosítással már rendelkeznek, nagy házuk van, vagy magas szociális osztályba tartoznak, másszóval a tehetősebb emberek.

Mellékelt **szabog-adb.tar.gz** file tartalma:

szabog-adb.odt	A feladat kidolgozott megoldása OpenOffice formátumban
szabog-adb.pdf	A feladat kidolgozott megoldása PDF formátumban
Coil original data/	A versenyen kiadott anyag
dictionary.txt	Adatszótár
targets.txt	Célértékek
ticdata.txt	Tanuló adatbázis
ticeval.txt	Teszt adatbázis
My data/	Saját adatok, az eredeti átalakítása adatbányászati eszközökkel
dictionary.txt	Adatszótár
ticdata.arff	Tanuló adatbázis weka formátumban
ticdataset.xls	Adatbázis átalakításához használt segéd file
ticeval.arff	Teszt adatbázis weka formátumban
ticeval.txt	Teszt adatbázis
ticvalid.arff	Validációs adatbázis weka formátumban
targets.txt	Célértékek
ticdata.csv	Tanuló adatbázis csv formátumban
ticdata.txt	Tanuló adatbázis
ticeval.csv	Teszt adatbázis csv formátumban
tictraining.arff	Tanuló adatbázis (kivéve belőle a validációs adatbázist) weka formátumban

Szabó Gábor

Programtervező Informatikus Msc.

Levelező 2. évfolyam