PËRDORIMI I NJOHURISË MAKINË NË FUSHËN E KRIMINALISTIKËS

OLTA PETRITAJ.[1], DENAJDA ÇIBUKU.[2]
[1]Kolegji Universitar "Logos" Tiranë**,** Fakulteti i Shkencave të Aplikuara,
Departamenti Informatikë e Zbatuar
[2]Kolegji Universitar "Logos" Tiranë, Fakulteti Ekonomik,
Departamenti Menaxhim Turizëm
e-mail: olta.petritaj@gmail.com

Përmbledhje

Data mining është një disiplinë e njohurisë makinë, e cila përdoret gjërësisht në shumë fusha të jetës. Një nga fushat në të cilën ka gjetur përdorim të madh është kriminalistika. Data mining siguron metodat dhe teknologjinë për transformimin e të dhënave në informacion të dobishëm për marrjen e vendimeve. Qëllimi i këtij artikulli është të vlerësojë metodat e klasifikimit data mining dhe performancën e tyre mbi të dhënat që rregjistrohen në lidhje me krimet e kryera. Metodat e klasifikimit data mining i krahasojmë teorikisht dhe praktikisht pas zbatimit të tyre mbi të dhënat e grumbulluara në ambjentet e paraburgimit në Kavajë dhe Tiranë. Disa atribute te datasetit janë gjinia, mosha, statusi i punësimit, lloji i krimit, vendi i krimit. Metodat do të zbatohen mbi këto të dhëna për të përcaktuar efektivitetin e tyre në zbulimin dhe parandalimin e krimit. Vlerësimet e realizuara mbi të dhënat treguan se metoda me performancën më të lartë është pema e vendimit. Këtë e dëshmuan disa tregues të performancës si: numri i instancave te klasifikuara saktë, saktësia ose preçizioni dhe recall, të cilat kishin vlerat më të larta në krahasim me metodat e tjera. Ne zbulojmë se metodat e klasifikimit data mining kontribuojnë në parashikimin për mundësinë e ndodhjes së krimit dhe si rezultat në parandalimin e tij.

Fjalëkyçe: Njohuri makinë, klasifikim, dataset, kriminalistikë.

Abstract

Data mining is a discipline of machine learning, which is widely used in many fields of our life. Data mining has become a fundamental methodology in criminology. It provides the methodology and technology to transform data into useful information for decision making. The purpose of this paper is to evaluate data mining classification methods and their performance that can be used for analyzing the collected data about the crimes committed. We will identify the most appropriate data mining classification methods to analyze the collected data from sources specialized in crime prevention by comparing theoretically and practically since their implementation on the collected data in detention areas in Kavaja and Tirana. Some attributes of this dataset are, gender, age, employment status, persons with whom they live, type of crime, crime place. Methods will be applied on these data to determine their effectiveness in detecting and preventing crime. Performed evaluations on the data showed that the method with a higher performance is decision tree. This was showed by some performance measures, as the number of instances correctly classified, accurancy or precision and recall, that has brought more high values compared with other methods. We find out that the data mining classification methods will contribute to the predictions on the possibility of occurrence of the crime and as a result in its prevention.

Key Words: Machine learning, classification, dataset, criminology.

Hyrje

Krimi, parashikimi dhe parandalimi i tij

Krimi është një fenomen kompleks social dhe kostoja e tij është rritur për shkak ndryshimeve të mëdha të shoqërisë, dhe kështu, agjensitë e zbatimit të ligjit, si ajo e policisë duhet të mësojnë faktorët që sjellin rritjen e tendencës së ndodhjes së krimit. Për të frenuar këtë të keqe sociale ka gjithmonë nevojë për strategji dhe politika të matura për parandalimin e krimit.

Si pasojë e zhvillimit të teknologjisë, shkencës dhe informacionit, data mining dhe mjetet e inteligjencës artificiale janë gjithnjë e më të përhapura në komunitetin e zbatimit të ligjit. Agjencitë e zbatimit të ligjit përballen me nje volum të madh të dhënash që duhen të përpunohen dhe të transformohen në informacione të dobishme dhe data mining mund të përmirësojë analizën e krimit duke ndihmuar në parashikimin dhe parandalimin e tij. Nga përpunimi dhe procesimi i të dhënave kriminale agjencitë e zbatimit të ligjit mund të ekstraktojnë modele krimi që mund të jenë të rëndësishëm në procesin e parandalimit të krimit. Mjetet e data mining rritin shpejtësinë e analizimit të të dhënave dhe nëpërmjet tyre analistët janë në gjendje që të shqyrtojnë data setet ekzistuese, për të identifikuar modelet vepruese dhe tendencat e krimit.

Baza të ndryshme të dhënash mund të përdoren në sistemin tonë të parashikimit dhe të parandalimit të krimit për të analizuar krimin në mënyrë inteligjente. Bazat e të dhënave mund të jenë të veçanta, të tilla si bazat e të dhënave për një vepër penale të caktuar si vjedhje, ose të shoqëruara me një model të krimeve. Sistemi i Menaxhimit i Bazës së të Dhënave është krijuar për menaxhimin e rasteve dhe numërimin e përgjithshëm të krimeve dhe jo për të kryer analizën e informacionit. Duke shtuar në sistemin e menaxhimit të bazës së të dhënave metodat data mining është bërë e mundur që të kemi një sistem, i cili të funksionojë si një ekspert në kriminalistikë.

Përdorimi i data mining në kriminalistikë

Kriminalistika është një fushë ku fokusohet studimi shkencor i krimit dhe sjellja kriminale. Ajo është një proces që synon të identifikojë karakteristikat e krimit. Kjo është një nga fushat më të rëndësishme ku aplikimi i teknikave data mining mund të prodhojë rezultate të rëndësishme. Data mining është një mjet shumë i dobishëm i cili mund të ndihmojë dhe të mbështesë agjencitë e zbatimit të ligjit.

Analiza e Krimit si një pjesë e kriminalistikës, ka si detyrë eksplorimin dhe zbulimin e krimit dhe marrëdhëniet e tij me kriminelë. Zbatimi i ligjit është një proces që synon të identifikojë karakteristikat e krimit. Identifikimi i karakteristikave të krimit është hapi i parë për zhvillimin e analizave të mëtejshme. Vëllimi i lartë i grupeve të të dhënave të krimit dhe gjithashtu kompleksiteti i marrëdhënieve ndërmjet këtyre të dhënave kanë bërë që

kriminalistika të jetë një fushë e përshtatshme për aplikimin e teknikave data mining.

Data mining, kur aplikohet për analizë taktike të krimit, është një zbulim, një mjet që mund të përdoret për të shqyrtuar shumë dataset-e të mëdha që përfshijnë një grup të madh të variablave, përtej asaj që një analist i vetëm, apo edhe një ekip analitik ose task forcë, mund të shqyrtojë në mënyrë të saktë. Ashtu si çdo problem tjetër për zgjidhjen e metodave, detyra e data mining fillon me një përkufizim të problemit. Identifikimi i problemit data mining mundëson përcaktimin e procesit data mining dhe teknikën e modelimit. Data mining u siguron agjencive të zbatimit të ligjit mundësinë për të mësuar mbi trendet e krimit, si dhe pse krimet janë kryer. Duke përdorur metodat data mining përmirësojmë analizën e krimit dhe ndihmojmë reduktimin dhe parandalimin e tij.

Metodologjia

Ne krahasojmë praktikisht metodat e klasifikimit të njohurisë makinë për të zbuluar metodën më të përshtatshme për natyrën e të dhënave tona. Metodat u krahasuan duke aplikuar algoritme të njohurisë makinë me të dhëna konkrete në mjedisin Weka. Metodat e klasifikimit që janë përdorur janë Pema e Vendimit, Naive Bayes, Rrjetat Neurale dhe Support Vector Machine. Të dhënat e analizuara janë mbledhur me anë të pyetësoreve në institucionet e vuajtjes së dënimit. Ato kanë të bëjnë me zonat ku ndodhin krimet dhe të dhënat e personave që i kryejnë ato. Disa nga tiparet që kemi marrë në shqyrtim janë mosha, gjinia, statusi i punësimit, lloji i krimit, vendi i krimit.

Klasifikimi

Klasifikimi është një teknikë data mining që i kategorizon të dhënat në mënyrë që të ndihmojë në parashikime dhe analiza më të sakta. Ai është një nga metodat data mining që synon të bëjë analizën e grupeve shumë të mëdha të të dhënave. Përdoret për nxjerrjen e modeleve që përcaktojnë me saktësi klasat e të dhënave të rëndësishme brenda data set-it. Klasifikimi konsiston në parashikimin e një rezultati të caktuar bazuar në një input të dhënë (Witten, *et.al*, 2011). Për të parashikuar rezultatin, algoritmi i klasifikimit përpunon një set trajnimi që përmban disa atribute dhe rezultatin perkatës, zakonisht i quajtur atribut i qëllimit ose parashikimit. Algoritmet e klasifikimit përpiqen të zbulojnë marrëdhëniet midis atributeve që do të bënin të mundur parashikimin e rezultatit. Ata analizojnë inputin dhe prodhojnë një parashikim. Një algoritëm klasifikimi përdor vërejtje të vlefshme që janë mbledhur në të kaluarën për të identifikuar një model që parashikon klasën e vërejtjeve të ardhshme për të cilat vlerat e atributeve nuyyk janë të njohura. Metodat e klasifikimit data mining përpunojnë një sasi të madhe të të dhënave. Për të bërë një klasifikim janë në dispozicion një sërë vëzhgimesh, zakonisht të përfaqësuara nga bashkësia e të dhënave (data set), ku klasa "output" e tyre nuk është e njohur (Witten, *et.al*, 2011; Gorunescu, 2011). Në shumicën e aplikimeve atributi output është i

përfaqësuar nga një ndryshore binare. Natyra kategorike e atributit output përcakton dallimin ndërmjet klasifikimit dhe regresionit. Metodat e klasifikimt janë krahasuar për të gjetur metodën më të përshtatshme për natyrën tonë të të dhënave. Një krahasim i këtyre metodave është paraqitur në vijim.

Pema e vendimit

Pema e vendimit është një metodë në të cilën të dhënat paraqiten në një strukturë pemë të bazuar në vlerat e atributeve të tyre. Ajo ndan të dhënat në bazën e të dhënave në subsets bazuar në vlerat e një ose më shumë fushave. Ky proces do të përsëritet për çdo nëngrup në mënyrë rekursive derisa të gjitha instancat të jenë një nyje në një klasë të vetme. Rezultati i pemës së vendimit është një strukturë në formë peme që përshkruan një seri të vendimit të dhënë në çdo hap (Witten, *et.al*, 2011). Pastaj këto vendime konsiderohen si rregulla për detyrën e klasifikimit.

Në një pemë vendimi shifrat në formën ovale tregojnë atributet dhe degezimet përshkruajnë vlera të ndryshme të atributeve dhe forma e rrethit në nyjet e gjetheve përfaqësojnë klasat. Nyja në krye të pemës së vendimit është e njohur si një nyje rrënjë. Nyjet në pjesën e poshtme të pemës së vendimit, të cilat portretizojnë klasa, quhen nyje fletë. Nyjet midis nyjes rrënjë dhe nyjes fletë quhen nyje të brendëshme (Phyu, 2009). Edhe pse ka një numër algoritmesh të pemës së vendimit të zhvilluara deri tani, ata të gjithë ndajnë karakteristika të dëshirueshme në paraqitjen dhe përshkrimet strukturore, dmth rregullat e kuptueshme. Algoritmet që përdoren zakonisht për ndërtimin e pemëve të vendimit janë; CART, CHAID dhe C4.5. CART (Classification And Regression Tree) ndërton një pemë binare me ndarjen e të dhënave në secilën nyje sipas një funksioni me të dhëna të një fushe të vetme. Nga ana tjetër, CHAID (Chi-squared Automatic Interaction Detection) është përdorur për zbulimin e marrëdhënieve statistikore ndërmjet variablave nëpërmjet ndërtimit të një pemë vendimi (Ibid). C4.5 është 'përça-dhe-sundo' afrohet tek pemët e vendimit të induksionit që është ID3 (Iterative Dichotomiser 3). Ajo është në dispozicion të paketave software-ike sepse ka më popullaritet dhe është lehtësisht e përdorshme.

Rrjetet neurale artificiale (Anns)

Rrjetet neurale artificiale (ANNs) janë lloje të arkitekturës kompjuterike të frymëzuar nga rrjetet nervore biologjike (sistemet nervore të trurit) dhe përdoren për të përafruar funksionet që mund të varen nga një numër i madh inputesh dhe përgjithësisht janë të panjohura (Nikam, 2015).

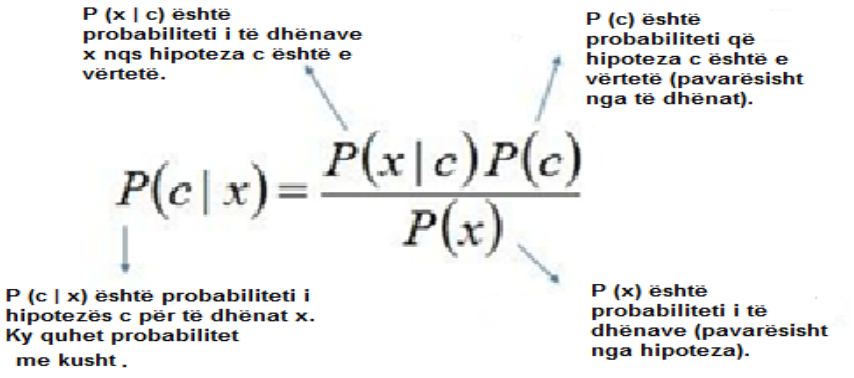
Rrjetet neurale konsiderohen të jenë kuti e zezë për shkak të sjelljeve jo-lineare të tyre dhe janë zakonisht më të komplikuara se teknikat e tjera. Trajnimi i një rrjeti neural është një sfidë që kërkon vendosjen e parametrave të shumta dhe prodhimi i një rrjeti neural nuk është aq i lehtë të kuptohet nga përdoruesi si output i parë në krahasim me atë të pemës së vendimit. Struktura e rrjetit neural është shumë e ngjashme me strukturën e neuroneve

në trurin e njeriut. Të gjitha proceset e një rrjeti neural kryhen nga ky grup i neuroneve apo njësive. Çdo neuron është një pajisje e veçantë komunikimi, duke bërë punën e saj relativisht të thjeshtë. Funksioni i një njësie është thjesht që të marrë të dhëna nga njësitë e tjera, si një funksion i inputeve që merr për të llogaritur një vlerë të prodhimit, të cilat ajo i dërgon njësive të tjera. Në rrjetet neurale artificiale, neuronet janë të grupuara në shtresa, klasifikuar shpesh si input të fshehur dhe shtresa output. Algoritmet e trajnimit klasifikohen si trajnim i mbikëqyrur dhe pa mbikëqyrje (Ibid). Në rastin e parë algoritmi i të mësuarit bën dallimin ndërmjet outputit të saktë dhe parashikimit të rrjeteve neurale në mënyrë që parashikimi i ardhshëm të jetë më afër përgjigjes së saktë (Witten, *et.al*, 2011; Gorunescu, 2011). Rrjeti neural ka për të dhënë shumë shembuj herë pas here në mënyrë që të mësojë dhe të bëjë parashikim të saktë. Ne rastin tjeter rrjeti është dhënë thjesht me një numër të inputeve dhe organizon veten në një mënyrë të tillë që të dalë me klasifikimin e vet të inputeve.

Klasifikimi Bayesian (Naive Bayes)

Teknika e Naive Bayes Classifier bazohet në teoremën Bayesian dhe përdoret veçanërisht kur dimensionaliteti i inputeve është i lartë. Klasifikuesi Bayesian është i aftë të llogarisë output-in më të mundshëm bazuar në input. Gjithashtu është e mundur të shtojmë të dhëna të papërpunuara në mënyrë të vazhdueshme dhe të kemi një klasifikues më të mirë probabilistik. Një klasifikues Naive Bayes konsideron se prania (ose mungesa) e një veçorie (atributi) të një klase nuk është e lidhur me praninë (ose mungesën) e ndonjë veçorie tjetër kur jepet ndryshorja e klasës. Edhe nëse këto karakteristika varen nga njëra-tjetra ose nga ekzistenca e karakteristikave të tjera të një klase, një klasifikues Naive Bayes konsideron të gjitha këto karakteristika të pavarura për të kontribuar në probabilitetin e një ngjarjeje (Nikam, 2015). Algoritmi funksionon si më poshtë.

Teorema e Bayes ofron një mënyrë për të llogaritur probabilitetin e një hipoteze të bazuar në njohuritë tona të mëparshme.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

• P (c | x) është probabiliteti i hipotezës (klasës) c për të dhënat x. Ky quhet probabilitet me kusht.

• P (c) është probabiliteti që hipoteza c është e vërtetë (pavarësisht nga të dhënat).

• P (x | c) është probabiliteti i të dhënave x nqs hipoteza c është e vërtetë.

• P (x) është probabiliteti i të dhënave (pavarësisht nga hipoteza).

Klasa (c) është e pavarur nga vlerat e parashikuesve të tjerë. Klasifikuesi Naïve Bayes mund të trajnohet në mënyrë efiçiente në të mësuarin e supervizuar[1]. Pas llogaritjes së probabilitetit me kusht për një numër të ndryshëm të hipotezave, ne mund të zgjidhim hipotezën (klasën) me probabilitetin më të lartë.

Support Vector Machine (SVM)

SVM-të kanë tërhequr një vëmendje të madhe në dekadën e fundit dhe aplikohen në mënyrë aktive në aplikacione të fushave të ndryshme. SVM-të zakonisht përdoren për klasifikimin e mësimit, regresionin ose funksionin e renditjes. SVM janë të bazuara në teorinë e të mësuarit statistikor dhe minimizimin e rrezikut strukturor dhe kanë për qëllim përcaktimin e vendndodhjes së kufijve të vendimeve të njohur gjithashtu si hyperplane që prodhojnë ndarjen optimale të klasave (Gorunescu, 2011; Nikam, 2015). Efikasiteti i klasifikimit të bazuar në SVM nuk varet direkt nga dimensioni i subjekteve të klasifikuara. SVM gjithashtu mund të zgjerohet për të mësuar funksionet e vendimeve jo-lineare duke projektuar së pari të dhënat hyrëse (input) në një hapësirë të një dimesioni të lartë duke përdorur funksionet e kernelit dhe duke formuluar një problem të klasifikimit linear në atë hapësirë. SMO (Support Vector Machine) zbaton algoritmin minimal optimizues minimal sekuencial të John C. Platt për trajnimin e një klasifikuesi Support Vector duke përdorur kernelët polinomial ose RBF. Ky zbatim në mënyrë globale zëvendëson të gjithë të vlerat e humbura dhe transformon atributet nominale në ato binare. Ai gjithashtu normalizon të gjithë atributet sipas parazgjedhjes (Jami, 2016).

Rezultatet eksperimentale

Për të kryer këtë studim kemi përdorur softuerin Weka duke u nisur nga qasja dhe familjariteti në përdorim që ai ka. Paketa softuerike Weka ka programe të ndryshme për teknika dhe algoritme të ndryshëm. Eksperimentet për vlerësimin e modelit janë bërë duke përdorur cross-validation.

Në tabelën 1 është paraqitur një krahasim i rezultateve të algoritmeve të aplikuara mbi të dhënat tona në Weka.

[1] Supervised Learning të mësuarit duke u bazuar nga të dhënat trajnuese.

Emri i algoritmit	CCI(Instancat e klasifikuara saktë)	ICI(Instancat të klasifikuara josaktë)	Recall	Precision
C4.5	70 (77.7778 %)	20 (22.2222 %)	0.778	0778
ANNs	65 (72.2222 %)	25 (27.7778 %)	0.722	0.730
Naive Bayes	65 (72.2222 %)	25 (27.7778 %)	0.722	0.730
SVM(SMO)	58 (64.4444 %)	32 (35.5556 %)	0.644	0.642

Tabela 1. Krahasimi i rezultateve të algoritmeve të aplikuara në Weka

Në këtë studim kemi përdorur algoritmin C4.5, i cili është një algoritëm i pemëve të vendimit. Ky algoritëm është përdorur sepse ai siguron një rezultat të qartë dhe të lehtë për tu interpretuar. Ndërtimi i modelit është bërë duke modifikuar vlerat e parametrave dhe ky algoritëm i klasifikon të dhënat mbi krimin me një saktësi më të lartë në krahasim me algortimet e tjerë të metodave të klasifikimit data mining. Ne konvertuam të dhënat tona në formatin CSV. Algoritmi C4.5 u zbatua në këto të dhëna. Ajo se çfarë doli nga ky algoritëm, vizualizimi dhe pema e vendimit paraqiten në figurën 1, figurën 2 dhe figurën 3.

```
=== Summary ===

Correctly Classified Instances          70               77.7778 %
Incorrectly Classified Instances        20               22.2222 %
Kappa statistic                          0.5411
Mean absolute error                      0.3012
Root mean squared error                  0.3881
Relative absolute error                 62.1711 %
Root relative squared error             78.876  %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.730     0.189     0.730      0.730     0.730      0.541     0.839     0.755     urbane
               0.811     0.270     0.811      0.811     0.811      0.541     0.839     0.849     rurale
Weighted Avg.  0.778     0.237     0.778      0.778     0.778      0.541     0.839     0.810

=== Confusion Matrix ===

  a  b   <-- classified as
 27 10 |  a = urbane
 10 43 |  b = rurale
```

Figura 1: Algoritmi C4.5

Numri i instancave të klasifikuara saktë është 70 ose 78%, ndërsa numri i instancave të klasifikuara josaktë është 20 instanca ose 22%.

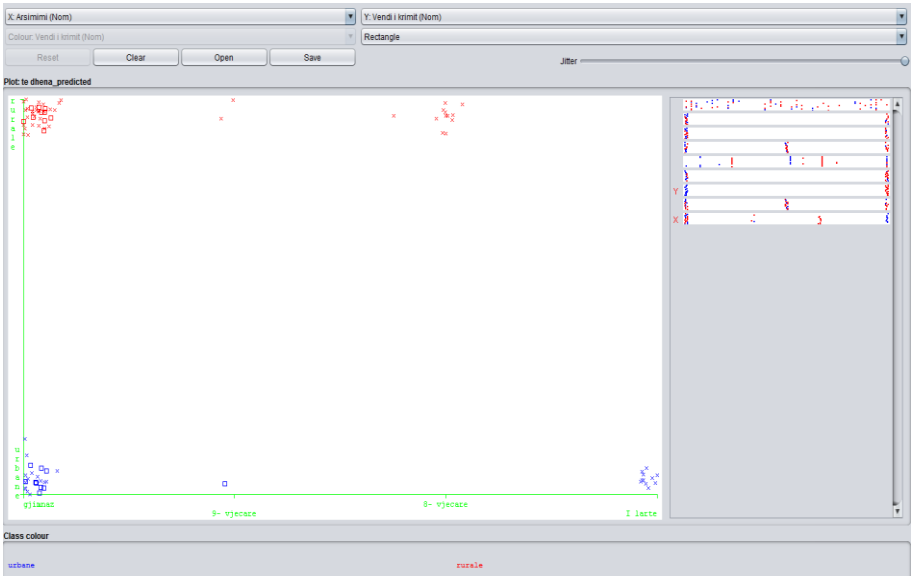


Figura 2: Gabimet e klasifikuesit në një rrjet kordinativ dy-dimensional

Figura 2 tregon gabimet e klasifikuesit në një rrjet kordinativ dydimensional. Ne mund të zgjedhim se cilat attribute të përdorim për X dhe Y duke përdorur kutitë e selektimit në krye.

Figura 3: Pema e vendimit

Zbatimi i këtij algoritmi ka klasifikuar të dhënat mbi krimin në bazë të atributeve të datasetit si psh. vendi ku ka ndodhur krimi(zonat urbane, zonat rurale) ku: numri i instancave të klasifikuara saktë, saktesia ose preçizioni

dhe recall kanë vlerat më të larta krahasuar me algoritmet e tjerë të metodave të klasifikimit.

Konkluzione

Qëllimi i këtij studimi ishte për të shqyrtuar zbatueshmërinë e metodave të klasifikimit data mining në procesin e parandalimit të krimit.

Rezultatet e eksperimenteve të kryera në këtë hulumtim duke përdorur pemën e vendimit kanë zbuluar se metodat e klasifikimit data mining janë të zbatueshme në procesin e parandalimit të krimit. Pema e vendimit si metodë klasifikimi data mining ka klasifikuar të dhënat e krimit në normë saktësie prej 78%.

Kjo metodë ka treguar rezultate premtuese për problemin e parandalimit të krimit pasi norma e saktësise është e lartë në eksperimentet e kryera. Për më tepër, pema e vendimit duket më e zbatueshme për shkak të faktit se në kontrast me algoritmet e tjera, ajo shpreh rregullat në mënyrë eksplicite. Këto rregulla mund të shprehen në gjuhën e njeriut në mënyrë që çdokush mund ti kuptojë. Pema e gjeneruar nga eksperimentet mbi pemën e vendimit ka treguar se atributet e shkelësve të tilla si gjinia, niveli arsimor, statusi civil dhe mosha ndikojnë për të përcaktuar nëse krimi është kryer në zonat rurale apo në zonat urbane.

Përdorimi i njohurisë makinë në fushën e kriminalistikës është i rëndësishëm sepse metodat e kalsifikimit data mining mund të përdoren në procesin e vendimmarrjes.

Literatura

Ian H. Witten, Eibe Frank, Mark A. Hall (2011): Data Mining Practical Machine Learning Tools and Techniques, Elsevier

Florin Gorunescu (2011): Data Mining: Concepts, Models and Techniques, Springer

Thair Nu Phyu, (2009): Survey of Classification Techniques in Data Mining Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol I IMECS, Hong Kong

Sagar S. Nikam, (2015): A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental Journal of Computer Science & Technology, ISSN: 0974-6471

Layla Safwat Jami, (2016): Data Analysis Based On Data Mining Algorithms Using Weka Workbench, International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655