ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო
უნივერსიტეტის ზუსტ
და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი

# ჯულია შავკაციშვილის

სამაგისტრო ნაშრომი

თემაზე

# მონაცემთა მოძიების მეთოდები SQL- ის რესურსების გამოყენებით

ნაშრომი შესრულებულია საინფორმაციო ტექნოლოგიების მაგისტრის აკადემიური ხარისხის მოსაპოვებლად

ხელმძღვანელი: ქ-ნი მაია არჩუაძე

2019 წელი

# ანოტაცია

SQL Server, ლიდერი პროგნოზირებად ანალიტიკაში 2000 წლიდან, უზრუნველყოფს მონაცემთა მოძიებას ანალიზში. SQL Server Data Mining წარმოადგენს ინტეგრირებლ პლატფორმას ანალიტიკაში, რომელიც მოიცავს მონაცემთა წმენდას და მომზადებას, მანქანურ სწავლებას და ანგარიშგებას. SQL Server Data Mining მოიცავს სხვადასხვა ალგორითმებს, მათ შორის EM (მონაცემთა მაქსიმიზაციის ალგორითმი) და K-საშუალოს კლასტერიზაციის მოდელებს, ნეირონულ ქსელებს, ლოჯისტიკურ და ხაზოვან რეგრესიას, ბაიესის კლასიფიკატორებს. ყველა მათგანს აქვს ინტეგრირებული ვიზუალი, რაც გვეხმარება მოდელების შემუშავებაში, დახვეწასა და შეფასებაში.

დღეისათვის არსებობს მრავალი ალგორითმი და მეთოდი, რომელიც გამოიყენება მონაცემთა ანალიზისთვის. ძირითადი პრობლემა სპეციფიური პრობლემის შემთხვევაში მონაცემთა ამოღებისათვის შესაფერისი ალგორითმის პოვნაა. ნაშრომში გამოკვლეული და გაანალიზებულია ის ფაქტორები, რომლებიც ზეგავლენას ახდენს შესაბამისი ალგორითმის შერჩევაზე.

სამაგისტრო ნაშრომი შეიძლება სასარგებლო იყოს სამეცნიერო მუშაკებისათვის, პედაგოგებისათვის, დოქტორანტებისათვის, მაგისტრანტებისათვის, ბაკალავრებისათვის და უმაღლესი სასწავლო დაწესებულებების სტუდენტებისათვის. მისი ძირითადი დასკვნები შეიძლება სხვადასხვა ორგანიზაციების შემდგომი კვლევების ჩატარებისას, აგრეთვე პედაგოგიურ პრაქტიკაში იქნეს გამოყენებული.

საკვანძო სიტყვები: SQL Server Data mining, k-საშუალოს კლასტერზაცია, გადაწყვეტილებათა მიღების ხეები.

# Abstract

SQL Server has been a leader in predictive analytics since the 2000 release, by providing data mining in Analysis Services. SQL Server Data Mining provides an integrated platform for predictive analytics that encompasses data cleansing and preparation, machine learning, and reporting. SQL Server Data Mining includes multiple standard algorithms, including EM and K-means clustering models, neural networks, logistic regression and linear regression, decision trees, and naive bayes classifiers. All models have integrated visualizations to help you develop, refine, and evaluate your models.

Today there are many algorithms and methods used for analysis of data. The main problem is to find a suitable algorithm for data retrieval in case of a particular problem. The work is examined and analyzed by the factors influencing the selection of the algorithm.

Master's thesis may be useful for scientific workers, teachers, doctoral students, graduate students, bachelors and students of higher education institutions. Its main conclusions can be used to conduct further studies in various organizations, as well as in pedagogic practice.



Keywords— SQL Server Data mining, k-means clustering, Decision Tree

# შინაარსი

# შესავალი

მონაცემთა ინტელექტუალურმა ანალიზმა ან მონაცემთა ბაზებში მონაცემთა მოპოვებამ (KDD), მიიპყრო უკვე მრავალი სამეცნიერო სფეროს ყურადღება. ბოლო ორი ათწლეულის განმავლობაში მონაცემები განსაკუთრებით გაფართოვდა, ამიტომ აუცილებელი გახდა გამოყენებულიყო მონაცემთა ინტელექტუალური ანალიზის მეთოდები და ალგორითმები, რომლებიც შემუშავებულ იქნა აქამდე. ყველა ეს მეთოდი და ალგორითმი გამოიყენება სწრაფად მზარდი მონაცემთა უზარმაზარი ბაზიდან ინფორმაციის მისაღებად.

მონაცემთა ბაზების ზრდას მივყავართ ახალი მეთოდებისა და კვლევების შემუშავებისაკენ, რათა მიღწეულ იქნას უკეთესი შედეგები ისეთი მსხვილი კომპანიების მიმდინარე და სამომავლო მოთხოვნილებების შესასრულებლად, რომლებიც ეძებენ სხვადასხვა გადაწყვეტილებებს. არის რამდენიმე სფერო, სადაც ეს მოთხოვნილება განსაკუთრებით იგრძნობა, მაგალითად გაყიდვა/მარკეტინგი, მყიდველთა ქცევები, თაღლითობის შემთხვევების გამოვლენა, საკრედიტო სკორინგი და ა.შ.

მონაცემთა ინტელექტუალური ანალიზი ძირითადად ეხმარება კლასიფიკაციასთან კლასტერიზაციასთან და ასოციაციის წესებთან დაკავშირებული პრობლემების გადაჭრაში. რამდენადაც რეალურ დროში წამოიჭრა მონაცემთა მოპოვების საკითხი, უამრავი მეთოდი და ალგორითმი იქნა აგებული მონაცემთა ბაზებიდან საჭირო ინფორმაციის გამოსაყოფად. მაგალითად: მონაცემთა მოძიების მეთოდებით შესაძლებელია დაკვირვება ვაწარმოოთ მომხმარებელთა ქცევაზე მათ მიერ შექმნილი მონაცემების საფუძველზე დროის გარკვეული პერიოდის განმავლობაში.

სხვა აღსანიშნავი ფაქტორია ის, რომ მონაცემები ინახება და ითვლება Oracle-ს მონაცემთა ბაზიდან, რაც ძალიან მნიშვნელოვანია, რადგან მონაცემთა ბაზა Oracle წარმოადგენს დეფაქტო სტანდარტს, რომელიც გამოყენებულია მთელს მსოფლიოში სხვადასხვა ორგანიზაციებში. ამრიგად, მოდელი შეიძლება განვიხილოთ, როგორც ზოგადი დანიშნულების და შეიძლება გამოყენებულ იქნას ნებისმიერ დაწესებულებაში საკითხების გადასაწყვეტად ან არსებულის ოპტიმიზაციისათვის.

# მონაცემთა მოძიების ალგორითმები

ალგორითმები  მონაცემთა მოძიებაში (ან მანქანურ სწავლებაში) არის ჰუმერისტიკის და გათვლების კომპლექტი, რომელიც ქმნის მოდელს ბაზიდან. მოდელის შესაქმნელად, ალგორითმი პირველად აანალიზებს მიწოდებული მონაცემებს, ეძებს კონკრეტული ტიპის ნიმუშებს ან ტენდენციებს. ალგორითმი ამ ანალიზის შედეგებს იყენებს, რათა იპოვოს ოპტიმალური პარამეტრები საძიებელი მოდელის შესაქმნელად. ეს პარამეტრები შემდეგ გამოიყენება მთელი მონაცემებისთის, რათა შეიქმნას ქმედითი ნიმუშები და დეტალური სტატისტიკა.

ბაზიდან ალგორითმის მიერ მიღებულმა საძებელმა მოდელმა შესაძლებელია მიიღოს სხვადასხვა ფორმა, მათ შორისაა:

- კლასტერთა კომპლექტი, რომელიც აღწერს, თუ როგორ არის დაკავშირებული მონაცემები ბაზაში.
- გადაწყვეტილების ხე, რომელიც პროგნოზირებს შედეგს და აღწერს, თუ რამდენად განსხვავდება კრიტერიუმები ამ შედეგზე.
- მათემატიკური მოდელი, რომელიც პროგნოზირებს გაყიდვებს.
- წესების კომპლექტი, რომელიც აღწერს როგორი პროდუქტებია დაჯგუფებული ერთად ტრანზაქციაში და ალბათობა, რომ ეს პროდუქტები შეძენილია ერთად.

SQL Server Data Mining-ში ალგორითმები არის ყველაზე პოპულარული და კარგად გამოკვლეული მეთოდების ნიმუშები. მაგალითად, K-საშუალოს კლასტერიცაზია არის ერთერთი ყველაზე ძველი კლასტერული ალგორითმი და ხელმისაწვდომია სხვადასხვა სახის იმპლემენტაციებით. თუმცა, K-means კლასტერული ალგორითმის პრაქტიკული გამოყენება  SQL Server Data Mining-ში შემუშავდა Microsoft Research-ის მიერ და შემდეგ ოპტიმიზირდა ანალიზის სერვისებით. მონაცემთა მოძიების ყველა ალგორითმი არის მოქნილი და სრულად პროგრამირებადი API-ს (Application programming interface) გამოყენებით.

## სწორი ალგორითმის შერჩევა

საუკეთესო ალგორითმის შერჩევა განსაკუთრებული ანალიტიკური საკითხისთვის არის გამოწვევა. განსხვავებულმა ალგორითმება ერთ საკითხზე შესაძლებელია მოგვცეს სხვადასხვა შედეგი, ან ერთმა ალგორითმა დაგვიბრუნოს სხვადახვა შედეგი. მაგალითად, გადაწყვეტილების ხეების (Decision Trees) ალგორითმი არ გამოიყენება მხოლოდ პროგნოზიებისთვის, ის, ასევე, არის გზა შემცირდეს სვეტების რაოდენობა მონაცემთა ბაზაში, რადგან Decision Trees-ს შეუძლია დააიდენტიფიციროს არასაჭირო სვეტი საბოლოო საძიებელი მოდელისთვის.

SQL Server Data Mining მოიცავს შემდეგი ტიპის ალგორითმებს:

- კლასიფიკაციის ალგორითმები (Classification algorithms) პროგნოზირებს ერთ ან მეტ დისკრეტულ ცვლადს მონაცემთა ბაზაში სხვა ატრიბუტების მიხედვით.
- რეგრესიული ალგორითმები (Regression algorithms ) პროგნოზირებს ერთ ან მეტ უწყვეტ რიცხვით ცვლადს, როგორიცაა მოგება ან ზარალი, მონაცემთა ბაზაში სხვა ატრიბუტების საფუძველზე.
- სეგმენტაციის ალგორითმები (Segmentation algorithms) ყოფს მონაცემებს, რომელთაც მსგავსი თვისებები აქვს, ჯგუფებად ან კლასტერებად.
- თანმიმდევრული ანალიზის ალგორითმები (Sequence analysis algorithms) აჯამებს თანმიმდევრულად ან კონკეტულ ეპიზოდებს მონაცემებში, როგორიცაა ვებ-გვერდებზე დაწკაპუნების ან წინა ჩანაწერების შესახებ(მაგ:საიტზე შესვლა/გამოსვლა) ინფორმაციას.

თუმცა, ეს არ ნიშნავს იმას რომ შეზღუდულია ალგორითმების გამოყენება საძიებელი მოდელის მისაღებად.გამოცდილი ანალიტიკოსები ხშირად იყენებენ ერთ ალგორითმს, რომ დაადგინონ ყველაზე ეფექტური საშუალებები (ანუ, ცვლადები) და შემდეგ იყენებენ სხვადასხვა ალგორითმებს კონკრეტული შედეგის პროგნოზირებისთვის ამ მონაცემებზე დაყრდნობით. SQL Server Data Mining-ით შესაძლებელია შეიქმნას რამდენიმე მოდელი ერთი საძიებო სტრუქტურით, ასე რომ შესალებელია გამოყენებულ იქნას როგორც კლასტერული ალგორთმი, ასევე, decision trees და ბაიესის მოდელები, რომ მივღოთ მონაცემთა განსხვავებული წამოდგენები. ასევე, ცალკეული ამოცანები შეიძლება შესრულდეს რამდენიმე ალგორითმით ერთად: მაგალითად, რეგრესიით ფინანსური პროგნოზების მიღებაა

შესაძლებელი და ნერვული ქსელის ალგორითმით, შეიძლება განხორციელდეს იმ ფაქტორების ანალიზი, რომლებიც გავლენას ახდენენ პროგნოზებზე.

## ასოციაციის ალგორითმი

ასოციაციის ალგორითმი არის ალგორითმი, რომელიც ხშირად გამოიყენება რეკომენდაციული სისტემებისთვის, რომელიც მომხმარებელს რეკომენდაციას უწევს საქონელზე, რაც მათ უკვე ნაყიდი აქვთ ან რომლითაც დაინტერესდნენ. ასოციაციის ალგორითმი ასევე სასარგებლოა ბაზრის ანალიზისთვის.

ასოციაციის ალგორითმის გამოყენებით შესაძლებელია განისაზღვროს ის პროდუქტები, რომელსაც მომხმარებელი ყიდულობს ერთად, ასე რომ განისაზღვრება მათ შორის ასოციაციური დამოკიდებულება, რაც საშუალებას იძლევა დადგინდეს ის პროდუქტები რომლებიც ხშირ შემთვხვევაში იყიდება ერთად. ძირითადი მარკეტინგული ამოცანაა, რაც შეიძლება მეტი პროდუქტი შეიძინოს მომხმარებელმა, აღნიშნული მეთოდის გამოყენება მარკეტინგის მენეჯერს უადვილებს მომხმარებლებისთვის შეარჩიოს ის პროდუქტი ან პროდუტები რაზეც იოლად მიიღებს შეძენის გადაწყვეტილებას.

ერთმანეთზე დამოკიდებულ პროდუქტებს ეწოდება ჯგუფი. ერთი ჯგუფის პროდუქტები, მაგალითად, არაჟანი, რძე, კარაქი ხშირად შეგვინიშნავს, რომ მარკეტში თაროებზე მოთავსებულია ერთმანეთის გვერდით. ორი რომელიმე სახის პროდუქტს, რომელზეც მომხმარებელი უმეტეს შემთხვევაში აკეთებს არჩევანს უწოდებენ ორელემენტიან ჯგუფს. როდესაც მონაცემთა ბაზა საკმაოდ დიდია, უფრო რთულია პროდუქტებს შორის დამოკიდებულების დანახვა, განსაკუთრებით მაშინ როდესაც ხდება სამელემენტიანი, ოთხი ან უფრო რთული ჯგუფის განსაზღვრა, სწორედ ასეთ შემთხვევებში გამოიყენება ეს ალგორითმი.

ასოციაციის მოდელი შედგება რიგი ელემენტების და წესებისგან, რომლებიც აღწერენ იმას, თუ როგორაა ელემენტები ერთმანეთთან დაჯგუფებული. ასოციაციის ალგორითმი პოტენციურად პოულობს მონაცემთა ნაკრებში მრავალ წესს. პირველ ეტაპზე უნდა განისაზღვროს ჯგუფის „ზომა“, გვინდა განვსაზღვროთ ორელემენტიანი, სამელემენტიანი თუ სხვა რომელიმე; ალგორითმი იყენებს ორ პარამეტრს, მხარდაჭერას (SUPPORT) და ალბათობას (PROBABILITY), რომელიც აღწერს ელემენტებს და დაგენერირგებულ წესებს. მაგალითად, თუ X და Y წარმოადგენს ორ ნივთს, რომელიც შეიძლება იყოს კალათაში, მხარდაჭერის პარამეტრია იმ შემთხვევების რაოდენობა, რომელიც შეიცავს X და Y-ის კომბინაციით მიღებულ მონაცემებს. მომხმარებელთა მიერ განსაზღვრული პარამეტრების კომბინაციით, MINIMUM_SUPPORT და MAXIMUM_SUPPORT- თან ერთად, ალგორითმი აკონტროლებს იმ

ელემენტების რაოდენობას, რომლებიც გენერირდება. ალბათობის პარამეტრი, რომელიც ასევე იწოდება როგორც საიმედოობა, წარმოადგენს ფრაქციებს მონაცემთა ნაკრებში, რომლებიც შეიცავს X-ს და, ასევე,Y-ს. ალბათობის პარამეტრით MINIMUM_PROBABILITY პარამეტრის კომბინაციით, ალგორითმი აკონტროლებს იმ წესების რაოდენობას, რომლებიც გენერირდება.

ასოციაციის ალგორითმი შედგება სამი ძირითადი ბიჯისგან:

1. გაერთიანება. მონაცემთა ბაზაში უნდა განისაზღვროს ცალკეული პროდუქტის განმეორების სიხშირე;

2. განცალკევება. ის ჯგუფები, რომელთაც აქვთ მხარდაჭერა და საიმედოობის საკმარისი მაჩვენებელი გადადის მომდევნო იტერაციაზე ორ კომპონენტიან ჯგუფში;

3. განმეორება. პირველი ორი ბიჯი მეორდება ჯგუფში შემავალი ყოველი სიდიდისთვის მანამ, სანამ არ მიიღება ადრე განსაზღვრული ზომა.

### ასოციაციის ალგორითმის მაგალითი

ერთ-ერთი კომპანია, „The Adventure Works Cycle“, თავისი ვებ-გვერდის ფუნქციონირების რედიზაინს გეგმავს. რედიზაინის მიზანი პროდუქციის გაყიდვის ზრდაა. იმის გამო, რომ კომპანია აწარმოებს თითოეული გაყიდვის აღრიცხვას ტრანზაქციულ მონაცემთა ბაზაში, მათ შეუძლიათ გამოიყენონ ასოციაციის ალგორითმი, რათა დაადგინონ პროდუქციის კომპლექტი, რომელიც შეძენილია ერთობლივად. შემდეგ მათ შეუძლიათ დააინტერესონ მომხმარებელი დამატებითი ელემენტის პროგნოზირებით, იმის საფუძველზე თუ რომელი ნივთები აქვს მომხმარებელს უკვე კალათაში.

### როგორ მუშაობს ასოციაციის ალგორითმი

ასოციაციის ალგორითმი მიმოიხილავს მონაცემთა ნაკრებს, რათა ნახოს ერთმანეთთან დაკავშირებული შემთხვევები. ალგორითმი შემდეგ ქმნის წესებს ელემენტებიდან. ეს წესები გამოიყენება მონაცემთა ბაზაში განთავსებული საგნების არსებობის პროგნოზირებაზე, იმ სხვა სპეციფიკური საგნების საფუძველზე, რომლითაც ალგორითმი ადგენს მნიშვნელოვან ინფორმაციას. მაგალითად, ალგორითმი თუ ნახავს, რომ ვებ-გვერდის კალათაში „ტურისთვის საჭირო 1000 წვრილმანი“ არის წყლის ბოთლის ჩასადები, მაშინ წყლის ბოთლიც სავარაუდოდ კალათაში იქნება.

### მონაცემები საჭიროა ასოციაციის მოდელებისათვის

როდესაც მზადდება მონაცემები ასოციაციის წესების მოდელის გამოყენებისას, უნდა გაანალიზდეს კონკრეტული ალგორითმის მოთხოვნები, მათ შორის, რამდენი მონაცემია საჭირო და როგორ გამოიყენება ეს მონაცემები.

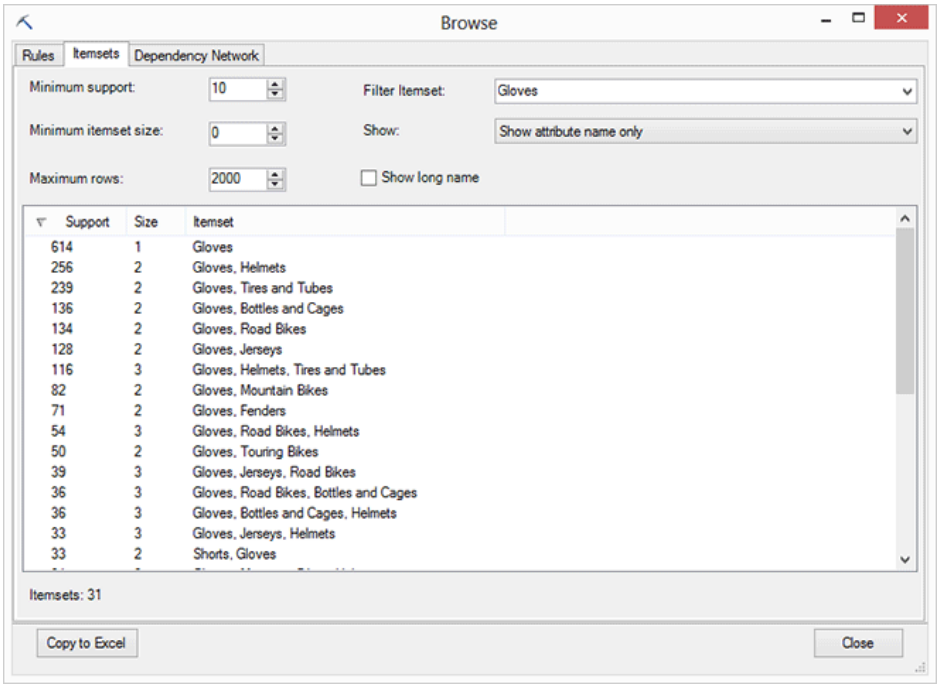
ასოციაციის წესების მოდელის მოთხოვნები შემდეგია:

- ერთი გასაღები სვეტი - თითოეული მოდელი უნდა შეიცავდეს ერთ რიცხვით ან ტექსტურ სვეტს, რომელიც ცალსახად განსაზღვრავს თითოეულ ჩანაწერს. რთული გასაღებები არ არის დაშვებული.
- ერთი პროგნოზირებადი სვეტი - ასოციაციის მოდელს შეიძლება ჰქონდეს მხოლოდ ერთი პროგნოზირებადი სვეტი. როგორც წესი, ეს არის შექმნილი (მაკავშირებელი) ცხრილის გასაღები სვეტი, სადაც შეტანილია პროდუქტები, რომელიც შეიძინა მომხმარებელმა. მნიშვნელობები უნდა იყოს დისკრეტული.
- შემომავალი სვეტები - შეყვანილი სვეტები უნდა იყოს დისკრეტული. შეყვანილი მონაცემები ასოციაციის მოდელისთვის ხშირად წამოღებულია ორ ცხრილდან. მაგალითად, ერთი ცხრილი შეიძლება შეიცავდეს მომხმარებლის ინფორმაციას, ხოლო მეორე მომხმარებლის შესყიდვებს. შესაძლებელია ეს მონაცემები მოდელში შევიტანოთ შექმნილი (მაკავშირებელი) ცხრილის გამოყენებით.

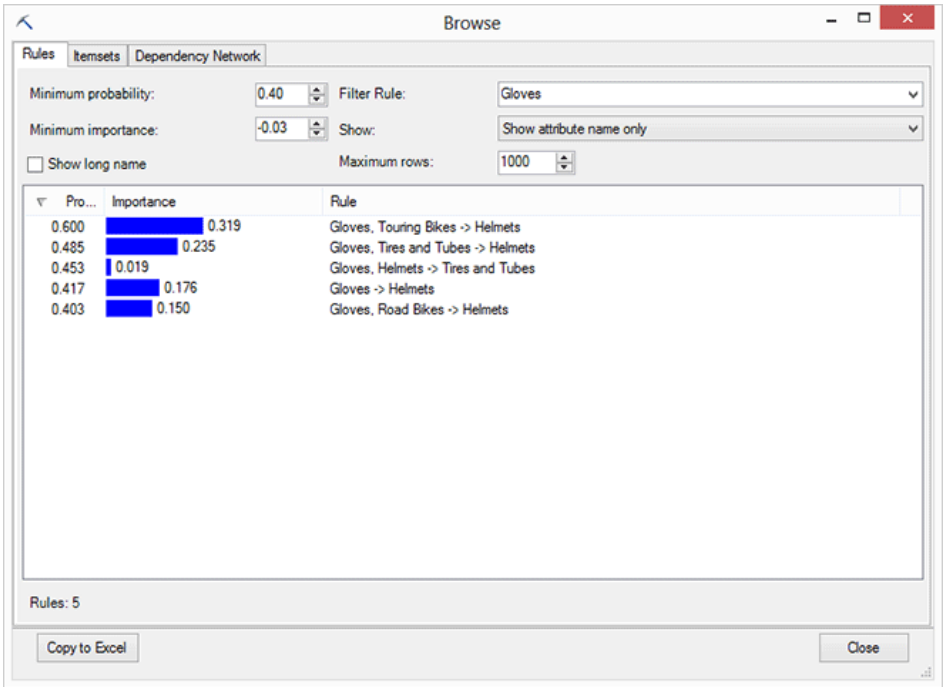
### ასოციაციის მოდელის ნახვა

მოდელის შესასწავლად შესაძლებელია გამოყენებული იქნას **Microsoft Association Viewer.** ასოციაციის მოდელის მიხედვით, ანალიზის სერვისები წარმოადგენენ კორელაციას სხვადასხვა კუთხით, ისე რომ უკეთესად იქნას გაგებული კავშირები და წესები, რომლებიც მონაცემებში მოიძებნა. ელემენტთა პანელი (**Itemset** pane) უზრუნველყოფს საერთო კომბინაციების დეტალების მიღებას, ხოლო წესების პანელი (Rules pane) წარმოგვიდგენს მონაცემების განზოგადების წესების ჩამონათვალს, ასევე, ითვლის ალბათობას და უსამაბებს წესებს იმის მიხედვით თუ რამდენად მნიშნველოვანია. შემდეგ Dependency Network პანელში შესაძლებელია ვიზუალურად ნახვა თუ როგორაა სხვადასხვა ნივთები აკავშირებული ერთმანეთთან. იხ. სურათი 1, სურათი 2, სურათი 3.

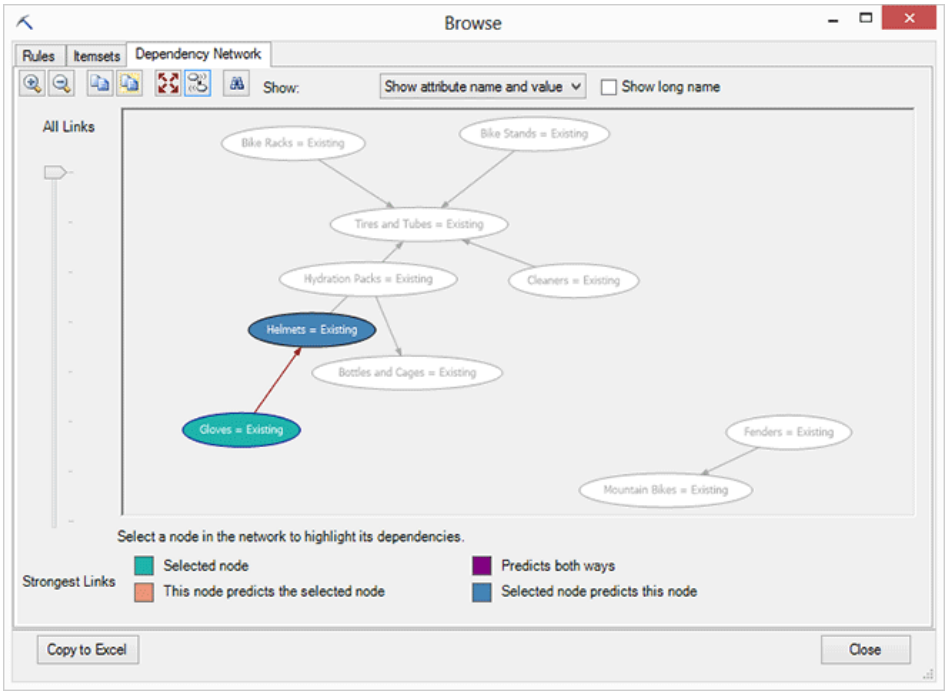
სურათი 1




სურათი 2




სურათი 3




**პროგნოზების გაკეთება**

მოდელის დამუშავების შემდეგ, შესაძლებელია წესების და ელემენტების გამოყენება პროგნოზების მისაღებად. ასოციაციის მოდელში, პრგონოზირება გულისხმობს სავარაუდოდ რა ნივთი იქნება წარმოდგენილი (მაგალითად, მომხმარებლის კალათაში) და პროგნოზირება, ასევე, შეიძლება მოიცავდეს ისეთ ინფორმაციას, როგორიცაა ალბათობა, მხარდაჭერა ან საგნის მნიშნელოვნება.

**ალგორითმის შესრულება**

მოდელის შექმნის პროცესი და კორელაციის დათვლა შეიძლება შრომატევადი იყოს. მიუხედავად იმისა, რომ ასოციაციის ალგორითმი იყენებს ოპტიმიზაციის ტექნიკას სივრცის შენახვისა და პროცესის სწრაფად მიღებისთვის, გათვალისწინებული უნდა იყოს შემდეგი პირობები, როგორიცაა:

- მონაცემთა ნაკრები დიდია მრავალი ინდივიდუალური ელემენტით.
- მინიმალური ელემენტიტის ზომა ძალიან დაბალია.

დამუშავების დროის მინიმუმამდე შემცირების და ელემენტების სირთულეების შემცირებსთვის, მონაცემების გაანგარიშებისას შესაძლებელია მოხდეს კატეგორიების მიხედვით დაჯგუფება სანამ გაანალიზდება ინფორმაცია.

# კლასტერული ალგორითმი

კლასტერული ალგორითმი არის იგივე სეგმენტაცია, რომელიც აჯგუფებს მონაცემთა ნაკრებში გამონაკლის შემთხვევებს კლასტერებად მსგავსი მახასიათებლების მიხედვით. ეს ჯგუფები გამოიყენება მონაცემების შესასწავლად, მონაცემებში არსებული ანომალიების დადგენასა და პროგნოზების შექმნისთვის.

კლასტერული მოდელები აიდენტიფიცირებენ მონაცემთა ბაზაში ისეთ კავშირებს, რომლებიც ლოგიკურად არ გამომდინარეობს ერთი შეხედვით. მაგალითად, ადვილი მისახვედრია, რომ ადამიანები რომლებიც სამსახურში მიდიან ველოსიპედით, არ უნდა ცხოვრობდნენ შორ მანძილზე სამსახურიდან. თუმცა, ალგორითმს შეუძლია იპოვოს ველოსპედით მოსარგებლე ადამიანებზე სხვა მახასითებლებიც, ისეთები რომლებიც აშკარა არაა.

კლასტერული ალგორითმი განსხვავდება სხვა ალგორითმებისგან, მაგალითად, როგორიცაა Decision Trees ალგორითმი, იმით რომ არ სჭირდება პროგნოზირებადი სვეტის დადგენა კლასტერული მოდელის შესაქმნელად.

### კლასტერული ალგორითმის მაგალითი

განვიხილოთ ადამიანების ჯგუფი, რომლებსაც გააჩნიათ ერთნაირი დემოგრაფიული ინფორმაცია და ვინც მსგავსს პროდუქტებს ყიდულობენ კომპანია, „The Adventure Works Cycle“ -დან. ხალხის ეს ჯგუფი წარმოადგენს მონაცემების კლასტერს. ასეთი კლასტერი შეიძლება რამდენიმე არსებობდეს მონაცემთა ბაზაში. კლასტერის შემაგდენელ სვეტებზე დაკვირვებით უფრო მკაფიოდ გამოჩნდება თუ როგორ არის ბაზაში მონაცემები ერთმანეთთან დაკავშირებული. კლასტერული ალგორითმის გამოყენების მაგალითებია, ასევე:

- მარკეტინგი: ეხმარება მარკეტინგის მომხმარებელს აღმოაჩინოს განსხვავებული ჯგუფები თავის მომხმარებლებში და შემდეგ ეს ცოდნა გამოიყენოს მიზნობრივი მარკეტინგის პროგრამების შემუშავებაში.
- მიწათსარგებლობა: დედამიწის სადამკვირვებლო მონაცემთა ბაზაში მსგავსი მიწის გამოყენების იდენტიფიცირება.
- სადაზღვევო: საავტომობილო სადაზღვევო პოლისის მფლობელების განსაზღვრა მაღალი საშუალო სადაზღვევო ღირებულებით.

- ქალაქის დაგეგმვა: სახლის ტიპის, ღირებულებისა და გეოგრაფიული ადგილმდებარეობის მიხედვით სახლების განსაზღვრა.

Clustering მრავალი დარგში გამოყენებული სტატისტიკური მონაცემების ანალიზის საერთო ტექნიკაა.

### როგორ მუშაობს ალგორითმი

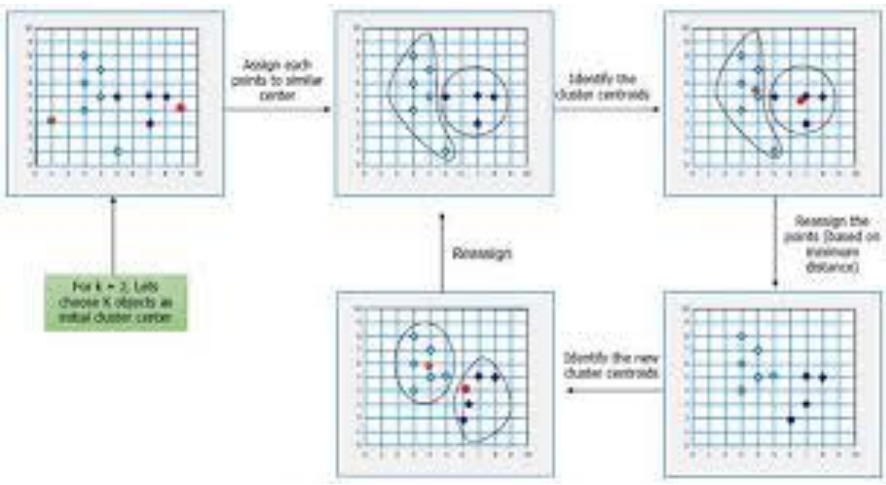
Clustering ალგორითმი პირველად განსაზღვრავს კავშირებს მონაცემთა ნაკრებში და შემდეგ ქმნის კლასტერთა სერიას, რომელიც ეფუძნება ამ ურთიერთობებს.

არსებობს რამდენიმე ცნობილი ალგორითმი, რომელთა გამოყენებით შესაძლებელი ხდება მონაცემთა ვიზუალიზაცია, დამუშავება და პრიორიტეტების განსაზღვრა. განვიხილოთ რამდენიმე ალგორითმი.

### მეთოდი „K-საშუალო“

კლასტერიზაციის ტიპის ალგორითმები მოიცავენ სხვხადასხვა ალგორითმებს, რომელთაგან ერთ-ერთია K-Means ალგორითმი. მოცემულია K კოეიფიცენტი და გარკვეულ წერტილთა სიმრავლე, რომელთაგან უნდა აიგოს კლასტერი. ალგორითმი შეიძლება იყოს წარმოდგენილი როგორც შემდეგი ბიჯების ერთობლიობა:

- დაიყოს ობიექტები k რაოდენობის არაცარიელ ქვესიმრავლეებად.
- დაყოფილი ქვესიმრავლლეების/კლასტერების ცენტროიდების პოვნა.
- თითოეული წერტილი/ობექტი მიეკუთვნოს კონკრეტულ კლასტერს.
- გამოთვლილ იქნას მანძილები თითოეული წერტილიდან და ამ კლასტერისთვის გამოყოფილი სხვა მრავალი წერტილიდან კლასტერამდე, სადაც მანძილი ცენტროიდამდე არის მინიმალური.
- წერტილების გადანაწილების შემდეგ ( წინა პუნქტის შედეგად ხდება გარკვეული წერტილის სხვა კლასტერზე მინიჭება, რადგან მის ცენტროიდთან უფრო ახლოს არის ეს წერტილი) ვიპოვთ ცენტროიდი ახალ კლასტერში.




სურათი 4. K-საშუალოს კლასტერიზაციის სიმულაცია

საბოლოოდ, ვიპოვით კლასტერებს, სადაც თითოეული მოცემული წერტილი განკუთვნილი იქნება კონკრეტული კლასტერისთვის. ეს წერტილები რა თქმა უნდა განსაზღვრული იქნება რაიმე სიდიდით ( მაგ. რიცხვითი სიდიდით), თუმცა უნდა მოხდეს ამ შედეგების გარკვეულ ფორმატში წარმოდგენა რათა შემდეგ ეს იფორმაცია გამოყენებულ იქნას როგორც სხვა ალგორითმის შემავალი ინფორმაცია ან/და საბოლოო შედეგების საჩვენებლად.

### EM-ალგორითმი

მონაცემთა მაინინგში გამოიყენება მოლოდინის მაქსიმიზაციის კლასტერული ალგორითმი (expectation-maximization - EM), რომლის დანიშნულებაცაა „ცოდნის“ აღმოჩენა.

მათემატიკურ სტატისტიკაში EM-ალგორითმი ითვლება იტერაციულად და გამოიყენება სტატისტიკური მოდელის მაქსიმალურად მსგავსი პარამეტრების გამოთვლისას, როდესაც ცვლადები დაფარულია. სტატისტიკური მოდელები არის ისეთი მოდელები, როდესაც აღწერილია უკვე ცნობილი, დადასტურებული მონაცემები, მაგალითად, გამოცდაზე მიღებული შეფასება შესაძლებელია წარმოვადგინოთ როგორც ნორმალური განაწილება, ამიტომ მოსალოდნელია, რომ შეფასებები დაგენერირდეს შესაბამისად ნორმალურ განაწილების მოდელში. განაწილება წარმოადგენს ყველა იმ ალბათურ შედეგს, რომელიც შეიძლება მიღებული იქნას გამოცდის შედეგად. მაგალითად, გამოცდაზე მიღებული შეფასება შეესაბამება ნორმალურ განაწილებას, სადაც არის ალბათობა იმისა, რომ წარმოდგენილი იქნება გამოცდაზე მიღებული ყველა სავარაუდო შედეგი. ე.ი. განაწილება გვეხმარება იმაში, გავიგოთ გამოცდაზე გასულმა რამდენმა ადამიანმა მიიღო ესა თუ ის შეფასება. მოდელის

მნიშვნელოვანი მახასიათებელია პარამეტრი, რომელიც აღწერს მოდელის ძირითად ნაწილს, განაწილებას. მაგალითად, საშუალო განაწილება აღიწერება საშუალო არითმეტიკულით და დისპერსიით. ზოგადად ნორმალური განაწილებისათვის აუცილებელია ორი პარამეტრის განსაზღვრა:

1. საშუალო არითმეტიკული;

2. დისპერსია.

იმ შემთხვევაში, როდესაც არ ვიცით ყველა შეფასების საშუალო არითმეტიკული ან დისპერია, შესაძლებელია შეფასდეს მხოლოდ ერთი მაგალითის მონაცემები.

EM-ალგორითმის ერთ-ერთი მნიშვნელოვანი მახასიათებელია, ხდომილება. ხდომილება – არის ალბათობა იმისა თუ რამდენად მოხდა გადახრა ნორმალური განაწილებიდან. ანუ ალბათობა იმისა, ნორმალური განაწილების მრუდი რამდენად სწორად აღწერს გამოცდაზე მიღებული შედეგების საშუალო არითმეტიკულს და დისპერსის.

მონაცემთა მაინინგსა და კლასტერიზაციაში მნიშვნელოვანია შეფასდეს კლასები მასში გაერთიანებული მონაცემების შესაბამისად, როგორც გამოტოვებული მონაცემები, რადგან თავიდან უცნობია კლასისი სახეობა, ამიტომ გამოტოვებული მონაცემების იტერაცია საკმაოდ მნიშვნელოვანი პროცესია კლასტერიზაციის ამოცანაში EM-ალგორითმის გამოყენებისას.

EM-ალგორითმი არის იტერაციული ალგორითმი, რომელიც გამოიყენება მაქსიმალურად თანხვედრი პარამეტრების მქონე სავარაუდო მოდელის შესაფასებლად. როდესაც მოდელში არის რამდენიმე არაცნობადი პარამეტრი, ფასდება მათი მაქსიმალური თანხვედრა სხვა მოდელთან, რის შედეგადაც EM - ალგორითმი ქმნის ახალ მოდელს. ახალ მონაცემთა ერთობლიობა აისახება როგორც კლასი. კლასტერიზაციის პროცესში კი სრულდება სამ ბიჯიანი იტერაციული პროცესი:

1. E - ბიჯი: აღნიშნულ ბიჯზე მოდელის ძირითადი პარამეტრების საფუძველზე გამოითვლება ალბათობა იმისა, ეკუთვნის თუ არა ეს მონაცემები მოცემულ კლასტერს;

2. M - ბიჯი: ხდება მოდელის პარამეტრების განახლება შესაბამის კლასტერულ განაწილებაში, რომელიც ჩატარდა E ბიჯზე;

3. პირველი ორი ბიჯი მეორდება მანამ, სანამ მოდელის პარამეტრები და კლასტერული განაწილება არ გათანაბრდება.

EM-ალგორითმის მთავარი ფასეულობაა გამოყენების სიმარტივე, ასევე მას შეუძლია მოახდინოს არა მარტო მოდელის პარამეტრების ოპტიმიზაცია, არამედ შეუძლია

განსაზღვროს რამდენად ღირებულია გამოტოვებული მონაცემები მოდელისათვის. EM - შესაძლებელია ჩავთვალოთ, როგორც საუკეთესო მეთოდი კლასტერიზაციისა და პარამეტრებზე დამოკიდებული მოდელის შექმნისათვის.

### მონაცემები საჭიროა კლასტერიზაციის მოდელებისათვის

კლასტერულ მოდელში გამოყენებული მონაცემების მომზადებისას უნდა გაანალიზდეს კონკრეტული ალგორითმის მოთხოვნები, მათ შორის, რამდენი მონაცემია საჭირო და როგორ გამოიყენება ეს მონაცემები.

კლასტერიზაციის მოდელის მოთხოვნები შემდეგია:

- ერთი გასაღები სვეტი - თითოეული მოდელი უნდა შეიცავდეს ერთ რიცხვით ან ტექსტურ სვეტს, რომელიც ცალსახად განსაზღვრავს თითოეულ ჩანაწერს. რთული გასაღებები არ არის დაშვებული.
- შემომავალი სვეტები - ყოველი მოდელი უნდა შეიცავდეს ერთ შემომავალ სვეტს მაინც, რომელიც შეიცავს მნიშვნელობას კლასტერის შესაქმნელად. შეყვანილი სვეტების რაოდენობა შეზღუდული არ არის, თუმცა დამოკიდებულია სვეტში მნიშვნელობების რაოდენობაზე, დამატებითი სვეტის შემოტანამ შესაძლებელია გაზარდოს მოდელის შემუშავების დრო.
- პირობითი პროგნოზირებადი სვეტი - ალგორითმს არ სჭირდება პროგნოზირებადი სვეტი მოდელის შესაქმნელად, მაგრამ შესაძლებელია დაემატოს ის თითქმის ყველა ტიპის მონაცემებში. პროგნოზირებადი სვეტი შეიძლება შემოღებულ იქნას როგორც შემომავალი ინფორმაცია კლასტერულ მოდელში, ან მიეთითოს რომ ის გამოიყენება მხოლოდ პროგნოზირებისთვის. მაგალითად, თუ ხდება სამომხმარებლო შემოსავლის პროგნოზირება კლასტერიზაციით დემოგრაფიის, როგორიცაა მაგალითად რეგიონის ან ასაკის მიხედვით, შემოსავალი განისაზღვრება როგორც PredictOnly(მხოლოდ პროგნოზისთვის) და ყველა სხვა სვეტი დაემატება, რეგიონი ან ასაკი და სხვა, როგორც შემომავალი ინფორმაცია.

## გადაწყვეტილებათა მიღების ხე

გადაწყვეტილებათა მიღების ხე (decision trees) მონაცემთა მაინინგში არის ერთ-ერთი ყველაზე განსაკუთრებული მიდგომა, სადაც ხის იერარქიულ სტრუქტურა კლასიფიცირებულია „თუ მაშინ“ (if-then) ლოგიკით. გადაწყვეტილების მიღებისას ხდება განსაზღვრა, თუ რომელ კლასს მიეკუთვნის რომელიმე ობიექტი ან არსებული სიტუაცია, ამავე დროს ხდება პასუხის გაცემა კითხვებზე, რომელიც ხის კვანძებზეა განთავსებული და ღებულობს დადებით ან უარყოფით პასუხს.

გადაწყვეტილებათა მიღებიის ხეების ალგორითმი კლასიფიკაციისა და რეგრესიის ალგორითმია, რომელიც გამოიყენება როგორც დისკრეტული, ასევე, უწყვეტი ატრიბუტების პროგნოზირების მოდელირებაში.

დისკრეტული ატრიბუტებისთვის, ალგორითმი ახდენს პროგნოზებს მონაცემთა ბაზაში შემომავალი ინფორმაციის კავშირების საფუძველზე. მაგალითად, გვაინტერესებს, რომელი მომხმარებელი შეიძენს ველოსიპედს, თუ ათი ახალგაზრდა მომხმარებლიდან ცხრა ყიდულობს ველოსიპედს, მაგრამ ათი ხანდაზმულიდან მხოლოდ ორი, ალგორითმი მიუთითებს, რომ ასაკი არის კარგი ატრიბუტი პროგნოზისთვის ველოსიპედის შეძენისას. გადაწყვეტილების ხე აკეთებს პროგნოზებს კონკრეტულ შედეგზე ამ ტენდენციის გათვალისწინებით.

უწყვეტი ატრიბუტებისთვის, ალგორითმი იყენებს ხაზოვანი რეგრესიას, რათა დადგინდეს, სად იყოფა გადაწყვეტილებათა ხე.

თუ ერთზე მეტი სვეტი განისაზღვრება როგორც პროგნოზირებადი, ან თუ შეყვანილი მონაცემები შეიცავს შექმნილ ცხრილს, რომელიც პროგნოზირებადია, ალგორითმი ადგენს ცალკეულ გადაწყვეტილებების ხეს თითოეული პროგნოზირებადი სვეტისათვის.

მარტივი მაგალითის საფუძველზე ეს პროცესი შეიძლება ასე ავხსნათ: მაგალითად, სატრანსპორტო გადაზიდვებში განსხვავებული ტვირთების შესახებ გვაქვს მონაცემთა ჯგუფები, რომელიც მოიცავს რამდენიმე ძირითად ატრიბუტს. ატრიბუტების საშუალებით ხდება ტვირთის აღწერა, მაგალითად, ლოკაციის ადგილი, ტრანსპორტირების საბოლოო ადგილი, ტვირთის სახეობა, ღირებულება, მოცულობა, წონა, ზომა, კონტეინერთან თავსებადობა და ა.შ. ჩვენ წინასწარ ვიცით განსხვავებული ტვირთების შესახებ ძირითადი მახასიათებელი პარამეტრები, რომელთაც ატრიბუტების სახით ვაერთიანებთ კლასებად, აღნიშნული ატრიბუტების საფუძველძე წინასწარ გვინდა განვსაზღვროთ - ტვირთი საშიშია

თუ არა. ტვირთი შესაძლებელია მოხვდეს ორი კლასიდან ერთში, საშიში ან უსაფრთხო. ალგორითმი შეტყონინებას უგზავნის თითოეული ტვირთის შესაბამის კლასს. ტვირთის განმსაზღვრელი ატრიბუტების და შესაბამისი კლასების საფუძველზე ალგორითმი აგებს გადაწყვეტილებათა ხეს, რომლის დახმარებით წინასწარ ატრიბუტების საფუძველზე მოხდება ტვირთის შესაბამისი კლასის განსაზღვრა.

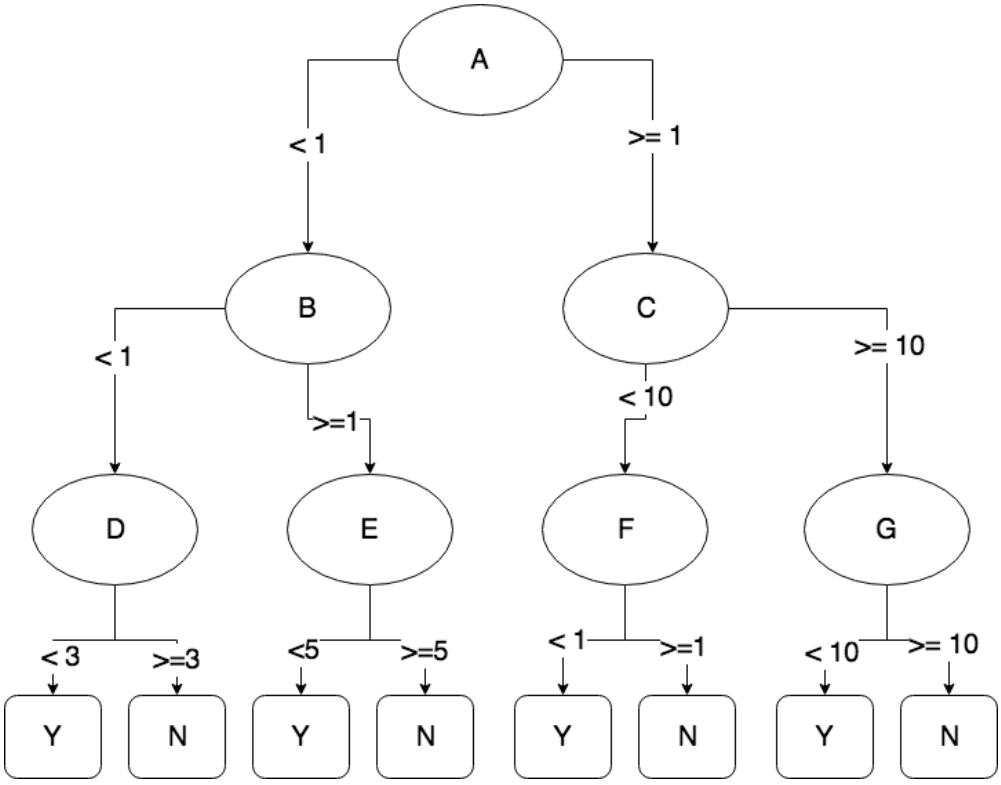
### მაგალითი

The Adventure Works Cycles კომპანიის მარკეტინგის განყოფილებას უნდა, რომ დააიდენტიფიციროს წინა მომხმარებელთა ის მახასიათებლები, რომლებიც შეიძლება მიუთითებდნენ შეიძენენ თუ არა ეს მომხმარებლები მომავალში პროდუქტს. AdventureWorks2012 მონაცემთა ბაზაში ინახება დემოგრაფიული ინფორმაცია, რომელიც აღწერს წინა მომხმარებლებს. ამ ინფორმაციას გაანალიზებით გადაწყვეტილების ხეების ალგორითმის გამოყენებით, მარკეტინგის დეპარტამენტს შეუძლია შექმნას მოდელი, რომელიც წინასწარ განსაზღვრავს შეიძენს თუ არა კონკრეტული მომხმარებელი პროდუქტებს, დემოგრაფიული ინფორმაციის ან წარსული შესყიდვების საფუძველზე.

### როგორ მუშაობს ალგორითმი

გადაწყვეტილება ხეების ალგორითმი აყალიბებს მონაცემთა მაინინგის მოდელს ხეების გაყოფით. ადგილს სადაც ხეები იყოფა ეწოდება კვანძები.

„A პარამეტრის მნიშვნელობა მეტია k -ზე“, თუ პასუხი დადებითია, მაშინ გადასასვლელი იხსნება ხის მარჯვენა კვანძის მომდევნო დონეზე, ხოლო უარყოფითი პასუხის შემთხვევაში მარცხენა მხარეს. შემდეგ ისევ დგება მომდევნო კვანძის შესაბამისი კითხვა და ა.შ. ხის სტრუქტურა საკმაოდ მარტივია და თვალსაჩინოა სურათი 5-ზე წარმოდგენილია გადაწყვეტილებათა მიღების ხის სტრუქტურის ფრაგმენტი.

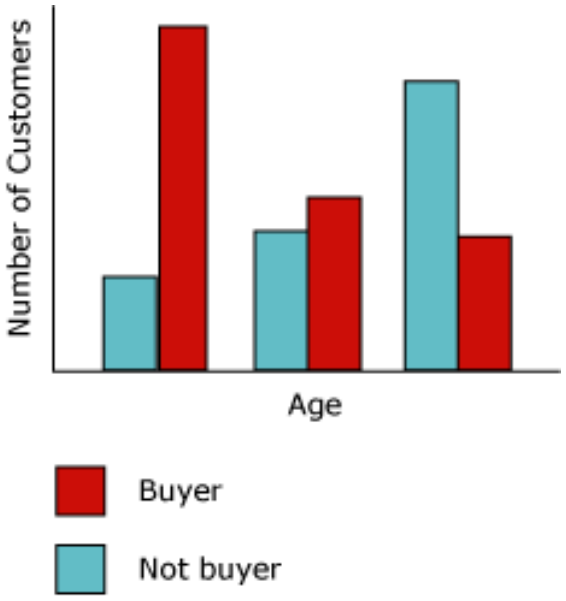



სურათი 5

გადაწყვეტილებათა ხე შესაძლებელია არ იყოს ერთადერთი საუკეთესო გამოსავალი მონაცემთა წარმოდგენისას, მაგრამ თანმიმდევრულად ხდება იმ ძირითადი მახასიათებლების წარმოდგენა და შეფასება, რომელიც უშუალოდ განსაზღვრავს მონაცემის შესაბამისობას ამა თუ იმ კლასის მახასიათებელ ძირითად პარამეტრებთან.

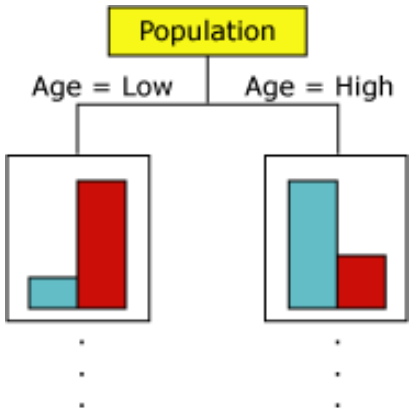
### დისკრეტული სვეტების პროგნოზირება

თუ როგორ ახდენს გადაწყვეტილების ხეების ალგორითმი დისკრეტული სვეტების პროგნოზირებას შესაძლებელია წარმოდგენილი იყოს ჰისტოგრამის სახით. შემდეგი დიაგრამა გვიჩვენებს თუ რა დამოკიდებულებაშია ასაკი ველოსიპედების მყიდველთა რაოდენობასთან. ჰისტოგრამა გვიჩვენებს რომ ადამიანის ასაკით შეგვიძლია განვსაზღვროთ შეიძენს თუ არა ის ველოსიპედს.იხ. სურათი 6.




სურათი 6.

კორელაციით, რომელიც ნაჩვენებია დიაგრამაზე, გადაწყვეტილების ხეების ალგორითმი შექმნის ახალ კვანძს მოდელში. იხ.სურათი 7.

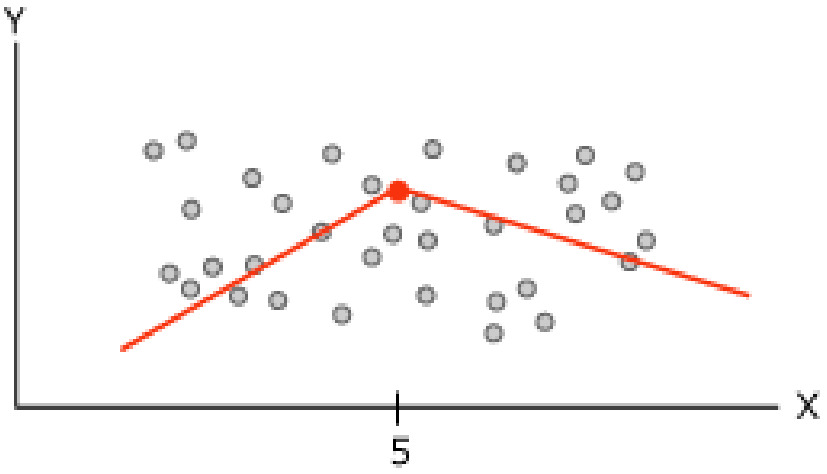



სურათი 7

ალგორითმის მიერ დაამატებული ახალი კვანძები მოდელში შექმნის ხის სტრუქტურას. თუ მოდელი გაიზრდება, ალგორითმი, ასევე, განიხილავს ყველა სვეტს.

### პროგნოზირების უწყვეტი სვეტები

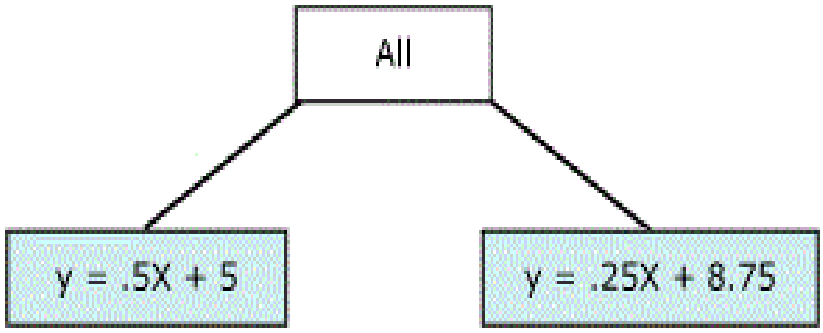
როდესაც გადაწყვეტილებათა ხეების ალგორითმი აშენებს ხეს, რომელიც ეფუძნება უწყვეტ პროგნოზირებად სვეტს, თითოეული კვანძი შეიცავს რეგრესიულ ფორმულას. ცვლილება ხდება რეგრესიულ ფორმულაში არაწრფივობის დროს. მაგალითად, განვიხილოთ შემდეგი დიაგრამა.იხ. სურათი 8.




სურათი 8

სტანდარტულ რეგრესიულ მოდელში, ცდილობენ მიიღონ მხოლოდ ერთი ფორმულა, რომელიც წარმოადგენს ტენდენციებს და კავშირებს მთლიან მონაცემთა ბაზაში. თუმცა, ერთმა ფორმულამ კომპექსურ მონაცემებში შესაძლებელია სამუშაო ცუდად შეასრულოს. ამიტომ, გადაწყვეტილებათა ხეების ალგორითმი უყურებს ხის ცალკეულ სეგმენტებს, რომლებიც ძირითადად ხაზოვანია და ქმნის ცალკეულ ფორმულებს ამ სეგმენტებისთვის. მონაცემთა დაყოფით განსხვავებულ სეგმენტებად, ალგორითმმა შესაძლებელია უკეთესად იმუშაობს.

ქვემოთ მოცემული ხის დიაგრამა (იხ. სურათი 9) წარმოადგენს სურათი 8-ზე მოცემული წერტილოვანი დიაგრამას ხის სახით. შედეგის პროგნოზირებისთვის მოდელი წარმოადგენს ორ განსხვავებულ ფორმულას: მარცხენა მხარეს მოცემულია ფორმულით y = .5x x 5 და მარჯვენა მხარეს მოცემულია y = .25x + 8.75. წერტილი სადაც ორი წრფე გაერთიანდა წერტილოვან დიაგრამაზე არის არაწრფივობის წერტილი და აღნიშნავს ადგილს, სადაც გაჩნდება კვანძი ხის მოდელში.




სურათი 9.

ეს არის მარტივი მოდელი მხოლოდ ორი ხაზოვანი განტოლებით. აქედან გამომდინარე, ხის მოდელში როცა ჩნდება კვანძი ხეც იყოფა ორ ტოტად. ხის გაყოფა ხდება ნებისმიერ დონეზე. ეს ნიშნავს იმას, რომ ხეში მრავალი დონე და კვანძია, სადაც თითოეული კვანძი დახასიათებულია ატრიბუტების სხვადასხვა კოლექციით, ფორმულა შეიძლება გავრცელდეს ბევრი კვანძებისვთის ან ვრცელდება მხოლოდ ერთი კვანძისთვის.

**მონაცემები საჭიროა გადაწყვეტილებათა ხეების მოდელებისათვის**

გადაწყვეტილებათა ხეების მოდელში გამოყენებული მონაცემების მომზადებისას უნდა გაანალიზდეს კონკრეტული ალგორითმის მოთხოვნები, მათ შორის, რამდენი მონაცემია საჭირო და როგორ გამოიყენება ეს მონაცემები.

გადაწყვეტილებათა ხეების მოდელის მოთხოვნები შემდეგია:

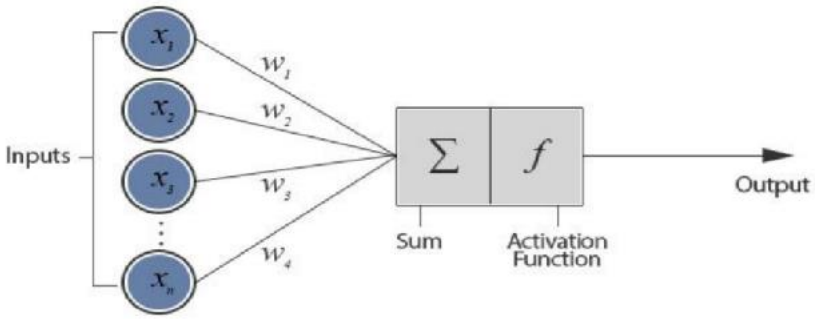
- ერთი გასაღები სვეტი - თითოეული მოდელი უნდა შეიცავდეს ერთ რიცხვით ან ტექსტურ სვეტს, რომელიც ცალსახად განსაზღვრავს თითოეულ ჩანაწერს. რთული გასაღებები არ არის დაშვებული.
- შემომავალი სვეტები - ყოველი მოდელი უნდა შეიცავდეს ერთ შემომავალ სვეტს მაინც, რომლის მნიშვნელობა შეიძლება იყოს დისკრეტული ან უწყვეტი. შეყვანილი ატრიბუტების რაოდენობა შეზღუდული არ არის, თუმცა დამატებითი ატრიბუტის შემოტანამ შესაძლებელია გაზარდოს მოდელის მუშაობის დრო.
- პროგნოზირებადი სვეტი - მოითხოვს ერთ პროგნოზირებად სვეტს მაინც. შესაძლებელია შეიცავდეს მრავალ პროგნოზირებად ატრიბუტს, და

თითოეულის ტიპი შეიძლება იყოს განსხვავებული, რიცხვითი ან დისკრეტული, თუმცა პროგნოზირებადი ატრიბუტების რაოდენობის ზრდა გაზრდის ალგორითმის მუშაობის დროს.

# ნეირონული ქსელის ალგორითმი

კომპიუტერები მოხერხებულია ისეთი პრობლემების გადასაწყვეტად, რომლებიც მათემატიკური ალგორითმებით წარმოიდგინებიან. მაგრამ არსებობენ პრობლემები, რომელთა წარმოდგენა მათემატიკური ალგორითმებით ძნელია, როგორიცაა სახეთა გამოცნობა და ბუნებრივი ენების კომპიუტერული მოდელირება. თუმცა, ამ პრობლემებს ადამიანის ტვინი ადვილად წყვეტს. ხელოვნური ნეირონული ქსელები რეალიზებულია კომპიუტერებზე ისე, რომ ისინი ინფორმაციას ამუშავებენ ადამიანის ტვინის ანალოგიურად, რაც საშულებას იძლევა გადავწყვიტოთ სახეთა გამოცნობისა და სხვა ანალოგიური პრობლემები კომპიუტერის საშუალებით ადამიანის ტვინის მსგავსად.

ხელოვნური ნეირონული მოდელები არიან ბიოლოგიურ ნეირონებზე დაფუძნებული გამარტივებული მოდელები. ისინი ახორციელებენ ბიოლოგიური ნეირონების არსებით ფუნქციებს. ჩვენ ვუწოდებთ ასეთ ხელოვნურ ნეირონებს პერსეპტრონებს (perceptrons ). ახლა, ვნახოთ როგორ გამოიყურება ასეთი პერსეპტრონი:




როგორც ნაჩვენებია დიაგრამაზე, პერსეპტრონს აქვს მრავალი შესასვლელი და თითოეულ შესასვლელს აქვს თავისი წონა. წონები ამცირებენ ან ზრდიან შესასვლელი სიგნალის მნიშვნელობას. მაგალითად, თუ შესასვლელის მნიშვნელობა არის ერთი და წონა 0.2., მაშინ მნიშვნელობა იქნება 0.2. ეს შეწონილი შესასვლელები შემდეგ იკრიბება და გადაეცემა აქტივაციის ფუნქციას. იგი გამოიყენება, რომ გადააქციოს შესასვლელი უფრო სასარგებლო გამოსასვლელად. არსებობენ სხვადასხვა ტიპის აქტივაციის ფუნქციები, მაგრამ ყველაზე მარტივია ნაბიჯი ფუნქცია. ამ ფუნქციას გამოაქვს 1, თუ შესასვლელი მეტია ზღვარზე და 0 წინააღმდეგ შემთხვევაში.

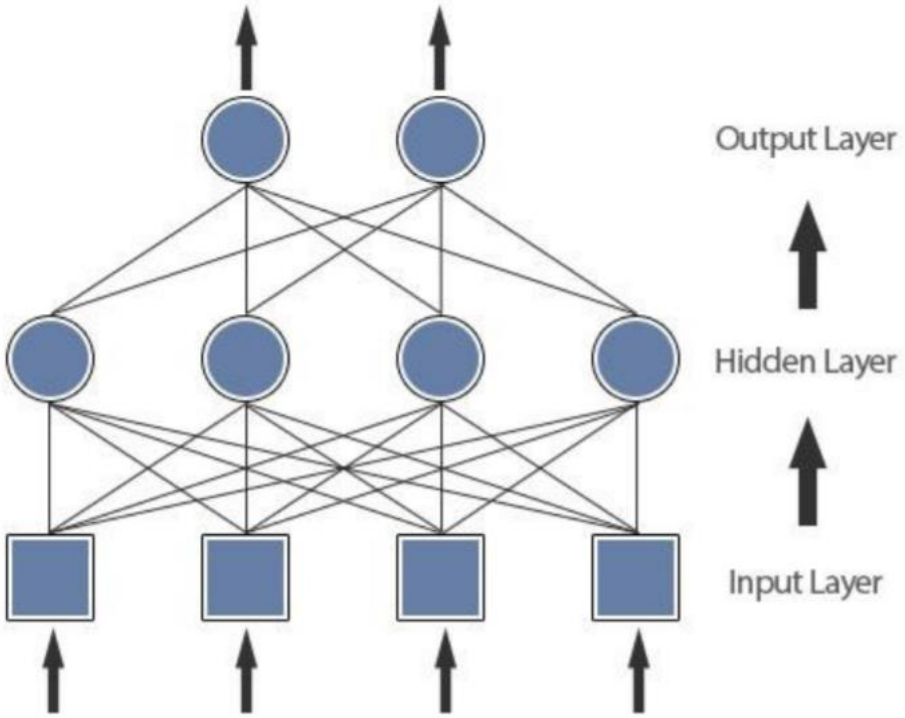
ნეირონული ქსელი არის ალგორითმების ჯაჭვი / სერია, რომლის მიზანია ამოიცნოს ინფორმაცია ჩვენთვის მოწოდებული ცნობილი მონაცემების მიხედვით, რომელიც ადამიანის ტვინის მუშაობის და ანალიზის იმიტაციას ახდენს. ნეირონულ ქსელებს შეუძლიათ ადაპტირება შეტანილი ინფორმაციის შეცვლით, ამგვარად ქსელის გენერირება საუკეთესო შედეგი იქნება. ეს მექანიზმი გადააკეთებს გამომავალი ინფორმაციის კრიტერიუმებს. ეს ტექნოლოგია სწრაფად ხდება პოპულარული სავაჭრო სისტემების სფეროში, სამედიცინო ოპერაციებში, ხელოვნურ ინტელექტში, სიგნალის მუშაობაში, ნიმუშის აღიარებაში და სხვა.

### მაგალითი

ნერვული ქსელის ალგორითმი სასარგებლოა საანგარიშო მონაცემების ანალიზისთვის, როგორიცაა წარმოება ან კომერციული პროცესი. დღეს არსებულ ყველაზე რთულ ქსელებს შეუძლიათ წარმატებით გაუმკლავდნენ ისეთ ამოცანებს როგორიცაა ფუნქციის აპროქსიმაცია, რეგრესიული ანალიზი, რიცხვით მიმდევრობაში კანონზომიერების აღმოჩენა, მონაცემთა კლასიფიკაცია, მონაცემთა ფილტრაცია, კლასტერიზაცია, მონაცემთა კომპრესია. შესაბამისად, საკმაოდ ფართოა ნეირონული ქსელების პრაქტიკაში გამოყენების არეალიც: სამედიცინო დიაგნოსტიკა, სამხედრო აღჭურვილობა, სახეთა ამოცნობა, დიდი მოცულობის ინფორმაციის კატეგორიზაცია, კრედიტების გაცემის მიზანშეწონილობის შეფასება, ბირჟების პროგნოზირება, ტექსტების დამუშავება, რობოტოტექნიკა.

### ხელოვნური ნეირონული ქსელების რეალიზაცია

ახლა ვნახოთ რა არის ხელოვნური ნეირონული ქსელი და როგორ ამუშავებს იგი ინფორმაციას. ჩვენ ვაპირებთ გავეცნოთ feedforward ქსელებს და როგორ არიან ისინი დაკავშირებული პერსეპტრონებთან. მაგრამ მანამდე, გავეცნოთ რა არის feedforward ქსელი.




როგორც დიაგრამიდან ჩანს, იგი შედგება 3 შრისაგან: შესასვლელი შრე; ფარული შრე და გამოსასვლელი შრე. შესასვლელი შრის ყოველი კვანძიდან ინფორმაცია გადაეცემა ფარული შრის ყოველ კვანძს და იქიდან გამოსასვლელი შრის ყოველ კვანძს. შეიძლება არსებობდეს მრავალი ფარული შრეები და შეიძლება, ასევე, სიგნალები მოძრაობდნენ ორივე მიმართულებით. ჩვენს შემთხვევაში სიგნალები მოძრაობენ ერთი მიმართულებით შესასვლელიდან გამოსასვლელის მიმართულებით, ამიტომაც მას ეწოდება feedforvard(წინკვება) ქსელი. ქსელებს, როცა სიგნალები მოძრაობენ ორივე მიმართულებით ეწოდებათ feedback(უკუკვება) ქსელები.

უმეტეს შემთხვევებში, გვჭირდება ერთი ან ორი შრე, მაგრამ საჭიროა ექსპერიმენტები, რომ ვიპოვოთ კვანძების ოპტიმალური რაოდენობა ყოველ კონკრეტულ შემთხვევაში.

ნეირონულ ქსელებს შეუძლიათ ამოიღონ და აჩვენონ ისეთი თვისებები, რომლებიც მიიღება სხვა ალგორითმებით დაჯგუფებისა და კლასიფიკაციისთვის. ასე რომ ნეირონული ქსელები შეიძლება განვიხილოთ, როგორც უფრო დიდი მანქანების შესწავლის აპლიკაციები, რომლებსაც აქვთ ალგორითმები გაძლიერებული სწავლისთვის, კლასიფიკაციისთვის და რეგრესიისთვის. ნეირონული ქსელების ყველაზე მნიშვნელოვანი თვისება არის ის, რომ ისინი ადაპტირებულია, რაც იმას ნიშნავს, რომ მათ შეუძლიათ შეცვლა ან ადაპტაცია.

## PageRank ალგორითმი

PageRank – არის ალგორითმი ე.წ. რანჟირებული ბმა. იგი განსაზღვრავს ობიექტის მნიშვნელობას და კავშირს ქსელში არსებულ სხვა ობიექტებთან. აღნიშნულ ალგორითმს ეფექტურად იყენებენ საძიებო სისტემები, მაგალითად, Google, Wikipeida და ა.შ. რანჟირებული ბმა არის ქსელური ანალიზის ტიპი, რომელიც განსაზღვრავს ობიექტებს შორის (წაიკითხე, დააკავშირე) ასოციაციურ კავშირს.

მეთოდი PageRank ვებ-გვერდებს ანიჭებს 0-დან 10-მდე პრიორიტეტს. შემდეგ ცხრილში მოცემულია კომპანია Google-ს მიერ გამოქვეყნებული რეიტინგი.

| Website | PageRank |
|---|---|
| twitter.com | 10 |
| facebook.com | 9 |
| reddit.com | 8 |
| stackoverflow.com | 7 |
| tumblr.com | 6 |
| crucial.com | 5 |
| programmingzen.com | 4 |
| dearblogger.org | 3 |

ცხრილში მაქსიმალური მნიშვნელობა (10 ქულა) ენიჭება ვებგვერდს, რომელიც ყველაზე პოპულარული და რელევანტურია. ალგორითმი PageRank სპეციალურადაა შექმნილი გლობალური ქსელისათვის, რომელსაც ადამიანების დამოკიდებულება გადაყავს ციფრებში. PageRank ალგორითმი არის სუპერეფექტური საშუალება, რომელიც ახდენს ბმების რანჟირებას. გასათვალისწინებელია ის გარემოება, რომ დასაკავშირებელი ობიექტები არაა აუცილებელი იყოს ვებ-გვერდები. PageRank ალგორითმის ინოვაციური გამოყენების სფეროებია:

1. ეკოლოგია, სადაც აღნიშნული ალგორითმის გამოყენებით განისაზღვრება ეკოსისტემების სასიცოცხლო ციკლი;

2. ტვიტერმა PageRank ალგორითმის გამოყენებით შეიმუშავა WTF (Who-to-Follow) – პერსონალიზებული სარეკომენდაციო ვარიანტები, სადაც ჩამონათვალში წარმოდგენილია იმ ადამიანთა სია, რომელთაც სჭირდებათ სხვადასხვა სახის შეთავაზებები და რეკომენდაციები;

3. PageRank ალგორითმი აქტიურად გამოყენება ჰონკონგის პოლიტექნიკური უნივერსიტეტის პროფესორის ბინ ჟენის (Bin Jiang) მიერ ტოპოლოგიურ ჩანაწერებში, სადაც წინასწარ ხდება ადამიანების აქტიური მოძრაობის განსაზღვრა. PageRank ალგორითმის მთავარი ღირებულება არის საიმედოობა, მიუხედავად იმ სირთულისა, რომელიც უკავშირდება რელევანტური ბმის პროცესს. გრაფიკული ან სქემური მონაცემების შესაბამისი პარამეტრების, პრიორიტეტებისა და რელევანტურობის განსაზღვრისათვის ასევე ყველაზე ეფექტური საშუალებაა PageRank -ის გამოყენება. საფირმო ნიშანი PageRank ეკუთვნის Google კომპანიას. ალგორითმი PageRank შეიქმნა და დაპატენტდა სტენფორდის უნივერსიტეტში. PageRank ალგორითმი ასევე რეალიზებულია შემდეგ პროგრამულ პაკეტებში:

1. C++ OpenSource PageRank;

2. Python PageRank;

3. ქსელური ანალიზის პაკეტი - igraph

ზოგადად შესაძლებელია ითქვას, რომ მონაცემთა მაინინგი არის მულტიდისციპლინარულ დარგი, სადაც მონაცემთა ბაზების შეფასება ხდება გამოყენებითი სტატისტიკის მეშვეობით, ხოლო ამოცნობის თვალსაზრისით გამოიყენება ხელოვნური ინტელექტის მეთოდები, მონაცემთა ბაზების თეორია და ა.შ.

# პრაქტიკული მაგალითი

The Adventure Works Cycles კომპანიის მარკეტინგის განყოფილებას სურს, რომ დააიდენტიფიციროს წინა მომხმარებელთა ის მახასიათებლები, რომლებიც შეიძლება მიუთითებდნენ შეიძენენ თუ არა ეს მომხმარებლები მომავალში პროდუქტს. ჩვენს შემთხვევაში, გვაინტერესებს მომხმარებელი შეიძენს თუ არა ველოსიპედს და ამ ფაქტის პროგნოზირებას მოვახდენთ გადაწყვეტილებათა მიღების ხეების ალგორითმის მიხედვით კლიენტის ასაკზე დაყრდნობით.

AdventureWorks2012 მონაცემთა ბაზაში ინახება დემოგრაფიული ინფორმაცია, რომელიც აღწერს წინა მომხმარებლებს, მათ სქესს, შემოსავალს, ოჯახურ მდგომარეობას, რეგიონს, ასაკს, საკუთრებას, განათლებას და ა.შ.

მაგალითი ბაზიდან:

| ID | Marital Status | Gender | Yearly Income | Children | Education | Occupation | Home Owner | Cars | Commute Distance | Region | Age | BikeBuyer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22711 | Single | Male | 30000 | 0 | Partial College | Clerical | No | 1 | 0-1 Miles | Europe | 33 | Yes |
| 13555 | Married | Female | 40000 | 0 | Graduate Degree | Clerical | Yes | 0 | 0-1 Miles | Europe | 37 | Yes |
| 28907 | Married | Male | 160000 | 5 | Partial College | Professional | No | 3 | 10+ Miles | Europe | 55 | No |
| 2 | Single | Male | 160000 | 0 | Graduate Degree | Management | Yes | 2 | 0-1 Miles | Pacific | 47 | No |
| 25410 | Single | Female | 70000 | 2 | Bachelors | Skilled Manual | No | 1 | 0-1 Miles | North America | 38 | Yes |

მონაცემები ამოღებულია შემდეგი სელექთით:

```
Select
    id,
    martial_status,
    gender,
    yearly_income,
    children,
    education,
    occupation,
    home_owner,
```

cars,

commute_distance,

region,

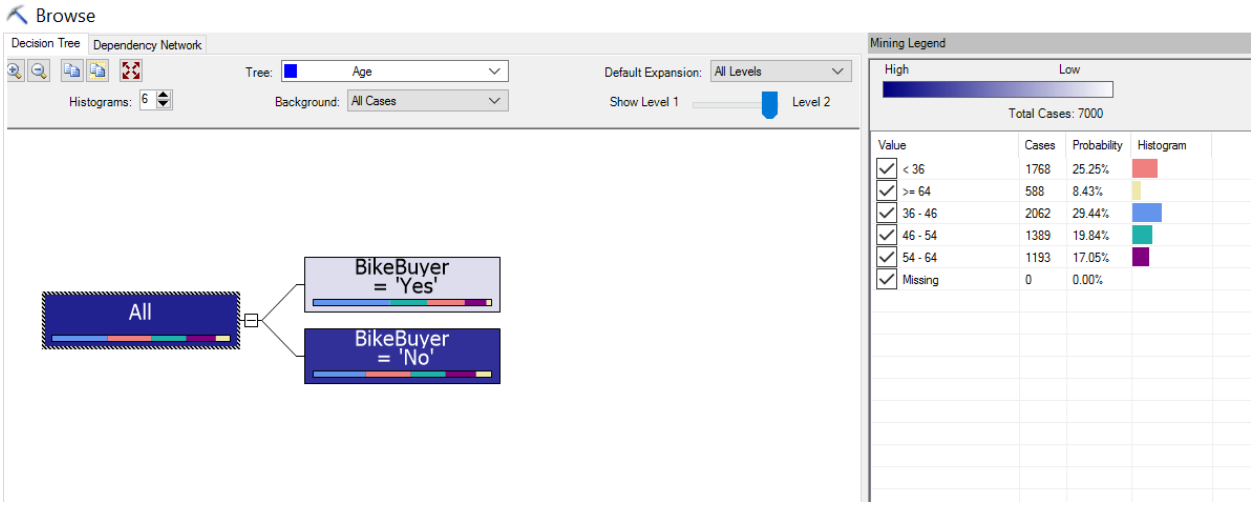
age,

bike_buyer

from j_customers

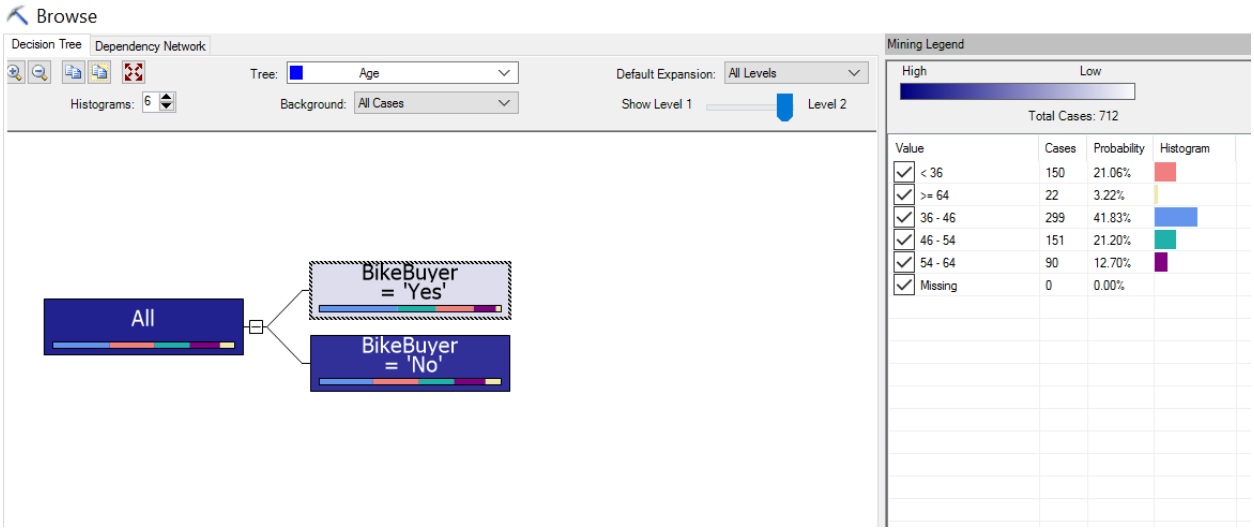
ჩვენ გამოვიყენებთ გადაწყვეტილებათა მიღებიის ხეების ალგორითმს. ვნახავთ, თუ როგორ ახდენს გადაწყვეტილების ხეების ალგორითმი დისკრეტული სვეტების პროგნოზირებას. შემდეგი დიაგრამა გვიჩვენებს თუ რა დამოკიდებულებაშია ასაკი ველოსიპედების მყიდველთა რაოდენობასთან. ჰისტოგრამა გვიჩვენებს რომ ადამიანის ასაკით შეგვიძლია განვსაზღვროთ შეიძენს თუ არა ის ველოსიპედს.

Decision tree:

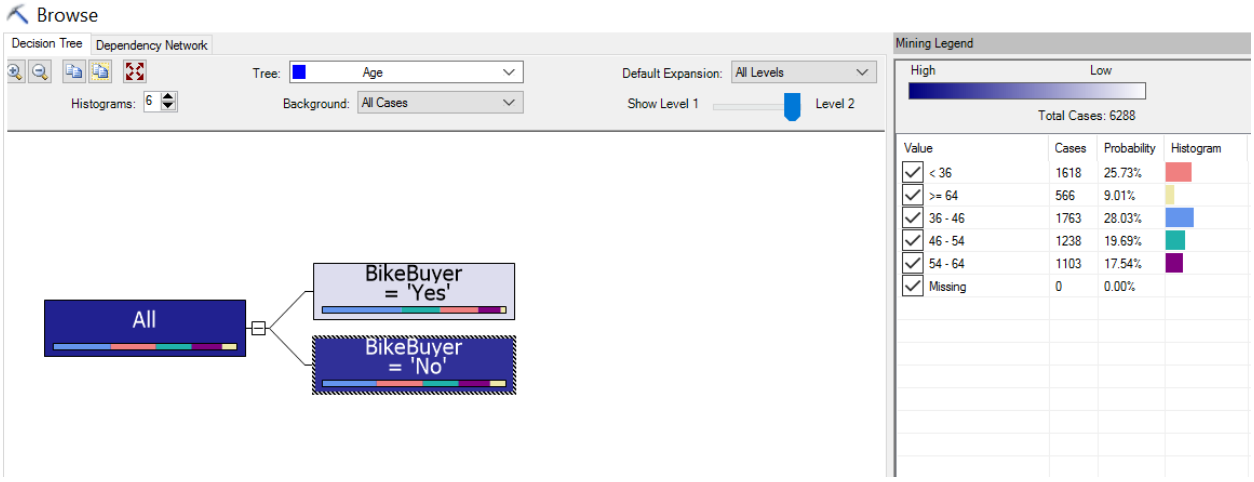



თითოეული node-ის აღწერა,

Node – Bikebuyer=”Yes”




Node - Bikebuyer=”No”




საბოლოოდ დისკრეტული ატრიბუტისთვის, ალგორითმის საშუალებით მოვახდინეთ პროგნოზირება ველოსიპედის შეძენაზე მონაცემთა ბაზაში არსებულ მომხმარებელთა ასაკის მიხედვით.

# დასკვნა

სამაგისტრო ნაშრომიში მონაცემთა მოძიების მეთოდების და ალგორითმების განხილვის შედეგად გამოიკვეთა, რომ არ არსებობს ინფორმაციის ძებნის უნივერსალური ალგორითმი, რომელიც ერთნაირად ეფექტური იქნება ვებ- სივრცეში, სხვადასხვა მონაცემთა საცავებში და ასევე მონაცემთა ბაზებში. განვიხლეთ კონკრეტული ალგორითმები, რომლის საშუალებითაც შეიძლება მოხდეს მონაცემთა მოძიება, რაც ანალიზის სერვისების შემადგენელი ნაწილია.

მოვახდინეთ სხვადასხვა ალგორითმების რეალიზაცია კონკრეტულ მაგალითებზე სრულიად შესაძლებელია არ დავეყრდნოთ განხილულ ალგორითმებს, მათ კომბინაციას და მოვახდინოთ კიდევ სხვა ალგორითმების გამოყენება.

მონაცემთა ინტელექტუალური ანალიზის მეთოდების თეორიულად განხილვის დასრულების შემდეგ, ჩვენ უნდა დაგვენახა როგორ გამოიყენება თითოეული მათგანი პრაქტიკულ ექსპერიმენტებში.

რამდენადაც ჩვენ გავაკეთეთ შერჩეული ალგორითმების ანალიზი, უნდა შეგვერჩია საუკეთესო ტექნიკა ჩვენი მოდელის ასაგებად და მონაცემების შესამოწმებლად. ჩვენ შევარჩიეთ გადაწყვეტებათა მიღების ხეების ალგორითმები, ჩვენი მოდელის ასაგებად. გამოვიყენეთ კოდის ფრაგმენტები იმისათვის, რათა უკეთ გაგვეგო რა სახის მონაცემები გვქონდა ალგორითმებისთვის და გამოვიყენეთ გრაფიკის რამდენიმე უბანი შედეგის მომზადებისა და ვიზუალიზაციისათვის.

კვლევის დასასრულს ჩვენ შევარჩიეთ საუკეთესო ალგორითმი, მაგრამ ეს არ ნიშნავს იმას, რომ სხვა ალგორითმები არ გამოდგება. მონაცემთა ინტელექტუალური ანალიზი და მანქანური სწავლება არის სიღრმისეული მეცნიერება და არ არსებობს საუკეთესო მოდელი ან მეთოდები ამ კონკრეტული ამოცანის ამოსახსნელად. ყველაფერი იმაზეა დამოკიდებული, თუ რა ამოცანას ასრულებთ, მონაცემთა ინტელექტუალური ანალიზის კონკრეტული ალგორითმის შესრულებაზე და როგორია თქვენი მონაცემთა ნაკრები.

ვინაიდან მხოლოდ ერთი ალგორთმით არ შეიძლება შემოვიფარგლოთ კონკრეტული საკითხის გადასაჭრელად და უმჯობესია რამდენიმე ალგორითმი გამოვიყენოთ და შემდეგ დავადგინოთ საუკეთესო მათ შორის, ან მათი კომბინაცია განვსაზღვროთ, სწორედ, აღნიშნული საკითხი წარმოადგენს სამომავლო კვლევის საგანს, რათა დადგინდეს რომელი ალგორითმები ან/და კომბინაციები არის ყველზე ეფექტური ამ და სხვა ამოცანებში.

# გამოყენებული ლიტერატურა


1. https://en.wikipedia.org/wiki/Data_mining
2. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017
3. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017
4. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-association-algorithm?view=sql-server-2017
5. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-clustering-algorithm?view=sql-server-2017
6. https://www.edureka.co/blog/introduction-to-clustering-in-mahout/
7. http://gtu.ge/book/PetriSurgu_DataManagmTechn.pdf
8. https://en.wikipedia.org/wiki/C4.5_algorithm
9. https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2017
10. https://ge.itstep.org/news-slider/%E1%83%9C%E1%83%94%E1%83%A0%E1%83%95%E1%83%A3%E1%83%9A%E1%83%98-%E1%83%A5%E1%83%A1%E1%83%94%E1%83%9A%E1%83%98%E1%83%A1-%E1%83%A8%E1%83%94%E1%83%A1%E1%83%90%E1%83%95%E1%83%90%E1%83%9A%E1%83%98/
11. http://amigo.ge/sangu/neural-networks.pdf
12. https://en.wikipedia.org/wiki/PageRank