მანქანური დასწავლის პროგრამული ტექნოლოგიების გამოყენებითი ანალიზი

ეკატერინე თურქია [1], ვალერიანე გელოვანი [1], მიხეილ გავაშელი [2]
1 - საქართველოს ტექნიკური უნივერსიტეტი
2 - ბოსტონის ჩრდილო-აღმოსავლური უნივერსიტეტი, აშშ

რეზიუმე

მანქანური დასწავლის ტექნოლოგია დღეს ერთ-ერთი პროგრესული და მაღალ-მოთხოვნადი მიმართულებაა. ამ ტექნოლოგიით ინტერესდება, ფაქტობრივად, ყველა ის დარგი, სადაც ანალიტიკური შედეგები მნიშველოვანი და ღირებულია, ხოლო გადაწყვეტილების მისაღებად საჭიროა დიდი რაოდენობის მონაცემების გაანალიზება. წინამდებარე სტატიაში განხილულია მიმდინარე პერიოდში აქტუალური მანქანური დასწავლის პროგრამული ენები და პროდუქტები: Python, R, Microsoft Azure Machine Learning Studio. პრაქტიკული მაგალითების სახით ნაჩვენებია ჩამოთვლილ საინფორმაციო სისტემებში მანქანური დასწავლის ალგორითმების მუშაობის პრინციპები, მაგალითები და შედეგები.

საკვანძო სიტყვები: მანქანური დასწავლა. მანქანური დასწავლის ენა. Python. R. Ms Azure Machine Learning Studio.

1. შესავალი

მანქანური დასწავლის ტექნოლოგიაში მოხდა სტატისტიკის, პროგრამული სისტემებისა და ხელოვნური ინტელექტის მეთოდების ინტეგრაცია. ამ მიდგომით უზრუნველყოფილ იქნა რთული ანალიტიკური ამოცანების რეალიზაცია, რაც დიდი მონაცემების ბაზაზე, კომბინირებული სტატისტიკური ალგორითმების მანიპულაციით ავტომატიზებულ დინამიკურ გადაწყვეტილებებს ახორციელებს.

მანქანური დასწავლის ალგორითმების ნუსხაში შედის კლასტერიზაციის, ასოციაციური ანალიზის, რეგრესიის, გადაწყვეტილებათა მიღების ხის, შემთხვევითი ტყის, კლასიფიკაციის მოდელების ჯგუფები, რომლებიც ზედამხედველურ და არაზედამ-ხედველურ ტიპებში ერთიანდება. ზედამხედველური ტიპი დაფუძნებულია ე.წ. „შაბლონის“ პრინციპზე, ახორციელებს მონაცემთა სიმრავლიდან აღებულ ე.წ. „სასწავლო მონაცემებზე“ მათემატიკური მოდელის შემუშავებას და შერჩეული მოდელის გავრცელებას მონაცემთა სრულ სიმრავლეზე; არაზედამხედველური ტიპი ორიენტირე-ბულია მონაცემების თვითკლასიფიკაციასა და გადახრების აღმოჩენაზე [1].

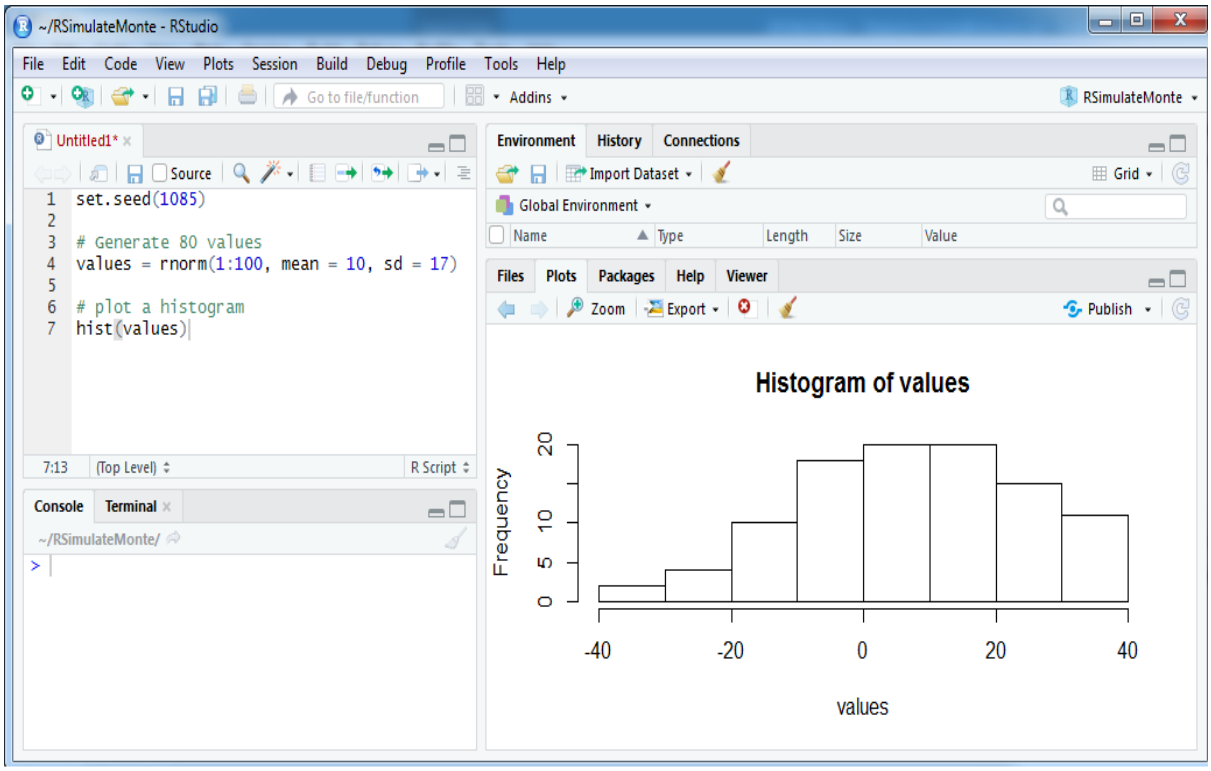
მანქანური დასწავლის მხარდამჭერი პროგრამული სისტემების მთავარი კრიტერიუმია მანქანური დასწავლის ალგორითმების ბიბლიოთეკის სიმდიდრე. ამ მიმართულებით პოპულარული, მოქნილი და ხელმისაწვდომი სკრიპტული პროგრამული სისტემებია „R“, „Python“, “Julia” და სხვ. მანქანური დასწავლის ალგორითმების გავრცელებული ბიბლიოთეკებია Scikit-Learn (Python), JSAT (Java), Accord Framework (.NET). წამყვანი პროგრამული პლატფორმების ინდუსტრიების მიერ შექმნილია Azure Machine Learning მწარმოებელი Microsoft, TensorFlow მწარმოებელი Google, Watson Machine Learning მწარმოებელი IBM, Amazon SageMaker მწარმოებელი Amazon და ა.შ.

მიმდინარე პერიოდში მანქანური დასწავლის გამოყენების მოთხოვნაზე გაზრდის კვალობაზე ტენდენციური გახდა პროგრამული სისტემების ინდუსტრიები, შესაბამისად პროგრამული სისტემების ბაზარი გაჯერებულია მანქანური დასწავლის მხარდამჭერი ინსტრუმენტების შექმნით, თანამედროვე Business Intelligence კლასის სისტემებში მანქანური დასწავლის მექანიზმებით აღჭურვით. გამომდინარე აქედან, საინტერესოა რა პრაქტიკულ შესაძლებლობებს იძლევა მოწინავე პროგრამული ტექნოლოგიები.

2. ძირითადი ნაწილი

სტატისტიკოსები და ანალიტიკოსები პრიორიტეტს ანიჭებენ პროგრამულ სისტემას „R“. პროგრამული კოდი „R“ (IDE - Rstudio) განკუთვნილია მათემატიკური მოდელირების ავტომატიზაციისთვის და მისი უპირატესობაა სტატისტიკური მოდელების მდიდარი ბიბლიოთეკა, მატრიცული ალგებრის მონაცემების სწრაფი დამუშავება [2].

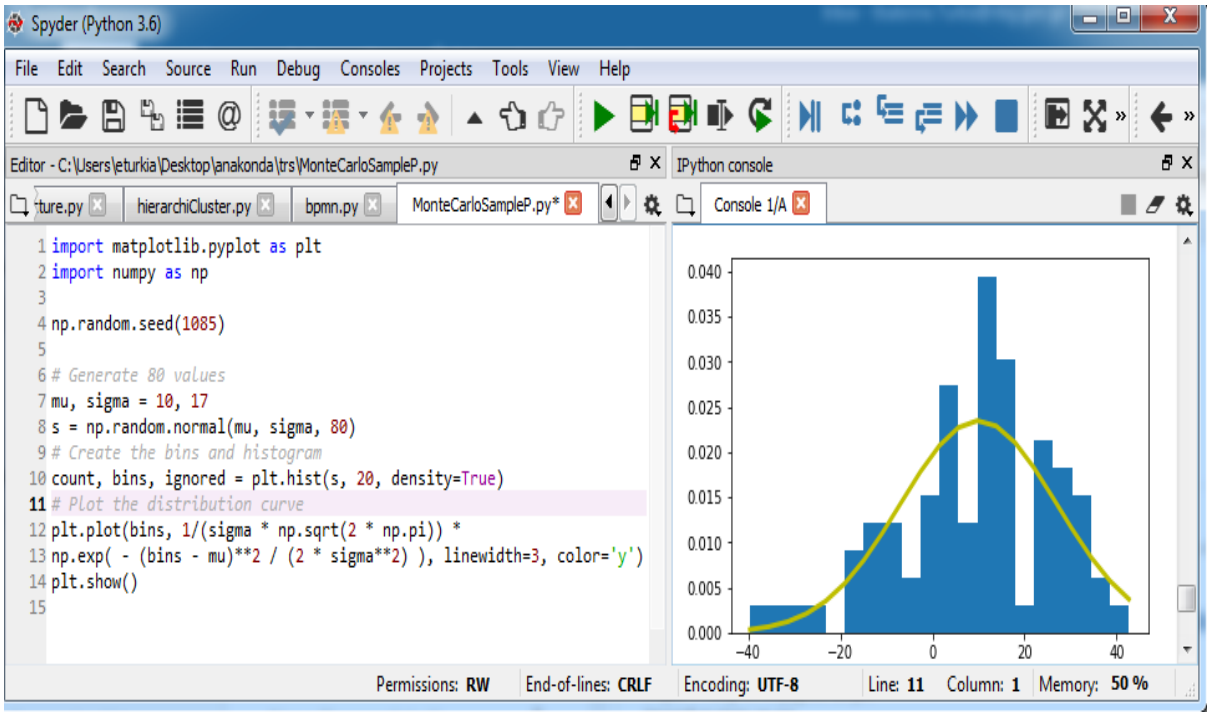
1-ელ ნახაზზე ნაჩვენებია R კოდის მაგალითი რეალიზებული სისტემაში Rstudio. მაგალითში განხილულია ნორმალურად განაწილებული შემთხვევითი რიცხვების ჰისტოგრამა. იგივე მაგალითი რეალიზებული Python კოდზე გარემოში „Spyder” მოცემულია მე-2 ნახაზზე.



ნახ.1. მაგალითის ფრაგმენტი „R“ პროგრამულ ენაში

Python პროგრამულ ენას ლიტერატურაში უნივერსალური და მრავალმიზნობრივი პროგრამირების ენის სტატუსი აქვს. მიუხედავად ამ შეფასებისა მას ჯერ-ჯერობით არ ყოფნის „R“ სისტემის ბიბლიოთეკის მნიშვნელოვანი პაკეტები [3].

R და Python ენების ხიბლია ღია პროგრამული კოდი და სპეციალისტების მიერ ბიბლიოთეკის გამდიდრების შესაძლებლობა.



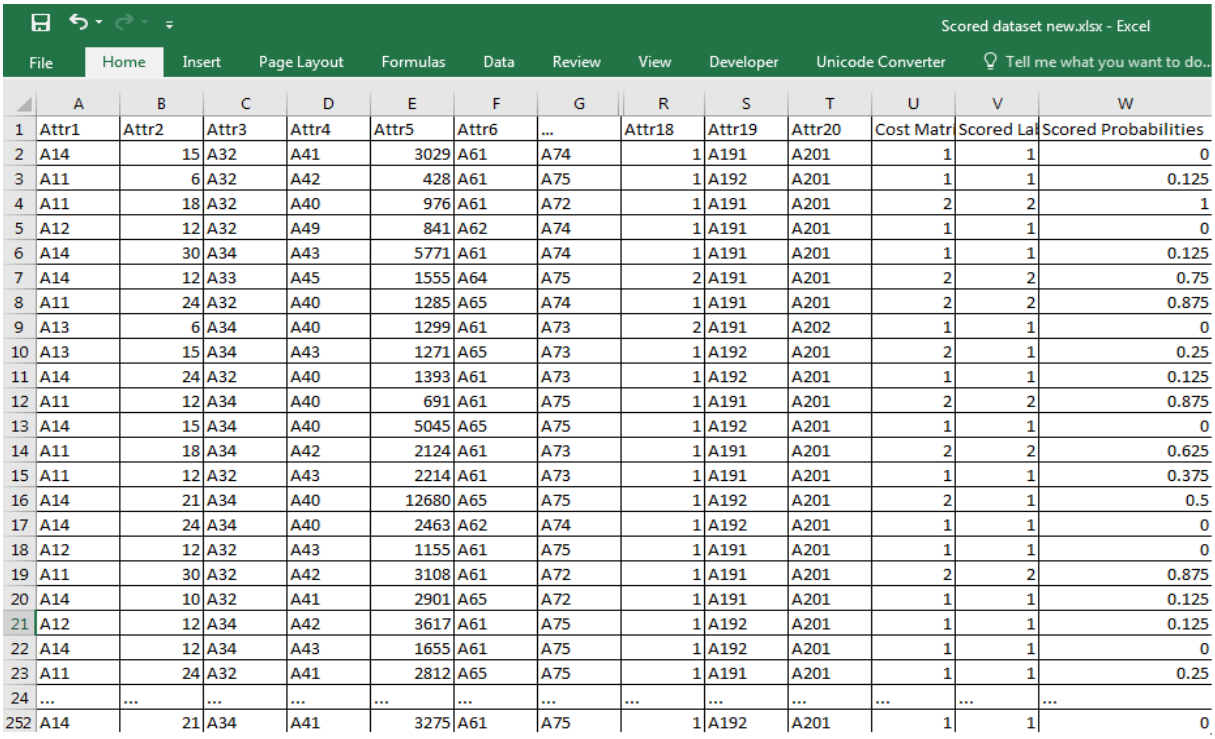
ნახ.2. მაგალითის ფრაგმენტი პროგრამულ ენაში „python“

მანქანური დასწავლის კიდევ ერთი საინტერესო პროდუქტია მაიკროსოფტის „Azure Machine Learning Studio“ (Azure ML), რომელიც გრაფიკულ ინტერფეისზე დაფუძნებული გარემოა მანქანური დასწავლის სამუშაო პროცესის შექმნისა და დანერგვისთვის. Azure ML შექმნილია მანქანური დასწავლის, მაიკროსოფტის პროდუქტების და სერვისების შესაძლებლობების საფუძველზე. დღესდღეისობით Azure საბაზრო სივრცეში განთავსებულია 25-ზე მეტი მანქანური დასწავლის API (Application programming interface).

Azure ML Studio-ში ანალიტიკური ამოცანების ექსპერიმენტული კვლევის მოდელი იწყობა გრაფიკულ-ინტერაქტიული დიზაინით ე.წ. „drag-and-drop“ ხელსაწყოს გამოყენებით, რითაც შესაძლებელია აიგოს, გაიტესტოს და განთავსდეს მანქანური დასწავლის ალგორითმით რეალიზებული ანალიტიკური გადაწყვეტილებები. რეალიზებული მოდელი სერვის-ორიენტირებული კონცეფციის საფუძველზე შესაძლებელია გამოყენებულ იქნას სხვადასხვა სამომხმარებლო აპლიკაციებში (მაგ., Excel, BI სისტემები და სხვ.) ვებ-სერვისის სახით [4].

მაგალითის სახით, განვიხილოთ Azure ML Studio-ში საკრედიტო რისკის პროგნოზირების ამოცანა მანქანური დასწავლის ზედამხედველური ანსამბლური ტიპის Two-Class Decision Forest ალგორითმის გამოყენებით. ზედამხედველური ტიპი ითვალისწინებს არსებული მონაცემების საფუძველზე ახალი მონაცემის პროგნოზირებას, ანუ არსებული საკრედიტო მონაცემების ბაზაზე ახალი კლიენტის გადახდისუნარიანობის განსაზღვრას. მონაცემთა კლასიფიკაცია ხდება ორი მნიშვნელობით - გადახდისუნარიანი (1) და გადახდისუუნარო (0), რომელიც საკრედიტო ისტორიის ბაზის ღირებულებათა მატრიცის ველში (Cost Matrix) აღირიცხება (ნახ.3).

მონაცემთა ბაზა, ზედამხედველური ტიპის გამოყენების დროს, უნდა დაიყოს ლოგიკური პროპორციით სასწავლო და სატესტო მონაცემებად. შესაბამისად, მონაცემთა ბაზა დავყოთ პროპორციით 75:25-ზე (75% - სასწავლო მონაცემი, 25% - სატესტო მონაცემი).



Scored dataset new.xlsx - Excel

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Developer | Unicode Converter | Tell me what you want to do..

	A	B	C	D	E	F	G	R	S	T	U	V	W
1	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	...	Attr18	Attr19	Attr20	Cost Matr	Scored Lal	Scored Probabilities
2	A14	15	A32	A41	3029	A61	A74	1	A191	A201	1	1	0
3	A11	6	A32	A42	428	A61	A75	1	A192	A201	1	1	0.125
4	A11	18	A32	A40	976	A61	A72	1	A191	A201	2	2	1
5	A12	12	A32	A49	841	A62	A74	1	A191	A201	1	1	0
6	A14	30	A34	A43	5771	A61	A74	1	A191	A201	1	1	0.125
7	A14	12	A33	A45	1555	A64	A75	2	A191	A201	2	2	0.75
8	A11	24	A32	A40	1285	A65	A74	1	A191	A201	2	2	0.875
9	A13	6	A34	A40	1299	A61	A73	2	A191	A202	1	1	0
10	A13	15	A34	A43	1271	A65	A73	1	A192	A201	2	1	0.25
11	A14	24	A32	A40	1393	A61	A73	1	A192	A201	1	1	0.125
12	A11	12	A34	A40	691	A61	A75	1	A191	A201	2	2	0.875
13	A14	15	A34	A40	5045	A65	A75	1	A192	A201	1	1	0
14	A11	18	A34	A42	2124	A61	A73	1	A191	A201	2	2	0.625
15	A11	12	A32	A43	2214	A61	A73	1	A191	A201	1	1	0.375
16	A14	21	A34	A40	12680	A65	A75	1	A192	A201	2	1	0.5
17	A14	24	A34	A40	2463	A62	A74	1	A192	A201	1	1	0
18	A12	12	A32	A43	1155	A61	A75	1	A191	A201	1	1	0
19	A11	30	A32	A42	3108	A61	A72	1	A191	A201	2	2	0.875
20	A14	10	A32	A41	2901	A65	A72	1	A191	A201	1	1	0.125
21	A12	12	A34	A42	3617	A61	A75	1	A192	A201	1	1	0.125
22	A14	12	A34	A43	1655	A61	A75	1	A192	A201	1	1	0
23	A11	24	A32	A41	2812	A65	A75	1	A191	A201	1	1	0.25
24	...	...	...	...	...	...	...	...	...	...	...	...	...
252	A14	21	A34	A41	3275	A61	A75	1	A192	A201	1	1	0

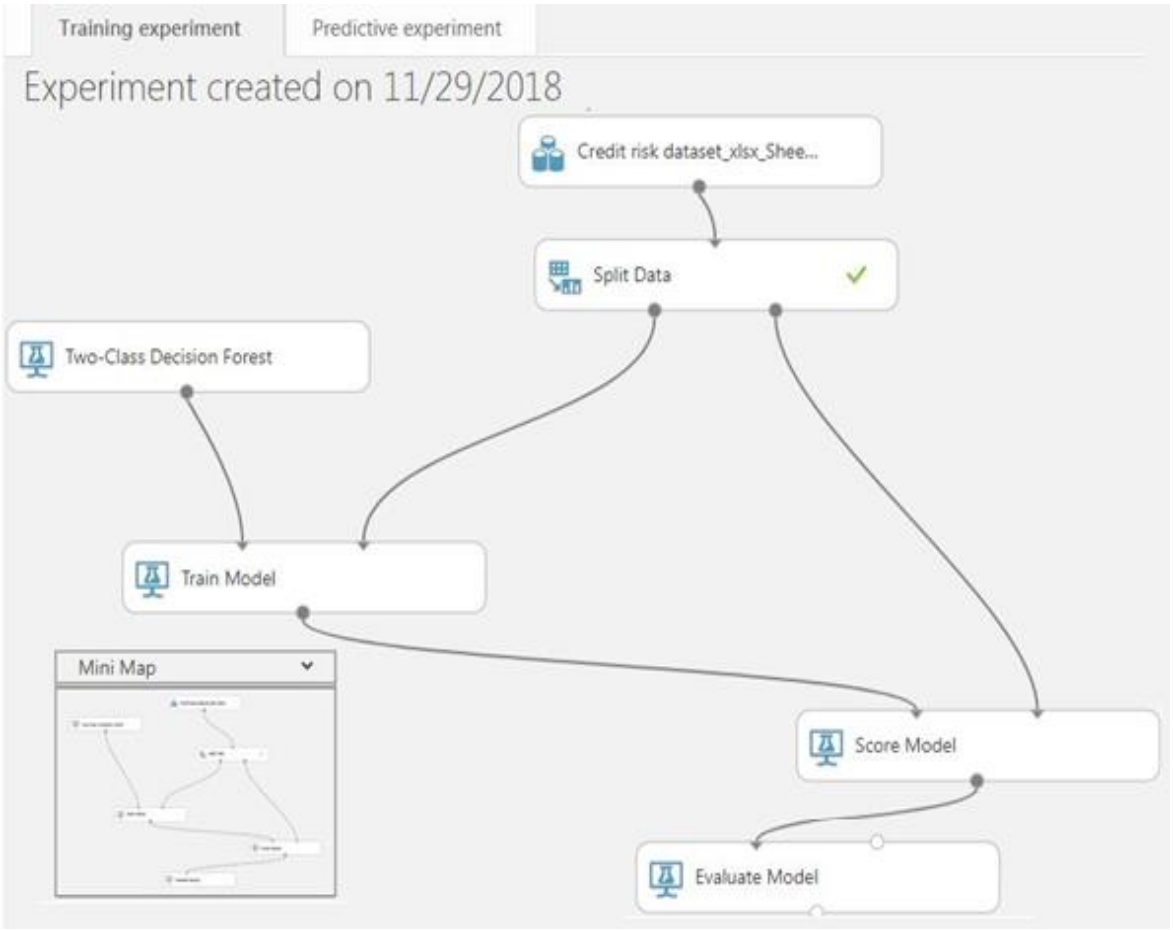
ნახ.3. საკრედიტო ისტორიის მონაცემები ფრაგმენტი

სამუშაო გარემოში იწყობა მანქანური დასწავლის მოდელი კომპონენტების თანმიმდევრობით – ძირითადი ბაზის მოდული უკავშირება მონაცემების პროპორციული დაყოფის (split Data) მოდულს, რომელიც განშტოვდება სასწავლო მონაცემებისა (Train Model) და სატესტო მონაცემების (Score Model) მოდულებში. სასწავლო მონაცემების მოდულთან ხორციელდება კლასიფიკაციის ალგორითმის (Classification) მოდულის დაკავშირება, შესაბამისი ალგორითმის მითითებით (ჩვენს შემთხვევაში ვირჩევთ Two-Class Decision Forest ალგორითმს) და უერთდება სატესტო მონაცემების მოდულს, რომელიც სრულდება მოდელის შეფასების (Evaluate Model) მოდულთან კავშირით. აღწერილი სქემის მიხედვით Azure ML Studio-ში აწყობილი საკრედიტო რისკის პროგნოზირების მოდელი ნაჩვენებია მე-4 ნახაზზე. მანქანური დასწავლის მიერ პროგნოზირებული შედეგები (Scored label) თავისი ალბათობებით (Scored Propabilities) მოცემულია მე-5 ნახაზზე.

Two-Class Decision Forest ალგორითმის გამოყენებით აგებული გადაწყვეტილების ხის მოდელის ფრაგმენტი და აგებული მოდელის სანდოობის შეფასების დიაგრამა ნაჩვენებია შესაბამისად მე-6, 7 ნახაზებზე.

3. დასკვნა

განხილული მაგალითის საფუძველზე შეგვიძლია დავასკვნათ, რომ მაიკროსოფტის მანქანური დასწავლის სტუდია არის ძალზე მოქნილი, მომხმარებელზე ორიენტირებული, გამარტივებული და უფასო ვებ-პლატფორმაზე დაფუძნებული claud ტექნოლოგია. მისი გამოყენება შეუძლია მანქანური დასწავლით დაინტერესებულ ნებისმიერ მომხმარებელს. არსებული ტექნოლოგია ამასთანავე არის ძალიან სწრაფი, შეუძლია დიდ მონაცემებთან გამკლავება მცირე დროში და დამატებით კომფორტს მომხმარებლისთვის ქმნის ისიც, რომ მასზე წვდომა შესაძლებელია ნებისმიერი წერტილიდან სადაც კი არის ინტერნეტი.



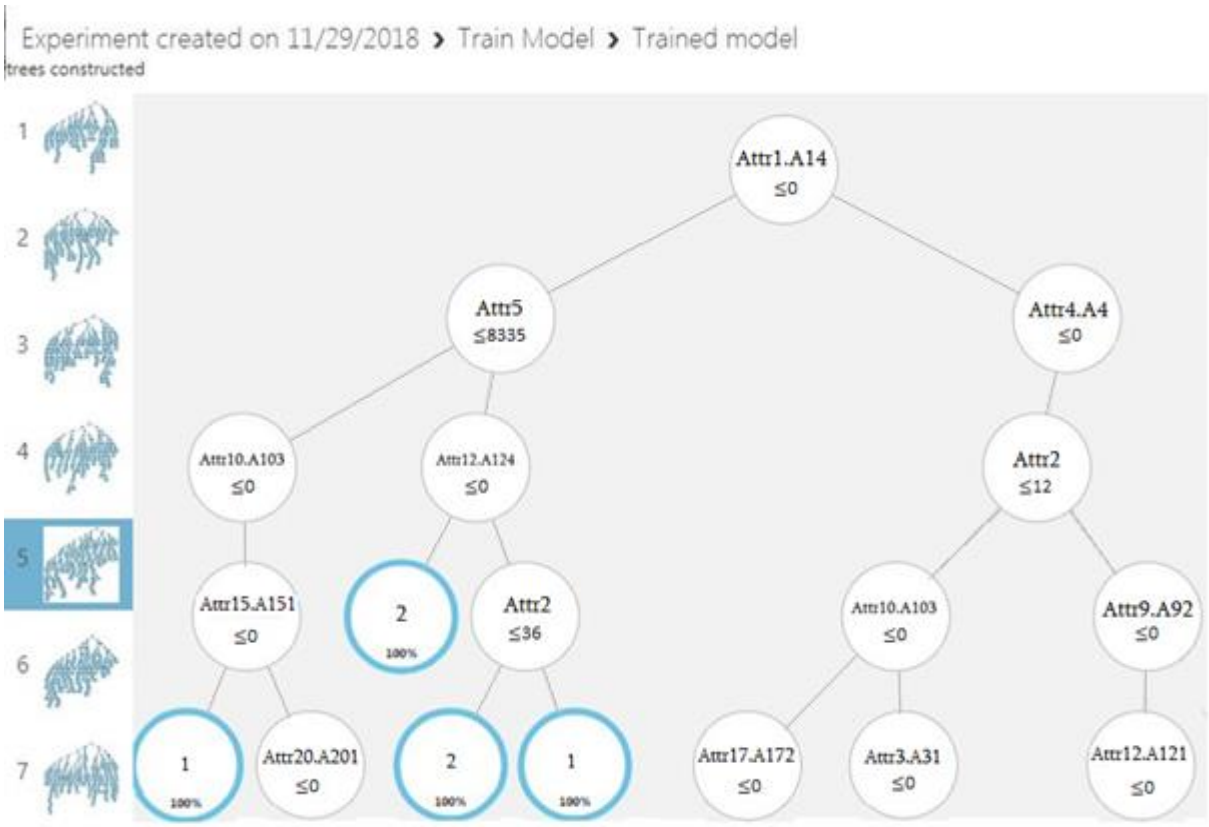
ნახ.4. საკრედიტო რისკის პროგნოზირების მოდელის ფრაგმენტი სისტემაში „MS Azure ML Studio“

Experiment created on 11/29/2018 › Score Model › Scored dataset

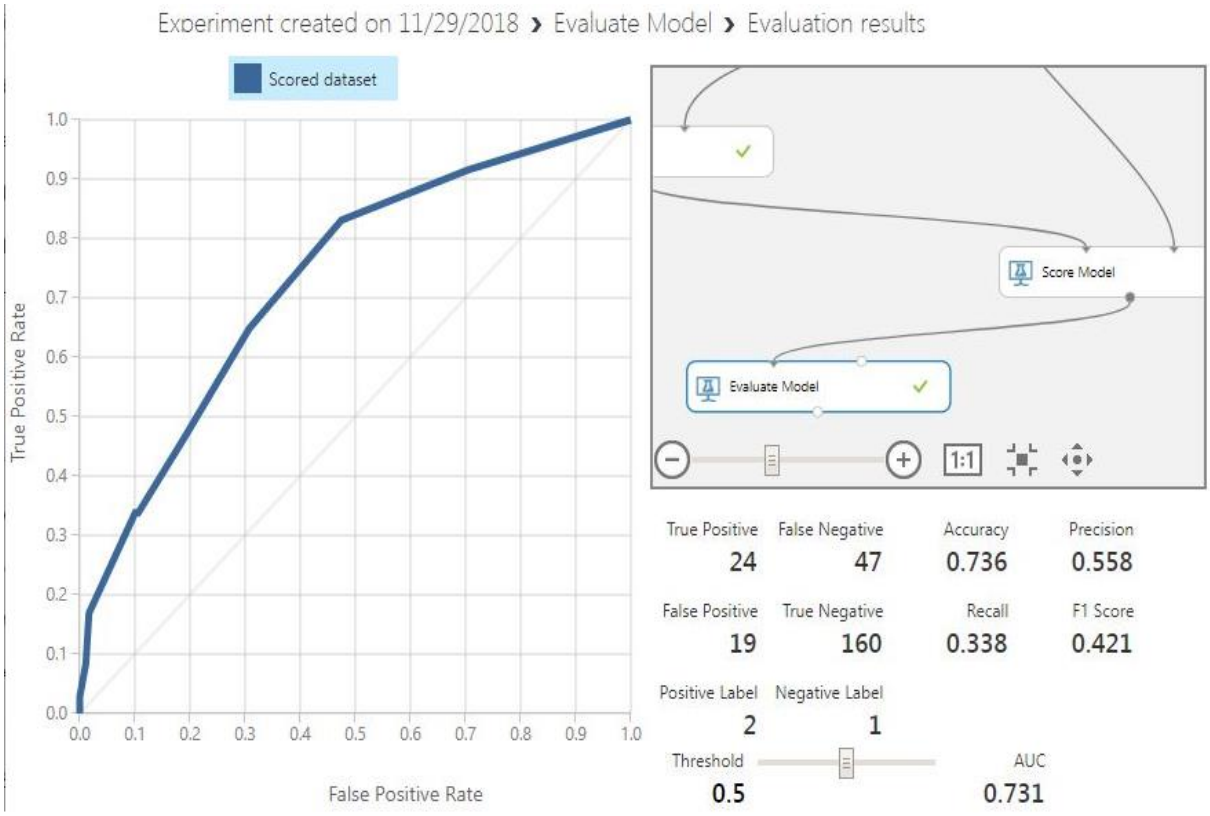
rows 250 columns 23

Attr8	Attr9	Attr10	Attr11	Attr12	Attr13	Attr14	Attr15 ...	Attr19	Attr20	Cost Matrix	Scored Labels	Scored Probabilities
1	A92	A101	4	A124	23	A143	A151	A191	A201	1	1	0.25
4	A93	A101	2	A121	50	A143	A152	A191	A201	1	1	0.25
4	A93	A101	4	A121	57	A142	A152	A191	A201	2	1	0.125
4	A92	A101	3	A121	36	A143	A152	A191	A201	1	1	0.125
1	A93	A101	2	A121	34	A143	A152	A191	A201	1	1	0.25
2	A92	A101	4	A123	25	A143	A152	A191	A201	1	1	0.375
2	A93	A101	3	A123	32	A141	A152	A191	A201	1	2	0.625
4	A92	A101	3	A122	22	A143	A152	A191	A201	2	1	0.25
2	A92	A101	4	A122	26	A143	A151	A192	A201	1	1	0.125
4	A92	A101	2	A122	33	A143	A152	A191	A201	2	1	0.375
4	A93	A101	4	A123	34	A143	A152	A191	A201	1	1	0.25
4	A93	A101	4	A123	28	A143	A152	A191	A201	1	1	0.25
1	A92	A101	2	A121	24	A143	A151	A191	A201	1	1	0.375

ნახ.5. მანქანური დასწავლის მიერ პროგნოზირებული შედეგების ფრაგმენტი



ნახ.6. Two-Class Decision Forest ალგორითმის გამოყენებით აგებული გადაწყვეტილების ხის მოდელის ფრაგმენტი



ნახ.7. მოდელის სანდოობის შეფასების დიაგრამა

მანქანური დასწავლის ავტომატიზებული სისტემები მნიშვნელოვანი მიღწევაა გადაწყვეტილების მიღებისა და ანალიტიკური ამოცანების დამუშავებისთვის. მნიშვნელოვანია, რომ მანქანური დასწავლის ალგორითმების გამოყენება ეფექტურია დიდი რაოდენობისა და გამართული მონაცემების არსებობის პირობებში. გამართული მონაცემები გულისხმობს სასწავლო მონაცემების ხარისხს და უტყუარობას. მანქანური დასწავლის მიმდინარე ტენდენციებისა და პროდუქტიული გამოყენების შედეგებიდან გამომდინარე, შეიძლება ითქვას, რომ მანქანური დასწავლის მოდული პროგრამულ პლატფორმებში დაიკავებს ერთ-ერთ ჩაშენებულ ინსტრუმენტულ საშუალებას.

ლიტერატურა – References – Литература:

1. Kelleher J.D., Namee B. M., D'Arcy A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. http://machinelearningbook.com/
2. Cotton R. Learning R. (2013). O'Reilly USA
3. Python Software Foundation. (). https://www.python.org/
4. https://azure.microsoft.com/en-us/services/machine-learning-studio/

APPLIED ANALYSIS OF MACHINE LEARNING TOOLS

Turkia Ekaterine[1], Gelovani Valeriane[1], Gavasheli Michael[2]
1 - Georgian Technical University
2 - Northeastern University, Boston, USA

Summary

Today machine learning is one of the highly demanded and developing field of IT and Statistics. Practically all areas of research that deal with large amounts of data and empirical analysis are turning to this technology. This paper reviews currently used programming languages and systems for machine learning applications. These are Python, R, and Microsoft Azure Machine Learning Studio. Practical examples of machine learning algorithms and coding principles of mentioned systems are shown.

ПРИКЛАДНОЙ АНАЛИЗ ПРОГРАММНЫХ ПРОДУКТОВ МАШИННОГО ОБУЧЕНИЯ

Туркия Е.[1], Геловани В.[1], Гавашели М.[2]
1 - Грузинский Технический Университет
2 - Северо-Восточный университет, Бостон, США

Резюме

Машинное обучение сегодня является одним из самых прогрессивных и высоко востребованных направлений. Этой технологией интересуются практически все те исследования, чьи аналитические результаты и решения зависят от анализа большого количества данных. В данной статье мы рассматриваем актуальные для машинного обучения в текущем периоде языки программирования и системы "Питон", "R", "Microsoft Azure Machine Learning Studio". Показаны практические примеры алгоритмов машинного обучения и принципы работы в указанных системах.