LĒMUMA KOKI KLASIFIKĀCIJAS UZDEVUMOS
DECISION TREES IN CLASSIFICATION TASKS

Autors: **Lauris STIRNA**, e-pasts: lavrencij@inbox.lv
Zinātniskā darba vadītājs: Dr.sc.ing., docents **Sergejs KODORS**
Rēzeknes Tehnoloģiju akadēmija,
Rēzekne, Atbrīvošanas aleja 115

Abstract. *The goal of the study is to develop a classification system for Iris dataset. The analytical development environment called KNIME and freely available Iris dataset were applied in the research. The developed classification is based on the decision tree application. The accuracy of decision tree is compared with artificial neural network.*

Keywords: *classification, decision trees, ID3.*

Ievads

Strauja informācijas tehnoloģiju attīstība, tajā skaitā, arī progress datu uzkrāšanā, uzglabāšanā un apstrādē ļāva lielam daudzumam organizāciju krāt lielu datu apjomu, kuru ir nepieciešams analizēt. Datu apjomi palika tik lieli, ka ekspertu (cilvēku) spēku to apstrādei ir palicis par maz, kas radījis pieprasījumu (līdz ar to arī problēmu) pēc automātiskās analīzes metodēm, kuru skaits ik gadu palielinās un turpina to darīt arī tagad. Lēmumu koki ir viena no tādām automātiskās analīzes metodēm. [4]

Informācijas jomā panāktais progress, jo īpaši datu apstrādes attīstība, rada milzīgu informācijas daudzumu. Tādu lielu informācijas apjomu analīzes rezultātā rodas grūtības iesniegt nepieciešamos datus analīzei piemērotā formā. Galvenā prasība informācijas sistēmai, kas vērsta uz datu analīzi, ir savlaicīga analītikas nodrošināšana ar visu informāciju, kas nepieciešama lēmuma pieņemšanai. Šādu modeļu konstruēšanas metodes parasti tiek attiecinātas uz mākslīgā intelekta jomu. [3] Datu ieguves algoritmi ietver: lēmumu kokus, tuvāko kaimiņu metodi, atbalsta vektora metodi, Baijesas tīklu un neironu tīklus.

Autora izvēle krita tieši uz lēmuma kokiem līdz ar to, ka lēmumu kokus ir vieglāk saprast un interpretēt, taču tie arvien biežāk tiek aizstāti ar daudz jaudīgākām metodēm. Atšķirībā no tām (metodēm) lēmuma koki nebūs labākā izvēle, ja formulētais uzdevums sastāv no pārāk lieliem datu masīviem vairākos līmeņos un/vai kategorijās, kuru gala izklāsts paliks nepilnīgs un daudz sliktāk optimizēts salīdzinājumā ar citiem alternatīviem variantiem. Plašu pielietojumu lēmuma koki ir guvuši rūpniecībā, medicīnā, molekulārajā bioloģijā un banku lietvedībā. [4]

Pētījuma mērķis ir pierādīt **hipotēzi**, ka lēmuma kokus var pielietot, lai automatizētu datu klasifikāciju.

Materiāli un metodes

Lai sasniegtu mērķi un pārbaudīt hipotēzi, autors pielietoja analītikas programmu *KNIME* un datu kopu *Iris*.

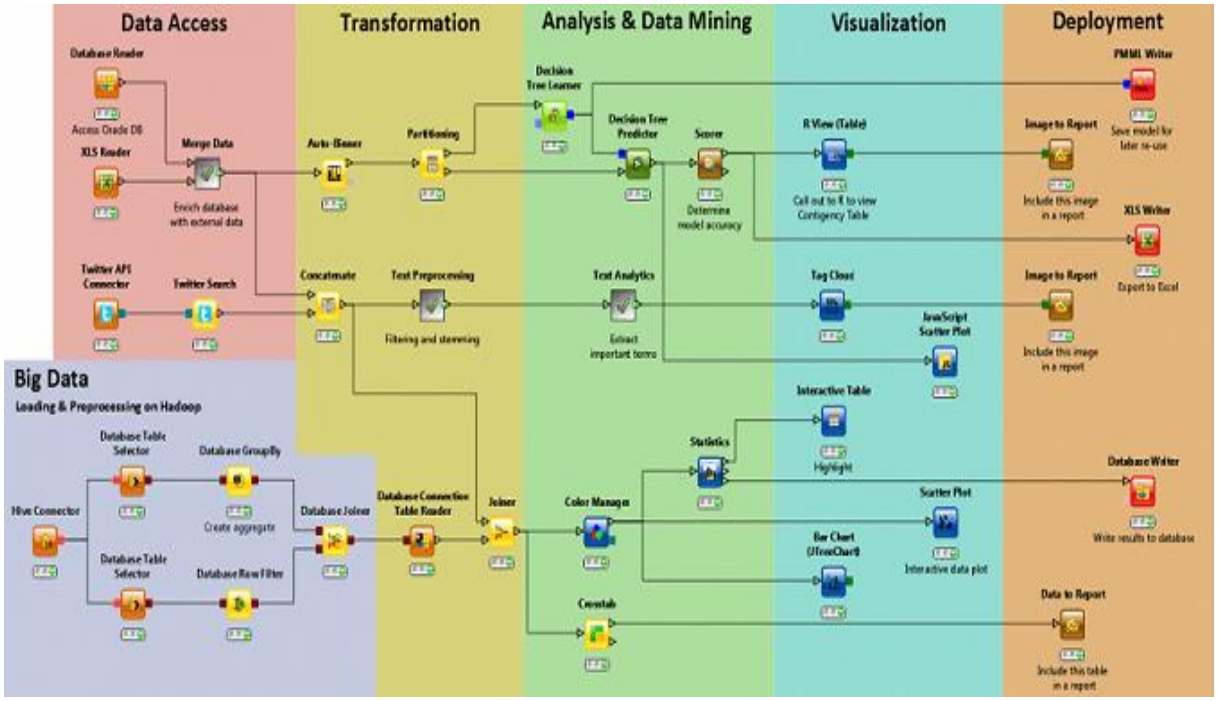
***KNIME* vai *Knime Analytics Platform* (tālāk *KAP*)** ļauj īstenot pilnīgu datu analīzes ciklu, kas ietver datu lasīšanu no dažādiem avotiem, konvertēšanu un filtrēšanu, analīzi, vizualizāciju un eksportu. *KAP* ir *open source* jeb brīvi pieejamā programmatūra, kura var būt noderīga, ja:

- lietotājs grib analizēt datus;
- lietotājs grib analizēt datus un viņam nav zināšanu programmēšanas nozarē;
- lietotājs grib pārskatīt jau realizēto algoritmu bibliotēku, un, iespējams, atrast ko jaunu.

https://doi.org/10.17770/het2021.25.6788

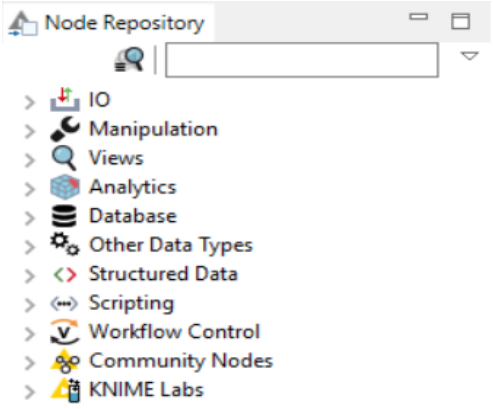
KAP programmēšanas loģikas process tiek veikts, izveidojot darbplūsmu (*workflow*). Darbplūsma sastāv no mezgliem, kas izpilda noteiktu funkciju (piemēram, datu nolasīšana no datu bāzes, to pārveidošana, vizualizācija). Atbilstoši mezgli ir savstarpēji savienoti ar bultiņām, kas norāda datu kustības (plūsmas) virzienu. [2]

KAP oficiālajā mājas lapā ir pieejama diagramma, kura labi atspoguļo, kā var komponēt dažādu tipu mezglus vienā plūsmā (skat. 1. attēlu).



1. att. Dažādu tipu mezglu komponēšanas iespējas [2]

Kā jau minēts, jebkura *KAP* darbplūsma sastāv no mezgliem (*nodes*), praktiski katram mezglam ir konfigurācijas dialogs, kur lietotājs var konfigurēt iestatījumus. *KAP* atbalsta šādu mezglu tipus (skat. 2. attēlu): *IO* - datu ievade/izvade (piemēram, *CSV reading*), *Manipulation* – datu pārveidošana (rindu un/vai kolonnu kārtošana, šķirošana), *Views* - datu vizualizācija (dažādu diagrammu un grafiku interpretācija, ieskaitot *Histogram*, *Pie Chart*, *Scatter Plot*, utt.), *Database* – iespēja veidot savienojumu ar datu bāzi, lasīt/rakstīt, *Workflow Control* – ciklu veidošana, grupu iterācijas *workflow* izpildes gaitā, u.c. No mezgliem, kas īsteno datu analīzi, ir pieejamas daudzas statistiskās analīzes metodes (t.s. lineārā korelācija, hipotēzes pārbaude) un *Data Mining* metodes (piemēram neironu tīklu, lēmuma koku, *cluster view* izveide). [2]



2.att. Pieejamo mezglu tipu klāsts

Datu kopa "*Iris*" ir viena no populārākajām un pazīstamākajām mašīnumācīšanas datu kopām, kas iegūta no *UCI* repozitorijas. *Iris* datu kopu izveidoja R. Fišers (*R.A. Fisher*). Datu kopa "*Iris*" satur trīs dažādas klases, pa 50 objektiem katrā. "*Iris*" esošo atribūtu sarakstu var raksturot kā kategoriskus, nominālus un nepārtrauktus. Eksperti ir minējuši, ka nevienā atribūtā nav trūkstošās vērtības no šī datu kopuma. Datu kopums ir pabeigts. [1]

Pirmā no klasēm ir lineāra kas atšķiras no pārējiem diviem, bet pārējās divas nav lineāli atdalāmas. Kopā 150 gadījumi, kas ir vienādi sadalīti starp trim klasēm, satur četrus kvantitatīvus atribūtus:

- „ziedlapas" garums – daļskaitlis (*double*);
- „ziedlapas" platums – daļskaitlis (*double*);
- „lapas" garums – daļskaitlis (*double*);
- „lapas" platums – daļskaitlis (*double*).

Piektais atribūts ir prognozes atribūts, kas ir klases atribūts, kas nozīmē katru gadījumu ietver arī identifikācijas klases nosaukumu, no kuriem katrs ir viens no trijiem:

- *IRIS Setosa;*
- *IRIS Versicolour;*
- *IRIS Virginica.*

Rezultāti un to izvērtējums

Uzsākot pētniecisko darbību, autors apmeklēja *KNIME* oficiālo māja lapu un lejupielādēja šo brīvi pieejamo analītisko vidi. Lejupielādei ir pieejami *.exe* fails (instalācijai) un *.zip* arhīvs, kurš satur jau izpakotu analītisko platformu, kuru autors arī izmantoja. Papildus tam, tiek lejupielādēta *Iris* datu kopa, kura tāpat ir brīvi pieejama pētījumu un citu darbību veikšanai. Neskatoties uz to, ka turpmāk pētījumā lietos *KNIME* pieejamo mezglu *CSV-reader,* kurš pēc noklusējuma „prasa" failu ar datiem attiecīgi *.csv* formātā, datu kopu "*Iris*" var lejupielādēt *.txt* formātā un tajā pašā formātā padot uz mezglu.

Tālāk, pēc *KNIME* palaišanas, tiek veidots jauns projekts caur pamata izvēlni: *File – New – New KNIME Workflow*, kas tulkojumā ir darbplūsma. Nākošais solis projekta izveidē piedāvā ievadīt darbplūsmas nosaukumu un direktoriju, kur to saglabāt.

Tālāk programma palaiž tukšu darba vidi, kurā autors paredzēja izvietot trīs mezglus, katrs no kuriem pilda attiecīgos uzdevumus:

- *CSV-reader* - mezgls, kurš nolasa *Iris* datus;
- *Decision Tree Learner* - mezgls, kurš mācīsies no iegūtajiem datiem;
- *Decision Tree Predictor* - mezgls, kurš veiks attiecīgo prognozēšanu un parādīs klasifikācijas rezultātu.

Pirms palaist analīzes procesu, tālāk seko *CSV Reader* mezgla konfigurēšana - mezglam ir jānorāda *Iris* datu kopa un jāveic lasīšanas parametru iestatīšana.

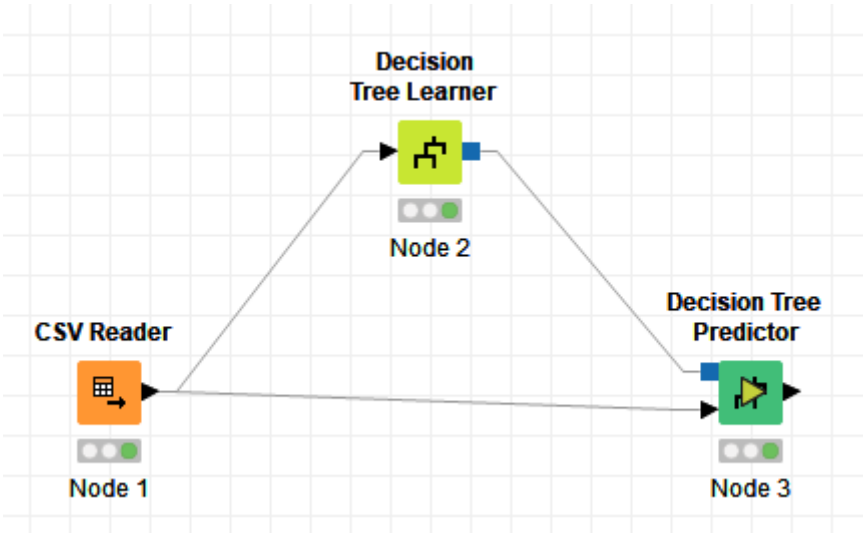
Ņemot vērā to, ka *Iris* datu kopa netiek nekādā veidā rediģēta, iestatījumos jāatslēdz ailes „*Has Column Header*" un „*Has Row Header*" līdz ar to, ka nemainot datu kopas saturu, tai nav nedz kolonnu, nedz rindu virsrakstu, taču tas nemainīs un nekādā veidā netraucēs turpmākai pētījuma veikšanai. Visi pārējie iestatījumi paliek pēc noklusējuma. Pielietojot konfigurāciju (*Apply* poga), *CSV Reader* mezgls iedegsies ar dzeltenu gaismu, un ar to ziņos, ka tas ir gatavs izpildei, tas arī tiek veikts ar *Execute* pogu.

Tālāk programmas vide ziņo par konfigurēšanas nepieciešamību nākošajam mezglam jeb *Decision Tree Learner* mezglam. To arī dara, taču konfigurēšanas logā priekš konkrētā eksperimenta, vajadzības pēc iestatījumu maiņas nav nekādas. Bet neskatoties uz to, ka kaut ko mainīt iestatījumos nav vajadzības, darbplūsmas tālāka izpilde nav iespējama pirms iestatījumu pielietošanas (*apply*) vismaz pēc to noklusējuma.

Pirms pēdējā mezgla konfigurēšanas, atkal ir jāpalaiž mezgla izpilde (*execute*). Rezultātā 2 no 3 mezgliem būs iedegušies ar zaļām gaismām, kas ziņo par veiksmīgu izpildi un par to, ka

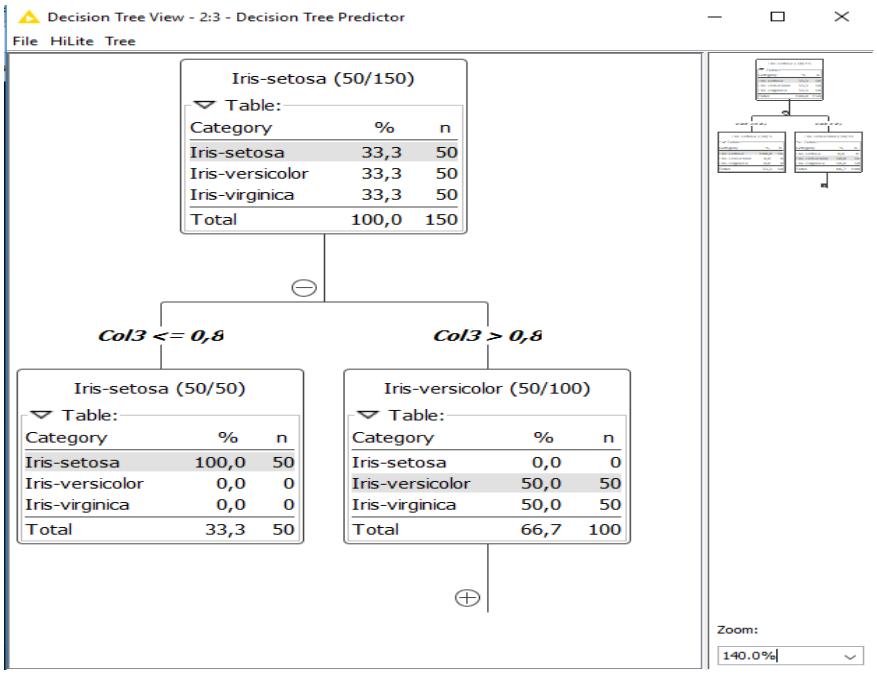
mezgla funkcija ir veiksmīgi pabeigta. Tālāk notiek *Decision Tree Predictor* mezgla konfigurēšana, pēc līdzīga scenārija - mainīt kaut ko iestatījumos nav nepieciešamības, bet to veiksmīga izpilde nav iespējama bez pielietošanas.

Pēc kārtējā pielietojuma (*apply*) un mezgla izpildes (*execute*), rezultātā darbplūsma izskatās šādi (skat. 3. attēlu).



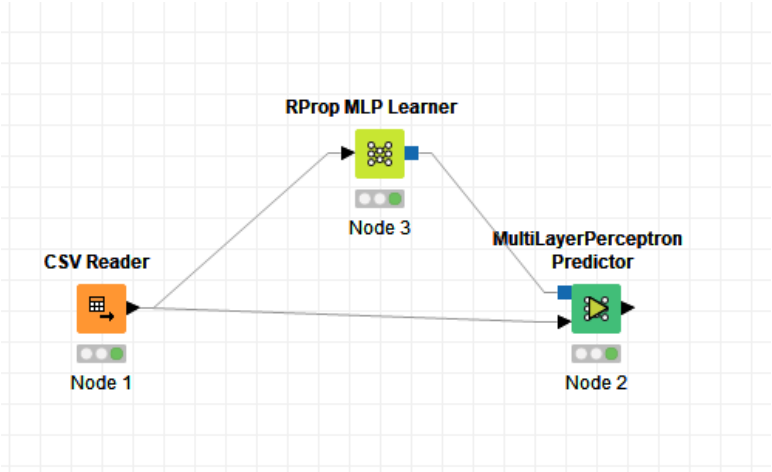
3.att. Lēmuma koka darbplūsmas gala izskats

Kad visi mezgli ir pareizi konfigurēti (šajā gadījumā gandrīz viss palicis pēc noklusējuma) un izpildīti, rezultātam jābūt tādam, kā tas ir redzams augstāk: katrs mezgls deg ar zaļu gaismu. Praktiskā daļa ir paveikta, un pēdējais, kas atliek, ir pārskatīt rezultātu, to var izdarīt, spiežot uz *Decision Tree Predictor* mezglu ar labo peles pogu, un izvēloties „*View: Decision Tree View*". Rezultātā tiek iegūts tipisks *ID3* algoritma lēmumu koka apraksts (skat. 4. attēlu).



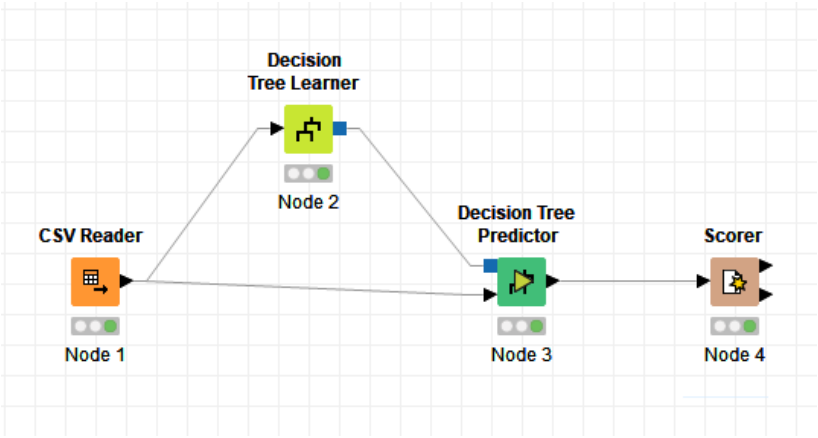
4.att. Lēmuma koka fragmenta ilustrācija

Autors šajā darbā nolēma izanalizēt klasifikatoru rezultātus priekš **atpazīšanas precizitātes** (zemāk), izmantojot salīdzināšanu starp paša klasifikācijas sistēmu ar analoģiski realizēto neironu tīkla variantu (skat. 5. attēlu).



5.att. Neironu tīkla darbplūsmas gala izskats

Mēģinot salīdzināt iegūtos rezultātus, parādījās vajadzība pieslēgt papildus *Scorer* mezglu (skat. 6. attēlu), kas ļaus noteikt atšķirību starp lēmuma koku klasifikatora un neirona tīkla klasifikatora rezultātiem procentuālā izteiksmē. Šo soli veica gan lēmuma koku klasifikatora darbplūsmai, gan neironu tīkla darbplūsmai.



6.att. *Scorer* mezgla pieslēgšanas ilustrācija

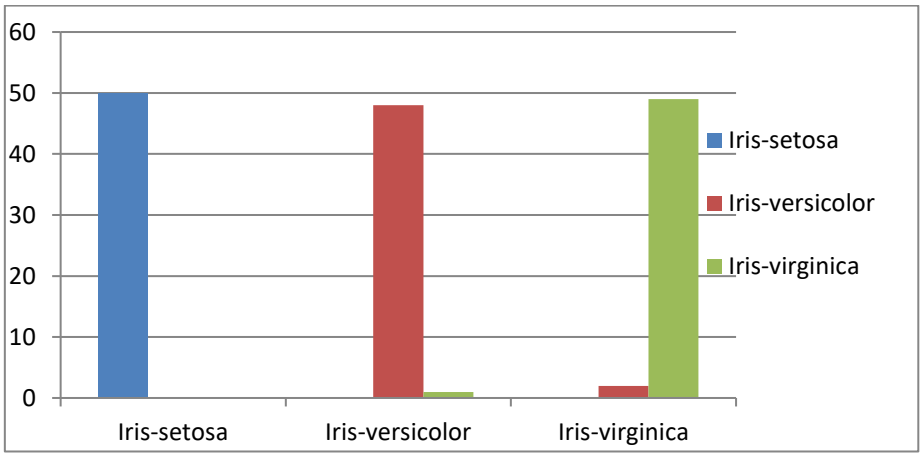
Scorer mezgls atrodams pēc ceļa: *Analytics – Mining – Scoring – Scorer*. Spiežot uz *Scorer* mezglu ar labo peles pogu, un izvēloties „*View: Confusion Matrix*", iegūst autoram nepieciešamo procentuālo izteiksmi, kā arī Kohena kappa koeficientu (*Cohen's kappa*), (skat. 7. attēlu).

Par pārsteigumu autoram, gan lēmumu koka, gan neironu tīkla rezultāti bija identiski:

- Nepareizi (*wrong classified*) klasificētas ir 3 gadījumi (skat. 7. attēlu);
- Klasifikācijas precizitāte (*accuracy*) ir 98%;
- Kohena kappa koeficients ir 0,97.

Pie precizitātes līdz 50% (ieskaitot) klasifikācija ir bezjēdzīga, jo to var salīdzināt ar metamo kauliņu. ~75% var uzskatīt par vidēji kvalitatīvu klasifikāciju, 90% un augstāk liecina par praktiski pielietojamo klasifikācijas modeli, uz kuru arī tiecas mūsdienu klasifikācijas attīstība.

Tipisks Kohena kappa kritērija pielietojums ir cilvēku vai priekšmetu novērtējums, ko veic divi eksperti. Šajā gadījumā *k* norāda ekspertu savstarpējo piekrišanas pakāpi. [3]



7.att. *Scorer* mezgla ilustrācija stabiņu diagrammas veidā (*Y* ir piemēru skaits)

Secinājumi

- Priekš darba veiksmīgas izpildes tika izdalīti trīs pamata soļi. Tie visi ir izpildīti, tāpēc mērķis ir sasniegts. Rezultāts (izstrādāt klasifikācijas sistēmu, kas lietos lēmumu kokus) ir pilnā mērā sasniegts;
- Izskanējusi hipotēze darba sākumā ir apstiprinājusies konkrētam gadījumam jeb salīdzinot ar neironu tīkliem tā nepiekāpās precizitāte izvēlētajos apstākļos, kaut arī tagad daudz biežāk jau runā par daudz jaudīgākām un populārākām klasifikācijas metodēm;
- Šī pētījumu niša jau ir pietiekami izpētīta un kaut ko pilnīgi novatorisku šeit atklāt ir objektīvi sarežģīti, taču tas neatceļ būtiskāko šī darba mērķi – iepazīstināt un personīgi veidot gan projektu ar pirms nepazīstamo programmatūru, gan veikt analīzi, gan sniegt rekomendācijas turpmākai šīs sfēras attīstībai, kura noteikti turpinās augt un paplašināties;
- Izvēlētā programmatūra, metodes un pieeja ļauj veikt padziļinātu ieskatu šajā virzienā bez jebkādas iepriekšējās specializācijas un/vai izglītības, tas liecina par dotā darba aktualitāti, lai lasītājs varētu ne tikai iepazīties ar saturu, bet arī veikt līdzīgus pētījumus, kādi tie ir aprakstīti šeit;
- Lēmuma koki un *KNIME* vide ir ļoti draudzīga priekš nepieredzējuša lietotāja un pie pareizām un ne visai sarežģītām darbībām izklāsta rezultātus ļoti ilustratīvā veidā, kas atkārtoti atvieglo sapratni par to, kas šajā darbā notiek;
- Izmantotā metodoloģija ar *KNIME* vidi var tikti izmantota daudz plašāk pētāmos jautājumos t.s. eksperimentējot ar citām klasifikācijas metodēm;
- Izmantotā pieeja ir gaužām vienkārša realizācijā un var tikt uzlabota, pielietojot vairāk mezglu darbplūsmā, kuri dažādos veidos var apstrādāt un/vai precizēt datus, to saturiskumu.

Summary

The author looked at the software base found, or the KNIME analytic environment, and the data set for the future Iris Dataset. The author concluded that the chosen environment is very friendly to users without any specification and additional skills and can be used for just knowing the theoretical part of the research to be performed. Both KNIME and Iris Dataset have freely available resources on the Internet and can be freely used by every user.

The work does not apply to the descriptions of through researches and to the full accuracy, as well as performs only the informative function, taking into account that the

direction of the chosen work has been in the active research environment for a long time, and with each passing year, the information on all this is renewed and expanded.

The content of the work emphasizes that decisions trees still remain relevant in some areas and, for the most part, their application remains inactive because of the emergence of increasingly new and more powerful classification methods, with the fact that this field of research is progressively developing now and will develop in the future. During the development of the study, all the hypotheses, objectives and tasks that have been put forward are successfully achieved by applying the methods that were originally planned. The obvious strength of the work is its ease of readability and transparency, as well as the very detailed description of all the work in the experimental part of the work.

The shortcomings can be noted in the not-too-perfect author's qualifications in the chosen direction, thus, in the course of reading, some inaccuracies can be detected in the translations of scientific works and in their personal sense. The main question of this research, whether the decision trees are not completely obsolete method in the field of classification, was clarified and practically proved.

Literatūra

1. M. Swain, S. Kumar Dash, S. Dash, A. Mohapatra. IRIS plant dataset. *International Journal on Soft Computing (IJSC)*, 2012. Vol. 3. No. 1. 80-81 pgs.
2. Владимиров А. Обзор *Knime Analytics Platform* — open source системы для анализа данных. Sk. Internetā (11.05.2018) https://habr.com/post/320500
3. Певченко С., Блужин В. Сравнительный анализ алгоритмов нейронной сети и деревьев принятия решений модели интеллектуального анализа данных. Sk. Internetā (24.05.2018) https://moluch.ru/archive/132/36999/
4. Шахиди А. Области применения деревьев решений. Sk. Internetā (16.04.2018) https://basegroup.ru/community/articles/description