VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Tiesioginio sklidimo DNT naudojant sistemą WEKA

Laboratorinis darbas

Atliko: 4 kurso 1 grupės studentas
Paulius Minajevas (SN: 2110599)

Vilnius – 2024

Turinys

Įvadas

Užduoties tikslas – Išmokyti neuroninį tinklą teisingai klasifikuoti duomenis naudojant sistemą WEKA. Užduoties variantas - 0, pasirinktas pagal studento numerio paskutinio skaitmens liekaną iš 3 (mod(9,3) = 0).

1. Klasifikavimo duomenys ir metodika

1.1. Klasifikavimo duomenys

Neuroniniam tinklui apmokyti ir testuoti naudojamas Irisų duomenų rinkinys. Šis rinkinys skirtas atpažinti gėlės rūšį pagal gėlės atributus. Viso rinkinį sudaro 150 eilučių, kurie padalinti į du rinkinius: vieną, po 40 eilučių kiekvienos klasės ir kitą po 10 kiekvienos klasės likusių eilučių. Antras rinkinys buvo naudojamas kaip modelio nematyti duomenys, klasifikavimo metrikoms tikrinti.

1.1.1. Irisų duomenų rinkinys

Irisų duomenų rinkinys susidaro iš atributų:

1. ”sepal length” [taurėlapio ilgis]
2. ”sepal width” [taurėlapio plotis]
3. ”petal length” [žiedlapio ilgis]
4. ”petal width” [žiedlapio plotis]
5. ”class” [klasė]

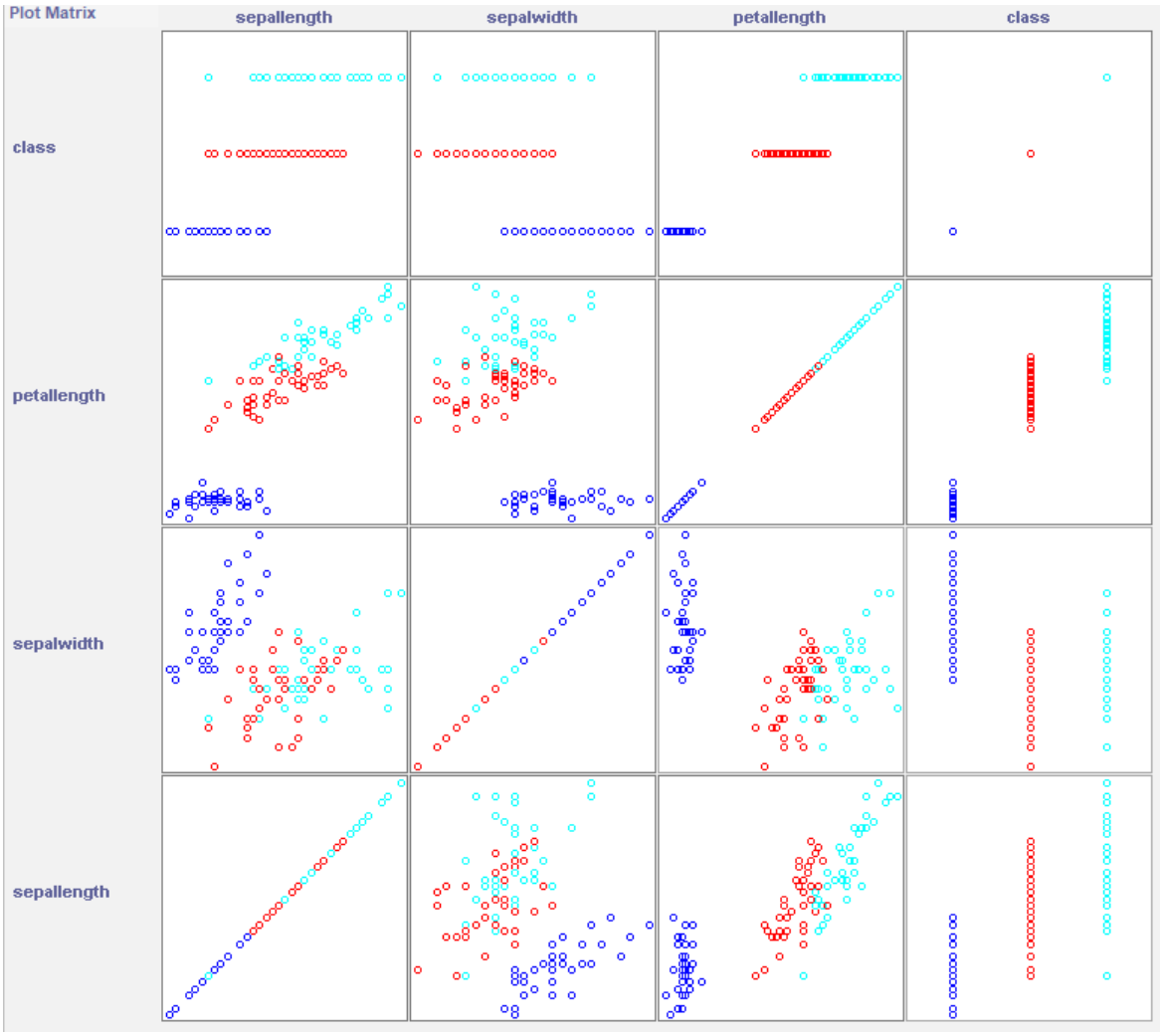
, iš kurių, šiame duomenų rinkinyje, galima atpažinti kelias gėlių rūšis:

1. Iris Setosa
2. Iris Versicolour
3. Iris Virginica

Kadangi pasirinktas 0 variantas, bus naudojami tik trys gėlių atributai: „sepal length”, „sepal width”, „petal length”. Modelis klasifikuos visas tris galimas klases pagal šiuos požymius.

1.1.2. Mokymo ir testavimo duomenų požymių porų vaizdas

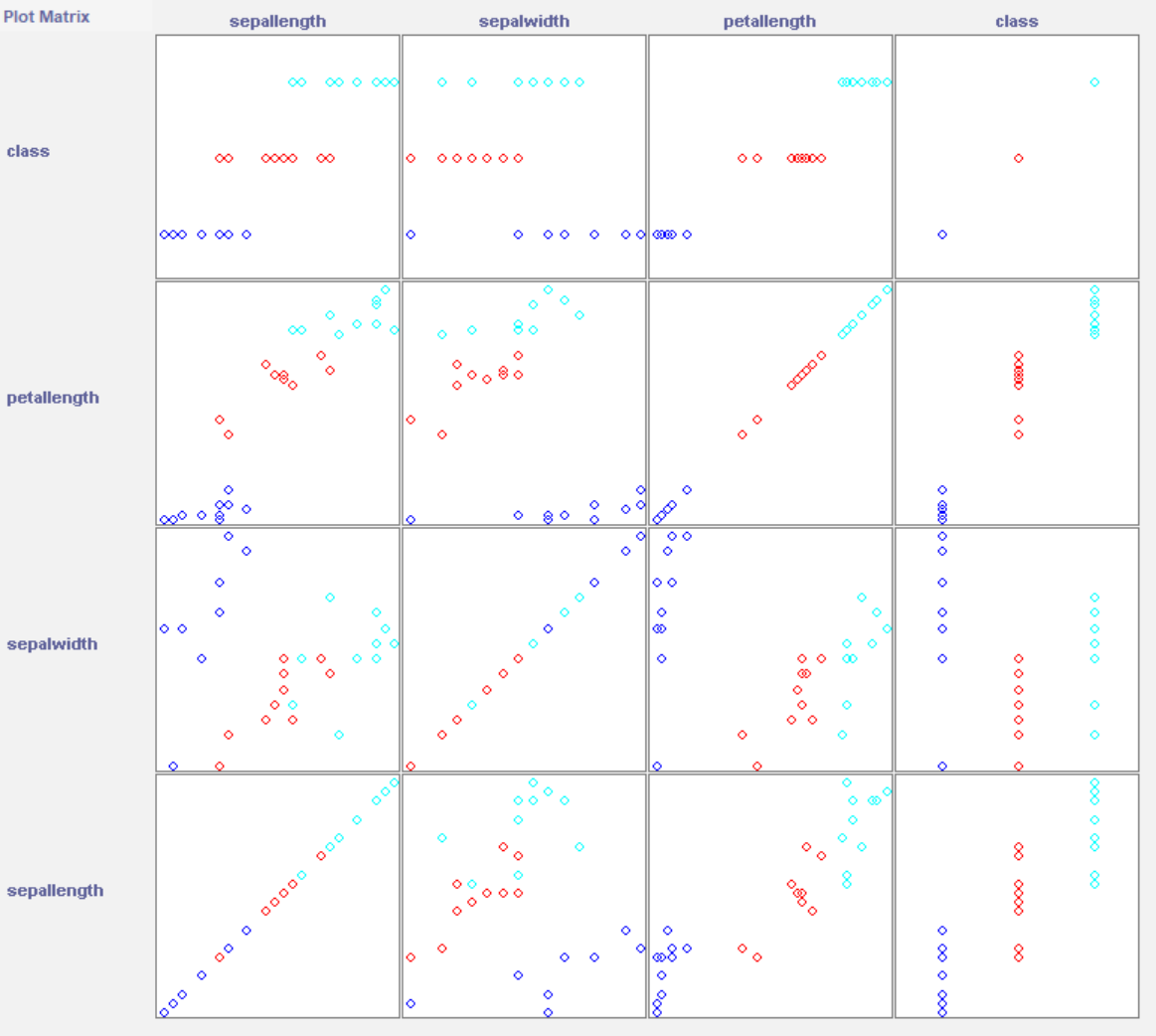
Žemiau pateikiama mokymo ir testavimo duomenų požymių porų vizualizacija Dekarto koordinačių sistemoje. Klasės atitinka šias spalvas: Iris-setosa - mėlyna, Iris-versicolor - raudona, Iris-virginica - žydra.



1 pav. Mokymo ir testavimo duomenų vizualizacija

1.1.3. Nematytų duomenų požymių porų vaizdas

Žemiau pateikiama modelio nematytų duomenų požymių porų vizualizacija Dekarto koordinačių sistemoje. Klasės atitinka šias spalvas: Iris-setosa - mėlyna, Iris-versicolor - raudona, Iris-virginica - žydra.



2 pav. Modelio nematytų duomenų vizualizacija

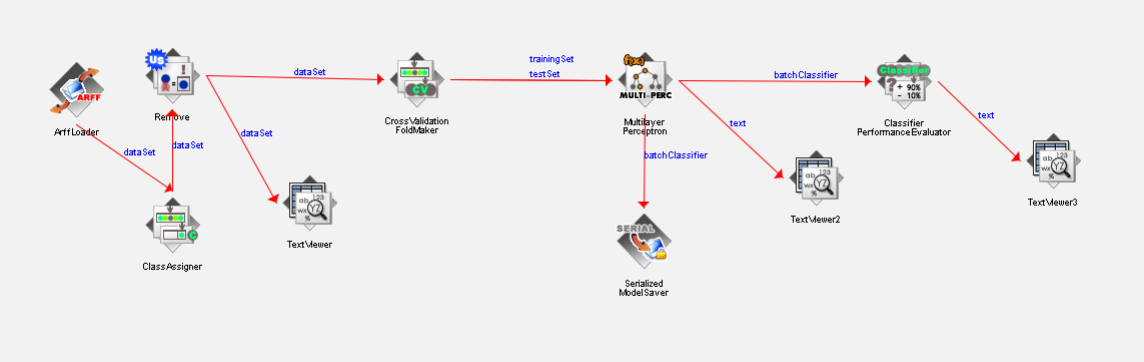
1.2. Metodika

Šiame laboratoriniame darbe bus konstruojamos kelios sistemos, naudojant „WEKA" programą. Žemiau poskyriuose pateikiamos programoje sudarytų sekų nuotraukos bei jų aprašymai.

1.2.1. Pirma užduočių seka

Pirma „WEKA" užduočių seka yra skirta daugiasluoksniam perceptronui apmokyti, naudojant treniravimo duomenis. Pirmiausia komponentėje „Remove" pašalinamas vienas iš duomenų požymių, paliekant tik tris klasifikacijai reikalingus požymius. Toliau komponentėje „SerializedModelSaver" nurodoma vieta, kur bus išsaugotas apmokytas modelis. Kryžminės patikros blokų skaičius (Number of folds) nustatomas į 5 naudojant „CrossValidation FoldMaker", o kompo-

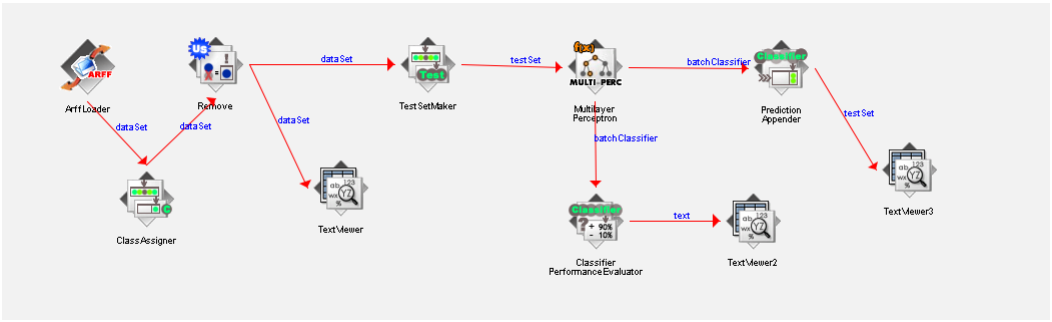
nentėje „MultiLayer Perceptron” paketo dydis (batchSize) pakeičiamas į 10. Šios sekos tikslas – optimaliai apmokyti neuroninį tinklą, kad jis galėtų tiksliai klasifikuoti duomenis.



3 pav. Daugiasluoksnio perceptrono apmokymo seka

1.2.2. Antra užduočių seka

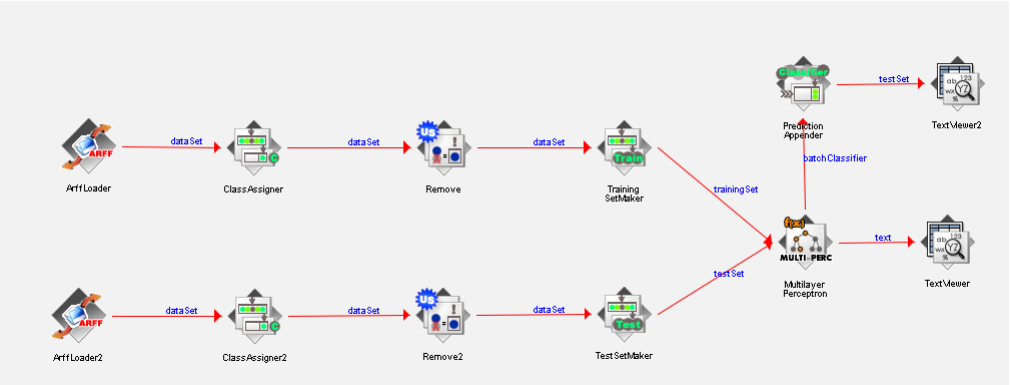
Antra „WEKA” užduočių seka skirta naujiems duomenims su nežinomomis klasėmis klasifikuoti, naudojant jau apmokytą ir išsaugotą modelį. Šioje sekoje įkeltas modelis priskiria klases duomenims iš failo, kuriame pateikti duomenys, neįtraukti į pradinį treniravimo rinkinį. Naudojant šią seką, nauji įrašai klasifikuojami remiantis ankstesnio apmokymo metu įgytomis žiniomis, leidžiant tinklui priskirti klases ir pateikti tikimybes, su kuriomis kiekvienas įrašas priklauso atitinkamai klasei.



4 pav. Naujų duomenų klasifikavimo seka

1.2.3. Trečia užduočių seka

Trečia „WEKA” užduočių seka naudojama tiek treniravimo, tiek testavimo duomenims klasifikuoti ir vertinti tinklo tikslumą. Šioje sekoje treniravimo duomenys ir testavimo duomenys leidžia įvertinti tinklo gebėjimą tiksliai klasifikuoti ne tik mokymosi metu matytus, bet ir nematytus duomenis. Nustatomas vieno paslėpto sluoksnio neuronų skaičius į 6, o komponentėje „PredictionAppender” parinktis „Append Probability” nustatoma į „True“, kad būtų rodomos ne tik priskirtos klasės, bet ir tikimybės. Ši seka padeda patikrinti, kaip gerai modelis priskiria klases ir kokiu tikslumu tai daro.



5 pav. Duomenų klasifikavimo ir testavimo seka

2. Tyrimas

2.1. Eksperimentai

Naudojant pirmą užduočių seką buvo atlikti eksperimentai geriausių mokymo parametrų radimui. Kintantys parametrai - paslėptų neuronų skaičius (toliau „neuronų sk.") (per kablelį atskiriami sluoksniai), mokymosi greitis bei „momentum", o optimizuojama - tikslumo reikšmė. Eksperimentų rezultatai pateikiami lentelėje.

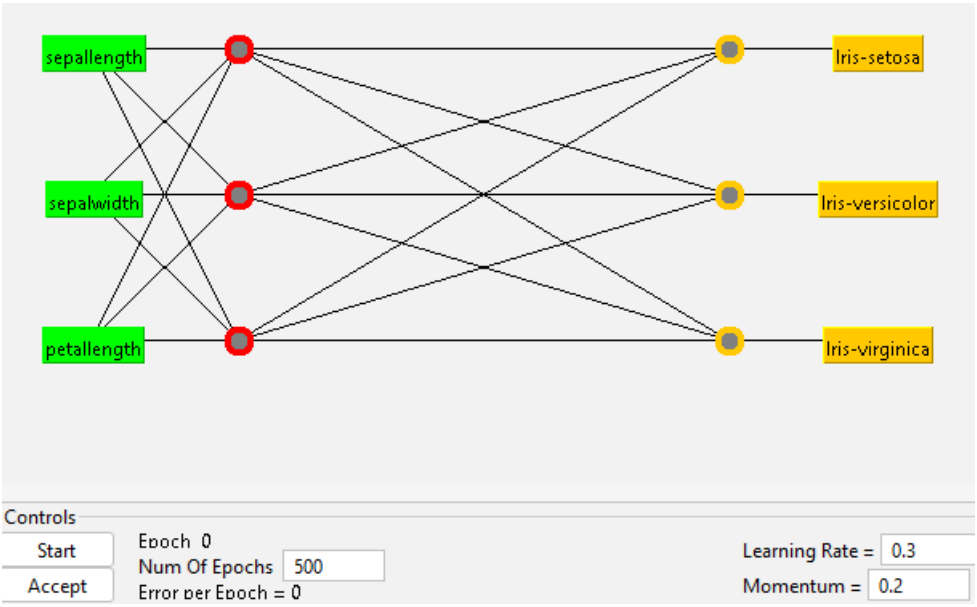
1 lentelė. Mokymosi parametrų ir klasifikavimo tikslumo lentelė

Neuronų sk.	Mokymosi greitis	Momentum	Tikslumas
3	0,3	0,3	94,1667%
3	0,2	0,3	94,1667%
3	0,1	0,3	94,1667%
3	0,01	0,3	88,3333%
3	0,3	0,2	95%
3	0,3	0,1	94,1667%
6	0,3	0,2	95%
9	0,3	0,2	95%
6, 3	0,3	0,2	93,3333%

2.2. Eksperimentų rezultatai

Atlikus tyrimus nustatyti optimalūs parametrai - 3; 0,3; 0,2 (atitinkamai - paslėptų neuronų skaičius; mokymosi greitis; momentum), kadangi su šiais parametrais modelis išnaudoja mažiausiai resursų, bei didesnis mokymosi greitis reiškia, kad mokymasis trunka trumpiau, lyginant su mažesniais dydžiais.

Žemiau pateikiamas neuroninio tinklo vaizdas su šiais parametrais.



6 pav. Neuroninio tinklo vaizdas

Lentelėse pateikiama informacija apie apmokyto modelio klasifikavimo rezultatus testavimo duomenų aibei, bei klasifikavimo tikslumo metrikas.

2 lentelė. Klasifikavimo rezultatai testavimo duomenų aibei

Tikroji klasė	Tikimybė, kad setosa	Tikimybė, kad versicolor	Tikimybė, kad virginica
setosa	0,987765	0,012183	0,000052
setosa	0,966651	0,033278	0,000071
setosa	0,987762	0,012186	0,000052
setosa	0,986613	0,013333	0,000054
setosa	0,986586	0,01336	0,000054
setosa	0,984388	0,015556	0,000056
setosa	0,987752	0,012196	0,000052
setosa	0,987017	0,01293	0,000053
setosa	0,987333	0,012614	0,000053
setosa	0,986405	0,013541	0,000054
versicolor	0,008121	0,937676	0,054203
versicolor	0,010328	0,965375	0,024296
versicolor	0,020745	0,977084	0,002171
versicolor	0,030384	0,968414	0,001201
versicolor	0,015665	0,97821	0,006125
versicolor	0,01863	0,977512	0,003858
versicolor	0,017812	0,977989	0,004199
versicolor	0,018664	0,978277	0,003059
versicolor	0,053061	0,946094	0,000845
versicolor	0,019575	0,977483	0,002942
virginica	0,000578	0,009329	0,990094
virginica	0,00504	0,784821	0,210139
virginica	0,000762	0,019451	0,979787
virginica	0,000409	0,003741	0,99585
virginica	0,000509	0,006663	0,992828
virginica	0,002089	0,238902	0,759009
virginica	0,001888	0,1941	0,804011
virginica	0,001532	0,117984	0,880484
virginica	0,000677	0,013867	0,985455
virginica	0,000992	0,03839	0,960618

3 lentelė. Klasifikavimo tikslumo metrikos testavimo duomenų aibei

TP	FP	Precizija	Atkūrimas	F1	Klasė
1	0	1	1	1	Iris-setosa
1	0,05	0,909	1	0,952	Iris-versicolor
0,9	0	1	0,9	0,947	Iris-virginica
0,967	0,017	0,97	0,967	0,967	Bendras vidurkis

2.3. „WEKA" ir „Excel" palyginimas

Trečios sekos sukurtas neuroninis tinklas buvo atkartotas ir „Excel" programoje, tam, kad būtų galima palyginti gautas klasifikavimo reikšmes tarp šių programų. Neuroninio tinklo svoriai buvo gauti apmokant modelį „WEKA" programoje ir po to šie svoriai buvo perkelti į „Excel" lenteles. Žemiau lentelėse pateikiami šie svoriai (ketvirta lentelė atitinka duomenų požymių įeičių sujungimų su paslėptų neuronų sluoksniu svorius, o penkta - paslėpto neuronų sluoksnio sujungimo su išeitimis svorius). Lentelėse svoriai suapvalinti, kad tilptų į dokumentą.

4 lentelė. Duomenų požymių svoriai

Įeina į	Paslėptą 1	Paslėptą 2	Paslėptą 3	Paslėptą 4	Paslėptą 5	Paslėptą 6
Poslinkis	-2,5795627	1,47802067	2,7763900	-2,7917824	0,4024159	-0,0763570
sepallength	-1,5184957	0,5399079	2,1486947	-1,6146352	0,8767623	-0,2071490
sepalwidth	-0,7902716	-2,0475960	0,5116013	2,7793697	-0,454892	0,1728768
petallength	8,6367778	4,141785	-9,6405367	-5,7469745	-2,4121471	0,7274917

5 lentelė. Paslėpto sluoksnio svoriai

Įeina į	Klasė setosa	Klasė versica	Klasė viginica
Poslinkis	-1,188644205	-1,513625126	-0,947152597
Paslėptas 1	-2,401757015	-6,436348722	4,870132506
Paslėptas 2	-4,534662128	2,900561005	3,011547873
Paslėptas 3	2,089272908	5,567931323	-7,398460406
Paslėptas 4	3,906902512	-8,039370152	-3,127113919
Paslėptas 5	0,226386384	0,149869911	-2,495367813
Paslėptas 6	-1,054413304	-0,935548402	0,266700405

Po to, modelio nematyti duomenys buvo atkartoti ir Excel programoje. Šie duomenys buvo normalizuoti tokiu pat principu kaip ir „WEKA" programoje, t. y. , reikšmės priskiriamos intervalui [-1; 1] naudojantis formule:

$$x_{ij} \leftarrow \frac{2x_{ij} - \min(x_{1j}, x_{2j}, \ldots, x_{mj}) - \max(x_{1j}, x_{2j}, \ldots, x_{mj})}{\max(x_{1j}, x_{2j}, \ldots, x_{mj}) - \min(x_{1j}, x_{2j}, \ldots, x_{mj})}$$

Vėliau paskaičiuotos duomenų įėjimo reikšmių ir svorių sandaugos ir gautiems rezultatams pritaikyta sigmoidinė funkcija. Šios sandaugos atitiko įeities reikšmes į kiekvieną paslėptą neuroną. Toliau paimtos reikšmės, gautos pritaikius sigmoidinę funkcija, kaip paslėptų neuronų įeities reikšmės ir sudaugintos su paslėptų neuronų svorių reikšmėmis. Šiam rezultatui, taip pat, pritaikyta sigmodinė funkcija, po kurios, gauti rezultatai atvaizdavo trijų skirtingų klasių pasirinkimo tikimybę. Šios tikimybės pateikiamos 6 lentelėje, o 7 lentelėje pateikiamos tikimybės gautos „WEKA" programoje. **„Excel" programoje įgyvendintas neuroninis tinklas prisegamas kaip papildomas failas prie darbo pavadinimu „neuralNetwork.xlsx".**

6 lentelė. „Excel" gautos tikimybės

Indeksas	Tikroji klasė	Tikimybė, kad setosa	Tikimybė, kad versicolor	Tikimybė, kad virginica
1	setosa	0,989523381	0,016189309	1,31938E-06
2	setosa	0,949819868	0,105229696	3,94182E-06
3	setosa	0,988770561	0,01675549	1,4656E-06
4	setosa	0,988462113	0,016873055	1,51744E-06
5	setosa	0,988230759	0,015705247	1,74789E-06
6	setosa	0,986112222	0,022237339	1,6007E-06
7	setosa	0,989208633	0,015320986	1,48813E-06
8	setosa	0,988213514	0,017677263	1,5029E-06
9	setosa	0,989353698	0,01600159	1,36544E-06
10	setosa	0,988491485	0,017960115	1,40744E-06
11	versicolor	0,000621194	0,07172942	0,948496065
12	versicolor	0,000951504	0,175821166	0,858024474
13	versicolor	0,00574948	0,953555025	0,034406219
14	versicolor	0,017632152	0,994329445	0,002455212
15	versicolor	0,001921779	0,544178842	0,465176911
16	versicolor	0,00352756	0,774385238	0,196952075
17	versicolor	0,00299529	0,740499111	0,241849824
18	versicolor	0,004015378	0,889843823	0,092532082
19	versicolor	0,045972891	0,995418474	0,000575852
20	versicolor	0,004177576	0,877874379	0,102272741
21	virginica	0,000191296	0,005041119	0,997777636
22	virginica	0,000461344	0,040089541	0,972110731
23	virginica	0,00020696	0,005756017	0,997327327
24	virginica	0,000171198	0,004102823	0,998403625
25	virginica	0,000183663	0,004628123	0,998113106
26	virginica	0,000290401	0,013056851	0,992433264
27	virginica	0,000273127	0,010932038	0,993074802
28	virginica	0,000258032	0,009738439	0,994825397
29	virginica	0,000206095	0,005265282	0,997785944
30	virginica	0,000225577	0,006743822	0,996892902

7 lentelė. „WEKA" gautos tikimybės

Indeksas	Tikroji klasė	Tikimybė, kad setosa	Tikimybė, kad versica	Tikimybė, kad virginica
1	setosa	0,980875	0,019124	0,000002
2	setosa	0,910571	0,089425	0,000004
3	setosa	0,979709	0,02029	0,000002
4	setosa	0,977942	0,022057	0,000002
5	setosa	0,978909	0,021089	0,000002
6	setosa	0,96988	0,030118	0,000002
7	setosa	0,981583	0,018415	0,000002
8	setosa	0,977508	0,02249	0,000002
9	setosa	0,980469	0,01953	0,000002
10	setosa	0,976841	0,023158	0,000002
11	versicolor	0,003598	0,813407	0,182995
12	versicolor	0,00467	0,889972	0,105358
13	versicolor	0,015096	0,980599	0,004305
14	versicolor	0,037135	0,962006	0,000858
15	versicolor	0,008796	0,969584	0,021619
16	versicolor	0,01347	0,975466	0,011064
17	versicolor	0,011808	0,975384	0,012807
18	versicolor	0,011736	0,979672	0,008592
19	versicolor	0,093161	0,90657	0,000269
20	versicolor	0,013937	0,978709	0,007354
21	virginica	0,000345	0,018304	0,981351
22	virginica	0,001757	0,503315	0,494928
23	virginica	0,000467	0,034932	0,964601
24	virginica	0,000244	0,008244	0,991511
25	virginica	0,000313	0,013875	0,985812
26	virginica	0,0009	0,167714	0,831387
27	virginica	0,000844	0,154337	0,844819
28	virginica	0,000742	0,106465	0,892793
29	virginica	0,000448	0,024983	0,974569
30	virginica	0,000598	0,054397	0,945005

Išvados

1. Pirmoje sekoje apmokytas modelis gerai atpažįsta pateiktas klases, tačiau skirtingų klasių tikslumo rodikliai skiriasi. setosa klasė buvo klasifikuojama itin tiksliai – tiek precizija, tiek atkūrimo rodiklis, tiek F1 reikšmė siekia 1, o tai reiškia, kad modelis šią klasę identifikuoja be klaidų. Versicolor klasė taip pat buvo klasifikuota gerai su pakankamai aukštomis metrikomis. Virginica klasifikavimas yra kiek mažiau tikslus žiūrint į metrikas, tai reiškia, kad modelis kartais netiksliai priskiria šiai klasei priklausančius pavyzdžius. Nepaisant to, bendras modelio tikslumas yra aukštas, su vidutine precizija ir atkūrimo rodikliu apie 0,967. Tai rodo, kad modelis yra patikimas ir gerai pritaikytas pateiktų duomenų klasifikavimui.
2. „Excel" ir „WEKA" gautos tikimybės yra panašios, tačiau tarp jų pastebimi tam tikri skirtumai. Šie skirtumai gali atsirasti dėl kelių priežasčių, įskaitant skirtingą tikslumo lygį ir apvalinimo taisykles, kuriuos naudoja abi platformos. „WEKA" dažnai naudoja didesnį skaičiavimo tikslumą nei „Excel aplinka". Be to, „WEKA" automatiškai normalizuoja duomenis, kad jie atitiktų tam tikrą intervalą, o „Excel" duomenys buvo normalizuoti rankiniu būdu, kas taip pat gali lemti skirtumus. Apibendrinant, nors „Excel" ir „WEKA" gauti rezultatai yra panašūs, skirtingi tikslumo lygiai, galimi normavimo skirtumai lemia šiuos nedidelius neatitikimus.