Otrais praktiskais darbs: Mašīnmācīšanās algoritmu lietojums

Darba autors:
Aleksejs Tihomirovs 211RDB173
7 gr. Automātika un datortehnika
https://github.com/werdillo/AI_pr2

Rīga, 2023

SATURA RADĪTĀJS

1. DATU PIRMAPSTRĀDE/IZPĒTE

Cukura diabēts ir nopietna un plaši izplatīta problēma visā pasaulē. Saskaņā ar Pasaules Veselības organizācijas datiem pašlaik pasaulē ar cukura diabētu slimo 422 miljoni cilvēku, kas ir aptuveni 8,5 % iedzīvotāju (1). Cukura diabēta problēma ir saistīta arī ar augstu mirstību un lielākām veselības aprūpes izmaksām.

1.1. Datu kopa

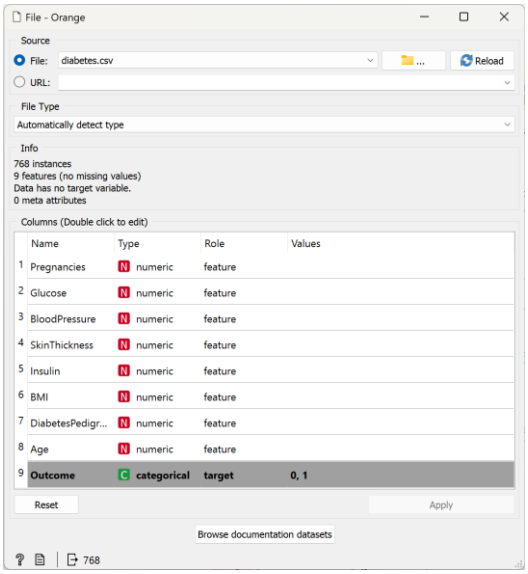
Tika izvēlēta datu bāze "Diabetes Dataset". Šo datu kopu sākotnēji sagatavoja "National Institute of Diabetes and Digestive and Kidney Diseases" un publicē Mehmet Akturk. Materiāls tika izplatīts ar "CC0: Public Domain" licenci un ir pieejams ņemts no kaggle mājas lapā: https://www.kaggle.com/datasets/mathchi/diabetes-data-set. Datu bāze ir viena csv datne, kuru darba autors ir lejupielādējis tālākai analīzei(2).

Autora aprkasts:

Šo gadījumu atlasei no lielākas datubāzes tika piemēroti vairāki ierobežojumi. Jo īpaši visi pacienti ir vismaz 21 gadu vecas sievietes, kas ir Pima indiāņu izcelsmes.

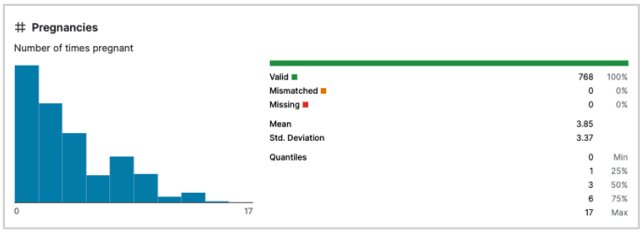
1.2. Datu kopas saturs

Darba autors lejupielādēja un instalēja Orange Data Mining programmatūru un augšupielādēja tajā CSV failu. Programmatūrā ir faila konfigurācijas logs, kurā var apskatīt datubāzes struktūru un īpašības (skat. 1.1 att.).



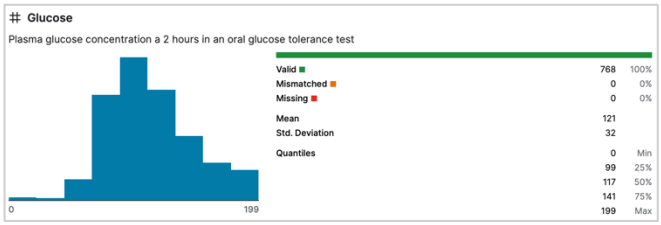
1.1. att. datu bāzes atskaite Orange Data mining programmatūrā

Kopā šajā datubāzē ir 768 objekti, to var redzēt, apskatot Info blokā 1.1 attēlā. Katram objektam ir 9 kritēriji. Datu bāzē ir tikai viena klases kritērijs "Outcome", tas nozīmē ka visi 768 objekti ir klasificēti saskaņā ar šo kritēriju. "Outcome" klases kritērijam ir divas vērtības, tas norāda, vai objektam ir diabēts. Vērtība 1 norāda, ka objetkam ir diabēts, 0 norāda, ka diabēta nav. Pēc darba autora domām, klases kritēriju nosaukums būtu informatīvs, ar nosaukumu "Diabēts", tomēr sākotnējā datubāze nav mainīta, un nosaukums "Outcome" paliek nemainīgs. Visiem atribūtiem, ir skaitliskas vērtības un nav trūkstošo vērtību, tāpēc dati nav jākonvertē. Ņemot vērā ja tas būtu nepieciešams, orange data minig ir atbilstoša funkcionalitāte, kas ļauj pārvērst neskaitliskos datus skaitliskos datos, aizstāt trūkstošās vērtības un izmantot ierobēžotus kritērijus turpmākajiem aprēķiniem.



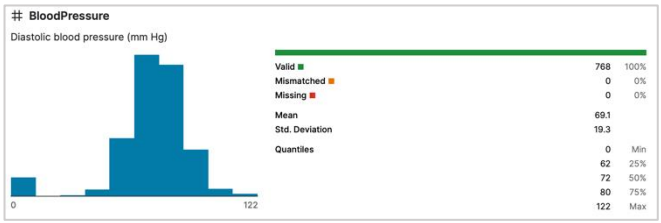
1.2. att. Grūtniecības kritērijs

Grūtniecības kritērijs raksturo sievietes grūtniecību skaitu no 0 līdz 17 reizes (skat. 1.2 att.).



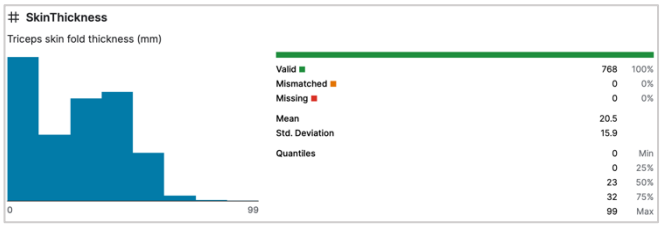
1.3. att. Glikozes kritērijs

Glikozes kritērijs atspoguļo glikozes koncentrācija plazmā 2 stundu laikā, veicot perorālo glikozes tolerances testu. Iegūti dati ir diapazonā no 0 līdz 199 (skat. 1.3 att.).



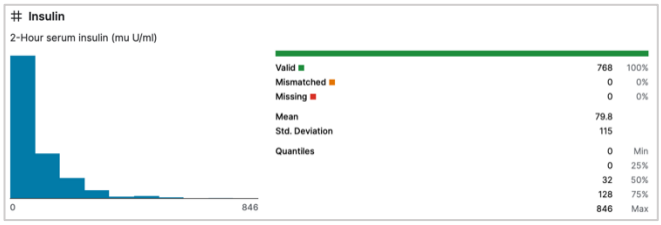
1.4. att. Asinsspiediena kritērijs

Asinsspiediena kritērijs rāda diastoliskais asinsspiedienu mm Hg diapazonā no 0 līdz 122 (skat. 1.4 att.).



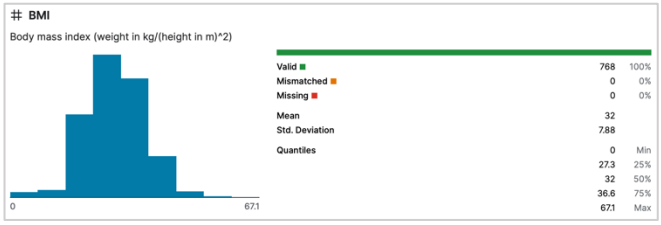
1.5. att. Ādas biezuma kritērijs

Ādas biezuma kritērijs rāda tricepsa ādas krokas biezumu milimetros. Mērījumu diapazons ir no o līdz 99 mm (skat. 1.5 att.).



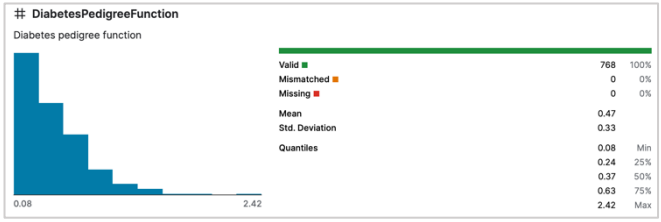
1.6. att. Insulīna kritērijs

Insulīna kritērijs rāda 2 stundu seruma insulīns (mu U/ml) diapazonā no 0 līdz 846 (skat. 1.6 att.).



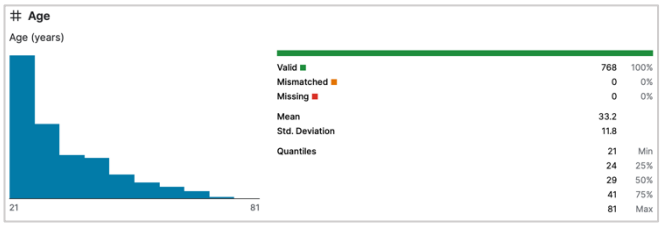
1.7. att. ĶMI jeb ķermeņa masas indekss kritērijs

ĶMI kritērijs ir ķermeņa masas indekss, ko aprēķina pēc formulas (svars kg/(augums m)^2) Šajā datubāzē ĶMI diapazons ir no 0 līdz 67.1 (skat. 1.7 att.).



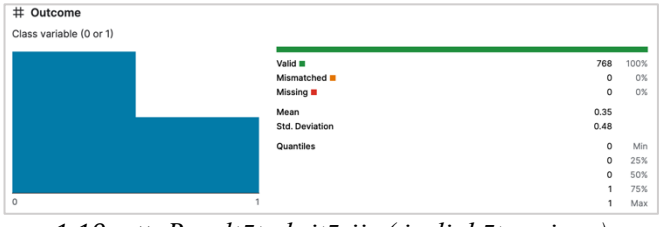
1.8. att. Diabēta ciltsgrāmatas funkcijas kritērijs

Diabēta ciltsdarba funkcijas kritērijs recenzentiem ir diapazonā no 0.08 līdz 2.42 (skat. 1.8 att.).



1.9. att. Vecuma kritērijs

Vecums ir no 21 līdz 81 gadiem (skat. 1.9 att.).



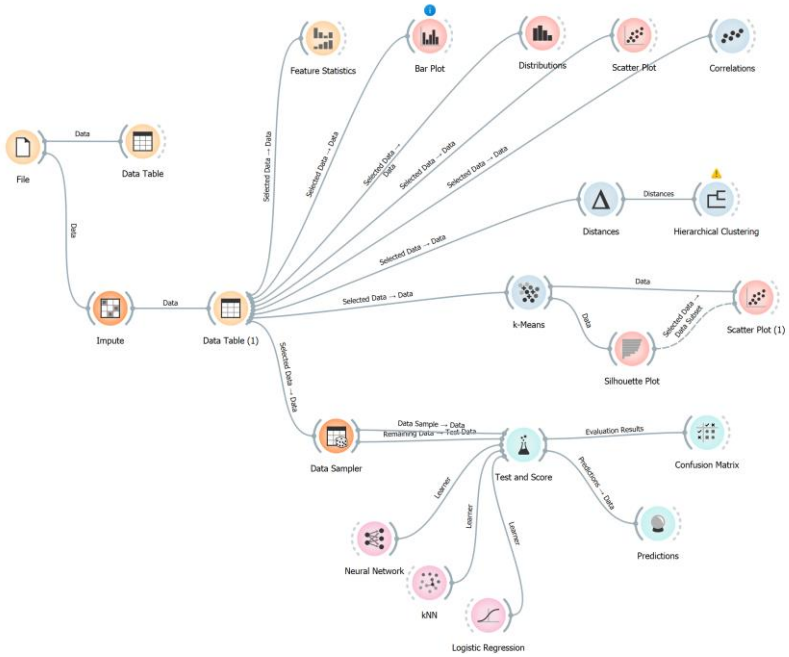
1.10. att. Rezultāta kritērijs (ir diabēts vai ne)

Rezultāta kritērijs rāda vai objektam ir diabēts vai ne (skat. 1.10 att.).

Lai parādītu csv datus, datubāzes fails tika atvērts programmā Excel (skat. 1.11 att.).

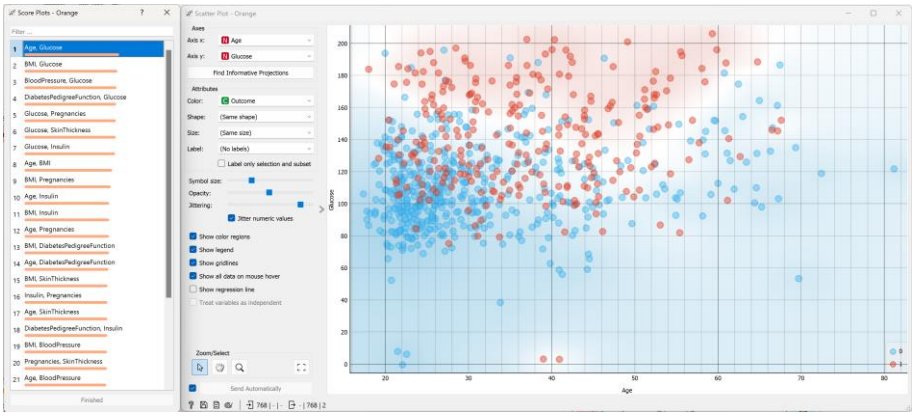
	A	B	C	D	E	F	G	H	I	J
1	Pregnancie	Glucose	BloodPress	SkinThickne	Insulin	BMI	DiabetesPe	Age	Outcome	
2	6	148	72	35	0	33.6	0.627	50	1	
3	1	85	66	29	0	26.6	0.351	31	0	
4	8	183	64	0	0	23.3	0.672	32	1	
5	1	89	66	23	94	28.1	0.167	21	0	
6	0	137	40	35	168	43.1	2.288	33	1	
7	5	116	74	0	0	25.6	0.201	30	0	
8	3	78	50	32	88	31	0.248	26	1	
9	10	115	0	0	0	35.3	0.134	29	0	
10	2	197	70	45	543	30.5	0.158	53	1	
11	8	125	96	0	0	0	0.232	54	1	
12	4	110	92	0	0	37.6	0.191	30	0	
13	10	168	74	0	0	38	0.537	34	1	
14	10	139	80	0	0	27.1	1.441	57	0	
15	1	189	60	23	846	30.1	0.398	59	1	
16	5	166	72	19	175	25.8	0.587	51	1	
17	7	100	0	0	0	30	0.484	32	1	
18	0	118	84	47	230	45.8	0.551	31	1	

1.11. att. datubāze atverta excel



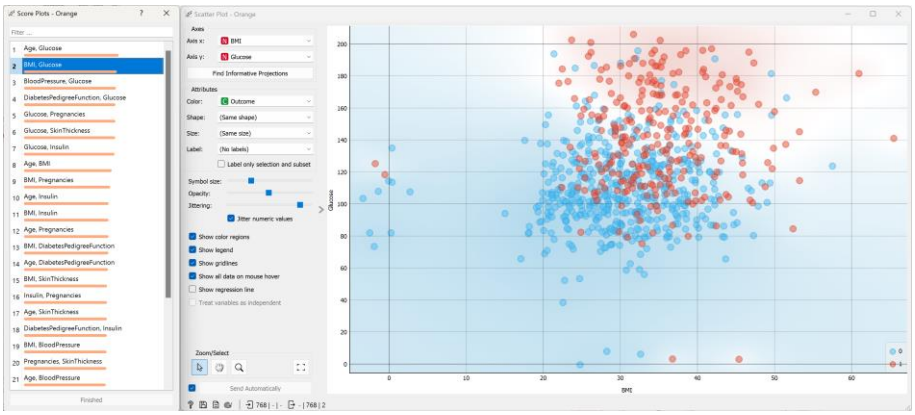
1.12. att.orange data mining

1.3. Datu kopas grafiki



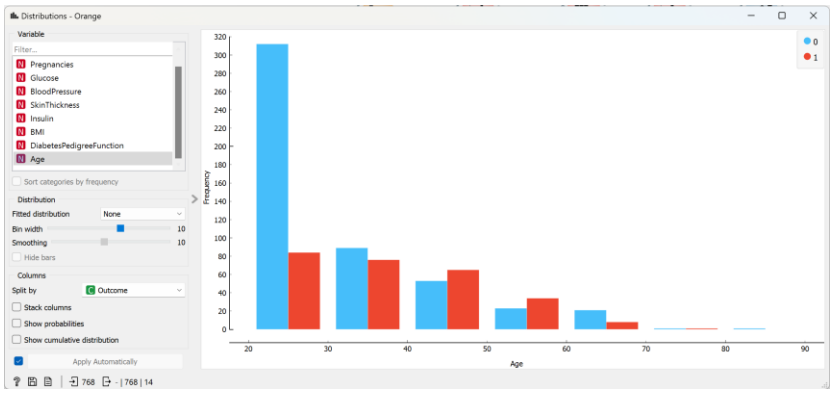
1.13. att. Vecuma un Glukozes Scatter plot grafiks

Izmantojot "Find Information Projection" pogu tika iegūts izkliedes grafiks ar vecuma vērtībām uz X ass un glikozes vērtībām uz Y ass (skat. 1.13 att.). Kā var redzat, dati ir diezgan slikti atdalīti viens no otra. Tuvāk aplūkojot grafiku, redzams, ka objekti ar vērtību 0 (bez diabeta) ir vairāk sagrupēti, bet objekti ar vērtību 1 ir ir mazāk pārpildīti. Objekti ar vērtību 1 (ar diabetu), vairāk atrodas diagrammas augšpusē, kas nozīmē, ka glikozes daudzums virs 120-160 var liecināt par lielu diabēta risku.



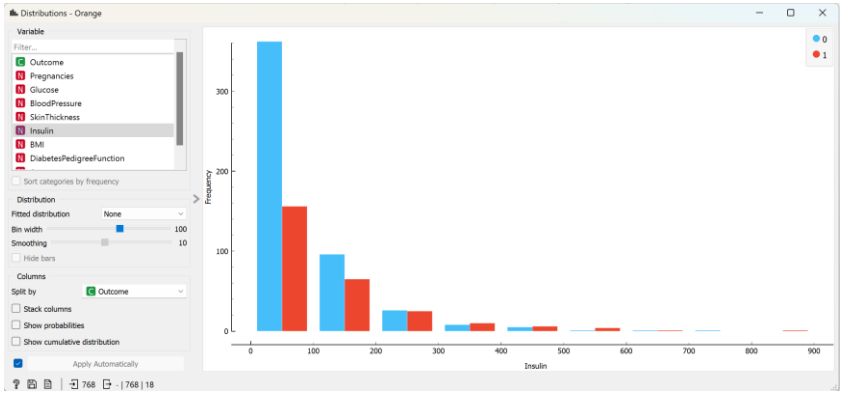
1.14. att. ĶMI un Glukozes korelācija

Ar "Find Information Projection" pogu tika iegūts izkliedes grafiks ar ĶMI vērtībām uz X ass un Glukozes vērtībām uz Y ass (skat. 1.14 att.). Kā var redzat, dati vēl mazāk atdalīti salīdzinājumā ar 1.13 att. grafiku, objekti ar vērtību 1 nav tik cieši izvietotas pēc 160 glikozes Y ass vērtības.



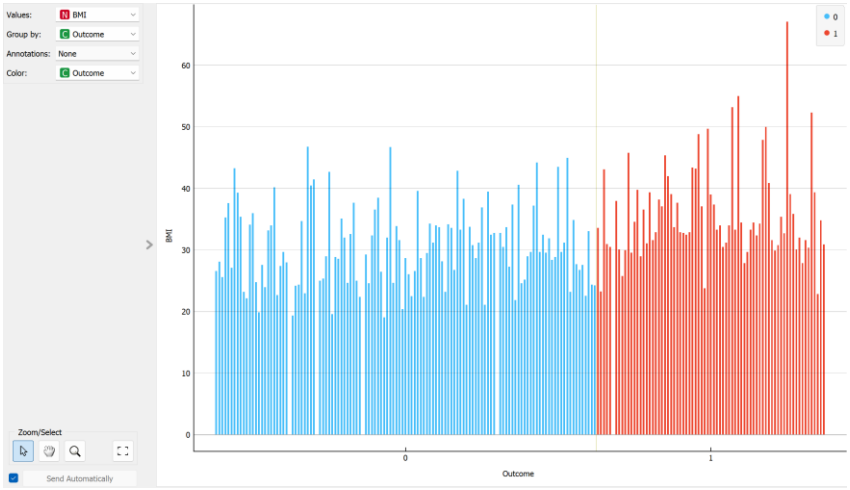
1.15. att. Vecuma histagrama

Ar "Distribution" orange programmā ir izstrādāta vecuma histogramma (skat. 1.15. att.). Varat redzēt objekti ar vērtību 0 dominē pār objektiem ar vērtību 1, jo objektu skaits ar vērtību 1 ir mazāk nekā objektu ar vērtību 0. No histogrammas var arī secināt, jo jaunāks ir objekts, jo mazāka ir iespēja, ka viņam ir diabēts. Tas ir īpaši redzams vecumā no 20 līdz 30 gadiem, kur objektu skaits bez diabēta ir 78,19%, bet objektu ar diabētu ir 21,21%. Var novērot tendenci, jo lielāks vecums, jo mazāka veselu cilvēku attieksme pret slimajiem ar diabetu.



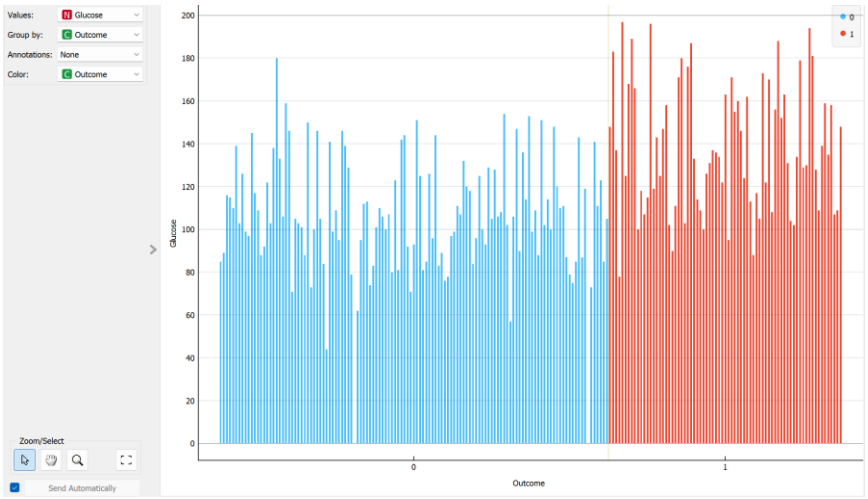
1.16. att. Insulina histograma

Skatoties uz insulina histogramu (skat. 1.16 att.), var arī secināt ka objekti ar vērtību 0 īpaši dominē diapazonā no 0 līdz 100 kopumā 67,45 % no visiem objektiem. Diapazonā no 100 līdz 200 kopumā ir 20,96 % no visiem objektiem. Diapazonā no 200 līdz 300 – 6,64%. Cilvēki, kuriem insulīna līmenis ir lielāks par 300 attiecība starp kopējo skaitu ir 4,95%. Ir tendence jo vairāk insulīna, jo lielāka varbūtība, ka objektam ir diabēts.



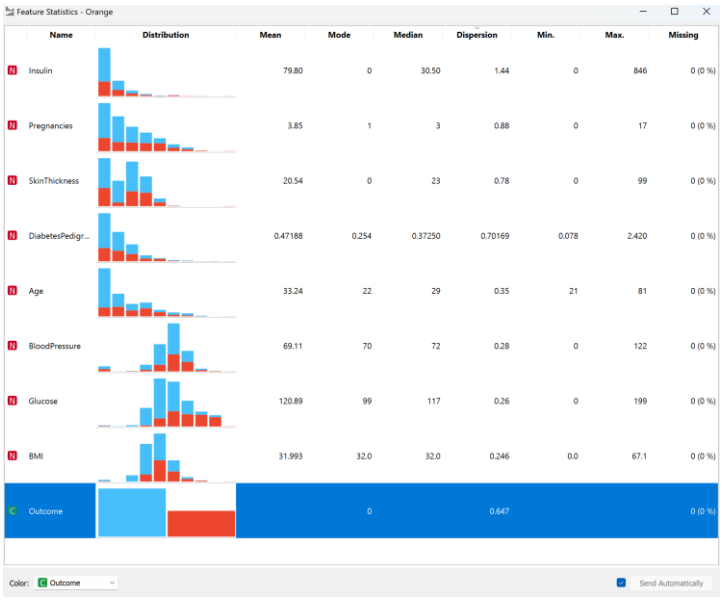
1.17. att. ĶMI sadalījums

Tika analizēts KMI sadalījums, un redzams, ka personām ar cukura diabētu ir augstāks KMI nekā personām bez cukura diabēta ar vērtību 0 (skat. 1.17 att.). Augsts ķermeņa masas indekss var netieši norādīt uz personas zemo fizisko aktivitāti, kas palielina diabēta iespējamību.

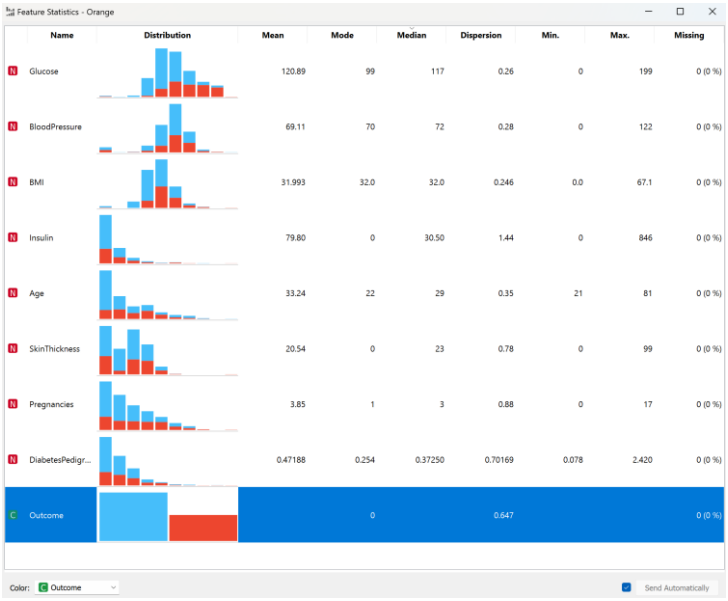


1.18. att. Glukozes sadalījums

Aplukojot Glukozes sadalījumu, ir redzams augstāks Glukozes līmenis personām ar diabetu un vertību 1 (skat. 1.18 att.). Augstais glukozes līmenis var būt saistīts ar diabēta slimības attīstību un var ietekmēt pacienta veselību un labklājību.



1.19. att. Dispersija



1.20. att. Mediana

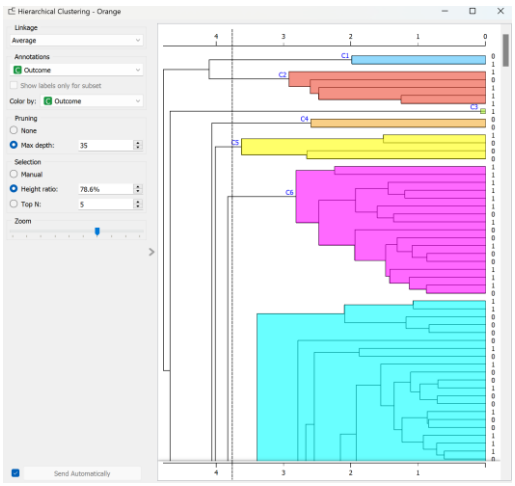
Secinājumi

Pēc datu izpētes un apstrādes var secināt, ka datu kopā dominē objekti ar vertību 0, kas nozīme datu kopā ir vairāk objektus bez diabetal (skat. 1.15, 1.16 att.). No 768 objektiem 500 nav diabēta un 268 diabēts ir. Objekti ir slikti atdalīti viens no otra, tāpēc ir grūti redzēt datu struktūru. Divas grupas var identificēt ar skaitliskajām vērtībām 0 un 1, 1 – diabets ir, 0 – diabeta nav. Abas grupas ir vidēji vizuāli atdalīti, bet atrodas tuvu viens otram ar pārklāšanās zona, kurā objekti ar dažādām vērtībām stipri pārklājas (skat. 1.13, 1.14 att.).

Grūtniecības mediānas vērtība ir 3, un mediānas vecums ir 29 gadi, tas raksturo datu izlasi (skat. 1.20 att.). Pēc darba autora domām, secinājumi nevar būt universāli visiem diabēta riska grupas cilvēkiem. DiabetesPedigreeFunction atribūta dispersija ir lielāks par mediānas parametru, lielāka dispersija var norādīt uz lielāku nevienlīdzību, atšķirībām vai svārstībām datu kopā.

2. NEPĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS

2.1 Hieritical Clustering Aprkasts



2.1. att. Hierarchical clustering piemērs

Hieritical Clustering hiperparametri (3):
Linkage

- **Single linkage** aprēķina attālumu starp abu kopu tuvākajiem elementiem
- **Average linkage** aprēķina vidējo attālumu divu klasteru elementiem
- **Weighted linkage** tiek izmantota WPGMA metode
- **Complete linkage** aprēķina attālumu starp klasteru visattālākajiem elementiem
- **Ward linkage** aprēķina kļūdas kvadrātu summas palielinājumu.

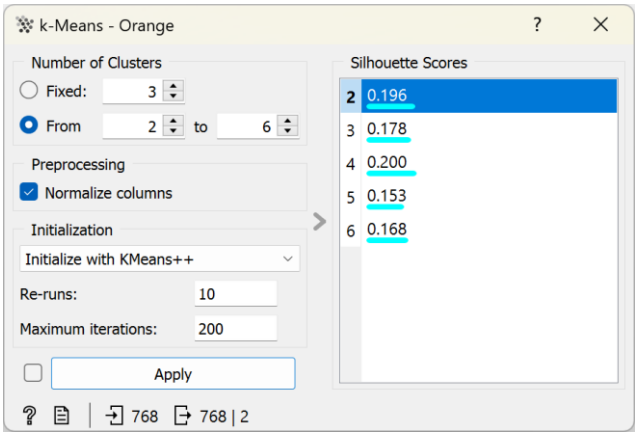
Annotation Dendrogrammas mezglu etiķetes var izvēlēties
Pruning:

None
maximum depth – Tas ietekmē tikai displeju, nevis faktisko sagrupēšanu.

Selection:

- **Manual** (noklikšķinot dendrogrammas iekšpusē, tiks atlasīts klasteris. Turot Ctrl/Cmd, var atlasīt vairākas kopas. Katrs izvēlētais klasteris tiek parādīts citā krāsā un izvadē tiek uzskatīts par atsevišķu klasteri.)
- **Height ratio** (Noklikšķinot uz dendrogrammas apakšējā vai augšējā lineāla, grafikā tiek parādīta robežlīnija. Tiek atlasīti vienumi pa labi no līnijas.)
- **Top N** (atlasa augšējo mezglu skaitu.)

2.2 k-Means Aprkasts



2.2. att. k-means piemērs

k-Means hiperparametri (4):

Fixed: algoritms grupē datus norādītajā klasteru skaitā.

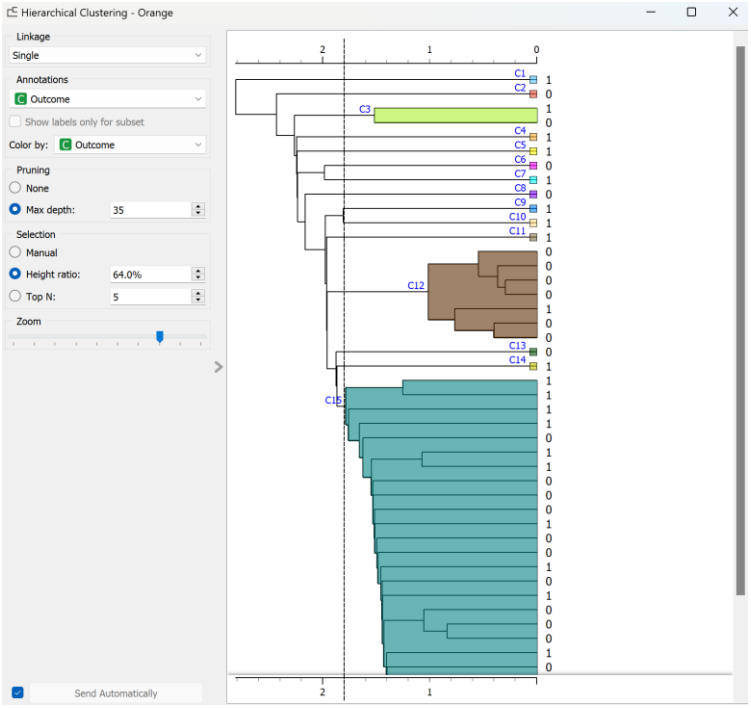
From X to Y: logrīks parāda klasterizācijas punktus atlasītajam klasteru diapazonam, izmantojot **Silhouette score** (Vidējais attālums līdz elementiem, kas atrodas tajā pašā klasterī, tiek salīdzināts ar vidējo attālumu līdz elementiem, kas atrodas citās kopās, lai iegūtu kontrastu starp tiem.).

Preprocessing: Ja tiek aktivizēta kolonnu normalizācijas opcija, tad kolonnu vidējā vērtība tiks nobīdīta uz nulli, un standarta novirze tiks pielāgota tā, lai tā būtu vienāda visās kolonnās un sasniegtu vērtību vienādu ar 1.

Initialization method (veids, kā algoritms sāk klasterizāciju):

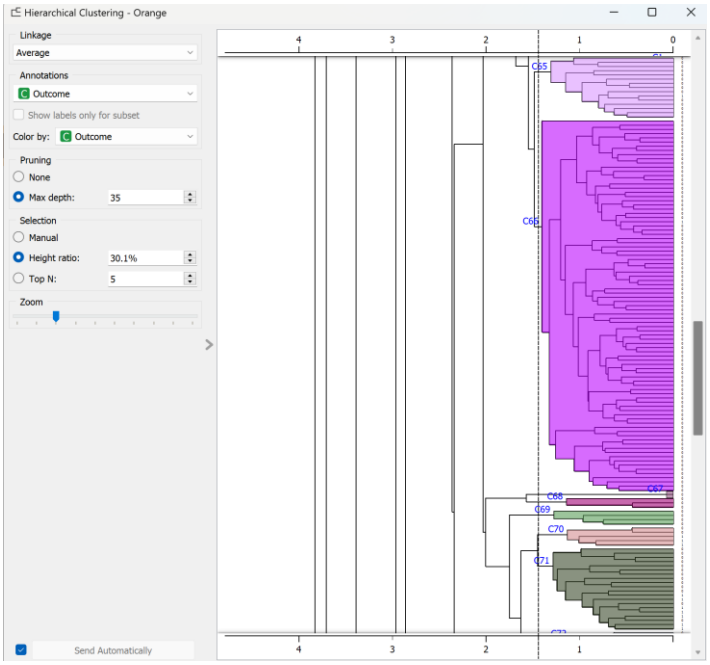
- **k-Means++** pirmais klastera centrs tiek izvēlēts nejauši. Nākamie klastera centri tiek izvēlēti no atlikušajiem punktiem, un to izvēle notiek ar varbūtību, kas proporcionāla kvadrātam attālumam no tuvākā klastera centra.
- **Random initialization** sākumā klasteriem tiek piešķirtas nejaušas vērtības, un pēc tam ar turpmākajām iterācijām tie tiek atjaunināti.
- **Re-runs** algoritms tiks palaists vairākas reizes no nejaušām sākotnējām pozīcijām, un tiks izmantots klasteru rezultāts ar zemāko kvadrātu summu. Maksimālais iterāciju skaits katrā algoritma izpildē var būt iestatīts manuāli.

2.3 Eksperimenti ar hierarhisko klasterizāciju.



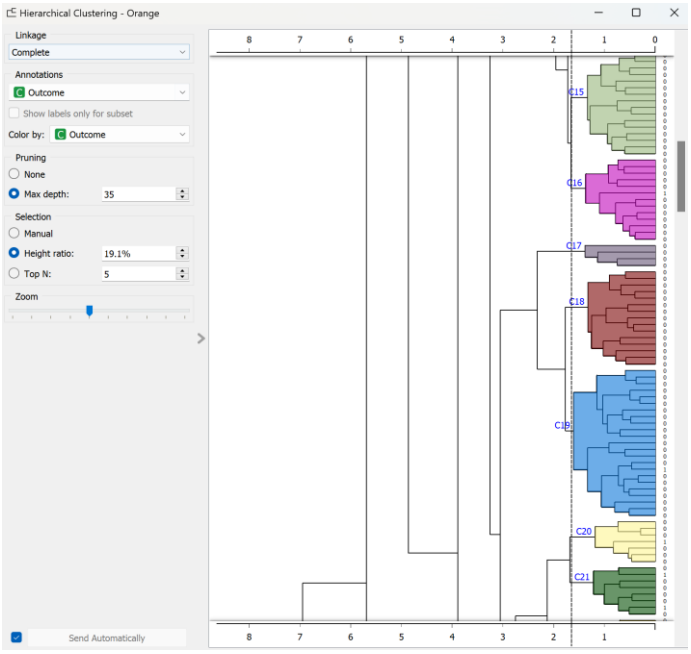
2.3. att. Hierarchical clustering. Linkage: Single.

Dendogrammai tika izvēlēti šādi parametri: Linkage – Single, max depth – 35, height ration: 64% (skat. 2.3. att.). Izmantojot šos parametrus, tiek veritas 15 klasteri. Klasteris C12 ir visveiksmīgākais, jo šajā klasterī dominē objekti ar vērtībām 0. Klasteris C15 ir vislielākais, bet klasterī nav viendabīgu datu. Klasterī C3 arī nav viendabīgu datu. Pārējiem klasteri ir mazi, tiem ir tikai viens parametrs.



2.4. att. Hierarchical clustering. Linkage: Avarage.

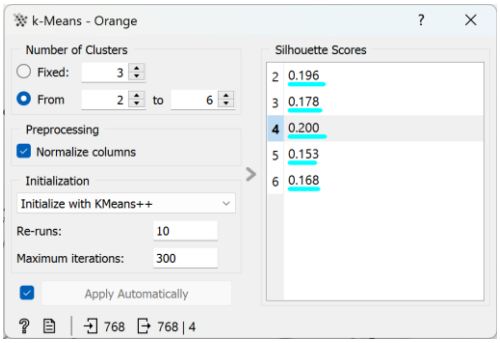
Dendogrammai tika izvēlēti šādi parametri: Linkage – Avarage, max depth – 35, height ration: 30.1% (skat. 2.4 att.). Izmantojot šos parametrus, tiek veritas 138 klasteri. Klasteri C66, C65, C66, C67, C68, C69, C71 satur viendabīgus datus galvenokārt ar vērtību 0. Klasteris C66 ir vislielākais, un 88 no 89 objektiem vērtība ir 0.



2.5. att. Hierarchical clustering. Linkage: Complete.

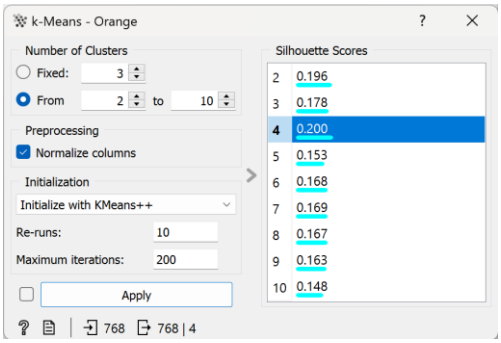
Dendogrammai tika izvēlēti šādi parametri: Linkage – Complete, max depth – 35, height ration: 19.1% (skat. 2.5 att.). Izmantojot šos parametrus, tiek ir saņemti 166 klasteri. Klasteriem C15 – C21 satur viendabīgus datus ar 0 vērtību, var secināt, ka 2.5. attēlā klasteri ir veiksmīgi atdalīti.

2.4 Eksperimenti ar k-means algoritmu.



2.6. att. k-means. Number of clusters 2 to 6

Algoritmam k-means tika atlasīti number of Cluster from 2 to 6 (skat 2.6. att.). Mainot klastera daudzumu, vislielākais rezultāts nemainījās un palika 0,2 (skat 2.7. att.). līmenī. Tas nozīmē, ka algoritms vislabāk darbojas ar 4. Klasteru.



2.7. att. k-means. Number of clusters 2 to 10

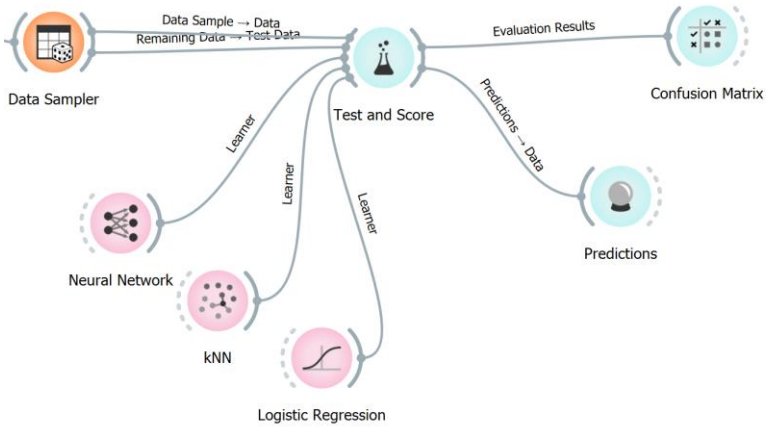
2.8. att. Silhouette plot

Izmantojot maksimālo parametru k-izlīdzināšanas metodi, iegūstam silueta diagrammu, kurā objekti, kas ir mazāk nekā 0 (pa kreisi) slikti klasterizēti, bet dati, kas ir lielāki par 0 (pa labi), ir labi klasterizēti (skat 2.8. att.). Diagrammā redzams, ka ir vairāk objektu, kas labi klasterizēti.

Secinājums

Pamatojoties uz iegūtajiem datiem, var secināt, ka k-means algoritms darbojas ar 20% efektivitāti, šis rādītājs ir salīdzinoši zems, jo tas ir mazāks par 50%, ko var iegūt, izmantojot citas datubāzes. Iegūti dati ir vidēji atdalami un klasterizēti, jo labi klasterizēto objektu skaits ir lielāks nekā neklasterizēto objektu skaits (skat. 2.8. att.).

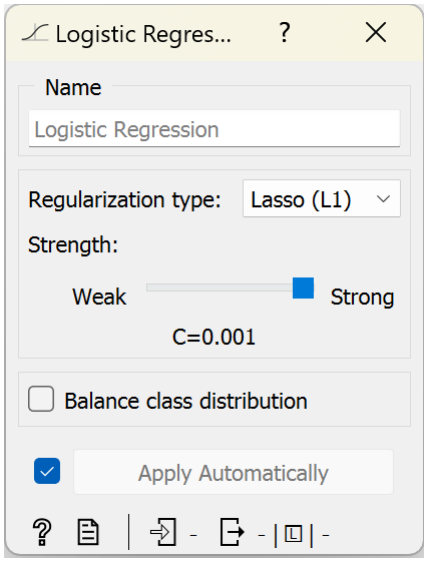
3. PĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS



3.1. att. Algoritmi Orange programmā

3.1 Logistic regression apraksts

Darba autors izmantoja loģistiskās regresijas algoritmu, jo tam ir vienkārša interpretācija, tas ir viegli piemērojams lielam pazīmju skaitam un to var izmantot, lai novērtētu piederības klases varbūtību (5).



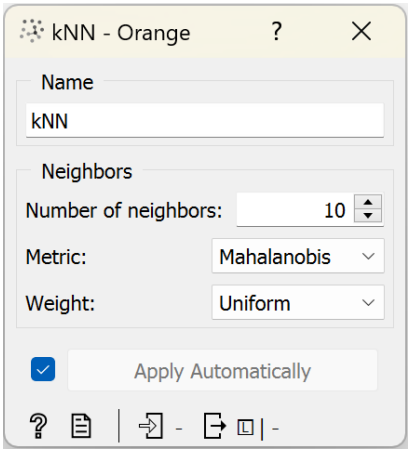
3.2. att. logistic regression algoritms

Regularization type – (L1 vai L2).

Strength: Iestatiet izmaksu stiprumu (noklusējums ir C = 1).

3.2 kNN algoritma apraksts

Darba autors izmanto KNN algoritmu jo tas ir vienkāršs, bet spēcīgs mašīnmācīšanās algoritms, ko autors studējis universitātē.Tā pamatā ir ideja, ka objektiem, kas ir tuvi pazīmju telpā, ir tendence piederēt tai pašai klasei vai tiem ir līdzīga mērķa mainīgā vērtība(6).



3.3. att. kNN algoritms

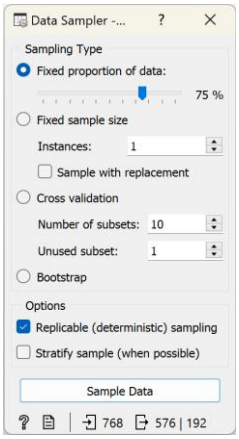
Number of neighbors- attāluma parametrs (metrika) un svari kā modeļa kritēriji.
Metric:

- **Euclidean** - "taisna līnija", attālums starp diviem punktiem
- **Manhattan** visu atribūtu absolūto atšķirību summa
- **Maximal** vislielākās absolūtās atšķirības starp atribūtiem
- **Mahalanobis** attālums starp punktu un sadalījumu.

Weight:

- Uniform- visi punkti katrā rajonā tiek svērti vienādi
- Distance- vaicājuma punkta tuvākajiem kaimiņiem ir lielāka ietekme nekā kaimiņiem tālāk.

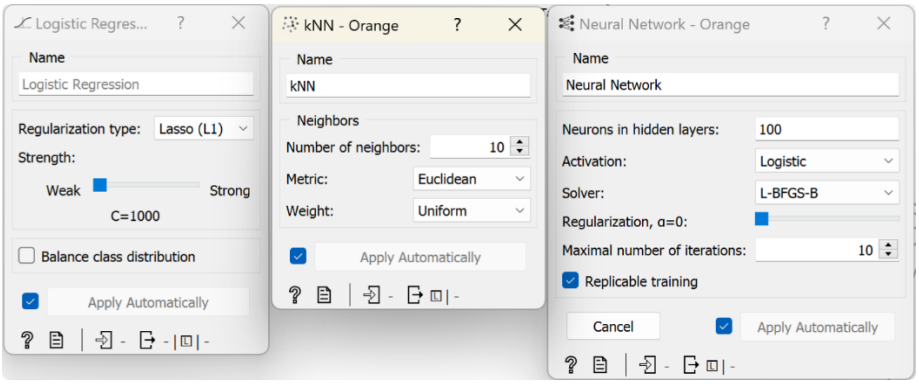
3.3 Algoritmu testēšana



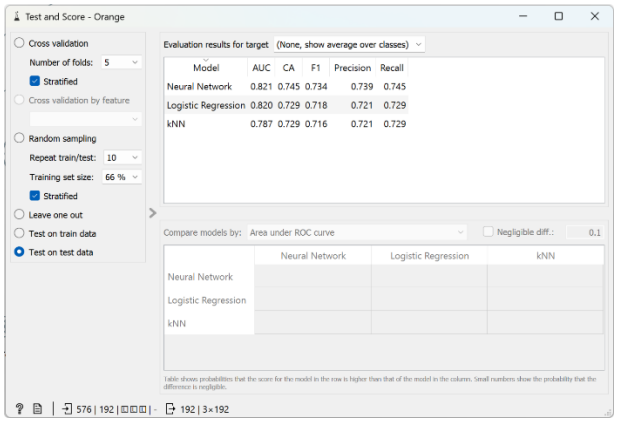
3.4. att. Data sampler

Lai pārbaudītu algoritmu, dati tika sadalīti divos veidos - testiem un algoritma darbībai. Izmantojot datu sampler, 75% tika izmantoti algoritma mācīšanai, bet 25% testiem (skat. 3.4. att.).

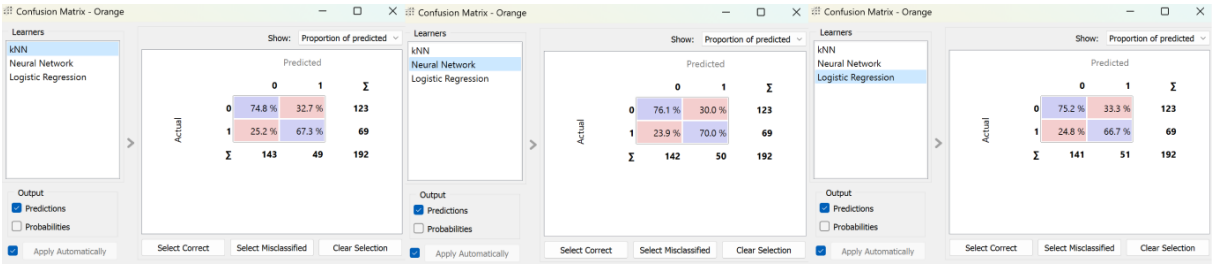
Tests 1



3.3. att. Testa 1 kNN, Logistic Regression, c parametri

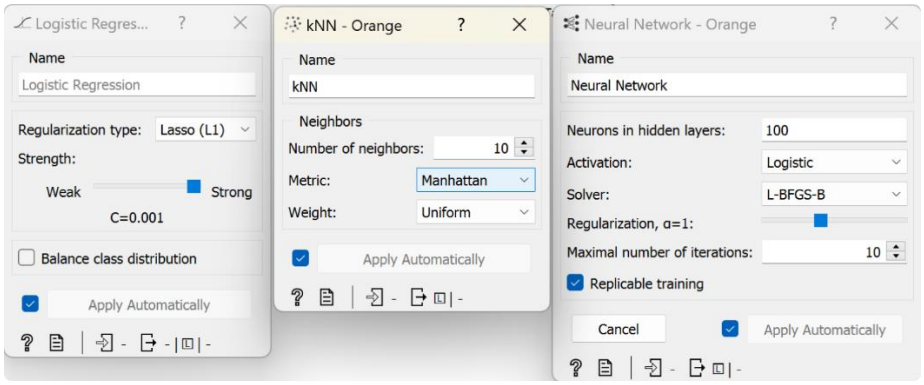


3.4. att. Testa 1 rezultāti

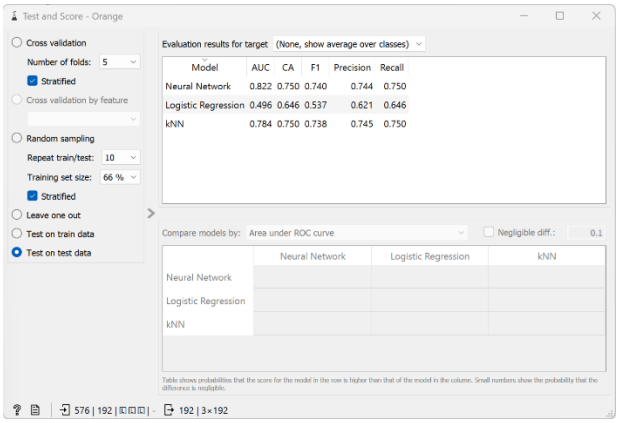


3.5. att. Testa 1 kNN, Logistic Regression, Neural Network proportion of prediction

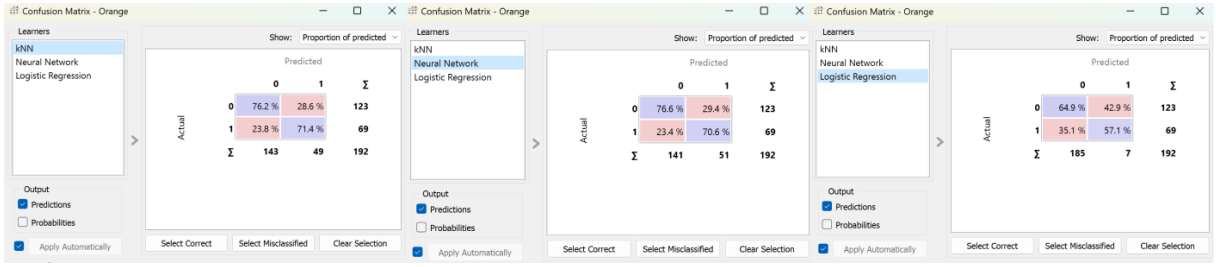
Tests 2



3.6. att. Testa 2 kNN, Logistic Regression, Neural Network parametri

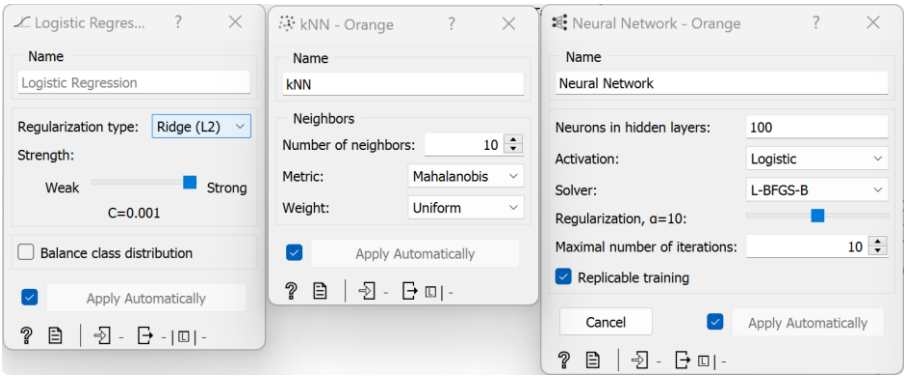


3.7. att. Testa 2 rezultāti

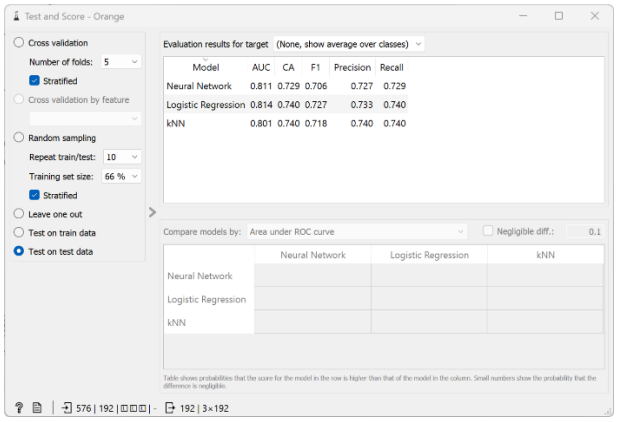


3.8. att. Testa 2 kNN, Logistic Regression, Neural Network proportion of prediction

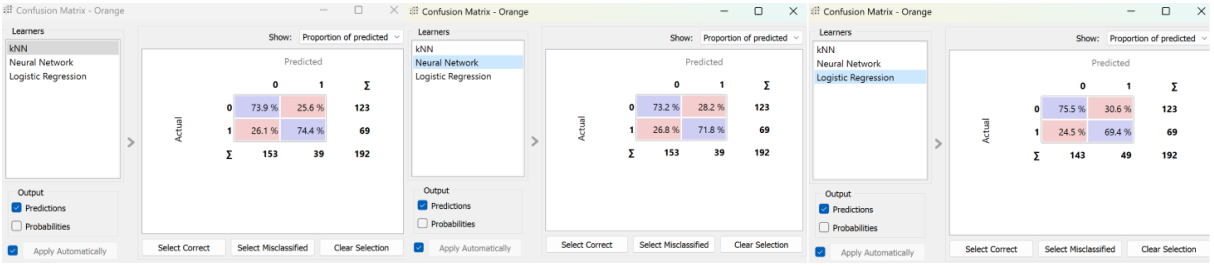
Tests 3



3.9. att. Testa 3 kNN, Logistic Regression, Neural Network parametri



3.10. att. Testa 3 rezultāti



3.11. att. Testa 3 kNN, Logistic Regression, Neural Network proportion of prediction

Rezultāti

Pēc testiem autors izmantoja datus “Confusion matrix” lai noteiktu algoritmu efektivitāti, ņemot divas pozitīvas vērtības izmantojot “Show: Propotion of predict” un atrodot vidējo aritmētisko (skat. 3.5, 3.8, 3.11 att.).

Pirmā testa skaitīšanas algoritma kNN piemērs (74,8% / 67,3%) / 2 = 71,05%.

Test	**kNN**	**Logistic Regression**	**Neural Network**
1	71,05%	70,95%	73,05%
2	73,8%	61%	73,6%
3	74.15%	72.45%	72.5%

Secinājumi

Pēc datu analīzes varam secināt, ka visefektīvākais bija algoritms 3. testā kNN, kas ieguva 74,15%. Var secināt, ka visefektīvākā knn metrika bija "Mahalanobis". Neironu tīklu visefektīvākā alfa vērtība bija 1 un logiska regresija bija visefektīvākais ar regulacijas tipu “Ridge” un c=0.001. Darba autors katrā algoritmā ir mainījis dažus parametrus, tāpēc, izmantojot citus parametrus vai citu datubāzi, efektīvākais algoritms var būt atšķirīgs.

SECINĀJUMI

- Orange data mining ir ļoti viegli apgūstama un jaudīga datu analīzes, apstrādes un pārvaldības programmatūra. Viena no galvenajām Orange Data Mining priekšrocībām ir tā, ka tajā ir iebūvēts plašs algoritmu klāsts, ko var piemērot dažādiem datu veidiem, kas ļauj iegūt augstas precizitātes rezultātus.
- Dažādu algoritmu testēšana palīdzēja praksē pārliecināties par to efektivitāti un iespējamiem pārmaiņu parametriem.
- Izpētot diagrammas, darba autors secināja, ka diabēta iespējamību ietekmē ķermeņa masas indekss, insulīna līmenis, glikoze un vecums.
- datu bāzei ir parametri, kas nav acīmredzami, piemēram, grupēšana un dati un to, cik lielā mērā tie ir savstarpēji nošķirami.

IZMANTOTIE INFORMĀCIJAS AVOTI

1. **https://www.who.int/health-topics/diabetes#tab=tab_1**
2. https://www.kaggle.com/datasets/mathchi/diabetes-data-set
3. https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/hierarchicalclustering.html
4. https://orangedatamining.com/widget-catalog/unsupervised/kmeans/
5. https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/logisticregression.html
6. https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html