LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

LIETOTĀJA UZVEDĪBAS PROGNOZĒŠANAS RĪKS CEĻOJUMU IZVĒLĒ

BAKALAURA DARBS

Autore: Līna Briņģe
Studenta apliecības Nr.: ls14033
Darba vadītājs: profesors, Dr. phil. Jurģis Šķilters

RĪGA 2018

ANOTĀCIJA

Bakalaura darbā "Lietotāja uzvedības prognozēšanas rīks ceļojumu izvēlē" izpētīts, vai un kā ceļotājiem ar patiesu un derīgu informāciju iespējams piedāvāt nākamo ceļojuma galamērķi, balstoties uz ceļotāja un citu ceļotāju iepriekšējās pieredzes.

Darbā detalizēti aprakstītas nozīmīgākās datu prognozēšanas metodes, klasifikācijas un regresijas metodes, kā arī to pielietojumi. Veikta ceļotāju uzvedības analīze, izpētot faktorus, kas var ietekmēt ceļotāju izvēli, izvēloties jaunu ceļojuma galamērķi. Darbā izpētīts, kā pareizi atlasīt, sagatavot, atspoguļot datus un izstrādāt datu prognozēšanas stratēģijas.

Veikta jaunas datu prognozēšanas stratēģijas izstrāde, izmantojot neironu tīklus, un izstrādāta lietotne datu sagatavošanai un prognozēšanai.

Atslēgvārdi: Datu prognozēšana, datu apstrāde, datu sagatavošana, cilvēku uzvedības prognozēšana, neironu tīkli.

ABSTRACT

USER BEHAVIOUR PREDICTION TOOL
FOR TRAVEL CHOICES

My BA thesis "User behaviour prediction tool for travel choices" explores whether and how it is possible to provide travellers with a usable information about possible destinations based on predictable results according to their and other travellers' previous experiences.

Most important data prediction, classification and regression methods and their usability are described. Analysis of the traveller behaviour is elaborated by studying the factors that might have impact on traveller's decision regarding the new prospective destination. My study also explores analysis, processing, preparation and visualisation of data for the purpose of elaboration of data prediction methods. Also proposes different strategies applicable to data prediction methods.

A new data prediction strategy is developed using neural networks as prediction model. Finally, an application is developed, which is further used for data preparation and prediction.

Keywords: Data prediction, data processing, data preparation, prediction of human behaviour, neural networks.

SATURS

APZĪMĒJUMI UN DEFINĪCIJAS

Jēdziens	Definīcija
API	Lietojumprogrammas saskarne. No angļu valodas *Application Programming Interface*.
Apmācība	Process kurā mašīnmācīšanās gaitā tiek ģenerēti datu modeļi, no pieejamajiem apmācības datiem.
Bibliotēka	Metožu un klašu kopums, kas izstrādāts ar noteiktu mērķi un tās iespējams pielietot atkārtoti un sistemātiski konkrētu problēmu risināšanā.
.csv	Faila formāts, ar noteiktu struktūru.
C++	Objektorientēta programmēšanas valoda.
CAFFE	Dziļās mašīnmācīšanās ietvars.
.excel	Faila formāts, ar noteiktu struktūru.
Eksabaits	Datu daudzuma mērvienība. Tā binārā vērtībā ir 2^{60}.
JAVA	Objektorientēta programmēšanas valoda.
.JSON	Faila formāts, ar noteiktu struktūru.
LUA	Programmēšanas valoda.
Modelis	Datu struktūra, kas satur noteiktu informāciju un tiek ģenerēta datu apmācības gaitā. Šī darba kontekstā modeli izmanto datu prognozēšanai.
OpenCV	Atvērtā pirmkoda datorredzes bibliotēka.
Qt	Objektorientētās programmēšanas valodas C++ ietvars.
Python	Programmēšanas valoda.
Zetabaits	Datu daudzuma mērvienība. Tā binārā vērtībā ir 2^{70}.

IEVADS

Veicot pirkumus internetā, informācija par pircēja veiktajiem meklējumiem un pirkumiem tiek saglabāta, lai meklētās vai jau nopirktās preces piedāvātu pircējam atkārtoti, pat apmeklējot citas interneta vietnes (kā sociālos tīklus, e-pastus vai, piemēram, ziņu portālus). Pircējam bieži šāda, jau neaktuāla, informācija var šķist kaitinoša un nederīga, tādējādi izraisot noraidošu attieksmi pret pakalpojuma sniedzēju. Tomēr, ja šie dati par lietotāju ir pieejami, tad kādēļ mērķtiecīgi nepiedāvāt lietotājam kādu citu, saturiski saistītu, preci vai pakalpojumu, kas lietotāju varētu interesēt, balstoties uz lietotāja paša un citu līdzīgu lietotāju vēsturiskajiem datiem?

Šobrīd daudzi reklāmu izvietotāji un pakalpojumu sniedzēji aizvien vairāk cenšas lietotājiem piedāvāt jaunus pakalpojumus un preces, balsoties uz iepriekšējo pirkumu informācijas, taču jaunie piedāvājumi bieži tiek iegūti patvaļīgi vai ar minimāliem ierobežojumiem vai atlases kritērijiem. Tomēr mūsdienu datu analīzes metodes piedāvā plašas iespējas, ar kuru palīdzību būtu iespējams veikt visai precīzu datu prognozēšanu, izmantojot jau konkrētā lietotāja vēsturiskos datus.

Domājot par to, kā šos vēsturiski pieejamos datus izmantot pircējam noderīgi, vērtīga šķiet datu izmantošana ceļojumu plānošanā. Cilvēki pirms došanās atvaļinājumā bieži nemaz skaidri nezina, kurp vēlētos doties atvaļinājumā, taču, zinot datus par lietotāju un viņa iepriekšējiem ceļojumiem, kā arī lietotāja vēlamajiem atlases kritērijiem, jau būtu iespējams piedāvāt jaunu ceļojumu klāstu. Jāņem vēra, ka piedāvājuma klāstā prognozēti tiktu vairāk nekā viens iespējamais galamērķis, lai sniegtu lietotājam plašākas izvēles iespējas.

Bakalaura darba mērķis ir apvienot lietotāju ceļojumu plānošanas izpēti un datu prognozēšanas rīka izstrādi, lai radītu prognozes nākotnes ceļojumu izvēlei. Kā rezultātā lietotājs varētu iegūt noderīgu ceļojumu piedāvājumu kopu, kurp doties nākamajā ceļojumā.

Mērķa sasniegšanai, ir jāveic sekojoši uzdevumi:

- jāizpēta datu prognozēšanas metodes un pieejamie rīki;
- jāveic lietotāju uzvedības analīze, plānojot ceļojumus;
- jāizpēta datu sagatavošanas metodes un noteicošie kritēriji;
- jāizstrādā datu prognozēšanas stratēģija un datu apstrādes un prognozēšanas rīks;
- jāveic datu prognozēšana un iegūto datu analīze.

Tiek plānots, ka ar korektiem datiem par citu ceļotāju pieredzi, ir iespējams veikt ceļojumu galamērķu prognozēšanu nākamajiem ceļojumiem. Veiksmīga rezultāta gadījumā tiktu iegūts priekšstats par datu prognozēšanas metodēm un stratēģijām, un izvēlētās metodes iespējām sniegt lietotājam noderīgu informāciju nākotnes ceļojumu plānošanā.

1. PROGNOZĒŠANAS STRATĒĢIJU IZPĒTE

1.1. Ievads

Nākotni paredzēt vēlas daudzi, taču bieži, lielāka daļa cilvēku pareģotajam tomēr netic, jo visbiežāk pareģotajam nav nekāda loģiska vai empīriska pamatojuma. Taču, ja veidojam nākotnes prognozes, izmantojot vēsturiskos datus, noteiktas stratēģijas un no pagātnes datiem izvirzītas likumsakarības, tad nākotnes prognozēšana ir ar lielāku patiesuma varbūtību.

Lai uzsāktu datu prognozēšanu, svarīgākais uzdevums ir izpētīt un izprast datu prognozēšanas metodes. Nodaļā tiek apskatītas dažādas datu prognozēšanas stratēģijas, kas tiek izmantotas datu prognozēšanas sistēmās, lai izvēlētos piemērotāko šīs problēmas risināšanai. Konkrētā problēma, paredzēt jeb piedāvāt ceļotājam nākamo iespējamo galamērķi, nav statistiski matemātiski triviāli aprakstāma, jo tās nākotnes prognozes ietekmē daudz un dažādi faktori, kam nav viennozīmīgu likumsakarību. Katram ceļotājam var būt savas intereses, paradumi un pieredze ceļošanā, kas sniedz lielu dažādību datu prognožu noteikšanā. Līdz ar to, datu prognozēšanā izmantojamā informācija ir plaša un tā jāspēj apkopot un klasificēt pēc noteiktiem kritērijiem, kas katram lietotājam var būt atšķirīgi.

1.2. Datu prognozēšanas attīstība

Strauji palielinoties pieejamajam datu apjomam, ar vien straujāk ir attīstījusies datu analīze. Pateicoties datu apjomam, laika gaitā ir radušās dažnedažādas metodes, kas pieejamos datus spēj padarīt noderīgus daudzās nozarēs, kā ekonomikā, medicīnā, tautsaimniecībā un citur. Tomēr, ir vērts nedaudz ieskatīties datu prognozēšanas attīstībā.

Lielapjoma datu analīzei un prognozēšanai nav pārāk senas pagātnes. Lielākā metožu pielietošana un straujā izaugsme novērojama tikai šajā gadsimtā. Lai gan ideja, tehniku un metožu pielietošana datu analīzei izmantota jau kopš pagājuša gadsimta[53].

Pirmās idejas par "lielo datu" būtības apzināšanos bijušas jau 1941. gadā, kad tika apzināta kvantitatīvo datu pieauguma prognozēšana. Pirmie lielapjoma datu novērojumi patiesībā saistīti ar datiem fiziskā veidā, kad pateicoties tehnoloģiju un zināšanu attīstībai, strauji pieauga, piemēram, grāmatu apjoms, taču datu analīze kā tāda, ir aktuāla jau kopš pirmo datoru parādīšanās[53]. Līdz ar to pirmie mēģinājumi datu prognozēšanā norit jau 1950. gadā, kad ENIAC datori ģenerē pirmos modeļus laikapstākļu prognozēšanai. Nedaudz vēlāk, 1958. gadā, datu analīzes modeļi tiek lietoti kredītsaistību piešķiršanas riska faktora noteikšanai. 1980. gadā tiek piedāvāts pirmais komerciālais rīks datu analīzes modeļu izstrādei, taču deviņdesmitajos gados, sākoties interneta ērai visā pasaulē, un pieaugot datu

apjomam un darījumiem internetā, kompānijas, kā Amazon, eBay un citi, sāk izmantot personalizētos piedāvājumus cilvēku pirkumiem internetā. Deviņdesmito gadu beigās kompānija Google izstrādā meklēšanas algoritmus, kas bāzēti uz datu analīzes algoritmu pamata, efektīvākai informācijas meklēšanai. Divtūkstošajos gados un līdz pat šodienai, katru dienu strauji pieaug datu apjoms, kā rezultātā arī strauji attīstās datu prognozēšanas tehnoloģijas. Parādās jēdziens "big data" jeb lielie dati, kas simbolizē lielapjoma datus, kas diendienā tiek uzkrāti par dažādiem procesiem. Tā rezultātā katru dienu tiek uzkrāts milzīgs datu apjoms, kas ļauj attīstīt un pētīt gan datu prognozēšanas metodes, gan datu analīzi, kā rezultātā ļauj pētīt un analizēt sabiedrību, dabu, ekonomiku, cilvēku rīcību un citus procesus[65].

Pateicoties lielapjoma datiem, šodien ļoti strauji attīstījušies ne tikai datu prognozēšanas algoritmi, bet arī valodu tulkošanas risinājumi, vizuālās un audiālās informācijas analīze un dažāda veida tīmeklī atspoguļojamā reklāmu informācija, piemēram, dinamiskās cenu reklāmas, un citi piedāvājumi. Tiek prognozēts, ka nākotnē datu analīzes nozīmība cilvēku dzīvēs tikai pieaugs [65].

1.3. Datu analīzes metodes

Uzsākot datu analīzes modeļa izstrādi, jāizvēlas pareiza datu analīzes metode. Tās sekmīga izvēle nodrošinās korektu datu analīzi, kā arī sniegs derīgo informāciju no pieejamajiem datiem, ar pietiekamu precizitāti. Pareiza datu izvēle ir pamats, derīgas informācijas iegūšanai no pieejamajiem datiem. Populārākās datu analīzes metodes ir klasifikācija (no angļu val. *classification*), prognozēšana(no angļu val. *prediction*), regresijas analīze (no angļu val. tulkots regression), kas ir prognozēšanas metode, un līdzības saskaņošanas analīze (no angļu val. tulkots similarity matching) [1]. Tālākajās nodaļās īsi apskatīta katra metode.

1.4. Klasifikācija

Klasifikācijas ideja ir iedalīt pētāmo problēmu noteiktās klasēs, tās kategorizējot pēc noteiktām iezīmēm. Piemēram, nepieciešams paredzēt, vai konkrētai personai ir droši dot aizdevumu bankā, vai nē. Attiecīgi, šajā gadījumā, katrs cilvēks pēc savām pazīmēm, atbilstoši klasifikācijas noteikumiem, atbilst vienai klasei [2].

Datu klasifikācijas process sastāv no diviem posmiem:

1) Datu klasificēšanas modeļa izstrādes, jeb apmācības fāze.

2) Datu klasificēšanas modeļa pielietošana datu prognozēšanas, testēšanas fāze [2, 5].

Datu modeļa prognozēšanas rezultāti var būt vai nu konkrēta klase, kā piemērā par aizdevumu piešķirt vai nepiešķirt aizdevumu, vai skaitliska vērtība, kas nosaka piederību katrai modeļa klasei [5]. Piemēram, ja būtu nepieciešams noteikt cilvēka izglītības līmeni, balstoties uz dažādiem cilvēka dzīves kvalitātes datiem, kā alga, dzīvesvieta, u.c., tad šī informācija nav tik viennozīmīgi nosakāma, līdz ar to vērtīgāk ir prognozēt iespējamību tam, cik ļoti persona pieder katrai izglītības klasei un pieņemt lēmumu balstoties uz to, kurai klasei persona pieder visvairāk. Šeit gan jāņem vērā, ka prognozēts tiktu maksimālais izglītības līmenis, un katra persona var piederēt tikai vienai klasei, tātad ja persona ieguvusi bakalaura grādu, tad šī persona piederēs pie bakalaura grāda klases un nepiederēs pie vidusskolas klases.

Datu klasificēšanas metodes datu prognozēšanai ir plaši izplatītas medicīnas nozarē, medicīnas datu prognozēšanai, tirdzniecības nozarē, personu iepirkšanās paradumu prognozēšanai, bioloģijas datu analīzē, sociālo tīklu analīzē, mēdiju analīzē, dokumentu kategorizēšanā un citur. Tas norāda uz klasifikācijas algoritmu nozīmību ne tikai datu prognozēšanā, bet arī citu, globālāku problēmu risināšanā.

1.4.1. Datu klasificēšanas modeļu izstrāde un pielietošanas būtība

Datu klasificēšanai nepieciešams iepriekš zināmu datu modelis, kas nodrošina jauno datu prognozēšanu, un noteikti algoritmi, kas nosaka, kā datus izmantot un klasificēt. Klasifikācijas modelim nepieciešami iepriekš zināmi dati, kuri tiek saukti par apmācības datiem, kas ir bāze zināšanu modelim un tiek pielietoti modeļa apmācībā [2, 4].

Piemēram, piemērā par aizdevuma izsniegšanu, apmācības datu kopa jeb zināšanu bāze sastāvētu no datiem, par lietotājiem, kuru dati aizdevuma piešķiršanas drošuma ziņā jau ir zināmi. Tātad katra lietotāja datiem jau ir noteikta klasifikācijas iezīme jeb klase - dot vai nedot aizdevumu. Statusa piešķiršanai dot vai nedot aizdevumu ir jāizstrādā noteikti algoritmi, kas var sniegt viennozīmīgu atbildi, kurai klašu grupai, pēc noteiktajiem parametriem lietotājs atbilst. Datu modeļus jeb klasifikatorus var izstrādāt pēc dažādām klasificēšanas metodēm, kuras izvēli ietekmē analizējamie dati. Kad apmācības dati ir iegūti un klasifikācijas modelis ir apmācīts pēc noteiktiem algoritmiem, tad šo modeli var pielietot datu prognozēšanā. Piemēram, ja zināšanu modelis tika apmācīts izvēlēties piešķirt vai nepiešķirt aizdevumu, tad zinot konkrētās personas datus, uz kādiem modelim ir izstrādāti noteikumi, var tikt noteikta šīs personas iespēja saņemt vai nesaņemt aizdevumu.

1.5. Klasifikācijas metodes

Datu klasifikācija ir plaši pielietota metode gan prognožu, gan citu problēmu risināšanai. Klasifikācijas metodes ietver dažādus algoritmus, kas, izmantojot noteiktas datu kopas, spēj izdarīt provizoriskus lēmumus jaunu datu klasificēšanai.

Nodaļā tiks apskatītas tādas klasifikācijas metodes, kā iezīmju atlases metode, varbūtības metode, lēmumu koku metode, instanču balstītā metode, SVM klasifikatori un neironu tīklu metode.

1.5.1. Iezīmju atlases metode

Iezīmju atlases metode (no angļu val. *Feature Selection method*) tiek saukta arī par pazīmju atlases metodi.

Pirmais un nozīmīgākais solis visām klasifikācijas metodēm ir pareiza iezīmju izvēle. Iezīmju atlases metode ir modeļu bāzēta pieeja, ar kuras palīdzību varam samazināt klasificējamās kopas nenoderīgos datus. Bieži klasificējamo datu kopa satur parametrus, kas tieši neietekmē klasificējamo objektu, kā rezultātā var tikt iegūti sarežģīti, nelietderīgi modeļi, kas var izraisīt datu neprecizitātes. Neveiksmīga un nepareiza iezīmju izvēle var novest pie nederīgiem un neprecīziem rezultātiem. Iezīmju jeb mainīgo atlasīšanas metode piedāvā samazināt šo nelietderīgo datu vērtību, tādējādi iegūstot precīzāku un vieglāk interpretējamu modeli [4].

Tiek izšķirtas divas nozīmīgākās iezīmju atlases metodes: filtrēšanas modeļi un saistītie modeļi. Filtrēšanas modeļu gadījumā tiek izmantots viens skaidrs parametrs pēc kā tieši iespējams noteikt piederību konkrētai klasifikācijas kopai. Šī metode nav atkarīga no pielietotā algoritma. Saistīto modeļu metodes gadījumā, iezīmes tiek izvēlētas balstoties pēc noteiktiem algoritmiem. Šīs metodes precizitāti un rezultātus iespējams uzlabot izvēloties dažādus algoritmus. Saistīto modeļu metodes gadījumā iespējams precīzāk konstatēt kas tieši ietekmē modeļu sniegtos rezultātus [4].

Lai veidotu filtrēšanas modeļus, tiek veikti dažādi mērījumi, kas nosaka modeļa atbilstību klasifikācijas procesam. Piemēram, Džīnī indekss, entropijas metode vai Fišera indekss.

Džīnī indekss nosaka atkāpi no vēlamā rezultāta. Piemēram, ja vēlamies ar kādu varbūtību noteiktā testa klase atbilst tās klasei, un mums ir n mērījumi, kur, katrs no p1..pi ir rezultāts, kas ar noteiktu varbūtību atbilst vēlamajai klasei, tad Džīnī indeksu var aprēķināt izmantojot sekojošu formulu:

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

kur G ir indekss, kura vērtība ir no 0 līdz 1/1-k. Visbiežāk mērījumu novērošanai tiek izmantots maksimālā Džīnī indeksa vērtība mērījumu rezultātā [4].

Entropijas metode nosaka gadījumlielumu nenoteiktību. Entropijas metode tiek uzdota šādā formā:

$$E = -\sum_{i=1}^{k} p_i \times \log(p_i)$$

kur E entropijas koeficents, p_i apmācības dati. Entropijas metode lieto tos pašus apzīmējumus, ko Džīnī indekss, tās vērtība ir robežās no 0 līdz $\log(p_i)$.

Fišera indeksa vienādojums nosaka izkliedes rādītāju starp klases atribūtiem. Fišera vienādojums tiek uzdots sekojošā veidā:

$$F = \frac{\sum_{j=1}^{k} p_{j\times(\mu_j * \mu)^2}}{\sum_{j=1}^{k} p_{j\times \sigma_j^2}}$$

kur F ir Fišera indekss, p_j ir apmācības dati, kas pieder klasei j, μ_j ir klases vidējā pazīme, klases globālā pazīme un σ mērījumu standartnovirze [4].

Šie visi mērījumi tiek izmantoti kā palīgmetodes, rezultātu precizitātes uzlabošanai vai klasifikācijas modeļa apmācības statusa noteikšanai.

Filtrēšanas modeļu gadījumā mainīgie bieži ir cieši saistīti viens ar otru, un gadījumos, kad daļa iezīmju jau ir atlasīta, jauna atribūta pievienošana esošo datu kopu ietekmē citādāk, nekā tā strādātu neatkarīgi [4]. Attiecīgi, filtrēšanas modeļu metodē svarīgi saprast izvēlēto iezīmju savstarpējo ietekmi uz kopējo rezultātu kvalitāti.

Saistīto modeļu gadījumā iezīmes tiek izvēlētas iteratīvā procesā atkarībā no izvēlētā klasifikācijas algoritma. Katrā solī tiek iegūta pazīmju kopa, kas tiek mākslīgi palielināta izmantojot noteiktu stratēģiju un testēta, lai pārbaudītu klasifikācijas modeļa kvalitātes uzlabojumus. Tā kā klasifikācijas algoritms tiek lietots iezīmju noteikšanai, iegūtais modelis it atkarīgs no klasifikācijas algoritma [4]. Saistītie modeļu algoritmi ir, piemēram SVM algoritms, kas aprakstīts zemāk nodaļā.

Šīs metodes galvenie plusi ir, ka tā ir modeļu bāzēta pieeja un parasti tās modeļi sniedz labu sniegumu, taču to mīnuss ir to skaitliskā dārdzība, jeb to skaitļošanas process var būt ļoti laikietilpīgs [59].

1.5.2. Varbūtības metode

Varbūtību teorēma pazīstama kā plaši lietota matemātikas statistikas metode, ko izmanto notikumu iespējamības noteikšanai. Par notikumu tiek uzskatīts jebkurš fakts vai darbība, kas var notikt izmēģinājuma vai novērojuma rezultātā [48]. Klasisks varbūtības piemērs ir kauliņu mešana. Piemēram, kāda varbūtība pastāv uzmest skaitli 1. Attiecīgi, to mēra iespējamo notikumu skaitu dalot ar visiem iespējamajiem notikumu skaitiem, šajā gadījumā, metot vienu kauliņu varbūtība ir 1/6 [6].

Varbūtību metode klasifikācijā un datu prognozēšanā tiek izmantota klašu varbūtību aprēķinam un tā ir nozīmīgākā datu klasifikācijas metode, jo nosaka klasifikācijas rezultātus. Varbūtības metodes algoritmi izmanto statistiskus secinājumus, lai noteiktu konkrētā mērījuma iespējamo piederību konkrētai klasei [4]. Attiecīgi tiek noteikta varbūtība, ar kādu konkrētie klasificējamie dati atbilst katrai klasifikatora klasei. Tā izmanto statistiskus secinājumus, lai noteiktu iespējamo rezultātu klasi.

Varbūtību metode izmanto *posterior* jeb "aizmugurējo" varbūtības rezultātu, kas tiek iegūta modeļa testa rezultātā, ievērojot tā īpatnības. Vienkāršāk sakot, šis rezultāts ir katra apmācības posma iegūto testu rezultāts, kas atspoguļo attiecīgā momenta datu modeļa precizitāti [4]. Šo varbūtību iespējams noteikt vairākos veidos.

Viens no variantiem, kā aprēķināt *posterior* varbūtību, ir aprēķināt katras klases varbūtību, prioritārās klases varbūtību un, pielietojot Beijesa teorēmu, atrast klases varbūtību[4]. Beijesa teorēma nosaka:

$$P(A|B) = \frac{\mathrm{P(B|A)P(A)}}{P(B)}$$

kur P(A) ir varbūtība vienam notikumam, bet P(B) varbūtība otram notikumam. Pielāgojot teorēmu konkrētajam gadījumam klasifikācijas vajadzībām, iegūstam:

$$P(Y(T) = i|x_1 \dots x_n) = P(Y(T) = i) \times \mathrm{P(x_1 \dots x_n\ |Y(T) = i)}$$

kur P(Y(T)=i) atbilst P(A) un P(x_1...x_n) atbilst P(B). P(A) šajā gadījumā ir notikuma varbūtība, kuru vēlamies noskaidrot, bet P(B) visi iespējamie notikumi. Formulā šajā gadījumā vēlamies noskaidrot *posterior* varbūtību, ka testa instance Y(T) atbilst klasei i, pēc

tam var pielieto Beijesa teorēmu, lai iegūtu iepriekš aprakstīto formulu. Tā kā Beijesa formulas saucējs P(B) šajā gadījumā vienmēr ir nemainīgs, jo vienmēr vēlamies noskaidrot maksimālo iestāšanās varbūtību, tad tas tiek noīsināts [4]. Varbūtību metode tiek izmantota gan kā galvenā klasifikatora metode, gan kā palīgmetode citām klasifikācijas metodēm.

Varbūtības metodes plusi ir tās nesarežģītums un skaitliski vienkāršie aprēķini. Kā mīnuss šai metodei ir tās būtība klasificēšanas problēmu risināšanā nav pielietojama tieši, bet gan kā papildus metode, kas nodrošina skaitlisko rezultātu noteikšanu.

1.5.3. Lēmumu koki

Lēmumu koki ir neparametriska uzraudzītā mācīšanās metode, kas tiek izmantota klasificēšanā. Lēmumu koku mērķis ir uzbūvēt modeli, kas balstoties uz objekta mainīgajiem, pēc noteiktiem noteikumiem var noteikt nezināmo objektu [4, 7]. Mašīnmācīšanās ir mākslīgā intelekta apakšnozare, kurā tiek pētīti un izstrādāti tādi algoritmi, kas spēj dot iespēju datoram "apmācīties" ar noteiktiem datiem, kas nav cieti iekodēti, dažādu lēmumu pieņemšanai. Neparametriskā uzraudzītā mācīšanās metode ir mašīnmācīšanās metode, kurā datu apmācībai un modeļu veidošanai izmanto tādus datus, kam ir zināmi gan to ievaddati, gan to atbildes. Tātad ir zināma vēlamā atbilde [3, 4].

Lēmumu koki veido hierarhisku datu sadalījumu, kur katra koka lapa atbilst noteiktiem nosacījumiem, kas piemēroti noteiktai klasei. Katrā koka līmenī notiek lēmumu pieņemšana, kuras rezultātā notiek zarošanās pēc noteiktiem algoritmiem.

Lēmumu kokiem piemīt vairākas priekšrocības, tos ir salīdzinoši vienkārši saprast un interpretēt, pateicoties to uzskatāmajai vizualizācijai. Tie neprasa lielu datu sagatavošanu, tomēr tie nepieļauj tukšo vērtību esamību [7]. Piemēram, citas metodes pieļauj datu prognozēšanu gadījumos, kad kāds no parametriem ir iztrūkstošs, taču lēmumu koku apmācībā šādus datus izmantot nevar. Vēl lēmumu koki spēj analizēt dažādu datu tipu datus, piemēram, gan skaitliskas, gan vārdiskas jeb kategorizētas vērtības.

Lēmumu koki ir izdevīgi no to sarežģītības līmeņa, jo to sarežģītība datu prognozēšanā ir logaritmiska datu punktu skaitam, ar kādu lēmumu koks tiek mācīts [4, 7].

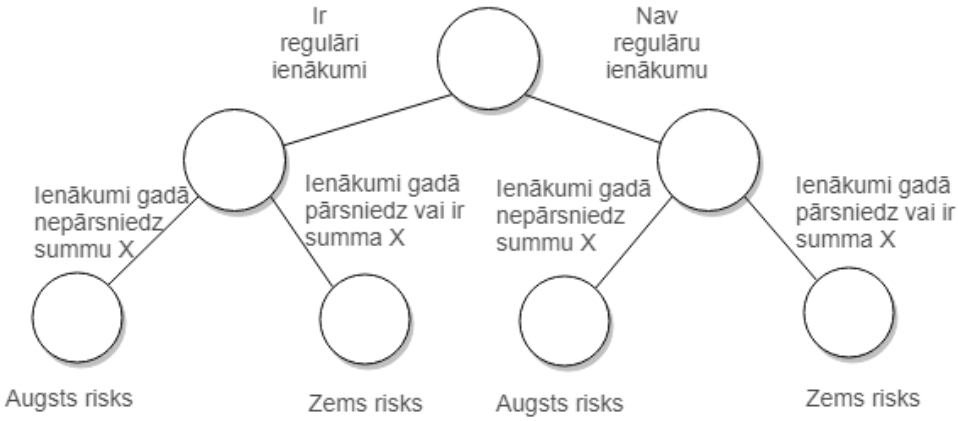
Klasiskā gadījumā lēmumu koku sarežģītības formula ir šāda:

$$O(n_{piemēri} n_{iezīmes} \log(n_{piemēri}))$$

kur $n_{piemēri}$ ir vienību skaits, bet $n_{iezīmes}$ ir īpašību skaits, kas tiek piemērots katrai vienībai.

Tomēr lēmumu kokiem ir arī būtisks mīnuss, tā veidošanas procesā nav iespējams prognozēt labāko brīdi, kad apturēt lēmuma koka augšanu, lai novērstu koka pārmērīgu mācīšanu. Līdz ar to, ir izdomāti vairāki veidi kā samazināt jeb laikus apgriezt lēmuma kokus. Viena no pieejām ir daļu no apmācības datiem nošķirt koka augšanas fāzē. Pēc tam tie tiek izmantoti testiem, lai noteiktu vai aizvietojot kādu no koka fragmentiem ar vienu mezglu, iespējams uzlabot klasifikācijas precizitāti. Ja kritērijs izpildās un koka klasifikācijas precizitāte ir uzlabojusies, tad koka apcirpšana ir izpildīta un koka fragments tiek aizstāts ar attiecīgo mezglu [4, 8].

Attēlā Nr. 1.1 redzams triviāla lēmuma koka piemērs, kurā atspoguļota riska iespējamība piešķirot klientam kredītu.



1.1. att. **Lēmumu koka piemērs**

Piemērā attēlā Nr.1.1. redzam lēmumu koku ar nosacījumiem, kuru rezultātā iespējams nonākt pie konkrētas klasifikācijas klases - augsts risks vai zems risks. Rezultātus iespējams iegūt, kokam pielietojot noteiktu klasifikācijas modeli, kurā jau ir zināmi, iepriekš kategorizēti dati. Piemēram, ja klientam Pēterim ir regulāri ienākumi un tie gadā pārsniedz summu X, tad klients Pēteris ir zemā riska kategorijā un viņam ir droši piešķirt kredītu.

Šajā piemērā nosacījumi abos mezglos ir vienādi, taču citos kokos tie var būt atšķirīgi, attiecīgi katrs zars var saturēt savus tālākos zarošanās nosacījumus.

Šīs metodes pluss viennozīmīgi ir iespēja veidot plašus datu atlases kritērijus, taču lielākais mīnuss reizē arī ir straujais datu analīzes pieaugums, gadījumos, kad pētāmā problēma kļūst ar vien lielāka.

1.5.4. Likumos balstītas metodes

Likumos balstītā metode ir mašīnmācīšanās metode, kas identificē, mācās vai nosaka likumus, lai tos tālāk izmantotu, piemēram, datu prognozēšanai. Pēc idejas, tā ir ļoti līdzīga lēmumu koku pieņemšanas metodei, taču tā neveido stingru mācību datu hierarhiju [4, 9].

Likumos balstīta metode ir bāzēta uz *if than* (no angļu val. ja tad) nosacījumu, kur vienā likumā aprakstīti var būt vairāki nosacījumi [10]. Tādējādi pretēji lēmumu pieņemšanas kokam likumos balstīta metode izvairās no secīguma un hierarhijas.

Piemēram, likumos balstītas metodes klasisks nosacījumu piemērs būtu izdarīt konkrētu darbību:

JA X vecums ir >=18 un X ir students **TAD** piešķirt X studenta atlaidi.

Piemēra nosacījuma jeb noteikumu daļa ir Ja vecums >= 18 un ir students, bet secinājums jeb sekas ir piešķirt studentam atlaidi. Ja nosacījuma vērtība ir patiesa, tad nosacījums tiek izpildīts un attiecīgajai personai X tiek piešķirta pazīme Y [19].

Likumos balstītās metodes noteikumi var tikt kombinēti, veidojot sarežģītāku struktūru. Tomēr, veidojot vairāku likumu modeli, jāizvērtē, vai struktūrā nevar veidoties pretrunas likumi. Piemēram,

JA X = students **TAD** X jāpērk dators

JA X = students, **TAD** X ienākumu līmenis ir zems.

Šajā piemērā objekts X atbilst abiem nosacījumiem, taču abu nosacījumu sekas izraisa pretrunu, jeb nereālu situāciju [19] . Attiecīgi, šajā gadījumā students nevar iegādāties datoru, taču viņam tas ir jādara. Viens no veidiem kā risināt šo problēmu ir piešķirt nosacījumiem prioritātes, tādējādi konfliktu situācijās izvēloties izpildīt to nosacījumu, kuram būs augstāka prioritāte [19]. Datu prognozēšanā ar nosacījumu palīdzību iespējams nonākt vai nu pie konkrētās klases vai nākamā nosacījuma. Gala rezultātā, balstoties uz nosacījumu bāzes būs iespējams veikt datu prognozēšanu.

Šī metode ir ērti pielietojama salīdzinoši mazu problēmu risināšanā. Tā ir relatīvi vienkārša, un strādā pēc noteikta likuma. Ārī šīs metodes problēma ir datu apjoma un dažādības pieauguma gadījumā tā kļūst skaitliski neizdevīga.

1.5.5. Instanču balstīta apmācības metode

Instanču balstīta apmācības metode ir mašīnmācīšanās metodes mācīšanās algoritmu klases pārstāve. Tā savā pamata darbībā izmanto testa kopas informāciju kā zināšanu bāzi, tā

vietā, lai veidotu optimizētu klasifikācijas modeli. Zināšanu bāze sastāv no apmācības gadījumu kopas. Jaunu instanču pievienošanas gadījumā, tiek meklēts tuvākais līdzīgais gadījums zināšanu bāzē, tādējādi papildinot zināšanu bāzi ar jaunu instanču vērtībām [11, 49].

Šī metode ir salīdzinoši sarežģīta, un noderīga gadījumiem, kad klasificējamie dati var saturēt vēl iepriekš neredzētas situācijas, pateicoties tās apmācības metodei. Instanču balstītās pieejas modeļiem ir liels pluss, tajā faktiski nav novērojami informācijas zudumi un tās, pateicoties apmācības modeļa metodei, modeļus ir viegli pielāgot [4].

Instanču balstītā apmācības metodei ir plašs pielietojums un vairāki apakš algoritmi, kas risina dažādas problēmas. Viens no piemēriem ir "rote" jeb atmiņas mācīšanās(no angļu val. rote learning). Šajā gadījumā ir sākotnējā zināšanu bāze, kurai tiek pievienoti jauni elementi.

Jauno elementu pievienošana notiek pievienojot jaunus elementus, ja to dati pilnībā sakrīt ar zināšanu bāzi [49].

Piemēram, ja zināšanu bāze sastāv no šādiem datiem(skatīt tabulā 1.1.).

1.1. tabula

Zināšanu bāze

Diena	Laikapstākļi	Vējainums	**Spēlēt futbolu**
05/07	Saulains	Nē	**Jā**
06/07	Lietains	Jā	Nē
07/08	Saulains	Jā	Jā

Tad zemāk redzamie jaunie dati zināšanu bāzē tiks pievienoti sekojoši:

Šodiena	Saulains	Nē	**?**

↓

Šodiena	Saulains	Nē	**Jā**

Redzams, ka jaunie dati, balstoties uz zināšanu bāzi, kurā klasificējamā klasei ir vai *spēlēt futbolu*, ir pievienoti klasei *jā*.

Kā vēl viens no piemēriem, ir tuvākā kaimiņa klasifikators (angļu val. nearest neighbor classifier). Šīs metodes klasifikators apmācības procesā kā apmācības datus izmanto jau klasificētas testa datu kopas tuvākos *k* kaimiņus, pēc noteiktām īpašībām. Konkrētās klases iezīme ar visbiežāko parādīšanos starp tuvākajiem *k* kaimiņiem tiek noteikta kā atbilstošākā klase [4, 11]. Ja apskatām iepriekšējo piemēru, tad atmiņas metode nepieļauj

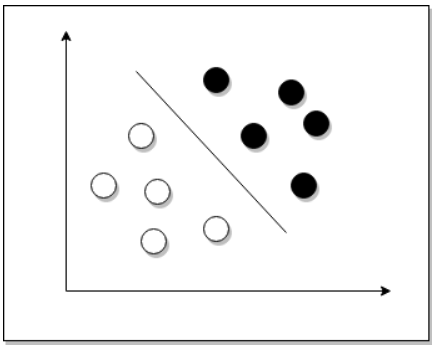
jaunu vērtību parādīšanos, taču tuvākā kaimiņa metode meklēs vairāk nekā vienu līdzīgo piemēru zināšanu bāzē, tādējādi pieļaujot jaunu vērtību ienākšanu zināšanu bāzē [49].

Instanču balstītā apmācības metode ir plaši izmantota, pateicoties tās augstajai precizitātei [49].

1.5.6. SVM klasifikatori

SVM klasifikatori jeb atbalsta vektoru tīklu klasifikatori ir uzraudzītās mašīnmācīšanās metožu kopums, kas tiek lietoti klasifikācijai regresijai un robežu noteikšanai. SVM metodes savā darbībā izmanto lineāri noteiktus kritērijus, nevis veido nošķirtas klases [4, 12, 13].

SVM klasifikatoru galvenais mērķis ir nodalīt dažādu pazīmju objektus vienu no otra. Zemāk attēlā Nr. 1.2 redzams piemērs.



1.2.att. **SVM klasifikatora paraugs**

Piemērā redzamais paraugs atspoguļo vienkāršu klasifikatora uzdevumu, nošķirt, jeb atrast robežu starp melnajiem un baltajiem apļiem, tātad objektu kopa tiek piekārtota tās klasei. Tomēr bieži reālās dzīves uzdevumi ir pavisam citas sarežģītības. SVM klasifikatora uzdevums ir risināt šos sarežģītākos uzdevumus, piemēram, nošķirt objektu gadījumā, kad tam nav viennozīmīgas lineāras robežas, bet gan objektu kopas ir vairāk saplūdušas.

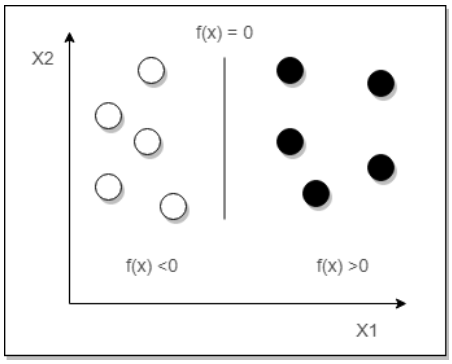
SVM klasifikatora modeļi tiek definēti binārām klasifikācijas problēmām, tātad klases mainīgais y_n n-tajai apmācības instancei X_n tiek mērīts robežās no {-1, +1}. To biežāk lietotais kritērijs robežu nodalīšanai ir tā saucama *maksimuma robežu hiperplakne.* SVM klasifikatori veic uzdevumu konstruējot šīs hiperplaknes daudzdimensionālā telpā, kas atdala dažādu klašu objektus [4, 14].

SVM lineārais klasifikators vispārinātā formā tiek uzdots sekojošā veidā:

$$f(x) = W \times X + b$$

kur W ir daudzdimensionālas vektors, kas satur hiperplaknes atdalīšanas koeficientus, X, ir apmācības dati un b ir konstante [4, 15].

Zemāk attēlā Nr. 1.3. redzams divdimensionāls SVM klasifikatora piemērs grafikā.



1.3. att.. **SVM klasifikatora paraugs**

Divdimensionālā gadījumā, kā redzams piemērā, robeža ir taisne, bet daudzdimensionālā gadījumā tā ir hiperplakne, kas uzdota W vektora formā. SVM klasifikators ir šīs vispārīgās formulas optimizācija, kas ievieš divus papildus kritērijus, tādējādi izpludinot robežu un panākot precīzāku rezultātu, gadījumos, kad objekti nav precīzi lineāri atdalāmi. Sekojošie kritēriji ir:

$$W^T \times X_i + b = 1$$
$$W^T \times X_i + b = -1$$

kur W^T hiperplaknes vektors, X_i konkrētā instance un b konstante. Tādējādi nosakot šādu kritēriju, robežas biezums tiek palielināts [4, 15].

SVM klasifikators plaši tiek izmantots tieši datorredzes gadījumos, kad attēlā nepieciešams atrast kādu konkrētu objektu, piemēram, cilvēka seju vai ķermeni. Tas iespējams, jo SVM izmanto vektoru, ko viegli pielietot attēla pikseļiem. Klasisks SVM pielietojuma scenārijs ir izveidot datu kopu, konkrētam objektam, piemēram, cilvēka ķermeņa atrašanai attēlā. Tajā attēlu pikseļi ar cilvēka ķermeņiem tiek uzdoti daudzdimensionālā īpašību vektorā, kas tiek pielietota SVM klasifikatora apmācībā. Pēc tam pielietojot klasifikatora vispārīgo funkciju testa attēlam, tiek noteikta objekta robeža [13, 15].

SVM klasifikatoru pluss ir tā precizitāte, taču to ieteicams lietot uz maza apjoma datiem. Ņemot vērā tā darbības principus, tā mīnuss ir lielu datu apjoma apstrāde, kuras rezultātā modeļu apmācības process var kļūt ļoti laikietilpīgs [4].

1.5.7. Neironu tīkli

Neironu tīkli ir mašīnmācīšanās informācijas apstrādes sistēma, kas radīta balstoties uz bioloģiskajām sistēmām, mēģinot imitēt cilvēka smadzeņu darbību. Cilvēka smadzenēs neironi ir savienoti ar sinapsēm. Bioloģiskās sistēmās mācīšanās process notiek reaģējot uz noteiktiem impulsiem, kuru rezultātā notiek sinapšu savienojumu spēku palielināšana. Šis mācīšanās process ir neironu tīklu pamata būtība [3, 4].

Neironu tīkls sastāv no liela skaitļošanas datu apjoma, un tā mērķis ir apmācības procesā spēt saņemto datu apjomu pārvērst noteiktās īpašības, un no tām spēt pielāgot noteiktus izejas datus. Visbiežāk neironu tīkli tiek izmantoti klasifikācijas problēmu risināšanai, jo tie, balstoties uz apmācības procesu, spēj veikt visai sarežģītus klasifikācijas uzdevumus. Neironu tīkls spēj strādāt ar bināriem datiem, tādēļ gadījumos, kad nepieciešama attēlu vai skaņu datu klasifikācija, tie jāpārveido skaitliskā formātā, kas var būtiski palielināt datu apjomu [3, 4].

Neironu tīklu vienkāršākā vienība ir neirons. Šie neironi var tikt savstarpēji savienoti dažādos veidos, izmantojot dažādus savienojumus un arhitektūras. Neironu tīkla uzbūve sastāv no svariem, summēšanas jeb izplatīšanas funkcijas un aktivitātes funkcijas [3]. Zemāk nodaļā ir aprakstīta katra neirona tīkla uzbūves sastāvdaļā sīkāk.

Visvienkāršākā neironu tīklu arhitektūra jeb modelis ir perceptrons. Perceptrons ir kalpojis par pamata bāzi lielākajai daļai neironu tīklu modeļu, tādēļ ir būtiski izprast tā uzbūvi. Apskatot perceptrona uzbūvi, ir iespējams labāk izprast neironu tīkla jēdzienu.

Perceptrons sastāv no viena vai vairākiem neironiem jeb mezgliem, kas satur vienādu skaitu ieejas datu, un izejas mezgla. Izejas mezgls saņem ieejas datu kopu, kas tiek iegūta no ieejas datiem. Izejas datu mezgls ir saistīts ar svaru kopumu W, kas tiek pielietots summēšanas funkcijai f(x) no tās ieejas datiem [4]. Katra svaru komponente svaru vektorā W ir saistīta kā ieejas un izejas datu savienojums. Šie svari var tikt uzskatīti kā analogi bioloģisko sistēmu sinaptiskajam spēkam. Svari ir būtiskākā sastāvdaļa, jo tie nodrošina apmācības procesu. Svari ir skaitliskas vērtības, kas tiek ietekmētas apmācības procesā un nodrošina neirona tīkla spēju gala rezultāta veikt kādu konkrētu uzdevumu [3]. Šajā gadījumā spēki neveic nekādus aprēķinus, taču tie tiek izmainīti kombinējot tos ar ieejas datiem un pielietojot tiem aktivitātes funkciju.

Summēšanas funkcija (no angļu val. propagation function) ir funkcija, kas pārvieto vērtības cauri neironu tīklu slāņiem. Tipiski summēšanas funkcija summē ienākošās vērtības ar svariem un tālāk nodod tās aktivitātes formulai [3, 16].

Aktivitātes funkcija (no angļu val. (activation function), ir galvenā funkcija, kas izrēķina izejas datu vērtību. Aktivitātes funkcijas var būt dažādas, tos izvēle atkarīga no neironu tīkla izstrādātāja.

Zemāk piemērā redzama summēšanas funkcija:

$$f(x) = \sum_{i=0}^{n} (W \times X_i + b)$$

kur izejas dati ir prognozētā vērtība konkrētajam klases mainīgajam, kas tiek padota aktivitātes funkcijai, b ir nobīde(*bias*) jeb papildus svars, X_i ieejas dati un W svaru vērtība. Svaru vektora un ieejas datu skaitam ir jāsakrīt [3, 4].

Kad summēšanas funkcija savu darbu ir veikusi, tālāk iegūtais rezultāts tiek pielietots aktivitātes funkcija, kas izrēķina izejas vērtību. Kā jau pirmīt minēts, tad aktivitātes funkcijas var būt dažādas. Tās varbūt lineāras triviālākiem gadījumiem un neliniāras sarežgītāku uzdevumu risināšanai. Viena no vienkāršākajām aktivitātes funkcijām ir sliekšņveida funkcija:

$$f(NET) = \begin{cases} B, & Net > T \\ A, & Net \leq T \end{cases}$$

kur T ir slieksnis, kas nosaka atbilstību vienam vai otram nosacījumam. Tomēr sliekšņveida funkcijas parasti tiek pielietotas bināru problēmu gadījumā, tādēļ populārāk tiek izmantotas pseidolineārās, sigmoidālās, hiperboliskās vai tangensiālās aktivitātes funkcijas[3].

Neironu tīklu izejas vērtības ir noderīgi iedalīt noteiktās robežās, piemēram, no 0 līdz 1, lai spētu nodrošināt precīzāku rezultātu interpretēšanu. Šādam mērķim vēlams izmantot sigmoidālās aktivitātes funkcijas [3, 4].

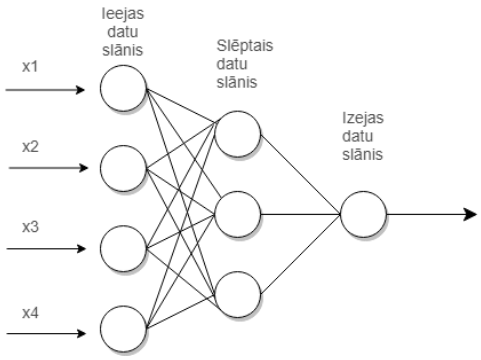
Neironu tīklu apmācības mērķis ir piemācīt svaru vektoru, pielietojot apmācības datu kopu. Apmācība ir process, kurā neironu tīkla svaru vērtību regulēšana, balstoties uz ieejas datiem, kuru mērķis ir risināt kādu noteiktu problēmu [3]. Idejiski pirmajā solī, svari tiek ģenerēti automātiski, vai citos gadījumos noteikti statiski. Pēc tam, apmācības procesā, gadījumos, kad klasificēšana ir bijusi kļūdaina, svari tiek pakāpeniski atjaunoti, piemērojot konkrētā brīža funkciju, apmācības datiem, ar kuriem tīkls kļūdījies. Svaru atjaunošanas apjoms tiek regulēts ar apmācības ātrumu λ, kas nosaka apmācības procesa ilgumu. Šajā

gadījumā λ veic kļūdu korekciju starp ieejas datiem un vēlamajiem izejas datiem[3, 4]. Vienkāršāk sakot, apmācības ātrums nosaka, cik ātri neironu tīkls ieraugot jaunu piemēru spēs saprast, ka arī šis piemērs ir pareizs. Piemēram, ja apmācīsim neironu tīklu atpazīt kaķus un tas desmit reizes pēc kārtas būs redzējis baltu kaķi, bet vienpadsmitajā reizē redzēs melnu kaķi, tad, jo λ būs lielāka, jo ātrāk tīkls spēs piemācīties uz jaunajām vērtībām. Jo mazāka būs šī vērtība, jo tīkls vairāk pieņems, ka melnais kaķis ir izņēmuma gadījums un neņems to vērā. Tipiski ir izdevīgi lietot pēc iespējas augstāku λ vērtību, taču ir gadījumi, kad zema λ vērtība var nākt par labu. Piemēram, tādos gadījumos, kad apmācības informācija ir ļoti vienāda. Neironu tīkla atjaunošanas funkcija ir sekojoša:

$$W^{t+1} = W^t + \lambda(\mathrm{y_i} - \mathrm{z_i})\mathrm{X_i}$$

kur W_t ir svaru vektors konkrētajā apmācības iterācijā, X_i apmācības dati, y_i-z_i ir starpība starp vēlamo un esošo rezultātu. Funkcijā iespējams saskatīt, ka ja, ieejas dati klasifikācijā būs pareizi, tad svaru korekcija nenotiks. Apmācības process notiek cikliski, un tā procesa laiku, kad vienu reizi neironu tīklam tiek padoti visi apmācības paraugi, tiek saukta par epohu [3]. Apmācības procesā epohu skaits parasti ir atkarīgs no problēmas, kā arī ieejas datu apjoma.

Zemāk attēlā Nr. 1.4. aplūkojums parasta vairākslāņu neirona tīkla piemērs.



1.4. att. **Daudzslāņu neirona tīkla paraugs**

Attēlā Nr. 1.4. redzamais piemērs satur četrus ieejas datu mezglus, trīs vidusslāņa jeb slēptā slāņa mezglus un vienu izejas mezglu, kas rezultātā dod atbildi, konkrētajai problēmai ar ieejas datiem.

Neironu tīkli spēj risināt daudz un dažādu sarežģītību uzdevumus, kā arī spēj prognozēt jaunus procesus, tos redzot pirmo reizi. Nodaļā apskatītais ir tikai paši pamati neironu tīklu būtībai, taču rada priekšstatu par to noderīgumu un vērtību. Taču to mīnuss ir to

sarežģītība. Neironu tīkli ir sarežģīti gan pēc to būtības, no algoritmu viedokļa un spējām tos saprast, gan no resursu patēriņa viedokļa. Neironu tīkli risina ļoti dārgas procedūras no skaitļošanas viedokļa, kā rezultātā tas var būt ļoti laikietilpīgs process apjomīgu un sarežģītu datu analīzē.

1.6. Prognozēšana

Datu prognozēšanas metodes mērķis ir prognozēt konkrētas vērtības salīdzinājumā ar klasifikācijas metodēm, kuras prognozē piederību noteiktām pazīmju grupām. Piemēram, mārketinga menedžerim ir nepieciešams prognozēt, cik daudz uzņēmums nopelnīs izpārdošanas laikā. Šajā gadījumā ir nepieciešams prognozēt konkrētu vērtību, pēc kādas noteiktas funkcijas, kas aprakstītas tālāk nodaļā.

Datu prognozēšana izmanto vēsturiskus datus, lai paredzētu nākotni. Parasti, vēsturiskie dati tiek apkopoti noteiktā matemātiskā modelī, kas tiek pielietots noteiktiem datiem, lai paredzētu, kas notiks nākotnē, vai lai ieteiktu noteiktas darbības nākotnes problēmu risināšanai [17] .

Datu prognozēšanas metožu pielietošana šobrīd ir kļuvusi ļoti aktīvi izmantota, kā arī pētīta, pateicoties pieejamajam datu apjomam un tehnoloģiskajai attīstībai, tieši mašīnmācīšanās algoritmu attīstībā. Tās popularitāte, pateicoties precizitātes pieaugumam, pieaug arī dažādās nozarēs, kā finanses, bizness, ekonomika un citas.

Datu prognozēšana atšķirībā no klasifikācijas, veic nezināmo vērtību noteikšanu, attiecīgi, piemēram, ja mārketinga menedžeris vēlas prognozēt kādu ienesīgumu nesīs konkrētais pircējs, tad tā būs nākotnes datu paredzēšana jau konkrētai vērtībai, taču, ja vēlēšanās ir saprast, vai konkrētais pircējs pirks vai nepirks datoru, tad tā jau ir konkrētu iezīmju piešķiršana un atbilst klasifikācijas problēmai [18]. Tātad galvenā atšķirība starp abām metodēm ir saistīta ar to, kādu problēmu vēlamies noskaidrot. Ja nepieciešams iedalīt grupās, piešķirt kādu pazīmi, tad jāizvēlas klasifikācija, ja nepieciešams prognozēt kādu nezināmu, piemēram, skaitlisku vērtību, tad tā ir prognozēšana. Tipiskākā prognozēšanas metode ir regresijas analīze [17].

1.6.1. Regresijas analīze

Regresijas analīzes mērķis ir pētīt sakarības starp noteiktiem datiem jeb lielumiem. Regresijas uzdevums ir pētīt un noskaidrot vai un kāda sakarība var pastāvēt starp atsevišķiem rādītājiem, kā arī veidot šo sakarību modeļus. Regresijas metodes tiek lietotas gadījumos, kad sakarības nav funkcionāli nosakāmas, bet ir statistiski nolasāmas [20]. Līdz ar to, problēmās,

ko pēta regresijas analīze, mainīgo vērtības var būt dažādas, tās var ietekmēt nejaušība vai kādi citi ārēji faktori.

Regresijas sakarības var būt gan lineāras, gan nelineāras. Tās iedala dažādās apakš metodēs, kā pāra regresija, saistītā regresija un nelineārā regresija [19, 20].

1.6.2. Pāra regresija

Pāra regresija ir sakritība starp diviem rādītājiem, kur abi lielumi ir mainīgi. Šādu sakarību var pierakstīt formā:

$$y = f(x) + e$$

kur y ir viens rādītājs, x otrs rādītājs un e nejaušs gadījuma lielums. Šādi var redzēt, ka y vērtībai piemīt sakarība, kur kādas funkcijas vērtībai no x pieskaitot e, iegūst y vērtību. Šajā gadījumā y ir atkarīgais mainīgais lielums, bet x ir neatkarīgais mainīgais [19]. Apskatīsim kādu piemēru. Piemēram, atkarīgais y mainīgā lielums būs produkcijas pašizmaksa, bet neatkarīgais mainīgais lielums x būs ražošanas apjoms vai cena.

Pāra regresijas matemātisko izteiksmi jeb modeli var pierakstīt sekojošā formā:

$$y = b_0 + b_1 x + e$$

kur b_0 , b_1 sakarības, kas noteikta lineārā formā, parametri un x, y, e, saglabā iepriekš aprakstīto nozīmi. b_0 tiekot dēvēts par vienādojumu brīvo locekli, kas parāda, y vērtību, ja x vienāds ar nulli, bet b_1 par regresijas koeficientu, kas atspoguļo y izmaiņas, ja c vērtība tiek palielināta par vienu vienību. Pāra regresijas gadījumā, veicot analīzi, tiek noteikts, kuras īpašības ietekmē viena otru [19, 23].

1.6.3. Saistīta regresija

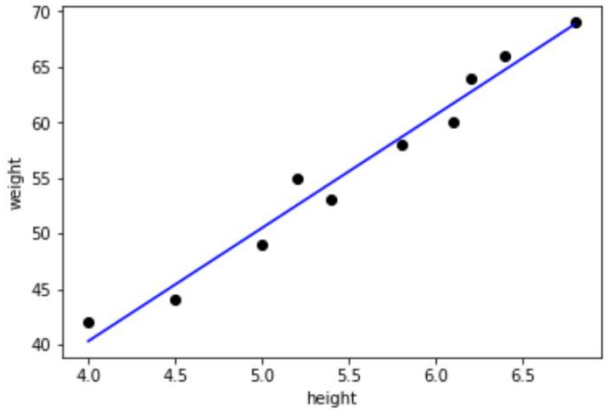
Dzīvē mēdz būt gadījumi, kad īsti nav iespējams noteikt kurš mainīgais ir atkarīgais un kurš neatkarīgais mainīgais. Abu ietekme vienam uz otru ir līdzvērtīga. Šādos gadījumos tiek izmantota saistītā regresija, kurā abi mainīgie tiek izteikti katrs savā vienādojumā. No tā izriet, ka vienai un tai pašai sakarībai var izveidot divus regresijas vienādojumus:

$$y = ax + b$$
$$x = ax + b$$

kur x viens atkarīgais mainīgais, y otrs atkarīgais mainīgais, a un b vienādojuma koeficienti.

No tā izriet, ka y = ax + b atbilstošā taisne grafikā ir saistīta ar x = ax + b taisni, un arī pretēji. Tomēr saistītajai regresijai ir būtisks mīnuss, piemēram, to ir grūti pareizi interpretēt [20, 21].

Gan pāra, gan saistītā regresija ir lineāras regresijas. Zemāk attēlā redzams korelācijas un lineāras regresijas piemērs.



1.5. att **Lineārā regresija[47]**

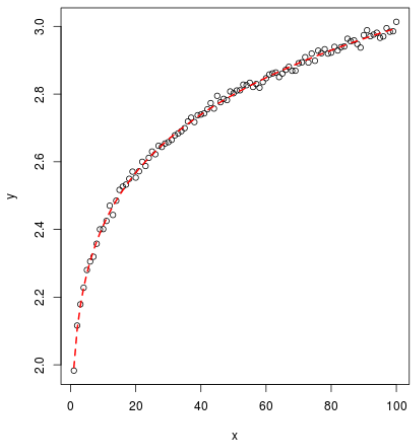
Attēlā Nr. 1.5. redzams tipisks regresijas piemērs, kurs starp diviem mainīgajiem novērojama lineāra sakarība, šajā gadījumā ķermeņa svars un garums. Pieaugot garumam pieaug arī svars. Līdz ar to, iespējams iegūt noteiktu lineāru vienādojumu, pēc kura iespējams prognozēt kāda cita objekta iztrūkstošos rezultātus.

1.6.4. Nelineārā regresija

Nelineārā regresija tā pat kā lineārā regresija, nosaka sakarību starp diviem mainīgajiem. Nelineāras regresijas gadījumā saistība starp mainīgajiem lielumiem nav nosakāma ar lineāras funkcijas sakarību. Šī regresija var pieņemt ļoti dažādas formas, bet visbiežāk tās ir kvadrātiskas vai eksponenciālas funkcijas [22, 26].

Vienkāršākais veids, kā saprast vai regresija ir nelineāra, ir burtiski vērot tās plūsmu, vai tā ir pielīdzināma taisnei, pretējā gadījumā, tā ir nelineāra regresija.

Pretēji lineāru sakarību gadījumā, nelineāru regresiju gadījumā vienam paredzamajam mainīgajam pretī var būt vairākas funkcijas vērtības. Piemēru nelineārai regresijai skatīt attēlā Nr. 1.6.



1.6.att. **Nelineāra regresija [27]**

Attēlā Nr. 1.6. redzams piemērs kurā attēlota logaritmiska sakarība starp regresijas datiem. Piemērā redzama regresija ir logaritmiska.

1.7. Līdzības saskaņošanas analīze

Kā vēl viena datu prognozēšanas metode, tiek izšķirta līdzības saskaņošanas analīze (no angļu val. similarity matching). Šī metode balstīta uz ideju, ka ja diviem objektiem ir līdzīgas kādas pazīmes, tad visticamāk, tiem būs līdzīgas arī citas pazīmes [24]. Šī metode ir noderīga, lai, piemēram, prognozētu nākotnes klientus kādā uzņēmumā, vai piedāvātu preču un pakalpojumu pircējiem internetā, nākotnes pirkumus [25].

Metodes svarīgākais uzdevums ir atrast raksturojošās pazīmes. Kad tās ir definētas, tad tās var pielietot dažādos datu apstrādes algoritmos, kas parasti ir balstīti uz noteiktiem noteikumiem, un tiek pielietoti lielam datu apjomam. Iztrūkstošu datu gadījumā, līdzības saskaņošanas metode, var tiek apvienota ar klasifikācijas algoritmiem, lai samazinātu iztrūkstošo datu objektus [24].

Šīs metodes svarīgākais atribūts ir datu kopa. Ir nepieciešams liels datu apjoms, kas tiek izmantots gan noteikumu sastādīšanai, gan datu prognozēšanai. Izplatītākās līdzību saskaņošanas metodes ir līdzību atlases metode, kas apskatīta iepriekšējā nodaļā, paraugu atpazīšanas metode, kas rezultātu noteikšanai izmanto iepriekš esošus paraugus un klasifikāciju veic pēc pilnīgas vai daļējas sakritības ar kādu paraugu. Ranga algoritmi, kuru notiekumu iespējamo varbūtību katrai klasei ietekmē iepriekš definēti rangi, un līdzības

noteikšanas metode, kas balstīta uz dažādu iezīmju savstarpējo algebrisko attālumu aprēķinu[23].

1.8. Kopsavilkums

Šīs nodaļas ietvaros ir izdevies apskatīt un izpētīt dažādas datu prognozēšanas metodes, kā klasifikācija, regresija, prognozēšana, datu salīdzināšana, kas katra paredzēta konkrētu problēmu risināšanai.

Tomēr, izvirzītā mērķa sasniegšanai, tiks pielietotas tās metodes, ar kuru palīdzību no vēsturiskiem datiem varēs prognozēt nākotnes piederību noteiktām klasēm, tātad klasifikācijas metode, un līdzīgo pazīmju pielīdzināšanas metode, lai dažādu lietotāju pieredzi varētu izmantot citu līdzīgu lietotāju datu prognozēšanā. Sīkāks pielietoto metožu apraksts konkrētās problēmas risinājumam aprakstīts nodaļā 4. PROGNOZĒŠANAS RĪKS.

2. PIRCĒJU UZVEDĪBAS ANALĪZE

Lai spētu prognozēt cilvēku potenciālos gala mērķus ceļojumos, ir nepieciešamas ne tikai tehniskas zināšanas, bet ir jānovērtē arī lietotāju uzvedība un paradumi konkrētajām dzīves situācijām. No tā, vai ir pareizi izprasta un izpētīta cilvēku uzvedība, konkrētās situācijās, ir atkarīgs, vai prognožu metožu izmantošanā tiek izmantoti korekti, noderīgi dati, kā arī vai tiek prognozēti nozīmīgi parametri. Šī darba ietvaros, paredzēts globāli konstatēt ceļotājus ierobežojošs faktorus.

Pētot cilvēku paradumus plānojot ceļojumus, ir svarīgi saprast galvenos cilvēka izvēli ietekmējošos faktorus, kas ietekmē gala mērķa izvēli. Kombinējot citu ceļotāju pieredzi, iespējams noteikt un prognozēt cita ceļotāja iespējamos galamērķus. Attiecīgi mērķis ir saprast, kādi ārēji ietekmējošie faktori var ļaut grupēt ceļotājus noteiktās grupās jeb kategorijās, pēc to pieredzes un ārēji ietekmējošiem faktoriem.

2.1. Izvēli ietekmējoši faktori

Cilvēka izvēli noteiktās dzīves situācijās vienmēr ietekmē dažādi ārējie un iekšējie faktori, kas balstās uz cilvēka iepriekšējo pieredzi. Līdz ar to, šie faktori tieši ietekmē cilvēka rīcību, tātad arī nākotnes rīcību. Šie ārējie faktori cilvēkam parasti jāsamēro ar savu gribu un interesēm, kā arī tie parasti būs noteicošie izvēles izdarīšanā. Izdarītās izvēles, var ietekmēt ģeogrāfiskais novietojums, ekonomiskā situācija, mērķis, laiks, cilvēka intereses, cilvēka iepriekšējā pieredze un citi.

Ceļojumu mērķu izvēle šī brīža ceļotājam ir kļuvusi ar vien pieejamāka un vieglāk realizējama, tieši ceļojumu galamērķu sasniegšanas ziņā, tātad cilvēka gribu vēl tiešāk ietekmēs ārēji ierobežojošie faktori. Šie faktori var būt ekonomiskie apstākļi, ģeogrāfiskais novietojums, ceļojuma mērķa faktors, laika ierobežojumi, cilvēka intereses un citi faktori[44].

2.2. Ekonomiskie apstākļi

Ekonomiskie apstākļi ir viens no primārajiem faktoriem, kas nosaka cilvēka izvēli un iespējas doties ceļojumā. Ekonomiskais stāvoklis visbiežāk ierobežos ceļotāju viņa izvēlē, kas attiecīgi katram ceļotājam var būt sava, atkarībā no ekonomiskā nodrošinājuma. Taču ir fundamentāla teorija, gadījuma lietderības maksimizēšana(RUM, no angļu val. Random utility maximization), kas nosaka, ka no pieejamo alternatīvu kopas, cilvēks vienmēr izvēlēsies sev izdevīgāko variantu, tieši pēc piedāvājuma lietderības faktora [44]. Šī teorija piemērojama arī ceļojumu izvēlē tieši no ekonomiskā viedokļa. Izdevīgumu, protams ne vienmēr var noteikt tik vienkārši, taču globāli skatoties, izvēle starp diviem ceļojumiem A un

B tiks izdarīta par labu izdevīgākajam tieši no ekonomiskā faktora. Faktiski ekonomisko izdevīgumu varam iedalīt divās apakškategorijās, pirmkārt, vai iespējams izdevīgi nokļūt iecerētajā galamērķi un otrkārt, vai ir izdevīgi dzīvošanas apstākļi iecerētajā galamērķī. To vai un kā katrs ceļotājs izvēlas kombinēt šīs izdevīguma kategorijas, nav nepieciešams zināt. Taču ir skaidrs, ka pēc savu konkrētu prasību atlases, izvēle tiks darīt par labu lētākajam piedāvājuma, kas atbilst ceļotāja izvirzītajām prasībām.

2.3. Ģeogrāfiskie faktori

Ģeogrāfiskajos faktoros ietilpst gan ceļotāja atrašanās vieta, gan attālums līdz iecerētajai ceļojuma vietai [44]. Piemēram, ceļotājam, kas dzīvo kāda no Amerikas Savienoto valstu štatiem, ir stipri lielāka varbūtība, ka nākamais ceļojuma galamērķis būs kādā no štatiem, nevis, piemēram, kāda no Eiropas valstīm. Ar to vēlos teikt, ka cilvēki daudz biežāk izvēlas sev ģeogrāfiski tuvākus ceļojumu galamērķus. Šo izvēli ietekmē gan ekonomiskais izdevīgums, gan dažādi apgrūtinājumi, kas saistīti ar nokļūšanu galamērķi, kā, piemēram, laiks, kas jāpavada ceļā.

Pie ģeogrāfiskajiem apstākļiem noteikti pieskaitāmi arī galamērķa klimats, attiecīgi vai valsts konkrētajā brīdī ir silta, auksta, vai valstī ir sniegs, karstums vai varbūt lietusgāzes. Klimats ir viens no svarīgiem izvēli ietekmējošiem faktoriem, kas tiek kombinēts gan ar konkrētās izvēles kalendāro laika periodu, gan ceļotāja ceļošanas mērķi [46].

2.4. Ceļojuma mērķa faktors

Lielu ietekmi uz mūsu ceļojumu plānošanu atstāj mūsu ceļojuma mērķis, ko ceļojumā vēlamies darīt, piemēram, doties pie dabas pārgājienos, vai apskatīt pilsētas. Veidojot ceļotāju nākotnes ceļojumu piedāvājumu ir svarīgi ņemt vēra ceļojuma mērķa faktoru, jo balstoties uz citu ceļotāju pieredzi ar līdzīgiem mērķiem, prognozes mērķa ceļotājs var uzzināt par sev vēl neatklātiem un neizzinātiem ceļojumu galamērķiem.

Ceļojuma mērķa faktors faktiski iedalāms divās lielās kategorijās, kā darba darīšanas un tūrisms. Tā kā mūs interesē izpētīt tieši atpūtas ceļojumus ietekmējošos faktorus, tad nepieciešams to iedalīt apakškategorijās kā atpūtas ceļojums, aktīvās atpūtas ceļojums, atpūta ar ģimeni, atpūta pie dabas, sporta brauciens, kulturālas izklaides brauciens un citi.

Izmantojot ceļojuma mērķi kā ārējo ietekmējošo faktoru, svarīgi saprast, ka ģenerējot nākotnes prognozes, nevar jaukt darba un atpūtas ceļojumus, jo darba braucienos ceļotāji dodas ne savas iniciatīvas vadīti, attiecīgi šāda informācija neko nepasaka par lietotāja gribas uzvedību, tādējādi veidojot apmācības datu kopu, šādi dati ir jāfiltrē.

Izvēloties papildus apakškategoriju, ceļotājs padara precīzāku potenciālo galamērķu piedāvājumu. Attiecīgi, tas būs tuvāks ceļotāj tā brīža interesēm un vēlmēm.

2.5. Laika perioda faktors

Ceļojuma laiks ir viens no ceļojuma raksturlielumiem, kas jo sevišķi ietekmē ceļotāja izvēli dodoties ceļojumā. Laiks ir izprotams gan kā jēdziens cik ilgi doties ceļojumā, gan kā jēdziens, kad doties ceļojumā. Gan ceļojuma ilgums, gan laika periods gada griezumā, parasti ir viens no pirmajiem nosakāmajiem ceļotāja kritērijiem, kurus ceļotājs nosaka pats, pirms uzsāk savu ceļojuma plānošanu [46].

Šis faktors faktiski ļauj iedalīt ceļojumus "īsajos" un "garajos" ceļojumos, kad tiek veikta ceļojumu atlase. Kā arī tas var būt kā papildus palīg kritērijs datu sagatavošanā, lai samazinātu iespējamo galamērķu skaitu. Tomēr jāatceras, ka laika periods faktiski nosaka tikai to, kad lietotājs vēlas doties ceļojumā, līdz ar to jaunu ceļojumu piedāvājumā šim faktoram nevajadzētu būtiski ierobežot ceļojumu piedāvājumu.

2.6. Intereses, īpašības

Nav noslēpums, ka līdzīgu interešu cilvēki mēdz daudz labāk atrast kopīgu valodu, kā tie, kuriem to ir visai maz vai nav vispār. Līdzīgi ir arī ceļotājiem. Ja kāds ceļotājs dodas ceļojumā uz kādu konkrētu galamērķi, tad citam ceļotājam ar līdzīgām interesēm varētu būt lielāka iespējamība arī doties uz šo galamērķi, kā citiem ceļotājiem [46].

Veidojot ceļotāja nākamā galamērķa prognozes, lieti noder zināt gan ceļotāja intereses, gan īpašības. Piemēram, kāds, kam ļoti patīk gozēties saulītē, neko nedarot, diez vai būs apmierināts ar aktīvu kalnos kāpšanu. Iespējams, kādreiz, taču tipiskākajos gadījumos, tā nebūs šāda ceļotāja pirmā izvēle. Tādējādi sastādot jaunus ceļojuma piedāvājuma datus, veidojot modeli, ja tas iespējams, svarīgi ņemt vērā arī ceļotāja intereses un īpašības, tādējādi samazinot iespējamo galamērķu daudzumu.

2.7. Nezināmu kategoriju lēmumi

Neskaitot iepriekšējās globālās kategorijas, katram no mums var būt neskaitāmi blakus faktori, kas var ietekmēt ceļojuma plānošanu. Piemēram, tas, kāda ir ceļojuma galamērķa virtuve vai laikapstākļi. Tas, kādā kompānija vēlamies doties ceļojumā, kādi būs pacelšanās un ielidošanas laiki, un vēl daudz citas kategorijas. Daudzus no šiem ietekmējošiem faktoriem sauc par "*Ad hoc*" kategorijām, kas nav primāri svarīgas, bet ir sasniedzamo mērķi ietekmējošas kategorijas. *Ad hoc* tulkojumā no latīņu valodas nozīmē *šim nolūkam*. Kas labi

paskaidro domu, ka konkrētie apstākļi faktori ir radušies tieši šajā reizē, plānojot tieši konkrēto mērķi. Tātad šim nolūkam.

Šīs kategorijas nav sistemātiskas, tās var būt nozīmīgas pēkšņi un tikai šajā reizē. Šīs kategorijas palīdz sasniegt mērķi, esošajā situācijā un pašas par sevi nav zināmas iepriekš. Piemēram, cilvēkam dodoties ceļojumā, pirmo reizi, lietas, kas jākrāmē koferī būs šīs kategorijas pārstāves. Cilvēkam nav iepriekš esošu paraugu, zināšanu uz kā balstīt savus izdarītos lēmumus, taču šīs kategorijas darbības tiešā mērā ietekmē sasniedzamo mērķi[28,29].

Raugoties uz šo kategoriju nedaudz no cita skata punkta, tad pētāmās problēmas gadījumā, jāņem vērā, ka dažādu maznozīmīgu faktoru kopums var būtiski ietekmēt ceļotāja lēmumus. Taču, šīs kategorijas var būt ļoti laba atslēga, līdzīgo pazīmju atrašanai savā starpā starp ceļotājiem. Attiecīgi kombinējot iepriekš apskatītos globālos ietekmējošos faktorus kā ģeogrāfiskie apstākļi, ekonomiskie apstākļi, u.c. ar *Ad hoc* kategorijām, visticamāk iespējams iegūt precīzākus rezultātus datu prognozēšanā. Tomēr, šai kategorijai datu prognozēšanā ir būtisks mīnuss, visticamāk iegūt datus par šīm kategorijām ir ļoti sarežģīti, attiecīgi tās pielietot un prognozēt nākotni var kļūt par ļoti sarežģītu uzdevumu.

2.8. Citi faktori

Starp citiem faktoriem ietilpst tādas kategorijas, kas nav ne pēkšņi parādošas, ne arī globālie ārējie faktori, tie drīzāk ir ceļotāju ierobežojoši faktori, piemēram, transporta ierobežojumi, vai specifiska ģeogrāfiskā atrašanās vieta, piemēram, kāda sala Norvēģijas ziemeļos, kuras iedzīvotājiem nav daudz līdzīgo ceļotāju.

Citi faktori vairāk jāņem vēra kā faktori, kas var būtiski ietekmēt ceļotāja galamērķi, taču nav iepriekš precīzi identificējami, kā arī otrajā gadījumā ir identificējami, bet nav ne lietotāja pieredzes, ne citu reālu datu, no kuriem piedāvājumu veikt. Šādi faktori, modelējot jaunu prognozi, faktiski var netikt ņemti vērā, līdz ar to piedāvājot lietotajam ne pārāk piemērotu galamērķi.

Svarīgs mūsdienu izvēli ietekmējošs faktors ir arī drošība ceļojot konkrētajā mērķa valstī. Drošību ietekmē gan dažādas slimības, gan militārās darbības valstī. Ļoti lielu ietekmi uz cilvēka ceļojuma galamērķi atstāj potenciālie riski, kas sagaida ceļotāju dodoties uz kādu valsti [45]. Šis gan ir mainīgs faktors, jo tas, vai konkrētajā valstī ceļot ir vai nav droši, laika gaitā mainās.

2.9. Kopsavilkums

Nodaļas ietvaros tika apskatīti svarīgākie cilvēka izvēli ietekmējošie faktori, kas var būt regulāri un var būt pēkšņi, pirmo reizi piedzīvoti. Ceļotāju galamērķu prognozēšanā pēc iespējas precīzākus rezultātus noteikti būs iespējams iegūt, ja būs pieejami pēc iespējas detalizētāki dati par ceļotāju. Taču ne vienmēr ceļojumu vietnēm tik detalizēta informācija ir pieejama. Līdz ar to ceļotāja nākotnes ceļojumu piedāvājums ir ļoti atkarīgs no tā, vai par lietotāju un citiem lietotājiem ir pieejami detalizēti dati un vai ceļotājs ir norādījis kādus globālos ārēji ietekmējošos faktorus, kā arī vai ceļotājam ir iepriekšēja pieredze ceļošanā. Pretējā gadījumā lietotāja piedāvājums var tikt ģenerēts uz minimāli pieejamo informāciju par lietotāju, piemēram, atrašanās vietu, ceļojuma laiku un ilgumu un var būt ļoti vispārīgs piedāvājums.

Galvenais secinājums - ceļojumu prognozei var būt tikai ieteicoša nozīmība, lai piedāvātu ceļotājam, iespējams, viņam atbilstošu ceļojumu klāstu, taču tas nevar būt strikti nosakošs, kā vienīgai pareizais variants. To ietekmē gan tas, ka ne vienmēr pieejama pamata informācija par ceļotāju, gan tas, ka nezinām *ad hoc* kategorijas faktorus, kam citreiz var būt ļoti būtiska nozīme.

3. DATU SAGATAVOŠANA

Datu prognozēšana nav iedomājama bez datiem, tas ir svarīgākais atribūts, prognožu veikšanā. Mūsdienās ikdienu tiek uzkrāts milzīgs datu apjoms - par cilvēku darbībām, ikdienu, finansēm, dabu, laikapstākļiem, ekonomiku un citām dzīvajai sistēmai svarīgajām lietām un procesiem. Tomēr dati var būt ne tikai skaitliskas tabulas, tie var būt teksts, attēli, arī skaņa. Svarīgi saprast ko tieši ar šiem datiem var izdarīt un kā tos pielietot un sagatavot, sava mērķa sasniegšanai. Piemēram, 2016. gada datos, Cisco pētījumā ir norādīts, ka datu apjoma plūsma internetā vienā mēnesī ir sasniegusi 90 exabaitus [29]. Un, tiek paredzēts, ka jau 2021. gadā globālā IP datu plūsma sasniegs 3.3 zettabitus [30]. Tātad datu apjoms ir milzīgs, līdz ar to spēja tos saprast un pielietot nākotnē būs nepieciešama aizvien vairāk.

Tiek uzskatīts, ka datu sagatavošana patērē līdz pat 80% no kopējā datu analīzes un prognozēšanas laika. Kā arī datu sagatavošanas procesā tiek pieļauta lielākā daļa kļūdu, nekorektai datu iegūšanai [29]. Tātad svarīgi novērtēt un izprast datu sagatavošanas procesu.

Nodaļas ietvaros tiek izpētīts kādu veidu dati var tikt pielietoti datu prognozēšanā, kā šos datus sagatavot, kā arī, kas jāņem vērā tos grupējot un veicot datu prognozes.

3.1. Datu tipi

Populārākais datu tips, ar ko ikdienā saskaramies ir teksts. Tekstuāli, ar burtiem vai skaitļiem fiksējam visdažādāko informāciju, piemēram, statistiskus datus. Šādā veidā fiksēti dati ir cilvēkam visnepārprotamākie, kā arī spēj skaidri un gaiši paust savu mērķi. Tomēr, arī skaņa un attēls ir neatņemama sastāvdaļa apkārtējās pasaules uztverei un izzināšanai. Arī datu prognozēšana var tikt veikta gan ar tekstuālu, gan vizuālu, gan pat ar audiālu informāciju.

Līdz ar to, nepieciešams izprast ko un kā katra tipa dati var tikt pielietoti datu prognozēšanā.

3.1.2. Tekstuāla informācija

Tekstuāla informācija, parasti konkrētu vienību veidā, ir populārākais datu prognozēšanas objekts. Tekstuāla informācija var saturēt gan personas datus, gan laiku, gan globālus tekstus. Šis informācijas uzglabāšanas veids ir sarežģītākais apstrādes objekts, jo parasti tekstuāla informācija ir lietotāja ievadīta un tās atkļūdošana var būt ļoti sarežģīts uzdevums. Kā arī tekstuāla informācija pati par sevi var būt ļoti dažāda.

Tekstuālas informācijas pirmais uzdevums ir datu kodēšana, attiecīgi pielāgošana noteiktām simbolu zīmēm. Nepareizi kodēts teksts, var radīt pilnīgi aplamu informāciju.

Tekstuālas informācijas problemātika var tikt risināta pēc sekojošām tehnikām:

1. Teksta virknes normalizācija;

2. Teksta virknes aptuvena atbilstība.

Teksta virknes normalizācijas uzdevums ir samazināt apstrādājamā teksta lieko simbolu skaitu. Tā sastāv no divām daļām, pirmkārt, aizvietojamās daļas noteikšana un atrašana, otrkārt atrastās frāzes aizvietošana. Piemēram, tā var būt lieko atstarpju aizvietošana ar vienu atstarpi. Teksta virknes aptuvenā atbilstība ir balstīta uz distances mērīšanu. Tekstā teikumi tiek savstarpēji salīdzināti pēc noteiktām metrikām. Ja to līdzība ir nosakāma, tad vienu no šiem teikumiem var neizmantot, tādējādi samazinot apstrādājamo datu apjomu [29].

Tomēr ļoti bieži tekstuāla informācija, kas tiek izmantota, ir nevis lielas tekstu kopas, bet gan konkrētas vērtības, tādēļ ir nozīmīgi saprast katru ievades lauku un izstrādāt tam atbilstošu pārbaudi. Piemēram, datums un laiks tekstuālā pieraksta formā tiek interpretēts ļoti dažādi. Ērta un laba prakse, ir glabāt datumu POSIX formā, kas ir laiks sekundēs no 1970. gada 1. janvāra pusnakts [29]. Šī ir skaitliska vērtība, ko ērti aprēķināt, atņemt, saskaitīt, kā arī izrēķināt gala rezultātu jauniegūtajam laikam. Šāda forma ļauj izvairīties no dažādības datuma un laika tekstuālajā pierakstā, kas var ieviest būtiskas kļūdas.

Skaitlisku lauku gadījumā vienmēr jāpārliecinās, vai skaitlis nesatur citus simbolus, kā arī, ja konkrētajai vērtībai ir nosakāmas vai zināmas robežas, tad jāpārliecinās vai šīs robežas netiek pārkāptas.

3.1.2. Vizuāla informācija

Vizuāla informācija ir attēls vai video fragments, kas ir attēlu virknējums. Attēlu parasti uztveram kā krāsainu vai melnbaltu objektu, kas viss kopā sniedz noteiktu informāciju. Piemēram, aplūkojot attēlu ar kaķi, jebkuram cilvēkam to aplūkojot būs skaidrs, kas attēlā attēlots, pēc redzamās formas, krāsas, detaļām. Taču digitāls attēls ir skaitliska informācija, kuru interpretējot dators spēj pārvērst cilvēkam saprotamā informācijā. Digitālu attēlu raksturo noteikti parametri, kas izvietoti noteiktā secībā. Attēlu raksturojošie parametri ir tā platums, un augstums, tā kanālu skaits un pikseļu vērtību pieraksta forma. Faktiski digitāls attēls ir skaitlisks daudzdimensionāls masīvs, kura informācijai iespējams piekļūt katram elementam, kā arī modificēt un analizēt tā elementus. Līdzīgi ir arī ar video fragmentiem. Video sastāv no daudz attēliem, kas noteiktā laika periodā tiek secīgi rādīti viens pēc otra. Tātad arī video fragmentus iespējams analizēt pēc to vienībām - attēla.

Datu prognozēšanas sakarā attēlu analīze tiek lietota, piemēram, cilvēka dzimuma un vecuma noteikšanai. Šādā tehnikā, ar noteiktu algoritmu palīdzību tiek atrastas cilvēku sejas attēlā, tās tiek sagrupētas pēc noteiktiem parametriem, piemēram, vīrietis vai sieviete. Tad tiek izstrādāts prognozēšanas modelis, kurš tiek apmācīts ar iepriekš saklasificētajiem datiem.

Kad modelis ir apmācīts, var tikt veikta datu, šajā gadījumā dzimuma, prognozēšana, jebkuram citam attēlam, kas satur cilvēka seju.

Ja iepriekš apskatītajās klasifikācijas metodēs saskārāmies ar skaitlisku datu prognozēšanu, tad faktiski arī attēls ir tā pati skaitliskā informācija, taču izvietota secīgās matricās, ar daudz vairāk parametriem, nekā klasiskos datu prognozēšanas uzdevumos. Šādas skaitliskas matricas iespējams pielietot datu prognožu algoritmiem, vistipiskāk tieši klasifikācijas metodēs.

3.1.3. Skaņas informācija

Skaņās informācija ir skaņa, ko dzirdam. Digitalizēta skaņa sastāv no digitalizētiem signāliem, kas izvietoti secīgi un kodēti noteiktā formā. Skaņa, ko dzird cilvēka auss, ir vibrācija, kurai iespējams piešķirt konkrētus raksturlielumus, kā frekvenci, viļņa garumu, skaņas ātrumu, periodu, u.c. Kad skaņa tiek ģenerēta, tā pārvietojas viļņu veidā, kas vibrācijas rezultātā rada noteiktu frekvenci. Frekvence ir skaņas svārstību skaits laika periodā T. Zemākām skaņām ir garāks svārstību vilnis, taču augstākām skaņām īsāks[32, 33]. Tātad digitalizēta skaņa sastāv no signāliem, kurus iespējams raksturot ar skaitlisku datu palīdzību. No tā iegūstam, ka arī skaņas informāciju iespējams analizēt un pielietot dažādu klasifikācijas uzdevumu veikšanai.

Skaņas informācija datu prognozēšanā tiek izmantota dažādu valodu risinājumu izpētē. Piemēram, runātā teksta pārvēršanai rakstītā tekstā, dažādu balss komandu realizēšanai, audio tulkošanai un citām metodē [31]. Šajā gadījumā to nelīdz galam var saukt par datu prognozēšanu, taču tā izmanto tās pašas vai līdzīgas metodes rezultātu atrašanai un noteikšanai, kā klasifikācijas metodes, kas tiek pielietotas datu prognozēšanā. Līdzīgi arī skaņas informācijas analīze, piemēram, runātā teksta atpazīšanai, strādā pēc līdzīga principa. Tā ir apmācīta ar noteiktu algoritmu, un balstoties uz iepriekš apmācītajiem datiem un atpazīšanas algoritma, nosaka jaunās informācijas piederību kādai klasei. Piemēram, valodas modeļi ir apmācīti ar dažādu balsu un izrunu variācijām, taču katrs cilvēks, kurš vienu un to pašu vārdu teiks no jauna, bet nav piedalījies modeļa apmācībā, būs kas jauns un nezināms, tas, kas būs jāprognozē vai jāklasificē. Skaņas signālu datus modeļu apmācībai arī izvieto matricās, jeb daudzdimensionālos masīvos, kur katrā masīvs satur informāciju par skaņas viļņiem noteiktā laika vienībā [34].

3.2. Datu sagatavošana

Jaunu datu prognožu veidošana būtu jāuztver kā jebkurš cits projekts, kas prasa sagatavošanos un analīzi. Arī datu prognozēšanā ir vairāki secīgi procesi, kas jāveic, veiksmīgu rezultātu sasniegšanai. Tie ir datu izprašana, datu saturiskās jēgas noteikšana, datu atlase un filtrēšana, kā arī datu apjoma analīze.

3.2.1. Datu interpretācija

Pirmais un svarīgākais solis datu sagatavošanā, ir mērķa nospraušana, jeb patiesībā biznesa analīzes veikšana [29]. Ir jāizprot datu analīzes nepieciešamība, kā arī jānosaka kā noteikt, vai rezultāts tiek vai netiek sasniegts. Šajā procesā ir svarīgi iesaistīt nozaru pārstāvjus, kas pārzina un spēj analizēt konkrētu problēmu datus, spēj sasaistīt sakarības starp pieejamajiem datiem.

Svarīgi ir izprast, ka datu prognozēšana ir tikai papildinājums, jeb palīgs mērķu sasniegšanai. Veidojot datu prognozēšanas stratēģiju, jādomā kritiski un visiem iesaistītajiem jāizprot, ka tā nevar sniegt atbildes uz tādiem jautājumiem, kā piemēram, kā mūsu uzņēmums sasniegs tādu peļņu [29]. Tā nenesīs atbildes uz jautājumiem kā sasniegt konkrētus mērķus, taču tā var palīdzēt uzzināt, kā konkrēti apstākļi var ietekmēt mērķu sasniegšanu.

Izstrādājot datu plānošanas tehnoloģiju, sākotnēji ir skaidri jāsaprot mērķi un jāspēj atbildēt uz šādiem jautājumiem:

1. Ko mēs vēlamies sasniegt ar šo tehnoloģijas izstrādi?
2. Uz kādu jautājumu tā dos atbildi?
3. Kā noteikt vai tehnoloģijas rezultāti ir snieguši veiksmīgu vai neveiksmīgu rezultātu?

Šie trīs jautājumi ir ābece, datu prognozēšanas sākuma posmam. Lai spētu noteikt rezultātu sasniegšanu, jānosaka arī hipotēze, kas norāda uz vēlamo mērķu sasniegšanu[29].

Vēl būtisks solis ir izvērtēt kā iegūtie dati būs pielietojami[29]. Var tikt izstrādāta ļoti vērtīga datu prognozēšanas un plānošanas stratēģija, taču tai nebūs nekādas jēgas, ja to nevarēs pielietot. Ir jāizstrādā stratēģija iegūto rezultātu pielietošanai tā, lai tos var ērti un vienkārši izmantot, piemēram, uzņēmuma darbinieki.

Veidojot lielas datu plānošanas stratēģijas, jāņem vērā faktors, ka jaunā sistēma jāspēj ērti integrēt esošajās sistēmās [29]. Kāpēc tas ir svarīgi? Visticamāk, ka katra šāda jauna datu plānošanas stratēģija tiek izstrādāta kāda uzņēmuma konkrētu mērķu sasniegšanai. Tātad šai kompānijai jau ir pieejami dati, iespējams pat no vairākām sistēmām. Dati ir apjomīgi un var būt izvietoti uz vairākām atsevišķām instancēm, kas risinājumā jāspēj apvienot un izdarīt nekļūdīgus secinājumus. Tātad šis ir ļoti svarīgs solis tieši stratēģijas izplānošanas procesā.

3.2.2. Datu uzglabāšanas veidi

Kā jau iepriekš nodaļā apskatīts, dati var būt dažādās formās, taču tie var būt arī dažādi uzglabāti. Tie var būt neapstrādāti (no angļu val. raw data), tehniski korekti dati (no angļu val. technically-correct data), agregāti dati, sakāroti dati, saspiesti dati vai formatēti dati[29].

Tehniski neapstrādāti dati ir tieši tādi, kādi tie tiek saņemti. Tie netiek apstrādāti, tie var saturēt kļūdas, piemēram, dažādus datu tipus. Šādi dati ir faktiski nelietojami, ja tie netiek apstrādāti, taču to pluss ir salīdzinoši ātra datu ielasīšana un rakstīšana [29].

Tehniski apstrādāti dati ir sakārtoti tehniski neapstrādātie dati. Brīdī, kad ienākošie dati ir ieguvuši datu tipus, vai sakārtoti, tie kļūst par tehniski apstrādātiem datiem, taču tas nenosaka, ka šie dati ir līdz galam korekti [29].

Sakāroti dati ir tādi dati, kuri tehniski sakāroti kādā noteiktā struktūrā, kuru iespējams ērti lasīt, interpretēt un pielietot citām sistēmām [29]. Šādus datus visvieglāk izprast kā .csv vai .excel datus. Dati ir sakārtoti rindās un kolonās, kas ir aizpildītas ar noteiktam vērtībām.

Datu analīzē mēdz tikt pielietoti arī agregāti dati. Gadījumos, kad datu apjoms ir liels vai rezultātus ietekmē noteikta laika līnija, dati tiek agregāti, daļa no vēsturiskajiem datiem tiek samazināta. Tā rezultātā tiek iegūti saspiesti dati [29].

Ir laba prakse datus sagatavot un saglabāt katru savā fāzē, tādējādi izvairoties un iespējamajām kļūdām, kā arī to gadījumā vieglāk atrast to cēloni.

3.2.3. Datu atlase

Datu atlases un sagatavošanas solim gala rezultātā būtu jābūt automatizētam, taču, lai spētu to automatizēt, jāizprot potenciālie datu sagatavošanas veidi.

Kad zināms kādi dati tiks izmantoti problēmas risināšanā, katram ievades laukam būtiski noteikt tā formu [29]. To, kā sagatavot un validēt datus noteiktām formām, var skatīt nodaļā 3.1.1 Tekstuāla informācija.

Faktiski datu atlases solis sevī ietver izanalizēto datu filtrēšanu pēc to lauku vērtībām, iztrūkstošajām vērtībām, un iegūšanu no to atrašanās vietām. Šis solis var būt ļoti laikietilpīgs, jo prasa korektu izvēlētās atlasāmās informācijas lauku validācijas mehānismu izstrādi, kā arī korektu datu interpretāciju. Protams, datu atlases laiks atkarīgs arī no atlasāmo datu apjoma.

3.2.4. Atlasāmo datu filtrēšana

Kad dati ir gatavi statistiskai analīzei, tie ir noteikti. Lai sasniegtu šādus datus, ir jāveic datu "tīrīšana". Tas nozīmē, ir jāizdzēš liekie simboli, un liekās vērtības. Ja iespējams, tad jāveic iztrūkstošo vērtību ievietošana [29]. Šajā solī dati ir tehniski korekti, tie ir sakārtoti,

taču tajos var būt nekorektas vērtības. Tas nozīmē, ka, ja vēlamies noteikt vai konkrētajam cilvēkam ir droši dot aizdevumu bankā, un kāda no atlases vērtībām ir iztrūkstoša, tad vai nu šis datu ieraksts ir jāizņem no apmācības kopas, vai arī tā tukšā vērtība jāizvieto, ja tas ir iespējams.

Datiem savā starpā konkrētajā atlasē jābūt noteiktiem un ar savu mērķi. Tie nedrīkst savstarpēji konfliktēt. Datu noteiktība ir iedalāma trīs tipos:

1. Ieraksta noteiktība (no angļu val. In-record consistency)
2. Savstarpējā noteiktā noteiktība (no angļu val. Cross-record consistency)
3. Savstarpējo datu kopu noteiktā noteiktība (no angļu val. Cross-data-set consistency)

Ieraksta noteiktība nozīmē to, ka analizējamā informācija ir stingri noteikta un glabājas konkrētā ierakstā. Savstarpējā noteiktā noteiktība nozīmē, ka statistiski dati konkrētajā datu kopā savā starpā nekonfliktē. Savstarpējo datu kopu noteiktā noteiktība nosaka, ka dažādu domēnu datu kopas savstarpēji ir atbilstošas un var tikt savstarpēji analizētas [29].

Viena no populārākajām problēmām datu kopās ir iztrūkstošie dati. Kādas vērtības iztrūkšana konkrētam ierakstam ir normāla un ikdienišķa parādība, kas izstrādājot datu prognozēšanas tehnoloģiju, jāņem vērā.

Vairākas datu apstrādes metodes spēj tikt galā ar iztrūkstošajām vērtībām, taču ne visas. Kā arī nozīmīgi ir noteikt šo iztrūkstošo vērtību ietekmi[29]. Viens no veidiem, kā risināt šo problēmu, ir noteikt noklusējuma vērtības, piemēram, izveidot "nezināms" kategoriju, taču ne vienmēr tas ir iespējams, tieši datu nozīmības dēļ.

Iztrūkstošās vērtības tiek iedalītas divās kategorijās - vērtība nav zināma datu kopai, vai vērtība ir zināma datu kopā, bet nav zināma konkrētajam ierakstam. Iztrūkstošo vērtību gadījumā jāņem vēra tieši pirmais gadījums. Tipiskākā un nekļūdīgākā metode iztūkstošo vērtību novēršanai ir šo ierakstu neiekļaušana datu kopā. Tomēr ja procentuāli iztrūkstošo datu ir vairāk kā neiztrūkstošo, tad būtu jāizvēlas piedāvājuma metode (no angļu val. Imputation Techniques) [29].

Piedāvājuma jeb inkriminēšanas metode nosaka iztrūkstošās vērtības, balstoties uz citām ieraksta vērtībām. Visvienkāršākais šīs metodes risinājums ir vidējā vērtība, taču tā ir ērti pielietojama tikai skaitlisku datu gadījumā. Otra metode ir koeficienta aprēķināšanas metode. Tā piedāvā aprēķināt nezināmo X_i vērtību, pēc vidējām X vērtībām un y attiecībām. Tipiski šī vērtība tiek iegūta summējot visus kopas x elementus un dalot ar visu kopas y elementu summu [29].

Cita piedāvājuma metode piedāvā iespēju aizpildīt nezināmo vērtību, pēc cita līdzīga ieraksta parauga [29]. Attiecīgi, ja vienam ieraksta nav zināms lielums Z, bet citam, ar līdzīgu saturu, ierakstam tas ir zināms, tad nezināmā vērtība tiek aizvietota ar zināmo Z lielumu.

Patiesībā šī metode šķiet ļoti interesanta un varētu tikt kombinēta ar vidējo vērtību no līdzīgajiem ierakstiem. Tomēr šajā metodē ir būtiska problēma, līdzīgo ierakstu noteikšana. Viens no risinājumiem ir izvēlēties sakārtot kopu, pēc kāda no zināmajiem parametriem un izvēlēties vienu no tuvumā esošajiem.

Vēl tiek piedāvāta tuvākā kaimiņa metode, kas tiek aprēķināta ar noteiktu distances formulu, izmantojot pieejamos ierakstu parametrus. Distances formulas var būt dažādas. Tās nosaka, kā izmantot konkrētos ieraksta laukus, lai aprēķinātu vienu līdzības koeficientu. Šo distances formulu pielieto tuvākajiem n kaimiņiem, tādējādi iegūstot lielāku datu kopu no kā izvēlēties iespējamo vērību. Tomēr arī šeit jāņem vērā skaitlisko datu ierobežojums. Šajos gadījumos, ja iztrūkstošā vērtība ir skaitliski dati, tad tiek ņemta vidējā vērtība no n tuvāko kaimiņu distanču aprēķina, gadījumos, ja tā ir kategorija, tad tiek izvēlēta tā kategorija, kuras ierakstam bija vismazākā distance ar konkrēto ierakstu.

Jau iepriekš nodaļā minēts, ka vērtības var būt kļūdainas, attiecīgi, var gadīties, ka vecums ir negatīva vērtība vai tieši nesamērīgi liela vērtība. Šādas kļūdas datu sagatavošanā ir nepieļaujamas un tās ir jānovērš. Datu nesakritības ir viegli konstatēt un filtrēt atbilstoši izstrādājot noteiktus noteikumus katram datu laukam. Automatizēta datu filtrēšana pēc noteiktiem nosacījumiem ļauj paātrināt datu sagatavošanu un ir relatīvi vienkārša, taču sarežģīti ir automatizēt noteikumu izveidi. Var gadīties arī tādi lauki, kuri ir savstarpēji saistīti un to robežvērtības ietekmē to savstarpējā attiecība [29]. Šie gadījumi ir jāizsver un jāaplūko datu analīzē jau pašā sākumā. Jāizpēta iespēja izveidot noteikumus tā, lai tie jāizstrādā vienreiz, un sistēmā ģenerējot jaunus mērījumus un datu prognozes, laukiem jau ir zināmi noteikti filtrēšanas nosacījumi.

Bieži ir sastopamas vērtības kā bezgalība vai NaN(nav skaitliska vērtība), šie gadījumi sastopami dažādu statistisku prognožu gadījumā. Ieraksti, kas satur šādas vērtības arī ir jāizņem no datu kopas. Konkrētu stratēģiju gadījumā vēlams no datu kopas izmest arī ekstrēmo gadījumu datu ierakstus. Šādiem ierakstiem var būt ļoti liela ietekme uz prognozējamo datu kvalitāti[29].

3.2.5. Datu apjoma limitēšana

Strādājot ar lielu datu apjomu, agrāk vai vēlāk rodas vajadzība ierobežot vai samazināt analizējamo datu apjomu. Jau iepriekš minēts, ka viena no metodēm ir pilnīga vai daļēja vēsturisko datu izmešana un datu kopas, taču tas nav vienīgais veids kā var un vajag samazināt apstrādājamo datu daudzumu. Uz šo jautājumu var skatīties arī no citas puses. Gadījumos, kad jāstrādā ar vairākām datu bāzēm, var būt situācijas, kad starp divām datu

bāzēm ir dati, kas atkārtojas. Šajos gadījumos apvienojot datus, ir jāveic datu tīrīšana likvidējot dublikātus, tādējādi samazinot datu apjomu.

Šim nolūkam var tikt veikta kartēšana (no angļu val. mapping), tas ir process, kurā tiek konstatētas sakarības starp datu atribūtiem un to vērtībām. Kā arī jāveic atribūtu atbilstības meklēšana (no angļu val. entity matching). Tas ir process, kurā tiek meklēti dažāda satura ieraksti, kas attiecas uz vienu un to pašu atribūtu. Šī metode ir jo sevišķi noderīga dublikātu likvidēšanā [29].

Tomēr veicot datu apjoma samazināšanu un tīrīšanu, ir jābūt ļoti uzmanīgam, jo palielinoties datu apjomam, palielinās arī varbūtība, ka dažādu uzņēmumu ietvaros vieni un tie paši dati tiek grupēti citādāk. Šo gadījumu detektēšana un tīrīšana var būt ļoti laikietilpīgs un dārgs process, taču, ja ir zināms, ka šāda situācija ir, tā ir jārisina.

3.3. Datu vizualizācija

Atlasītie un sakārotie dati tālāk tiek nodoti kādam no izvēlētajiem datu prognozēšanas un apmācības algoritmiem. Katrs no šiem algoritmiem atgriež rezultātus, taču šie rezultāti ir definēti specifiskās formās un visbiežāk sistēmas gala lietotājam nesaprotamā formātā. Tādēļ, veicot datu prognozēšanas sistēmas izstrādi, ļoti svarīgi ir iegūto rezultātu attēlot lietotājam viegli uztveramu un saprotamu, tā, lai datu analīzes speciālisti varētu nekļūdīgi novērtēt rezultātus.

Sistēmas analīzes procesā vēl viens svarīgs jautājums ir - kā dati tiks atspoguļoti? Kā arī, kas lietos iegūtos datus? Svarīgi zināt kāda līmeņa lietotāji datus lietos. Vai datus nepieciešams integrēt kādā citā sistēmā noteiktā formā, kas jau pati veiks datu vizualizāciju, vai arī jāparedz datu vizualizācija pašiem. Var gadīties, ka sistēmas lietotājiem nepieciešama ļoti detalizēta datu vizualizācija, lai veiktu nepieciešamo analīzi, bet var gadīties, ka sistēmas lietotājam vajadzīga tikai īsa kodolīga atbilde[29]. Tātad veicot sistēmas plānošanu, šiem jautājumiem ir jābūt skaidri definētiem, pretējā gadījumā iegūtā informācija var būt vērtīga, taču nekam nederīga, jo neviens nevarēs to pielietot.

Gadījumos, kad rezultātus nepieciešams vizualizēt, svarīgākais rezultātu atspoguļošanā ir to uztveramība. Bieži, lai cik laba būtu izstrādātā datu prognozēšanas un analīzes metode, dati tiek atspoguļoti nesaprotami vai sarežģīti uztverami. Gala lietotājam nevajadzētu pavadīt daudz laika, lai iedziļinātos un uztvertu informāciju. Tai jābūt atspoguļotai tieši un nepārprotami.

Datus iespējams atspoguļot ar dažādu tehniku palīdzību, galvenais ir laikus konstatēt atspoguļojamo datu nepieciešamību. Dati var tikt atspoguļoti tikai atbildes gadījumā, piemēram, ja vēlamies noskaidrot, vai dot aizdevumu klientam, tad mūs parasti interesē tikai

atbilde jā vai nē. Šādā gadījumā nav nepieciešama sarežģīta datu atspoguļošana, taču dati var būt arī sarežģītāki, piemēram, dažādu ekonomisko datu prognožu sakarā, tad nepieciešams veikt gan vēsturisku, gan prognožu datu vizualizāciju.

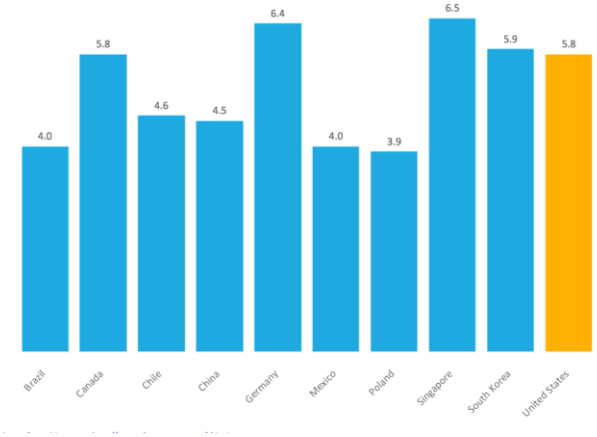
Populārākās datu atspoguļošanas metodes ir līniju diagrammas, sektorveida diagrammas, stabiņu diagrammas, tabulas, savienojumu diagrammas, simbolu kartes[35,36].

Līniju diagrammas labi atspoguļot laika līniju gadījumos, kad noteiktā laika perioda notikušas noteiktas izmaiņas[36].

3.1. att. **Līniju diagramma[37]**

Attēlā Nr. 3.1. redzams klasisks līniju diagrammas piemērs, kurā saprotami nolasāma laika un mērījumu sakarība un vērtības.

Joslu jeb stabiņu diagrammas ieteicams izmantot gadījumos, kad nepieciešamas parādīt noteiktu skaitu pret kādu grupu vai klasi[36].

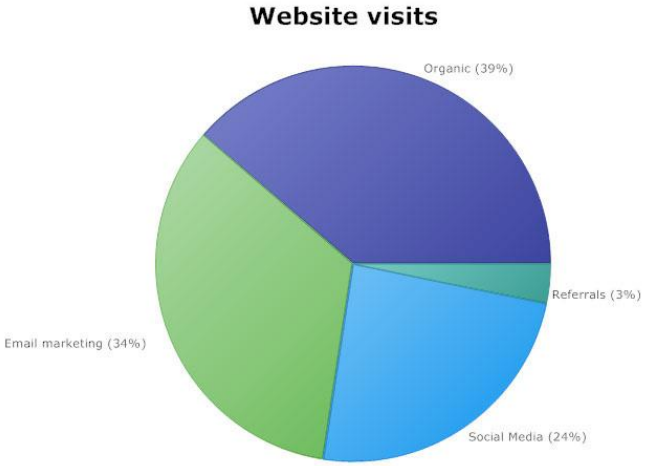


3.2. att. **Stabiņu diagramma[38]**

Kā redzams attēlā Nr. 3.2., katra valsts ir kā kategorija, kurai ir noteikts skaits vienību. Uzskatāmi un saprotami atspoguļota informācija. Šāda tipa diagrammas var izmantot arī ja

nepieciešams vienlaicīgi atspoguļot dažādu vienību skaitu konkrētā laika periodā. Piemēram, atspoguļot sieviešu un vīriešu skaitu valstī pa gadiem.

Sektorveida diagrammas tipiski tiek izmantotas, lai parādītu proporcionālo sadalījumu starp vairākām vienībām, kas attiecas uz vienu atribūtu[36].

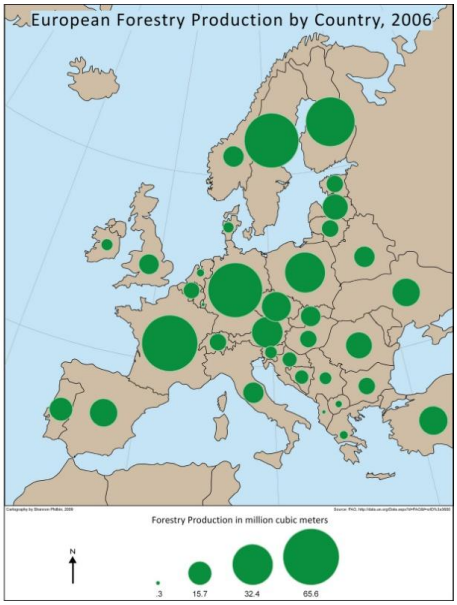


3.3.att. **Sektorveida diagramma[39]**

Sektorveida diagramma uzskatāmi atspoguļo procentuālo sadalījumu starp kādām vienībām. Piemēram, kā redzams attēlā Nr. 3.3., atspoguļots procentuālais sadalījums starp interneta vietņu apmeklējuma iemeslu. Sektorveida diagramma var būt riskanta izvēle gadījumos, kad tā sadalīta ļoti sīkās vienībās. Šādos gadījumos labāk izvēlēties citu metodi, piemēram, stabiņu diagrammu.

Tabulas ir klasisks datu atspoguļošanas veids, taču lielu datu apstrādes gadījumā tabulā atspoguļojamos datus nepieciešams iedalīt datu grupās, tādējādi samazinot atspoguļojamo rezultātu. Tabulas ir diezgan detalizēta līmeņa datu atspoguļošanas veids, tādēļ tās labāk izvēlēties gadījumos, kad zināms, ka auditorija tās spēs uztvert.

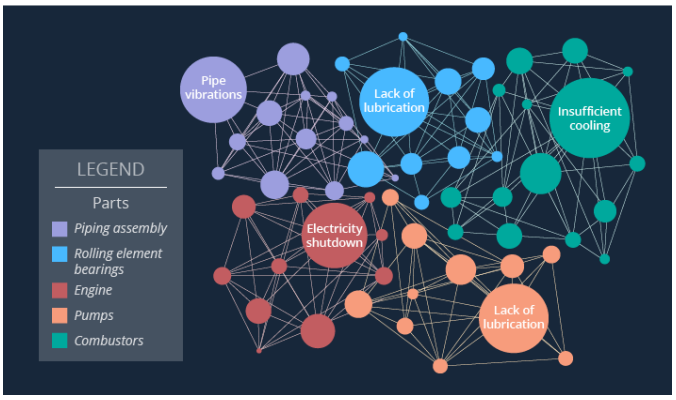
Simbolu kartes parasti izvēlas, lai atspoguļotu noteiktu īpašību sadalījumu ģeogrāfiskos reģionos[36].



3.4.att. **Simbolu diagramma[40]**

Kā redzams attēlā Nr. 3.4., tad katri Eiropas valstij tiek piemērota noteikta daudzuma īpašība. Šādas diagrammas ir ļoti ērti uztveramas kā arī informāciju saturošas. Tās vizuāli ļauj skatītajam ātri uztvert kurām valstīm konkrētā īpašība ir vairāk vai mazāk, tātad savā ziņā ļauj ātri, neiedziļinoties klasificēt daudzuma grupas un tai pat laikā, nedaudz iedziļinoties, skatītājs var ātri uztvert arī detalizētāku īpašības daudzuma informāciju.

Savienojumu diagrammas izmanto datu atspoguļošanā gadījumos, kad nepieciešams redzēt sakarību starp datiem [36].



3.5.att. **Savienojumu diagramma[36]**

Attēlā Nr. 3.5. redzama savienojumu diagramma. Šī ir tipiski mazāk lietotā metode datu atspoguļošanā, taču tā var būt noderīgi. Šīs metodes izvēlē jābūt uzmanīgam, lai

nesamudžinātu datus, kā arī padarītu relatīvi sarežģīto vizualizācijas veidu interesantu skatītājam.

Izvēloties atspoguļošanas metodi, svarīgi to izvēlēties atbilstoši biznesa vajadzībām. Attiecīgi, jāizvērtē gan skatītāju auditorija, gan mērķis, ko galu galā ar šo datu prognozēšanu un atspoguļošanu vēlamies pateikt.

3.4. Datu prognozēšanu ietekmējošie faktori

Datu prognozēšanas procesā ir daudz ietekmējošu faktoru. Pirmkārt, ļoti svarīga ir datu kvalitāte, kuras uzlabošana un nodrošināšana aprakstīta nodaļa jau iepriekš. Taču kā jau zināms, tad datu prognozēšanas pamata atribūts ir paši dati. Ir svarīga to tīrība, taču vēl svarīgāka ir to pieejamība. Datu pieejamības problēmas var būt apskatāma no dažādiem skata punktiem, tie var būt maza apjoma, tie var būt nepieejami, vai arī tie var būt daļēji pieejami.

Piemēram, kompānija Google savā rakstā 2009. gadā uzsver, ka viņu straujā attīstība norisinās pateicoties datu apjoma pieaugumam un pieejamībai. Viņi spējuši strauji attīstīt dažādus valodu risinājumus, gan tulkošanu, gan runātā teksta atpazīšanu daudz labāk, lielākoties nevis tieši algoritmu uzlabošanās dēļ, bet pateicoties lielākai datu pieejamībai [31].

Tātad ir skaidrs, ka pēc iespējas efektīgāku datu analīzi iespējams veikt ar lielu datu apjomu. Līdz ar to datu apjoms un pieejamība kļūst par ierobežošu faktoru datu prognozēšanā. Taču tā labai un jaunai idejai nekomerciālos apstākļos var būt liela problēma, iegūt datus, kas nav publiskie pieejami.

Lielākoties konkrētos uzņēmumos, kas izvēlējušies izstrādāt savu datu analīzes un prognozēšanas mehānismu, datu pieejamība nav šķērslis, jo šie uzņēmumi var veikt datu analīzi balstoties uz sev pieejamajiem klientu datiem. Tomēr, arī tad var parādīties dažādi datu ierobežojumi, piemēram, uzņēmumam ir vēlēšanās risināt un prognozēt kādu konkrētu problēmu, taču tās risināšanai vai nu nepietiek datu vai nemaz nav atbilstošu datu. Tad atliek vien izstrādāt stratēģiju vai, kā un cik ilgā iespējams iegūt vēlamos datus.

3.5. Datu analīze noteiktā laika periodā

Veicot datu prognozēšanu, parasti vēlamies no vēsturiskiem datiem prognozēt jaunus, potenciālos datus, taču, rodas jautājums, kā analizēt datus, kas laika līnijā ir nepārtraukti, kā arī cik senus datus drīkstam analizēt?

Pirmkārt, jāsaprot, ka datu prognozēšanas problēmai nav universāla risinājuma. Dažādiem biznesa mērķiem būs jāizstrādā savas datu analīzes stratēģijas. Šo stratēģiju rezultātā tiek veikta analizējamo datu analīze ne tikai no datorikas viedokļa, bet arī no biznesa viedokļa. Tātad viens no veidiem, kā noteikt kādu laika periodu ņemt vērā, ir atkarīgs no tā,

kā nozares analītiķi to nosaka, vai nu balstoties uz iepriekšējiem novērojumiem, vai balstoties uz noteiktām prasībām.

Tomēr ne vienmēr biznesa prasības var noteikt, kādi ir analizējamie laika periodi. Datiem, kas tiek prognozēti pēc notiktiem laika intervāliem, un norit nepārtraukti, jāveic laika sērijas analīze (no angļu val. time series analysis)[42]. Laika sēriju dati parasti ir tādi dati, kuriem viens atribūts ir laiks, bet otrs ir mērāmais atribūts noteiktā laika periodā. Šajā gadījumā dati ir nepārtraukti, tātad katram laika momentam ir pielāgojams konkrēts atribūts [29]. Piemēram, cilvēka veselības noteikšanai, pēc viņa sirdsdarbības ritma, analizējamie dati ir labs laika sērijas datu gadījuma piemērs[43]. Vēl līdzīgi gadījumi ir laikapstākļu prognozēšana, vai cilvēku veiktie peles klikšķi kādā interneta vietnē. Laika sērijas dati ir tādi dati, kas korektai datu prognozēšanai jāanalizē pēc noteiktiem laika periodiem, piemēram, diena, mēnesis vai gads.

3.6. Datu proporcionalitāte

Strādājot ar lielu datu apjomu, relatīvi sarežģīts uzdevums ir datu proporcionalitātes sadalījuma ievērošana. Ievācot analizējamos datus, ir būtiski piedomāt par datu proporcionalitāti prognozējamo datu kvalitātes nodrošināšanai[41]. Ir svarīgi, lai prognozējamo datu, piemēram, konkrētu klasifikācijas problēmu gadījumā, konkrētu klašu pārstāvju skaits būtu proporcionāls. Piemēram, ja vēlamies apmācīt kādu no prognozēšanas algoritmiem atpazīt cilvēka dzimumu pēc to sejām, tad gan sieviešu, gan vīriešu klasēs, datu kopām jābūt līdzvērtīgām - vienlīdz lielām. Pretējā gadījumā iespējams iegūt neprecīzu prognožu modeli, kas datu apmācības procesā būs vairāk apmācījies atpazīt vienu vai otru klasi, tādējādi radot iespēju, ka prognožu modelis kļūdīsies daudz biežāk.

Tomēr ir skaidrs, ka ir gadījumi, kad vai nu datu nepietiekamības dēļ, vai konkrētu novērojamu sakarību dēļ dati nav proporcionāli. Šādos gadījumos jāveic papildus dziļāka analīze un jāizsver vai nu konkrētu klašu izmešana no analīzes, vai jāveic mākslīga datu palielināšana, vai tieši pretēji, samazināšana.

Datu proporcionalitāti svarīgi ievērot tieši pret to lielumu, kuru vēlamies prognozēt[41].

3.7. Kopsavilkums

Izstrādājot datu prognozēšanas stratēģiju, pirmais solis ir datu izpēte un analīze. Lai izdarītu veiksmīgu un derīgu datu prognozēšanu, svarīgi datus pareizi sagatavot. Jāievēro virkne svarīgu noteikumu, piemēram, datu korektums, vērtību atbilstība konkrētiem nosacījumiem, jāveic dublikātu dzēšana, jāveic mākslīga datu apjoma palielināšana vai tieši pretēji - samazināšana, nepieciešamības gadījumā, un citi datu kvalitāti ietekmējoši faktori.

Svarīgākais datu sagatavošanā, ir saprast ko tieši ar šiem datiem vēlamies prognozēt un kādam mērķim tie tiks lietoti, tad arī būs iespējams izstrādāt gan korektu prognozēšanas stratēģiju, gan pareizi sagatavot prognozēm pielietojamos datus. Kad dati iegūti, jāatceras arī par pareizu datu atspoguļošanu, tā, lai tie ir uztverami un saprotami lietotāju auditorijai, pretējā gadījumā izstrādātā stratēģijas metode ir nederīga, jo nesniedz iecerēto.

4. PROGNOZĒŠANAS RĪKS

Iepriekšējās nodaļās ir apskatītas dažādas datu prognozēšanas metodes un algoritmi, kurus iespējams pielietot dažādu problēmu risināšanā. Piemēram, bankas aizdevumu noteikšanai, cilvēku uzvedības prognozēšanai iepērkoties un citām problēmām. Ir izpētīti galvenie faktori cilvēka izvēļu izdarīšanā, dodoties ceļojumā, kā ekonomiskie apstākļi, ģeogrāfiskie apstākļi, ceļojuma laika izvēle, dažādi *ad hoc* iemesli. Kā arī teorētiski izpētīts, kā sagatavot un grupēt apjomīgus datus datu apstrādei.

Šīs nodaļas ietvaros tiks aprakstītas praktiskajā darbā pielietotās metodes un stratēģijas, prognožu rīka izstrādei, kā arī alternatīvu prognožu rīku pārskats. Praktiskās daļas datu analīze tiks veikta, izmantojot iepriekš izpētītās teorijas - *ad hoc* un lēmumu pieņemšanas teoriju, prognožu veikšanai tiks pielietoti neironu tīklu modeļi, kā arī datu sagatavošanā tiks pielietotas iepriekš nodaļā aprakstītās datu sagatavošanas metodes.

4.1. Datu prognozēšanas rīki

Šobrīd pasaulē, jau ir pieejami dažādi datu prognozēšanas rīki, kas koncentrējās gan uz statistisku datu prognozēšanu, gan biznesa un finanšu lauciņu datu prognozēšanu. Lai izvērtētu sava rīka izstrādi, vispirms nepieciešams apskatīt jau esošos risinājumus, kas risina līdzīgas problēmas. Nodaļas ietvaros tiks apskatīti daži no populārākajiem datu prognozēšanas rīkiem, kā korporācijas *Facebook* atvērtā pirmkoda rīks *Prophet*, korporācijas *IBM* datu prognožu platforma *SPSS* un kompānijas *SAS analytics* piedāvātos risinājumus. Tieši šie rīki izvēlēti salīdzināšanai, jo tie ir vieni no populārākajiem, kā arī pieder pasaulē spējīgākajām datu analītikas un prognozēšanas nozares kompānijām.

4.1.1. *Facebook* datu prognozes rīks *Prophet*

Prophet ir datu prognozēšanas rīks, kas paredzēts laika rindu datu tipa datu prognozēšanai. Tā bāzi veido aditīvs modelis, kurā iespējams prognozēt nelineāras tendences, kas norisinās nemitīgi, ik gadu, nedēļu vai dienu. Šis rīks spēj risināt iztrūkstošo datu problēmas, datu noviržu problēmas, kā arī lielas datu nobīžu problēmas. Šis rīks ir atvērtā pirmkoda rīks, kas pieejams jebkuram interesentam gan lietošanai, gan papildināšanai un izmantošanai savos risinājumos. *Prophet* rīks ir ātrs, spēj risināt dažādas problēmas pat sekunžu laikā. *Facebook* to izmanto lielākajai daļai savu datu prognozēšanas risinājumu problēmām. Rīks piedāvā plašu datu vizualizāciju un aprakstošo dokumentāciju, kas ļauj lietotājam ērti lietot rīku un interpretēt datus [50].

Aplūkojot *Prophet* internetā pieejamo dokumentāciju, tā tiešām ir gana plaši aprakstīta, lai bez liekiem sarežģījumiem uzsāktu tā izmantošanu. Protams, jāņem vērā, ka šis rīks prasa

programmēšana zināšanas, jo tas tiek lietots kā API risinājums. Pieejami arī modeļu apmācībai sagatavojamo datu paraugi, taču to apraksti ir vāji.

Prophet rīku noteikti ir vērts izmantot tādu datu prognozēšanā, kā laikapstākļi, vai, piemēra, valūtu svārstības. Šie dati ir mainīgi noteiktā laika līnijā, kā arī katram laika momentam iespējams noteikt konkrētu vērtību. *Facebook* to izmanto interneta vietnes lietotāju datu plūsmas analīzei un prognozēšanai. Arī šiem datiem katrā sekundē ir noteikta vērtība.

Kā mīnusu šai metodei varētu vērtēt tā pieejamību noteiktai auditorijai, taču jāņem vērā, ka šis rīks ir pieejams bez maksas. Konkrēti šī darba ietvaros šis rīks nav izmantojams, jo tas paredzēts laika rindu datiem, tātad tādiem datiem, kas norisinās nepārtraukti, taču darba izvirzītās problēmas risināšanā šāda tipa dati netiek izmantoti.

4.1.2. *IBM* datu prognozēšanas platforma *SPSS*

IBM piedāvā datu analīzes rīku *SPSS statistics,* kas ir viens no vadošajiem pasaules statistikas rīkiem, ko izmanto dažādu biznesa un izpētes problēmu risināšanā. Tas izmanto *ad hoc* analīzi, hipotēžu analīzi, ģeotelpisko un datu prognožu analīzi. Rīks piedāvā populārākās datu analīzes un prognožu metodes, kā lēmumu koku izmantošanu, regresijas analīzi, un citas. Tajā iespējams ērti aplūkot datus, dažādās formās un interpretācijās. Tam pieejama arī lietotāju saskarne, kurā salīdzinoši ērti iespējams aplūkot un analizēt datus [51].

Rīks pieejams kā API risinājums un to iespējams integrēt ar Python un R programmēšanas valodu risinājumiem. Risinājumam pieejama plaša dokumentācija, gan tā lietošanā, gan datu sagatavošanā. Tajā ir iestrādāta automātiskā datu apstrāde, piemēram, iztrūkstošo vērtību problēmu risināšanā.

SPSS rīks izmanto datu klasifikācijas metodes, lai noteiktu, piemēram, datu aizdevumu drošumu un apjomu konkrētiem klientiem. Šo rīku ērti lietot dažādu statistisku datu analīzei un pat datu prognozēšanai.

Šis ir maksas risinājums, līdz ar to tas ir mīnuss šī rīka izmantošanai. Tādēļ šī darba ietvaros šis risinājums netiks izmantos. Tomēr, tā plašais sadistisko analīžu piedāvājums un uzticamība *IBM* risinājumiem ir labs iedrošinājums, lai to izvēlētos kā iespējamo risinājumu kādai salīdzinoši lielai kompānijai.

4.1.3. *SAS analytics* risinājumi

SAS analytics ir viena no vadošajām kompānijām pasaulē, kas piedāvā plašu datu analīzes rīku klāstu datu analīzei un prognozēšanai, izmantojot mašīnmācīšanās algoritmus. *SAS* piedāvā dažādus rīkus un metodes datu analīzei un prognozēšanai, dažādu statistikas,

ekonomikas problēmu risināšanai, izmantojot dziļo mašīnmācīšanos, valodas apstrādes risinājumus, attēlu apstrādes risinājumus un citas metodes. Tehnoloģijas iespējams izmantot un integrēt kā atvērtā pirmkoda risinājums Python, LUA un JAVA programmēšanas valodu risinājumos. [52]

SAS piedāvātie risinājumi spēj analizēt un prognozēt dažādas biznesu tirgus problēmas, prognozējot iespējamos kāpumus un kritumus, dažādu medicīnas diagnostiku prognozēšanai, valdības problēmu risināšanai, saistībā ar naudas plūsmu un likumu analīzi un daudz citus sarežģītus risinājumus. Risinājumus ir iespējams pielāgot konkrētām darījuma prasībām [52].

Apskatot SAS piedāvātos risinājumus, tie tik tiešām spēj risināt dažādas, sarežģītas problēmas, taču lielākā daļa risinājumi ir izstrādāti konkrētām vajadzībām, tātad tie izstrādāti kā konkrēti produkti, kas spēj risināt vienu problēmu ar noteiktiem datiem. Līdz ar to, tā kā risinājumi ir komerciāli, to apraksti un lietošanas pamācības nav plaši pieejamas. Kompānijas piedāvātie risinājumi ir maksas, kas ir mīnuss šo rīku izmantošanai. Ir pieejami daži, ļoti specifiski bezmaksas pagaidu risinājumi, kas ir pieejami noteiktu dienu skaitu un nesatur pilno funkcionalitāti. Līdz ar to šī darba ietvaros *SAS* piedāvātie risinājumi nesniedz nepieciešamo funkcionalitāti vai pieejamību, lai tos izmantot ceļojumu prognožu rīka izstrādē.

4.1.4. Secinājumi

Apskatot populārākos datu prognožu rīku piedāvājumus, nākas secināt, ka tie ir ļoti iespaidīgi un katrs savu mērķi noteikti spēj risināt ļoti kvalitatīvi, taču, diemžēl tie vai nu funkcionalitātes vai to ierobežotās piekļuves dēļ, nedod iespēju tos izmantot ceļojumu prognožu rīka izstrādē.

Ceļojumu prognozēšanai ir nepieciešami specifiski ievaddati, kas prasa īpašu datu sagatavošanu, izpēti un sakarību noteikšanu. Esošie risinājumi sniedz universālu, statistisku datu analīzes un prognozēšanas iespējas.

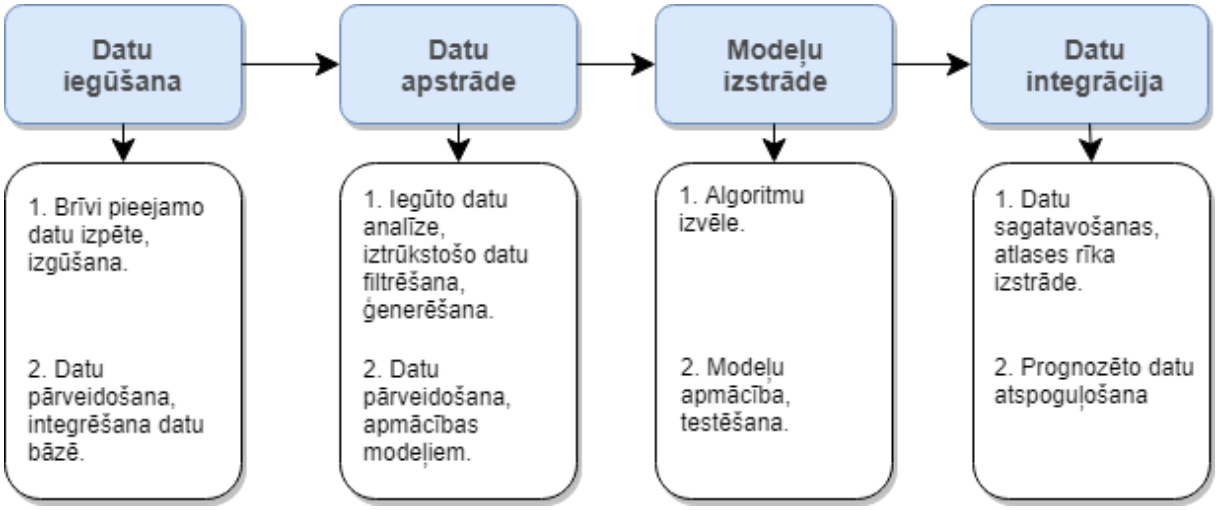
Līdz ar to, nākas secināt, ka brīvi pieejami rīki ceļojumu prognozēšanas problēmu risināšanai vai tam līdzvērtīgu problēmu risināšanai nav pieejami, tātad ir pamatots iemesls izstrādāt jaunu datu prognozes stratēģiju ceļojumu prognozēšanai.

4.2. Prognozēšanas problēma

Šī darba mērķis ir prognozēt ceļotāju nākamos galamērķus. Lai veiktu nākamo ceļojumu galamērķu prognozēšanu, nepieciešami vairāki parametri, pirmkārt, ceļotāju dati, ar kuru palīdzību varam veikt līdzības pazīmju meklēšanu, kā arī ceļotāja papildus ārējos ietekmējošos faktorus, piemēram, ceļotāja dzīves vieta, vai ceļojuma ilgums.

4.3. Izstrādes gaita

Datu prognozēšana, veiksmīgai rezultāta sasniegšanai, iedalāma noteiktos posmos (skatīt attēlā Nr. 4.1.)[17].



4.1.att. **Praktiskās daļas gaita**

Attēlā Nr. 4.1. redzama datu prognozēšanas soļu plūsma. Pirmais posms ir pieejamo ceļojumu datu izpēte un iegūšana. Pēc tam iegūtie dati jāpārveido izmantojamā formā un jāieimplementē datu bāzē, tādējādi iegūstot vieglāk un ātrāk apstrādājamus datus. Nākamais solis ir iegūto datu izpēte un analīze, kā arī nederīgo datu izmešana no datu kopas. Trešajā solī tiek veikta piemērotāko algoritmu apvienošana un datu modeļu izstrāde, tālākai datu prognozēšanai. Gala rezultātā tiek izstrādāts robusts rīks datu atlasei un nākotnes datu ģenerēšanai, kas nodrošinās iespēju lietotājam veikt dažus atlases kritērijus, piemēram, savu atrašanās vietu vai ceļojuma ilgumu.

4.4. Datu aizsardzība

Strādājot ar personas datiem un tos analizējot, ir jābūt informētam un jāievēro personas datu aizsardzības regula. It sevišķi lielapjoma datu analīzē, datu analītiķiem ir pieejama plaša personas datu informācija. Vispirms būtu jādefinē kas tieši ir personas dati?

Personas dati ir vārds uzvārds, personas kods, kontaktinformācija, attēls, biometriskie dati, interneta protokola adrese, personas paradumi, atrašanās vieta, veselības stāvoklis, un citi [54]. Latvijas Republikas personas datu aizsardzības likums nosaka, ka personas dati ir visa informācija, kas attiecas uz identificējamo personu[55]. Idejiski nozīme ir tam, vai konkrētos iegūtos datus var saistīt ar personu jeb datu subjektu. Ja tas ir iespējams, tad šī informācija ir

personas dati. Tātad, ja mums ir pieejams tikai personas vārds un uzvārds, piemēram Jānis Kalniņš, tad tie vēl nebūt nav konkrētas personas dati, jo nav zināms par kuru Jāni Kalniņu ir runa. Taču, ja pieejams Jāņa Kalniņa personas kods, tad jau šī un visa saistītā informācija kļūst par personas datiem. Tiek izšķirti konkrēti dati, kurus ir aizliegts apstrādāt, izņemot gadījumus, kad persona pati to ir atļāvusi, persona pati tos ir publiskojusi vai likumiskā kārtā ir iegūta nepieciešamība datus apstrādāt. Šie dati ir dati, kas atklāj rasi vai etnisko piederību, filozofiskos vai reliģiskos uzskatus, dalību arodbiedrībās, kā arī ģenētiskā personas informācija, kas izmantojama fiziskas personas identifikācijai.[55,56] Galvenais noteikums, analizējot personas datus, ir nodrošināt to, ka personas ir informētas vai devušas piekrišanu par konkrēto datu lietošanu kā arī apzinās kādam mērķim dati tiek lietoti. Izstrādājot savas datu prognozēšanas stratēģijas, kas saistītas ar personas datu aizsardzību rūpīgi jāizpēta, kādus datus izstrādājamā metode izmantos, un vai iegūtie dati, ja persona ir identificējama, ir apstiprināti no personas puses.

4.5. Datu iegūšana

Dati ir svarīgākais nākotnes datu prognozēšanas instruments. Un jāatzīst, ka šīs tēmas izvēlē brīvi pieejamu datu nav daudz. Šāda informācija parasti ir konkrētu tirgotāju rīcībā un netiek brīvi izpausta. Tomēr, veicot nopietnu brīvi pieejamo datu meklēšanu un izpēti, ir izdevies atrast populāras ceļojumu plānošanas vietnes *TripAdvisor* (www.tripadvisor.com) lietotāju atsauksmju datus konkrēta laika perioda ietvaros. Dati pieejami šeit[60]. Šajā vietnē lietotāji brīvi, pēc savas gribas publicē savas atsauksmes un viedokļus par noteiktām tēmām. Nav pilnīgas garantijas, ka viedokļi šajā vietnē ir patiešām patiesi un nav mākslīgi ģenerēti, tomēr jāņem vērā faktors, ka šī vietne ir ceļotāju iespēja paust savu viedokli par kādu vietu, kurā ceļotājs ir bijis, neatkarīgi no tā vai šī vieta ir vai nav patikusi. Šīs vietnes mērķis nav izplatīt un pārdot kādu īpašu piedāvājumu, kā tas varētu būt, piemēram, dzīvesvietu piedāvājumu vietnēs, līdz ar to nepieciešamība mākslīgi ģenerēt atsauksmes nav tik aktuāla. Katra atsauksme ir ievienota no kāda lietotāja profila, līdz ar to, viedokļi ir jāizsaka no kāda profila, kas samazina datu automātisku ģenerēšanas iespēju.

Ja uz šo problēmu raugās no biznesa situāciju skata punkta, kā, piemēram, lidojumu vai naktsmītņu piedāvājuma vietņu skata punkta, tad visi pieejamie dati ir iegūto no lietotāju reālām darbībām. Attiecīgi datu prognozēšanai vārētu tikt izmantoti tikai tie dati, par kuriem reāli ceļotājs ir veicis pirkumus, tādejādi visticamāk, ka iepriekšējā ceļotāja pieredze būs patiesa.

4.5.1. Iegūto datu pārskats

Vietnes *Tripadvisor* iegūtie dati satur lietotāju atsauksmes par viesnīcām visā pasaulē. Dati ir angļu valodā. Iegūtie dati satur informāciju par lietotājvārdu, viedokļa paudēja dzīvesvietas valsti, galamērķi, kur ceļotājs devies, laiku, kad ceļotājs devies, ceļotāja atsauksme un vērtējums par dažādām pozīcijām. Kopumā ir iegūti 1 430 342 ieraksti. Par katru ierakstu ir pieejami noteikti dati(skatīt tabulā 4.1.).

4.1. tabula

Vietnes *TripAdvisor* datu piemērs

ID	Lietotājs	Dzīvesvieta	Galamērķis	Atsauksme	Serviss	Tīrība	Cena	Komforts	Ceļošanas Laiks
308530	Nadine R	Jacksonville, Florida	Baltimore	Usually stay near the airport, but this trip we ha...	5	5	5	5	April 19, 2012
308531	blt3116	Chillicothe, Ohio	Baltimore	Stayed at this Hilton for 2 nights...	4	4	4	4	January 30, 2012
339250	CharityK13	Portland, Oregon	London	We stayed at the City Stay Hotel near Bow Road...	5	4	0	5	October 16, 2009

Kā redzams tabulā, tad par katru lietotāju nav pieejama pārāk plaša informācija. Nav iespējams identificēt personu, kas paudusi savu viedokli, jo pieejams tikai personas lietotājvārds. Kā *arī www.tripadvisor.com* vietnē lietotāja profilos nav pausts lietotāja vārds uzvārds vai cita konkrētu personu saistoša informācija, kas varētu noteikti tieši par kuru personu ir runa. Visas atsauksmes lietotāji ir paši ievietojuši publiskai apskatei ceļojumu vietnē, tātad katrs pats ir devis iespēju visiem aplūkot viņa viedokli un informāciju par ceļošanu.

Tomēr ir pieejama pietiekama pamata informācija, lai mēģinātu veikt eksperimentālu rezultātu prognozēšanu. Datos ir pieejami gan lietotāji, kas šajā gadījumā ir subjekti kā x, vai y, jo nav zināms par kuru personu ir runa, kuri ceļojuši vairākas reizes uz dažādām vietām, gan lietotāji, kuri ceļojuši tikai vienu reizi. Pieejama informācija par to no kurienes uz kurieni ir notikusi ceļošana, kā arī kādā laika periodā. Tādējādi konkrētā darba ietvaros ir iespējams izstrādāt prognozēšanas metodi, kas ļoti robusti spēj noteikt iespējamo gala mērķi, balstoties uz limitētiem datiem.

Ja šādu datu piedāvājuma tehnoloģiju sniegtu kāda no ceļošanas kompānijām, tad šīs kompānijas, ar lietotāju atļauju, varētu veikt stipri apjomīgāku rīka izstrādi, izmantojot iepriekš darbā izpētītos ceļotāju ceļojuma izvēli ietekmējošos faktorus, tādējādi padarot piedāvājumu daudz pielāgotāku ceļotājam.

Dati tika iegūti .JSON formātā (.JSON faila piemērs aplūkojams pielikumā Nr.1.), kas lielapjoma datu apstrādei nav pārāk ērts formāts. Tādēļ datu prognožu rīkā ir izstrādāta papildus funkcionalitāte datu pārveidošanai un ievietošanai datu bāzē. .JSON faili ir strukturēti pa mapēm, kur katrā mapē ir notiktas valsts viesnīcu atsauksmes. Katrā mapē viens .JSON fails ir par konkrētu viesnīcu un tajā apkopoti pieejamie viedokļi un vērtējumi par konkrēto viesnīcu. Datu apstrādes rīks ielasa .JSON failus un ievieto tos datu bāzes tabulā.

4.6. Datu analīze

Iepriekšējā nodaļā redzami iegūto datu piemēri. Noteiktu problēmu gadījumā, kā, piemēram, lietotāja nākotnes pirkumu prognozēšana, jeb piedāvāšana, ir svarīgs plašs pieejamo datu klāsts. Ir būtisks gan pieejamo datu apjoms, gan pieejamo datu kvalitāte un nozīmīgums.

Kā iepriekš tika izpētīts, ceļotāja izvēli ļoti bieži ietekmē tieši ekonomiskie un ģeogrāfiskie apsvērumi. No pieejamās informācijas datu prognožu veikšanai, ir pieejama katra ceļotāja dzīvesvieta un iepriekšējie galamērķi. Tātad skaidri iespējams prognozēt analizēt tieši ceļotāju galamērķu datu informāciju.

Ekonomiskie dati šajā datu kopā nav pieejami, tādēļ nav iespējams veikt datu analīzi un prognozēšanu pēc lietotāja turības līmeņa vai citām ekonomiskām iezīmēm. Tā kā darba ietvaros netiek izmantoti reālu ceļojumu piedāvājumi, tad nav iespējams veikt arī datu atlasi pēc ekonomisko piedāvājumu grupas. Taču, ja uzņēmumiem informācija par ceļotāja ekonomisko stāvokli, iepriekšējo pirkumu cenām vai ceļojumu cenu kategorijām ir pieejama, tad tā noteikti ir ļoti vērtīga datu prognozēšanai. Gadījumos, ja šādas informācijas nav, tad papildus datu atlasei, var lietotājam ļaut norādīt iespēju atzīmēt cenu kategoriju kurā

iekļauties, un tad pēc tās un prognozētajiem datiem dot lietotājam piedāvājumu. Šī darba ietvaros, datu trūkuma dēļ to nav iespējams veikt.

No pieejamās informācijas vēl tiks izmantots datums, kas norāda uz ceļotāja izvēli doties ceļojumā noteiktā laika periodā uz noteiktu valsti. Kā arī iespējamie servisu vērtējumi.

Datu prognozēšanā netiks izmantota atsauksmes informācija, jo tās informāciju nav iespējams izmantot automatizēti. Atsauksme satur viedokli un tās informācija var būt ļoti dažāda. Tomēr, jāņem vēra, ka tā varētu būt noderīga un atklāt dažādas ceļotāja iezīmes un īpašības, kā piemēram ierastās nodarbes ceļojot, ierastās vēlmes un prasības pret ceļošanas dzīvesvietām u.c. Šādai informācijas analīzei potenciāli varētu pielietot gan atslēgvārdu meklēšanu, gan kopējo teksta nokrāsas analīzi. Šī darba ietvaros tā netiks veikta, taču potenciāli varētu tikt apskatīta kā nākotnes pētījuma ideja.

Ir redzams, ka pieejamie dati nav pārāk plaši, līdz ar to nav pieejamas specifiskas informācijas par lietotāja ieradumiem, vēlmēm, izglītību, ceļošanas paradumiem un cita veida ārējo faktoru informācija. Tā rezultātā praktiskā darba ietvaros iespējams veikt diezgan robustu datu prognozēšanu, kas tiek veikta balstoties uz ierobežotu informāciju, taču ar to ir pietiekami, lai pārbaudītu, vai potenciāli un cik lielā mērā ir iespējams veikt ceļojumu piedāvājumus.

4.7. Datu sagatavošana

Iepriekš aprakstītie dati pirms to izmantošanas ir jāfiltrē un jāpārbauda, vai nav pārlieku liels daudzums tukšo un nederīgo vērtību. Šajā gadījumā tukšās vērtības ir tādi ieraksti, kuros kāda no vērtībām nav norādīta, bet nederīgie dati ir tādi ieraksti, kur kāda no vērtībām satur neatbilstošu informāciju. Pretējā gadījumā, datu prognozēšanā tiktu izmantoti "netīri" jeb nekorekti dati, kā rezultātā prognozēšanas rezultāti būtu nekam nederīgi.

Kā jau iepriekš minēts, tad datu analīzei tiks izmantoti sekojoši lauki, skatīt tabulā 4.2.

4.2. tabula

Datu lauki

Lauka nosaukums	**Veids**	**Informācija**
Lietotāja vārds	String - tekstuāla informācija	Tekstuāla informācija, nav nepieciešams filtrēt, nav nozīmes saturam. Norāda uz konkrētu lietotāju.
Dzīvesvieta	String - tekstuāla informācija	Tekstuāla informācija. Ir nozīme saturam, nepieciešams filtrēt tās datus.
Galamērķis	String - tekstuāla informācija	Tekstuāla informācija. Ir nozīme saturam, nepieciešams filtrēt tās datus.
Datums	String - tekstuāla informācija, formā diena-mēnesis-gads	Konstanta informācija, vienmēr uzdota konkrētā formā. Nepieciešams pārbaudīt vai informācija uzdota pareizā formā un nav tukšuma vērtību.
Serviss	Int - skaitliska vērtība	Skaitliska vērtībā, jāpārbauda minimālās, maksimālās robežas, tukšie dati.
Tīrība	Int - skaitliska vērtība	Skaitliska vērtībā, jāpārbauda minimālās, maksimālās robežas, tukšie dati.
Komforts	Int - skaitliska vērtība	Skaitliska vērtībā, jāpārbauda minimālās, maksimālās robežas, tukšie dati.
Vērtība	Int - skaitliska vērtība	Skaitliska vērtībā, jāpārbauda minimālās, maksimālās robežas, tukšie dati.

Kā redzams tabulā, tad lietotāja vārds ir lauks, kuru nav nepieciešams filtrēt. Jāpārbauda vai šis lauks nav tukšuma vērtība, ja lauks ir tukšs, tad šis datu ieraksts netiks ņemts vērā, jo potenciāli var būt mākslīgi ģenerēts.

Dzīvesvieta un galamērķis, abi ir tekstuāli lauki, kas ir nozīmīgākās vērtības šī prognožu rīka izstrādē. Šīs vērtības ir ļoti dažādas, kā arī ir nepieciešams izstrādāt ģeogrāfisko sakarību saikni, lai nodrošinātu iespēju, pēc ģeogrāfiskās atrašanās vietas prognozēt iespējamos galamērķus. Šim nolūkam, ir izveidota papildus tabula datu bāzē, kas kalpo kā sava veida datu vārdnīca, lai nodrošinātu korektu valstu analīzi. Tabula ir ģenerēta, balstoties pēc starptautiskā standarta ISO-3166, kurā aprakstītas valstis, to saīsinājumu kodi, piederība konkrētam reģionam, kā, piemēram, Bulgārijas kods ir BG, kontinents ir Eiropa un reģions Eiropā ir Austrumu Eiropa [57]. Datu piemēru skatīt pielikumā Nr.2. Lai sasaistītu konkrētas pilsētas, valstu reģionus un štatus, papildus katrai valstij ir savs pilsētu saraksts, kas ģenerēts no pieejamās informācijas par valstu un pilsētu saikni[58]. Datu piemēru skatīt pielikumā Nr.3.

Rezultātā tika iegūta datu tabula, kas satur sekojošu informāciju, skatīt tabulā 4.3.

4.3. tabula

Valstu informācija

Pilsēta	**Reģions**	**Valsts**	**Pasaules daļa, reģions**	**Valsts kods**
SIJUA	JHARKHAND	INDIA	ASIA	VI

Tabulā Nr. 4.3. redzams datu piemērs, kā izskatās pilsētu, reģionu un valstu ģeogrāfiskās saistības tabula. Tabulas informācija ir angļu valodā, tā pat, kā datu tabula, tādējādi izvairoties no nepieciešamības veidot papildus valodu vārdnīcu datu analīzei. Burti ir ģenerēti kā lielie burti, lai atvieglotu pilsētu un reģionu meklēšanu katram vērtējuma ierakstam. Tādējādi tiek nodrošināta izvairīšanās no pārbaudēm ar mazajiem burtiem. Pirms meklēšanas datu bāzē, katra ieraksta ģeogrāfiskā informācija arī tiek pārvērsta uz lielajiem burtiem.

Filtrējot analizējamos galamērķu un dzīvesvietu datus, tika konstatēts, ka galamērķu informācija vienmēr ir korekta, jo tā ievadīta no vietnes piedāvātajām vērtībām, kā arī nav tādu ierakstu, kam nebūtu norādīts vēlamais galamērķis.

Toties lietotāju dzīvesvietai gan ir ļoti dažādas vērtības, sākot no tukšumiem un neloģiskām simbolu virknēm, līdz pat trīs vārdiem. Daļa no ierakstu vērtībām tika labotas manuāli, piemēram, drukas kļūdu gadījumā, tādējādi padarot datu apstrādi kā ļoti apjomīgu un darbietilpīgu procesu. Ģeogrāfisko datu pārbaudes procesā tika pārbaudīts, vai, pirmkārt, ir ievadīta gan dzīvesvietas vērtība, gan galamērķa vērtība. Pretējā gadījumā šādi dati netika izmantoti datu prognozēšanā. Jo nav iespējams ģenerēt informāciju par to, kur cilvēks dzīvo, ja tā datos nav pieejama. Kā arī šo abu lauku korekta esamība analizējamos datos faktiski ir svarīgākā, jo rada pamata sakarību starp populārākajiem ceļojumu galamērķiem. Rezultātā

ģenerējot datus datu modelim, tiek izmantoti tikai tie lauki, kas satur gan galamērķi, gan dzīvesvietu, un ievadītā dzīvesvieta ir korekta un atrodama valstu tabulā. Šīs datu filtrācijas rezultātā tika pazaudēts apjomīgs datu daudzums, to skatīt zemāk, nodaļā tabulā 4.4.

Datums ir uzdots noteiktā formā mēnesis vārdiem, datums un gads ciparos. Šī informācija ir nozīmīga, jo sniedz informāciju par aptuveno laiku, kad lietotājs ir izvēlējies doties ceļojumā. Šo informāciju ir nepieciešams pārbaudīt, lai noteiktu, vai tās vērtības atbilst reālām datumu vērtībām. Tika konstatēts, ka šīm vērtībām nav nevienas tukšuma vērtības, tātad papildus datu ģenerēšanu nevajadzēja veikt. Datu analīzē tika izpētīts un nolemts, ka izmantota tiks tikai informācija par mēnesi un gadu, jo dienas analīze nedod nozīmīgu ietekmi datu prognozēšanā. Tai būtu nozīme, ja būtu zināms precīzs ceļojuma laiks un garums, tad tiktu analizēts un datu prognozē izmantots arī ceļojuma ilgums dienās.

Serviss, tīrība, komforts un vērtība ir skaitliskas vērtības robežās no 1 - 5 un 0, gadījumos, ja vērtība nav norādīta. Filtrēšanas procesā tiek pārbaudīts vai vērtības atbilst skaitliskās informācijas noteikumiem. Filtrēšanas rezultātā tika noskaidrots, ka šīm vērtībām nav tukšuma vērtību, bet ir daudz gadījumi, kad norādītie lauki satur 0 vērtību. Tika izlemts, ka šīs rindas tiks atstātas kā derīgas vērtības. Šie papildus lauki, šķietami būtiski neietekmē ceļojuma izvēli, taču hipotētiski to līdzība var norādīt uz ceļotāju līdzīgu domāšanu vai attieksmi pret konkrētām lietām, līdz ar to, iespējams arī šo ceļotāju galamērķu izvēle varētu būt līdzīgāka. Protams, gadījumā, ja datos būtu pieejama specifiskāka informācija, tad iespējams šo vērtējumu informāciju nebūtu nepieciešams izmantot, taču no pieejamajiem datiem, tā ir vienīgā papildus informācija.

Papildus tika pārbaudīts vai datu tabula nesatur dublikātus, kā rezultātā dublikāti tika likvidēti, atstājot tikai vienu unikālo vērtību. Tika izņemtas arī tādas vērtības, kurām viens un tas pats lietotājs ir sniedzis atsauksmes vienā laikā dažādām viesnīcām, kuras atrodas vienā valstī. Tādējādi izvairoties no mākslīga konkrētu valstu apmeklējumu palielinājuma.

Datu sagatavošanas un filtrēšanas rezultātā tika ievērojami samazināta analizējamo datu kopa. Tabulā Nr. 4.4. iespējams apskatīt derīgo datu skaitu.

4.4. tabula

Datu skaits

Sākotnējais datu ierakstu skaits:	1 430 342
Vērtību skaits pēc dublikātu vai vienlaicīgo ierakstu dzēšanas:	988335
Vērtību skaits pēc tukšo un nederīgo dzīvesvietu vērtību dzēšanas.	756944
Unikālo lietotāju skaits.	685005

Tabulā Nr. 4.4. redzams ka rezultātā izmantojami ir tikai puse no datu ierakstiem datu bāzē.

Datu filtrācija, neskaitot manuālos labojumus, fiziski noris izstrādātajā rikā. Viena no rīka funkcionalitātēm ir datu ievietošana, izgūšana no datu bāzes un datu filtrēšana pēc noteiktiem noteikumiem.

Rīks izgūst no datu bāzes informāciju, veic valstu pielīdzināšanu, iepriekš aprakstīto filtrēšanas procesu un pārveido iegūto derīgo informāciju .csv faila formātā. Šāds formāts ir ērti nolasāms un izmantojams jebkāda veida datu analīzei.

4.8. Algoritmu izvēle un alternatīvas

Izvirzītā problēma, ir sava veida klasifikācijas problēma, kad balstoties uz noteiktiem ieejas datiem par ceļotāju, vēlamies noteikt procentuālu piederību kādai no galamērķu valstu grupām. Pieejamie dati ir ļoti dažādi un iespējamas visdažādākās to kombinācijas.

Šajā gadījumā ir ļoti dārgi, no sarežģītības un resursu viedokļa izmantot lēmumu koku vai likumos balstītās metodes, turklāt, ņemot vērā datu lielumu, izstrādāt šo metožu noteikumus var būt ļoti sarežģīti, un pastāv ļoti liela iespēja kļūdīties to izstrādes gaitā. Taču, iespējams no visām alternatīvām, likumos balstītās metodes vai lēmumu koki ir ļoti labs papildinājums citiem algoritmiem, piemēram, lai samazinātu gala rezultātā analizējamo datu apjomu.

Iespējams, precīzi varētu būt SVM algoritmi, jo tie spēj labi tikt galā ar situācijām, ka tie ierauga vēl neredzētus datus, taču kā iepriekš aprakstīts, tad lielapjoma datu gadījumā, SVM klasifikatori var būt ļoti laikietilpīgi ne tikai apmācības procesā, bet arī klasificēšanas procesā. Tomēr arī SVM klasifikators varētu būt viena no labākajām alternatīvām, prognožu veikšanai.

Arī instanču balstītā pieeja var būt nelietderīga, jo tās pamatā ir jaunu noteikumu pievienošana apmācības laikā, līdz ar to tā nepārāk labi strādās uz jauniem, vēl neredzētiem datiem, kas šī darba problēmas gadījumā var būt bieža parādība.

Līdz ar to izpētot pieejamos datu prognožu risinājumus, kā arī alternatīvo risinājumu metodes, viena no piemērotākajām metodēm problēmas risināšanai, ir neironu tīklu izmantošana prognožu noteikšanai, jeb jaunu valstu piedāvājumu izteikšanai. Neironu tīkli ir salīdzinoši sarežģīta metode no to izprašanas viedokļa, taču tā ir ļoti spējīga, precīza un spēj atrast sakarības starp vēl neredzētiem datu piemēriem. Neironu tīklu modeļus var saukt par sava veida melnās kastes principa modeļiem, jo ir zināmi ievaddati un izvaddati, taču to precīzi pamatojumi katra soļa izvēļu izdarīšanā liela datu apjoma gadījumā, piemēram, attēlu analīzes gadījumos nav nosakāmi. Arī šīs ceļojumu prognožu noteikšanas stratēģijas gadījumā izstrādāt precīzus noteikumus ir ļoti sarežģīti un laikietilpīgi, lai neteiktu, ka neiespējami. Vēl sarežģītāk tas kļūst, kad par katru lietotāju pieejama plaša informācija ar tā īpašībām, pazīmēm un iepriekšējiem ceļojumiem. Šajā gadījumā katra lietotāja informācija kalpo kā ievaddati un kāds no galamērķiem ir izvaddati, taču kādas sakarības pie tā noved mēs nezinām.

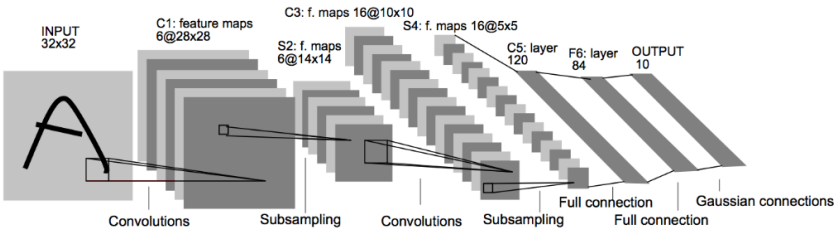
Darba ietvaros nav plānots izstrādāt jaunas neironu tīklu struktūras vai risinājumus. Izlasot un izpētot pieejamo informāciju par neironu tīkliem, šī problēma salīdzinoši nav sarežģīta no datu apjoma viedokļa un neironu tīklu risinājumiem ar to potenciāli būtu jātiek galā.

4.8.1. Neironu tīkla struktūra

Izpētot esošās publiski pieejamās neironu tīklu struktūras, pētīju to pielietojumu citiem mērķiem. Tā rezultātā daudzas no neironu tīklu struktūrām, piemēram, populārais AlexNet vai GoogleNet spēj risināt daudz sarežģītākus uzdevumus, piemēram, objektu atpazīšanu attēlos [61]. Šādos gadījumos informācija ir ļoti apjomīga, viena objekta iezīmes ir daudz un ļoti dažādas. Tā rezultātā izvēlējos šī darba ietvaros izmantot konvolūciju neironu tīkla struktūru LeNet.

LeNet ir konvolūciju neironu tīkls, kas izstrādāts 1998. gadā. Konvolūciju neironu tīkli ir ļoti līdzīgi parastajiem neironu tīkliem. Tie ir veidoti no neironiem, kam ir apmācāmi svari un novirzes. Katrs neirons saņem noteiktus datus, veic to apstrādi un tālāk pēc noteikta principa, padod datus nākamajam slānim. Šīs struktūras pamat atšķirība ir tā, ka tā nosaka, ka datus iespējams specificēt jau apmācības laikā, tādējādi samazinot kopējo datu apstrādes lielumu un ilgumu. Tā otrs pluss ir iespēja izmantot šāda tipa neironu tīklu struktūrās trīsdimensionālus datus, dažādos izmēros[62,63].

LeNet ir specifisks daudzslāņu neironu tīkls, kura pamata struktūra aplūkojama zemāk attēlā Nr 4.2.[62].



4.2.att. **CNN LeNet tīkla struktūra[64]**

LeNet tīkla struktūra šī darba ietvaros sastāv no ieejas datu slāņa, konvolūciju slāņa ar kodola izmēru 5 un soli 1, apvienošanas slāņa ar kodolu 2 un soli 2, vēl viena secīga konvolūcijas un apvienošanas slāņa, viena slēptā slāņa, un rezultātu slāņa, kas veic rezultātu atgriešanu. Konvolūciju slānis ir pamata kodols neironu tīklam. Tas veic pamata skaitļošanu. Šis slānis sastāv no filtriem, kas tiek pielietoti datiem apmācības procesā. Katram filtram ir noteikts kodola izmērs un dziļums. Datiem tiek secīgi pielietoti filtri, tādējādi pēc noteikta principa pārvēršot daļu datu citā vērtībā. Filtri tiek pielietots datiem pēc noteikta soļa ieejas datu augstumā un platumā. Konvolūciju slāņu rezultātā tiek iegūta apjomīga informācija, kas atspoguļo apmācības procesu. Intuitīvi, konvolūciju slāņa filtri tiek apmācīti tīklā aktivizēties pēc noteiktiem datiem. Tā rezultātā tiek iegūtas aktivācijas kartes, kas sarindojas secīgi pa dimensijām. Apvienošanas slāņa funkcija vistiešākajā mērā ir samazināt datu apjomu. Šie slāņi parasti ir starp konvolūciju slāņiem. Parasti apvienošanas slāņi samazina datu apjomu par 2x2 matricām. Tādējādi samazinot īpašību skaitu un risinot pēkšņo "izlēcēju" vērtību gadījumus. Filtru pielietošana un datu samazināšana notiek ar matricu reizināšanu palīdzību[63].

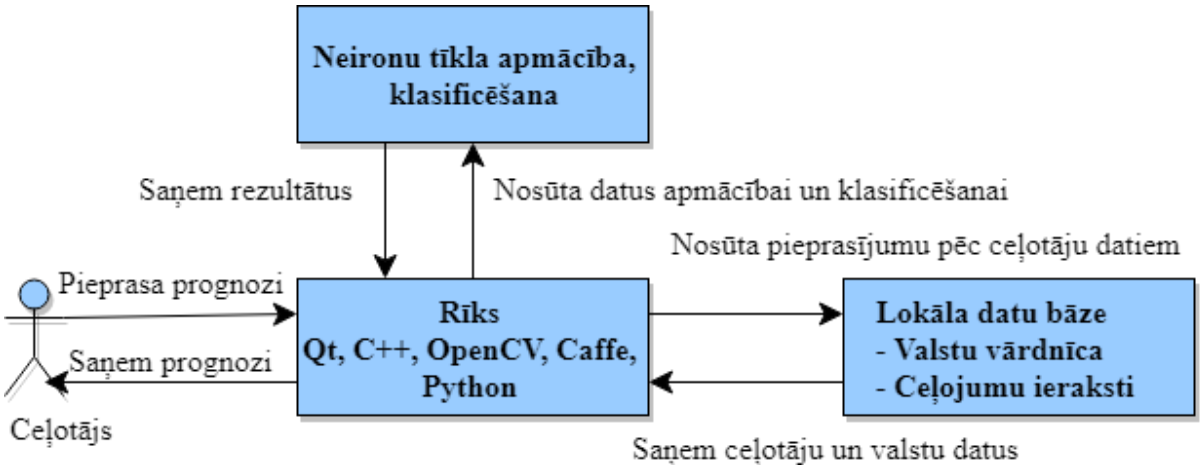
LeNet populārākais pielietojuma piemērs ir ar roku rakstītu skaitļu atpazīšana un klasificēšana. Taču šī tīkla struktūra ir pielietota arī neskaitāmu citu problēmu pielietošanā, kas saistītas ne tikai ar vizuālu attēlu problēmu risināšanu. Tas ticis izmantots arī uzņēmumu tirgus analīzē, un nākotnes prognozēšanā, konkrētiem laika periodiem.

4.9. Rīka izstrāde

Datu prognozēšana sastāv no vairākām daļām - datu apstrāde, datu atlase, datu sagatavošanu apmācībai, apmācības modeļa izstrāde, un datu prognozēšanas modeļa pielietošana. Visas šīs komponentes izstrādātas robustā rīkā, kas strādā kā neliela patstāvīga lietojumprogramma. Tai ir vienkārša un tehniska lietojuma saskarne.

4.9.1. Tehniskie parametri

Rīks ir izstrādāts C++ programmēšanas valodā, izmantojot QT ietvaru. Tehnisko shēmu skatīt attēlā Nr. 4.3.



4.3.att. **Tehniskā shēma**

Attēlā Nr. 4.3. redzama rīka darbību plūsma. Rīka pamatā ir Qt ietvarā C++ programmēšanas valodā izstrādāts rīks, kas veic gan lietotāja ievades datu apstrādi, gan komunikāciju ar datu bāzi un klasifikācijas mehānismu. C++ risinājums veic datu filtrēšanu, priekš algoritmus datu atlasei no datu bāzes, veic pieprasījumu pēc datiem un saņem tos no datu bāzes. Datu bāze izvietota lokāli. Saņemtos datus risinājums pārvērš vajadzīgajā formā neironu tīklu apmācībai. Tie tiek ģenerēti un saglabāti lokāli uz lietotāja datora. C++ risinājums tālāk izsauc PYTHON skriptu, kas veic modeļa apmācību un datu prognozēšanu. Neironu tīkla realizācijai tiek izmantots Caffe ietvars.

Kad modelis ir izveidots un dati prognozēti, C++ risinājums saņem atbildes datus, kas tiek pārveidoti lietotājam lasāmā formātā. Pēc datu atspoguļošanas nevajadzīgie dati un modeļi tiek izdzēsti. Nav nepieciešamas glabāt esošos datus, jo tie tiek ģenerēti katram lietotājam no jauna, balstoties tieši uz viņam līdzīgajiem lietotājiem.

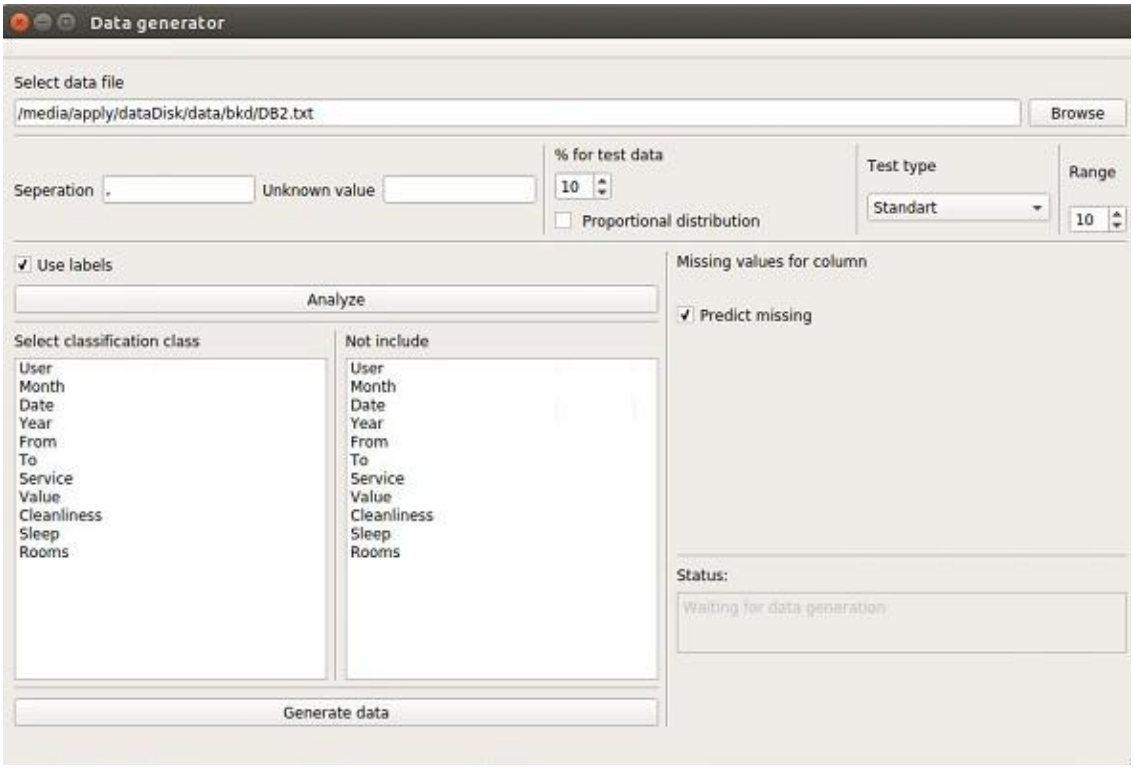
4.9.2. Tehnoloģiskie parametri

Neironu tīklu izmantošana ir resursu ietilpīga un prasa daudz laika. Tā apmācību un darbību iespējams paātrināt izmantojot video karti. Šī darba ietvaros darba paātrināšanai ir izmantota GeForce GTX 970 videokarte ar 4GB atmiņu.

Rīks izstrādāts uz Linux Ubuntum 14.04, taču rīku var darbināt arī uz jaunākām Ubuntum versijām. Neironu tīklu struktūras izmantošanai nepieciešama Caffe ietvara bibliotēka, kā arī neironu tīklu apmācības procesā ir izmantots DIGITS ietvars. Datu bāzes nodrošināšanai tiek izmantota lokālā datu bāze.

4.9.3. Izstrādātais rīks

Attēlā 4.4. iespējams aplūkot rīka tehnisko grafisko lietotāja saskarni.



4.4.att. **Apstrādes rīks**

Attēlā redzams, ka rīkā iespējams ielasīt datu failu, izvēlēties datu faila vērtību atstarpju simbolu, kā šajā gadījumā tas ir komats. Iespējams norādīt, cik lielu daudzumu no datiem atvēlēt testa datu kopas izstrādei. Validācijas dati tiek nodalīti atsevišķi no apmācības datiem modeļa apmācības gaitā. Iespējams norādīt kādu datu sadalījumu veidot. Tas nozīmē, ka ir iespējams atstāt datus tādu kādi tie ir, un ir iespējams datus samazināt, tādējādi izveidojot vienlīdzīgu sadalījumu starp klašu grupām. Klases, kurām nepieciešama datu izņemšana, lai nodrošinātu vienlīdzīgu datu sadalījumu, tiek samazinātas izmantojot gadījuma funkciju. Iespējams norādīt nezināmās vērtības apzīmējumu, ja tāds ir.

Zemāk divos logos iespējams apskatīt pieejamos laukus un atlasīt tos pēc kādiem kritērijiem veikt klasifikāciju un kuras no klasēm neiekļaut prognozes rīkā.

Papildus iespējams veikt iztrūkstošo vērtību meklēšanu un izņemšanu no datu kopas. Malā aplūkojams arī statusa un datu rezultātu logs. Rīks ir tehnisks, tādēļ šī darba ietvaros tam izstrādāta tehniska lietotāja saskarne.

4.10. Modeļu apmācības gaita

Kad dati ir atlasīti, un "iztīrīti", tie ir gatavi to pārveidošanai neironu tīkla modelim. Idejiski katru ierakstu iespējams pārvērst kā vektoru, kas satur noteiktu informāciju par katru ierakstu. Katrai datu kolonnai ir iespējams dažāds vērtību skaits, bet tikai viena no tām katram ierakstam būs patiesa.

Tā kā datu apjoms šī darba ietvaros ir ierobežots, nav pieejami ne ekonomiskie dati, ne reāli piedāvājumi, tad eksperimenti datu prognozēšanā tiks veikti, lai noteiktu kur konkrētam ceļotājam doties ceļojumā, bez papildu atlases kritērijiem. Attiecīgi sākotnējā iecere bija izstrādāt datu analīzes rīku, kas spētu prognozēt iespējamos galamērķus konkrētam ceļotājam, taču datu trūkuma dēļ, lai noskaidrotu, vai vispār iespējams veikt ģeogrāfisko datu prognozēšanu ceļojumu izvēlē, tiks pārbaudīts vai ar esošajiem datiem iespējams veikt jebkādas ceļojumu galamērķu prognozes. Līdz ar to datiem kā klasificējamā klase tiks nodalīta ceļojuma galamērķis, kas šo datu gadījumā ir atsauksmes viesnīcas vieta.

Iepriekš aprakstītajā nodaļā redzamajā datu rīkā iespējams redzēt, kā lietotājam iespējams atlasīt klasi, uz kurieni doties un izvēlēties neņemt vērā konkrētus ierakstus. Pēc šo parametru atlases tiek ģenerēti dati globāla modeļa izveidei. Tātad, izstrādājot modeli, ir zināmas lietotāju dzīvesvietas, ceļojumu galamērķi, aptuvenais laiks, vērtējumi kā arī lietotāja vārds. Citi dati prognozēšanai nav pieejami, tādējādi padarot šo prognozēšanu ļoti robustu.

Dati tiek ģenerēti vektoru formā, ņemot vērā iespējamās vērtību variācijas, kas tālāk tiek padoti neironu tīklam modeļa apmācībai.

Modeļa apmācībā, kā iepriekš jau minēts, tiek izmantota LeNet tīkla struktūra. Tīkla vizuālo shēmu iespējams aplūkot pielikumā Nr.4. Visiem apmācības modeļiem datu apmācības procesā tika izmantoti 100 soļi jeb epohas, validācija tiek veikta pēc katra pilna soļa. Apmācībai tiek izmantota eksponenciālā samazinājuma funkcija apmācības ātrumam ar vērtību 0.95, šāda vērtība un funkcijas izvēle nodrošina lēnāku neirona tīkla apmācību un samazina potenciālos informācijas zudumus.

Modeļu sagatavošanas process ir laikietilpīgs. Pirmais solis ir datu ģenerēšana, kas vidēji aizņem piecas līdz sešas stundas. Kad dati ir uzģenerēti, tie tiek nodalīti apmācības, validēšanas un testēšanas kopās. Tālāk tiek veikta modeļu apmācība. Viena modeļa apmācība ar iepriekš aprakstītajiem tehniskajiem parametriem, aizņem aptuveni 5 līdz 7 stundas.

4.11. Testēšana un rezultāti

Darba ietvaros tika veikta vairāku modeļu ģenerēšana, balstoties uz dažādiem atlases kritērijiem.

Modeļi tiek ģenerēti no vieniem un tiem pašiem datiem, taču atšķiras to saturs, atbilstoši atlases kritērijiem. Modeļi tika testēti, tiem izmantojot testa datu kopas datus. Testa dati tika nošķirti jau pirms modeļa ģenerēšanas, tādējādi nepieļaujot iespējamību, ka tie var iekļūt apmācības procesā.

Katrs modelis tika apmācīts, lai klasificētu iespējamo ceļojuma galamērķi. Testa datu kopa sastāv no ierakstiem, kam zināmi iespējamie ceļojumu galamērķi, tādējādi darbinot klasifikācijas modeli uz šiem piemēriem iespējams noteikt vai modelis spēj vai nespēj prognozēt reālus datus. Tālāk nodaļā apskatāmi modeļu testēšanas rezultāti un secinājumi.

Testēšanas procesā katram modelim tika izveidots noteikts testa datu skaits, kas pielietots konkrētajam modelim. Testa dati tika ielasīti reizē, un padoti neironu tīklam atpazīšanai. Rezultāti tiek atgriezti teksta failā, lai atvieglotu to analīzi.

Kā pirmais tika ģenerēts modelis, kas nosaka uz kuru kontinentu doties. Šajā datu kopā netiek ņemta vērā no kuras uz kuru valsti ir devies vai vēlas doties ceļotājs bet gan tiek noteikt kontinents, uz kuru doties ceļotājam. Tā kā datu par katru konkrēto lietotāju ir ļoti maz, tad modeļa ģenerēšanas procesā netika ņemts vērā konkrēts lietotājvārds. Tātad izmantojot izstrādāto rīka funkcionalitāti datu sagatavošanai, datu ģenerēšanai tika atlasīta kā klasificējamā klase - kontinents un kā klase, ko neņemt vērā datu analīzē - lietotājvārds.

Vēl būtiska piezīme ir tā, ka pateicoties datu nevienlīdzībai un laikietilpīgajai modeļu ģenerēšanai, to analizējamo datu skaits tika vēl būtiski samazināts, lai nodrošinātu augstāku precizitāti.

Zemāk tabulā aplūkojama modeļa apmācības dati.

4.5. tabula

Modeļa dati

Veids	**Klašu skaits**	**Validācijas datu apjoms**	**Modeļa precizitāte**	**Apmācības datu apjoms**
Kontinentu modelis	6	8%	99%	~ 40 000 ierakstu katrai klasei

Šī modeļa ietvaros nācās veikt manuālu datu samazināšanu, jo savstarpējā proporcionalitāte starp kontinentiem bija ļoti nevienmērīga. Šis modelis satur 6 klases, pēc

kontinentiem - Eiropa, Āzija, Āfrika, Austrālija un Okeānija, Ziemeļu un Dienvidu Amerikas. Datos vislielākais galamērķu apjoms ir tieši Ziemeļu un Dienvidu Amerikas, kā rezultātā klases abas Amerikas satur gandrīz 70 % no datu apjoma. Tādēļ, lai izlīdzinātu datus un veiktu saturīgu analīzi, datu apjoms katrā kontinentu klasē ir aptuveni 40 000.

Modelis tika testēts, tam padodot iepriekš atlasītos testa datus. Testa dati ir 1 % no katras klases ievaddatiem. Tabulā 4.6. skatīt atbildes rezultātus.

4.6. tabula

Kontinenta modeļa rezultāti

Sagaidāmā vērtība	Atbildes vērtība	Otrā atbildes vērtība
Dienvidamerika	Dienvidamerika - 98.08 %	Āfrika - 1.8%
Āfrika	Āfrika - 99.01 %	Dienvidamerika - 0.8%
Eiropa	Eiropa - 100 %	Ziemeļamerika - 0 %
Āzija	Āzija - 99.95 %	Eiropa - 0.04 %
Austrālija un Okeānija	Austrālija un Okeānija - 100 %	Ziemeļamerika - 0.0%
Ziemeļamerika	Ziemeļamerika - 99.89%	Eiropa - 0.07%

Tabula atspoguļo fragmentu no testēšanas piemēra. Šajā gadījumā ir redzams pa vienam ierakstam no katras klases ar sagaidāmo vērtību, atbildes vērtību un tās atbilstību procentos un otro labāko alternatīvas vērtību. Modeļa testēšanā tika norādīts izvēlēties top 3 vērtības, attiecīgi noteikt 3 atbilstošākās klases. Šīs klases kopā var neveidot 100% sadalījumu, jo 100% ir summa visām klasēm, starp kurām tiek izvēlēta klasificējamā vērtība. Tā klase, kurai ir vislielākā procentuālā vērtība ir klasificēšanas rezultāts. Šajā gadījumā tabulā atspoguļotas tikai pirmā un otrā vērtība, jo modeļa vērtību noteikšana ir ļoti izteikta un precīza, tādējādi 3 klase satur vai nu 0 % vai ļoti niecīgu daļu no procenta.

Katra klase tika testēta ar 1% no ieejas datiem kas ir aptuveni 400 testa ieraksti vienai klasei. Vērojot un analizējot rezultātus tika novērotas interesantas sakarības. Piemēram, gandrīz vienmēr Ziemeļamerikas otrā izvēle ir Dienvidamerika, bet Austrālijas un Okeānijas klasei gandrīz vienmēr otrā izvēle ir Ziemeļamerika. Sakarība novērojama arī starp Dienvidameriku un Āfriku, tās savstarpēji ir populārākā otrā izvēle viena otrai. Eiropas un Āzijas gadījumā abiem kontinentiem bieži kā otrā izvēle ir vienam starp otru un Ziemeļameriku. Patiesībā saskatāmas loģiskas sakarības, jo sakarības veidojas starp siltajām zemēm un tuvajiem kontinentiem. Testa datu kopā nebija nevienas vērtības, kas neatbilstu sagaidāmajam rezultātam. Līdz ar to, apskatot testa rezultātus var ņemt vērā, ka šī modeļa apmācība ir noritējusi veiksmīgi.

Kā otrais tika ģenerēts modelis, kas nosaka uz kādu pilsētu, nevis reģionu doties ceļotājam. Tātad klasificējamā problēma tiek palielināta un jau specificēta. Līdzīgi kā iepriekš modeļa klasificējamā klase tiek atlasīta pilsēta, bet netiek ņemti vērā lietotāju vārdi. Zemāk tabulā 4.7. modeļa apmācības dati.

4.7. tabula

Modeļa dati

Veids	**Klašu skaits**	**Validācijas datu apjoms**	**Modeļa precizitāte**	**Apmācības datu apjoms**
Pilsētu modelis	54	8%	45 %	~ 4000 ierakstu katrai klasei

Šajā modelī ir 54 klases. Tā pat kā iepriekš, pirms modeļa ģenerēšanas tika nodalīti validācijas un testēšanas dati. Šajā gadījumā datu nevienlīdzība ir vēl izteiktāka, līdz ar to, nākas noteikt, ka minimālais datu apjoms klasē ir aptuveni 3500, bet maksimālais 4500 ieraksti. Kā rezultātā populārāko pilsētu dati tiek samazināti pēc nejaušības metodes, bet klases, kurās apmeklējumu skaits ir neliels, netiek iekļautas modelī.

Šī modeļa ietvaros novērojama straujš precizitātes kritums. Visticamāk, tas saistīts ar samazināto datu daudzumu un informācijas nepietiekamību.

Šie iespējamās ierakstu variācijas ir ļoti daudz, līdz ar to vidēji 4000 ierakstu katrai klasei ir ļoti maz, lai spētu nodrošinātu precīzu modeļa apmācību. Tabulā 4.8. redzami testēšanas rezultāti.

4.8. tabula

Pilsētu modeļa rezultāti

Sagaidāmā vērtība	**Atbildes vērtība**	**Otrā atbildes vērtība**	**Trešā atbildes vērtība**
Barselona	Parīze - 79.73%	**Barselona** - 14.56 %	Amsterdama - 5.71 %
Honkonga	Singapūra - 57.60 %	Dubaija - 22.36%	**Honkonga** - 20.04%
Berlīne	Kalistoga - 94.95 %	Kapalua - 3.56%	**Berlīne** - 1.40 %
Dubaija	Napa - 75.98 %	Londona - 12.23 %	Hana - 11,79 %
Losandželosa	Monreāla - 85.01%	Madride - 10.44%	Losandželosa - 4.55 %
Sidneja	**Sidneja** - 65.74 %	Rinkona - 24.65 %	Monreāla - 9.61 %

Testa datos ir 1% datu no apmācības datu apjoma katrai klasei. Kā redzams rezultātu tabulā, tad gadījumi, kad modelis precīzi atrastu rezultātus ir ļoti maz. Šis, protams, ir tikai

neliels fragments no testa rezultātiem, taču tas atspoguļo reālo situāciju. Datu katru klašu ietvaros ir maz un kļūdīšanās koeficients ir ļoti augsts. Taču ir vairākas pazīmes, kas norāda uz to, ka potenciāli ar pareizu datu apjomu modeli būtu iespējams ģenerēt precīzāku. Pirmkārt, modelis bieži top 3 iekļauj savu pilsētu. Otrkārt, reti kad atbildes vērtība ir ar izteikti augstu procentuālo vērtību. Šīs pazīmes norāda uz to, ka modelis mācās, bet vēl nav apmācījies.

Interesanti novērojumi rezultātu analīzē saistāmi ar to, ka bieži atbildes vērtība ir vai nu tā paša kontinenta ietvaros, vai proporcionāli līdzīgā laika joslā novietota pilsēta. Patiesībā šāds daļēji apmācīts modelis būtu interesants tieši pavisam necerētu piedāvājumu radīšanai, kas acīmredzot veido kaut kādas kopsakarības.

Tā kā analizējamie dati ir ļoti neproporcionāli, tad testēšanas gaitā tika izlemts izstrādāt vēl trešo modeli, kas satur tikai abu Ameriku pilsētas kā potenciālo galamērķi. Zemāk tabulā modeļa parametri.

4.9. tabula

Amerikas modelis

Veids	**Klašu skaits**	**Validācijas datu apjoms**	**Modeļa precizitāte**	**Apmācības datu apjoms**
Amerikas pilsētu modelis	36	8%	56 %	~ 7000 ierakstu katrai klasei

Tā kā Ameriku gadījumā datu ir daudz vairāk, tad , lai izlīdzinātu rezultātus, kā augšējā robeža tika izvēlēta 7000 ierakstu kā maksimālais ierakstu skaits. Arī šajā gadījumā rezultāti tika dzēsti pēc nejaušības principa. Rezultātā kritērijiem atsilstošas ir 36 klases jeb 36 Ameriku pilsētas. Zemāk tabulā 4.10. aplūkojami testu rezultāti.

4.10. tabula

Amerikas pilsētu modeļa rezultāti

Sagaidāmā vērtība	**Atbildes vērtība**	**Otrā atbildes vērtība**	**Trešā atbildes vērtība**
Čikāga	**Čikāga** - 45.33%	Sietla - 34.86	Lasvegasa - 19.81
Dalasa	Vankūvera - 27.60 %	**Dalasa** - 12.31%	Orlando - 8.43%
Kapalua	Lasvegasa - 15.27 %	**Kapalua** - 8.76%	Sietla- 3.54 %
Lasvegasa	Vankūvera - 34.45%	**Lasvegasa** - 28.56%	Čikāga - 12.89%

Modelis tika testēts ar datiem, kur no katras klases iekļauts 1 % datu no apmācības kopas datu apjoma. Tabulā redzams maz fragments no testēšanas, kas atspoguļo reālo

situāciju. Šis modelis ir kļuvis mazliet precīzāks, taču, tā precizitāte un atbilžu vērību procentuālā vērtība vienalga ir ļoti zema. Tas norāda uz to, ka modelim būtu nepieciešams vairāk vai arī specifiskāki dati. Tomēr atkal ir novērojams tas, ka sagaidāmā vērtība gandrīz vienmēr ir starp top 3 piedāvājumiem. Šajā modelī daudz grūtāk saskatīt sakarības starp piedāvātajiem variantiem.

Jāpiebilst, ka arī pašas Amerikas ietvaros datu proporcionalitāte ir nosacīta, jo, piemēram, jau Ņujorkai vai Lasvegasai vien ir pāri 30 000 ierakstu. Tādējādi, veicot datu samazināšanu noteikti tiek zaudēta arī nepieciešamā informācija.

Kopumā redzams, ka ar esošajiem datiem nepietiek lai sniegtu precīzas, detalizētas prognozes specificētos galamērķos, taču pietiek, lai noteiktu kontinentu. Lai sniegtu drošas un tik tiešām ticamas, prognozes, nevis apšaubāmas vai patvaļīgas, kā to mēdz piedāvāt jau esošie ceļojumu portāli, ir nepieciešams gan detalizētāks datu apjoms, gan papildus datu priekšatlases filtri, kas veidotu lietotāju grupas, pēc to vienojošajiem faktoriem, tādējādi samazinot apstrādājamo un analizējamo datu skaitu.

Šī praktiskā darba rezultāti ir snieguši atbildi, ka ar pareiziem un detalizētiem datiem, nākotnes ceļojumus ir iespējams prognozēt individuāli, pēc konkrētā un citu lietotāju iezīmēm.

4. 12. Problēmas, izstrādātā modeļa risinājumā, ierobežojumi

Izstrādājot konkrēto risinājumu, nācās secināt, ka šādu problēmu risinājumā ir daudz problēmu un ierobežojumu, kas saistīti ar kvalitatīvu datu iegūšanu.

Viena no galvenajām problēmām ir datu trūkums un to *netīrība*. Datu trūkums neļauj veikt pilnīgus eksperimentus un noteikt precīzas sakarības, kā rezultātā iegūtie rezultāti var tikt pielīdzināmi minējumiem. Datu netīrība toties prasa lielu laika ieguldījumu to sakārtošanā. No datu aspekta nākamais punkts ir datu neproporcionalitāte. Neironu tīkla apmācības procesā, līdzīgi kā citos klasifikatoru apmācības procesos ir nepieciešama proporcionāla datu kopa. Pretējā gadījumā neironu tīkls iemācās atpazīt pārsvarā tikai tās klases, kas ir vairumā, tādējādi arī nekorekti atrodot citas klases.

No cita aspekta raugoties liels trūkums ir datu neesamība, jeb nepieejamība, konkrētāk tieši ietekmējošo faktoru datu nepieejamība. Daudz precīzākas prognozes būtu iespējams noteikt, ja prognozējamā lietotāja datus varētu analizēt pēc šī lietotāja personīgajiem datiem, kas saistīti ar interesēm, citiem pirkumiem, tādējādi iegūstot pieredzi no citiem līdzīgajiem ceļotājiem. Šī darba ietvaros nebija iespējams to realizēt praktiski. Konkrētu organizāciju gadījumos gan tā nav tik liela problēma, jo dati jau ir pieejami vai salīdzinoši viegli, ar lietotāju atļauju iegūstami.

No datu "tīrības" viedokļa jau iepriekš nodaļā apskatīta iespējamība, ka dati ir mākslīgi ģenerēti. Nodarbojoties ar pētniecību un brīvi pieejamiem datiem tas noteikti ir jāņem vērā un ir jāizvērtē iespējamā datu ģenerācija. Gadījumos, kad dati ir kāda konkrēta uzņēmuma rokās, tad datu mākslīgās ģenerēšanas iespējamība ir faktiski neiespējama vai arī ir salīdzinoši reāli iespējams noteikt vai dati ir vai nav ģenerēti.

Vēl viens faktors datu prognozēšanā ir datu patiesums. Ar to domāts vai paustie dati atbilst reālai cilvēka rīcībai. Šīs ceļojumu problēmas gadījumā, ja datu analīzi veic kāda kompānija, tad tās rīcībā ir reāli dati par cilvēku rīcību, taču, piemēram, intereses un citas papildu norādītās vērtības var neatspoguļot reālās cilvēka intereses. Šis jautājums veidojot datu analīzes stratēģijas noteikti ir jāņem vērā.

Iespējams laika gaitā, uzkrājoties lielam datu apjomam, ceļotāji cits citam var kļūt tik līdzīgi, ka kāds, kurš vienmēr izvēlas aktīvo atpūtu, pēkšņi gribētu doties atpūtā uz Ēģipti, visticamāk savā piedāvājumā šādu opciju neredzētu. Līdz ar to izriet secinājums, ka ir vērts kategorijām un ietekmējošiem faktoriem piešķirt kādus papildus koeficentus, kas noteiktu konkrēto atlases datu nozīmīgumu.

Vēl viens faktors, kas jāņem vērā, ir, kā iekļaut tās valstis, kas datu kopā netiek lietots, vai tieši pretēji tiek lietotas ļoti maz. Kā viens no iespējamajiem risinājumiem, būtu vai nu speciālu koeficentu pielietošana vai papildus datu ģenerēšana, taču tas būtu jāizpēta nopietnāk, un jāizvērtē iespējamie risinājumi.

Ievērojot tendenci, ka modelis ir precīzs, ja tiek izvērtēta kontinentu prognozēšana, bet valstu gadījumā tā precizitāte ir vāja, rodas ideja, ka iespējams šo neprecizitāti varētu uzlabot pielietojot vairāku tīklu mehānismu. Vispirms iegūstot kontinentu un tad atsevišķi izstrādājot katram kontinentam savu valstu modeli, noteikt iespējamās valstis. Tādējādi tiktu samazināts viena modeļa datu apjoms, un uzlabota precizitāte katra modeļa ietvaros. It sevišķi vērtīgi tas ir gadījumos, ja pieejami vairāk dati.

Protams, strādājot ar personas datiem, šobrīd ļoti aktuālā tēma un reizē arī ierobežojums ir personas datu aizsardzības regula. Strādājot ar personas datiem gan zinātnē, gan uzņēmējdarbībā ir jāievēro personas datu aizsardzības likums, tādējādi analīzei izmantojot tikai tādus datus, kas atbilst regulas prasībām.

Kopumā vērtējot rīka izstrādes gaitā radušos secinājumus par ierobežojumiem un problēmām, šis uzdevums un problēma ir laikietilpīgi risināmi jautājumi, kas lai tos atrisinātu prasa gan spēcīgas zināšanas, gan laiku šo problēmu risināšanā.

REZULTĀTI

Darba ietvaros tika izpētīti dažādi datu prognozēšanas metožu algoritmi, kas tiek plaši pielietoti gan datu prognozēšanā, gan dažādu klasifikācijas problēmu risināšanā. Darba autore izpētīja algoritmus teorētiskā līmenī, kā arī rada izpratni par algoritmu pielietošanas mērķiem praktiskajā darbībā.

Viens no darba uzdevumiem bija izprast ceļotāju izdarīto izvēļu ietekmējošos faktorus. Šī mērķa sasniegšanai veikta cilvēku uzvedības analīze, izvēloties ceļojumu galamērķus. Rezultātā noskaidrots, ka cilvēka izvēli dodoties ceļojumā ietekmē ne tikai ekonomiskie apstākļi, kā tas šķiet visbiežāk, bet gan arī daudzi citi faktori, kā, piemēram, ģeogrāfiskie aspekti, drošība, laikapstākļi, intereses un citi. Iepriekš uzskaitītie faktori parasti ir sistemātiski un katram ceļotājam nosakāmi vienmēr, tātad faktori, kuri var tikt izmantoti prognožu noteikšanai. Interesanti, ka pastāv arī kategorija, kuru sauc par *ad hoc* kategoriju. Šīs kategorijas faktori nav iepriekš prognozējami un katram ceļotājam var būt noteicoši tikai konkrētajā reizē. Šādi faktori ir faktiski neprognozējami. Tiem var būt un ļoti noteicoša nozīme ceļojuma galamērķa izvēlē konkrētajā reizē. Rezultātā secināts, ka ceļojumu prognozēm var būt tikai ieteicoša nevis galēja nozīme, jo skaidri zināms, ka nav zināmi visi būtiskie ierobežojošie faktori *ad hoc* kategorijas dēļ.

Datu prognozēšanas svarīgākais objekts ir dati un to kvalitāte. Darba ietvaros darba autore izzināja kvalitatīvu datu nozīmību, un izpētīja un uzzināja, kā pareizi sagatavot datus korektu prognožu veikšanai. Tā kā datu analīzē būtiska ir arī pareiza datu atspoguļošana, tika apskatīti dažādi pamata datu atspoguļošanas veidi.

Darba praktiskajā daļā veikta esošo datu prognozēšanas rīku izpēte un analīze. Tā rezultātā izdarīts secinājums, ka nepieciešams izstrādāt jaunu datu prognozēšanas stratēģiju ceļojumu mērķu noteikšanai. Kā stratēģijas pamata prognozēšanas modelis izvēlēts neironu tīklu apmācība. Tālāk veikta brīvi pieejamo ceļojumu datu izpēte, analīze un sagatavošana. Datu apstrādei izstrādāts speciāls rīks, kas veic datu atlasi un sagatavošanu neironu tīkla apmācībai. Pēc datu sagatavošanas, tika veikta trīs dažādu prognožu modeļu apmācība, testēšana un iegūto datu analīze.

Rezultātā tika sasniegts izvirzītais mērķis, izpētīt un izprast datu prognozēšanas procesus un metodes, un izstrādāt nelielu rīku ceļojumu galamērķu prognozēšanai. Izstrādātā prognožu modeļa rezultāti sniedza vērtīgus secinājumus par izvēlētās stratēģijas pielietošanu ceļojumu prognožu izvēlē. Rezultātā secināts, ka ar neironu tīklu modeļiem ir iespējams veikt šādas prognozes, taču galvenais faktors ir gan korektu, gan atbilstošu datu esamība.

SECINĀJUMI

Datu daudzums internetā ik dienas pieaug lielos apjomos, vienas dienas laikā tas tiek palielināts par 10^{18} baitiem, līdz ar to pieaug arī iespējas šos datus analizēt un izmantot lietderīgi. Ir pieejamas dažādas datu prognozēšanas metodes dažādu problēmu risināšanai. Viens no secinājumiem, izpētot datu prognozēšanas metodes, ir prognozes mērķa un idejas nozīmīguma izprašana. Ar esošajām datu prognozes metodēm, tās pareizi pielietojot, iespējams risināt pat ļoti komplicētas problēmas, kā prognozēt laikapstākļus, uzņēmuma peļņu un citas problēmas.

Izstrādājot konkrētas problēmas prognozēšanas stratēģiju ir jāizprot dati ne tikai virspusējā līmenī, bet ir jāizprot to jēga un ietekme. Piemēram, ja vēlamies veikt cilvēku uzvedības prognozēšanu, tad vispirms jāsāk ar cilvēku uzvedības izpēti, faktoriem un ārējiem apstākļiem, kas ietekmē cilvēka uzvedību konkrētās dzīves situācijās. Tikai pēc tam varam izvēlēties datus, ar kuriem veikt prognozes. Svarīgi apzināties arī pieejamo datu patiesumu un izvērtēt prognozējamo problēmu, kā arī apzināties, vai šādu problēmu vispār ir iespējams prognozēt, izmantojot rīcībā esošās metodes un rīkus.

Datu pareiza saturiskā izvēle un to kvalitāte ir nozīmīgākie ietekmējošie faktori, datu prognožu stratēģijas un modeļa izstrādē. Šis secinājums nosaka, ka datu tīrība, datu korektums, to atbilstība prasītajam un datu tipiem, un, pats galvenais, datu pieejamība ir viens no svarīgākajiem aspektiem datu prognozēšanā. Šī darba ietvaros tika izstrādāta prognozēšanas stratēģija un rīks ceļotāju galamērķu prognozēšanai. Izstrādājot stratēģijas modeļus, tika secināts, ka pieejamie dati ir nepietiekami, lai varētu veikt detalizēti pielāgotus ceļojumu piedāvājumus. Taču balstoties uz modeļu testu rezultātiem tika secināts, ka ar detalizētākiem datiem datu prognozes ceļojumu izvēlē ir iespējams veikt.

Izstrādātā metode un rīks ir labs sākums tālākajiem pētījumiem, ko pilnveidot un papildināt precīzākām prognozēm. Viens no veidiem, kā padarīt rīku precīzāku, būtu veikt datu priekšatlases pēc noteiktiem likumiem, tādējādi samazinot analizējamo datu apjomu. Kā arī, balstoties uz iegūtajiem rezultātiem modeļu testēšanas gaitā, noderīga šķiet iespējamā modeļu kombinēšana, piemēram, vispirms atlasot kontinentu, vai vēlamo mēnesi, vai kādu no interesēm un pēc tam prognozējot konkrētus galamērķus.

Kopumā secinu, ka darbs ir izstrādāts sekmīgi, jo sasniegti izvirzītie mērķi, ir iegūts vispārīgs priekšstats par datu prognozēšanas metodēm un izstrādāts praktisks datu prognozēšanas rīks.

IZMANTOTĀ LITERATŪRA

1) Gregory Piatetsky-Shapiro, Ph.D, *KDnuggets*, blogs; https://www.kdnuggets.com/, tiešsaiste, skatīts [04.04.2018.].

2) "Data Mining - Classification & Prediction", blogs; https://www.tutorialspoint.com/data_mining/dm_classification_prediction, tiešsaiste, skatīts [06.04.2018.].

3) Jānis Zuters Dr.dat, "Neironu tīkli"; http://home.lu.lv/~janiszu/courses/eanns/eanns.pdf, tiešsaiste, skatīts [12.05.2018].

4) Charu C. Aggarwal, "*Data Classification Algorithms and Applications*", CRC Press, 2015.

5) Isabelle Guyon, Andre Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, V-3, Nr.1, 2003, lpp 1157-1182.

6) Kārlis Podnieks, "Varbūtību teorija un matemātiskā statistika ", lekciju materiāli, 2015; http://podnieks.id.lv/slides/mining/varbut2.pdf, tiešsaiste, skatīts [07.04.2018].

7) "Decision Trees"; http://scikit-learn.org/stable/modules/tree.html, tiešsaiste, skatīts [07.04.2018].

8) Prashant Gupta, "Decision Trees in Machine Learning", maijs, 2017; https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052, tiešsaiste, skatīts [05.05.2018].

9) Sholom M. Weiss, Nitin Indurkhya, "Rule-based machine learning methods for functional prediction", *Journal of Artificial Intelligence Research*, V-3 Nr.1, 1995 , lpp 383-403.

10) "Data Mining - Rule Based Classification "; https://www.tutorialspoint.com/data_mining/dm_rbc.htm, tiešsaiste, skatīts [08.04.2018].

11) David W. AHA, Denis Kibler , Marc K. Albert , "Instance-Based Learning Algorithms", *Journal of Machine Learning*, V-6, Nr.1, 1991., lpp 37-66.

12) " Support Vector Machines "; http://scikit-learn.org/stable/modules/svm.html, tiešsaiste, skatīts [15.04.2018.].

13) Sunil Ray, "Understanding Support Vector Machine algorithm from examples", septembirs, 2017; https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/, tiešsaiste, skatīts [15.04.2018.].

14) "Support Vector Machines (SVM) Introductory Overview" ; http://www.statsoft.com/Textbook/Support-Vector-Machines, tiešsaiste, skatīts [15.04.2018.].

15) A. Zisserman, "The SVM classifier", lekciju materiāli, 2015; http://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf, tiešsaiste, skatīts [15.04.2018].

16) "Propagation function", blogs; https://www.nnwj.de/, tiešsaiste, skatīts [18.04.2018.].
17) MathWorks, "Predictive Analytics 3 Things You Need to Know", blogs; https://se.mathworks.com/, tiešsaiste, skatīts [18.04.2018.].
18) "Classification & Prediction, Basic Concepts of Classification and Prediction", lekciju materiāli, 2012; http://www.inf.unibz.it/dis/teaching/DWDM/slides2012/lesson9-Classification1.pdf, tiešsaiste, skatīts [20.04.2018.].
19) "Classification & Prediction, Rule Based Classification ", lekciju materiāli, 2012; http://www.inf.unibz.it/dis/teaching/DWDM/slides2012/lesson11-12-Classification3.pdf, tiešsaiste, skatīts [20.04.2018.].
20) Počs R. "Kvantitatīvās metodes ekonomikā un vadīšanā", 2003.
21) "Vienkāršā lineārā regresija un korelācija", lekciju materiāli; https://estudijas.rtu.lv/file.php/63844/Matematiska_statistika/regresija_un_korelacija.pdf, tiešsaiste, skatīts [03.04.2018.].
22) "Nelineārā regresija", http://ezis.appspot.com/Statistika/d.10.htm, tiešsaiste, skatīts [15.04.2018.].
23) "Korelācijas un regresijas analīzes elementi", lekciju materiāli, 2016; http://www.politeh.lv/elibrary/DATA/MT000002/Korelac_2016.pdf, tiešsaiste, skatīts [15.04.2018.].
24) Dr. Mashud Kabir, "Similarity matching techniques for fault diagnosis in automotive infotainment electronics", *IJCSI International Journal of Computer Science Issues*, V-3, 2009, lpp 1694-0784.
25) "3 Data Science Methods and 10 Algorithms for Big Data Experts", decembris, 2017; https://datafloq.com/read/data-science-methods-and-algorithms-for-big-data/2500, tiešsaiste, skatīts [17.04.2018.].
26) The minitab blogs, "What Is the Difference between Linear and Nonlinear Equations in Regression Analysis", jūlijs, 2017, blogs; http://blog.minitab.com/blog/adventures-in-statistics-2/what-is-the-difference-between-linear-and-nonlinear-equations-in-regression-analysis, tiešsaiste, skatīts [17.04.2018.].
27) "Non-linear Regression in R for biologist"; http://rstudio-pubs-static.s3.amazonaws.com/7812_5327615eb0044cf29420b955ddaa6173.html, tiešsaiste, skatīts[17.04.2018.].
28) Lawrence W.Barsalou, "Deriving Categories to Achieve Goals", *Psychology of Learning and Motivation*, V-27, 1991, lpp 1-64.
29) Federico Castanedo, "Data Preparation in the Big Data Era", *O'Reilly Media*, 2015.

30) CISCO,"*The Zettabyte Era*", blogs; https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html, tiešsaiste, skatīts [12.04.2018.].
31) Alon Halevy, Peter Norvig, and Fernando Pereira ,"The Unreasonable Effectiveness of Data ", *IEEE Intelligent Systems*, V-24, I-2, 2009, lpp 1541-1672.
32) "Digital Audio"; https://www.techopedia.com/definition/226/digital-audio, tiešsaiste, skatīts [21.04.2018.].
33) "Sound"; https://www.computerhope.com/jargon/s/sound.htm", tiešsaiste, skatīts [21.04.2018.].
34) Arthur Juliani "Recognizing Sounds", aprīlis, 2016; https://medium.com/@awjuliani/recognizing-sounds-a-deep-learning-case-study-1bc37444d44d, tiešsaiste, skatīts [21.04.2018.].
35) Fadi Elawar, "Improving Predictive Analytics with Data Visualization"; https://data-visualization.cioreview.com/cxoinsight/improving-predictive-analytics-with-data-visualization-nid-15311-cid-163.html, tiešsaiste, skatīts [20.04.2018.].
36) ScienceSoft, "Big data visualization techniques: a quick intro", blogs; https://www.scnsoft.com
37) "Line graph"; https://www.dr-aart.nl/Statistics-frequency-polygon.html, tiešsaiste, skatīts [20.04.2018.].
38) "Bar chart examples"; http://www.conceptdraw.com/How-To-Guide/bar-chart-examples, tiešsaiste, skatīts [20.04.2018.].
39) SmartDraw, "Pie chart", blogs; https://www.smartdraw.com/pie-chart/
40) Josh Parker's Environmental Biology Blog, blogs; http://enb-2011f-jp.blogspot.com/2011/09/proportional-symbol-map.html, tiešsaiste, skatīts [20.04.2018.].
41) Jen Underwood, "The Art & Science of Data Preparation for Predictive Analytics", jūlijs, 2016; http://www.impactanalytix.com/samples/DataPrepforPredictiveAnalytics.pdf, tiešsaiste, skatīts [25.04.2018.].
42) StatisticSolutions, "Time Series Analysis" , blogs; http://www.statisticssolutions.com/, tiešsaiste, skatīts [25.04.2018.].
43)"Classification" , https://www.cs.sfu.ca/~jpei/publications/Sequence%20Classification.pdfs, tiešsaiste, skatīts[25.04.2018.].
44) Singleton, Patrick Allen, "A Theory of Travel Decision-Making with Applications for Modeling Active Travel Demand", 2013, doktora disertācija.

45) Anshul Garg, "Travel Risks vs Tourist Decision Making: A Tourist Perspective", jūlijs, 2015.
46) Jeffrey J. LaMondia, Tara Snell, Chandra Bhat, "Traveler Behavior and Values Analysis in the Context of Vacation Destination and Travel Mode Choices", *Transportation Research Record: Journal of the Transportation Research Board* , V-2156, 2016.
47) Vinay Kumar, "Linear Regression Using Python scikit-learn", novembris, 2017; https://dzone.com/articles/linear-regression-using-python-scikit-learn, tiešsaiste, skatīts [07.05.2018.].
48) "Gadījumnotikumu varbūtības un darbības ar varbūtībām"; http://ezis.appspot.com/Statistika/d.02.htm, tiešsaiste, skatīts [07.05.2018.].
49) "Instance-Based Learning", lekciju materiāli; https://pdfs.semanticscholar.org/presentation/cc26/a24c9cc725b6f15295e98da004c8d68c9887.pdf, tiešsaiste, skatīts [07.05.2018.].
50) Facebook, "Forecasting at scale"; https://facebook.github.io/prophet/, tiešsaiste, skatīts[17.04.2018.].
51) IBM, "IBM SPSS Software"; https://www.ibm.com/, tiešsaiste, skatīts [17.04.2018.].
52) SAS, "SAS reimagines its data science portfolio"; https://www.sas.com/en_us/news/analyst-viewpoints/forrester-names-sas-leader-in-predictive-analytics-machine-learning.html, tiešsaiste, skatīts [17.04.2018.].
53) Gill Press, "A Very Short History Of Big Data", blog; https://www.forbes.com.
54) Edīte Brikmane, "Kas ir personas dati? Vispārīgā datu aizsardzības regula I", aprīlis, 2018; https://lvportals.lv/skaidrojumi/294871-kas-ir-personas-dati-vispariga-datu-aizsardzibas-regula-i-2018, tiešsaiste, skatīts [05.05.2018.].
55) Latvijas likumdošana, Fizisko personu datu aizsardzības likums, I nodaļa, Vispārīgie noteikumi.
56) "Eiropas parlamenta un padomes regula (es) 2016/679" , Eiropas Savienības Oficiālais Vēstnesis, maijs, 2016.
57) Pasaules valstu apzīmējumu standarts; https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv, tiešsaiste, skatīts [04.03.2018.].
58) Pasaules valstu apzīmējumu standarts; https://datahub.io/core/world-cities, tiešsaiste, skatīts [04.03.2018.].
59) Gavin Brown, "Feature Selection", lekciju materiāli; http://www.cs.man.ac.uk/~nogueirs/files/SLIDES-master.pdf, tiešsaiste, skatīts[04.05.2018.].
60) Datu kopas; http://times.cs.uiuc.edu/~wang296/Data/, tiešsaiste, skatīts [04.03.2018.].

61) Siddharth Das, "CNNs Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more", novembris, 2015; https://medium.com/@siddharthdas_32104/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5, tiešsaiste, skatīts [04.05.2018.].
62) Yann, "LeNet-5", blogs; http://yann.lecun.com/exdb/lenet/, tiešsaiste, skatīts[04.05.2018].
63) "Convolutional Neural Networks"; http://cs231n.github.io/convolutional-networks/, tiešsaiste, skatīts [04.05.2018.].
64) Adil Moujahid, "A Practical Introduction to Deep Learning with Caffe and Python", jūnijs, 2016; http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/, tiešsaiste, skatīts [05.05.2018.].
65) "The History Of Predictive Analytics - Infographic", jūnijs, 2017; https://datafloq.com/read/history-predictive-analytics-infographic/438, tiešsaiste, skatīts[05.05.2018.].

PIELIKUMS

1. pielikums. TripAdvisor datu piemērs


```
{
   "Ratings":{
      "Overall":"4.0"
   },
   "AuthorLocation":"80241",
   "Title":"\u201cNice hotel for the Pioneer Square Area\u201d",
   "Author":"Janko",
   "ReviewID":"UR5767903",
   "Content":"We stayed here in late August. This hotel is a decent stay for a decent price for this time of year. The service is awesome. They wo
   "Date":"September 4, 2006"
}
],
"HotelInfo":{
   "Name":"BEST WESTERN PLUS Pioneer Square Hotel",
   "HotelURL":"/ShowUserReviews-g60878-d72572-Reviews-BEST_WESTERN_PLUS_Pioneer_Square_Hotel-Seattle_Washington.html",
   "Price":"$117 - $189*",
   "Address":"<address class=\"addressReset\"> <span rel=\"v:address\"> <span dir=\"ltr\"><span class=\"street-address\" property=\"v:street-
address\">77 Yesler Way</span>, <span class=\"locality\">
<span property=\"v:locality\">Seattle</span>, <span property=\"v:region\">WA</span> <span property=\"v:postal-code\">98104-2530</span>
</span> </span> </span> </address>",
   "HotelID":"72572",
   "ImgURL":"http://media-cdn.tripadvisor.com/media/ProviderThumbnails/dirs/51/f5/51f5d5761c9d693626e59f8178be15442large.jpg"
}
```


2. pielikums. Valstu kodu standarta datu fragments


```
name,alpha-2,alpha-3,country-code,iso_3166-2,region,sub-region,intermediate-region,region-code,sub-region-code,intermediate-region-code
Afghanistan,AF,AFG,004,ISO 3166-2:AF,Asia,Southern Asia,"",142,034,""
Åland Islands,AX,ALA,248,ISO 3166-2:AX,Europe,Northern Europe,"",150,154,""
Albania,AL,ALB,008,ISO 3166-2:AL,Europe,Southern Europe,"",150,039,""
Algeria,DZ,DZA,012,ISO 3166-2:DZ,Africa,Northern Africa,"",002,015,""
American Samoa,AS,ASM,016,ISO 3166-2:AS,Oceania,Polynesia,"",009,061,""
Andorra,AD,AND,020,ISO 3166-2:AD,Europe,Southern Europe,"",150,039,""
Angola,AO,AGO,024,ISO 3166-2:AO,Africa,Sub-Saharan Africa,Middle Africa,002,202,017
Anguilla,AI,AIA,660,ISO 3166-2:AI,Americas,Latin America and the Caribbean,Caribbean,019,419,029
Antarctica,AQ,ATA,010,ISO 3166-2:AQ,"","","","","",""
Antigua and Barbuda,AG,ATG,028,ISO 3166-2:AG,Americas,Latin America and the Caribbean,Caribbean,019,419,029
Argentina,AR,ARG,032,ISO 3166-2:AR,Americas,Latin America and the Caribbean,South America,019,419,005
Armenia,AM,ARM,051,ISO 3166-2:AM,Asia,Western Asia,"",142,145,""
Aruba,AW,ABW,533,ISO 3166-2:AW,Americas,Latin America and the Caribbean,Caribbean,019,419,029
Australia,AU,AUS,036,ISO 3166-2:AU,Oceania,Australia and New Zealand,"",009,053,""
```


...

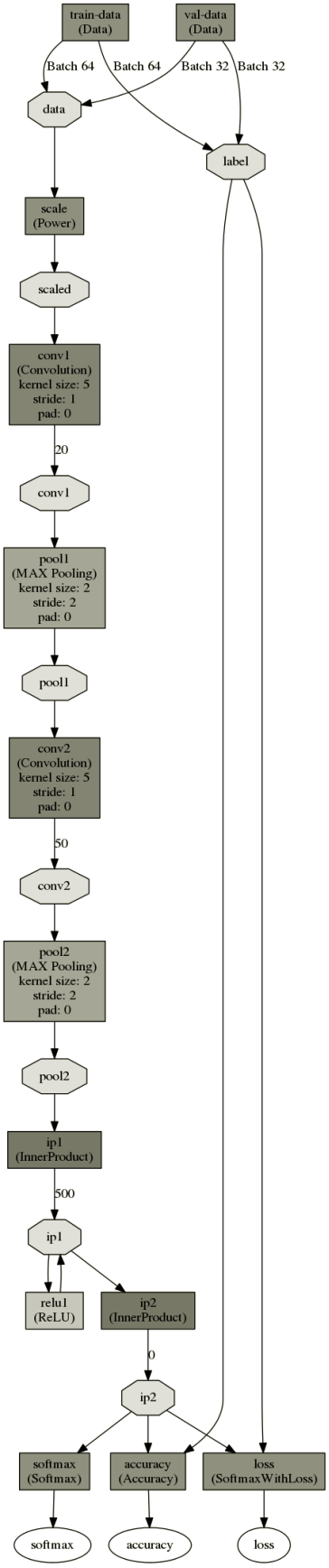

```
AL|Albania
DZ|Algeria
AS|American Samoa
AD|Andorra
AO|Angola
AI|Anguilla
AQ|Antarctica
AG|Antigua And Barbuda
AR|Argentina
AM|Armenia
AW|Aruba
```


...

3. pielikums Valstu un pilsētu savstarpējās piederības datu piemēra fragments

```
name,country,subcountry,geonameid
les Escaldes,Andorra,Escaldes-Engordany,3040051
Andorra la Vella,Andorra,Andorra la Vella,3041563
Umm al Qaywayn,United Arab Emirates,Umm al Qaywayn,290594
Ras al-Khaimah,United Arab Emirates,RaÊ¼s al Khaymah,291074
Khawr FakkÄn,United Arab Emirates,Ash ShÄriqah,291696
Dubai,United Arab Emirates,Dubai,292223
Dibba Al-Fujairah,United Arab Emirates,Al Fujayrah,292231
Dibba Al-Hisn,United Arab Emirates,Al Fujayrah,292239
Sharjah,United Arab Emirates,Ash ShÄriqah,292672
Ar Ruways,United Arab Emirates,Abu Dhabi,292688
Al Fujayrah,United Arab Emirates,Al Fujayrah,292878
Al Ain,United Arab Emirates,Abu Dhabi,292913
Ajman,United Arab Emirates,Ajman,292932
...
```

4. pielikums. Neironu tīklu struktūra



Bakalaura darbs „**Lietotāja uzvedības prognozēšanas rīks ceļojumu izvēlē**” izstrādāts LU Datorikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autore: _____ Līna Briņģe __.05.2018.

Rekomendēju/nerekomendēju darbu aizstāvēšanai

Vadītājs: profesors, Dr. phil. Jurģis Šķilters _____ __.05.2018.

Recenzents: docents, Dr. dat. Viesturs Vēzis

Darbs iesniegts Datorikas fakultātē __.05.2018.

Dekāna pilnvarotā persona:

vecākā metodiķe Ārija Sproģe _____

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē

__.06.2018. prot. Nr. ___.

Komisijas sekretārs/e: _____