

**KAUNO TECHNOLOGIJOS UNIVERSITETAS
MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

Ugnė Kardauskaitė

**PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO PAGAL
DUOMENIS GAUTUS IŠ PALYDOVŲ MODELIS**

Baigiamasis magistro projektas

Vadovas

Doc. dr. Vytautas Janilionis

KAUNAS, 2017

**KAUNO TECHNOLOGIJOS UNIVERSITETAS
MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

**PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO PAGAL DUOMENIS
GAUTUS IŠ PALYDOVŲ MODELIS**

Baigiamasis magistro projektas
Taikomoji matematika (621G10003)

Vadovas

Doc. dr. Vytautas Janilionis

Recenzentas

Doc. dr. Tomas Ruzgas

Projektą atliko

Ugnė Kardauskaitė

KAUNAS, 2017



KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ

(Fakultetas)

Ugnė Kardauskaitė

(Studento vardas, pavardė)

Taikomoji matematika, 621G10003

(Studijų programos pavadinimas, kodas)

AKADEMINIO SAŽININGUMO DEKLARACIJA

20 17 m. birželio 1 d.
Kaunas

Patvirtinu, kad mano, **Ugnės Kardauskaitės**, baigiamasis projektas tema „PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO PAGAL DUOMENIS GAUTUS IŠ PALYDOVŲ MODELIS“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

(vardą ir pavardę įrašyti ranka)

(parašas)

Kardauskaitė U. Pastatų pamatų nusėdimo vertinimo pagal duomenis gautus iš palydovų modelis: Taikomosios matematikos magistro darbas / vadovas doc. dr. V. Janilionis; Taikomosios matematikos katedra, Matematikos ir gamtos mokslų fakultetas, Kauno technologijų universitetas. – Kaunas, 2017. – 50 p

SANTRAUKA

Vykstant potvyniams ar atoslūgiams užfiksuojama gruntinio vandens kaita, kuri sukelią pastatų pamatų nusėdimo problema. Ši problema sukelią didelę žalą pastato struktūrai. Siekiant išvengti galimos žalos sudaromi modeliai gebantis naudojantis palydovinėmis matavimų technologijomis gautais pastatų pamatų nusėdimo pokyčio matavimais, bei žinomais ekspertiniais vertinimais, įvertinti pastatų pamatų nusėdimo riziką.

Uždavinio sprendimui pasirinktą sudaryti pastatų pamatų nusėdimo rizikos vertinimo modelį, paremtą laiko eilučių klasifikavimo rezultatų analizę. Atlikus bandymus nustatyta, kad analizuojamų duomenų pastatų pamatų nusėdimo vertinimo modelio sudarymui geriausia taikyti atraminių vektorių klasifikatorių grįsta DBT spektro koeficientais.

Pritaikius sudarytą pastatų pamatų nusėdimo rizikos vertinimo modelį gautą, jog iš 412 analizuojamų pastatų su tikimybe didesnę arba lygią 0,6, 7 pastatai buvo priskiriami aukščiausios rizikos grupei, 384 pastatai buvo priskirti vidutinės rizikos grupei, o dešimt pastatų žemiausios rizikos grupei.

Kardauskaitė U. Model of building foundation subsidence assessment using satellite measurements data: Master's work in applied mathematics / supervisor assoc. doc. dr. V. Janilionis; Department of Mathematical Modelling, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology. – Kaunas, 2017. – 50 p.

SUMMARY

During the floods and ebbs the change of ground water is observed that cause buildings foundation subsidence. This problem has huge damage to structure of the building. In order to avoid the damage, there are created models, that can sort buildings in different risk groups by data obtained from satellite and expert evaluation. The satellite measure the change of subsidence of buildings foundations.

To solve building foundation subsidence risk assessment problem was chosen to create model based on machine learning model of subsidence measurement classification results analysis.

Created model application show that out of 412 analyzed buildings with a probability greater or equal to 0.6, 7 buildings were classified as high risk, 384 buildings have been assigned a medium risk, while the ten buildings in the lowest risk group.

TURINYS

SANTRUMPOS	7
LENTELIŲ SĄRAŠAS	8
PAVEIKSLĖLIŲ SĄRAŠAS	9
ĮVADAS	11
1. TEORINĖ DALIS.....	12
1.1. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO UŽDAVINYS IR JO SPRENDIMO METODŲ APŽVALGA.....	12
1.1.1. PASTATŲ PAMATŲ NUSĖDIMAS	12
1.1.2. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO UŽDAVINYS.....	12
1.2. KLASIFIKAVIMO METODŲ APŽVALGA IR PALYGINIMAS	14
1.2.1. KLASIFIKAVIMO METODO PARINKIMAS.....	15
1.3. LAIKO EILUČIŲ TRANSFORMACIJŲ APŽVALGA.....	21
1.4. DARBE SPRENDŽIAMŲ UŽDAVINIAI	24
2. TIRIAMOJI DALIS IR REZULTATAI.....	25
2.1. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIS	25
2.1.1. ANALIZUOJAMŲ DUOMENŲ PARUOŠIMAS.....	26
2.1.2. ARTIMIAUSIŲ KAIMYNŲ MODELIS GRĮSTAS DWT METRIKA.....	26
2.1.3. DISKREČIŲJŲ BANGELIŲ TRANSFORMACIJOS TAIKYMAS LAIKO EILUTĖMS IR SPEKTRO KOEFICIENTŲ ATRINIKIMAS.....	27
2.1.4. ATRAMINIŲ VEKTORIŲ MODELIO SUDARYMAS	28
2.1.5. KLASIFIKAVIMO MODELIO VERTINIMAS	29
2.2. PROGRAMINĖ MODELIO SUDARYMO REALIZACIJA.....	31
2.3. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIO SUDARYMAS IR TAIKYMAS.....	32
2.3.1. ANALIZUOJAMI DUOMENYS.....	32
2.3.2. DBT TAIKYMAS LAIKO EILUTĖMS IR KOEFICIENTŲ ATRINKIMAS...	34
2.3.3. ATRAMINIŲ VEKTORIŲ MODELIS GRĮSTAS DBT KOEFICIENTAIS.	35

2.3.4. ARTIMIAUSIŲ KAIMYNŲ MODELIS GRĮSTAS DBT KOEFICIENTAIS ..	38
2.3.5. ARTIMIAUSIŲ KAIMYNŲ MODELIS GRĮSTAS DWT METRIKA	40
2.3.6. KLASIFIKAVIMO MODELIŲ PALYGINIMAS.....	43
2.3.7. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIO PROGRAMINĖ REALIZACIJA IR TAIKYMO REZULTATAI	46
IŠVADOS	48
LITERATŪROS SĄRAŠAS	49
1 PRIEDAS. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIO TAIKYMO REZULTATAI	51
2 PRIEDAS. PASTATŲ PAMATŲ NUSĖDIMO MODELIO SUDARYMO SUDARYMO PROGRAMINĖ REALIZACIJA	60

SANTRUMPOS

DTW Dinaminis laiko skalės iškraipymas (angl. *Dynamic Time warping*)

DBT Diskrečioji bangelių transformacija (angl. *Discrete wavelet transform*)

LENTELIŲ SĄRAŠAS

1.1 lentelė Klasifikavimo metodų palyginimas (**** žymi aukščiausią vertinimą * žemiausią)	20
2.1 lentelė Sumaišymo matrica	30
2.2 lentelė Klasifikavimo kokybės vertinimo metrikos	30
2.3 lentelė F1 metrikos reikšmės prie skirtingų parametrų γ ir C reikšmių (aukščiausios rizikos grupė)	36
2.4 lentelė F1 metrikos reikšmės prie skirtingų parametrų γ ir C reikšmių (vidutinės rizikos grupė)	37
2.5 lentelė F1 metrikos reikšmės prie skirtingų parametrų γ ir C reikšmių (žemiausios rizikos grupė)	37
2.6 lentelė Apibendrinta F1 metrikos reikšmės prie skirtingų parametrų γ ir C reikšmių.....	37
2.7 lentelė Atraminių vektorių modelio grįsto DBT koeficientais sumaišymo matrica	43
2.8 lentelė Atraminių vektorių modelio grįsto DBT koeficientais klasifikavimo kokybės metrikos	44
2.9 lentelė Artimiausių kaimynų modelio grįsto DBT koeficientais sumaišymo matrica.....	44
2.10 lentelė Artimiausių kaimynų modelio grįsto DBT koeficientais klasifikavimo kokybės metrikos	44
2.11 lentelė Artimiausių kaimynų modelio grįsto DTW metrika sumaišymo matrica.	45
2.12 lentelė Artimiausių kaimynų modelio grįsto DTW metrika klasifikavimo kokybės metrikos	45

PAVEIKSLĖLIŲ SĄRAŠAS

1.1 pav. Palydovo taikymas nusėdimo vertinimui	13
1.2 pav. Klasifikavimo modelio sudarymo etapų schema	15
1.3 pav. Sprendimų medžio pavyzdys	16
1.4 pav. Neuroninis tinklas.....	17
1.5 pav. Tiesė atskiriamų plokštumos taškų pavyzdys.....	18
1.6 pav. Trajektorijos minimizuojančios astumą tarp nagrinėjamų eilučių pavyzdys	21
1.7 pav. PAA transformacijos pavyzdys	24
2.1 pav. Pastatų pamatų nusėdimo vertinimo modelio sudarymo schema.....	25
2.2 pav. „Vienas-prieš-vieną“ metodo gautas 3 klasių klasifikatorius.....	28
2.3 pav. Aukščiausios rizikos grupės pastatų pamatų nusėdimo matavimų laiko eilutės pavyzdys.....	33
2.4 pav. Vidutinės rizikos grupės pastatų pamatų nusėdimo matavimų laiko eilutės pavyzdys.....	33
2.5 pav. Žemiausios rizikos grupės pastatų pamatų nusėdimo matavimų laiko eilutės pavyzdys.....	34
2.6 pav. Pamatų nusėdimo matavimų laiko eilutės skaidymo DBT spektro koeficientais pavyzdys.....	34
2.7 pav. Fisher skiriamumo santykio DBT koeficientams grafikas	35
2.8 pav. F1 metrikos priklausomybė nuo atraminių vektorių klasifikatoriaus parametro C	36
2.9 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (aukščiausios rizikos grupė).....	38
2.10 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (vidutinės rizikos grupė)	39
2.11 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (žemiausios rizikos grupė)	39
2.12 pav. F1 metrikos priklausomybė nuo parametro k	40
2.13 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (aukščiausios rizikos grupė).....	41
2.14 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (vidutinės rizikos grupė)	42
2.15 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (žemiausios rizikos grupė)	42

2.16 pav. F1 metrikos priklausomybė nuo parametro k	43
--	-----------

ĮVADAS

Netolygiai kintant vandens kiekiui grunte esančiame po pastato pamatais, susiduriama su pastatų pamatų nusėdimo problema. Ši problema dažniausiai pastebima jūrinėse valstybėse: Nyderlanduose, Didžiojoje Britanijoje, Italijoje ir kitose valstybėse. Pastatų pamatų nusėdimas gali sukelti didelę žalą pastato struktūrai. Siekiant išvengti galimos žalos ir iš anksto informuoti apie gresiantį pavojų, paskutiniu metu pradėtos naudoti palydovinės matavimų technologijos pastatų pamatų nusėdimo pokyčiams matuoti. Šiuo metu aktualus uždavinys - pagal šiuos matavimo duomenis įvertinti pastatų pamatų nusėdimo riziką.

Darbo tikslas – pagal pastatų pamatų nusėdimo matavimus gautus iš palydovų ir dalies pastatų rizikos ekspertinio vertinimo duomenis sudaryti pastatų rizikos vertinimo modelį.

Teorinėje dalyje pateikta sprendžiamos problemos analizė, klasifikavimo metodų ir laiko eilučių požymių išskyrimo metodų analizė bei aptarti darbe sprendžiami uždaviniai.

Tiriamąjame ir rezultatų dalyje pateikta pastatų pamatų nusėdimo vertinimo modelio sudarymo shema. Kiekvieno modelio sudarymo etapo vykdymui pasiūlyti metodai. Modelio sudarymo schema realizuota programiškai. Darbe pateikti rezultatai gauti sudarant pastatų pamatų nusėdimo vertinimo modelį pagal Nyderlandų Zanstado regiono duomenis gautus iš palydovų.

Pritaikius sudarytą modelį atsitiktinai atrinktų 412 pastatų imties vetinimui, gauta, kad 7 pastatai priskirti aukščiausios rizikos grupei, 384 pastatai priskirti vidutinei rizikos grupei.

Darbo tematika buvo perskaityti pranešimai dvejose konferencijose:

- Konferencija „Matematika ir matematikos dėstymas” – KTU, 2017;
- XV tarptautinė studentų konferencija „Matematika ir gamtos mokslai: teorija ir taikymas” – KTU, 2017;

1. TEORINĖ DALIS

1.1. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO UŽDAVINYS IR JO SPRENDIMO METODŲ APŽVALGA

1.1.1. PASTATŲ PAMATŲ NUSĖDIMAS

Pastatų pamatų nusėdimu vadinamas procesas, kai dėl grunto sąlygų pasikeitimo pastato pamatai juda žemyn. Kai kada nusėdimu vadinamas ir apatinio grunto sluoksniu po pamatu pakilimas. Nusėdimas dažnai yra susijęs su gruntinio vandens kiekiu po pamatais esančiame grunte pokyčiu. Skirtingo tipo gruntai pasižymi nevienoda nusėdimo atsiradimo tikimybe. Ypač pavojingi yra molingi gruntai, nes jie traukiasi ir plečiasi priklausomai nuo besikeičiančio juose esančios drėgmės kiekio. Arti pamatų augantys augalai taip pat gali būti nusėdimo priežastis, kadangi jų šaknys iš grunto siurbia vandenį ir gruntas susitraukia. Didesnis pavojus kyla, kai užsitęsia sausas laikotarpis, nes augalų šaknys, ieškodamos vandens, skverbiasi pro pamatus. Kartais iš pažeistos nuotekų sistemos besisunkiantis vanduo gali suminkštinti po pamatais esantį gruntą arba jį visai išplauti. Tai sukelia grunto poslinkį ir tuo pačiu pastato poslinkį žemyn [2]. Remiantis literatūros šaltiniu [3] galime paminėti, jog pamatų nusėdimą be pokyčių grunte, taip pat gali sukelti: įdubimai virš požeminių ertmių, staigus gruntinio vandens lygio pakylimas išplaunantis gruntą esantį po pamatais.

Vienas didžiausių nusėdimo požymių yra pastate ir jo sienose atsiradę plyšiai. Dėl nusėdimo atsiradę plyšiai paprastai eina įstrižai iš viršaus į apačią, jie staiga atsiranda pastato sienuose. Dėl nusėdimo atsiradę plyšiai paprastai yra platesni nei dėl kitų priežasčių atsiradę plyšiai, be to, dažniausiai viršuje jie būna platesni nei apačioje. Galimų nusėdimų požymiai taip pat gali susiformuoti, prie pastato durų ir langų. Dėl pamatų nusėdimo durys arba langai nustoja laisvai atsidarinti/užsidarinti [8].

1.1.2. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO UŽDAVINYS

Pastatų pamatų nusėdimo vertinimo uždavinio sprendimo metodiką galima skaidyti į dvi dalis: vizualinę analizę, periodinių matavimų analizę. Vizualiniai analizėi atlikti ekspertai vyksta į stebimame regione esančius pastatus ir ištiria visus su nusėdimo susijusius požymius ir tikėtinas jų atsiradimo priežastis. Periodinių matavimų analizę galime išskaidyti į du etapus: pamatų nusėdimo periodinis matavimas ir surinktų matavimų analizę. Pamatų nusėdimo matavimams atlikti naudojamos metodikos literatūroje dažniausiai skirstomos į dvi grupes: grunto pokyčio matavimai naudojantis geodeziniais prietaisais ir palydovinėmis technologijomis paremti matavimai [19].

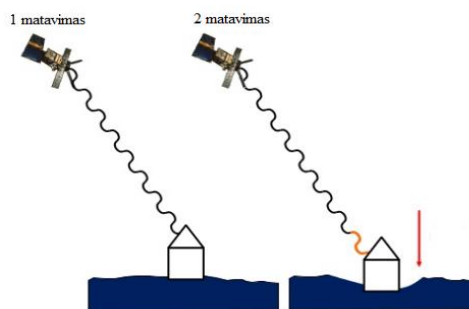
Grunto pokyčio matavimai naudojant geodezinius prietaisus.

Šiai grupei galime priskirti įvairias sausumos matavimo sistemas, kurios yra naudojamos stebėti ir kiekybiškai vertinti žemės paviršiaus deformacijas: optinis gulstainis (angl. *Optical levelling*), elektroninis atstumo matavimo (angl. *Electronic distance measurement (EDM)*) horizontalaus judėjimo vertinimo gulstainis. Taikant šiuos prietaisus pastatų pamatų nusėdimo vertinimui pasiekiamas ~ 10 mm tikslumas. Pagrindinis visų šių prietaisų veikimo principas yra pasirinktų pastatų taškų aukščio virš jūros lygio pokyčio periodinis vertinimas. Norint panaudoti šiuos prietaisus pastatų nusėdimo vertinimui reikia nemažai žmogiškųjų resursų, dėl to galima įvertinti tik nedidelės teritorijos pastatų pamatų nusėdimą. [19].

Palydovinėmis technologijomis grįsti matavimai.

Palydovinėmis technologijomis grįsti matavimai atlieka nusėdimų matavimus nereikalaujant žmogaus įsikišimo matavimų metu, bei užtikrinant *mm* matavimų tikslumą. Dažniausiai literatūroje minimos [19],[14] šios grupės matavimų technologijos:

- **Globalios padėties nustatymo sistema** (angl. *Global Positioning System*) arba (GPS) stebėjimai, nusėdimai atskiruose taškuose gali būti matuojamas kasdien matavimo tikslumas $<5\text{ mm}$. GPS tinklas turi būti įdiegtas visame stebimame regione ir stebimas pakartotinai. Pageidautina, kad tinklas būtų eksploatuojamas nepertraukiamai.
- **InSAR** (angl. *Interferometric Synthetic Aperture Radar*) palydovas užtikrina aukštą erdvinę skiriamąją gebą ir *mm* matavimų tikslumą, bei suteikia galimybę išmatuoti dešimčių km^2 dydžio teritorijas. Matavimas atlikti nereikalingi papildomi įrenginiai žemėje. O priklausomai nuo palydovinės misijos nusėdimo vertinimą galima atlikti kelių dienų ar kelių savaitių periodiškumu. Šio palydovo veikimo technologija paremta dviem etapais:
 1. Pirmo matavimo metu išsaugomas atspindžio nuo matuojamo taško bangos ilgis.
 2. Tolesnių matavimų metu 1 matavimas laikomas atskaitos tašku pastatų pamatų nusėdimo vertinimui. Šiems matavimas apskaičiuojamas skirtumas tarp gauto atspindžio bangos ilgio ir pirmojo matavimo atspindžio bangos ilgio [14].



1.1 pav. Palydovo taikymas nusėdimo vertinimui

Atlikus matavimus įmonėms, kurios stebi ir atlieka pastatų pamatų rekonstravimą dėl pamatų nusėdimo, vis dar lieka klausimų: Ar galima prognozuoti tolesnį pastato nusėdimą, kokią pastato nusėdimo riba yra rizikinga? Kada pastatui reikalinga pamatų renovacija, o kada būtina atlikti vizualinę pastato analizę ir jį stebėti?

Siekiant atsakyti į keletą iš šių klausimų, literatūroje rasti keli siūlomi sprendimo būdai. Literatūroje [13], analizuojama priklausomybė tarp inžinerinių pastatų parametrų ir pamatų nusėdimo įvertinimo. Analizuojami didelio telekomunikacijų pastato nusėdimo matavimai ir inžineriniai parametrai: pamatų gylis, pamatų polių ilgis, polių skaičius, išoriniai vietos faktoriai. Naudojant Gauso proceso mašinio mokymo modelį sudaromos nusėdimo vertinimo priklausomybės nuo matavimų ir išorinių faktorių. Gautas modelis yra tikslus (paklaida 5,3 %).

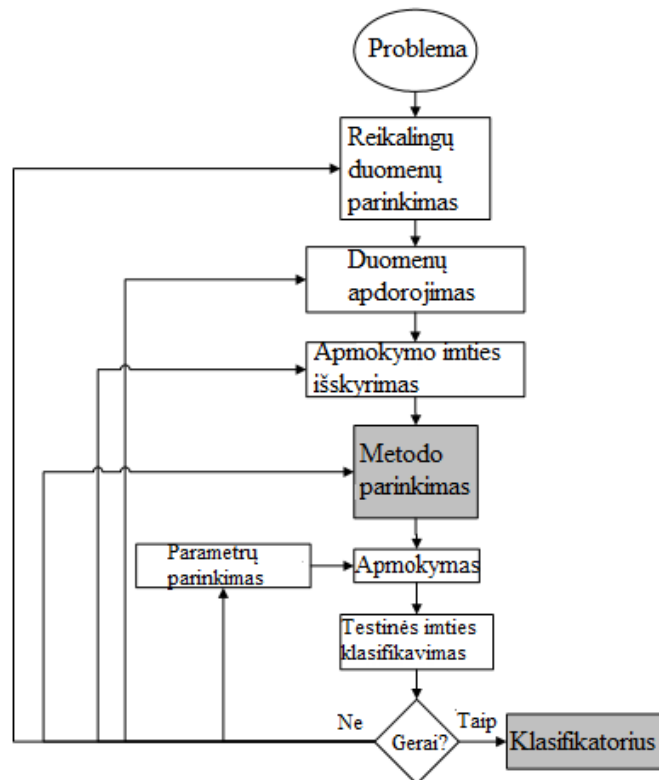
Kitas literatūroje [11] pateiktas metodas nusėdimo matavimų eilutes aprašo skirstiniais ir gautus skirstinius naudoja prognozavimui. Darbe nagrinėjami Gauso, Beta, Veibulo, Rayleigh, Nakagami skirstinių tinkamaumas nusėdimo eilutėms aprašyti. Tinkamiausiu pripažintas Beta skirstinys.

Įvertinti nagrinėjamų pastatų pamatų nusėdimo riziką, literatūroje [3], buvo atliekama gautų palydovo matavimų laiko eilučių analizė. Išskyrus iš gautų laiko eilučių tendą, buvo atliekamas eilučių tendą aprašančių kreivių klasifikavimas. Jei nėra trendo, tuomet daroma prielaida, kad pastatas nepasirizikuoja pamatų nusėdimo rizika. Jei eilutės trendas pasirodo staigiomis mažėjimo ir ar pakilimo atkarpomis, tokiam pastatui galime priskirti vidutinės rizikos lygį, o tolygiai mažėjančio trendo eilutėms apibūdinami pastatai priskiriamas aukštas rizikos lygis.

Atlikus straipsnių apie pastatų pamatų nusėdimo problemą ir/ar gautų pastatų pamatų nusėdimų matavimų analizę, galime apibendrinti jog su šios problemos sprendimu susijusioje literatūroje dažniausiai analizuojama matavimų metodika ir tikslumas, o ne patys matavimai. Atsižvelgiant į sprendžiamo uždavinio specifiką ir turimus duomenis darome prielaidą, kad pastatų pamatų nusėdimo rizikos vertinimui galima siūlyti taikyti klasifikavimo metodus ir mašininio mokymo metodiką. Tolesniuose literatūros apžvalgos skyriuose apžvelgsime klasifikavimo metodus ir jų taikymą laiko eilučių analizei.

1.2. KLASIFIKAVIMO METODŲ APŽVALGA IR PALYGINIMAS

Klasifikavimu vadinama duomenų tyrybos metodų grupė, kurie naudojami objektų priskyrimui iš anksto žinomoms klasėms. Šie metodai dar vadinami mokymosi su mokytoju metodais (angl. *supervised learning technique*). Klasės identifikuojantis kintamasis statistikoje vadinamas priklausomu kintamuoju. Po apmokymo, klasifikavimo modelis (klasifikatorius), taikomas naujų duomenų klasifikavimui. Klasifikavimo procesą apibūdina 1.1 pav. pateikta schema [9].



1.2 pav. Klasifikavimo modelio sudarymo etapų schema

Pirmame klasifikavimo modelio sudarymo etape svarbu tinkamai parinkti duomenis, kurie apibūdina sprendžiamą problemą. Tam dažniausiai konsultuojamasi su sprendžiamos problemos srities ekspertais arba naudojamosi statistiniais ar algoritminiais savybių parinkimo metodais. Kitame etape atliekamas atrinktų duomenų apdorojimas: išsiskiriančių stebėjimų šalinimas, trūkstančių reikšmių įvertinimas. Apdorotą duomenis, sudarome modelio mokymui naudojamą imtį. Tolesnis ir pats svarbiausias klasifikavimo modelio sudarymo etapas yra klasifikavimo metodo parinkimas. Parinkus klasifikavimo metodą, atliekamas mokymas bei siekiama parinkti tokius modelio parametrus, kad gautume kuo mažesnę mokymo imties paklaidą. Paskutiniame etape po mokymo gautą modelį panaudojame testinės imties klasifikavimui, t.y. atliekame sudarytojo klasifikavimo modelio tikslumo įvertinimą. Jei gautas tikslumas tenkina procesą, stabdome, jei ne – nuo pasirinktojo etapo kartojame procesą dar kartą [1].

1.2.1. KLASIFIKAVIMO METODO PARINKIMAS

Literatūros šaltiniuose [9,1,17], klasifikavimo metodo parinkimo etapas išskiriamas kaip vienas iš svarbiausių ir turinčių didelę įtaką siekiant gero klasifikavimo tikslumo. Apžvelgdami klasifikavimo metodų gausą, galime išskirti keletą metodų grupių: statistiniai mokymosi metodai, loginių taisyklių metodai, pavyzdžiais grindžiami metodai, neuroninių tinklų metodai, atraminių vektorių klafikatoriai [9].

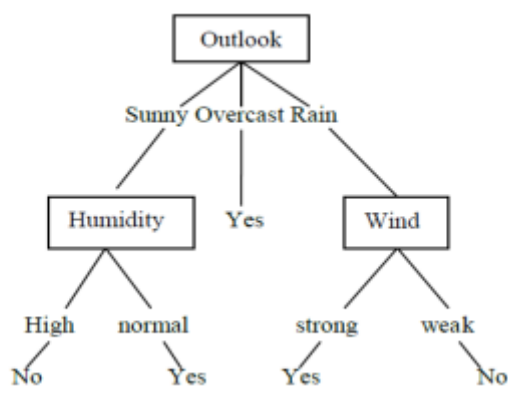
Statistiniai mokymosi metodai

Statistinio mokymosi metodai yra pirmieji mašininio mokymosi metodai. Šie metodai sukurti naudojantis statistikos ir funkcinės analizės teorija. Statistiniai mokymosi metodai, naudodamiesi duomenų rinkiniu, siekia sudaryti prognozavimo funkciją. Naudojantis šios grupės metodais, sudaryta funkcija skiriasi nuo kitų metodų tuo, kad jos rezultatas nėra konkrečios klasės objektui priskyrimas, o tikimybių, kad objektas priklauso bet kuriai iš galimų klasių, rinkinys. Praktikoje dažniausiai taikomi ir literatūroje minimi šios grupės metodai yra: Naivusis Bajeso klasifikatorius, Bajeso tinklai, tiesinė diskriminantinė analizė.

Loginių taisyklių metodai

Šie metodai dažniausiai yra skirstomi į dvi pagrindines grupes – sprendimo medžius ir taisyklėmis paremtus klasifikatorius.

Sprendimų medžiai (angl. *Decision tree*) – yra statistinio modeliavimo metodas, naudojamas klasifikuoti objektus pagal juos aprašančius vektorius (X_n, y_n) , kur X_n – objektą aprašančių požymių vektorius, o y_n – žymi objekto klasę, $n = 1, 2, \dots, N$, kur N - objektų skaičius. Sprendimo medžiai pasižymi struktūrinės diagramos (medžio) tipo struktūra (žr. 1.3 pav.), kur kiekvienas vidinis mazgas žymi objekto požymio testą, o kiekviena šaka žymi testo rezultatus, ir kiekvienas išorinis mazgas (lapas) žymi klasę, kuriai priskiriamas objektas [10].



1.3 pav. Sprendimų medžio pavyzdys

Sprendimo medžio sudarymą galime išskaidyti į du etapus:

1. medžio augimo (angl. *tree building*). Naudojantis pasirinktą sprendimo medžio sudarymo metodiką kiekviename iš mazgų atliekant, testą sudaromas sprendimo medis;
2. medžio genėjimo (angl. *tree pruning*). Siekiant sumažinti klasifikatoriaus sudėtingumą, atliekami medžio struktūros pakeitimai, neturintys įtakos prognozavimo tikslumui [10].

Taisyklėmis paremtais klasifikatoriais aprašoma kiekviena iš klasių, naudojant disjunkcinę normalinę formą (DNF). k -DNF išraiškos forma $(X_1 \wedge X_2 \wedge \dots \wedge X_n) \vee (X_{n+1} \wedge X_{n+2} \wedge \dots \wedge X_{2n}) \vee \dots \vee (X_{(k-1)n+1} \wedge X_{(k-1)n+2} \wedge \dots \wedge X_{kn})$, kur k yra disjunkcijų skaičius, o n yra konjunkcijų skaičius

kiekviename iš disjunktų. Modelio mokymo metu siekiame, naudodamiesi mokymo imtimis, sudaryti kiek galima mažesnę taisyklių rinkinį, skirtą kiekvienai iš klasių aprašyti [1].

Pavyzdžiais grindžiami metodai

Šie metodai yra laikomi tingaus mokymosi metodais, nes jie neturi išskirto mokymo etapo. Kita vertus, kiekvienam naujai klasifikuojamam objektui atliekamas atskiras mokymas pagal tai, į kuriuos mokymo imties objektus jis panašiausias.

Vienas iš šios grupės populiariausių metodų yra k artimiausių kaimynų metodas. Šis metodas priklauso vektorinių metodų grupei. Jo esmė yra duomenų atvaizdavimas specialioje vektorinėje erdvėje. Vektorinės erdvės dimensijos atitinka klasifikavimui naudojamų požymių aibę. Pagrindiniai metodo etapai:

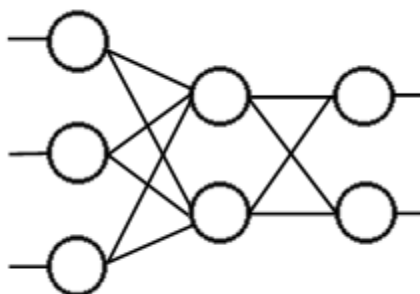
1. Objektui iš mokymo aibės parenkami k artimiausių objektų pagal tam tikrą apibrėžtą atstumo arba panašumo matą.
2. Objektas priskiriamas klasei, kurios objektų yra daugiausia tarp artimiausių k kaimynų.

Modelių, gautų naudojantis šiuo metodu, tikslumui daugiausiai įtakos turi atstumo arba panašumo matas bei parametras k [1].

Neuroninių tinklų metodai

Neuroniniai tinklai buvo sukurti remiantis žmogaus smegenų veikimo principais (t. y. neuronų tarpusavio sąveika). Neuronai tarpusavyje yra sujungiami jungtimis, kuriomis yra perduodami signalai (žr. 1.4 pav.). Kiekviena jungtis turi svorį, kuris nurodo jungties stiprumą ir modifikuoja perduodamą signalą. Pagrindinės neuroną sudarančios dalys: sumatoriaus ir aktyvavimo funkcijos. Sumatorius sumuoja visus ateinančius signalus iš kitų neuronų, prieš tai juos padauginęs iš jungčių svorių, o aktyvavimo funkcija transformuoja gautą rezultatą.

Dažniausiai naudojama aktyvavimo funkcija yra sigmoidinė $\sigma(x) = \frac{1}{1 + e^{-x}}$.



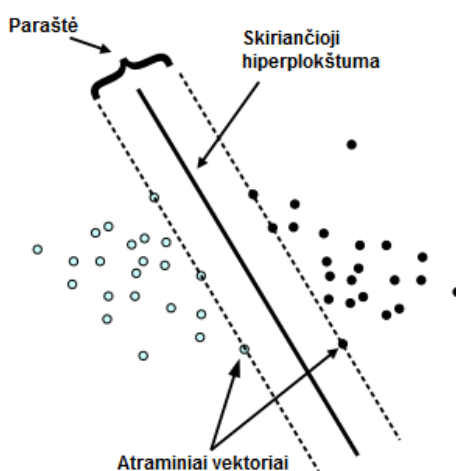
1.4 pav. Neuroninis tinklas

Dažniausiai neuronai yra grupuojami į sluoksnius ir kiekvienas sluoksnio neuronas yra sujungiamas su gretimų sluoksnių visais neuronais. Pirmasis sluoksnis vadinamas įvedimo, o visi

kiti, išskyrus paskutinįjį, vadinami paslėptaisiais, o paskutinis vadinamas išvedimo. Į pirmąjį sluoksnį yra perduodami požymių vektoriai, kuriuos reikia apdoroti. Tada tinklas yra įvykdomas, ir išvedimo sluoksnyje gaunamas apskaičiuotas rezultatas [1].

Atraminių vektorių klasifikatoriai

Pagrindinis atraminių vektorių klasifikatoriaus (angl. *Support vector machines*) metodo tikslas yra sudaryti optimalią hiperplokštumą su kiek galima didesniu atstumu atskiriančią klasifikuojamas grupes. Tinkamai parinkus optimalią hiperplokštumą, nagrinėjami duomenys dažniausiai yra transformuojami į didesnio matavimo dimensijos erdvę ir tik tada ieškoma optimalios skiriančios hiperplokštumos. Transformuotoje erdvėje arčiausiai parinktos optimalios hiperplokštumos esantys duomenų vektoriai vadinami atraminiais vektoriais (žr. 1.5 pav.) [18].



1.5 pav. Tiesė atskiriamų plokštumos taškų pavyzdys

Metodų palyginimas

Siekdami apibendrinti ir palyginti atrinktų klasifikavimo metodų grupių dažniausiai naudojamus metodus, sudarėme 1.1 lentelę bei toliau apžvelgėme pagrindinius metodų privalumus ir trūkumus.

Vertindami metodų jautrumą duomenų rinkinio požymių tipams, galime išskirti atraminių vektorių klasifikatorių ir neuroninių tinklų metodus, kaip gerbūčius pasiekti gerą tikslumą daugiamatiams, tolydžiū požymiū rinkiniams. Tuo tarpu loginiū taisykliū metodai dažniausiai pasiekia geriausius rezultatus sprendžiant kategoriniais/diskrečiais požymiais aprašomas problemas. Neuroninių tinklū ir atraminių vektoriū mašinū metodū geram klasifikavimo tikslumui pasiekti, reikalinga mokymo imtis, sudaryta iš didelio skaičiaus stebėjimū, tuo tarpu aukšto tikslumo Naiviojo Bajeso klasifikatoriaus sudarymas galimas su sąlyginai maža imtimi.

Vertindami mokymui reikalingą laiką, galime išskirti Naivųjį Bajeso klasifikatoriū, kaip vieną iš mažiausių laiko resursū reikalaujančių metodū. Taip pat galime paminėti vienmatį

sprendimo medžių modelį, kurio mokymui reikalingas laikas net kelis kartus mažesnis už neuroninių tinklų ar atraminių vektorių.

Išsiskiriantys stebėjimai daugiausiai gali sumažinti artimiausių kaimynų klasifikatoriaus tikslumą. Šis metodas paprastai laikomas išskiriančių stebėjimų netoleruojančiu. Priešingai nei artimiausių kaimynų metodas, dauguma sprendimų medžių metodų yra laikomi toleruojančiais išsiskiriančius stebėjimus.

Sprendžiant klasifikavimo problemą, kai duomenų rinkinio požymiai pasižymi multikolinearumu, mažiausiai jautrūs ir tinkamiausi būtų neuroninių tinklų ir atraminių vektorių mašinų metodai.

Mokymui naudojant duomenų rinkinį su įtrauktais nereikšmingais požymiais, pastebimas artimiausio kaimyno metodo klasifikavimo jautrumas, šis bruožas gali būti paaiškintas tuo, kaip algoritmas veikia. Be to, nereikšmingų savybių buvimas mokymo imtyje gali padaryti neuroninių tinklų mokymą labai neefektyvų, o gautą rezultatą nepraktišką.

Klasifikavimo metodų palyginimas (**** žymi aukščiausią vertinimą * žemiausią)

	Sprendimo medžiai	Neuroniniai tinklai	Naivusis Bajeso	k artimiausių kaimynų	Atraminų vektorių
Apibendrintas modelio tikslumas	**	***	*	**	****
Klasifikavimo greitis	****	****	*	****	****
Mokymosi greitis atsižvelgiant į požymių ir stebėjimų skaičių	***	*	****	****	*
Trūkstatų reikšmių toleravimas	***	*	****	*	**
Koreliuojančių požymių toleravimas	**	***	*	*	***
Nesusijusių požymių toleravimas	***	*	**	**	****
Išsiskiriančių stebėjimų toleravimas	**	**	**	**	****
Galimybė panaudoti diskrečiuosius/binariusius/polydžiuosius požymius	****	***(ne diskretus)	***(ne tolydus)	***	** (ne diskretus)
Permokymo rizikos vertinimas	**	*	***	***	**
Galimybė interpretuoti rezultatus	****	*	****	**	*
Modelio jautrumas parametrui	***	*	****	***	*

Dažniausiai parenkant tinkamiausią klasifikavimo metodą nagrinėjamai problemai spręsti, remiantis duomenų savybėmis atrenkama keletas klasifikavimo metodų. Atlikus atrinktų metodų mokymą su ta pačia mokomąja imtimi, atliekamas klasifikavimo tikslumo įvertinimas,

panaudojant sudarytuosius modelius testinės imties klasifikavimui. Metodas, kurį taikant gautas aukščiausias klasifikavimo tikslumas, laikomas tinkamiausiu nagrinėjamai klasifikavimo problemai spręsti.

1.3. LAIKO EILUČIŲ TRANSFORMACIJŲ APŽVALGA

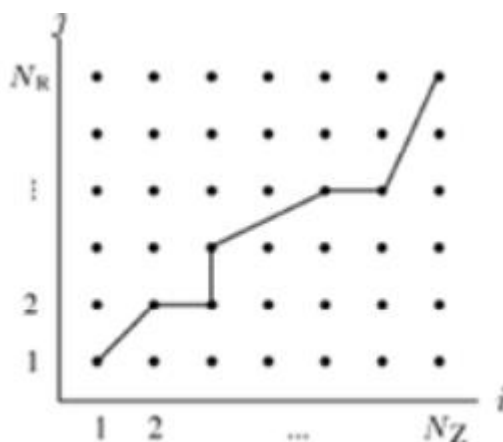
Laiko eilutėmis vadinami statistiniai duomenys, surinkti reguliariais laiko intervalais [9]. Vieną iš svarbesnių laiko eilučių savybių yra egzistuojantis koreliacinis ryšys tarp matavimų atliktų skirtingais laiko momentais. Remiantis klasifikavimo modelio sudarymo metodika yra žinoma, jog klasifikavimo naudojami objektai turi būti aprašyti nekoreliuojančiais požymiais. Taip pat laiko eilutės dažniausiai aprašomi ilgą laiką stebimi procesai, dėl to klasifikavimo modelio sudarymui reiktų pateikti didelės dimensijos požymių vektorius, kas lėtina klasifikavimo metodo veikimą, gali turėti neigiamos įtakos klasifikavimo tikslumui ir rezultatų interpretavimui.

Siekiant išvengti paminėtų laiko eilučių klasifikavimo problemų prieš atliekant klasifikavimą, nagrinėjamos laiko eilutės taikomi transformavimo ar panašumo įvertinimo metodai, kurie panaikina stebėjimų koreliaciją ir sumažina matavimo dimensiją.

Toliau aptarsime keletą dažniausiai literatūroje sutinkamų metodų.

Dinaminis laiko skalės iškrypimas (angl. *Dynamic Time warping*)

Dinaminis laiko skalės iškrypimas (DTW) tai metodas naudojamas vertinti laiko eilučių panašumą. Šis metodas eilučių panašumo vertinimui naudoja atstumo metriką. Atstumams apskaičiuoti siūlomi įvairūs metodai: tikėtinumų santykio logaritmo matas, Čebyševio atstumas, tačiau dažniausiai naudojamas Euklido atstumas [18]. Nagrinėjant laiko eilutes $R = (r_1, r_2, \dots, r_N)$, $N \in \mathbb{N}$ ir $Z = (z_1, z_2, \dots, z_N)$, $N \in \mathbb{N}$ pirmiausia suformuojama $N \times M$ matrica, kurios (i, j) elementas yra Euklido atstumo tarp stebėjimų kvadratas $d(r_i, z_j) = (r_i - z_j)^2$. Iš gautos matricos siekiant kuo tiksliau įvertinti panašumą tarp nagrinėjamų eilučių naudojantis dinaminio programavimo ieškoma trajektorija minimizuojanti suminį atstumą tarp taškų [6].



1.6 pav. Trajektorijos minimizuojančios atstumą tarp nagrinėjamų eilučių pavyzdys

Laiko eilučių klasifikavime šis metodas dažniausiai naudojamas kaip artimiausių kaimynų klasifikavimo metodo atstumo matas. Išskiriant šio metodo minusus ir plusus galime paminėti didelių laiko ir skaičiavimo resursų reikalingumą, tačiau galimybę pasiekti aukštą klasifikavimo tikslumą naudojant DTW kaip 1 – artimiausių kaimynų klasifikatoriaus atstumo matą [6].

Diskrečioji bangelių transformacija (angl. *Discret wavelet transform*)

Diskrečioji bangelių transformacija (DBT) transformuoja laiko eilutę į laiko-dažnio skalę. Šią transformaciją apdorojant laiko eilute, formuojamas kintamo dydžio (mastelio) langas, kuris slenkamas išilgai laiko (erdvės) ašies, ir kiekvienai lango pozicijai apskaičiuojamas bangelių spektras (transformacija). Transformacijos rezultatas – signalo išraiškų (vaizdavimų pagal dažnį ir erdvėje) rinkinys. Kiekvienoje diskrečiosios bangelių transformacijos (DBT) iteracijoje panaudojama taip vadinama mastelio funkcija laiko eilutės apdorojimui. Jeigu pradinis signalas X turi N reikšmių, tai mastelio funkcija bus panaudota tam, kad būtų apskaičiuotos $N/2$ suvidurkintos reikšmės. Be mastelio funkcijos laiko eilutės transformavimui kiekvienoje iteracijoje panaudojama bangelių funkcija. Jei pradinis signalas X turi N reikšmių, tai bangelių funkcija bus pritaikyta tam, kad apskaičiuoti $N/2$ skirtumines (atspindinčias informacijos pasikeitimus signale X) reikšmes. Sekančiose iteracijose abi funkcijos (mastelio ir bangelių) taikomos tikrai suvidurkintų reikšmių vektoriui, gautam prieš tai atliktoje iteracijoje. Po baigtinio iteracijų skaičiaus gaunamas laiko eilutės transformacijos koeficientų rinkinys sudarytas iš suvidurkintos reikšmės (gautos n – toje iteracijoje) ir skirtuminių reikšmių rinkinio (gauto ankstesnėse iteracijose) [15-4].

Transformavimui naudojamos mastelio ir bangelių funkcijų sistemos: Haaro, Daubechies, Meyer, Hermitian, Morlet, Le Gall. Praktikoje laiko eilučių transformavimui dažniausiai naudojama Harro funkcijų sistema.

Literatūroje [6] iškiriamos šios pagrindinės DBT savybės:

- Transformacija tinka nestacionarioms ir neilgoms laiko eilutėms.
- Apdoroja tik 2^n ilgio eilutes.
- Haaro transformacija pasižymi lokališkumo savybe, t.y. kuo didesnis transformacijos koeficiento numeris, tuo mažesnę eilutės dalį jis atspindi.
- Bangelių transformacijų apskaičiavimo algoritmas yra $O(n)$ sudėtingumo.

Klasifikavimo modelio apmokymui naudojant diskrečiųjų bangelių transformaciją, siekiant sumažinti matavimų dimensiją reikalinga atlikti reikšmingų koeficientų parinkimą. Remiantis šaltiniais [4,20] galime išskirti bendrojoje bangelių naudojimo teorijoje dažniausiai naudojama k pirmų reikšmių parinkimo metodika. Tačiau ji neatsižvelgia į klasių atskiriamumą, siekianti

atrinkti tuos koeficientus, kurie pasižymi reikšmių skirtumu tarp klasių dažniausiai naudojamas Fišerio skiriamumo santykio vertinimas.

Diskrečioji kosinusinė transformacija (angl. *Discrete cosine transform*)

Diskrečioji kosinusinė transformacija (DCT) kaip ir diskrečioji Furjė transformacija skirta siekiant transformuoti laiko eilutės stebėjimų priklausomybę nuo laiko į stebėjimų priklausomybę nuo dažnių. Skirtingai nei diskrečioji Furjė transformacija DCT naudojasi tik kosinusinėmis funkcijomis transformacijai atlikti. Diskrečioji kosinusinė transformacija laiko eilutei x , kurios ilgis n apibrėžiama kaip :

$$X_f = K(f) \sum_{i=1}^n x_i \cos \frac{\pi f(i-0.5)}{n} \quad f = 0, \dots, n-1 \quad (1.1)$$

kur $K(f) = \frac{1}{\sqrt{n}}$, kai $f = 0$ ir $K(f) = \sqrt{\frac{2}{n}}$ kai $1 \leq f \leq n-1$.

Literatūroje [6] išskiriamos šios pagrindinės DCT transformacijos savybės:

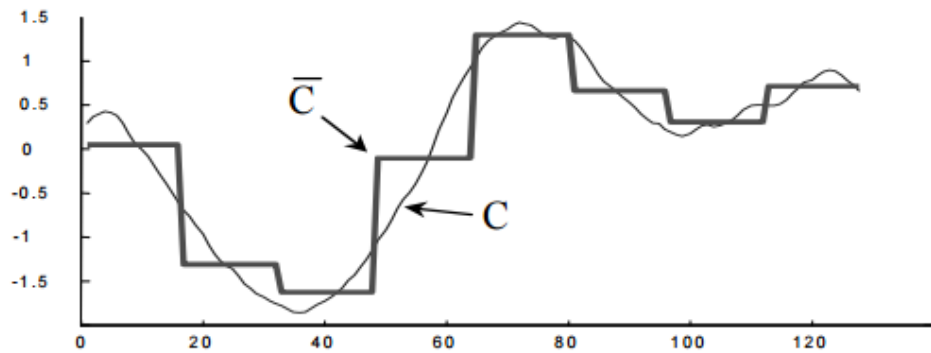
- Naudojantis greitu DCT koeficientu apskaičiavimo algoritmu koeficientu apskaičiavimo uždavinys tampa $O(n \log(n))$ sudėtingumo.
- Apskaičiuoti koeficientai visada yra realus skaičiai.
- Geba apdoroti laiko eilutes su trendo komponentę.
- Transformuojamos eilutės turi pasižymėti periodiškumu (sezoniškumu).

Aproksimavimas dalinių sekų agregavimo metrikų reikšmėmis (angl. *Piecewise Aggregate Approximation*)

Laiko eilutė C , kurios ilgis n gali būti tarnsformuojama į W -matmenų erdvę kaip vektorius $\bar{C} = \bar{c}_1, \dots, \bar{c}_w$. Vektoriaus \bar{C} i -tasis elementas apskaičiuojamas naudojantis formulę:

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \quad (1.2)$$

Pirmame transformacijos etape pasirinkus parametro w reikšmę nagrinėjama eilutė išskaidoma į w vienodo ilgio atkarpų. Sekančiame etape apskaičiuojama laiko eilutės stebėjimų patenkančių į atkarpos ribas vidurkis ir gauta reikšmė išsaugoma kaip transformacijos koeficientas \bar{c}_1 . Apskaičiavus vidurkius visoms atkarpoms gaunamas transformacijos vektorius. Transformacijos pavyzdys pateikiamas 1.7 pav. Lyginant šį metodą su kitais galime paminėti jo realizavimo ir interpretavimo paprastumą, bei mažą skaičiavimo resursų reikalingumą [5].



1.7 pav. PAA transformacijos pavyzdys

1.4. DARBE SPRENDŽIAMU UŽDAVINIAI

Remiantis atlikta apžvalga galime teigti, kad su pastatų pamatų nusėdimo problema susiduria dauguma jūrinių valstybių. Šios problemos žalai spręsti per paskutinius 10 metų vien Didžiojoje Britanijoje buvo išleista ~50 mln. svarų. Dėl to svarbu sudaryti su šia problema susijusių matavimų analizės modelių sudarymo metodiką, kuri remiantis atliktais ekspertiniais pastatų pamatų nusėdimo vertinimais gebėtų iš anksto informuoti apie rizikingą pastatui pamatų nusėdimą, kol dar nėra išoriškai pastebimų jo požymių.

Baigiamojo projekto metu naudojantis Nyderlandų Zaanstand regiono 2009-02-05 iki 2015-04-21 laikotarpio InSAR palydovų matavimais, ir žinomais 394 pastatų mokomosios imties ekspertiniais vertinimais siekiama sudaryti pastatų pamatų nusėdimo vertinimo modelį. Atsižvelgiant į tai, jog pastatams yra saugomas skirtingas matavimų skaičius darbo tikslo įgyvendinimui pirmiausia sprendžiami nusėdimo matavimų laiko eilučių apdorojimo ir jų klasifikavimo modelių sudarymo uždaviniai, o paskutiniame žingsnyje sprendžiamas pastatų pamatų nusėdimo vertinimo uždavinys.

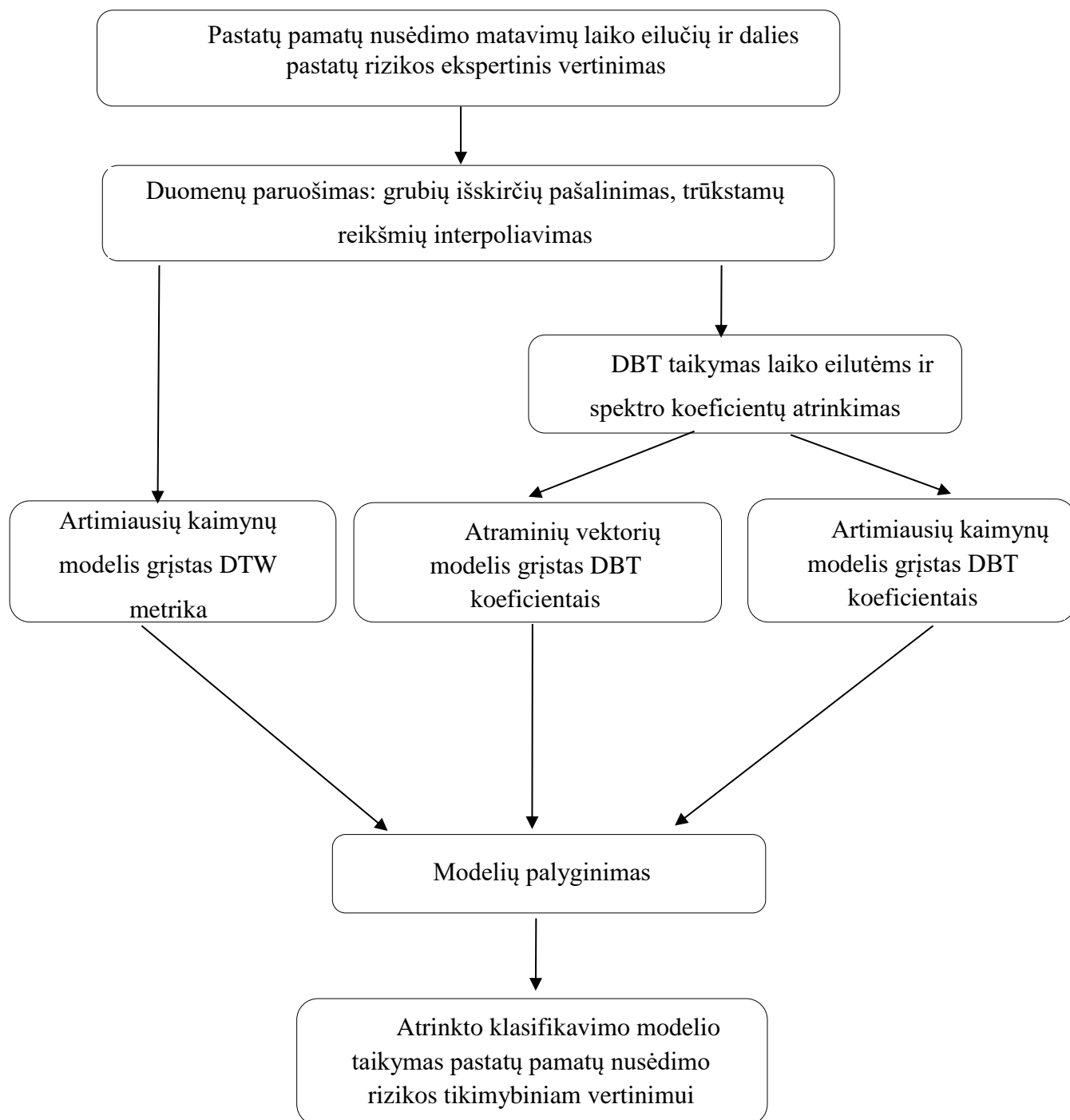
Darbe sprendžiami uždaviniai:

- Pasirinktų laiko eilučių klasifikavimo modelių sudarymo metodikos realizavimas R programos funkcijomis.
- Laiko eilučių transformacijų ar panašumo metrikų taikymas palydovų matavimų eilučių analizei.
- Klasifikavimui naudojamų laiko eilučių požymių (transformacijos koeficientų) atrinkimas.
- Klasifikavimo modelio sudarymas ir klasifikavimo tikslumo vertinimas.
- Skirtingų modelių palyginimas ir tinkamiausio modelio parinkimas pastatų pamatų nusėdimo matavimų laiko eilučių klasifikavimui.
- Pastatų nusėdimo rizikos vertinimo modelio realizavimas R funkcija ir taikymas naujų pastatų pamatų nusėdimo rizikos vertinimui

2. TIRIAMOJI DALIS IR REZULTATAI

2.1. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIS

Šiame skyriuje pateikiama pastatų pamatų nusėdimo vertinimo modelio sudarymo schema bei aprašoma modelio sudarymo etapuose naudojama metodika. Modelio sudarymo schema pavaizduota 2.1 paveiksle.



2.1 pav. Pastatų pamatų nusėdimo vertinimo modelio sudarymo schema

2.1.1. ANALIZUOJAMŲ DUOMENŲ PARUOŠIMAS

Prieš atliekant laiko eilučių transformavimą ir klasifikavimą buvo atliktas išskirčių šalinimas ir trūkstamų reikšmių interpoliavimas.

Išskirčių šalinimas.

Išskiriančių reikšmių pašalinimui ir pašalintų reikšmių interpoliavimui buvo panaudota Loess kreivių [21] sudarymo ir periodinio laiko eilutės skaidymo į trendo, sezoninės ir liekanų komponentės metodika. Stebėjimas pagal šią metodiką laikomas išskirtimi, jei liekanos patenka į intervalą $\pm 2(q_{0.9} - q_{0.1})$, kur q_p žymi p liekanų kvantilį.

Trūkstamų reikšmių interpoliavimas.

Trūkstamų reikšmių interpoliavimui bus naudojamas paprastojo slenkančio vidurkio metodas. Taikant šį metodą trūkstamų reikšmių interpoliavimui pasirinkus glodinimo lango ilgį 5 trūkstamos reikšmės buvo gautos naudojantis (2.1) formule, žinomos stebėjimų reikšmės nebuvo vidurkinamos.

$$c_l = \frac{1}{n} \sum_{i=1}^n c_{l-i+1} \quad (2.1)$$

čia n žymi glodinimo lango ilgį, c_l eilutės stebėjimą l - tuoju laiko momentu.

2.1.2. ARTIMIAUSIŲ KAIMYNŲ MODELIS GRĮSTAS DWT METRIKA

K -artimiausių kaimynų (angl. *k-Nearest Neighbours*) metodas yra klasifikavimo be mokytojo metodas. Prieš pradėdant klasifikavimą, metodui reikia pateikti apmokymo imtį, t.y. matavimų laiko eilutes, bei informaciją apie jų susikirstymą į grupes. Tada kiekvienos naujos laiko eilutės klasifikavimui reikia įvykdyti šiuos žingsnius:

- Naudojantis (DTW) atstumo metrika nagrinėjamai eilutei iš apmokymo imties parenkama k artimiausių eilučių.
- Atrinkus k artimiausių eilučių nustatoma, kokiai grupei jų priklauso daugiausiai, ir tai grupei yra priskiriama klasifikuojama eilutė.

Matematiškai šis algoritmas gali būti aprašomas kaip: naujos laiko eilutės X priskyrimas grupei c_i jei c_i turi didžiausią panašumo rodiklį tarp visų kategorijų. Panašumo rodiklis eilutei X grupei c_i apskaičiuojamas pagal formulę:

$$s(X, c_i) = \sum_{x_i \in k-NN} sim(X, X_i) y(X_i, c_i), \quad (2.2)$$

kur $sim(X, X_i)$ – DTW atstumas tarp eilutės X ir apmokymo eilutės X_i ;

$X_i \in k - NN$ – reiškia, kad eilutė X_i yra iš k artimiausių eilutės X kaimynų DTW atstumo metrikos atžvilgiu; $\gamma(X_i, c_j)$ – funkcija, įgyjanti reikšmę 1, kai eilutė X_i priklauso kategorijai c_j ir 0 kitu atveju.

DTW atstumo tarp eilutės $X = (x_1, \dots, x_n)$ ir apmokymo eilutės $X_i = (x_{i1}, \dots, x_{in})$ apskaičiavimui naudojama metodika paremta šiais žingsniais:

- Sudaroma $n \times n$ atstumų matrica, kurios (i, j) elementas yra Euklido atstumo tarp x_i ir x_{ij} kvadratas: $d(x_i, x_{ij}) = (x_i - x_{ij})^2$.
- DTW atstumas tarp eilučių X ir X_i randamas minimizuojant suminį iškraipymo kelio tarp eilučių taškų atstumą: $DTW(X, X_i) = \min \sqrt{\sum_{k=1}^K w_k}$. Čia w_k yra k -tasis iškraipymo kelio W elementas, kuris sudarytas iš atstumo matricos $(i, j)_k$ elementų. $\max(n, m) \leq K \leq m + n - 1$.
- Optimalaus iškraipymo kelio W sudarymo (atstumo tarp eilučių minimizavimo) uždavinys sprendžiamas naudojant dinaminę programavimą rekurentinei formulei apskaičiuoti: $\gamma(i, j) = d(x_i, x_{ij}) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}$, kur $\gamma(i, j)$ yra suminis atstumas tarp $d(x_i, x_{ij})$.

Apskaičiavus panašumo rodiklius eilutės priskyrimas grupei vykdomas pagal formulę:

$$\arg \max_{j=1, \dots, m} (s(X, c_j)), \quad (2.3)$$

kur c_1, \dots, c_m – iš anksto apibrėžtos grupės [6].

2.1.3. DISKREČIŲJŲ BANGELIŲ TRANSFORMACIJOS TAIKYMAS LAIKO EILUTĖMS IR SPEKTRO KOEFICIENTŲ ATRINIKIMAS

Siekiant apskaičiuoti nagrinėjamų laiko eilučių diskrečiųjų bangelių transformacijos spektro koeficientus pirmame žingsnyje eilutės papildomos nulniais matavimais taip, kad naujų eilučių matavimų skaičius būtų 2^n . Sekančiame žingsnyje parinkus Haaro mastelio funkciją:

$$\phi(x) = \begin{cases} 1, & \text{kai } 0 \leq x < 1 \\ 0, & \text{kitu atveju} \end{cases} \quad (2.4)$$

ir Haaro motininę bangelę:

$$\psi(x) = \begin{cases} 1, & \text{kai } 0 \leq x < 1/2 \\ -1, & \text{kai } 1/2 \leq x < 1 \\ 0, & \text{kitu atveju} \end{cases} \quad (2.5)$$

Gauname, jog diskrečiųjų bangelių Haaro transformacijos koeficientus nagrinėjamų nusėdimo matavimo eilučių transformavimui galime rekurentiškai apskaičiuoti naudojantis išraiškomis:

$$d_{1,i} = \frac{1}{\sqrt{2}} (s_{1-1,2i} - s_{1-1,2i+1}), \quad (2.6)$$

$$s_{l,i} = \frac{1}{\sqrt{2}}(s_{l-1,2i} + s_{l-1,2i+1}), \quad (2.7)$$

čia $l = 1, \dots, \log_2(n)$, $i = 1, \dots, \frac{n}{2^l}$ ir $s_{0,i} = x_i$.

Po baigtinio iteracijų skaičiaus gaunamas laiko eilutės transformacijos koeficientų rinkinys sudarytas iš suvidurkintos reikšmės $s_{\log(n),1}$ ir skirtuminių reikšmių rinkinio $\{d_{l,i}, \text{kai } l = 1, \dots, \log_2(n) - 1, o i = 1, \dots, \frac{n}{2^l}\}$ [4].

Siekiant sumažinti matavimų dimensiją, spektro koeficientų naudojamų klasifikavimui atrinkimui apskaičiuosime koeficientų Fišerio skiriamumo santykį (angl. *Fisher's discriminant ratio*).

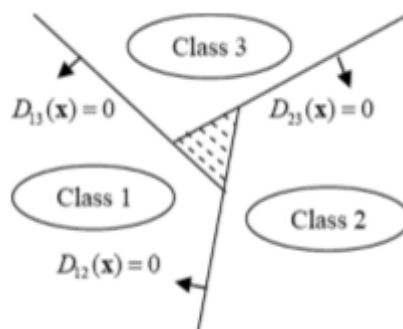
$$FDR = \sum_i^c \sum_{j \neq i}^c \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}, \quad (2.8)$$

kur c žymi klasių skaičių, μ_i – nagrinėjamo koeficiento vidurkį i -tosios klasės atžvilgiu, σ_i^2 – nagrinėjamo koeficiento dispersija i -tosios klasės atžvilgiu. Siekiant parinkti koeficientus, pagal kuriuos klasifikatorius galėtų lengvai atskirti nagrinėjamas grupes, rekomenduojama parinkti $k < n$ pirmų koeficientų išskiriančių aukštą Fišerio skiriamumo santykio reikšmę [20].

2.1.4. ATRAMINIŲ VEKTORIŲ MODELIO SUDARYMAS

Pagal sprendžiamo uždavinio specifiką žinome, kad šis uždavinys priskiriamas ne binariojo klasifikavimo uždaviniui. Siekiant tokio uždavinio sprendimui sudaryti atraminių vektorių klasifikatorių, jo apmokymui dažniausiai taikomos „vienas-prieš-vieną“ (angl. *one-against-one*) arba „vienas-prieš-visus“ (angl. *one-against-all*) metodikos. Darbe dėl programinio realizavimo paprastumo pasirinktą taikyti „vienas-prieš-vieną“ metodika.

Taikant šią metodiką darome prielaidą, kad $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ yra apmokymo aibė, kur $x_i \in R^m$ ir $y_i \in \{1, 2, \dots, c\}$. Remiantis „vienas-prieš-vieną“ metodika, sprendžiant c klasių problemą reikia sudaryti $c(c-1)/2$ klasifikatorių (trijų klasių klasifikatoriaus pavyzdys 2.2 pav.)



2.2 pav. „Vienas-prieš-vieną“ metodo gautas 3 klasių klasifikatorius

Optimali klasę i nuo klasės j skirianti hiperplokštuma aprašoma lygtimi:

$$D_{ij}(x) = w_{ij}^T \phi(x) + b_{ij} = 0, \quad (2.9)$$

kur $i < j, 1 < j \leq c, 1 \leq i < c, b_{ij}$ skaliaras ir $\phi(x)$ atvaizdžio funkcija.

Optimalios hiperplokštumos krypties parinkimas apibrėžiamas lygtimi:

$$D_{ij}(x) = -D_{ji}(x) \quad (2.10)$$

Sudaryti optimalios hiperplokštumos lygtį sprendžiamas minimizavimo uždavinys:

$$\min \frac{1}{2} w_{ij}^T w_{ij} + C \frac{1}{2} \sum_{i=1}^l \xi_i, \quad (2.11)$$

kur C apibrėžia kompromisą tarp maksimalios skiriančios ribos (paraštės) ir minimalios klasifikavimo klaidos, o ξ_i paraštės dydį apibrėžiantis parametras.

Naudojantis Lagrandžo daugianarių metodu, bei įvedus branduolio funkcijos sąvoką $K(x, x_i) = \phi(x)^T \cdot \phi(x_i)$. po baigtinio žingsnių skaičiaus gauname, jog optimali hiperplokštumos lygtis gali būti randama naudojantis išraiška:

$$f(x) = \text{sgn}\left\{\sum_{i=1}^l \alpha_i y_i \cdot K(x, x_i) + b\right\}, \quad (2.12)$$

kur α_i žymi Lagrandžo daugianarius. Praktikoje dažniausiai naudojamos branduolio funkcijos:

- tiesinė: $K(x, x_i) = x * x_i$.
- polinominė: $K(x, x_i) = (\gamma * x * x_i + \text{const})^p$.
- radialinių bazių: $K(x, x_i) = e^{-\gamma * |x - x_i|^2}$

Apskaičiavus optimalių hiperplokštumų lygtis, naujo vektoriaus x klasifikavimui, apskaičiuojame:

$$D_i(x) = \sum_{j \neq i, j=1}^c \text{sgn}(D_{ij}(x)) \quad (2.13)$$

visoms $i = 1, \dots, c$ ir priskiriame vektorių x klasei i su kuria gaunama maksimali D_i reikšmė [16].

2.1.5. KLASIFIKAVIMO MODELIO VERTINIMAS

Siekiant įvertinti sudaryto klasifikavimo modelio tinkamumą nagrinėjamiems duomenims pirmiausiai sudaromos sumaišymo matricos (angl. *Confusion matrix*). Sumaišymo matrica parodo teisingai ir neteisingai suklasifikuotų objektų skaičių. Dažnai pasitaiko, kad modelis gali būti labai geras vienu objektų klasifikatorius ir tuo pačiu visai neišskiria kitų klasių. Sumaišymo matrica trijų klasių atveju susideda iš 9 celių (žr. 2.1 lentelė) [15].

2.1 lentelė

Sumaišymo matrica

		Prognozuojamos klasės		
		A	B	C
Tikrosios klasės	A	TP_A	e_{AB}	e_{AC}
	B	e_{BA}	TP_B	e_{BC}
	C	e_{CA}	e_{CB}	TP_C

Sumaišymo matricoje:

- TP_A – klasei A priskirtų jos objektų skaičius;
- TP_B – klasei B priskirtų jos objektų skaičius;
- TP_C – klasei C priskirtų jos objektų skaičius;
- e_{AB} – klasei B priskirtų klasės A objektų skaičius;
- e_{AC} – klasei C priskirtų klasės A objektų skaičius;
- e_{BA} – klasei A priskirtų klasės B objektų skaičius;
- e_{BC} – klasei C priskirtų klasės B objektų skaičius;
- e_{CA} – klasei A priskirtų klasės C objektų skaičius;
- e_{CB} – klasei B priskirtų klasės C objektų skaičius;

Kadangi sprendžiamame uždavinyje stebėjimų skaičius klasėse nėra pasiskirstęs tolygiai, klasifikavimo modelio įvertinimui pasinaudojanti sudarytosios sumaišymo matricos reikšmėmis apskaičiuosime 2.2 lentelėje pateiktas metrikas. Metrikos pateikiamos klasės A klasifikavimo vertinimui, tačiau atitinkamos formulės taip naudojamos B ir C klasių klasifikavimo vertinimui.

2.2 lentelė

Klasifikavimo kokybės vertinimo metrikos

Metrikos pavadinimas	Formulė	Aprašymas
Tikslumas (angl. <i>precision</i>)	$\frac{TP_A}{TP_A + e_{BA} + e_{CA}}$	Klasei A priskirtų objektų skaičiaus, kurie iš tiesų yra A klasės objektai santykis su suminiu A klasei priskirtų objektų skaičiumi.
Jautrumas (angl. <i>recall</i>)	$\frac{TP_A}{TP_A + e_{AB} + e_{AC}}$	A klasės objektų dalis, kuri buvo teisingai su klasifikuota
F1	$2 * precision * recall / (precision + recall)$	Harmoninis arba svertinis tikslumo ir jautrumo vidurkis

Apibendrintą sudarytojo klasifikavimo modelio F1 metrikos vertę gaunama apskaičius aritmetinį F1 metrikų gautų nagrinėjamos klasės vidurkį. Siekiant nagrinėjamiems duomenims parinkti kuo tinkamesnį klasifikavimo modelį metrikos F1 reikšmė turi būti maksimizuojama ($F1_{\max} = 1$).

2.2. PROGRAMINĖ MODELIO SUDARYMO REALIZACIJA

Pasirinktos modelio sudarymo metodikos realizavimui buvo panaudota atviro kodo R programinė įranga. Šį pasirinkimą nulėmė R realizuotų diskrečių bangelių transformacijos koeficientų apskaičiavimo ir klasifikavimo modelių sudarymo algoritmų gausa.

Modelio sudarymo žingsnių realizavimui buvo sukurtos šios funkcijos:

- *DwtTransform*, *FDR* – DBT spektro apskaičiavimui ir koeficientų atrinkimui.
- *euclideanDist*, *knn_predict2* – Artimiausių kaimynų modelio grįsto DBT koeficientais sudarymui.
- *DtwDist*, *knn_predict* – artimiausių kaimynų modelio grįsto DTW atstumo metrika sudarymui.
- *metrixConfusion* – klasifikavimo vertinimo metrikų apskaičiavimui.

Atraminių vektorių modelio grįsto DBT spektro koeficientais sudarymui buvo panaudota R paketo *e1071* funkcija *svm* ir paketo *stat* funkcija *predict*.

Paminėtų funkcijų bei duomenų apdorojimo žingsnių programinis kodas pateiktas 2 priede.

2.3. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIO SUDARYMAS IR TAIKYMAS

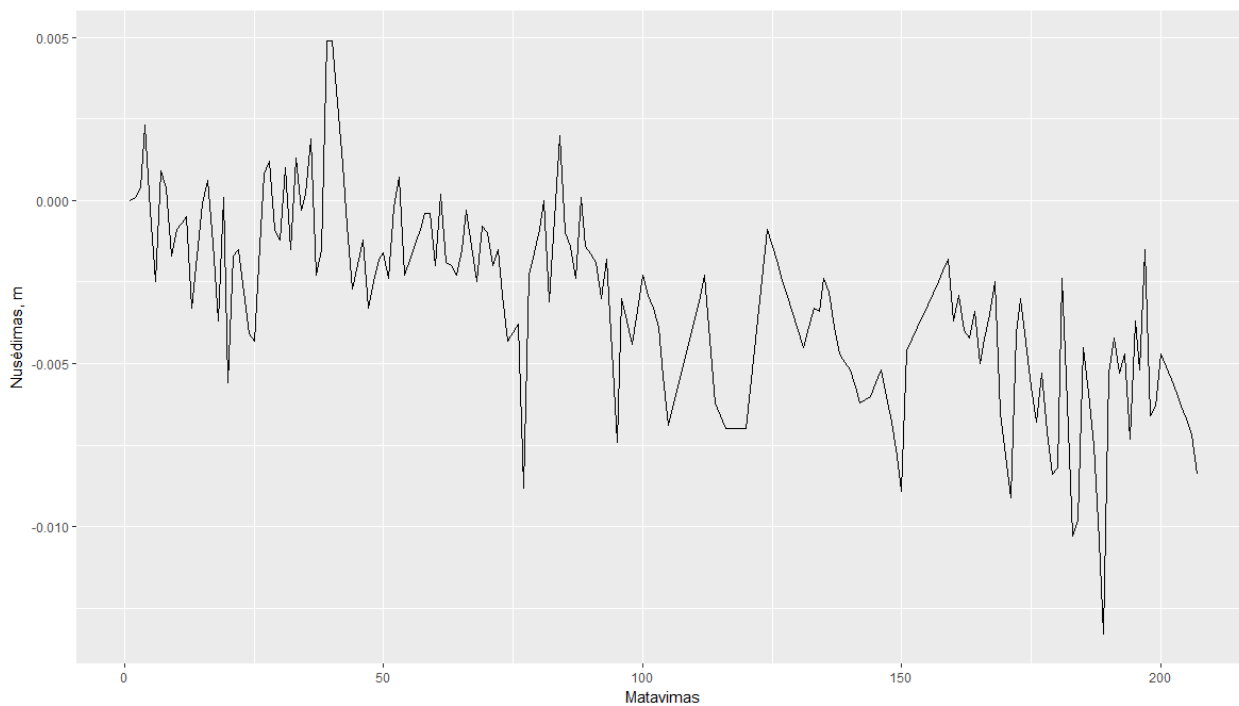
2.3.1. ANALIZUOJAMI DUOMENYS

Analizuojamos pastatų pamatų nusėdimo matavimų laiko eilutės gautos atlikus matavimus, panaudojant palydovus. Matavimai atlikti laikotarpyje nuo 2009-02-09 iki 2015-04-21 (iš viso 207 matavimai). Mokymo imtyje vieną pastatą apibūdinčių laiko eilučių skaičius yra nuo 1 iki 17. Jų skaičius priklauso nuo palydovo signalą reguliariai atspindinčių taškų. Mokymo imtyje 100 pastatų buvo matuojamas vienas taškas, 291 pastatams matuojamų taškų skaičius yra nuo 2 iki 13, dviem pastatams - 15 taškų ir vienam - pastatui 17 taškų. Aptarus problema su ekspertais buvo sutarta, kad prielaida, jog ekspertų įvertintų pastatų rizikos grupės vertinimas pastato taškams bus vienodas, yra teisinga. Mokymo imtis sudaryta iš 1305 matavimų taškų.

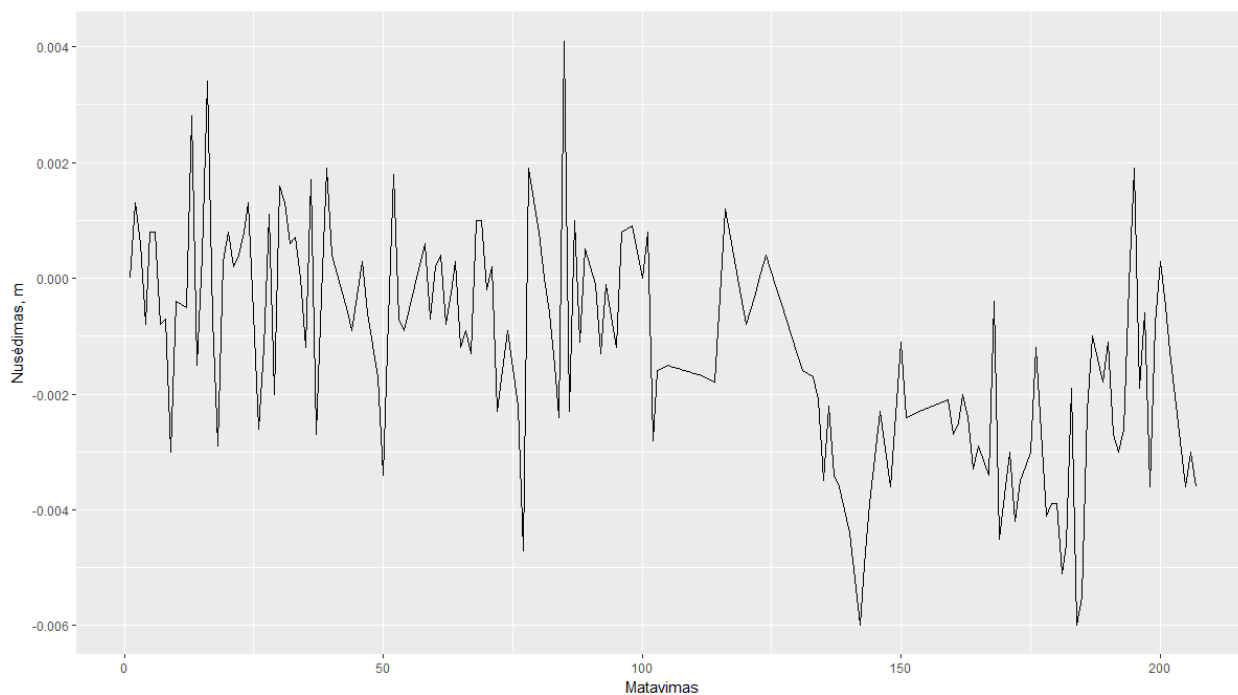
Modelio mokymui naudojama ekspertų į tris pastatų pamatų nusėdimo rizikos grupes (mažiausia, vidutinė, aukščiausia) suskirstyta matavimų imtis.

- Žemiausios rizikos grupei priskirti 48 pastatai (174 taškai).
- Vidutinės rizikos grupei priskirti 332 pastatai (1090 taškai).
- Aukščiausios rizikos grupei priskirti 14 pastatų (41 taškai).

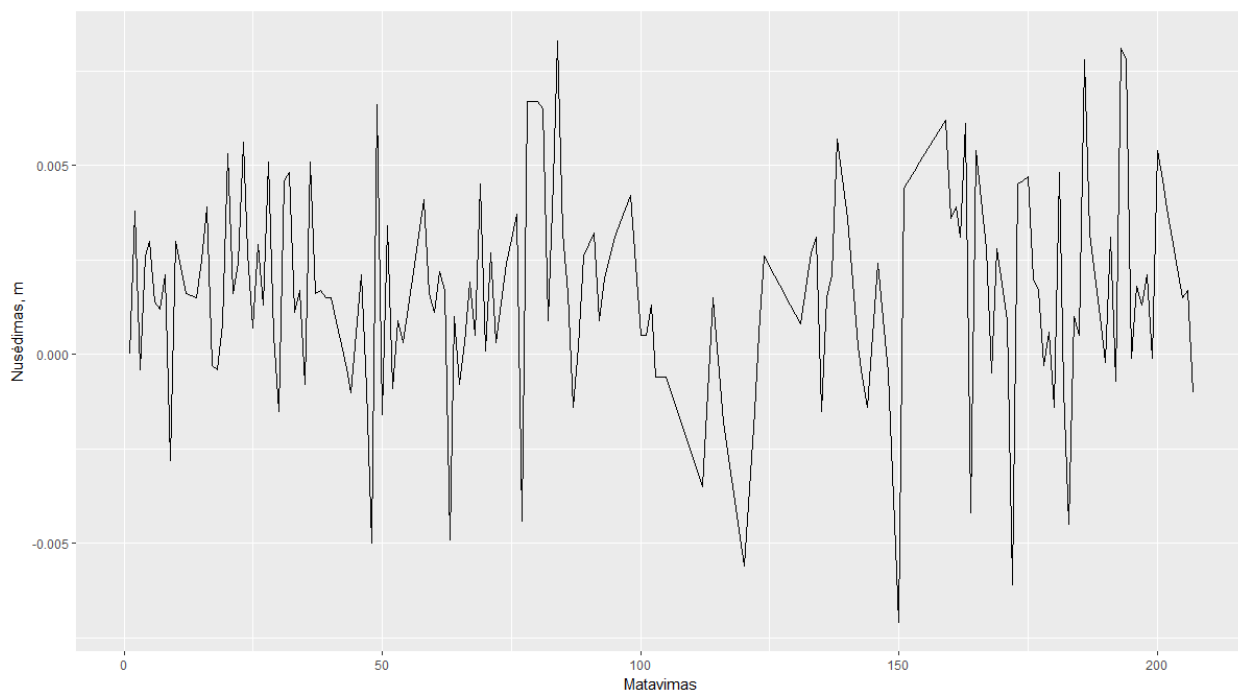
Atlikus duomenų tiriamąją analizę buvo pašalintos grubios išskirtys ir atliktas trūkstamų reikšmių interpoliavimas panaudojant slenkančio vidurkio metodą. Gautų matavimų laiko eilučių pavyzdžiai kiekvienai rizikos grupei pateikti 2.3-2.5 pav.



2.3 pav. Aukščiausios rizikos grupės pastatų pamatų nusėdimo matavimų laiko eilutės pavyzdys



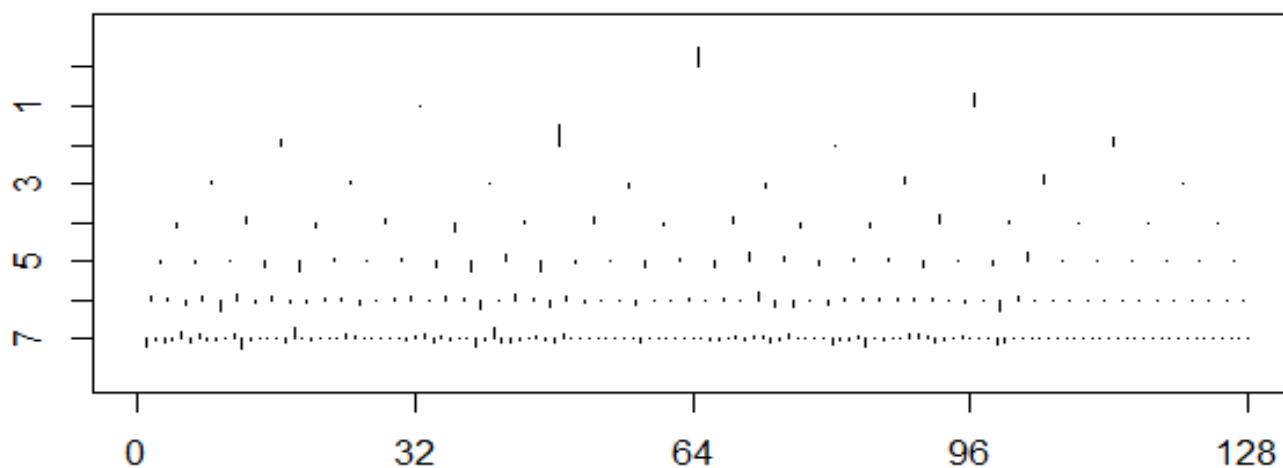
2.4 pav. Vidutinės rizikos grupės pastatų pamatų nusėdimo matavimų laiko eilutės pavyzdys



2.5 pav. Žemiausios rizikos grupės pastatų pamatų nusėdimo matavimų laiko eilutės pavyzdys

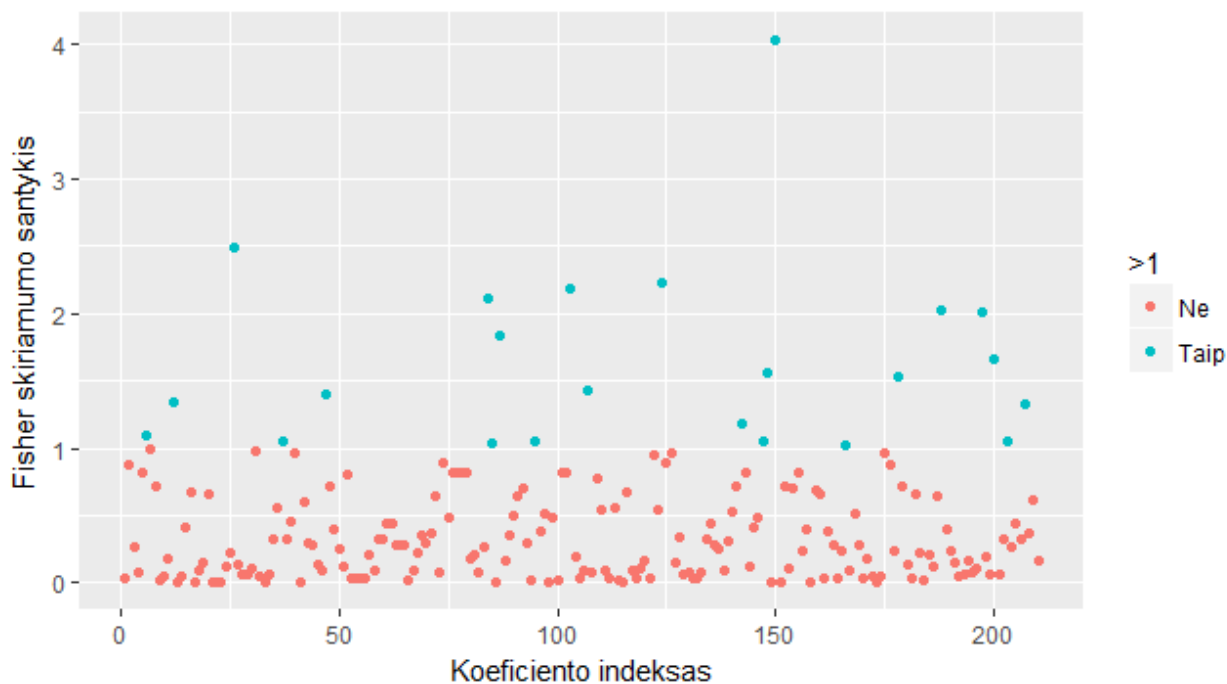
2.3.2. DBT TAIKYMAS LAIKO EILUTĖMS IR KOEFICIENTŲ ATRINKIMAS

Pagal 2.1.3 poskyryje pateiktą DBT koeficientų skaičiavimo metodiką analizuojamas matavimų laiko eilutės (stebėjimų skaičius lygūs 207), papildėme 49 nulinėmis reikšmėmis ir gavome laiko eilutes, kurių matavimų skaičius $2^8=256$. Gautas eilutes išskaidėme į 8 lygių spektrą, kurio pirmuose 7 lygiuose saugomos 255 skirtuminės reikšmės ir paskutiniame 0 lygyje suvidurkinta reikšmė.



2.6 pav. Pamatų nusėdimo matavimų laiko eilutės skaidymo DBT spektro koeficientais pavyzdys

Tinkamiausių klasifikavimui koeficientų parinkimui apskaičiavome kiekvieno iš jų Fišerio skiriamumo santykį (angl. *Fisher's discriminant ratio*). Gavome, kad daugumos koeficientų Fišerio skiriamumo santykis yra mažesnis už vienetą. Siekiant parinkti tuos koeficientus, kurie leistų geriau atskirti nagrinėjamų pamatų nusėdimo rizikos grupes, pasirinkome koeficientų rinkinį sudarytą iš 23 koeficientų, kuriems gautasis Fišerio skiriamumo santykis yra didesnis už 1 (žr. 2.7 pav.).



2.7 pav. Fisher skiriamumo santykio DBT koeficientams grafikas

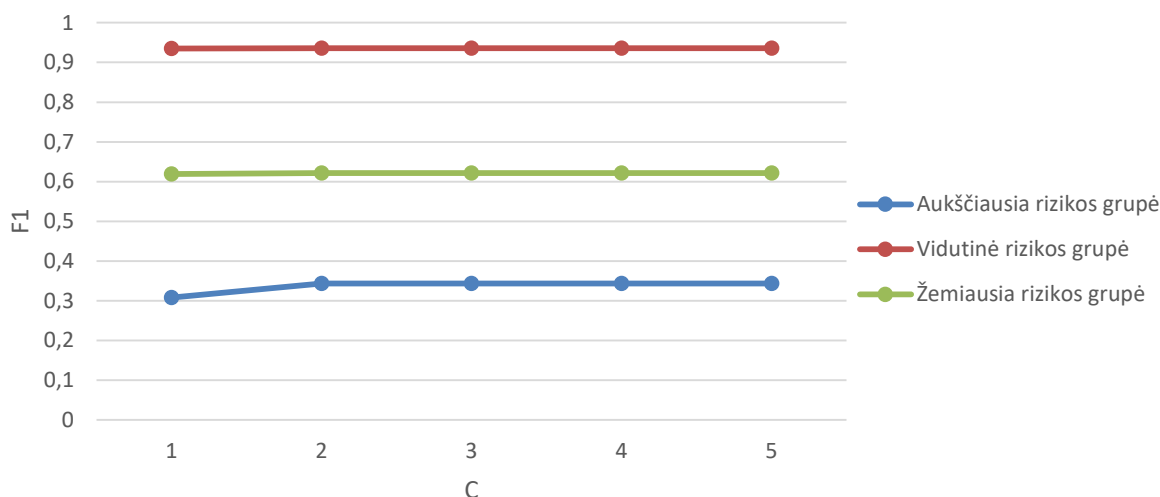
Sekančiame etape naudojant atrinktus DBT koeficientus sudarysime atraminių vektorių ir artimiausių kaimynų klasifikavimo modelius.

2.3.3. ATRAMINIŲ VEKTORIŲ MODELIS GRĮSTAS DBT KOEFICIENTAIS

Remiantis atraminių vektorių klasifikatoriaus specifika, sudarant modelį svarbu tinkamai parinkti branduolio funkcija, optimizavimo uždavinio parametą C , kuris apibrėžia kompromisą tarp maksimalios ribos ir minimalios klasifikavimo klaidos, bei pasirinktosios branduolio funkcijos parametų rinkinį.

Modelio sudarymui gautąją imtį išskaidėme į mokymo 0,8 % (1044 taškai) ir 0,2% (261 taškai) testinę imtis. Mokymo imtį išskaidėme į 5 dalis, kiekvienoje dalyje buvo siekiama išsaugoti vienodas klasių proporcijas. Naudojantis kryžminio patikrinimo metodika, keičiant modelio parametų rinkinius kiekvienam iš jų atlikti 5 modelio apmokymai bei klasifikavimo metrių vertinimai. Kiekvienoje mokymo iteracijoje buvo keičiama iš mokymo imties išskiriama testinė imtis, o likusios 4 dalys naudojamos modelio mokymui.

Pirmiausia sudarėme atraminių vektorių klasifikavimo modelį su tiesinę branduolio funkcija. Kadangi ji neturi savo vidinių parametru, sudarėme priklausomybių tarp parametro C ir vidutinės $F1$ klasifikavimo metrikos nagrinėjamos rizikos grupėms grafikus. Iš 2.8 pav. grafiko matome, kad vidutinės ir žemiausios rizikos grupių atžvilgiu gauta $F1$ metrikos vertė nepriklauso nuo parametro C (žemiausios rizikos grupės $F1 = 0,62$, o vidutinio rizikos grupės $F1 = 0,93$). Aukščiausios rizikos grupės atžvilgiu $F1$ metrikos vertė $0,04$ vieneto padidėja keičiant parametru C iš 1 į 2 ir pasiekia didžiausią reikšmę $0,34$, kai $C \geq 2$. $F1$ metrikos vertė parametru C kintant intervale $[1;5]$, kinta nuo $0,62$ iki $0,633$.



2.8 pav. F1 metrikos priklausomybė nuo atraminių vektorių klasifikatoriaus parametro C

Siekiant parinkti tinkamus nagrinėjamo modelio parametrus sudarėme modelį su radialinių bazių branduolio funkcija. Atraminių vektorių modelio mokymui parinkus šią branduolio funkciją sudarėme $F1$ metrikos reikšmių lenteles keičiant parametru γ ir C reikšmes.

2.3 lentelė

F1 metrikos reikšmės prie skirtingų parametru γ ir C reikšmių (aukščiausios rizikos grupė)

$\gamma \setminus C$	5	10	15	20
0,01	0,47	0,58	0,55	0,60
0,02	0,45	0,61	0,63	0,58
0,03	0,59	0,58	0,54	0,54
0,04	0,54	0,54	0,54	0,54
0,05	0,54	0,54	0,54	0,54

Išanalizavus 2.3 lentelėje pateiktus rezultatus pastebėta, kad aukščiausios rizikos grupės klasifikavimo atžvilgiu $F1$ metrikos reikšmė neviršija $0,63$ ir parinkus parametru $\gamma \geq 0,04$, $F1$ metrikos reikšmė nepriklauso nuo parametro C reikšmės.

2.4 lentelė

F1 metrikos reikšmės prie skirtingų parametru γ ir C reikšmių (vidutinės rizikos grupė)

$\gamma \backslash C$	5	10	15	20
0,01	0,94	0,95	0,95	0,95
0,02	0,95	0,95	0,96	0,95
0,03	0,95	0,96	0,95	0,95
0,04	0,96	0,95	0,96	0,96
0,05	0,96	0,96	0,96	0,96

Vidutinio rizikos grupės klasifikavimo atžvilgiu gautus F1 metrikos reikšmės kinta intervale [0,94; 0,96]. Kaip ir aukščiausios rizikos grupės atžvilgiu parinkus parametru $\gamma \geq 0,04$ gavome, jog F1 metrikos reikšmės šios grupės atžvilgiu nepriklauso nuo parametro C .

2.5 lentelė

F1 metrikos reikšmės prie skirtingų parametru γ ir C reikšmių (žemiausios rizikos grupė)

$\gamma \backslash C$	5	10	15	20
0,01	0,69	0,59	0,74	0,74
0,02	0,75	0,75	0,77	0,78
0,03	0,76	0,79	0,79	0,80
0,04	0,77	0,79	0,80	0,80
0,05	0,82	0,80	0,80	0,80

2.5 lentelėje pateikti rezultatai rodo, kad žemiausios rizikos grupės klasifikavimo atžvilgiu maksimali F1 metrikos reikšmė 0,82 gauta parinkus parametru rinkinį $(\gamma; C) = (0,05; 5)$. Su kitais parametru rinkiniais F1 metrikos reikšmės kinta intervale [0,69;0,8].

2.6 lentelė

Apibendrinta F1 metrikos reikšmės prie skirtingų parametru γ ir C reikšmių

$\gamma \backslash C$	5	10	15	20
0,01	0,70	0,71	0,74	0,76
0,02	0,72	0,77	0,79	0,77
0,03	0,77	0,77	0,76	0,77
0,04	0,76	0,76	0,77	0,77
0,05	0,77	0,77	0,77	0,77

Apskaičiavus apibendrintas F1 metrikos reikšmes sudarytiems modeliams gavome, jog maksimali F1 metrikos reikšmė 0,79, gauta modeliu su parametru rinkiniu $(\gamma, C) = (0,02; 15)$. Atžvelgiant į tai jog aukščiausios ir vidutinės rizikos grupių maksimali F1 metrikos vertė taip pat įgyjama su šiuo parametru rinkiniu, parenkame jį kaip optimalų.

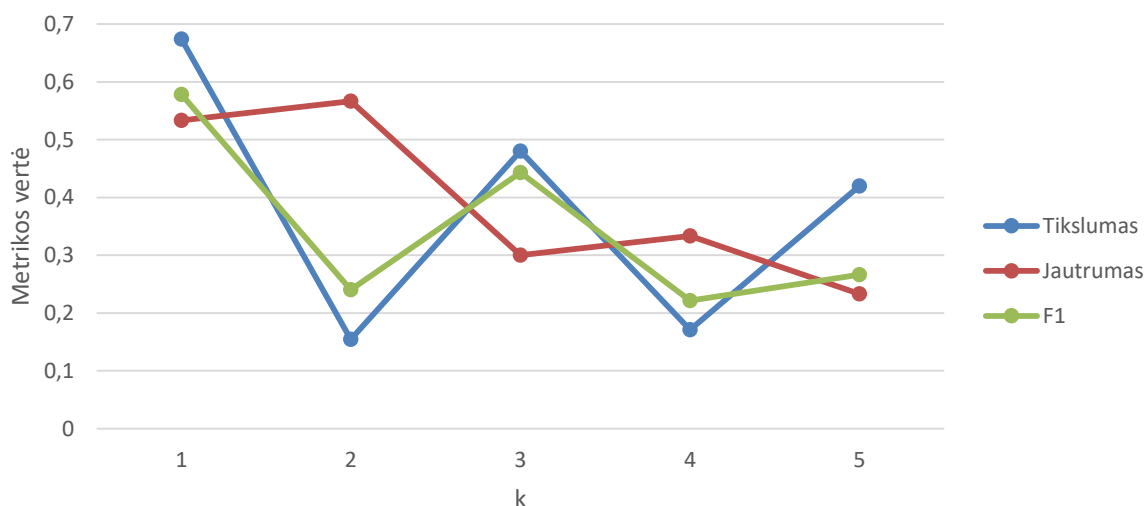
Palyginus rezultatus gautus naudojant atraminių vektorių klasifikatorių su tiesinio branduolio ir radialinių bazių branduolio funkcijomis galime teigti, kad pastatų nusėdimo matavimų eilučių klasifikavimui geriausiai tinka radialinių bazių branduolio funkcija su parametrais $(\gamma, C) = (0,02; 15)$, nes apibendrinta F1 metrika 0,15 vieneto didesnė, nei naudojant tiesinio branduolio funkcija.

2.3.4. ARTIMIAUSIŲ KAIMYNŲ MODELIS GRĮSTAS DBT KOEFICIENTAIS

Siekiant atrinkti geriausią klasifikavimo modelį matavimų laiko eilučių klasifikavimui, taip pat sudarėme artimiausio kaimynų klasifikatorių grįstą 2.3.2 skyriuje atrinktais DBT koeficientais.

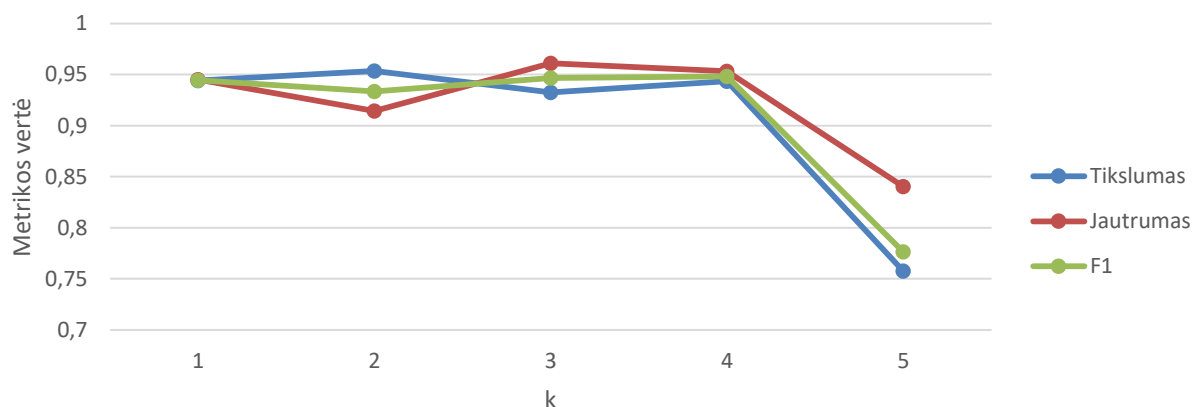
Modelio mokymui taikėme 2.1.2. skyriuje pateiktą artimiausių kaimynų modelio sudarymo metodiką, tik šiuo atveju atstumui tarp skirtingų eilučių aprašomu DBT spektro koeficientu rinkiniu, buvo naudojama Euklido atstumo metrika. Rizikos lygio parinkimo metodika sutampa su 2.1.2 skyriuje pateikta metodika. Modelio sudarymui gautąją imtį išskaidėme į mokymo 0.8 % (1044 taškai) ir 0.2% (261 taškai) testinę imtis. Mokymo imtį išskaidėme į 5 dalis, kiekvienoje dalyje buvo siekiama išsaugoti vienodas klasių proporcijas. Naudojantis kryžminio patikrinimo metodika buvo keičiamos artimiausio kaimyno metodo parametro k reikšmės nuo 1 iki 5 ir kiekvienai iš jų atlikti 5 modelio mokymai bei klasifikavimo metrikų vertinimai. Kiekvienoje mokymo iteracijoje buvo keičiama testinė imtis, o likusios 4 dalys naudojamos modelio mokymui.

Artimiausių kaimynų modelio grįsto DBT koeficientais klasifikavimo kokybės metrikų priklausomybės nuo parametro k pateiktos 2.9-2.11 pav.



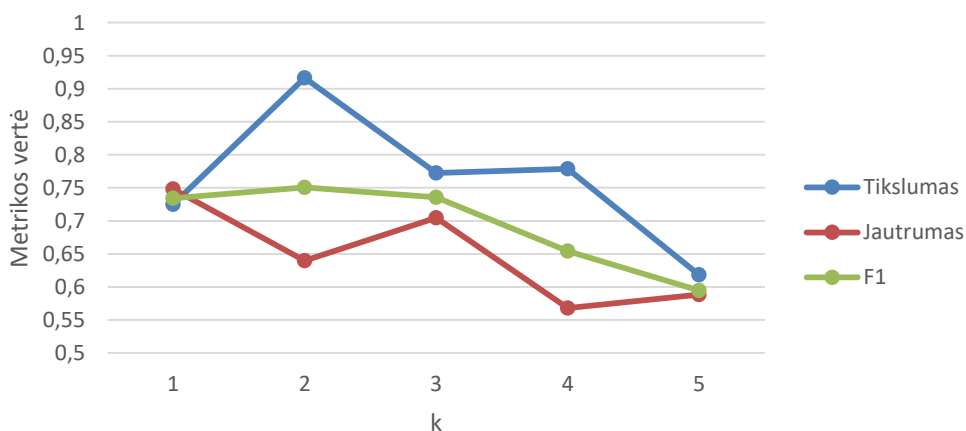
2.9 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (aukščiausios rizikos grupė)

2.9 pav. pateiktas grafikas rodo, kad parinkus artimiausių kaimynų skaičių lygų 1 buvo gautos aukščiausios harmoninis tikslumo ir jautrumo vidurkis $F1 = 0,58$. Parinkus parametro $k = 1$ gautas artimiausio kaimynų klasifikavimo modelis, vidutiniškai 53% aukščiausios rizikos grupės stebėjimų suklasifikavo teisingai. O vidutinis santykis, tarp teisingai suklasifikuotų aukščiausios rizikos grupės taškų ir neteisingai aukščiausios rizikos grupei priskirtų taškų skaičiaus, kai $k = 1$, yra 0,67. Didėjant parametro k vertėm harmoninis tikslumo ir jautrumo vidurkis kinta intervale $[0,22;0,45]$.



2.10 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (vidutinės rizikos grupė)

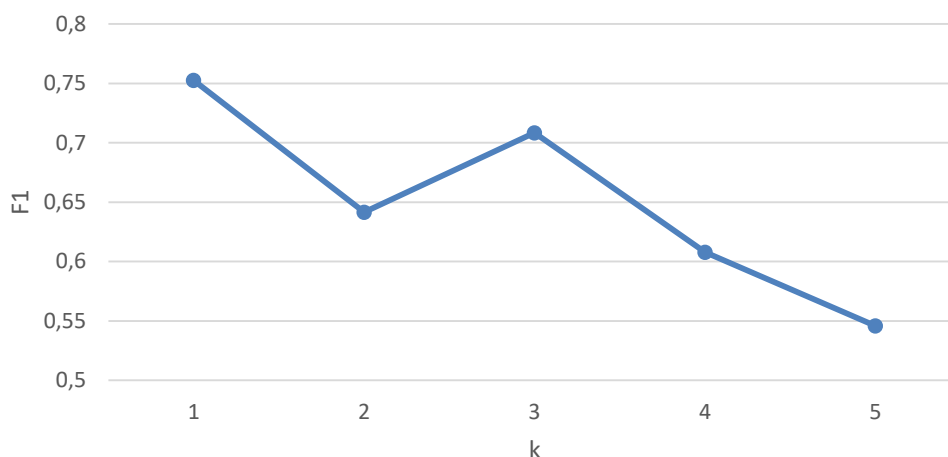
Analizuojat 2.10 pav. pateiktas vidutinės rizikos grupės klasifikavimo kokybės metrikų priklausomybes nuo parametro k galime pastebėti, kad maksimali vidutinė tikslumo metrikos reikšmė 0,95 šioje grupėje gauta parinkus parametą $k = 2$, tuo tarpu didžiausia jautrumo reikšmė 0,96 gauta parinkus $k = 3$. Didžiausias harmoninis klasifikavimo tikslumo ir jautrumo vidurkis ($F1 = 0,948$) vidutinės rizikos grupės atžvilgiu pasiektas klasifikatoriu gautu parinkus parametą $k = 4$.



2.11 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (žemiausios rizikos grupė)

Apibendrinant gautas žemiausios rizikos grupės klasifikavimo kokybės metrikų reikšmes galime pastebėti, kad parinkus parametą $k = 2$ gauname, kad vidutinis santykis, tarp teisingai suklasifikuotu žemiausios rizikos grupės taškų ir neteisingai klasifikatoriaus žemiausios rizikos grupei priskirtų taškų skaičiaus yra 0,92. Jautrumo metrikos reikšmė, kai $k = 2$ parodo, kad vidutiniškai 64% šios grupės stebėjimų buvo suklasifikuoti teisingai. Aukščiausias harmoninis klasifikavimo tikslumo ir jautrumo vidurkis ($F1 = 0,75$) šios grupės atžvilgiu taip pat pasiektas klasifikatoriui parinkus pametrą $k = 2$.

Dėl pastebėtų skirtingų F1 metrikos priklausomybių analizuojamų grupių atžvilgiu ir siekiant parinkti tinkamiausią artimiausių kaimynų modelio parametro k vertę, sudarytą apibendrintos F1 metrikos priklausomybė nuo parametro k . Remiantis gautais rezultatais žr. 2.12 pav., galime išskirti, jog maksimali apibendrinta F1 metrikos reikšmė 0,75 gauta parinkus parametą $k = 1$. Toliau dėl 2.12 pav. pateiktų rezultatų, bei dėl 2.9 pav. išskirtos didžiausios F1 vertės, kai $k = 1$, optimaliu artimiausio kaimynų modelio grįsto DBT koeficientais parametru k laikomas 1.



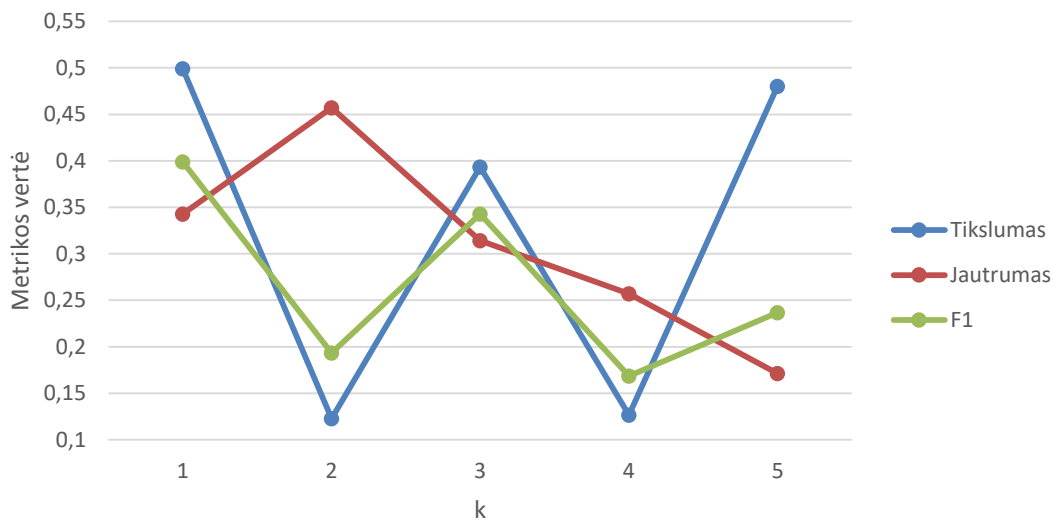
2.12 pav. F1 metrikos priklausomybė nuo parametro k

2.3.5. ARTIMIAUSIŲ KAIMYŲ MODELIS GRĮSTAS DWT METRIKA

Pirmame modelio sudarymo žingsnyje dėl DTW metodo specifikos nagrinėjamos eilutės buvo standartizuojamos (atėmus vidurkį ir padalinus iš standartinio nuokrypio). Modelio sudarymui gautąją imtį išskaidėme į mokymo 0.8 % (1044 taškai) ir 0.2% (261 taškai) testinę imtis. Mokymo imtį išskaidėme į 5 dalis, kiekvienoje dalyje buvo siekiama išsaugoti vienodas klasių proporcijas. Naudojantis kryžminio patikrinimo metodika buvo keičiamos artimiausio kaimyno metodo parametro k reikšmes nuo 1 iki 5 ir kiekvienai iš jų atlikti 5 modelio mokymai bei klasifikavimo metrikų vertinimai. Kiekvienoje mokymo iteracijoje buvo keičiama testinė imtis, o likusios 4 dalys naudojamos modelio mokymui.

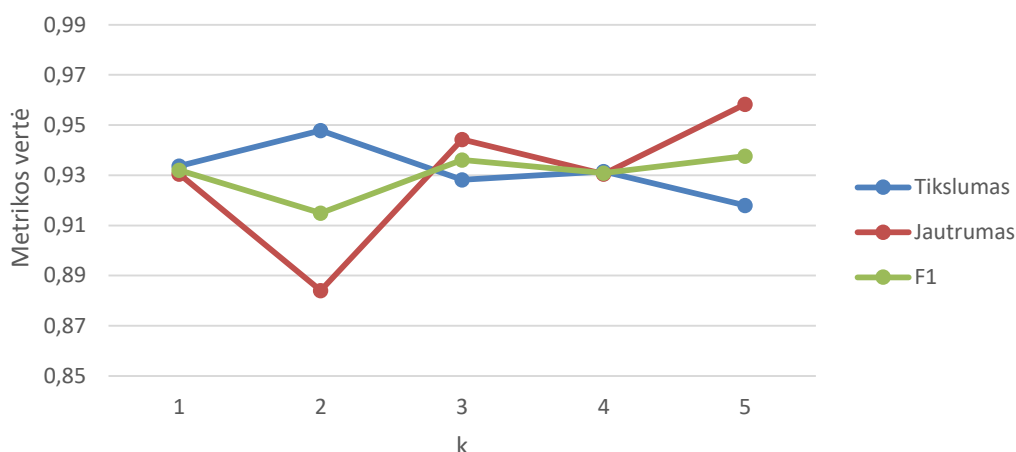
Remiantis artimiausio kaimyno sudarymo metodika kiekvienoje iteracijoje nagrinėjamos testinės imties eilutėms naudojantis DTW atstumo metrika buvo apskaičiuojamas atstumas iki apmokymo imtyje esančių eilučių, pagal pasirinktą modelio parametro k reikšmę parenkama k artimiausių kaimynų ir paskutiniame etape eilutė priskiriama tai grupei, kuriai priskirta dauguma artimiausių kaimynų.

Panaudojus aprašytąją kryžminio patikrinimo metodika apskaičiuotos vidutinės klasifikavimo vertinimo metrikos kiekvienai iš klasių. Metrikų priklausomybių nuo modelio parametro k grafikai pateikti 2.13 - 2.15 pav.



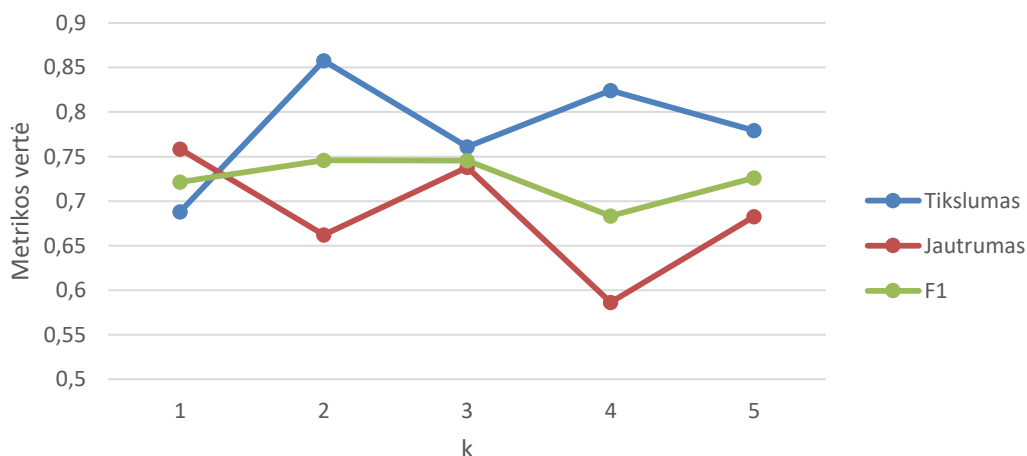
2.13 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (aukščiausios rizikos grupė)

2.13 pav. pateikti rezultatai rodo, kad parinkus parametą $k = 1$, vidutinis santykis tarp teisingai klasifikuotu aukščiausios rizikos grupės taškų ir neteisingai klasifikatoriaus aukščiausios rizikos grupei priskirtų taškų skaičiaus yra 0,5. O klasifikatoriaus jautrumas, kai $k = 1$ aukščiausios rizikos lygio atžvilgiu yra 0,34, t.y. vidutiniškai 34 % aukščiausios rizikos grupės stebėjimų buvo suklasifikuota teisingai. Apskaičiavus harmoninį jautrumo ir tikslumo vidurkį nagrinėjamos rizikos grupės atžvilgiu galime išskirti, jog maksimali F1 vertė pasiekta, kai $k = 1$.



2.14 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (vidutinės rizikos grupė)

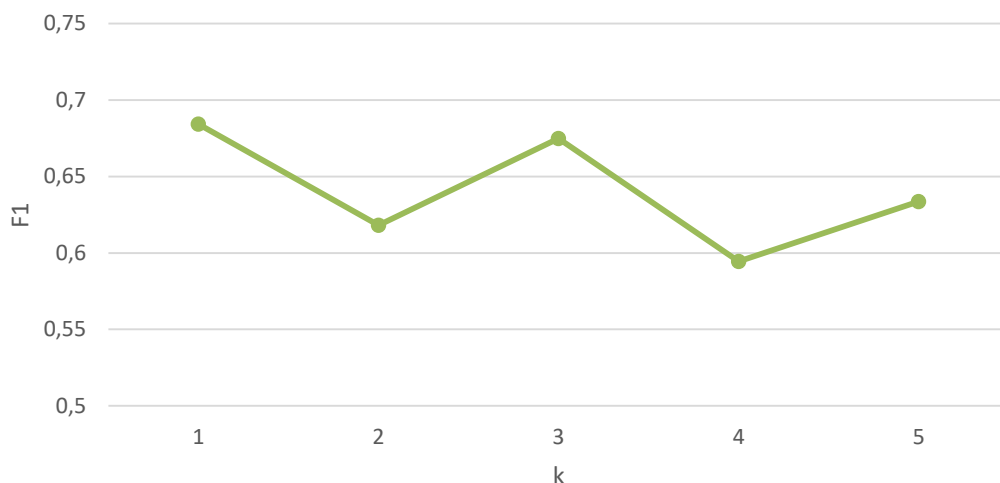
Nagrinėjant vidutinės rizikos grupei priskirtų taškų klasifikavimo metrikų priklausomybę nuo nagrinėjamų artimiausių kaimynų skaičiaus k galime pastebėti, jog maksimali F1 metrikos reikšmė 0,937 gauta parinkus parametą $k = 5$. Jautrumas šiuo atveju lygus 0,96, t.y. vidutiniškai 96% vidutinės rizikos grupės stebėjimų buvo klasifikuoti kaip vidutinės rizikos grupės stebėjimai. O gautas tikslumas vidutinės rizikos grupės atžvilgiu, kai $k = 5$ lygus 0,92, t.y. 0.27 mažesnis už maksimalų tikslumą.



2.15 pav. Klasifikavimo kokybės metrikų priklausomybės nuo parametro k (žemiausios rizikos grupė)

Žemiausios rizikos grupės klasifikavimo metrikų priklausomybės grafike galime pastebėti, jog maksimali F1 metrikos reikšmė 0,746 gauta parinkus parametą $k = 2$. Jautrumas šiuo atveju lygus 0,66, t.y. vidutiniškai 66% žemiausios rizikos grupės stebėjimų buvo suklasifikuoti kaip žemiausios rizikos grupės stebėjimai, t.y. 0,09 mažesnis už maksimalų jautrumą gautą, kai $k = 1$. Maksimalus tikslumas 0,86 gautas parinkus parametą $k = 2$.

Iš 2.13-2.15 pav. pateiktų klasifikavimo kokybės vertinimo metrikų priklausomybių nuo parametro k , galime pastebėti, kad skirtingos rizikos klasės pasižymi nevienodu metrikų verčių svyravimu. Siekiant apibendrinti klasifikatoriaus vertinimą sudaryta apibendrinto harmoninio vidurkio metrikos priklausomybė nuo parametro k . Remiantis 2.16 pav. pateikta apibendrintos F1 metrikos priklausomybę nuo parametro k galime išskirti, jog didžiausia vertė 0,68 buvo gauta, kai $k = 1$. Atsižvelgiant į tai, kad parinkus parametą $k = 1$ aukščiausios rizikos grupės klasifikavimo atžvilgiu taip pat gautą maksimali F1 metrikos reikšmė, 1 toliau laikomas optimalią parametro k vertę.



2.16 pav. F1 metrikos priklausomybė nuo parametro k

2.3.6. KLASIFIKAVIMO MODELIŲ PALYGINIMAS

Nustačius tinkamiausius pasirinktų klasifikavimo modelių parametrus analizuojamai imčiai klasifikuoti atlikome modelių apmokymą naudojant visą apmokymo imtį (1044 taškai) ir klasifikavimo kokybės vertinimą testinės imties klasifikavimui (261 taškai). Gauti rezultatai pateikti 2.7-2.12 lentelėse.

2.7 lentelė

Atramiųjų vektorių modelio grįšto DBT koeficientais sumaišymo matrica

		Prognozuojamos rizikos grupės		
		Aukščiausia	Vidutinė	Žemiausia
Tikrosios rizikos grupės	Aukščiausia	5	1	0
	Vidutinė	1	214	5
	Žemiausia	0	7	28

2.8 lentelė

Atraminų vektorių modelio grįsto DBT koeficientais klasifikavimo kokybės metrikos

Rizikos grupė	Tikslumas	Jautrumas	F1
Aukščiausia	0,83	0,83	0,83
Vidutinė	0,96	0,97	0,97
Žemiausia	0,85	0,8	0,82

Apibendrinus 2.7-2.8 lentelėse pateiktus rezultatus galime pastebėti, jog aukščiausios rizikos grupės tikslumo, jautrumo ir atitinkamai harmoninio vidurkio metrikų vertės yra lygios. Kaip matome iš 2.7 lentelėje pateiktos sumaišymo matricos toks rezultatas gautas dėl to, kad neteisingai suklasifikuotų eilučių skaičius sutampa su neteisingai aukščiausios rizikos grupei priskirtų rinkinių skaičiumi. Apskaičiavus apibendrintą F1 metrikos reikšmę visų grupių atžvilgiu gavome, jog bendras klasifikatoriaus jautrumo ir tikslumo harmoninis vidurkis testiniai imčiai klasifikuoti naudojantis parinktu atraminu vektorių klasifikatoriumi grįstu DBT koeficientais su jo optimaliais parametrais yra 0,87.

2.9 lentelė

Artimiausių kaimynų modelio grįsto DBT koeficientais sumaišymo matrica

		Prognozuojamos rizikos grupės		
		Aukščiausia	Vidutinė	Žemiausia
Tikrosios rizikos grupės	Aukščiausia	4	2	0
	Vidutinė	3	213	4
	Žemiausia	0	5	30

2.10 lentelė

Artimiausių kaimynų modelio grįsto DBT koeficientais klasifikavimo kokybės metrikos

Rizikos grupė	Tikslumas	Jautrumas	F1
Aukščiausia	0,57	0,67	0,62
Vidutinė	0,97	0,97	0,97
Žemiausia	0,88	0,86	0,87

Remiantis 2.9-2.10 lentelėse pateiktus rezultatus galime pastebėti, jog iš 220 vidutinės rizikos grupės matavimų eilučių 3 buvo suklasifikuotos kaip aukščiausios rizikos grupės rinkiniai, o 4 kaip žemiausios, dėl to gautas jautrumas 0,97. Apskaičiavus apibendrintą F1 metrikos reikšmę visų grupių atžvilgiu gavome, jog bendras klasifikatoriaus jautrumo ir tikslumo harmoninis

vidurkis testiniai imčiai klasifikuoti naudojantis parinktu atraminiu vektorių modeliu grįstu DBT koeficientais su jo optimaliais parametrais yra 0,82.

2.11 lentelė

Artimiausių kaimynų modelio grįsto DTW metrika sumaišymo matrica

		Prognozuojamos rizikos grupės		
		Aukščiausia	Vidutinė	Žemiausia
Tikrosios rizikos grupės	Aukščiausia	5	1	0
	Vidutinė	3	205	18
	Žemiausia	1	6	22

2.12 lentelė

Artimiausių kaimynų modelio grįsto DTW metrika klasifikavimo kokybės metrikos

Rizikos grupė	Tikslumas	Jautrumas	F1
Aukščiausia	0,55	0,83	0,66
Vidutinė	0,97	0,9	0,94
Žemiausia	0,55	0,76	0,64

Apibendrinus 2.11-2.12 lentelėse pateiktus rezultatus galime pastebėti, jog aukščiausios ir žemiausios rizikos grupės klasifikavimo tikslumai nagrinėjamu atveju sutapo, nes teisingai šioms grupėms priskirtų laiko eilučių skaičius sudarė 55% visų tai grupei priskirtų laiko eilučių skaičiaus. Apskaičiavus apibendrintą F1 metrikos reikšmę visų grupių atžvilgiu gavome, jog bendras klasifikatoriaus jautrumo ir tikslumo harmoninis vidurkis testiniai imčiai naudojantis šiuo klasifikavimo modeliu, su parinktais optimaliais parametrais yra 0,75.

Palyginus 2.7-2.12 lentelėse gautus klasifikatorių taikymo testinės imties klasifikavimui rezultatus galime pastebėti, jog taikant artimiausių kaimynų klasifikavimo modelį grįsta DTW metrika pasiektas 0,75 harmoninio vidurkio tarp klasifikavimo jautrumo ir tikslumo rezultatas yra mažiausias iš analizuojamų modelių. Taip pat galime išskirti, kad klasifikuojant matavimų eilutes šiuo modeliu lyginant su kitais daugiausia laiko eilučių buvo klaidingai priskirtą žemiausios ir aukščiausios rizikos grupėms, tai gali neigiamai įtakoti pastatų nusėdimo vertinimo rezultatus. Tuo tarpu taikant artimiausių kaimynų klasifikavimo modelį grįsta DBT koeficientais pasiektas 0,82 harmoninio vidurkio tarp klasifikavimo jautrumo ir tikslumo rezultatas. Tačiau klasifikuojant šiuo modeliu aukščiausios rizikos grupės atžvilgiu buvo gautas tik 0,57 tikslumas (klaidingai aukščiausios rizikos grupei priskirta 43% klasifikuojamų eilučių) ir 0,67 jautrumas (teisingai aukščiausios rizikos grupei priskirta 67% aukščiausios rizikos grupės eilučių). Dėl to galime daryti

prielaidą apie klaidingą aukščiausios rizikos grupės vertinimą taikant šį modelį pastatų pamatų nusėdimo vertinime.

Iš išanalizuotų modelių geriausi klasifikavimo rezultatai buvo pasiekti atraminių vektorių klasifikavimo modeliu grįstu DBT koeficientais apibendrintas jautrumo ir tikslumo harmoninis vidurkis yra 0,87, aukščiausios rizikos grupės klasifikavimo tikslumas ir jautrumas 0,83, dėl to galutinio pastatų pamatų rizikos vertinimo modeliui siūloma taikyti atraminių vektorių klasifikatorių grįsta 2.3.2 skyriuje atrinktais DBT koeficientais, parinkus radialinių bazių branduolio funkcija su parametrais $(\gamma, C) = (0,02; 15)$.

2.3.7. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO MODELIO PROGRAMINĖ REALIZACIJA IR TAIKYMO REZULTATAI

Parinkus optimalų pastatų pamatų nusėdimo matavimų eilučių klasifikavimo modelį, buvo sudaryta tikimybinio pastatų nusėdimo rizikos vertinimo funkcija *BuildingSubsidenceEvaluation* su parametrais (*TrainPoinCoeffe, NewData*):

TrainPoinCoeffe - atrinktų apmokymo imties DBT koeficientų aibė, su koeficientų vektoriams priskirtu ekspertiniu vertinimu.

Test – nusėdimo matavimų eilučių sąrašas gautas atlikus pirminius duomenų paruošimo žingsnius. Failas sudarytas iš matavimų, kuriems nebuvo atliktas ekspertinis vertinimas.

Sudarytoji funkcija pirmiausia apskaičiuoja ir suformuoja DBT koeficientų rinkinius. Tam naudojama modelio sudarymo metu sukurta funkcija „*DwtTransform*“. Sekančiame etape funkcija atrenka iš testinės imties koeficientų rinkinio tik tuos koeficientus, kurie sudaro apmokymo imtį. Naudojantis R paketo *e1071* funkcija *svm* sudaromas atraminių vektorių klasifikavimo modelis, parinkus radialinių bazių branduolio funkcija su parametrais $(\gamma, C) = (0,02; 15)$ ir naudojantis R paketo *stat* funkcija *predict* atliekamas matavimo taškų suskirstymas į rizikos grupes. Apibendrinti gauti rezultatai pateikiami lentelėje su pastato ID, rizikos lygio kodu ir tikimybinio pastato rizikos lygio vertinimu. Gauti rezultatai išsaugomi csv formatu, bei išvedami R konsolėje. Funkcijos R kodas pateiktas 2 priede.

Iš pastatų pamatų nusėdimo matavimų bazės atsitiktinai atrinkus 412 pastatų nusėdimo matavimų eilutes, naudojantis sukurta *BuildingSubsidenceEvaluation* funkcija buvo atliktas tikimybinis pastatų pamatų nusėdimo rizikos vertinimas. Gauti rezultatai pateikiami 1 priede.

Iš gautų rezultatų galime pastebėti, kad naudojantis sudarytu pastatų pamatų nusėdimo modeliu atrinktai imčiai su tikimybe lygie 1, 5 pastatai su buvo priskirti aukščiausios rizikos grupei, 9 pastatai buvo priskirti žemiausios rizikos grupei, o 364 vidutinės rizikos grupei. Iš likusių 34 pastatų, 20 pastatų su tikimybe iš intervalo $[0,6; 0,95]$ buvo priskirtą vidutinės rizikos grupei. Aukščiausio rizikos grupei su tikimybėmis 0,6 ir 0,7 buvo priskirta po vieną pastatą. Ir vienas

pastatas su tikimybe lygia 0,6 buvo priskirtas žemiausios rizikos pastatų grupei. Keturi pastatai su vienoda 50 % tikimybe buvo priskirti aukščiausios ir vidutinės rizikos grupėms. O likę 7 pastatai su vienoda 50 % tikimybe buvo priskirti vidutinės ir žemiausios rizikos lygio pastatų grupėms.

IŠVADOS

1. Apžvelgus literatūrą pastatų pamatų nusėdimo vertinimo tematika nustatyta, kad literatūroje dažniausiai analizuojama matavimų metodika ir tikslumas, o ne matavimų analizės problemos, apsiribojama matavimų laiko eilutės trendo išskyrimu.
2. Pasiūlyta pastatų pamatų nusėdimo rizikos vertinimui pagal duomenis gautus iš palydovų taikyti laiko eilučių transformacijas ir klasifikavimo metodus.
3. Sudarius pastatų pamatų nusėdimo matavimų eilučių artimiausių kaimynų klasifikavimo modelį grįsta DTW metrika gauta, kad didžiausia F1 metrikos vertė 0,68, kai $k = 1$. Taikant atraminių vektorių modelį grįstą DBT koeficientais gauta didžiausia F1 metrikos vertė yra 0,79 parinkus radialinių bazių branduolio funkciją ir parametrų rinkinį $(\gamma, C) = (0,02; 15)$. Taikant artimiausių kaimynų modelį grįstą DBT koeficientais gauta, kad didžiausia F1 metrikos vertė 0,75., kai $k = 1$.
4. Atlikus testinės imties klasifikavimą didžiausia F1 metrikos vertė 0,87 gauta naudojant atraminių vektorių modelį grįstą DBT koeficientais ir parinkus radialinių bazių branduolio funkciją bei parametrų rinkinį $(\gamma, C) = (0,02; 15)$.
5. Pritaikius sudarytą modelį atsitiktinai atrinktų 412 pastatų rizikos vertinimui, gauta, kad 7 pastatai priskirti aukščiausios rizikos grupei, 384 pastatai priskirti vidutinės rizikos grupei.

Rekomendacijos. Klasifikavimo modelių tobulinimui siūloma rinkti ir įtraukti į modelį papildomą informaciją apie matavimų taškų geometrinį išsidėstymą bei kitus pastatus charakterizuojančius parametrus. Panaudoti daugiau klasifikavimo metodų ir įvertinti jų klasifikavimo kokybę.

LITERATŪROS SĄRAŠAS

1. Bramer, M. Principles of Data Mining 2007 p.89 – 247.
2. Bresch,D. Corti, T. Muccione, V, P. Kollner-Heck, S. Seneviratne, I. Simulating past droughts and associated building damages in France/ Hydrol. Earth Syst. Sci., 2009, 13,p. 177–179.
3. Bresch,D.,S. Corti,T ,Wuest,M., and Seneviratne, I.Drought-induced building damages from simulations at regional scale/ Earth Syst. Sci., 2011,11,p 335–342.
4. Chaovalit, P., Gangopadhyay, A., Karabatis, G., and Chen, Z. 2011. Discrete wavelet transform-based time series analysis and mining. ACM Comput. Surv. 43, 2, Article 6 (January 2011), p. 37-39.
5. Chiu, B, Jessica, L. KeoghE.,and Lonardi,S., A Symbolic Representation of Time Series, with Implications for Streaming Algorithms/University of California - Riverside Computer Science & Engineering Department Riverside, CA 92521.
6. Chotira, AR.,Keogh, E. Everything you know about Dynamic Time Warping is Wrong /Department of Computer Science and Engineering University of California, Riverside . Priegą per internetą: http://wearables.cc.gatech.edu/paper_of_week/DTW_myths.pdf.
7. Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium. Opening black box Data Mining models using Sensitivity Analysis. Priegą per internetą: <http://ieeexplore.ieee.org/document/5949423/?reload=true>.
8. Cooper,A.H The classification, recording, databasing and use of information about building damage caused by subsidence and landslides./ British Geological Survey. Priegą per inernetą : http://sgil.isprambiente.it/sinkhole/pubblicazioni/internazionali/Cooper_2008_building_damage_classification_and_recording.pdf.
9. Cortez,P.Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In: Advances in Data Mining -- Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining. Berlin Germany, 2010,p. 572-583.
10. David L. Dursun, D and Olson,D. Advanced Data Mining Techniques.2009, 180p.
11. Fu,K. Wang, H. Xian Sun, X, Xu,G. and Zhang,Y.Model accuracy comparison for high resolution insar coherence statistics over urban areas./The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLI-B7, 2016.
12. Frank,F Witten, I.H ir Frank,F. Data Mining Practical Machine Learning Tools and Techniques 2005, 525 p.
13. Guoshao, SU. Huanling, Z and Keshi, Z. Deformation Prediction of Foundation Pit Using Gaussian Process Machine Learning. Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications. 2009, p. 99-102.
14. Hyypä, J ,Karila,K.and Karjalainen,M. . Urban land subsidence studies in Finland using synthetic aperture radar images and coherent targets. *The Photogrammetric Journal of Finland*, 2010, 19, p. 43-53.
15. Makhtar, M., Neagu,DC , Ridley,MJ. Comparing Multi-class Classifiers: On the Similarity of Confusion Matrices for Predictive Toxicology Applications/ 12th International Conference, Norwich, UK, September 7-9, 2011.
16. Makhtar,M.Daniel Mick,NC. Ridley,J. Nesting One-Against-One Algorithm Based on SVMs for Pattern Classification/ Browse Journals & Magazines 18 November 2008.
17. Mitchell, T. Machine learning. McGraw Hill. 1996.
18. Tomas R.,Hibridinių kalbos atpažinimo metodų tyrimas, Mokslinė ataskaita MII-DS-07T-14-13, 2014. Priegą per internetą: http://old.mii.lt/files/07t_rasymas_ataskaita2014.pdf.
19. The Australian National University Canberra /Research School of Earth Sciences. Priegą per internetą: http://www.chiefscientist.nsw.gov.au/__data/assets/pdf_file/0016/33028/Subsidence-Monitoring_McClusky-Tregoning_ANU.pdf .

20. Zhang H., Ho T.B., Lin MS., Liang X. (2006) Feature Extraction for Time Series Classification Using Discriminating Wavelet Coefficients. In: Wang J., Yi Z., Zurada J.M., Lu BL., Yin H. (eds) Advances in Neural Networks - ISNN 2006. ISNN 2006. Lecture Notes in Computer Science, vol 3971. Springer, Berlin, Heidelberg.
21. Prieigą per internetą: <http://r-statistics.co/Outlier-Treatment-With-R.html>.

**1 PRIEDAS. PASTATŲ PAMATŲ NUSĖDIMO VERTINIMO
MODELIO TAIKYMO REZULTATAI**

ID	Rizikos lygis	Tikimybė
479100000000001	Medium	1,00
479100000000003	Medium	1,00
479100000000004	Medium	1,00
479100000000005	Medium	1,00
479100000000007	Medium	1,00
479100000000008	Medium	1,00
479100000000011	Medium	1,00
479100000000012	Highest	0,09
479100000000012	Medium	0,91
479100000000013	Highest	0,08
479100000000013	Medium	0,92
479100000000014	Medium	1,00
479100000000015	Medium	1,00
479100000000016	Medium	1,00
479100000000017	Medium	1,00
479100000000018	Medium	1,00
479100000000019	Medium	1,00
479100000000020	Medium	1,00
479100000000021	Medium	1,00
479100000000022	Medium	1,00
479100000000023	Medium	1,00
479100000000024	Medium	1,00
479100000000025	Highest	0,06
479100000000025	Lowest	0,06
479100000000025	Medium	0,88
479100000000026	Medium	1,00
479100000000029	Medium	1,00
479100000000030	Medium	1,00
479100000000031	Lowest	0,60
479100000000031	Medium	0,40
479100000000032	Lowest	1,00
479100000000033	Medium	1,00
479100000000034	Medium	1,00
479100000000035	Medium	1,00
479100000000036	Highest	1,00
479100000000037	Medium	1,00
479100000000038	Medium	1,00
479100000000040	Medium	1,00
479100000000041	Medium	1,00
479100000000042	Lowest	0,17
479100000000042	Medium	0,83

479100000000043	Medium	1,00
479100000000045	Medium	1,00
479100000000046	Medium	1,00
479100000000047	Medium	1,00
479100000000048	Medium	1,00
479100000000049	Medium	1,00
479100000000050	Medium	1,00
479100000000052	Medium	1,00
479100000000053	Lowest	0,50
479100000000053	Medium	0,50
479100000000055	Medium	1,00
479100000000056	Lowest	0,50
479100000000056	Medium	0,50
479100000000058	Medium	1,00
479100000000059	Medium	1,00
479100000000060	Medium	1,00
479100000000061	Medium	1,00
479100000000064	Medium	1,00
479100000000066	Medium	1,00
479100000000071	Medium	1,00
479100000000072	Medium	1,00
479100000000079	Medium	1,00
479100000000080	Medium	1,00
479100000000081	Lowest	1,00
479100000000083	Medium	1,00
479100000000084	Medium	1,00
479100000000085	Medium	1,00
479100000000089	Medium	1,00
479100000000090	Medium	1,00
479100000000091	Medium	1,00
479100000000092	Medium	1,00
479100000000094	Medium	1,00
479100000000095	Medium	1,00
479100000000096	Medium	1,00
479100000000097	Medium	1,00
479100000000098	Medium	1,00
479100000000100	Medium	1,00
479100000000101	Medium	1,00
479100000000103	Medium	1,00
479100000000104	Medium	1,00
479100000000105	Medium	1,00
479100000000106	Medium	1,00
479100000000107	Medium	1,00
479100000000108	Medium	1,00
479100000000109	Lowest	0,50
479100000000109	Medium	0,50

479100000000110	Medium	1,00
479100000000111	Highest	0,33
479100000000111	Medium	0,67
479100000000112	Medium	1,00
479100000000113	Medium	1,00
479100000000114	Medium	1,00
479100000000115	Medium	1,00
479100000000117	Medium	1,00
479100000000123	Medium	1,00
479100000000124	Medium	1,00
479100000000126	Lowest	0,40
479100000000126	Medium	0,60
479100000000128	Medium	1,00
479100000000131	Medium	1,00
479100000000133	Medium	1,00
479100000000136	Medium	1,00
479100000000138	Medium	1,00
479100000000140	Highest	0,11
479100000000140	Lowest	0,11
479100000000140	Medium	0,79
479100000000143	Medium	1,00
479100000000144	Medium	1,00
479100000000146	Medium	1,00
479100000000147	Medium	1,00
479100000000148	Medium	1,00
479100000000149	Medium	1,00
479100000000151	Medium	1,00
479100000000153	Medium	1,00
479100000000156	Medium	1,00
479100000000157	Medium	1,00
479100000000160	Lowest	0,17
479100000000160	Medium	0,83
479100000000161	Lowest	0,03
479100000000161	Medium	0,97
479100000000162	Medium	1,00
479100000000163	Medium	1,00
479100000000166	Medium	1,00
479100000000167	Medium	1,00
479100000000168	Medium	1,00
479100000000170	Medium	1,00
479100000000172	Highest	0,50
479100000000172	Medium	0,50
479100000000176	Medium	1,00
479100000000178	Highest	0,50
479100000000178	Medium	0,50
479100000000185	Medium	1,00

479100000000186	Medium	1,00
479100000000187	Medium	1,00
479100000000188	Lowest	0,50
479100000000188	Medium	0,50
479100000000192	Medium	1,00
479100000000193	Medium	1,00
479100000000196	Medium	1,00
479100000000197	Medium	1,00
479100000000198	Medium	1,00
479100000000204	Medium	1,00
479100000000209	Medium	1,00
479100000000211	Medium	1,00
479100000000213	Medium	1,00
479100000000214	Medium	1,00
479100000000215	Lowest	1,00
479100000000217	Medium	1,00
479100000000218	Medium	1,00
479100000000228	Medium	1,00
479100000000229	Medium	1,00
479100000000230	Medium	1,00
479100000000232	Medium	1,00
479100000000233	Medium	1,00
479100000000238	Medium	1,00
479100000000244	Medium	1,00
479100000000247	Medium	1,00
479100000000251	Medium	1,00
479100000000252	Medium	1,00
479100000000253	Medium	1,00
479100000000258	Medium	1,00
479100000000260	Medium	1,00
479100000000262	Medium	1,00
479100000000263	Medium	1,00
479100000000264	Medium	1,00
479100000000266	Medium	1,00
479100000000269	Medium	1,00
479100000000270	Medium	1,00
479100000000272	Medium	1,00
479100000000273	Medium	1,00
479100000000278	Medium	1,00
479100000000279	Lowest	0,50
479100000000279	Medium	0,50
479100000000280	Medium	1,00
479100000000281	Medium	1,00
479100000000282	Medium	1,00
479100000000283	Medium	1,00
479100000000285	Medium	1,00

479100000000286	Medium	1,00
479100000000287	Medium	1,00
479100000000288	Medium	1,00
479100000000289	Medium	1,00
479100000000292	Medium	1,00
479100000000293	Medium	1,00
479100000000294	Medium	1,00
479100000000295	Medium	1,00
479100000000296	Highest	0,33
479100000000296	Medium	0,67
479100000000297	Medium	1,00
479100000000301	Medium	1,00
479100000000302	Medium	1,00
479100000000306	Medium	1,00
479100000000308	Lowest	0,50
479100000000308	Medium	0,50
479100000000309	Medium	1,00
479100000000311	Highest	0,50
479100000000311	Medium	0,50
479100000000312	Medium	1,00
479100000000314	Medium	1,00
479100000000320	Medium	1,00
479100000000322	Medium	1,00
479100000000335	Medium	1,00
479100000000338	Medium	1,00
479100000000339	Medium	1,00
479100000000341	Medium	1,00
479100000000344	Medium	1,00
479100000000345	Medium	1,00
479100000000348	Medium	1,00
479100000000349	Lowest	1,00
479100000000350	Medium	1,00
479100000000351	Medium	1,00
479100000000352	Medium	1,00
479100000000353	Highest	0,60
479100000000353	Medium	0,40
479100000000354	Highest	0,40
479100000000354	Medium	0,60
479100000000355	Medium	1,00
479100000000356	Medium	1,00
479100000000361	Medium	1,00
479100000000364	Medium	1,00
479100000000366	Medium	1,00
479100000000367	Medium	1,00
479100000000368	Medium	1,00
479100000000371	Medium	1,00

479100000000372	Medium	1,00
479100000000375	Medium	1,00
479100000000377	Medium	1,00
479100000000378	Lowest	1,00
479100000000379	Medium	1,00
479100000000383	Medium	1,00
479100000000384	Medium	1,00
479100000000388	Medium	1,00
479100000000392	Medium	1,00
479100000000393	Medium	1,00
479100000000394	Medium	1,00
479100000000395	Medium	1,00
479100000000396	Medium	1,00
479100000000398	Medium	1,00
479100000000400	Medium	1,00
479100000000401	Lowest	0,33
479100000000401	Medium	0,67
479100000000403	Medium	1,00
479100000000404	Medium	1,00
479100000000408	Medium	1,00
479100000000411	Highest	0,13
479100000000411	Medium	0,88
479100000000413	Medium	1,00
479100000000414	Medium	1,00
479100000000415	Medium	1,00
479100000000416	Lowest	0,50
479100000000416	Medium	0,50
479100000000419	Medium	1,00
479100000000423	Medium	1,00
479100000000427	Medium	1,00
479100000000431	Highest	1,00
479100000000434	Medium	1,00
479100000000436	Medium	1,00
479100000000437	Medium	1,00
479100000000438	Medium	1,00
479100000000439	Medium	1,00
479100000000444	Medium	1,00
479100000000445	Medium	1,00
479100000000446	Medium	1,00
479100000000453	Medium	1,00
479100000000454	Medium	1,00
479100000000456	Medium	1,00
479100000000457	Medium	1,00
479100000000459	Lowest	0,33
479100000000459	Medium	0,67
479100000000462	Medium	1,00

479100000000464	Medium	1,00
479100000000465	Medium	1,00
479100000000466	Medium	1,00
479100000000468	Medium	1,00
479100000000472	Medium	1,00
479100000000478	Medium	1,00
479100000000479	Highest	0,67
479100000000479	Medium	0,33
479100000000480	Lowest	0,40
479100000000480	Medium	0,60
479100000000483	Medium	1,00
479100000000488	Medium	1,00
479100000000490	Medium	1,00
479100000000491	Medium	1,00
479100000000499	Medium	1,00
479100000000507	Medium	1,00
479100000000508	Medium	1,00
479100000000509	Medium	1,00
479100000000511	Medium	1,00
479100000000512	Medium	1,00
479100000000513	Medium	1,00
479100000000515	Medium	1,00
479100000000518	Medium	1,00
479100000000519	Medium	1,00
479100000000522	Medium	1,00
479100000000523	Medium	1,00
479100000000524	Medium	1,00
479100000000528	Medium	1,00
479100000000533	Medium	1,00
479100000000535	Medium	1,00
479100000000536	Medium	1,00
479100000000539	Medium	1,00
479100000000545	Medium	1,00
479100000000546	Medium	1,00
479100000000547	Lowest	0,33
479100000000547	Medium	0,67
479100000000549	Medium	1,00
479100000000550	Medium	1,00
479100000000552	Medium	1,00
479100000000555	Medium	1,00
479100000000563	Medium	1,00
479100000000565	Medium	1,00
479100000000568	Medium	1,00
479100000000570	Medium	1,00
479100000000575	Medium	1,00
479100000000576	Medium	1,00

479100000000579	Highest	1,00
479100000000580	Medium	1,00
479100000000581	Medium	1,00
479100000000587	Medium	1,00
479100000000589	Medium	1,00
479100000000590	Medium	1,00
479100000000598	Medium	1,00
479100000000601	Highest	0,25
479100000000601	Medium	0,75
479100000000603	Medium	1,00
479100000000604	Medium	1,00
479100000000607	Medium	1,00
479100000000615	Medium	1,00
479100000000618	Medium	1,00
479100000000619	Medium	1,00
479100000000623	Medium	1,00
479100000000626	Medium	1,00
479100000000629	Medium	1,00
479100000000636	Medium	1,00
479100000000637	Medium	1,00
479100000000640	Medium	1,00
479100000000642	Medium	1,00
479100000000643	Medium	1,00
479100000000645	Medium	1,00
479100000000646	Medium	1,00
479100000000647	Medium	1,00
479100000000648	Medium	1,00
479100000000649	Medium	1,00
479100000000653	Medium	1,00
479100000000657	Medium	1,00
479100000000659	Medium	1,00
479100000000662	Lowest	0,33
479100000000662	Medium	0,67
479100000000666	Medium	1,00
479100000000670	Medium	1,00
479100000000677	Medium	1,00
479100000000684	Medium	1,00
479100000000685	Medium	1,00
479100000000700	Medium	1,00
479100000000702	Medium	1,00
479100000000703	Lowest	1,00
479100000000704	Medium	1,00
479100000000705	Medium	1,00
479100000000709	Medium	1,00
479100000000710	Medium	1,00
479100000000711	Medium	1,00

479100000000712	Medium	1,00
479100000000718	Medium	1,00
479100000000720	Medium	1,00
479100000000723	Medium	1,00
479100000000726	Medium	1,00
479100000000728	Medium	1,00
479100000000732	Medium	1,00
479100000000734	Medium	1,00
479100000000740	Lowest	0,20
479100000000740	Medium	0,80
479100000000742	Highest	0,05
479100000000742	Medium	0,95
479100000000744	Medium	1,00
479100000000753	Medium	1,00
479100000000762	Medium	1,00
479100000000764	Medium	1,00
479100000000768	Medium	1,00
479100000000772	Medium	1,00
479100000000774	Medium	1,00
479100000000783	Medium	1,00
479100000000784	Medium	1,00
479100000000787	Medium	1,00
479100000000788	Medium	1,00
479100000000789	Medium	1,00
479100000000797	Medium	1,00
479100000000802	Medium	1,00
479100000000804	Lowest	1,00
479100000000808	Medium	1,00
479100000000812	Medium	1,00
479100000000816	Medium	1,00
479100000000817	Medium	1,00
479100000000818	Medium	1,00
479100000000819	Medium	1,00
479100000000826	Medium	1,00
479100000000828	Medium	1,00
479100000000840	Medium	1,00
479100000000841	Medium	1,00
479100000000844	Medium	1,00
479100000000845	Medium	1,00
479100000000848	Medium	1,00
479100000000853	Medium	1,00
479100000000857	Lowest	1,00
479100000000858	Medium	1,00
479100000000859	Medium	1,00
479100000000860	Medium	1,00
479100000000861	Medium	1,00

479100000000864	Medium	1,00
479100000000865	Medium	1,00
479100000000867	Medium	1,00
479100000000869	Medium	1,00
479100000000872	Medium	1,00
479100000000875	Medium	1,00
479100000000884	Highest	1,00
479100000000897	Medium	1,00
479100000000905	Medium	1,00
479100000000906	Medium	1,00
479100000000914	Medium	1,00
479100000000930	Medium	1,00
479100000000936	Medium	1,00
479100000000944	Medium	1,00
479100000000946	Medium	1,00
479100000000950	Medium	1,00
479100000000954	Medium	1,00
479100000000955	Lowest	1,00
479100000000956	Medium	1,00
479100000000960	Medium	1,00
479100000000971	Medium	1,00
479100000000972	Medium	1,00
479100000000975	Medium	1,00
479100000000984	Medium	1,00
479100000000986	Medium	1,00
479100000000992	Medium	1,00
479100000000994	Medium	1,00
479100000000996	Medium	1,00
479100000001005	Medium	1,00
479100000001008	Medium	1,00
479100000001011	Medium	1,00
479100000001013	Medium	1,00
479100000001022	Highest	0,50
479100000001022	Medium	0,50
479100000001032	Medium	1,00
479100000001033	Medium	1,00
479100000001034	Medium	1,00
479100000001041	Highest	1,00
479100000001043	Medium	1,00

2 PRIEDAS. PASTATŲ PAMATŲ NUSĖDIMO MODELIO SUDARYMO SUDARYMO PROGRAMINĖ REALIZACIJA

Duomenų apdorojimo žingsnis.

```

TrainPoint<-read.csv("C:/Users/ugne/Documents/MAGISTRAS/IV
kursas/Duome/MyData1.csv")
TrainPoint$ID<-as.factor (TrainPoint$ID)
TrainPointSet<-TrainPoint[,c(2,3,5,6)]
#ClassID
classID<-LearningSet[!duplicated(LearningSet$ID),c(1,10)]
classID1<-classID[classID$RiskCodesUsedCurrenty=='ingrijpen',]
classID2<-classID[classID$RiskCodesUsedCurrenty=='Ingrijpen',]
classID1$RiskCodesUsedCurrenty<-'Highest'
classID2$RiskCodesUsedCurrenty<-'Highest'
classID3<-classID[classID$RiskCodesUsedCurrenty=='inspectie',]
classID4<-classID[classID$RiskCodesUsedCurrenty=='monitoren',]
classID3$RiskCodesUsedCurrenty<-'Medium'
classID4$RiskCodesUsedCurrenty<-'Lowest'
classID<-as.data.frame(rbind(classID1,classID2,classID3,classID4))
classID$RiskCodesUsedCurrenty<-
as.factor(classID$RiskCodesUsedCurrenty)
classID<-classID[!duplicated(classID),]
TrainPointSet<-
merge (TrainPointSet,classID,by.x="ID",by.y="ID",all.x =TRUE)
#-----##
TrainPointSet<-TrainPointSet[!duplicated (TrainPointSet),]
p1<-as.data.frame (table (unlist (TrainPointSet$SatelitePointID)))
p1$Freq<-as.factor (p1$Freq)
p12<-as.data.frame (p1$Freq)
TrainPointSet<-
TrainPointSet [TrainPointSet$SatelitePointID!='L93690P42954',]
p2<-
as.data.frame (table (unlist (TrainPointSet$RiskCodesUsedCurrenty)))
TrainPointSet<-
TrainPointSet [order (TrainPointSet$ID,TrainPointSet$SatelitePointID,TrainPointSet$Date),]
a <-
ts (TrainPointSet [TrainPointSet$SatelitePointID=='L85145P47155',4])
library (forecast)
plot (a,main="Aukščiausia rizikos klasė",xaxt="n",ylab="Pokytis, m",
xlab='Stebėjimas')
#-----#

```

```

#Missing matavimu parinkimas
MissingDate<-read.csv("C:/Users/ugne/Documents/MAGISTRAS/IV
kursas/Duome/MissingDate.csv")
rezUniq<-sqldf("SELECT ID, SatelitePointID,RiskCodesUsedCurrenty,
COUNT(*) as Freq
FROM TrainPointSet
GROUP BY ID, SatelitePointID")
rezpoint<-sqldf("SELECT ID, COUNT(*) as Freq
FROM rezUniq
GROUP BY ID")
rezHouse <- sqldf("SELECT ID, RiskCodesUsedCurrenty FROM rezUniq
GROUP BY ID ")
rezClass<-sqldf("SELECT RiskCodesUsedCurrenty, COUNT(*) as Freq FROM
rezHouse
GROUP BY RiskCodesUsedCurrenty")
summary(rezpoint)

kl<-NULL
for(i in 1:nrow(rezUniq)){
  set<-rezUniq[i,-4]
  for (j in 1:nrow(MissingDate)){
    set1
data.frame(set$ID,set$SatelitePointID,MissingDate[j,],',',set$RiskCodes
UsedCurrenty)
kl <-rbind(kl,set1)
}

}
names(kl)<-names(TrainPointSet)
TrainPointSet<-rbind(TrainPointSet,kl)
TrainPointSet$SateliteMeasure<-
as.double(TrainPointSet$SateliteMeasure)
summary(TrainPointSet)
TrainPointSet<-
TrainPointSet[order(TrainPointSet$ID,TrainPointSet$SatelitePointID,as.
Date(TrainPointSet$Date)),]
setwd("C:/Users/ugne/Documents/MAGISTRAS/IV kursas/Duome")
write.csv(TrainPointSet, file = "TrainPointSet.csv")

```

```

TrainPointSetP<-read.csv("C:/Users/ugne/Documents/MAGISTRAS/IV
kursas/Duome/TrainPointSet.csv")
TrainPointSetP$ID<-as.factor(TrainPointSetP$ID)
TrainPointSetP<-TrainPointSetP[,c(-1)]
summary(TrainPointSetP)
p1<-as.data.frame(table(unlist(TrainPointSetP$SatelitePointID)))
p13<-TrainPointSetP[,c(2,5)]
p14<-p13[!duplicated(p13),]
summary(p14)
TrainPointSet1<-TrainPointSetP
rezUniq<-sqldf("SELECT ID, SatelitePointID,RiskCodesUsedCurrenty,
COUNT(*) as Freq
FROM TrainPointSet1
GROUP BY ID, SatelitePointID")
rezpointP<-sqldf("SELECT ID, COUNT(*) as Freq
FROM rezUniq
GROUP BY ID")
rezpointP2<-sqldf("SELECT Freq,COUNT(*) as Freq1
FROM rezpointP
GROUP BY Freq")
#Missing reikšmių interpoliavimas
for (i in 1:nrow(rezUniq)){
xs<-
ts(TrainPointSet1[TrainPointSet1$SatelitePointID==rezUniq[i,2],4])
TrainPointSet1$Index<-c()
plot(ts(xs))
xsClean<-tsclean(xs)
index<-c(1:207)
l<-na.ma(xsClean[1:207], k = 5, weighting = "simple")

TrainPointSet1[TrainPointSet1$SatelitePointID==rezUniq[i,2],4]<-l

TrainPointSet1[TrainPointSet1$SatelitePointID==rezUniq[i,2],6]<-index
}
TrainPointSet3<-TrainPointSet1
TrainPointSet3<-
TrainPointSet3[!duplicated(TrainPointSet3[c(2,3)]),]
summary(TrainPointSet1)

```

```

#-----#
#Summary Statistic by class
#-----#
statisticByClass<-TrainPointSet[,c(3,4,5)]
statisticByClassH<-
statisticByClass[statisticByClass$RiskCodesUsedCurrenty=="Highest",]
summary(statisticByClassH)
sd(statisticByClassH$SateliteMeasure)
statisticByClassL<-
statisticByClass[statisticByClass$RiskCodesUsedCurrenty=="Lowest",]
summary(statisticByClassL)
sd(statisticByClassL$SateliteMeasure)
statisticByClassM<-
statisticByClass[statisticByClass$RiskCodesUsedCurrenty=="Medium",]
summary(statisticByClassM)
sd(statisticByClassM$SateliteMeasure)
#-----#
#Plot after missing value interpolation #
#-----#
library('ggplot2')
ggplot(data=TrainPointSet1[TrainPointSet1$SatelitePointID=='L99185
P38064',], aes(x=V6, y=SateliteMeasure, group=SatelitePointID)) +
geom_line()+labs(x = "Matavimas", y="Nusédimas, m")
ggplot(data=TrainPointSet1[TrainPointSet1$SatelitePointID=='L12399
5P26605',], aes(x=V6, y=SateliteMeasure, group=SatelitePointID)) +
geom_line()+labs(x = "Matavimas", y="Nusédimas, m")
ggplot(data=TrainPointSet1[TrainPointSet1$SatelitePointID=='L85520
P43565',], aes(x=V6, y=SateliteMeasure, group=SatelitePointID)) +
geom_line()+labs(x = "Matavimas", y="Nusédimas, m")
TrainPointSet2<-TrainPointSet1
TrainPointSet2<-
TrainPointSet2[!duplicated(TrainPointSet2[c(2,3)]),]
c2<-as.data.frame(table(unlist(TrainPointSet2$SatelitePointID)))
c3<-c2[c2$Freq>1,]
columData<-NULL
for (i in 1:nrow(c3)) {
c<-c3[i,1]
filtr<-TrainPointSet2[TrainPointSet2$SatelitePointID==c,4]

```



```

    set<- (filtr-mean(filtr))/sd(filtr)
    columData<-cbind(columData,set)
  }
  columData<-as.data.frame(columData)
  colnames(columData)<-colnames(c3$Var1)
  columData1<-as.data.frame(t(columData))
  columData1<-cbind(columData1,c3$Var1)
  classID2<-TrainPointSet1[,c(2,5)]
  classID2<-classID2[!duplicated(classID2),]
  columData1<-
merge(columData1,classID2,by.x="c3$Var1",by.y="SatelitePointID",all.x
=TRUE)

```

Artimiausio kaimyno metodo sudarymas taikant DTW atstumo vertinimo metodika.

```

DtwDist <- function(a, b){
  library("dtw")
  d <-proxy::dist(a, b, method="DTW")
  return(d)
}

knn_predict <- function(test_data, train_data, k_value){
  pred <- c() #empty pred vector
  #LOOP-1
  for(i in c(1:nrow(test_data))){
    Dist =c() #dist & char empty vector
    char = c()
    Lowest = 0
    Medium = 0
    Highest = 0
    #LOOP-2-looping over train data
    for(j in c(1:nrow(train_data))){
      Dist <- c(Dist, DtwDist(test_data[i,2:208],
train_data[j,2:208]))
      char <- c(char, as.character(train_data[j,209]))
    }

    eu <- data.frame(char, Dist)

```

```

eu <- eu[order(eu$Dist),]
eu <- eu[1:k_value,]

#Loop 3: loops over eu and counts classes of neighbors.
for(k in c(1:nrow(eu))){
  if(as.character(eu[k,"char"]) == "Lowest"){
    Lowest = Lowest + 1
  }else if(as.character(eu[k,"char"]) == "Medium"){
    Medium = Medium + 1
  }else{
    Highest = Highest + 1
  }
}
# Compares the no. of neighbors
if((Medium>Lowest)&(Medium>Highest)){
  pred <- c(pred, "Medium")
}else if((Lowest>Medium)&(Lowest>Highest)){
  pred <- c(pred, "Lowest")
} else{
  pred<-c(pred,"Highest")
}
}
return(pred) #return pred vector
}

matrixConfusion<-function(testdata){
  testdata[,210]<-as.factor(testdata[,210])
  re<-levels(testdata[,210])
  if (length(re)==2){
    if(length(re[re=="Lowest"])==0){
      levels(testdata[,210])<-c(levels(testdata[,210]),"Lowest")
    }
    if(length(re[re=="Medium"])==0){
      levels(testdata[,210])<-c(levels(testdata[,210]),"Medium")
    }
    if(length(re[re=="Highest"])==0){
      levels(testdata[,210])<-c(levels(testdata[,210]),"Highest")
    }
  }
}

```

```

    if (length(re)==1){
      levels(testdata[,210])<-
c(levels(testdata[,210]),"Highest","Lowest")
    }
    cm = as.matrix(table(Actual = testdata[,209], Predicted =
testdata[,210]))
    diag = diag(cm) # number of correctly classified instances per
class
    rowsums = apply(cm, 1, sum) # number of instances per class
    colsums = apply(cm, 2, sum) # number of predictions per class
    precision = diag / colsums
    recall = diag / rowsums
    f1 = 2 * precision * recall / (precision + recall)
    p<-data.frame(precision, recall, f1)
    return(p)
} # create the confusion matrix
train_size<-floor(0.8*nrow(columnData1))
train_ind<-sample(1:nrow(columnData1),size=train_size)
train<-columnData1[train_ind,]
test<-columnData1[-train_ind,]
r<-createFolds(train$RiskCodesUsedCurrenty,k=5)
split_up <- lapply(r, function(ind, dat) dat[ind,], dat = train)
tr<-c()
PrecRecF1<-NULL
library(foreach)
library(doSNOW)
cl <- makeCluster(4)
registerDoSNOW(cl)
foreach(i =c(1:5)) %dopar%{
  #Segement your data by fold using the which() function
  p<-c(1:5)
  j<-p[-i]
  testData <- split_up[[i]]
  trainData<-NULL
  for(s in j){
    trainData<-rbind(trainData,split_up[[s]])
  }
  predictions<- knn_predict(testData,trainData,K)
}

```

```

testData[,210]<-predictions
tr<-c(tr,accuracy(testData))
PrecRecF1<-rbind(PrecRecF1,metrixConfusion(testData))
}

```

```

predictions<- knn_predict(test,train,1)
test[,210]<-predictions
metrixConfusion(test)

```

Diskrečiosios bangelių transformacijos taikymas ir Fišerio skiriamumo santykio apskaičiavimas.

```

#Išskiriame DBT koeficientus (naudojantis Haar filter)
DwtTransform<-function(TrainPointData){
library(wavelets)
library(wavethresh)
c2<-as.data.frame(table(unlist(TrainPointData$SatelitePointID)))
c3<-c2[c2$Freq>1,]
wtData <- NULL
for (i in 1:nrow(c3)) {
  c<-c3[i,1]
  x<-vector(mode='numeric',length = 49)
  filtr<-TrainPointData[TrainPointData$SatelitePointID==c,4]
  a<-c(filtr,x)
  a <- ts(a)
  wt<-wd(a,filter.number=1,family="DaubExPhase")
  plot(wt)
  wtData <- rbind(wtData,c(unlist(as.character(c)),wt$D,accessC(wt,level=0)))
}
wtData <- as.data.frame(wtData)
tf<-TrainPointData[,c(2,5)]
tf<-tf[!duplicated(tf),]
for (i in 2:ncol(wtData)){
  wtData[,i]<-as.double(as.character(wtData[,i]))
}
wtData12<-merge(wtData,tf,by.x="V1",by.y="SatelitePointID",all.x
=TRUE)
return(wtData12)

```

```

}
#FDR calculation
FDR<-function(wtData) {
fd<-NULL
for(i in 2:(ncol(wtData)-1)) {
  r<-colnames(wtData)
  r1<-wtData[wtData$RiskCodesUsedCurrenty=="Highest",i]
  r2<-wtData[wtData$RiskCodesUsedCurrenty=="Lowest",i]
  r3<-wtData[wtData$RiskCodesUsedCurrenty=="Medium",i]
  meanr1<-mean(r1)
  meanr2<-mean(r2)
  meanr3<-mean(r3)
  sdr1<-var(r1)
  sdr2<-var(r2)
  sdr3<-var(r3)
  fd<-rbind(fd,c(r[i],(((meanr1-meanr2)^2/(sdr1+sdr2))+((meanr1-
meanr3)^2/(sdr1+sdr3)))+(((meanr2-meanr1)^2/(sdr2+sdr1))+((meanr2-
meanr3)^2/(sdr2+sdr3)))+(((meanr3-meanr1)^2/(sdr3+sdr1))+((meanr3-
meanr2)^2/(sdr3+sdr2))))))
}
fd<-as.data.frame(fd)
fd1<-fd[!is.na(fd$V2),]
fd1$V2<-as.double(as.character(fd1$V2))
#fd1<-fd1[fd1$V2>1,]
return(fd1)
}
WtDataCoff<-DwtTransform(TrainPointSet3)
FDRCoff<-FDR(WtDataCoff)
FDRCoff<-FDRCoff[!is.na(FDRCoff$V2),]
index<-as.data.frame(c(1:210))
FDRCoff<-cbind(FDRCoff,index)
colnames(FDRCoff)[3]<-"Index"
for(i in 1:nrow(FDRCoff)) {
  if (FDRCoff[i,2]>1) {
    FDRCoff[i,4]<-"Taip"
  }else{
    FDRCoff[i,4]<-"Ne"
  }
}

```

```

}
ggplot(data=FDRCoeff, aes(x=Index,
y=V2,color=V4))+labs(x="Koefficiento indeksas", y="Fisher skiriamumo
santykis", colour=">1")+geom_point()

```

SVM ir KNN modelių sudarymas naudojant atrinktus koeficientus.

```

FDRCoeff<-FDRCoeff[FDRCoeff$V2>1,]
filtr<-as.character(unlist(FDRCoeff$V1))
ClassificationData<-
WtDataCoff[,c("V1",filtr,"RiskCodesUsedCurrenty")]
train_size<-floor(0.8*nrow(ClassificationData))
train_ind<-sample(1:nrow(ClassificationData),size=train_size)
train<-ClassificationData[train_ind,]
test<-ClassificationData[-train_ind,]
r<-createFolds(train$RiskCodesUsedCurrenty,k=5)
split_up <- lapply(r, function(ind, dat) dat[ind,], dat = train)
PrecRecF1<-NULL
cosV=0.1
g=0.0005
library(e1071)
for(i in c(1:5)){
  #Segment your data by fold using the which() function
  p<-c(1:5)
  j<-p[-i]
  testData <- split_up[[i]]
  testData<-testData[,-1]
  trainData<-NULL
  for(s in j){
    trainData<-rbind(trainData,split_up[[s]])
  }
  tr<-trainData[,-1]
  svm.model <- svm(RiskCodesUsedCurrenty~ ., data= tr,cost=10,
kernel="linear")
  svm.pred <- predict(svm.model, testData)
  prediction<-NULL
  for(l in 1:nrow(testData)){
    prediction<-
rbind(prediction,as.character(unlist(svm.pred[[l]])))

```

```

    }
    testData<-cbind(testData,prediction)
    PrecRecF1<-rbind(PrecRecF1,metrixConfusion(testData))
  }
  print(PrecRecF1)
  tr<-train[,-1]
  test1<-test[,-1]
  svm.model <- svm(RiskCodesUsedCurrenty~ ., data= tr,cost =
300,gamma=0.0005,kernel="radial")
  svm.pred <- predict(svm.model, test1)
  prediction<-NULL
  for(l in 1:nrow(test)){
    prediction<-
rbind(prediction,as.character(unlist(svm.pred[[l]])))
  }
  test1<-cbind(test1,prediction)
  metrixConfusion(test1)
  #knn -----#
  euclideanDist <- function(a, b){
    d = 0
    for(i in c(1:(length(a)-1) ))
    {
      d = d + (a[[i]]-b[[i]])^2
    }
    d = sqrt(d)
    return(d)
  }
  knn_predict2 <- function(test_data, train_data, k_value){
    pred <- c() #empty pred vector
    #LOOP-1
    for(i in c(1:nrow(test_data))){ #looping over each record of
test data
      Dist =c()
      char = c()
      Lowest = 0
      Medium = 0
      Highest = 0
      #LOOP-2-looping over train data

```

```

for(j in c(1:nrow(train_data))){

    #adding euclidean distance b/w test data point and train data
to eu_dist vector
    Dist      <-      c(Dist,      euclideanDist(test_data[i,2:33],
train_data[j,2:33])
    char <- c(char, as.character(train_data[j,33]))
}

eu <- data.frame(char, Dist)
eu <- eu[order(eu$Dist),]
eu <- eu[1:k_value,]

#Loop 3: loops over eu and counts classes of neighbors.
for(k in c(1:nrow(eu))){
    if(as.character(eu[k,"char"]) == "Lowest"){
        Lowest = Lowest + 1
    }else if(as.character(eu[k,"char"]) == "Medium"){
        Medium = Medium + 1
    }else{
        Highest = Highest + 1
    }
}

# Compares the no. of neighbors
if((Medium>Lowest)&(Medium>Highest)){
    pred <- c(pred, "Medium")
}else if((Lowest>Medium)&(Lowest>Highest)){
    pred <- c(pred, "Lowest")
} else{
    pred<-c(pred,"Highest")
}

}
return(pred) #return pred vector
}
K =2
PrecRecF<-NULL

```



```

for(i in c(1:5)){
  #Segment your data by fold using the which() function
  p<-c(1:5)
  j<-p[-i]
  testData <- split_up[[i]]
  trainData<-NULL
  for(s in j){
    trainData<-rbind(trainData,split_up[[s]])
  }
  predictions<- knn_predict2(testData,trainData,K)
  testData[,34]<-predictions
  PrecRecF<-rbind(PrecRecF,metrixConfusion(testData))
}
predictions<- knn_predict2(test,train,2)
test[,34]<-predictions
cm = as.matrix(table(Actual = test[,33], Predicted = test[,34]))
metrixConfusion(test)

```