TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

**Kaarel Tamuri**

# Masinõppe meetodite võrdlus vardas esinevate pragude iseloomustamiseks

**Bakalaureusetöö (9 EAP)**

Juhendaja: Ljubov Jaanuska, PhD

Tartu 2025

## Masinõppe meetodite võrdlus vardas esinevate pragude iseloomustamiseks

**Lühikokkuvõte:**

Vardad on olulised konstruktsioonielemendid, kus pragude olemasolu võib vähendada terve süsteemi kandevõimet. Bakalaureusetöö hindab erinevate masinõppemeetodite sobivust vardas esinevate pragude iseloomustamiseks. Võrreldakse kolme masinõppe mudeli, lineaarne regressiooni, juhumetsa ja XGBoosti, võimet prao sügavuse ja asukoha ennustamiseks mitme mõõdiku alusel. Samuti soovitakse leida, millised sisendandmed sobivad kõige paremini mudelite treenimiseks. Selleks kasutatakse varda omasagedusi, Haari lainikute 16 koefitsienti ning Haari lainikute 32 koefitsienti. Andmetele lisatakse müra mõõtemääramatuse simuleerimiseks.

Tulemused näitasid, et sisendandmetena annavad parima tulemuse Haari lainikute koefitsiendid, eriti prao asukoha ennustamisel. Võrreldud mudelitest andis parima tulemuse XGBoost, saavutades Haari lainikute 32 koefitsiendiga andmestikul kombineeritud ennustusülesandes $R^2$ väärtuseks 0.896. Juhumets näitas samuti häid tulemusi, samas kui lineaarne regressioon oli täpsuselt selgelt tagasihoidlikum. Kokkuvõttes osutus XGBoost koos 32 Haari lainiku koefitsiendiga kõige tõhusamaks lahenduseks varda pragude iseloomustamiseks.

**Võtmesõnad:**

Masinõpe, varras, pragude tuvastamine, regressioon, omasagedus, Haari lainik

**CERCS:** P176 Tehisintellekt

## Comparison of Machine Learning Methods for Cracks Identification in Rods

**Abstract:**

Rods are common components in engineering, where the presence of cracks can significantly reduce load-bearing capacity and pose safety risks. This bachelor’s thesis evaluates the suitability of different machine learning methods for crack identification in rods. Three machine learning models, linear regression, random forest and XGBoost, are compared in their ability to predict crack depth and location using input features from three different

domains: natural frequencies, 16 Haar wavelet coefficients and 32 Haar wavelet coefficients. The models are assessed using standard regression metrics and noise is introduced to simulate real-world measurement uncertainty.

The results show that the Haar wavelet coefficients outperform natural frequencies, especially in predicting crack location. Among the models, XGBoost consistently delivers the highest accuracy, achieving $R^2$ up to 0.896 in the combined prediction task using the dataset of 32 Haar wavelet coefficients. Random forest also performs well, while linear regression provides fast but less accurate results. The study concludes that XGBoost trained on 32 Haar wavelet coefficients is the most effective approach for open crack identification in rods.

**Keywords:**

Machine learning, rod, crack identification, regression, natural frequency, Haar wavelet

**CERCS:** P176 Artificial intelligence

## Sisukord

# 1. Sissejuhatus

Vardad on olulised konstruktsioonielemendid, mis suudavad tõhusalt kanda erinevaid koormusi. Vardaid kasutatakse peaaegu kõikjal, näiteks hoonetes, masinates ja sildades. Kui vardasse tekib pragu, väheneb selle kandevõime. Selle tagajärjel väheneb ka kogu süsteemi kandevõime. Kandevõime langus võib põhjustada mitmeid rikkeid ning halvimal juhul õnnetusi. Seetõttu on vardas esinevate pragude tuvastamine kriitilise tähtsusega. Kuigi pragude tuvastamiseks on olemas erinevaid lahendusi, ei ole need alati piisavalt täpsed, ressursitõhusad või töökindlad.

Alternatiivina võib pragude tuvastamiseks kasutada masinõpet. Terve ja kahjustatud varda füüsikalised omadused, sealhulgas omasagedused ja moodi kujust tuletatud Haari lainikute koefitsiendid, on erinevad. Need erinevused on mõõdetavad ja masinõppe mudelites rakendatavad. Paraku ei ole olemas ühte universaalselt parimat masinõppe mudelit, mis suudaks igat ülesannet kõige paremini lahendada. Samuti ei ole teada, milline masinõppe mudel sobib antud ülesande lahendamiseks kõige paremini.

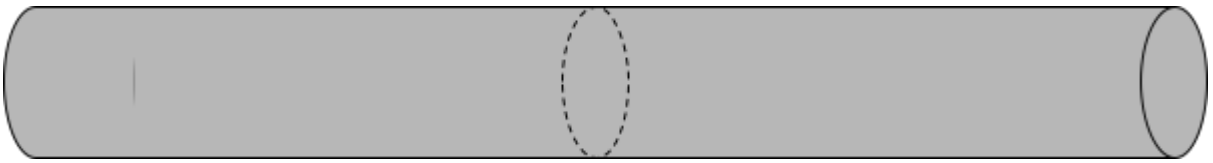
Selle töö eesmärk on hinnata erinevate masinõppe meetodite sobivust vardas esinevate pragude iseloomustamiseks. Soovime leida, milline mudel suudab kõige paremini ennustada prao sügavust varda sees ning selle asukohta piki varda telge. Samuti tahame teada, millised varda omadused sobivad kõige paremini sisendandmeteks mudelite treenimisel.

Töö käigus võrreldakse kolme masinõppe meetodit: lineaarset regressiooni, juhumetsa ja XGBoosti. Sisendandmetena kasutatakse varda omasagedusi ja Haari lainikute koefitsiente. Iga mudeli ja andmestiku puhul koostatakse kolm eraldi mudelit, mille eesmärk on identifitseerida prao sügavust, asukohta ning mõlemat korraga. Mudelite sooritust hinnatakse mitme mõõdiku alusel, et leida, milline kombinatsioon sobib kõige paremini antud ülesande lahendamiseks.

# 2. Teoreetiline taust

## 2.1. Varraste omadused ja nende tähtsus

Varras (ingl *rod*) on ruumiline keha, mille pikkus on oluliselt suurem selle ülejäänud kahest mõõtmest ehk laiusest ja kõrgusest (Laigna, 2000). Varda kuju iseloomustatakse selle ristlõike ja teljega. Varda telg on joon, mis läbib ristlõikepindade keskmeid, ja ristlõikepind on varda teljega ristuv tasandkujund. Vastavalt pikitelje kujule eristatakse sirget, kõverjoonelist ja murdjoonelist varrast. Ristlõikepinna järgi liigitatakse vardaid ühtlase ja muutuva ristlõikepinnaga varrasteks. Varras võib olla mistahes materjalist, näiteks metallist, puidust jne. Samuti ei ole varda pikkusele antud muid piiranguid. Sellest töös keskendutakse sirgetele ja ühtlase ristlõikepinnaga varrastele (vt joonis 1), sest need on üldlevinud ehituselemendid.



Joonis 1. Ühtlase ristlõikepinnaga sirge varras. Autori koostatud.

Vardad on ehituses ja inseneriteaduses tähtsad komponendid, kuna nad kannavad tõhusalt erinevaid koormuseid, sealhulgas surve- ja tõmbekoormusi (Klauson jt, 2012). Vardaid kasutatakse peaaegu kõikjal, olgu selleks hooned, masinad, rippsillad jne. Erinevad varda omadused, nagu jäikus ja materjal, määravad ära, kuidas varras koormuseid talub. Erinevates konstruktsioonides on oluline valida sobiv varda tüüp, kuna vale materjal või mõõdud võivad viia konstruktsiooni kahjustumiseni.

Vardas esinevad praod (ingl *cracks*) vähendavad konstruktsiooni kandevõimet või põhjustavad muid rikkeid (Jaanuska, 2021). Praod võivad tekkida erinevatel põhjustel, olgu selleks disaini- ja tootmisvead või keskkonnatingimused. Nende kiire tuvastamine ja hindamine on hädavajalik, et vältida edasiste kahjude tekkimist. Keerukates konstruktsioonides, nagu kõrghooned või sillad, võib varraste kahjustuste õigeaegne avastamine hoida ära tõsiseid õnnetusi. Seetõttu on pragu tuvastavate tehnoloogiate arendamine äärmiselt oluline.

## 2.2. Alternatiivsed lahendused pragude tuvastamiseks

Vardas esinevate pragude tuvastamiseks kasutatakse mitmeid meetodeid, millest igaühel on oma eelised ja puudused. Paljud meetodid võivad osutuda teatud tingimustel ebaefektiivseks, sõltudes rakendustingimustest, inspektorite kogemusest ja kasutatava tehnoloogia piirangutest. Sageli kombineeritakse erinevaid lahendusi, et parandada tulemuste täpsust ja usaldusväärsust. Järgnevalt antakse ülevaade enim levinud lahendustest.

Visuaalne kontroll on kahjustuste tuvastamise meetod, mille käigus hindab inspektor struktuuri seisundit palja silma või erinevate suurendusseadmete, näiteks luubi või mikroskoobiga (Ooijevaar, 2014). Meetodi tulemuslikkus sõltub suuresti inspektori oskustest ning kogemusest, muutes tulemused subjektiivseks ning raskesti võrreldavateks. Lisaks on visuaalne kontroll aeganõudev ning vaadeldavale objektile peab olema tagatud kerge ligipääs, mis ei pruugi keerukate süsteemide puhul olla võimalik. Samuti ei ole võimalik tuvastada sisemisi kahjustusi.

Röntgenkiirguse kasutamine on üks täpsemaid pragude tuvastamise meetodeid ning sellega on võimalik näha nii pindmisi kui ka sisemisi kahjustusi. Samas märgivad Peng jt (2022), et antud lahendusel on mitmeid puudusi, sealhulgas kõrge kulu ja ranged ohutusnõuded. Lisaks võivad tihedast materjalist objekti korral tekkida tulemuste moonutused, mis vähendavad meetodi usaldusväärsust.

Ultraheli meetod põhineb helilainete peegeldumisel ja murdumisel, kui need kohtavad objekti eripärasid, näiteks pragusid (Peng jt, 2022). Seda kasutatakse peamiselt sisemiste, kuid ka pindmiste pragude tuvastamiseks. Meetodi kasutamisel on siiski mitmeid piiranguid. Objekti pinna lähedal esinevate pragude nägemine on keeruline, sest tagasi peegeldub ka objekti piirjoon. Samuti on tähtis andurite õige asetus, mis võib teatud tingimustel olla keeruline, näiteks tulenevalt ruumipuudusest või inspektori oskustest. Lisaks on kvaliteetsete tulemuste jaoks hädavajalik, et helilained jõuaksid üldse objektini.

# 3. Masinõppepõhine lähenemine pragude tuvastamiseks

## 3.1. Masinõppe põhimõtted

Masinõpe on andmeteaduse valdkond, mille eesmärk on arendada mudeleid, mis suudavad andmete põhjal teha ennustusi ilma, et neid selleks otseselt programmeeritaks (Sügis jt, 2025). Paljusid ülesandeid, mida inimene suudab lahendada intuitiivselt, ei ole võimalik samm-sammulise loogikana kirjeldada. Kuna selliseid protsesse on keeruline programmeerida, võimaldab masinõpe luua lahendusi, mis õpivad andmetest, mitte inimeste koostatud reeglitest. See tähendab, et traditsioonilise reeglipõhise programmeerimise asemel esitatakse arvutile andmestik ning algoritm, mille abil luuakse probleemile sobiv lahendus ehk masinõppe mudel.

Masinõppe rakendamine eeldab sobiva algoritmi valikut, treeningandmestikku ja treeningprotsessi läbiviimist (Sügis jt, 2025). Tänapäevased tööriistad võimaldavad seda teha lihtsalt ja efektiivselt, algoritme ise nullist ehitamata. On oluline, et masinõpe suudaks üldistada andmetest saadud teadmisi ka varem nägemata sisenditele.

Masinõppe mudel on matemaatiliste tehete ja võrduste loetelu, mille abil on võimalik andmetest mustreid avastada (Sügis jt, 2025). Nende mustrite põhjal saab arvutada väljundi, mis sõltuvalt ülesandest võib olla näiteks arvuline prognoos või hinnang kindlasse klassi kuulumise kohta. Masinõppe mudelid erinevad üksteisest mustrite avastamise viisist ning väljundi tüübist.

Juhendatud masinõppe (ingl *supervised machine learning*) ülesanded jagunevaks peamiselt kaheks: klassifikatsioon (ingl *classification*) ja regressioon (ingl *regression*) (GeeksforGeeks, 2025). Klassifikatsioon määrab sisendandmete kuuluvust etteantud kategooriatesse. Regressioon keskendub pidevate väärtuste ennustamisele. Seega seisneb regressiooni ja klassifikatsiooni põhiline erinevus väljundi tüübis.

Regressioonimudelite eesmärk on leida seos sisendmuutujate ja väljundi vahel (IMSL, 2021). See võimaldab hinnata, kuidas muutused sõltumatutes muutujates mõjutavad ennustatavat väärtust. Selline seoste leidmine on oluline, sest see aitab andmetest tuletada üldistatavaid teadmisi. Lõpptulemuseks luuakse üldine reeglistik, mille abil saab prognoosida ka varem mitte nähtud väärtusi.

## 3.2. Omasagedused

Omasagedus (ingl *natural frequency*) on füüsikaline omadus, mis iseloomustab, millisel sagedusel keha pärast selle häirimist vibreerib väliste keskkonna takistuse puudumisel (Aroeira, 2024). Igal süsteemil võib olla mitu omasagedust, millel see loomulikult võngub, kusjuures iga sagedus vastab konkreetsele võnkemoodile (ingl *mood*). Koos annavad need võnkemoodid tervikliku ülevaate keha käitumisest.

Praktikas leitakse objekti omasagedused sageli eksperimentaalse modaalanalüüsi teel (Aroeira, 2024). Üheks võimaluseks on teha löögikatse, mille käigus lüüakse objekti kontrollitud jõuga ning seejärel selle vibratsioon mõõdetakse. Saadud vibratsioonisignaalist arvutatakse sagedusreaktsioonifunktsioon (ingl *frequency response function*), mille resonantstipud näitavad ära objekti omasagedused.

Objekti omasagedused on tundlikud selle seisundi ja vigastuste suhtes (Aroeira, 2024). Konstruktsiooni jäikuse vähenemine, näiteks prao tekkimisel, muudab selle omasagedusi. Seetõttu on omasageduste muutused märgiks, et objekt võib olla kahjustatud. Kuna muutused on mõõdetavad, on neid võimalik kasutada sisendandmetena masinõppe mudelites, mis suudavad hinnata pragude asukohta ja sügavust.

## 3.3. Haari lainikud

Haari lainik (ingl *Haar wavelet*) on lihtne matemaatiline funktsioon, mida kasutatakse signaalitöötluses signaalide analüüsimiseks ja tihendamiseks (Brainforge). See põhineb väikestel lainekujulistel funktsioonidel, millega saab uurida signaali erinevaid detaile ja sageduskomponente. Haari lainik jagab signaali väiksemateks osadeks ehk alasagedusteks ning arvutab iga osa kohta koefitsiendid. Tulemuseks on kompaktne esitus, mis säilitab olulise info, võimaldades signaali tõhusalt tihendada ilma olulist kvaliteeti kaotamata.

Haari lainikute koefitsiendid on tundlikud võnkevormide lokaalsete muutuste suhtes (Jaanuska, 2021). Muutused, mis on põhjustatud näiteks prao tekkimisest, mõjutavad objekti võnkemoodi (ingl *mode shape*) ning vastavaid Haari lainikute koefitsiente. Seetõttu saab neid masinõppe mudelites sisendandmetena kasutada.

## 3.4. Masinõppe rakendamine pragude tuvastamisel

Masinõppe mudelid on saanud oluliseks tööriistaks pragude tuvastamiseks. Informatiivsete sisendandmestikena kasutatakse elemendi omasagedusi ja Haari lainikute koefitsiente (Jaanuska, 2019). Omasagedused ja esimesest võnkemoodist tuletatud Haari lainikute koefitsiendid sõltuvad objekti ehitusest, näiteks jäikusest ning kujust. Kui objektis esineb pragu, siis muutub selle jäikus ning seeläbi ka süsteemi omasagedused ja Haari lainikute koefitsiendid. Muutus on mõõdetav ning masinõppemeetodite abil analüüsitav. Tulemusena on võimalik identifitseerida pragude asukohta ning nende sügavust.

Pragude tuvastamisel masinõppega on mitmeid eeliseid. Võrreldes teiste meetoditega, on antud lahendus võrdlemisi kiire ja odav (Hernández-Díaze jt, 2024). Lisaks annavad hästi treenitud mudelid usaldusväärseid tulemusi. Hernández-Díaze jt töö näitel on võimalik saavutada väga kõrge determinatsioonikordaja ($R^2$) ja madal ruutkeskmise vea (*RMSE*), mis viitab mudeli kõrgele usaldusväärsusele.

Omasagedust ja Haari lainikute koefitsiente kasutatakse tihti pragude sügavuse ja asukoha määramiseks. Jaanuska ja Hein (2019) uurisid erinevaid masinõppe mudeleid pragude lokaliseerimiseks, sealhulgas tagasilevi tehisnärvivõrke (ingl *backpropagation neural network*) ja juhumetsi. Töös kasutati kahte andmekogumit: esimeses oli aluseks objekti omasageduse kaheksa esimest parameetrit ja teises Haari lainikute koefitsiendid. Tulemused näitasid, et tehisnärvivõrkude mudelid andsid natuke väiksema keskmise ruutvea (*MSE*) kui juhumetsade mudelid. Samuti leiti, et omasageduste kasutamine annab täpsema ülevaate pragude sügavusest. Lisaks selgus, et pragude asukoha määramine on oluliselt lihtsam ülesanne kui nende sügavuse hindamine.

Jaanuska ja Hein (2022) avastasid Pasternaki alusel paiknevate talade ehk elastsel pinnal paiknevatel talade pragude uurimisel, et juhumetsad võivad teatavatel tingimustel olla efektiivsemad, kuna nende hüperparameetrite seadistamine on lihtsam, treeningprotsess kiirem ning täpsus veidi parem. Kinnitati ka varasemat järeldust, et pragude sügavuse tuvastamine on keerulisem kui nende asukoha määramine ning sügavuse hindamisel on omasageduste kasutamine tõhusam kui Haari lainikute koefitsientide rakendamine.

# 4. Kasutatavad masinõppe mudelid

Erinevate masinõppe mudelite võrdlemine on lõputöö raames oluline, et selgitada, milline meetod sobib kõige paremini varraste pragude iseloomustamiseks. Igal masinõppe mudelil on oma eelised ja puudused, mistõttu puudub universaalselt parim mudel kõikvõimalike ülesannete jaoks. Pragude asukoha ja sügavuse hindamisel võivad erinevad algoritmid anda erinevaid tulemusi.

## 4.1. Lineaarne regressioon

Lineaarne regressioon on üks vanimaid ja lihtsamaid masinõppe meetodeid, mida kasutatakse pidevate väärtuste ennustamiseks (IBM, 2021). Mudel eeldab, et ennustatav väärtus on sisendandmetega lineaarses ehk sirgjoonelises seoses. See tähendab, et kui üks sisendtunnus suureneb, muutub ka väljund ennustatav viisil, kas suurenedes või vähenedes.

Lineaarne regressioon üritab leida sirget, mis kirjeldab võimalikult täpselt, kuidas sisendtunnused väljundit mõjutavad (IBM, 2021). Treenimise käigus otsib mudel tunnustele kaale ja nihet, mis minimeeriks ennustusvigu. Kui mudel on treenitud, saab seda kasutada uute väärtuste ennustamiseks.

Lineaarse regressiooni suurim eelis on selle lihtsus ja kiirus (IBM, 2021). See on hästi arusaadav ning sobib hästi juhtudeks, kus seos andmete vahel on lineaarne. Samuti võimaldab see mõista, kui palju iga sisendtunnus väljundit mõjutab. Kui seos andmete vahel on keerulisem või mittelineaarne, ei suuda lineaarne regressioon seda hästi kirjeldada. Samuti võib tulemus olla ebatäpne kui andmetes on palju müra.

Lineaarne regressioon valiti võrdlusesse baasvõrdlusenda. Selle abil saab hinnata, mil määral on pragude sügavuse ja asukoha seos sisendandmetega lineaarne ja kui hästi lihtne mudel antud andmetel toime tuleb. Lineaarse regressiooni tulemused annavad lähtetaseme, millega võrrelda keerukamaid mudeleid. Kui lineaarne mudel suudab saavutada kõrge täpsuse, viitab see sellele, et seosed on valdavalt lineaarsed ning keerukamatel meetoditel võib olla madal lisaväärtus.

## 4.2. Juhumets

Juhumets (ingl *random forest*) on masinõppe mudel, mis ühendab mitme otsustuspuu tulemused (IBM). See kuulub ansambelmeetodite (ingl *ensemble method*) hulka, kus mitu

mudelit töötavad koos, et parandada üldist ennustuse tulemust. Juhumetsade puhul luuakse iga puu jaoks juhuslik andmevalim ning igas otsustussõlmes kaalutakse juhuslikku tunnuste alamkomplekti. Selline lähenemine vähendab puude omavahelist korrelatsiooni ja aitab vältida üleõppimist, mis on üksikute otsustuspuude jaoks probleemne.

Juhumetsal on kolm peamist hüperparameetrit (IBM), milleks on sõlme suurus, puude arv ja tunnuste arv. Iga puu treenitakse erineva andmevalimi peal, kasutades nn *bootstrap* meetodit, kus andmepunkte valitakse juhusliku asendamisega. Lisaks valitakse ka tunnused juhuslikult, et tagada puude mitmekesisus. Lõpuks kombineeritakse kõigi puude ennustused, kus regressiooniülesannetes arvutatakse nende keskmine.

Juhumetsal on mitmeid eeliseid (IBM). Juhumets hõlmab mittelineaarset õppimisvõimet, mis on vajalik, kui seosed osutuvad keerukaks. Tänu puude kombineerimisele vähendatakse üleõppimise riski. Samuti võimaldab juhumets mõista, kui palju iga sisendtunnus väljundit mõjutab. Siiski esineb mudelil ka puudusi. Mudeli treenimine võib suure andmehulga korral olla aeganõudev. Lisaks vajab see rohkem arvutusressursse kui lihtsamad mudelid.

Juhumets valiti võrdlusesse, kuna see on tõhus ja tõestatud meetod konstruktsioonide kahjustuste tuvastamiseks. Näiteks kasutasid Jaanuska ja Hein (2019) juhumetsi talades pragude lokaliseerimiseks ja leidsid, et see andis häid tulemusi. Seega annab see töös võrdlusmomendi varasemate töödega. Lisaks on mudel praktikas suhteliselt stabiilne ja üleõppimiskindel ka mõõtemüra korral. Juhumetsal põhinev analüüs vastab küsimustele, kas mudel suudab kirjeldada pragude sügavust ja asukohta paremini kui lihtne lineaarne regressioon ning kui palju jääb puudu kaasaegsemast lähenemisest.

## 4.3. XGBoost

XGBoost on masinõppe algoritm, mis põhineb gradiendi võimendamise (ingl *gradient boosting*) meetodil (IBM, 2024). Selle eesmärk on suurendada mudeli täpsust, ühendades järjestikku mitmeid otsustuspuudel põhinevaid nõrgemaid mudeleid.

Erinevalt juhumetsast, kus puud treenitakse paralleelselt, treenib XGBoost puud järjestikku (IBM, 2024). Iga uus puu keskendub sellele, mida eelnevad puud valesti ennustasid. Mudel suudab käsitleda ka puudulikke andmeid. Kui andmestikus esineb tühje väärtusi, suudab XGBoost ise õppida, milline suund puus toob parima tulemuse. Lisaks toetab XGBoost hajutatud ja paralleelset arvutust, võimaldades kasutust väga suurte andmemahtude korral.

Samuti sisaldab XGBoost sisseehitatud karistusfunktsioone (ingl *penalty function*), mis aitab vältida üleõppimist ning toetab paremat üldistamist uute andmete puhul.

XGBoosti eelised tulenevad selle kõrgest täpsusest, efektiivsusest ja võimest õppida keerulisi andmemustreid (IBM, 2024). Samas on XGBoost lihtsamatest mudelitest, nagu lineaarne regressioon või isegi juhumets, märgatavalt keerukam. Selle häälestamine nõuab rohkem teadmisi ja arvutusressursse. Samuti võib mudeli sisemine loogika osutuda keeruliselt tõlgendatavaks.

XGBoost on käesoleva töö võrdluses kõige arenenum mudel. XGBoosti kaasamine võimaldab kontrollida, kas keerukam võimendamise meetod suudab seosed täpsemalt modelleerida kui juhumets või lineaarne regressioon. Meetodit ei ole pragude tuvastamisel teadaolevalt varem kasutatud, seega annab XGBoosti kasutamine võimaluse demonstreerida uuemate masinõppetehnikate rakendamist inseneerias. Kokkuvõttes toob XGBoost esile, kas maksimaalsete tulemuste nimel tasub panustada keerukamasse mudelisse või on lihtsamad alternatiivid pragude tuvastamisel praktilisemad.

# 5. Mudelite hindamiskriteeriumid

Masinõppe mudelite hindamine on oluline, et mõista nende sobivust ülesande lahendamiseks. Erinevad hindamiskriteeriumid aitavad analüüsida mudeli sooritust eri aspektidest. Masinõppega pragude asukoha ja sügavuse ennustamine on regressiooniülesanne, seetõttu peame kasutama selle jaoks sobivaid mõõdikuid.

## 5.1. Keskmine absoluutviga

Keskmine absoluutviga (ingl *mean absolute error, MAE*) mõõdab mudeli ennustuste ja tegelike väärtuste keskmist erinevust (Agrawal, 2025). Keskmine absoluutviga arvutatakse valemiga

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - y_i\right|,$$

kus $n$ on andmepunktide hulk, $x_i$ andmepunkti $i$ tegelik väärtus ja $y_i$ andmepunkti $i$ ennustatud väärtus.

*MAE* eeliseks on selle väljund samas ühikus kui ennustatav muutuja (Agrawal, 2025). Lisaks ei mõjuta üksikud väga valed ennustused selle väärtust nii palju kui teised mõõdikud, mistõttu sobib *MAE* hästi olukordadesse, kus andmetes võib esineda äärmuslikke väärtusi.

## 5.2. Keskmine ruutviga

Keskmine ruutviga (ingl *mean squared error, MSE*) on üks enim kasutatud hindamiskriteeriume (Agrawal, 2025). Ehituselt sarnaneb see keskmise absoluutveaga, kuid erinevusena leiab see ennustuste ja tegelike väärtuste ruudu. Keskmine ruutviga arvutatakse valemiga

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2,$$

kus $n$ on andmepunktide hulk, $x_i$ andmepunkti $i$ tegelik väärtus ja $y_i$ andmepunkti $i$ ennustatud väärtus.

Siiski tekitab ruudu võtmine ennustuste ja väljundi vahel ühikute erinevuse (Agrawal, 2025). Samuti karistab see andmetes olevaid äärmuslikke väärtusi. Selle tulemusel hindab ta rohkem andmeid kui mudelit ennast.

## 5.3. Ruutkeskmine viga

Ruutkeskmine viga (ingl *root mean squared error, RMSE*) sarnaneb ehituselt omakorda *MSE*-le, sest tegemist on selle ruutjuurega (Agrawal, 2025). Ruutkeskmine viga arvutatakse valemiga

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2},$$

kus $n$ on andmepunktide hulk, $x_i$ andmepunkti $i$ tegelik väärtus ja $y_i$ andmepunkti $i$ ennustatud väärtus.

*RMSE* parandab *MSE* ühikute mittekattuvuse probleemi, sest sisendandmete ühikud kattuvad väljundiga (Agrawal, 2025). Siiski karistab ka tema andmetes esinevaid äärmuslikke väärtusi.

## 5.4. Determinatsioonikordaja

Determinatsioonikordaja, masinõppes tuntud ka kui $R^2$, näitab, kui suure osa ennustatava muutuja dispersioonist suudab regressioonimudel selgitada võrreldes lihtsa keskmise ennustusega (Agrawal, 2025). $R^2$ on dimensioonitu ning kontekstist sõltumatu mõõdik, mis seab null-mudeliks alati keskmise ennustuse. $R^2$ arvutatakse valemiga

$$R^2 = 1 - \frac{SSR}{SST},$$

$$SSR = \sum_{i=1}^{n}\left(x_i - y_i\right)^2,$$

$$SST = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2,$$

kus $n$ on andmepunktide arv, $x_i$ andmepunkti $i$ tegelik väärtus, $\bar{x}$ tegelike väärtuste aritmeetiline keskmine ja $y_i$ andmepunkti $i$ ennustatud väärtus.

$R^2$ eelisteks on võrreldavus, selge protsendiline tõlgendus ning sõltumatus ühikutest ja andmete skaalast (Agrawal, 2025). Puudusena võib see ületreenitud mudelite korral anda liiga positiivseid tulemusi ning samuti puudub info parameetrite statistilise tähenduse kohta.

# 6. Mudelite rakendamine ja võrdlus

## 6.1. Andmestike kirjeldus

Kuna eksperimentaalsete andmete kogumine on tehniliselt keeruline, kallis ja aeganõudev, otsustati kasutada arvutuslikul meetodil saadud andmeid, mis põhinevad varda modaalanalüüsil ning juhendaja varasematel töödel[1].

Kõik kolm andmestikku kirjeldavad sama tüüpi konstruktsiooni: sirget, ühtlase ristlõikega varrast, mille üks ots on jäigalt kinnitatud ja teine ots on vaba. Iga mõõtepunkt vastab konkreetsele prao asukohale ja sügavusele. Andmestikud erinevad üksteisest sisendparameetrite tüübi ning andmete hulga poolest. Esimene andmestik sisaldab varda omasagedusi ning teised kaks Haari lainikute koefitsiente.

Kõik andmestikud koosnevad 5525 kirjest (vt tabel 1). Andmed on täielikud ehk neis ei esine puudulikke ega korduvaid väärtusi. Mõõtepunktid katavad ühtlaselt pragude asukohti ja sügavusi, mis tagab andmete esinduslikkuse.

Tabel 1. Andmestike põhiomadused.

| Andmestik | Kirjete arv | Atribuutide arv | Puuduvad väärtused | Korduvad kirjed |
|---|---|---|---|---|
| Omasagedused | 5525 | 12 | 0 | 0 |
| Haari lainikute 16 koefitsienti | 5525 | 19 | 0 | 0 |
| Haari lainikute 32 koefitsienti | 5525 | 35 | 0 | 0 |

### 6.1.1. Omasageduste andmestik

Esimene andmestik koosneb varda esimese kümne omasageduse väärtustest ning neile vastava prao sügavusest ja asukohast. Omasagedused iseloomustavad objekti loomulikke võnkesagedusi ning muutuvad juhul, kui konstruktsioonis esineb vigastus, näiteks pragu. Prao asukoht ja sügavus on skaleeritud, et tagada võrreldavus ja toetada stabiilset treenimist.

[1] Andmed on kättesaadavad Kaggle platvormil: https://www.kaggle.com/datasets/ljubovjaanuska/open-crack-in-longitudinal-vibrations-of-rods

Kõik omasagedused on esitatud kuue kümnendkoha täpsusega ning nende väärtused jäävad vahemikku 1.42 kuni 29.85. Näitajate keskmised väärtused suurenevad sageduse järjestuses. Standardhälbed näitavad, et kõrgemate võnkesageduste hajuvus on suurem, mis võib viidata, et need sagedused on tundlikumad konstruktsiooni muutustele. Tegu on kompaktse ja mõõdukalt informatiivse andmestikuga, mille abil on võimalik uurida, kuivõrd hästi saab nende abil iseloomustada varraste kahjustusi.

### 6.1.2. Haari lainikute 16 koefitsiendi andmestik

Teine andmestik koosneb 16 Haari lainikute konfitsientidest, esimesest omasagedusest ning neile vastava prao sügavusest ja asukohast. Koefitsiendid on tuletatud esimesest võnkemoodist. Asukoht ja sügavus on ka selles andmestikus skaleeritud samasse vahemikku nagu eelnevas andmestikus.

Kõik koefitsiendid ja lisatud esimene omasagedus on esitatud kuue kümnendkoha täpsusega. Koefitsientide väärtused sisaldavad nii positiivseid kui negatiivseid arve ning hajuvus erinevate tunnuste lõikes varieerub. Jällegi on tegemist on mõõdukalt detailse andmestikuga, mille abil on võimalik uurida, kuivõrd hästi saab iseloomustada varraste kahjustusi.

### 6.1.3. Haari lainikute 32 koefitsiendi andmestik

Kolmas andmestik koosneb 32 Haari lainikute koefitsiendist, esimesest omasagedusest ning neile vastava prao sügavusest ja asukohast. Koefitsiendid on arvutatud esimese võnkemoodi põhjal. Võrreldes eelmise andmestikuga võimaldab suurem koefitsientide arv kirjeldada võnkemoodi detailsemalt. See tähendab, et muutused signaali kujus, mis on põhjustatud erineva asukoha ja sügavusega pragudest, on paremini esile tõstetud. Selle tulemusel võib mudelil olla suurem võime eristada näiliselt sarnaseid, kuid siiski erinevad juhtumeid. Asukoht ja sügavus on ka siin skaleeritud samasse vahemikku nagu eelnevates andmestikes.

Kõik väärtused on esitatud kuue kümnendkoha täpsusega. Suurem tunnuste hulk suurendab mudeli käsutuses olevat infot, kuid samas eeldab ka hoolikamat häälestamist ja kontrolli, et vältida üleõppimist. Andmestikus on kokku 35 tunnust ning see on sobilik juhtudeks, kus soovitakse saavutada võimalikult täpseid ennustusi ja ollakse valmis suurema arvutusliku keerukuse arvestamiseks.

## 6.2. Katse ülesehitus ja mudelite treenimine

Selles peatükis kirjeldatakse, kuidas viidi läbi praktiline katse, mille eesmärk oli hinnata lineaarse regressiooni, juhumetsa ja XGBoosti sobivust varrastes pragude asukoha ja sügavuse ennustamisel. Katse ülesehitus oli võimalikult sarnane kõigi mudelite ja andmestike lõikes, et tagada tulemuste võrreldavus ja usaldusväärsus.

Kogu töö viidi läbi Google Colab veebikeskonnas, mis on laialdaselt levinud platvorm andmeteaduse ja masinõppe ülesannete lahendamiseks. Keskkonna valik oli tingitud selle lihtsast kasutatavusest, ligipääsust arvutusressurssidele ning vajalike töövahendite olemasolust. Töö käigus kasutati mitmeid levinud teeke, sealhulgas andmetabelite käsitlemiseks ja analüüsiks, mudelite koostamiseks ning tulemuste hindamiseks ja visualiseerimiseks. Kõik arvutused tehti samas keskkonnas, mis tagas sujuva töövoo ning võimaldas andmestikke käsitleda ühtsel viisil. See lõi eeldused, et kõiki mudeleid sai võrrelda samadel alustel ja iga tööetapp oli kergesti korratav. Töö lõpptulemus on Jupyter Notebook fail, mis on kättesaadav GitHubis[2].

### 6.2.1. Andmete eeltöötlus

Kõik kolm andmestikku muudeti *pandas* DataFrame andmestruktuurideks. Igale reale määrati vastavad veerunimed, et tagada nende korrektne käsitlus. Esmase kontrolli käigus veenduti, et andmestikes ei esineks puudulikke ega korduvaid kirjed. Selliseid kirjeid andmestikus ei leitud.

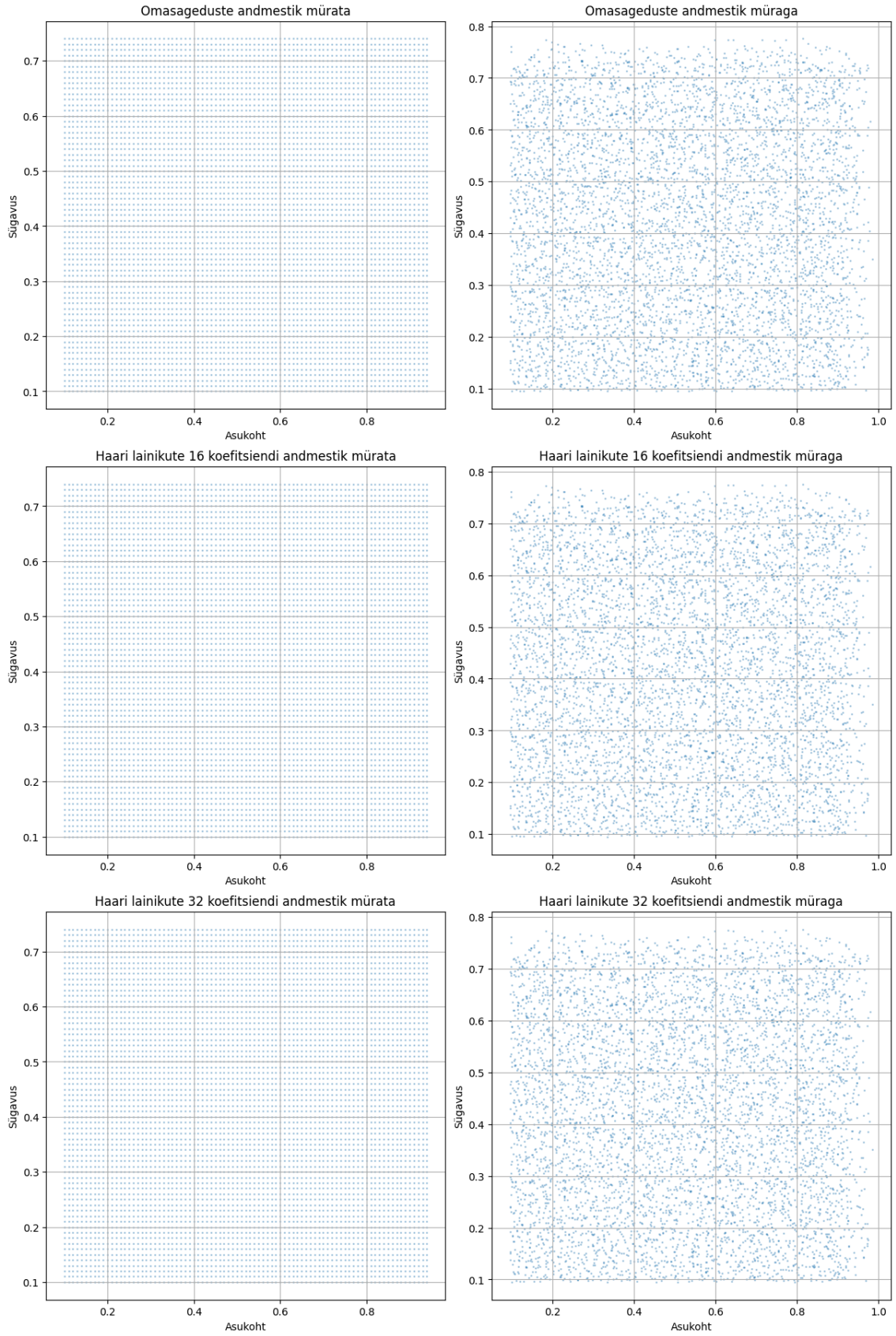
Prao asukoha ja sügavuse omavahelist jaotumist hinnati, et veenduda andmete piisavas hajutatuses. Eesmärk oli kontrollida, et andmed ei oleks kallutatud kitsale piirkonnale. Kuna väärtused katavad ühtlaselt kogu asukoha ja sügavuse skaala, ei olnud vajadust täiendavaks tasakaalustamiseks.

Et uurida mudelite tundlikkust reaalsetes oludes esineva mõõtemääramatuse suhtes, lisati igale andmestikule juhuslik müra (vt joonis 2). Müra lisati nii sisendtunnustele kui ka ennustatavatele muutujatele. Selle jaoks korrutati iga väärtust juhuslikult valitud teguriga vahemikus 0.95 kuni 1.05. Müra lisamise eesmärk oli simuleerida mõõtemääramatust, mis võib andmekogumisel tekkida. See võimaldas hinnata mudelite sobivust mittetäiuslike andmete korral.

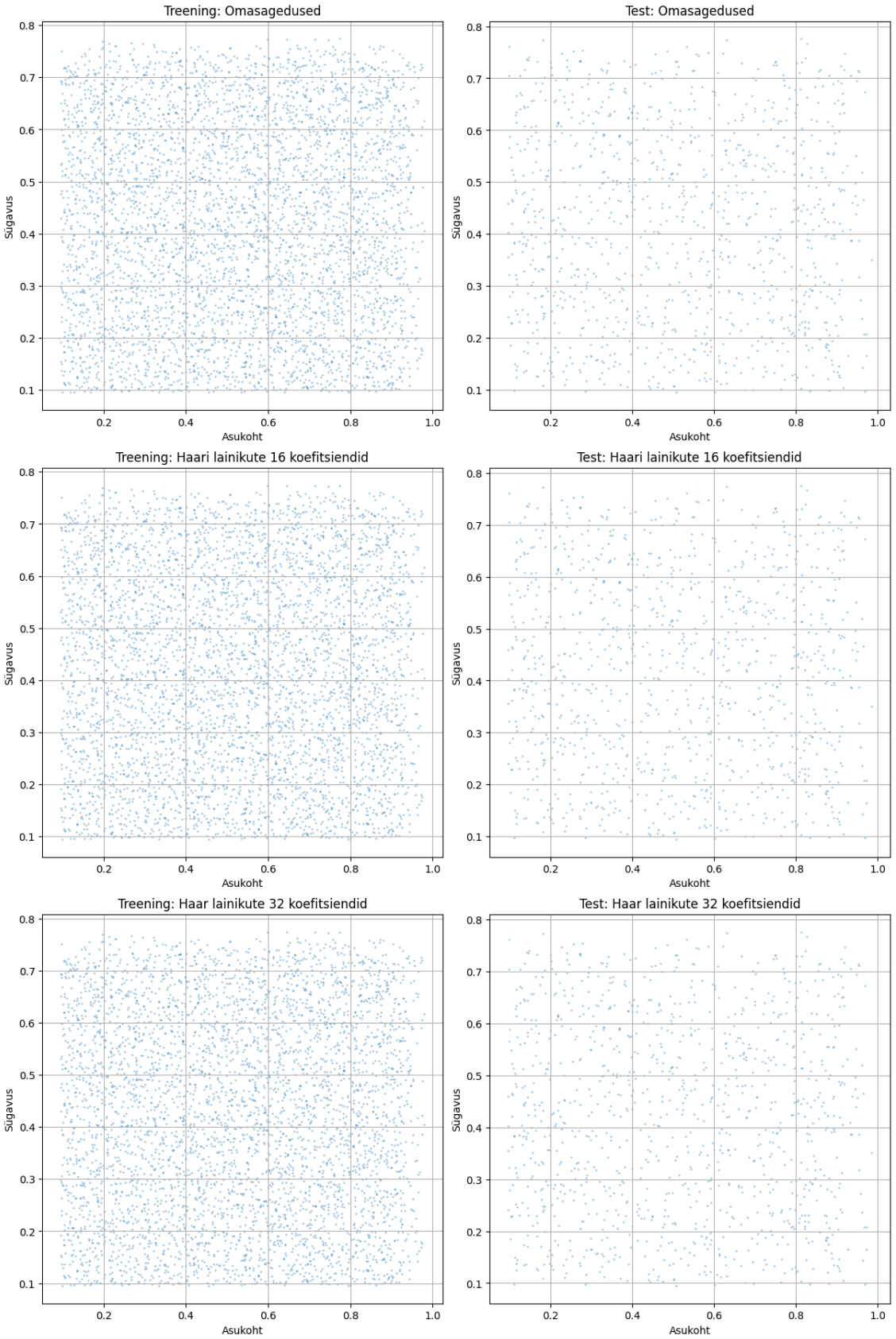
[2] Töö on kättesaadav GitHubis: https://github.com/KaarelTamuri/Vardas_esinevate_pragude_tuvastamine

Seejärel jaotati iga andmestik treening- ja testandmestikeks. Andmed jaotati treening- ja testandmestikeks selliselt, et 80 protsenti läks mudelite treenimiseks ning 20 protsenti testimiseks. Jaotamine tehti kolme eesmärgi jaoks: eraldi prao asukoha ennustamiseks, eraldi sügavuse ennustamiseks ning mõlema muutuja samaaegseks ennustamiseks. Andmete jaotust kontrolliti visuaalselt, et vältida olulisi erinevusi treening- ja testandmete vahel. Joonisel 3 on näitena toodud asukoha ja sügavuse jaotus treening- ja testandmestikes sügavuse ennustamiseks.

Kõik eeltöödeldud andmestikud olid seejärel valmis regressioonimudelite treenimiseks, kus neid rakendati ühtse metoodika alusel kõigi ennustusülesannete puhul.




Joonis 2. Andmestikes sügavuse ja asukoha muutus müra lisamisel.




Joonis 3. Treening- ja testandmete jaotus sügavuse ennustamiseks.

### 6.2.2. Mudelite häälestus ja treenimine

Töös keskenduti kolmele erinevale masinõppe mudelile: lineaarne regressioon, juhumets ja XGBoost. Lineaarse regressiooni ja juhumetsa realiseerimiseks kasutati *scikit-learni* klasse LinearRegression ja RandomForestRegressor. XGBoosti realiseerimiseks kasutati teegi *xgboost* klassi XGBRegressor. Parimate tulemuste saavutamiseks tehti iga andmestiku ja ülesande puhul mudelitele süsteemne hüperparameetrite häälestamine. Mudelite konfiguratsioonid ja hüperparameetrite valim on esitatud lisas 1.

Esmalt määratleti iga mudeli jaoks hüperparameetrite valim. Lineaarse regressiooni puhul uuriti kahe parameetri, *fit_intercept* ja *positive,* mõju. Juhumetsa puhul häälestati kolm peamist parameetrit: *n_estimators*, *max_depth* ja *max_features*. XGBoostil häälestati järgmised parameetrid: *n_estimators*, *max_depth* ja *learning_rate*. Mudelite hüperparameetrite häälestamiseks kasutati *scikit-learni* klassi GridSearchCV. Iga parameetrite kombinatsiooni hindamiseks kasutati viiekordset ristvalideerimist, mis aitas parandada mudelite üldistusvõimet.

Erilist tähelepanu pöörati juhtumitele, kus väljund koosnes kahest komponendist ehk asukohast ja sügavusest korraga. Kuna GridSearchCV ei toeta mitme väljundiga mudelite otsest treenimist, tehti hüperparameetrite häälestus mõlemale osale eraldi. Seejärel kasutati leitud parameetreid mitme väljundiga mudelis. Selline lähenemine tagas, et ka kahemõõtmelised ennustusülesanded lähtusid samast häälestusloogikast nagu ühe väljundiga mudelid.

Kokkuvõttes võimaldas mudelite süsteemne häälestamine kasutada iga mudeli maksimaalset potentsiaali. Kuna andmestikud sisaldasid ka lisatud müra, oli eriti oluline valida hüperparameetrid, mis võimaldavad head üldistusvõimet. Häälestusprotsessis salvestati parimate parameetrite kombinatsioonid, et neid saaks tulevikus taas kasutada.

### 6.2.3. Tulemuste mõõtmine ja salvestamine

Pärast mudelite treenimist hinnati nende sooritust regressiooni mõõdikutega. Iga mudeli, andmestiku ja ennustusülesande korral mõõdeti neli põhinäitajat: keskmine absoluutviga, keskmine ruutviga, ruutkeskmine viga ja determinatsioonikordaja. Kui hinnati nii prao asukohta kui sügavust korraga, arvutati mõõdikud mõlema kohta eraldi ning võeti nende

keskmine väärtus. Lisaks mõõdeti mudeli testandmestiku ennustamise aeg, et hinnata nende ajakulu.

Tulemused salvestati struktureeritud tabeli kujul, kus iga rida vastab konkreetsele kombinatsioonile: mudel, andmestik, ennustatav muutuja. Selle kaudu oli võimalik hiljem tulemusi süstemaatiliselt võrrelda ja visualiseerida. Täielik tulemuste tabel on esitatud lisas 2. Samuti salvestati iga treenitud mudeli parimad hüperparameetrid ning valminud mudelid ise, et neid oleks võimalik tulevikus ilma täiendava treeninguta uuesti kasutada ja võrrelda.

## 6.3. Tulemused ja võrdlus

### 6.3.1. Tulemused omasageduste andmestikuga

Tulemused näitavad, et sügavuse ennustamine oli edukam kui asukoha määramine, mis kehtis kõigi mudelite puhul. Seda kinnitavad nii $R^2$ väärtused kui ka teised headuse mõõdikud (vt joonis 4 ja joonis 5)

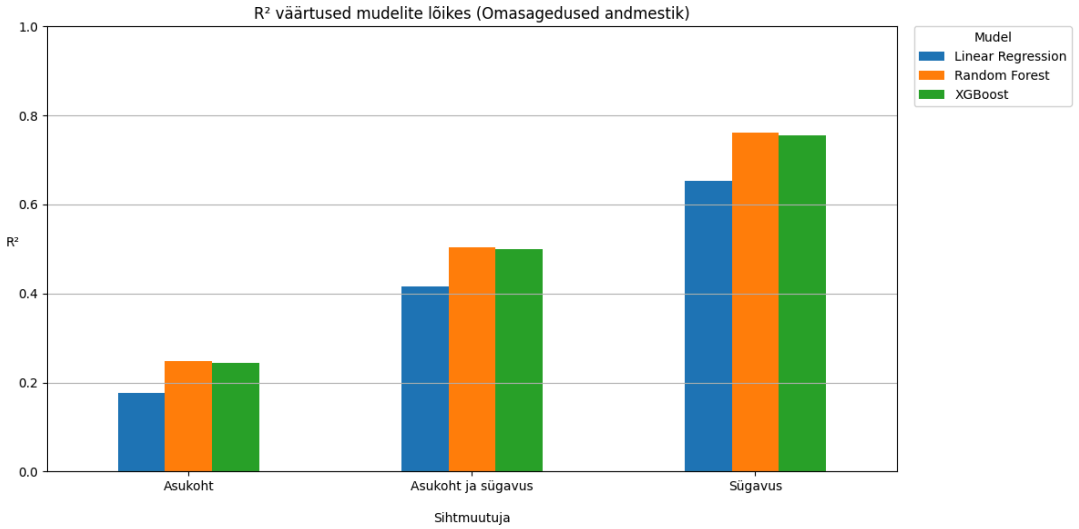
Lineaarne regressioon saavutas sügavuse ennustamisel *MAE* 0.087, *MSE* 0.012 ja *RMSE* 0.108. Samal ajal oli $R^2$ 0.654, mis tähendab, et mudel suutis selgitada enam kui poole sügavuse varieeruvusest. Asukoha ennustamine oli vähem edukas: *MAE* 0.192, *MSE* 0.052, *RMSE* 0.228 ja $R^2$ 0.178. Kombineeritud ülesandes, kus mõõdeti asukohta ja sügavust korraga, oli *MAE* 0.140, *MSE* 0.032, *RMSE* 0.179 ja $R^2$ 0.416. Ennustusaja poolest oli lineaarne regressioon siiski väga tõhus, vajades testandmestiku ennustamiseks vaid 0.002 kuni 0.014 sekundit. Siiski jääb selle mudeli üldine kasutuskõlblikkus piiratuks.

Juhumets parandas tulemusi kõikides mõõdikutes. Sügavuse puhul langes *MAE* 0.072-le, *MSE* 0.008-le, *RMSE* 0.090-le ja $R^2$ tõusis 0.761 juurde. Asukoha ennustuse puhul olid vastavad väärtused: *MAE* 0.182, *MSE* 0.0477, *RMSE* 0.219 ja $R^2$ 0.248. Kombineeritud ennustamisel saavutati *MAE* 0.127, *MSE* 0.028, *RMSE* 0.167 ja $R^2$ 0.505. Võrreldes lineaarse regressiooniga oli ennustusaeg siiski pikem, ulatudes 0.045 kuni 0.176 sekundini. Tulemuste põhjal on selge, et juhumets suudab leida keerukamaid seoseid ja annab usaldusväärsema tulemuse.

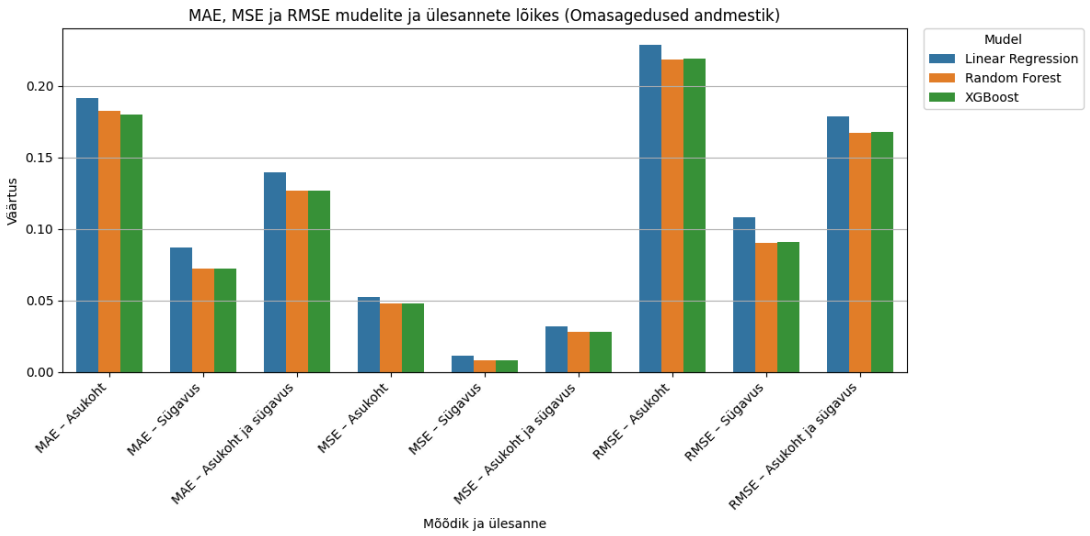
XGBoost saavutas juhumetsale väga sarnased tulemused. Sügavuse ennustamisel oli *MAE* 0.072, *MSE* 0.008, *RMSE* 0.091 ja $R^2$ 0.755. Asukoha puhul olid vastavad näitajad *MAE* 0.180, *MSE* 0.0480, *RMSE* 0.219 ja $R^2$ 0.244. Kombineeritud ülesandes saavutati *MAE* 0.127,

*MSE* 0.028, *RMSE* 0.168 ja $R^2$ 0.499. Kuigi tulemused sarnanesid juhumetsale, oli selle ennustusaeg oluliselt parem, jäädes 0.008 kuni 0.015 sekundi vahele

Kokkuvõttes ennustasid kõik kolm mudelit prao sügavust paremini kui selle asukohta. XGBoost andis sügavuse ennustamisel kõige täpsemad tulemused. Kombineeritud ülesandes saavutas parima tulemuse juhumets. Lineaarne regressioon jäi täpsuselt kõigis ülesannetes teistest maha. Ennustuskiirus oli suurim lineaarse regressiooni puhul, järgnemas XGBoost ja seejärel juhumets.




Joonis 4. $R^2$ väärtused mudelite lõikes (sisendvektoris on omasagedued).




Joonis 5. *MAE*, *MSE* ja *RMSE* väärtused mudelite ja ülesannete lõikes (sisendvektoris on omasagedused).

### 6.3.2. Tulemused Haari lainikute 16 koefitsiendiga

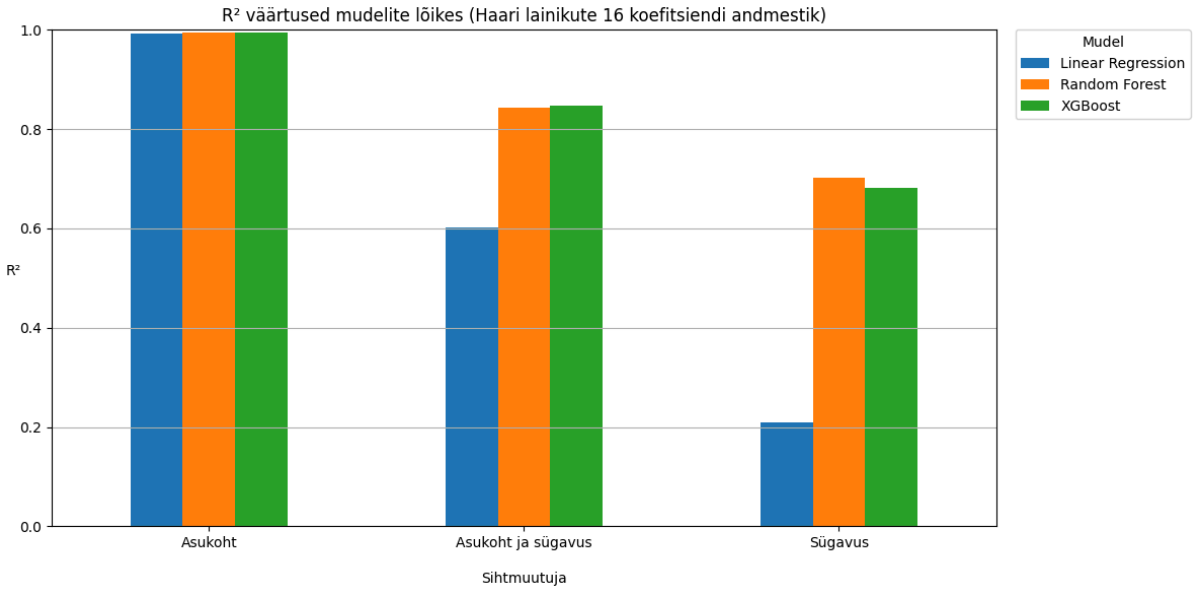
Haari lainikute esimesed 16 koefitsienti andsid asukoha määramisel omasagedustega võrreldes oluliselt paremad tulemused. Kõik kolm mudelit saavutasid asukoha ennustamisel väga madala vea ja kõrge determinatsioonikordaja (vt joonis 6 ja joonis 7). Sügavuse hindamisel sarnanesid tulemused siiski omasageduste omadega.

Lineaarne regressioon saavutas sügavuse ennustamisel *MAE* 0.135, *MSE* 0.027, *RMSE* 0.164 ja $R^2$ 0.210. Asukoha määramisel olid tulemused oluliselt paremad: *MAE* 0.018, *MSE* 0.00050, *RMSE* 0.022 ja $R^2$ 0.992. Kombineeritud ülesandes saavutati *MAE* 0.076, *MSE* 0.0137, *RMSE* 0.117 ja $R^2$ 0.601. Ennustusaeg jäi vahemikku 0.0025 kuni 0.0047 sekundit. Tulemustest ilmneb, et kuigi asukoha määramine oli väga täpne, jäi sügavuse ennustamine selgelt nõrgemaks.

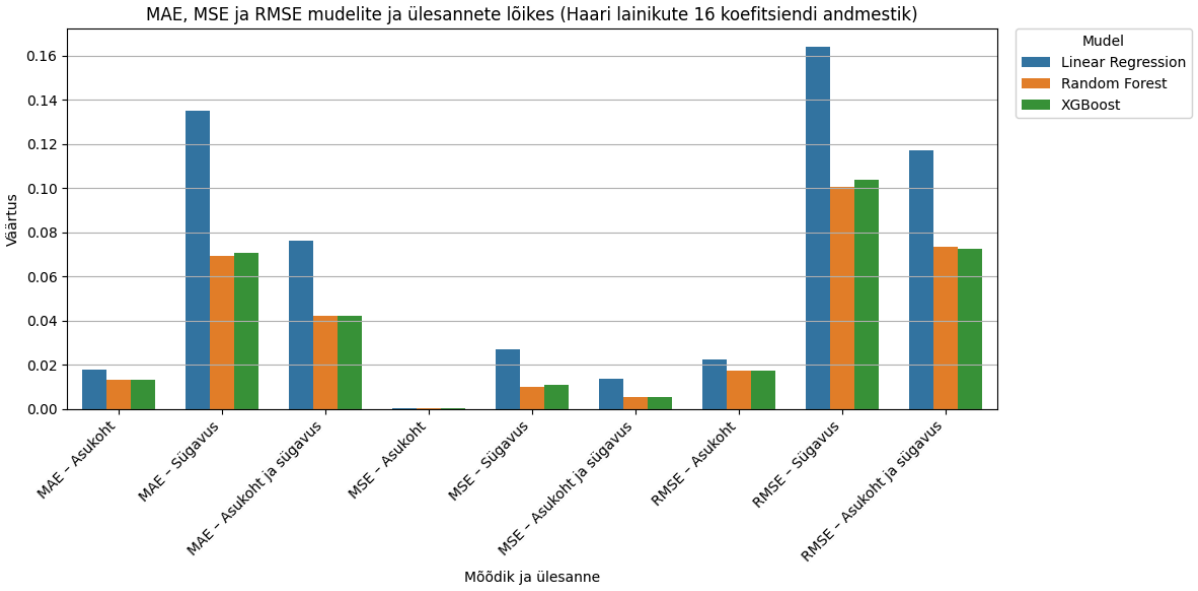
Juhumets parandas tulemusi märgatavalt ka sügavuse ennustamisel. Sügavuse korral saavutati *MAE* 0.069, *MSE* 0.010, *RMSE* 0.101 ja $R^2$ 0.702. Asukoha puhul olid tulemused väga head: *MAE* 0.013, *MSE* 0.0003, *RMSE* 0.017 ja $R^2$ 0.995. Kombineeritud ülesandes saadi *MAE* 0.042, *MSE* 0.005, *RMSE* 0.073. $R^2$ kasvas väärtuseni 0.844, mis on selgelt kõrgem kui omasageduste korral. Ennustamiseks kulus aega 0.058 kuni 0.248 sekundit. Kokkuvõttes olid tulemused stabiilselt head nii ühe kui kahe väärtuse ennustamisel.

XGBoost oli tulemuste poolest juhumetsaga võrreldav. Sügavuse ennustamisel saavutati *MAE* 0.071, *MSE* 0.011, *RMSE* 0.104 ja $R^2$ 0.683. Asukoha ennustamise tulemused olid veelgi paremad: *MAE* 0.013, *MSE* 0.0003, *RMSE* 0.017 ja $R^2$ 0.995. Kombineeritud ülesandes saavutati *MAE* 0.042, *MSE* 0.005, *RMSE* 0.073 ja $R^2$ 0.995, mis oli kõigist kolmest mudelist parim. Ennustamisaeg jäi vahemikku 0.010 kuni 0.022 sekundit. Tulemused kinnitasid, et mudel on täpne ja samas ka kiire.

Kokkuvõttes näitasid kõik mudelid, et Haari 16 koefitsiendid on väga tõhusad prao asukoha määramisel. Sügavuse ennustamisel oli täpsus madalam, kuid siiski sarnane omasagedustega võrreldes. XGBoost ja juhumets andsid kõige paremad tulemused kombineeritud ülesandes, samas kui lineaarne regressioon jäi sügavuse puhul selgelt nõrgemaks. Ennustamise kiiruse poolest oli XGBoost mõõdukalt kiirem kui juhumets.




Joonis 6. $R^2$ väärtused mudelite lõikes (sisendvektoris on Haari lainikute 16 koefitsienti).




Joonis 7. *MAE*, *MSE* ja *RMSE* väärtused mudelite ja ülesannete lõikes (sisendvektoris on Haari lainikute 16 koefitsienti).

### 6.3.3. Tulemused Haari lainikute 32 koefitsiendiga

Haari lainikute 32 koefitsiendiga andmestik andis kõigi mudelite puhul häid tulemusi ning jätkas 16 koefitsiendiga saadud trendi. Erinevused mudelite vahel olid sarnased 16 koefitsiendi andmestikuga, kuid kombineeritud ülesannetes saavutati mõnevõrra paremad näitajad. Kõik mudelid saavutasid asukoha ennustamisel väga väikese vea ja kõrge

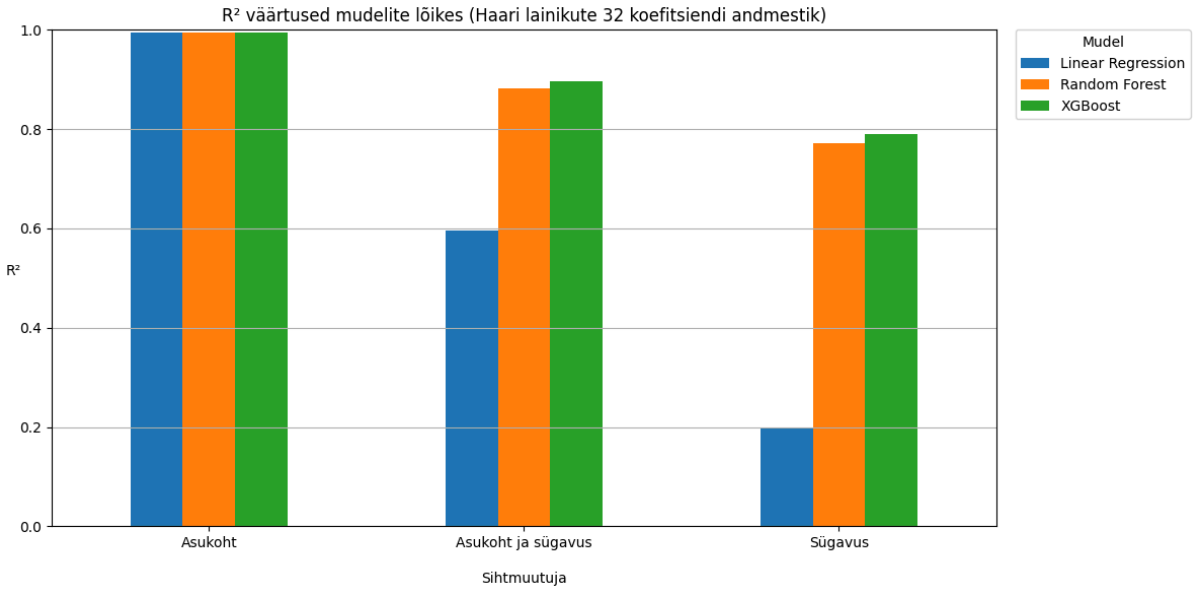
determinatsioonikordaja (vt Joonis 8 ja Joonis 9). Sügavuse korral olid tulemused mudelite lõikes eristuvamad.

Lineaarne regressioon saavutas sügavuse ennustamisel *MAE* 0.137, *MSE* 0.027, *RMSE* 0.165 ja $R^2$ 0.199, mis oli väga sarnane 16 koefitsiendi tulemustele ning jääb alla teistele mudelitele. Asukoha ennustamisel olid tulemused taas väga täpsed: *MAE* 0.016, *MSE* 0.00038, *RMSE* 0.*020* ja $R^2$ 0.*994*. Kombineeritud ülesandes saavutati *MAE* 0.076, *MSE* 0.0138, *RMSE* 0.118 ja $R^2$ 0.596. Ennustusaeg jäi vahemikku 0.0031 kuni 0.0199 sekundit.

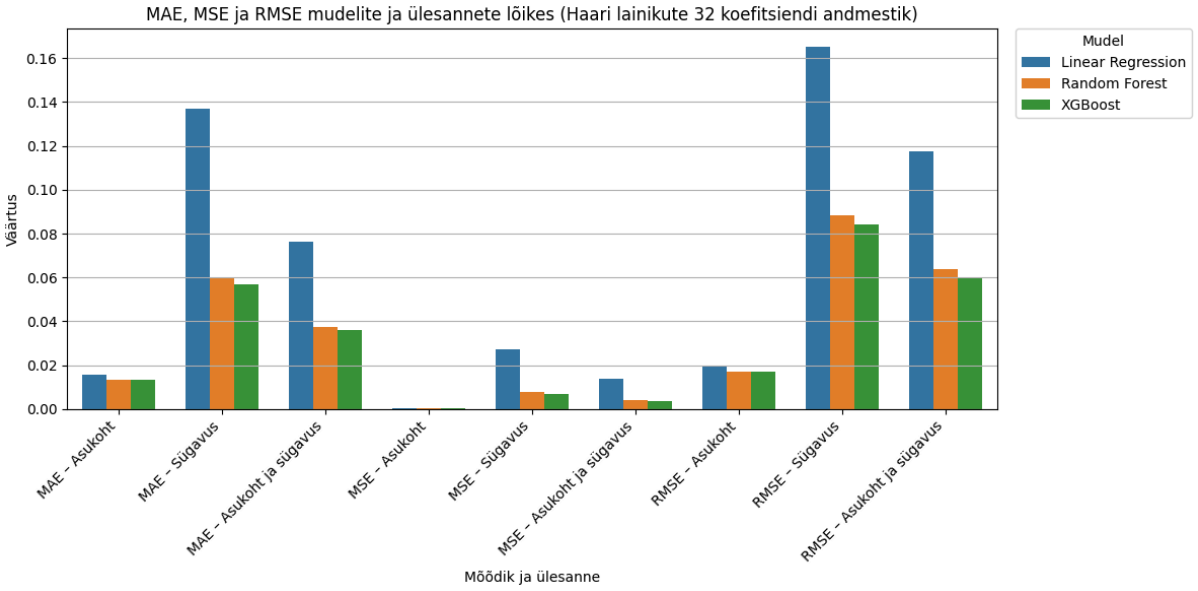
Juhumets parandas sügavuse ennustamise tulemusi märgatavalt, saavutades *MAE* 0.060, *MSE* 0.0078, *RMSE* 0.088 ja $R^2$ 0.771. Asukoha puhul olid tulemused väga täpsed: *MAE* 0.013, *MSE* 0.00029, *RMSE* 0.017 ja $R^2$ 0.995. Kombineeritud ennustamisel saavutati *MAE* 0.037, *MSE* 0.0041, *RMSE* 0.064 ja $R^2$ 0.883. Ennustamiskiirus oli aeglasem kui teistel mudelitel, jäädes vahemikku 0.062 kuni 0.241 sekundit. Jällegi olid tulemused järjekindlalt tugevad kõigis kolmes ülesandes.

XGBoost andis selle andmestiku puhul kõige paremad tulemused. Sügavuse ennustamisel saavutati *MAE* 0.057, *MSE* 0.0071, *RMSE* 0.084 ja $R^2$ 0.791. Asukoha ennustamisel olid mõõdikud *MAE* 0.013, *MSE* 0.00029, *RMSE* 0.017 ja $R^2$ 0.995. Kombineeritud ülesandes *MAE* 0.036, *MSE* 0.0036, *RMSE* 0.060 ja $R^2$ 0.896. Ennustusaeg jäi vahemikku 0.0178 kuni 0.0310 sekundit. Tulemused olid stabiilselt kõrged kõikides mõõdikutes.

Kokkuvõttes andis Haari 32 koefitsiendi andmestik kõige täpsemad tulemused kõigis ülesannetes. Parima soorituse saavutas XGBoost, mille tulemused olid järjekindlalt täpsemad kui teistel mudelitel. Juhumets jäi väga lähedale, samas kui lineaarne regressioon jäi sügavuse ennustamisel märkimisväärselt nõrgemaks. Võrreldes 16 koefitsiendiga andmestikuga oli tulemuste paranemine tagasihoidlik, kuid siiski märgatav, mis viitab sellele, et suurem tunnuste arv aitab paremini eristada peenemaid erinevusi.




Joonis 8. $R^2$ väärtused mudelite lõikes (sisendvektoris on Haari lainikute 32 koefitsienti).




Joonis 9. *MAE*, *MSE* ja *RMSE* väärtused mudelite ja ülesannete lõikes (sisendvektoris on Haari lainikute 32 koefitsienti).

### 6.3.4. Mudelite võrdlus

Erinevate masinõppe mudelite võrdlemine näitas selgelt, et mudelite sooritusvõime sõltub oluliselt nii andmestikust kui ka ülesandest. Kõik kolm mudelit käitusid erinevalt sõltuvalt sellest, kas tuli ennustada prao asukohta, sügavust või mõlemat korraga. Samuti ilmnes, et mudelid käituvad erinevatel andmestikel erinevalt. Järgnevalt antakse ülevaade iga mudeli

tugevustest ja nõrkustest ning analüüsitakse, milline andmestik osutus kõige tõhusamaks. Lõpuks tehakse kokkuvõte, millist mudelit võiks antud ülesandes eelistada.

Lineaarne regressioon näitas stabiilseid tulemusi asukoha ennustamisel, kuid jäi sügavuse ja kombineeritud ülesannetes teistest oluliselt alla. Parimad tulemused saavutati Haari 32 koefitsiendiga andmestikul, kus asukoha määramisel oli $R^2$ 0.994 ja kombineeritud ülesandes 0.596. Mudel sobib olukordadesse, kus võib leppida väiksema täpsusega ja oluline on lihtsus.

Juhumets oli tugev kandidaat kõikides ülesannetes, eriti Haari andmestike korral. Haari lainikute 32 koefitsiendi andmestiku puhul saavutati kombineeritud ülesandes $R^2$ väärtus 0.883. Asukoha määramisel ületas $R^2$ pidevalt 0.995. Mudel jäi täpsuselt XGBoostile veidi alla, kuid andis stabiilselt häid tulemusi kõigis mõõdikutes.

XGBoost oli võrreldud mudelitest kõige täpsem. Parim tulemus saavutati Haari lainikute 32 koefitsiendi andmestikuga, kus kombineeritud ülesande $R^2$ ulatus 0.896-ni. Ka sügavuse ennustamisel saavutati maksimaalseks $R^2$ 0.791 ning asukoha korral 0.995. Tulemused olid järjekindlalt parimad kõigis ülesannetes, mis kinnitab mudeli sobivust pragude identifitseerimiseks.

Kõigi mudelite puhul jäid ennustusajad alla murdosa sekundi, mistõttu ei ole kiirus praktilises kasutuses piirav tegur. Lineaarne regressioon oli kõige kiirem, järgnemas XGBoost ja seejärel juhumets. Siiski olid erinevused väikesed. Kuna mudelid treeniti eelnevalt ja ennustamine toimub vaid tulemuste põhjal, ei mõjuta kiirus oluliselt mudeli valikut.

Kokkuvõttes oli parim mudel XGBoost, mis saavutas kõige täpsemad tulemused nii sügavuse, asukoha kui ka kombineeritud ülesannetes. Parim andmestik oli Haari lainikute 32 koefitsiendi andmestik; juhumets jäi tulemustelt XGBoostile väga lähedale, kuid siiski veidi alla. Lineaarne regressioon sobib kiireks ennustamiseks, kuid ei paku piisavat täpsust keerukamate mustrite puhul.

.

# 7. Kokkuvõte

Selles bakalaureuse töös võrreldi lineaarset regressiooni, juhumetsa ja XGBoosti sobivust vardas esinevate pragude ennustamiseks. Eesmärk oli identifitseerida prao asukoht ja sügavus, kasutades kolme tüüpi sisendandmeid: omasagedused, 16 Haari lainikute koefitsienti ja 32 Haari lainikute koefitsienti. Mudelid treeniti arvutusliku meetodi abil saadud andmestikel ning nende sooritust hinnati mitme regressioonmõõdikute abil. Lisaks lisati andmetele müra, et simuleerida mõõtemääramatust.

Tulemused näitasid, et prao sügavuse ennustamine oli üldiselt täpsem kui asukoha määramine. Kõige paremaid tulemusi andsid Haari lainikute koefitsiendid, eriti 32 koefitsiendiga andmestik. Lineaarne regressioon oli ennustusajalt kõige kiirem, ent jäi täpsuselt teistest nõrgemaks. Juhumets saavutas igas ülesandes stabiilselt häid tulemusi, kuid kõige silmapaistvama sooritusega oli XGBoost. Prao sügavuse ja asukoha koos hindamisel ulatus Haari lainikute 32 koefitsiendil treenitud mudeli determinatsioonikordaja 0.896-ni. Juhumets sai samal andmestikul tulemuseks 0.883 ja lineaarne regressioon 0.596. Võib järeldada, et XGBoost koos 32 Haari lainiku koefitsiendiga on antud probleemi lahendamiseks kõige sobivam.

Töö näitas, milline mudel ja milline andmestik sobivad kõige paremini varraste pragude ennustamiseks. Saadud teadmisi saab kasutada praktiliste süsteemide arendamisel, mis aitavad jälgida konstruktsioonide seisukorda. Tulevikusuunad võiksid hõlmata täiendavate masinõppe mudelite rakendamist, et uurida nende võimekust pragude tuvastamisel. Samuti võiks kasutada uusi sisendvektoreid, sealhulgas reaalsetel mõõtmistel põhinevaid andmeid.

## 8. Viidatud kirjandus


1. Agrawal, R. (2025). Know The Best Evaluation Metrics for Your Regression Model! Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/ (27.04.2025)
2. Aroeira, C. (2024). Identification of natural frequencies. https://www.dmc.pt/en/identificacao-de-frequencias-naturais/ (13.04.2025)
3. Brainforge. Haar Wavelet. https://www.brainforge.ai/glossary/haar-wavelet (11.04.2025)
4. Donges, N. Random Forest: A Complete Guide for Machine Learning. Built In. https://builtin.com/data-science/random-forest-algorithm (14.04.2025)
5. GeeksforGeeks. (2025). Classification vs Regression in Machine Learning. https://www.geeksforgeeks.org/ml-classification-vs-regression/ (01.05.2025)
6. Hernández-Díaz, A. M., Pérez-Aracil, J., Lorente-Ramos, E., Marina, C. M., Peláez-Rodríguez, C., & Salcedo-Sanz, S. (2024). Machine learning as an alternative strategy for the numerical prediction of the shear response in reinforced and prestressed concrete beams. Results in Engineering, 22, 102139. https://doi.org/10.1016/j.rineng.2024.102139
7. IBM. What is random forest? https://www.ibm.com/think/topics/random-forest (13.04.2025)
8. IBM. (2021) What is linear regression? https://www.ibm.com/think/topics/linear-regression (13.04.2025)
9. IBM. (2024) What is XGBoost? https://www.ibm.com/think/topics/xgboost (13.04.2025)
10. Jaanuska, L. (2021). Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification, TÜ arvutiteaduse instituudi doktoritöö, https://hdl.handle.net/10062/71083

11. Jaanuska, L., Hein, H. (2019). Comparison of machine learning methods for crack localization. Acta et Commentationes Universitatis Tartuensis de Mathematica, 23(1), 125-143, https://doi.org/10.12697/ACUTM.2019.23.13

12. Jaanuska, L., Hein, H. (2022). Quantification of cracks in beams on the Pasternak foundation using Haar wavelets and machine learning. Proceedings of the Estonian Academy of Sciences, 71(1), 16-29, https://doi.org/10.3176/proc.2022.1.02

13. Klauson, A., Metsaveer, J., Põdra, P., Raukas (2012). Tugevusõpetus. Tallinn: Tallinna Tehnikaülikooli Kirjastus.

14. Laigna, K. (2000). Tugevusõpetus. Tallinn: Eesti Mereakadeemia.

15. Ooijevaar, T. (2014). Vibration based structural health monitoring of composite skin-stiffener structures, University of Twente doktoritöö, https://doi.org/10.3990/1.9789036536240

16. Peng, K., Zhang, Y., Xu, X., Han, J., Luo, Y. (2022). Crack Detection of Threaded Steel Rods Based on Ultrasonic Guided Waves. SENSORS, 22(18), 6885. https://doi-org.ezproxy.utlib.ut.ee/10.3390/s22186885

17. Pfaller, M. 6 important methods for crack testing in non-destructive testing https://blog.foerstergroup.com/en/component-testing/the-6-most-important-methods-for-crack-testing-in-non-destructive-material-testing (05.12.2024)

18. Sügis, E., Tampuu, A., Aljanaki, A., Fišel, M., Kull, M. (2025). Praktiline andmeteadus. Tartu Ülikooli arvutiteaduse instituut.

# Lisad

## I. Kasutatud mudelid ja hüperparameetrite valim

Alljärgnev koodilõik iseloomustab kasutatud masinõppe mudelite konfiguratsiooni ja nende hüperparameetrite valimit.

```
models = {
    "Linear Regression": {
        "model": LinearRegression(),
        "params": {
            "fit_intercept": [True, False],
            "positive": [True, False]
        }
    },
    "Random Forest": {
        "model": RandomForestRegressor(random_state=42),
        "params": {
            "n_estimators": [30, 100, 200],
            "max_depth": [None, 2,10],
            "max_features": [1.0, None, "sqrt", "log2"]
        }
    },
    "XGBoost": {
        "model": xgb.XGBRegressor(
            random_state=42,
            objective='reg:squarederror',
            tree_method='auto'
        ),
        "params": {
            "n_estimators": [50,100, 200],
            "max_depth": [3, 6, 9],
            "learning_rate": [0.01, 0.05, 0.1]
        }
    }
}
```

## II. Mudelite tulemuste võrdlustabel

Alljärgnev tabel koondab kõigil kolmel andmestikul ja kolmel muutujal treenitud mudelite tulemused. Tabelis on esitatud veamõõdikud ning testandmestiku ennustamiseks kulunud aeg.

| Mudel | Andmestik | Ennustatav väärtus | MAE | MSE | RMSE | $R^2$ | Ennustuse aeg (s) |
|---|---|---|---|---|---|---|---|
| Linear Regression | omasagedused | asukoht | 0.19153 | 0.052179 | 0.228427 | 0.177747 | 0.003386 |
| Linear Regression | omasagedused | sügavus | 0.087134 | 0.01176 | 0.108444 | 0.654074 | 0.002245 |
| Linear Regression | omasagedused | sügavus ja asukoht | 0.139332 | 0.031969 | 0.1788 | 0.41591 | 0.004425 |
| Linear Regression | Haari lainikute 16 koefitsienti | asukoht | 0.017742 | 0.000495 | 0.022246 | 0.992201 | 0.002862 |
| Linear Regression | Haari lainikute 16 koefitsienti | sügavus | 0.135017 | 0.026845 | 0.163845 | 0.210342 | 0.002101 |
| Linear Regression | Haari lainikute 16 koefitsienti | sügavus ja asukoht | 0.07638 | 0.01367 | 0.116919 | 0.601272 | 0.007988 |
| Linear Regression | Haari lainikute 32 koefitsienti | asukoht | 0.015779 | 0.000382 | 0.019532 | 0.993988 | 0.010256 |
| Linear Regression | Haari lainikute 32 koefitsienti | sügavus | 0.136823 | 0.027247 | 0.165068 | 0.198515 | 0.010332 |
| Linear Regression | Haari lainikute 32 koefitsienti | sügavus ja asukoht | 0.076301 | 0.013814 | 0.117535 | 0.596251 | 0.007827 |
| Random Forest | omasagedused | asukoht | 0.182228 | 0.047746 | 0.218508 | 0.247603 | 0.040707 |
| Random Forest | omasagedused | sügavus | 0.072211 | 0.00814 | 0.09022 | 0.760573 | 0.089862 |
| Random Forest | omasagedused | sügavus ja asukoht | 0.126706 | 0.027903 | 0.167042 | 0.504711 | 0.182751 |
| Random Forest | Haari lainikute 16 koefitsienti | asukoht | 0.013324 | 0.000297 | 0.017242 | 0.995315 | 0.069875 |
| Random Forest | Haari lainikute 16 koefitsienti | sügavus | 0.069484 | 0.010138 | 0.10069 | 0.701775 | 0.084992 |
| Random Forest | Haari lainikute 16 koefitsienti | sügavus ja asukoht | 0.042319 | 0.005391 | 0.073423 | 0.843526 | 0.173099 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Random Forest | Haari lainikute 32 koefitsienti | asukoht | 0.013213 | 0.000291 | 0.017046 | 0.995421 | 0.061694 |
| Random Forest | Haari lainikute 32 koefitsienti | sügavus | 0.059707 | 0.007775 | 0.088179 | 0.771283 | 0.0948 |
| Random Forest | Haari lainikute 32 koefitsienti | sügavus ja asukoht | 0.037216 | 0.004063 | 0.063743 | 0.8825 | 0.196164 |
| XGBoost | omasagedused | asukoht | 0.179989 | 0.047973 | 0.219027 | 0.244024 | 0.023216 |
| XGBoost | omasagedused | sügavus | 0.072329 | 0.008315 | 0.091188 | 0.755407 | 0.009441 |
| XGBoost | omasagedused | sügavus ja asukoht | 0.12659 | 0.028153 | 0.167788 | 0.499457 | 0.014396 |
| XGBoost | Haari lainikute 16 koefitsienti | asukoht | 0.013477 | 0.000301 | 0.017342 | 0.995261 | 0.008637 |
| XGBoost | Haari lainikute 16 koefitsienti | sügavus | 0.070718 | 0.01079 | 0.103874 | 0.682618 | 0.014914 |
| XGBoost | Haari lainikute 16 koefitsienti | sügavus ja asukoht | 0.042004 | 0.005265 | 0.072561 | 0.847341 | 0.021913 |
| XGBoost | Haari lainikute 32 koefitsienti | asukoht | 0.013394 | 0.00029 | 0.017033 | 0.995428 | 0.013063 |
| XGBoost | Haari lainikute 32 koefitsienti | sügavus | 0.056776 | 0.007114 | 0.084345 | 0.790739 | 0.018973 |
| XGBoost | Haari lainikute 32 koefitsienti | sügavus ja asukoht | 0.036024 | 0.003621 | 0.060178 | 0.895603 | 0.029031 |

## III. Litsents

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Kaarel Tamuri,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   “Masinõppe meetodite võrdlus vardas esinevate pragude iseloomustamiseks”,

   mille juhendaja on Ljubov Jaanuska, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

*Kaarel Tamuri*

***15.05.2025***