Tallinna Tehnikaülikool

Kaur Matthias Raavel, Tanel Sõerd, Rainer Randmaa

Tuuliku tootmisvõimsuse ennustamine

Statistilised meetodid masinõppes

Tallinn 2023

Sisukord

1 Uurimisprobleem

Taastuvenergia on maailmas järjest rohkem oluline valdkond. Rohelistesse energiaallikatesse investeeritakse igal aastal järjest rohkem ning roheenergia osakaal meie igapäevases elektritarbimises on järjest suurem. Näiteks Enefit Greenile kuulub 2023 aasta seisuga Eestis ja Leedus 22 tuuleparki ning neis töötab kokku 165 tuulikut. Tuuleparkide koguvõimsus on 398 MW ning need toodavad aastaga ligi 1 teravatt-tundi elektrit, selle kogusega saaks varustada aasta jooksul rohkem kui 330 000 keskmise tarbimisega kodu [1]. Võrdluseks, Eesti suurima Auvere elektrijaama installeeritud võimsus Nord Pool andmetel on 274 MW [2].

Roheenergia tootmisel on aga üks probleem. Kui elektrijaamad, mille tooraineks on näiteks põlevkivi, on juhitava võimsusega, siis rohelist energiat tootvad jaamad, näiteks tuulepargid, ei ole. Tuulepargi tootmisvõimsus sõltub näiteks tuule kiirusest ja suunast ning tuuliku tootmisvõimsust ei ole võimalik vastavalt vajadusele suunata.

Siiski on oluline tootmist kuidagi prognoosida ja juhtida. Elektrivõrk on mõeldud töötama mingisuguse kindla sageduse peal (Eestis on see näiteks 50 Hz [3]) ning seda sagedust on võimalik säilitada siis, kui tarbimine ja tootmine on tasakaalus. Kui tahta tarbimist katta taastuvenergiat tootvate allikatega, on oluline võimalikult täpselt prognoosida nende tootmisvõimsust. Elektrikauplejad kasutavad prognoose ja ostavad või müüvad iga päev järgmiseks päevaks elektribörsil tunnipõhise täpsusega tootmise ja tarbimise vahe katteks elektrit. Ebakõlad tootmise ja tarbimise vahel tekitavad olukorra, kus elektrivõrgu stabiilsuse tagamiseks on vaja viimasel hetkel elektrit turult juurde osta. Päevasiseselt elektribörsilt ostetud elekter on oluliselt kallim, kui päev-ette börsilt.

Toodetud energiat kaubeldakse rahvusvahelisel turul. Põhjamaades, Baltikumis ja Ühendkuningriikides pakub energiakauplemise võimalust Nord Pool energiaturg [4].

Seega on tuuleparkides asuvate turbiinide tootmisvõimsuse ennustamine äriliselt väga oluline probleem. Et elektrivõrgus tarbimine ja pakkumine tasakaalus hoida, tuleb tootmist planeerida võimalikult täpselt. Kui võrk muutub ebastabiilseks, rikub see kõiki seadmeid, mis sellega ühendatud on, seehulgas võrgu enda infrastruktuuri kui ka tarbijate seadmeid. Kui planeeritud tootmises ja tarbimises on ebakõlad, on võrgu stabiilsuse tagamiseks elektrienergia sisse ostmisel tehtavad kulutused vältimatud, kuid ülikulukad.

Käesolev töö proovib vastata järgmistele uurimisküsimustele:

- Millised on olulised parameetrid tuuliku energiatootluse ennustamiseks? (järeldamise probleem)
- Kuidas kasutada ajaloolisi andmeid, et treenida võimalikult täpse ennustusvõimega mudel ning kui täpseks on võimalik mudel teha? (prognoosimise probleem)

Proovime prognoosida justnimelt tuuliku tootmisvõimsust, mitte tuulepargi tootmisvõimsust, sest tuulepargi tootmisvõimsuse ennustamisel on tuulikupõhine granulaarsus oluline. Tuuleparkide juhtimine ja haldus kui protsess toimub tuuliku granulaarsusega ning tootmisvõimsuse ennustamine kui äriline protsess peab seda toetama. Vahel võib olla võimalik kasutada ainult ükskuid tuulikuid ning sellisel juhul oleks tuulikupõhisest ennustamisest rohkem kasu. Selline olukord võib tekkida siis, kui tuulepargis on mõnel tuulikul toimumas näiteks hooldustööd.

2 Andmed

Selle projekti andmed pärinevad Zenodo avatud andmekogust. Andmekogu uusim versioon on avaldatud 2023 aasta augustis ettevõtte Cubico Sustainable Investments Ltd poolt ning autoriks on Charlei Plumley. Tegemist on Ühendkuningriikides asuva Kelmarshi tuulepargis asuva kuue tuuliku kohta käivate andmetega. Andmekogu sisaldab nii staatilisi andmeid tuulikute kohta (koordinaadid, hinnatud võimsus, rootori diameeter) kui ka 10 minutilise granulaarsusega aegridu tuulikute SCADA kontrolleritelt. [5]

Andmekogu sisaldab andmeid vahemikus 2016 aasta algus kuni 2022 aasta lõpp. Käesolevas projektis on peamisteks andmeallikateks kõige uuemad, 2021–2022 aastate andmed. Selle perioodi andmed on valitud nende ajakohasuse ja kvaliteedi tõttu, mis tagab töö suurema usaldusväärsuse ja täpsuse. Samas, eksperimentaalsel eesmärgil treenisime mudeli ka ajavahemiku 2017–2022 andmete põhjal, et uurida andmestiku pikema perioodi mõju mudeli täpsusele ja robustsusele. Siinkohal tuleb märkida, et 2016. aasta andmeid nende ebakvaliteetse iseloomu tõttu käesolevas projektis ei kasutatud.

Aegridadel on kokku 303 muutujat, kuid nende hulgas on mitmeid selliseid muutujaid, mis järelduvad toodangust ning neid prognoosimisel kasutada ei saa, näiteks nagu investeeringu tasuvus, pinge, voolutugevus jne. Lisaks on mitmeid muutujaid, millel on N/A väärtuste osakaal suur. Järgnev tabel loetleb üles 20 SCADA andmete muutujat, mida on autorite hinnangul võimalik tootmise ennustamiseks kasutada. Minimaalsete ja maksimaalsete muutujate puhul on oluline silmas pidada, et tegemist on ajaperioodi (10 minuti) miinimumi ja maksimumiga.

Muutuja nimi andmestikus	Selgitus
Wind.speed..m.s.	Tuule kiirus, meetrit sekundis
Wind.speed..Standard.deviation..m.s.	Tuule kiiruse standardhälve, meetrit sekundis
Wind.speed..Minimum..m.s.	Tuule kiirus minimaalne, meetrit sekundis
Wind.speed..Maximum..m.s.	Tuule kiirus maksimaalne, meetrit sekundis
Long.Term.Wind..m.s.	Pikaajaline tuule kiirus, meetrit sekundis
Density.adjusted.wind.speed..m.s.	Õhutihedusega kohandatud tuule kiirus, meetrit sekundis
Wind.direction....	Tuule suund, kraadides

Wind.direction..Standard.deviation....	Tuule suund, standardhälve, kraadides
Wind.direction..Minimum....	Tuule suund, minimaalne, kraadides
Wind.direction..Maximum....	Tuule suund, maksimaalne, kraadides
Lost.Production.Total..kWh.	Kaotatud toodang, kokku, kWh – näitab palju kaotati toodangut erinevatel põhjustel, põhjuseks võivad olla linnud, lepingulised põhjused, võrgu tasakaalu hoidmine jne.
Nacelle.ambient.temperature...C.	Õhutemperatuur turbiini koja ümber, kraadi Celsiust
Ambient.temperature..converter....C.	Õhutemperatuur trafo juures, kraadi Celsiust
Ambient.temperature..converter...Max...C.	Õhutemperatuur trafo juures, maksimaalne, kraadi Celsiust
Ambient.temperature..converter...StdDev...C.	Õhutemperatuur trafo juures, standardhälve, kraadi Celsiust
Nacelle.ambient.temperature..Max...C.	Õhutemperatuur turbiini koja ümber, maksimaalne, kraadi Celsiust
Nacelle.ambient.temperature..Min...C.	Õhutemperatuur turbiini koja ümber, minimaalne, kraadi Celsiust
Nacelle.ambient.temperature..StdDev...C.	Õhutemperatuur turbiini koja ümber, standardhälve, kraadi Celsiust
Date.and.time	Kuupäev ja kellaaeg, yyyy-MM-dd HH:mm:ss formaat, UTC
Power..kW. (sõltuv tunnus)	Turbiini keskmine võimsus ajaperioodil, kilovatti

Tabel 1. Kasutatud SCADA andmete muutujad.

Sõltuvaks muutujaks on turbiini keskmine tootmisvõimsus ajaperioodil. Kuna tegemist on aegreaga, ei sobi kuupäev ja kellaaeg sellisel kujul mudelile sisendiks. Metoodika kirjelduses on seletatud, kuidas kuupäeva ja kellaaja muutujat mudeli sisendiks kohaldati.

Valitud muutujate hulga kirjeldav statistika on järgnev (v.a kuupäev).

Muutuja	Keskmine	Mediaan	Standardhälve	Min	Max
Wind.speed..m.s.	6,55	6,22	2,33	3,00	21,94
Wind.speed..Standard.deviation..m.s.	0,94	0,87	0,43	0,05	5,86
Wind.speed..Minimum..m.s.	4,72	4,54	1,89	-14,7	17,77
Wind.speed..Maximum..m.s.	8,34	7,83	2,95	3,29	30,03
Long.Term.Wind..m.s.	6,11	6,14	0,66	5,26	7,21
Density.adjusted.wind.speed..m.s.	6,52	6,18	2,33	2,90	21,84
Wind.direction....	204,40	216,87	87,30	0,01	360,0
Wind.direction..Standard.deviation....	8,90	8,22	3,95	1,23	59,83
Wind.direction..Minimum....	179,36	196,50	88,92	0,00	351,0
Wind.direction..Maximum....	228.00	236,90	85,07	11,82	360,0
Lost.Production.Total..kWh.	0,32	0,57	13,08	-170,8	255,7
Nacelle.ambient.temperature...C.	11,98	11,45	5,88	-2,47	39,33
Ambient.temperature..converter.. ..C.	14,96	14,42	6,33	0,20	45,46
Ambient.temperature..converter ...Max...C.	15,36	14,80	6,34	0,50	46,05
Ambient.temperature..converter ...StdDev...C.	0,20	0,20	0,06	0,04	1,02
Nacelle.ambient.temperature..Max...C.	12,12	11,50	5,92	-2,40	39,60
Nacelle.ambient.temperature..Min...C.	11,84	11,40	5,85	-2,60	39,00
Nacelle.ambient.temperature..StdDev ...C.	0,09	0,07	0,09	0,00	1,72
Power..kW. (sõltuv tunnus)	688,58	493,08	592,58	-10,62	2082

Tabel 2. Kasutatud SCADA muutujate kirjeldav statistika.

3 Metoodika

Mudeli tegemisel otsustasime kasutada juhusliku metsa algoritmi, mis vähendab üksikute otsustuspuude kallutatust ja korrelatsiooni ning tavaliselt parandab võrreldes tavalise otsustuspuuga mudeli üldist ennustustäpsust ja robustsust. Lisaks suudab juhusliku metsa algoritm efektiivselt hallata suurte mõõtmetega andmestikke ja toime tulla korreleeritud tunnustega. Selleks, et leida üles, milliseid tunnuseid kasutada, siis võtsime abiks LASSO [6, p. 241]. Seda meetodit kasutatakse, et leida üles olulised tunnused ning seeläbi saavutada mudeli parem ennustusvõime ja tõlgendatavus. LASSO abiga otsustasime esialgse andmestiku 303 tunnuse hulgast ebaolulised tunnused mudelist eemaldada. LASSO on eriti kasulik suuremõõtmeliste andmekogumite korral, kus paljud tunnused võivad olla korreleeritud või mitteolulised [6]. Meie andmekogu iseloomustavad nii suured mõõtmed kui tugev korrelatsioon. Korrelatsioon väljendub selles, et mitmed tunnused on tegelikult ühe nähtuse kohta. Näiteks tuule kiirust iseloomustavad ajaperioodi keskmine, maksimaalne, minimaalne ja õhutihedusega kohandatud tuule kiirus.

Kasutatud andmetes oli olemas kuupäev ning kellaaeg, kuid see oli algselt *character* tüüpi muutuja ning seda ei osanud juhusliku metsa algoritm kasutada. Esimese sammuna tegime sellest kuupäeva, kasutades lubridate [7] teeki ning lisasime andmeraami juurde päeva (1-365) ning kellaaja (0-23) tulbad. Põhjendame otsust sellega, et aastaaeg ja kellaaeg võivad olla tootmisvõimsuse ennustamisel oluline informatsioon. Tuleb tähele panna, et andmed päeva ja tunni kohta on tsüklilised, kuid numbriliselt on näiteks tund 0 ja 23 üksteisega võrreldes skaala erinevates otstes, samas tegelikult on nende vahe üks tund. Lootuses teha see informatsioon mudelile paremini arusaadavaks, kasutasime andmete kodeerimiseks siinust ja koosinust [8]. Selle tulemusena on väärtuste üleminek otspunktidel sujuvam ning see võiks olla mudeli ehitamisel algoritmile paremini arusaadav.

Andmete analüüsimiseks kasutasime R keelt [9] ning RStudio Desktop tarkvara [10]. R funktsionaalsuste laiendamiseks kasutasime randomForest [11], glmnet [12] lubridate [7] ning plyr [13] teeke.

4 Tulemused ja järeldused

4.1 LASSO kasutamine

Et leida üles olulised tunnused, kasutasime LASSO-t. Esialgu proovisime rakendada LASSO-t kõikide andmestiku 303 tunnuse peal, kuid selgus, et see hindab oluliseks tunnuseid, mis järelduvad otseselt tegelikust tootmisvõimsusest – näiteks pinge ja voolutugevus. Kuna tunnuseid oli andmetes väga palju, siis oli vajalik enne korjata välja sellised tunnused, mis otseselt on seotud toodanguga ning mida ei ole adekvaatselt võimalik mudelile sisendiks anda, kui on soov reaalselt toodangut prognoosida. Enamasti jäid andmetesse sisse meteoroloogilised andmed, nagu temperatuur ja tuule kiirus, aga ka kaotatud toodang. Tehes mitu iteratsiooni LASSO-ga, korjasime mõned tunnused veel välja ning lõpptulemus jäi järgnev (9 tunnust).

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                              s1
(Intercept)                          547.208497
Wind.speed..m.s.                     548.408742
Long.Term.Wind..m.s.                  32.926853
Density.adjusted.wind.speed..m.s.     41.346022
Wind.direction....                    -2.701662
Lost.Production.Total..kWh.         -177.368410
Available.Capacity.for.Production..kW.  13.318029
Nacelle.ambient.temperature...C.        14.469544
Ambient.temperature..converter....C.     5.814048
```

Pilt 1. LASSO rakendamise väljund.

LASSO iteratsioonide vahepeal proovisime mudeleid treenida ka 20 tunnusega (sisaldas erinevaid *min*, *max* ja standardhälbe tunnuseid), kuid nende mudelite ennustustulemused olid kehvemad kui vähemate tunnustega treenitud mudelitel.

4.2 Informatsioon päeva ja tunni kohta

Peale 20 esimese olulise tunnuse välja selgitamist lisasime andmetesse info kuupäeva ja tunni kohta. LASSO tulemus näitas, et lisatud tunnused on olulised ning kasutades samasid andmeid treenimiseks ja testimiseks, selgus hoopis, et keskmine viga on suurem kui enne.

Peale seda kohandasime päeva ja tunni andmeid siinuse ja koosinusega ning tegime andmed tsükliliseks, lootuses teha need andmed algoritmile paremini arusaadavaks. LASSO tulemus hindas endiselt need tunnused oluliseks, välja arvatud päeva siinus ja samade andmetega treenides ning testides saime keskmiseks vea veel suurema kui lihtsalt päeva ja tunni info lisades.

Seega võib järeldada, et neid andmeid ei ole mõtet lisada, ilmselt on hooajalisuse informatsioon oluline pigem ilma ennustamisel, kuid toodangu ennustamisel ei ole vahet, mis aastaaeg või kellaaeg on, olulised on siiski muud tunnused. Huvitaval kombel, kui algselt treenisime ja testisime sama tuuliku 2022 andmetega, siis prognoosi tulemus paranes. Tulemus paranes siis, kui lisasime päeva ja tunni andmed ning kui lisasime siinuse ja koosinuse, paranes veelgi, kuid kui treenisime 2021 andmetega ja testisime 2022 andmetega, siis andis see vastupidise tulemuse.

4.3 Juhusliku metsa mudeli treenimine

Juhusliku metsa mudeli treenimisel ning testimisel otsustasime proovida erinevaid lähenemisi, kasutades treenimisel ja/või testimisel kas kõiki andmeid või andmeid ühe turbiini kohta. Kuna meie mudeli eesmärk on tuleviku tootmist prognoosida, siis kasutasime treenimiseks andmeid maksimaalselt 2021. aastast ning testimiseks andmeid 2022. aastast.

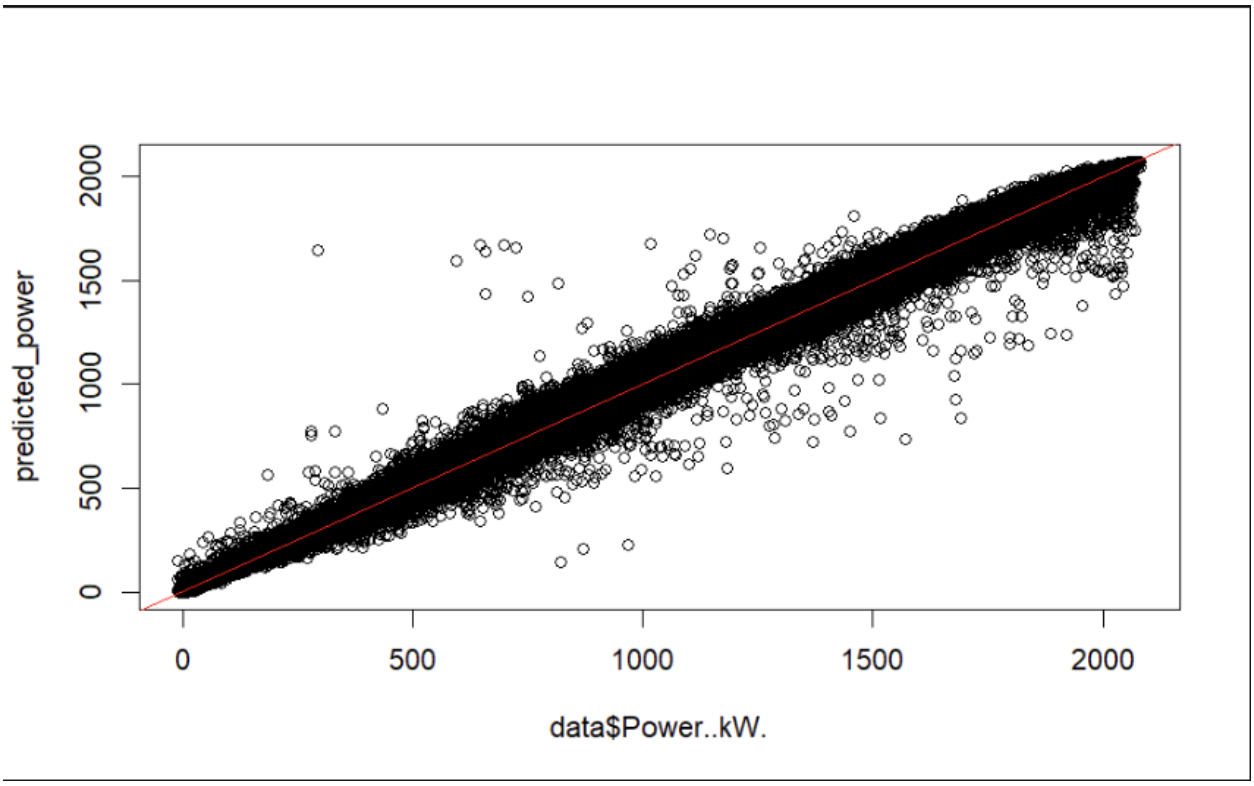
Esimese eksperimendina treenisime 2021. aasta kõikide turbiinide andmete peal mudeli ning hindasime mudeli täpsust 2022. aasta kõikide turbiinide andmete peal. Järgnevalt on esitatud pildid ja joonised viie puuga, kümne puuga ja viieteistkümne puuga juhusliku metsa mudeli karakteristikutest.

```
[1] "RMSE"
[1] 35.93882

Call:
 randomForest(formula = formula, data = train_data, method = "parRF",      ntree =
ntree)
               Type of random forest: regression
                     Number of trees: 5
No. of variables tried at each split: 3

          Mean of squared residuals: 1253.582
                    % Var explained: 99.7
```

Pilt 2. Esimese eksperimendi 5 puu väljund.



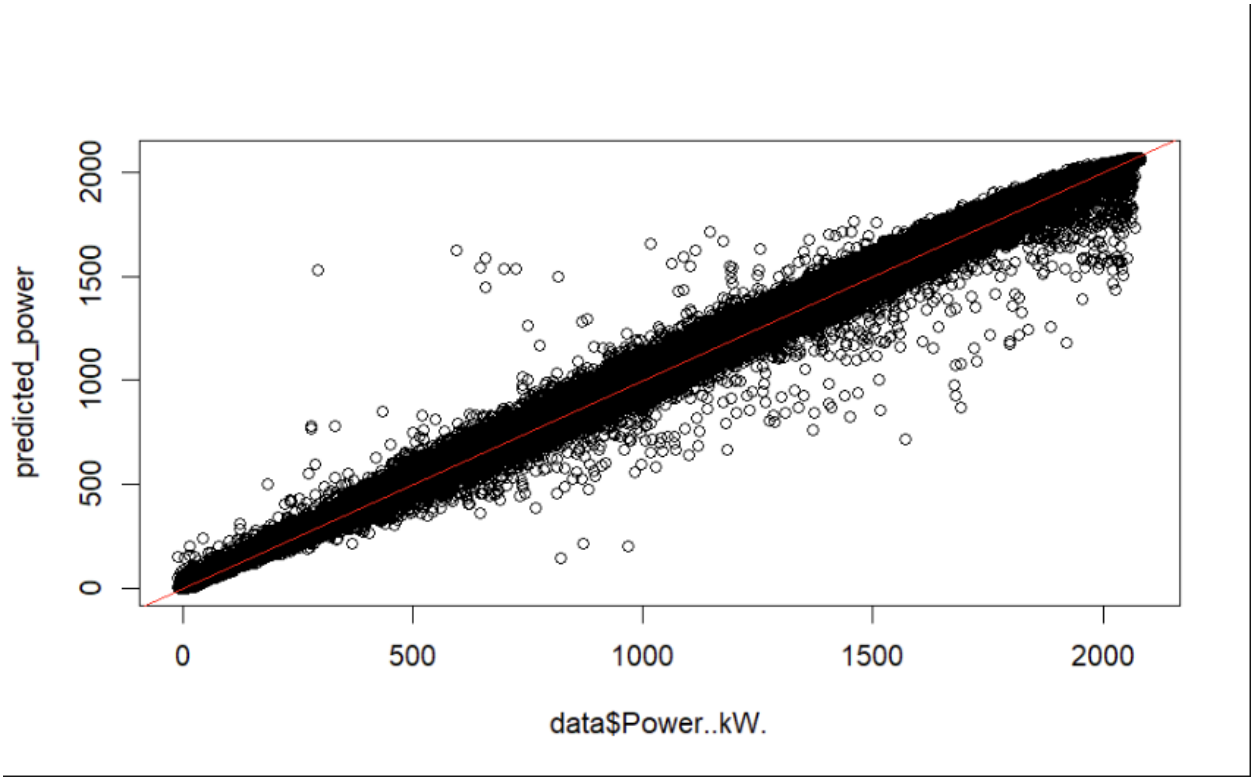
Joonis 1. Esimese eksperimendi 5 puu ennustustäpsus. y-teljel prognoos, x-teljel tegelik võimsus

```
[1] "RMSE"
[1] 33.17422

Call:
 randomForest(formula = formula, data = train_data, method = "parRF",      ntree =
ntree)
               Type of random forest: regression
                     Number of trees: 10
No. of variables tried at each split: 3

          Mean of squared residuals: 952.775
                    % Var explained: 99.77
```

Pilt 3. Esimese eksperimendi 10 puu väljund.



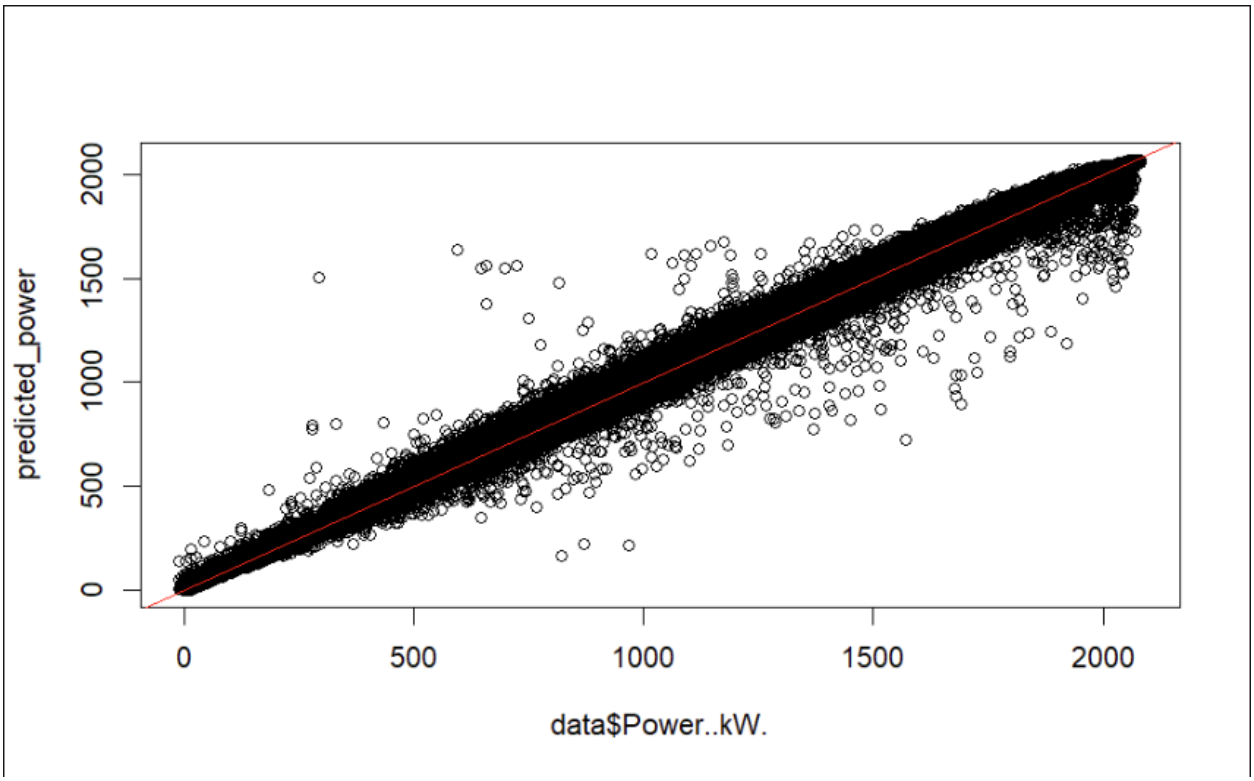
Joonis 2. Esimese eksperimendi 10 puu ennustustäpsus. y-teljel prognoos, x-teljel tegelik võimsus

```
[1] "RMSE"
[1] 32.53913

Call:
 randomForest(formula = formula, data = train_data, method = "parRF",      ntree =
ntree)
               Type of random forest: regression
                     Number of trees: 15
No. of variables tried at each split: 3

          Mean of squared residuals: 827.7931
                    % Var explained: 99.8
```

Pilt 4. Esimese eksperimendi 15 puu väljund.



Joonis 3. Esimese eksperimendi 15 puu ennustustäpsus. y-teljel prognoos, x-teljel tegelik võimsus

Näeme, et ennustustäpsus selliste treening ja testandmetega on küllaltki hea. Keskmine viga on isegi viia puuga vaid ~36 kW, mis on tootmisvõimsuse keskmise ja mediaanväärtusega võrreldes küllaltki väike suurus. Samuti näeme, et ennustustäpsuse erinevused sõltuvad puude arvust juhuslikus metsas. Kümne puuga juhusliku metsa RMSE on parem, kui viie puuga juhuslikus metsas. Samas on paranemine suhteliselt marginaalne. 15 puuga juhusliku metsa RMSE on vaid natuke parem, kui kümne puuga, kuid väga marginaalselt. Arvestades mudeli treenimise arvutuslikku keerukust (ja seeläbi ka mudeli treenimiseks kulunud aega), ei ole 15 puuga juhusliku metsa treenimisele kulunud ressursid õigustatud võrreldes 10 puuga juhusliku metsaga. Tulemused on pea et identsed. Võib ka mainida, et variatsiooni seletab treenitud mudel väga ilusti, lausa ~99,8%.

Teise eksperimendina proovime, kui täpne on mudel, mis on treenitud 2021. aasta kõikide turbiinide andmete peal, kui sellega proovida ennustada ühe turbiini 2022. aasta tootmisvõimsust. Selle mudeli treenimisel kasutame vaid kümne puuga juhusliku metsa mudelit, lähtuvalt eelmise

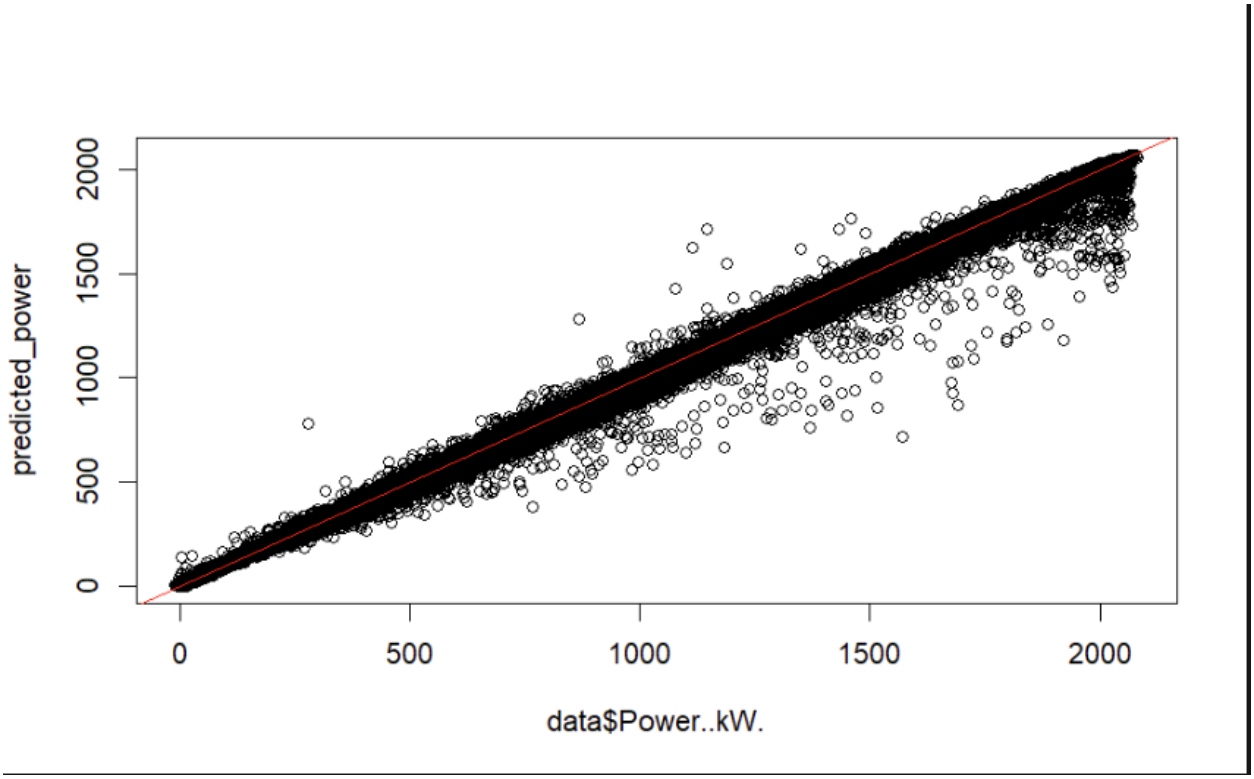
eksperimendi tulemustest. Järgnevalt on esitatud joonised ja pildid eksperimendi käigus treenitud ja testitud juhusliku metsa mudeli karakteristikutest.

```
[1] "RMSE"
[1] 47.86893

Call:
 randomForest(formula = formula, data = train_data, method = "parRF",      ntree =
ntree)
               Type of random forest: regression
                     Number of trees: 10
No. of variables tried at each split: 3

          Mean of squared residuals: 952.775
                    % Var explained: 99.77
```

Pilt 5. Teise eksperimendi juhusliku metsa väljund.



Joonis 4. Teise eksperimendi juhusliku metsa ennustustäpsus. y-teljel prognoos, x-teljel tegelik võimsus

Teise eksperimendi käigus oli treeningandmestik sama, mis esimese eksperimendi korral. Erinev testimisandmestik andis aga hoopis halvema tulemuse. Selles eksperimendis kasutatud treeningandmestik on lähedasem olukord sellele, kuidas meie näeme treenitavat mudelit reaalses

kasutuses. Seetõttu on RMSE, keskmine viga ~47,87 kW reaalsem hinnang mudeli täpsusele. Seletatud variatsiooni osakaal säilib suur, lausa 99,77%.

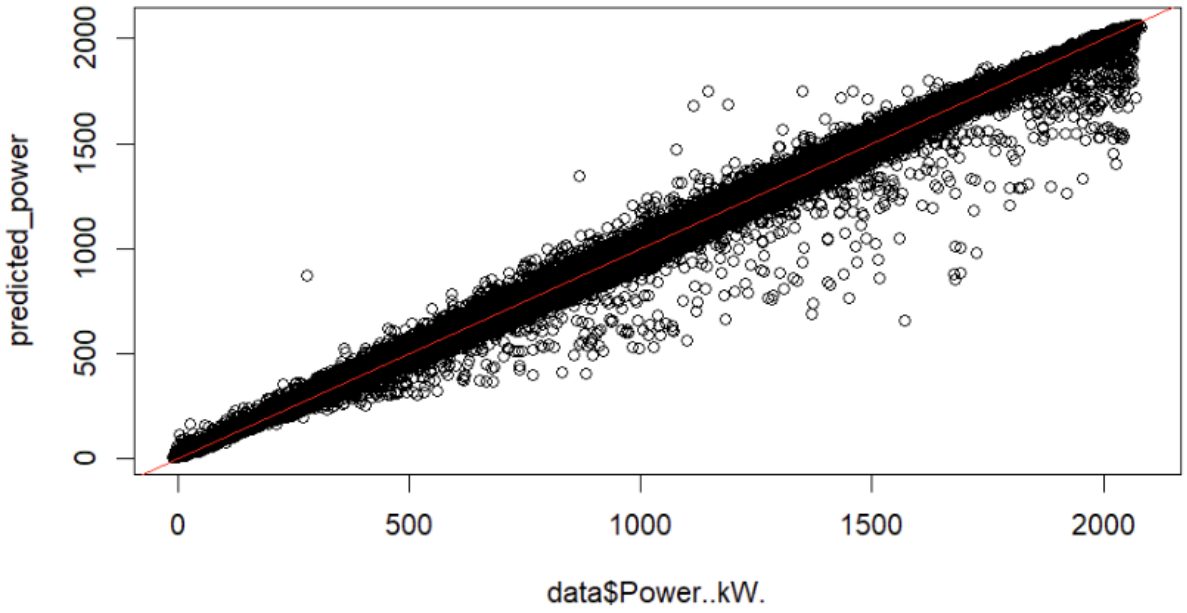
Kolmanda eksperimendina proovime parandada teises eksperimendis saadud tulemust sellega, et kasutame ühe konkreetse tuuliku andmeid mudeli treenimiseks ning ennustame selle sama tuuliku tootmisvõimsust. Kasutame treenimiseks turbiin 1 2021. aasta andmeid ning testimiseks turbiin 1 2022. aasta andmeid. Taaskord kasvatame juhuslikus metsas kümmet puud. Järgnevalt on esitatud joonised ja pildid eksperimendi käigus treenitud ja testitud juhusliku metsa mudeli karakteristikutest.

```
[1] "RMSE"
[1] 49.11983

Call:
 randomForest(formula = formula, data = train_data, method = "parRF",      ntree =
ntree)
               Type of random forest: regression
                     Number of trees: 10
No. of variables tried at each split: 3

          Mean of squared residuals: 663.9022
                    % Var explained: 99.84
```

Pilt 6. Kolmanda eksperimendi juhusliku metsa väljund.



Joonis 5. Kolmanda eksperimendi juhusliku metsa ennustustäpsus. y-teljel prognoos, x-teljel tegelik võimsus

Näeme, et kui ühe kindla tuuliku 2021. aasta andmetel treenida mudel ning testida sama tuuliku 2022. aasta andmetel, on selle mudeli ennustustäpsuse hinnang nõrgem, kui mudeli puhul, kus treenimiseks kasutati kõikide tuulikute andmeid. Keskmine viga on kolmanda eksperimendi puhul lausa ~49,12 kW.

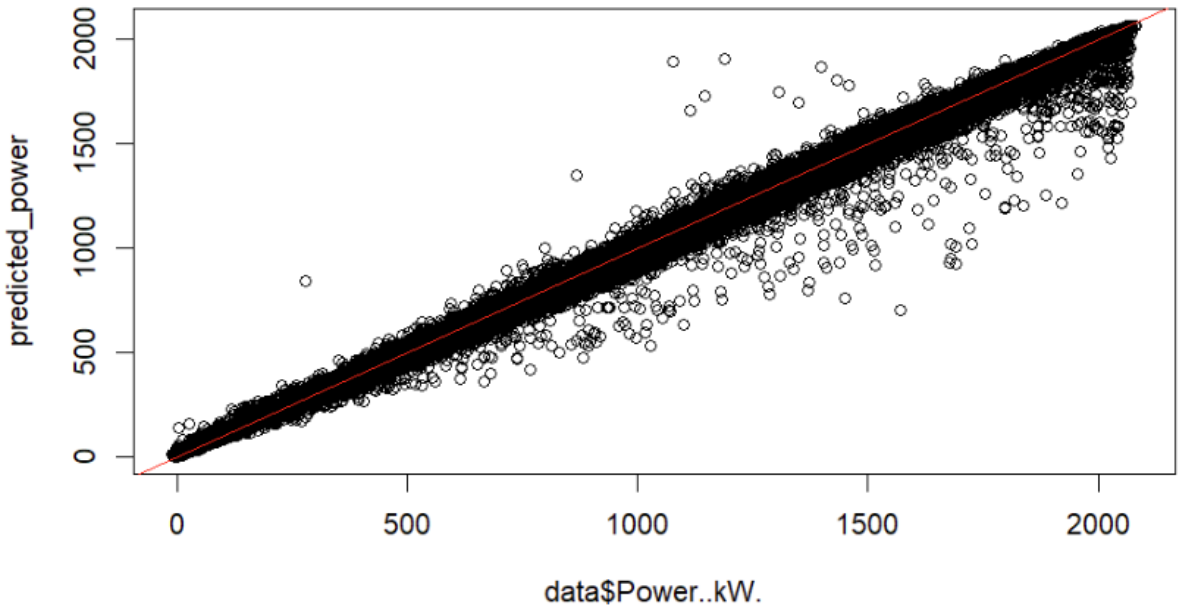
Neljanda eksperimendina proovime, kas kolmandas eksperimendis treenitud mudeli ennustustäpsus paraneb, kui treeningandmestikku suurendada. Proovime treeningandmestikuna kasutada turbiin 1 andmeid aastate vahemikust 2017 kuni 2021 ning kasutame testandmetena taaskord 2022. aasta andmeid. Kasvatame juhuslikus metsas jälle kümme puud. Järgnevalt on esitatud joonised ja pildid eksperimendi käigus treenitud ja testitud juhusliku metsa mudeli karakteristikutest.

```
[1] "RMSE"
[1] 49.20979

Call:
 randomForest(formula = formula, data = train_data, method = "parRF",      ntree =
ntree)
               Type of random forest: regression
                     Number of trees: 10
No. of variables tried at each split: 3

          Mean of squared residuals: 930.3327
                    % Var explained: 99.77
```

Pilt 7. Neljanda eksperimendi juhusliku metsa väljund.



Joonis 6. Neljanda eksperimendi juhusliku metsa ennustustäpsus. y-teljel prognoos, x-teljel tegelik võimsus

Näeme, et suurema hulga andmete kasutamine parandab tulemust, kuid mitte oluliselt. Ühe turbiini 2017 kuni 2021 aasta andmete kasutamine ei tee tulemust paremaks, kui kõikide turbiinide 2021 andmeid kasutades.

4.4 Analüüs

Tulemuste põhjal saavad püstitatud uurimisküsimused järgnevad vastused:

Millised on olulised parameetrid tuuliku energiatootluse ennustamiseks?

Olulisteks parameetriteks osutusid LASSO kasutamise ja valdkonnateadmiste rakendamise tulemusena põhiliselt meteoroloogilised andmed. Lisaks leidsime katsetamise tulemusena, et mõistlik oli kasutada meteoroloogiliste andmete puhul ainult ajaperioodi keskmist ning välja jätta ajaperioodi maksimumid, miinimumid, standardhälbed. Lisaks meteoroloogilistele andmetele on üks oluline tunnus ka kaotatud tootmine. See on oluline tunnus, sest mudel mõistab selle abil, miks tootmine mingil hetkel madalam on, kui teised tunnused samad on.

Üks käesoleva projekti käigus arendatud mudeli puudustest on see, et andmestikus oli saadaolev tootmisvõimekus enamasti konstantne. Väärtus oli kas 0 või 2050 ning vahepealseid väärtuseid oli vähe. Kui seda mudelit soovida kasutada mõne teistsuguse tootmisvõimekusega turbiini jaoks, tuleks seda mudelit vastavalt kohaldada. Samuti, kui turbiin oleks saadaval vaid osaliselt, näiteks poole oma tootmisvõimekusega, siis kuna mudelit on vähe selliste andmetega treenitud, võib prognoos olla seetõttu ebatäpne.

Lisaks avastasime, et hooajalisuse mudelisse kodeerimine ei parandanud tulemust. Päevade ja tundide lisamine mudelisse tegi mudeli ennustustäpsust halvemaks.

Kuidas kasutada ajaloolisi andmeid, et treenida võimalikult täpse ennustusvõimega mudel ning kui täpseks on võimalik mudel teha?

Reaalne kasutusjuht meie mudelile on ühe konkreetse tuuliku tootmisvõimsuse prognoosimine mingite parameetrite juures. Seda arvesse võttes näitavad meie tulemused, et mudelit on mõistlik treenida nii suure koguse andmete peal, kui võimalik (kasutades juhusliku metsa meetodit). Kasutada tuleks tuulepargi kõikide tuulikute andmeid treeningandmetena.

Suurim mudeli täpsus 2022 aasta esimese turbiini andmeid testandmetega kasutades, mis me suutsime saavutada oli RMSE 44,65 kW. Teoreetiliselt on võimalik seda tulemust parandada, kui kasutada treenimisel lisaks 2021 aasta kõikide turbiinide andmetele ka varasemate aastate andmeid. Selle projekti käigus sellist mudelit ei treenitud, sest andmekogused läksid niivõrd suureks, et sülearvutites mudeli treenimine polnud enam mõeldav. Näiteks kõikide turbiinide 2021. aasta

andmete põhjal 15 puuga juhusliku metsa treenimine võttis aega ca 25 minutit, sealjuures ühe tuuliku ühe aasta puhastatud andmed koosnevad ~40K vaatlusest. Kui anda hinnang saavutatud täpsusele, siis oleme tulemusega rahul, kuna 45 kW viga on võrreldes keskmise ning mediaantootmisvõimsusega küllaltki väike. Keskmisest tootmisvõimsusest moodustab mudeli keskmine viga umbes 6,5%.

Peame mainima kindlasti ka seda, et meie mudeli tulemused, mis käesolevas töös on esitatud, eeldavad täiesti täpseid ilmaprognoose. Reaalses elus on ka meteoroloogiliste andmete prognoosides määramatused, mis omakorda võimendavad meie treenitud mudeli eksimist.

Olenemata sellest aitab mudel päris täpselt anda arusaama sellest, kui palju toodangut võib tuulikult oodata.

5 Kasutatud kirjandus

[1] Enefit Green, [Võrgumaterjal]. Available: https://enefitgreen.ee/tuuleenergia/tootmine. [Kasutatud 26 12 2023].

[2] Nord Pool, „REMIT UMM,“ 2023. [Võrgumaterjal]. Available: https://umm.nordpoolgroup.com/#/messages/2d24b020-9602-4add-9f45-98144a6a8410/4. [Kasutatud 26 12 2023].

[3] Vabariigi Valitsus, „Riigi Teataja - Võrgueeskiri,“ [Võrgumaterjal]. Available: https://www.riigiteataja.ee/akt/12831412?leiaKehtiv. [Kasutatud 26 12 2023].

[4] Nord Pool, „Nord Pool Group,“ 2023. [Võrgumaterjal]. Available: https://www.nordpoolgroup.com/en/trading/.

[5] Zenodo, „Kelmarsh wind farm data,“ 2022. [Võrgumaterjal]. Available: https://zenodo.org/records/8252025?fbclid=IwAR3UeYhHWDjhPSzcBbhCsccceO_fPDqzGD4uK7EVU1bEwDGKNZRLS4T95_8.

[6] G. James, D. Witten, T. Hastie ja R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 2023.

[7] V. Spinu, G. Grolemund ja H. Wickham, „lubridate,“ [Võrgumaterjal]. Available: https://lubridate.tidyverse.org/. [Kasutatud 26 12 2023].

[8] A. v. Wyk, „Encoding Cyclical Features for Deep Learning,“ [Võrgumaterjal]. Available: https://www.avanwyk.com/encoding-cyclical-features-for-deep-learning/. [Kasutatud 26 12 2023].

[9] The R Foundation, „What is R?,“ [Võrgumaterjal]. Available: https://www.r-project.org/about.html. [Kasutatud 26 12 2023].

[10] Posit Software, „RStudio Desktop,“ [Võrgumaterjal]. Available: https://posit.co/download/rstudio-desktop/. [Kasutatud 26 12 2023].

[11] L. Breiman, A. Cutler, A. Liaw ja M. Wiener, „randomForest,“ [Võrgumaterjal]. Available: https://cran.r-project.org/web/packages/randomForest/index.html. [Kasutatud 26 12 2023].

[12] Stanford, „Lasso and Elastic-Net Regularized Generalized Linear Models,“ [Võrgumaterjal]. Available: https://glmnet.stanford.edu/. [Kasutatud 26 12 2023].

[13] H. Wickham, „plyr: Tools for Splitting, Applying and Combining Data,“ [Võrgumaterjal]. Available: https://cran.r-project.org/web/packages/plyr/index.html. [Kasutatud 27 12 2023].

[14] G. James, D. Witten, T. Hastie ja R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2023.