TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Vladislav Zaitsev 192296IABM

CFB katla õigeaegne lekke tuvastamine masinõppe abil

Magistritöö

Juhendaja: Olga Ruban
PhD

Tallinn 2021

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Vladislav Zaitsev

04.04.2021

Annotatsioon

Lõputöös kirjeldatakse kogu protsessi alates vajalike andmete otsimisest ja töötlemisest kuni masinõppe mudeli treenimiseni, et vähendada katla lekete õigeaegsest avastamisega Enefit Power AS ettenägematuid kulusid. 215 MWe plokis, mis sisaldab kaht tsirkuleeriva keevkihiga (CFB) katelt koos ühe vaheülekuumendiga oli perioodil 01.2020-03.2021 fikseeritud 7 ootamatut seisakut auru-veekontuuri lekke tõttu esimeses katlas. Ühe lekke õigeaegne ennustamine säästab kulutusi kuni 58000 €. Masinõpet kasutatakse kahes etapis. Otsustusmetsa mudel taastab ja täidab puuduvad või käsitsi valesti andmestikku sisestatud ajaloolised andmed täpsusega 97%, Logit mudel ennustab leket täpsusega 81%. Eesmärgi saavutamiseks kasutatakse standardseid statistika ja masinõppe meetodeid IT-vaatenurgast, süvenemata protsessi tehnoloogiasse.

Lõputöö on kirjutatud eesti keeles ning sisaldab 8 peatükki, 17 joonist, 6 tabelit.

Abstract

Early detection of leaks in CFB boiler using machine learning.

The object of the research is the Enefit Power unit containing two boilers with circulating fluidized bed with one intermediate heater. Full unit capacity is 215 MW. In the period 01.2020-03-2021 in the first boiler there was 7 unexpected shutdowns due to leaks in the steam-water circuit. The aim of the work is to develop a machine learning model that can be used for timely prediction of boiler leaks to prevent or to minimize financial losses of production. To achieve this goal, standard statistical and machine learning methods are used. Preparing data and training machine learning from an IT perspective without going deep into technological processes understanding. This kind of approach should help find irrational dependencies and train the model with better accuracy. In addition, it is necessary to solve the problem when there is too less data for training models, and manual records of an existing leak cannot be used as 100% reliable source. Used machine learning model algorithm which is suitable for integration into the existing system process control. RW describe necessary tools, programming language and machine learning algorithm selection suitable for this specific research.

Predicting one leak in time will save up to €58,000 in costs. Machine learning is used in two stages. The Random Forest model restores and fills in missing or incorrectly manually entered historical data in the dataset with an accuracy of 97%, the Logit model predicts a leak with an accuracy of 81%. To achieve this objective, standard statistics and machine learning methods are used from an IT perspective without delving into process technology. The achieved result exported into model formula, consisting of simple mathematical operations ready to import into existing control system.

The thesis is in Estonian language and contains 30 pages of text, 8 chapters, 17 figures, 6 tables.

Lühendite ja mõistete sõnastik

CFB	(ingl Circulating fluidized bed) tsirkuleeriv keevkiht
Enefit Power AS	Eesti suurenergeetika ettevõte
Valmet	Soome automatiseerimise teenuste pakkuja
Valmet DNA	Valmet hajutatud juhtimissüsteem
DNAHistorian	Valmet DNA andmebaas
Python	üldotstarbeline interpreteeritav programmeerimiskeel
R	statistilise andmetöötluse ja graafika programmeerimiskeel
Logit	Logistiline regressioon, masinõppe algoritm
Otsustusmets	Masinõppe algoritm, mis kuulub ansambelõppe meetodite hulka
TVM	(ingl SVM) Tugivektor-masin masinõppe algoritm
KKS	(saksa - Kraftwerk-Kennzeichensystem) elektrijaama identifitseerimise süsteem
STEP	(ingl Stepwise regression) regressioonimudeli sobitamise meetod
AIC	Akaike information criterion
eksimismaatriks	Maatriks, milles registreeritakse katselistele näidetele mingi reeglistiku rakendamisel saadavate õigete ja väärade liigitusjuhtude arv
OOB	(ingl Out-of-bag) vea ennustamise mõõtmismeetod
MSE	mean squared error
ROC-kõver	receiver operating characteristic curve
AUROC	The area under the receiver operating characteristic
Nordpool	ettevõte, mis peab Norra, Taani, Rootsi, Soome, Eesti ja Leedu ühist elektribörsi

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Uurimisobjekt on Enefit Power plokk, mis sisaldab kahte tsirkuleeriva keevkihiga (CFB) katelt koos ühe vaheülekuumendiga. Ploki brutovõimus on 215MWe. Perioodil 01.2020 - 03.2021 fikseeriti esimeses katlas 7 ootamatut seisakut auru-veekontuuri lekke tõttu.

Töö eesmärk on koostada masinõppe mudel, mille abil saab õigeaegselt ennustada leket katlas, et vältida või minimiseerida tootmise rahalist kaotust. Masinõppe mudeli algoritm peab sobima integreerimiseks protsessijuhtimissüsteemi.

Töö koosneb 4 põhiosast.

Metoodika osas uurib autor detailsemalt objekti ja saada olevaid andmeallikaid. Teostatakse tööriistade, programmeerimiskeele ja masinõppe algoritmi valik.

Andmekaevandamise osas kogutakse, töödeldakse, optimeeritakse andmed mudeli treenimiseks, kasutades statistilise analüüsi meetodeid.

Masinõppe mudeli koostamise osas treenitakse ja testitakse lõplik mudel. Tulemusena peab olema eksporditud masinõppe mudeli valem, mis koosneb lihtsatest matemaatikaoperatsioonidest ja toetab uurimisobjekti juhtimissüsteemi.

Analüüsi ja järelduse osas hindab autor treenitud mudeli olulisust, kirjeldab töö käigus leitud tugevusi ja nõrkusi, arvutab majandusliku kasu ja projekti kogumaksumuse. Samuti tuvastab ja juhib tähelepanu ettevõtte nõrkustele masinõppe rakendamise valmisolekus. Koostab edasise töö arenguplaani.

Eesmärgi saavutamiseks kasutatakse standardseid statistika ja masinõppe meetodeid IT-vaatenurgast, süvenemata protsessi tehnoloogiasse. Selline lähenemine aitab leida mitteilmseid sõltuvusi ja treenida hea täpsusega mudelit.

Lisaks lahendatakse töö käigus probleem, kus andmestikus on liiga vähe lähteandmeid mudeli treenimiseks ja olemasoleva lekke juhtumi käsitsi tehtud sisestusi ei saa kasutada usaldusväärse allikana. Antud töös on leitud kõige tugevam masinõppe algoritm ja seda

on kasutatud koos käsitsi järeltöötlusega, et tõsta andmestiku kvaliteeti, leida kõik lekke juhtumid. Isegi need, mis on logides fikseerimata, ja suurendada andmestiku mahtu lõpliku mudeli täpsemaks treenimiseks.

2 Metoodika

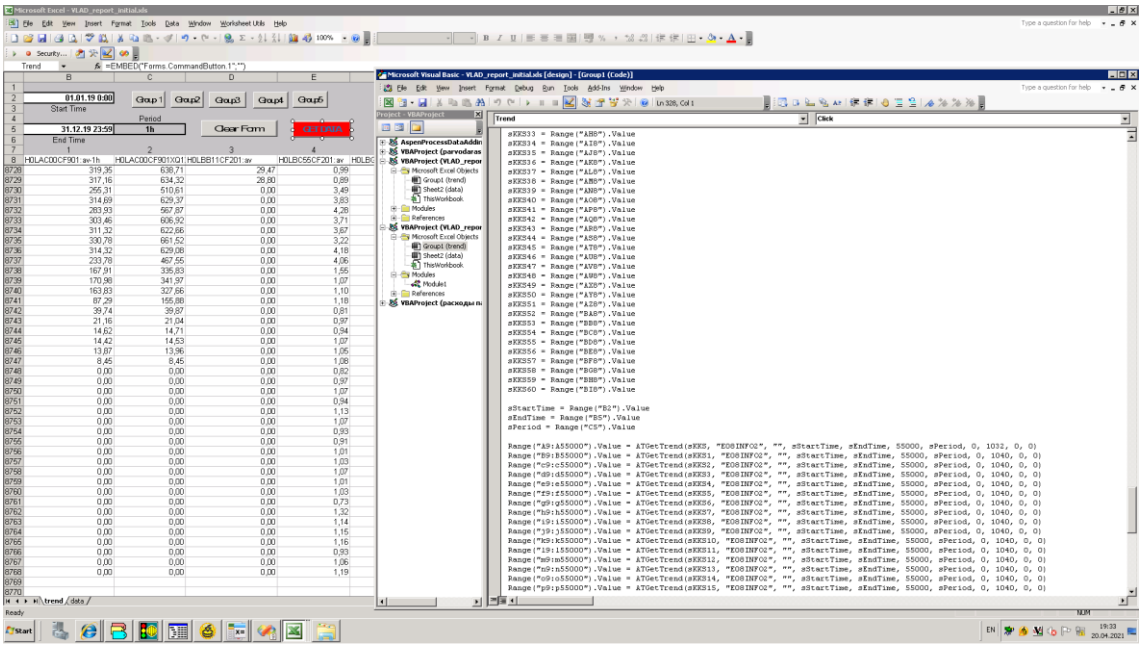
2.1 Objekti kirjeldamine

Uurimisobjekt on Enefit Power elektrijaama tsirkuleeriva keevkihiga plokk, mis sisaldab kahte tsirkuleeriva keevkihiga katelt koos ühe vaheülekuumendiga. Energeetiline osa koosneb kolmeastmelisest auruturbiinist ja vesinikjahutusega generaatorist. Jaama tööd juhitakse vastavalt võrgu vajadustele ning jaam on ettenähtud töötamiseks pideval täiskoormusel või osaliselt koormatuna. Ploki genereeritav projektvõimsus on 215 MW (brutovõimsus). Pärast peatransformaatorit kujutab endast võrku antav energia kolmefaasilist 50 Hz vahelduvvoolu nominaalpingega 330 kV. Kütusena kasutakse Eesti põlevkivi.

Ploki kaks katelt töötavad paralleelselt ja on ühendatud paralleelselt. Toitevesi jaotatakse kummalegi katlale pärast kõrgrõhu eelsoojendeid ning katelde aurutsüklid võivad töötada üksteisest sõltumatult. [1]

2.2 Andmeallikas ja andmed

Uuritav plokk juhitakse *Valmet DNA* automaatjuhtimissüsteemi abil ja kõik olulised ajaloo andmed säilitatakse *Valmet DNAHistorian* serveris. Andmebaas on koostatud *AspenTech InfoPlus.21* baasil. Vajalikud uurimisandmed on välja võetud andmebaasist kasutades EXCEL *Aspen Process Data Add-In* tarkvara. Töö lihtsustamiseks on koostatud Valmet DNA-s Excel-i Macrod. Joonis 1 illustreerib andmekaevandamise protsessi.



Joonis 1. Andmekaevandamine Valmet DNA ajalooserverist Excel makro abil.

Uuritava objekti lekke ajaloo juhtumid aastatel 2019 - 2021 on leitud lähteandmete läbitöötamisel käsitsi. Lähteandmetena on kasutatud seadmete olekulogi, mida täidab vahetuse ülem ja operatiivpäevikut, mida täidab katla-turbiini seadmete vahetuse vanem.

2.3 Programmeerimiskeele valik

Mudeli treenimise programmeerimiskeelena valiti kaks kõige populaarsemat programmeerimiskeelt statistiliseks analüüsiks, andmetöötluseks ja masinõppeks avatud lähtekoodiga: R ja Python, mis asuvad vastavalt 2021. aasta aprilli 1. ja 7. rea programmeerimiskeele populaarsuse edetabeli 1. ja 7. rea vahel. Joonis 2 kajastab keele üldist populaarsust edetabelis kõigis IT-valdkondades, mitte ainult statistilises analüüsis või masinõppes. [2]

Worldwide, Apr 2021 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	29.5 %	-1.0 %
2		Java	17.51 %	-0.6 %
3		JavaScript	8.19 %	+0.2 %
4		C#	7.05 %	-0.2 %
5	↑	C/C++	6.73 %	+1.0 %
6	↓	PHP	6.23 %	+0.0 %
7		R	3.86 %	+0.0 %
8		Objective-C	2.77 %	+0.3 %

Joonis 2. *PopularitY of Programming Language*. [2]

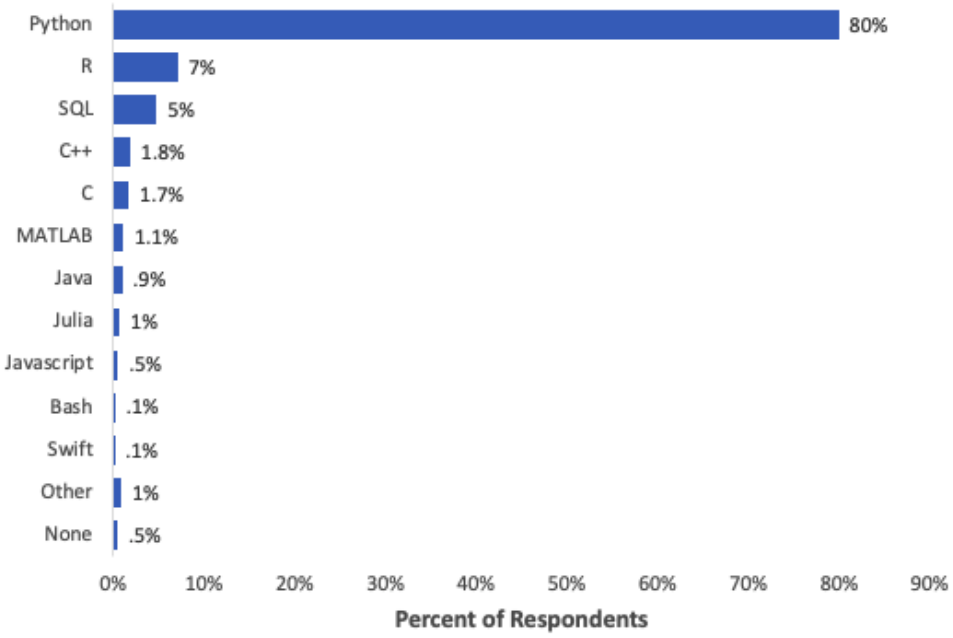
Python on üldotstarbeline programmeerimiskeel, mis sobib lisaks statistilisele analüüsile ja masinõppe ülesannetele paljude ülesannete täitmiseks. Python on populaarne nii veebirakenduste arendamiseks kui ka andmete töötlemiseks tänu selle jaoks kirjutatud suurele hulgale teekidele. Pythonit on lihtne õppida ja järk-järgult tõsta teadmiste taset. Python toetab piiramatut visualiseerimisvõimalust, kuid graafikameetodid on keerulised. Keerulisele küsimusele saab leida alati vastuse.

R on statistilise andmetöötluse ja graafika programmeerimiskeel. R sobib inimestele, kellel puudub programmeerimise kogemus. R-i teegi CRAN-i hoidlas on üle 10 000 paketi [3]. Suur hulk neist võimaldab lihtsalt visualiseerida mahukat ja hästi loetavat graafikat. R sobib nende ülesannete lahendamiseks, mis ei kuulu statistika ja masinõppe valdkonda. Selle tõttu kasutatakse R-i rohkem teaduslikel eesmärkidel, suur hulk saadaolevaid artikleid ja teadusraamatuid on kirjutatud R-keele kohta.

Arvestades, et mõlemaid keeli saab lihtsalt kasutada *Logit Regressiooni*, Otsustusmetsa ja TVMi (inglise keeles *SVM* - Tugivektor-masinaid) masinõppe mudeli treenimiseks, tuleb otsuse tegemiseks lähtuda programmeerimiskogemusest.

Juhul kui puudub kogemus nii R kui ka Python keeles programmeerimisel, on otstarbekas valida Python, mis annab sama keerukusega rohkem võimalusi.

Joonisel 3 on *Business over Broadway* koostatud ja *Kaggle'i* 2020. aastal küsitletud 20 000 spetsialisti soovituste koondtabel andmeteaduse keeleõppe valikute kohta. Valdav enamus vastanutest (80%) soovitab andmekaevandamise ja masinõppe jaoks esimese programmeerimiskeelena Pythonit. [4] [5]



Joonis 3. Professionaalne arvamus esimese programmeerimiskeele kohta andmeteaduses. [5]

Teadustöö kirjutamiseks valib autor statistilise andmetöötluse ja graafika programmeerimiskeele R. Autoril puudub kogemus Python keeles programmeerimisel, kuid on üheaastane R-keeles programmeerimise kogemus TalTech äriinfotehnoloogia magistriõppe raames.

2.4 Arengukeskkonna valimine

R-programmeerimiskeele jaoks on arenenud lähtekoodiga RStudio arenduskeskkond, mis sisaldab konsooli, koodi süntaksit ja valmis tööriistu visualiseerimiseks ja silumiseks.

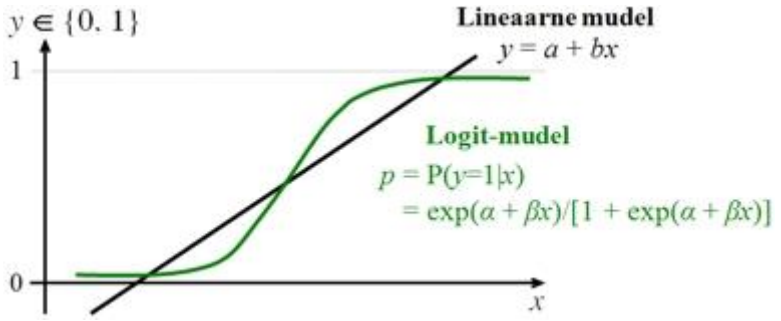
Keskkonda on integreeritud ka R-keele tugi ja juhised ning süntaks *Markdown*, mis hõlbustab oluliselt andmetöötluse ja masinõppe kirjutamist ja publitseerimist. [6]

Seoses RStudio laia funktsionaalsusega ja kasutusmugavusega ei ole põhjust kasutada kolmandat arengukeskkonda. Teistes teadustöödes kasutatakse maailmas valdavalt R-keelt koos RStudio arengukeskkonnaga.

2.5 Klassifitseerimise mudeli valik

Binaarse klassifikatsiooni jaoks, kus kategooria valimiseks on ainult kaks väärtust, on kõige sobivamad masinõppe mudelid: Logit Regressiooni, Otsustusmets ja Tugivektor-masinad [7]. Töö eesmärk on saavutada masinõppe valem, millega saab tulevikus programmeerida olemasolevat juhtimissüsteemi Valmet DNA. Kuna programmeerivate loogikakontrollerite võimalused on piiratud ja programmeeritakse vastavalt IEC 61131-3 standardile, on parim lahendus kasutada Logit mudelit, mida on võimalik lihtsalt sisse ehitada Valmet DNA loogikasse, kasutades MATH funktsioonplokki [8].

Logistiline regressioon (ingl *logistic regression*) või üldisemalt logistiline mudel ehk logit-mudel prognoosib uuritava sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt pideva argumenttunnuse väärtuse muutumisest. [9]. Joonisel 4 on Logiti ja lineaarse mudeli graafik.



Joonis 4. Logit Mudel versus Lineaarne. [7]

2.6 Mudeli efektiivsuse hinnang

Mudeli efektiivsust hinnatakse kasutades eksimismaatriksit. Eksimismaatriks näitab tabeli 1 kujul eksimuse ja õiget klassifikatsiooni ning aitab visualiseerida mudeli

efektiivsust. Tabeli järgi on kohe nähtav, kui palju on õigesti ja valesti kvalifitseeritud klasse võrreldes tegelikega.

Tabel 1. Eksimismaatriks.

		Tegelik	
		1	0
Mudeli tulemus	1	Õiged “1”(TP)	Valed „1“(FP)
	0	Valed „0“(FN)	Õiged “0”(TN)

Eksimusmaatriksi järgi on võimalik arvutada ja hinnata mudeli erinevaid omadusi. Selleks et omavahel võrrelda erinevaid mudeli tulemusi, arvutab autor mudeli ennustamise täpsuse. Täpsus näitab õige ennustatud ja kogu ennustatud suhet ja arvutatakse valemi järgi:

Täpsus=(TP+TN)/kõik

3 Andmekaevandamine

Võttes arvesse, et uurimistöö eesmärk on koostada ennustav algoritm, mis aitab ennustada katla leket, on otsustatud kasutada mudeli parameetrina kõiki auru ja veekulude mõõtmisi. Kuna auru-veetsükkel ühendab tervet plokki (kaks katelt ja turbiin), oli arvesse võetud mõlema katla ja turbiini kulumõõtmised.

Ploki seadmed on kodeeritud vastavalt Foster Wheeler KKS (*Kraftwerk - Kennzeichen – System*) identifitseerimissüsteemile. Näiteks seade H0LCA40CF201 saab lahti mõtestada: “H”- 8 plokk, „0“- üldine osa, „LCA“- peamine kondensaadi torujuhe, „CF“- vooluhulga mõõtmine, „201“- analoogsignaal. Vastavalt *KKS* kodeerimisreeglitele on üles leitud kõik vee- ja auru vooluhulga mõõtmised, mis salvestatakse *Valmet DNAHistoriani* maski järgi: “H?L*CF*AV“. Algtunnuste nimekirja on lisatud ploki brutovõimsus MW ja mõlema katla peamise auruklapi asend. Auruklapi asendid on olulised, et tuvastada katla seisund. Võimsus on katla üks peaparameetritest, mis suurendab tulevase mudeli kirjeldavat jõudu. Tabelis 2 on esitatud algtunnuste nimekiri.

Tabel 2. Algtunnused.

H0LAC00CF901:av	H1LAE70CF201:av
H0LAC00CF901XQ12:av	H1LAF10CF201:av
H0LBB11CF201:av	H1LAF20CF201:av
H0LBC55CF201:av	H1LBA10CF201:av
H0LBG01CF201:av	H1LBA10CF202:av
H0LBG01CF202:av	H1LBA10CF901:av
H0LBG35CF201:av	H1LBB10CF901:av
H0LCA40CF201:av	H1LCQ10CF201:av
H0LCA40CF901:av	H2LAB80CF201:av
H0LCA45CF201:av	H2LAB80CF202:av
H0LCE10CF201:av	H2LAB80CF901:av
H0LCE20CF201:av	H2LAE50CF201:av
H0LCH10CF201:av	H2LAE60CF201:av
H0LCJ30CF201:av	H2LAF10CF201:av
H1LAB80CF201:av	H2LBA10CF201:av
H1LAB80CF202:av	H2LBA10CF202:av
H1LAB80CF901:av	H2LBA10CF901:av
H1LAE50CF201:av	H2LBB10CF901:av
H1LAE60CF201:av	H2LCQ10CF201:av
H1LBA10AA101:sclose	H2LBA10AA101:sclose
H0MKA__CE203XQ13:av-1h	

RStuudiosse on imporditud andmestik ajavahemikul 01.2020 - 03.2021. Esimese tegevusena kontrolliti ja kustutati andmestikust ära kõik read, kus andmed on puudu ja kus on tunnused, mille kohta andmed ei ole leitavad *DNAHistorianis*. Andmetüübi tunnus kuupäev ja kellaaeg on vormindatud String andmetüübist sobivasse kuupäeva ja kellaaja formaati:

as.POSIXct(CF$date,format="%d.%m.%y %H:%M", tz=Sys.timezone())

On lisatud ajutine kunstlik faktor „*BoilerOFF*“ tunnus, mis näitab katla seisundit. Katelt peetakse seiskunuks, kui peamise auruklapi asend on kinni või kõrgsurve auru tarbimine on vähem kui 27 kg/s. *BoilerOFF* tunnus on leitud, kasutades IFELSE funktsiooni:

BoilerOFF = ifelse((CF$H1LBA10AA101==1 | CF$H1LBA10CF901<=27) ,1,0))

Kustutatud on kõik andmed, kus uuritav katel seisis, ehk „*BoilerOFF*“ tunnus oli võrdne ühega.

Faktor tunnus „*H1LEAK*“, mis kirjeldab lekke olemasolu katlas, on lisatud kasutades elektrijaama operatiivpäeviku andmeid.
Operatiivpäeviku kande näidis 1:

“23.08.2020 13:30 При прослушивании ГХ обнаружен свищ по тракту ВД средний сепаратор ,отм.40-45м) Подозение на свищ: Расход ОП 73кг/сек ;Тпода 865-870 гр.С; загрузка Д-100% ; Подпитка 15 кг/сек (54т/ч), Расход Пара на ЗМ 15т/ч”
Näidis 2 opinfo kanne: *„Неплановый ремонт Расхолаживание 23.08.2020 11:20 06.09.2020 1:00 Свищ в первичном П/П“.*
Neid andmeid täidab mitu inimest ja andmete kvaliteet ja struktuur sõltub erinevatest faktoritest alates inimfaktorist kuni elektrijaama hetkeolukorra ja koormuseni.

Antud töös on kehtestatud järgmised lekke olemasolu kvalifitseerimise põhireeglid:

- „1“ tähendab lekke olemasolu katlas
- Lekke olemasolu loetakse tuvastatuks, kui pärast lekke avastamist läheb katel viivitamatult avariiseiskamisse. Juhtum määratletakse kriitilise lekkena. Leke võis tekkida kolm päeva enne avastamist.

- Lekke olemasolu loetakse tuvastatuks, kui pärast lekke avastamist katel ei lähe avariiseiskamisse ja lekke kõrvaldamiseks planeeritakse lähiajal tavaline seisak. Juhtum määratletakse õigeaegse avastamisena. Leke võis tekkida üks päev enne avastamist.

- Lekke olemasolu loetakse tuvastatuks, kui pärast katla remonti on remondi põhjuseks määratletud leke. Lisatingimuseks on kvaliteetne lekke kõrvaldamine, mille tulemusena katel on töökorras järgmisel 14 töös oleku päeval.

- Lekke olemasolu ei loeta tuvastatuks, kui 14 päeva jooksul pärast remonti ilmub leke uuesti või katel läheb mingil teisel põhjusel plaanivälisesse seiskamisse. See tähendab, et tõenäoliselt oli remont tehtud mittekvaliteetselt ja selle juhtumiga seotud andmeid ei ole võimalik kasutada.

- Kõik kvalifitseerimata ajad loetakse määratlematuks ja kustutakse andmestikust.

Pärast lekke kvalifitseerimist koosneb andmestik:

- 1303 “null” kvalifitseeritud andmetest,

- 844 “üks” kvalifitseeritud andmetest.

Järgmisel etapil on kustutatud andmestikust aja tunnus, et välistada mudelil võimalus otsida sõltuvust ajast. Null dispersiooni lähedal olevate ennustajate tuvastamine. Need on ennustajad, millel on üks ainulaadne väärtus või väga vähe kordumatuid väärtusi (vähem kui 5%). Leitud null dispersiooni ennustajad on kujutatud joonisel 5.

```
##  H0LBG35CF201.av.1h H0LCA45CF201.av.1h H0LCE10CF201.av.1h H2LAE50CF201.av.1h
##  Min.   :0.00000    Min.   :0.00000    Min.   :0.00000    Min.   :0.00000
##  1st Qu.:0.00000    1st Qu.:0.00000    1st Qu.:0.00000    1st Qu.:0.00000
##  Median :0.00000    Median :0.00000    Median :0.00000    Median :0.00000
##  Mean   :0.01194    Mean   :0.00225    Mean   :0.03109    Mean   :0.01121
##  3rd Qu.:0.00000    3rd Qu.:0.00000    3rd Qu.:0.00000    3rd Qu.:0.00000
##  Max.   :2.00000    Max.   :4.73000    Max.   :7.71000    Max.   :1.92000
##  H2LAE60CF201.av.1h H1LBA10AA101.sclose H2LBA10AA101.sclose BoilersOFF
##  Min.   :0.000      Min.   :0           Min.   :0.00000     0:2147
##  1st Qu.:0.000      1st Qu.:0           1st Qu.:0.00000     1:   0
##  Median :0.000      Median :0           Median :0.00000
##  Mean   :0.151      Mean   :0           Mean   :0.03866
##  3rd Qu.:0.380      3rd Qu.:0           3rd Qu.:0.00000
##  Max.   :0.870      Max.   :0           Max.   :1.00000
```

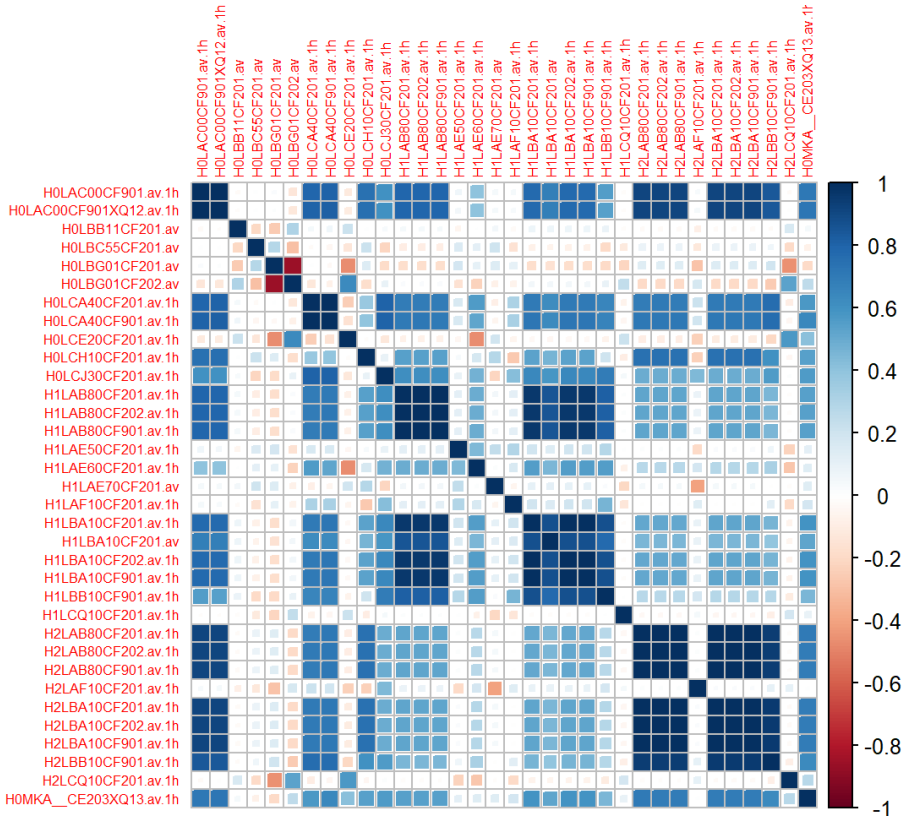
Joonis 5. Nulli lähedased dispersiooni ennustajad

Kogutud andmed on kontrollitud ja sealhulgas on domineerivad teise katla ja üldosa seadmete mõõtmised. Ainulaadset esimese katla „H1LBA10AA101.sclose“ signaali kasutatakse selleks, et defineerida katla tööseisundit. Null dispersiooni lähedaste ennustajate kustutamine tõstab mudeli ennustusvõimekust ja lihtsustab mudeli arvutamise keerukust.

Selleks, et tuvastada erinevate tunnuste vaheline seos ja seose iseloom, on koostatud korrelatsioonimaatriks. Joonisel 7 kuvatakse kõikvõimalike paaride vahelisi korrelatsioone. Töös on maatriksi koostamisel kasutatud lineaarset korrelatsioonikordajat ehk Pearsoni korrelatsioonikordajat, mis arvutatakse valemiga (1) [10].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Mida lähedam on korrelatsioonikordaja absoluutväärtus ühele, seda tugevam on kahe tunnuse seos (Joonis 6).

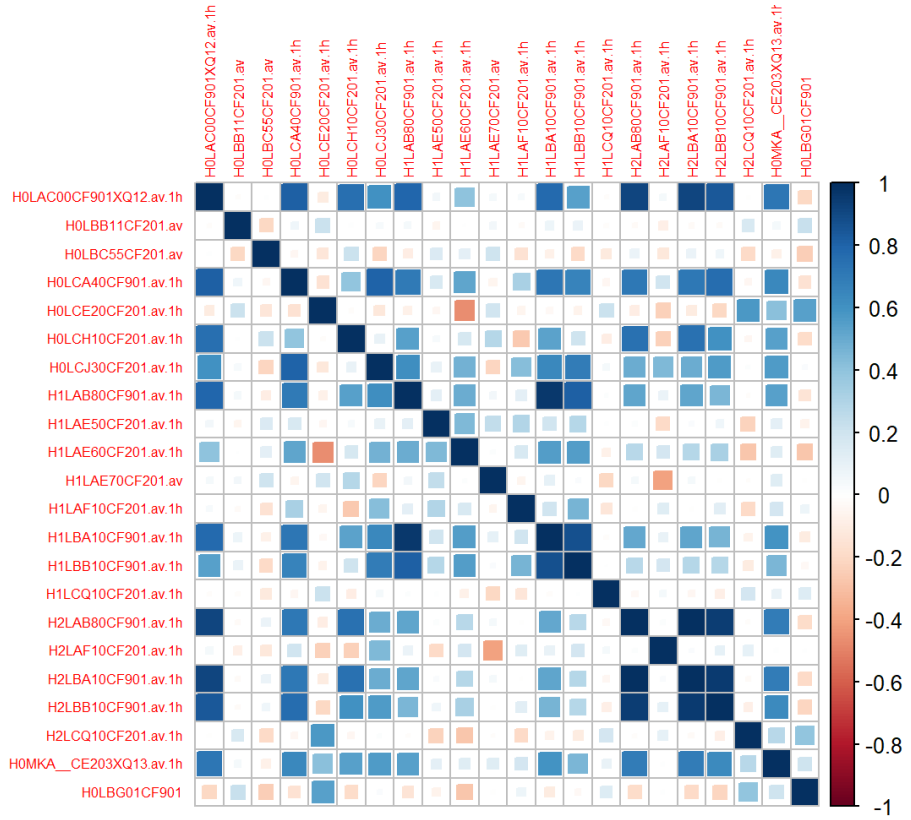


Joonis 6. Korrelatsioonimaatriksi algandmed.

Korrelatsioonimaatriks näitab, et esineb väga suur sõltuvus „CF901“ ja sama algusega „CF20“ vooluhulkade tunnuste vahel. *KKS* reeglite järgi tähendab „CF901“ arvutatud väärtust. Seda tõendab ka automaatjuhtimissüsteemide loogika, kus “CF901” on kahe või kolme füüsilise mõõtmise arvutatud tingimuslik keskmine. Seda skeemi kasutatakse ploki töökindluse tõstmiseks ja nimetakse 2-3st skeemiks või 2-2st skeemiks. Kuna arvutatud 2-3st tunnuse üks omadusest on töökindlus ja kehtetu mõõtmise väärtus on loogika järgi välistatud, on antud töös otsustatud vabaneda multikollineaarsusest, kustutades kõik dubleeritud mõõtmised, millel esineb arvutatud alternatiiv.

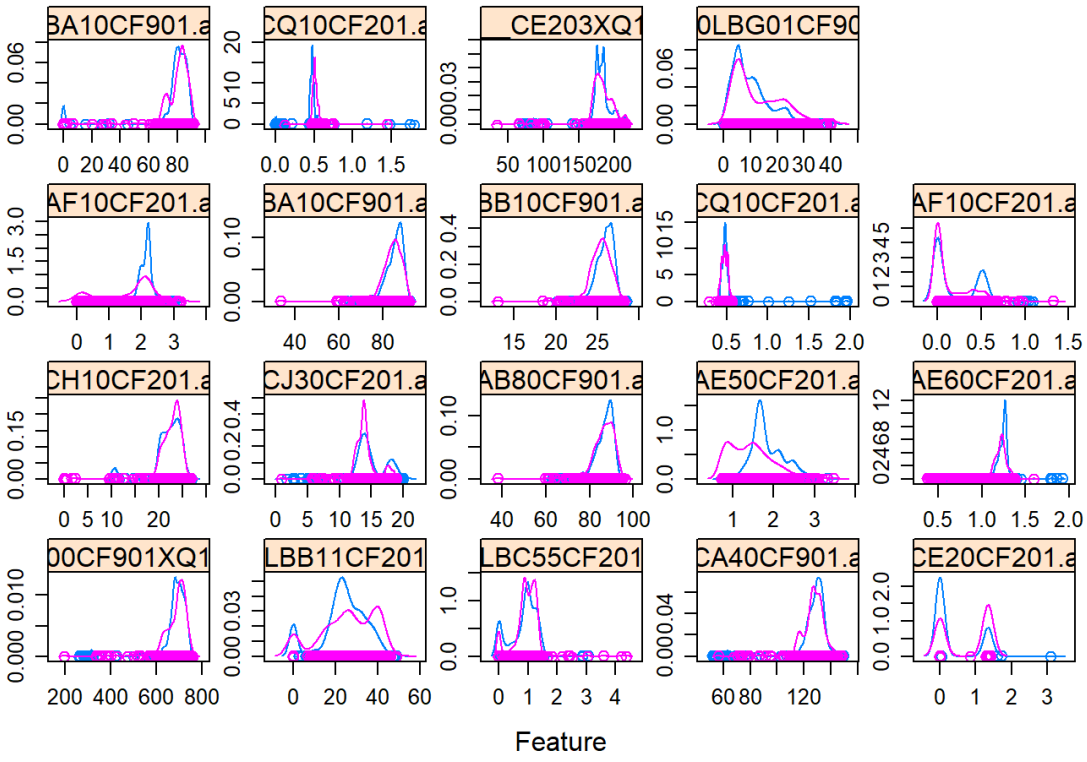
Lisaks sellele esineb suur pöördvõrdeline seos H0LBG01CF201:av ja H0LBG01CF202:av tunnuste vahel. Need tunnused on vastavalt ploki auru eksport ja import. Et vältida seda sõltuvust, on nende tunnuste asemele lisatud arvutatud tunnus H0LBG01CF901, mis võrdub H0LBG01CF201:av ja H0LBG01CF202:av vahega.

Lõppandmestiku korrelatsioonimaatriks on toodud joonisel 7.



Joonis 7. Lõppandmestiku korrelatsioonimaatriks

Jaotuste graafik näitab, et leket ei ole võimalik lihtsalt klassifitseerida kasutades ühte tugeva tunnuse parameetrit. See tähendab, et masinõppe algoritmi kasutamine on õigustatud (Joonis 8).



Joonis 8.Tunnuste jaotuse graafik

4 Testmudeli koostamine

Pärast andmekaevandamist on eraldatud 2020. aasta andmed ja nende põhjal on treenitud LOGIT regressiooni mudelit. Järgmise tegevusena on mudeli andmetele rakendatud STEP R funktsiooni. STEP funktsiooni iga iteratsioon lisab mudelis ühe ja kustutab ühe tunnuse ning võrdleb saavutatud Akaike informatsioonikriteeriumiga (AIC). AIC on hindaja ennustusviga ja näitab antud andmekogumi statistilise mudeli suhtelist kvaliteeti. Mida väiksem AIC, seda paremini on mudel sobitatud. [11] AIC vähenes pärast STEP meetodi kasutamist 960-lt kuni 955-ni ja 22-st tunnusest jäi 19. Tavalise LOGIT mudeli ja STEP mudeli võrdlus on esitatud võrdlustabelina 3.

Tabel 3. LOGIT ja STEP eksimismaatriks koos täpsusega.

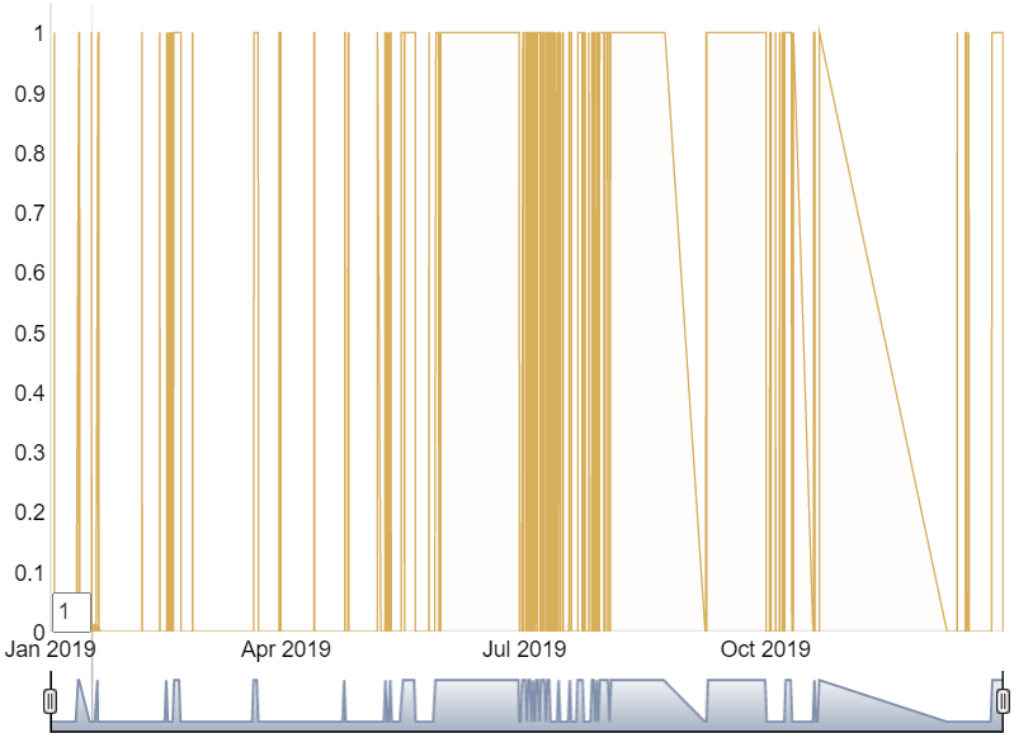
LOGIT		**Tegelik**	
		1	0
Mudeli tulemus	1	1236	67
	0	99	745

STEP		**Tegelik**	
		1	0
Mudeli tulemus	1	1235	68
	0	101	743

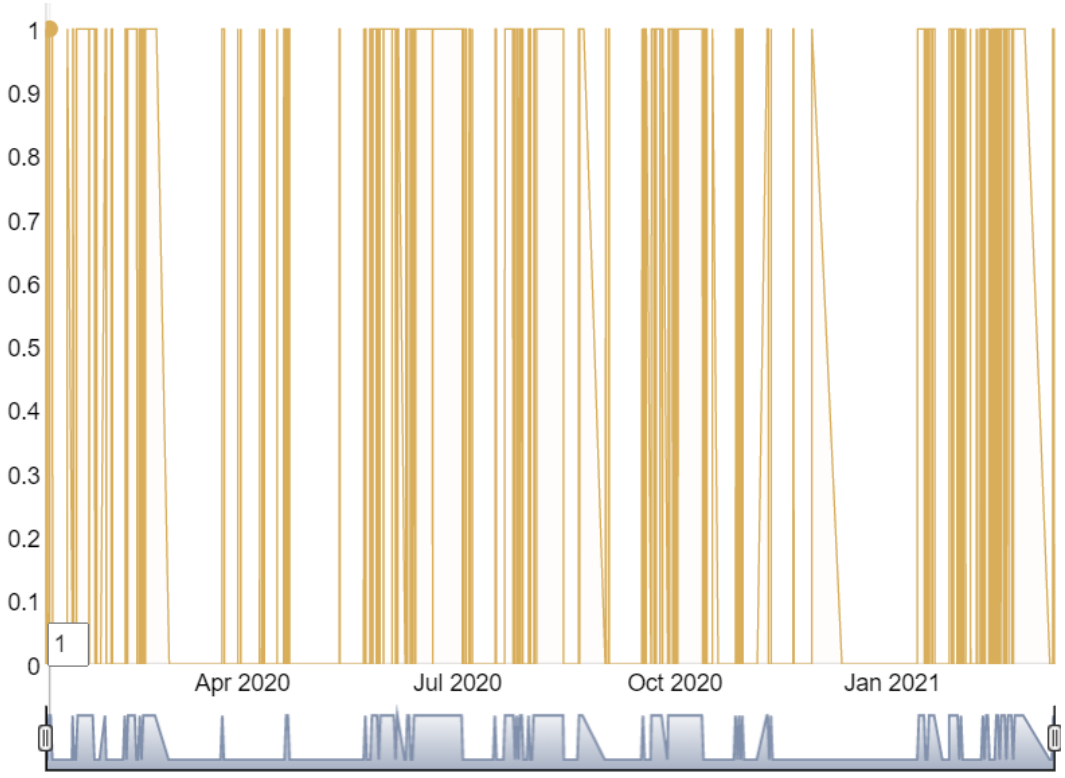
Täpsus:	0.923	0.921

Pärast mudeli optimeerimist STEP meetodiga, on R-meetodi mudeli ennustamisvõime täpsuse parameetri järgi jäänud praktiliselt samale tasemele. Mida väiksem tunnuste arv, seda lihtsam on seda rakendada vastava loogika algoritmis. Sellest lähtudes on antud töös edaspidi kasutatud ennustamiseks STEP mudelit.

Järgnevalt testiti mudelit, kasutades kõiki 2020. aastal saadaolevaid andmeid. Ennustatud lekkeid aastal 2019 ja 2020 näitavad joonise 9 ja joonise 10 graafikud.



Joonis 9. Ennustatud lekke graafik 2019.



Joonis 10. Ennustatud lekke graafik 2020

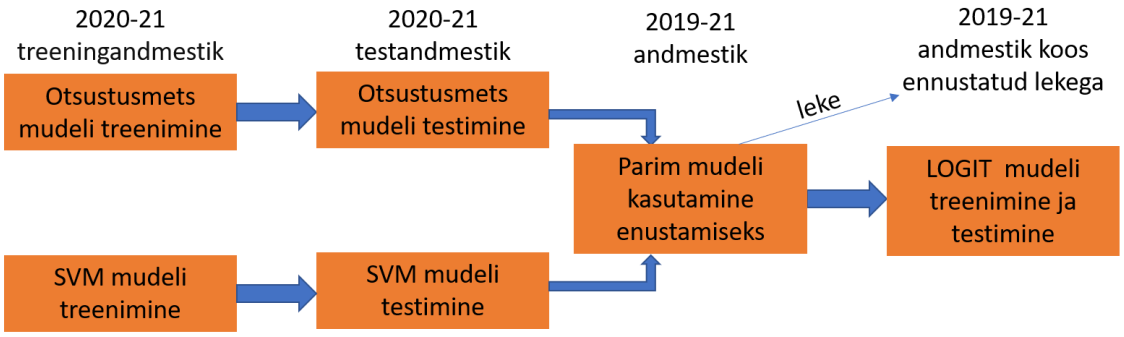
5 Vahepealsed järeldused

Analüüsides mudeliga arvutatud andmeid ja võrreldes neid seadmete olekulogiga, mida täidab vahetuse ülem ning operatiivpäevikuga, mida täidab katla-turbiini seadmete vahetuse vanem, **saab teha** järgmised järeldused:

- Mudel leidis kõik 2019. aastal fikseeritud lekked.
- Mudel leidis 2019. ja 2020. aastate ennustavad lekked, mille kohta ei eksisteeri operatiivteenuse kinnitust.
- Teoreetiliselt on võimalik, et operatiivteenuse poolt logis mitte registreeritud leke oli põhjustatud reguleerimise ventiilide mitte 100% tihedusest või parandatud katelt seiskamata.
- Ei saa välistada ka mudeli eksimuse võimalust. Seda tõendavad ka lühiajalised lekke ennustamise ajavahemikud. Teoorias algab leke väikesest kulust ja kasvab aja jooksul kriitiliste kuludeni. On väga väike tõenäosus, et leke ilmub ja taastub perioodiliselt.
- Kuni 10.2020 on operatiivpäeviku andmetele piiratud ligipääs, mis on seotud tarkvara platvormi uuendusega. Logide kvaliteet ja maht kuni 10.2020 on võrreldes uue versiooniga madalam.
- Olekulogi tabel, kus registreeritakse seadme olek ja seiskamise põhjus aastal 2019, ei ole informatiivne võrreldes aastaga 2020. Tihti on täitmata seiskamise põhjus detailselt ja ei ole arusaadav, kas põhjuseks on katla leke või mitte. Alates 2020. aasta keskpaigast on olukord seoses ettevõtte struktuurimuutustega muutunud palju paremaks.
- Usaldusväärseid andmeid on liiga vähe, et treenida kvaliteetselt katla lekke ennustuse LOGIT mudelit.

Seoses eeltoodud järeldustega otsustati proovida ennustada ajavahemikus 01.2019-03.2020 toimunud lekkeid, kasutades kõige populaarsemaid binaarse klassifikatsiooni mudeleid Otsustusmetsa ja SVMi, valida parim mudel ja kasutada LOGIT mudeli treenimiseks ja testimiseks. Sel juhul välditakse inimtegurit ja käsitsi

sisestamise võimalikku ebausaldusväärsust ning tuginetakse ainult masina intelligentsusele. Joonis 11 illustreerib lõppmudeli ehitusskeemi.



Joonis 11. Lõpliku mudeli ehitusskeem

6 Masinõppe mudeli koostamine

6.1 TVMi ja Otsustusmetsa mudelite võrdlus

Selleks, et omavahel maksimaalselt täpselt võrrelda TVM ja Otsustusmetsa mudeleid, valmistati ette balansseeritud 2020. aasta andmestik. Balansseeritud andmestik on andmestik, kus klassifikaatori variatsioonide kogus on omavahel võrdne. Kõik tunnused on skaleeritud ja tsentreeritud, kasutades scale R funktsiooni. Skaleerimine ei ole oluline Otsustusmetsa mudeli treenimisel võib aga oluliselt tõsta TVM mudeli ennustustugevust. Andmestik oli jagatud kaheks osaks: 80% andmetest on kasutatud mudeli treenimiseks ja 20% mudeli testimiseks ja võrdluseks.

Lõpptulemus on esitatud tabelis 4.

Tabel 4. Otsustusmetsa ja TVMi võrdlustabel.

Otsustusmets		**Tegelik**		**TVM**		**Tegelik**	
		1	0			1	0
Mudeli tulemus	1	162	7	**Mudeli tulemus**	1	143	26
	0	1	168		0	20	149
Täpsus:		0.976				0.864	

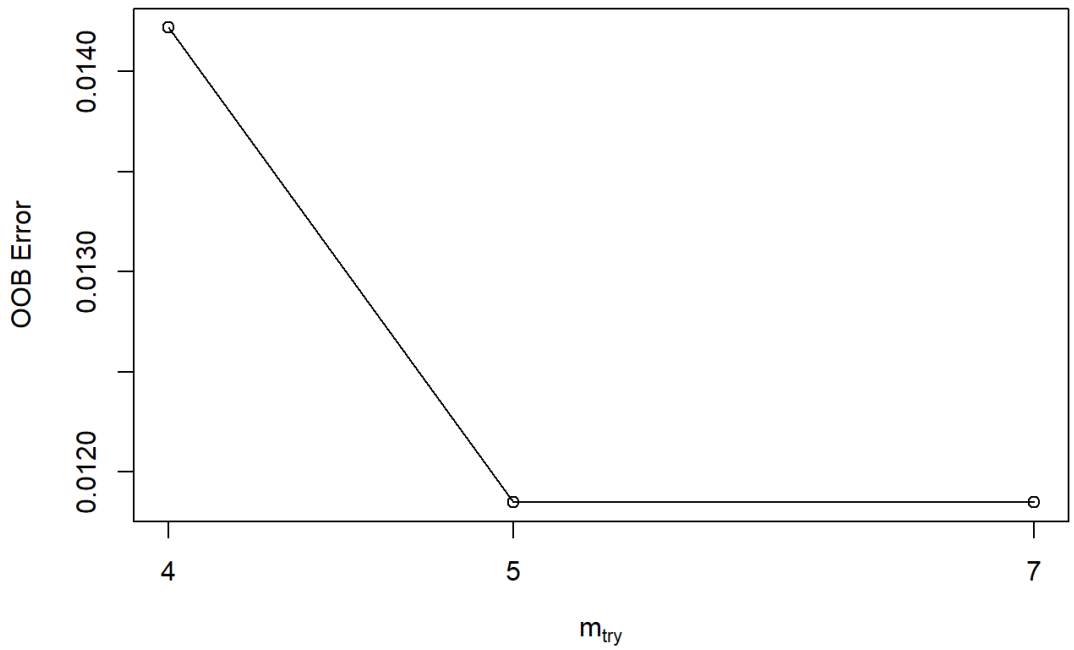
Otsustumetsa mudel annab nende andmetega parema tulemuse kui TVM ja on antud töös valitud edasisteks tegevusteks.

6.2 Otsustusmetsa mudeli optimeerimine

Seoses punktis 6.1 tõendatuga, et Otsustusmetsa mudel annab hea tulemuse täpsusega 97% testandmetel, oli parema mudeli koostamiseks võetud kasutusele kõik balansseeritud andmed aastast 2020 ilma testi ja treenigu andmetele jagamata. Andmestik koosneb 844 reast, kus leke esineb ja 844 reast, kus leke ei esine.

Otsustusmetsa mudelit on võimalik optimeerida kasutades parameetreid Mtry ja Ntree. Mtry näitab igas puusõlmes jagamiseks saadaolevate muutujate arvu. Kasutades tuneRF R funktsiooni funktsiooni oli leitud optimaalne Mtry=5. TuneRF otsib optimaalse Mtry,

võrreldes *Out-of-Bag* veahinnanguga [12]. Joonisel 12 on tuneRF tulemus esitatud graafikuna.

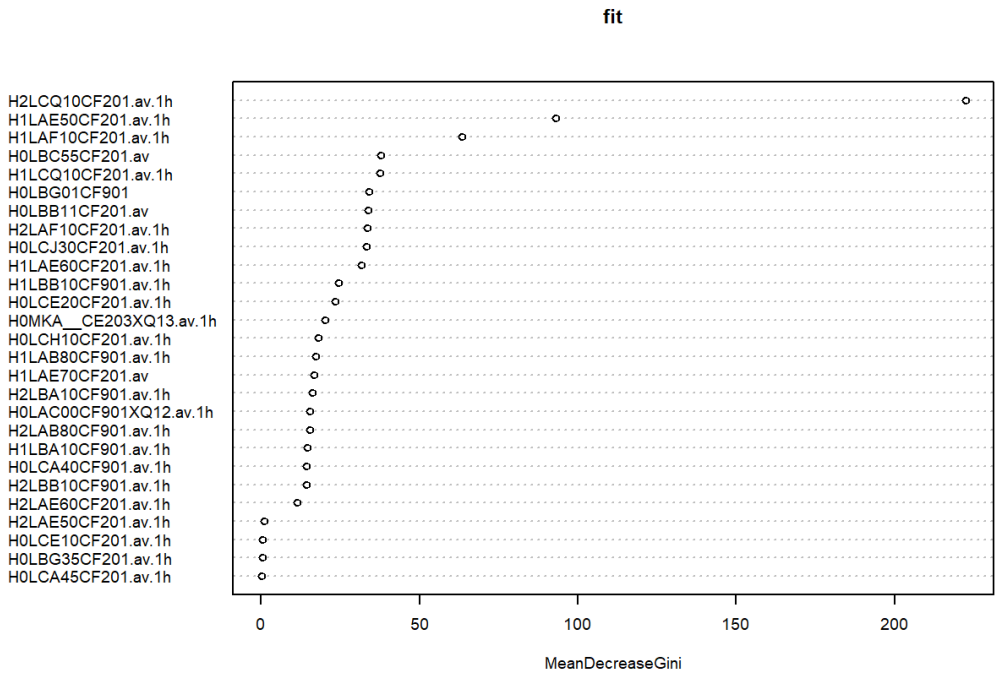


Joonis 12. TuneRF graafik

Ntree näitab kasvavate puude arvu. Optimaalne ntree = 1000 oli leitud valikumeetodina võrreldes mudeli täpsust.

Otsustusmetsa mudel võimaldab tuvastada tunnuste tähtsust ja visualiseerida, kasutades *importance() ja varImpPlot()* R funktsioone.

Esimene mõõt arvutatakse *OOB*-andmete permutiseerimise põhjal: iga puu jaoks registreeritakse andmete *out-of-bag* osa ennustusviga (klassifitseerimise veamäär, regressiooni korral *MSE*). Seejärel tehakse sama pärast iga ennustaja muutuja permutatsiooni. Seejärel arvutatakse nende kahe vahe kõigi puude keskmiseks ja normaliseeritakse erinevuste standardhälbega. Kui erinevuste standardhälve on muutuja jaoks võrdne 0-ga, jagamist ei tehta (kuid keskmine on sel juhul peaaegu alati võrdne 0-ga). Teine mõõt on sõlmede lisandite vähenemine muutujaga jagamisel, mis arvutatakse kõigi puude keskmisena. Klassifitseerimiseks mõõdetakse sõlme lisandit Gini indeksiga. [10] Joonisel 13 on graafik, mis näitab tunnuste olulisust. Mida suurem on „*MeanDecreaseGini*“, seda tähtsam on vastav näitaja.



Joonis 13. Tunnuste tähtsus.

H2LCQ10CF201 tunnus on teise katla mõõde ja tähtsus mudelile, mis ennustab esimese katla vee või auru leket, on liialdatud ning see tunnus on otsustatud kustutada mudeli treenimisest koos nende tunnustega, mille tähtsus on nulli lähedane.

Saadud Otsustusmetsa mudeli täpsus on 97%. Tulemus on esitatud tabelina 5.

Tabel 5. Otsustusmetsa lõpliku mudeli eksimismaatriks

Otsustusmets		**Tegelik**	
		1	0
Mudeli tulemus	1	814	30
	0	15	829

täpsus: 0.973

6.3 Andmestiku taastamine kasutades Otsustusmetsa mudelit

Eelnevalt saadud mudelit kasutati lekete olemasolu ennustamiseks ajavahemiku 01.2019-03.2021 andmete põhjal. Selleks et vabaneda mürast, kus leke ilmub ja kaob lühikese aja jooksul, on otsustatud kasutada 24 tundi liikuvat keskmist. Kõik 24 tunni liikuv-keskmised lekke ennustused, mis on suuremad kui 0,6, võrdsustati ühega ehk leke on

olemas ja ülejäänud nulliga ehk leke puudub. Keskmised ja filtreerimisparameetrid valiti eksperimentaalselt, analüüsides saadud lekkegraafikut ning võrreldes esinemise ja kõrvaldamise aega operatsioonilogiga.

Andmestikku töödeldi võttes arvesse lõigus 3 märgitud analüüsi. Andmestik jagati kaheks osaks. 75% andmetest kasutati mudeli treenimiseks ja jäänud 25% mudeli testimiseks.

6.4 Logit mudeli treenimine ja testimine

Lõpptreenitud Logit mudeli valem on:

H1LEKKE = -0.091570342
*+ H1LAE50CF201.av * -2.594089826 +*
*+ H1LAF10CF201.av * -1.01012989 +*
*+ H0LCJ30CF201.av * 0.267352318 +*
*+ H2LAF10CF201.av * -3.026367989 +*
*+ H0LBC55CF201.av * 0.487359665 +*
*+ H0LBG01CF901 * -0.017021171 +*
*+ H1LCQ10CF201.av * -1.95875418 +*
*+ H1LBB10CF901.av * -0.632698247 +*
*+ H1LAE60CF201.av * 1.052166265 +*
*+ H0MKA__CE203XQ13.av * 0.00096767 +*
*+ H0LC+ H10CF201.av * -0.117488671 +*
*+ H0LCE20CF201.av * 0.495062174 +*
*+ H1LAB80CF901.av * -0.03446414 +*
*+ H0LCA40CF901.av * 0.06180533 +*
*+ H0LAC00CF901XQ12.av * 0.125157848 +*
*+ H2LBA10CF901.av * 0.028837113 +*
*+ H1LBA10CF901.av * -0.285148399 +*
*+ H2LAB80CF901.av * -0.522211265 +*
*+ H2LBB10CF901.av * -0.237361717 +*
*+ H2LAE60CF201.av * -1.152462154 +*
*+ H2LAE50CF201.av * 2.582788892* (2)

Valem (2) arvutab lekke tõenäosuse vahemikus 0-1.

OptimalCutoff() R funktsioon võimaldab arvutada optimaalse tõenäosuse piirväärtuse, lähtudes kasutaja määratletud eesmärgist. Selle uuringu mudeli jaoks on see 0.48. Kui arvutatud valemi (2) järgi on tulemus rohkem kui 0.48, tähendab, et leke on olemas.

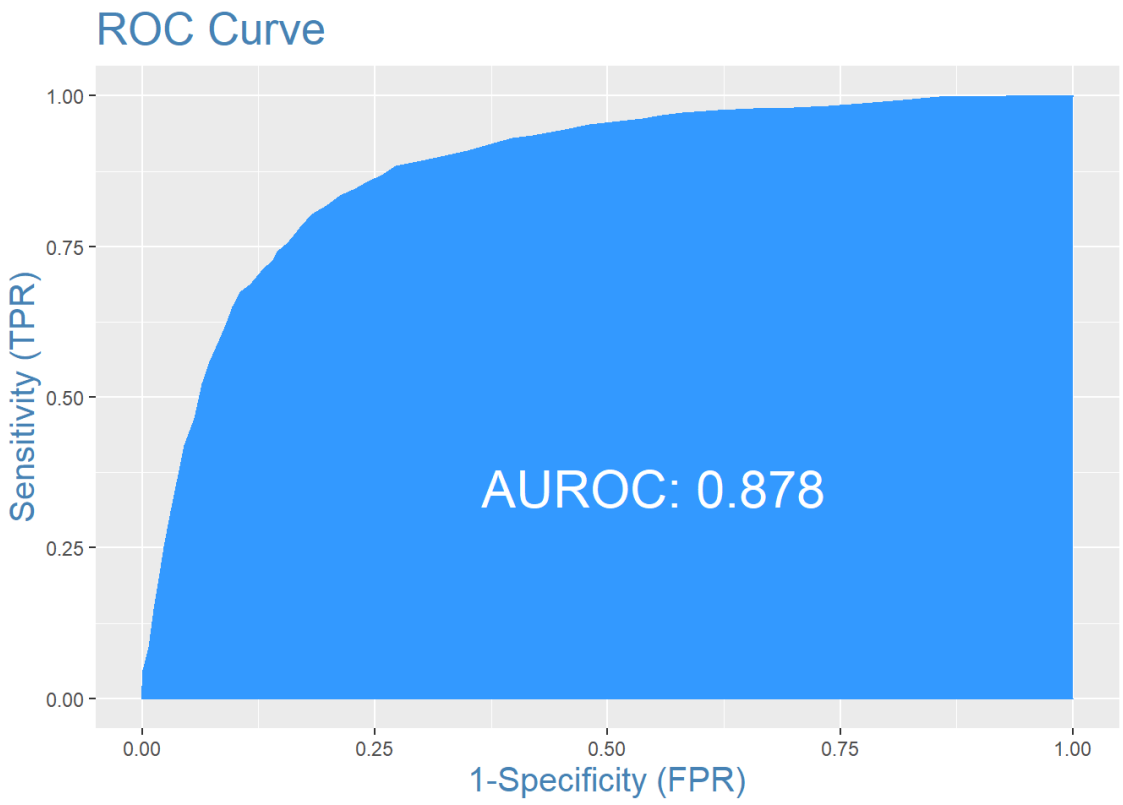
Logit mudeli tulemuse täpsuse eksimusemaatriks on esitatud tabelis 6.

Tabel 6. Lõpliku mudeli eksimusmaatriks ja täpsus.

Logit		Tegelik	
		1	0
Mudeli tulemus	1	1059	230
	0	286	1153

Täpsus: 0.811

ROC-kõvera alune pindala plot joonisel 14 näitab mudeli headuse mõõte. *AUROC* rohkem kui 0.8 hindab mudelit heaks. [14]



Joonis 14 Logit mudeli *ROC* diagramm

7 Analüüs ja järeldus

Lõputöös on saavutatud põhieesmärk - koostada masinõppe mudel, mis ennustab lekke olemasolu katlas õigeaegselt ning hea täpsusega. Üks tingimus oli luua mudel, mida saaks tulevikus sisse ehitada Valmet DNA automaatjuhtimissüsteemi loogikasse vastavalt IEC 61131-3 standardile. Antud töö põhineb oletusel, et katla leket on võimalik ennustada, kasutades kõiki olemasolevaid vee ja auru hetkekulusid, kuna ploki koosseisus olevate katelde 1 ja 2 ning turbiini vee-auru tsükkel on suletud. Eesmärgiks ei olnud tehnoloogia põhjalikum uurimine ja lekke põhjuse leidmine, vaid statistilise analüüsi ja masinõppe meetodite kasutamine lekke ennustamiseks. Valitud andmed lubasid hea, 81% täpsusega ennustada uuritava katla leket. Ennustamise täpsust saab parandada, kui enne andmestiku vormistamist uurida üksikasjalikumalt tehnoloogiat. Koostöös tehnoloogidega võib koostada lekke põhjuste teistsuguse algandmestiku ja ennustamise täpsus läheb paremaks. Mudeli edasiarendamiseks on plaanis lisaks kulumise mõõtmistele lisada andmestikku ka rõhumõõtmised.

Masinõppe mudeli treenimiseks ja testimiseks peab olema piisavalt suur hulk ajaloolisi andmeid, mis kirjeldavad maksimaalselt kõiki võimalikke lekkeid kogu protsessi vältel. Üks peatingimus on, et treenimise andmestik sisaldab katla seisundi kohta ainult 100% usaldusväärset teavet. Uuritava objekti puhul tuvastati kaks tõsist probleemi. Esimene on see, et andmete allika katla seisundi ja seiskamise põhjuste kohta täidavad katla käitajad käsitsi. Lisaks kirjutatakse need andmed vabas vormis ja kasutajasõbralikus keeles. Kõik see takistab andmete kogumist algoritmi abil. Teine probleem on andmete usaldusväärsus. Kasutades käsitsi lisatud andmeid ei saa usaldada, et kirjas on kõik uuritava seisundi juhtumid ja nende algus on õigesti määratud. Lisaks sellele on ootus, et pärast avariiremonti on lekke põhjus kõrvaldatud. Kui leke jääb kvaliteetselt parandamata või on veel lekkeid, mida pole lokaliseeritud, siis sellise juhtumi andmeid ei saa või ei tohi mudeli koostamisel kasutada. Keeruliseks muudab tarkvara platvormi asendamine aastal 2020, mistõttu on kättesaadav ainult piiratud kogus teavet, millest ei piisa hea mudeli treenimiseks. Kõik need tingimused mõjutavad andmestiku usaldusväärtust, selle tulemusena treenitakse mudelit mittekorrektsete andmetega ning väheneb lekke ennustamise täpsus. Selleks, et vältida neid probleeme tulevikus, on kõigepealt vaja automatiseerida olekulogi täitmine. Tuleb kehtestada logide sissekannete tegemiseks

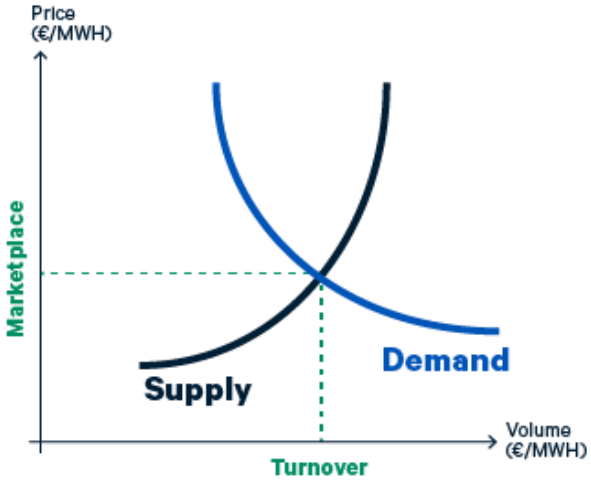
kindel vorm ning lubada ainult loendist valimine. Logis peab olema kindlasti määratud lekke tuvastamise ja kõrvaldamise ajavahemikud ja täpne koht, kus leke oli leitud.

Antud töö eripära on selles, et masinõppe meetodeid on kasutatud mitte ainult lõpliku mudeli loomiseks, vaid ka andmestiku ettevalmistamise etapis. Kõigepealt ennustati kogu ajavahemiku 2019 – märts 2021 lekke olemasolu, kasutades kõige täpsemat mudeli klassifitseerimist. Selles uuringus oli Otsustusmets kõige täpsem mudel (ennustamise tulemus 97%). Otsustusmetsa kasutati puuduvate andmete taastamiseks. Enne andmestiku taastamist oli treenimiseks saadaval vaid 2147 eksemplari. Pärast andmekaevandamist masinõppe abil kasvas eksemplaride arv viis korda ja saavutati 10910 rida. Selline maht lubab jagada andmed treenimiseks ja testimiseks ning veenduda, et kõik uurimisperioodil saadaolevad lekke variatsioonid on mudeli järgi treenitud ja testitud. Tuleb aga arvestada, et kahe masinõppemudeli järjestikune kasutamine vähendab lõplikku täpsust. Lõplik täpsus ei ole midagi muud kui kahe mudeli täpsuse korrutis ja selle konkreetse ülesande jaoks on see 97%. Otsustusmetsa täpsus korrutatud 81% Logit täpsusega annab lõplikuks täpsuseks 79%. Tänu Otsustusmetsa mudeli väga heale täpsusele, langes lõpliku mudeli täpsus ainult 2%, võrreldes teoreetilise andmestikuga, kus oleks sama arv eksemplare 100% usaldusväärsest algallikast.

Saavutatud mudeli täpsus on 79%, mis tähendab, et potentsiaalselt saab mudel õigeaegselt tuvastada neli katla leket viiest. Edasine finantsarvestus põhineb avatud allikate andmetel ning võib tegelikust olukorrast oluliselt erineda. Tootmiskulud ja muud finantsnäitajad on konfidentsiaalsed ja ei ole arvutamiseks kättesaadavad. Enefit Power on NordPool Baltic turu klient. Nord Pool pakub tõhusat, lihtsat ja turvalist järgmise päeva kauplemist Põhjamaades, Baltikumis, Kesk-Lääne-Euroopas ja Suurbritannias [15]. Päev ette kauplemine tähendab, et kuni kella 12-ni fikseeritakse järgmise päeva elektrikogus ja hind. Sel juhul on õigeaegne katla lekke tuvastamine siis, kui leke on avastatud enne kl 12 päeval ja lekke kriitilisus võimaldab katlas edasi toota piisavalt auru, et genereerida ettemüüdud elektrienergiat järgmise päeva lõpuni kuni kella 23:59.

Et arvutada loodud mudeli rahaline kasu, on võetud halvim võimalik teoreetiline stsenaarium ja oletatud, et õigeaegse lekke mittetuvastamise põhjal on üks katel peatatud 11.02.2021 kell 12:05. Oletame, et järgmiseks päevaks on turul fikseeritud ploki täisvõimsus 190MW neto. NordPool andmetel oli keskmine päev ette hind 11.02.2021 99,72 euro/MWh ja 12.02.2021 82,48 euro/MWh. Statistika järgi on ploki

tootmisvõimsus ühe katlaga maksimaalselt 75MW neto. See tähendab, et ettevõte peab puuduva osa 190-75=115MW ostma sel hetkel turult turuhinnaga ja müüma turule tagasi päev ette fikseeritud hinnaga. Niisugused mitteplaneeritud ostud päeva keskel ühe tehinguga vähendavad pakkumist ja tõstavad nõudlust, mis omalt poolt tõstab turul hinda. Pakkumise ja nõudluse sõltuvust näitab joonis 15.



Joonis 15. Pakkumise ja nõudluse kõver. [16]

Oletame, et hind tõuseb 15% päeva ettefikseeritud hinnast. Sel juhul ettevõtte kahjum arvutatakse valemiga 2.

$$Kahjum = \\ = (24t - 12t) * \frac{99{,}72€}{MWt} * 115MWt * 15\% + 24t * \frac{82{,}48€}{MWt} * 115MWt * 15\% \approx \\ \approx 54800\ €$$

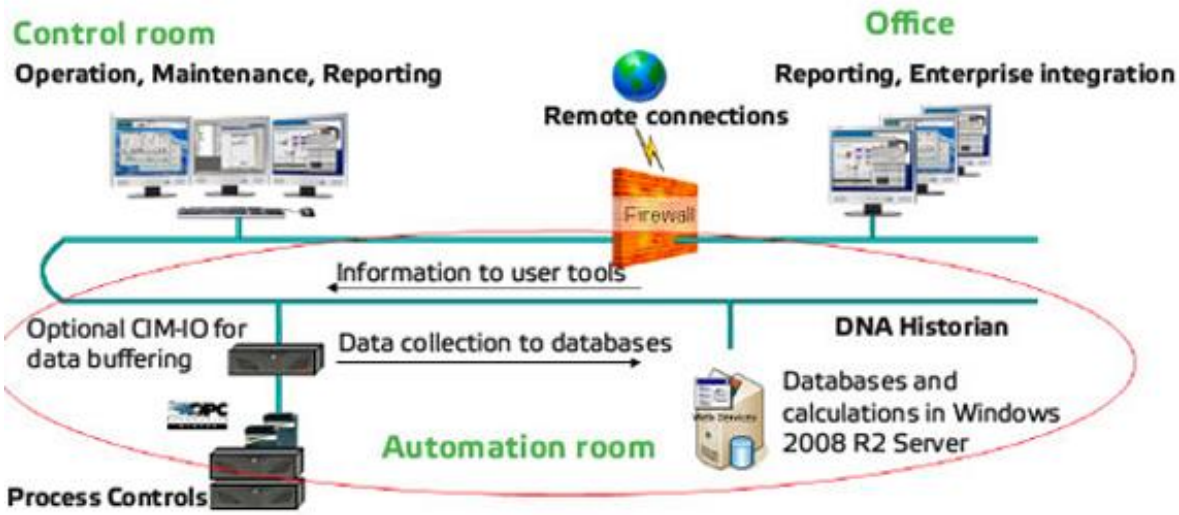
(2)

Kuna saadud mudel ei paranda leket, vaid ainult teavitab ette ja katel on vaja lähiajal remontida, siis saamata jäänud kasum on olemas mõlema olukorra puhul nii koos masinõppe ennustamisega kui ka ilma. Ei võeta arvesse ka tööjõukulude vähenemist, kuna objekti on vaja mõlemal juhul üle vaadata. Otsuse teeb inimene objekti ülevaatuse põhjal, mudel on ainult lisameetod, mis aitab seda otsust teha.

Töö kirjutamiseks on kulunud umbes 500 töötundi, mis vastab 3 kuule esmaspäevast reedeni 8 tundi päevas. Vastavalt palga.ee uuringule 90% R ja Pythoni programmeerijatest saab neto palka 2980 euro/kuus. [17] Tööandja kulud kokku on 5170 euro/kuus. Kulu kokku on 15510 eurot. Kogenud programmeerija, masinõppe ja

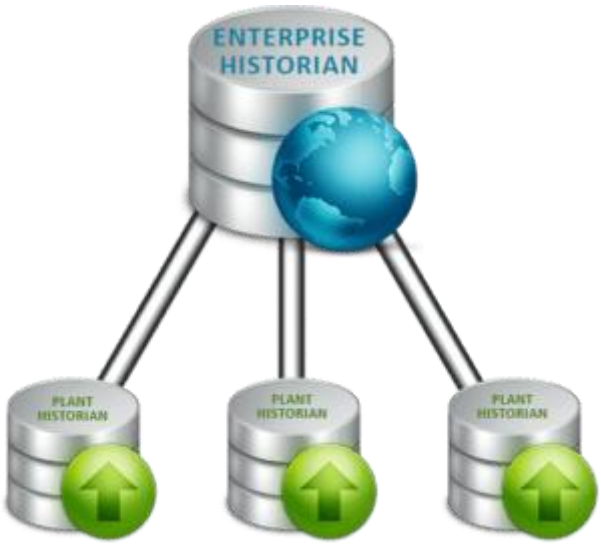
statistikateadlane valmistab nõutud mudeli veel kiiremini. Kulud tasuvad ära juba esimesel lekke ennustamisel.

Esialgne plaan ehitada treenitud masinaõpe mudeli algoritm sisse juhtimissüsteemi kontrollerite tasemele, piiras tõsiselt masinõppe mudeli valikut. Vaatamata sellele, et Logit mudel annab testandmetel hea tulemuse, oleks Otsustusmetsa mudeli kasutamine täpsusega 97% eelistatavam. Autor lähtus oma otsuses seadmete olekust, mille järgi *Valmet DNA* protsessijaam või kontroller on kõige uuem ja võimsam komponent uurimisobjekti juhtimissüsteemil ja on protsessiga otseselt seotud. See tähendab, et ennustamise teave tuleb operatiivselt süsteemihaldurile. Töö käigus on leitud, et õigeaegne lekke tuvastamine tähendab ennustamist vähemalt poolteist päeva enne avariiseiskamist. Sellest võib järeldada, et mudeliarvutuse kasutamine protsessijaama baasil reaalajas ei ole otstarbekas ja jääb piisavalt aega ennustamiseks, kui kasutada masinõppe algoritmi ploki ajalooserveri põhjal *DNA Historian* (joonis 16).



Joonis 16. Valmet DNA andmekogu. [18]

Teine variant, mis toob kaasa lisakulutusi, kuid samas ka võimalusi, on kasutada ettevõtte üldist ajalooserverit *Enterprice Historian.* Näidis ettevõtte ajaloo serverite hierarhiast on toodud joonisel 17.



Joonis 17. Ettevõtte andmeserverite hierarhia. [19]

Ettevõtte andmekogu serverit kasutatakse statistika, ettevõtte KPI arvutuseks ja asub väljaspool protsessivõrku. Sellel tasemel saab võrrelda kaht samasugust plokki omavahel ja kasutada sisseehitatud masinõppe ja statistika meetodeid. Eriti raskete ülesannete lahendamiseks on võimalik ühenduda *Enterprise Historian* masinõppe pilveteenusega. Sellist teenust pakub täna *Google Cloud, Amazon AWS, Microsoft Azure* ja teised.

8 Kokkuvõte

Enefit Power AS plokis, mis sisaldab kaht tsirkuleeriva keevkihiga (CFB) katelt koos ühe vaheülekuumendiga, oli perioodil 01.2020 - 03.2021 esimeses katlas auru/veekontuuri lekke tõttu fikseeritud 7 ootamatut seisakut. Töö eesmärk oli koostada masinõppe mudel, mille abil saab katlas õigeaegselt ennustada leket, et vältida või minimiseerida tootmise rahalist kaotust. Kasutatud oli statistilise andmetöötluse ja graafika programmeerimiskeelt R ja RStudio arengukeskkonda. Andmestiku taastamiseks ja täitmiseks on treenitud Otsustusmetsa masinõppe mudel täpsusega 97%. Leket ennustav lõplik Logit mudel on treenitud täpsusega 81%. Ühe lekke õigeaegne ennustamine poolteist päeva enne avariiseiskamist säästab ootamatuid kulutusi kuni 58000€. Masinõppe edasiarendamiseks ettevõttes on vaja investeerida andmete kogumise parandamisse. Koguda ajalooserverisse eranditult kõik protsessi mõõtmised, parameetrid ja muutujad ning vältida andmebaaside täitmist käsitsi vabas vormis.

Kasutatud kirjandus

[1] F. Wheeler, *KATLA KASUTUSJUHEND. EESTI ELEKTRIJAAMA PLOKK 08,* 2004.

[2] „PYPL PopularitY of Programming Language,“ [Võrgumaterjal]. Available: https://pypl.github.io/PYPL.html. [Kasutatud 09 04 2021].

[3] „Available CRAN Packages By Name,“ [Võrgumaterjal]. Available: https://cran.r-project.org/web/packages/available_packages_by_name.html. [Kasutatud 21 04 2021].

[4] „For Data Professionals, Python Remains Top Programming Language while R Continues to Decline,“ [Võrgumaterjal]. Available: http://businessoverbroadway.com/2021/01/11/for-data-professionals-python-remains-top-programming-language-while-r-continues-to-decline/. [Kasutatud 09 04 2020].

[5] „Kaggle Machine Learning & Data Science Survey,“ [Võrgumaterjal]. Available: https://www.kaggle.com/c/kaggle-survey-2020/data. [Kasutatud 09 04 2021].

[6] „R Studio,“ [Võrgumaterjal]. Available: https://www.rstudio.com/products/rstudio/. [Kasutatud 09 04 2021].

[7] K. J. Max Kuhn, Applied Predictive Modeling, 2013.

[8] 61131-3:2013 International Standart. Programmable controllers – Part 3:Programming languages, IEC, 2013 .

[9] T. Kaart, „Binaarsete tunnuste analüüsimeetodid,“ 2012. [Võrgumaterjal]. Available: http://www.eau.ee/~ktanel/bin_tunnuste_analyys/pt31.php. [Kasutatud 11 04 2021].

[10] A. L. Edwards, An introduction to linear regression and correlation, 1976.

[11] R. McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2019.

[12] G. James, D. Witten, T. Hastie ja R. Tibshirani, An Introduction to Statistical Learning, 2013.

[13] „RDocumentation. importance: Extract variable importance measure,“ [Võrgumaterjal]. Available: https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/importance. [Kasutatud 28 04 2021].

[14] T. Fawcett, „An introduction to ROC analysis,“ 2005.

[15] [Võrgumaterjal]. Available: https://www.nordpoolgroup.com/trading/Day-ahead-trading/. [Kasutatud 09 05 2021].

[16] „Day-ahead market,“ [Võrgumaterjal]. Available: https://www.nordpoolgroup.com/the-power-market/Day-ahead-market/. [Kasutatud 06 05 2021].

[17] [Võrgumaterjal]. Available: https://www.palgad.ee/salaryinfo/infotehnoloogia-it/python-programmeerija. [Kasutatud 06 05 2021].

[18] „Valmet DNA Historian,“ [Võrgumaterjal]. Available: https://www.valmet.com/automation/distributed-control-system/information-management/valmet-dna-historian/. [Kasutatud 08 05 2021].

[19] „Operational Historian vs. Enterprise Historian: What’s the Difference?,“ Parasyn, [Võrgumaterjal]. Available: https://www.parasyn.com.au/article/operational-historian-vs-enterprise-historian-whats-the-difference/. [Kasutatud 08 05 2021].

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Vladislav Zaitsev

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „CFB katla õigeaegne lekke tuvastamine masinõppe abil“, mille juhendaja on Olga Ruban
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

09.05.2021

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.