TALLINNA TEHNIKAÜLIKOOL
INSENERITEADUSKOND
Elektroenergeetika ja mehhatroonika instituut

TALLINNA TEHNIKAÜLIKOOL

ELEKTRITARBIMISEL NING MASINÕPPEL PÕHINEVATE ELAMU KASUTUSE HINDAMISMEETODITE TEHNOLOOGILINE ANALÜÜS

TECHNOLOGICAL ANALYSIS OF RESIDENTIAL BUILDING USAGE ESTIMATION METHODS BASED ON ELECTRICITY CONSUMPTION AND MACHINE LEARNING

BAKALAUREUSETÖÖ

Üliõpilane: Oliver Kütt

Üliõpilaskood 179822EAAB

Juhendaja: Vahur Maask

Tallinn, 2021

(Tiitellehe pöördel)

AUTORIDEKLARATSIOON

Olen koostanud lõputöö iseseisvalt.
Lõputöö alusel ei ole varem kutse- või teaduskraadi või inseneridiplomit taotletud.
Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

"....." ..... 202…..

Autor: .....
/ allkiri /

Töö vastab bakalaureusetöö esitatud nõuetele

"....." ..... 202…..

Juhendaja: .....
/ allkiri /

Kaitsmisele lubatud

"....."..... 202… .

Kaitsmiskomisjoni esimees .....
/ nimi ja allkiri /

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Oliver Kütt

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose Elektritarbimisel ning masinõppel põhinevate elamu kasutuse hindamismeetodite tehnoloogiline analüüs,

mille juhendaja on Vahur Maask,

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.

18.05.2021

---

[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.*

LÕPUTÖÖ LÜHIKOKKUVÕTE

Autor: Oliver Kütt *Lõputöö liik:* Bakalaureusetöö

Töö pealkiri: Elektritarbimisel ning masinõppel põhinevate elamu kasutuse hindamismeetodite tehnoloogiline analüüs

*Kuupäev:*18.05.2021 42 *lk*

Ülikool: Tallinna Tehnikaülikool

Teaduskond: Inseneriteaduskond

Instituut: Elektroenergeetika ja mehhatroonika instituut

Töö juhendaja(d): doktorant-nooremteadur Vahur Maask

Sisu kirjeldus:

Antud töö eesmärk oli teostada elektritarbimisel ning masinõppel põhinevate elamu kasutuse hindamismeetodite analüüs ning rakendamine päriselulistel andmetel. Selleks koguti 18 päeva jooksul andmed neljaliikmelise perekonnaga korteri kasutuse ja elektritarbimise kohta. Kasutuse hindamiseks ühe inimese täpsusega valiti kolm erinevat masinõppe algoritmi: k-lähimat naabrit, otsustusmets ja tugivektor-masin. Valitud algoritme analüüsiti põhjalikumalt ning seejärel loodi kogutud andmestiku abil mudelid kasutuse hindamiseks elektritarbimise põhjal, kasutades Scikit-learn masinõppe teeki. Valideerimisandmetel saavutati mudelite puhul täpsused vahemikus 63-65%. Mudelite täpsus ei erinenud palju, kuid oli selge, et tunnuste nagu elektritarbimise hulk kasutamine tõstis kasutuse hindamise täpsust. Töö edasiarendamisel pakutakse lahendusteks väärtuslike andmete kogumise lihtsustamise, et pikema andmekogumise perioodi abil mudeli treenimise kvaliteeti tõsta.

Märksõnad: elamu kasutus, hindamine, masinõpe, elektritarbimine, algoritmid.

ABSTRACT

Author: Oliver Kütt *Type of the work:* Bachelor Thesis

Title: Technological analysis of residential building usage estimation methods based on electricity consumption and machine learning

Date: 18.05.2021 42 *pages*

University: Tallinn University of Technology

School: School of Engineering

Department: Department of Electrical Power Engineering and Mechatronics

Supervisor(s) of the thesis: early stage researcher Vahur Maask

Abstract:

The aim of this work was to analyse residential building usage estimation methods based on electricity consumption data and to apply them on real world data. For this purpose, an apartment usage and electricity data was being collected during 18 days. The apartment had four inhabitants. Therefore to precisely estimate the number of people present, three machine learning algorithms were chosen: k-nearest neighbors, random forest and support-vector machine. The chosen algorithms were then more thoroughly analysed. Algorithms were implemented with Scikit-learn machine learning library. For created models accuracies between 63-65% were achieved on validation data. The models did not differ by a large margin but it was clear that using electricity consumption data improved the accuracy of the model. For future development a longer data collecting timeframe was suggested to improve the training of the model.

Keywords: residential building usage, estimation, machine learning, electricity consumption, algorithms.

LÕPUTÖÖ ÜLESANNE

Lõputöö teema:	**Elektritarbimise analüüsil põhineva elamu kasutuse ennustusmeetodite tehnoloogiline analüüs**
Lõputöö teema inglise keeles:	**Technological analysis of occupancy prediction methods for residential building by using electrical energy consumption data**
Üliõpilane:	**Oliver Kütt, 179822**
Eriala:	**Elektroenergeetika ja mehhatroonika, elektroenergeetika peaeriala**
Lõputöö liik:	**bakalaureusetöö**
Lõputöö juhendaja:	**Vahur Maask**
Lõputöö ülesande kehtivusaeg:	**16.06.2021**
Lõputöö esitamise tähtaeg:	**18.05.2021**

/Allkirjastatud digitaalselt/	*/Allkirjastatud digitaalselt/*	*/Allkirjastatud digitaalselt/*
Üliõpilane (allkiri)	Juhendaja (allkiri)	Õppekava juht (allkiri)

1. Teema põhjendus

Järjest enam pööratakse tähelepanu hoonete energiakasutusele ja sisekliima kvaliteedile. Hoonete energiatarbimine kokku moodustab Euroopa Komisjoni andmetel üle 40% toodetavast energiast, seega juba vähesel määral hoonete paindlik juhtimine avaldab suurt mõju kogu elektrivõrgule ning üldisele energiatarbimisele. Mida paremini on teada hoone kasutuse profiil, seda täpsemini oleks võimalik rakendada

energiapaindlikku juhtimist hoone tehnosüsteemidel, võimaldades süsteeme näiteks elektrihinna põhiselt juhtida, sealjuures tagades nõutud sisekliima kvaliteedi taseme. Elamute kasutust on mõistlik hinnata elektriarvestitega, sest Eestis on kõik elamud ühendatud elektrivõrguga läbi kaugloetavate arvestite, mistõttu puudub lisaseadmete paigaldamise vajadus. Antud lõputöö teeb ülevaate elektriarvestitega elamute kasutuse hindamise võimalustest ning annab soovituse taolise meetodi rakendamiseks.

2. Töö eesmärk

Töö eesmärgiks on uurida ja katsetada elektritarbimise analüüsil põhinevaid elamu kasutuse ennustamise meetodeid ning anda soovitus sobivaimale meetodile.

3. Lahendamisele kuuluvate küsimuste loetelu:

1) Missugused on tüüpiliste elamute energiatarbimise profiilid?
2) Millised on levinuimad elamu kasutuse ennustusmeetodid?
3) Mille poolest erinevad elamu kasutuse ennustusmeetodid?
4) Mis meetodit oleks mõistlik rakendada reaalsel elamul, kasutades reaalseid elektritarbimise andmeid?

4. Lähteandmed

- Teemaga seotud teadusartiklid ja raamatud
- Elamu tarbimisajalugu Elektrilevi iseteenindusest

5. Uurimismeetodid

Elamute tüüpiliste energiatarbimise profiilide ja elamu kasutuse profiilide uurimiseks kasutatakse kirjandust ja vastavaid standardeid. Elektritarbimisel põhineval elamu kasutuse ennustusmeetodite võrdlus toimub kirjanduse alusel.

Katsetamine reaalsetel andmetel plaanitakse teostada andmetöötluskeskkonnas.

Töö tulemusena oskame valida sobiva meetodi elamu kasutuse ennustamiseks ja hinnata selle rakendatavust.

6. Graafiline osa

Töös kasutatakse tabeleid ja graafikuid tarbimise/kasutuse profiilide kujutamiseks ning ennustusmeetodite iseloomustamiseks. Simuleerimiskeskkonnas katsete jaoks kirjutatud lähtekoodi saab vajadusel välja tuua töö lisades.

Täpsem info tabelite, graafikute ja jooniste kohta saadakse töö käigus

7. Töö struktuur

1. Lõputöö lühikokkuvõte
2. Lõputöö ülesanne
3. Sisukord
4. Eessõna
5. Sissejuhatus
6. Elamu kasutus
7. Elamu kasutuse ennustusmeetodid
8. Meetodite katsetamine reaalsetel andmetel
9. Kokkuvõte

8. Kasutatud kirjanduse allikad

Allikateks on raamatud, teadusartiklid, standardid ning seadusandlikud aktid.

9. Lõputöö konsultandid

Vajadus konsultantide järele selgub töö käigus.

10. Töö etapid ja ajakava

- Elamu tüübid ning nende tüüpilised tarbimise/kasutuse profiilid (25.01.2021)
- Ülevaade ennustusmeetoditest ja nendest peamiste valik (16.02.2021)
- Ennustusmeetodite võrdlev analüüs (16.03.2021)
- Ülevaade simuleerimiskeskkonnast ja katsete ülesehitus (16.04.2021)
- Katsetulemuste analüüs ning rakendatavuse hinnang (26.04.2021)
- Kokkuvõtte koostamine (01.05.2021)
- Juhendajale läbilugemiseks saatmine (01.05.2021)
- Paranduste sisseviimine (07.05.2021)
- Töö lõplik versioon valmis (17.05.2021)

SISUKORD

EESSÕNA

Tänan oma juhendajat Vahur Maaski, kelle poolt oli välja pakutud antud teemal lõputöö koostamine. Tänan ka Kaarel Hendrikut ja tema perekonda, kes lahkelt olid nõus nende liikumise ja elektritarbimise kohta andmete kogumisega.

LÜHENDITE JA TÄHISTE LOETELU

CO_2 – süsinikdioksiid

HVAC – kliimatehnika

PIR – passiivne infrapuna

Wi-Fi – traadita kohtvõrgu seadme tähis

GPS – üleilmne asukoha määramise süsteem

SISSEJUHATUS

Tarbitud energia elamusektoris moodustab Euroopa Liidus energia lõpptarbimisest ligi 26%, millest omakorda 78,1% moodustavad kütmine, jahutus, valgustus, ventilatsioon ja muud elektriseadmed [1]. Kasutades tõhusamalt kulutatavat energiat, väheneb ka koormus keskkonnale ning elamusektori suure mastaabi tõttu omavad suurt mõju ka väikesed energia kokkuhoiud ühe elamu kohta. Samuti annab väiksem energianõudlus paremad võimalused varakult katta nõudlus taastuvenergia tootmistega, et saavutada ambitsioonikad kliimaeesmärgid.

Kui eesmärgiks on elamu energiatarvet optimiseerida tehnosüsteemide targema juhtimise abil, tuleb samal ajal tagada elamu sisekliima kvaliteet. Energiatarbe vähendamine sisekliima kvaliteedi arvelt mõjub halvasti inimeste tervisele ja vähendab nende produktiivsust ehk saavutatud energiavõit ei ole õigustatud. Mida paremini on teada hoone kasutuse profiil, seda täpsemini oleks võimalik rakendada energiapaindlikku juhtimist hoone tehnosüsteemidel, sealjuures tagades nõutud sisekliima kvaliteedi taseme.

Antud töö eesmärgiks on elamu kasutuse hindamine kaudsete meetodite abil, täpsemalt elektritarbimise ja masinõppe meetodite põhjal. See tähendab, et teave kasutuse kohta saavutatakse ilma lisaseadmetesse investeerimata. Hoone kasutuse täpne hindamine annab võimaluse tehnosüsteemide juhtimise sisendites arvestada CO_2 taseme tõusuga, vabasoojustega jm inimestega kaasnevate muutustega.

Töö sisaldab nii elamu kasutuse hindamismeetodite analüüsi kui ka vastavate mudelite rakendamist spetsiifilise elamu näitel. Selleks viidi läbi andmete kogumine 18 päeva vältel ühes nelja elanikuga korteris. Enamik masinõppe probleeme selles valdkonnas on püstitatud binaarsena ehk elamu kasutust vaadeldakse kas hõivatuna või mitte hõivatuna. Selles töös on võetud eesmärgiks täpne inimeste arvu hindamine ehk ülesanne on püstitatud mitmeklassilise probleemina.

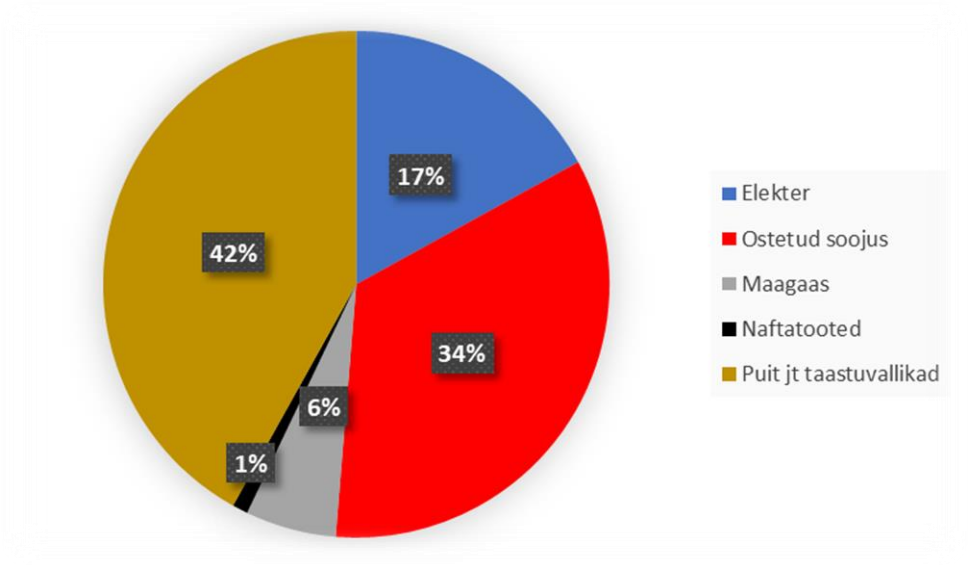
Töö põhiosa algab tüüpilise elamu kasutuse ja elektritarbimise uurimisest tüüpgraafikute näitel. Teises peatükis tutvutakse nii olemasolevate otseste meetoditega kui ka kaudsete meetoditega, mida kasutatakse kasutusandmete kogumiseks. Seejärel tehakse valik kolme masinõppe algoritmi kasuks, tuginedes uuritava elamu andmete iseloomule. Valitud masinõppe algoritmideks on k-lähimat naabrit, otsustusmets ja tugivektor-masin, mis järgmises peatükis rakendati töödeldud andmetel kasutades masinõppe teeki Scikit-learn programmeerimiskeelele Python. Lõpetuseks tehakse järeldused katsetulemusteks ning antakse hinnang võimalike rakendusvaldkondade kohta.

1. TÜÜPILINE ELAMU KASUTUS JA ENERGIA TARBIMINE

1.1 Energia tarbimine elamutes

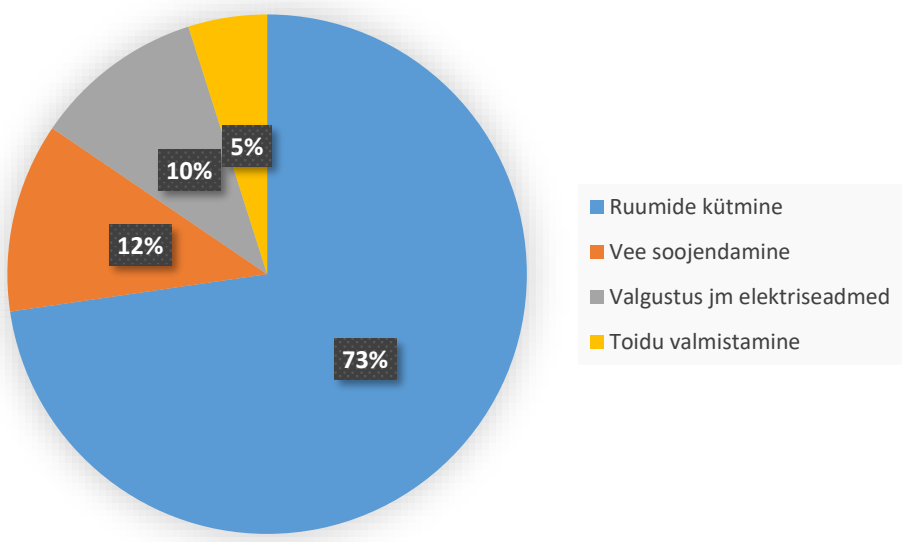
Elamu energiatarbimine oleneb peamiselt elanike arvust hoones ning nende elanike tarbimisharjumustest. Ühe leibkonna keskkonnajalajälge on võimalik vähendada ka tarbimisharjumusi muutmata, kui muuta targemaks hoone tehnosüsteemide automaatika. Kuigi elamu võimalikult säästlik energiatarbimine on tähtis eesmärk, näitab mõju keskkonnale ka kasutatav primaarenergia. Eestis on olukord keskmisest parem, sest levinud on keskküttevõrgud, mille soojus toodetakse koostootmisjaamades, mille kogukasutegur võib küündida üle 80% ning mille primaarenergiaks saab kasutada ka jäätmeid. Lisaks väheneb Eestis järjepidevalt ka selle elektri osakaal, mis on toodetud põlevkivist.

Eesti elamusektoris kasutatud energiaallikate osakaalu (Joonis 1) analüüsides ongi võimalik jõuda otseselt või kaudselt primaarenergia liigini ja selle keskkonnamõjuni. Jooniselt 1 selgub, et näiteks maagaasi osakaal energiaallikana Eesti elamusektoris on ainult 5,8%, mis on oluliselt väiksem Euroopa Liidu üle 30%-sest keskmisest [1].



Joonis 1.1 Elamusektori energiaallikate osakaalud

Energia tarbimise liikidest kulub Eesti elamusektoris enim energiat ruumide kütmiseks, mille osakaal ka kogu Euroopa mõistes on üks suurimaid [1]. Umbes kümnendiku moodustavad nii vee soojendamine kui ka valgustus jm elektriseadmed.



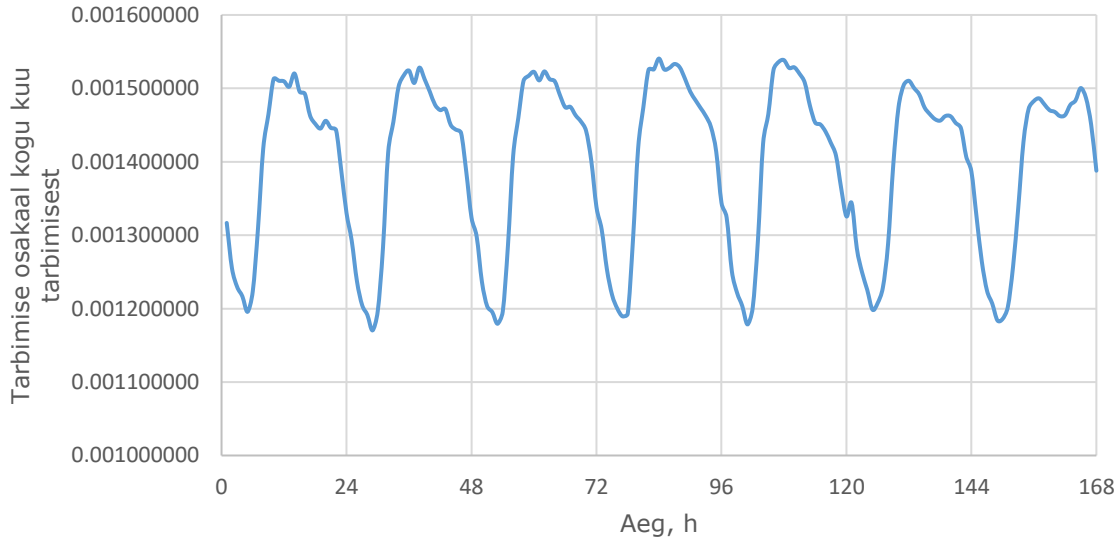
Joonis 1.2 Elamusektori energiatarbimise liikide osakaalud [1]

On ilmne, et hoonete tehnosüsteemide töö optimeerimisega on võimalik vähendada elamu energiatarbimist märgatavalt, vähendades sellega nii koormust loodusele kui ka lõpptarbija energiaarveid. Selle teostamiseks on üheks vajalikuks sisendiks teadmine, millal ei ole hoone kasutuses, et vähendada seadmete asjatut töös olekut. HVAC (kliimatehnika) seadmeid võib paindlikult kasutada, kuna kui hoone ei ole kasutuses, ei pea selle sisekliimat hoidma tingimata elaniku mugavuse piirides. Selleks ongi tähtis jõuda aina täpsemate ennustusmeetoditeni, mis näitavad kuna hoone on kasutuses.

Elektrienergiaga varustatus on Eestis 100% lähedal. Keskmiselt tarbis Eestis üks kodumajapidamine 2010. aastal 3465 kWh elektrienergiat [2]. See number kasvab iga aastaga vaatamata energiasäästlikematele elektriseadmetele, sest erinevaid seadmeid kasutatakse aina rohkem. HVAC seadmete hulgast kasutavad elektrit näiteks ventilaatorid, elektriradiaatorid jm elektriküttekehad ning eriti soojemates kliimades on suureks energiakuluks õhukonditsioneerid. USA lõunaranniku läheduses on elamute keskmine energiakulu õhu konditsioneerimiseks 27% [3]. Kui massiliselt kasutatakse seadmeid hoolimatult ja juhtimist optimiseerimata, võib see näiteks kuumalaine ajal elektrivõrgule väga koormavaks osutuda. Teised tüüpilised elektriseadmed kodudes, mida kasutatakse näiteks valgustuseks ja söögi tegemiseks, on vajaduspõhised ehk neid ei saa vastavalt elamu kasutusele juhtida.

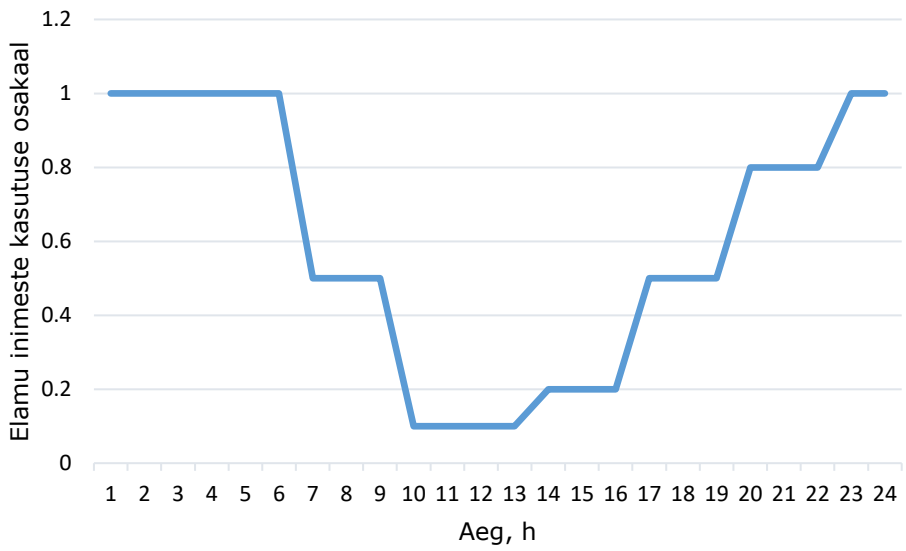
1.2 Tüüpgraafikud

Vajalike andmete puudumisel saab erinevaid arvutusi teha tüüpgraafikute abil. Tüüpkoormusgraafik on elektrienergia jaotamise ajagraafik ehk n-ö teisendustabel, milles jaotatakse klientide kuu jooksul tarbitud elektrikogus tundide lõikes laiali [4]. Tüüpkoormusgraafikud kasutatakse näiteks siis, kui millegipärast ei ole tunnipõhised tarbimisandmed jõudnud võrguettevõtjani. Graafik vastab mingile piirkonnale, seega ei saa selle põhjal kindlaid järeldusi teha üksikule elamule.



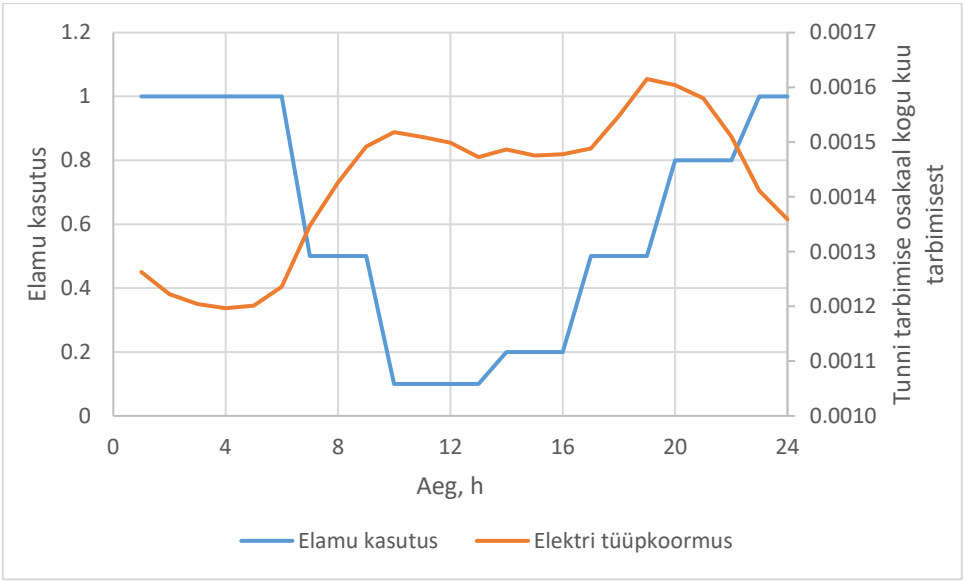
Joonis 1.3 2020. aasta juuni esimese nädala tüüpkoormusgraafik kodu- ja väikeärikliendile Elektrilevi teeninduspiirkonnas [5]

Lisaks elektri tarbimisele saab tüüpgraafiku koostada ka selle kohta, kui suurel määral on elamu kasutuses inimeste poolt. Näiteks peab sellega arvestatama hoone energiatõhususe arvutustes [6]. Antud kasutusprofiil on siiski puudustega kuna selline profiil ei vasta väga hästi nädalavahetuste käitumisele ja ei pruugi sarnaneda üksiku leibkonna käitumismustrile. Paremaks võiks osutada näiteks mudel, mis näitab elamu kasutust kogu nädala vältel ning põhineb varasemale kasutusmustrile.



Joonis 1.4 Elamu inimeste kasutusprofiil ühe ööpäeva jooksul [6]

Võrreldes ühe ööpäeva jooksul elamu kasutusprofiili ja tüüpkoormusgraafikut, saame järgneva graafiku:



Joonis 1.5 Elamu kasutuse ja elektritarbimise ööpäevane tüüpgraafik

Selgub, et elektri tarbimisandmete analüüsil tuleb arvestada mitmeid tegureid. Öösiti ei tähenda madal elektri tarbimine, et hoonet ei kasutata. On näha ka seda, et võetud andmed ei ühildu kui analüüsida päeva keskmist osa. Osaliselt on selle põhjuseks asjaolu, et kokku on võetud nii eramute kui ka väikeäride elektritarbimine.

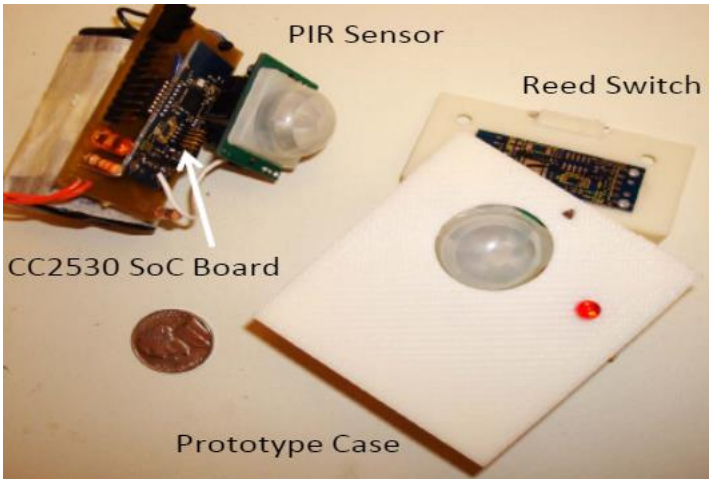
2. KASUTUSANDMETE KOGUMISE MEETODID

Iga pere käitumismustrid on erinevad, seega tagab parima info elamu kasutuse kohta individuaalne lähenemine. Individuaalse kasutusmustri määramise meetodid võib jagada otsesteks ja kaudseteks meetoditeks.

2.1 Otsesed meetodid

Otseste meetodite puhul paigutatakse hoonesse vastavalt valikule erinevaid andureid. Kasutatavate seadmete hulka kuuluvad liikumisandurid, ukseandurid, kaamerad, CO_2 andurid ja kontaktituvastusseadmed. Kasutatavate seadmete skaala sõltub tavaliselt sellest kui suure eksimusega tegelikust kasutusest ollakse nõus töötama. Täpsuse tõstmiseks võib andureid kasutada ka koos kaudsete meetodite abil kasutust analüüsides.

Kaameraid kasutatakse keerukamates süsteemides koos masinnägemise algoritmidega. Laiatarbelist kasutust piiravad paigaldamine, rahaline kulu ja privaatsusküsimused. CO_2 andurid võivad olla juba hoonesse paigutatud ventilatsiooni juhtimiseks, kuid kasutuse tuvastamiseks on sellise süsteemi puuduseks CO_2 taseme liiga aeglane muutumine jälgimaks kiireid hoone kasutuse muutusi. Passiivne liikumisandur (PIR-andur) on hoonetes kõige laialdasemalt levinud kuna seda kasutatakse enim valgustuse juhtimiseks. PIR-anduri põhine süsteem kasutuse tuvastamiseks ei anna siiski väga häid tulemusi kuna sellega on raske määrata inimeste arvu. Uurimuses [7] kasutati kontoorihoone kasutuse tuvastamiseks PIR-andureid koos magnetiliste ukseanduritega. Saadud kasutuse informatsiooni abil saavutati 10-15% energiasääst.



Joonis 2.1 Uurimuses kasutatud liikumis- ning ukseandur [7]

Antud meetodite puuduseks on see, et täpsuse tõstmiseks tõuseb ka jälgimissüsteemi hind ja keerukus. Sel põhjusel on kodumajapidamistele mõeldud nutikad termostaadid tihti müügil koos üksiku PIR-anduriga [8]. See tõstab aga vigaste juhtimiskäskude andmist HVAC süsteemile. Andurite kasutamisega tekivad mitmed lisakohustused nagu paigaldus, kalibreerimine, hooldus ja nende toite tagamine. Selle tõttu on perspektiivikam tuletada infot kasutuse kohta kaudsel teel.

2.2 Kaudsed meetodid

Kaudsete meetodite korral suudetakse mingist kontekstist eraldada hoone kasutuse muutumisest tekkinud mõjutused. Näitena võib tuua Wi-Fi võrgu kasutamise seostamine hoone kasutusega. Selline lähenemisviis tagab sujuva andmehõive kuna vajalikud seadmed on tihtipeale juba paigaldatud. Eestis ongi sellisteks seadmeteks näiteks kaugloetavad arvestid, mille edastatavat tarbimisinfot analüüsides on võimalik leida hoone kasutusest tingitud mõjutused.

Alternatiiviks elektritarbimise analüüsile on uuritud ka GPS andmete kasutamist elamu HVAC süsteemi optimaalseks juhtimiseks. 8 osalejaga uurimuses kasutati telefoni asukohainfo põhjal ennustatud koju jõudmise aega termostaadi sisendina ning suudeti säästa kuni 7% leibkonna energiatarbimisest [9]. Selle lähenemise teeb võimalikuks nutitelefonide laialdane kasutus.

Elamu kasutuse ennustamine elektritarbimise analüüsi põhjal on aina teostatavam kuna kaugloetavad elektriarvestid paigaldatakse aina rohkematesse kodutesse. Euroopa Liidu eesmärk 2020. aastaks oli jõuda selleni, et 80% tarbijatest kasutaksid kaugloetavat arvestit [10]. Uurimuses [8] saavutati juhendamisega masinõppe abil kasutuse tuvastamise täpsusvahemikuks 83-94%. Juhendatud õppe rakendamiseks on tähtis omada koos nii elektri tarbimise kui ka tegeliku hoone kasutuse andmeid.

Selles uurimistöös keskendutakse elamu kasutusinfo saamisele elektritarbimise analüüsi põhjal, kasutades selleks just juhendatud masinõppe algoritme, mis on laialt rakendatavad ja head tuvastamaks seoseid elektritarbimise ja kohal olevate inimeste arvu vahel.

3. SOBIVATE MASINÕPPE ALGORITMIDE VALIK ELAMU KASUTUSE HINDAMISEKS NING NENDE ANALÜÜS

3.1 Andmed

Andmed on masinõppe valdkonna aluseks. Teatud objekti kirjeldavaid omadusi kutsutakse selle objekti tunnusteks (*features*). Tunnused võivad olla näiteks arvulised (kaal, pikkus jne) või kategoorilised (nimi, sugu vm piiratud väärtuste hulgaga omadus). Objektide tunnuste väärtustest moodustub andmestik, mida saab kujutada tabelina. Tabel organiseeritakse nii, et iga rida vastab ühele juhtumile ning iga veerg vastab teatud tunnusele. Masinõppe algoritm otsib seoseid võrreldes erinevate juhtumite tunnuste väärtusi. Selleks, et algoritmid töötaksid, peavad kõik tunnused olema kirja pandud arvudena.

Mida suurem on andmestik ridade arvult, seda parem on algoritmil õppides leida seoseid. Seevastu suurem veergude arv ei tähenda alati, et andmestik on parem. Väheinformatiivsed ja dubleerivad tunnused vähendavad andmestiku kvaliteeti. Suur ridade hulk iseenesest ei taga loodud mudeli täpsust. Kogutud näited peavad olema ka esinduslikud ehk sisaldama kõiki võimaluste vahemikke, mis tulevikus võivad ette tulla ning eelistatult samas proportsioonis. Kallutatud andmete kasutamisel ei peegelda tulemus seda kui usaldusväärne oleks mudel tegelikult pärismaailmas rakendades. Lisaks tuleb andmestiku kvaliteedi tagamiseks eemaldada puuduvad või vigased väärtused.

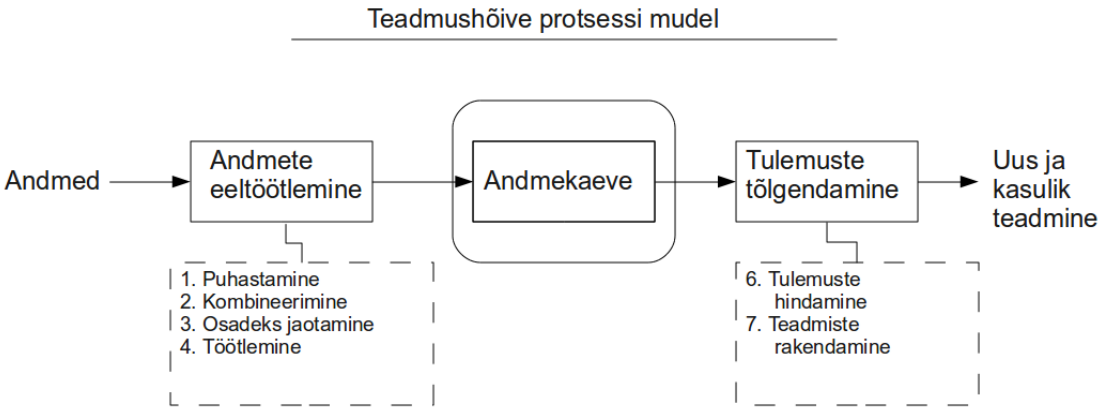
Mõningatel juhtudel on üks tunnustest teistest erilisem. Näiteks on selle väärtusi olla raske või kulukas koguda, väärtused saab teada tagantjärgi või on see oma olemuselt teistest erinev. Sel juhul oleks väärtuslik teiste tunnuste väärtuste põhjal hinnata erilisema tunnuse väärtus. Sellist tunnust, millele kasutaja määrab eristaatuse, nimetatakse märgendiks (*label*) ning saadud andmestikku märgendatud andmestikuks. Masinõppe algoritmi eesmärgiks on õppida tunnuste väärtuste põhjal hindama märgendite väärtusi.

3.2 Andmekaeve

Elamu kasutuse tuvastamise algoritmes on oma osa ka andmekaevandamisel. Elektritarbimisest otsitakse mustreid, mis viitaksid hoone kasutusele ehk paljastatakse kasulikke mustreid suurtest andmehulkadest. Masinõppe puhul kasutatakse mudeli

optimeerimiseks meetodeid andmekaevest. Traditsiooniliselt loetakse andmekaevandust teadmushõive mudeli üheks etapiks (vt Joonis 3.1).

Kuigi masinõppimine keskendub rohkem treeningandmetes olemasolevate omaduste põhjal ennustuste tegemisele ning andmekaeve rohkem andmetest uute omaduste leidmisele, siis kattuvad need tehnikad tihtipeale. Kuna masinõppe mudeli arendamiseks võidakse kasutada samu võtteid, seega on asjakohased ka etapid, mis läbitakse teadmushõive protsessi käigus.



Joonis 3.1 Teadmushõive protsess [11]

Andmekaevele eelneb andmete eeltöötlemine (puhastamine, jaotamine, normeerimine) ja järgneb järelanalüüs (tulemuste hindamine, teadmiste rakendamine). Andmekaevandamise protsessi all mõeldakse sobivate tehnikate valimist ja rakendamist ettevalmistatud andmetel. Antud töös kasutatakse kirjeldava analüüsi tehnikat, täpsemini klassifitseerimist, kus ajalooliste andmete põhjal määratakse tundmatutele väärtustele klassid. Kõikide teadmushõive protsessi etappide läbimine on osa ka antud uurimusest.

3.3 Uuritava elamu iseloomustus

Probleemülesande lahendamiseks tuleb valida vastavalt olukorrale sobiv algoritm, et saada parimaid tulemusi. Antud töös analüüsitakse ühe perekonna käitumist, et treenida juhendatud masinõppemudeleid. Juhendamine tähendab seda, et treenimise ajaks peavad väärtused teada olema nii sisendi (elektritarbimine) kui ka väljundi (elamu kasutus) kohta. Uurimise all on üks neljatoaline korter korterelamus, kuid võib eeldada, et tulemused on laiendatavad kõikidele eluasemetüüpidele.

Korteris elab igapäevaselt neli inimest, neist kaks on õpilased ning kaks omavad töökohta. Andmete kogumise perioodil kehtisid üleriigilised piirangud viiruse leviku tõttu, mis põhjustas tavapärasest erinevaid liikumismustreid. Perioodi vältel olid piirangud siiski muutumatud, seega tulemustele see mõju ei avalda. Mudelit luues on huvi pakkuv igaühe käitumine, sest tulemusena tahetakse tuvastada inimeste arv ühekohalise täpsusega.

Elektritarbimine korteris ei ole üheselt vastavuses seal viibivate inimeste arvuga. Esineb seadmeid nagu nõudepesumasin, veeboiler ja konditsioneer, mis tarbivad elektrit vastavalt seadesuurusele või oma ajagraafikule ning ei indikeeri alati inimeste kohalolekut. Samas on ka seadmeid nagu ahi või teler, mis on otseses seoses inimeste viibimisega ruumides.

Teadmised eluaseme elektritarvitite ning leibkonnaliikmete kohta on vajalikud andmete õigsuse kontrollis ja mudeli koostamisel.

3.4 Masinõpe ja selle jaotumine

Masinõppimise (*machine learning*) puhul on arvuti ise võimeline looma algoritmi, mis suudaks piisava täpsusega näidisandmetest eraldada meid huvitava info. Masinõpe 'ründab' tavaliselt selliseid probleeme, kus eesmärk on hästi teada (à la pildilt nägude leidmine), aga ei ole klassikalist meetodit kuidas sinna jõuda. Klassikaliselt jaotatakse masinõppe ülesanded juhendatud õppeks (*supervised learning*), juhendamata õppeks (*unsupervised learning*) ja stiimulõppeks(*reinforcement learning*) [12].

3.4.1 Juhendatud õpe

Juhendatud masinõpe genereerib funktsiooni, mis teisendab sisendandmed soovitud väljundandmeteks. Juhendamisega õppel peab masinale olema etteantud märgistatud treeninguandmete kogu, mille põhjal tehakse uusi järeldusi [11]. Märgend (*label*) on tunnus mille ennustamine meid huvitab, lõpliku hulga märgendi väärtuste korral nimetatakse seda klassiks ja vastavat tegevust klassifitseerimiseks. Selle töö kontekstis on klassideks inimeste arvud. Väljundi tüübi järgi saame juhendatud õppe algoritmid omakorda jaotada regressioonanalüüsi ja klassifitseerimise algoritmideks.

Regressioon on kasutusel pidevate väljundite ennustamisel. Saadavateks väljunditeks on suurused, mis on paindlikult määratletavad ning ei ole piiratud märgendite komplektiga. Näitena võib tuua taime kõrguse ennustamine sõltuvalt sademete hulgast või auto kütusekulu ennustamine sõltuvalt mudelist.

Klassifitseerimine on kasutusel diskreetse märgendi (klassi) ennustamisel. Paljudel juhtudel on situatsioonil ainult kaks võimalikku tulemust ning sellist olukorda nimetatakse binaarseks klassifitseerimiseks (*binary classification*). Võimalike tulemuste järgi on levinud veel mitmeklassiline klassifitseerimine (*multiclass classification*) ja *multi-label classification*. Need on vastavalt kolme või rohkemate klasside võimalusega klassifitseerimine ja klassifitseerimine, kus väljundiks võib olla mitu klassi korraga.

Õige mudeli valimine on oluline etapp masinõppe protsessis. On oluline, et mudel sobiks vastava probleemi ja andmekoguga. Elamu kasutuse tuletamine elektritarbimise põhjal esindab klassifitseerimise probleemi. Töö autori eesmärgiks on inimeste arvu täpne tuvastamine, seega on neljaliikmelise leibkonna puhul täpsemalt tegu mitmeklassilise klassifitseerimise ülesandega.

Paljud klassifitseerimisalgoritmid võimaldavad rohkema kui kahe klassi kasutamist, kuid mõned on oma loomult binaarsed algoritmid. Viimaste puhul on siiski võimalik kasutada erinevaid strateegiaid lisaklasside võimaldamiseks.

Kõiki tingimusi arvesse võttes valiti antud töös tulemuste võrdlemiseks kolm levinud juhendatud masinõppe algoritmi, mis sobiksid uuritava elamu kasutuse hindamiseks elektri tarbimisandmetest. Kaks neist toetavad loomupäraselt rohkemate klasside kasutust ning ühe puhul tuleb rakendada vajalikku strateegiat.

3.4.2 Juhendamata õpe

Juhendamata masinõppe korral proovitakse leida märgistamata andmete kogust uusi struktuure. Sel juhul ei üritatagi ennustada konkreetset märgendit. Otsitakse näiteks anomaaliaid või gruppidesse jaotumist (klasterdamine). Kui lihtsate probleemide korral on see teostatav ka inimintuitsiooniga, siis rohkemate tunnustega andmepunktide korral on vajalik juba automaatne tuvastus.

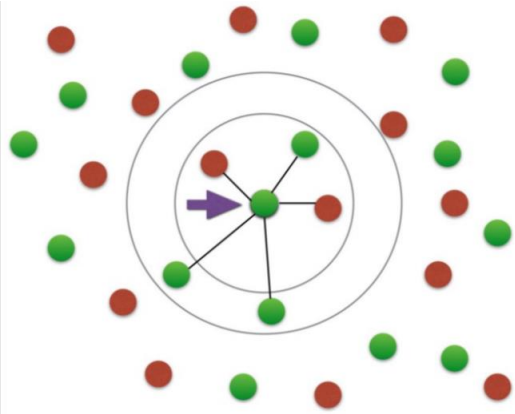
Antud töös ei ole mindud juhendamata õppe suunas kuna probleemülesande klassid on üheselt mõistetavad ning tegelikud klasside väärtused andmete kogumisel on kergesti kättesaadavad.

3.5 K-lähima naabri algoritm

K-lähima naabri algoritmi (*k-nearest neighbors algorithm*) aluseks on oletus, et sarnaste tunnustega andmepunktid kuuluvad enamasti samasse klassi. Kui on kättesaadav andmestik koos klasside väärtustega (treeninghulk), siis võttes uue andmepunkti ilma klassita, leiab algoritm uue punkti lähimad naabrid ja määrab klassi vastavalt naabrite

seas enamlevinud klassi järgi. See tähendab, et sisendandmetest ei looda üldistatud mudelit.

Klassifitseerija täpsuse hindamiseks eraldatakse kogu andmestikust lisaks treenimisele veel valideerimishulk ja antakse mudelile justkui võõrad andmepunktid, kuid hiljem võrreldakse mudeli ennustusi valideerimisandmepunktide tegelike klassidega. Täpsuse optimeerimiseks on võimalik muuta naabrite arvu. Klassifitseerimise käigus on *k* kasutaja defineeritud konstant, mis määrab naabrite arvu uuele punktile klassi määramisel. Konstandi väärtuse olulisust illustreerib Joonis 3.2, kus *k* väärtuse muutmisel muutub ka uuele punktile määratud klass. Sisemise ringi ($k = 3$) valikul määratakse „punane" klass ning välimise ringi ($k = 5$) valikul „roheline" klass.



Joonis 3.2 K-lähima naabri algoritmi rakendamine kahedimensioonilise andmestikuga [13]

Parim naabrite arv sõltub andmestikust. Üldiselt toob väga väike naabrite arv kaasa ülesobitamise (*overfitting*), mis tähendab liiga tugevat sõltuvust treeningandmetest ehk üksikud erandlikud treeningpunktid võivad olla otsustavaks faktoriks võõra punkti klassi määramisel. Liiga suur naabrite arv toob omakorda kaasa alasobitamise (*underfitting*), kus klassifitseerija ei suuda arvestada väiksemate trendidega treenimishulgas.

Algoritmi toimimiseks on vajalik matemaatiliselt defineerida, mida mõeldakse lähimate naabrite all. Laialt on levinud ning ka antud töös simuleerides kasutatakse naabrite kauguste hindamiseks eukleidilist kaugust. Klassifitseeritav andmepunkt kujutatakse objektina n-dimensionaalses vektorruumis. Seejärel arvutatakse kahe punkti vahelised ruutkaugused igas dimensioonis ning summeeritakse need. Võttes summast ruutjuure saadaksegi eukleidiline kaugus kahe punkti vahel (Valem 3.1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3.1)$$

kus $x_1,...,x_n$ tähistavad vektori $x = (x_1,x_2,...,x_n)$ koordinaate. [14]

3.6 Otsustusmets

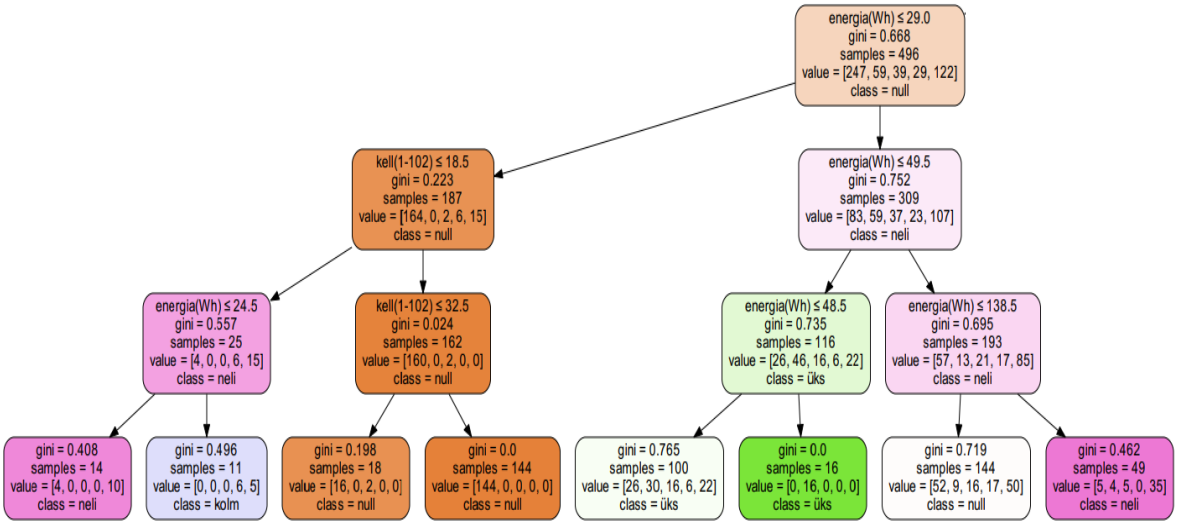
Otsustusmetsa (*random forest*) algoritm kuulub ansambelõppe meetodite hulka. Ansambelmeetodi mõte on kasutada koos paljusid „nõrku õppijaid" (siinkohal otsustuspuu), et moodustada nendest üks „tugev õppija" [15]. Üksik otsustuspuu on aldis ülesobitamisele ning otsustusmets minimeerib ülesobitamise ohu, andes võõra andmepunkti igale puule klassifitsserimiseks ja seejärel määratakse klass populaarseima vastuse põhjal. Suure arvu otsustuspuude puhul ei oma mõned ülesobitatud otsustuspuud enam suurt rolli.

Kuigi otsustuspuu loomine on deterministlik, siis saavutatakse erinevate puude loomine näiteks *bagging*-algoritmi abil, kus treeninghulgast võetakse juhuslikult alamhulk andmeid otsustuspuu loomiseks ning korratakse protsessi, seejuures ei eemaldata juba valitud andmeid treeninghulgast. Klassi määramisest süvitsi arusaamiseks on vajalik mõista, kuidas iga otsustuspuu eraldi jõuab klassifitseerimise tulemuseni.

3.6.1 Otsustuspuu

Otsustuspuu (*decision tree*) on järjestikustest „kas" küsimustest ning „jah"/"ei" jagunemistest koosnev „puu". Otsustuspuu sõlmedeks on tunnused ning harudeks tunnuste võimalikud väärtused. Iga haru märgib võimalikku alternatiivi, kusjuures alternatiivid peavad olema üksteist välistavad ja ammendavad. Jagunemisi valitakse üldjuhul tunnuse põhjal, mis eristab klasse kõige paremini. Klassifitseerimistäpsuse paranemist hinnatakse näiteks Gini indeksiga, mis võimaldab eristada häid ja halbu jagunemisi. Tüüpiliselt otsitakse puu tipus iga tunnuse jaoks parimat õpiandmete jagunemist ning lõpuks valitakse välja tunnus, mille parim jagunemine on teiste tunnustega võrreldes omakorda parim [16]. Ülesobitamise vastu ühe puu raames kasutatakse tagasilõikamist (*pruning*), mis aitab samuti kaasa mudeli täpsuse parandamisele. Puud loetakse ülalt alla – andmepunkti võrreldakse esmalt kõige üleval oleva küsimusega ja liigutakse vastavalt vastusele kuni jõutakse puu „lehte", millele on määratud kindel klass.

Tegu on ühe intuitiivsema masinõppe algoritmiga. Ühe otsustuspuu struktuuri näitlikustamiseks on Joonisel 3.3 toodud antud töös kogutud andmete põhjal loodud tagasilõikamistega otsustuspuu väike osa, mis lõppeb „lehtedega".



Joonis 3.3 Üks osa otsustuspuust, mille loomiseks kasutati antud töö treeningandmeid

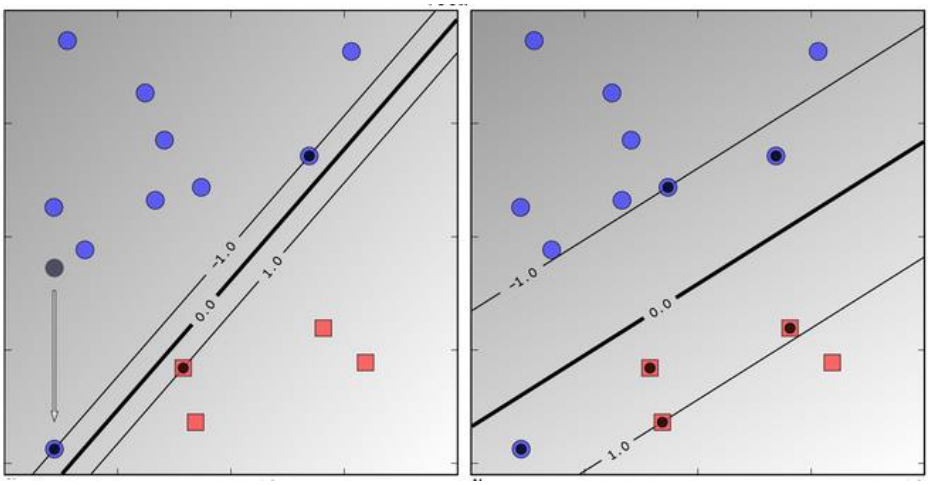
On näha, kuidas tehakse uue andmepunkti üle otsuseid vastavalt energia ja kellaaja väärtusele ning jõudes leheni, määratakse lõplik klass. Antud juhul on puud kärbitud ning osade lehtede Gini koefitsient on lähemal ühele kui nullile, mis tähendab et lõplik klass ei määratud suure ülekaaluga. Seda saab parandada otsustuspuu sügavuse suurendamine, kuid sellega tõuseb omakorda ülesobitamise risk.

3.7 Tugivektor-masin

Juhendatud masinõppes on levinud klassifitseerimisalgoritmiks ka tugivektor-masin (*support-vector machine*). Klasside eraldamiseks treenitakse hüpertasand ning uue andmepunkti klass määratakse selle järgi, kummale poole antud vektor hüpertasandist jääb. Klassid eraldatakse lineaarselt ja mittelineaarselt eralduvate klasside korral kujutatakse klassid kõrgema dimensiooniga ruumi ning lineaarne eraldamine teostatakse seal [17].

Nimi tugivektor tuleneb sellest, et leitakse vektorruumis klasside äärealadel olevad punktid ning defineeritakse need vektorina algpunktist. Neid vektoreid kasutatakse „toena" separaatori loomiseks. Lõplik separaator valitakse selliselt, et tugivektorite ja separaatori vaheline kaugus oleks maksimaalne. Kuna mudeli loomiseks kasutatakse tugivektoreid on võimalik paljusid treeningpunkte ignoreerida, mis on algoritmi üheks tugevuseks.

Mõnikord võib tugivektori ja hüpertasandi maksimaalse vahekauguse kasutamine mudeli täpsust hoopiski vähendada, kui treeningpunktides esinevad üksikud erandlikud väärtused. Selle mõju vähendamiseks võidakse lubada separaatoril mõne treeningpunkti valesti klassifitseerimist ehk kasutada pehme äärega tugivektor-masinat. Erinevuste hoomamiseks on parim kasutada kahe tunnusega andmestikku, kus kahedimensioonilisel graafikul on separaatoriks sirge. Joonisel 3.4 on toodud selline olukord, kus parempoolsel graafikul on küll ignoreeritud ühte „sinist" märgendit, kuid tänu sellele saavutatakse üldiselt parem klasside eraldatus. Lisaks on mõlemal juhul märgistatud tugivektorid sisemise musta ringiga.



Joonis 3.4 Pehme ääre konstandi *C* muutmise efekt simuleerimiskeskkonnas [18]

3.7.1 *one-vs-one* strateegia mitmeklassiliseks klassifitseerimiseks

Kui jagada mitmeklassilise klassifitseerimise probleem mitmeteks binaarse klassifitseerimise probleemideks, on võimalik rakendada tugivektor-masin algoritme. Antud töö mudelis kasutatakse kõikide klassipaaride võrdlusi (*one-vs-one*). Selle strateegia puhul iga klassipaaride võrdluse mudel ennustab võõra andmepunkti klassi ning lõplik klass määratakse enamlevinud ennustuse põhjal. Kasutatud võrdluste arv *n* leitakse Valem 3.2 järgi.

$$n = \frac{K(K-1)}{2} \tag{3.2}$$

kus K – võimalike klasside arv. [19]

4. ANDMETE KOGUMINE JA MUDELITE RAKENDAMINE

4.1 Andmed ja nende töötlemine

Töö fookuses on elektri tarbimisandmed ning elamu kasutuse andmed. Autor on otsustanud kasutada ka koos elektriandmetega salvestatavat ajalist väärtust kuna elamu kasutus on samuti sõltuv kellaajast. Kellaaja salvestamine ja edastamine toimub arvesti poolt ehk selleks ei ole vaja lisaseadmeid paigaldada. Andmestiku üks rida (andmepunkt) on valitud 10-minutilisele vahemikuna, kus on tunnusteks kellaaeg, elektritarbimine, kas tegemist on tööpäevaga ja elamus viibivate inimeste arv. Päevasiseselt kogutakse andmeid vahemikus 06.00-23.00 kuna mudeli eesmärgiks ei ole võetud magamisperioodil elamu kasutuse hindamine.

Andmete kogumine toimus vahemikus 15.04.-02.05.2021 igal nädalapäeval. Peale puudulike väärtuste eemaldamist on lõpliku kogutud andmestiku tabeli suuruseks 1705 rida ja 4 veergu.

4.1.1 Ajast tuletatavad andmed

Andmepunktidele lisatunnuste leidmine annab võimaluse mudeli täpsemaks tegemisele. Lisades ajalise info, leiab algoritm seosed ka päevasisese kellaaja ja elamu kasutuse vahel. Kellaja kasutamine ei sea andmete märkimise tihedusele piiri, kuid valitud 10-minutiline tihedus tagab kasutuse märkimise ebatäpsuste silumise, samal ajal säilitades liikumiste üldise trendi. Lisaks leiab autor, et 10-minutline samm annab laiemad võimalused hiljem mudeli rakendatavuse uurimisel.

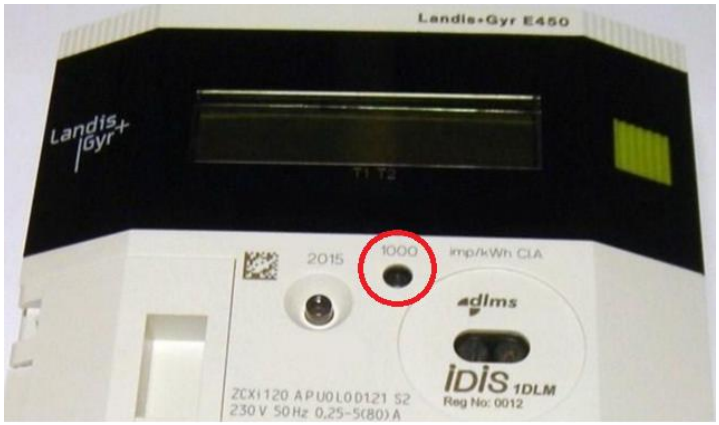
Algoritmi sisenditeks sobivad arvulised väärtused. Kasutades 10-minutiliseid vahemikke, saame seega ühe päeva kohta ajalistele vahemikele määratud väärtusteks 1-102.

Mudeli täpsuse tõstmiseks leiab autor, et on õigustatud ka tööpäeva ja nädalavahetuse päeva eristamine, kuna tüüpiliselt on liikumisharjumused nädalavahetuseti teistsugused. Selleks on tunnusena lisatud ka „tööpäev" veerg, mille väärtuseks on üks tööpäeva andmepunktides ning null nädalavahetuse puhul.

4.1.2 Elukoha elektritarbimise andmed

Elektriarvesti kaudu kättesaadava elamu elektri tarbimisajaloo ajasamm on üks tund. Selles töös on otsustatud mõõta tarbimist tihedamalt, et oleks võimalik 10-minutilise ajavahemiku puhul kasutada tarbitud elektrienergia tegelikku väärtust, mitte tunnitarbimisest tuletatud keskmist energiakogust. Tarbimine salvestatakse spetsiaalse

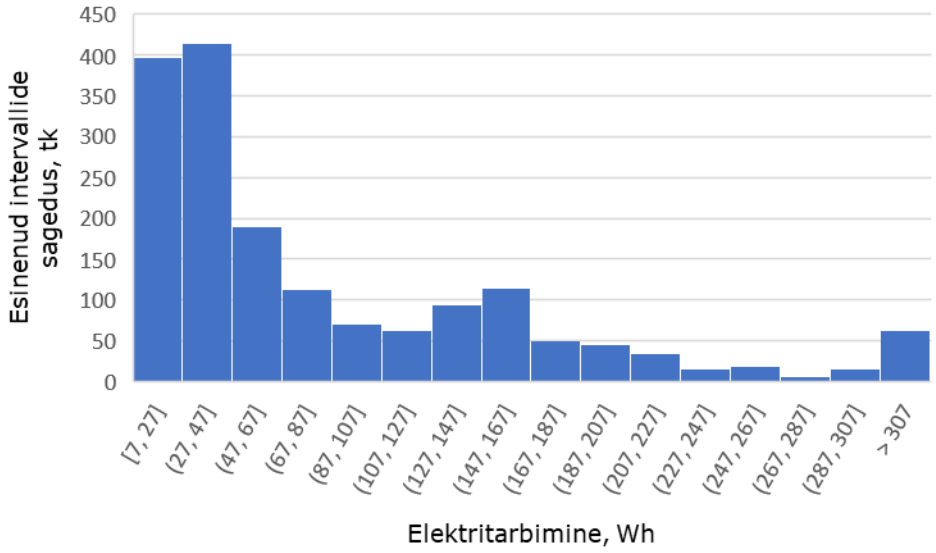
fototakistiga seadmega, lugedes vilkumiste arvu arvesti vastava valgusdioodi poolt (Joonis 4.1).



Joonis 4.1 Ringiga tähistatud valgusdiood tüüpilisel kaugloetaval arvestil [20]

Valgusimpulss antakse arvesti poolt iga vatt-tunni tarbimise järel. Kasutatud seadmega loendati kümne minuti impulsid kokku ning salvestati mälukaardile. Salvestatud vatt-tunnid on andmepunktide üheks tunnuseks.

Tarbimisharjumuste analüüsimiseks graafilise kujutisega annab võimaluse histogrammi loomine, mis annab ülevaate tarbitud energiasuuruste jaotumisest sageduse järgi. Joonisel 4.2 on horisontaalteljel toodud tunnuse väärtused (kümne minuti elektritarbimine) ja vertikaalteljel esinemissagedused.



Joonis 4.2 Uuritava elamu elektritarbimise histogramm

Elamu elektritarbimine on kallutatud vähesema elektritarbimise poole. Selgub, et isegi kui elektriseadmeid aktiivselt ei kasutata, esineb elamus konstantne tarbimine vähemalt 7 Wh kümne minuti jooksul.

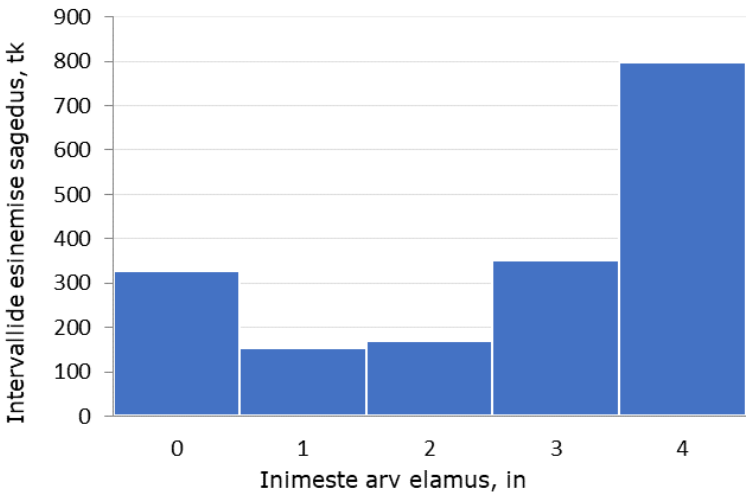
4.1.3 Elukoha kasutuse andmed

Nagu eelnevalt välja toodud, on korteris püsivaid elanikke neli. Koos võimalusega, et kohal ei viibi ühtegi inimest, saadakse viis erinevat klassi (kohal inimeste arv), mille hindamine on eesmärgiks. Masinõppe mudeli treenimiseks ning hiljem valideerimisel saavutatud täpsuse määramiseks on vajalik koguda ka klasside tegelikud väärtused. Selleks otsustati kasutada inimeste liikumise paberile märkimist, mille rakendamine on kiire ning täpsus hea. Sellega välditi lisasüsteemide paigaldamist, mis võivad olla ka küsitava täpsusega. Elanikele anti info märkida saabumisel/lahkumisel hetke kellaaeg ning numbriga inimeste hulk märke tegemise järgsel ajal. Märkimise jaoks loodud tabeli ülesehitus on näidatud Tabel 4.1 abil.

Tabel 4.1 Elamu kasutuse märkimiseks prinditava tabeli ülesehitus viie tööpäeva näitel

Päev (E-P)	Kellaeg ja liikumine	Päev (E-P)	Kellaeg ja liikumine	Päev (E-P)	Kellaeg ja liikumine	Päev (E-P)	Kellaeg ja liikumine	Päev (E-P)	Kellaeg ja liikumine
E		T		K		N		R	

Andmete sisestamisel täideti kasutuse tunnuse veerg vastavalt märgitud andmetele, eelnevalt ümardades liikumishetked kümne minutilise täpsuse peale. Inimeste arvu jaotumise kirjeldamiseks 10-minutiliste vahemike peale, kasutatakse sarnaselt eelmisele punktile histogrammi (Joonis 4.3).

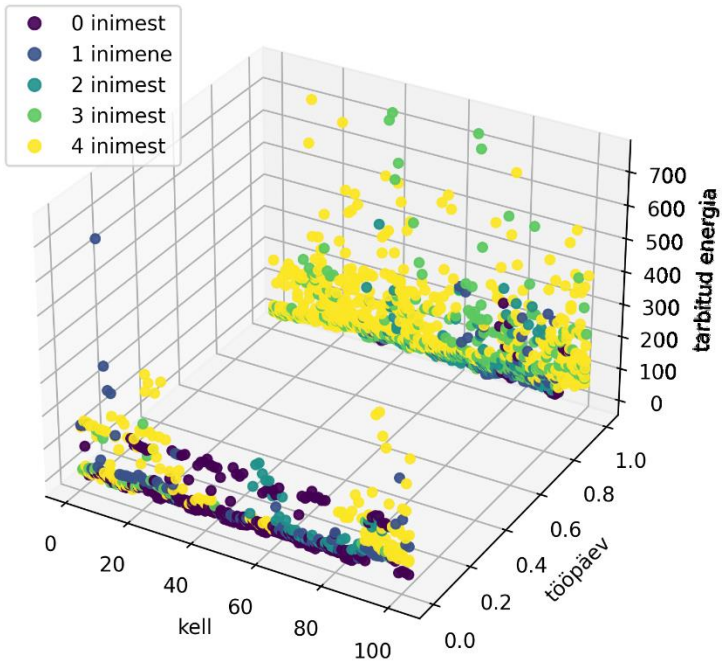


Joonis 4.3 Uuritava elamu kasutuse histogramm

Histogramm on kausja kujuga ehk levinum on kolme ja nelja pereliikme kohalolu ja ka olukord, kus elamus ei viibi kedagi.

4.1.4 Andmepunktide visualiseerimine

Andmestikku, millel on kolm tunnust, on võimalik kujutada kolmemõõtmelise teljestiku abil. Joonisel 4.4 on koondatud kogutud andmepunktid ning klassiline kuuluvus kuvatakse värvide kaupa.



Joonis 4.4 Kogutud andmepunktid teisendamata kujul

Teljestike mõõtkavad erinevad oluliselt ehk andmed on erinevates suurusjärkudes. See segab meie valitud algoritmidest k-lähima naabri ja tugivektor-masina algoritmide tööd, mis kasutavad võrdsete mõõtkavadega vektorruume. Suurte väärtuste erinevusega tunnused domineerivad seoste leidmisel ning trendid väiksemate väärtusvahemikega tunnustes jäävad algoritmi poolt märkamata. Sarnaste väärtusvahemike loomiseks tunnuste lõikes kasutatakse andmete peal näiteks standardiseerimist või *min-max* normaliseerimist. Antud töös kasutatakse kahe algoritmi puhul standardiseerimist.

4.1.5 Standardiseerimine

Standardiseerimine tähendab väärtuste tsentreerimist tunnuse keskmise suhtes selles kogumis ja tsentreeritud väärtuse väljendamist tunnuse standardhälbe ühikutes [21]. Sel juhul tunnuse väärtus x on negatiivne kui see jääb alla kogu hulga keskmise ja vastupidi. Teisendatud väärtus z leitakse Valem 4.1 alusel.

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

kus μ – andmehulga aritmeetiline keskmine,

σ – andmehulga standardhälve. [22]

4.2 Scikit-learn

Scikit-learn on vaba lähtekoodiga masinõppe teek programmeerimiskeelele Python, mida kasutatakse selles töös algoritmide rakendamiseks, täpsuse hindamiseks ja muudeks funktsioonideks. See sisaldab suurel hulgal valmiskujul vahendeid ja algoritme, mis on koondatud vastavatesse moodulitesse ja klassidesse. Scikit-learn omab head ühilduvust ka selliste teekidega nagu Pandas ja Matplotlib, millega antud töös teostatakse andmete manipuleerimist ja graafikute loomist.

4.3 Algoritmide testimine

Erinevate algoritmide omavaheliseks võrdlemiseks on vajalik hinnata neid võrdsetel alustel. Juhendatud õppe puhul tähendab see valitud algoritmidele sama treeninghulga ning valideerimishulga (*validation set*) sisendiks andmist. Mudeleid ei tohi treenida valideerimiseks mõeldud andmete abil. Valideerimisandmeid kasutatakse loodud mudeli täpsuse hindamiseks, jättes esialgu klasside väärtused kõrvale ning hiljem võrreldes neid tunnuste kaudu ennustatud tulemustega. Kui mudeli täpsus on testimisel väga kõrge, võib viidata see andmete „lekkimisele" treeninghulka.

Selles töös on treeninghulk ja valideerimishulk koostatud kogu andmestikust scikit-learn mooduli abil, mis juhuslikkuse alusel teostab jaotuse kasutaja poolt määratud osakaalude alusel. Antud töös on treeningandmete osakaal 80% ja valideerimisandmetel vastavalt 20%. Seejärel luuakse nende andmete põhjal valitud kolme algoritmi mudelid ning leitakse iga mudeli täpsus.

4.3.1 Tunnuste valik

Tunnuste valiku mõju hindamiseks võrreldakse iga algoritmi korral kolme stsenaariumi, kus vastavalt on mudeli loomiseks kasutatud kellaaega; kellaaega ja tööpäeva infot; kellaaega, tööpäeva infot ja elektritarbimist (Tabel 4.2).

Tabel 4.2 Masinõppe mudelite loomisel kasutatavad tunnused

Tunnuste grupp	**Tunnused**
1	Kellaaeg
2	Kellaaeg, tööpäev
3	Kellaaeg, tööpäev, elektritarbimine

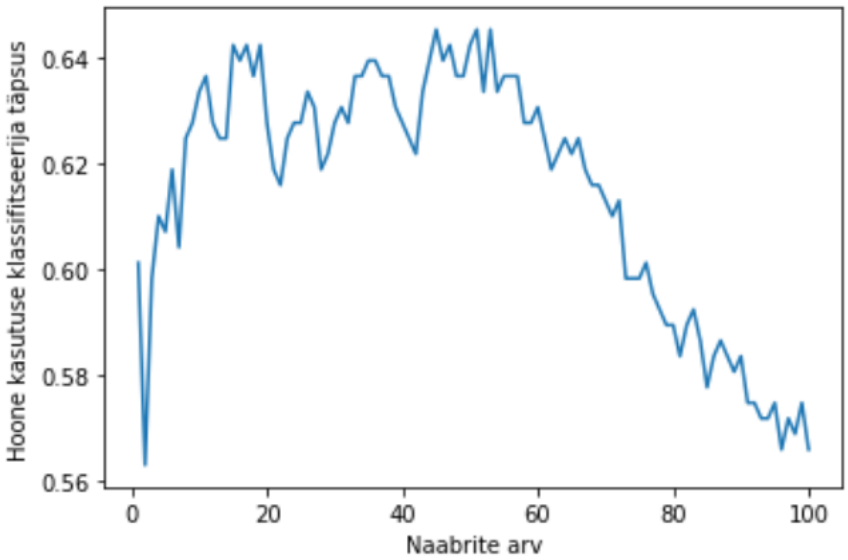
4.3.2 Tulemuste hindamine

Mudelit luues on huvipakkuv võimekus korrektselt ennustada märgendeid andmetele, mille märgendit me ei tea. Levinuim on klassifitseerimise puhul täpsuse (*accuracy*) arvutamine ehk kui suurel protsendil oli ennustatud märgend õige.

Tulemused saab esitada ka eksimismaatriksina (*confusion matrix*), kus on iga klassi kohta toodud tõeselt positiivsete, tõeselt negatiivsete, valenegatiivsete ja valepositiivsete ennustuste arv. Õigesti ennustatud märgendite arvud asuvad peadiagonaalil.

4.3.3 K-lähima naabri mudel

K-lähima naabri algoritmi korral on muudetavaks parameetriks naabrite arv. Naabrite arvu optimeerimise abil saavutatakse parim täpsus, mis on esindatud Joonisel 4.5 kui kolme kõrgeima tipuna.



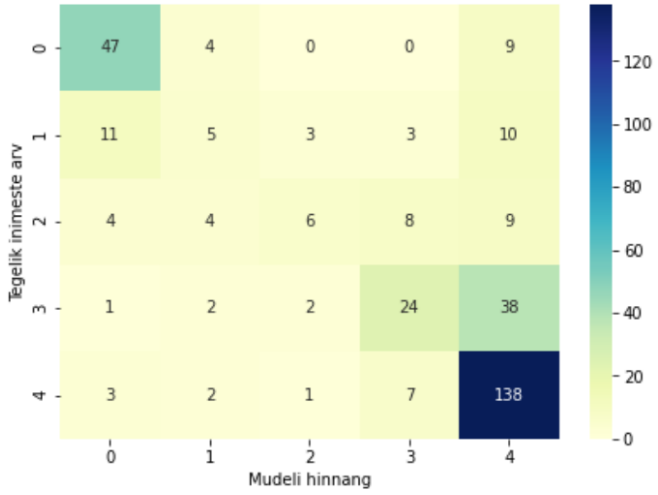
Joonis 4.5 K-lähima naabri mudeli täpsus sõltuvalt valitud konstandist *k*

Parimaid tulemusi andnud naabrite arvu seast otsustati modelleerimisel kasutada naabrite arvu 51. Saadud tulemused tunnuste valiku kaupa on toodud Tabelis 4.3

Tabel 4.3 K-lähima naabri mudeli ennustustäpsuse sõltuvus valitud tunnustest

Tunnuste grupp	**Täpsus**
1	46,9%
2	59,5%
3	64,5%

Mudeli parameetri valiku mõju vastab tüüpilisele olukorrale, kus liiga väike naabrite arv tähendab ülesobitatud mudelit ning liiga suur naabrite arv põhjustab mudeli alasobitatust. Parima tunnuste grupi ja parameetriga mudeli eksimismaatriks on toodud Joonisel 4.6. Valideerimisandmepunktide arvuks on 341 ning mudel hindas õigesti 220 klassi.



Joonis 4.6 K-lähima naabri mudeli eksimismaatriks

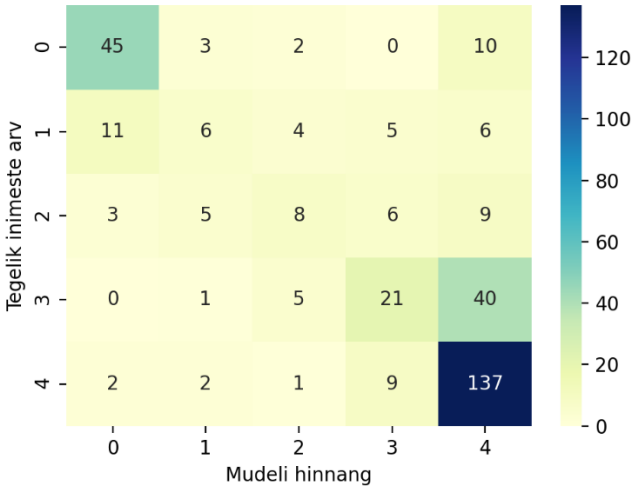
4.3.4 Otsustusmetsa mudel

Scikit-learn mooduli abil mudeli parameetrite optimeerimisel jõuti mudelini, kus on 800 otsustuspuud; otsustuspuus tehakse sõlmes jaotus vähemalt kümne andmepunkti olemasolul; lõplik „leht" peab sisaldama vähemalt neli andmepunkti. Saadud täpsused tunnuste kaupa on toodud Tabelis 4.4.

Tabel 4.4 Otsustusmetsa mudeli ennustustäpsuse sõltuvus valitud tunnustest

Tunnuste grupp	**Täpsus**
1	43,3%
2	54,5%
3	63,6%

Mudeli eksimismaatriks (Joonis 4.7) sarnaneb k-lähima naabri mudeli omale. On selge, et kõrge täpsuse saavutamiseks peab mudel eelkõige õigesti hindama klasside null, kolm ja neli esinemist, mis on eeldatav ka klasside histogrammi põhjal.



Joonis 4.7 Otsustusmetsa mudeli eksimismaatriks

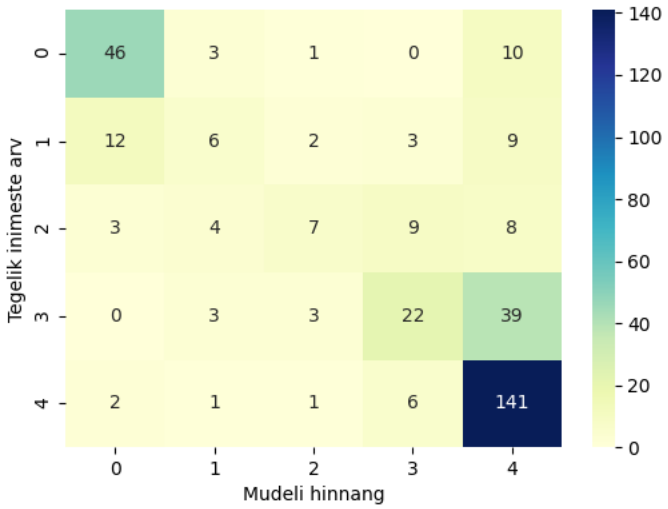
4.3.5 Tugivektor-masina mudel

Mittelineaarse separaator luuakse *radial basis function* alusel. Mudeli parameetrite optimeerimisel saadi parim täpsus *C* = 1 ja *gamma* = 4 korral. Kõrgem gamma väärtus tekitab rohkem eraldiseisvaid nn regioone vektorruumis selle asemel, et määratleda klassiline kuuluvus ühe eralduspiiriga.

Tabel 4.5 Tugivektor-masina mudeli ennustustäpsuse sõltuvus valitud tunnustest

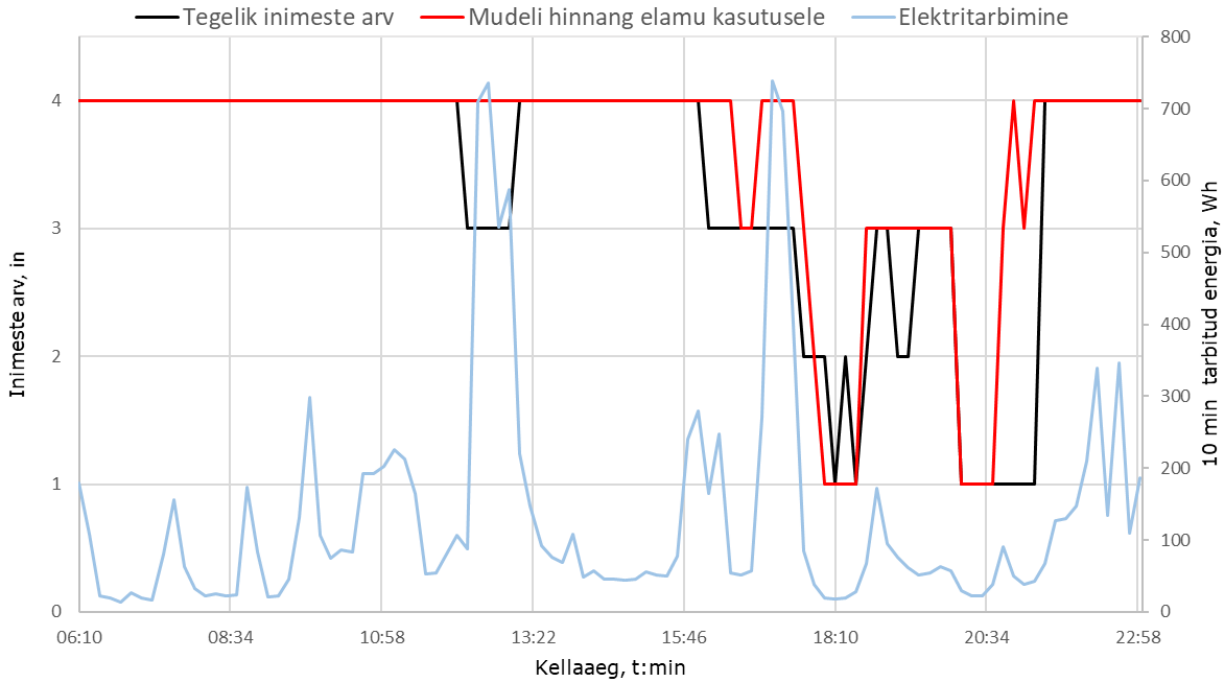
Tunnuste grupp	**Täpsus**
1	46,3%
2	59.2%
3	65,1%

Joonisel 4.8 loodud eksimismaatriksi alusel on näha, et õigesti määrati klass 222 korral. Mudel saavutas napilt parima tulemuse valideerimisandmetel valitud algoritmide hulgast. Enamasti erinevad eksimused vaid ühe klassi võrra tegelikust. Võttes seda arvesse, võib täpsuse tõstmiseks seda määratleda ka kui tõenäosust, et hinnang ei erine kahe või rohkema klassi võrra. Sel juhul esineb nn eksimusi 41 korral ehk „täpsuseks" saavutatakse 88%. Selline võte võib olla õigustatud kui rakendusvaldkond ei ole tundlik mõningase eksimise vastu.



Joonis 4.8 Tugivektor-masina eksimismaatriks

Juhuslikkuse alusel treening- ning valideerimisandmete loomisel ei saa luua ühtlast mudeli väljundit, mis vastaks päriselulisele olukorrale. Selleks, et visualiseerida paremini mudeli toimimist, on võetud valideerimiseks üks kindel päev (28. aprill) ning ülejäänud andmete põhjal toimub mudeli treenimine. Tulemustest joonistub välja muster, kus päeva esimeses pooles olenemata elektritarbimisest on kasulik hinnata kõiki elanikke elamus viibivateks (Joonis 4.9).



Joonis 4.9 Masinõppe mudeli väljund 28. aprilli näitel

5. JÄRELDUSED KATSETULEMUSTEST

Algoritmide vahel suuri erinevusi kasutuse hindamise täpsusi ei esinenud. Täpsed mudeli täpsused tunnuste kaupa on toodud Tabelis 5.1, kus on rõhutatud ka iga tunnuste grupi parim täpsus. Iga tunnus sisaldas kasulikku infot ehk lisatunnuste kasutuselevõtt kõrgendas selgelt mudelite täpsust. Autori hinnangul on tulemused aktsepteeritavad, kuid põhjalikuma ettevalmistusega korralikult parandatavad.

Tabel 5.1 Mudelitevaheline täpsuse võrdlus

Mudel	Tunnuste grupp	Täpsus
K-lähimat naabrit	1	**46,9%**
	2	**59,5%**
	3	64,5%
Otsustusmets	1	43,3%
	2	54,5%
	3	63,6%
Tugivektor-masin	1	46,3%
	2	59.2%
	3	**65,1%**

Masinõppe ülesannetes on peamisteks mudeli arendamise viisideks andmete kogumise kestuse pikendamine ja väärtuslike tunnuste valiku suurendamine. Mudeli valik ning selle parameetrite optimiseerimine annavad juba väiksemad edasiminekud. Väärtuslike tunnustena, mille mõõtmise võimekust tulevates töödes tasuks lisada, võib välja tuua näiteks maksimumvõimsuse mõõtmist ja võimsuste/energia varieerumise arvestamist.

Andmete kogumise kestuse pikendamisele aitaksid kaasa sellised lihtsustused nagu elektriarvesti poolt tihedam tarbimise edastamine ja standardse seadme loomine, mis võimaldab passiivselt elamu kasutust tuvastada ja salvestada suure täpsusega. Sellisel juhul ei oleks elanikele barjääri lisategevuste näol, et koguda põhjalik andmestik antud elamu kohta ning luua isikustatud masinõppe mudel.

Läbitehtud protsessi võib tulevikus võtta alusena, et teostada sarnast kasutuse hindamist kontorihoonete puhul. Võttes täpse inimeste arvu asemel kasutusele inimeste arvu vahemikud ning luues hoone sektsioonides eraldi elektrienergia mõõtmise võimekuse, on võimalikud rakendusvaldkonnad laiad.

KOKKUVÕTE

Antud töö eesmärgiks oli elamu kasutuse hindamine elektritarbimise ja masinõppe meetodite põhjal. Töö sisuks on nii elamu kasutuse hindamismeetodite analüüs kui ka vastavate mudelite rakendamine spetsiifilise elamu näitel. Töö spetsiifika seisnes inimeste arvu täpses hindamises.

Töö põhiosa algas tüüpilise elamu kasutuse ja elektritarbimise uurimisest tüüpgraafikute näitel. Energia lõpptarbimise osakaale uurides selgus, et enamus hoone energiakulust kulub tehnosüsteemide tööle ehk potentsiaal kulude vähendamisele on suur. Tutvudes üldistatud elamu kasutuse ja energiatarbimise tüüpgraafikutega selgus, et kasutuses ei eristata tööpäevi ning elektritarbimise keskmine profiil on koostatud koos väikeäride tarbimisega ehk saadud info ei ole piisavalt täpne, et kasutada seda individuaalses majapidamises.

Tutvudes nii olemasolevate otseste meetoditega kui ka kaudsete meetoditega, mida kasutatakse kasutusandmete kogumiseks, leiti et paljude füüsiliste seadmete paigaldamine kasutuse tuvastamiseks tõstab kiiresti süsteemi keerukust ja alginvesteeringut. Kaudsete meetodite kasutamise teeb võimalikuks teatud seadmete nagu telefonide või kaugloetavate elektriarvestite lai kasutuselevõtt. Leiti, et kasutades elektritarbimise analüüsi masinõppe meetodite abil, on erinevates uurimustes saavutatud elamu hõivatuse tuvastamine ligi 90% täpsusega. Saadi kinnitust masinõppe meetodite efektiivsuses.

Seejärel tehti valik kolme masinõppe algoritmi kasuks, tuginedes uuritava elamu andmete iseloomule ja masinõppe liikide analüüsile. Võttes aluseks vajaduse ennustada viite erinevat olukorda, otsustati k-lähima naabri algoritmi, otsustusmetsa algoritmi ja tugivektor-masin algoritmi kasuks, mis on laialt levinud ning samuti võimelised lihtsasti hakkama saada mitmeklassilise klassifitseerimisega. Kõigi kolme algoritmi kohta tehti analüüs, mis viisil saavutatakse klasside (inimeste arv) hindamine andmepunktide põhjal.

Kindla elamu kohta andmete kogumisel suudeti luua andmestik 1705 andmepunktiga, millel oli neli tunnust, millest omakorda üks tähistas hoone kasutust (klass). Andmestiku ja algoritmi põhjal masinõppe mudelite loomiseks kasutati Scikit-learn masinõppe teeki. Mudelite rakendamisel saavutati valideerimistulemused vahemikus 63,6-65,1%. Lubades eksimusi ühe klassi võrra, saavutati täpsus kuni 88%.

Saavutati töö eesmärk ehk veenduti elektriandmete analüüsi kasulikkuses elamu kasutuse hindamisel, kasutades masinõppe meetodeid. Suuri erinevusi mudelitel

omavahel täpsuses ei esinenud. Samuti leiti potentsiaali töö edasiseks arendamiseks. Kogudes laiapõhjalisema andmestiku on võimekus luua veelgi täpsem ja laiemate rakendusvõimalustega elamu- või ka suurema hoone kasutuse hindamise mudel.

KASUTATUD KIRJANDUS

[1] „Energy consumption in households," Eurostat, 2020. [Võrgumaterjal]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php/Energy_consumption_in_households#Energy_consumption_in_households_by_type_of_end-use. [Kasutatud 14.05.2021.].

[2] R. Raudjärv ja L. Kuskova, „Energiatarbimine kodumajapidamistes," 2011.

[3] „Air conditioning accounts for about 12% of U.S. home energy expenditures," EIA, 2018. [Võrgumaterjal]. Available: https://www.eia.gov/todayinenergy/detail.php?id=36692. [Kasutatud 14.05.2021.].

[4] „Tüüpkoormusgraafikud," Konkurentsiamet, [Võrgumaterjal]. Available: https://www.konkurentsiamet.ee/et/elekter-maagaas/elekter/tuupkoormusgraafikud. [Kasutatud 14.05.2021.].

[5] „Elektrilevi tüüpkoormus," Elektrilevi, [Võrgumaterjal]. Available: https://www.elektrilevi.ee/ettevottest/elektriturg?modal=elektriturul-isainfo&tabgroup_1=electricity_market. [Kasutatud 14.05.2021.].

[6] „Hoone energiatõhususe arvutamise metoodika," Riigi Teataja, [Võrgumaterjal]. Available: https://www.riigiteataja.ee/akt/119012018007?leiaKehtiv. [Kasutatud 14.05.2021.].

[7] Y. Agatwal, B. Balaji, R. Gupta, J. Lyles, M. Wei ja T. Weng, „Occupancy-Driven Energy Management for Smart Building Automation," 2010.

[8] K. W., C. Beckel ja S. Santini, „Household Occupancy Monitoring Using Electricity Meters," 2015.

[9] M. Gupta, S. Intille ja K. Larson, „Adding GPS-Control to Traditional Thermostats: An Exploration of Potential Energy Savings and Design Challenges," 2009.

[10] European Data Protection Supervisor, „Smart Meters in Smart Homes," European data protection supervisor, 2019. [Võrgumaterjal]. Available: https://edps.europa.eu/data-protection/our-

work/publications/techdispatch/techdispatch-2-smart-meters-smart-homes_en. [Kasutatud 14.05.2021.].

[11] Vikipeedia, „Andmekaeve," [Võrgumaterjal]. Available: https://et.wikipedia.org/wiki/Andmekaeve. [Kasutatud 14.05.2021.].

[12] T. Pungas, „Masinõpe: mittetehniline ülevaade," [Võrgumaterjal]. Available: https://pungas.ee/masinope-mittetehniline-ulevaade/. [Kasutatud 14.05.2021.].

[13] E. Allibhai, „Building a k-Nearest-Neighbors (k-NN) Model with Scikit-learn," *Towards data science.*

[14] Vikipeedia, „Kaugus," [Võrgumaterjal]. Available: https://et.wikipedia.org/wiki/Kaugus. [Kasutatud 14.05.2021.].

[15] Vikipeedia, „Otsustusmets," [Võrgumaterjal]. Available: https://et.wikipedia.org/wiki/Otsustusmets. [Kasutatud 14.05.2021.].

[16] K. Käärmann, „Otsustuspuudega klassifitseerimine," 2003.

[17] H. Tint, „Sissejuhatus tugivektor-masinatesse," 2003.

[18] A. Ben-Hur, S. Sonnenburg, B. Schölkopf ja C. S. Ong, „Support vector machines and kernels for computational biology," 2008.

[19] A. Band, „Multi-class Classification - One-vs-All & One-vs-One," 2020.

[20] „1-faasilised voolumõõtjad," Aleksander Siilbaum Elektrikaup, [Võrgumaterjal]. Available: https://www.elektrikaup.ee/1-faasilised-elektrienergia-arvestid-ja-voolumootjad/5508/kaugloetav-elektriarvesti-1-faasiline-2-tariifne-5-80a-zcxi120apu0l0d1-21-s2-landis-gyr-e450.html. [Kasutatud 14.05.2021.].

[21] L.-M. Tooding, „Regressioonimudelid," 2014. [Võrgumaterjal]. Available: http://samm.ut.ee/regressioonanalyys. [Kasutatud 14.05.2021.].

[22] Vikipeedia, „Standard score," [Võrgumaterjal]. Available: https://en.wikipedia.org/wiki/Standard_score. [Kasutatud 14.05.2021.].