TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele õppekava

Mihkel Järveoja

Põllukultuuride tuvastamise masinõppe mudeli tunnuste olulisuse hindamine

Magistritöö (15 EAP)

Juhendajad: Kaupo Voormansik (PhD)
Tambet Matiisen (MSc)

Tartu 2021

Põllukultuuride tuvastamise masinõppe mudeli tunnuste olulisuse hindamine

Lühikokkuvõte:

Kaugseireandmete ulatuslikum kasutuselevõtt tsiviilelus on eriti silmapaistev põllumaade seires. Suurte ja tihti inimsilmale keeruliste andmekihtide töötlemisel kasutatakse erinevaid masinõppepõhiseid lähenemisi. Magistritöös sobitati juhumetsa masinõppe mudel 28 Eesti põllukultuurigrupi automaatseks tuvastamiseks, kasutades treeningandmetena 2018. ja 2019. aasta kohta arvutatud Sentinel-1 radarsatelliidi, Sentinel-2 optilise satelliidi, mullastiku ja ilmastiku tunnuseid. Hinnati erinevate sisendtunnuste absoluutset olulisust ja nende olulisuse ajalist muutlikkust klassifitseerimise tulemustesse. Mudeli klassideülene kaalutud F1-skoor 2018. aasta testkogul oli 0,82 ja 2019. aasta testkogul 0,85. Olulisemate tunnuste hulka kuulusid Sentinel-1 erinevate polarisatsioonide (VH ja VV) tagasihajumise tunnused ja Sentinel-2 indeksid PSRI, NDVI ja TC_Vegetation. Ilmnes selge sesoonne varieeruvus eri tüüpi tunnuste olulisuses. Sentinel-2 tunnused osutusid määravamateks hooaja alguses (mai) ja lõpus (august), samas kui hooaja keskel (juuni, juuli) vähenes nende olulisus märgatavalt. Sentinel-1 tagasihajumise tunnused olid olulisemad kesksuvel.

Võtmesõnad:

Põllukultuuride automaatne tuvastamine, masinõpe, juhumets, tunnuste olulisus, Sentinel-1, Sentinel-2

CERCS: P176

Feature Importance in Crop Classification Machine Learning Model

Abstract:

Remotely sensed, in particular satellite data, is already widely used in agricultural parcels monitoring, and this trend is not showing signs of diminishing. Wide range of machine learning algorithms have significantly reduced the burden to interpret bulky and often complex satellite data, contributing to the exploration of new use-cases and services. In this study Random Forest classification model is used to separate 28 crop type classes in Estonia. Input data consisted of two seasons (2018, 2019) of Estonian agricultural parcels and features calculated from Sentinel-1 and Sentinel-2 satellite images, meteorological records and soil maps. Achieved multiclass weighted F1 score for year 2018 test set was 0.82 and for year 2019 0.85. Among most important features were Sentinel-1 VH and VV polarization back-scatter intensities and Sentinel-2 PSRI, NDVI and TC-vegetation indices. It was discovered that Sentinel-2 features were more prominent in early (May) and late season (August), but during mid-season (June, July) their importance decreased significantly. Sentinel-1 back-scatter features were more important during mid-season. It was concluded, that using both radar and optical satellite data ensure better classification result than using any of them separately, since they complement each other.

Keywords:

Crop classification, machine learning, Random Forest, feature importance, Sentinel-1, Sentinel-2

CERCS: P176

Sisukord

1. Sissejuhatus

Satelliitseire on tegemas tsiviilelus uut revolutsiooni, mis ei seisne niivõrd tehnoloogia täiustumises, kuivõrd suurte ja erinevate andmekogude paremas kättesaadavuses ning nende laialdases kasutuselevõtus igapäevaelu probleemide lahendamiseks. 2021. aasta alguses ennustas ajakiri Science oma esikaaneloos (Rosen, 2021), et tulevikus põhinevad maapinna seiresüsteemid radarsatelliitide andmestikul.

Põllumajandus on üks valdkondadest, kus see uute andmeallikate algatatud revolutsioon on juba käimas ja mitmed reaalajas töötavad satelliitseire rakendused on juba kasutuses. Seejuures Euroopa Liit (EL) ja Eesti on ühed automaatse põllumajandusseire vedurid. Olemasolevad rakendused on aga veel väike osa suuremast täisautomaatsest põllumajandusmaade seire kontseptsioonist, kuhu oma pisikese panuse püüab anda ka käesolev uurimistöö.

Põllukultuuride automaatne tuvastamine on põllumajandusmaade seire vundament, millele toetuvad järgmised, juba spetsiifilisemat väärtust või funktsionaalsust loovad rakendused, mis peavad otsuseid tegema põllukultuuri põhiselt. Eesti Teadusagentuuri rahastatud RITA Kaugseire projekti raames hakkas 2018. aastal OÜ KappaZeta juhitud töörühm arendama Eesti oludesse sobivat kaugseirepõhist põllukultuuride tuvastamise metoodikat. Töö üheks alamosaks oli tuvastusmudeli sisendtunnuste olulisuse hindamine. Mudeli tunnusruumi sisuline tundmine võimaldab teha tulevikus teadlikumaid ja ressursisäästlikumaid valikuid selle rakendamisel operatiivteenustes.

Magistritöö eesmärk on hinnata, kui suur panus on erinevatel tuvastusmudelis kasutatavatel satelliiditunnustel ja milline on tunnuste olulisuse ajaline muutlikkus kasvuperioodi jooksul. Kasutatud sisendandmestik hõlmab endas kõiki Eesti põlde 2018. ja 2019. aasta kasvuperioodidest. Tunnuste olulisuse hindamisel kasutati juhumetsa algoritmi, mis erineb RITA projektis rakendatud närvivõrkude mudelist. Alternatiivse masinõppealgoritmi tulemused on hea võrdlus ka eelnimetatud projekti ja metoodika raames arendatud mudelile ja kinnitavad valitud lähenemise usaldusväärsust.

Uurimistöö on jaotud kolme suuremasse mõttelisse ossa. Taustainfo peatükis antakse esmalt ülevaade põllukultuuride kaugseire tähtsusest, selle võimaldajatest ja olulisematest initsiatiividest selles valdkonnas. Samas peatükis kirjeldatakse ka maakattetüüpide klassifitseerimise kaugseirepõhiseid lähenemisi, mille lahutamatuks komponendiks on masinõpe ja erinevad klassifitseerimisalgoritmid. Töö praktilises osas selgitatakse andmete eeltöötlust ja juhumetsa mudeli sobitamist, mille järel antakse ülevaade klassifitseerimise ja tunnuste olulisuse hindamise tulemustest. Arutelu osas tõlgendatakse eelmises osas leitud tulemusi ja võrreldakse neid teiste uurimistöödega. Töö neljas lisas on esitatud mahukamad võrdlustabelid töö tulemuste kohta.

Andmete eeltöötluse, mudeli sobitamise ja tunnuste hindamise lähtekoodi on kirjeldatud lisas V.

2. Taustainfo

Kuigi selle uurimistöö sisuline osa käsitleb kitsast masinõppespetsiifilist probleemi, paigutub see ometi väga selgelt laiemasse konteksti, mille mõistmine aitab töö tulemusi paremini tõlgendada. Kaugseirerakendused tungivad inimeste igapäevaellu ning koos masinõppealgoritmidega lihtsustavad mitmeid tegevusi. Kuigi need kaks valdkonda on algusest peale hästi läbi saanud, toimub muutusi, mis seda suhet veel harmoonilisemaks muudavad.

Käesolevas peatükis liigutakse üldisematest taustateemadest detailsete meetodite kirjeldusteni, püüdes vastata küsimustele, miks see uurimisteema oluline on ja mil moel on võimalik püstitatud probleemi lahendada.

2.1 Põllukultuuride kaugseire vajalikkus

Põllumaa ja -kultuuride seiramisel on erinevaid ajendeid. Viimase statistika (FAO, 2020) kohaselt moodustab põllumajandusmaa 37% kogu Maa pinnast ja peab ühel või teisel moel toitma ära 7,7-miljardilise (2019) elanikkonna, mis prognooside põhjal (United Nations jt, 2019) kasvab aastaks 2050 juba 9,7 miljardile. Samal ajal, kui inimkond ja toidu tootmine on kasvanud, on põllumajandusmaa pindala viimastel kümnenditel (2000-2018) aga hoopis 2% võrra vähenenud (FAO, 2020), mis näitab, et olemasolevat maad kasutatakse intensiivsemalt ja selle väärtus tõuseb. Olles piiratud ressurss, tuleb põllumaad kasutada võimalikult säästlikult, ent samal ajal rahuldada kasvavat nõudlust toodangule. Need olemuselt vastuolulised nõudmised tingivad ulatusliku seire vajaduse, et ennetada väärkasutust ja olla valmis kliimamuutuste põhjustatud ootamatusteks, millel võivad olla väga tõsised tagajärjed paljude inimeste eludele.

Euroopa Liidu (EL) pinnast on 38,2% (EUROSTAT, 2020) kasutuses põllumajandusmaana, millele rakendub EL-i ühine põllumajanduspoliitika (EPP) oma hüvede ja reeglitega. Ühise põllumajanduspoliitika, millega alustati juba 1962. aastal, peamisteks eesmärkideks on toetada põllumajandustootjaid ning soodustada säästvat loodusvarade kasutust. 2019. aastal oli EPP toetuste maht 58,82 miljardit eurot, mis on 35% kogu EL-i eelarvest (Euroopa Komisjon, 2021). Üheks EPP toetuseks on ühtne pindalatoetus, mida makstakse igal aastal konkreetse taotluse ja põllu kohta. Raha kasutamist ja poliitika tulemuslikkust kontrollitakse ühise seire- ja hindamisraamistikuga, mille hulka kuulub ka pindalatoetuste seire. Toetuste erisused ja neile kohanduvad tingimused on keeruline bürokraatlik rägastik, ent lühidalt kokkuvõetult tahab EL pindalatoetuste seire käigus teada, kas taotleja esitatud andmed (ja hooaja jooksul tehtud tegevused) oma põllumajandusmaa kohta vastavad sellele maale makstava toetuse nõuetele ja eesmärkidele.

Seni on pindalatoetuste seiret viidud läbi inspektorite valikuliste kontrollkäikudega, mis, arvestades, et EL-is oli 2016. aastal 10,3 miljonit põllumajandusettevõtet ja 156,7 miljonit hektarit põllumaad (EUROSTAT, 2020), on väga kallis ja aeganõudev ning suudab lõpuks katta väga väikse osa kogu taotletud pinnast. 2018. aastal võeti aga Euroopa Komisjoni poolt vastu uued reeglid, mis lubavad pindalatoetuste seires kasutada uusi tehnoloogiad, eelkõige automaatset satelliitseiret (Euroopa Komisjon, 2018). Ühtlasi tähendab see satelliidipiltide aktsepteerimist juriidiliselt pädeva tõendina reeglite rikkujate sanktsioneerimiseks.

Tõuke ja aluse nimetatud kontrollimetoodika muutuseks on andnud Maa seire programm Copernicus, mis alates 2014. aastast toodab erineva suunitlusega satelliitidelt kõigile kättesaadavat regulaarset infovoogu. Ülesvõtete piisavalt hea lahutus, vaba ja avatud andmejaotuspoliitika, kontrollitud ja ühtlane kvaliteet ning tihe kordustsükkel võimaldavad

ehitada Copernicuse andmetele mitmeid teenuseid eri valdkondades, sealhulgas põllumajanduses.

2.2 Copernicuse programm ja Sentineli satelliidid

Copernicuse programmis on kokku kuus satelliidiperekonda. Põllumajandusmaade seires kasutatakse neist peamiselt kahe: optilise satelliidi Sentinel-2 ja radarsatelliidi Sentinel-1 andmeid.

Sentinel-2 missioon koosneb kahest identsest satelliidist (A ja B), mis tiirlevad 786 km kõrgusel polaarorbiidil, tagades süstemaatilise katvuse pea kogu Maa maismaa osale. Satelliitide kordustsükkel on ekvaatoril 5 päeva ja keskmistel laiuskraadidel 2-3 päeva (Drusch jt, 2012). See tähendab, et pea kogu maismaa pinna kohta saadakse identsete vaatlustingimustega võrreldavad pildid minimaalselt iga 5 päeva tagant, Eestis umbes iga 2 päeva tagant.

Sentinel-2 mõlema satelliidi pardal on multispektraalne instrument (MSI), mille 13 spektraalkanalit on lainepikkuste vahemikus 443 - 2202 nm (spektri nähtavast osast kuni lühiinfrapunaseni) (Drusch jt, 2012). Piltide ruumiline lahutus sõltub spektraalkanalist ja varieerub 10 meetrist kuni 60 meetrini.

Laia vaateala (290 km), tiheda kordustsükli ja hea multispektraalse ning ruumilise lahutuse tõttu on Sentinel-2 andmestik kasutatav maapinna muutuste jälgimiseks ülemaailmselt ning võimaldab luua vastavaid rakendusi. Optilise sensorina on ta aga tundlik ilmastikuoludele ja ta ei näe läbi pilvede. Hoolimata tihedast kordustsüklist võivad seetõttu pilvistes piirkondades Sentinel-2 piltidest loodud aegread jääda väga lünklikuks. Näiteks on hinnatud, et maailma peamistes põllumajanduspiirkondades on kasvuperioodil vähem kui pooled 8-päevalised komposiitpildid vähemalt 70% pilvevabad (Whitcraft jt, 2015). Kullamaa (2015) on uurinud kaugseireks sobivate päevade arvu ka Eesti alal ning parimal juhul on võimalik saada mai- ja juunikuus keskmiselt üks pilvevaba pilt nädalas. Suve lõpus ja sügisel aga pigem kaks või veel vähem pilvevaba pilti kuus.

Küll aga näevad läbi pilvede ja ei vaja päevavalgust Sentinel-1 missiooni kaks identset polaarorbiidil tiirlevat satelliiti, mille kordusülesvõtete tiheduseks on Euroopas minimaalselt 6 ja mujal 12 päeva. C-laineala tehisavaradar (C-SAR) töötab nende pardal erinevates polarisatsioonides ja neljas vaatlusrežiimis, mille vaateulatused (80 – 400 km) ja ruumilised lahutused (5 – 40 m) on erinevad (Torres jt, 2012). Sentinel-1 radar saadab ise välja mikrolaineid, mis läbistavad pilvi ja võimaldavad teha ülesvõtteid nii päeval kui öösel. Radarsatelliidi erinevad polarisatsioonid tähendavad selle poolt saadetavate ja vastu võetavate elektromagnetlainete võnkumissuunda. Sentinel-1 satelliit suudab välja saata ja maapinnalt tagasipeegeldunud laineid vastu võtta horisontaalses (H) ja vertikaalses (V) polarisatsioonis. Tagasihajumise tunnus „*s0vh*“ tähendab sel juhul, et välja saadeti vertikaalselt võnkuv signaal ja vastu võeti tagasipeegeldunud horisontaalselt võnkuv signaal. Asjaolu, et erinevad maakattetüübid hajutavad mikrolaineid erinevates polarisatsioonides tagasi erineva intensiivsusega, teeb need tunnused väga kasulikuks erinevates kaugseire rakendustes, sealhulgas põllukultuuride eristamises.

2.3 Kaugseire põllumajandusrakendused Euroopa Liidus ja Eestis

Juba varem mainitud EPP reformi ootuses ja tuules on käivitatud mitu EL-i rahastatud pilootprojekti ja teenust, kus kaugseire ja eelkõige Copernicuse andmeallikad mängivad võtmerolli. Sen4CAP (*Sentinels for Common Agricultural Policy*) on neist ehk kõige tuntum ja ambitsioonikam, soovides pakkuda kõigile Euroopa Liidu makseagentuuridele (esialgu kuuele pilootprojektis osalevale riigile) valideeritud algoritme ja töövooge, et edukalt,

lihtsalt ja kuluefektiivselt seirata põllumajandusmaid (Sen4CAP konsortsium, 2017). Projekti tarkvara ja algoritmide portfellist leiab Sentinel-1 ja -2 andmestikul põhinevaid kultuuride tuvastamise, rohumaade niitmise tuvastamise, taimkatte seisukorra hindamise ja mitmesuguste põllumajandustööde tuvastamise algoritme.

Sen4CAP-ga on tihedalt seotud NIVA (*New IACS Vision in Action*) projekt, mis püüab rakendada EPP ühtse haldus- ja kontrollisüsteemi uut visiooni praktikas, kasutades ära digitaalseid tehnoloogiad. Kaugseire rakendustest väärib selles projektis äramärkimist automaatne põllupiiride tuvastamine, mille masinõppe algoritmides kasutakse sisendina Sentinel-2 pilte (NIVA, 2019).

Kahe eelmisega üsna sarnase kaugseire komponendiga on ka DIONE projekt, mille raames arendatakse terviklikku põllumajandusmaade seire tööriistakasti ja kus lähtutakse samuti EPP nõuetest ja vajadustest. Lisaks satelliidiandmetele püütakse süsteemi integreerida ka droonipilte (DIONE konsortsium, 2019).

Valmis ja täielikult testitud põllumajandusmaade satelliitseire teenust, mis lähtuks EPP vajadustest ja oleks lihtsasti integreeritav EL-i makseagentuuride kohalike infosüsteemidega, hetkel olemas ei ole. Erinevate projektide raames arendavad osade riikide makseagentuurid oma süsteeme. Vähem aktiivsete riikide makseagentuurid on aga äraootaval seisukohal, lootes, et millalgi tekib valmis satelliitseireteenus, mille saab võimalikult väikeste kuludega kasutusele võtta.

Eestis korraldab EL-i põllumajandustoetuste jagamist ja sihipärase kasutuse kontrolli Põllumajanduse Registrite ja Informatsiooni Amet (PRIA). Eesti riik oli üks esimestest Euroopas, kes üleriigiliselt käivitas satelliitseirel põhineva rohumaade hooldamise kontrolli süsteemi. 2018. aastal operatiivkasutuses tööle hakanud seiresüsteem SATIKAS kasutab Sentinel-1 ja -2 andmeid ja masinõppe mudelit, et anda iga rohumaa kohta hinnang – tähtajaks hooldatud või mitte. Nüüdseks kolm aastat kasutatud rakendus on vähendanud PRIA inspektorite tööd ning lihtsustanud toetuste väärkasutajate leidmist. SATIKA-s kasutatav niitmise tuvastamise mudel ja metoodika põhinevad suuresti Eesti teadlaste uurimistöödel (Voormansik jt, 2016; Zalite jt, 2016; Tamm jt, 2016).

Kuigi SATIKA arendus oli ühelt poolt tingitud EPP nõuetest, on teisalt taoline tehisintellektil põhinev protsesside automatiseerimine ka Eesti riigi prioriteet. 2019. aastal võeti vastu Eesti riiklik tehisintellektialane tegevuskava aastani 2021 (Majandus- ja Kommunikatsiooniministeerium, 2019), milles kirjeldatud tegevustega edendatakse tehisintellektil põhinevate rakenduste ehk krattide kasutuselevõttu. SATIKAS oli üks esimesi avaliku sektori kratte, mis leidis kasutust üleriigilises operatiivsüsteemis. Eesti Teadusagentuuri RITA Kaugseire projekti (Noorma jt, 2020) raames on praeguseks valminud ka kaugseirel põhinev automaatne põllukultuuride tuvastamise metoodika, mis võiks tulevikus saada SATIKA uueks funktsionaalsuseks ja sobituda hästi nii PRIA kui ka Eesti riigi eespool nimetatud eesmärkidega. Ka see magistritöö valmis suuresti RITA projekti raames ning selles kasutati samu eeltöödeldud andmeid ja panustati uue teabega projekti lõpparuandesse.

2.4 Põllukultuuride ja maakattetüüpide klassifitseerimise viisid

Satelliidipiltidelt põllukultuuride ja maakattetüüpide automaatne tuvastamine on kaugseire teaduses pika ajalooga. Alates LANDSAT-i satelliidiseeria esimese multispektraalse skanneri orbiidile saatmisest 1972. aastal on kogu Maad katvad pildid olnud kättesaadavad erinevateks kasutusjuhtudeks. Spektraalse info klassifitseerimine on seejuures olnud üks peamine analüüsimeetod.

Peamiselt lähtutakse satelliidipiltide klassifitseerimisel kas piksli- või objektipõhisest lähenemisest. Kui pikslipõhine pildianalüüs on piltlikult öeldes olnud olemas alates esimestest salvestatud pikslitest, siis objektipõhisele lähenemisele on andnud hoogu piltide ruumilise lahutuse paranemine, mille tulemusena on huvipakkuvad objektid (nt põllud) suuremad kui üksikud pikslid. Lisaks on paremad piltide segmentimise algoritmid teinud võimalikuks sarnaste pikslite grupeerimise ja automaatse objektide tuvastuse (Blaschke, 2010). Objektide tuvastust ei pea aga tingimata segmentimise teel tegema, kui on olemas muu usaldusväärne allikas huvipakkuvate objektide piiritlemiseks ja neile ruumilise dimensiooni lisamiseks – näiteks juba teadaolevad põllupiirid antud töö kontekstis. Sel juhul on objektipõhise lähenemise suurim kasu üldistava spektraalse info loomine, sest pikslikogumitest koosnevatele objektidele saab arvutada erinevate kanalite statistilisi väärtusi, mida üksikutele pikslitele teha ei saa: näiteks keskväärtus, mediaan, miinimum, maksimum ja standardhälve. Lisaks on leitud, et SAR-piltide klassifitseerimisel on objektipõhine lähenemine pikslipõhisest täpsem, sest pikslite keskmistamisel väheneb haavelmüra (Blaes jt, 2005). Haavelmüra on radariandmetele iseloomulik kõrgsageduslik müra.

Cai jt (2018) on konkreetsemalt põllukultuuride klassifitseerimisel kirjeldanud kahte peamist teadusmaailmas levinud strateegiat. Esimesel juhul kasutatakse vaid spektraalset infot ühest hooajasisesest satelliidipildist, teisel juhul lisatakse spektraalsele dimensioonile ka ajaline mõõde ehk kasutatakse erinevate ajahetkede pilte kas ühest või isegi mitmest hooajast. Nimetatud strateegiad põhinevad erinevatel eeldustel. Ühe pildi kasutamisel eeldatakse, et erinevatel põllukultuuridel on konkreetsel ajahetkel iseloomulikud ja üksteisest eristuvad spektraalsed omadused. Piltide aegrea analüüsil lähtutakse aga asjaolust, et igal kultuuril on hooaja jooksul iseloomulik kasvukäik, mis väljendub spektraalsete parameetrite aegrea kujus. Terve kasvuperioodi piltide töötlemine nõuab küll rohkem ressursse, ent on üldiselt parandanud klassifitseerimise täpsust (Gómez jt, 2016; McNairn jt, 2009a).

Järgmine suur dilemma põllukultuuride klassifitseerimisel on andmestiku valik. Jättes kõrvale spetsiifilised satelliidimissioonid ja sensorite platvormid (droon, lennuk, satelliit), võib üldistades eristada optilist ja radarandmestikku. Mõlemal andmestikul on selged eelised ja puudused, mida lühidalt kirjeldati juba Copernicuse programmist rääkides. Optilised pildid sisaldavad väga rikkalikku spektraalset infot, ent on häiritud pilvedest ja sõltuvad valgustingimustest. Radari andmestik on saadaval sõltumata ilmaoludest ja valgustingimustest, ent on mürarikkam ja raskemini interpreteeritav. Mitmed uuringud (Van Tricht jt, 2018; Inglada jt, 2016; Blaes jt, 2005) on leidnud, et kõige paremaid tulemusi annab mõlema andmestiku kombineerimine, ja seda eriti operatiivsüsteemides, kus on oluline peaaegu reaalajaliste tulemuste esitamine. Kaugseire andmestiku täienduseks on kasutatud ka taimede fenoloogilist infot (Bargiel, 2017), ent sellise meetodi puuduseks on autorite enda sõnul universaalsuse puudumine, kuna ilma tugeva piirkonna ja põllukultuuride taustateadmiseta pole see klimaatiliselt ja fenoloogiliselt erinevatel aladel kasutatav.

Viimaseks, ent mitte vähem oluliseks kaalutluseks satelliidipiltide klassifitseerimisel on klassifitseerimisalgoritmi valik. Digitaalsed klassifitseerimisalgoritmid võib jagada kaheks peamiseks grupiks – parameetrilised, mis eeldavad andmete normaaljaotust, ja mitteparameetrilised, mis normaaljaotust ei eelda (Aunap jt, 2014: 63). Viimaste hulka kuuluvad ka erinevad masinõppe algoritmid, mis viimastel kümnenditel on satelliidipiltide klassifitseerimisel üha enam kasutust leidnud ja annavad üldiselt paremaid tulemusi kui traditsioonilised parameetrilised klassifikaatorid, eriti kui tegu on keerulise mitmemõõtmelise tunnusruumiga andmestikuga (Maxwell jt, 2018).

2.5 Masinõppe kasutamine kaugseire andmete klassifitseerimisel

Maxwell jt (2018) on välja toonud kuus nii-öelda küpset masinõppealgoritmi, mida on laialdaselt testitud ja kasutatud nii teadusmaailmas kui ka operatiivsetes kaugseire rakendustes: tugivektorklassifitseerija (*support vector machine*, SVM), otsustuspuu (*single decision tree*, DT), juhumets (*random forest*, RF), võimendatud otsustuspuu (*boosted decision tree*, boosted DT), tehisnärvivõrk (*artificial neural networks*, ANN) ja k-lähim naaber (*k-nearest neighbour*, k-NN). Tehisnägemises laialdast kasutust leidnud sügavad närvivõrgud (*deep neural networks*, DNN) on jõudnud ka põllukultuuride eristamiseni ja näidanud kohati isegi paremaid tulemusi kui eelnimetatud „küpsed" algoritmid (Kussul jt, 2017; Cai jt, 2018).

Sobiva klassifikaatori valimine võib olla keeruline, sest teaduskirjanduses avaldatud algoritmide võrdlused on andnud vastuolulisi tulemusi. Maxwelli jt (2018) arvates võib vastuolude üheks põhjuseks olla kasutatud protseduuride võrreldamatus. Samas on ka väga kindlate protseduurireeglite ja erinevate andmestike peal tehtud võrdlustes selgunud, et masinõppe algoritmide universaalset superstaari, mis annaks kaugseire andmete klassifitseerimisel alati parimaid tulemusi, ei ole olemas. Mõned üldised soovitused siiski on. Esiteks, ansambelmeetodid (näiteks RF, boosted DT) toimivad üldjuhul paremini kui üksikud klassifikaatorid (näiteks DT). Teiseks, kuna universaalset „parimat" algoritmi pole, tuleks katsetada alati mitme erineva algoritmiga. Suure tõenäosusega määrab optimaalse mudeli klassifitseerimisülesande spetsiifika (näiteks klasside arv, treeningandmete iseloom ja tunnusruum) (Maxwell jt, 2018).

Algoritmi valikuga klassifitseerimisülesande lahendamisel tehtavad valikud veel ei lõpe. Kõigi eespool nimetatud masinõppe algoritmide rakendamisel tuleb kasutajal määrata hüperparameetrid, mis kontrollivad algoritmi tööd ja võivad mõjutada klassifitseerimistäpsust. Kuigi kasutada saab väljakujunenud vaikeväärtusi, on empiiriline erinevate väärtuste läbikatsetamine reeglina vajalik, kui tahta saavutada parimaid tulemusi. Erinevate klassifikaatorite hüperparameetrite hulk ja optimeerimise keerukus on tihti kaalukeeleks, mille põhjal klassifikaatorit valitakse (Maxwell jt, 2018). Hüperparameetrite lihtne optimeerimine on näiteks RFi eeliseks ANNi ja SVMi ees.

Lisaks hüperparameetritele on võimalik optimeerida klassifitseerimise tunnuseid ehk karakteristikuid (*features, predictor variables*), mille põhjal algoritmid otsuseid vastu võtavad. Võiks ju arvata, et mida rohkem sisendtunnuseid, seda parem? See provokatiivne küsimus väärib antud magistritöö teemapüstituses eraldi peatükki.

2.6 Tunnuste valimise teoreetilised lähtekohad ja vajadus

Potentsiaalselt sisaldab iga lisandunud tunnus uut infot klassifitseerimisotsuse tegemiseks. Näiteks multispektraalsete satelliidipiltide puhul võib iga spektraalkanal täpsustada mingile klassile iseloomulikku spektraalset signatuuri. Hüperspektraalsete piltide puhul on eristatavaid lainepikkuse vahemikke ehk tunnuseid juba mitu korda rohkem.

Tegelikkuses ei ole aga kõik tunnused klassifitseerimisel võrdväärsed ja neid jagatakse olulisteks (*relevant*), mitteolulisteks (*irrelevant*), ülearusteks (*redundant*) ja eksitavateks (*misleading*) (Khalid jt, 2014). Tunnusruumi keerukuse ja dimensionaalsuse kasv võib hoopis vähendada klassifitseerimistäpsust. Seda probleemi tuntakse ka kui „dimensionaalsuse needust" või – spetsiifilisemalt kaugseires – Hughesi fenomeni. Eelkõige kummitab see parameetrilisi klassifikaatoreid, ent esineb ka masinõppe algoritmide puhul. Väike treeningandmete arv on samuti Hughesi fenomeni soodustavaks asjaoluks. Üldiselt saab kõige paremini keeruliste tunnusruumidega hakkama RF (Maxwell jt, 2018).

Teine põhjus kasutatavate tunnuste optimeerimiseks peitub mudeli lihtsustamises. Kui vähemate tunnustega on võimalik saavutada samaväärne täpsus, siis milleks kulutada arvutus- ja salvestusressursse mitteoluliste, ülearuste (rääkimata eksitavate!) tunnuste arvutamiseks, hoiustamiseks ja töötlemiseks? Duro jt (2012) töös selgus, et maakattetüüpide klassifitseerimisel RF algoritmiga ei muutunud klassifitseerimistäpsus, kui esialgset tunnusruumi (418 tunnust) vähendati 60% võrra.

Khalid jt (2014) on jaganud tunnusruumi optimeerimise meetodid kahte peamisse gruppi: tunnustest esindusliku alamhulga valimine (*feature selection*) ja tunnuste kombineerimine/transformeerimine (*feature extraction/transformation*). Erinevatel meetoditel on eeliseid ja puudusi. Tunnustest alamhulga valimisel võib valikust välja jäänud tunnuste info kaotsi minna, samas info iga valikusse pääsenud konkreetse tunnuse olulisusest jääb alles. Tunnuste kombineerimisel jääb alles kogu tunnusruumi info, samas muutub ebaselgemaks iga algtunnuse olulisus. Valik, millist teed tunnusruumi optimeerimisel minna, tuleb teha andmestiku iseloomu ja valdkonna põhjal, kus neid meetodeid rakendatakse.

2.7 Juhumetsa algoritm ja Gini ebapuhtus

Juhumetsa algoritm on kaugseire andmete klassifitseerimisel väga populaarne ja seda mitmel põhjusel. Võrreldes teiste masinõppe algoritmidega on see näidanud häid klassifitseerimistäpsusi, on kiire, vajab vähe kasutajapoolseid parameetreid, on suhteliselt tundetu treeningandmestiku vigade ja ülesobitamise suhtes, saab hästi hakkama keerukate tunnusruumidega ja võimaldab lihtsalt tunnuste olulisust tuvastata (Belgiu ja Drăguţ, 2016). Nagu varem kirjeldatud, ei ole nimetatud omadused absoluutsed, kuid sage juhumetsa kasutamine nii teadustöödes kui rakendustes lubab selle algoritmi tõsta masinõppe algoritmide hüpoteetilisele poodiumile.

Juhumets on juhendatud ansambelõppe algoritm, kus mitmed otsustuspuud (*decision trees*) nii-öelda hääletavad kõige populaarsema klassi poolt. Otsustuspuude loomisel valitakse reeglina juhuslikud treeningnäidiste alamhulgad ja juhuslikud tunnuste alamhulgad. Paljude üksteisest sõltumatute otsustuspuude koostöö ja „enamuse hääl“ vähendavad üksikute puude kallutatust, ebatäpsust ja ülesobitamist (Breiman, 2001).

Minimaalselt peab kasutaja määrama kaks parameetrit: kui palju eraldiseisvaid otsustuspuid luuakse ja kui suure hulga tunnuste põhjal otsustuspuu sõlmedes otsuseid tehakse. Otsustuspuu sõlmeks (*node*) nimetatakse puu elementi, kus tunnuste väärtuste põhjal toimub näidiste lahknemine erinevateks harudeks ehk oksadeks. Eristatakse juur- (*root*), sise- (*internal*) ja terminaalsõlmi (*leaf*). Terminaalsõlmest enam harusid ei välju, vaid antakse klassifitseerimisotsus.

Otsustuspuude sisesõlmedes parimate lahknemisotsuste tegemiseks püüab algoritm leida tunnuse, mis eristab klasse kõige paremini. Otsuse ehk lahknemise kvaliteedi hindamiseks kasutab ta jagunenud näidiste alamhulkade ebapuhtuse (*impurity*) määra (tuntud ka kui Gini ebapuhtus või indeks). Lõppeesmärk on ebapuhtusest täielikult lahti saada, et alamhulgas oleks vaid ühte klassi kuuluvad näidised. Sellest otsustuspuude ja juhumetsa algoritmi omadusest (või ka treenimise kõrvalproduktist) saab infot iga tunnuse suhtelise olulisuse kohta, sest mida rohkem mingi lahknemise puhul ebapuhtus väheneb, seda olulisem kasutatud tunnus oli. Mõõdikut, mis Gini indeksi põhjal tunnuste suhtelist olulisust illustreerib, nimetatakse Gini olulisuseks. Lihtsustades võib öelda, et Gini olulisus näitab, kui tihti mingit tunnust otsuste tegemisel kasutati ja kui suur oli selle eristav mõju konkreetses klassifitseerimisülesandes (Breiman, 2001; Menze jt, 2009).

3. Andmestik ja eeltöötlus

Kasutatud sisendandmestik pärineb RITA1/02-52 põllumaade alamprojekti (Voormansik jt, 2020) käigus loodud andmekogust, mis sisaldab kahe põllumajandushooaja (2018, 2019) kohta Eesti põldudele arvutatud tunnuseid. Andmestik on avalikult kättesaadav siit: http://datadoi.ee/handle/33/316.

Selles peatükis kirjeldatud andmetöötluse etapid hõlmavad vaid spetsiifilist juhumetsa mudeli treenimiseks läbiviidud sisendandmestiku eeltöötlust. „Andmete päritolu" alampeatükis kirjeldatakse lühidalt sisendandmete allikaid, kuid ei laskuta selle andmestiku loomise detailidesse. Sisendandmete tootmise ja töötluse kohta leiab rohkem infot RITA1/02-52 põllumaade alamprojekti lõpparuandest (Voormansik jt, 2020).

3.1 Andmete päritolu

Eesti põldude andmekiht pärineb Põllumajanduse Registrite ja Informatsiooni Ametist (PRIA). Infot iga põllul kasvava kultuuri kohta saab PRIA reeglina põllumeestelt, kes on kohustatud seda iga-aastaselt pindalatoetusi taotledes edastama. Osaliselt on seda infot kontrollinud ja parandanud PRIA inspektorid. Andmestik hõlmab kahte aastat – 2018 ja 2019.

Põllukultuuride klassifikatsiooni pani samuti paika PRIA, kes enda vajadustest ja kultuuride sarnasusest lähtuvalt jaotas kõik Eestis kasvatatavad põllukultuurid 28 klassi. Põllugeomeetria ja kultuuri omavahelise sidumise tegi Tartu Ülikooli (TÜ) geograafia osakond, kes arvutas ka põldude normeeritud asukoha koordinaadid.

Kõik kaugseire parameetrid Sentinel-1 ja Sentinel-2 satelliidipiltidelt on arvutatud OÜ KappaZeta poolt, kasutades Eesti riikliku satelliidiandmete keskuse ESTHub infrastruktuuri.

Põldude mullatüübi leidmiseks on kasutatud 1:10 000 mõõtkavas Eesti mullastiku kaarti (Maa-amet, 2021). Muldade rühmitamiseks on erinevaid viise. Selles töös kasutatud lähenemine põhineb litoloogilis-geneetilise ja niiskuse maatriksil ning mullad on jagatud 23 erinevasse klassi, mille hulgast alati üks, pindalaliselt valdav, on omistatud antud põllu mullatüübiks. Töötluse viis läbi TÜ geograafia osakond.

Põllupõhised sademesummad arvutas TÜ geograafia osakond, kasutades selleks TÜ füüsika instituudi poolt RITA1/02-52 meteoroloogia alamteema (Post jt, 2020) raames välja töötatud täppissademete kaarte, mis kombineerivad ilmaradarite ja mõõtejaamade andmeid. Ööpäeva keskmised temperatuurid pärinevad samuti TÜ füüsika instituudi poolt loodud Harmonie ilmamudeli kaartidelt, mida töötles geograafia osakond.

KappaZeta kogus kõik erinevatest allikatest pärit andmed ühte andmekogusse ja viis läbi kvaliteedikontrolli.

3.2 Tunnuste kirjeldus

Tunnuskomplekti saab jagada viide loogilisse gruppi, mis omakorda jagunevad veel alamgruppidesse. Kõik satelliiditunnused on arvutatud rasterpiltidelt (pikslite väärtused) ja üldistatud põllu geomeetria piires – antud juhul on võetud mediaanväärtus. Tunnuste täpsed nimetused on kujutatud kursiivis.

1) Sentinel-1 radarsatelliidi piltidelt arvutatud tunnused:

- *s1_cohvh_median* ja *s1_cohvv_median* – VH- ja VV-polarisatsiooni 6 päeva interferomeetriline koherentsus. Koherentsus on kompleksse radari pildipaaril

arvutatud normeeritud sarnasuse mõõde, kus väärtus 0 tähendab täiesti erineva tagasihajumise mustriga pildipaari ja väärtus 1 tähendab ideaalselt kokkulangeva tagasihajumise mustriga pildipaari. Taimkatte korral vastab kõrge koherentsus madalale taimestikule või selle puudumisele ning madal koherentsus lopsakale taimestikule (Tamm jt, 2016).

- *s1_s0vh_median* ja *s1_s0vv_median* – VH- ja VV-polarisatsiooni tagasihajumine.
- *s1_vhvv_median* – VH- ja VV-polarisatsiooni tagasihajumiste suhe.

2) Sentinel-2 optilise satelliidi piltidelt arvutatud tunnused:

- üksikute spektrikanalite peegeldustegurid (*s2_b2_median, s2_b3_median, s2_b4_median, s2_b5_median, s2_b6_median, s2_b7_median, s2_b8_median, s2_b8a_median, s2_b11_median, s2_b12_median*)
- erinevate spektrikanalite kombineerimisel arvutatud indeksid (*s2_ndvi_median, s2_ndwi_median, s2_ndvire_median, s2_tc_wetness_median, s2_tc_vegetation_median, s2_tc_brightness_median, s2_misra_yellow_vegetation_median, s2_psri_median, s2_wri_median*)

3) põllu asukoha tunnused:

- *x_norm_loc* ja *y_norm_loc* – põllu normeeritud asukohakoordinaadid.

4) mullastiku tunnus:

- *soil_type4* – põllu domineeriv mullatüüp.

5) ilmastiku tunnused:

- *temp_mean* – päeva keskmine temperatuur
- *precip_mean_3h* ja *precip_mean_24h* – Sentinel-1 satelliidi ülelennule eelneva 3 ja 24 tunni sademesumma (mm)

Põllu asukoha ja mullastiku tunnused olid üksikparameetrid – iga põllu kohta parajasti üks väärtus. Sentinel-1 ja -2 ning ilmastiku tunnused aegreana. Sentinel-1 tunnused iga ülelennu kuupäeva kohta, Sentinel-2 tunnused iga pilvevaba ülelennu kuupäeva kohta ning ilmastiku tunnused iga päeva kohta.

3.3 Kasutatud tarkvara ja tööriistad

Kogu töötlusahel ehitati üles Pythonis, kasutades peamiselt järgmisi mooduleid ja teeke:

- psycopg2[1] – andmebaasiühenduste loomiseks ja päringute tegemiseks
- pandas[2] – andmete korrastamiseks klassifikaatorile sobivasse formaati
- scikit-learn[3] – klassifitseerimisülesande sooritamiseks
- seaborn[4] – andmete visualiseerimiseks

Eeltöötluseks ja klassifitseerimiseks kasutati riistvara, mille protsessoriks oli Intel® Core™ i7-8700K (6 tuuma, 12 lõime) ja muutmälu 64 GiB. Kogu Eesti ühe aasta andmete (*ca* 100 000 põldu) eeltöötlus (sealhulgas ka andmebaasist andmete laadimine ja lõpptulemuse

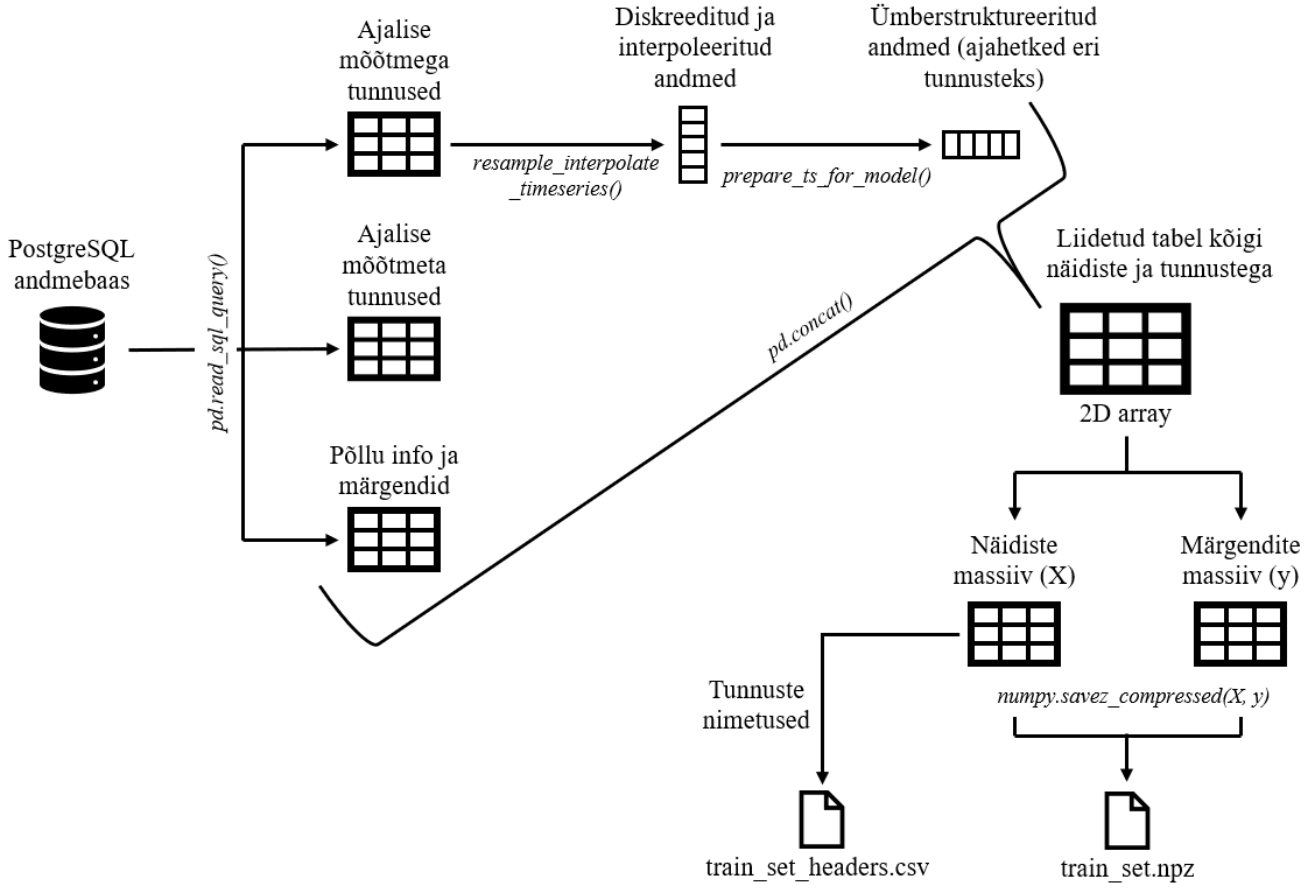
[1] https://pypi.org/project/psycopg2/
[2] https://pandas.pydata.org/
[3] https://scikit-learn.org/stable/
[4] https://seaborn.pydata.org/index.html

väljundfaili salvestamine) võttis aega u 1h 20 minutit ja mudeli sobitamine koos hinnangute andmisega (sealhulgas andmefaili sisse lugemine ja tunnuste olulisuste määramine) 2 minutit.

3.4 Spetsiifiline eeltöötlus juhumetsa mudeli jaoks

Kogu sisendandmestik paiknes PostgreSQL andmebaasis, kust eeltöötluse esimese sammuna andmed päriti. Andmebaasis oli tehtud juba osa eeltöötlusest ja koondatud andmed eri tabelitest vaadetesse, et lihtsustada päringute tegemist. Siiski tuli teha päring kuuest erinevast andmebaasi vaatest.

Andmebaasist pärides piirati aegrea tüüpi andmete ulatust 1. maist 31. augustini (123 päeva). Lisaks jäeti kõrvale need põllud, millel oli vähem kui 10 aegreapunkti ükskõik millisest Sentinel-1 (S1) või Sentinel-2 (S2) tunnusest (selliseid põlde oli 2018: 355 ja 2019: 300). Kuna S1 ja S2 mõõtmisi ei ole kõigi päevade kohta, interpoleeriti mõõtmiste vahele jäävatele päevadele väärtused lineaarselt (meetod: *slinear*) ning tühjadele päevadele aegrea alguses ja lõpus omistati 0 väärtus. Temperatuuri ja sademete aegread olid täielikud ja mõõtmistulemused iga päeva kohta olemas. Juhul, kui seal peaks siiski esinema puuduvaid mõõtmisi, siis temperatuuri puhul interpoleeritakse need sarnaselt satelliitmõõtmiste aegridadega ja sademete puhul omistatakse neile lihtsalt väärtus 0.



Joonis 1. Andmete eeltöötluse peamised vaheetapid ja olulisemad funktsioonid.

Scikit-learn'i funktsioonide jaoks on vaja sisendandmestik viia kahemõõtmelise massiivi (*2D array*) kujule, kus ridadel on näidised (põllud) ja veergudes tunnused. Kuna antud juhul on osal tunnustest ka kolmas, ajaline mõõde, tuli igast tunnusest teha 123 (päevade arv) eraldiseisvat tunnust. Nii sai näiteks temperatuuri tunnusest temp_121, temp_122...temp_243 (kus arv on aasta päev, 1=1. jaanuar ja 365=31. detsember). Pärast tabeli „lamendamist" tekkis iga põllu kohta kokku 3322 tunnust. Tunnuste hulk kujunes järgmise valemi järgi:

3 ajaliselt muutumatut tunnust + 27 ajas muutuvat tunnust x 123 päeva = ***3324 tunnust****.*

Lihtsustatud skeem massiivi (*2D array*) ülesehitusest on toodud tabelis 1.

Tabel 1. Korrastatud andmete struktuuri näidis. Kahe aasta maatriksite suuruseks oli 111 685 rida, 3324 veergu (2018) ja 112 177 rida, 3324 veergu (2019).

	märgend	**x_norm_loc**	**y_norm_loc**	**mullatüüp**	**temp_121**	**...**	**temp_243**	**...**	**s2_wri_121**	**...**	**s2_wri_243**
põllu_id											
1	2	0.439542	0.448561	45	7.34	...	13.23	...	0.194852	...	0.195024
2	4	0.395841	0.384254	21	6.45	...	12.45	...	0.218569	...	0.22893
…	...	...	...	...	...	...	...	...	...	...	...
112127	3	0.352769	0.382563	10	7.32	...	15.78	...	0.329528	...	0.332308

Põldudel kasvavate kultuuride klassid (ehk märgendid) eraldati iseseisvasse *Numpy* massiivi *(array*) ning eemaldati andmete üldmassiivist. Eeltöötluse tulemusel tekkinud korrastatud näidiste/tunnuste massiiv koos märgendite massiiviga salvestati pakitud NPZ-faili. (suurusega 2GB), kust mudeli sobitamise tarkvara selle uuesti sisse sai lugeda. Sel viisil on mudeli sobitamise sammu võimalik teostada ajakulukat andmete ettevalmistamise osa läbimata. Eeltöötluse olulisemad etapid ja lähtekoodi (lisa V) funktsioonid on kujutatud joonisel 1.

4. Juhumetsa (*Random Forest*) mudeli sobitamine

Magistritöö peamine eesmärk ei olnud leida parimat põllukultuuride klassifitseerimismudelit, vaid hinnata klassifitseerimisel osalevate tunnuste olulisust. Tunnuste olulisuse infoni jõudmiseks on aga mudeli sobitamine siiski vajalik. Mida paremini mudel töötab, seda usaldusväärsem on ka tunnuste olulisuse info.

4.1 Algoritmi valik ja andmestiku jaotamine

Scikit-learn teegis on kaks klassifitseerimisalgoritmi, mis baseeruvad juhusliku otsustuspuude juhendatud õppe meetodil (*randomized supervised learning method*): juhumets (*Random Forest*) ja eriti juhuslikud otsustuspuud (*Extremely Randomized Trees*). Klassifitseerimisülesandes kasutati esimest ehk juhumetsa[5], mida on täpsemalt kirjeldatud taustainfo peatükis.

Andmestik jagati treening- (80 %) ja testandmestikuks (20 %), millest esimest kasutati klassifitseerija sobitamiseks ja teist selle täpsuse ning universaalsuse hindamiseks. Andmestiku jaotust aastate kaupa illustreerib tabel 2.

Tabel 2. Kogu andmestiku jagunemine treening- ja testandmestikuks (näidiste arv).

	2018	**2019**
Treeningandmestiku suurus	89348 (80 %)	89738 (80 %)
Testandmestiku suurus	22337 (20 %)	22435 (20 %)
Kogu andmestik	111685 (100 %)	112173 (100 %)

Kasutatud klassifitseerija (*RandomForestClassifier)*[6] hüperparameetrid koos lühiselgitustega on toodud tabelis 3.

Sobivaimate parameetrite määramisel kasutati *scikit-learn'*i funktsionaalsust *RandomizedSearchCV*[7], millega katsetati 2019. aasta treeningandmete peal erinevaid variatsioone järgmistest hüperparameetritest: *n_estimators, max_features, max_leaf_nodes, min_sample_split* ja *bootstrap*. Lõplik otsus võeti vastu mudeli kaalutud keskmise F1-skoori (*f1_weighted*) hindamisel. F1-skoori selgitus on toodud peatükis 4.2. Etteantud piirides muutuvate juhuslike parameetritega läbi viidud 20 testi tulemused on esitatud lisas I.

Testiti järgmisi hüperparameetrite väärtusi:

- *n_estimator* – 100, 200, 300, 400, 500.
- *max_features* – 'sqrt', 'log2'.
- *max_leaf_nodes* – None, 100, 200, 300, 400, 500.
- *min_sample_split* – 2, 5, 10.
- *bootstrap* – True, False.

[5] https://scikit-learn.org/stable/modules/ensemble.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Tabel 3. Juhumetsa algoritmi hüperparameetrite väärtused koos selgitustega. Jämedas kirjas on parameetrid, mida püüti optimeerida. Teiste puhul kasutati vaikeväärtusi.

Hüperparameeter	**Väärtus**	**Selgitus**
bootstrap	False	Kas kasutatakse juhuslikke ja duplitseeritud näidiseid puude loomisel. Antud juhul kasutatakse kogu andmestikku.
ccp_alpha	0.0	Parameeter puu kärpimise (*Minimal Cost-Complexity Pruning*) kontrollimiseks. Antud juhul seda ei teostata.
class_weight	None	Klasside kaalud. Kui *None*, siis kõik klassid kaaluga 1.
criterion	'gini'	Funktsioon, mis mõõdab sõlme hargnemise kvaliteeti. Kaks võimalust – *'gini'* või *'entropy'*.
max_depth	None	Otsustuspuude maksimaalne sügavus. Antud juhul laiendatakse otsustuspuid seni, kuni kõik lehed on puhtad või sisaldavad vähem näidiseid kui *min_samples_split.*
max_features	'log2'	Tunnuste arv parima hargnemisotsuse tegemisel. Antud juhul siis leitakse nii: *log2(kõik tunnused).*
max_leaf_nodes	None	Maksimaalne terminaalsõlmede arv. Antud juhul piiramatu arv.
max_samples	None	Näidiste arv, mida kasutatakse iga otsustuspuu treenimisel. Antud juhul kasutatakse kõiki näidiseid, aga kui *bootstrap=True*, tehtaks selle arvu piires valik kõigi hulgast.
min_impurity_decrease	0.0	Parameeter puu hargnemiste ebapuhtuse piiritlemiseks.
min_impurity_split	None	Ebapuhtuse-põhine piirmäär hargnemiste peatamiseks.
min_samples_leaf	1	Minimaalne näidiste arv terminaalsõlmes (*leaf node*).
min_samples_split	2	Minimaalne näidiste arv, et sisesõlm (*internal node*) hargneks edasi.
min_weight_fraction_leaf	0.0	Minimaalne osa kaalutud näidiste summast terminaalsõlmes. Oluline siis, kui kasutatakse klasside kaalumist.
n_estimators	300	Otsustuspuude arv juhumetsas.
n_jobs	-1	Paralleelselt jooksutatavate tööde arv. Antud juhul kasutatakse kõiki saadaolevaid protsessoreid.
oob_score	False	*Out-of-bag (OOB)* näidiste kasutamine täpsuse hindamiseks.
random_state	42	Juhuslikkust (kui seda kasutatakse) kontrolliv parameeter.
verbose	0	Kontrollib, kui palju infot väljastatakse sobitamise ja ennustamise käigus.
warm_start	False	Iga kord sobitatakse täiesti uus mudel ja ei taaskasutata eelmise sobitamise infot.

4.2 Klassifitseerimise tulemused

Mudeli täpsuse ja sobivuse hindamiseks kasutatakse erinevaid mõõdikuid. Kuigi allpool kirjeldatud mõõdikud on tavaliselt kasutuses binaarsetes klassifitseerimisülesannetes (jah/ei probleemid), saab neid siiski kasutada ka mitmeklassiliste tulemuste hindamiseks:

- **täpsus** (*precision*) = konkreetse klassi õigesti klassifitseeritud näidised / kõik selleks klassiks klassifitseeritud näidised (TP/(TP + FP))*
- **saagis** (*recall*) = konkreetse klassi õigesti klassifitseeritud näidised / kõik selle klassi tegelikud näidised (TP/(TP + FN))
- **F1-skoor** (*F1 Score*) = 2 × (täpsus × saagis)/(täpsus + saagis). Kombineerib täpsuse ja saagise üheks mõõdikuks.

* TP – õigepositiivne *(True Positive);* FP – valepositiivne *(False Positive);* TN – õigenegatiivne *(True Negative);* FN – valenegatiivne *(False Negative).*

Kui püüda eelnev panna põllukultuuride konteksti, võiks näide olla järgmine:

<u>Täpsus</u> näitab, kui palju oli tegelikult kartulipõlde kõigi mudeli poolt kartulipõlluks klassifitseeritud põldude seas. Juhtub, et mudel klassifitseerib ka näiteks mõned peedi- või porgandipõllud kartuliks (valepositiivsed).

<u>Saagis</u> näitab, kui suure osa tegelikest kartulipõldudest mudel üles leidis. Juhtub, et vahel mudel määrab kartulipõllu hoopis peedipõlluks (valenegatiivsed).

Mõõdikute arvutuskäigud on illustreeritud tabelis 4, mis kujutab hüpoteetilist eksimismaatriksit nelja kultuuriga klassifikatsioonile, kus andmestikus oli igat kultuuri 10 põldu.

Tabel 4. Mõõdikute arvutamise põhimõtted eksimismaatriksil. Maatriksi ridadel on kultuuride tegelikud ja tulpades ennustatud klassid/märgendid.

		Ennustatud *(predicted)*								
		Kultuur 1	Kultuur 2	Kultuur 3	Kultuur 4	**Täpsus** *(precision)*		**Saagis** *(recall)*		**F1-skoor** *(F1 score)*
Tegelik *(true)*	Kultuur 1	9	1	0	0	9/(9+0+0+1)=	0,90	9/(9+1+0+0)=	0,90	0,90
	Kultuur 2	0	7	3	0	7/(1+7+2+1)=	0,64	7/(0+7+3+0)=	0,70	0,67
	Kultuur 3	0	2	7	1	7/(0+3+7+3)=	0,54	7/(0+2+7+1)=	0,70	0,61
	Kultuur 4	1	1	3	5	5/(0+0+1+5)=	0,83	5/(1+1+3+5)=	0,50	0,63

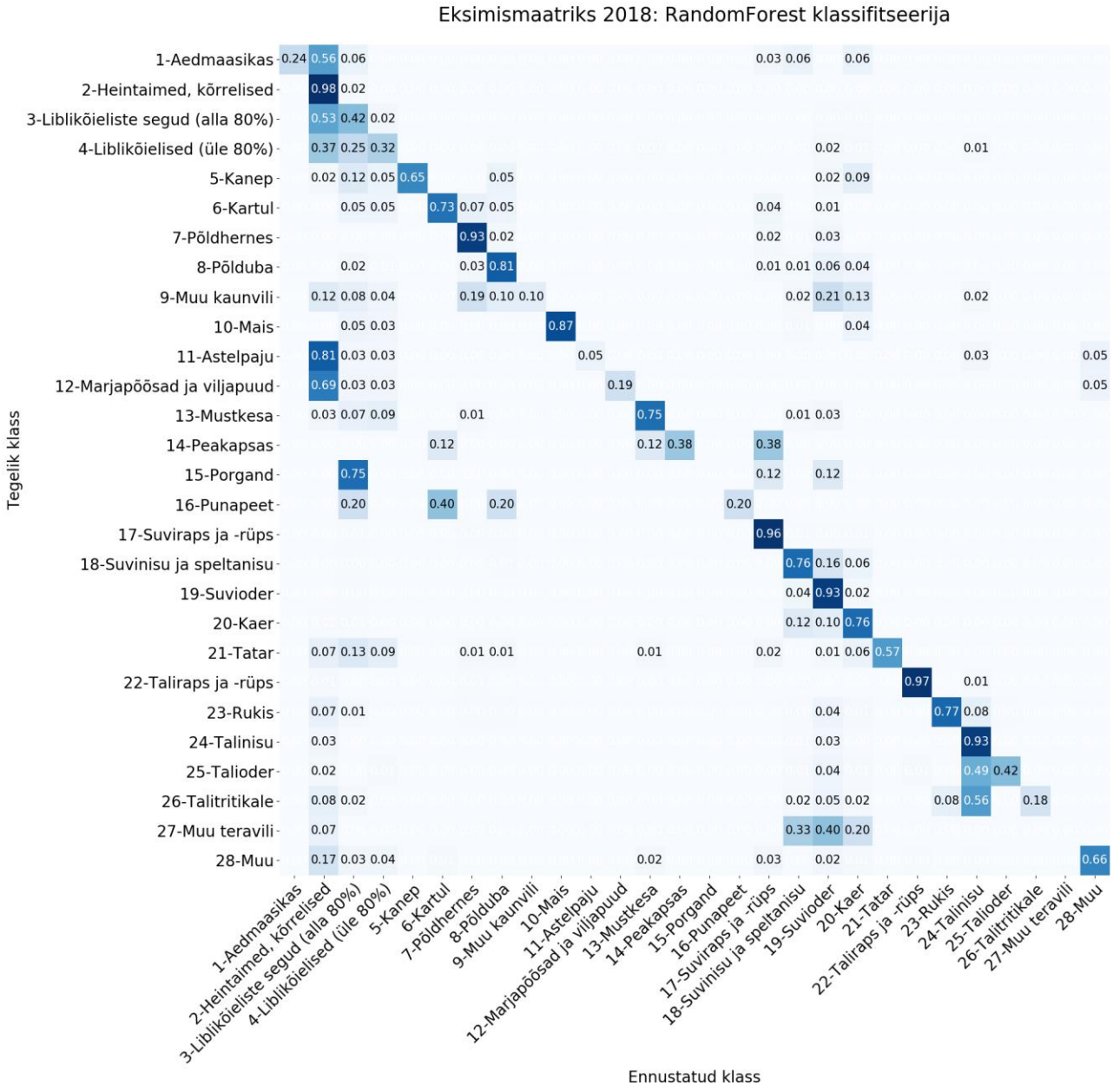
Küsimusele, millisest mõõdikust lähtuda mudeli soorituse hindamisel, pole ühest vastust ning mõõdiku valik sõltub suuresti klassifitseerimisülesande ja andmestiku iseloomust.

Mõõdikuid saab ka klasside üleselt keskmistada või kaaluda. Kui tavalise keskmistamise puhul (*macro average*) liidetakse kõikide klasside mõõdikud kokku ja jagatakse klasside arvuga, siis kaalutud keskmistamisel (*weighted average*) kaalutakse iga klassi mõõdik enne

läbi selle klassi näidiste arvuga. Siin töös valiti sobivaim mudel kaalutud F1-skoori (*f1_weighted)* põhjal.

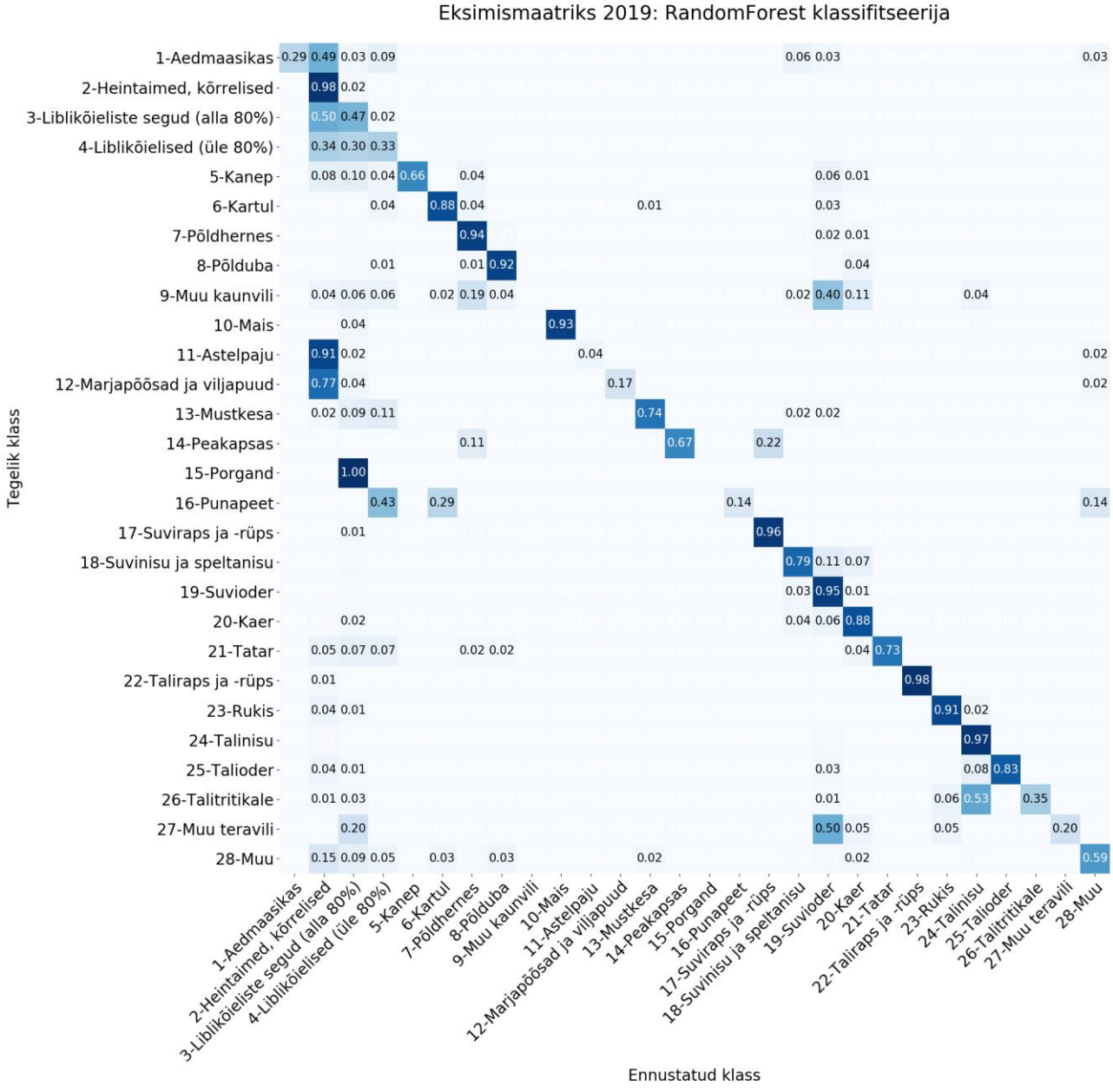
Joonisel 2 ja 3 on kujutatud mõlema aasta eksimismaatriksid ning klassipõhised mõõdikud on toodud lisas II.

Klassideülene kaalutud F1-skoor oli 2018. aasta testandmestiku puhul 0,82 ja 2019. aastal 0,85. See on heas kooskõlas RITA projekti raames väga sarnasel lähteandmestikul, ent närvivõrkudel treenitud kahe aasta mudeli testandmestiku kaalutud F1-skooriga (0,85). Võrreldud närvivõrkude mudel koosnes sisendkihist (153 kuupäeva x 25 tunnust), sellele järgnevast tensori mõõtmeid lamendavast kihist (*flatten layer*) ja kahest täissidusast



Joonis 2. Juhumetsa 2018. aasta testandmestiku klassifitseerimise eksimismaatriks. Maatriksi peadiagonaalil on iga klassi saagised.

vahekihist (*dense*). Mõlema täissidusa kihi järele oli lisatud ka andmeid normeeriv kiht (*batch normalization layer*). Esimeses täissidusas kihis kasutati aktivatsiooniks mittenegatiivset lineaarfunktsiooni (*Rectified Linear Unit*), teises sigmoid funktsiooni ja väljundkihis softmax aktivatsioonifunktsiooni (Voormansik jt, 2020). Tuleb rõhutada, et kuigi klassifitseerimisülesanne ja väljundklassid on nende kahe mudeli puhul samad, pole nende testandmestikud identsed, mistõttu tulemuste väga otseses võrdlemises tuleb ettevaatlik olla. Sarnased tulemused on pigem indikaatoriks, et juhumetsa mudeli andmetöötluses ja sobitamises pole esinenud suuri süstemaatilisi vigu.



Joonis 3. Juhumetsa 2019. aasta testandmestiku klassifitseerimise eksimismaatriks. Maatriksi peadiagonaalil on iga klassi saagised.

5. Tunnuste olulisuse hindamine

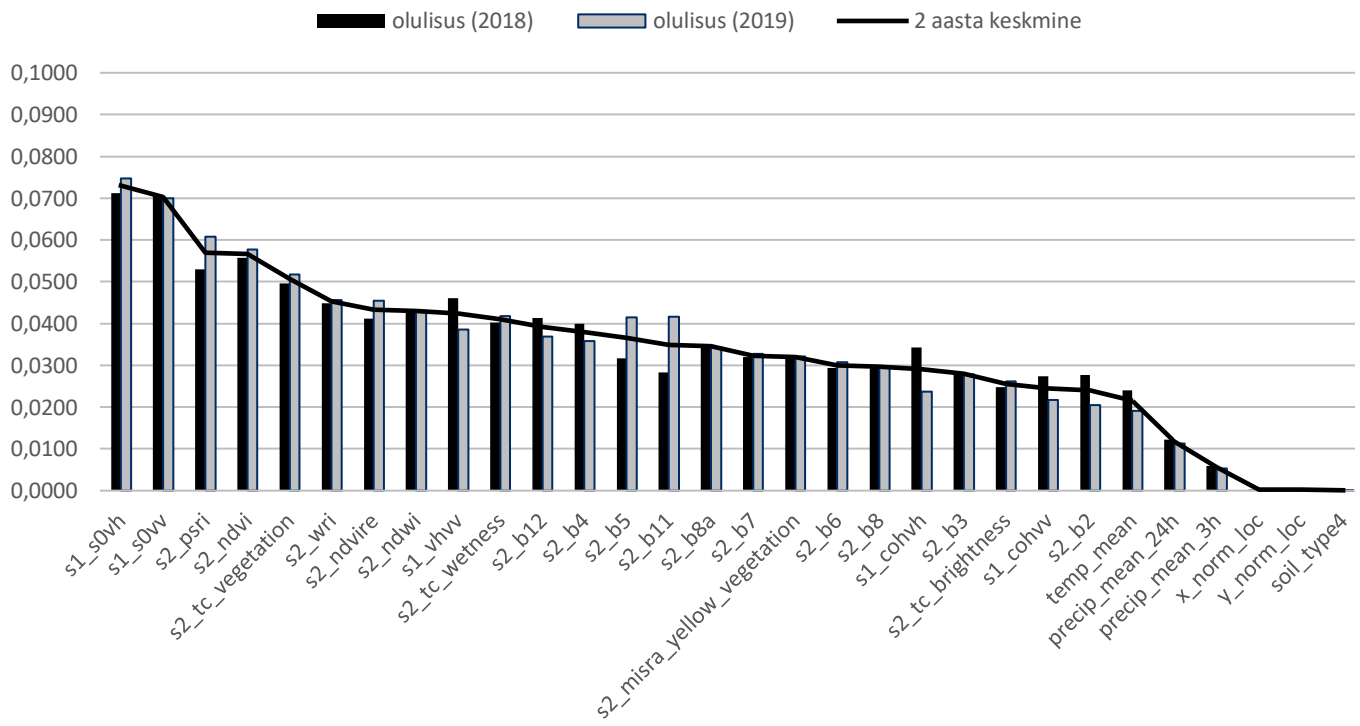
Klassifitseerimisel kasutati kõiki olemasolevaid tunnuseid, kuid iga tunnuse panus lõpptulemusse ei olnud tõenäoliselt sama. Selles peatükis kirjeldatakse juhumetsa treenimise kõrvalproduktina tekkinud tunnuste pingerea töötlust ja visualiseeritakse töötluse tulemused.

5.1 Kõigi tunnuste olulisuse töötlus ja hindamine

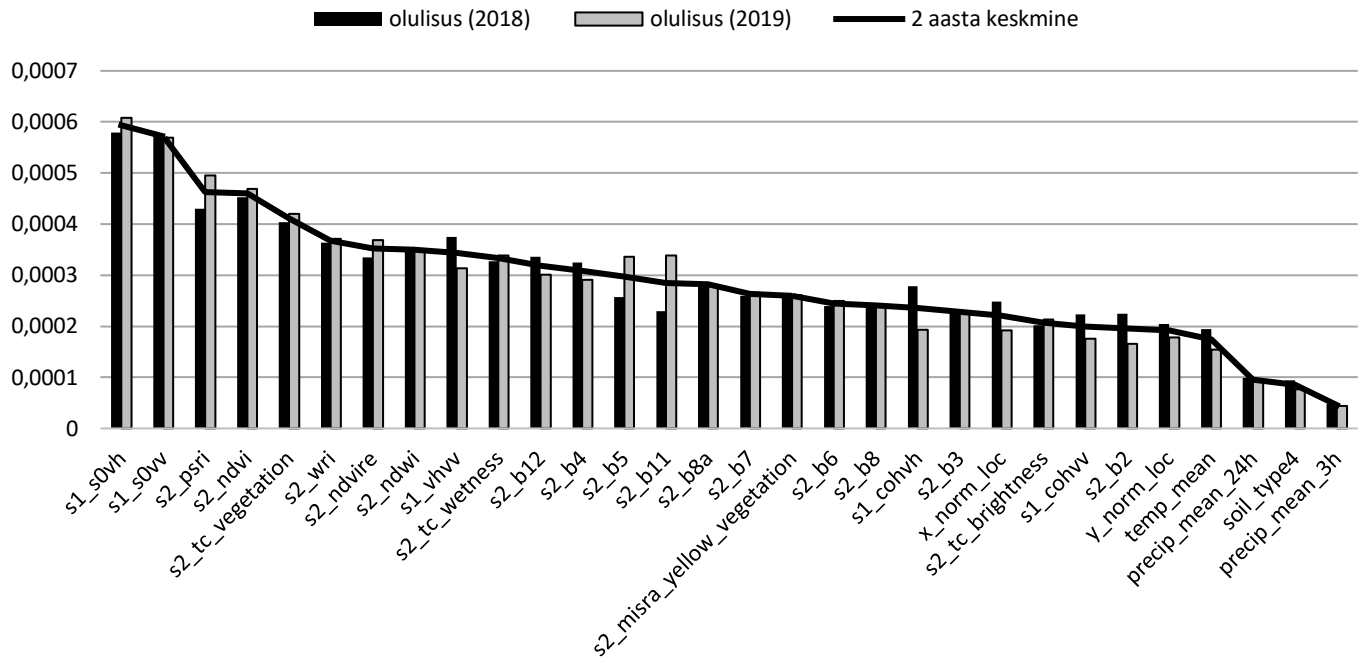
Juhumetsa klassifikaatorilt on ühe omadusena võimalik pärida tunnuste olulisuse hinnanguid (*feature_importances_)*. Tuntud ka kui Gini-olulisus või ebapuhtusepõhine tunnuste olulisus (*impurity based*). Täpsemalt on seda lähenemist kirjeldatud peatükis 2.7.

Vaikimisi leidis klassifikaator kõigi 3324 tunnuse olulisuse eraldi (väljastatakse listina), mis polnud aga antud juhul, kus ajaline mõõde oli ka eraldi tunnusteks tehtud, väga informatiivne. Lihtsustamaks tunnuste grupeerimist ja sorteerimist, viidi kogu tunnuste olulisuse info tabelisse (*Pandas DataFrame*). Selleks tekitati esimese sammuna sõnastik, kuhu kahest eraldi nimekirjast pandi kokku tunnuse nimetus (võti) ja olulisuse hinnang (väärtus), ning seejärel loodi sellest sõnastikust tabel, kus eraldi veergudesse eraldati tunnuse nimetusest „päeva“ ja „tunnuse“ osad. Nii tekkiski tabel, kus lisaks indeksile said veergudeks tunnuse nimi, päev ja olulisuse väärtus (lähtekoodis funktsioon *calculate_feature_importance()*).

Saadud tabelist grupeeriti read tunnuse nime põhjal, summeeriti antud tunnuse kõigi päevade olulisused (joonis 4) ja leiti ka tunnuste kõigi päevade keskmised olulisused (joonis 5).



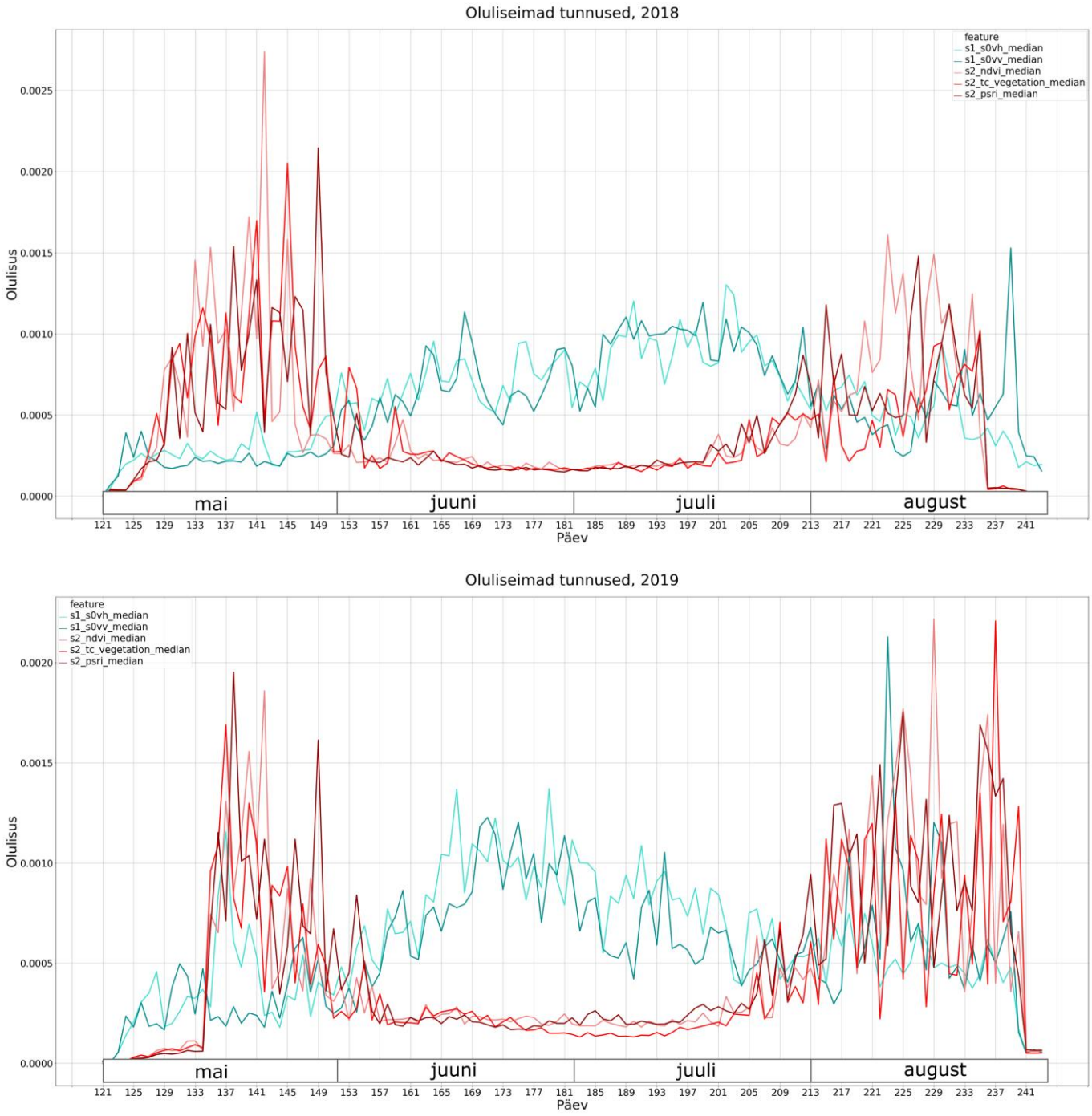
Joonis 4. Tunnuste kõigi päevade summeeritud Gini-olulisus 2018. ja 2019. aasta treeningandmete sobitamisel. Järjestatud kahe aasta keskmise väärtuse järgi, mis on kujutatud musta katkendjoonega. Mida suurem väärtus, seda olulisem tunnus.



Joonis 5. Tunnuste kõigi päevade keskmine Gini-olulisus 2018. ja 2019. aasta treeningandmete sobitamisel. Järjestatud kahe aasta keskmise väärtuse järgi, mis on ka kujutatud musta katkendjoonega. Mida suurem väärtus, seda olulisem tunnus.

Arusaadavalt on ajaliselt summeeritud graafikul viimased tunnused, millel ajaline mõõde puudus. Samas 5 kõige olulisemat tunnust on samad nii ajaliselt summeeritud kui keskmistatud lähenemisega – Sentinel-1 eri kanalite tagasihajumised ja Sentinel-2 kolm indeksit, mis kombineerivaid erinevaid spektrikanaleid.

Lisaks iga tunnuse erinevale olulisusele võivad tunnuste olulisused ka ajas muutuda. Kuna loodud tabelis oli meil eraldi veeruna eraldatud päevade indeksid, oli võimalik visualiseerida iga tunnuse olulisust hooaja vältel või ka kõiki tunnuseid koos hooaja jooksul. Kui vaadata 5 olulisima tunnuse olulisust ajas (joonis 6), siis on märgata, et optilise satelliidi Sentinel-2 tunnused (graafikul punaka värviga) on olulisemad hooaja alguses ja lõpus, samas kui radarsatelliidi Sentinel-1 tunnused (graafikul rohekad) omavad suuremat mõju hooaja keskel. Tulemusi interpreteeritakse põhjalikumalt arutelu peatükis.



Joonis 6. Viie olulisima tunnuse olulisus päevade lõikes. Punakate toonidega on kujutatud Sentinel-2 tunnuseid ja sinakasrohekate toonidega Sentinel-1 tunnuseid. 2018. aasta üleval ja 2019. aasta all.

6. Arutelu

Sobitatud juhumetsa mudelit oleks võimalik veel optimeerida. 20 juhuslikku katsetust 5 hüperparameetri erinevate väärtustega (vt lisa I) ei ole piisav, et leida parimat kombinatsiooni. Hetkel parima tulemuse andnud kombinatsioon oli ka ainuke, kus muutuja *max_leaf_nodes* ehk maksimaalne terminaalsõlmede arv oli piiramatu, mistõttu ei ole teada, milline oleks tulemus, kui see hüperparameeter paika jätta ja lasta teistel muutuda. Kuna mudeli optimeerimine polnud selle töö fookus, siis pigem tuleb seda primitiivset hüperparameetrite kohandamise katsetust võtta kui näidet kasutatud teegi funktsionaalsusest ja võimalustest.

Samas, juhumetsa eeliseks paljude teiste klassifikaatorite ees ongi väheste kasutajapoolsete parameetrite seadistamine. Üldiselt tuleks optimeerida vaid otsustuspuude (*n_estimators*) ja kasutatavate tunnuste arvu (*max_features*). On leitud (Maxwell jt, 2018), et puude arv ei oma tegelikult lõpptulemusele suurt mõju seni, kuni puid on vähemalt 100. Seda arvu veel suurendades jõuab lõpptäpsus tihti platoole ja enam ei parane. Samast artiklist lähtub ka, et kasutatavatel tunnuste arvul on klassifitseerimistäpsusele vaid keskpärane mõju. Rodriguez-Galiano jt (2012) on sõnastanud põhimõtte, et kui optimeerimine pole võimalik, tuleks kasutada palju otsustuspuid ja vähe tunnuseid. Liiga suure arvu tunnuste (rohkem kui ruutjuur kõigist tunnustest) kasutamisel täheldasid ka Deschamps jt (2012) olulist klassifitseerimistäpsuse langust.

Kasutatud treeningandmestiku klassid ehk põllukultuurid on selgelt tasakaalust väljas. Näiteks kõrrelised heintaimed moodustavad *ca* 40% kogu andmestikust, samas kui väiksemate klasside kohta on vaid mõned üksikud näidised (vt lisa II). Sellise andmestiku kasutamisel hakkab klassifikaator eelistama ülekaalus klasse ja eirama väikseid. Klasside tasakaalustamiseks on tavaliselt kaks moodust – arvukaid klasse kärpida või vähearvukatele kopeerimise abil näidiseid juurde luua (Maxwell jt, 2018). Selle töö käigus klasside tasakaalustamist läbi ei viidud, küll aga katsetati seda nii-öelda emaprojekti raames (Voormansik jt, 2020) ning vähearvukate klasside testkogu saagised olid seal valdavalt paremad kui selle töö juhumetsa mudelil. Samas tuleb rõhutada, et saadud mudelid pole üks ühele võrreldavad, kuna treeningandmestikud pole identsed.

Operatiivteenustes rakendataks varasemate aastate andmetel treenitud mudelit järgmise hooaja kultuuride tuvastamisel. Selle peamise kasutusjuhu läbimängimiseks lasti 2018. aasta andmetel treenitud mudelil anda hinnanguid ka 2019. aasta testkogul ja tulemused läksid kehvemaks (vt eksimismaatriksit lisas III). Klassideülene kaalutud F1-skoor oli 0,70 (sama aasta andmetel treenitul mudeli puhul oli see olnud 0,85) ja üle 90% saagiseid saavutasid mõned üksikud klassid (suviraps ja -rüps, kõrrelised heintaimed ning taliraps). See test kinnitab lihtsat tõdemust, et konkreetse aasta ilmastikutingimused mõjutavad taimede fenoloogiat ja kasvukäiku. Mudeli jaoks, mis suudaks ühtmoodi hästi ennustada erinevatel hooaegadel, tuleks selle treeningandmestikku kaasata näidiseid mitmest aastast.

Joonistel 4 ja 5 on näha, et väga selgelt teistest olulisuse poolest eristuvaid tunnuseid ei ole ning oma väikse panuse on andnud pea kõik satelliiditunnused. Viie kõige olulisemaks hinnatud tunnuse hulgast leiab nii Sentinel-1 (*s1_s0vh, s1_s0vv*) kui Sentinel-2 (*s2_psri, s2_ndvi, s2_tc_vegetation*) tunnuseid. Mitmed uurimistööd (Blaes jt, 2005; Veloso jt, 2017; Inglada jt, 2016) on tõestanud, et radarandmete kasutamine suurendab põllukultuuride eristamise edukust. Seda kinnitab ka antud töö, kus Sentinel-1 tagasihajumise parameetrid olid olulisimate tunnuste hulgas. Veel paremini illustreeriks seda katse, kus mudelit sobitataks eraldi Sentinel-1 ja Sentinel-2 tunnustega ja võrreldaks tulemusi, ent see ei mahtunud antud töö raamidesse.

Tähelepanuväärne on, et Sentinel-1 koherentsuse parameetrid, mis mängivad määravat rolli näiteks niitmissündmuste tuvastamisel (Tamm jt, 2016), on kultuuride tuvastamisel olulisuse pingerea lõpuosas. Näib, et järsud muutused taimestiku struktuuris ja hulgas, mida koherentsuse muutused hästi kajastavad, ei ole omased taimede järkjärgulisele arengule. Tamm jt (2016) on leidnud, et suurel osal aastast on enamikul rohelise ja vett sisaldava taimkattega põldudel koherentsus nullilähedane, vähe muutuv ja täiendavat infot mittelisav.

Sentinel-2 oluliste tunnuste hulgast ei ole üllatav leida NDVI indeksit, mis on üks levinumaid taimestikuindekseid ja rohelise biomassi hindamise standard. NDVI koos PSRI ja TC-VEGETATION indeksitega hõlmavad endas infot spektraalkanalitest B2, B3, B4, B6, B8, B11 ja B12 ehk katavad ära suuresti kogu Sentinel-2 spektriulatuse, mis võib ka olla põhjuseks, miks nii-öelda toorkanalid olulisuse pingereas tahapoole jäävad.

Uurides eelnimetatud viie tunnuse olulisusi ajas (joonis 6), joonistuvad selgelt välja optiliste ja radaripõhiste tunnuste erinevused. Optilised taimkatteindeksid on olulisemad kevadel (mais) ja suve lõpus (augustis). Juunis ja juulis nad klassifitseerimisel olulist rolli ei mängi. Erinevates polarisatsioonides radari tagasihajumise tunnused on aga märgatavalt kasulikumad just kesksuvel (juuni ja juuli), jäädes muul ajal optiliste tunnuste varju. Sesoonseid erinevusi tunnuste olulisuses on täheldatud mitmes uurimistöös. Sarnast mustrit on kirjeldanud Van Tricht jt (2018), kes klassifitseerisid kaheksat Belgia põllukultuuri Sentinel-1 ja -2 piltidelt. NDVI indeksil tuvastati kaks väga selgelt olulisemat perioodi kevadel, kui taliviljad arenevad kiirelt ja suviviljad tärkavad, ning vahetult enne koristusaega, kui taimede fenoloogilised erinevused on kõige suuremad. Sentinel-1 olulisus hakkas tõusma mais ja varieerus siis ilma selge mustrita. Sentinel-1 tunnuste vähest kevadist mõju on täheldanud ka Veloso jt (2017), kes viitas ka NDVI küllastumisele hooaja keskel. Üleüldse ei soovita nad kasutada NDVI-d näiteks suviteraviljade eristamiseks.

Kuna kasvuperioodi pikkused sõltuvad regiooni klimaatilistest tingimustest, ei ole mõistlik tunnuste olulisusi tõlgendada ainult kalendrikuude kontekstis, vaid lähtuvalt taime kasvufaasidest ja fenoloogiast. Varasemad autorid on leidnud, et Sentinel-1 parameetrid eristavad paremini taimede strukturaalseid muutusi, näiteks õitsemine ja viljapeade areng, aga ka seemnete areng ja küpsemine (McNairn jt, 2009b), mis võiksid olla eristuvateks tunnusteks ka siis, kui kesksuvel enamik kultuure ei erine värvuselt (on ühtlaselt rohekad) ja näivad optilisele sensorile sarnastena. Lisaks on radariparameetrid oluliselt tundlikumad taimede veesisaldusele.

Siit jõuame veel ühe olulise muutujani, mis võib põhjustada erinevat tüüpi tunnuste sesoonset muutlikkust – ilmastik. Eestis on mais reeglina rohkem pilvevabu päevi (Kullamaa, 2015) kui järgnevatel kuudel, mistõttu sel ajal saab optiline satelliit ka rohkem kasulikke pilte. Samas ei saa ilmastiku mõjuga seletada Sentinel-2 augusti parameetrite olulisust, sest suve lõpus on meie taevas tihti veel pilvisem kui juunis ja juulis. Täpsemate vastuste saamiseks oleks vaja uurida 2018. ja 2019. aasta pilvisuse andmeid – kui palju oli maist augustini igas kuus pilvedest rikkumata kasutatavaid Sentinel-2 andmepunkte.

Kasutatud ilmastiku, mullastiku ja põldude asukoha tunnused (joonis 5) osutusid pigem ebaolulisteks. Ilmaparameetrite puhul oleks ilmselt informatiivsem kasutada kumulatiivseid tunnuseid, näiteks sademe- ja temperatuurisummad hooaja algusest, mis iseloomustaksid konkreetset hooaega ja võiksid osutada kasulikuks andmestikes, kus on näidiseid mitmest erinevast aastast.

Loomulikult võib tekkida mõte, et kui palju kaotaks mudel oma sooritusvõimest, kui kasutada ainult kõige olulisemaid tunnuseid. Selleks prooviti samade parameetritega klassifitseerijat sobitada 2019. aasta andmestikul, kus olid vaid 5 olulisimaks hinnatud

tunnust (*s1_s0vh*, *s1_s0vv*, *s2_psri*, *s2_ndvi*, *s2_tc_vegetation)* varasema 27 asemel. Testkogul tehtud hinnangute klassideülene kaalutud keskmine F1-skoor oli 0,83 ehk veidi halvem kui kõigi tunnuste puhul (0,85). Kõigi tunnustega mudeli ja 5 valitud tunnusega mudeli hinnangute võrdlus on toodud lisas IV. Samas treeningandmestiku maht vähenes 2GB-lt 465 MB-ni, eeltöötlus kestis varasema 1h 20 minuti asemel 30 minutit ning mudeli sobitamine koos hinnangute andmisega võttis varasema 2 minuti asemele aega 1 minut ja 24 sekundit. Antud töö eesmärgiks ei olnud luua võimalikult efektiivset eeltöötluse koodi, ent mingi indikaatori see võrdlus siiski annab. Küsimus, kas minimaalne klassifitseerimistäpsuse langus ressursside kokkuhoiu nimel on aktsepteeritav, vajab iga konkreetse kasutusjuhu puhul eraldi analüüsimist.

7. Kokkuvõte

Töös seati eesmärgiks juhumetsa masinõppemudeli abil klassifitseerida Eestis kasvavad põllukultuurid 28-sse eelnevalt defineeritud klassi. Sisendandmetena kasutati RITA Kaugseire projekti andmestikku, mis koondab endas kahe hooaja (2018, 2019) kõigi Eesti põllumaade info, Sentinel-1, Sentinel-2 satelliidipiltidelt arvutatud tunnused ja ilmastiku andmed. Teostati ka spetsiifiline andmete eeltöötlus, et neid saaks kasutada konkreetse klassifikaatori sobitamisel *scikit-learn* masinõppe teegis.

2018. aasta testkogu klassideülene kaalutud F1-skoor oli 0,82 ja 2019. aastal 0,85. Tulemused on sarnased RITA Kaugseire projekti raames närvivõrkudel treenitud mudeliga (kaalutud F1-skoor 0,85), mis kasutas mõlema aasta andmeid koos. 2018. aasta andmetel treenitud mudel andis 2019. aasta testkogul halvemaid tulemusi (kaalutud F1-skoor 0,7) kui sama aasta andmetel treenitud mudel.

Tunnuste olulisuse analüüsis selgus, et kõige olulisemate tunnuste hulgas esineb nii Sentinel-1 kui Sentinel-2 tunnuseid, kuid ilmastiku, mullastiku ja põllu suhtelise asukoha info nii oluline pole. Kõige suurema panuse põllukultuuride eristamisse andsid radarsatelliidi erinevate polarisatsioonide tagasihajumise tunnused (*s1_s0vh, s1_s0vv)* ja optilise satelliidi taimestikuindeksid (*s2_psri, s2_ndvi, s2_tc_vegetation).*

Ilmnes ka selge sesoonne varieeruvus eri tüüpi tunnuste olulisuses. Sentinel-2 tunnused osutusid määravamateks hooaja alguses (mai) ja lõpus (august), samas kui hooaja keskel (juuni, juuli) vähenes nende olulisus märgatavalt. Sentinel-1 tagasihajumise tunnused olid aga olulisemad just kesksuvel.

Vaid viie olulisima tunnusega klassifitseerides, oli 2019. aasta testkogu F1-skoor 0,83, kuid eeltöötlusele kulunud aeg ja treeningandmete maht vähenes mitu korda.

Töö tõestas, et põllukultuuride tuvastamisel mängivad olulist rolli nii radar- kui optilise satelliidi andmed, kuid nende panus hooaja jooksul varieerub. Seetõttu on soovituslik kasutada praktilistes rakendustes mõlemaid andmeallikaid, sest nad täiendavad teineteist.

8. Viidatud kirjandus

Aunap, R., Frey, J., Hang, T., Jaagus, J., Järvet, A., Kalm, V., Kanal, A., Karro, E., Kirs, J., Kirsimäe, K., Mander, Ü., Mardiste, H., Oja, T., Peterson, U., Pärtel, M., Rubel, M., Sepp, M., Tuuling, I., Mander, Ü., … Tartu Ülikool (Toim). (2014). *Üldmaateadus: Õpik kõrgkoolidele*. Eesti Loodusfoto.

Bargiel, D. (2017). A new method for crop classification combining time series of radar images and crop phenology information. *Remote Sensing of Environment*, *198*, 369–383. https://doi.org/10.1016/j.rse.2017.06.022

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Blaes, X., Vanhalle, L., & Defourny, P. (2005). Efficiency of crop identification based on optical and SAR image time series. *Remote Sensing of Environment*, *96*(3), 352–365. https://doi.org/10.1016/j.rse.2005.03.010

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(1), 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sensing of Environment*, *210*, 35–47. https://doi.org/10.1016/j.rse.2018.02.045

Deschamps, B., McNairn, H., Shang, J., & Jiao, X. (2012). Towards operational radar-only crop type classification: Comparison of a traditional decision tree with a random forest classifier. *Canadian Journal of Remote Sensing*, *38*(1), 60–68. https://doi.org/10.5589/m12-012

DIONE konsortsium. (2019). *An Integrated EO-based Toolbox for Modernising Cap Area-Based Compliance Checks and Assessing The Respective Environmental Impact*. https://dione-project.eu/ (30.04.2021)

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, *120*, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Duro, D. C., Franklin, S. E., & Dube, M. G. (2012). Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests. *International Journal of Remote Sensing*, *33*(14), 4502–4526. https://doi.org/10.1080/01431161.2011.649864

Euroopa Komisjon. (2018). *Modernising the CAP: satellite data authorised to replace on-farm checks* [News]. https://ec.europa.eu/info/news/modernising-cap-satellite-data-authorised-replace-farm-checks-2018-may-25_en (30.04.2021)

Euroopa Komisjon. (2021). *Ühise põllumajanduspoliitika lühitutvustus*. https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/cap-glance_et (30.04.2021)

EUROSTAT. (2020). *Agriculture, forestry and fishery statistics—2020 edition*. European Union. https://ec.europa.eu/eurostat/en/web/products-statistical-books/-/KS-FK-20-001

FAO. (2020). *World Food and Agriculture—Statistical Yearbook 2020*. FAO. https://doi.org/10.4060/cb1329en

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *116*, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

Inglada, J., Vincent, A., Arias, M., & Marais-Sicre, C. (2016). Improved Early Crop Type Identification By Joint Use of High Temporal Resolution SAR And Optical Image Time Series. *Remote Sensing*, *8*(5), 362. https://doi.org/10.3390/rs8050362

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *2014 Science and Information Conference*, 372–378. https://doi.org/10.1109/SAI.2014.6918213

Kullamaa, M. (2015). *Maapinna ja vee optiliseks kaugseireks sobivate aastaaegade analüüs Eesti alal METEOSAT satelliitpilvisuse andmete põhjal* [Thesis, Tartu Ülikool]. https://dspace.ut.ee/handle/10062/47385

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 778–782. https://doi.org/10.1109/LGRS.2017.2681128

Maa-amet. (2021). *Mullastiku kaart*. https://geoportaal.maaamet.ee/est/Ruumiandmed/Mullastiku-kaart-p33.html (30.04.2021)

Majandus- ja Kommunikatsiooniministeerium. (2019). *Eesti riiklik tehisintellekti alane tegevuskava 2019-2021*. https://www.mkm.ee/sites/default/files/eesti_kratikava_juuli2019.pdf

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

McNairn, H., Champagne, C., Shang, J., Holmstrom, D., & Reichert, G. (2009a). Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*(5), 434–449. https://doi.org/10.1016/j.isprsjprs.2008.07.006

McNairn, H., Shang, J., Jiao, X., & Champagne, C. (2009b). The Contribution of ALOS PALSAR Multipolarization and Polarimetric Data to Crop Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(12), 3981–3992. https://doi.org/10.1109/TGRS.2009.2026052

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*(1), 213. https://doi.org/10.1186/1471-2105-10-213

NIVA. (2019). *New IACS Vision in Action (NIVA)*. https://www.niva4cap.eu/ (30.04.2021)

Noorma, A., Jakobson, L., Lang, M., Kutser, T., Oja, T., Uiboupin, R., Voormansik, K., Puust, R., Post, P., Sepp, K., & Liibusk, A. (2020). *Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel (projekti RITA1 KAUGSEIRE lõpparuanne)*. https://datadoi.ee/handle/33/310

Post, P., Toll, V., Rahu, J., & Voormansik, T. (2020). *Sademete täppiskaardistamine. Lõpparuanne*. https://datadoi.ee/handle/33/312

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Rosen, J. (2021). Shifting ground. *Science*, *371*(6532), 876–880. https://doi.org/10.1126/science.371.6532.876

Sen4CAP konsortsium. (2017). *Sen4CAP - Sentinels for Common Agriculture Policy*. http://esa-sen4cap.org/ (30.04.2021)

Zalite, K., Antropov, O., Praks, J., Voormansik, K., & Noorma, M. (2016). Monitoring of Agricultural Grasslands With Time Series of X-Band Repeat-Pass Interferometric SAR. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(8), 3687–3697. https://doi.org/10.1109/JSTARS.2015.2478120

Tamm, T., Zalite, K., Voormansik, K., & Talgre, L. (2016). Relating Sentinel-1 Interferometric Coherence to Mowing Events on Grasslands. *Remote Sensing*, *8*(10), 802. https://doi.org/10.3390/rs8100802

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., … Rostan, F. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24. https://doi.org/10.1016/j.rse.2011.05.028

United Nations, Department of Economic and Social Affairs, & Population Division. (2019). *World population prospects Highlights, 2019 revision Highlights, 2019 revision*.

Van Tricht, K., Gobin, A., Gilliams, S., & Piccard, I. (2018). Synergistic Use of Radar Sentinel-1 and Optical Sentinel-2 Imagery for Crop Mapping: A Case Study for Belgium. *Remote Sensing*, *10*(10), 1642. https://doi.org/10.3390/rs10101642

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., & Ceschia, E. (2017). Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. *Remote Sensing of Environment*, *199*, 415–426. https://doi.org/10.1016/j.rse.2017.07.015

Voormansik, K., Jagdhuber, T., Zalite, K., Noorma, M., & Hajnsek, I. (2016). Observations of Cutting Practices in Agricultural Grasslands Using Polarimetric SAR. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(4), 1382–1396. https://doi.org/10.1109/JSTARS.2015.2503773

Voormansik, K., Järveoja, M., Domnich, M., Sünter, I., Tamm, T., Lang, M., Sagris, V., Oja, T., & Sepp, K. (2020). *Põllumajandusmaade kasutuse seire*. https://doi.org/10.23673/re-259

Whitcraft, A. K., Vermote, E. F., Becker-Reshef, I., & Justice, C. O. (2015). Cloud cover throughout the agricultural growing season: Impacts on passive optical earth observations. *Remote Sensing of Environment*, *156*, 438–447. https://doi.org/10.1016/j.rse.2014.10.009

Lisad

I. Mudeli parameetrite leidmiseks sooritatud testid ja tulemused

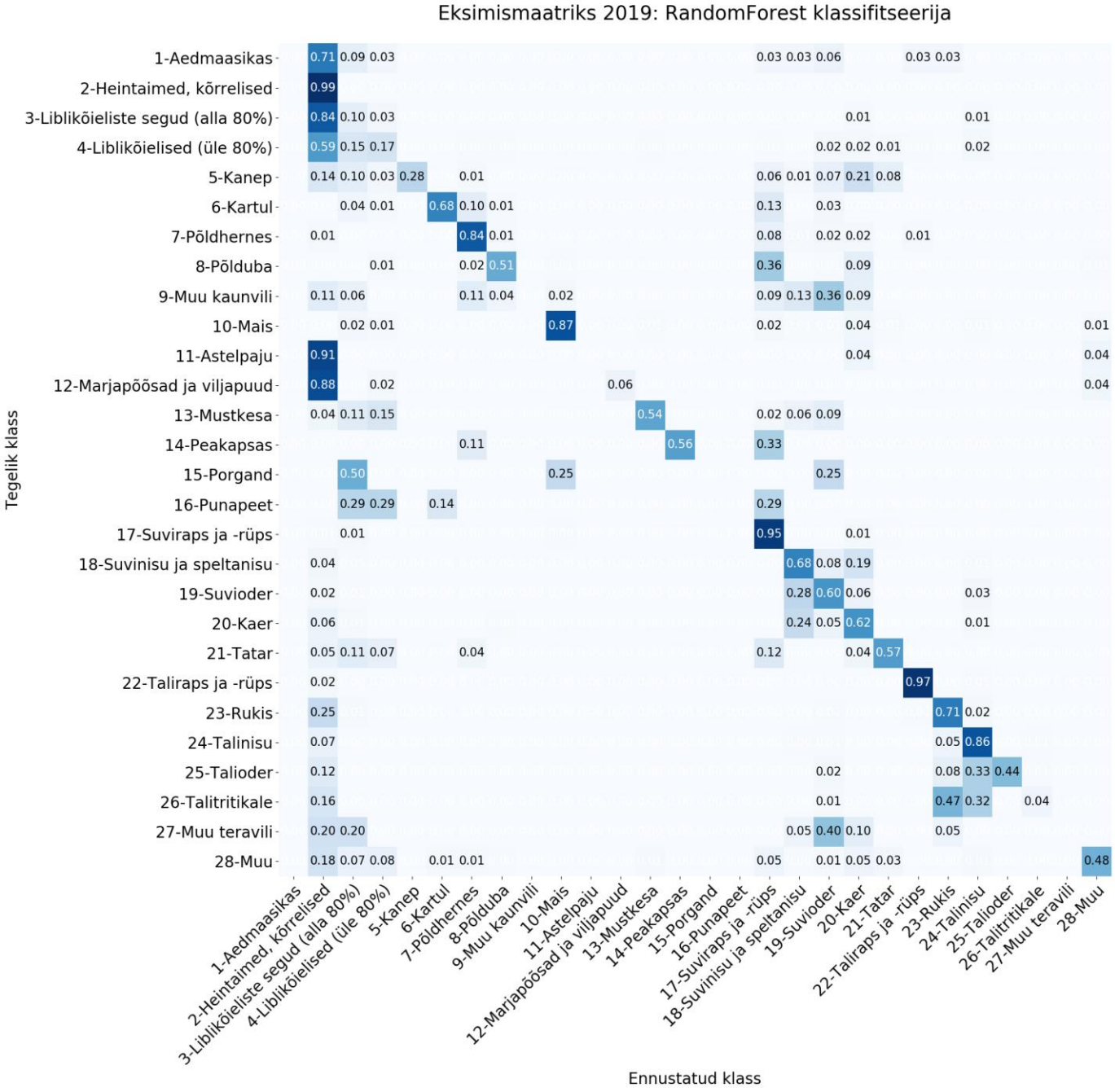
Parimaks osutunud test märgitud rohelise värviga.

	param_n_estimators	param_min_samples_split	param_max_leaf_nodes	param_max_features	param_criterion	param_bootstrap	split0_test_score	split1_test_score	split2_test_score	mean_test_score	std_test_score	rank_test_score	split0_train_score	split1_train_score	split2_train_score	mean_train_score	std_train_score
0	500	10	200	sqrt	gini	FALSE	0,7770	0,7793	0,7754	0,7773	0,0016	7	0,7967	0,7964	0,7963	0,7965	0,0002
1	300	10	200	sqrt	gini	TRUE	0,7791	0,7807	0,7775	0,7791	0,0013	6	0,7984	0,7978	0,7982	0,7981	0,0002
2	100	2	100	log2	gini	FALSE	0,7202	0,7184	0,7183	0,7190	0,0009	20	0,7281	0,7266	0,7284	0,7277	0,0008
3	300	2	200	log2	gini	FALSE	0,7443	0,7451	0,7439	0,7444	0,0005	17	0,7582	0,7595	0,7606	0,7594	0,0010
4	200	10	300	sqrt	gini	TRUE	0,7929	0,7938	0,7913	0,7927	0,0010	5	0,8219	0,8211	0,8214	0,8215	0,0003
5	300	2	NONE	log2	gini	FALSE	0,8382	0,8390	0,8338	0,8370	0,0023	1	1,0000	1,0000	1,0000	1,0000	0,0000
6	100	2	100	sqrt	gini	TRUE	0,7524	0,7533	0,7504	0,7520	0,0012	13	0,7606	0,7632	0,7619	0,7619	0,0011
7	300	10	300	sqrt	gini	TRUE	0,7930	0,7939	0,7913	0,7927	0,0011	4	0,8223	0,8210	0,8216	0,8217	0,0005
8	400	5	400	sqrt	gini	FALSE	0,8014	0,8030	0,7983	0,8009	0,0020	3	0,8381	0,8376	0,8368	0,8375	0,0005
9	500	2	200	log2	gini	TRUE	0,7461	0,7465	0,7456	0,7461	0,0004	15	0,7596	0,7599	0,7615	0,7604	0,0008
10	300	5	500	sqrt	gini	TRUE	0,8080	0,8085	0,8053	0,8073	0,0014	2	0,8539	0,8538	0,8535	0,8537	0,0002
11	500	5	500	log2	gini	FALSE	0,7746	0,7739	0,7716	0,7734	0,0013	9	0,8114	0,8104	0,8095	0,8104	0,0008
12	100	2	200	sqrt	gini	FALSE	0,7768	0,7799	0,7745	0,7771	0,0022	8	0,7969	0,7965	0,7955	0,7963	0,0006
13	500	5	400	log2	gini	FALSE	0,7676	0,7659	0,7653	0,7662	0,0010	11	0,7964	0,7958	0,7952	0,7958	0,0005
14	500	10	100	log2	gini	TRUE	0,7222	0,7197	0,7215	0,7211	0,0011	19	0,7294	0,7278	0,7291	0,7288	0,0007
15	200	5	200	log2	gini	TRUE	0,7455	0,7461	0,7445	0,7454	0,0007	16	0,7599	0,7605	0,7605	0,7603	0,0003
16	400	5	100	sqrt	gini	FALSE	0,7507	0,7510	0,7494	0,7504	0,0007	14	0,7597	0,7614	0,7608	0,7606	0,0007
17	100	2	200	log2	gini	FALSE	0,7438	0,7431	0,7445	0,7438	0,0006	18	0,7593	0,7589	0,7607	0,7596	0,0008
18	500	2	300	log2	gini	TRUE	0,7593	0,7590	0,7584	0,7589	0,0004	12	0,7815	0,7806	0,7808	0,7810	0,0004
19	200	10	500	log2	gini	FALSE	0,7755	0,7722	0,7717	0,7732	0,0017	10	0,8115	0,8098	0,8092	0,8102	0,0010

II. 2018. ja 2019. aasta testandmestiku klassifitseerimise tulemused juhumetsa mudeliga

Klass	Kultuur	2018				2019			
		täpsus	saagis	F1-skoor	põlde	täpsus	saagis	F1-skoor	põlde
1	Aedmaasikas	1	0,24	0,38	34	1	0,29	0,44	35
2	Heintaimed, kõrrelised	0,84	0,98	0,91	9619	0,85	0,98	0,91	9525
3	Liblikõieliste segud (alla 80%)	0,68	0,42	0,52	2291	0,68	0,47	0,56	2377
4	Liblikõielised (üle 80%)	0,73	0,32	0,45	894	0,74	0,33	0,46	748
5	Kanep	1	0,65	0,79	57	0,98	0,66	0,79	71
6	Kartul	0,9	0,73	0,81	82	0,84	0,88	0,86	77
7	Põldhernes	0,92	0,93	0,92	554	0,95	0,94	0,94	625
8	Põlduba	0,84	0,81	0,82	335	0,87	0,92	0,89	180
9	Muu kaunvili	1	0,1	0,18	52	0	0	0	47
10	Mais	0,96	0,87	0,92	124	0,99	0,93	0,96	164
11	Astelpaju	1	0,05	0,1	37	1	0,04	0,09	45
12	Marjapõõsad ja viljapuud	1	0,19	0,32	62	1	0,17	0,3	52
13	Mustkesa	0,69	0,75	0,72	67	0,75	0,74	0,75	54
14	Peakapsas	1	0,38	0,55	8	1	0,67	0,8	9
15	Porgand	0	0	0	8	0	0	0	4
16	Punapeet	1	0,2	0,33	5	1	0,14	0,25	7
17	Suviraps ja -rüps	0,93	0,96	0,94	704	0,99	0,96	0,97	401
18	Suvinisu ja speltanisu	0,82	0,76	0,79	1537	0,88	0,79	0,84	1112
19	Suvioder	0,82	0,93	0,87	2441	0,87	0,95	0,91	2003
20	Kaer	0,78	0,76	0,77	1071	0,86	0,88	0,87	1009
21	Tatar	0,92	0,57	0,71	82	0,98	0,73	0,84	56
22	Taliraps ja -rüps	0,99	0,97	0,98	407	0,99	0,98	0,99	770
23	Rukis	0,96	0,77	0,86	247	0,97	0,91	0,94	521
24	Talinisu	0,87	0,93	0,9	1200	0,95	0,97	0,96	1917
25	Talioder	0,93	0,42	0,58	130	0,98	0,83	0,9	310
26	Talitritikale	1	0,18	0,3	62	0,96	0,35	0,51	77
27	Muu teravili	0	0	0	15	1	0,2	0,33	20
28	Muu	0,94	0,66	0,78	212	0,91	0,59	0,72	219
Põldude arvuga kaalutud keskmised		**0,83**	**0,83**	**0,82**	**22337**	**0,86**	**0,86**	**0,85**	**22435**

III. 2018. aasta andmetel treenitud juhumetsa mudeli eksimismaatriks 2019. aasta testkogul



IV. 2019. aasta kõigi tunnustega ja valitud viie tunnusega mudeli hinnangute võrdlus testkogul

Klass	Kultuur	2019 kõik tunnused			2019 viis olulisemat tunnust			Muutus		
		täpsus	saagis	F1-skoor	täpsus	saagis	F1-skoor	täpsus	saagis	F1-skoor
1	Aedmaasikas	1	0,29	0,44	0,91	0,26	0,41	-0,09	-0,03	-0,03
2	Heintaimed, kõrrelised	0,85	0,98	0,91	0,84	0,98	0,9	-0,01	0	-0,01
3	Liblikõieliste segud (alla 80%)	0,68	0,47	0,56	0,7	0,43	0,53	0,02	-0,04	-0,03
4	Liblikõielised (üle 80%)	0,74	0,33	0,46	0,74	0,34	0,47	0	0,01	0,01
5	Kanep	0,98	0,66	0,79	0,94	0,53	0,68	-0,04	-0,13	-0,11
6	Kartul	0,84	0,88	0,86	0,79	0,85	0,82	-0,05	-0,03	-0,04
7	Põldhernes	0,95	0,94	0,94	0,94	0,92	0,93	-0,01	-0,02	-0,01
8	Põlduba	0,87	0,92	0,89	0,77	0,88	0,82	-0,1	-0,04	-0,07
9	Muu kaunvili	0	0	0	1	0,02	0,05	1	0,02	0,05
10	Mais	0,99	0,93	0,96	0,83	0,96	0,89	-0,16	0,03	-0,07
11	Astelpaju	1	0,04	0,09	1	0,05	0,1	0	0,01	0,01
12	Marjapõõsad ja viljapuud	1	0,17	0,3	0,56	0,11	0,18	-0,44	-0,06	-0,12
13	Mustkesa	0,75	0,74	0,75	0,78	0,89	0,83	0,03	0,15	0,08
14	Peakapsas	1	0,67	0,8	1	0,73	0,84	0	0,06	0,04
15	Porgand	0	0	0	0	0	0	0	0	0
16	Punapeet	1	0,14	0,25	1	0,5	0,67	0	0,36	0,42
17	Suviraps ja -rüps	0,99	0,96	0,97	0,97	0,95	0,96	-0,02	-0,01	-0,01
18	Suvinisu ja speltanisu	0,88	0,79	0,84	0,79	0,74	0,76	-0,09	-0,05	-0,08
19	Suvioder	0,87	0,95	0,91	0,87	0,93	0,89	0	-0,02	-0,02
20	Kaer	0,86	0,88	0,87	0,77	0,83	0,8	-0,09	-0,05	-0,07
21	Tatar	0,98	0,73	0,84	0,97	0,57	0,71	-0,01	-0,16	-0,13
22	Taliraps ja -rüps	0,99	0,98	0,99	0,99	0,98	0,98	0	0	-0,01
23	Rukis	0,97	0,91	0,94	0,94	0,92	0,93	-0,03	0,01	-0,01
24	Talinisu	0,95	0,97	0,96	0,94	0,97	0,96	-0,01	0	0
25	Talioder	0,98	0,83	0,9	0,98	0,84	0,91	0	0,01	0,01
26	Talitritikale	0,96	0,35	0,51	0,93	0,36	0,52	-0,03	0,01	0,01
27	Muu teravili	1	0,2	0,33	1	0,18	0,31	0	-0,02	-0,02
28	Muu	0,91	0,59	0,72	0,81	0,55	0,65	-0,1	-0,04	-0,07
Põldude arvuga kaalutud keskmised		**0,86**	**0,86**	**0,85**	**0,84**	**0,85**	**0,83**	-0,02	-0,01	-0,02

V. Lähtekood

Töö lähtekood on lisatud failis *RandomForest_CropClassification_repo.rar* ja kättesaadav ka siit: https://bitbucket.org/mihkeljarveoja/randomforest_cropclassification/src/master/.

VI. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Mihkel Järveoja, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

„Põllukultuuride tuvastamise masinõppe mudeli tunnuste olulisuse hindamine“,

mille juhendajad on Kaupo Voormansik ja Tambet Matiisen, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

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

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

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

Mihkel Järveoja

10.05.2021