

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Kätlin Eveli Epner, Diana Kaldlaur, Elina Sirk

ESG HINNANGUTE MÕJU ETF-IDE TOOTLUSELE

Statistilised meetodid masinõppes grupitöö projekt

Tallinn 2023

Sisukord

1	Uurimisprobleem	3
2	Andmed.....	5
3	Metoodika	14
3.1	Peakomponentide analüüs.....	14
3.2	Otsustuspuud.....	14
4	Tulemused ja järeldused	16
4.1	Tulemused.....	16
4.1.1	Peakomponentide analüüsi tulemused	16
4.1.2	Otsustuspuude tulemused.....	18
4.2	Järeldused.....	22
	Kasutatud allikad	24
	Lisa 1.....	25

1 Uurimisprobleem

Tänapäeval on üha kasvav huvi ettevõtete jätkusuutlikkuse vastu, seega võib ka paljude pikaajaliste investorite huvides olla oma raha paigutamine fondidesse, mis järgivad ka ESG põhimõtteid. Loomulikult on investeerimisotsuse tegemisel oluline samuti fondi tootlus. Mistõttu ESG hinnangute seos fondi tootlusega võiks olla huvipakkuv teema investoritele, fondijuhtidele, analüütikutele ning teistele kasusaajatele.

Varasemas kirjanduses Norra ettevõtete näitel on leitud, et kõrge või madala ESG hinnanguga fondidesse investeerimisel ei ole praegu rahalisi eeliseid ega puudusi [1]. Küll aga teine sarnane uurimus jõudis järeldusele, analüüsides Euroopa fonde eraldi, et kõrge ESG hindegaga fondide puhul on oluliselt kõrgem tootlus. Leiti, et kõrgema ESG skooriga kaasneb ka suurem risk. [2] Positiivse seose ESG tulemuslikkuse ja finantstulemuste vahel järeldasid ka [3]. Analüüsitud varasema kirjanduse põhjal saab eeldada, et kõrgemate ESG hinnanguga fondide ja tootluse vahel siiski leitakse positiivne korrelatsioon.

Erinevate ESG kategooriate seost fondi tootlusega on uuritud vähem. Kategooriate seost ettevõtete kasumlikkusega on uuritud ning leitud, et ESG individuaalsete ja kombineeritud tegurite ning ettevõtte kasumlikkuse ja ettevõtte väärtuse vahel puudub oluline seos [4]. Sarnasele tulemusele jõudis ka teine uurimus, et ei järeldatud mingit erinevust nende ettevõtete tulemuslikkuses, kellel on eraldi kõrge või madal keskkonna-, sotsiaal- või juhtimisalane hinnang [5].

Selle põhjal otsustati käesolevas töös keskenduda fondide ESG hinnangutele üldiselt ning kategooriate kaupa ja nende seostele fondi tootlusega, milleks moodustati järgmised uurimisküsimused:

1. Kas ESG hinnangud on seotud fondi tootlusega?
2. Kas ESG-kategooriate (keskkondlik, sotsiaalne vastutus, juhtimine) riskitasemete põhjal saab prognoosida fondide tootlust? Kui jah, siis kuidas need tootlust mõjutavad?

Seatud küsimustest esimene uurimisküsimus on avastuslik, et leida, kas ESG hinnangud on ikkagi seotud ETF-ide tootlusega. Teine uurimisküsimus on ETF-ide tootluse prognoosimine ESG-kategooriate põhjal.

2 Andmed

Käesolevas töös valitud andmestik koosneb 9495-st erinevast Euroopa ETF-ist (börsil kaubeldavad fondid). Fondide kohta on kogutud 132 erinevat näitajat, mille hulgas on kõik olulised tehnilised andmed, nagu ISIN kood, ETF-i nimi, fondi loomise kuupäev, kategooria, erinevate sektorite osakaal investeringutes ja palju muid muutujaid. Lisaks on välja toodud fundamentaalsed andmed, näiteks Morningstar hinnangud, iga-aastane tootlus alates 2015. aastast kuni 2020. aastani, haldustasu, ROE, P/B, ESG hinnangud ning paljud teised näitajad. Andmed on kogutud algselt 3 aastat tagasi Morningstar veebilehelt. Morningstar on USA finantsteenuste pakkuja. [6] Seega valitud andmestik sisaldab püstitatud uurimisküsimustele vastamiseks vajalikke andmeid suure hulga fondide kohta, võimaldades analüüsida ESG hinnangute mõju, erinevate ETF-ide tootlusi aja jooksul ja teha nende põhjal üldistatavaid järeldusi.

Eelnevalt püstitatud uurimisküsimustele vastuste leidmiseks valitakse andmestikust 9 mõõdikut (vt Tabel 1). Valituks osutusid vaid 3 ja 5 aasta tootlus, kuna ESG olulisus on esile tulnud vaid mõne viimase aasta jooksul.

Tabel 1. Valimisse kuuluvad muutujad

Muutuja	Seletus
3Y trailing return (<i>fund_trailing_return_3years</i>)	Numbriline muutuja, näitab fondi 3 aasta tootlust %
5Y trailing return (<i>fund_trailing_return_5years</i>)	Numbriline muutuja, näitab fondi 5 aasta tootlust %
Environmental score (<i>environmental_score</i>)	Numbriline muutuja skaalal 0–100, mis näitab fondi keskkonna riski hinnangut, kus 1 on madalaim ning 100 kõrgeim tulemus
Social score (<i>social_score</i>)	Numbriline muutuja skaalal 0–100, mis näitab sotsiaalset riski hinnangut, kus 0 on madalaim ning 100 kõrgeim tulemus

Governance score (<i>governance_score</i>)	Numbriline muutuja skaalal 0–100, mis näitab juhtimise riski hinnangut, kus 0 on madalaim ning 100 kõrgeim tulemus
Sustainability score (<i>sustainability_score</i>)	Numbriline muutuja skaalal 0–100, mis näitab jätkusuutlikkuse ehk ESG riski hinnangut, kus 0 on madalaim ning 100 kõrgeim tulemus
Sustainability rank (<i>sustainability_rank</i>)	Numbriline muutuja skaalal 1–5, mis näitab jätkusuutlikkuse ehk ESG järku, kus 1 on madalaim ning 5 kõrgeim tulemus
Category (<i>category</i>)	Tekstiline muutuja, näitab millist tüüpi fondiga on tegu, viidates fondi sektorile, tüübile, geograafilisele asukohale või strateegiale.
Fund size (<i>fund_size</i>)	Numbriline muutuja, näitab fondi varade väärtust

Morningstar-i fondide ESG hinne on abiks investoritele, et mõõta portfelli tasemel keskkonna-, sotsiaal- ja juhtimis- ehk ESG-teguritest tulenevat riski. Morningstar kasutab portfelli hindamiseks Sustainalyticsi hinnanguid, mis mõõdavad ettevõtte olulist ESG-riski. Seejärel arvutatakse ettevõtte tasandil hinded varade kaalutlemise alusel, et saada portfelli koondhinnang. Kõrge ESG riskiga fondid saavad hindeks 1, samas kui madala ESG riskiga fondidele antakse hindeks 5. [7]

ESG skoori hinnatakse skaalal 0–100, jaotades nelja kvartiili vastavalt jätkusuutlikkuse ja läbipaistvuse tasemele. Esimesse kvartiili kuuluvad ettevõtted, kellel skoor jääb alla 25 ning neid vaadeldakse kui ettevõtteid kehvaga ESG suutlikkuse ning ebapiisava läbipaistvusega ESG aruandlusel. Viimasesse ehk neljandasse kvartiili kuuluvad ettevõtted skooriga üle 75 ning neid nähakse suurepärase ESG suutlikkusega ja kõrge läbipaistvusega. [8]

Järgnevalt kirjeldava statistika abil antakse ülevaade andmestikust, et paremini mõista võimalikke esinevaid seoseid ja mustreid. Paremaks andmeanalüüsiks eelnevalt eemaldatakse andmestikust kõik puudulike väärtustega read, seega valimisse jääb 3027 ETF-i, millel olid jätkusuutlikkuse skoorid olemas.

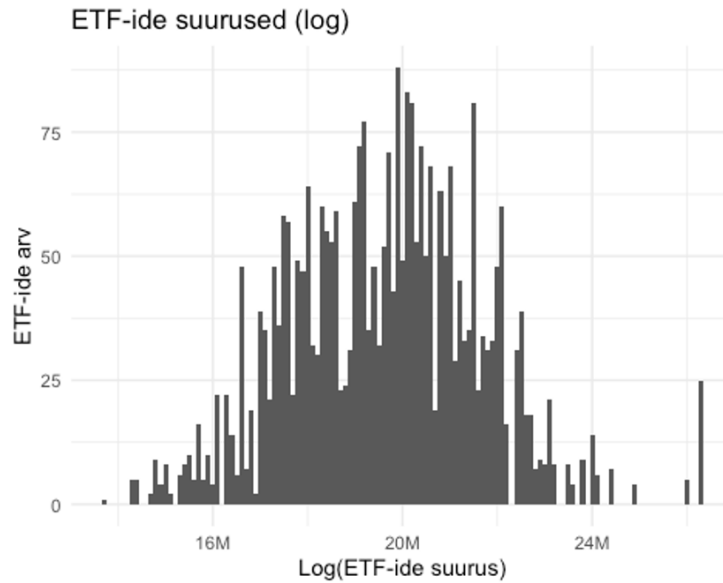
Järgnevast kirjeldava statistika tabelis (vt Tabel 2) on ülevaade valitud muutujate kohta. Tabelis pole välja toodud fondi kategooriat, kuna tegu ei ole numbrilise muutujaga. Tulemustest paistab,

et fondide suurused erinevad üsna suurel määral ning väikseim fond on vaid 890 000 ning suurim 272 miljardit. Kuigi ESG skoorid peaksid jagunema skaalal 0–100, siis käesolevas andmestikus on fondide maksimaalne ESG skoor vaid 62,7.

Tabel 2. Kirjeldav statistika valimis olevatest andmetest, n = 3027

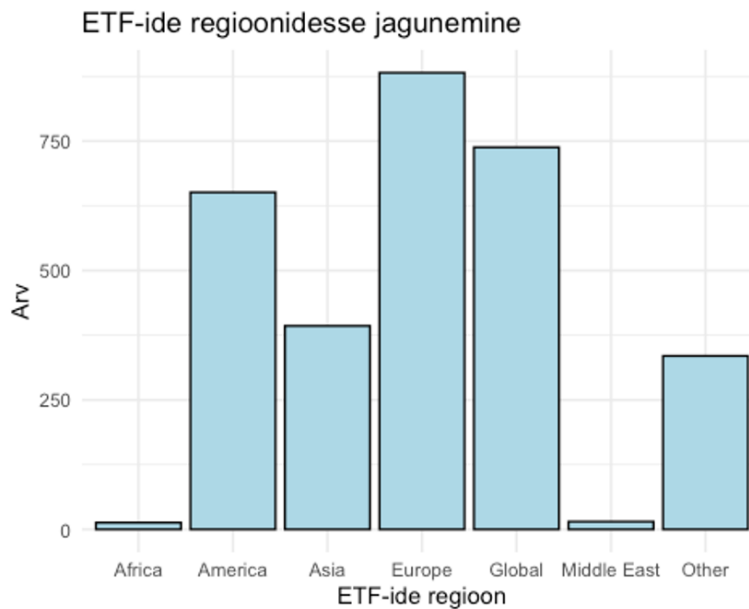
	Mean	Sd	Median	Mad	Min	Max	Range	Skew	Kurtosis	Se
3Y trailing return	3,59	7,93	3,36	5,0	-66,39	25,05	91,44	-2,33	16,93	0,14
5Y trailing return	10,23	7,37	10,08	4,69	-49,76	112,33	162,09	-0,04	37,79	0,13
Environmental score	5,52	3,67	4,81	1,45	0,07	62,7	62,63	8,39	117,12	0,07
Social score	9,81	2,95	9,92	1,32	2,98	63,68	60,70	11,92	217,6	0,05
Governance score	8,13	2,59	7,90	0,87	4,53	59,07	54,54	15,13	294,82	0,05
Sustainability score	23,9	4,10	23,28	2,30	10,43	55,91	45,48	2,03	11,65	0,07
Sustainability rank	3,05	1,00	3,00	1,48	1,00	5,00	4,00	0,07	-0,19	0,02
Fund size (in millions)	4114	25941	368	490	0,89	271530	271530	9,64	93,97	472

Kuna fondide suurused on väga laias vahemikus, siis on mõistlik nende visualiseerimiseks kasutada logaritmilist skaalat. Kui vaadata ETF-ide jagunemist suuruse järgi graafikult (vt Joonis 1), siis on näha, et fondi suurused on üsna normaaljaotuse kujuga, kuid mõned erandid on paremal pool. Logartimilise skaala puhul graafikul kujutatud x-teljel olev väärtus x on tegelikult 10^x fondi suurust. Keskmine fondi suurus on valimis 4,1 miljardit, väikseim fond suuruselt umbes 890 000 ning suurim fond on ligi 272 miljardit.



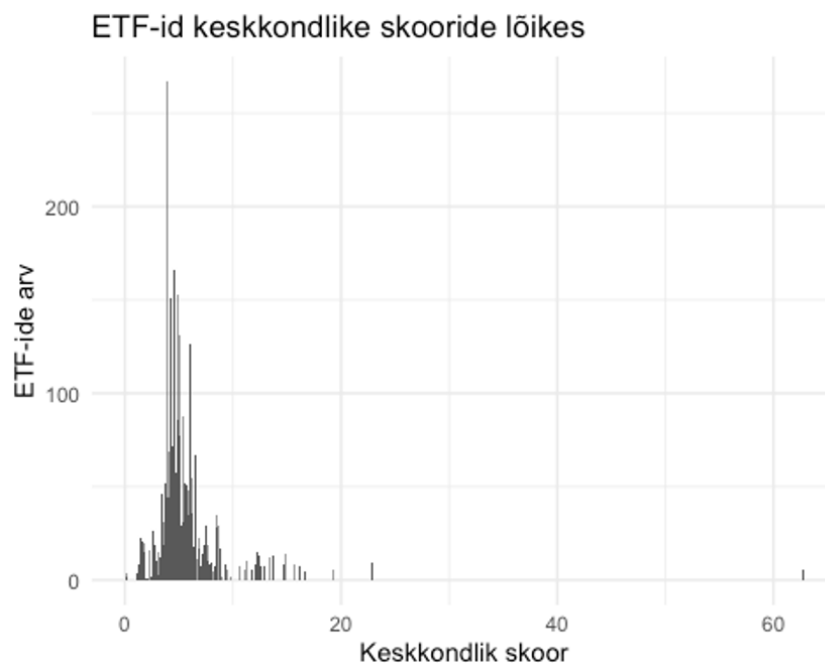
Joonis 1. Histogramm fondide suurustest logaritmilisel skaalal

Vaadates fondide jagunemist piirkonniti (vt Joonis 2), siis paistab, et põhiliselt on valimis olevate ETF-ide regiooniks Euroopa ning teisena globaalne fookus. Regiooni puhul mõeldakse fondi investeeringute fookust või strateegiat ehk suurem osa ETF-idest valimis on fookuseks valinud Euroopa või globaalse strateegia.



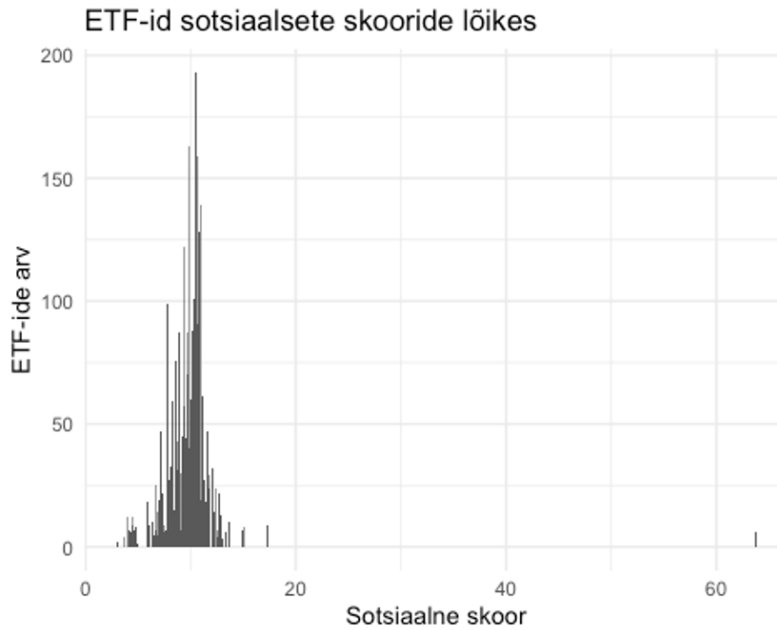
Joonis 2. Histogramm fondide jagunemisest regiooniti

Kuna käesolev uurimus keskendub ESG hinnangutele ja ETF-ide tootlusele, siis tuleb andmestiku paremaks mõistmiseks vaadata lähemalt, kuidas jagunevad erinevad skoorid fondide lõikes. Esmalt on võimalik näha (vt Joonis 3), et enamus fondidel on keskkonna skoor üsna madal, olles vaid 0 ja 10 vahel, ning mõndadel üksikutel fondidel on see väärtus üle 20. Samuti paistab, et tegu on üsna püstaka graafikuga väikese kaldega paremale, mis tähendab, et enamus tulemusi on koondunud aritmeetilise keskmise ümber, kuid mõned erandid siiski on. See graafiline tulemus kinnitab ka eelnevalt Tabelis 1 välja toodud aritmeetilist keskmist 5,52.



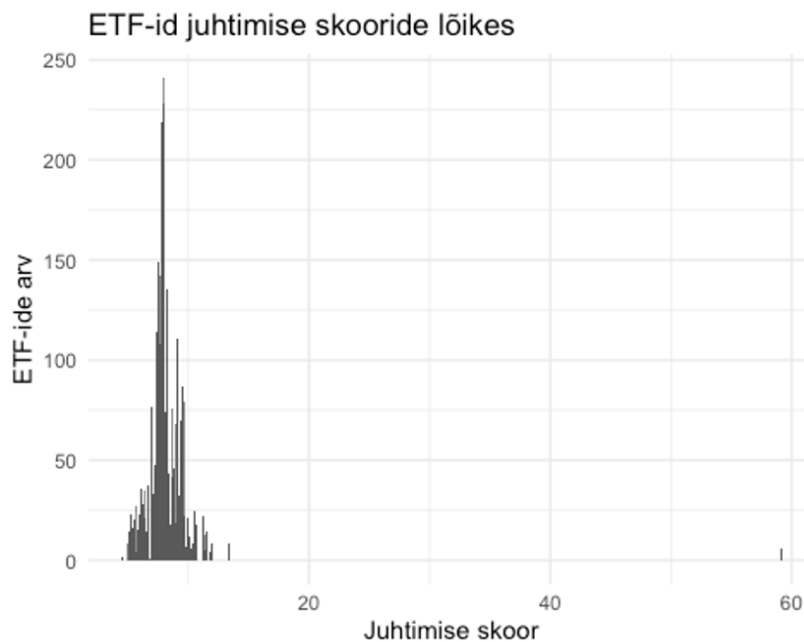
Joonis 3. Histogramm fondide keskkonna skooridest

Sarnaselt keskkonna skooridele jagunevad ka sotsiaalsed skoorid (vt Joonis 4) fondide vahel ümber aritmeetilise keskmise ning antud juhul on tegu veelgi püstakama graafikuga. Sotsiaalsete skooride aritmeetiline keskmine on aga veidi kõrgem võrreldes keskkondlike skooride väärtusega 9,81 ning ka sel juhul on mõned üksikud ETF-id üle 60.



Joonis 4. Histogramm fondide sotsiaalsetest skooridest

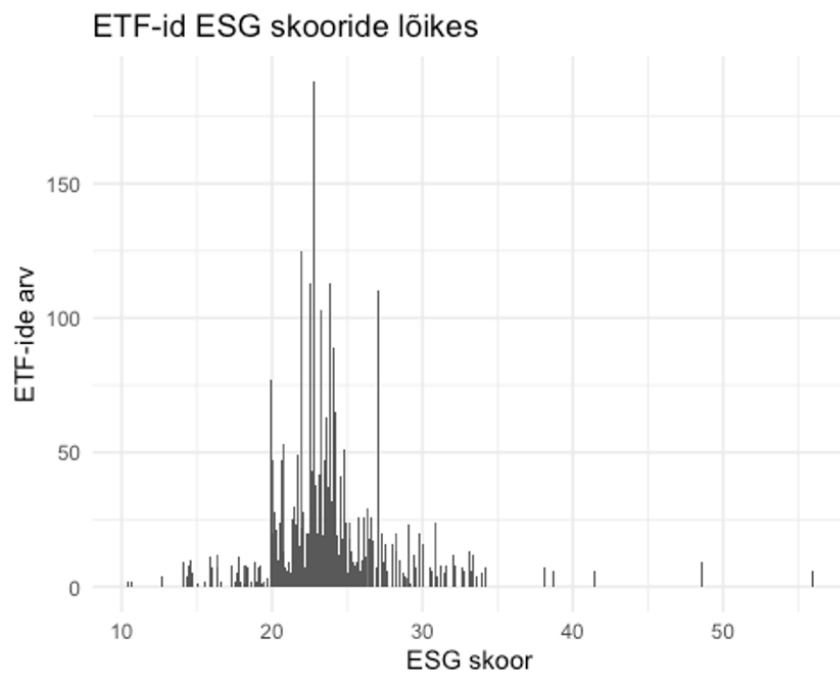
Kahele eelnevale sarnaselt jagunevad ka juhtimise skoorid fondide vahel väga püstakalt (vt Joonis 5). Juhtimise skooride aritmeetiline keskmine on 8,13 ning maksimum veidi alla 60.



Joonis 5. Histogramm fondide juhtimise skooridest

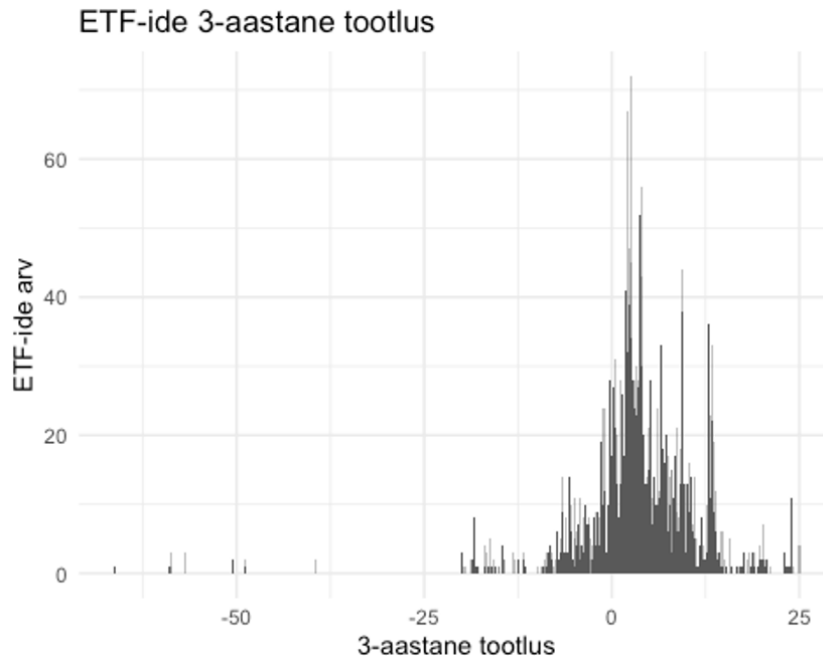
Eelneva kolme skoori põhjal on võimalik moodustada terviklik ESG hinnang (vt Joonis 6) ehk jätkusuutlikkuse skoor ja järk. Antud juhul ei ole enam tegu niivõrd püstaka graafikuga ning

tulemused on rohkem hajutatud. Jätkusuutlikkuse aritmeetiline keskmine on 2,37 ning ka maksimum tulemus jääb antud juhul vaid 50 lähedale.



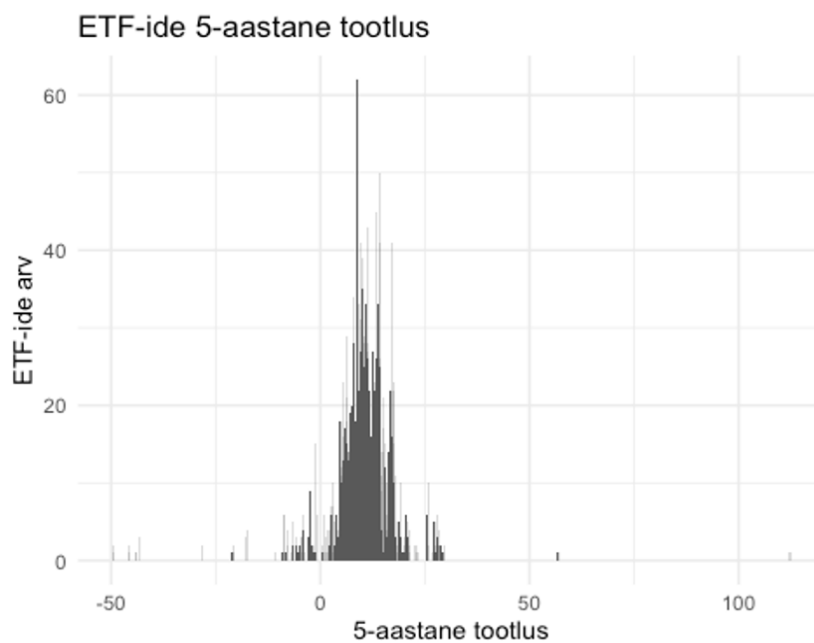
Joonis 6. Histogramm fondide ESG skooridest

Viimaks on sobilik vaadata ka fondide tootluseid (vt Joonis 7), mis andmestikus on kogutud 3 ja 5 aasta vältel. Kõigepealt vaadates fondi 3-aastase tootluse histogrammi, mille puhul on tegu üsna püstaka graafikuga ning tootlused on veidi paremale poole kaldu ehk tootlused on enamasti positiivsed.



Joonis 7. Histogramm fondide 3 aasta tootlusest

Järgnevalt võime näha 5 aasta tootluse puhul (vt Joonis 8), et tootlused on eelnevast veidi rohkem paremale poole kaldu ning pole enam nii püstakat graafikut. Suurem osa tootluseid jääb graafikul nullist paremale, viidates rohkemate fondide positiivsetele tootlustele.



Joonis 8. Histogramm fondide 5 aasta tootlusest

Kõrgeima mediaan tootlusega on 5-aastane tootlus, mille väärtus on veidi üle 10% ning madalaim mediaan on 3 aasta tootlusel väärtusega 3,4% (vt Tabel 1). Kõige suurem tootluste ulatus on samuti 3-aastaste tootluste hulgas, mille puhul on palju negatiivse tootlusega fonde, kuid on ka palju väga kõrge tootlusega ETF-e. Kuid kogu valimi kõrgeim tootlus on 5 aasta tootluse hulgas 112%.

3 Metoodika

Käesolevas peatükis antakse ülevaade uurimisküsimustele vastuste leidmiseks kasutatavatest meetoditest. ESG hinnangute ja ETF-ide tootluse seose uurimiseks kasutatakse peakomponentide analüüsi. Teise küsimuse puhul võetakse kasutusele otsustuspuud, üritades näha, kas ESG hinnangute põhjal saaks fondide tootlust prognoosida ning kuidas hinnangud võivad mõjutada fondide tootlust.

3.1 Peakomponentide analüüs

Peakomponentide analüüsi (PCA) eesmärgiks oli uurida, milline on seos ESG hinnangute ja tootluse vahel. Peakomponentide analüüsi puhul on tegu juhendamata õppe metoodikaga ning see aitab iseloomustada ka andmete sarnasusi ja erinevusi. Graafikule kujutatud noolte pikkuste põhjal saame teada, kui tugevat mõju antud teema peakomponendile avaldab. Noolte suund annab mõista komponentide vahelisest korrelatsioonist. See tähendab, et kui suund on ühtne, siis on tegu positiivse korrelatsiooniga.

Peakomponentide analüüsi kohta luuakse kaks graafikut: ühe 3-aastase ja teise 5-aastase fondi tootluse kohta. Analüüsis ei võeta arvesse kõiki komponente vaid leitakse kõige suurema mõjuga komponendid ning kasutatakse neid PCA analüüsi läbiviimiseks. Andmete analüüsimiseks saab kasutada R programmeerimiskeelt. R keele teegid tidyverse ja factoextra aitavad andmeid korrastada, viia läbi ja visualiseerida PCA analüüsi. Lisaks saab ggplot2 abil luua andmetest selgemad visuaalid, et saada parem ülevaade tulemustest.

3.2 Otsustuspuud

Otsustuspuu meetodit kasutatakse teise uurimisküsimuse vastuse leidmiseks, mille eesmärk on prognoosida fondide tootlust ESG-kategooriate näitajate põhjal. Otsustuspuu meetod aitab selgitada, millised ESG kategooriad võivad mõjutada fondide tootlust. Käesolevas töös

kasutatakse prognoosimiseks regressiooni puud ja juhusliku metsa meetodit, mis on vastavalt juhendamata ja juhendatud masinõppe meetod.

Esialgselt koostatakse treening andmestik, mis kasutab 60% valimi andmetest. Regressioonipuu sõltuvateks muutujateks olid valitud järgnevad muutujad: *fund_trailing_return_3years*, *environmental_score*, *social_score*, *governance_score*, *fund_size*, *sustainability_score* ja *sustainability_rank*. Treening andmete põhjal koostatakse esialgne otsustuspuu. Puu visualiseerimiseks kasutati graafikut, kus igal harul on näidatud tingimus, mille alusel andmeid jaotati. Seejärel arvutatakse keskmine ruutviga (MSE), mis näitab kui palju keskmiselt ennustatud väärtused erinevad tegelikest väärtustest ning koostatakse tegelike väärtuste ja ennustatud väärtuste graafik. Vastavas metoodikas kasutati kahte R keele teeki: *tree* ja *tidyverse*.

Selleks, et uurida, millised muutujad on otsustuspuu mudeli jaoks olulised, rakendatakse täiendavalt juhuslik metsa meetodit (*ingl. Random Forest*). Antud meetod hindab muutujate tähtsust fondi tootluse ennustamisel. Tulemused kuvatakse graafiliselt, et näidata, millised ESG hinnangud mõjutavad enim fondi tootlust. Antud metoodikas kasutati liskas R keeleteeki *randomForest*.

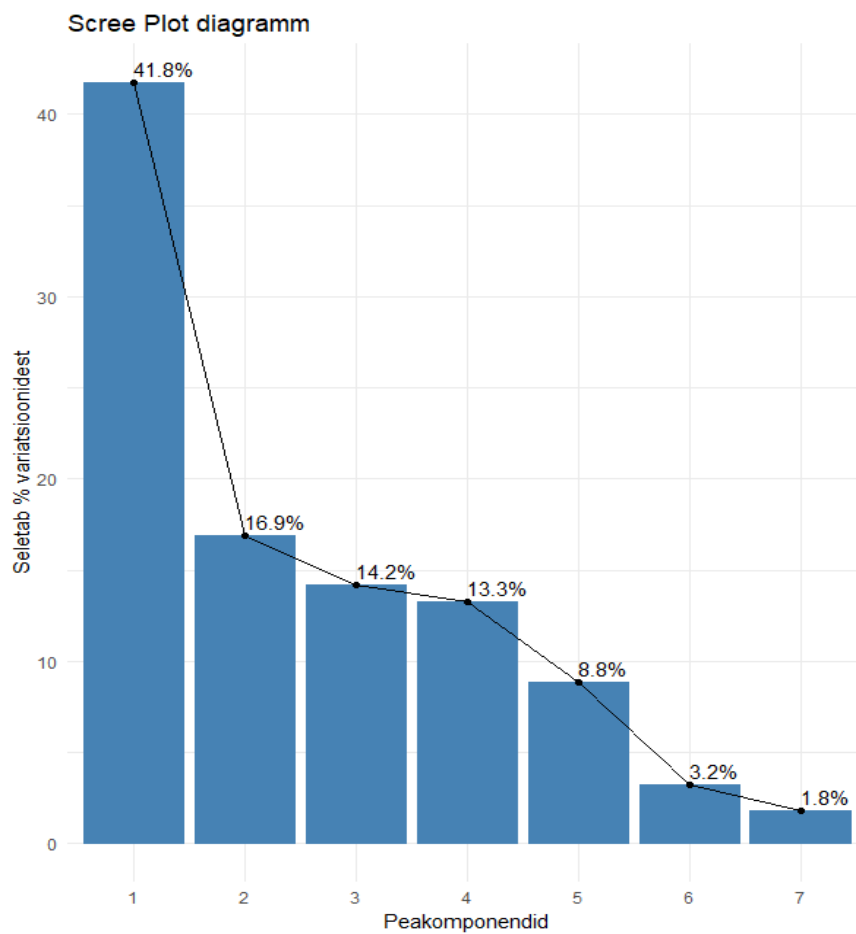
4 Tulemused ja järeldused

Käesolevas peatükis antakse ülevaade kasutatud metoodika tulemustest ning tehakse analüüsisist järeldused vastavalt uurimisküsimustele.

4.1 Tulemused

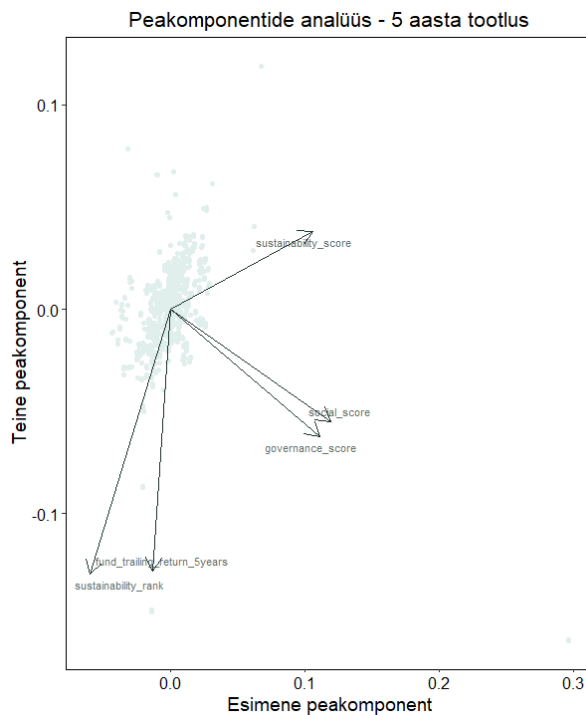
4.1.1 Peakomponentide analüüsi tulemused

PCA tulemuste põhjal sai loodud *Scree Plot* graafik (vt Joonis 9) kõige suurema mõjuga peakomponentide leidmiseks.

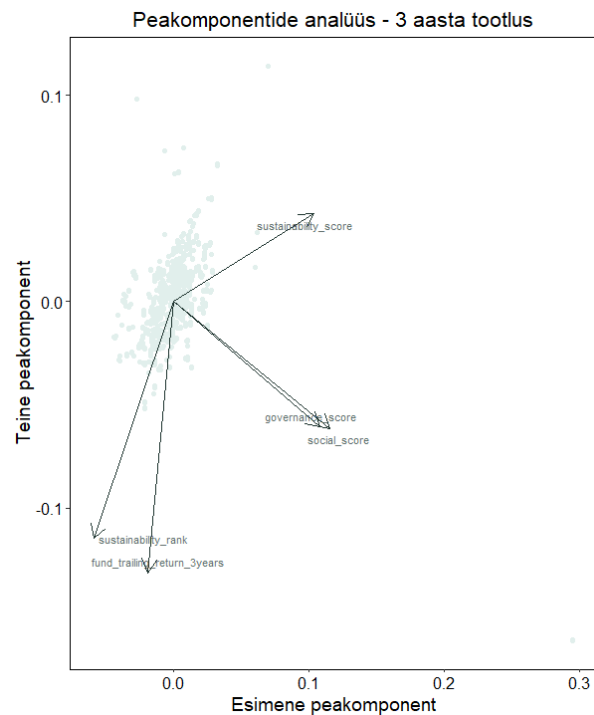


Joonis 9. Scree Plot diagramm 3-aastase tootluse korral

Selle põhjal otsustati analüüsida viis peakomponenti, millel oli kõige suurem mõju. ETF-i tootluse kõrval olid kõige suurema mõjuga jätkusuutlikkuse hinnang, valitsemise skoor sotsiaalne skoor ja üldine jätkusuutlikkuse skoor. Nende viie näitaja kohta loodi täpsemad graafikud, kus on välja toodud ka nende laadungvektorid.



Joonis 10. 5 aasta tootluse peakomponentide analüüs

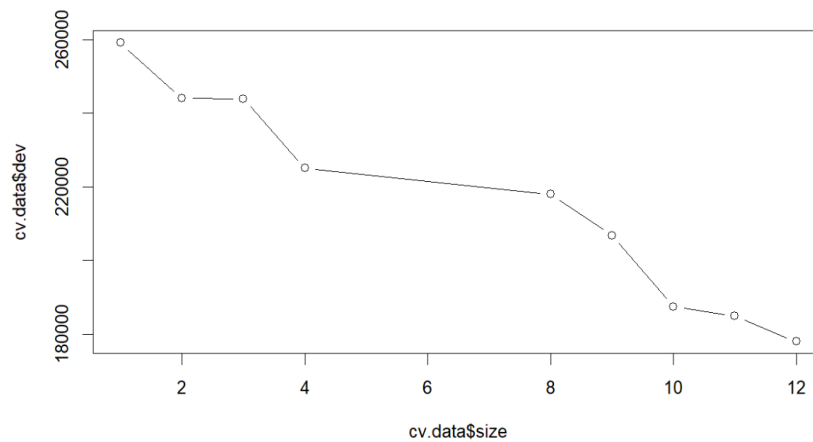


Joonis 11. 3 aasta tootluse peakomponentide analüüs

Graafikutelt (vt Joonis 10 ja Joonis 11) võib välja lugeda, et eriti tugevalt on tootlusega seotud ETF-i jätkusuutlikkuse hinnang. Vähest positiivset mõju avaldavad valitsemise skoor ja sotsiaalne skoor. Fondi tootlikkusele negatiivse mõjuga on jätkusuutlikkuse skoor. Jätkusuutlikkuse skoor väljendab riski fondi tootlikkusele, ehk siit järeldub, et suur jätkusuutlikkuse riski protsent mõjub negatiivselt fondi tootlikkusele.

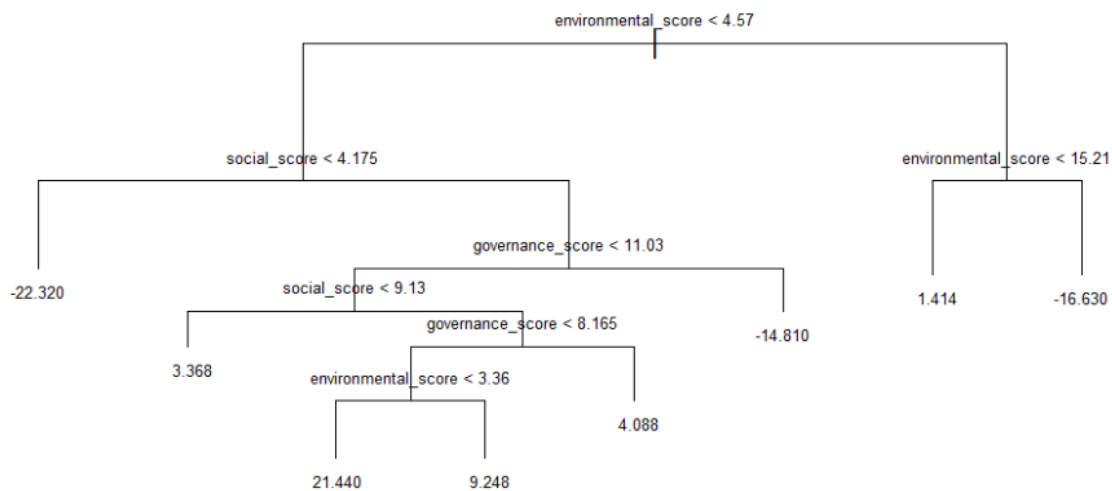
4.1.2 Otsustuspuude tulemused

Regressioonipuu, mis oli koostatud treening andmete põhjal, koosnes 12 lehest ja 10 sisesõlmest. Otsustuspuus esines kokku vaid 3 muutujat: keskkonna riski hinnang, sotsiaalse riski hinnang ja juhtimise riski hinnang.



Joonis 12. Otsustuspuu ristvalideerimise tulemused

Joonisel 12 on näha ristvalideerimise tulemusi, mis näitab puu suuruse ja mudeli deviatsiooni (hälbe) vahelist suhet. Selles olukorras valiti ristvalideerimise põhjal kõige keerukam puu. Vastav puu annab paremaid tulemusi treening andmetel, kuid võib olla üle sobitatud. Paremaks tõlgendamiseks otsustuspuu pügatakse. Pügatud otsustuspuu on välja toodud joonisel 13. Pügatud puul on 8 lehte ja 6 sisesõlme.

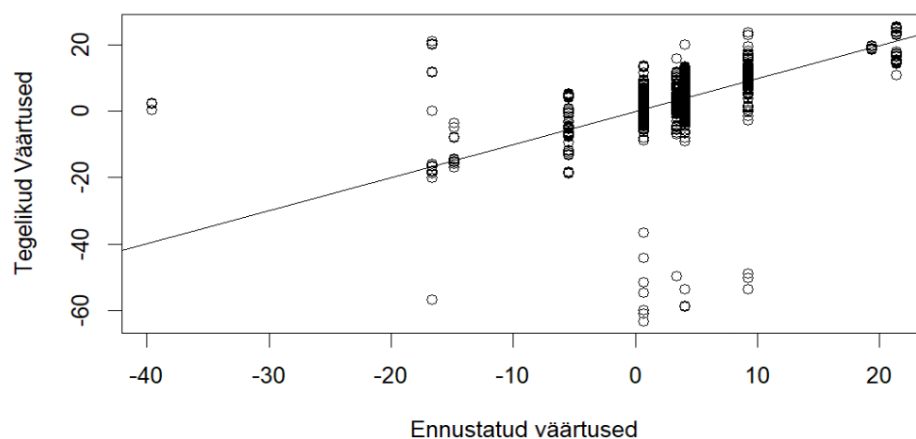


Joonis 13. Pügatud otsustuspuu

Pügatud otsustuspuu tulemustest saame välja lugeda, millistel tingimustel fondi tootlikkus suureneb või väheneb. Fondi tootlikkust mõjutab kõige enam keskkonna riski hinnang, mis asetseb esimeses sisesõlmes. Järgnevas sisesõlmes hargneb puu sotsiaalse riski hinnang ja keskkonna riski

hinnangu alusel. Kui fondi keskkonna riski hinnang on suurem või võrdne 15,21, siis on fondi tootlikuseks keskmiselt -16,63 %.

Vaadates otsustuspuu vasakut haru, on näha, et fondi tootlus on keskmiselt kõige madalam, kui fondil on keskkonna riski hinnang väiksem kui 4,57% ja sotsiaalse riski hinnang on väiksem kui 4,175%. Keskmise fondi tootlikkus (21,44%) on kõige suurem otsustuspuu lehel, kus keskkonna riski hinnang on alla 3,36%, juhtimise riski hinnang on alla 8,165% ja sotsiaalne riski hinnang on kõrgem kui 9,93%.



Joonis 14. Tegelike väärtuste ja ennustatud väärtuste võrdluse graafik

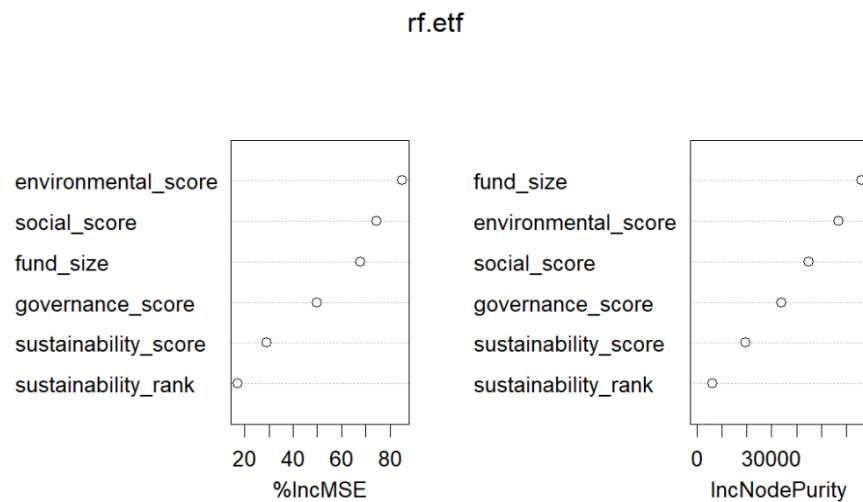
Keskmise ruutvea arvutamiseks kasutatakse pügamata otsustuspuid, et vältida suuremat nihet. Samal ajal võetakse arvesse, et pügamata otsustuspuu võib olla ülesobitatud. Joonisel 14 on väljatoodud tegelike väärtuste ja ennustatud väärtuste võrldemise graafik. Keskmiseks ruutveaks (*ingl. MSE*) saadi 67,76, mis näitab kui palju keskmiselt ennustatud väärtused erinevad tegelikest väärtustest. Ruutjuur 67,76-st on ligikaudu 8,23. Seega võib öelda, et mudel ennustab testandmestiku põhjal keskmiselt umbes 8% veaga fondi tootlikkuse keskmisest.

Käesolevas töös teostati ka juhusliku metsa mudel. Mudel luuakse testandmestiku põhjal, mida kasutati regressiooni puu tegemisel. Juhusliku metsa mudeli põhjal saadi keskmise ruutvea väärtuseks ligikaudu 26,44, mis on oluliselt madalam võrreldes regressiooni puu mudeli MSE väärtusest.

Tabel 3. Muutujate tähtsus mudeli ennustuste tegemisel

	%IncMSE	IncNodePurity
environmental_score	85.120	57033.67
social_score	74.45	44951.96
governance_score	49.90	33937.36
fund_size	67.73	66008.54
sustainability_score	29.04	19587.88
sustainability_rank	17.04	6411.07

Lisaks teostati juhusliku metsa mudeli muutujate tähtsuse analüüs, mida teostati R keele importance funktsiooni abil. Testi tulemused on välja toodud tabelis 3 ja joonisel 15. %IncMSE (*Mean Squared Error Increase*) näitab, kui palju iga muutuja suurendab mudeli ennustuste täpsust ning IncNodePurity näitab muutuja panust sõlme puhtuse (*Node Purity Increase*) suurendamisel.



Joonis 15. Juhusliku metsa mudel valideerimisandmete põhjal

Mida suurem on väärtus mõlemas kategoorias, seda olulisem on muutuja mudeli jaoks. Seega juhuslike metsade arvesse võetud puude puhul on kõige suurema mõjuga anntud mudelis järgnevad muutujad: keskkonna riski hinnang, sotsiaalse riski hinnang ja fondi suurus.

4.2 Järeldused

Käesoleva projekti eesmärgiks oli uurida ESG hinnangute ja Euroopa ETF-ide tootluse vahelist seost. Selleks püstitati kaks uurimisküsimust:

1. Kas ESG hinnangud on seotud fondi tootlusega?
2. Kas ESG-kategooriate (keskkond, sotsiaalne vastutus, juhtimine) riskitasemete põhjal saab prognoosida fondide tootlust? Kui jah, siis kuidas need tootlust mõjutavad?

Analüüsist selgus, et fondide tootlusele on tugev mõju jätkusuutlikkuse hinnangul (*sustainability_rank*). Lisaks on vähenenud positiivne mõju ka sotsiaalsel skooril. Kui võtta arvesse ka väiksema mõjuga ESG-kategooriaid, siis selgub, et näiteks keskkonnategur ei mõjuta fondide tootlust kummaski suunas. Ka varasem kirjandus toob välja, et sotsiaalsel ja juhtimise skooril on positiivne mõju fondi tootlusele, keskkonnategur aga ei mõjuta tootlust [2]. Uurimisküsimuse vastusena võib välja tuua, et kuigi jätkusuutlikkusel on selgelt positiivne mõju fondi tootlusele, siis teiste ESG-kategooriate mõju tootlusele on väike ning seosed liiga nõrgad lõplike järelduste tegemiseks.

ESG-kategooriate riskiandmete põhjal koostati otsustuspuu mudel, mille keskmiseks ruutveaks osutus 67,76. Seega ESG-kategooriate põhjal ennustab mudel keskmiselt 8% veaga fondi tootlikkuse keskmisest. Samuti selgus, et fondi keskmine fondi tootlikkus on kõige madalam kui keskkonna riski hinnang on suurem või võrdne 15,21%-ga või keskkonna ja sotsiaalse riski hinnang on madalamad kui 4,57% ja 4,175%. Sarnaselt PCA analüüsiga ilmnes, et kõrgema sotsiaalse riski hinnanguga on kõrgem ETF fondi tootlus. Juhusliku metsa mudeli põhjal selgus, et prognoosimisel on suurema mõjuga keskkonna riski hinnang, sotsiaalse riski hinnang ja fondi suurus.

Käesoleva analüüsi piirangutena saab välja tuua andmete puhul, et suuremal osal andmestikus olevatest ETF-idest oli vajalikud andmed puudu, seega ei saa hästi üldistada analüüsitulemusi kõikidele Euroopa ETF-idele. Rakendatud meetoditest on PCA mõeldud ennekõike andmete eeltöötluks ning täpsema seose uurimiseks oleks mõistlik lisaks kasutada ka mõnda juhendatud õppe meetodit. Otsustuspuude rakendamisel on ohuks, et mudelit üle kerge ülesobitada.

Kasutatud allikad

- [1] J. Taghawi Moussawi, ‘Sustainable Investing : is there a relationship between ESG ratings and fund performance?’, Master thesis, Norwegian University of Life Sciences, Ås, 2019. Accessed: Dec. 26, 2023. [Online]. Available: <https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/2608324>
- [2] M. Steen, J. T. Moussawi, and O. Gjølberg, ‘Is there a relationship between Morningstar’s ESG ratings and mutual fund performance?’, *Journal of Sustainable Finance & Investment*, vol. 10, no. 4, pp. 349–370, Oct. 2020, doi: 10.1080/20430795.2019.1700065.
- [3] G. Abate, I. Basile, and P. Ferrari, ‘The level of sustainability and mutual fund performance in Europe: An empirical analysis using ESG ratings’, *Corporate Social Responsibility and Environmental Management*, vol. 28, no. 5, pp. 1446–1455, 2021, doi: 10.1002/csr.2175.
- [4] R. Atan, Md. M. Alam, J. Said, and M. Zamri, ‘The Impacts of Environmental, Social, and Governance Factors on Firm Performance: Panel Study on Malaysian Companies’, *Management of Environmental Quality An International Journal*, vol. 29, pp. 182–194, Jan. 2018, doi: 10.1108/MEQ-03-2017-0033.
- [5] ‘The independent effects of environmental, social and governance initiatives on the performance of UK firms - Jacquelyn E Humphrey, Darren D Lee, Yaokan Shen, 2012’. Accessed: Dec. 26, 2023. [Online]. Available: <https://journals.sagepub.com/doi/10.1177/0312896211410081>
- [6] ‘European Funds dataset from Morningstar’, Kaggle. Accessed: Dec. 20, 2023. [Online]. Available: <https://www.kaggle.com/datasets/stefanoleone992/european-funds-dataset-from-morningstar?select=Morningstar++European+Mutual+Funds.csv>
- [7] ‘The Morningstar Sustainability Rating, Explained | Morningstar’, Morningstar. Accessed: Dec. 22, 2023. [Online]. Available: <https://www.morningstar.com/sustainable-investing/morningstar-sustainability-rating-explained>
- [8] ‘ESG Scores’, LSEG Data & Analytics. Accessed: Dec. 22, 2023. [Online]. Available: <https://www.lseg.com/en/data-analytics/sustainable-finance/esg-scores>

Lisa 1

```
library(tidyverse)

# Loeme sisse andmed

data <- read.csv("Morningstar - European ETFs.csv")

# Valime nende seast välja uuritavad mõõdikud

data2 <- data %>%

select(fund_trailing_return_3years,      fund_trailing_return_5years,      environmental_score,
social_score, governance_score, sustainability_score, sustainability_rank, category, fund_size)

# Eemaldame puuduvad väärtused

data2 <- na.omit(data2)

# Fondide regioonide jaotus

data2$fund_region <- ifelse(grepl("(US|America|Canada|Brazil)", data2$category, ignore.case =
TRUE),                                     "America",
ifelse(grepl("(Eur|EUR|Nordic|GBP|UK|Norway|France|SEK|Sweden|CHF|Denmark|Switzerland
|NOK|Spain|Italy)", data2$category, ignore.case = TRUE), "Europe",

ifelse(grepl("(Asia|Thailand|China|Singapore|Indonesia|Taiwan|Hong
Kong|India|Japan|Islamic|Korea|ASEAN)", data2$category, ignore.case = TRUE), "Asia",

ifelse(grepl("(Russia)", data2$category, ignore.case = TRUE),
"Russia",

ifelse(grepl("(Global|Sector)", data2$category, ignore.case =
TRUE), "Global",

ifelse(grepl("(Australia)", data2$category, ignore.case =
TRUE), "Australasia",
```



```

    ifelse(grepl("(Africa)", data2$category, ignore.case =
TRUE), "Africa",

```

```

    ifelse(grepl("(Turkey)", data2$category, ignore.case
= TRUE), "Middle East", "Other"))))))))

```

```

#Kirjeldav statistika
library(dplyr)
library(ggplot2)
library(psych)

```

```

summary_data <- describe(data2)
ggplot(data2, aes(x = environmental_score)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-id keskkondlike skooride lõikes",
    x = "Keskkondlik skoor",
    y = "ETF-ide arv") +
  theme_minimal()

```

```

ggplot(data2, aes(x = social_score)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-id sotsiaalsete skooride lõikes",
    x = "Sotsiaalne skoor",
    y = "ETF-ide arv") +
  theme_minimal()

```

```

ggplot(data2, aes(x = governance_score)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-id juhtimise skooride lõikes",
    x = "Juhtimise skoor",
    y = "ETF-ide arv") +
  theme_minimal()

```

```

ggplot(data2, aes(x = sustainability_score)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-id ESG skooride lõikes",
    x = "ESG skoor",
    y = "ETF-ide arv") +
  theme_minimal()

```

```

ggplot(data2, aes(x = fund_trailing_return_3years)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-ide 3-aastane tootlus",

```

```

    x = "3-aastane tootlus",
    y = "ETF-ide arv") +
  theme_minimal()

ggplot(data2, aes(x = fund_trailing_return_5years)) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-ide 5-aastane tootlus",
    x = "5-aastane tootlus",
    y = "ETF-ide arv") +
  theme_minimal()

summary(data2$fund_size)
hist(data2$fund_size, breaks = 30, main = "ETF-ide suurus")

library(scales)

ggplot(data2, aes(x = log(fund_size))) +
  geom_histogram(binwidth = 0.1) +
  labs(title = "ETF-ide suurus (log)",
    x = "Log(ETF-ide suurus)",
    y = "ETF-ide arv") +
  scale_x_continuous(labels = scales::comma_format(scale = 1, suffix = "M")) +
  theme_minimal()

ggplot(data2, aes(x = fund_region)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "ETF-ide regioonidesse jagunemine", x = "ETF-ide regioon", y = "Arv") +
  theme_minimal()

library(tidyr)

long_data <- gather(data2, key = "Variable", value = "Return",
  fund_trailing_return_3years,
  fund_trailing_return_5years)

long_data$Variable <- factor(long_data$Variable, levels = c("fund_trailing_return_3years",
"fund_trailing_return_5years"))

#PCA osa

library(factoextra)

# Valime välja ainult 3 või 5 aasta andmed - vastavalt saab muuta select() sees 3years -> 5years
data_pca3 <- data2 %>%
  select(fund_trailing_return_3years, social_score, governance_score, sustainability_score,
sustainability_rank)

```

```

data_pca_3 <- na.omit(data_pca3)

# Võtame arvesse ainult numbreid
data_numeric <- data_pca3 %>%
  select_if(is.numeric)

# Skaleerime kõik andmed
scaled_data <- scale(data_numeric)
pca <- prcomp(scaled_data)

# Andmete ülevaade
pr.out <- prcomp(data_pca3, scale = TRUE)

# Loo me visualiseerimiseks graafiku
autoplot(pca, data = data2,
  loadings = TRUE, loadings.colour = "#1f3530",
  loadings.label = TRUE, loadings.label.size = 3,
  loadings.label.color = "#1f3530", loadings.label.alpha = 0.7,
  loadings.label.repel = TRUE, loadings.label.nudge = 0.1,
  loadings.arrow = arrow(length = unit(0.03, "npc")),
  loadings.arrow.fill = "#1f3530", loadings.arrow.alpha = 0.7,
  loadings.arrow.position = position_jitter(0.2),
  geom = "point", colour = "#e1efec")+
labs(
  x = "Esimene peakomponent",
  y = "Teine peakomponent",
  title = "Peakomponentide analüüs - 3 aasta tootlus"
) +
theme_classic() +
theme(
  axis.line = element_line(size = 0.5, colour = "black"),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.border = element_rect(colour = "black", fill = NA, size = 0.5),
  plot.title = element_text(size = 16, hjust = 0.5),
  legend.position = "top",
  legend.key.size = unit(0.4, "cm"),
  legend.key.width = unit(0.4, "cm"),
  legend.title = element_blank(),
  axis.title.x = element_text(size = 14),
  axis.title.y = element_text(size = 14),
  axis.text.x = element_text(size = 12),
  axis.text.y = element_text(size = 12),
  axis.text = element_text(color = "black")
)

```

```

# Scree Plot graafik
plot <- fviz_eig(pca_result, addlabels = TRUE) +
  labs(
    x = "Peakomponendid",
    y = "Seletab % variatsioonidest",
    title = "Scree Plot diagramm"
  ) +
  theme_minimal()

print(plot)

#Otsustsupuud

library(tree)
library(tidyverse)

european_funds_data <- read.csv("Morningstar - European ETFs.csv")

data <- european_funds_data %>%
  select(fund_trailing_return_3years, environmental_score, social_score,
    governance_score, fund_size, sustainability_score, sustainability_rank)

data <- na.omit(data)

# Looime treeningandmete valimi (60%).
set.seed(1)
training_index <- sample(c(1:dim(data)[1]), dim(data)[1]*0.6)
training_data <- data[training_index, ]

tree.training_data <- tree(fund_trailing_return_3years ~ ., training_data)
summary(tree.training_data)

par(mar = c(0, 0, 0, 0))
plot(tree.training_data)
text(tree.training_data, cex = 0.45)

cv.data <- cv.tree(tree.training_data)
plot(cv.data$size, cv.data$dev, type = "b")

prune.data <- prune.tree(tree.training_data, best = 5)
plot(prune.data)
text(prune.data, cex = 0.5)

yhat_data <- predict(tree.training_data, newdata = data[-training_index, ])
data.test <- data[-training_index, "fund_trailing_return_3years"]

```

```

plot(yhat_data, data.test$fund_trailing_return_3years,
     xlab = "Ennustatud väärtused", ylab = "Tegelikud Väärtused")
abline(0, 1)

mean((yhat_data - data.test$fund_trailing_return_3years)^2)

library(randomForest)

set.seed(1)
rf.etf <- randomForest(fund_trailing_return_3years ~ ., data = data,
                      subset = training_index, mtry = 6, importance = TRUE)
yhat.rf <- predict(rf.etf, newdata = data[-training_index, ])
mean((yhat.rf - data.test$fund_trailing_return_3years)^2)

importance(rf.etf)

varImpPlot(rf.etf)

```