Suomen Pelastusalan Keskusjärjestö
SPEK

27

Laura Kuurne ja Kari Telaranta

Ennakoiva analytiikka ja ohjattu koneoppiminen rakennuspaloriskien mallintamisessa

Laura Kuurne & Kari Telaranta

Ennakoiva analytiikka ja ohjattu koneoppiminen rakennuspaloriskien mallintamisessa

Ennakoivan analytiikan tutkimushanke 2020/2021

SPEK tutkii on julkaisusarja, jossa julkaistaan SPEKin omia tutkimus- ja selvityshankkeita. Julkaisut ovat saatavilla osoitteesta spek.fi/vaikuttaminen/tutkimukset/julkaisut/

Kansikuva
Shutterstock

Kansien suunnittelu ja ulkoasu
Johanna Kuittinen, SPEK

Taitto
Leena Huhmarniemi, SPEK

ISBN 978-951-797-715-717-3(pdf)
ISSN 2242-1653 (pdf)
Helsinki 2022

Julkaisija
Suomen Pelastusalan Keskusjärjestö SPEK
Ratamestarinkatu 11, 00520 Helsinki
Puhelin (09) 476 112, spekinfo@spek.fi
www.spek.fi

Sisällysluettelo

Tiivistelmä 5
Referat 8
Abstract 11
Keskeiset käsitteet 14
1. Johdanto 15
1.1 Tutkimuksen tavoitteet 17
1.2 Tutkimuksen teoreettinen viitekehys 17
1.3 Tutkimusongelmat- ja menetelmät 18
1.4 Tulosten hyödynnettävyys 19
1.5 Tutkimuksen eettisyys 20
2. Ennakoivan analytiikan käyttö rakennuspalojen ennustamisessa 22
2.1 Ennakoivan analytiikan käyttö pelastusalalla 22
2.2 Ulkomailla käytössä olevat ennakoivaa analytiikka hyödyntävät alustat 24
3. Laskennallinen viitekehys 28
3.1 Tarvittavan tiedon kartoitus 28
3.2 Tiedon keräys 34
3.3 Mallinnusmenetelmien kokeileminen käytännössä 35
4. Yhteenveto 62
5. Hankkeen viestintä 66
6. Lähteet 67
LIITE 1. Aineistojen kuvaus ja hakuprosessi 75

Kuvailulehti

Julkaisija	Suomen Pelastusalan Keskusjärjestö		29.4.2022
Tekijät	Laura Kuurne ja Kari Telaranta		
Julkaisun nimi	Ennakoiva analytiikka ja ohjattu koneoppiminen rakennuspaloriskien mallintamisessa		
Julkaisusarjan nimi ja numero	SPEK tutkii 27		
ISBN PDF	978-951-797-709-717-3	**ISSN PDF**	2242-1653
URN-osoite			
Sivumäärä	84	**Kieli**	FI
Asiasanat	Ennakoiva analytiikka, tekoäly, koneoppiminen, big data, mallinnus, laskennallinen viitekehys, rakennuspalo, rakennuspalovaara		
Kustantaja	Suomen Pelastusalan Keskusjärjestö		
Julkaisun myynti/ jakaja	Suomen Pelastusalan Keskusjärjestö		

Tiivistelmä

Ennakoivan analytiikan ja big datan käyttö onnettomuuksien ennakoinnissa ja pelastustoimen palveluiden kohdentamisessa on yleistynyt kansainvälisesti viimeisen viiden vuoden aikana. Erityisesti ohjatun koneoppimisen algoritmien hyödyntäminen alueellisten rakennuspaloriskien mallintamisessa sekä tapauskohtaisesti, että laajemmin integroituna erilaisiin ohjelmistoihin, on yleistynyt eri puolilla maailmaa. Ohjatun koneoppimisen menetelmistä saatava hyöty on kytköksissä saatavilla olevan tiedon määrään ja laatuun. Aiemmissa tutkimuksissa ennakoivan analytiikan menetelmien soveltamisen suurin rajoittava tekijä pelastustoimessa on ollut kattavien ja laadukkaiden paikkaan sidottujen tietoaineistojen saatavuus.

Suomessa pelastustoimen riskiarvio ja suunnittelu perustuu yhden neliökilometrin riskiruutuihin, joihin laskettu rakennuspalojen riskiarvio perustuu ruutujen asukaslukuun, kerrosalaan sekä niiden yhteisvaikutukseen. Aiemman kansainvälisen ja kotimaisen tutkimuksen perusteella asukasluvun ja kerrosalan lisäksi myös rakennuksen teknisillä ominaisuuksilla, ympäristötekijöillä ja erityisesti kiinteistöissä toimivilla tai asuvilla henkilöillä, on olennainen vaikutus tulipaloriskiin.

Tällä hetkellä lupaavimmalta pelastustoimen riskianalyysin toteuttamiseksi näyttäytyy erilaiset ohjattuun koneoppimiseen perustuvat riskianalyysimallit, jotka hyödyntävät useista eri lähteistä saatavaa tietoa. Näiden menetelmien hyödyntämistä tulipalojen ja muiden onnettomuuksien ennustamiseen on tutkittu useissa maissa, saaden hyviä tuloksia, ja niiden soveltamista Suomen pelastustoimessa tulisi tutkia kansallisessa viitekehyksessä.

Toistaiseksi ennakoivan analytiikan menetelmien soveltuvuudesta Suomen pelastustoimen käyttöön on vain vähän tietoa esimerkiksi oikeudelliseen toimintaympäristöön ja rakennustekniikkaan liittyvistä syistä. Tämän tutkimuksen keskeinen tutkimusongelma voidaan määrittää seuraavasti: Voidaanko viranomaisten hallussa olevia tietoja sekä ohjatun koneoppimisen algoritmeja hyödyntäen mallintaa rakennuskohtaista tulipaloriskiä Suomessa?

Tämä tutkimus on luonteeltaan soveltavaa tutkimusta. Kokeellisen tietojenkäsittelytieteen näkökulmasta tietojen käsittelyyn liittyviä ongelmia lähestyttiin ensin tunnistamalla mahdollisia aineistoja ja niiden analysoimiseen soveltuvia algoritmeja, jonka jälkeen ratkaisujen soveltuvuutta, luotettavuutta ja tehokkuutta arvioitiin käytännön kokeilujen avulla.

Tutkimushankkeen aikana tehtiin kaksi käytännön kokeilua, jonka mallinnuksissa hyödynnettiin eri ohjatun koneoppimisen algoritmeja sekä opetusaineistoa, joissa oli tietoa rakennuksista, onnettomuuksista ja sosioekonomista tietoa väestöstä. Ensimmäinen kokeilu toteutettiin avoimilla aineistoilla ja Pelastustoimen toimenpide- ja resurssirekisteri PRONTOn onnettomuustiedoilla pääkaupunkiseudun alueelta. Toinen kokeilu toteutettiin koko Suomen alueella Tilastokeskuksen ja PRONTOn aineistoja hyödyntäen.

Kokeilujen perusteella päätöspuumalleihin perustuvat ohjatun koneoppimisen menetelmät, nk. XGBoost ja satunnaismetsä eli nk. Random Forest, soveltuivat tarkkuuden ja tehokkuuden puolesta käytettävissä olevien aineis-

tojen mallintamiseen parhaiten. Koko Suomen rakennuspaloriskiarvojen mallintamisessa asukasmäärä osoittautui kaikista painavammaksi rakennuspaloja selittäväksi tekijäksi. Toiseksi painavin muuttuja oli autolliset asukaskunnat. Työttömyyteen, toimeentuloon ja eläkkeeseen liittyvät muuttujat olivat myös keskeisiä. Aineistojen saaminen tutkimuskäyttöön osoittautui hankkeen suurimmaksi haasteeksi.

Tutkimuslupien ja tietopyyntöjen vaativuuden takia hakemusprosessit olivat pitkiä ja edellyttivät paljon resursseja. Jotta vastaavanlaista aineistokokonaisuutta voitaisiin hyödyntää jatkuvassa arvioinnissa ja päätöksenteossa, tulisi keskeiset aineistot tulevaisuudessa koota yhteen päivitettävään tietojärjestelmään, kuten Sisäministeriön parhaillaan kehittämään onnettomuuksien ehkäisyn tietojärjestelmään. Onnettomuustietojen lisäksi näihin keskeisiin aineistoihin kuuluvat esimerkiksi Tilastokeskuksen FOLK-henkilötietoaineiston moduulit.

Presentationsblad

Utgivare	Räddningsbranschens Centralorganisation i Finland		29.4.2022
Författare	Laura Kuurne och Kari Telaranta		
Publikationens titel	Prediktiv analys och guidad maskininlärning i byggnadsbrandriskmodellering		
Publikationsseriens namn och nummer	SPEK undersöker 27		
ISBN PDF	978-951-797-709-717-3	**ISSN PDF**	2242-1653
URN-adress			
Sidantal	84	**Språk**	FI
Nyckelord	Prediktiv analys, artificiell intelligens, maskininlärning, big data, modellering, beräkningsramverk, byggnadsbrand, byggnadsbrandrisk		
Förläggare/ Distribution	Räddningsbranschens Centralorganisation i Finland		

Referat

Användningen av proaktiv analys och big data för att förutsäga olyckor och förbättring av räddningstjänster har blivit mer utbredd internationellt under de senaste fem åren. I synnerhet har användningen av styrda maskininlärningsalgoritmer för att modellera regionala brandrisker i byggnader, både från fall till fall och mer integrerat i olika programvaror, blivit utbredd över hela världen. Fördelarna med maskinstyrda inlärningsmetoder är kopplade till mängden och kvaliteten på tillgänglig information. I tidigare studier har den främsta begränsande faktorn vid tillämpningen av proaktiva analysmetoder i räddningsinsatser varit tillgången till omfattande och högkvalitativ platsspecifik data.

I Finland baseras räddningstjänstens riskbedömning och planering på en kvadratkilometer riskrutor, där den beräknade brandriskbedömningen ba-

seras på torgens befolkning, golvyta och deras sammanlagda effekt. Baserat på tidigare internationell och inhemsk forskning har, förutom antalet invånare och golvyta, byggnadens tekniska egenskaper, miljöfaktorer och i synnerhet de personer som arbetar eller bor i byggnaderna en betydande inverkan på brandrisk.

I dagsläget verkar det mest lovande sättet att genomföra en räddningsriskanalys vara olika riskanalysmodeller baserade på maskinstyrt lärande, som använder sig av information från en mängd olika källor. Användningen av dessa metoder för att förutsäga bränder och andra olyckor har studerats över hela världen, med goda resultat, och deras tillämpning i finländska räddningsinsatser bör studeras inom det nationella ramverket.

Än så länge är lite känt om tillämpbarheten av metoderna för prediktiv analys i Finland av skäl relaterade till till exempel den juridiska verksamhetsmiljön och byggtekniken. Det huvudsakliga forskningsproblemet i denna studie kan fastställas enligt följande: Är det möjligt att modellera den byggnadsspecifika brandrisken i Finland med hjälp av myndigheternas data och de guidade maskininlärningsalgoritmerna?

Denna studie är inom grenen tillämpad vet enskap. Ur experimentell datavetenskaps synvinkel togs problemen relaterade till databehandling först upp genom att identifiera möjliga data och algoritmer lämpliga för deras analys, varefter lämpligheten, tillförlitligheten och effektiviteten av lösningarna utvärderades genom praktiska experiment. Under forskningsprojektet genomfördes två praktiska experiment, vars modellering utnyttjade olika algoritmer för guidad maskininlärning samt läromedel med information om byggnader, olyckor och socioekonomisk information om befolkningen. Det första experimentet genomfördes med öppna data och PRONTO-olycksdata i huvudstadsregionen. Det andra experimentet genomfördes i hela Finland med hjälp av uppgifter från Statistikcentralen och PRONTO.

Baserat på experimenten var de guidade maskininlärningsmetoderna baserade på beslutsträdsmodeller, den så kallade XGBoost, och den slumpmässiga skogen, den så kallade Random Forest, bäst lämpade för att modellera tillgängliga data för noggrannhet och effektivitet. Vid modellering av byggnadsbrandriskvärden för Finland som helhet visade sig antalet invånare vara den viktigaste faktorn för att förklara byggnadsbränder. Den näst viktigaste variabeln var om det finns en bil i bostaden. Variabler relaterade

till arbetslöshet, försörjning och pensioner var också viktiga. Att få fram materialet för forskning visade sig vara projektets största utmaning.

På grund av den krävande karaktären av undersökningstillstånd och förfrågningar om information var ansökningsprocesserna långa och resurskrävande. För att kunna använda en liknande datauppsättning i det fortlöpande beslutsfattandet bör nyckeluppgifterna i fortsättningen sammanställas till ett aktuellt informationssystem, såsom det olycksförebyggande informationssystem som för närvarande utvecklas av inrikesministeriet. Dessa nyckeluppgifter omfattar förutom olycksdata till exempel modulerna i Statistikcentralens FOLK-personuppgifter.

Description sheet

Published by	The Finnish National Rescue Association		29.4.2022
Authors	Laura Kuurne and Kari Telaranta		
Title of publication	Predictive modeling applied in accident analysis and fire prevention		
Series and publication number	SPEK research 27		
ISBN PDF	978-951-797-709-717-3	**ISSN PDF**	2242-1653
Website address URN			
Pages	84	**Language**	FI
Keywords	Predictive modeling, AI, Machine learning, Big data, Statistical framework, Building fire, risk analysis		
Publisher/ Distributed by	The Finnish National Rescue Association		

Abstract

Over the past five years, the use of predictive modeling and big data in accident prediction and prevention has become increasingly popular in public safety industry. In particular, the use of supervised machine learning algorithms in regional building fire risk modelling, both on a case-by-case basis and more widely integrated into different software's, has become more common in different parts of the world. However, one of the requirements to use these types of analyses and to get the best outcome from them is to have access to big data, meaning large amount of structured data collected by different organizations and public agencies. In previous studies, access to high-quality and comprehensive information has been the biggest challenge when applying these techniques.

In Finland, the national risk assessment used in planning by rescue services is based on a regression model that utilize one square kilometer of risk

squares, on which the calculated risk assessment of building blocks is based on the number of inhabitants of the squares, the floor area, and their combined impact. However, according to previous international and domestic studies, technical characteristics of the building, environmental factors and the persons working or living in the buildings, have a significant impact on the risk of fire. According to previous studies, supervised machine learning algorithms utilizing information from multiple data sources seem to be the most promising methodology to conduct Rescue Services' risk analysis. These methods have been successfully applied in predicting fires and other accidents in several case studies and their application by the Finnish rescue services should be investigated in the national framework.

So far, there is little information on the suitability of predictive modelling methods in Rescue Services in Finland, mainly due to reasons related to the legal environment and construction technology. The main research problem of this study can be defined as follows: Can the data held by the authorities and supervised machine learning algorithms be used to model the building-specific fire risk in Finland?

This study is of an applied nature. From the experimental computing science point of view, problems related to data processing were first approached by identifying possible data and algorithms suitable for their analysis, after which the suitability, reliability and efficiency of the solutions were evaluated through practical experiments.

During the research project, two practical experiments were conducted using different supervised machine learning algorithms and datasets containing information on buildings, accidents, and socio-economic information on the population. The first experiment was conducted with open data and PRONTO (database of the fire and rescue service's official register of incidents) data from the Helsinki Metropolitan Area. The second experiment was conducted in the whole territory of Finland by using PRONTO data and Statistics Finland (national statistical institution in Finland) data.

The experiments indicate that decision tree models, such as XGBoost and Random Forest are best suited for modelling the data available in terms of accuracy and efficiency. In modelling the building fire risk values for the whole of Finland, the number of inhabitants proved to be the most important variable explaining building fires. The second-most important variable

was households that own a car. The variables related to unemployment, income and pensions were also key in explaining the prevalence of fire occurrences. Access to data proved to be the biggest challenge of the project. Due to the complexity of research permits and requests for information, the application processes were lengthy and required a lot of resources. The findings of this study indicate that key data should be compiled into an information system, e.g., the database currently developed by the ministry of the interior, so that these techniques and data can be utilized in continuous evaluation and decision-making. In addition to accident data, FOLK-data from Statistics Finland should be included.

Keskeiset käsitteet

Ennakoiva analytiikka: Analyysimenetelmä, jolla ennakoidaan tulevaa, käyttämällä isoa määrää tietoa menneisyydestä. Ennakoivan analytiikan menetelmiä voidaan myös käyttää nykyhetken tilannekuvan tai riskiarvion muodostamiseen aikaisempien tietojen perusteella. (Abbot 2014)

Tekoäly: Kuvaa tekniikoita, joiden avulla tietokoneet saadaan ratkomaan ongelmia ja tekemään älyllisiä johtopäätöksiä itsenäisesti.

Koneoppiminen: Tekoälyn osa-alue, joka kuvaa toimintaa, jossa algoritmit oppivat tekemään päätöksiä ja jäsentämään dataa itsenäisesti. (Kotsiantis ym. 2007)

Big data: Kuvaa suuria tietomassoja, joita tulee monesta eri lähteestä, lisääntyy jatkuvasti ja on muodoltaan vaihtelevaa.

Mallinnus: Tämän tutkimuksen yhteydessä mallinnuksella viitataan laskennalliseen mallintamiseen, jonka avulla analysoidaan ilmiöitä, joiden syntyyn vaikuttaa monia välittömiä ja välillisiä tekijöitä.

Laskennallinen viitekehys: Yhdistelmä algoritmeja ja ohjelmistokoodeja, jonka avulla pystytään tuottamaan monivaiheisia ja moniulotteisia analyysejä.

Rakennuspalo: Pelastusopiston määritelmän mukaan ”*palo, jossa palo on levinnyt syttymiskohdastaan sytyttäen rakennuksen rakenteet tai irtaimiston palamaan, joko liekehtien tai kytemällä*” (Pelastusopisto 2013a).

Rakennuspalovaara: Pelastusopiston määritelmään mukaan ”*palosta, kuumenemisesta tai kärähtämisestä on syntynyt näkyvää savua. Tilanteesta olisi ollut mahdollista kehittyä rakennuspalo, mutta se ei ole levinnyt kuumenemis- tai syttymiskohdasta rakennuksen rakenteisiin tai irtaimistoon*.” (Pelastusopisto 2013b.)

1. Johdanto

Big datan eli suurten, järjestelmättömien ja jatkuvasti lisääntyvien tietomassojen hyödyntäminen sisältää uusia mahdollisuuksia kohdentaa viranomaisen toimintaa riskiperusteisesti. Jaetun tietopohjan mahdollistama parempi riskiprofilointi, missä hyödynnetään tietoa, analytiikkaa ja tekoälyä, onkin tunnistettu keskeiseksi viranomaistoiminnan vaikuttavuutta parantavaksi tekijäksi esimerkiksi työ- ja elinkeinoministeriön ”Luvat ja valvonta” -hankkeessa (Tommila 2017). Kansainvälisesti on jo tehty lupaavia havaintoja ennakoivan analytiikan hyödyntämisestä pelastustoimen valvonnan kohdentamisessa. Esimerkiksi Yhdysvalloissa pelastustoimessa on otettu käyttöön erinäisiä ennakoivaa analytiikkaa ja big dataa hyödyntäviä ohjelmistoja ja laskennallisia viitekehyksiä, kuten Firebird, Firecast, ja Firecare (Madaio ym. 2016; Roman 2014; Gilbert & Gilbert 2018).

New Yorkin pelastuslaitosten käytössä olevaan Firecast-ohjelmistoon on integroitu ohjattua koneoppimista hyödyntäviä mallinnustekniikoita ja jatkuvasti päivittyviä tietovarantoja, mahdollistaen reaaliaikaisen riskianalyysin tuottamisen pelastuslaitosten päivittäisen suunnittelun tueksi. Paremman valvonnan kohdentamisen seurauksena valvonnan yhteydessä havaittujen puutteiden lukumäärä on noussut lähes 20 prosenttia (Apolitical 2017). Viime vuosina ulkomailla rakennuspalojen ennakoimiseen käytettyjä ohjatun koneoppimisen algoritmeja on testattu myös tapauskohtaisesti esimerkiksi Kiinassa, Etelä-Koreassa ja Britanniassa (Wang ym. 2019, Seon-gwan Hon & Seung-ryul Jeong 2018, Dang 2019).

Suomessa monella hallinnonalalla on parhaillaan käynnissä tietojärjestelmien uudistustyö. Hallitusohjelman 2019 mukaan Suomi tunnetaan edelläkävijänä, jossa digitalisaation ja teknologisen kehityksen luomia mahdollisuuksia kehitetään ja otetaan käyttöön yli hallinto- ja toimialarajojen. Hallitusohjelman mukaan Suomeen luodaan rakennetun ympäristön valtakunnallinen digitaalinen rekisteri ja tietoalusta, joihin maankäyttöä ja rakentamista koskevat päätökset ja prosessit tukeutuvat. (Valtioneuvosto 2019.) Vastaavasti pelastustoimen onnettomuuksien ehkäisyn tietojärjestelmä on kehitteillä, kuten myös lukuisia muita tietojärjestelmähankkeita, esimerkiksi Turvallisuus- ja kemikaaliviraston lupien ja valvonnan digitaalisen hallinnan tietojärjestelmä ”VALLU”.

Tiedon luokittelu parantaa keskeisesti tiedon hyödynnettävyyttä. Tietojärjestelmien rakentamisvaiheessa niihin on mahdollisuus vaikuttaa, mutta myöhemmin se voi olla vaikeampaa, ellei jopa mahdotonta. Tietojärjestelmien välisten rajapintojen kannalta on keskeistä selvittää niitä tietoja, jotka ovat keskeisiä pelastustoimen riskianalyysin tekemiseksi tulevaisuudessa, jotta nämä rajapinnat voidaan huomioida niin eri tietojärjestelmien suunnittelussa kuin pelastustoimen tiedonsaantioikeuksia koskevan lainsäädännön valmistelussa.

Pelastustoimen lakisääteisenä tehtävänä on kohdentaa pelastustoimen tavoitteet, voimavarat, palvelut ja niiden taso sekä kohdennettava valvontaa riskiperusteisesti. Tällä hetkellä käytössä oleva pelastustoimen riskianalyysi perustuu regressiomalliin, jossa muuttujina ovat ruudun asukasluku, kerrosala sekä niiden yhteisvaikutus (Tillander ym. 2010). Tämän lisäksi pelastustoimen toimintavalmiuden suunnitteluohjeen mukaisesti riskiluokan määräytymiseen vaikuttaa myös ruudussa (viiden vuoden seurantajaksolla) tapahtuneiden onnettomuuksien määrä. Riskiluokan määrittämisessä ei tällä hetkellä huomioida esimerkiksi ruudussa uhattuna olevia arvoja tai tapahtuneita vahinkoja. Pelastustoimen tällä hetkellä käyttämää regressiopohjaista riskianalyysimallia on arvosteltu muun muassa sote-maakuntalakipaketin lausunnoissa.

Paikkatiedollinen big data antaa pelastustoimelle uusia mahdollisuuksia ennakoida, missä erilaisia onnettomuuksia todennäköisesti tapahtuu, minkälaiset seuraukset olisivat todennäköisiä ja miten resursseja tulisi kohdentaa. Tällä hetkellä lupaavimmalta pelastustoimen riskianalyysin toteuttamiseksi näyttäytyy erilaiset algoritmipohjaiset riskianalyysimallit, jotka hyödyntävät useista eri lähteistä saatavaa tietoa. Koska näiden menetelmien hyödyntämisestä tulipalojen ja muiden onnettomuuksien ennakoimisessa on saatu hyviä tuloksia maailmalla, tulisi tutkia niiden soveltuvuutta kansallisessa viitekehyksessä.

Pelastustoimen riskiarvio, suunnittelu, seuranta, tutkimus ja kehitystoiminta pohjautuvat jatkossa pelastustoimen valtakunnalliseen tilasto- ja paikkatietoaineistoon. Aineistoon on koottu valtakunnallisiin tietoaineistoihin pohjautuvia valmiita tilasto- ja paikkatietoaineistoja, pelastustoimen alueittain ja riskiarvion alueyksiköittäin (Sisäministeriö, 2020). Uuden alustan olisi kuitenkin mahdollistettava huomattavasti Pelastustoimen toimenpide- ja resurssirekisteriä (PRONTO) monipuolisempia raportointi- ja analyysitoiminnallisuuksia erityisesti paikkatietoaineistojen ja paikkatietomuotoon ja-

lostettavien tilastoaineistojen osalta, kuten riskiruutu-perusteisesta tarkastelusta, onnettomuuksien lukumäärästä ja onnettomuusprofiilista, onnettomuusvahingoista ja laskennallisista riskeistä.

1.1 Tutkimuksen tavoitteet

Tämän tutkimuksen tavoitteena on tuottaa tietoa, joka tukee pelastustoimen riskianalyysin kehittämistä. Tutkimuksen avulla luodaan kokonaiskuva ennakoivan analytiikan tarjoamista mahdollisuuksista, käyttötavoista ja rajoituksista, jotka tulisi huomioida Suomen pelastustoimen riskianalyysitoiminnassa. Tarkoituksena on selvittää jo olemassa olevien riskianalyysialustojen ja analyysimenetelmien soveltuvuutta ja koostaa analyysimenetelmiä, joilla voidaan kohdentaa pelastustoimen palveluita aiempaa riskiperusteisemmin. Tutkimuksen tavoitteena on myös edistää tiedon tehokkaampaa hyödyntämistä kartoittamalla uusia aineistoja, joista on pelastustoimelle hyötyä valtakunnallisella tasolla.

1.2 Tutkimuksen teoreettinen viitekehys

Resurssit, esimerkiksi data- ja IT-infrastruktuuriratkaisut sekä toiminnot, kuten big data -analyysi, on kuvattu olennaisina mekanismeina, jotka edistävät organisaatioiden arvon luomista. On tärkeää, että organisaatio tunnistaa ja ymmärtää tietoon perustuvan arvon luomisen merkityksen kilpailukyvyn saavuttamiseksi ja parempien palvelujen tarjoamiseksi. Resurssipohjaisen näkökulman mukaan yritykset, jotka hankkivat arvokkaita resursseja ja jalostavat niitä tarkoituksenmukaisesti, voivat luoda ainutlaatuisia arvoja tai ominaisuuksia, mitkä antavat niille kilpailuetua. Tämä on yleisimpiä organisaatioteoreettisia lähestymistapoja big data -tutkimukseen (Galetsi ym. 2020).

Ennakoivan analytiikan ja ohjatun koneoppimisen algoritmien käytön yleistyminen rakennuspalojen ennakoimisessa kansainvälisesti pelastusalla on selitettävissä Galetsin ym. kuvaamalla organisaatioteoreettisella lähestymistavalla big data-analyyseihin. Kykenemällä luomaan tarkkoja rakennuspalojen riskiarvioita, pystytään ennakoimaan tulipaloja ja parantamaan organisaation palvelukykyä kohdentamalla palvelut sinne, missä niitä eniten tarvitaan.

Tämä tutkimus on luonteeltaan soveltavaa tutkimusta, jonka päämääränä ovat käytännön sovellutukset. Soveltavassa tutkimuksessa sovelletaan perustutkimuksen tuottamaa tieteellistä tietoa johonkin käytännölliseen ongelmaan. Soveltavaa tutkimusta on esimerkiksi sovellusten etsiminen perustutkimuksen tuloksille tai uusien menetelmien ja keinojen luominen tietyn ongelman ratkaisemiseksi. Tieteenalojen välisessä jaottelussa tämä tutkimus sitoutuu ensisijaisesti tietojenkäsittelytieteen tutkimustraditioihin. Tietojärjestelmien rakenteiden ja niiden eroavaisuuksien johdosta big data -tutkimus vaatii usein edistyksellisiä tietojenkäsittelymenetelmiä. Tässä tutkimuksessa sovelletaan sekä teoreettista, että kokeellista tietojenkäsittelytiedettä.

Teoreettisen tietojenkäsittelytieteen näkökulmasta pyritään syventämään tietämystä algoritmien ja tietojenkäsittelyn rajoituksista ja potentiaalista sekä löytämään mahdollisimman tehokkaita ja luotettavia ratkaisuja ongelmiin, jotka liittyvät big data –analyysien soveltamiseen riskianalyysityössä. Kokeellisen tietojenkäsittelytieteen näkökulmasta tietojenkäsittelyyn liittyviä ongelmia lähestytään ensin tunnistamalla mahdollisia ratkaisuja ja sitten arvioimalla ratkaisujen soveltuvuutta, luotettavuutta ja tehokkuutta. Tämän avulla pyritään löytämään ratkaisuja moniulotteisiin ongelmiin, joissa esimerkiksi käytössä olevan tiedon analysoiminen vaatii uusien laskennallisten menetelmien kehittämistä tai soveltamista uusiin tarkoituksiin. Tietojenkäsittelytieteessä ja big data-analyyseissä hyödynnetään tyypillisesti ohjatun koneoppimisen menetelmiä ja niitä sovelletaan eri toimialojen riskianalyysityössä yhä enenemissä määrin. (Dodig-Crnkovic 2002)

1.3 Tutkimusongelmat- ja menetelmät

Ulkomailla tehdyissä tutkimuksissa on havaittu, että big data-analytiikalla, eli laajoilla tietoaineistoilla ja niihin kohdistuvien algoritmipohjaisten analyysien avulla on mahdollista ennustaa tulipaloja melko hyvilläkin tarkkuuksilla. Toistaiseksi ei ole tietoa näiden mallien soveltuvuudesta Suomeen esimerkiksi kulttuuriin, oikeudelliseen toimintaympäristöön tai rakennustekniikkaan liittyvistä syistä.

Tämän tutkimuksen keskeinen tutkimusongelma voidaan määrittää seuraavasti:

Voidaanko eri viranomaisten hallussa olevia tietoja sekä keinoälyalgoritmeja hyödyntäen ennustaa rakennuskohtaista tulipaloriskiä Suomessa?

Tästä muodostuu tutkimuksen hypoteesi. Hypoteesin mukaan rakennusten teknisillä ominaisuuksilla sekä siellä asuvien ihmisten käyttäytymisellä ja taustoilla on keskeinen vaikutus tulipaloriskiin. Tämän tutkimuksen tavoitteena on kokeellisesti testata, pitääkö edellä mainittu hypoteesi paikkansa. Tämä toteutetaan käyttämällä ohjattua koneoppimista hyödyntäviä mallinnusmenetelmiä.

Mallinnuksessa hyödynnetään opetusaineistoa, joka koostuu tutkittavasta ilmiöstä ja sitä mahdollisesti selittävistä tekijöistä. Aiemmissa tutkimuksissa, joissa ohjatun koneoppimisen menetelmiä on käytetty rakennuspalojen ennakoimiseen, ilmiötä selittävinä tekijöinä on käytetty tietoa rakennuksista, onnettomuuksista ja sosioekonomista tietoa väestöstä. Selittäviä tekijöitä kartoitetaan koostamalla aiemmissa tapaustutkimuksissa käytettyjä muuttujia. Muuttujien olennaisuus varmistetaan ohjausryhmälle suunnatulla kyselyllä. Kun opetusaineistoon on koottu kaikki tunnistetut tekijät, joilla saattaa olla välitön tai välillinen vaikutus rakennuspalon syntyyn, mallia koulutetaan aineiston avulla ja se muodostaa funktion, eli koneoppimisratkaisun, jolla se luokittelee tutkittavan aineiston.

Koneoppimisratkaisun suorituskykyä arvioidaan k-kertaisella ristiinvalidiointi-menetelmällä, jossa esimerkiksi 80 prosentin aineistosta käytetään mallin koulutukseen ja 20 prosenttia mallin testaamiseen (Refaeilzadeh ym. 2009). Suorituskyvyn mittareina käytetään esimerkiksi Receiver Operating Character (ROC) ja Area Under the ROC curve (AUC) käyriä, jotka visualisoivat binääriluokittelijan oikea positiivisuus ja väärä positiivisuus -asteiden välistä vaihtuvuutta ja kuvaa mallin tarkkuutta (Huang & Ling 2005). Rakennuspaloja keskeisesti selittäviä tekijöitä kartoitetaan tutkimalla muuttujien painoarvoa mallinnuksessa.

1.4 Tulosten hyödynnettävyys

Hankkeessa syvennetään tietoa menetelmistä, joiden tavoitteena on tukea kansallisia analytiikan potentiaalin hyödyntämisen tavoitteita käyttämällä ennakoivaa analytiikkaa, big dataa ja koneoppimista. Tämän tutkimuksen tuloksista saadaan tietoa pelastustoimen tiedolla johtamisen tarpeita varten, jota voidaan hyödyntää tiedolla johtamisen työkalujen suunnittelussa. Tavoitteena on tunnistaa keskeiset tiedontuottajat, jotka ovat välttämättömiä algoritmipohjaisen riskianalyysin tekemiseksi, ja siten mahdollistaa niiden kytkeminen pelastustoimen tietojärjestelmiin.

Tuotetut menetelmät auttavat kohdentamaan pelastustoimen palveluita aiempaa riskiperusteisemmin. Tutkimustuloksia hyödyntämällä esimerkiksi paloasemien sijainnit ja onnettomuuksien ehkäisyyn liittyvät tehtävät voidaan kohdentaa aiempaa paremmin, mikä lisää yhteiskunnan turvallisuutta ja tehostaa varojen käyttöä.

Hankkeen tulokset ovat suoraan hyödynnettävissä kansallisessa kehittämistyössä, ja hankkeen tavoitteena on palvella kansallisia analytiikan potentiaalin hyödyntämisen tavoitteita turvallisuuden edistämisessä laajemminkin. Algoritmipohjainen lähestyminen mahdollistaa hyvin erilaisten ilmiöiden ennustamisen, joten hankkeen löydökset voivat hyödyttää myös muita viranomaisia ja julkishallinnon toimijoita.

1.5 Tutkimuksen eettisyys

Tutkimusta tehdessä on huomioitu Tutkimuseettisen neuvottelukunnan (TENK) yleiset tutkimuseettiset periaatteet (Kohonen ym. 2019). Tekoälyjärjestelmien ja koneoppimisen käytön eettisyydestä tiedolla johtamisessa ja päätöksen teossa on käyty paljon yhteiskunnallista keskustelua viime vuosina. Koneoppimiseen liittyvät eettiset ongelmat liittyvät usein järjestelmien ohjelmointiin, niihin asetettuihin sääntöihin, käytettyihin aineistoihin ja käyttötarkoituksiin (Nieminen ym. 2019). Väärinkäytöksien ehkäisemiseksi aiemmin mainitut seikat on otettu huomioon.

Nämä seikat on otettu huomioon siten, että tutkimuksella on asianmukainen ja julkinen tutkimussuunnitelma, aineistonhallintasuunnitelma ja henkilötietojen käsittelyä koskeva seloste. Tutkimukselle on nimetty vastuuhenkilöt ja tutkimukseen haettuja aineistoja käytettiin vain tieteellistä tutkimusta varten, tietosuojalainsäädäntöä noudattaen ja siten, että tiettyä henkilöä koskevat tiedot eivät paljastu ulkopuolisille.

Riskiprofiloinnissa käytetyn opetusaineiston muuttujat on valittu huolella, aiemman tutkimuksen ja asiantuntijoiden konsultaatioihin perustuen. Menetelmillä muodostettuja riskiarvioita ei käytetä syrjintään tai alueiden tai ihmisten arvottamiseen, vaan tulevaisuudessa mahdollisesti pelastustoimen palveluiden kohdentamiseen, mikä lisää yhteiskunnallista turvallisuutta ja on kaikkien kansalaisten etujen mukaista.

Tutkimuksen tavoitteena ei ole tutkia tulipaloriskiä yksilötasolla. Ihminen on kuitenkin merkittävin muuttuja, jolla on vaikutusta tulipalon syttymistaajuuteen. Ilman ihmistä tulipaloja olisi huomattavasti nykyistä vähemmän. Tutkimuksessa yksilötason tiedot tarvitaan vain sitä varten, että inhimillinen tekijä saadaan yhdistettyä osaksi rakennuskohtaista tietoa. Tutkimushankkeessa ei tämän yhdistämisen jälkeen käsitellä yksilökohtaisia tietoja. Tulosten raportoinnissa ruutuihin tai muihin maantieteellisiin alueisiin perustuvia tietoja voidaan julkaista silloin, kun yksittäisiä henkilöitä tai asuntokuntia ei voi tunnistaa. Tulosten raportoinnissa aggregoimme tulokset suuremmalle aluetasolle siten, että suora ja epäsuora tunnistaminen on estetty.

Hankkeelle nimettiin ohjausryhmä, johon kuuluu seuraavat jäsenet ja varajäsenet:

Alisa Puustinen, erikoistutkija, Pelastusopisto
Esa Kokki, erityisasiantuntija, sisäministeriö
Hanna Rekola, tutkimuspäällikkö, Helsingin pelastuslaitos
Heidi Huuskonen, kehittämispäällikkö, Etelä-Karjalan pelastuslaitos
Heidi Liukkonen, suunnittelija, Pelastusopisto
Kati Tillander, riskienhallintapäällikkö, Keski-Uudenmaan pelastuslaitos
Petteri Mustonen, ylitarkastaja, Turvallisuus- ja kemikaalivirasto
Rami Ruuska, pelastusylitarkastaja, sisäministeriö
Riku Leppänen, johtava valvonnan suunnittelija, Helsingin pelastuslaitos
Sara Lax, prosessiturvallisuusasiantuntija, Turvallisuus- ja kemikaalivirasto
Simo Hostikka, professori, Aalto-yliopisto
Teija Mankkinen, erityisasiantuntija, sisäministeriö
Thomas Nyqvist, riskienhallintapäällikkö, Pohjanmaan pelastuslaitos

Hankkeessa tehtiin myös yhteistyötä sisäministeriön Pelastustoimen ja siviilivalmiuden suorituskyky ja suunnitteluperusteet -hankkeen, kumppanuusverkoston riskianalyysityöryhmän sekä pelastustoimen onnettomuuksien ehkäisyn toimintaohjelmahankkeen koordinaatioryhmän kanssa. Vuonna 2020 valmistuneessa KAT3 hankkeessa ja samana vuonna alkaneessa KAT-TIME hankkeissa on yhteneväisyyksiä ennakoivan analytiikan tutkimushankkeen kanssa. Hankkeista vastaavat Etelä-Karjalan pelastuslaitos, Eksote, Lappeenrannan-Lahden teknillinen yliopisto (LUT), Pelastusopisto sekä LAB-ammattikorkeakoulu. Hankkeen työntekijöiden kanssa käytiin palavereja hankkeiden kytköksiin liittyen ja jaettiin keskenään tietoa tutkimusten löydöksistä ja tavoitteista, jolla myös varmistettiin, ettei hankkeissa tehdä päällekkäistä työtä.

2. Ennakoivan analytiikan käyttö rakennuspalojen ennustamisessa

Ennakoivalla analytiikalla kuvataan prosessia, missä suuresta määrästä tietoa louhitaan merkityksellisiä kaavoja menneistä tapahtumista ja niihin liittyvistä tekijöistä, joiden avulla voidaan luoda tarkkoja riskiarvioita ja ennakoida tulevaa (Abbot 2014). Käytännössä ennakoiva analytiikka on todennäköisyyslaskentaa, jonka avulla pyritään ennakoimaan eri ilmiöiden ja tulevien tapahtumien ilmentymistä. Ennakoivaa analytiikkaa, koneoppimista ja big dataa on jo pitkään hyödynnetty monilla sovellusalueilla, kuten teollisuudessa huoltotarpeiden ennakoinnissa, terveydenhuollossa hoitosuosituksien tekemisessä ja liikemaailmassa markkinoinnin kohdentamisessa (Dubei ym. 2019, Dinov 2018, Bradlow ym. 2017).

Ennakoiva analytiikka on keskeinen osa tiedolla johtamisen datavetoista toteutumista, minkä avulla päätökset tehdään algoritmien löydöksien, ei olettamuksien perusteella. Tutkittaessa ilmiötä, jonka kohdemuuttujan esiintymistä halutaan ennakoida, käytetään usein ohjatun koneoppimisen menetelmiä. Tässä tapauksessa ilmiö voi olla esimerkiksi onnettomuus, jolloin kohdemuuttuja kuvaa sitä, tapahtuiko tietyssä ympäristössä onnettomuutta vai ei. Kohdemuuttujaa verrataan erilaisiin tapahtumaan liittyviin ympäristötekijöihin. Ohjatussa koneoppimisessa algoritmien luokittelija muodostetaan opetusaineiston avulla, joka sisältää suuren määrän tietoa ilmiöstä, eli kohdemuuttujasta ja sitä mahdollisesti selittävistä tekijöistä (Kotsiantis ym. 2007). Ohjatun koneoppimisen avulla voidaan automatisoida tiedon luokittelua ja prosessoida suuria määriä tietoa.

2.1 Ennakoivan analytiikan käyttö pelastusalalla

Ennakoivan analytiikan ja koneoppimisen hyödyntäminen pelastusalan riskianalyysitoiminnassa on kasvava ilmiö. Tämän hankkeen alkaessa Suomen pelastustoimessa oli käynnissä muutamia riskianalyysimenetelmiä tutkivia hankkeita, kuten LAB-ammattikorkeakoulun hallinnoima KAT-TIME hanke ja sisäministeriön Pelastustoimen ja siviilivalmiuden suorituskyky ja suunnitte-

luperusteet –hanke, jatkossa Suorituskyky-hanke, jonka työpaketin 2 tarkoituksena on uudistaa riskimallit ja onnettomuusennusteet.

KAT-TIME hankkeessa kehitetään ennakoiviin mallinnustekniikoihin perustuvaa viitekehystä, jonka tarkoituksena on tehostaa pelastustoimen verkoston palveluvastetta ja ennakoida riskien alueellista jakaantumista (KAT kotona asumisen turvallisuus, 2022). Hankkeessa kehitettyä viitekehystä kokeiltiin Etelä-Karjalan alueella ja hanke toteutettiin yhteistyössä Etelä-Karjalan pelastuslaitoksen, Eksoten, Lappeenrannan-Lahden teknillinen yliopiston, Pelastusopiston sekä LAB-ammattikorkeakoulun kanssa. Aineistojen saatavuuden takia viitekehystä ei ole kokeiltu valtakunnallisesti. Suorituskyky-hankkeen työpaketissa 2 selvitetään ennakoivan analytiikan kytköksiä tiedolla johtamisen järjestelmään, laaditaan uusittu malli palvelutason mitoittamiseen sekä tehdään esitys paloasemaverkoston suunnitteluun. Hankkeet luovat pohjaa kehitystyölle, jonka saralla riittää vielä paljon tutkittavaa.

Ulkomailla pelastustoimessa ennakoivan analytiikan menetelmiä on käytetty erityisesti metsäpaloennusteissa (Safi & Bouroumi 2013, Castelli ym. 2015, Satir ym. 2016, Pham ym. 2020), vasteaikojen optimoinnissa (Photis & Grekousis 2012, Kiran ym. 2018, Aldabbas ym. 2018, Grekousis & Liu 2019) sekä liikenneonnettomuuksien ennakoinnissa (Awad, & Janson 1998, Mabrouk 2016, Dutta & Singhal 2019). Ennakoivan analytiikan teho perustuu kuitenkin tiedon määrään, joten siitä saatava hyöty on riippuvainen tiedon saatavuudesta. Aiemmissa tutkimuksissa suurin haaste ja rajoittava tekijä on ollut tiedon saatavuus, erityisesti jos aineistoissa hyödynnetään sosioekonomisia tekijöitä.

Ennakoivan analytiikan ja algoritmipohjaisten riskianalyysimallien käyttö rakennuspalojen ennustamisessa on yleistynyt kansainvälisesti pelastusalla viimeisen viiden vuoden aikana. Erilaisia kertaluontoisia kokeiluja on tehty ympäri maailmaa ja esimerkiksi USA:n pelastustoimessa on myös otettu käyttöön ohjelmistoja, joihin on integroitu tekoälyä hyödyntäviä mallinnustekniikoita ja tietovarantoja mahdollistaen reaaliaikaisen riskianalyysin tuottamisen pelastuslaitosten päivittäisen suunnittelun tueksi.

2.2 Ulkomailla käytössä olevat ennakoivaa analytiikka hyödyntävät alustat

Paljon kansainvälistä huomiota saanut New Yorkissa käyttöön otettu **Firecast** on ennakoivaa analytiikkaa ja suurta tietomassaa hyödyntävä algoritmi, joka mallintaa 330 000 asuinrakennusta ja liiketilaa päivittäin ja luokittelee rakennukset paloriskin perusteella. Algoritmi on integroitu osaksi Risk-Based Inspection System (RBIS) nimistä ohjelmistoa, joka hyödyntää Oracle -tietokantaa ja sen tiedonlouhintaominaisuuksia (Heaton 2015). RBIS ja Firecast on New Yorkin palolaitoksen (FDNY) kehittämä.

Ohjelmisto on yhdistetty tietovarantoon, johon paikalliset tiedonkerääjätahot päivittävät tietoja rakennuksista, onnettomuuksista ja sosioekonomista tietoa väestöstä. Algoritmia on kehitetty vuodesta 2010 ja siitä on vuosien varrella tehty useampia eri versioita. Yksi ohjelmakehityksen suurimpia ja aikaa eniten vieviä haasteita on ollut tiedon saamisen vaativuus.

Firecast 2.0 hyödynsi viiden eri tiedonkerääjätahon tietokantoja, joista oli poimittu 60 eri rakennuspaloja selittävää tekijää analyysimallia varten. Firecastin versio 3.0 hyödyntää tietoa 17 eri tiedonkerääjätaholta ja analyysimallia varten on poimittu 7500 eri riskitekijää (Roman, 2014). Algoritmin luokittelujen perusteella ohjelma luo New Yorkin palotarkastajille listan, jossa on priorisoitu korkean paloriskiarvon saaneet rakennukset. Ohjelman käyttöönoton jälkeen New Yorkin palotarkastuksia on kohdennettu jopa 20 prosenttia tarkemmin kuin ennen (Apolitical 2017).

Firecare on 2017 USA:ssa käyttöön otettu verkossa toimiva ohjelmisto, jonka avulla pelastuslaitokset voivat laskea alueellisen paloriskiarvon yksikön toiminta-alueella. Pelastuslaitokset voivat syöttää ohjelmistoon paikkatietoa alueellisista onnettomuuksista, rakennuksista ja tietoa väestöstä. Firecare laskee aineistojen perusteella kolme keskeistä turvallisuuteen liittyvää luokitusta; yhteisöjen yleinen rakenteellinen paloturvallisuusriski, paloasemien suorituskyky suhteutettuna USA:n pelastustoimen vaatimuksiin ja resurssien riittävyys riskiennusteisiin nähden. (Treibitz 2017)

Firecare -ohjelmisto on kehitetty usean eri organisaation yhteistyönä, ja työn on rahoittanut Federal Emergency Management Agencyn (FEMA)

Assistance to Firefighter -rahasto. Mukana työkalua kehittämässä on ollut muun muassa FEMA, Commission on Fire Accreditation International, the International Association of Fire Fighters, the International Association of Fire Chiefs, the Metropolitan Fire Chiefs, Underwriters Laboratory, the National Institute of Standards and Technology, Urban Institute ja University of Texas at Austin. USA:n paloasemat saavat käyttää ohjelmistoa veloituksetta. (International Public Safety Data Institute, 2022)

Firebird -viitekehys kehitettiin Atlantan pelastustoimen palotarkastustyön priorisoimisen tueksi. Se on kolmiosainen avoin lähdekoodi, joka on kehitetty liiketilakohtaisen paloriskiarvion mallintamiseen käyttämällä koneoppimisen mallinnustekniikoita. Lähdekoodin ensimmäinen osa liittää tiedot yhteen, toinen osa luokittelee tiedon ja mallintaa paloriskin ja kolmas osa visualisoi tulokset. Lähdekoodi, sen käyttöön liittyvät ohjeet ja testikäyttöön soveltuva materiaali ovat avoimesti käytettävissä GitHub verkkosivustolla (https://github.com/dssg-firebird).

Mallin testaamiseen käytettiin alun perin 252 rakennuspaloja selittävää tekijää. Lopulliseen malliin valittiin 58 muuttujaa sen perusteella, kuinka paljon ne vaikuttivat paloriskiin mallien testivaiheessa. Lähdekoodin testaamiseen käytettiin aineistoja, joissa on tietoja 8223 liiketilasta. Opetusaineistoon sisältyi Googlen API paikkatietoa, ARFD:n onnettomuustietoja, tietoa liiketilojen tulenkäsittelyluvista, anniskeluluvista, tietoa rakennusten kunnosta, sekä alueen väestöstä ja rikollisuudesta. Testien perusteella nk. tukivektorikone eli SVM ja satunnaismetsä -algoritmit suoriutuvat parhaiten, ja pystyivät ennustamaan noin 70 prosenttia tapahtuneista rakennuspaloista mallien testivaiheessa. Tuloksia käytettiin Atlantan tulipaloihin erikoistuneen pelastuslaitoksen ARFD:n palotarkastusten priorisointiin. (Madaio ym. 2016)

Firebird on alun perin kehitetty liiketilojen mallintamiseen, mutta lähdekoodia on myöhemmin hyödynnetty eri tarkoituksiin. Esimerkiksi Los Angelesissa tehtiin vertaisarvioimaton kokeilu, jossa Firebird lähdekoodia käyttämällä mallinnettiin Louisianan osavaltion pääkaupunki Baton Rougen asuinrakennukset ja liiketilat avoimella datalla (Jay 2017). Myöhemmin myös Pennsylvaniassa kokeiltiin lähdekoodia asuinrakennusten ja liiketilojen mallintamiseen (Singh ym. 2018). Molemmissa kokeilussa testattiin myös algoritmeja, jotka eivät sisälly alkuperäiseen lähdekoodiin.

Tapaustutkimukset

Ympäri maailmaa on tehty kertaluontoisia tapaustutkimuksia, joissa erilaisia algoritmipohjaisia ennakoivaa analytiikkaa hyödyntäviä riskianalyysimalleja on kokeiltu sekä asuinrakennusten että liiketilojen paloriskin kartoittamiseksi. Ulkomaisissa tapaustutkimuksissa rakennusten paloriskien mallintamiseen on yleisimmin käytetty satunnaismetsä- ja tukivektorikonealgoritmeja, neuroverkkoja, monikerroksista perseptroniverkkoa, sekä erilaisia aggregointi algoritmeja kuten Adaptive Boosting ja Extreme Gradient Boosting -algoritmeja (Taulukko 1). Tapaustutkimuksissa käytettyjä muuttujia on eritelty tämän raportin kohdassa 3.2.1.

Taulukko 1. Kooste ulkomaisista tapaustutkimuksista, joissa ohjatun koneoppimisen algoritmeja on kokeiltu rakennuspalojen ennakoimiseen. Algoritmien lyhenteet: satunnaismetsä (RF), tukivektorikone (SVM), Gradient-Boosted Trees (GBT) Logistic Regression (LR), Ada Boost (AB), XG Boost (XGB), Artificial Neural Network (ANN), decision trees (DT), Naive Bayesian (NB), monikerroksinen perseptroniverkko (MLP).

Tapaustutkimus	Aineisto	Käytetyt mallit ja mallien tarkkuus (AUC)
USA, Atlanta 2016 (Madaio ym. 2016)	8223 liikehuoneistoa, 58 muuttujaa	RF (0.76–0.82), SVM (0.76–0.81)
USA, Los Angeles 2017 * (Jay 2017)	120 000 asuinrakennusta ja liiketilaa, 12 muuttujaa (avoin data)	RF (0.81), GBT (0.79)
USA, Pennsylvania 2018 (Singh ym. 2018)	20 636 liiketilaa ja 350 korttelia, 226 muuttujaa	LR (0.53), AB (0.63–0.72), RF (0.70–0.74), XGB (0.73–0.75)
Etelä-Korea, Seoul 2018 (Hong & Jeong 2018)	7000 rakennusta, 33 muuttujaa	ANN (0.61–0.96), DT (0.62–0.64), SVM (0.63–0.68), NB (0.62-0.63)
Englanti, Lancaster 2019 (Dang ym. 2019)	54 680 liiketilaa, 69 muuttujaa	AB (0.88–0.91), XGB (0.89–0.93), RF (0.82–0.99), MLP (0.78–0.88)

Tapaustutkimus	Aineisto	Käytetyt mallit ja mallien tarkkuus (AUC)
Kiina, Zhengzhou, 2019 (Wang ym. 2019)	Zhengzhoun alue 1x1 km riskiruutuina	LASSO (0.45), CRF (0.63), LR (0.66), LR-P (0.57), SVM (0.56), SVM-P (0.63), GRU-LR (0.67), GRU-CRF (0.67), DeepST (0.74), Neuro-Fire (0.76)
Portugal, Lisbon, 2020 (Teixeira 2020)	Lissabonin kaupunki jaettuna 3662 osa-alueeseen, 39 muuttujaa	RF(0.68), DT(0.54),LR(0.63)

*Ei vertaisarvioitu

3. Laskennallinen viitekehys

Laskennallisella viitekehyksellä tarkoitetaan tässä yhteydessä yhdistelmää algoritmeja ja ohjelmistokoodeja, jonka avulla pystytään tuottamaan rakennuspalojen riskiarvioita. Riskiarvioiden muodostaminen koostuu monista eri vaiheista, joten mallinnusprosessia varten luodaan viitekehys, jota voidaan hyödyntää jatkossa toimivaksi havaittujen menetelmien soveltamisessa.

Hankkeen alussa kokeiltiin Atlantassa kehitettyä Firebird viitekehystä, jonka lähdekoodi on avoimesti saatavilla GitHubissa (https://github.com/dssg-firebird). Ohjelmistokoodit ajettiin sen yhteydessä tarjottujen testiaineistojen avulla. Kokeilun perusteella tultiin siihen tulokseen, että Suomessa saatavilla olevat aineistot ovat siinä määrin erilaisia verrattuna Atlantassa käytettyihin aineistoihin, on parempi luoda vastaavanlainen viitekehys, mutta soveltaa sitä paikallisiin aineistoihin sopivaksi. Firebirdin mallinnuksessa käytetään satunnaismetsä ja tukivektorikone -algoritmeja. Mallinnuksen algoritmit valitaan käyttötarkoituksen ja mallinnettavien aineistojen perusteella, joten paikallisten aineistojen mallintamiseen tulee kokeilla useampaa eri algoritmia.

Tässä tutkimuksessa kehitetty viitekehys on samankaltainen kuin Atlantan Firebird. Se koostuu aineistojen siivoamisesta, standardisoinnista, yhdistämisestä opetusaineistoksi, algoritmien koulutuksesta opetusaineistolla ja rakennuspalojen riskiarvion muodostamisesta. Tässä tutkimuksessa riskiarvioita ei visualisoitu toisin kuin Atlantassa, koska tässä hankkeessa ensisijainen tavoite oli testata menetelmien toimivuutta ja rakennuspalojen riskiarvioiden tarkkuutta. Tulosten visualisointi tulee tehdä pelastuslaitosten tarpeiden mukaan ja se on ajankohtaista vasta, jos menetelmät otetaan käyttöön kentällä. Tarkkoja alueellisia riskiarvioita ei myöskään eettisistä ja tietosuojasyistä julkaista.

3.1 Tarvittavan tiedon kartoitus

Laskennallisen viitekehyksen suunnittelu aloitettiin tarvittavan tiedon kartoituksella. Ennen tiedonkeruuta tehtiin selvitys muuttujista, joita on käytetty aiemmissa tutkimuksissa rakennusten paloriskin mallintamisessa (taulukko 2). Taulukon muuttujat on kuvattu englanniksi, koska osalle muuttujista ei ole suoria käännöksiä ja alkuperäiset termit kuvaavat muuttujia

parhaiten. Aiemmassa tutkimuksessa rakennuspalojen riskiarvioiden mallinnuksessa on käytetty tietoa asuinrakennuksista, liiketiloista, rakennuksia käyttävistä ihmisistä ja yrityksistä, alueista ja niiden demografiasta, ympäristöstä, häiriötekijöistä ja tietoa pelastuslaitosten valvontarekistereistä.

Taulukko 2. Aiemmissa rakennuspalojen riskiarvioita käsittelevissä tutkimuksessa käytettyjä muuttujia.

	Tietokanta	Muuttujat	Lähde
Atlanta 2016 (Madaio ym. 2016)	Fire incidents	Incident ID, call type, address, propery type, number of building, building occupation status, coordinates (+SIC count data; number of fires for the business type)	Atlanta Fire Department
	Fire Inspection Permits	Address, occupancy type, inspection date, permit issue data, number of inspections in last 5 years, percentage of inspection for the business type, coordinates (+SIC count data; percentage of business types inspected)	
	Liquor licenses	Liquor ID, company name, business name and type, address	
	Parcel Data	Structure condition (e.g. fair, deteriorated)	City of Atlanta
	Business Licenses (Atlanta)	Business type, license code, more specific business type and associated code, coordinates	
	Community Investment Report	N/A	
	List of properties that have not been inspected	Property type, address, coordinates, contact details	City of Atlanta and Google API's
	Neighborhood Planning Unit	Polygon dataset of the neighbourhood areas	Atlanta Regional Commission
	Demographic Data	Household number, population by race and age	U.S. Census Bureau
	Socio-economic Data	Household median income	
	CoStar Property Report	Commercial property information	CoStar Group, Inc
	Business Location Data	Property type, address, coordinates, contact details	Google API's

	Tietokanta	Muuttujat	Lähde
Atlanta 2016 (Madaio ym. 2016)	Fire incidents	Incident ID, call type, address, propery type, number of building, building occupation status, coordinates (+SIC count data; number of fires for the business type)	Atlanta Fire Department
	Fire Inspection Permits	Address, occupancy type, inspection date, permit issue data, number of inspections in last 5 years, percentage of inspection for the business type, coordinates (+SIC count data; percentage of business types inspected)	
	Liquor licenses	Liquor ID, company name, business name and type, address	
	Parcel Data	Structure condition (e.g. fair, deteriorated)	City of Atlanta
	Business Licenses (Atlanta)	Business type, license code, more specific business type and associated code, coordinates	
	Community Investment Report	N/A	
	List of properties that have not been inspected	Property type, address, coordinates, contact details	City of Atlanta and Google API's
	Neighborhood Planning Unit	Polygon dataset of the neighbourhood areas	Atlanta Regional Commission
	Demographic Data	Household number, population by race and age	U.S. Census Bureau
	Socio-economic Data	Household median income	
	CoStar Property Report	Commercial property information	CoStar Group, Inc
	Business Location Data	Property type, address, coordinates, contact details	Google API's
	Street Address Listing	Address, coordinates	Baton Rouge open data portal
	U.S. Census block group land area (derived from lat/long)	Census block value	

	Tietokanta	Muuttujat	Lähde
Atlanta 2016 (Madaio ym. 2016)	Combined property information	Lot acreage, businesses: any registered at lot (yes/no), current use category, future use category, design level category, no. of lot database entries at address, no. of addresses at same lot entry	Baton Rouge open data portal
	Tax roll 2016	Fair market value, value of improvements, no. of tax parcel entries at address	
	Public safety open data of crimes in LA	1+ crime occurrence 2011 (yes/no), 1+ crime occurrence 2012-2015 (yes/no), 1+ crime occurrence 2012-2015 (yes/no)	
	Public safety open data of fire incidents in LA	1+ fire occurrence 2012-2015, 1+ fire occurrence 2016	
	Population density, residential building density, education, income, employment, family structure, household tenure	Population density per km2, residential buildings per km2, information about education, median annual income, information about employment, proportion of households with children/only adults/at least one retired resident, proportion of households living in owner-occupied dwellings/rented and household right of occupancy dwellings	Tilastokeskus (2019)
	Building year and household crowdness (Tietoa Rakennuksista)	Construction years, average floor area per resident in m2, average number of rooms per resident	YKR (2019)
	Fire incidents	27 features (not specified)	Bureau of Fire
	Violations	20 features (not specified)	
	Parcels	46 features (not specified)	OPA
	Property data	80 features (not specified)	OPA
	Tax Liens	11 features (not specified)	DCR
	American Community Survey	42 features (not specified)	Census Bureau

	Tietokanta	Muuttujat	Lähde
	Fire incident information	Type of response, duration of response, outcome of response, scale of response, location of response, identities and response order of units, identities of regularly assigned units	N/A
	Elevation above sea level to response location	N/A	
	Fire code violations	N/A	
	311 service requests	N/A	
	Department of Building data	Certificate of occupancy, presence and nature of violations, and any open work permits assigned to the particular occupancy.	
	Weather	N/A	National Oceanic and Atmospheric Administration
	Property information	Total number of floors, number of ground floor, architectural style, number of underground floors, building pillar form, building Roof Type, total area of the building, year of construction, number of dangerous goods facilities, Number of electrical facilities, Number of fire facilities, Number of gas facilities, Building main use (Apartment)	
	Fire safety information	Combustion expansion prevention (fire protection compartment), Fire History, automatic fire detection system, Automatic fire alarm system, Emergency alarm system, Emergency broadcasting facility, Other digestion facilities _ Connection water pipe facility, Other digestion facilities _ Connection sprinkling facilities, Other digestion facilities _ Emergency outlet facilities, Other firefighting equipment _ Wireless communication auxiliary equipment, Other digestion facilities _ Fire hydrants Evacuation facility Affiliate ventilation	
	Fire safety information	Water-based extinguishing equipment _ Manual, Water-based extinguishing equipment_ Automatic, Fire extinguisher, Gas system fire extinguishing equipment	

*Ei vertaisarvioitu julkaisu

Ennen tiedon keruuta myös hankkeen ohjausryhmää konsultointiin eri muuttujien tarpeellisuudesta rakennuspaloriskimallien testaamista varten. Tätä varten tehtiin Webropol-kysely, johon koottiin muuttujia eri tahojen olemassa olevan datan metatietojen perusteella. Kyselyyn valitut tekijät

valittiin sillä perusteella, mitä tekijöitä aiemmissa tutkimuksissa on käytetty ja mitä tietoa on saatavilla. Kysely sisälsi 165 eri muuttujaa seitsemästä eri lähteestä (taulukko 3).

Taulukko 3. Konsultaatiossa arvioitujen muuttajien lähde ja määrä.

Lähde	Kategoria	Selittäjätekijöiden määrä
Tilastokeskus	Auskastiedot	55*
Findata	Asukastiedot	46*
Digi- ja väestötieto-virasto	Rakennus- ja huoneisto-tiedot	45
Pelastuslaitokset	Pelastuslaitosten tiedot	15
PRONTO	Onnettomuustiedot	12
Avoin data	Asuinalue	3
Poliisihallitus/hätä-keskuslaitos	Järjestyshäiriöt	3

**päällekkäisiä muuttujia*

Tutkimuksen tarkoituksena on saada kaikki kyselyssä mainitut muuttujat mallin testaamista varten, mutta tiedon saamisen haastavuuden takia kyselyllä pyrittiin kartoittamaan eri muuttujien olennaisuutta ja priorisoimaan haettavien tietojen tarvetta. Kyselyn lisäksi ohjausryhmän jäsenten avulla pyrittiin kartoittamaan kaikki mahdolliset tahot, jotka tuottavat tietoa onnettomuuksista, rakennuksista tai sosioekonomisista tekijöistä valtakunnallisella tasolla.

Muuttujien painoarvoa kysyttiin muodossa ”Asteikolla 1–10, kuinka todennäköisesti seuraavilla muuttujilla on joko välitön tai välillinen vaikutus rakennuspalon syntyyn? 0= Ei todennäköistä, 10= Erittäin todennäköistä”. Kyselyn perusteella 10 merkittävintä muuttujaa sisältää:

- rakennukseen on tehty ilmoitus palovaarasta tai muusta onnettomuusriskistä
- asukkaan ikä
- toimintarajoitteisuuteen liittyvä tuki (asuntokunta)
- rakennuksen käyttötarkoitus
- toimintakelpoisuus
- huumausainerikokset
- vanhuuden turva (asuntokunta)

- taloudessa asukas työttömyyseläkkeellä
- asukkaalla työkyvyttömyyden lisäsairaus
- työkyvytön asukas

3.2 Tiedon keräys

Tiedon keräyksen ja hallinnoinnin tueksi laadittiin tietosuojaseloste ja aineistohallintasuunnitelma. Koska usea eri taho kysyi samankaltaisia lisäkysymyksiä hankkeen tietosuojaan ja tiedon tarpeellisuuteen liittyen, tutkimuslupahakemusten tueksi luotiin myös tiedosto, johon koottiin vastauksia usein kysyttyihin kysymyksiin.

Aineistoja haettiin seuraavilta tahoilta:

- **Pelastustoimen toimenpide- ja resurssirekisteri PRONTO** (Käyttöoikeus kaikkiin PRONTOn aineistoihin)
- **6 Pelastuslaitosten tiedot** (valvontarekisterin tiedot poistumisturvallisuusselvityksistä, palotarkastuksista, pelastuslain 42 §:n mukaisista paloriski-ilmoituksista, kiinteistöjen palontorjuntatekniikasta sekä muusta kiinteistökohtaisesta, strukturoidusta tiedosta koskien kiinteistön paloturvallisuutta. Pelastuslaitoksiin kuuluvat Helsingin, Pohjanmaan, Etelä-Karjalan sekä Itä-, Länsi-, ja Keski-Uudenmaan pelastuslaitokset)
- **Digi- ja väestötietovirasto** (rakennus- ja huoneistorekisteri)
- **Hätäkeskuslaitos ja Poliisihallitus** (alueelliset melu- ja häiriöilmoitukset)
- **Avoin data** (Väestötietoruudukko, mutta vain pääkaupunkiseudun alueella)
- **Findata** (KELAn, THL:n ja Eläketurvakeskuksen sosioekonomiset aineistot sekä DVV:n henkilötiedot)
- **Tilastokeskus** (FOLK-moduulit yhdistettynä sijaintitietoruudukkoon)

Näistä tahoista kaikki paitsi Findata myönsi hankkeelle haetuille aineistoille tutkimusluvan. Tutkimuslupa prosessit olivat monen tahon kohdalla pitkiä, monimutkaisia ja vaativat paljon resursseja. Tarkempi prosessi on kuvailtu tämän raportin liitteessä (LIITE 1).

Keskeisten aineistojen lisäksi kartoitettiin myös Turvallisuus- ja kemikaalivirasto Tukesin Kemu -rekisteri, Rakennusvalvonnan Facta-tietojärjestelmä,

Säteilyturvakeskuksen tietokannat ja Tilannehuoneen aineistot. Nämä eivät kuitenkaan osoittautuneet tutkimuksen kannalta olennaisiksi.

3.3 Mallinnusmenetelmien kokeileminen käytännössä

Pääkaupunkiseudun rakennuspalojen mallinnus

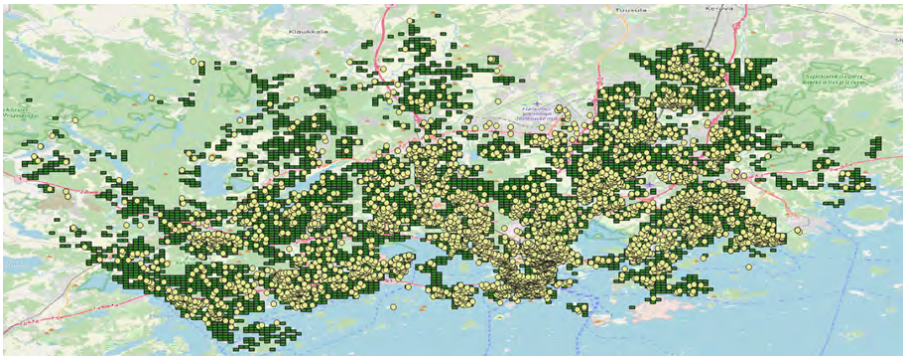
Tietopyyntöjen käsittelyaikana kokeiltiin eri ohjatun koneoppimisen algoritmeja rakennuspalojen riskiarvioiden muodostamiseen pääkaupunkiseudulla avointa dataa hyödyntämällä. Tapaustutkimuksessa käytettyihin aineistoihin kuuluvat Pelastustoimen toimenpide- ja resurssirekisterin PRONTOn tiedot, Helsinki Region Infosharen (HRI) osoitetiedot ja Helsingin seudun ympäristöpalveluiden (HSY) pääkaupunkiseudun väestötietoruudukko. Kokeilun alueeksi valittiin pääkaupunkiseutu, koska alueelle löytyi tarpeeksi tarkat rakennuskohtaiset koordinaatit sekä väestötietoa avoimista tietolähteistä. Kokeiluun käytettiin aineistoja vuosilta 2013–2019, koska väestötietoruudukossa sitä vanhemmat tiedot on kirjattu isomassa resoluutiossa kuin näinä vuosina.

Aineistot

PRONTOsta haettiin sekä rakennuspalot että rakennuspalovaarat vuosilta 2013-2019. Alueeksi rajattiin Helsinki, Espoo, Vantaa ja Kauniainen. Onnettomuustyypin, vuoden ja tapahtumakunnan lisäksi muuttujiin sisältyi pohjoiskoordinaatit, itäkoordinaatit ja osoite. Rajattu aineisto sisältää 5384 rakennuspaloa ja rakennuspalovaaraa. (Kuva 1)

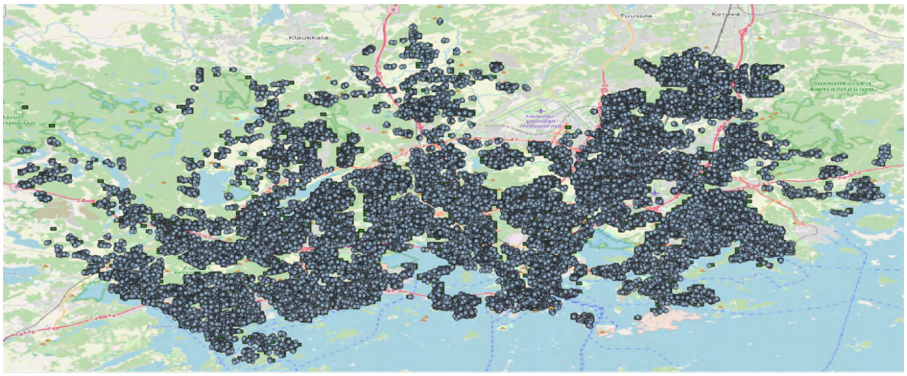
Opetusaineiston muuttujat muodostettiin Pääkaupunkiseudun väestötietoruudukosta, joka on Helsingin seudun ympäristöpalveluiden (HSY) ylläpitämä paikkatietoaineisto (Helsingin seudun ympäristöpalvelut, 2021). Aineiston muuttajat ilmaistaan ruuduittain, joiden koko on 250 m x 250 m vuodesta 2013 alkaen. Sitä vanhemmat tiedostot on ilmaistu 500 m x 500 m tarkkuudella. Muuttujiin sisältyy ruudun sisällä olevien rakennusten vakinaisten asukkaiden määrä, asumisväljyys (huoneistoala jaettuna asukaslukumäärällä) ja ikä, joka ilmaisee asukkaiden ikäjakauman, yhdeksään ikä-

luokkaan jaettuna. Ikäluokat ovat 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79- vuotiaat sekä 80 vuotta täyttäneet ja sitä vanhemmat. Aineistot ladattiin Shapefile-tiedostomuodossa vuosilta 2013–2019. (Kuva 1)



Kuva 1. Keltaiset pisteet kuvaavat pääkaupunkiseudulla tapahtuneita rakennuspaloja ja rakennuspalovaaroja vuosina 2013-2019. Taustalla oleva vihreä ruudukko on HSY:n väestötietoruudukko. Pohjakartta: OpenStreetMap®, lisenssi: Open Data Commons Open Database License (ODbL), OpenStreetMap Foundation (OSMF).

Helsinki Region Infosharen (HRI) sivulta ladattiin Pääkaupunkiseudun osoiteluettelo (Helsinki Region Infoshare, 2021), johon on kerätty tietoja Helsingin, Espoon, Vantaan ja Kauniaisten kaupunkien kiinteistöistä ja kaavayksiköiden osoitteista. HRI on avoimen datan sivusto, joka sisätää pääkaupunkiseudun kaupunkien julkisia tietovarantoja. Aineisto on julkaistu 2013 ja sitä päivitetään jatkuvasti. Aineisto ladattiin CSV -muodossa. Aineiston muuttujista suodatettiin: kadunnimi, osoitenumero, osoitenumero2, osoitekirjain, n (pohjoiskoordinaatit), e (itäkoordinaatit), ja kaupunki. Aineisto sisältää yhteensä 139 338 osoitetta. (Kuva 2)



Kuva 2. HRI pääkaupunkiseudun osoitetiedot visualisoituna. Pohjakartta: OpenStreetMap®, lisenssi: Open Data Commons Open Database License (ODbL), OpenStreetMap Foundation (OSMF).

Osoitetietojen ja koordinaattien standardisointi

PRONTO-aineistoissa koordinaatit on merkitty 7-numeroisilla luvuilla, KKJ3 -koordinaatiston mukaisesti. HRI:n osoitetiedoston pohjoiskoordinaatit (n) on merkitty 7-numeron tarkkuudella ja itäkoordinaatit (e) 8-numeron tarkkuudella. Sekä osoitetietojen että väestötietoruudukon koordinaattijärjestelmä on ETRS-GK25 (EPSG:3879). PRONTO-aineistossa koordinaatteja ei ole kirjattu standardisoidulla tavalla, esimerkiksi joissain tapauksissa samalle huoneistolle kirjatut koordinaatit eivät olleet identtisiä. Koska HRI:n koordinaatit vastaavat kaupunkien paikkatietopalvelusivustojen koordinaatteja, PRONTOn rakennuspalojen ja palovaarojen koordinaatit korjattiin vastaamaan osoitetietoaineiston koordinaatteja.

Ainoa yhdistävä tekijä PRONTOn ja HRIn aineistoissa on osoite, mutta se on kirjattu eri muodossa aineistojen välillä. Molempien aineistojen osoitetiedot standardisoitiin koordinaattien linkittämistä varten. Linkitys tehtiin kaupunkikohtaisesti, koska eri kaupungeissa on samoja osoitteita. Kaikille PRONTO tiedostossa oleville osoitteille ei löytynyt paria osoitetiedostosta. Tämä johtui osittain siitä, että PRONTOn osoitetietoihin oli myös kirjattu muita tietoja kuin osoitetietoja. Vaikka molemmissa lähteissä on myös huoneistokohtaisia osoitteita, rakennuskohtaisia pareja löytyi enemmän (4905), joten niitä käytettiin aineiston tarkkuutena.

Väestötietojen yhdistäminen rakennus- ja onnettomuustietoihin

HSY:n väestötietoruudukon ikäluokissa on ilmoitettu ikäluokkaan kuuluvien asukkaiden määrä. Aineistojen analyysiä varten ikäluokkien asukasmäärät muutettiin asukkaiden osuudeksi asukasmäärästä. Ruudut, joissa asuu alle 100 asukasta poistettiin, koska niihin ei ollut kirjattu ikätietoja tietosuojasyistä. Muuttujien keskenäistä korrelaatiota tutkittiin, koska korrelaatio saattaa vaikuttaa mallinnuksen virhemarginaaliin (Aliverti ym. 2018). Sitä varten käytettiin R-ohjelmistokielen Hmisc -laajennuspaketin (Harrel Jr 2019) *rcorr* argumenttia, joka luo korrelaatiomatriisin, jossa on sekä tutkittavien tekijöiden korrelaatiokerroin että niiden tilastollinen merkitsevyys (p-arvo). Muuttujien välillä ei löytynyt merkittäviä korrelaatioita.
Aineistojen analyysiä varten väestötiedot yhdistettiin osoitetietoihin koordinaattien perusteella. PRONTO -aineistoihin muuttujat haettiin vuosi kerrallaan niin, että väestötietoruudukon (VTR) vuosi vastasi rakennuspalon tai palovaaran tapahtumavuotta. Muihin osoitteisiin väestötiedot haettiin vuoden 2019 VTR -tiedostosta. Yhteensä 101 908 rakennusta sekä 3173 rakennuspaloa tai palovaaraa osui VTR -ruudukkoon. Lopuksi aineistot yhdistettiin ja luotiin ylimääräinen sarake, biäärimuuttuja osoittamaan aiemmin tapahtuneita onnettomuuksia. Sarakkeessa kohteiden, joissa on tapahtunut rakennuspalo tai rakennuspalovaara vastasi numeroa 1 ja loput kohteet numeroa 0.

Käytetyt mallit ja mallien koulutus

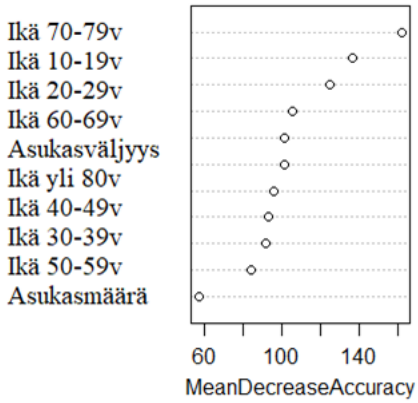
Testejä varten valittiin kaksi rakennuspalojen ennustamisessa yleisimmin käytettyä algoritmia, satunnaismetsä ja tukivektorikone, sekä yleistettyihin lineaarisiin malleihin kuuluva logistinen regressio. Satunnaismetsä on päätöspuumalli, mikä on joukko-oppimista soveltava metodi, jossa yksinkertaisena algoritmikomponenttina on päätöksentekopuu (Telivuo 2018). Päätöksentekopuita on useita ja ne luokittelevat aineiston muuttujien avulla, oksien haaroissa olevien kysymysten perusteella. Tukivektorikone on lineaarinen luokitinmalli, mikä yksinkertaistettuna jakaa tiedon eri luokkiin ja tekee päätelmiä niiden etäisyyksistä marginaalitasoihin (Noble 2016). Välimatkojen etäisyyksiä rajoittavia tekijöitä kutsutaan tukivektoreiksi, jotka määrittävät lopullisen luokittelun. Logistinen regressio mallintaa vastemuuttujan ja selittävien muuttujien suhdetta arvioimalla todennäköisyydet logit-funktion avulla (Kleinbaum ym. 2020).

Mallinnuksissa käytettiin RStudiota, joka on tilastollisen laskennan ohjelmisto, jossa hyödynnetään R-ohjelmointikieltä. Satunnaismetsä -kokeilussa malli koulutettiin *caTools* -laajennuspaketin (Tuszynski & Khachatryan 2013) ohjelmistokoodeilla. Ennustuksessa käytettiin *random forest* -laajennuspakettia (RColorBrewer & Liaw 2018) ja mallin suorituskyvyn arviointiin käytettiin *ROCR*-laajennuspakettia (Sing ym. 2007). Tukivektorikonekokeiluissa hyödynnettiin sekä *Caret* (Kuhn ym. 2020) että *kernlab* (Karatzoglou ym. 2019) -laajennuspaketteja. Logistinen regressiomalli koulutettiin ja arvioitiin *dismo* -laajennuspaketin ohjelmistokoodeilla (Hijmans ym. 2017). Mallien luotettavuuden arvioimiseksi käytettiin 10-kertaista ristivalidointimenetelmää, jossa 80 % aineistosta käytettiin mallin koulutukseen ja 20 % mallin testaamiseen.

Tulokset

Mallien suorituskykyä arvioitiin käyttämällä ROC-käyrän AUC arvoa. ROC-käyrässä tutkitaan mallin oikea- ja väärä-positiivisuusasteiden arvoja (Provost & Fawcett, 1998). Käyrän alle jäävää pinta-alaa kutsutaan AUC:ksi ja se kuvaa luokittelijan tarkkuutta asteikolla 0.5-1. Mitä korkeampi arvo, sitä tarkempi ennustus. Lisäksi, satunnaismetsäalgoritmin arvioinnissa otettiin huomioon OOB error -arvo. OOB (out-of-bag) -error laskee päätöspuiden ennustusvirheitä käyttämällä bootstrap-aggregaatiota (Breiman 1996). Parhaiten malleista suoriutui satunnaismetsäalgoritmi, jonka AUC arvo oli 0,91 (OOB 1,9%). Toiseksi parhaiten suoriutui tukivektorikone (AUC 0,82) ja kolmanneksi logistinen regressiomalli (AUC 0,79).

Muuttujien painoarvoa arvioitiin kuvaamalla muuttujien keskimääräistä laskutarkkuutta (Mean Decrease Accuracy). MDA tutkii miten paljon tarkkuutta malli menettää, jos yksittäisen muuttujan ottaa pois selittäjätekijöiden joukosta. Tulosten perusteella ikäluokat 70-79, 10-19 ja 20-29 vuotta olivat merkittävimpiä muuttujia. Vähinten painoarvoa oli asukasmäärällä. (Kuva 3)



Kuva 3. Muuttujien painoarvo mallinnuksessa satunnaismetsäalgoritmiä käyttämällä.

Kokeilun yhteenveto ja johtopäätökset

Tulosten perusteella rakennuspalojen riskiarvioissa satunnaismetsäalgoritmi soveltuu parhaiten saatavilla olevan tietokokonaisuuden analysoimiseen. Tässä tapaustutkimuksessa mallinnetut riskiarvot ovat rakennuskohtaisia, mutta ennustuksien todellinen maantieteellinen tarkkuus on 250m^2. Tätä tarkempia paikkatietoanalyysiin soveltuvia muuttujia ei ollut saatavilla avoimista tietolähteistä. Kokeillut menetelmät tarjoavat kuitenkin tarkempia riskiarvioita, kuin parhaillaan Pelastustoimessa käytössä oleva regressiomalli.

Tutkittaessa eri muuttujien painoarvoa malleissa, ikäluokalla 70-79v oli suurin vaikutus rakennuspalojen esiintyvyyttä kuvaavaan riskiarvioon. Tulee kuitenkin muistaa, että malli tekee riskiarviot vain annettujen muuttujien perusteella eli sen luotettavuus todellisuudessa riippuu siitä, miten hyvin valitut tekijät kuvaavat tutkittua ilmiötä. Mallinnuksia varten tulisi testata useita eri muuttujia, joista voidaan valita keskeisimmät laajempaa riskianalyysiä varten.

Mallien tuloksia tulkitessa tulee myös ottaa huomioon, että kyseessä on rakennuspalojen määrän suhteen pieni otos. Pienien aineistojen mallintamisessa on suurempi ylisovittumisen riski. Ylisovittuminen tarkoittaa sitä, että algoritmi kehittää sellaisen luokittelijan, joka sopii täydellisesti käytettyyn opetusjoukkoon ja opettelee ulkoa opetusalkioiden suhteen kohdemuuttujaan. (Krzywinski & Altman 2016)

Koko Suomen rakennuspaloriskien mallinnus

Syksyyn 2021 mennessä tutkimusta varten saatiin tarpeeksi aineistoja, jotta pystyttiin toteuttamaan laajempi rakennuspalojen riskiarvioiden mallinnus ennakoivan analytiikan menetelmiä hyödyntämällä. Mallinnuksen tavoitteena on selvittää, mitkä analyysimenetelmät soveltuvat parhaiten rakennusten tulipaloriskin mallintamiseen ja mitkä rakennuspaloja selittävistä tekijöistä ovat keskeisimpiä rakennusten tulipaloriskin kannalta. Tutkimusta varten saadut aineistot olivat koko Suomen kattavia, joten tutkimusalueeksi rajattiin koko Suomi, mukaan lukien Ahvenanmaa.

Aineistot ja aineistojen käsittely

Tilastokeskuksen teknisten haasteiden vuoksi kaikkia ulkopuolisia aineistoja ei saatu vietyä etäkäyttöympäristöön hankkeen aikana. Koska Tilastokeskukselle toimitetuista aineistoista vietiin vain murto-osa tiedostojen riveistä, ei suurinta osaa aineistoista pystytty analysoimaan. Marras/joulukuussa 2021 analysoitiin opetusaineisto, joka muodostettiin Tilastokeskuksen FOLK-moduuleista ja PRONTOn rakennuspalo ja rakennuspalovaara -tietoriveistä, jotka olivat käytettävissä etäkäyttöympäristössä (14 868 rakennuspaloa/palovaaraa vuosilta 2008–2020).

Tilastokeskuksen aineistot

Tilastokeskuksen aineistosta poimittiin tilastovuosien 2008–2020 tiedot. Aineistojen perusteella muodostettiin binäärisiä muuttujia, jotka liitettiin myöhemmässä vaiheessa 250 m^2 paikkatietoruudukkoon. Muuttujista muodostettiin arvot ruudukkoihin sen perusteella, mikä niiden osuus on ruudussa olevasta asukasmäärästä. Analyysissä ruutuja vertaillaan keskenään rakennuspalojen esiintyvyyden ja muuttujien osuuksien perusteella.

1. FOLK-henkilöaineiston asuinliitot-moduuli (FOLK-Asuinliitot)

Aineistosta valittiin alla olevassa taulukossa kuvatut muuttujat. Muuttujien tarkemmat kuvaukset löytyvät Tilastokeskuksen aineistokatalogista.

Taulukko 4. Taulukon muuttujat on poimittu Tilastokeskuksen FOLK-henkilöaineiston asuinliitot-moduulista. (Tilastokeskus, 2021)

Muuttuja	Kuvaus (Tilastokeskus)
Tilastovuosi	Aineistovuosi
Suojattu TK:n henkilönumero	Suojattu TK:n henkilönumero, joka on sama koko FOLK-aineistossa.
Erilleen muuttopäivä	Erilleen muuttopäivätiedon lähteenä on henkilötilastojen tietovaraston muuttotilastoaineisto.
Avioliiton solmimispäivä	Vihkipäivä-tiedot Väestörekisterikeskuksen väestötietojärjestelmästä.
Asumuseropäivä	Asumuseropäivä-tiedot Väestörekisterikeskuksen väestötietojärjestelmästä
Avioeropäivä	Avioeropäivä-tiedot Väestörekisterikeskuksen väestötietojärjestelmästä.
Puolison kuolinpäivä	Kuolinpäivä-tiedot Väestörekisterikeskuksen väestötietojärjestelmästä
Vanhuksen laitosasuminen	Asuinliittoparista jompikumpi puoliso asuu erilleen muuton jälkeen laitoksessa tai huoltolaitosrakennuksessa. Tieto koskee vain 70 vuotta täyttäneitä.

Muuttujien perusteella luotiin jokaiselle tutkittavalle tilastovuodelle (2008–2020) alla olevat yhdeksän binääristä muuttujaa.

1. Henkilön puoliso on kuollut tilastovuonna (laskettu puolison kuolinpäivän perusteella).
2. Henkilön puoliso on kuollut tilastovuotta edeltävänä vuotena (laskettu puolison kuolinpäivän perusteella).
3. Henkilö on muuttanut erilleen kumppanista tilastovuonna (laskettu erilleen muuttopäivän perusteella. Tiedostosta poistettu rivit, jossa henkilön tai henkilön puolison kuolinpäivä on samana vuonna).
4. Henkilö on muuttanut erilleen kumppanista tilastovuotta edeltävänä vuotena (laskettu erilleen muuttopäivän perusteella. Tiedostosta poistettu rivit, jossa henkilön tai henkilön puolison kuolinpäivä on samana vuonna).
5. Henkilö on avioeronnut tilastovuonna (laskettu avioeropäivän perusteella. Tiedostosta poistettu rivit, jossa henkilön tai henkilön puolison kuolinpäivä on samana vuonna).
6. Henkilö on avioeronnut tilastovuotta edeltävänä vuotena (laskettu avioeropäivän perusteella. Tiedostosta poistettu rivit, jossa henkilön

tai henkilön puolison kuolinpäivä on samana vuonna).

7. Puoliso siirtynyt laitokseen tilastovuonna (laskettu erilleen muuttopäivän ja vanhuksen laitosasumis muuttujan perusteella).
8. Puoliso siirtynyt laitokseen tilastovuotta edeltävänä vuotena (laskettu erilleen muuttopäivän ja vanhuksen laitosasumis muuttujan perusteella).
9. Henkilö on asuinliitossa tilastovuonna (laskettu erilleen muuttopäivän perusteella. Tiedostosta poistettu rivit, jossa henkilön tai henkilön puolison kuolinpäivä on samana vuonna).

Muuttujat luotiin vuosien perusteella erillisiin tiedostoihin, johon sisällytettiin muuttujat, tapahtumavuosi ja suojattu henkilönumero. Tiedostot yhdistetään myöhemmässä vaiheessa sijaintitietoihin suojatun henkilönumeron perusteella.

2. FOLK-henkilöaineiston asuntokuntamoduuli (FOLK-Asuntokunta)

Aineistosta valittiin alla olevassa taulukossa olevat muuttujat. Muuttujien tarkemmat kuvaukset löytyvät Tilastokeskuksen aineistokatalogista.

Taulukko 5. Taulukon muuttujat on poimittu Tilastokeskuksen FOLK-henkilöaineiston asuntokuntamoduulista. (Tilastokeskus, 2021)

Muuttuja	Kuvaus (Tilastokeskus)
Tilastovuosi	Aineistovuosi
Suojattu TK:n henkilönumero	Suojattu TK:n henkilönumero, joka on sama koko FOLK-aineistossa.
Asuntokunnan koko	Vakinaisesti asuvien henkilöiden lukumäärä.
Asuntokuntarakenne	0= laitosväestö ja yli 25 hengen asuntokunnat 1= perhe, ei muita 2= 1 perhe ja muita 3= 2 perhettä, ei lisäksi muita 4= 2 perhettä, lisäksi muita 5=vähintään 3 perhettä, ei lisäksi muita 6=vähintään 3 perhettä, lisäksi muita 7= ei perhettä, asuntokunnan koko=1

Muuttuja	Kuvaus (Tilastokeskus)
Asuntokuntarakenne	8= ei perhettä, asuntokunnan koko=2 joilla sama sukupuoli 9= ei perhettä, asuntokunnan koko=2 joilla eri sukupuoli 10= ei perhettä, asuntokunnan koko >2; kaikilla sama sukupuoli 11= ei perhettä, asuntokunnan koko >2; joilla eri sukupuoli
Asumisväljyys	1=Tilava. 1-5 henkilön asuntokunta, jonka käytössä on vähintään kolme asuinhuonetta enemmän kuin asuntokunnan henkilömäärä (keittiötä ei lueta huonelukuun) 2=Normaali 3=Ahdas. Enemmän kuin yksi henkilö huonetta kohti (keittiötä ei lueta huonelukuun) 4=Tuntematon
Autojen lukumäärä	Asuntokunnille summataan asuntokunnan hallinnassa tai omistuksessa olevat autot.
Sairausajan toimeentuloturva	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset sairauteen liittyvät etuudet.
Toimintarajoitteisuuteen liittyvä tuki	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset työkyvyttömyys ja vammaisetuudet.
Vanhuuden turva	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset vanhuuden turvaan luettavat etuudet.
Lesken ja muiden omaisten tuki	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset lesken ja orpojen lasten tuet.
Lapsiperheiden etuudet	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset lapsiperheiden tuet.
Työttömyysturva	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset työttömyyteen liittyvät etuudet.
Asumisen tuet	Asuntokunnan jäsenten yhteenlasketut ESSPROS-luokituksen mukaiset asumisen tuet.
Opintoetuudet	Asuntokunnan jäsenten yhteenlasketut opintoetuudet.

Muuttuja	Kuvaus (Tilastokeskus)
Muu toimeentuloturva	Asuntokunnan jäsenten yhteenlasketut muut toimeentuloturvaetuudet ESSPROS-luokituksen mukaan.
Käytettävissä oleva rahatulo	Asuntokunnan jäsenten yhteenlasketut käytettävissä oleva rahatulot.
Velat yhteensä	Asuntokunnan jäsenten yhteenlasketut velat.

Aineistojen perusteella muodostettiin seuraavat binääriset muuttujat:

1. Asuntokunnan rakenne = 1 perhe, ei muita
2. Asuntokunnan rakenne = 1 perhe, lisäksi muita asukkaita
3. Asuntokunnan rakenne = 2 perhettä, ei lisäksi muita
4. Asuntokunnan rakenne = 2 perhettä, lisäksi muita asukkaita
5. Asuntokunnan rakenne = vähintään 3 perhettä, ei lisäksi muita
6. Asuntokunnan rakenne = vähintään 3 perhettä, lisäksi muita asukkaita
7. Asuntokunnan rakenne = ei perhettä, asuntokunnan koko 1 henkilö
8. Asuntokunnan rakenne = ei perhettä, asuntokunnan koko 2 henkilöä, joilla sama sukupuoli
9. Asuntokunnan rakenne = ei perhettä, asuntokunnan koko 2 henkilöä, joilla eri sukupuoli
10. Asuntokunnan rakenne = ei perhettä, asuntokunnan koko >2; kaikilla sama sukupuoli
11. Asuntokunnan rakenne = ei perhettä, asuntokunnan koko >2; joilla eri sukupuoli
12. Asumisväljyys = Tilava. 1–5 henkilön asuntokunta, jonka käytössä on vähintään kolme asuinhuonetta enemmän kuin asuntokunnan henkilömäärä (keittiötä ei lueta huonelukuun)
13. Asumisväljyys = Normaali (tilavan ja ahtaan välille sijoittuvat luvut)
14. Asumisväljyys = Ahdas. Enemmän kuin yksi henkilö huonetta kohti (keittiötä ei lueta huonelukuun)
15. Asuntokunnassa on auto
16. Asuntokunnassa on henkilö, joka saa sairausajan toimeentuloturvaa
17. Asuntokunnassa on henkilö, joka saa toimintarajoitteisuuteen liittyvää tukea
18. Asuntokunnassa on henkilö, joka saa vanhuuden turvaan liittyviä etuuksia
19. Asuntokunnassa on henkilö, joka saa lesken ja muiden omaisten tukia
20. Asuntokunnassa on henkilö, joka saa lapsiperheiden etuuksia
21. Asuntokunnassa on henkilö, joka saa työttömyysturvaa
22. Asuntokunnassa on henkilö, joka saa asumisen tukia
23. Asuntokunnassa on henkilö, joka saa opintoetuuksia

24. Asuntokunnassa on henkilö, joka saa muuta toimeentuloturvaetuuksia
25. Asuntokunnan jäsenten yhteenlasketut käytettävissä oleva rahatulo vähemmän kuin mediaani tulot -10 000 €
26. Asuntokunnan jäsenten yhteenlasketut käytettävissä oleva rahatulo mediaani ± 10 000 €
27. Asuntokunnan jäsenten yhteenlasketut käytettävissä oleva rahatulot enemmän kuin mediaani +10 000 €
28. Asuntokunnassa ei ole velkaa
29. Asuntokunnan jäsenten yhteenlasketut velat vähemmän kuin mediaani +10 000 €
30. Asuntokunnan jäsenten yhteenlasketut velat enemmän kuin mediaani +10 000 €

3. FOLK (jaksotieto) eläke (FOLK-eläkejaksot)

Aineistosta valittiin eläkelaji sekä tunnistetiedot (suojattu henkilönumero ja aikamuuttuja). Eläkelajit on muodostettu Tilastokeskuksessa alkuperäisen aineiston koodistosta seuraavan luokituksen mukaisiksi muuttujiksi:

1. Henkilö on vanhuuseläkkeellä
2. Henkilö on työkyvyttömyyseläkkeellä
3. Henkilö on yksilöllisellä varhaiseläkkeellä
4. Henkilö on työttömyyseläkkeellä
5. Henkilö on maatalouden erityiseläkkeellä
6. Henkilö on osa-aikaeläkkeellä
7. Henkilö on perhe-eläkkeellä
8. Henkilö on takuueläkkeellä

4. FOLK-henkilöaineiston perustietomoduuli (FOLK-Perus)

Aineistosta valittiin alla olevassa taulukossa olevat muuttujat. Muuttujien tarkemmat kuvaukset löytyvät Tilastokeskuksen aineistokatalogista.

Taulukko 6. Taulukon muuttujat on poimittu Tilastokeskuksen FOLK- henkilöaineiston perustietomoduulista. (Tilastokeskus, 2021)

Muuttuja	Kuvaus
Sukupuoli	Henkilön viimeisin sukupuolitieto.
Syntymävaltio	Henkilön syntymävaltio.
Kansalaisuus	Henkilön viimeisin kansalaisuustieto.
Syntyperä	Henkilön syntyperä
Kieli	Henkilön viimeisin kielitieto
Ikä	Henkilön ikä vuosissa vuoden lopussa.

Muuttuja	Kuvaus
Maaseutu-kaupunki-luokitus	SYKE:n aluetypologian mukainen maaseutu-kaupunki-luokitus.
Ammattiasema (TVM)	Ammattiasema päätyösuhteen mukaan
Koulutuslaji	Opiskelijan koulutuslaji
Pääasialllinen toiminta (TVM)	Pääasiallinen toiminta
Ammattikoodi, 3-nrotaso	Tieto henkilön ammatista on tuotettu 18-74-vuotiaille työllisille, siis palkansaajille ja yrittäjille.
Työttömyyskuukausia	Henkilön työttömyyskuukaudet -tieto vuosittain
Hallintaperuste	Asunnon hallintaperuste on pääsääntöisesti päätelty vakinaisesti asutuille asunnoille eli asuntokunnille
Huoneluku keittiö pl.	Asunnon huoneiden lukumäärä ilman keittiötä.
Talotyyppi	Talotyyppi. 1 Erillinen pientalo 2 Rivi- tai ketjutalo 3 Asuinkerrostalo 4 Muu rakennus (liike-, toimisto- ym.)
Varustetaso	Asunnon varustelutaso.

Aineistojen perusteella muodostettiin seuraavat binääriset muuttujat:

1. Sukupuoli: mies
2. Sukupuoli: nainen
3. Suomalaistaustainen, syntynyt Suomessa
4. Suomalaistaustainen, syntynyt ulkomailla
5. Ulkomaalaistaustainen, syntynyt Suomessa
6. Ulkomaalaistaustainen, syntynyt ulkomailla
7. Henkilön ensisijainen kieli on suomi
8. Henkilön ensisijainen kieli on ruotsi
9. Henkilön ensisijainen kieli on muu kuin suomi tai ruotsi
10. Henkilön ikä vuosissa vuoden lopussa 0–9
11. Henkilön ikä vuosissa vuoden lopussa 10–19
12. Henkilön ikä vuosissa vuoden lopussa 20–29
13. Henkilön ikä vuosissa vuoden lopussa 30–39
14. Henkilön ikä vuosissa vuoden lopussa 40–49
15. Henkilön ikä vuosissa vuoden lopussa 50–59
16. Henkilön ikä vuosissa vuoden lopussa 60–69
17. Henkilön ikä vuosissa vuoden lopussa 70–79
18. Henkilön ikä vuosissa vuoden lopussa >79

19. Sijainti: sisempi kaupunkialue
20. Sijainti: ulompi kaupunkialue
21. Sijainti: kaupungin kehysalue
22. Sijainti: maaseudun paikalliskeskus
23. Sijainti: kaupungin läheinen maaseutu
24. Sijainti: ydinmaaseutu
25. Sijainti: harvaan asuttu maaseutu
26. Ammattiasema päätyösuhteen mukaan palkansaaja
27. Ammattiasema päätyösuhteen mukaan yrittäjä
28. Opiskelijan koulutuslaji: lukiokoulutus
29. Opiskelijan koulutuslaji: AMK
30. Opiskelijan koulutuslaji: Ammatillinen koulutus
31. Pääasiallinen toiminta: Työllinen
32. Pääasiallinen toiminta: Työtön
33. Pääasiallinen toiminta: Opiskelija
34. Pääasiallinen toiminta: Eläkeläinen
35. Pääasiallinen toiminta: Varusmies, siviilipalvelumies
36. Pääasiallinen toiminta: Muu työvoiman ulkopuolella oleva
37. Henkilö on ollut työtön pidempään kuin puoli vuotta tilastovuodesta
38. Asunnon hallintaperuste: Omistaa talon
39. Asunnon hallintaperuste: Omistaa asunnon osakkeet
40. Asunnon hallintaperuste: Aravavuokra-asunto
41. Asunnon hallintaperuste: Korkotukivuokra-asunto
42. Asunnon hallintaperuste: Muu vuokra-asunto
43. Asunnon hallintaperuste: Asumisoikeusasunnot
44. Talotyyppi: Erillinen pientalo
45. Talotyyppi: Rivi- tai ketjutalo
46. Talotyyppi: Asuinkerrostalo
47. Talotyyppi: Muu rakennus (liike-, toimisto- ym.)
48. Asunnon varustelutaso: hyvin varustettu
49. Asunnon varustelutaso: puutteellisesti/erittäin puutteellisesti varustettu

5. FOLK (jaksotieto) työnhaku (FOLK-työnhakujaksot)

Aineistosta poimittiin aikamuuttujien ja tunnistetietojen lisäksi työnhakijan työllisyyskoodi, joka kuvaa henkilön työllisyystilannetta tarkasteltavana tilastovuotena. Aineistojen perusteella muodostettiin seuraavat binääriset muuttujat:

1. Henkilö on työtön
2. Henkilö on lomautettu
3. Henkilö on lyhennetyllä työviikolla
4. Henkilö on työvoiman ulkopuolella
5. Henkilö on työttömyyseläkkeellä
6. Henkilö on työllistymistä edeltävässä palvelussa
6. FOLK (jaksotieto) työvoimakoulutus (FOLK-työvoimakoulutusjaksot)

Aineistosta luotiin binäärinen muuttuja, joka kuvaa onko henkilö osallistunut työvoimakoulutukseen tarkasteltavana tilastovuonna. Muuttujan luomiseen käytettiin tilastovuotta sekä työvoimakoulutuksen alku- ja lopetuspäivämäärää.

Pelastustoimen toimenpide- ja resurssirekisteri PRONTOn aineistot

PRONTO-aineistoista haettiin sekä rakennuspalot että rakennuspalovaarat vuosilta 2008–2020. Muuttujiksi valittiin onnettomuustyyppi, vuosi ja koordinaattitiedot. PRONTO-aineistojen osoitetiedot ja rakennustunnukset olivat puutteelliset, joten onnettomuustiedot olisi liitetty muihin aineistoihin koordinaattien perusteella. Jos onnettomuustiedot liitettäisiin muihin aineistoihin rakennuksien tarkkuudella, tulisi ne liittää osoitetietojen perusteella, koska koordinaattitiedot eivät ole identtisiä muiden aineistojen, kuten DVV:n rakennus- ja huoneistorekisterin kanssa. Jollain alueilla tietojen yhdistäminen rakennuksen tarkkuudella pelkästään koordinaattitietojen perusteella on mahdollista, mutta tiheästi asutuilla seuduilla tiedot eivät ole riittävän tarkkoja.

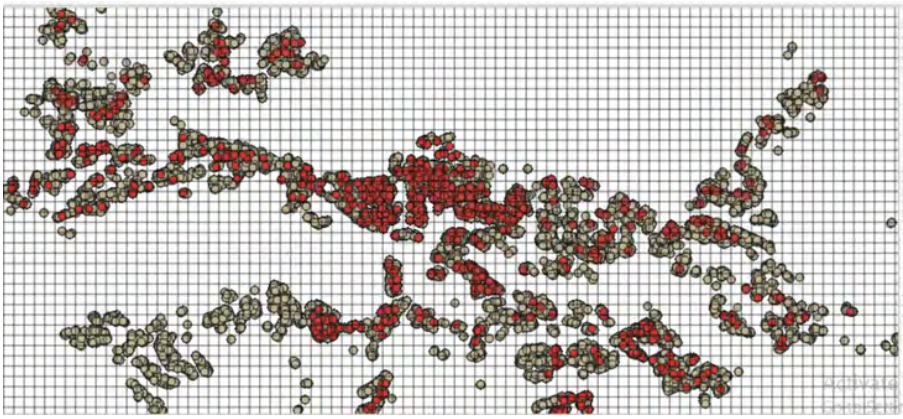
Koska onnettomuustiedot liitetään Tilastokeskuksen aineistoihin 250m^2 tarkkuudella, koordinaattitiedot olivat tarpeeksi tarkkoja tähän tarkoitukseen. Suurin osa PRONTOn koordinaateista on kirjattu KKJ3 -koordinaatiston mukaisesti. Koordinaatit muunnettiin ETRS-TM35FIN tasokoordinaattijärjestelmän mukaisiksi, jotta tiedot saadaan yhdistettyä muihin aineistoihin. Aineisto sisältää 72 092 rakennuspaloa ja rakennuspalovaaraa. Tilastokeskus siirsi aineistosta vain osan etäkäyttöjärjestelmään, joten käytössä oli 14 868 rakennuspalon/rakennuspalovaaran tiedot (20,6 %).

Tiedon yhdistäminen ja opetusaineiston luonti

Tilastokeskuksen FOLK-moduulien suuren koon vuoksi tiedostot pilkottiin tilastovuosien perusteella erillisiksi tiedostoiksi. Opetusaineiston muuttujat luotiin FOLK-moduulien tietojen perusteella RStudio -ohjelmistoa käyttämällä. Muuttujat luotiin jokaiselle tilastovuodelle aikamuuttujia hyödyntämällä. Muuttujista muodostettiin binäärisiä siten, että arvo on aina joko 0 tai 1. Esimerkiksi ”Henkilö on työkyvyttömyyseläkkeellä” muuttujan arvo on 0 jos henkilö ei tarkasteltavana tilastovuotena ole työkyvyttömyyseläkkeellä ja arvo on 1, jos henkilö on. FOLK-moduulien tiedot yhdistettiin keskenään tilastovuosien ja suojatun henkilönumeron perusteella.

Tilastokeskuksen aineistoista poimittiin ne tietorivit, joissa oli koordinaattitiedot (noin 2,6 miljoonaa tietoriviä per tilastovuosi). Tilastokeskuksen sijaintitietomoduulissa on henkilöiden sijaintitiedot 250 m^2 tarkkuudella alueilla, joilla asuu enemmän kuin kolme henkilöä. FOLK-moduulien tiedot yhdistettiin sijaintitietomoduuliin tilastovuoden ja suojatun henkilönumeron perusteella. Jokaiselle 250m2 ruudulle laskettiin ruudulla olevien pisteiden määrä ja muuttujien 1 arvojen prosenttiosuus. Jos esimerkiksi ruudulle osuu 50 pistettä (henkilöä), prosenttiosuus kuvaa kuinka suurella osuudella henkilöistä ”Henkilö on työkyvyttömyyseläkkeellä” muuttujan arvo on 1. Laskelmien tekemiseen käytettiin QGIS -paikkatieto-ohjelmaa.

Rakennuspalot lisättiin samaan sijaintitiedostoon siten, että rakennuspalon vuosi vastaa kyseisen tilastovuoden muuttujia. Yhdistäminen tehtiin vuosi kerrallaan, jotta muuttujat kuvaavat rakennuspalon aikaista ympäristöä. Lopuksi tiedostot yhdistettiin yhdeksi opetusaineistoksi. Lopullinen opetusaineisto sisältää ruudun tunnistetiedot, muuttujat ja tiedon onko ruudussa sattunut rakennuspaloa (0 tai 1). (Kuva 4)



Kuva 4. Havainnollinen esimerkki tietojen yhdistämisestä QGIS paikkatieto-ohjelmistolla. Ruudut kuvaavat 250m^2 alueita, harmaat pisteet Tilastokeskuksen muuttujia ja punaiset pisteet rakennuspaloja tai rakennuspalovaaroja.

Muuttujien analysointi mallinnusta varten

Koska muuttujien välinen korrelaatio saattaa vaikuttaa niiden painoarvoon mallinnuksessa, opetusaineiston muuttujien keskinäistä korrelaatiota tutkittiin korrelaatiomatriisin avulla. Sitä varten käytettiin R-ohjelmistokielen

Hmisc -laajennuspaketin (Harrel Jr 2019) rcorr -argumenttia, joka luo korrelaatiomatriisin, jossa on sekä tutkittavien tekijöiden korrelaatiokerroin että niiden tilastollinen merkitsevyys (p-arvo). Muuttujia, joilla oli vahva positiivinen korrelaatio (>0.8) tarkasteltiin ja osa muuttujista poistettiin sen perusteella. Aineistoissa oli esimerkiksi useampi vanhuuseläkettä, työttömyyttä ja koulutusastetta kuvaavaa muuttuja, joista säilytettiin vain yhdet tekijää kuvastavat muuttujat. Syntymävaltiota, kansalaisuutta ja syntyperää kuvaavat muuttujat korreloivat myös vahvasti keskenään, joten syntymävaltio ja kansalaisuus -muuttujat poistettiin.

Lopullinen opetusaineisto sisältää seuraavat muuttujat:

1. Henkilön puoliso on kuollut samana vuonna (kuin rakennuspalo tapahtui)
2. Henkilön puoliso on kuollut edellisenä vuotena
3. Henkilö on muuttanut erilleen kumppanista samana vuonna
4. Henkilö on muuttanut erilleen kumppanista edellisenä vuotena
5. Henkilö on avioeronnut samana vuonna
6. Henkilö on avioeronnut edellisenä vuotena
7. Puoliso on siirtynyt laitokseen samana vuonna
8. Puoliso on siirtynyt laitokseen edellisenä vuotena
9. Henkilö on asuinliitossa tarkasteltavana tilastovuotena
10. Asuntokuntarakenne: 1 perhe, ei muita
11. Asuntokuntarakenne: 1 perhe ja muita
12. Asuntokuntarakenne: 2 perhettä, ei lisäksi muita
13. Asuntokuntarakenne: 2 perhettä, lisäksi muita
14. Asuntokuntarakenne: vähintään 3 perhettä, ei lisäksi muita
15. Asuntokuntarakenne: vähintään 3 perhettä, lisäksi muita
16. Asuntokuntarakenne: ei perhettä, asuntokunnan koko=1
17. Asuntokuntarakenne: ei perhettä, asuntokunnan koko 2 henkilöä, joilla sama sukupuoli
18. Asuntokuntarakenne: ei perhettä, asuntokunnan koko 2 henkilöä, joilla eri sukupuoli
19. Asuntokuntarakenne: ei perhettä, asuntokunnan koko >2; kaikilla sama sukupuoli
20. Asuntokuntarakenne: ei perhettä, asuntokunnan koko >2; joilla eri sukupuoli
21. Asumisväljyys: Tilava. 1–5 henkilön asuntokunta, jonka käytössä on vähintään kolme asuinhuonetta enemmän kuin asuntokunnan henkilömäärä (keittiötä ei lueta huonelukuun)
22. Asumisväljyys: Normaali
23. Asumisväljyys: Ahdas. Enemmän kuin yksi henkilö huonetta kohti (keittiötä ei lueta huonelukuun)
24. Asuntokunnassa on auto
25. Asuntokunnassa on henkilö, joka saa sairausajan toimeentuloturvaa

26. Asuntokunnassa on henkilö, joka saa toimintarajoitteisuuteen liittyvää tukea
27. Asuntokunnassa on henkilö, joka saa vanhuuden turvaan luettavia etuuksia
28. Asuntokunnassa on henkilö, joka saa lesken ja muiden omaisten tukea
29. Asuntokunnassa on henkilö, joka saa lapsiperheiden etuuksia
30. Asuntokunnassa on henkilö, joka saa työttömyysturvaa
31. Asuntokunnassa on henkilö, joka saa asumisen tukia
32. Asuntokunnassa on henkilö, joka saa opintoetuuksia
33. Asuntokunnassa on henkilö, joka saa toimeentuloturvaa
34. Asuntokunnan käytettävissä oleva rahatulo 10 000 € vähemmän kuin kaikkien asuntokuntien mediaani tulo
35. Asuntokunnan käytettävissä oleva rahatulo on kaikkien asuntokuntien mediaani tulo ±10 000 €
36. Asuntokunnan käytettävissä oleva rahatulo on enemmän kuin kaikkien asuntokuntien mediaani tulo + 10 000 €
37. Asuntokunnassa ei ole velkaa
38. Asuntokunnan jäsenten yhteenlasketut velat ovat vähemmän kuin kaikkien asuntokuntien velkojen mediaani + 10 000 €
39. Asuntokunnan jäsenten yhteenlasketut velat ovat enemmän kuin kaikkien asuntokuntien velkojen mediaani + 10 000 €
40. Asukkaan sukupuoli on mies
41. Asukkaan sukupuoli on nainen
42. Asukas on syntyperältään suomalaistaustainen, syntynyt Suomessa
43. Asukas on syntyperältään suomalaistaustainen, syntynyt ulkomailla
44. Asukas on syntyperältään Ulkomaalaistaustainen, syntynyt Suomessa
45. Asukas on syntyperältään Ulkomaalaistaustainen, syntynyt ulkomailla
46. Asukkaan ensisijainen kieli on suomi
47. Asukkaan ensisijainen kieli on ruotsi
48. Asukkaan ensisijainen kieli on muu kuin suomi tai ruotsi
49. Henkilön ikä vuoden lopussa on 0–9 vuotta
50. Henkilön ikä vuoden lopussa on 10–19 vuotta
51. Henkilön ikä vuoden lopussa on 20–29 vuotta
52. Henkilön ikä vuoden lopussa on 30–39 vuotta
53. Henkilön ikä vuoden lopussa on 40–49 vuotta
54. Henkilön ikä vuoden lopussa on 50–59 vuotta
55. Henkilön ikä vuoden lopussa on 60–69 vuotta
56. Henkilön ikä vuoden lopussa on 70–79 vuotta
57. Henkilön ikä vuoden lopussa on >79 vuotta
58. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Sisempi kaupunkialue.
59. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Ulompi kaupunkialue

60. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Kaupungin kehysalue
61. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Maaseudun paikalliskeskus
62. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Kaupungin läheinen maaseutu
63. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Ydinmaaseutu
64. Henkilön asuinpaikan maaseutu-kaupunki -luokitus: Harvaan asuttu maaseutu
65. Henkilön ammattiasema: palkansaaja
66. Henkilön ammattiasema: yrittäjä
67. Opiskelijan koulutuslaji: lukiokoulutus
68. Opiskelijan koulutuslaji: ammattikorkeakoulu
69. Opiskelijan koulutuslaji: ammatillinen koulutus
70. Henkilön pääasiallinen toiminta: työllinen
71. Henkilön pääasiallinen toiminta: työtön
72. Henkilön pääasiallinen toiminta: opiskelija, koululainen
73. Henkilön pääasiallinen toiminta: varusmies, siviilipalvelusmies
74. Henkilön pääasiallinen toiminta: muu työvoiman ulkopuolella oleva
75. Henkilön asunnon hallintaperuste: omistaa talon
76. Henkilön asunnon hallintaperuste: omistaa asunnon osakkeet
77. Henkilön asunnon hallintaperuste: Aravavuokra-asunto
78. Henkilön asunnon hallintaperuste: korkotukivuokra-asunto
79. Henkilön asunnon hallintaperuste: muu vuokra-asunto
80. Henkilön asunnon hallintaperuste: asumisoikeusasunnot
81. Henkilön asunnon hallintaperuste: muu hallintaperuste (sukulaisuus ym.)
82. Henkilön asunnon talotyyppi: Erillinen pientalo
83. Henkilön asunnon talotyyppi: Rivi- tai ketjutalo
84. Henkilön asunnon talotyyppi: Asuinkerrostalo
85. Henkilön asunnon talotyyppi: Muu rakennus (liike-, toimisto- ym.)
86. Henkilön asunnon varustelutaso: hyvin varustettu
87. Henkilön asunnon varustelutaso: puutteellisesti/erittäin puutteellisesti varustettu
88. Henkilö on työkyvyttömyyseläkkeellä
89. Henkilö on työttömyyseläkkeellä
90. Henkilö on maatalouden erityiseläkkeellä
91. Henkilö on osa-aikaeläkkeellä
92. Henkilö on perhe-eläkkeellä
93. Henkilö on takuueläkkeellä
94. Henkilö on lomautettu
95. Henkilö työskentelee lyhennetyllä työviikolla
96. Henkilö on työvoiman ulkopuolella
97. Henkilö on työllistymistä edistävässä palvelussa
98. Henkilö on työvoimakoulutuksessa

Rakennuspalojen riskiarvojen mallinnus

Algoritmit

Koko Suomen rakennuspaloriskien mallintamisessa kokeiltiin neljää eri ohjatun koneoppimisen algoritmia, johon kuuluu, satunnaismetsä, XGboost (extreme gradient boosting), tukivektorikone sekä yleistetyn lineaarisen mallin logistinen regressio.

Satunnaismetsä on päätöspuumalli, mikä on joukko-oppimista soveltava metodi, jossa yksinkertaisena algoritmikomponenttina on päätöksentekopuu. Koulutusprosessissa algoritmi rakentaa useita päätöksentekopuita, jotka luokittelevat aineiston oksien haaroissa olevien kysymysten perusteella, jotka on muodostettu opetusaineiston perusteella (Telivuo 2018). Satunnaismetsä -algoritmia on yleisesti hyödynnetty aikaisemmissa tapaustutkimuksissa, joissa on mallinnettu eri alueiden rakennuspaloriskiä ohjatun koneoppimisen algoritmeja hyödyntämällä (Madaio ym. 2016, Jay 2017, Singh ym. 2018, Dang ym. 2019, Teixeira 2020). Satunnaismetsän vahvuuksiin kuuluu sen tarkkuus, ylisovittumisen vähäinen riski ja tehokkuus mallin koulutusprosessissa. Algoritmi myös sietää hyvin puuttuvia arvoja. Tosin suuria aineistoja analysoidessa satunnaismetsä ei välttämättä ole tehokkain algoritmi reaaliaikaisen analyysien tuottamissa, vaikka sen kouluttaminen onkin verrattain nopeaa (Liu ym. 2013).

XGBoost on päätöspuihin perustuva algoritmi, joka hyödyntää gradientin laskeutumisarkkitehtuuria puumenetelmissä (Chen ym. 2015). Se on suunniteltu tehostamaan ohjatun koneoppimismallien nopeutta ja tarkkuutta. XGBoostin käyttö on yleistynyt huomattavasti eri tieteenaloilla viime vuosien aikana (Nielsen 2016) ja sitä on hyödynnetty myös muutamassa rakennuspaloihin liittyvässä tapaustutkimuksessa (Singh ym. 2018, Dang ym. 2019). XGBoostin on sanottu olevan yksi tehokkaimmista koneoppimisen algoritmeista ja sen vahvuuksiin kuuluu kyky käsitellä suuria aineistoja tehokkaasti ja tunnistaa monimutkaisia vuorovaikutuksia (Nielsen 2016). Aiemmassa tutkimuksessa XGBoost ei ole aina ollut tarkin mahdollinen algoritmi, mutta kiistattomasti nopein (Ramraj ym. 2016).

Tukivektorikone on lineaarinen luokitinmalli, mikä yksinkertaistettuna jakaa tiedon kahteen luokkaan ja tekee päätelmiä niiden etäisyyksistä marginaalitasoihin (Noble 2006). Välimatkojen etäisyyksiä rajoittavia tekijöitä

kutsutaan tukivektoreiksi, jotka määrittävät lopullisen luokittelun. Tukivektorikonetta on myös kokeiltu muutamissa rakennuspaloriskejä käsittelevissä tapaustutkimuksissa, esimerkiksi USA:ssa (Madaio ym. 2016), Etelä-Koreassa (Hong & Jeong 2018) ja Kiinassa (Wang ym. 2019). Aiemmissa rakennuspaloja mallintavissa tapaustutkimuksissa tukivektorikoneen suorituskyky on ollut samankaltainen kuin satunnaismetsäalgoritmilla. Tukivektorikoneen vahvuuksia ovat sen kyky käsitellä epälineaarisia aineistoja tehokkaasti ja sen kyky ratkaista sekä luokitteluun että regressioon liittyviä ongelmia (Guo & Bai 2009). Tukivektorin suorituskyky on kuitenkin yleisesti parempi pienempiä aineistoja analysoidessa ja algoritmin pyörittäminen vaatii suuren muistikapasiteetin. Algoritmin koulutus suurilla aineistoilla on hidasta.

Yleistetyn lineaarisen mallin logistinen regressio on luokittelualgoritmi, joka mallintaa vastemuuttujan ja selittävien muuttujien suhdetta arvioimalla todennäköisyydet logit-funktion avulla (Kleinbaum ym. 2020). Logistinen regressio on regressioanalyysin erityistyyppi, jota käytetään binäärisen lopputuleman todennäköisyyden laskemiseen. Logistista regressiomallia on käytetty aikaisemmissa vastaavissa tapaustutkimuksissa esimerkiksi USA:ssa (Madaio ym. 2016), Kiinassa (Wang ym. 2019) ja Portugalissa (Teixeira 2020). Logistinen regressio on yleisesti käytetty algoritmi erityyppisten onnettomuuksien, kuten liikenneonnettomuuksien (Lu ym. 2015, Jin & Lee 2017) ja kaivosonnettomuuksien (Maiti & Bhattacherjee 2001, Palei & Das 2009) riskien ennakoimisessa. Aiemmissa rakennuspaloihin keskittyneissä kokeiluissa logistisen regression suorituskyky on ollut alhaisempi verrattuna muihin algoritmeihin. Sen vahvuuksiin kuuluvat sen tehokkuus koulutusvaiheessa sekä sen soveltamisen ja tulosten tulkittavuuden helppous (Healy 2006), mutta se on heikko tunnistamaan monimutkaisempia muuttujien välisiä suhteita (Tu 1996).

Mallien koulutus

Mallien koulutuksessa, rakennuspalojen riskiarvojen muodostamisessa sekä suorituskyvyn arvioimisessa käytettiin RStudio -ohjelmistoa (RStudio Team 2021). Koska jokaisella algoritmilla on useita erilaisia hyperparametrejä, malleja testattiin eri asetuksilla ja lopulliset asetukset määräytyivät parhaan suorituskyvyn mukaan. Satunnaismetsään perustuvissa mallinnuksissa hyödynnettiin *random forest* -laajennuspakettia (RColorBrewer & Liaw 2018) ja puiden määräksi valittiin 500. Aiemmassa tutkimuksessa puiden määrä on vaihdellut, esimerkiksi Atlantassa (Madaio ym. 2016) puiden määräksi va-

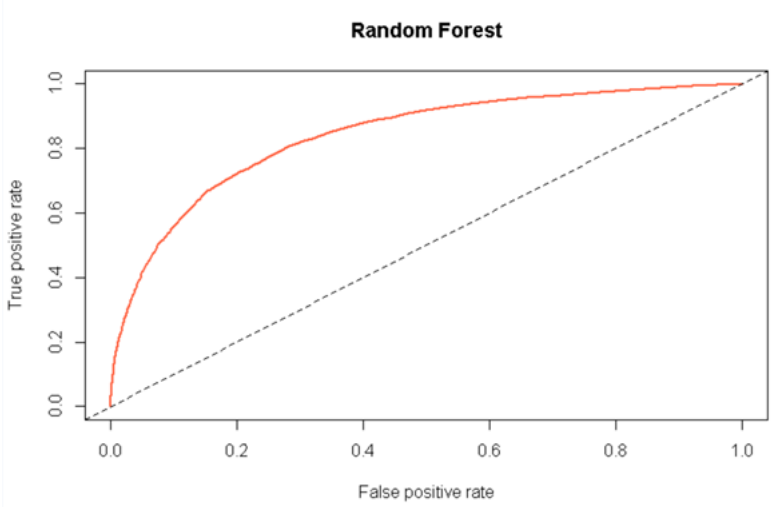
littiin 200, koska suuremmalla määrällä puita ei ollut olennaista vaikutusta mallin suorituskyvyn kannalta.

XGBoostiin perustuvassa mallinnuksessa hyödynnettiin Matrix ja Xgboost -laajennuspakettien ohjelmistokoodeja (Chen 2019, Bates & Maechler 2021). XGBoost -algoritmissa käytettiin seuraavia asetuksia: max_depth=5, eta=0.3, colsample_bytree=0.5, nrounds=200. Logistisessa regressiomallissa käytettiin glm2 laajennuspaketin ohjelmistokoodeja (Marschner & Donoghoe 2018), family=logit asetuksilla. Tukivektorikoneen osalta käytettiin *kernlab* -laajennuspakettia (Karatzoglou ym. 2019). Tukivektorikonemallissa käytettiin *kernlab* – laajennuspaketin *kvsm*-funktion oletusasetuksia, eikä vaihtoehtoisia asetuksia ehditty testata. Tilastokeskuksen etäkäyttöliittymän 31.12.2021 sulkeutumisen takia. Tukivektorikoneen osalta tuloksia pystyttäisiin mahdollisesti parantamaan säätämällä mallin asetuksia.

Mallien luotettavuuden arvioimiseksi käytettiin 10-kertaista ristivalidointimenetelmää, jossa 75 % aineistosta käytettiin mallin koulutukseen ja 25 % mallin testaamiseen, hyödyntämällä *caTools* -laajennuspaketin ohjelmistokoodeja (Tuszynski & Khachatryan 2013). Mallin suorituskykyä arvioidessa koulutusprosessi ajettiin viisi kertaa jokaisen algoritmin kohdalla ja laskettiin suorituskykyä mittaava AUC-arvojen keskiarvo. Mallin suorituskyvyn arviointiin käytettiin pääosin *ROCR*-laajennuspakettia (Sing ym. 2007)

Mallien suorituskyky

Mallien suorituskykyjen vertailuun käytettiin ROC-käyrän (receiver operating characteristic) AUC-arvoa (area under ROC-curve). ROC-käyrä kuvaa mallin kykyä luokitella käytettävissä olevaa aineistoa mittaamalla mallin oikea- ja väärä-positiivisuusasteiden suhdetta (Ling ym. 2003). Käyrän alle jäävää pinta-alaa kutsutaan AUC:ksi. Se kuvaa luokittelijan tarkkuutta asteikolla 0.5-1 siten, että 0.5 tarkoittaa erottelukyvyltään kelvotonta mallia ja 1 tarkoittaa täydellistä suoritusta. Yleisesti ottaen 0.9-1 arvoa pidetään erittäin hyvänä, 0.8-0.9 hyvänä, 0.7-0.8 tyydyttävinä ja 0.6-0.7 heikkona. Alle 0.6 arvoa pidetään epäonnistuneena tuloksena.



Kuva 5. Satunnaismetsän AUC käyrä (AUC 0.839)

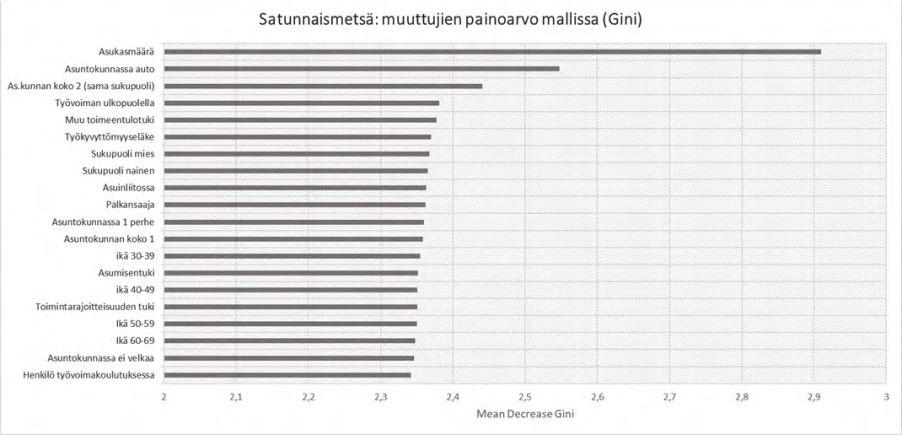
Testatut mallit olivat suorituskyvyltään samankaltaisia ja AUC-arvot vaihtelivat 0.730 ja 0.842 välillä. Testatuista malleista parhaiten suoriutui XGBoost, jonka AUC-arvojen keskiarvo oli 0.842. Toiseksi parhaiten suoriutui satunnaismetsä (AUC 0.839, kuva 5), kolmanneksi logistinen regressio (AUC 0.827) ja heikoimpana tukivektorikone (0.730). Malleista XGBoost oli ylivoimaisesti nopein kouluttaa, algoritmin latausaika oli n.5-10 minuuttia. Logistisen regression latausaika oli noin 15-30 minuuttia ja satunnaismetsän noin 1-2 tuntia. Tukivektorikone oli ylivoimaisesti hitain kouluttaa (noin 16-18 tuntia).

Muuttujien painoarvo mallinnuksessa

Mallien kouluttamisen jälkeen arvioitiin muuttujien painoarvoa mallinnuksissa, jonka avulla selvitettiin, mitkä tekijät vaikuttavat eniten rakennuspalojen todennäköisyyteen. Rakennuspaloja keskeisesti selittäviä tekijöitä arvioitiin kahden parhaiten suoriutuneen algoritmin muuttujien painoarvojen perusteella. Tarkastelussa on 20 painavinta muuttujaa.

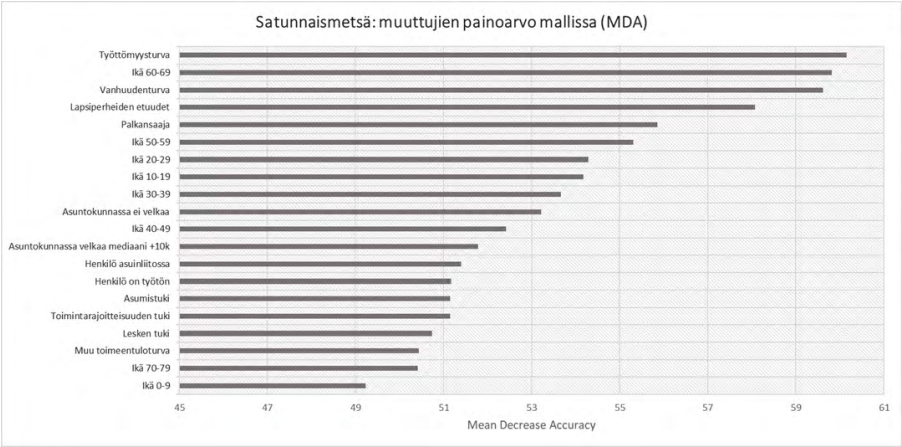
Satunnaismetsään perustuvan mallinnuksen muuttujien painoarvoja arvioitiin vähenevän Gini-arvon (Mean Decrease Gini) ja keskivähenemän tarkkuuden (Mean Decrease Accuracy) perusteella (Han ym. 2016). Vähenevä Gini-arvo mittaa päätöspuun haarojen oikein luokiteltujen havaintojen osuuksia ja laskee väärin luokitelluiden havaintojen todennäköisyyksiä. Tätä

todennäköisyyttä kutsutaan Gini-epäpuhtaudeksi ja niiden vähenevät arvot mittaavat muuttujien tärkeyttä luokittelussa. Gini-arvojen mukaan asukasmäärä on selkeästi painavin muuttuja mallinnuksessa (kuva 6). Toiseksi painavin on ”asuntokunnassa auto” muuttuja, joka viittaa siihen, että asuntokunta omistaa vähintään yhden auton. Kolmanneksi painavin muuttuja on asuntokunnan koko (kaksi ihmistä). Tulokset viittaavat siihen, että mikä suurempi taulukossa kuvattujen muuttujien osuus alueella on, sitä todennäköisempää on, että ruudussa sattuu rakennuspalo.



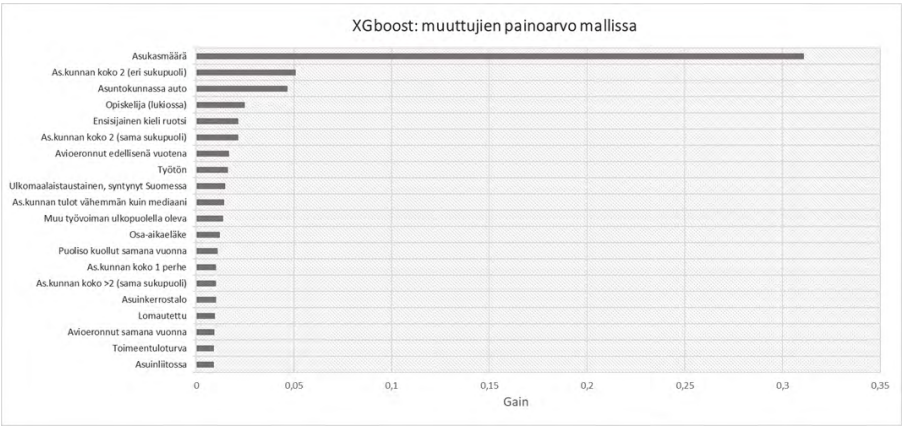
Kuva 6. Satunnaismetsän vähenevät Gini-arvot, 20 painavinta muuttujaa. Mitä suurempi luku, sitä merkittävämpi muuttuja on mallissa.

Satunnaismetsän muuttujien keskivähenemän tarkkuus mittaa kuinka paljon tarkkuutta malli menettää, jos yksittäisen muuttujan ottaa pois opetusaineistosta. Mitä suurempi keskivähenemän tarkkuus arvo on, sitä painavampi muuttuja on mallinnuksessa. Keskivähenemän tarkkuusarvion perusteella mallin tarkkuuden kannalta painavin muuttuja on ”Työttömyysturva”, toiseksi painavin ”ikä 60–69” ja kolmanneksi painavin ”Vanhuudenturva”. (Kuva 7)



Kuva 7. Satunnaismetsän muuttujien keskivähenemän tarkkuus, 20 painavinta muuttujaa. Mitä suurempi luku, sitä merkittävämpi muuttuja on mallissa.

XGBoostiin perustuvassa mallinnuksessa muuttujien painoarvoja mitataan vahvistus (gain) arvolla, joka mittaa kuinka paljon yksittäinen muuttuja vaikuttaa mallin päätöspuihin. Mitä korkeampi muuttujan arvo on, sitä suurempi vaikutus sillä on päätöksenteossa verrattuna muihin muuttujiin. Arvo lasketaan samankaltaisella menetelmällä kuin satunnaismetsän vähenevä Gini-arvo. Vahvistus arvon perusteella mallinnuksen ylivoimaisesti painavin muuttuja on asukasmäärä, toiseksi painavin asuntokunnan koko (2 henkilöä) ja kolmanneksi ”asuntokunnassa auto”-muuttuja. (Kuva 8)



Kuva 8. XGBoost-mallinnuksen muuttujien vahvistusarvo, 20 painavinta muuttujaa. Mitä suurempi luku, sitä merkittävämpi muuttuja on mallissa.

Rakennuspaloriskiarvon muodostaminen

Rakennuspaloriskiarvo muodostetaan syöttämällä koulutettuun malliin aineisto, joka kuvaa viimeisintä tietoa rakennusten, tai tässä tapauksessa ruudukkojen muuttujista ja niiden sijainneista. Riskiarvo on luku 0–1 välillä. Mitä lähempänä arvo on ykköstä, sitä alttiimpi kyseinen alue on rakennuspaloille.

Käyttötarkoituksesta riippuen, riskiarvot voidaan jakaa eri luokkiin. Esimerkiksi Dang ym. 2019 käyttivät samankaltaista menetelmää rakennusten paloriskien määrittelyyn ja he jakoivat rakennukset viiteen eri tasoon, 0–4, riskiarvon perusteella. Taso 4 sisälsi 1 % listan rakennuksista, joilla on korkein riskiarvo. Taso 3 sisälsi listan seuraavat 4 % rakennuksista, taso 2 seuraavat 12 %, taso 1 seuraavat 45 % ja taso 0 loput rakennuksista. Singh ym. toteutti samankaltaisen riskiarvion Pennsylvaniassa 2018 ja jakoi rakennukset kolmeen eri riskiluokkaan, Pittsburghin palontorjunnallisten tarpeiden mukaisesti.

Tämän tutkimuksen keskeisenä tarkoituksena on kartoittaa valtakunnallisia aineistoja ja menetelmiä, joita voidaan hyödyntää rakennuspalojen ennakoinnissa. Rakennusten ja alueiden kategoriointi tulee tehdä yhteistyössä niitä hyödyntävien tahojen kanssa, kuten pelastuslaitosten kanssa.

Analyysin yhteenveto

Testatut algoritmit suoriutuivat hyvin samankaltaisesti, mallien AUC-arvojen välillä ei ollut suuria eroja. XGBoost suoriutui parhaiten ja mallin koulutus oli myös huomattavasti nopeampaa verrattuna muihin algoritmeihin. Satunnaismetsän ja XGBoostin tarkkuusero oli häviävän pieni, 0.003 yksikköä. Satunnaismetsä oli kuitenkin hitaampi kouluttaa ja sen latausaika oli moninkertainen verrattuna XGBoostiin. Tukivektorikone suoriutui testeistä heikoiten, tosin mallia ei ehditty tuunata parhaimmaksi mahdolliseksi aineistojen käyttöajan umpeutumisen takia. Tukivektorikoneen latausaika oli ylivoimaisesti pisin.

Tarkastellessa muuttujien painoarvoa mallinnuksissa, mallien 20 painavimman muuttujien välillä oli paljon samankaltaisuuksia. Asukasmäärä oli selkeästi painavin muuttuja ja autolliset asuinkunnat toiseksi painavin. Sekä XGBoost että satunnaismetsä -mallinnuksissa kahden ihmisen asuntokunnat olivat painavimpien muuttujien kärjessä. Työttömyyteen ja toimeentu-

loon liittyvät muuttujat olivat molemmissa mallinnuksissa 20 painavimman muuttujan joukossa (työtön, muu työvoiman ulkopuolella, henkilö työvoimakoulutuksessa, toimeentuloturva, muu toimeentulotuki, asumisentuki). Myös asuinliitto -muuttuja sekä eläkkeeseen liittyvät muuttujat (työkyvyttömyyseläke, takuueläke, vanhuuden turva ja osa-aika eläke) toistuivat molempien mallien 20 painavimman joukossa.

4. Yhteenveto

Ennakoivan analytiikan tutkimushankkeessa selvitettiin, miten ennakoivan analytiikan menetelmiä, esimerkiksi koneoppimista voidaan hyödyntää pelastustoimen riskianalyysitoiminnan tukena, erityisesti rakennuspalojen mallintamisessa Suomessa. Kirjallisuuskatsauksen avulla tunnistettiin keskeisiä ohjatun koneoppimisen algoritmeja, joilla on mahdollista kohdentaa pelastustoimen palveluita aiempaa riskiperusteisemmin. Aiemmissa ulkomaisissa tapaustutkimuksissa algoritmeja on hyödynnetty pääsääntöisesti tapauskohtaisesti, mutta niitä on mahdollista soveltaa myös laajemmin, esimerkiksi laskennallisen viitekehyksen avulla.

Ohjatun koneoppimisen hyödyntäminen pelastustoimen riskianalytiikassa on viime vuosina yleistynyt kansainvälisesti. Rakennuspalojen mallinnuksessa hyödynnetään yleisesti erilaisia päätöspuumalleja, lineaarisia luokitinmalleja ja neuroverkkoja. Yhdysvalloissa ohjatun koneoppimisen algoritmeja on integroitu erilaisiin ohjelmistoalustoihin, jotka mahdollistavat menetelmien käytön esimerkiksi pelastustoimen toimintavalmiuden suunnittelun tukena. Hankkeessa kokeiltiin Atlantassa kehitettyä Firebird viitekehystä sen avoimen lähdekoodin yhteydessä tarjotuilla testiaineistoilla. Viitekehys on kuitenkin suunniteltu erityyppisille aineistoille, kuin mitä Suomessa on saatavilla, joten alusta ei ole suoraan sovellettavissa Suomen pelastustoimen tarpeisiin.

Kokeilun perusteella kuitenkin todettiin, että vastaavanlaisella ohjatun koneoppimisen algoritmeihin perustuvalla laskennallisella viitekehyksellä on potentiaalia korvata perinteiset regressiomallit, johon tämänhetkinen valtakunnallinen pelastustoimen riskianalyysi perustuu. Viitekehyksen kehittämistä varten kerättiin tietoa aiemmissa vastaavanlaisissa tutkimuksissa käytetyistä algoritmeista ja kartoitettiin Suomessa valtakunnallisesti saatavilla olevia paikkatietoaineistoja, joita voidaan hyödyntää rakennuspalojen alueellisten riskiarvioiden tekemisessä. Aiemman tutkimuksen ja asiantuntijoiden konsultaation perusteella tunnistettiin keskeisiä aineistoja, jotka kuvaavat tietoa esimerkiksi asukkaista, rakennuksista, asuinalueesta sekä aiemmista onnettomuuksista ja häiriötilanteista.

Aineistojen saaminen tutkimuskäyttöön osoittautui hankkeen suurimmaksi haasteeksi, koska käytännön kokeilua varten haettiin laajoja aineistoja,

jotka sisälsivät arkaluonteisia tietoja, kuten tarkkoja sijaintitietoja ja sosioekonomisia muuttujia. Tutkimuslupien ja tietopyyntöjen vaativuuden takia hakemusprosessit olivat pitkiä ja niiden käsittelyprosessit ja käsittelyajat vaihtelivat paljon eri tahojen välillä. Aineistojen koonti ja niiden saamiseen liittyvät kustannukset vaativat huomattavasti resursseja. Jotta vastaavanlaista aineistokokonaisuutta voitaisiin hyödyntää jatkuvassa päätöksenteossa, tulisi keskeiset aineistot tulevaisuudessa koota yhteen päivitettävään tietojärjestelmään, kuten sisäministeriön parhaillaan kehittämään onnettomuuksien ehkäisyn tietojärjestelmään. Onnettomuustietojen lisäksi näihin keskeisin aineistoihin kuuluvat esimerkiksi Tilastokeskuksen FOLK-moduulit.

Pääkaupunkiseudun avointen aineistojen kokeilun perusteella satunnaismetsäalgoritmi suoriutui parhaiten. Kokeilun tuloksena saatu rakennuspaloriskiarvio kuvaa 250m^2 alueita, joissa asuu yli 100 henkilöä, mutta harvemmin asuttuja alueita ei otettu riskiarviossa huomioon. Lähtökohtaisesti avoimilla aineistoilla ei pystytä tuottamaan tarkempia analyyseja ja muilta kuin pääkaupunkiseudun alueelta ei löytynyt avoimesti saatavilla olevaa väestötietoruudukkoa tarpeeksi pienessä resoluutiossa, jotta tuloksista olisi varsinaista hyötyä pelastustoimelle. Vastaavanlaisia ja kattavampia väestötietoa sisältäviä paikkatietoaineistoja on kuitenkin valtakunnallisesti saatavilla 250m^2 alueilta, esimerkiksi Tilastokeskukselta, jos aineistojen käyttöön saa luvan ja niitä käsitellään tietoturvallisessa etäkäyttöympäristössä.

Koko Suomen rakennuspaloriskiarvojen mallintamiseen luotiin laskennallinen viitekehys, joka koostui kahdesta osasta. Ensimmäinen lähdekoodi järjestelee, siivoa, standardisoi ja yhdistää aineistot opetusaineistoksi. Toinen osa kouluttaa algoritmit opetusaineistoilla ja luo rakennuspaloriskiarvion aineistojen viimeisimmän vuoden tietojen perusteella. Testien perusteella XGBoost suoriutui mallinnuksista parhaiten ja soveltuisi hyvin pelastustoimen riskianalyysien muodostamiseen. Laskennallisen viitekehyksen tarkoituksena on luoda valmis ohjelmistokoodikokonaisuus, joka nopeuttaa riskiarvioiden tekemistä, jos menetelmää sovelletaan useaan kertaan tai eri paikoissa.

On kuitenkin epätodennäköistä, että vastaavanlaista aineistokokonaisuutta tullaan hyödyntämään pelastuslaitosten jatkuvassa käytössä, joten lopullinen viitekehys tulee jatkokehittää käytettävissä olevia aineistoja varten. Tältä osin viitekehyksen ensimmäinen osa toimii enemminkin esimerkkinä, kuin varsinaisena työkaluna. Viitekehyksen toista osaa pystyy suoraan soveltamaan erilaisten aineistokokonaisuuksien käyttöön. Jatkokehittelyssä

tulee myös ottaa huomioon, että kaikilla pelastuslaitoksilla ole välttämättä edellytyksiä itsenäisesti tehdä analyysiä, joten ohjelmistokoodit tulisi integroida käyttäjäystävälliseen ohjelmistoalustaan.

Hankkeessa tuotetun käytännön kokeilun tarkoituksena oli viitekehyksen muodostamisen lisäksi kartoittaa uusia aineistoja ja löytää rakennuspaloja keskeisesti selittäviä tekijöitä, sekä aiemmin tunnistettuja, mutta myös sellaisia, joita ei tällä hetkellä hyödynnetä pelastustoimen riskianalyyseissä. Koko Suomen kattavaan analyysiin ei saatu sisällytettyä kaikkia myönnettyjä aineistoja Tilastokeskuksen teknisten ongelmien vuoksi, mutta testit osoittivat sosioekonomisten muuttujien tärkeyden rakennuspalojen riskiarvioiden tekemisessä. Tekniset ongelmat saatiin ratkaistua Tilastokeskuksen kanssa hankkeen päätyttyä, mutta analyysi päätettiin toistaa erillisenä projektina siten, että DVV:n rakennus- ja huoneistorekisterin sekä Hätäkeskuslaitoksen aineistot otetaan mukaan analyysiin. Tästä julkaistaan erillinen artikkeli.

Tutkimuksen alussa rakennuspaloriskiarviot oli tarkoitus muodostaa kuuden eri pelastuslaitoksen toiminta-alueella, koska koko Suomen kattavien aineistojen saanti tutkimuskäyttöön oli epävarmaa. Hanketta varten saatiin kuitenkin Suomen kattavat aineistot DVV:ltä, Tilastokeskukselta ja Hätäkeskuslaitokselta, joten pelastuslaitosten valvontarekisterien aineistoja ei pystytty käyttämään. Kaikkien Suomen pelastuslaitosten aineistojen kerääminen ja standardisointi analyysejä varten olisi ollut tälle hankkeelle varattujen resurssien puitteissa mahdotonta. Koska Tilastokeskuksen aineistot saatiin käyttöön vasta hankkeen viimeisten viikkojen aikana, ei erillisen kuuden pelastuslaitoksen analyysiä ehditty tehdä erikseen.

Koko Suomen rakennuspaloriskiarvojen mallintamisessa asukasmäärä osoittautui kaikista painavammaksi rakennuspaloja selittäväksi tekijäksi ja toiseksi painavin muuttuja oli autolliset asukaskunnat. Työttömyyteen, toimeentuloon ja eläkkeeseen liittyvät muuttujat olivat myös keskeisiä mallinnuksessa. Aiempi kotimainen ja kansainvälinen tutkimus tukee analyysin tuloksia, sillä aiemman tutkimuksen perusteella sosioekonomisten tekijöiden vaikutus rakennusten tulipaloriskiin on kiistaton (Nilson & Bonander 2020; Ojala 2017; Kaukonen 2011; Kokki 2011; Kokki 2014). Esimerkiksi Kaukonen tutki rakennuspalojen lukumääriin vaikuttavia tekijöitä Suomessa (2011) lineaarisella kasvukäyrämallilla. Mallinnuksessa käytetyistä muuttujista väestön korkea ikääntyneisyys ja sosiaali- ja terveydenhuollon asukaskohtaiset nettokustannukset vaikuttivat keskimääräistä korkeampaan rakennuspalojen määrään kunnissa. Kokki (2011, 51–61; 2014) on puolestaan arvioinut pa-

lokuolleiden profiilia, jossa on myös havaittavissa selkeitä sosioekonomisia tausta- tai riskitekijöitä.

Suomessa tällä hetkellä käytössä oleva pelastustoimen riskianalyysissä perustuu yhden neliökilometrin (1 km^2) kokoisiin riskiruutuihin, joihin laskettu rakennuspalojen riskiarvio perustuu ruutujen asukaslukuun, kerrosalaan sekä niiden yhteisvaikutukseen. Aiemmin VTT:n tekemien testien perusteella 1 km^2 ruudut ennakoivat rakennuspaloja tarkemmin kuin 250m^2 ruudut. Riskianalyysimallien kehityksessä kokeiltiin kolmea eri lähestymistapaa, joihin kuuluvat yksinkertainen regressiomalli, spatiaalinen regressiomalli ja itseorganisoituva kartta (SOM). Testien perusteella yksinkertainen regressio malli soveltui parhaiten haettuun käyttötarkoitukseen. Malli selitti noin 50 % rakennuspalomäärien vaihtelusta. (Tillander ym. 2011.)

Tässä tutkimushankkeessa testatut algoritmit suoriutuivat huomattavasti edellä mainittua, tällä hetkellä käytössä olevaa regressiomallia paremmin. Kun käytössä olevaa riskianalyysimallia kehitettiin, kokeiltiin myös neuroverkkoihin perustuvia itseorganisoituvia karttoja sekä Bayesilaiseen tilastotieteeseen perustuvaa spatiaalista regressiomallia. Itseorganisoituvat kartat antoivat oikeansuuntaisia tuloksia, mutta ne arvioivat rakennuspalojen määrät alakanttiin. Ennustuksien epätarkkuuden arveltiin johtuvan käytettävissä olevien muuttujien määrästä ja todettiin, että ennustuskyvyn parantamisen kannalta olisi keskeistä tunnistaa onnettomuuksien kannalta tärkeimpiä taustamuuttuja ja hyödyntää niitä verkon rakentamisessa. Myöskään spatiaalisilla regressiomalleilla ei päästy yhtä hyviin tuloksiin kuin yksinkertaisemmalla regressiomallilla ja rajoittavana tekijänä pidettiin mallin laskennallista raskautta ja mallinnuksen vaativuutta.

Tässä tutkimuksessa ohjatun koneoppimisen algoritmien avulla saatujen riskiarvioiden hyvään tarkkuuteen vaikuttaa se, että aineistoja oli saatavilla laajalta alueelta usealta eri vuodelta. Menetelmät ovat myös kehittyneet siltä ajalta, kun pelastustoimen riskianalyysimalleja on viimeksi kehitetty. Ohjatun koneoppimisen menetelmiä on ruvettu hyödyntämään eri tieteenaloilla ja käyttötarkoituksissa yhä enenemissä määrin viime vuosien aikana ja niiden käyttöä tulisi harkita myös pelastustoimen riskianalytiikassa. Ohjatun koneoppimisen menetelmiä on helppo soveltaa hyvin erityyppisten ilmiöiden mallintamiseen ja ennakoimiseen, joten tässä tutkimuksessa käytetyt ja kuvatut menetelmät voivat hyödyttää paitsi pelastustointa, myös sosiaali- ja terveydenhuollon hallinnonalaa koko maassa ja tulevilla hyvinvointialueilla, mutta myös muita valtakunnallisesti toimivia viranomaisia ja julkishallinnon toimijoita.

5. Hankkeen viestintä

Hankkeesta on viestitty aktiivisesti koko tutkimuksen ajan. Viestinnän kautta on laajennettu hankkeen sidosryhmiä, tunnistettu ja hyödynnetty yhteistyömahdollisuuksia ja pyritty edesauttamaan tulosten jalkauttamista tutkimuksen jälkeen.

Esitelmät/luennot:

- Onnettomuuksien ehkäisyn ja rakenteellisen paloturvallisuuden opintopäivät 2020
- Pelastuslaitosten TKI-verkoston seminaari 2020
- PSCE:n konferenssi tammikuussa 2021
- Pelastustoimen analytiikka webinaari 17.5.2021
- Pelastusopiston tutkimuspäivä 9.6.2021
- Palotutkimuksen päivät 8.9.2021
- SPEKin hallituksen kokous 28.10.2021
- SPEKin valtuuston kokous 29.11.2021
- Yritysturvallisuusfoorumi 24.-26.11.2021, työpaja “Tekoäly tulipalojen ennakoinnissa”
- ISES seminaari

Artikkelit, uutiset ja tiedotteet:

- Uutinen Suorituskyky hankkeen uutiskirjeessä 2020
- Uutinen Feeniks-tiedotteessa (Pelastuslaitosten kumppanuusverkoston uutiskirje) 2020
- Artikkeli Pelastustieto lehdessä 2020
- Artikkeli SPEKin ajankohtaisissa uutisissa 2020
- Artikkeli Palotutkimuksen päivien seminaari julkaisussa 2021

6. Lähteet

Abbott, D., 2014. Applied predictive analytics: Principles and techniques for the professional data analyst. John Wiley & Sons.

Agarwal, P., Tang, J., Narayanan, A. N. L., & Zhuang, J., 2020. Big data and predictive analytics in fire risk using weather data. Risk analysis.

Aktaş, E., Özaydın, Ö., Bozkaya, B., Ülengin, F., & Önsel, Ş., 2013. Optimizing fire station locations for the Istanbul metropolitan municipality. Interfaces, 43(3), 240-255.

Aldabbas, M., Venteicher, F., Gerber, L., & Widmer, M., 2018. Finding the adequate location scenario after the merger of fire brigades thanks to Multiple Criteria Decision Analysis Methods. Foundations of Computing and Decision Sciences, 43(2), 69-88.

Aliverti, E., Lum, K., Johndrow, J. E., & Dunson, D. B., 2018. Removing the influence of a group variable in high-dimensional predictive modelling. arXiv preprint arXiv:1810.08255.

Allgeier, T., 2012. Embedding strategic social marketing in Fire and Rescue services. *Social Business*, *2*(4), pp.338-361.

Apolitical – the global earning platform for government. New York City is saving lives by predicting where fires will break out. https://apolitical.co/en/solution_article/new-york-city-saving-lives-predicting-fires-will-break [4.2.2022]

Ardianto, R., & Chhetri, P., 2019. Modeling spatial–temporal dynamics of urban residential fire risk using a Markov chain technique. International Journal of Disaster Risk Science, 10(1), 57-73.

Awad, W. H., & Janson, B. N., 1998. Prediction models for truck accidents at freeway ramps in Washington State using regression and artificial intelligence techniques. Transportation Research Record, 1635(1), 30-36.

Bahri, S., 2016. 0-1 integer linear programming model for location selection of fire station: A case study in Indonesia. In AIP Conference Proceedings (Vol. 1723, No. 1, p. 030004). AIP Publishing LLC.

Bates, D. & Maechler, M. M., 2021. Package 'Matrix'.

Beaulieu, E., Smith, J., Zheng, A., & Pike, I., 2019. Association between neighborhood socioeconomic features and residential fire incidence, related casualties and children: a cross-sectional population-based study in 4 Canadian provinces. CMAJ open, 7(3), E562.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S., 2017. The role of big data and predictive analytics in retailing. Journal of Retailing, 93(1), 79-95.

Breiman, L., 1996. Out-of-bag estimation. Berkeley, California: Department of Statistics, University of California. https://www.stat.berkeley.edu/pub/users/breiman/OOBestimation.pdf [4.2.2022]

Castelli, M., Vanneschi, L., & Popovič, A., 2015. Predicting burned areas of forest fires: an artificial intelligence approach. Fire ecology, 11(1), 106-118.

Çatay, B., 2011. Siting new fire stations in Istanbul: A risk-based optimization approach. OR insight, 24(2), 77-89.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K., 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

Chhetri, P., Corcoran, J., Ahmad, S., & Kiran, K. C., 2018. Examining spatio-temporal patterns, drivers and trends of residential fires in South East Queensland, Australia. *Disaster Prevention and Management: An International Journal*.

Chevalier, P., Thomas, I., Geraets, D., Goetghebeur, E., Janssens, O., Peeters, D., & Plastria, F., 2012. Locating fire stations: An integrated approach for Belgium. Socio-Economic Planning Sciences, 46(2), 173-182.

Chhetri, P., Corcoran, J., Stimson, R.J. and Inbakaran, R., 2010. Modelling potential Socio-economic determinants of building fires in southeast Queensland. Geographical Research, 48(1), pp.75-85.

Dang, T. T., Cheng, Y., Mann, J., Hawick, K., & Li, Q., 2019. Fire Risk Prediction Using Multi-Source Data: A case study in Humberside area. In 2019 25th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

Dinov, I. D., 2018. Data Science and Predictive Analytics. Springer Science and Business Media LLC: Berlin/Heidelberg, Germany

Dodig-Crnkovic, G., 2002. Scientific methods in computer science. In Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia (pp. 126-130).

Dubey, R., Gunasekaran, A., Childe, S. J., Blome, C., & Papadopoulos, T., 2019. Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. British Journal of Management, 30(2), 341-361.

Duncanson, M., Woodward, A., & Reid, P., 2002. Socioeconomic deprivation and fatal unintentional domestic fire incidents in New Zealand 1993–1998. Fire Safety Journal, 37(2), 165-179

Dutta, C., & Singhal, D., 2019. A Hybridization of Artificial Neural Network and Support Vector Machine for Prevention of Road Accidents in Vanet. International Journal of Computer Engineering and Technology, 10(01).

ePressi, Uudella valvontasovelluksella ehkäistään onnettomuuksia ja sujuvoitetaan palotarkastuksia. https://www.epressi.com/tiedotteet/hallitus-ja-valtio/uudella-valvontasovelluksella-ehkaistaan-onnettomuuksia-ja-sujuvoitetaan-palotarkastuksia.html [4.2.2022]

Grekousis, G., & Liu, Y., 2019. Where will the next emergency event occur? Predicting ambulance demand in emergency medical services using artificial intelligence. *Computers, Environment and Urban Systems*, *76*, 110-122.

Guo, Z., & Bai, G., 2009. Application of least squares support vector machine for regression to reliability analysis. *Chinese Journal of Aeronautics*, *22*(2), 160-166.

Han, H., Guo, X., & Yu, H., 2016. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)* (pp. 219-224). IEEE.

Harrell Jr, F. E., & Harrell Jr, M. F. E., 2019. Package 'hmisc'. CRAN2018, 2019, 235-236.

Hastie, C., & Searle, R., 2016. Socio-economic and demographic predictors of accidental dwelling fire rates. Fire Safety Journal, 84, 50-56.

Healy, L. M., 2006. Logistic regression: An overview. *Eastern Michighan College of Technology*.

Heaton, B. 2015. New York City Fights Fire with Data. Government Technology and Emergency Management magazine. https://www.govtech.com/public-safety/new-york-city-fights-fire-with-data.html [8.2.2022]

Helsingin seudun ympäristöpalvelut, 2021. Pääkaupunkiseudun väestötietoruudukko. https://www.hsy.fi/ymparistotieto/avoindata/avoin-data---sivut/paakaupunkiseudun-vaestotietoruudukko/ [4.2.2022].

Helsinki Region Infoshare (HRI), 2021. Pääkaupunkiseudun osoiteluettelo. Aineiston ylläpitäjä on Helsingin kaupunkiympäristön toimiala / Kaupunkimittauspalvelut ja alkuperäinen tekijä Helsingin, Espoon, Vantaan ja Kauniaisten mittausorganisaatiot. Aineisto on ladattu Helsinki Region Infoshare -palvelusta 26.11.2020 lisenssillä Creative Commons Attribution 4.0. https://hri.fi/data/fi/dataset/seudullinen-osoiteluettelo [4.2.2022]

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., & Hijmans, M. R. J., 2017. Pack-age 'dismo'. Circles, 9(1), 1-68.

Hong, S. G., & Jeong, S. R., 2018. Development and Comparison of Data Mining-based Prediction Models of Building Fire Probability. Journal of Internet Computing and Services, 19(6), 101-112. https://medium.com/homeland-security/big-data-providing-fire-service-and-emergency-response-communities-with-tools-to-predict-and-fecacba466cc

Huang, J., & Ling, C. X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310.

International Public Safety Data Institute, 2022. https://firecares.org/media [8.2.2022]

Jay J., 2017. How we predicted building fires in Baton Rouge, LA – working version. Urban Analytics Blog (2017). https://scholar.harvard.edu/jonjay/blog/ how-we-predicted-building-fires-baton-rouge-la-working-version [15.5.2021]

Jin, S. B., & Lee, J. W., 2017. Study on Accident Prediction Models in Urban Railway Casualty Accidents Using Logistic Regression Analysis Model. *Journal of the Korean Society for Railway*, *20*(4), 482-490.

Karatzoglou, A., Smola, A., Hornik, K., & Karatzoglou, M. A., 2019. Package 'kernlab'. CRAN R Project.

KAT kotona asumisen turvallisuus, 2021. https://www.kotonaasumisenturvallisuus.fi/hanke/kat-time-hanke

Kaukonen, E., 2011. Rakennuspalojen lukumäärään vaikuttavia tekijöitä Suomessa -Kuntakohtaisten rakennuspalojen lukumäärien ja niiden muutosten mallintaminen. Pelastusopiston julkaisu B-sarja: Tutkimusraportit 5/2011.

Kc, K., Corcoran, J., & Chhetri, P., 2018. Spatial optimisation of fire service coverage: a case study of Brisbane, Australia. Geographical research, 56(3), 270-284.

Kiran, K. C., & Corcoran, J., 2017. Modelling residential fire incident response times: A spatial analytic approach. Applied Geography, 84, 64-74.3

Kiran, K. C., Corcoran, J., & Chhetri, P., 2020. Measuring the spatial accessibility to fire stations using enhanced floating catchment method. Socio-Economic Planning Sciences, 69, 100673.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M., 2002. Logistic regres-sion. New York: Springer-Verlag.

Kokki, E. (2011). Palokuolemat ja ihmisen pelastamiset tulipaloissa 2007-2010. Pelastusopisto: Kuopio. http://info.smedu.fi/kirjasto/Sarja_B/B3_2011.pdf

Kokki, E. (2014). Palokuolemat vähentyneet. Suomen palokuolematilastot 2007-2013. Pelastusopisto. https://www.pelastusopisto.fi/wp-content/uploads/2016/12/52749_B2_2014.pdf

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P.,2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

Kohonen, I., Kuula-Luumi, A., & Spoof, S. K. (2019). Ihmiseen kohdistuvan tutkimuksen eettiset periaatteet ja ihmistieteiden eettinen ennakkoarviointi Suomessa. Tutkimuseettisen neuvottelukunnan ohje, 2021-01.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A.,…Benesty, M., 2020. Package 'caret'. The R Journal, 223.

Lehmus, Niina, 2015. Big datasta ja analytiikasta apua tiedolla johtamiseen, opinnäytetyö, Haaga-Helia ammattikorkeakoulu

Lever, J., Krzywinski, M., & Altman, N., 2016. Points of significance: model selection and overfitting.

Ling, C. X., Huang, J., & Zhang, H., 2003. AUC: a better measure than accuracy in comparing learning algorithms. In Conference of the canadian society for computational studies of intelligence (pp. 329-341). Springer, Berlin, Heidelberg.

Liu, M., Wang, M., Wang, J., & Li, D., 2013. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators B: Chemical*, *177*, 970-980.

Lizhong, Y., Heng, C., Yong, Y., & Tingyong, F., 2005. The effect of socioeconomic factors on fire in China. Journal of fire sciences, 23(6), 451-467.

Lu, T., Dunyao, Z. H. U., Lixin, Y., & Pan, Z., 2015. The traffic accident hotspot prediction: Based on the logistic regression method. In *2015 International Conference on Transportation Information and Safety (ICTIS)* (pp. 107-110). IEEE.

Mabrouk, H. H., 2016. Machine learning from experience feedback on accidents in transport. In 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) (pp. 246-251). IEEE.

Madaio, M., Chen, S.T., Haimson, O.L., Zhang, W., Cheng, X., Hinds-Aldrich, M., Chau, D.H. and Dilkina, B.,, 2016. Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 185-194).

Maiti, J., Bhattacherjee, A., 2001. Predicting accident susceptibility: a logistic regression analysis of underground coal mine workers. *Journal of the Southern African Institute of Mining and Metallurgy*, *101*(4), 203-208.

Marschner, I. & Donoghoe, M. M. W., 2018. Package 'glm2'. *Journal, Vol*, *3*(2), 12-15.

Morello, T. F., Ramos, R. M., Anderson, L. O., Owen, N., Rosan, T. M., & Steil, L., 2020. Predicting fires for policy making: Improving accuracy of fire brigade allocation in the Brazilian Amazon. Ecological Economics, 169, 106501.

Murray, A. T., 2013. Optimising the spatial location of urban fire stations. Fire safety journal, 62, 64-71.

Nielsen, D. (2016). *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* (Master›s thesis, NTNU).

Nilson, F., & Bonander, C., 2020. Household fire protection practices in relation to sociodemographic characteristics: evidence from a Swedish national survey. Fire technology, 56(3), 1077-1098.

Nieminen, M., Gotcheva, N., Leikas, J., Koivisto, R., 2019. Ethical AI for the governance of the Society: Challenges and opportunities. In *CEUR Workshop Proceedings* (Vol. 2505, pp. 20-26).

Noble, W. S., 2006. What is a support vector machine? Nature biotechnology, 24(12), 1565-1567.

O'Neill, D., Pierson, V., Joe Russo, J., Wolfe, A., Demaise, C., Wilson, T., 2017. Big Data: Providing Fire Service and Emergency Response Communities with Tools to Predict and Respond. *Homeland Security platform*. https://medium.com/homeland-security/big-data-providing-fire-service-and-emergency-response-communities-with-tools-to-predict-and-fecacba466cc [8.2.2022]

Ojala, T., 2017. Iäkkäiden ja muistisairaiden arjen turvallisuus. OTE – Osallistava Turvallisuus Erityisryhmille. Suomen Pelastusalan Keskusjärjestö. https://issuu.com/spek_ry/docs/spek_tutkii_18 [5.2.2022]

Palei, S. K., & Das, S. K., 2009. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Safety science*, *47*(1), 88-96.

Pedersen, R. D., & Lemelson, J. H., 2001. U.S. Patent No. 6,289,331. Washington, DC: U.S. Patent and Trademark Office.

Pelastusopisto, 2013a. PRONTO – Dynaaminen koulutuskansio: rakennuspalo. https://www.pelastusopisto.fi/wp-content/uploads/2017/02/46315_d_rakennuspalo.pdf [28.3.2022]

Pelastusopisto, 2013b. PRONTO – Dynaaminen koulutuskansio: rakennuspalovaara. https://www.pelastusopisto.fi/wp-content/uploads/ 2017/02/46317_x_rakennuspalovaara.pdf [28.3.2022]

Pham, B.T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H.P.H., Phong, T.V., Nguyen, D.H., Le, H.V., Mafi-Gholami, D. and Prakash, I., 2020. Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry, 12(6), p.1022.

Photis, Y. N., & Grekousis, G., 2012. Locational planning for emergency management and response: An artificial intelligence approach. International Journal of Sustainable Development and Planning, 7(3), 372-384.

Provost F., Fawcett T., 1998. The case against accuracy estimation for comparing induction algorithms. Proc. 15th Intl. Conf. On Machine Learning, s. 445-453

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications, 9, 651-662.

RColorBrewer, S., & Liaw, M. A., 2018. Package 'randomForest'. University of California, Berkeley: Berkeley, CA, USA.

Refaeilzadeh, P., Tang, L., & Liu, H., 2009. Cross-validation. Encyclopedia of database systems, 5, 532-538.

Roman, J., 2014. Data driven - IN PURSUIT OF SMART. NFPA Journal. https://nfpa.org/-/media/Files/News-and-Research/Resources/Research-Foundation/Current-projects/Smart-FF/SmartFFNFPAJournalDataDriven. [04.01.2022]

Roman, J. 2014. In pursuit of smart. NFPA journal, November/December 2014, s.40-50.

RStudio Team, 2021. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Safi, Y., & Bouroumi, A., 2013. Prediction of forest fires using artificial neural networks. Applied Mathematical Sciences, 7(6), 271-286.

Satir, O., Berberoglu, S., & Donmez, C., 2016. Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. Geomatics, Natural Hazards and Risk, 7(5), 1645-1658.

Seong-gwan Hong & Seung-ryul Jeong, 2018. Development and Comparison of Data Mining-based Prediction Models of Building Fire probability. Internet Information Society, 19(6), pp.101-112.

Shahparvari, S., Fadaki, M., & Chhetri, P., 2020. Spatial accessibility of fire stations for enhancing operational response in melbourne. Fire Safety Journal, 103149.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T., 2007. The ROCR package. R Vignette.

Singh Walia, B., Hu, Q., Chen, J., Chen, F., Lee, J., Kuo, N., Narang, P., Batts, J., Arnold, G. and Madaio, M., 2018. A dynamic pipeline for spatio-temporal fire risk prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 764-773)

Sisäministeriö, 2020. Alustava kilpailutettavan hankinnan kohteen kuvaus – Onnettomuuksien ehkäisyn tietojärjestelmä, Sisäministeriö, Pelastusosasto. http://dynastyjulkaisu.pohjoiskarjala.net/siunsote/kokous/2021136-7-25738.PDF [4.2.2022]

Suomen Pelastusalan Keskusjärjestö, 2020. Ennakoivan analytiikan tutkimushanke. Hankesuunnitelma. Lue lisää: SPEK ajankohtaiset uutiset https://www.spek.fi/tekoalylla-voi-ennustaa-tulipaloja/ [4.2.2022]

Teixeira, L. P. P. B., 2020. Predictive analytics applied to firefighter response, a practical approach (Doctoral dissertation).

Telivuo, Suvi, 2018. Satunnaismetsä-koneoppimismenetelmä, teoria ja soveltaminen. Pro gradu tutkielma, Helsingin yliopisto.

Tillander, K., Matala, A., Hostikka, S., Tiittanen, P., Kokki, E., & Taskinen, O., 2010. Pelastustoimen riskianalyysimallien kehittäminen. Espoo, VTT Tiedotteita, 2530.

Todorovic, S., 2020. Modelling risk factors in urban residential fires in Helsinki. Pro gradu -tutkielma, Helsingin yliopisto. https://www.researchgate.net/publication/342353779_Modelling_risk_factors_in_urban_residential_fires_in_Helsinki [14.1.2022]

Tommila, P., 2017. Työ- ja elinkeinoministeriö, Luvat ja valvonta -kärkihanke. Tausta, tavoitteet, eteneminen. https://tem.fi/documents/1410877/7148138/Seminaari24112017_final.pdf/e183ddfd-3471-4b48-bb6a-b51610164ff2/Seminaari24112017_final.pdf. [04.01.2022]

Treibitz, S. 2017. Big Data Comes to Public Safety. *Firecare media release*. https://firecares-static-prod.firecares.org/fc_release.pdf [8.2.2022]

Tu, J. V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225-1231.

Tuszynski, J., & Khachatryan, M. H., 2013. Package 'caTools'.

Valtioneuvosto, 2019. Pääministeri Sanna Marinin hallituksen ohjelma 10.12.2019. Osallistava ja osaava Suomi – sosiaalisesti, taloudellisesti ja ekologisesti kestävä yhteiskunta. Valtioneuvoston julkaisuja 2019:31. https://julkaisut.valtioneuvosto.fi/handle/10024/161931 [14.4.2022]

Wang, Q., Zhang, J., Guo, B., Hao, Z., Zhou, Y., Sun, J., Yu, Z. and Zheng, Y., 2019. CityGuard: Citywide Fire Risk Forecasting Using A Machine Learning Approach. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(4), pp.1-21.

Yao, J., & Zhang, X., 2016. Spatial-temporal dynamics of urban fire incidents: a case study of Nanjing, China. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 63-69.

Ympäristöministeriö, 2021. Rakennetun ympäristön tietojärjestelmä. https://ym.fi/ryhti [4.2.2022]

Yu, Y. F., 2012. Optimization for Fire Station Location Planning Based on Travel Time of Fire Vehicles. In 2012 Eighth International Conference on Computational Intelligence and Security (pp. 328-331). IEEE.

LIITE 1. Aineistojen kuvaus ja hakuprosessi

Pelastustoimen toimenpide- ja resurssirekisteri PRONTO

PRONTO on sisäministeriön (SM) järjestelmä, jonka yleinen ohjaaminen ja kehitys on SM:n pelastusosaston vastuulla ja tekninen ylläpito ja kehittämisvastuu on Pelastusopistolla. PRONTO-järjestelmä sisältää koko valtakunnan kattavia onnettomuustietoja, sisältäen hälytys-, onnettomuus-, rakennus- ja palontutkintaselosteet vuodesta 1996 alkaen.

Hankkeeseen haettiin ja saatiin PRONTOn käyttöoikeus, joka kattoivat kaikki PRONTOn tiedot. PRONTOn vapaapoiminta -toiminnolla raakadataa pystyy lataamaan vain rajatun määrän kerralla, joten hankkeen alkuvaiheessa selvitettiin raakadatapoiminnan mahdollisuutta, joka kattaisi koko tarvittavan tietomäärän. Koska PRONTO-rekisterin tiedot sijaitsevat Tieto Oyj:n palvelinlaitteistoilla, kertaluontoinen laajempi raakadatapoiminta olisi aiheuttanut huomattavat kustannukset, joten todettiin, että erilliset poiminnat vapaapoiminnalla on kustannustehokkaampi vaihtoehto.

Tarvittavat tiedot ladattiin järjestelmästä ja tiedon käsittelyn sujuvoittamiseksi tiedostoja varten luotiin MS Access tietokanta, jonne rakennuspalojen ja rakennuspalovaarojen hälytys-, onnettomuus-, ja rakennusselosteiden sisällöt tallennettiin. Tiedot ladattiin vuosilta 2008–2019. Latausmäärärajoitusten vuoksi tiedot haettiin palasina ja yhdistettiin tietokannassa.

Pelastuslaitosten tiedot

Pelastuslaitoksilta haettiin hanketta varten Pelastuslaitosten valvontarekisterin tiedot poistumisturvallisuusselvityksistä, palotarkastuksista, pelastuslain 42 §:n mukaisista paloriski-ilmoituksista, kiinteistöjen palontorjuntatekniikasta sekä muusta kiinteistökohtaisesta, strukturoidusta tiedosta koskien kiinteistön paloturvallisuutta. Yhteistyötä varten kuudelle eri pelastuslaitokselle lähetettiin tutkimuslupahakemukset, joiden perusteella saatiin tutkimusluvat aineistoihin. Näihin kuuluvat Helsingin, Pohjanmaan, Etelä-Karjalan sekä Itä-, Länsi-, ja Keski-Uudenmaan pelastuslaitokset.

Ensimmäinen tietojen poiminta tehtiin Keski-Uudenmaan pelastuslaitoksen kanssa. Poiminnan eri vaiheista tehtiin yksityiskohtainen ohje, jota seuraa-

malla muut Pelastuslaitokset pystyivät poimimaan tiedot kustannustehokkaasti ja standardisoidulla tavalla. Tietoja haettiin vuosilta 2008–2020.

Digi- ja väestötietovirasto
Digi- ja väestötietovirastolta (DVV) haettiin tutkimuslupaa koko valtakunnan kattavaan rakennus- ja huoneistorekisteriin (RHR). Koska DVV toimittaa RHR:n veloituksetta pelastuslaitoksille kerran vuodessa pelastuslain edellyttämällä tavalla, selvitettiin mahdollisuutta käyttää aineistoja hankkeen yhteistyöpelastuslaitosten kautta, jolloin DVV:n ei tarvitsisi poimia tietoja vaan pelastuslaitokset toimittaisivat tiedot erillisellä DVV:n myöntämällä tutkimusluvalla.

Digi- ja väestötietovirastoon otettiin ensimmäisen kerran yhteyttä lokakuun 2020 alussa ja asiasta käytiin viestintää joulukuuhun 2020 asti. 22.12.2020 DVV päätyi sisäisten keskustelujen jälkeen ratkaisuun, jossa tutkimuksesta vastaava SPEK hakee aineistot suoraan DVV:ltä tavanomaisen tietolupaprosessin kautta. Koko Suomen kattava RHR -aineisto saatiin hankkeen käyttöön 15.3.2021.

Hätäkeskuslaitos ja Poliisihallitus
Hätäkeskuslaitoksen yleiseen sähköpostiin lähetettiin tietopyyntö 16.11.2020, jossa haettiin tietoja melu- ja häiriöilmoituksista. Myöhemmin erillisen puhelinkeskustelun perusteella tietopyyntö lähetettiin Hätäkeskuslaitoksen poliisitoimintapäällikölle, joka välitti pyynnön kehityspäällikölle. Kehityspäällikön mukaan tietoja kannattaa tiedustella suoraan poliisihallitukselta, sillä heidän tietojärjestelmästänsä ei saa välitettyä raakadataa. Raakadata pitää erikseen tilata tietojärjestelmätoimittajalta, mikä tulisi maksamaan arvion mukaan muutamia tuhansia euroja. Keskustelujen perusteella tietosisällön luovutusmahdollisuus oli myös juridisesti epäselvää.

Hätäkeskuslaitoksen yleisestä sähköpostista toimitettiin erillisessä keskustelussa tutkimuslupalomake ja ei mainittu mahdollisista esteistä tiedon hakemiseen liittyen. Tutkimuslupahakemus toimitettiin joulukuussa 2020. Hakemus kuitattiin vastaanotetuksi ja seuraavia seikkoja tarkennettiin:

- Meluun ja häiriöön liittyvien hätäilmoituksien osalta rekisterinpitäjänä toimii Poliisihallitus ja tutkimuslupaa kyseisten tietojen osalta tulee hakea suoraan sieltä.
- Hätäilmoituksia säilytetään lakisääteisesti viisi vuotta ilmoitushetkestä. Yksityiskohtaisia tehtävätietoja kyetään täten hakemaan viimeisten viiden vuoden ajalta.

- Hätäkeskuslaitoksen raportointijärjestelmien kautta voidaan asiaa hakea vuodesta 2010 lähtien tehtävälajin näkökulmasta mutta tällöin mukana ei ole mitään tarkkoja osoitteita, tai edes kuntaan liittyvää tietoa. Hätäkeskuslaitos tilastoi omia toimintojaan vain hätäkeskusalue -näkökulmasta.
- Hätäkeskuslaitos ei voi poimia tietolistausta, jossa näkyisi tehtävistä rivitietoina esim. päivämäärä, kellonaika ja tarkka tehtäväosoite. Työ tulee tilata erikseen tietojärjestelmän tekniseltä ylläpitäjältä, joka laskuttaa arviolta 1000 € / työpäivä.
- Kyseiset tehtävät, jotka ovat johtaneet riskinarvion perusteella tehtävän välittämiseen, on välitetty yksilöivine tietoineen poliisille. Tiedot saattavat löytyä suoraan heidän tietojärjestelmistään

Annettujen tietojen valossa tietoja päädyttiin hakemaan suoraan Poliisihallitukselta.

Poliisihallitukseen oltiin ensimmäisen kerran yhteydessä elokuussa 2020 ja selvitettiin mahdollisuutta saada alueelliseen häiriötoimintaan liittyviä aineistoja tutkimuskäyttöön. Poliisihallituksen analyysitoiminnon edustajat arvioivat marraskuussa tutkimuslupa-asian käsittelyn vievän noin kuusi kuukautta. Tietopyyntöä pyydettiin tarkentamaan, mutta poliisihallitus ei pystynyt tarjoamaan tarkempaa tietoa poliisin keräämästä aineistosta. Tutkija tiedusteli mahdollisuutta neuvotella tietojärjestelmien sisällöstä tapaamisen merkeissä, jotta lupahakemusta voidaan tarkentaa. Ehdotuksesta kieltäydyttiin resurssien puutteen vuoksi. Käytännössä tutkimuslupahakemuksessa piti tarkkaan eritellä aineistot, joita haetaan, mutta metadata-tiedostoa varten tulisi tehdä erillinen tutkimuslupahakemus.

Tutkimuslupahakemus tehtiin ilman metadatatiedostoa. Aineistot ja muuttujat eriteltiin sen julkisen tiedon perusteella, mitä on tiedettävissä Hätäkeskuslaitoksen keräämistä aineistoista. Hakemuksen viimeinen versio lähetettiin maaliskuussa 2021. Poliisihallitus myönsi tutkimusluvan huhtikuussa 2021, ja Hätäkeskuslaitos tulisi suorittamaan tietojen poiminnan.

Hätäkeskuslaitokselle lähetettiin tietojen poimintapyyntö ja tutkimuslupa 14.4.2021. Hätäkeskuslaitos toimitti vuosien 2019–2021 melu- ja häiriöilmoitukset 16.9.2021 ja Digia toimitti Hätäkeskuslaitoksen puolesta vuosien 2016–2018 aineistot 28.10.2021.

Avoin data
Hankkeen alussa kartoitettiin, voidaanko netistä löytyvää avointa paikkatietoa hyödyntää valtakunnallisissa riskianalyysimalleissa. Yleisesti ottaen vapaasti verkosta saatavissa oleva avoin data on harvoin koko maan kattavaa ja tieto on resoluutioltaan liian epätarkkaa varsinaisen hyödyn saamiseksi.

Avoindata.fi -sivu on Digi- ja väestötietoviraston ylläpitämä sivusto, joka on tarkoitettu avoimen datan julkaisemiseen ja hyödyntämiseen. Sivulta löytyvään paikkatietoaineistoihin on koottu useamman eri palveluntuottajan tarjoamia avoimia paikkatiedostoja. Sivulta löytyvä 250m² resoluution väestötietoruudukko katsottiin rakennuspaloennustemallin kannalta hyödylliseksi, mutta se kattaa vain pääkaupunkiseudun alueen. Väestötietoruudukko on Helsingin seudun ympäristöpalveluiden (HSY) ylläpitämä tiedosto ja se sisältää ruutukohtaista tietoa pääkaupunkiseudun väestön lukumäärästä, ikäjakaumasta ja asumisväljyydestä. Tietosuojasyistä tiedostosta on poistettu ruudut, joissa on 0–4 asukasta. Myös väestönrakenteen tiedot ovat poistettu ruuduista, joissa on 0–99 asukasta.

Suomen ympärikeskuksen (SYKE) Elinympäristön tietopalvelu Liiterissä on myös hyödyllistä paikkatietoa, mutta avoin tieto on resoluutioltaan niin epätarkkaa, että tiedostoja ei voi hyödyntää sellaisenaan hankkeen tarkoituksiin.

Turvallisuus- ja kemikaalivirasto Tukes, Kemu -rekisteri
Hankkeen alussa käytiin palaveri Tukesin edustajan kanssa, jossa esiteltiin KemuDigi tietojärjestelmä ja sen sisältämät aineistot. Tietokannassa ei ollut kuitenkaan tämän tutkimushankkeen kannalta olennaisia paikkatietoaineistoja, koska tiedot koskevat pääosin laitoksia. KemuDigissä on tosin yritysten kemikaalirekistereitä, jotka ovat liiketilojen kannalta oleellisia. Jos riskianalyysimalleja sovelletaan yritysten ja laitosten rakennuspaloriskien kartoittamiseen, KemuDigin aineistot tulisi ottaa huomioon.

Rakennusvalvonta
Hankkeen alussa pidettiin palaveri Helsingin rakennusvalvonnan Facta -tietojärjestelmän pääkäyttäjän ja tarkastusinsinöörin kanssa ja käytiin yhdessä läpi mitä Facta -tietojärjestelmä sisältää. Tietojärjestelmä ei ole valtakunnallinen ja sen sisällössä on alueellista vaihtelevuutta. Selvisi, että rakennusvalvonnan tietojärjestelmät pohjautuvat pitkälti DVVn RHR aineistoihin, eikä tietojärjestelmässä ole muita aineistoja, joista olisi hyötyä tutkimuksessa.

Säteilyturvakeskuksen (STUK) tietokannat
Säteilyturvakeskuksen tietokantojen sisällöstä keskusteltiin STUKin edustajan kanssa elokuussa 2020. Keskustelun perusteella tultiin siihen tulokseen, että hanketta varten asuntojen radonpitoisuus mittausten tiedot voisivat olla hyödyllisiä. Mittaukset tehdään asukkaan tai taloyhtiössä hallituksen omasta aloitteesta, jolla saattaa olla välitön vaikutus kiinteistön turvallisuuskulttuuriin ja välillinen vaikutus paloturvallisuuteen. Tietopyyntö välitettiin STUKin kirjaamoon ja sieltä STUKin lakimiehen harkintaan.

STUKin lakimies antoi kieltävän päätöksen tietopyynnölle lokakuussa 2020. STUKin näkemyksen mukaan pyydetyn aineiston otos ei ole riittävän kattava tai edustava tutkimusta ajatellen, pyydetyn datan ja tulipalojen syntymekanismin välinen syy-yhteys jää tutkimussuunnitelman perusteella avoimeksi eikä tutkimus tuottaisi viranomaisille tai yhteiskunnalle merkittävää lisäarvoa tai säteilyturvallisuutta edistäviä vaikutuksia.

STUKille toimitettiin lisätietoja, joilla oikaistiin harkinnassa käytettyjä seikkoja ja pyydettiin ottamaan tietopyyntö uudelleen harkintaan. STUKin lakimies ilmoitti marraskuussa, että toimitettujen tietojen perusteella asia voidaan ottaa uudelleen harkintaan. STUK antoi myönteisen päätöksen tietopyynnölle joulukuussa. Vastauksessa eriteltiin, että tietojen antamisesta peritään julkisuuslain 34 §:n mukaisesti maksu. Tutkimushankkeen myöhemmässä vaiheessa tultiin kuitenkin siihen tulokseen, että aineistojen kustannushyöty ei olisi tarpeeksi kannattava tutkimuksen kokonaisuutta ajatellen, eikä niitä otettu mukaan tutkimusaineistoon.

Tilannehuone (Helsinki GSE:n, VATT:n ja Tilastokeskus)
Tilannehuone on Helsinki GSEn perustama tietovaranto koronan taloudellisten vaikutusten seuraamista varten. Tietovarantoon on kerätty julkisia ja yksityisiä aineistoja usean eri tiedontuottajatahon ja instituution kanssa. Näihin kuuluvat muun muassa Tilastokeskus, valtiovarainministeriö, työ- ja elinkeinoministeriö, Suomen Pankki, veroviranomaiset ja KELA.

Hankkeen tutkija keskusteli Tilannehuoneen projektipäällikön kanssa elokuussa 2020 tietovarannon tietosisällöstä ja mahdollisesta yhteistyöstä tietojen hyödyntämiseen liittyen. Projektipäällikkö lupasi selvittää tietojen hyödyntämismahdollisuuksia ja palata asiaan. Asiasta tiedusteltiin uudestaan syyskuun lopussa, mutta siihen ei lopulta vastattu. Tultiin siihen tulokseen, että aineistoja varten tarvitaan joka tapauksessa tutkimusluvat niitä hallinoivilta tahoilta, joten aineistot haetaan erikseen.

Findata
Findata on sosiaali- ja terveysalan tietolupaviranomainen. Findatan kautta voi hakea sosiaali- ja terveysalan tietoja toissijaiseen käyttöön ja organisaation kautta voi hakea tietoa useammalta viranomaiselta samaan aikaan. Findatan tietolupahakemusta varten käyttäjä tarvitsee suomi.fi tunnukset. Findata veloittaa (2020) tietolupahakemuksesta koskevasta päätöksestä 1000 euroa. Aineiston käsittelystä peritään erillinen maksu.

Osa Findatalta haettavista aineistoa sisältää tietoa, jota voi käsitellä vain organisaation ylläpitämässä tietoturvallisessa etäkäyttöympäristössä. Etäkäyttöympäristöön ei voi viedä eikä sieltä voi tuoda tiedostoja, mutta Findatalle voi toimittaa ulkoisia aineistoja käyttöympäristöön vietäväksi. Etäkäyttöjärjestelmän käytöstä peritään kuukausittainen maksu, joka riippuu käyttöympäristön koosta ja mahdollisesti tarvittavista teknisistä räätälöinneistä.

Hanketta varten Findatalta haettiin KELAn, THLn, Eläketurvakeskuksen ja Digi- ja väestötietoviraston aineistoja. KELAlta haettiin eläkkeensaajan asumistuen saajat ja maksut (EAMAB), yleisen asumistuen saajat ja maksut, Kelan maksetut eläke-etuudet, Kelan eläke-etuuksien saajat, kuntoutuspalvelujen saajat ja maksut (täydennetty perustetiedoilla), lastenhoidon tukien maksut, lapsilisän maksut (BMA), opintotuen maksut (OMA), perustoimeentulotuen saajat (henkilöt) (F_SAAJAT_PC), työttömyysturvan maksutiedot (FMA) ja vammaistukien saajat. THL:tä haettiin toimeentulotukirekisteri ja alkoholijuomien kulutus -rekisteri sekä Eläketurvakeskukselta eläkerekisterin ja ansaintarekisterin tiedot. Digi- ja väestötietoviraston henkilötiedot haettiin Findatan kautta.

Tutkimuslupahakemus lähetettiin Findatalle 14.1.2021. Keväällä ja kesällä 2021 hakemuksen käsittelyajoista tiedusteltaessa Findata vastasi automaattisella sähköpostiviestillä. Viestissä kerrottiin, että hakemuskäsittelyssämme on ruuhkaa, eikä yksittäisiin hakemusten käsittelyaikaa koskeviin tiedusteluihin pystytä vastaamaan, koska kaikki resurssit käytetään hakemusruuhkan purkamiseen.

Findata otti hakemuksen tiimoilta yhteyttä ensimmäisen kerran 20.9.2021, jolloin pyydettiin lisätäydennyksiä hakemuksen tueksi. Pyynnön mukaan hakemusta tarkennettiin, muokattiin ja hakemuksen tueksi luotiin päivitetty tutkimussuunnitelma, joka vastaa yksityiskohtaisemmin Findatan kysymyksiin. Hakemuksen uudelleen lähettämisen jälkeen Findata pyysi useita lisätäydennyksiä loka- ja marraskuun aikana. Lisätäydennyspyynnöt liittyi-

vät eri tietolupaviranomaisten kysymyksiin. Lisätäydennyksiä lähetettiin sitä mukaa, kun niitä pyydettiin. 10.12.2021 Findata ilmoitti, että tekoälyä hyödyntävissä hankkeissa on Findata hiljattain alkanut edellyttää hakemuksen liitteeksi myös yleisen tietosuoja-asetuksen 35 artiklan 1 kohdan mukaista vaikutustenarviointia.

Hakemusprosessin hitauden takia aineistoja ei ehditty saada tätä tutkimushanketta varten, mutta hakemusta muokattiin siten, että samalla hakemuksella haetaan aineistoja tämän tutkimushankkeen jatkohankkeelle 2022 varten. Findatan hakemusprosessi keskeytettiin toistaiseksi ja sitä jatketaan jatkohankkeen aikana.

Tilastokeskus
Tilastokeskukselta haettiin samoja ja samankaltaisia sosioekonomisia muuttujia, kuin Findatalta. Koska Tilastokeskuksen hakemuskäsittely on maksuton, tutkimuslupaa haettiin samaan aikaan sekä Tilastokeskukselta että Findatalta.

Hankkeen tutkija keskusteli Tilastokeskuksen lakimiehen kanssa mahdollisuudesta saada Tilastokeskukselta henkilötason tietoja ja vaihtoehtoisesti mahdollisuudesta saada tiedot pseudonyymina yhdistettynä osoitetietoihin. Tämä ei ole tietosuojasyistä mahdollista eikä asiaan myönnetä poikkeuksia. Tilastokeskukselta voidaan kuitenkin hakea sosioekonomisia tietoja, jotka voi yhdistää 250 m^2 resoluution tarkkuudella olevaan sijaintitietoruudukkoon.

Tilastokeskukselle lähetettiin tutkimuslupahakemus, jossa oli eritelty tarvittavat muuttujatekijät FIRM_EMPENT ja FOLK valmisaineisto moduuleista. Haettavia aineistoja voi käsitellä vain Tilastokeskuksen tietoturvallisessa etäkäyttöympäristössä FIONAssa. Hakemus lähetettiin 30.11.2020. Hakemus otettiin käsittelyyn ja Tilastokeskus pyysi lisätarkennuksia, johon vastattiin. Valmisaineistopyyntöjen käsittelyajaksi arvioitiin noin kuukausi. Tilastokeskus kirjasi viimeisimmän version hakemuksesta käsittelyyn 28.12.2020.

Tilastokeskukselta otettiin uudelleen yhteyttä 9.2.2021 ja hanketta varten nimitetty yhteyshenkilö kysyi tarkentavia lisäkysymyksiä, joihin vastattiin. Tilastokeskus kysyi lisäkysymyksiä ja täydennyspyyntöjä tasaisin välein helmi-, maalis- ja huhtikuun aikana. Pyyntöihin vastattiin sitä mukaan, kun niitä esitettiin. 25.5.2021 yhteyshenkilö ilmoitti, että hakemus oli tilastoeettisen lautakunnan käsittelyssä, jossa todettiin, että koska aineistopyyntö on

laaja, tarvitaan vielä yksityiskohtaisempi perustelu, joka avaa aineistojen tarpeellisuutta jokaiselta pyydetyltä aineistovuodelta. Pyynnön perusteella tehtiin laaja selvitys seikoista, joiden osalta pyydettiin tarkennuksia ja muokattu hakemus lisätietoineen lähetettiin Tilastokeskukselle 8.7.2021.

Käyttölupa hyväksyttiin 1.9.2021. Tilastokeskus toimitti organisaatiosopimuksen allekirjoitettavaksi ja tietoturvaselvityksen täytettäväksi 29.9.2021. Sopimukset ja selvitykset toimitettiin lokakuun alussa. Tilastokeskus avasi etäkäyttöympäristön tutkijalle 3.11.2021, noin 11 kuukautta hakemuksen jättämisen jälkeen.

Tilastokeskuksen kanssa tehdyn sopimuksen mukaan etäkäyttöjärjestelmään toimitetaan Digi- ja väestötietoviraston rakennus- ja huoneistorekisteri, Hätäkeskuslaitoksen melu- ja häiriöilmoitukset sekä PRONTOn aineistot. Sopimuksen mukaan Tilastokeskus pseudonymisoi aineistot ja varmistaa, ettei etäkäyttöjärjestelmään viedä paikkatietoa, joka osuu 250m² alueille, joissa asuu alle kolme henkilöä. Tilastokeskus ei kuitenkaan pystynyt teknisten haasteiden takia itse suodattamaan aineistoista kaikkia tietoja, jotka osuivat yli kolmen henkilön ruudukoille. Joulukuun aikana Tilastokeskus toimitti ulkoisista aineistoista seuraavat:

- 20 prosenttia rakennuspaloista ja palovaaroista (14 868 kpl, 2008–2020)
- 30 prosenttia rakennus- ja huoneistorekisterin tiedoista (847 981 kpl)
- 58 prosenttia melu- ja häiriöilmoituksista vuosilta 2016–2018 (107 449 kpl)
- 0 prosenttia melu- ja häiriöilmoituksista vuosilta 2019–2020 (0kpl)

Hankkeen tutkija tarjoutui useampaan otteeseen tekemään aineistojen linkityksen Tilastokeskuksen puolesta, jotta ulkopuolisista aineistoista saataisiin vietyä enemmän tietorivejä. Linkitystä varten tarvitaan tieto siitä, missä alle kolmen henkilön ruudut sijaitsevat, jotta ne voidaan poistaa ulkoisista aineistoista. Useista yrityksistä huolimatta Tilastokeskus ei toimittanut ruudukkoa. Etäkäyttöympäristö sulkeutui 31.12.2021. Tilastokeskus ilmoitti tammikuun 2022 puolivälissä, että linkitykseen vaaditun ruudukon voisi sittenkin ostaa Tilastokeskukselta tutkimuskäyttöön.

Aineistojen käytölle haettiin jatkoaikaa helmi/maaliskuun ajaksi, jotta ulkopuoliset aineistot saataisiin vietyä etäkäyttöympäristöön ja otettua huomioon analyyseissä. Sopimuksen mukaan Tilastokeskus tekee alle kolmen

henkilön ruututietojen karsinnan, mutta sitä ei pystytty tekemään kohtuullisessa ajassa, joten päädyttiin ostamaan linkitykseen vaadittava ruututietoaineisto ja tietorivien karsinta tehtiin Tilastokeskuksen puolesta. Tämän hankkeen rakennuspalojen riskimallinnukset tehtiin niiden aineistojen perusteella, jotka Tilastokeskus sai vietyä etäkäyttöympäristöön 2021 lopussa. Kevään 2022 kokeiluista julkaistaan erillinen raportti.

27

SUOMEN PELASTUSALAN KESKUSJÄRJESTÖ
SPEK TUTKII
SPEK

Ennakoivan analytiikan tutkimushankkeessa selvitettiin, miten ennakoivan analytiikan menetelmiä, esimerkiksi koneoppimista voidaan hyödyntää pelastustoimen riskianalyysitoiminnan tukena, erityisesti rakennuspalojen mallintamisessa Suomessa.

Kirjallisuuskatsauksen ja käytännön kokeilujen avulla tunnistettiin keskeisiä ohjatun koneoppimisen algoritmeja, joilla on mahdollista kohdentaa pelastustoimen palveluita aiempaa riskiperusteisemmin. Samalla kartoitettiin uusia aineistoja ja tunnistettiin rakennuspaloja keskeisesti selittäviä tekijöitä, joita tulisi hyödyntää pelastustoimen riskianalyyseissä tulevaisuudessa.

Pääkaupunkiseudun avointen aineistojen kokeilun perusteella satunnaismetsä-algoritmi suoriutui parhaiten. Koko Suomen rakennuspaloriskiarvojen mallintamisessa asukasmäärä osoittautui kaikista painavammaksi rakennuspaloja selittäväksi tekijäksi ja toiseksi painavin muuttuja oli autolliset asukaskunnat. Työttömyyteen, toimeentuloon ja eläkkeeseen liittyvät muuttujat olivat myös keskeisiä mallinnuksessa. Koko Suomen rakennuspaloriskiarvojen mallintamiseen luotiin kaksiosainen laskennallinen viitekehys. Testien perusteella XGBoost suoriutui mallinnuksista parhaiten ja soveltuisi hyvin pelastustoimen riskianalyysien muodostamiseen.

Suomen Pelastusalan Keskusjärjestö
SPEK

Suomen Pelastusalan Keskusjärjestö SPEK
Ratamestarinkatu 11, 00520 Helsinki
p. 09 476 112 spekinfo@spek.fi
www.spek.fi