Sydänsairauksien ennakointi koneoppimisen avulla

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Toukokuu 2025
Santeri Louna

TURUN YLIOPISTO
Tietotekniikan laitos

Santeri Louna: Sydänsairauksien ennakointi koneoppimisen avulla

LuK-tutkielma, 31 s.
Tietojenkäsittelytiede
Toukokuu 2025

Sydän- ja verisuonisairaudet ovat merkittävä kansanterveydellinen haaste, ja niiden varhainen havaitseminen on keskeistä potilaiden hoidon ja ennusteen parantamiseksi. Tämä kandidaatintutkielma tarkastelee koneoppimisen menetelmien soveltamista sydänsairauksien havaitsemisessa. Työssä käsitellään yleisimpiä sydänsairauksia ja niiden diagnosointia, sekä esitellään keskeisiä koneoppimisen lähestymistapoja, kuten päätöspuut, tukivektorikoneet ja satunnaismetsät. Tulosten perusteella tehokkaimmat algoritmit saavuttivat jopa yli 85 % tarkkuuden. Koneoppiminen tarjoaa lupaavan työkalun sydänsairauksien havaitsemisen tueksi, mutta sen käyttö vaatii edelleen tarkkaa mallien validointia ja eettisten näkökulmien huomiointia kliinisessä käytössä.

Asiasanat: koneoppiminen, sydänsairaudet

UNIVERSITY OF TURKU
Department of Computing

Santeri Louna: Sydänsairauksien ennakointi koneoppimisen avulla

Bachelor's Thesis, 31 p.
Laboratory
May 2025

Cardiovascular diseases are a major public health challenge, and their early detection is essential to improve patient care and prognosis. This thesis explores the application of machine learning methods to the detection of heart disease. It discusses the most common heart diseases and their diagnosis, and introduces key machine learning approaches such as decision trees, support vector machines and random forests. The results show that the most efficient algorithms achieved accuracies of up to more than 85 %. Machine learning offers a promising tool to support the detection of heart disease, but its use still requires rigorous model validation and consideration of ethical aspects in clinical applications.

Keywords: machine learning, heart disease

Sisällys

Kuvat

Taulukot

1 Johdanto

1.1 Aiheen taustoitus

Sydän- ja verisuonisairaudet ovat maailmanlaajuisesti merkittävä kuolleisuuden ja sairastavuuden aiheuttaja. Maailman terveysjärjestön (WHO) mukaan sydän- ja verisuonisairauksiin kuolee vuosittain noin 17,9 miljoonaa ihmistä, mikä vastaa noin 32 % kaikista maailman kuolemista [1]. Tavallisia sydänsairauksia ovat esimerkiksi sepelvaltimotauti, sydämen vajaatoiminta ja rytmihäiriöt. Monet näistä sairauksista kehittyvät hitaasti ja voivat olla pitkään oireettomia, minkä vuoksi niiden varhainen toteaminen ja riskitekijöiden tunnistaminen on keskeistä sairauksien ehkäisyssä ja hoidossa.

Koneoppiminen (engl. machine learning) on tekoälyn osa-alue, jossa tietokoneohjelmat oppivat tunnistamaan malleja ja tekemään ennusteita datan perusteella ilman tarkkaa ennalta määriteltyä ohjelmointia. Koneoppimismenetelmiä on hyödynnetty laajasti erilaisissa sovelluksissa, kuten kuvantunnistuksessa, kielimallinnuksessa ja käyttäytymisanalytiikassa. Viime vuosina koneoppiminen on noussut tärkeäksi työkaluksi myös lääketieteellisessä tutkimuksessa, erityisesti diagnostiikassa ja riskinarvioinnissa [2].

Sydänsairauksien kohdalla koneoppimista voidaan hyödyntää esimerkiksi potilastietojen, laboratoriotulosten ja kuvantamisdatan perusteella tapahtuvaan sairauksien ennustamiseen ja diagnosointiin. Koneoppimismallit voivat auttaa tunnista-

maan sellaisia potilaita, joilla on kohonnut riski sairastua vakavaan sydänsairauteen, ja näin tukea terveydenhuollon ammattilaisten päätöksentekoa [3]. Koneoppimisen avulla on myös mahdollista löytää uusia yhteyksiä, joita ei perinteisillä menetelmillä ole havaittu [4].

1.2 Tutkimuskysymykset ja rajaukset

Tutkielmassa on tarkoitus vastata seuraaviin tutkimuskysymyksiin:

- Tutkimuskysymys 1: Mitä haasteita sydänsairauksien havaitsemiseen liittyy?
- Tutkimuskysymys 2: Mitä koneoppimismenetelmiä voidaan hyödyntää sydänsairauksien havainnoinnissa?

Tutkimus on toteutettu kirjallisuuskatsauksena. Tutkimuksen tulos -osion lähteitä on rajattu englanninkielisiin, vuoden 2015 jälkeen julkaistuihin artikkeleihin. Rajauksen avulla pyritään mahdollisimman ajankohtaiseen tutkimustietoon.

Aineistonhaku suoritettiin ScienceDirectin tietokannasta. Hakulauseena toimi ("machine learning"OR ml) AND "heart disease"AND "detection".

Lopulliseen tarkasteluun valikoitui kaksi tutkimusta, Pathan ym. [3] ja Singh ym. [5], sillä ne vastasivat parhaiten tutkimuskysymyksiin, perustuivat samaan laajasti hyväksyttyyn Framinghamin sydänsairausaineistoon ja hyödynsivät erilaisia koneoppimismenetelmiä. Tämä mahdollistaa sekä tulosten vertailtavuuden että menetelmien monipuolisen tarkastelun.

1.3 Tutkielman rakenne

Tutkielma rakentuu neljästä pääluvusta. Ensimmäisessä luvussa käsitellään sydänsairauksia yleisellä tasolla. Luvussa esitellään, mitä sydänsairaudet ovat, mitkä ovat

niiden yleisimmät riskitekijät ja miksi niiden varhainen havaitseminen on tärkeää yksilön terveyden ja terveydenhuoltojärjestelmän näkökulmasta.

Toinen luku keskittyy koneoppimiseen. Siinä selitetään, mitä koneoppiminen tarkoittaa ja miten sitä voidaan hyödyntää lääketieteellisessä kontekstissa, erityisesti sydänsairauksien havaitsemisessa. Luvussa esitellään myös keskeiset koneoppimisen menetelmät, jotka ovat keskiössä tutkielman tarkastelussa.

Kolmannessa luvussa tarkastellaan tarkemmin kahta valittua tutkimusta: Pathan ym. [3] ja Singh ym. [5]. Luvussa esitellään näiden tutkimusten käyttämä aineisto, menetelmät sekä saavutetut tulokset. Lisäksi vertaillaan tutkimusten lähestymistapoja ja arvioidaan niiden vahvuuksia sekä rajoitteita.

Viimeisessä eli neljännessä luvussa kootaan yhteen aiempien lukujen keskeiset havainnot. Luvussa pohditaan, miten koneoppiminen voi konkreettisesti tukea sydänsairauksien varhaista tunnistamista, ja vastataan asetettuihin tutkimuskysymyksiin. Lisäksi arvioidaan tutkimusten perusteella koneoppimisen hyötyjä ja haasteita osana tulevaisuuden terveydenhuoltoa.

Generatiivisen tekoälyn käyttö

Tutkielmassa on käytetty apuna generatiivista tekoälyä. Chat-GPT:tä käytettiin apuna tutkielman kielen tarkistuksessa, jossa sovellukselta pyydettiin yksittäisten lauseiden muotoilemista luontevammaksi. Lisäksi tekoälyä käytettiin käännösten etsimisessä, sillä osalle tutkielmassa käytetystä sanastosta ei ole vakiintunutta suomennosta. Chat-GPT auttoi myös muuntamaan Pathanin tutkimuksessa esitetyt taulukot LaTex muotoon. Generatiivista tekoälyä käytettiin apuna taulukoiden 4.1, 4.2, 4.3, 4.4, 4.5 ja 4.6 luomisessa.

2 Sydänsairaudet

Sydän- ja verisuonisairaudet ovat kansainvälisen terveysorganisaatio WHO:n mukaan maailman yleisin kuolinsyy. Sydän- ja verisuonisairauksiin kuolee vuosittain 17,9 miljoonaa ihmistä. Yleisimpiä tauteja ovat sepelvaltimotauti, erilaiset reumaattiset sydänsairaudet sekä aivoverenkierronhäiriöt. 80 % sydän- ja verisuonisairauksien kuolemista aiheutuu sydänkohtauksesta sekä aivohalvauksesta. Näistä 1/3 tapahtuu ennenaikaisesti alle 70-vuotiaille henkilöille. [1]

2.1 Yleisimmät sydänsairaudet

Sepelvaltimotauti

Sepelvaltimotauti johtuu sepelvaltimoiden ahtautumisesta, mikä rajoittaa sydämen hapensaantia [6]. Tämä voi johtaa rintakipuun (angina pectoris) ja pahimmillaan sydäninfarktiin. Taudin riskitekijöitä ovat korkea verenpaine, korkea kolesteroli, tupakointi ja diabetes [7].

Sydämen vajaatoiminta

Sydämen vajaatoiminta tarkoittaa sydämen kyvyttömyyttä pumpata verta tehokkaasti. Se voi johtua pitkäaikaisista sydänsairauksista, kuten verenpainetaudista tai sepelvaltimotaudista. Tyypillisiä oireita ovat hengenahdistus, turvotus ja väsymys [8].

Rytmihäiriöt

Rytmihäiriöt ovat sydämen sähköisen toiminnan häiriöitä, jotka voivat aiheuttaa epäsäännöllistä sykettä. Yleisimmät rytmihäiriöt ovat eteisvärinä ja kammioperäiset rytmihäiriöt. Eteisvärinä lisää merkittävästi aivohalvauksen riskiä [9].

Sydäninfarkti

Sydäninfarkti tapahtuu, kun sepelvaltimon tukkeutuminen estää verenkierron sydänlihakseen, aiheuttaen kudosvaurioita. Nopea hoito, kuten liuotushoito tai pallolaajennus, on elintärkeää. Sydäninfarktin riskitekijöitä ovat korkea verenpaine, tupakointi, diabetes ja korkea kolesteroli [10].

Läppäviat

Sydänläppäviat voivat olla synnynnäisiä tai kehittyä esimerkiksi reumakuumeen tai ikääntymisen seurauksena. Läppäviat voivat aiheuttaa sydämen vajaatoimintaa, mikäli niitä ei hoideta ajoissa [11].

Sydän- ja verisuonisairaudet voivat usein olla pitkään oireettomia. Ensimmäinen havaittu oire voikin olla sydänkohtaus tai aivohalvaus. [1] Muita oireita ovat esimerkiksi rintakipu sekä hengitysvaikeudet. [12]

Monet sydänsairauksista ovat osittain perinnöllisiä [13]. Myös ulkoiset tekijät vaikuttavat sairauksien syntyyn. Riskiä aiheuttavia ulkoisia tekijöitä ovat esimerkiksi liikkumattomuus, tupakointi, liiallinen alkoholin käyttö sekä huono ruokavalio. Riskien aiheuttamat vaikutukset voivat näkyä yksilössä esimerkiksi ylipainona, kohonneena verenpaineena sekä kasvaneena veren kolesteroli- tai glukoosipitoisuutena. [1]

Sydän- ja verisuonisairauksia hoidetaan monin eri tavoin. Kohonnutta verenpainetta sekä sepelvaltimotautia hoidettaessa hoito aloitetaan usein elämäntapojen kuntoon laittamisella. Huomioon otettavia asioita ovat riittävä liikunta, terveellinen

ruokavalio sekä erityisesti tupakoinnin lopettaminen. [14]

Myös lääkitys on tärkeä osa hoitoa sydän- ja verisuonisairauksissa. Sepelvaltimotaudissa lääkitys voi olla joko sairauden pysäyttävä tai oireita helpottava lääkitys. Sydämen vajaatoimintaa hoidetaan ensisijaisesti lääkityksen avulla. [14]

Muita mahdollisia hoitoja sydänsairauksissa on esimerkiksi erilaiset leikkaushoidot. Sepelvaltimotautia voidaan hoitaa pallolaajennuksella. Pallolaajennuksessa valtimoiden tukkeumia avataan suonen sisältä päin. Sydämentahdistinta käytetään apuna sydämen vajaatoiminnassa sekä rytmihäiriöissä. [14]

2.2 Sydänsairauksien diagnosointi

Sydänsairaudet voivat pitkään olla oireettomia, ja paljastua sattumalta [15]. Sydänsairauksia pyritään diagnosoimaan mahdollisimman varhaisessa vaiheessa. Perinteiset sydänsairauksien havainnointimenetelmät ovat usein paljon aikaa vieviä ja kallita [4]. Aikaisin diagnosoituna sydänsairauksien hoito ja niistä parantuminen helpottuvat. Perinteisiä sydänsairauksien diagnosointimenetelmiä ovat:

Elektrokardiografia (EKG)

EKG on yksi tärkeimmistä ja yleisimmistä menetelmistä sydänsairauksien diagnosoinnissa. Se mittaa sydämen sähköistä toimintaa ja voi paljastaa rytmihäiriöitä, sydänlihaksen iskemiaa ja muita sydämen toiminnan poikkeavuuksia [16].

Kaikukardiografia (ECHO)

Kaikukardiografia on ultraäänitutkimus, joka antaa kuvan sydämen rakenteista ja toiminnasta. Se auttaa arvioimaan sydänläppien toimintaa sekä sydänlihaksen pumppaustehokkuutta. [17]

Sydänkoe ja rasitustesti

Sydämen rasituskoe tehdään yleensä juoksumatolla tai polkupyörällä samalla seuraten EKG:tä. Se auttaa tunnistamaan sepelvaltimotaudin ja arvioimaan sydämen suorituskykyä. [18]

Verikokeet

Sydänsairauksien diagnosoinnissa käytetään erilaisia verikokeita, kuten troponiinitestiä, joka mittaa sydänlihassolujen vaurioita ja on erityisen tärkeä sydäninfarktin diagnosoinnissa [10].

Sepelvaltimoiden varjoainekuvaus (angiografia)

Varjoainekuvausta käytetään erityisesti sepelvaltimotaudin diagnosointiin. Varjoaine ruiskutetaan sepelvaltimoihin, ja röntgenkuvien avulla nähdään mahdolliset tukokset. [19]

3 Koneoppiminen

Tekoälyllä tarkoitetaan laitteita tai ohjelmia, jotka pystyvät oppimaan ja auttavat ihmisiä niissä toiminnoissa, joita varten se on suunniteltu [20]. Tekoäly suorittaa monimutkaisia tehtäviä hyvin samankaltaisesti kuin ihmisetkin [21]. Neuroverkot, jotka pyrkivät jäljittelemään ihmisten aivojen toimintaa, keksittiin jo 1940-luvulla. Kiinnostus neuroverkkoja kohtaan kohosi 1990-luvulla, mutta sen aikaiset tietokoneet eivät pystyneet käsittelemään koulutuksessa vaadittua valtavaa datamäärää. 2010-luvulla tietokoneet olivat jo huomattavasti edistyneempiä, sekä dataa oli saatavilla entistä enemmän, joten kiinnostus neuroverkkoja kohtaan kasvoi taas. [20] Deloitten vuoden 2020 kyselytutkimuksen mukaan jo 67 % yrityksistä käytti koneoppimista apunaan, ja 97 % vastanneista yrityksistä aikoi käyttää sitä tulevana vuonna [21].

Nykyisten tekoälysovellusten oppiminen perustuu suurelta osin datasta oppimiseen. Ihmisille helppojen tehtävien ratkaisu voi olla tekoälylle hankalaa. Esimerkiksi kuvista, äänistä ja erilaisista tapahtumaketjuista ihmisen on helppo oppia tarvitsemansa informaatio. Tekoälylle nämä asiat voivat tuottaa hankaluuksia, sillä dataa voi olla vaikea saada tietokoneelle ymmärrettävään muotoon. [20]

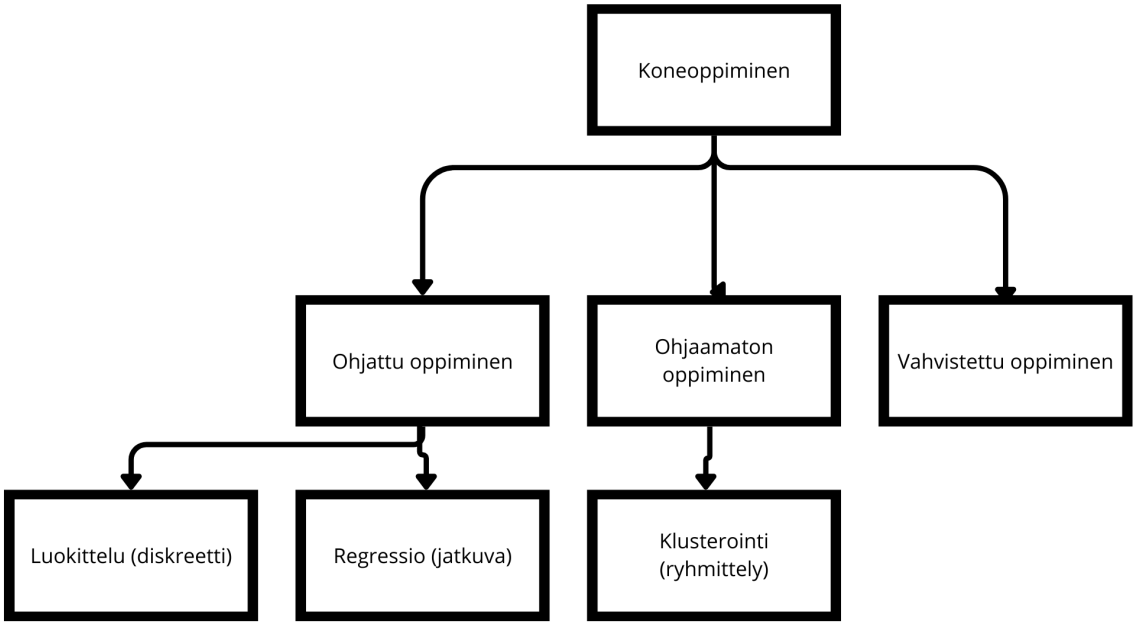
Koneoppiminen on yksi tekoälyn osa-alueista. Koneoppimisessa kone saadaan itsenäisesti oppimaan toistuvista tapahtumista. Itsenäisellä oppimisella tarkoitetaan, ettei ihmisen tarvitse opettaa konetta erikseen. Tavoitteena koneoppimisessa on ohjelmiston parempi toiminta saadun datan ja mahdollisen käyttäjän toiminnan pohjalta. Tavoitteina koneoppimisessa voidaan pitää tiedon tulkitsemisen automatisoin-

tia, sekä koneiden älykkyyden lisäämistä. [20]

Koneoppimista hyödyntävä ohjelma voi olla kuvaileva, ennustava tai ohjaava. Kuvaileva ohjelma selittää syötetyn datan avulla mitä on tapahtunut. Ennustava ohjelma pyrkii ennakoimaan saadun datan perusteella, mitä tulee tapahtumaan. Ohjaavan ohjelman tarkoitus on suositella toimenpiteitä, joita olisi syytä tehdä seuraavaksi, ja jotka sen mielestä auttaisivat pääsemään toivottuun lopputulokseen. [21]

Koneoppiminen jaetaan tyypillisesti kolmeen eri alakategoriaan

- Ohjattu oppiminen
- Ohjaamaton oppiminen
- Vahvistettu oppiminen



Kuva 3.1: Koneoppimisen jaottelu.

Ohjatussa oppimisessa konetta pyritään opettamaan valmiiksi luokitellun datan avulla. Esimerkiksi tietokoneelle voidaan antaa kuvia, joista osa on luokiteltu kissakuviksi, ja loput kuvista on muita kuvia. Kone oppii tunnistamaan luokitelluista

kuvista kissoille tyypillisiä piirteitä, joiden avulla se pystyy tekemään luokittelua jatkossakin. Ohjattu oppiminen on yleisin käytössä olevista koneoppimisen kategorioista. [21]

Ohjattu oppiminen voidaan lisäksi jakaa kahteen alakategoriaan, luokitteluun ja regressioon, tavoitteensa perusteella. Luokittelussa tavoitedata on tarkoitus jakaa erilaisiin ryhmiin, kuten edeltävässä kissakuvaesimerkissä. Regressiossa data on jatkuvaa, kuten lämpötila, eikä ohjelman ole tarkoitus luokitella dataa. [20]

Ohjaamattomassa oppimisessa ohjelma etsii samankaltaisuuksia annetusta luokittelemattomasta syötteestä [21]. Samankaltaisuuden perusteella ohjelma jakaa datan ryhmiin. Tätä ryhmiinjakotekniikkaa kutsutaan klusteroinniksi [22]. Ohjaamaton oppiminen muistuttaa paljon ihmisen tapaa oppia ja tunnistaa datan piirteitä[20]. Ohjaamattoman oppimisen etuna on sen kyky löytää myös samankaltaisuuksia, joita ihminen ei itse välttämättä tietoisesti edes etsi [21]. Esimerkki ohjaamattomasta oppimisesta olisi sydänsairauspotilaiden ryhmittely heidän oireprofiiliensa perusteella.

Vahvistetussa oppimisessa kone oppii kokeilemalla. Onnistuneesta päätöksestä kone saa positiivista palautetta, kun taas epäonnistuneesta päätöksestä palaute on negatiivista. Ohjelmisto tekee valinnat aiemmin positiivista palautetta tuottaneiden, ja vielä tuntemattomien vaihtoehtojen välillä. Laitteen tavoitteena on kasvattaa positiivisen palautteen määrää, sekä vähentää negatiivista palautetta. Vahvistettua oppimista käytetään muun muassa itseohjautuvissa autoissa. [20]

3.1 Ohjatun oppimisen menetelmiä

Tässä tutkielmassa keskitytään erityisesti ohjatun oppimisen menetelmiin. Valintaan päädyttiin, koska suurin osa löydetystä lähdemateriaalista käytti apunaan ohjatun oppimisen menetelmiä.

Päätöspuu

Päätöspuut (engl. decision trees) ovat yksinkertaisia ohjatussa koneoppimisessa käytettyjä malleja, joita voidaan käyttää niin luokittelu-, kuin regressio-ongelmissakin. Päätöspuu koostuu juurisolmusta, oksista, sisäsolmuista ja lehtisolmuista. Sisäsolmut tekevät päätöksiä, joiden lopputulokset kuvataan lehtisolmuilla. Päätöspuun on tarkoitus esittää kaikki mahdolliset lopputulokset. Puiden kasvaessa liian suuriksi, on vaarana, että sisäsolmulle saapuva aineisto on liian vähäistä, jolloin osapuun (engl. subtree) data voi fragmentoitua. Tämä taas voi johtaa helposti mallin ylisovittamiseen. [23] Päätöspuut ovat helposti toteutettavia ja käyttävät vain vähän muistia [20].

Tukivektorikone

Tukivektorikonetta (engl. support vector machine) käytetään binääriseen luokitteluongelmaan. Se etsii aineistosta lineaarista päätöspintaa, jonka avulla pystytään erottamaan kahteen eri luokkaan kuuluvat datapisteet toisistaan. Datan ollessa lineaarisesti erotettava, paras päätöspinta on se joka erottelee luokat toisistaan suurimmalla marginaalilla. Mikäli aineisto ei ole lineaarisesti erotettava, väärälle puolelle päätöspintaa oleville pisteille käytetään virhefunktiota. [20]

Satunnaismetsä

Satunnaismetsät (engl. random forests) yhdistelevät useita päätöspuita ennusteen tuottamiseksi. Kutakin päätöspuuta koulutetaan erillisellä osalla aineistoa, ja lopuksi nämä päätökset yhdistetään yhdeksi tarkaksi ennusteeksi. Satunnaismetsät kykenevät havaitsemaan moniulotteista aineistoa ja monimutkaisia piirteiden vuorovaikutuksia. Luokitteluongelmissa jokainen puu tekee päätöksen aineiston perusteella, jonka jälkeen eniten kannatusta saanut päätös on satunnaismetsän lopullinen päätös. Regressio-ongelmissa päätöspuiden vastausten keskiarvo on satunnaismetsän

lopullinen päätös. [5]

3.2 Aineiston esikäsittely

Koneoppimisessa käytetty aineisto voi usein olla esimerkiksi alun perin virheellistä, se voi tulla useasta eri lähteestä tai tietoa on voitu syöttää useaan kertaan. Jotta aineisto olisi käyttökelpoista algoritmin opettamiseen, aineisto täytyy esikäsitellä (engl. data preprocessing). [20]

Ensimmäiset esikäsittelyn vaiheet ovat aineiston siivous (engl. data cleaning) ja aineiston yhdistäminen (engl. data integration). Siivouksessa korjataan virheellisesti tai useaan kertaan syötettyjä tietoja, sekä puutteellisia tietoja. Aineiston yhdistämisessä yhdistetään useasta eri lähteestä olevat tiedot samassa muodossa olevaksi kokonaisuudeksi. [20]

Aineiston vähentämisellä pyritään analysoinnin nopeuttamiseen sekä selkeyttämiseen. Tarkoitus on säilyttää mahdollisimman paljon olennaista tietoa alkuperäisestä aineistosta. Aineiston vähentämisen menetelmiä ovat piirteiden valinta (engl. feature selection), piirreirrotus (engl. feature extraction) ja näytteenotto (engl. sampling). Piirteiden valinnalla (engl. feature selection) tarkoitetaan prosessia, jossa valitaan aineistosta oleellisimmat, tulokseen eniten vaikuttavat piirteet omaksi osajoukoksi [24]. Piirteiden valinnan hyötyjä on datan laadun parantuminen, datan keräämisen tehostaminen, ja laskenta-ajan väheneminen [24]. Piirreirrotuksessa pyritään esittämään alkuperäinen muuttujajoukko pienemmällä määrällä ulottuvuuksia. Yleisin piirreirrotuksen menetelmä on pääkomponenttianalyysi (engl. principal component analysis), jossa luodaan uusi joukko muuttujia, joihin pyritään saamaan alkuperäisen aineiston oleellisimmat olennaisuudet. Näytteenotossa aineistosta valitaan osajoukko analyysia varten. [20]

Aineiston muutoksen menetelmiä ovat normalisointi (engl. normalisation), diskretisointi (engl. discretation) sekä ominaisuuksien luonti (engl. feature generation).

Aineiston muutoksilla pyritään valmistelemaan muuttujat sopivaan muotoon analysointia varten. Normalisoinnissa eri muuttujien arvot skaalataan keskenään vertailukelpoisiksi. Diskretoinnissa jatkuvat muuttujat muutetaan diskreeteiksi. Syynä tähän voi olla esimerkiksi, että ennustusmenetelmät osaavat käyttää vain diskreettejä muuttujia. Ominaisuuksien luonnissa aineistoon tehdään kokonaan uusia ominaisuuksia jo olemassa olevien muuttujien pohjalta. [20]

3.3 Suorituksen arviointi

Luokittelumallien suorituskykyä arvioidaan usein sekaannusmatriisin (engl. confusion matrix) avulla, jossa mallin tekemät luokitukset jaotellaan neljään päätyyppiin. Aidosti positiivinen (engl. true positive, TP) tarkoittaa tapausta, jossa algoritmi tunnistaa sydänsairauden oikein sairaalla potilaalla. Aidosti negatiivinen (engl. true negative, TN) puolestaan kuvaa tilannetta, jossa algoritmi päättelee oikein, että potilaalla ei ole sydänsairautta. Väärä negatiivinen (engl. false negative, FN) tarkoittaa virhettä, jossa algoritmi ei tunnista sydänsairautta potilaalla, jolla sellainen todellisuudessa on. Väärä positiivinen (engl. false positive, FP) taas viittaa siihen, että algoritmi virheellisesti luokittelee terveen potilaan sydänsairaaksi. Lääketieteellisessä kontekstissa väärä negatiivinen luokitus on usein vakavin, sillä se voi johtaa siihen, että potilas ei saa ajoissa tarvitsemaansa hoitoa [3].

Yleisimmin käytetyt arviointikriteerit koneoppimismalleille ovat:

- Täsmällisyys (engl. accuracy), joka mittaa oikein ennustettujen luokittelujen määrää [25].
- Tarkkuus (engl. precision), mittaa aidosti positiivisten määrää, kaikkiin positiivisesti luokiteltuihin [25].
- Herkkyys (engl. recall), jolla mitataan oikein luokiteltujen aidosti positiivisten määrää suhteessa kaikkiin aidosti positiivisiin [25].

- Väärien positiivisten määrä (engl. false positive rate), joka mittaa väärien negatiivisten määrää suhteessa kaikkiin aidosti negatiivisiin [25].
- ROC-arvo (engl. receiver operating characteristic curve) saadaan laskemalla todellisen positiivisen tuloksen osuus ja väärän positiivisen tuloksen osuus jokaisella mahdollisella kynnysarvolla [26].
- F1-piste (engl. F1-score), joka käyttää tarkkuuden ja herkkyyden harmonista keskiarvoa [27]

Täsmällisyyttä mittaava kaava on:

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{3.1}$$

F1-piste saadaan tarkkuuden ja herkkyyden harmonisena keskiarvona:

$$Precision = TP/(TP + FP) \tag{3.2}$$

$$Recall = TP/(TP + FN) \tag{3.3}$$

$$F1 - Score = 2(PrecisionxRecall)/(Precision + Recall) \tag{3.4}$$

ROC-arvo arvioi herkkyyden eli aidosti positiivisten (TPR), sekä väärien positiivisten suhdetta (FPR).

$$FPR = FP/(FP + FN) \tag{3.5}$$

Muita suorituksen arviointiin hyödyllisiä työkaluja ovat korrelaatioanalyysi, varianssianalyysi, sekä f-testi. Korrelaatioanalyysiä käytetään kahden muuttujan välisen riippuvuuden tutkimiseen laskemalla niille korrelaatiokerroin. Korrelaatioker-

rointen arvot sijoittuvat välille (-1, 1). Mitä suurempi on kertoimen poikkeama nollasta, sitä voimakkaampi on muuttujien välinen riippuvuus. [28] Varianssianalyysillä pyritään määrittämään onko useiden otosten keskiarvot peräisin samasta jakaumasta. F-testiä käytetään kahden tilastollisen suureen, esimerkiksi juuri varianssin, suhteen laskemiseen [29].

4 Tarkastellut tutkimukset

Tutkielmassa käsiteltävät kaksi tutkimusta, Pathan ym. [3] ja Singh ym. [5], on valittu erityisesti siksi, että molemmat hyödyntävät samaa, laajasti tunnettua ja validoitua Framinghamin sydänsairausdataa. Framinghamin aineisto on yksi tunnetuimmista sydän- ja verisuonitautien riskitekijöiden tutkimuksessa käytetyistä aineistoista, mikä tekee tutkimuksista keskenään vertailukelpoisia ja tieteellisesti merkityksellisiä. Valitut tutkimukset edustavat myös monipuolisesti koneoppimisen lähestymistapoja sydänsairauksien ennustamiseen: Pathan ym. painottavat logistista regressiota ja hermoverkkoja, kun taas Singh ym. tarkastelevat erityisesti puupohjaisia menetelmiä kuten päätöspuita ja satunnaismetsiä. Tämä menetelmällinen vaihtelu tarjoaa hyvän pohjan arvioida erilaisten koneoppimisalgoritmien soveltuvuutta sydänsairauksien havaitsemiseen sekä vertailla niiden suorituskykyä saman aineiston pohjalta.

Pathanin ym. [3] tutkimuksessa koulutusmateriaalina käytettiin Framingham Heart Studyn (FHS) aineistoa, sekä McKinseyn & Companyn CVD-aineistoa. CVD-aineisto sisälsi 29 072 potilaasta kerätyt tiedot. Jokaisesta potilaasta oli kerätty 12 attribuuttia potilaan terveydentilan määrittelyyn. Framinghamin aineisto sisälsi tietoja Framignhamin (Massachutes, Yhdysvallat) asukkaista, jotka osallistuivat sydänsairauksien tutkimukseen. Aineisto sisälsi tiedot 4 240 potilaasta, joille kullekin oli määritelty 15 attribuuttia. Kyseistä pakettia käytetään usein luokittelutehtävissä, joissa on tarkoitus määritellä potilaan todennäköisyyttä sairastua sepelvaltimo-

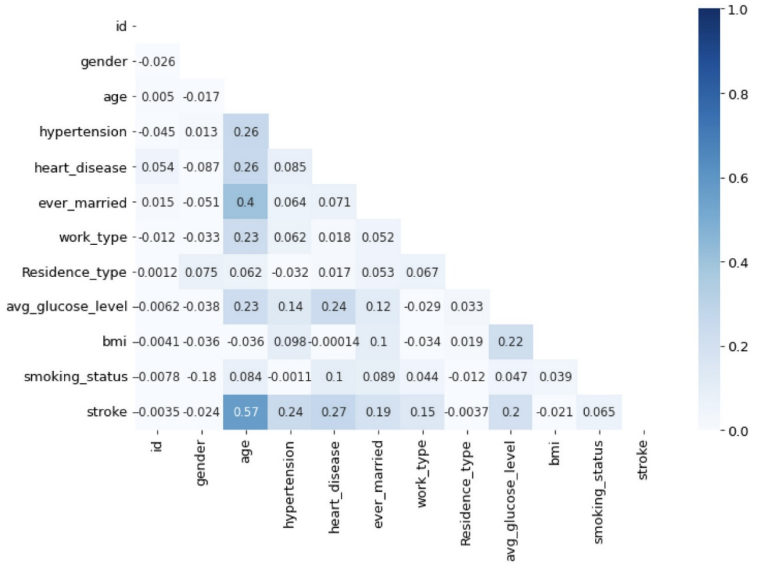
tautiin seuraavan kymmenen vuoden kuluessa. Molemmat aineistot sisälsivät monia yhteisiä attribuutteja kuten ikä, verensokeri, sukupuoli sekä potilaan tupakointi, joiden tiedetään olevan sidoksissa sydänsairauksiin. [3]

Muuttuja	**Selite**
i.d	potilaan tunniste
gender	sukupuoli ("mies": 0, "nainen": 1, "muu": 2)
age	potilaan ikä (jatkuva)
hypertension	korkea verenpaine ("kyllä":1, "ei":0)
heart_disease	sydänsairaus ("kyllä":1, "ei":0)
ever_married	siviilisääty ("kyllä":1, "ei":0)
work_type	työn tyyppi ("lapsi":0, "valtion työ":1, "ei ole työskennellyt":2, "yksityinen":3, "itse työllistetty":4)
residence_type	asuinalueen tyyppi ("maaseutu":0, "kaupunki":1)
avg_glucose_level	veren keskimääräinen glukoositaso (jatkuva)
bmi	painoindeksi (desimaaliluku)
smoking_status	tupakointitila ("ei koskaan tupakoinut":0, "aiemmin tupakoinut":1, "tupakoi":2)
stroke	aivohalvaus ("kyllä":1, "ei":0)

Taulukko 4.1: CVD-aineiston muuttujat ja niiden kuvaukset [3].

Muuttuja	**Selite**
age	potilaan ikä (jatkuva)
male	sukupuoli ("mies": 0, "nainen": 1)
education	koulutustaso (1–4)
currentSmoker	nykyinen tupakoija ("kyllä":1, "ei":0)
CigsPerDay	keskimääräinen päivittäinen savukemäärä (jatkuva)
BPMeds	käyttää verenpainelääkkeitä ("kyllä":1, "ei":0)
prevalentStroke	aiempi aivohalvaus ("kyllä":1, "ei":0)
prevalenHyp	korkea verenpaine ("kyllä":1, "ei":0)
diabetes	diabetes ("kyllä":1, "ei":0)
totChol	kokonaiskolesteroli (jatkuva)
sysBP	systolinen verenpaine (desimaali)
diaBP	diastolinen verenpaine (desimaali)
BMI	painoindeksi (desimaali)
HeartRate	syke (jatkuva)
glucose	glukoositaso (jatkuva)
TenYearCHD	riski sairastua sepelvaltimotautiin 10 vuoden sisällä ("kyllä":1, "ei":0)

Taulukko 4.2: Framingham-aineiston muuttujat ja niiden kuvaukset [3].

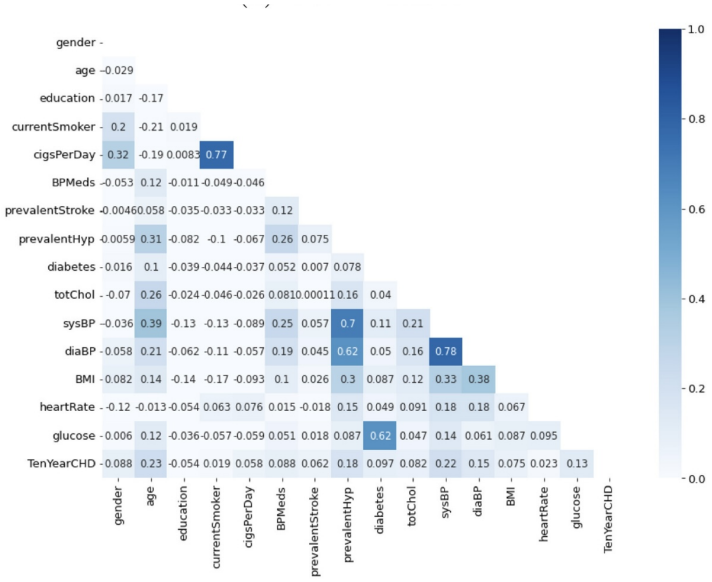


Kuva 4.1: CVD-aineiston korrelaatiokertoimet [3].

Aineiston käsittely

Aineistoja esikäsiteltiin tutkimusta varten. Datan esikäsittely helpottaa datan analysointia, sekä tehostaa koneoppimisalgoritmien tarkkuutta ja nopeutta. Esikäsittelyssä aineistosta poistettiin tyhjät arvot, sillä ne voivat vaikuttaa koneoppimismallien tarkkuuteen [30]. Lisäksi aineistoissa oli selkeää epätasapainoa sydänsairauksista kärsineiden ja terveiden välillä. CVD-aineistossa vain 548 potilasta oli kokenut aivoinfarktin oireita, ja FHS-aineistosta vain 557 potilastietoa osoitti kohonnutta sepelvaltimotaudin riskiä. Tutkijat käyttivät satunnaista näytteenottoa (engl. Random Down Sampling), jossa satunnaiset data pisteet poistetaan aineistosta datan tasapainottamiseksi, ja jakoivat aineistot enemmistö- ja vähemmistöluokkiin edellä mainituin perustein. [3] Aineistoissa olevien potilaiden attribuuteille tehtiin myös korrelaatioanalyysi [3].

Kuvan 4.1 taulukosta voidaan havaita, että CVD-aineistossa iällä ("age"), verenpaineella ("hypertension"), sydänsiraudella("heart_disease") sekä verensokerilla("avg_glucose_lvl") on positiivinen korrelaatio aivoinfraktin("stroke") kanssa.

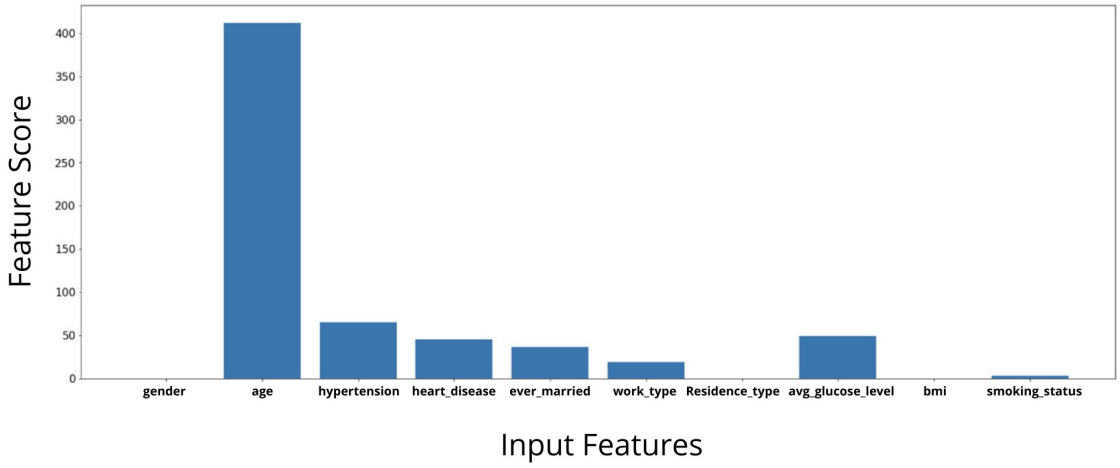


Kuva 4.2: Framingham-aineiston korrelaatiokertoimet. [3].

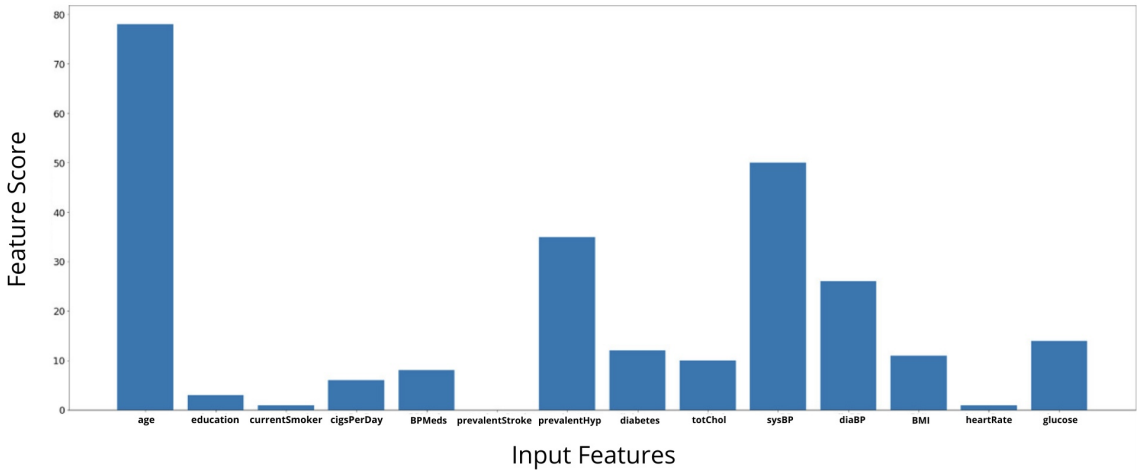
Samoin Framinghamin aineistosta (Kuva 4.2) huomataan positiivista korrelaatiota iän ("age"), verenpaineen ("sysBP", "prevalentHyp", "diabBP"), ja verensokerin ("glucose") kohdalla, kun niitä verrataan todennäköisyyteen sairastua sepelvaltimotautiin ("TenYearCHD"). Muilla muuttujilla korrelaatiot olivat vähäisempiä tai täysin mitättömiä.

Tutkimuksen päätavoitteena oli löytää piirteitä, joiden avulla sydänsairauksien ennustamisesta tulisi tarkempaa [3]. Pathan ym. käyttivät tutkimuksessa varianssianalyysia tunnistamaan tärkeimmät piirteet aineistoista [3]. Pathan ym. ovat käyttäneet Python-ohjelmointikielen scikit-kirjastoa, joka tarjoaa funktiot tärkeimpien piirteiden valitsemiseen (f_classif()) [3].

Kuvissa 4.3 ja 4.4 nähdään molempien aineistojen piirteet ANOVA-F -testillä arvioituna. Kuvan 4.3 tilastojen mukaan, tärkeimmät piirteet aivoinfarktin ("stroke") määrittelyyn ovat ikä, verenpaine, sydänsairaus ja verensokeri. Kuvan 4.4 tilastosta voidaan nähdä, että seuraavan kymmenen vuoden aikana sepelvaltimotaudin saaminen tuottaa suuret f-testin pisteet iän, diabeteksen, verenpaineen, sekä veren-



Kuva 4.3: CVD-aineiston piirteet arvioituna [3].



Kuva 4.4: Framingham aineiston piirteet arvioituna. [3].

sokerin kanssa. Voidaan siis todeta tulosten olevan yhteneväisiä aiemmin esitettyjen korrelatiokerrointen kanssa.

Pathan ym. [3] tarkastelivat koneoppimisalgoritmeja täysille aineistoille arvioidakseen mitkä mallit toimisivat parhaiten molemmille aineistoille. Seuraavaksi tutkittiin, miten mallit suoriutuivat aineistoista muodostetuille osajoukoille, piirteiden valintatekniikan vaikutuksen analysoimiseksi. Koneoppimismallien onnistumista mitattiin täsmällisyydellä, F1-scorella ja ROC:lla. Aineistot jaettiin koulutus- (80 %) ja testidataan (20 %).

Malli	**Tarkkuus**	**Tasapainotettu tarkkuus**	**ROC AUC**	**F1-piste**
Perceptron	0.73	0.74	0.74	0.73
SGD Classifier	0.72	0.73	0.73	0.71
Logistic Regression	0.73	0.73	0.73	0.73
Quadratic Discriminant Analysis	0.72	0.73	0.73	0.72
Linear SVC	0.72	0.72	0.72	0.72
SVC	0.71	0.72	0.72	0.71
Nu SVC	0.71	0.72	0.72	0.71
Nearest Centroid	0.71	0.72	0.72	0.71
Calibrated Classifier CV	0.71	0.72	0.72	0.71
Bernoulli NB	0.71	0.72	0.72	0.71
Gaussian NB	0.71	0.71	0.71	0.71
Passive Aggressive Classifier	0.71	0.71	0.71	0.71
Ridge Classifier CV	0.70	0.71	0.71	0.70
Ridge Classifier	0.70	0.71	0.71	0.70
Linear Discriminant Analysis	0.70	0.71	0.71	0.70
Random Forest Classifier	0.70	0.70	0.70	0.70
AdaBoost Classifier	0.70	0.70	0.70	0.70
KNeighbors Classifier	0.69	0.69	0.69	0.69
Bagging Classifier	0.68	0.68	0.68	0.68
Extra Tree Classifier	0.66	0.66	0.66	0.66
LGBM Classifier	0.65	0.66	0.66	0.65
XGB Classifier	0.65	0.65	0.65	0.65
Decision Tree Classifier	0.61	0.61	0.61	0.61
Label Spreading	0.60	0.60	0.60	0.60
Label Propagation	0.60	0.59	0.59	0.60
Dummy Classifier	0.46	0.46	0.46	0.46

Taulukko 4.3: Mallien suorituskykymittarit CVD-aineistolla: tarkkuus, tasapainotettu tarkkuus, ROC AUC ja F1-pisteet [3].

Malli	**Tarkkuus**	**Tasapainotettu tarkkuus**	**ROC AUC**	**F1-piste**
Linear SVC	0.66	0.67	0.67	0.66
Linear Discriminant Analysis	0.66	0.67	0.67	0.66
Calibrated Classifier CV	0.66	0.67	0.67	0.66
Ridge Classifier CV	0.66	0.67	0.67	0.66
Ridge Classifier	0.66	0.67	0.67	0.66
Logistic Regression	0.65	0.66	0.66	0.65
Nearest Centroid	0.64	0.66	0.66	0.64
KNeighbors Classifier	0.64	0.65	0.65	0.64
Random Forest Classifier	0.64	0.65	0.65	0.64
Bernoulli NB	0.63	0.64	0.64	0.63
LGBM Classifier	0.63	0.64	0.64	0.63
AdaBoost Classifier	0.63	0.64	0.64	0.63
Extra Tree Classifier	0.62	0.63	0.63	0.62
XGB Classifier	0.62	0.63	0.63	0.61
Decision Tree Classifier	0.61	0.62	0.62	0.61
Bagging Classifier	0.60	0.61	0.61	0.59
SGD Classifier	0.60	0.60	0.60	0.60
Gaussian NB	0.57	0.60	0.60	0.53
Passive Aggressive Classifier	0.58	0.59	0.59	0.57
Nu SVC	0.59	0.59	0.59	0.59
Extra Tree Classifier	0.59	0.59	0.59	0.59
SVC	0.58	0.58	0.58	0.58
Quadratic Discriminant Analysis	0.55	0.58	0.58	0.51
Perceptron	0.57	0.55	0.55	0.55
Label Propagation	0.54	0.54	0.54	0.53
Label Spreading	0.53	0.53	0.53	0.53
Dummy Classifier	0.52	0.52	0.52	0.52

Taulukko 4.4: Mallien suorituskykymittarit Framingham-aineistolla: tarkkuus, tasapainotettu tarkkuus, ROC AUC ja F1-pisteet [3].

Kokonaisille aineistoille kouluttamisessa käytetty laskenta-aika oli CVD:n osalta 10,98 iteraatiota sekunnissa (it/s) ja Framinghamin aineistolle 24,20 iteraatiota

sekunnissa.

Taulukosta 4.3 voidaan nähdä, että parhaiten koko CVD-aineistolla suoriutui multilayer perceptron-malli (MLP), eli monikerroksinen neuroverkko, joka saavutti 0,73 täsmällisyyden, 0,74 ROC-arvon ja 0,73 F1-pisteen. MLP:n hyvä suoriutuminen perustuu sen kykyyn löytää kaavoja (engl. pattern) monimutkaisesta lääketieteellisestä aineistosta. Muita hyvin suoriutuneita malleja olivat Logistinen regressio (LR), tukivektoriluokittelija (SVC) ja satunnaismetsä (Random forest). Ne pystyivät tuottamaan kohtuullisen tarkan ennusteen täydellä aineistolla. Huonoiten testissä suoriutui Dummy Classifier. Tutkijoiden mukaan syynä tähän on todennäköisesti sen käyttämät yksinkertaiset säännöt. Tämän takia Dummy Classifier on huono malli käytettäväksi oikean maailman ongelmiin. [3]

Framinghamin kokonaisessa aineistossa (Taulukko 4.4) tulokset eivät olleet kovin hyviä, sillä paras saavutettu tarkkuus oli 0,66, ROC-arvo 0,67 ja F1-pisteet 0,66. Myös muut algoritmit, kuten lineaarinen diskriminanttianalyysi (LDA), logistinen regressio (LR) ja ridge-luokitin, tuottivat samankaltaisia tuloksia. Heikkojen tulosten syynä saattaa tutkijoiden mukaan olla datan ominaisuuksien arvojen laaja vaihteluväli. Ominaisuuksien skaalaus voisi parantaa mallien suorituskykyä normalisoimalla datan tietyille vaihteluväleille. Kuitenkin lääketieteellisissä tutkimuksissa datan muokkaaminen voi aiheuttaa merkittäviä harhoja, minkä vuoksi tutkijat pitivät kaikki ominaisuusarvot ennallaan. [3]

Tutkimuksen tavoitteena oli arvioida piirteiden valinnan vaikutusta luokittelun tarkkuuteen. Merkittävimmät piirteet valittiin koko piirrejoukosta yksittäisten piirteiden pisteytyksen perusteella ANOVA-F -testin avulla. CVD-datasta valittiin neljä keskeistä ominaisuutta (ikä, verenpainetauti, sydänsairaus, keskimääräinen glukoositaso) ja Framingham-datasta viisi ominaisuutta (ikä, korkea verenpaine, systolinen ja diastolinen verenpaine, glukoosi). [3]

Malli	**Tarkkuus**	**Tasapainotettu tarkkuus**	**ROC AUC**	**F1-piste**
SVC	0.74	0.75	0.74	0.74
Nearest Centroid	0.74	0.75	0.74	0.74
Logistic Regression	0.73	0.74	0.73	0.73
SGD Classifier	0.73	0.73	0.73	0.73
Linear SVC	0.72	0.73	0.72	0.73
Linear Discriminant Analysis	0.72	0.73	0.72	0.73
Ridge Classifier CV	0.72	0.73	0.72	0.73
Ridge Classifier	0.72	0.73	0.72	0.73
Quadratic Discriminant Analysis	0.72	0.73	0.72	0.72
Calibrated Classifier CV	0.72	0.73	0.72	0.72
Label Spreading	0.71	0.71	0.71	0.71
Bagging Classifier	0.70	0.70	0.70	0.70
AdaBoost Classifier	0.70	0.71	0.70	0.71
Label Propagation	0.70	0.70	0.70	0.70
Bernoulli NB	0.70	0.70	0.70	0.70
Nu SVC	0.70	0.70	0.70	0.70
LGBM Classifier	0.70	0.70	0.70	0.70
Extra Trees Classifier	0.69	0.70	0.69	0.69
XGB Classifier	0.69	0.70	0.69	0.69
Random Forest Classifier	0.69	0.70	0.69	0.69
Gaussian NB	0.69	0.69	0.69	0.69
Decision Tree Classifier	0.68	0.69	0.68	0.69
Extra Tree Classifier	0.68	0.69	0.68	0.69
KNeighbors Classifier	0.67	0.68	0.67	0.68
Perceptron	0.64	0.64	0.64	0.64
Passive Aggressive Classifier	0.63	0.63	0.63	0.63
Dummy Classifier	0.50	0.50	0.50	0.50

Taulukko 4.5: Mallien suorituskyky CVD-aineistolla piirrevalinnan jälkeen [3].

Malli	**Tarkkuus**	**Tasapainotettu tarkkuus**	**ROC AUC**	**F1-piste**
Perceptron	0.71	0.72	0.72	0.71
AdaBoost Classifier	0.71	0.71	0.71	0.71
SGD Classifier	0.69	0.69	0.69	0.69
Logistic Regression	0.69	0.69	0.69	0.69
Bernoulli NB	0.68	0.69	0.69	0.68
Linear Discriminant Analysis	0.68	0.68	0.68	0.68
Ridge Classifier CV	0.68	0.68	0.68	0.68
Ridge Classifier	0.68	0.68	0.68	0.68
Linear SVC	0.68	0.68	0.68	0.68
Calibrated Classifier CV	0.68	0.68	0.68	0.68
Gaussian NB	0.66	0.68	0.68	0.65
SVC	0.67	0.67	0.67	0.67
Nearest Centroid	0.66	0.67	0.67	0.66
KNeighbors Classifier	0.65	0.66	0.66	0.65
Bagging Classifier	0.63	0.64	0.64	0.63
Quadratic Discriminant Analysis	0.62	0.64	0.64	0.58
Decision Tree Classifier	0.62	0.62	0.62	0.61
Extra Tree Classifier	0.62	0.62	0.62	0.62
Random Forest Classifier	0.62	0.62	0.62	0.62
Label Spreading	0.59	0.59	0.59	0.59
Label Propagation	0.58	0.58	0.58	0.58
LGBM Classifier	0.57	0.58	0.58	0.57
Nu SVC	0.57	0.58	0.58	0.57
XGB Classifier	0.57	0.57	0.57	0.57
Passive Aggressive Classifier	0.57	0.56	0.56	0.57
Dummy Classifier	0.55	0.56	0.56	0.55
Extra Tree Classifier	0.51	0.51	0.51	0.51

Taulukko 4.6: Mallien suorituskyky Framingham-aineistolla piirrevalinnan jälkeen [3].

Mallien suorituskyky arvioitiin käyttäen vain valittuja piirteitä, ja tulokset osoittivat, että koneoppimismallit toimivat paremmin kuin käyttäessään koko piirre-

joukkoa. Korkein saavutettu täsmällisyys CVD-datassa oli 0,74 SVC-mallilla, ja Framingham-datassa 0,71, mikä oli parempi kuin täyttä piirrejoukkoa käyttävien mallien tarkkuus. Lisäksi mallien laskennallinen tehokkuus parani, sillä ne vaativat vähemmän laskennallisia iteraatioita sekunnissa (CVD 3,86 it/s ja Framingham 15,52 it/s). Tulokset vahvistettiin vertaamalla niitä aiempiin tutkimuksiin, joissa käytettiin samoja aineistoja. Yhteenvetona havaittiin, että piirrevalinnan avulla koneoppimismallien suorituskykyä voidaan parantaa merkittävästi, samalla kun laskennallista kuormitusta vähennetään. [3]

Singh ym. [5] tutkimuksen tavoitteena on parantaa sydänsairauksien diagnosointia ja havainnointia koneoppimisen avulla. Tutkimuksessa aineistona käytettiin myös Framinghamin dataa samoin kuin Pathanin ym. tutkimuksessa. Aineistoa esikäsiteltiin, jotta data saatiin käyttökelpoisempaan muotoon koneoppimismallien kouluttamiseen. Esikäsittelymenetelmänä käytettiin satunnaista yliotantaa (engl. random oversampling).

Tutkimuksessa keskityttiin erityisesti ohjatun oppimisen algoritmeihin kuten päätöspuut, satunnaismetsä, pääkomponenttianalyysi (engl. Principal component analysis), sekä gradienttitehostus (engl. Gradient boosting). Tutkijoiden mukaan kullakin menetelmällä on omat yksilölliset ominaisuutensa ja vahvuutensa sydänsairauksien tunnistamisessa. [5]

Singh ym. onnistuivat tutkimuksessaan saavuttamaan peräti 97 % täsmällisyyden käyttäessään satunnaismetsää. Päätöspuulla saavutettu arvo oli 84 %, PCA-arvo 79 % ja tukivektorilla 68 %. [5] Täsmällisyys mittaa mallin ennustamia oikeita positiivisia tuloksia. Mallien herkkyydet olivat identtisiä mallien täsmällisyyden kanssa [5]. Herkkyydellä mitataan oikein ennustettujen positiivisten suhdetta kaikkiin positiivisiin tuloksiin. Myös F1-piste tuotti kaikilla aineistoilla hyvin saman kaltaisia tuloksia.

Tutkimuksen johtopäätöksenä todetaan koneoppimismallien olevan hyödyllisiä

sydänsairauksien ennakoinnissa. Niiden vahvuudeksi mainitaan muun muassa diagnoosien teon tarkkuus, sekä resurssien jakamisen virtaviivaistuminen. Koneoppimismallien suoritusten arviointiin käytetyt työkalut sekä ristiin validointi takaavat luotettavat ja tarkat ennusteet. [5]

5 Johtopäätökset

Tässä kirjallisuuskatsauksessa tarkasteltiin koneoppimismenetelmien käyttöä sydänsairauksien tunnistamisessa kahden tutkimuksen kautta. Molemmat tutkimukset käyttivät Framinghamin sydänsairausaineistoa, mutta lähestyivät mallintamista hieman eri näkökulmista.

Pathanin ym. [3] tutkimus korosti piirrevalinnan merkitystä ennustusmallien tehokkuuden kannalta. Heidän käyttämänsä ANOVA-F -testin avulla valitut muuttujat, kuten ikä, verenpainetaso ja verensokeri, osoittautuivat erityisen merkittäviksi mallien suorituskyvyn kannalta. Merkittävä havainto oli, että piirrevalinnan jälkeen koneoppimismallit, kuten gradienttitehostus, saavuttivat jopa korkeampia suoritusarvoja kuin käyttäessään koko alkuperäistä piirrejoukkoa. Tämä osoittaa, että tiettyjen keskeisten muuttujien korostaminen voi parantaa mallin tarkkuutta ja vähentää ylisovittamisen riskiä.

Singh ym. [5] puolestaan keskittyivät koneoppimismallien vertailuun ja osoittivat, kuinka tärkeää esikäsittely, kuten satunnainen yliotanta, voi olla mallien suorituskyvyn kannalta. Heidän tutkimuksessaan satunnaismetsä (engl. random forest) saavutti vaikuttavan 97 % täsmällisyyden, mikä oli korkein kaikkien testattujen mallien joukossa. Myös päätöspuu ja PCA suoriutuivat kohtalaisesti, mutta tukivektorikoneen tulokset jäivät heikommiksi. Mallien arviointiin käytetyt mittarit, täsmällisyys, herkkyys ja F1-score, tukivat johtopäätöstä, jonka mukaan koneoppimismallit voivat merkittävästi parantaa sydänsairauksien varhaista tunnistamista.

Molemmat tutkimukset osoittavat, että koneoppiminen tarjoaa lupaavia mahdollisuuksia sydänsairauksien diagnosoinnin tueksi. Tarkkuuden lisäksi mainittiin myös resurssien virtaviivaistuminen ja mahdollisuus nopeampiin päätöksentekoprosesseihin. Kuitenkin tulokset myös korostavat, kuinka tärkeää on valita sopivat algoritmit, suorittaa huolellinen esikäsittely ja hyödyntää tehokkaita piirrevalintamenetelmiä.

Jatkossa koneoppimismalleja olisi hyvä testata monipuolisemmilla ja laajemmilla kliinisillä aineistoilla. Lisäksi olisi tärkeää tutkia mallien käytettävyyttä oikeassa terveydenhuollon ympäristössä, jotta voidaan varmistaa mallien luotettavuus ja eettinen soveltuvuus päätöksenteon tueksi.

Tutkielmalle oli asetettu kaksi tutkimuskysymystä. Ensimmäinen kysymys oli, mitä haasteita sydänsairauksien havaitsemiseen liittyy. Kuten luvussa 2 todettiin, sydänsairauksien havaitsemiseen liittyy useita merkittäviä haasteita, kuten sydänsairauksien oireet voivat olla epämääräisiä ja vaikeasti tunnistettavia varhaisessa vaiheessa. Tämä vaikeuttaa oikea-aikaista diagnoosia. Lisäksi kuten luvussa 4 käsitellyistä tutkimuksista huomattiin, käytettävissä olevat terveystiedot voivat olla puutteellisia, epätasapainossa olevia tai sisältää häiriöitä, jotka voivat heikentää analyysin tarkkuutta. Tämä tekee datan esikäsittelystä, kuten puuttuvien arvojen käsittelystä ja luokkien epätasapainon korjaamisesta, kriittisen osan ennustemallien kehittämistä [3]. Myös oikeiden piirteiden (muuttujien) valinta on haastavaa, sillä liian moni tai epäolennainen muuttuja voi johtaa ylisovittamiseen ja heikentää mallin yleistettävyyttä, kuten luvussa 3 todettiin.

Toisena tutkimuskysymyksenä oli, mitä koneoppimismenetelmiä voidaan hyödyntää sydänsairauksien havainnoinnissa. Sydänsairauksien havainnointiin voidaan hyödyntää useita koneoppimismenetelmiä. Tässä tutkielmassa keskityttiin erityisesti ohjatun oppimisen menetelmiin. Tarkastelluissa tutkimuksissa käytettiin muun muassa seuraavia menetelmiä:

- Satunnaismetsä: Erittäin suorituskykyinen menetelmä, joka saavutti parhaim-

millaan jopa 97 % täsmällisyyden.

- Päätöspuu: Yksinkertainen ja tulkittava menetelmä, joka soveltuu hyvin ennustamiseen.
- Pääkomponenttianalyysi: Käytettiin piirreulottuvuuden pienentämiseen, mikä voi parantaa mallin tehokkuutta ja tulkittavuutta.
- Tukivektorikone: Suoriutui Pathanin tutkimuksessa hyvin myös ennen piirrevalintaa.

Pathanin ym. tutkimuksesta huomattiin, että koneoppimismenetelmien tehokkuutta voidaan parantaa esikäsittelyvaiheilla, kuten satunnaisella yliotannalla (engl. random oversampling), sekä huolellisella piirrevalinnalla [3]. Valittu menetelmä, käytetty aineisto ja ennustettavat muuttujat vaikuttavat kaikki merkittävästi lopputulokseen [3][5].

Lähdeluettelo

[1] *Cardiovascular diseases*, en. url: https://www.who.int/health-topics/cardiovascular-diseases (viitattu 22. 10. 2024).

[2] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman ja S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications", *International Journal of Intelligent Networks*, vol. 3, s. 58–73, tammikuu 2022, ISSN: 2666-6030. DOI: 10.1016/j.ijin.2022.05.002. url: https://www.sciencedirect.com/science/article/pii/S2666603022000069 (viitattu 20. 04. 2025).

[3] M. S. Pathan, A. Nag, M. M. Pathan ja S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction", *Healthcare Analytics*, vol. 2, s. 100 060, marraskuu 2022, ISSN: 2772-4425. DOI: 10.1016/j.health.2022.100060. url: https://www.sciencedirect.com/science/article/pii/S2772442522000235 (viitattu 02. 04. 2025).

[4] M. A. Bouqentar, O. Terrada, S. Hamida et al., "Early heart disease prediction using feature engineering and machine learning algorithms", *Heliyon*, vol. 10, nro 19, e38731, lokakuu 2024, ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2024.e38731. url: https://www.sciencedirect.com/science/article/pii/S2405844024147627 (viitattu 09. 10. 2024).

[5] A. Singh, H. Mahapatra, A. K. Biswal, M. Mahapatra, D. Singh ja M. Samantaray, "Heart Disease Detection Using Machine Learning Models", *Procedia*

Computer Science, International Conference on Machine Learning and Data Engineering (ICMLDE 2023), vol. 235, s. 937–947, tammikuu 2024, ISSN: 1877-0509. DOI: 10.1016/j.procs.2024.04.089. url: https://www.sciencedirect.com/science/article/pii/S1877050924007658 (viitattu 09. 10. 2024).

[6] *Sepelvaltimotauti*, fi. url: https://www.terveyskirjasto.fi/dlk00077 (viitattu 20. 05. 2025).

[7] E. J. Benjamin, P. Muntner, A. Alonso et al., "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association", eng, *Circulation*, vol. 139, nro 10, e56–e528, maaliskuu 2019, ISSN: 1524-4539. DOI: 10.1161/CIR.0000000000000659.

[8] P. Ponikowski, A. A. Voors, S. D. Anker et al., "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC", en, *European Heart Journal*, vol. 37, nro 27, s. 2129–2200, heinäkuu 2016, ISSN: 0195-668X, 1522-9645. DOI: 10.1093/eurheartj/ehw128. url: https://academic.oup.com/eurheartj/article-lookup/doi/10.1093/eurheartj/ehw128 (viitattu 17. 03. 2025).

[9] C. T. January, L. S. Wann, H. Calkins et al., "2019 AHA/ACC/HRS Focused Update of the 2014 AHA/ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society in Collaboration With the Society of Thoracic Surgeons", *Circulation*, vol. 140, nro 2, e125–e151, heinäkuu 2019, Publisher: American Heart Association. DOI: 10.1161/CIR.0000000000000665. url:

https://www.ahajournals.org/doi/10.1161/CIR.0000000000000665 (viitattu 17. 03. 2025).

[10] K. Thygesen, J. S. Alpert, A. S. Jaffe et al., "Fourth Universal Definition of Myocardial Infarction (2018)", *Circulation*, vol. 138, nro 20, e618–e651, marraskuu 2018, Publisher: American Heart Association. DOI: 10.1161/CIR.0000000000000617. url: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000617 (viitattu 17. 03. 2025).

[11] R. A. Nishimura, C. M. Otto, R. O. Bonow et al., "2017 AHA/ACC Focused Update of the 2014 AHA/ACC Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines", *Circulation*, vol. 135, nro 25, e1159–e1195, kesäkuu 2017, Publisher: American Heart Association. DOI: 10.1161/CIR.0000000000000503. url: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000503 (viitattu 17. 03. 2025).

[12] R. Yılmaz ja F. H. Yağın, "Early Detection of Coronary Heart Disease Based on Machine Learning Methods", en, *Medical Records*, vol. 4, nro 1, s. 1–6, tammikuu 2022, Number: 1 Publisher: Tıbbi Kayıtlar Derneği, ISSN: 2687-4555. DOI: 10.37990/medr.1011924. url: https://dergipark.org.tr/en/pub/medr/issue/67333/1011924 (viitattu 29. 10. 2024).

[13] *Sairauksien periytyvyys*, fi. url: https://www.terveyskirjasto.fi/dlk00985 (viitattu 31. 10. 2024).

[14] *Sydän- ja verisuonitautien hoito - THL*, fi, joulukuu 2023. url: https://thl.fi/aiheet/kansantaudit/sydan-ja-verisuonitaudit/sydan-ja-verisuonitautien-hoito (viitattu 01. 11. 2024).

[15] *Sydänlihassairaus (kardiomyopatia)*, fi. url: https://www.terveyskirjasto.fi/dlk00634 (viitattu 20. 05. 2025).

[16] S. Al-Zaiti, L. Besomi, Z. Bouzid et al., "Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram", en, *Nature Communications*, vol. 11, nro 1, s. 3966, elokuu 2020, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-020-17804-2. url: https://www.nature.com/articles/s41467-020-17804-2 (viitattu 17. 03. 2025).

[17] *Sydämen kaikukuvaus terveyskeskuslääkärin käytössä*, fi, maaliskuu 1999. url: https://www.laakarilehti.fi/tieteessa/alkuperaistutkimukset/sydamen-kaikukuvaus-terveyskeskuslaakarin-kaytossa/?public=07fa555d822d0d3 (viitattu 20. 05. 2025).

[18] *Rasituskoe*, fi. url: https://sydan.fi/fakta/rasituskoe/ (viitattu 20. 05. 2025).

[19] *Sepelvaltimoiden varjoainekuvaus*, fi. url: https://sydan.fi/fakta/sepelvaltimoiden-varjoainekuvaus/ (viitattu 20. 05. 2025).

[20] H. Tuominen, P. Neittaanmäki, E. Niinimäki et al., *Tekoälyn perusteita ja sovelluksia*, fin. 2019, Accepted: 2019-07-03T12:15:07Z, ISBN: 978-951-39-7796-2. url: https://jyx.jyu.fi/handle/123456789/64975 (viitattu 21. 10. 2024).

[21] *Machine learning, explained | MIT Sloan*, en, lokakuu 2024. url: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (viitattu 01. 11. 2024).

[22] *What is clustering? | Machine Learning*, en. url: https://developers.google.com/machine-learning/clustering/overview (viitattu 01. 11. 2024).

[23] *What is a Decision Tree? | IBM*, en, marraskuu 2021. url: https://www.ibm.com/think/topics/decision-trees (viitattu 08. 04. 2025).

[24] *1.13. Feature selection*, en. url: https://scikit-learn/stable/modules/feature_selection.html (viitattu 03. 04. 2025).

[25] *Classification: Accuracy, recall, precision, and related metrics | Machine Learning*, en. url: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall (viitattu 20. 05. 2025).

[26] *Classification: ROC and AUC | Machine Learning*, en. url: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (viitattu 20. 05. 2025).

[27] *F1 Score in Machine Learning: Intro & Calculation*, en. url: https://www.v7labs.com/blog/f1-score-guide (viitattu 20. 05. 2025).

[28] Tietoarkisto, *Kovarianssi ja korrelaatio - Tietoarkisto*, fi. url: https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/korrelaatio/korrelaatio/ (viitattu 20. 05. 2025).

[29] Tietoarkisto, *Varianssianalyysi - Tietoarkisto*, fi. url: https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/varianssi/anova/ (viitattu 20. 05. 2025).

[30] M. R. Stavseth, T. Clausen ja J. Røislien, "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data", *SAGE Open Medicine*, vol. 7, s. 2 050 312 118 822 912, tammikuu 2019, ISSN: 2050-3121. DOI: 10.1177/2050312118822912. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6329020/ (viitattu 02. 04. 2025).