K-means-algoritmin käyttö asiakassegmentoinnissa

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Toukokuu 2026
Roope Lehtinen

TURUN YLIOPISTO
Tietotekniikan laitos

Roope Lehtinen: K-means-algoritmin käyttö asiakassegmentoinnissa

TkK-tutkielma, 22 s.
Tietotekniikka
Toukokuu 2026

Asiakassegmentointi on keskeinen osa nykyajan liiketoimintaa ja markkinointia, sillä sen avulla yritykset voivat tunnistaa asiakasdatasta asiakasryhmiä ja kohdentaa toimintaansa tehokkaammin. Tutkielman tavoitteena on tarkastella K-means-algoritmin käyttöä asiakassegmentoinnissa ja selvittää, miksi se on yksi yleisimmistä asiakassegmentointimenetelmistä, ja millaisia etuja ja haasteita sen käyttöön liittyy. Tutkielma on toteutettu kirjallisuuskatsauksena.
K-means-algoritmin suosio perustuu sen laskennalliseen tehokkuuteen ja helppoon tulkittavuuteen. K-means on perusmenetelmä, jota voidaan soveltaa eri toimialoilla. Se on suhteellisen helppokäyttöinen ja laskennallisesti tehokas klusterointialgoritmi, ja sillä saadut segmentointitulokset ovat helppo tulkita ja hyödyntää markkinoinnissa ja liiketoiminnan päätöksenteossa. K-means-algoritmin käyttöön liittyy myös haasteita, kuten klusterien määrän valinta, algoritmin herkkyys alustukselle ja datan ominaisuuksien vaikutus segmentointituloksiin. Tutkielma osoitti, että vaikka K-means-algoritmilla on rajoitteita, sen merkittävät vahvuudet varmistavat sen aseman yhtenä keskeisimmistä asiakassegmentoinnin menetelmistä.

Asiasanat: K-means-algoritmi, asiakassegmentointi, klusterointi, data-analytiikka

Sisällys

Kuvat

Taulukot

1 Johdanto

Asiakassegmentointi on keskeinen osa liiketoiminnan ja markkinoinnin päätöksentekoa. Sen avulla yritykset voivat paremmin ymmärtää asiakkaitaan, kohdentaa markkinointiaan ja parantaa asiakaskokemusta. Asiakassegmentoinnissa asiakkaat jaetaan ryhmiin samankaltaisuuksien ja eroavaisuuksien perusteella, jotta voidaan kehittää entistä tehokkaampia ja yksilöllisempiä markkinointistrategioita. Tekoäly- ja koneoppimismenetelmien kehityksen myötä yritykset käyttävät yhä useammin näitä menetelmiä apunaan asiakassegmentoinnissa. Algoritmisesta asiakassegmentoinnista on tullut hallitseva lähestymistapa, ja sen tutkimuskirjallisuudessa on havaittu vuodesta 2000 lähtien 46 erilaista algoritmia. Näistä yleisin on K-means-algoritmi. [1] [2]

Asiakassegmentointia on tärkeää tutkia, sillä ilman tärkeimpien asiakasryhmien ymmärrystä yritykset voivat käyttää resurssejaan tehottomasti [2]. K-means-algoritmi ei ole ainoastaan yleisin asiakassegmentoinnin menetelmä, vaan se on myös yksinkertainen ja helposti toteutettava, minkä perusteella se valikoitui tutkielman aiheeksi [1] [3]. K-means-algoritmin lisäksi asiakassegmentoinnissa käytetään monia muita menetelmiä. K-meansin lisäksi suosittuja algoritmeja ovat fuzzy-algoritmit, geneettiset algoritmit ja K-means-algoritmin pohjalta kehitetyt variantit [1].

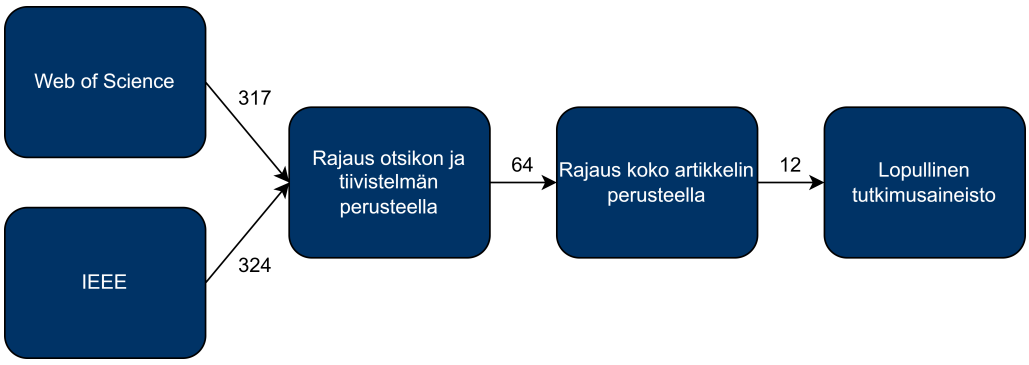
Tämän tutkielman tavoitteena on tutkia syitä K-means-algoritmin suosioon asiakassegmentoinnin tehtävissä sekä selvittää, millaisia etuja ja haasteita K-means-algoritmin käyttöön liittyy. Tutkimalla K-means-algoritmin suosion syitä voidaan

ymmärtää paremmin, mistä syystä sen käyttö on vakiintunut. Tutkielma on toteutettu kirjallisuuskatsauksena. Tutkimuskysymykset ovat:

TK1: Miksi K-means on yksi yleisimmistä asiakassegmentoinnissa käytetyistä menetelmistä?

TK2: Millaisia etuja ja haasteita K-means-algoritmin käyttöön liittyy?

Aineiston hakutietokannoiksi valikoituivat Web of Science ja IEEE. Nämä tietokannat valittiin, koska ne tuottivat osuvimmat hakutulokset. Hakulausekkeena käytettiin *("customer segmentation"* OR *"market segmentation")* AND *("k-means"* OR *"k means"* OR *"kmeans")* AND *("cluster*")* AND *("algorithm"* OR *"machine learning"* OR *"data mining")*. Aineistonhakuprosessi esitetään Kuvassa 1.1.



Kuva 1.1: Aineistonhakuprosessi

Hakulausekkeella löytyi Web of Science -hakutietokannasta 317 tulosta ja IEEE -hakutietokannasta 324 tulosta. Tuloksista käytiin läpi otsikot ja tiivistelmät, minkä perusteella rajattiin pois tulokset, jotka eivät liity aiheeseen tai vastaa tutkimuskysymyksiin. Ensimmäisen rajauksen jälkeen valikoitui 64 tulosta, jotka luettiin kokonaan. Näiden joukosta valittiin lopulliseksi tutkimusaineistoksi 12 artikkelia, joilla vastataan tutkimuskysymyksiin.

Tutkielma koostuu johdannosta sekä luvuista 2, 3 ja 4. Luku 2 on taustaluku, jossa käsitellään asiakassegmentointia ja klusteroinnin taustaa. Luvussa käsitellään

myös yleisesti K-means-algoritmin toimintaperiaatetta sekä datan roolia asiakassegmentoinnissa. Luvussa 3 keskitytään syvällisemmin K-means-algoritmin käyttöön asiakassegmentoinnissa tarkastelemalla tutkimusaineistoa tutkimuskysymysten TK1 ja TK2 näkökulmasta. Luku 4 on yhteenveto tutkielmasta, joka sisältää pohdintaa, yhteenvedon tutkielman tuloksista sekä vastaa asetettuihin tutkimuskysymyksiin.

2 Tausta

2.1 Asiakassegmentointi

Asiakassegmentointi (engl. *customer segmentation*) on menetelmä, jossa asiakaskunta jaetaan ryhmiin. Tavoitteena on, että ryhmät ovat sisäisesti mahdollisimman samankaltaisia ja toisiin ryhmiin verrattuna mahdollisimman erilaisia. Asiakassegmentoinnilla ryhmiä luodaan asiakkaiden eri ominaisuuksien, kuten ostokäyttäytymisen, tarpeiden ja arvojen perusteella. Segmentoinnin tavoitteena on auttaa ymmärtämään asiakaskunnan toimintaa ja tarpeita. Kun asiakkaiden tarpeita ymmärretään paremmin, on mahdollista kehittää eri asiakasryhmille kohdistettuja palveluita, tuotteita ja markkinointistrategioita. [4]

Historiallisesti segmentointi on perustunut helposti saatavilla oleviin muuttujiin, kuten esimerkiksi ikään, sukupuoleen ja tulotasoon. Viime vuosikymmeninä data-analytiikan ja koneoppimisen kehityksen myötä pystytään hyödyntämään entistä monimutkaisempia malleja, jotka perustuvat käyttäytymisdataan. Tämän ansiosta segmentointia käytetään nykyään keskeisenä osana data-analytiikan ja liiketoiminnan prosessia. [5]

Asiakassegmentointia voidaan soveltaa monilla eri toimialoilla, kuten terveydenhuollossa, verkkokaupoissa, telekommunikaatiossa, kaupan alalla, vakuutusalalla, energia-alalla ja finanssiteknologiassa. Asiakassegmentoinnista on myös apua päätöksenteossa, asiakassuhteiden hallinnassa sekä markkinointistrategioiden optimoin-

nissa. [5] Asiakassegmentointia käytetään laajasti monilla eri aloilla, mikä osoittaa sen monipuolisuuden ja joustavuuden.

Datan määrän kasvaessa on syntynyt tarve klusterointimenetelmille, joilla voidaan analysoida suuria määriä dataa tehokkaammin. Tähän tarkoitukseen käytetään monia erilaisia klusterointimenetelmiä, kuten esimerkiksi K-means-algoritmia, hierarkkista klusterointia, DBSCAN-algoritmia ja fuzzy C-meansia [1] [6]. Näillä klusterointimenetelmillä on mahdollista suorittaa segmentointi ilman, että etukäteen olisi määritelty ryhmiä. Klusteroinnilla muodostetaan klustereita eli ryhmiä, jotka perustuvat todellisiin käyttäytymis- ja ostomalleihin. [5] [7]

Kaikilla klusterointimenetelmillä on hyötyjä ja haittoja. Näiden hyötyjen ja haittojen perusteella jotkut klusterointimenetelmät soveltuvat paremmin tiettyihin käyttötarkoituksiin kuin toiset klusterointimenetelmät. Hierarkkinen klusterointi ryhmittelee samankaltaisia havaintoja yhteen ja muodostaa hierarkkisen puumaisen rakenteen perustuen havaintojen välisiin suhteisiin. Se ei vaadi klusterien määrän määrittämistä etukäteen, ja klusterointituloksista näkee selkeästi klusterien väliset suhteet. Se on kuitenkin hidas käsittelemään suuria määriä dataa, toisin kuin K-means-algoritmi. [6] DBSCAN muodostaa klustereita datapisteiden tiheyden perusteella. DBSCAN ei myöskään vaadi klusterien määrän määrittämistä etukäteen, ja se on erityisen hyvä löytämään epäsäännöllisen muotoisia klustereita. DBSCAN-algoritmia käyttäessä parametrien valinta voi kuitenkin olla vaikeaa, ja se toimii hyvin vain tietynlaisissa dataseteissä. [6] Fuzzy C-means puolestaan ryhmittelee datapisteet siten, että yksittäinen datapiste voi kuulua osittain moneen eri klusteriin eri jäsenyysasteiden perusteella. Sen avulla voidaan kuvata paremmin klusterien päällekkäisyyksiä, mutta se on toisaalta myös laskennallisesti raskaampi toteuttaa kuin K-means-algoritmi. [6] [8]

Yksi K-means-algoritmin merkittävimmistä eduista on sen käyttökelpoisuus monissa eri segmentointitehtävissä. Tämän vuoksi K-means on säilyttänyt asemansa

laajasti käytettynä klusterointimenetelmänä. [1] K-meansin kaltaisten perinteisten klusterointimenetelmien lisäksi on myös olemassa uudempia syväoppimiseen perustuvia klusterointimenetelmiä. Nämä menetelmät voivat tunnistaa rakenteita suurista ja monimutkaisista dataseteistä hyvin tehokkaasti. Syväoppimismenetelmät vaativat kuitenkin huomattavasti enemmän laskentatehoa ja parametrien optimointia. [9] Tämän lisäksi monimutkaisten mallien tulkinta voi olla haastavampaa liiketoiminnan näkökulmasta, minkä vuoksi perinteiset klusterointimenetelmät ovat edelleen laajasti käytössä asiakassegmentoinnin tehtävissä. [4] [5] [7] [9] Hyvä esimerkki perinteisestä asiakassegmentointimenetelmästä, joka toimii hyvin yksinkertaisuutensa vuoksi, mutta tuottaa silti relevantteja ja tulkittavia tuloksia, on K-means-algoritmi, jonka toimintaa käsitellään seuraavaksi [7] [8].

2.2 K-means-algoritmi ja klusterointi

K-means-algoritmi on klusterointialgoritmi, eli se perustuu havaintojen jakamiseen klustereihin siten, että saman klusterin havainnot ovat keskenään mahdollisimman samankaltaisia ja eri klustereihin kuuluvat mahdollisimman erilaisia [10]. Se julkaistiin ensimmäisen kerran jo 1950-luvulla [11]. Vaikka K-means-algoritmi kehitettiin 70 vuotta sitten ja sen jälkeen on julkaistu tuhansia klusterointialgoritmeja, sitä käytetään yhä laajasti monilla eri aloilla. Tämä kertoo siitä, kuinka vaikeaa on suunnitella klusterointialgoritmi, joka toimii hyvin kaikissa klusterointia vaativissa tehtävissä. [11]

Klusterointi on yksi tärkeimmistä ohjaamattoman oppimisen (engl. *unsupervised learning*) menetelmistä. Klusteroinnilla pyritään löytämään datasta luonnollisia ryhmiä siten, että ryhmiä ei määritellä luokkiin etukäteen. Sen sijaan ohjatun oppimisen (engl. *supervised learning*) menetelmillä pyritään tunnistamaan valmiiksi määriteltyjä luokkia, jotka on opetettu etukäteen luokkiin jaetun datan avulla. Klusterointi

on siis hyödyllinen apukeino löytämään rakenteita ja samankaltaisuuksia datasta, jonka luokkia ei vielä tunneta. [10] [11]

Klusterointi on lähtökohtaisesti haastava ongelma laskennallisesti, sillä täysin optimaalisen ratkaisun löytäminen on käytännössä mahdotonta käsiteltäessä suuria aineistoja [8] [11]. Tämän takia klusteroinnissa käytetään menetelmiä, jotka antavat hyviä ja laskennallisesti tehokkaita ratkaisuja, vaikka ratkaisut eivät olisikaan aina optimaalisia. Hyvä esimerkki tällaisesta klusterointimenetelmästä on K-means-algoritmi. [8]

K-means-algoritmi pyrkii löytämään klusterikeskukset niin, että datapisteet ovat sijoittuneet mahdollisimman lähelle oman klusterin keskipistettä. Algoritmi perustuu tyypillisesti datapisteiden ja klusterikeskusten välisiin neliöllisiin etäisyyksiin, eli euklidisen etäisyyden neliöön, jota minimoimalla klusterikeskukset määritetään. [8] [11] Neliöllisen etäisyyden käyttö K-means-algoritmissa on hyödyllistä siksi, että se korostaa kaukana toisistaan olevien datapisteiden vaikutusta tehden algoritmista herkän poikkeaville havainnoille samalla pitäen samankaltaiset datapisteet tiiviissä klustereissa [8].

K-means-algoritmi koostuu kahdesta peräkkäisestä vaiheesta. Ensimmäinen vaihe on pisteiden liittäminen lähimpään klusterikeskukseen. Vaiheessa valitaan tietty määrä klusterikeskuksia, jonka jälkeen jokaiselle pisteelle lasketaan etäisyydet klusterikeskuksiin. Pisteet liitetään niihin klustereihin, joilla on pienin etäisyys pisteeseen. Toinen vaihe on klusterikeskusten päivittäminen. Tämä tapahtuu laskemalla kunkin klusterin datapisteiden keskiarvo, joka asetetaan uudeksi klusterikeskukseksi, jolloin se kuvaa entistä tarkemmin klusterin rakennetta. Iteraatiota jatketaan, kunnes klusterikeskusten sijainnit eivät muutu enää merkittävästi. Vaikka K-means-algoritmilla saadut ratkaisut eivät aina ole täysin optimaalisia, ratkaisu löytyy usein nopeasti, vain muutamassa kymmenessä iteraatiossa. Tämä laskennallinen helppous

tekee K-means-algoritmista hyvän menetelmän suurten tietomassojen analysointiin. [8] [11]

K-means-algoritmissa klusterien lukumäärä täytyy määrittää etukäteen, minkä takia algoritmin toimivuuden kannalta on tärkeää valita oikea määrä klustereita. Optimaalinen klusterien määrä riippuu klusteroitavan datan rakenteesta, klusterien muodosta sekä klusteroinnin tavoitteista. Ei ole olemassa universaalia sääntöä, joka olisi ratkaisu kaikkiin tapauksiin. [8] [10] [11] Klusterien optimaalisen määrän arviointiin voidaan käyttää erilaisia menetelmiä, kuten esimerkiksi siluetti-indeksiä ja gap-statistiikkaa. Siluetti-indeksi mittaa klusterin sisäistä yhtenäisyyttä ja klusterien välistä erottelua. Gap-statistiikka puolestaan vertailee K-means-algoritmilla muodostetun klusterointirakenteen laatua satunnaisesti muodostettuun vertailudataan. Sen avulla pyritään arvioimaan, kuinka selkeä luonnollinen klusterirakenne aineistossa on eri klusterimäärillä. Jos klusterointi tuottaa selvästi paremman rakenteen kuin satunnainen vertailudata, klusterien määrää voidaan pitää sopivana [12] [13].

Vaikka K-means-algoritmi on yksinkertainen ja laskennallisesti tehokas klusterointialgoritmi, sillä on myös rajoitteita. K-means-algoritmi on herkkä poikkeaville havainnoille, sillä se käyttää neliöllisiä etäisyyksiä, jolloin poikkeavan havainnon etäisyys klusterikeskuksesta kasvaa nopeasti suureksi [8]. Algoritmilla on myös mahdollisuus päätyä paikalliseen optimiin, sillä globaalia optimia ei aina löydetä [11]. Lisäksi se, että K-means-algoritmissa klusterien määrä täytyy määrittää etukäteen, tekee klusterointiprosessista huomattavasti haastavampaa. [8] [10] [11]

2.3 Datan rooli asiakassegmentoinnissa

Asiakassegmentoinnin onnistuminen perustuu merkittävissä määrin käytössä olevaan dataan. Segmentoinnin onnistuminen ei ole kiinni pelkästään algoritmin valinnasta, vaan käytettävän datan laatu, määrä ja sisältö vaikuttaa siihen merkittäväs-

ti. Käytetty data määrittää, millaisia asiakasryhmiä voidaan muodostaa ja kuinka tarkasti asiakasryhmät vastaavat todellista asiakaskäyttäytymistä. Segmentointitulokset jäävät pinnallisiksi ja harhaanjohtaviksi, jos data ei ole tarpeeksi tarkkaa tai sisältää paljon virheitä. [1] [7] [14]

Datan laatu on yksi ratkaisevista tekijöistä asiakassegmentoinnissa. Datassa olevat virheet, puutteet ja merkittävät datan määrän vaihtelut eri muuttujien välillä heikentävät klusterointimenetelmien toimintaa ja tuottavat epäluotettavia sekä vaikeasti tulkittavia tuloksia. [7] [14] K-means-algoritmi sekä muut klusterointimenetelmät, jotka pohjautuvat datapisteiden etäisyyksiin, ovat herkkiä tällaisille ongelmille [14]. Tämän takia näiden ongelmien tunnistaminen ja korjaaminen on tärkeä osa asiakassegmentointiprosessia.

Myös käytettävien muuttujien määrä vaikuttaa merkittävästi asiakassegmentoinnin tuloksiin. Jos käytetään liian vähän muuttujia, syntyy liian yleistäviä asiakasryhmiä. Liian suuri määrä muuttujia puolestaan lisää melua ja tekee klusterointitulosten tulkinnasta haastavampaa. Tämän takia piirrevalinta on tärkeä osa asiakassegmentointia. Piirrevalinnan tavoite on varmistaa, että asiakassegmentointiprosessissa hyödynnetään sellaisia asiakkaiden ominaisuuksia, jotka ovat yrityksen näkökulmasta merkityksellisiä. [1] [7]

Kokonaisuudessaan data toimii siis asiakassegmentoinnin perustana. Datan laatu ja käsittely vaikuttavat suoraan siihen, kuinka laadukkaita tuloksia asiakassegmentoinnilla saadaan aikaiseksi. K-means-algoritmin toiminta perustuu vahvasti datan määrään ja rakenteeseen, joten datan laatu ja oikeanlainen käsittely ovat erityisen tärkeitä sitä käytettäessä. [1] [7] [14]

3 K-means asiakassegmentoinnissa

Tässä luvussa tarkastellaan tutkimusaineistoa, jonka pohjalta vastataan tutkimuskysymyksiin TK1 ja TK2. Tutkimusaineisto on esitetty Taulukossa 3.1, ja siihen on merkitty, mihin keskeisiin aiheisiin kukin aineisto keskittyy.

Taulukko 3.1: Aineistojen aihealueet

	K-means päämenetelmänä	K-means vertailumenetelmänä	Algoritmin parantaminen	Käytännön soveltaminen	Liiketoimintanäkökulma	K-meansin rajoitteet	Klusterien määrän arviointi
Salminen et al. [1]		x			x	x	x
Bhatia et al. [2]	x			x			
Mirantika ja Rijanto [3]		x		x			x
Tabianan et al. [15]	x			x	x		
Li et al. [16]		x		x	x		
Wilbert et al. [17]	x	x		x	x	x	x
Sivaguru ja Punniyamoorthy [18]		x	x	x	x	x	x
Khan et al. [19]	x		x	x	x		
Gupta et al. [20]		x		x			
Chilla et al. [21]	x		x	x	x		x
Kansal et al. [22]		x		x	x		x
Han et al. [23]	x		x	x	x	x	

3.1 K-means-algoritmin yleisyys asiakassegmentoinnissa

K-means-algoritmi on yksi käytetyimmistä asiakassegmentointimenetelmistä, mikä käy ilmi myös aineistosta sekä monista alan tutkimuksista. Salmisen et al. [1] tekemän 172 artikkelia kattavan systemaattisen kirjallisuuskatsauksen mukaan K-means on ylivoimaisesti eniten käytetty klusterointialgoritmi asiakassegmentoinnin tehtävissä. K-meansin käyttö esiintyy koko aineistossa joko päämenetelmänä tai vertailumenetelmänä uusille tai vaihtoehtoisille klusterointimenetelmille. Uudempien alalla käytettävien algoritmien suorituskykyä verrataan monesti K-meansiin, ja näistä algoritmeista usean toiminta perustuu pohjimmiltaan K-means-algoritmiin [16] [18]. K-meansin käyttö ikään kuin oletusmenetelmänä kertoo sen vakiintuneisuudesta asiakassegmentoinnin tehtävissä.

K-means-algoritmin vakiintuneisuudelle on monia syitä, kuten sen yksinkertaisuus ja laskennallinen tehokkuus sekä segmentointitulosten tulkittavuus. K-means on hyvä valinta liiketoiminnan näkökulmasta ja varsinkin big dataa käsiteltäessä algoritmin tehokkuuden sekä nopeuden puolesta. [15] [23] Segmentointitulosten tulkittavuuden helppous tekee puolestaan K-meansista hyvän työkalun liiketoiminnan tehtäviin. K-meansilla voidaan segmentoida dataa selkeästi määriteltyihin klustereihin. K-means-klusteroinnin avulla voidaan siis löytää esimerkiksi asiakasdatasta suhteellisen vaivattomasti eri asiakasryhmiä ja hyödyntää tietoa löydetyistä ryhmistä esimerkiksi markkinoinnin päätöksenteossa. Useissa aineiston tutkimuksissa painotetaan sitä, kuinka segmentointitulosten tulkittavuus liiketoiminnan kannalta on yksi keskeisimpiä tekijöitä algoritmin valinnassa. [15] [17] [19]

Lisäksi K-meansin yleisyydestä kertoo sen laaja soveltaminen monilla eri toimialoilla. Aineistossa K-meansia hyödynnetään esimerkiksi vähittäiskaupan ja asiakkaan ostokäyttäytymisen tutkimuksessa [2] [17] [22], verkkokaupan ja digitaalisten palve-

luiden asiakasdatan analysoinnissa [15] [19] [20] sekä lentoliikenteen asiakassegmentoinnissa [23]. Myös Salmisen et al. [1] systemaattinen kirjallisuuskatsaus vahvistaa K-meansin laajan käytön eri toimialoilla. Tämä osoittaa, että K-meansin käyttö ei ole rajoittunut vain yhteen tai muutamaan sovelluskohteeseen, vaan sitä voidaan hyödyntää laajasti monissa eri asiakassegmentoinnin tehtävissä. K-means-algoritmi soveltuu käytettäväksi monille eri toimialoille ilman, että algoritmin perusrakennetta joudutaan muuttamaan. K-meansin asema perusmenetelmänä, joka on käyttökelpoinen moniin eri tehtäviin selittää, miksi K-means on säilyttänyt jo pitkään asemansa yhtenä suosituimmista asiakassegmentoinnin menetelmistä.

3.2 K-means-algoritmin vahvuudet

Vaikka K-means-algoritmilla on myös rajoitteita ja heikkouksia, sen vahvuudet tulevat selkeästi esille aineiston tutkimuksissa. Aineiston mukaan K-meansin merkittävimmät vahvuudet ovat laskennallinen tehokkuus, hyvä skaalautuvuus sekä segmentointitulosten tulkittavuus liiketoiminnan näkökulmasta [1] [15] [23].

Yksi K-means-algoritmin merkittävimmistä vahvuuksista on sen kyky käsitellä runsaasti dataa sisältäviä suuria aineistoja tehokkaasti. Tämä tulee esille aineistossa varsinkin tutkimuksissa, joissa käsitellään laajoja asiakasdatakokonaisuuksia. Tästä hyviä esimerkkejä on Han et al. [23], jossa käsitellään lentoyhtiön asiakasdataa sekä Tabianan et al. [15], jossa käsitellään ostokäyttäytymiseen perustuvaa asiakasdataa. Kun näihin tutkimuksiin on valittu algoritmi asiakassegmentointia varten, on K-meansin laskennallinen yksinkertaisuus otettu huomioon. Vaikka käsiteltävää dataa on paljon, segmentoinnin laskennallinen vaativuus pysyy silti kohtuullisena. [15] [23]

K-means-algoritmin tehokkuus perustuu sen tekniseen rakenteeseen. K-meansin toiminta on kaksivaiheinen iteratiivinen prosessi, jossa datapisteet liitetään aina lähimpään klusterikeskukseen, jonka jälkeen klusterikeskukset päivitetään havaintojen keskiarvona. Tämä toistuva ja yksinkertainen prosessi ei vaadi monimutkaista mal-

lin parametrien estimointia tai jakaumaoletuksia, minkä takia K-means-algoritmi on kevyempi toteuttaa kuin moni muu segmentointialgoritmi, kuten esimerkiksi K-medoids, fuzzy C-means ja hierarkkinen klusterointi [3] [18] [20]. [22] [23]

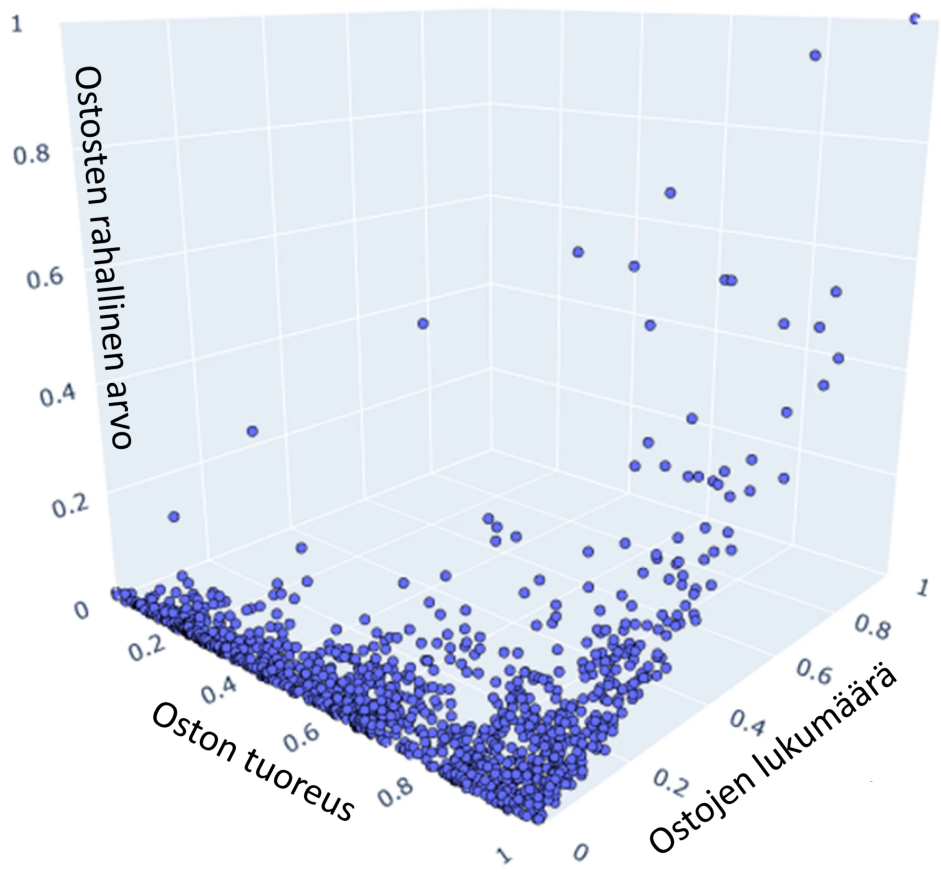
K-meansin nopeudesta ja yksinkertaisuudesta on hyötyä varsinkin liiketoimintaympäristössä, jossa segmentointia tehdään toistuvasti käyttäen hyväksi ajantasaista dataa. K-means-algoritmilla voidaan saada varsin laadukkaita segmentointituloksia suhteessa siihen, kuinka vähän laskentatehoa sen käyttäminen vaatii. [22] [23]

Toinen K-means-algoritmin merkittävimmistä vahvuuksista asiakassegmentoinnin tehtävissä on segmentointitulosten hyvä tulkittavuus ja käytettävyys liiketoiminnan päätöksenteossa. Jokaisella K-means-algoritmilla muodostetulla klusterilla on klusterikeskus, joka kuvaa klusteriin kuuluvien datapisteiden muuttujien keskiarvoja. Tämän ansiosta klusterien ominaisuuksia on helppo tarkastella suoraan alkuperäisten muuttujien tasolla, kunhan muuttujat ovat ymmärrettäviä ja skaalattu oikein. Alkuperäisillä muuttujilla tarkoitetaan asiakkaan alkuperäisiä tietoja tai ominaisuuksia, kuten esimerkiksi ikää, sukupuolta ja ostotapahtumien päivämääriä.[15] [17]

K-means-klusteroinnin käyttö ei kuitenkaan rajoitu vain alkuperäisiin muuttujiin, vaan klusteroinnissa voidaan käyttää yhtä hyvin johdettuja muuttujia, kuten RFM-arvoja. RFM-arvoilla tarkoitetaan oston tuoreutta (engl. *recency*), ostojen lukumäärää (engl. *frequency*) ja ostosten rahallista arvoa (engl. *monetary*). Nämä ovat helposti ymmärrettäviä johdettuja muuttujia. Johdettujen muuttujien käyttäminen ei myöskään välttämättä tarkoita, että segmentointituloksia olisi vaikeampi tarkastella kuin alkuperäisten muuttujien pohjalta tehdyssä asiakassegmentoinnissa. Merkittävin tekijä on muuttujien ymmärrettävyys. Johdetut muuttujat eivät kuitenkaan välttämättä heikennä segmenttien tulkittavuutta, jos ne ovat helposti ymmärrettäviä. Tämä on nähtävissä Wilbertin et al. [17] tutkimuksessa. Tutkimuk-

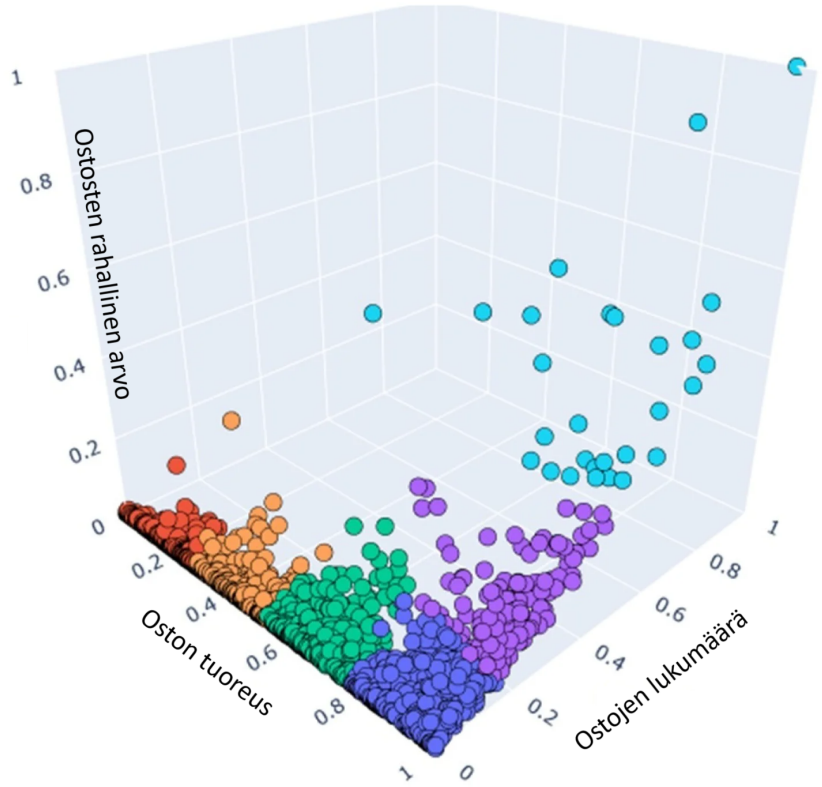
sessa hyödynnetään valmista datasettiä, jonka datapisteiden muuttujiksi on asetettu RFM-menetelmän mukaisesti. [17]

Wilbertin et al. [17] tutkimukseen perustuvassa Kuvassa 3.1 esitetään vaatekaupan myyntidataa segmentoimattomassa muodossa. Datasetissä on 1845 asiakkaan ostodataa vuoden 2016 tammikuusta vuoden 2021 joulukuuhun saakka. Segmentointidatasta on poistettu 97 asiakasta, jotka edustivat yhteensopimatonta dataa, kuten esimerkiksi asiakkaat ilman yhtäkään ostosta. Datapisteet esitetään kolmen muuttujan avulla RFM-menetelmän mukaisesti. Lähellä yhtä olevat arvot kuvaavat, että kyseisen ominaisuuden arvo on korkea suhteessa muihin asiakkaisiin, kun taas lähellä nollaa olevat arvot kuvaavat, että kyseinen ominaisuuden arvo on matala suhteessa muihin asiakkaisiin. [3] [17]



Kuva 3.1: Kuvaus asiakkaiden ostokäyttäytymisestä, perustuen lähteeseen [17]

Kuvan 3.1 3D-pistekaavion datapisteet eivät muodosta selkeitä toisistaan erillisiä ryhmittymiä, joten ilman klusterointia asiakasryhmiä ja yksittäisten datapisteiden kuuluvuutta tiettyyn asiakasryhmään on vaikea määrittää. Kuvassa 3.2 asiakasdata on segmentoitu käyttäen K-means-klusterointia. Parhaaksi klusterien määräksi löydettiin Wilbertin et al. [17] tutkimuksessa kuusi, eli K = 6. Klustereista voidaan tunnistaa kuusi asiakassegmenttiä, jotka kuvaavat niiden sisältämien asiakkaiden ostokäyttäytymistä. Nämä asiakassegmentit ovat menetetyt asiakkaat, joita kuvaa punainen väri, sitoutumattomat asiakkaat, joita kuvaa oranssi väri, viimeaikaiset asiakkaat, joita kuvaa sininen väri, vähemmän viimeaikaiset asiakkaat, joita kuvaa vihreä väri, uskolliset asiakkaat, joita kuvaa violetti väri sekä parhaat asiakkaat, joita kuvaa vaaleansininen väri. [17]



Kuva 3.2: Kuvaus asiakkaiden ostokäyttäytymisestä K-means-klusteroinnin jälkeen, perustuen lähteeseen [17]

Näin datasta, josta ei ole helposti tunnistettavissa selkeitä ryhmittymiä muodostetaan K-means-klusteroinnin avulla toisistaan selkeästi erottuvia asiakasryhmiä, joita voidaan käyttää suoraan hyväksi markkinoinnin suunnittelussa. Jokainen asiakas kuuluu yksiselitteisesti vain yhteen klusteriin, eli asiakasryhmään. Tämä tekee segmentointituloksista selkeitä ja helposti ymmärrettäviä, kun niitä halutaan käyttää liiketoiminnan päätöksenteossa. Myös ne, joiden ymmärrys klusteroinnin toiminnasta on rajallista tai olematonta, voivat helposti ymmärtää K-means-klusteroinnilla muodostettujen asiakasryhmien merkitykset liiketoiminnan kontekstissa klusterointitulosten hyvän ymmärrettävyyden ansiosta. [1] [15] [17]

3.3 K-means-algoritmin rajoitteet ja haasteet

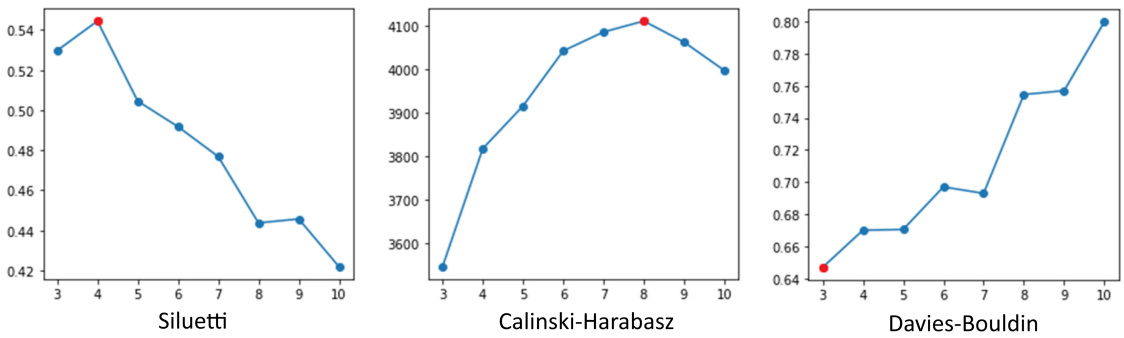
K-means-algoritmilla on sen suosiosta ja monista vahvuuksista huolimatta myös merkittäviä rajoitteita ja haasteita. Aineiston perusteella K-meansin keskeisimmät haasteet ovat klusterien lukumäärän valinta, algoritmin herkkyys alustukselle sekä datan ominaisuuksien vaikutus segmentointituloksiin. Nämä tekijät vaikuttavat K-means-algoritmin soveltuvuuteen ja luotettavuuteen. Tämän takia on tärkeää, että K-means-algoritmin käyttö on huolellista ja suunnitelmallista. [1] [18] [21]

K-means-algoritmin käyttöä rajoittaa se, että klusterien määrä K täytyy määrittää etukäteen ennen kuin algoritmi voidaan suorittaa. Väärin valittu klusterien määrä voi johtaa siihen, että segmentointitulokset ovat harhaanjohtavia. Tämän takia algoritmin käyttäjällä on suuri vastuu siitä, että asiakassegmentointia varten on valittu oikea määrä klustereita. Klusterien määrää valittaessa ei kuitenkaan ole olemassa yhtä universaalia vastausta. Datan rakenne, muuttujat sekä segmentoinnin tavoitteet vaikuttavat kaikki klusterien määrän valintaan. [1] [17]

Sopivan klusterien määrän selvittämiseen voidaan käyttää erilaisia arviointimenetelmiä, kuten esimerkiksi siluetti-indeksiä, joka mittaa kuinka hyvin datapisteet sopivat omaan klusteriinsa verrattuna muihin klustereihin, Calinski–Harabasz-

indeksiä, joka arvioi kuinka selkeästi klusterit erottuvat toisistaan ja kuinka tiiviitä ne ovat sisäisesti, sekä Davies–Bouldin-indeksiä, joka mittaa kuinka samankaltaisia ja lähellä toisiaan klusterit ovat. Näillä menetelmillä voidaan arvioida klusteroinnin laatua K:n eri arvoilla. Nämä kolme menetelmää ovat kaikki tyypiltään sisäisiä validointi-indeksejä, eli ne mittaavat klusteroinnin laatua aineiston oman datan perusteella. [17] [21]

Wilbertin et al. [17] tutkimuksessa käytetään näitä kolmea klusterien lukumäärän arviointimenetelmää apuna K:n optimaalisen arvon löytämiseen. Tutkimuksessa klusteroinnin tavoitteena on löytää vaatekaupan myyntidatasta liiketoiminnan kannalta järkevät asiakassegmentit. Arviointia varten asiakasdata segmentoitiin K-means-algoritmilla käyttäen eri klusterien lukumääriä arvosta K = 3 arvoon K = 10. Kuvassa 3.3 on segmentoinnin perusteella kunkin indeksin arvio parhaasta klusterien lukumäärästä. Siluetti-indeksi antoi arvioksi neljä klusteria, Calinski-Harabasz-indeksi antoi arvioksi kahdeksan klusteria ja Davies-Bouldin-indeksi antoi arvioksi kolme klusteria. Siluetti-indeksin ja Calinski-Harabasz-indeksin tulkinnassa valitaan korkein arvo, kun taas Davies-Bouldin-indeksin tulkinnassa valitaan matalin arvo. [17] [21]

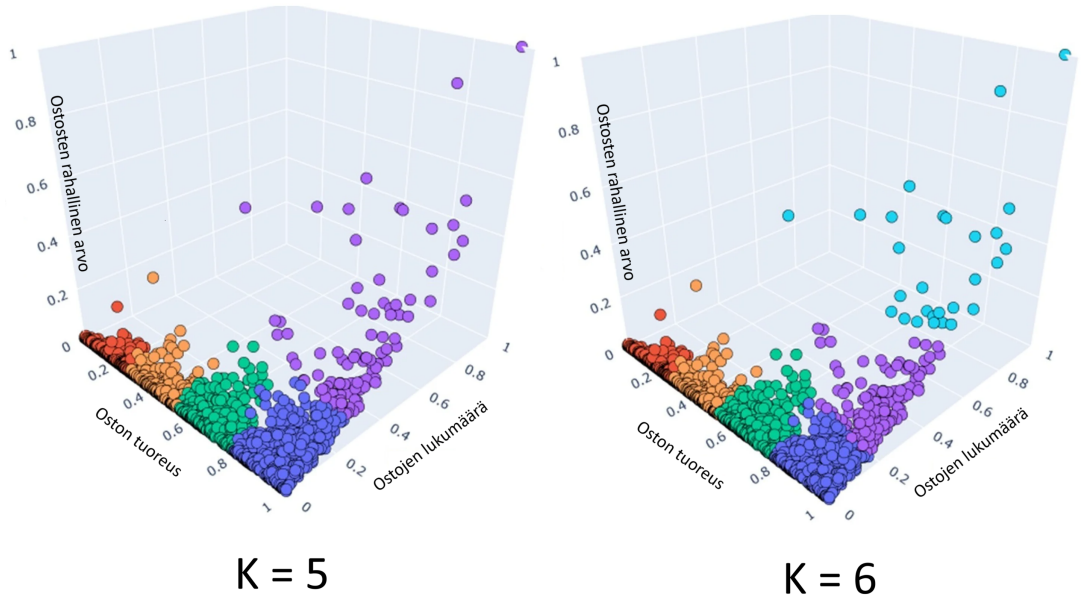


Kuva 3.3: Arviot optimaalisesta klusterien lukumäärästä, perustuen lähteeseen [17]

Ongelmana on, että eri indeksien arviot optimaalisesta klusterien määrästä vaihtelevat merkittävästi. Tämä on yleistä varsinkin dataseteissä, jossa ei ole selkeitä

helposti toisistaan erottuvia klustereita. Selkeästi erottuvien klustereiden puuttuminen on tyypillistä asiakasdatalle, mikä tekee sisäisten validointi-indeksien avulla optimaalisen klusterien lukumäärän selvittämisestä vaikeaa. Tällaisissa tilanteissa voidaan käyttää apuna ulkoisia validointi-indeksejä. Tämä voi kuitenkin olla joissain tilanteissa haastavampaa, sillä ulkoiset validointi-indeksit mittaavat klusteroinnin laatua aineiston ulkopuolisen datan perusteella. Sopivaa ulkoista dataa pitää siis olla saatavilla, jotta näitä menetelmiä voidaan hyödyntää. [17] [21]

K-means-algoritmia käyttäessä lopullinen klustereiden lukumäärä K tulee valita manuaalisesti. Wilbertin et al. [17] tutkimuksessa lopulliseksi klustereiden määräksi valittiin K = 6. Kuvassa 3.4 vertaillaan asiakasdatan klusterointituloksia K:n arvoilla K = 5 ja K = 6. Erona näiden kahden välillä on vaaleansininen klusteri, joka kuvaa asiakkaita, joiden kaikki kolme RFM-arvoa ovat korkeat. Klustereiden määräksi valittiin K = 6, jotta kaikenlaisia asiakkaita kuvaavat segmentit löytyisivät asiakasdatasta. [17]



Kuva 3.4: Asiakasdatan vertailu K-means-klusteroinnin jälkeen K:n arvoilla K = 5 ja K = 6, perustuen lähteeseen [17]

K-means-algoritmia käytettäessä on otettava huomioon sen herkkyys alustukselle. K-means vaatii, että klusterikeskuksilla on alkuarvot. Nämä satunnaiset alkuarvot vaikuttavat siihen, mihin lopputulokseen algoritmi lopulta päätyy. Vaikka klusteroitavana on aina sama aineisto ja käytetään samaa klusterien määrää, klusterointitulokset voivat vaihdella huomattavasti eri alustusten vaikutuksesta. Tämä tarkoittaa, että yksittäinen ajokerta ei välttämättä tuota parasta mahdollista klusterointitulosta. [18] [23]

K-means-algoritmin toiminta perustuu iteratiiviseen prosessiin, jossa klusterointi etenee askeleittain kohti parempaa tilaa. Tämän takia on mahdollista, että algoritmi juuttuu tilaan, josta ei enää ole askelta parempaan tilaan, vaikka tämä tila ei olisikaan kokonaisuudessa paras mahdollinen tila. Tällaista tilaa kutsutaan paikalliseksi optimiksi. Aineistossa käsitellään tätä rajoitetta laajasti, ja sille esitetään mahdollisia ratkaisuja, kuten erilaiset alustustekniikat ja algoritmin ajaminen useampaan kertaan, jotta voitaisiin löytää mahdollisimman optimaalinen lopputulos. Nämä ratkaisut kuitenkin lisäävät klusteroinnin suorittamiseen tarvittavaa laskennallista työmäärää. [18] [23]

Klusteroitavan datan ominaisuudet voivat myös aiheuttaa haasteita klusterointiprosessissa. K-means-algoritmi laskee klusterikeskusten sijainnit klusteriin kuuluvien datapisteiden keskiarvona, minkä vuoksi se on herkkä muusta datasta eroaville datapisteille. Yksittäinen datapiste, joka on kaukana kaikista muista datapisteistä voi vaikuttaa klusterikeskusten sijaintiin merkittävästi ja aiheuttaa vääristyneitä segmentointituloksia. On siis tärkeää, että segmentoitavasta datasta voidaan tunnistaa poikkeavat havainnot ja käsitellä ne ennen klusteroinnin suorittamista. [18] [3]

Haasteita aiheuttaa myös se, että K-means-algoritmi tekee oletuksia klustereiden muodosta. K-means toimii parhaiten, kun klusterit ovat muodoltaan jotakuinkin pallomaisia ja kooltaan keskenään samankokoisia. On kuitenkin yleistä, että segmentoitava asiakasdata ei noudata näitä oletuksia. Näissä tapauksissa segmentointitulokset

jäävät monesti epäoptimaalisiksi, kun tuloksia vertaillaan muihin klusterointialgoritmeihin. [1] [20]

Klusterointiprosessissa tulee myös huomioida segmentoitavan datan muuttujien mittakaavat. K-means-algoritmin toiminta perustuu etäisyyksien laskemiseen, joten eri mittayksiköissä olevat muuttujat voivat vaikuttaa segmentointiin negatiivisesti. Tämän takia segmentoitava data kannattaa skaalata ennen kuin klusterointia aletaan suorittamaan. [15] [17]

4 Yhteenveto

Tässä tutkielmassa tarkasteltiin K-means-algoritmin käyttöä asiakassegmentoinnissa. Tutkielma toteutettiin kirjallisuuskatsauksena, jonka tavoitteena on vastata asetettuihin tutkimuskysymyksiin aineiston pohjalta. Asiakassegmentointia on tärkeää tutkia, koska sen paremman ymmärryksen avulla voidaan oppia lisää eri asiakaskuntien tarpeista ja ominaisuuksista, jolloin resursseja voidaan kohdentaa ja käyttää tehokkaammin. Asiakassegmentointia tutkittiin keskittyen K-means-algoritmin rooliin asiakassegmentoinnissa, sillä se on yksi käytetyimmistä ja keskeisimmistä alan menetelmistä. Aineiston saamiseksi suoritettiin aineistonhakuprosessi, joka dokumentoitiin mahdollisimman selkeästi, jotta prosessi olisi mahdollista toistaa. Tämä luku sisältää vastaukset tutkimuskysymyksiin, pohdintaa työn rajoitteista sekä ehdotuksia mahdollisista jatkotutkimuksista.

TK1: *Miksi K-means on yksi yleisimmistä asiakassegmentoinnissa käytetyistä menetelmistä?* K-means-algoritmi on yksinkertainen, mutta laskennallisesti tehokas klusterointialgoritmi. K-meansilla tuotetut segmentointitulokset ovat helposti tulkittavia, mikä tekee siitä helposti hyödynnettävän ja tehokkaan asiakassegmentoinnin työkalun varsinkin liiketoimintanäkökulmasta. K-means on yksi käytetyimmistä asiakassegmentointimenetelmistä, ja sitä voidaan soveltaa laajasti monilla eri toimialoilla, minkä vuoksi se on saavuttanut asemansa perusmenetelmänä.

TK2: *Millaisia etuja ja haasteita K-means-algoritmin käyttöön liittyy?* K-means-algoritmin merkittävimmät vahvuudet ovat laskennallinen tehokkuus, hyvä skaalau-

tuvuus sekä segmentointitulosten helppo tulkittavuus. K-meansin iteratiivinen toimintaperiaate mahdollistaa tehokkaan klusteroinnin vaatimatta runsasta laskentatehoa. K-means on myös yksinkertainen, suhteellisen helppokäyttöinen ja sillä saadut segmentointitulokset ovat helppo tulkita ja hyödyntää markkinoinnissa sekä liiketoiminnan päätöksenteossa.

K-means-algoritmilla on sen suuresta suosiosta huolimatta myös merkittäviä rajoitteita ja haasteita. Klusterien lukumäärä täytyy määrittää etukäteen, eikä optimaalisen määrän valintaan ole yksiselitteistä menetelmää. Algoritmi on herkkä alustukselle, minkä vuoksi klusteroinnin tulokset voivat vaihdella eri ajokertojen välillä, ja se voi myös jäädä paikalliseen optimiin. Haasteena on myös klusteroitavan datan ominaisuudet, kuten poikkeukselliset havainnot, muodoltaan ja kooltaan oletuksesta poikkeavat klusterit sekä eri mittakaavoissa olevat muuttujien arvot.

Tutkielman rajoitteena oli aineiston rajallinen koko, mikä johtui kandidaatintyön laajuudesta. Hakua jouduttiin rajaamaan voimakkaasti, minkä vuoksi lähteitä, jotka olisivat muuten soveltuneet hyvin tutkielman aineistoksi, saattoi jäädä löytämättä. Aineiston tutkimusten välillä ei havaittu merkittäviä ristiriitoja, mikä voi johtua siitä, että K-means-algoritmi on vakiintunut asiakassegmentoinnin menetelmä, jota on tutkittu jo pitkään. Sen sijaan algoritmin parannuksiin keskittyvissä tutkimuksissa voisi esiintyä enemmän eroavaisuuksia, mutta rajallisen aineiston vuoksi sitä ei voitu havaita.

Jatkotutkimuksessa olisi hyödyllistä tarkastella laajempaa aineistoa tai keskittyä K-means-algoritmin parannuksiin. Lisäksi voisi olla perusteltua vertailla K-meansia muihin klusterointimenetelmiin, jotta saataisiin parempi käsitys eri menetelmien soveltuvuudesta asiakassegmentointiin. Asiakassegmentointi algoritmeja hyödyntäen tulee todennäköisesti jatkossakin olemaan keskeinen osa data-analytiikan tutkimusta, sillä organisaatioiden tarve ymmärtää asiakaskäyttäytymistä ja kohdentaa toimenpiteitä tehokkaasti ei ole katoamassa.

Lähdeluettelo

[1] J. Salminen, M. Mustak, M. Sufyan ja B. J. Jansen, "How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation", *JOURNAL OF MARKETING ANALYTICS*, vol. 11, nro 4, SI, s. 677–692, joulukuu 2023, ISSN: 2050-3318. DOI: 10.1057/s41270-023-00235-5.

[2] T. K. Bhatia, S. Gupta ja A. Sharma, "Analysis of Customer Segmentation Model through K-Means Clustering", teoksessa *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2022, s. 1–6. DOI: 10.1109/ICRITO56286.2022.9965157.

[3] N. Mirantika ja E. Rijanto, "Comparative Analysis of K-Means and K-Medoids Algorithms in Determining Customer Segmentation Using RFM Model", *JOURNAL OF ENGINEERING SCIENCE AND TECHNOLOGY*, vol. 18, nro 5, s. 2340–2351, lokakuu 2023.

[4] J. Joung ja H. Kim, "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews", *INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT*, vol. 70, kesäkuu 2023, ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2023.102641.

[5] R. Hayat Khan, D. Fabian Dofadar, M. G. R. Alam, M. Siraj, M. Rafiul Hassan ja M. Mehedi Hassan, "LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model", *IEEE ACCESS*, vol. 12, s. 96 462–96 480, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3420221.

[6] S. S. Ling, C. W. Too, W. Y. Wong ja M. H. Hoo, "Customer Relationship Management System for Retail Stores Using Unsupervised Clustering Algorithms with RFM Modeling for Customer Segmentation", teoksessa *2024 IEEE 14th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2024, s. 1–6. DOI: 10.1109/ISCAIE61308.2024.10576353.

[7] X. Li ja Y. S. Lee, "Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm", *JOURNAL OF CASES ON INFORMATION TECHNOLOGY*, vol. 26, nro 1, 2024, ISSN: 1548-7717. DOI: 10.4018/JCIT.336916.

[8] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija ja J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data", *INFORMATION SCIENCES*, vol. 622, s. 178–210, huhtikuu 2023, ISSN: 0020-0255. DOI: 10.1016/j.ins.2022.11.139.

[9] N. R. T. Vadrevu, P. Doshi, S. Shrivastava ja H. Dandu, "Deep Learning-Driven Dynamic Clustering for Intelligent Customer Segmentation", teoksessa *2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, 2025, s. 449–455. DOI: 10.1109/ICICI65870.2025.11069888.

[10] K. P. Sinaga ja M.-S. Yang, "Unsupervised K-Means Clustering Algorithm", *IEEE ACCESS*, vol. 8, s. 80 716–80 727, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988796.

[11] A. K. Jain, "Data clustering: 50 years beyond K-means", *PATTERN RECOGNITION LETTERS*, vol. 31, nro 8, SI, s. 651–666, kesäkuu 2010, 19th Inter-

national Conference on Pattern Recognition (ICPR 2008), Tampa, FL, DEC 08-11, 2008, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2009.09.011.

[12] A. Starczewski ja A. Krzyzak, "Performance Evaluation of the Silhouette Index", teoksessa *ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, PT II (ICAISC 2015)*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh ja J. Zurada, toim., sarja Lecture Notes in Computer Science, 14th International Conference on Artificial Intelligence and Soft Computing (ICAISC), Zakopane, POLAND, JUN 14-18, 2015, Polish Neural Network Soc; Univ Social Sci; Czestochowa Univ Technol, Inst Computat Intelligence; IEEE Computat Intelligence Soc, Poland Chapter, vol. 9120, 2015, s. 49–58, ISBN: 978-3-319-19369-4. DOI: 10.1007/978-3-319-19369-4_5.

[13] A. M. El-Mandouh, H. A. Mahmoud, L. A. Abd-Elmegid ja M. H. Haggag, "Optimized K-Means Clustering Model based on Gap Statistic", *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 10, nro 1, s. 183–188, tammikuu 2019, ISSN: 2158-107X.

[14] A. Wasilewski, "Customer segmentation in e-commerce: a context-aware quality framework for comparing clustering algorithms", *JOURNAL OF INTERNET SERVICES AND APPLICATIONS*, vol. 15, nro 1, 2024, ISSN: 1867-4828. DOI: 10.5753/jisa.2024.3851.

[15] K. Tabianan, S. Velu ja V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data", *SUSTAINABILITY*, vol. 14, nro 12, kesäkuu 2022. DOI: 10.3390/su14127243.

[16] Y. Li, X. Chu, D. Tian, J. Feng ja W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm", *APPLIED SOFT COMPUTING*, vol. 113, nro B, joulukuu 2021, ISSN: 1568-4946. DOI: 10.1016/j.asoc.2021.107924.

[17] H. J. Wilbert, A. F. Hoppe, A. Sartori, S. F. Stefenon ja L. A. Silva, "Recency, Frequency, Monetary Value, Clustering, and Internal and External Indices for Customer Segmentation from Retail Data", *ALGORITHMS*, vol. 16, nro 9, syyskuu 2023. DOI: 10.3390/a16090396.

[18] M. Sivaguru ja M. Punniyamoorthy, "Performance-enhanced rough k-means clustering algorithm", *SOFT COMPUTING*, vol. 25, nro 2, s. 1595–1616, tammikuu 2021, ISSN: 1432-7643. DOI: 10.1007/s00500-020-05247-2.

[19] R. H. Khan, D. F. Dofadar ja M. G. Rabiul Alam, "Explainable Customer Segmentation Using K-means Clustering", teoksessa *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2021, s. 0639–0643. DOI: 10.1109/UEMCON53757.2021.9666609.

[20] R. Gupta, S. Subedi, A. Singh ja S. K. Singh, "Comparative Study of Unsupervised Learning Algorithms for Customer Segmentation", teoksessa *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2024, s. 1054–1059. DOI: 10.23919/INDIACom61295.2024.10498443.

[21] P. K. Chilla, S. K. Nahak, M. V. Venkata Deepika, B. S. Rohit, T. Hemalatha ja A. K. Dalai, "Efficient Customer Segmentation Using Silhouette Based K-Means Algorithm", teoksessa *2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP)*, 2024, s. 1–5. DOI: 10.1109/AISP61711.2024.10870813.

[22] T. Kansal, S. Bahuguna, V. Singh ja T. Choudhury, "Customer Segmentation using K-means Clustering", teoksessa *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, s. 135–139. DOI: 10.1109/CTEMS.2018.8769171.

[23] C. Han et al., ”Optimization of K-means algorithm and its application in airline industry customer segmentation”, teoksessa *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, 2022, s. 1370–1373. DOI: 10.1109/ICDSCA56264.2022.9988380.