Marja Satukangas

Klusterointimenetelmät asiakassegmentoinnissa

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Tieto- ja viestintätekniikka

Insinöörityö

30.11.2022

Tiivistelmä

Tekijä:	Marja Satukangas
Otsikko:	Klusterointimenetelmät asiakassegmentoinnissa
Sivumäärä:	55 sivua
Aika:	30.11.2022
Tutkinto:	Insinööri (AMK)
Tutkinto-ohjelma:	Tieto- ja viestintätekniikka
Ammatillinen pääaine:	Ohjelmistotuotanto
Ohjaajat:	Lehtori Vesa Ollikainen

Asiakassegmentointi on tärkeä osa yritysten toimintaa. Segmentoinnin avulla voidaan palveluita, markkinointia ja viestintää suunnitella eri kohderyhmien tarpeet huomioiden ja sitä kautta sitouttaa asiakkaat palveluiden käyttöön. Asiakassegmentit voidaan muodostaa erilaisia menetelmiä hyödyntäen ja erilaisin kriteerein. Tässä opinnäytetyössä selvitettiin, miten klusterointimenetelmiä voidaan hyödyntää asiakassegmentoinnissa ja saadaanko klusteroinnin avulla UniSportin asiakastietojärjestelmän datasta muodostettua toiminnan kehittämisen kannalta hyödylliset segmentit, joita voidaan helposti päivittää.

Klusterointitehtävän aineisto koottiin UniSportin asiakastietojärjestelmästä saatavilla olevasta datasta, joka käsiteltiin sellaiseen muotoon, että klusterointialgoritmi voitiin suorittaa. Aineisto rajattiin kausikorttiasiakkaisiin, ja se koostui palveluiden käyttöön ja asiakkuuden vaiheisiin liittyvistä muuttujista. Klusterointi suoritettiin k-means- ja k-medoids-algoritmeja hyödyntäen ja erilaisia muuttujien yhdistelmiä ja datamuunnoksia testattiin prosessin aikana.

Työn tuloksena saatiin muodostettua asiakassegmentit, joita voidaan hyödyntää UniSportin toiminnassa ja ne voidaan helposti päivittää. Palveluiden käytön perusteella asiakkaista voitiin erottaa aktiivisesti ja passiivisesti palveluja käyttävät asiakkaat, uudet asiakkaat ja pitkäaikaiset asiakkaat sekä mahdollisesti palveluiden käytön lopettaneet tai lopettamisvaarassa olevat asiakkaat. Data-aineistosta voitiin myös löytää tiettyjä kausikorttipalveluita preferoivat ryhmät. Huomionarvoista on, että vaikka työssä löytyi eri segmenteille tyypillinen tapa käyttää palveluita, oli segmenttien sisällä kuitenkin jonkin verran vaihtelua, eivätkä segmentit erottuneet erityisen selkeästi toisistaan.

Työn tuloksia voidaan jatkossa hyödyntää viestinnässä ja markkinoinnissa sekä palveluiden suunnittelussa. Syntyneitä segmenttejä voidaan myös tarkentaa tai syitä tietynlaiseen käyttäytymiseen selvittää esimerkiksi kohdentamalla segmenteille asiakaskyselyitä.

Avainsanat:	klusterointi, asiakassegmentointi, ohjaamaton oppiminen, asiakasprofiili

Abstract

Author:	Marja Satukangas
Title:	Clustering in Customer Segmentation
Number of Pages:	55 pages
Date:	30 November 2022
Degree:	Bachelor of Engineering
Degree Programme:	Information and Communication Technology
Professional Major:	Software Engineering
Supervisors:	Vesa Ollikainen, Senior Lecturer

Customer segmentation is an effective tool for companies to better understand their customers, target marketing actions and communication and offer services that their customers need. Customer segmentation can be done by many different ways and criterion. In the present study the goal was to investigate how clustering algorithms can be used for customer segmentation problems and whether it is possible to create sensible customer segments based on UniSport customer data by using clustering methods.

Data for the clustering task was gathered from the UniSport's client data system. Before it could be passed for the algorithm, it was preprocessed in an appropriate way. Others than training card customers were filtered out and features such as frequency of using services and duration of customership were included in the analysis. Clustering algorithms used in the analysis were k-means algorithm and k-medoids algorithm. Different combinations of features and data transformations were examined during the process.

As a result, segments which can be useful for UniSport and which can be easily updated, were formed. It was recognized that there were segments of active and passive customers, new and long-standing customers and those that may be at risk of quitting using services. Also segments that were profiled for using certain types of services were found. It is worth consideration that even though the generalized manner of using services of each segment was found, there was some variance among customers within the segments and the segments were not that clearly separated from each other.

The results of the study can be used in the future for example to target marketing actions and communication to specific segments. To make the segments more accurate, it is possible to expand the research for example by customer surveys.

Keywords: clustering, customer segmentation, unsupervised learning, customer profile

Sisällys

Lyhenteet

1 Johdanto 1

2 Asiakassegmentointi 2

2.1 Asiakassegmentoinnin ja -profiloinnin käsitteet 2
2.2 Asiakassegmentointimallit 3
2.3 Asiakassegmentointi liikunta-alalla 5
2.4 UniSportin asiakasprofiilit - aikaisempi tutkimus 6

3 Klusterointi asiakassegmentoinnissa 7

3.1 Klusteroinnin hyödyntäminen asiakassegmentoinnissa 7
3.2 Klusteroinnin määritelmä 8
3.3 Klusterointiprosessi 9
3.3.1 Muuttujien valinta ja esikäsittely 9
3.3.2 Läheisyysmitan valinta 11
3.3.3 Klusterointikriteerit 12
3.3.4 Klusterointialgoritmin valinta 12
3.3.5 Tulosten validointi 13
3.4 K-means-algoritmi 15
3.5 K-means-algoritmin variaatioita 18
3.5.1 K-median 18
3.5.2 K-medoids 18
3.5.3 K-modes ja K-prototypes 18

4 Analyysin toteutus 19

4.1 Analyysin lähtökohdat, tavoitteet ja rajaus 19
4.2 Analyysissa käytetyt työkalut ja menetelmät 20
4.3 Data-aineisto 21
4.4 Muuttujien valinta 21
4.5 Aineiston esikäsittely 22
4.5.1 Aineiston kokoaminen 23
4.5.2 Muuttujien jakaumat ja poikkeavien havaintojen käsittely 26
4.6 Mallinnus 34
4.6.1 Aineiston standardointi 35

4.6.2 Algoritmin suoritus 36
4.7 Tulokset 40
4.8 Arviointi 47
5 Yhteenveto 49
Lähteet 52

Lyhenteet

RFM: *Recency, frequency, monetary*. Asiakkuuden arvoon perustuva segmentointimalli, jossa muuttujina ovat ostosten viimeaikaisuus, ostotiheys ja ostosten suuruus.

SSE: *Sum of squared error.* Virheiden neliösumma, eli havaintojen ja havaintoryhmän keskiarvon etäisyyden neliöiden summa. Funktiota voidaan käyttää klusterin sisäisen hajonnan mittaamiseen.

1 Johdanto

Asiakassegmentointi eli asiakkaiden jakaminen ryhmiin tiettyjen ominaispiirteiden, kiinnostuksen kohteiden tai käyttäytymisen perusteella on tärkeä työkalu yrityksille asiakasymmärryksen kasvattamiseksi ja sitä kautta liiketoiminnan kehittämiseksi. Asiakassegmentoinnin myötä palveluita, viestintää ja markkinointia osataan suunnitella paremmin erilaisten kohderyhmien ominaispiirteet, tarpeet ja mielenkiinnon kohteet huomioon ottaen ja sitouttaa asiakkaita palveluiden käyttöön.

Asiakassegmentit voidaan muodostaa erilaisin menetelmin. Yksi menetelmä on hyödyntää asiakassegmentointiin klusterointia, minkä avulla voidaan automatisoidusti löytää toisistaan eroavat ryhmät isosta datamassasta useisiin eri muuttujiin perustuen. Tämän opinnäytetyön tavoitteena on selvittää, voidaanko klusteroinnin avulla löytää UniSportin asiakastietojärjestelmään kertyneestä datasta toiminnan kehittämisen kannalta hyödylliset asiakassegmentit.

UniSport on Helsingin yliopiston, Aalto-yliopiston, Hankenin ja Taideyliopiston yhteinen liikunta- ja hyvinvointipalveluorganisaatio. UniSportin asiakkaat koostuvat yllä mainittujen yliopistojen henkilökunnasta ja opiskelijoista sekä muista asiakkaista. Asiakkaille tuotetaan mm. ryhmäliikunta- ja kuntosalipalveluita, pallopelivuoroja, hierontaa, valmennusta ja kursseja kuudessa eri toimipisteessä pääkaupunkiseudulla. UniSportin tehtävänä on tehdä liikunnasta ja hyvinvoinnista yhä kiinteämpi osa ihmisten arkea ja sitä kautta edistää hyvinvointia yliopistoyhteisössä ja yhteiskunnassa. Tavoitteiden saavuttamiseksi on yhtenä tärkeänä toimenpiteenä tunnistettu olevan asiakasymmärryksen kasvattaminen data-analytiikan keinoin.

Aiemmin UniSportissa on segmentoitu asiakkaita olemassa olevaa dataa hyödyntäen ristiintaulukoimalla yksittäisiä muuttujia. UniSportin asiakkaista on luotu myös asiakasprofiilit laadulliseen tutkimukseen perustuen. Näistä on saatu paljon hyödyllistä tietoa, mutta jotta asiakastietojärjestelmään kertynyttä moniulot-

teista dataa asiakkaiden taustoista, ostokäyttäytymisestä ja palveluiden käytöstä voidaan hyödyntää tehokkaammin ja jotta asiakassegmentit voidaan automatisoidusti päivittää, on segmentointiin käytettävä yksinkertaista ristiintaulukointia kehittyneempää menetelmää. Näistä lähtökohdista lähdettiin tässä opinnäytetyössä selvittämään, voisiko klusterointimenetelmien avulla saada muodostettua asiakassegmentit.

Opinnäytetyön luvussa 2 syvennytään asiakassegmentoinnin käsitteeseen, hyötyihin ja menetelmiin sekä otetaan lyhyt katsaus siihen, miten asiakassegmentointia on liikunta-alalla ja UniSportissa aikaisemmin tehty. Seuraavassa luvussa tutustutaan klusteroinnin teoriaan ja klusterointiprosessin eri vaiheisiin sekä tutkitaan erilaisten klusterointimenetelmien soveltuvuutta erilaisiin data-aineistoihin. Lisäksi katsotaan, miten klusterointia on aiemmin hyödynnetty asiakassegmentointitehtävissä. Tämän jälkeen luvussa 4 käydään tarkemmin läpi työn lähtökohdat ja tavoitteet sekä selvitetään, minkälaista dataa on saatavilla, miten sitä tulee muokata ja minkälaista lopputulosta tavoitellaan. Tämän perusteella suoritetaan klusterointitehtävä, jonka jälkeen arvioidaan tulokset. Luvussa 5 vedetään yhteen tulokset sekä pohditaan, mitkä olivat tutkimuksen hyödyt ja rajoitteet ja mitkä ovat mahdolliset jatkotutkimuksen aiheet.

Opinnäytetyöstä on hyötyä UniSportin lisäksi myös lukijoille, jotka ovat kiinnostuneita asiakassegmentoinnista ja data-analytiikasta liikunta-alan toimintaympäristössä.

2 Asiakassegmentointi

2.1 Asiakassegmentoinnin ja -profiloinnin käsitteet

Asiakassegmentointi on markkinoinnin kirjallisuudessa laajalti käsitelty aihe. Sillä tarkoitetaan nykyisten ja/tai potentiaalisten asiakkaiden jakamista osiin tiettyjen piirteiden perusteella. Asiakassegmentointia voidaan tehdä eri tavoin riippuen siitä, mitä segmentoinnilla tavoitellaan. Ryhmän jäsenten yhteiset piirteet voivat liittyä esimerkiksi asiakkaiden ominaisuuksiin, käyttäytymiseen, arvostuksiin tai mielenkiinnon kohteisiin. (Kotler & Armstrong 2018.)

Asiakassegmentoinnin avulla voidaan markkinointia ja viestintää kohdentaa erilaisille ryhmille heitä puhuttelevalla tavalla ja heidän käyttämissään kanavissa sekä suunnitella palveluita heidän tarpeet huomioon ottaen. Lopputuloksena saadaan tyytyväisempi asiakas, sitoutetaan asiakasta palveluiden käyttöön ja lisätään myyntiä. (Bergström & Leppänen 2021.)

Palvelumuotoilun alalla puhutaan tyypillisesti asiakasprofiloinnista. Asiakasprofiililla tarkoitetaan tietynlaisen asiakasryhmän edustajan perusteellista kuvausta. Asiakasprofiilia voidaan käyttää kuvaamaan ja konkretisoimaan asiakassegmenttejä. Profiili sisältää yleensä kuvitteellisen asiakkaan nimen, valokuvan, demografiset tiedot, elämäntyylin, kiinnostuksen kohteet ja motiivit käyttää yrityksen palvelua. Ennen kuin profiilin kuvausta tehdään, pyritään syvälliseen asiakasymmärrykseen. Profiilien luomisessa käytetäänkin yleensä kvalitatiivisia menetelmiä, kuten haastattelututkimuksia, mutta myös esimerkiksi dataa verkkokäyttäytymisestä on hyödynnetty joissakin tutkimuksissa. (Tuulaniemi 2011; Asiakasprofiili tehostaa viestintää ja markkinointia.)

2.2 Asiakassegmentointimallit

Segmentointimallit jaetaan usein demografiseen, psykografiseen ja käyttäytymiseen perustuvaan segmentointiin.

Demografisessa segmentoinnissa ryhmitellään asiakkaat heidän perustietojensa kuten iän, sukupuolen, koulutuksen, tulotason ja postinumeron perusteella. Psykografisia muuttujia ovat esimerkiksi asenteet, mielipiteet, arvot, persoonallisuus ja kiinnostuksen kohteet. Käyttäytymiseen perustuvassa segmentoinnissa taas keskitytään asiakkaiden ostamiin tuotteisiin tai käyttämiin palveluihin, osto- tai käyntitiheyteen tai käytettyihin palvelukanaviin. (Bergström & Leppänen 2021; Baker.) Segmentointi voidaan tehdä myös asiakkuuden arvoon perustuen. Tällainen on esimerkiksi RFM-malli, jossa asiakkaan arvo lasketaan kolmeen muuttujaan perustuen: kuinka äskettäin asiakas on ostanut (recency), kuinka usein asiakas ostaa (frequency) ja kuinka paljon rahaa asiakas käyttää (monetary). (Huttunen 2021.)

Demografisiin muuttujiin pohjautuva segmentointi on helppoa, mutta sitä on kritisoitu eri lähteissä. Muun muassa Ahonen & Luoto (2015) kritisoivat demografista segmentointia. Heidän mukaansa demografisilla tekijöillä ei ole minkäänlaista korrelaatiota tai kausaliteettia kulutuskäyttäytymisen kanssa. Kuitenkin esimerkiksi Bergström & Leppänen (2021) kirjoittavat, että demografiset muuttujat ovat tärkeitä asiakasegmentoinnissa, mutta ne eivät yksinään riitä. Niiden ohelle tarvitaan myös muita tekijöitä, kuten asenteet ja persoonallisuus, jotta voidaan esimerkiksi suunnitella asiakasryhmälle omat markkinointitoimenpiteet. On myös havaittu, että iän perusteella voidaan markkinointia kohdentaa eri ikäisille asiakkaille erilaisissa sosiaalisen median kanavissa (Agnes 2022).

Psykografisiin muuttujiin perustuvaa segmentointia pidetään hyvänä menetelmänä, koska se antaa kattavan kokonaiskäsityksen asiakaskunnasta sekä elää muuttuvien ympäristötekijöiden mukana. Psykografinen segmentointi vie kuitenkin aikaa, koska se vaatii useimmiten kyselytutkimuksen. (Huttunen 2021.)

Käyttäytymiseen perustuvaa segmentointia käytetään yhä enemmän erilaisten dataa keräävien ohjelmistojen yleistyessä. Asiakkaiden käyttäytymisestä verkkosivuilla ja sosiaalisessa mediassa sekä heidän ostamistaan tuotteista ja asiakkuuksistaan saadaan erilaisten työkalujen avulla entistä enemmän tietoa ja dataa voidaan helposti hyödyntää segmentoinnissa. (Huttunen 2021.)

Segmentointimallin valinta riippuu segmentoinnin tavoitteesta, käytettävissä olevasta datasta ja toiminnan luonteesta. Harvemmin segmentointiin käytetään kuitenkaan vain yhtä segmentointimallia, vaan yleensä erilaisia segmentointimalleja yhdistelemällä saadaan paras tulos. (Bergström & Leppänen 2021.) Cooil ym. (2006) mukaan on kuusi tekijää, joista voi tunnistaa hyvän segmentoinnin. Näitä ovat segmenttien stabiilius, kasvupotentiaali, koko, saavutettavuus, responsiivisuus ja se, että yrityksellä on edellytykset toimia ryhmittelyn perusteella ja se auttaa yritystä pääsemään tavoitteisiin (Cooil ym. 2006).

2.3 Asiakassegmentointi liikunta-alalla

Katsaus liikunta-alan asiakassegmentointitutkimuksiin osoitti, että useissa tutkimuksissa asiakassegmentointi on toteutettu motiiveihin ja asenteisiin perustuen. Niistä monissa asiakassegmenteistä on luotu profiilikuvaukset, joiden avulla segmenteille tyypillisiä arvoja ja toimintaa on kuvattu. Seuraavaksi kuvataan yksi esimerkki tällaisesta tutkimuksesta.

Vuonna 2021 Falardeaun ym. (2021) laatimassa raportissa kirjoitetaan, että asiakassegmentit ovat tällä hetkellä erottuvampia kuin koskaan. Heidän tekemässään tutkimuksessa liikuntapalveluiden kuluttajat jaetaan psykografisten ja käyttäytymiseen perustuvien muuttujien mukaan neljään ryhmään: hyvinvointi-intoilija, tutkija-kokeilija, perinteisiin tukeutuja ja passiivinen osallistuja. Hyvinvointi-intoilijoiden osuus kuluttajista on 23 %, ja he ovat kiinnostuneita kokonaisvaltaisesta hyvinvoinnista. Tutkija-kokeilijat etsivät jatkuvasti uutta ja mielenkiintoista eivätkä välttämättä ole lojaaleja asiakkaita. Perinteisiin tukeutujille on tärkeää tiettyyn aikaan ja tietyssä paikassa tapahtuva treeni. Heitä on vaikea saada asiakkaaksi, mutta kun heidät saadaan, kestää asiakassuhde usein pitkään. Heidän osuus asiakkaista on 11 %. Raportin mukaan 55 % asiakkaista on passiivisia osallistujia, jotka harrastavat liikuntaa epäsäännöllisesti. (Falardeau 2021.)

Club owners -nimisessä blogissa Agnes kirjoittaa asiakassegmentoinnin hyödyistä liikunta-alalla sekä segmentoi liikunta-alan kuluttajat viiteen segmenttiin. Tärkeimpänä tässä pidetään psykografisiin ja käyttäytymiseen perustuvia muuttujia, mutta mainitaan, että ikään perustuva segmentointi auttaa esimerkiksi oikeiden viestintä- ja markkinointikanavien valinnassa. Kirjoittajan mukaan on myös hyödyllistä tietää, mitä palveluita asiakkaat käyttävät ja minkä verran, jotta heille osataan suunnitella oikeanlaisia jäsenyysmalleja ilman, että asiakkaat kokevat maksavansa turhasta. Asiakkaiden segmentointi käyntitiheyden mukaan auttaa asiakkaiden sitouttamisessa. Tämän tiedon pohjalta voidaan esimerkiksi asiakasta lähestyä tietynlaisella viestillä silloin, kun näyttää siltä, että asiakas ei käytä palveluita. (Agnes 2022.)

2.4 UniSportin asiakasprofiilit - aikaisempi tutkimus

UniSportissa on aikaisemmin selvitetty asiakastietojärjestelmän datan perusteella muun muassa palveluiden käytön jakautumista ja asiakassuhteen pituutta. Lisäksi on selvitetty laadullisin tutkimusmenetelmin eri asiakasryhmille ominaisia tapoja käyttää palveluita sekä sitä, mitä eri asiakasryhmät arvostavat ja mikä heitä motivoi.

UniSportin vuonna 2021 tekemässä analyysissa selvisi, että noin 70 % UniSportin asiakkaista käyttää kuntosalipalveluita, noin 43 % käyttää ryhmäliikuntapalveluita ja noin 3 % osallistuu pallopelivuoroille. Noin 50 % asiakkaista käy vain kuntosalilla. (Palveluiden analysointi syksy 2021.) Asiakassuhteen pituutta tarkastelevissa selvityksissä on havaittu, että asiakkaissa on paljon vaihtuvuutta ja palveluiden käytön lopettaneista noin 40 % on ollut asiakkaina alle vuoden (UniSport x Frankly -loppumateriaali 2020).

Vuonna 2020 tehtiin laadullinen tutkimus, jossa asiakkaista tunnistettiin neljä eri asiakasprofiilia: aloitteleva liikkuja, tavoitteellinen treenaaja, ryhmässä treenaava ja sosiaalinen ryhmässä liikkuja. Aloitteleville liikkujille liikunta ei ole muodostunut rutiiniksi. Heillä on tahtoa saada liikunta osaksi arkea, mutta syystä tai toisesta se on jäänyt. Tavoitteelliselle treenaajalle taas liikkuminen on tärkeä osa arkea. Liikkuminen on säännöllistä, monipuolista ja tavoitteellista. Se ei ole erityisen sosiaalista. Sosiaalista ryhmässä liikkujaa taas motivoi yhdessä tekeminen. Tällä ryhmällä liikkuminen voi olla viikoittaista, mutta se saattaa myös välillä jäädä. Ryhmässä treenaava treenaa tavoitteellisesti ja rutinoituneesti ryhmässä. Heille esimerkiksi ryhmäliikunta on tärkeää. (UniSport x Frankly -loppumateriaali 2020.)

Yllä esiteltyjen lukujen analyysi ja asiakasprofiilitutkimus ovat olleet hyödyllisiä UniSportin kehittämistyössä, mutta asiakasymmärryksen kasvattamiseksi edelleen on myös laajamittaisemmalla ja reaaliaikaisemmalla data-analytiikalla nähty arvoa. On mietitty, voisiko klusterointimenetelmiä hyödyntämällä saada ymmärrystä asiakkaiden käyttäytymisestä niin, että myös muutoksia käyttäytymisessä voitaisiin helposti seurata. Seuraavissa luvuissa käydään läpi, miten

klusterointimenetelmiä on hyödynnetty asiakassegmentointitehtävissä yleensä ja miten klusterointiprosessi etenee. Tämän jälkeen testataan, miten klusterointimenetelmät soveltuvat UniSportin käyttötapaukseen.

3 Klusterointi asiakassegmentoinnissa

3.1 Klusteroinnin hyödyntäminen asiakassegmentoinnissa

Asiakassegmentointi on mahdollista toteuttaa eri menetelmin. Asiakkaat voidaan esimerkiksi luokitella ennalta määriteltyjen sääntöjen pohjalta erilaisiin ryhmiin. Datamäärän lisääntyessä ollaan kuitenkin kokonaisvaltaisemman asiakasymmärryksen lisäämiseksi oltu kiinnostuneita käyttämään asiakassegmenttien muodostamisessa useampia muuttujia kahden sijaan, jolloin sääntöperustainen luokittelu muuttuu haastavaksi. Sääntöperustainen luokittelu vaatii myös ennakkotietoa tutkittavasta datasta, jollaista aina ei ole tarjolla. Algoritmien avulla voidaan monidimensionaalisesta datasta löytää ennestään tuntemattomia rakenteita ja muodostaa asiakassegmentit joustavasti ilman rajoittavia sääntöjä. (Customer Segmentation via Cluster Analysis 2022.)

Kun asiakassegmentoinnissa on tavoitteena jakaa asiakkaat ryhmiin mielekkäällä tavalla, on klusteroinnissa tavoitteena löytää datasta siinä luonnollisesti esiintyviä samankaltaisia ryhmiä. Näin ollen klusteroinnin on huomattu olevan hyödyllinen menetelmä asiakassegmenttien muodostamisessa (Collica 2011). Klusterointia on jo kauan hyödynnetty esimerkiksi lääketieteessä, mutta sen hyödyntämisen tutkiminen asiakassegmentointitehtävissä on viimeisten vuosikymmenien aikana lisääntynyt (Jain 2010). Hakusanalla ”customer segmentation and clustering” löytyy Google Scholar -palvelusta tuhansia tieteellisiä artikkeleita, mikä kertoo aiheen suosiosta.

Katsanto asiakassegmentointitutkimuksiin, joissa on hyödynnetty klusterointia, osoitti, että klusterointimenetelmät ovat moninaisia. Tutkimukset pääasiallisesti käsittelevät rajattuja aihepiirejä, ja menetelmät on valittu soveltuvaksi tiettyyn tarkoin rajattuun data-aineistoon ja tutkimusongelmaan. Näin ollen tutkimuksia on suoraan hankala soveltaa kulloinkin käsillä olevaan segmentointitehtävään.

Huomionarvoista on kuitenkin se, että useissa asiakassegmentointitehtävissä klusteroinnissa on hyödynnetty k-means-klusterointialgoritmia, jonka avulla asiakkaat on segmentoitu RFM-malliin perustuen.

3.2 Klusteroinnin määritelmä

Koneoppiminen jaetaan perinteisesti ohjattuun oppimiseen ja ohjaamattomaan oppimiseen. Ohjatussa oppimisessa järjestelmä opetetaan toimimaan annetun mallin perusteella, kun taas ohjaamattomassa oppimisessa ei ole ennestään tiedossa olevaa mallia, vaan tavoitteena on löytää ennestään tuntemattomasta datasta jonkinlainen rakenne, jossa tietyt kaavamaisuudet esiintyvät useammin kuin toiset. (Alpaydin & Francis 2014.)

Klusterointi on ohjaamattoman oppimisen menetelmä, jossa tavoitteena on osittaa ennestään tuntemattomalla tavalla rakentunut data ryhmiin siten, että samaan ryhmään kuuluvat havainnot ovat mahdollisimman samankaltaisia toistensa kanssa, mutta eroavat muihin ryhmiin kuuluvista havainnoista mahdollisimman paljon. Kirjallisuudessa klusteri on määritelty muun muassa pisteiden ryhmittymäksi avaruudessa, jossa etäisyys minkä tahansa kahden pisteen välillä saman klusterin sisällä on pienempi kuin pisteiden etäisyys mihin tahansa toisen klusterin pisteisiin. Toisessa määritelmässä klusteri määritellään sellaiseksi pisteiden joukoksi, jossa pisteet esiintyvät tiheästi ja joka erottuu toisista tiheiden pisteiden joukoista vähemmän tiheiden alueiden avulla. Nämä määritelmät ovat hyvin yleisluontoisia, eivätkä anna vastausta siihen, miten samankaltaisuus määritellään ja miten sitä mitataan. Nämä riippuvat tutkimusongelmasta ja ovat osa klusterointitehtävää. Siksi tarkkaa yleispätevää määritelmää klusteroinnille ja klusterille on haastava löytää. (Jain & Dubes 1988; Theodoridis & Koutroumbas 2006.)

Eri sovellustarkoituksiin on kehitetty suuri määrä erilaisia klusterointimenetelmiä, joita tutkijat voivat soveltaa omiin tutkimusongelmiinsa. Riippuu käytettävissä olevasta data-aineistosta ja tutkimusongelmasta, mitä menetelmiä käyte-

tään. Seuraavissa aliluvuissa katsotaan, minkälaisista vaiheista klusterointiprosessi koostuu ja minkälaisia valintoja sen aikana tehdään. Valinnoista riippuu, minkälainen tulos klusteroinnista saadaan.

3.3 Klusterointiprosessi

Theodoris & Koutroumbas (2006) jakaa klusterointitehtävän seuraaviin vaiheisiin: muuttujien valintaan ja esikäsittelyyn, läheisyysmitan valintaan, klusterointikriteerien valintaan, klusterointialgoritmin valintaan, tulosten validointiin ja tulosten tulkintaan.

3.3.1 Muuttujien valinta ja esikäsittely

Muuttujien valinnalla voidaan vaikuttaa oleellisesti klusteroinnin lopputulokseen ja laskentatehoon. Muuttujien valinnassa mahdollisimman paljon informaatiota kiinnostuksen kohteesta säilytetään, mutta turha ja päällekkäinen tieto eliminoidaan. (Theodoris & Koutroumbas 2006.) Muuttujista voidaan valita vain kaikista tärkeimmät muuttujat varsinaiseen analyysiin ja käyttää muita muuttujia jälkianalyysissa. Muuttujia voidaan myös yhdistää laskemalla niille uusi arvo tähän tarkoitettuja tekniikoita hyväksi käyttäen. Yksi esimerkki tällaisesta menetelmästä on pääkomponenttianalyysi (PCA), jota käytetään usein, jos muuttujia on paljon. (Jain 2010.) Klusteroinnin yhteydessä puhutaan monidimensionaalisuuden kirouksesta. Siinä dimensionaalisuudesta johtuen melkein kaikki pisteparit ovat yhtä kaukana toisistaan. Tätä ilmiötä voidaan välttää muuttujia valikoimalla. (Leskovec ym. 2011.)

Esikäsittelyssä data muunnetaan sellaiseen muotoon, että sitä voidaan käsitellä seuraavissa vaiheissa tarkoituksenmukaisesti. Lisäksi data standardoidaan ja puuttuvat, virheelliset ja poikkeavat havainnot käsitellään tarpeen mukaan. (Theodoris & Koutroumbas 2006.)

Standardoinnilla tarkoitetaan muuttujien saattamista yhteneväiseen muotoon. Useat klusterointimenetelmät vaativat, että muuttujat noudattavat samaa skaalaa. Muutoin muuttujat, joissa skaala on suurempi, painottuvat analyysissa toisia

muuttujia enemmän. Standardoinnilla voidaan myös muokata aineiston jakaumaa sekä vaikuttaa poikkeaviin havaintoihin. Yleisiä tapoja standardoida aineisto on Z-score-standardointi ja Min-Max-skaalaus. (Karani 2022.) Z-score-standardoinnissa kustakin havainnosta vähennetään kaikkien havaintojen keskiarvo ja tulos jaetaan havaintojen keskihajonnalla (kaava 1). Tällä menetelmällä havaintojen keskiarvoksi tulee 0 ja keskihajonnaksi 1. Aivan yhtenäistä skaalaa ei näin saavuteta, mutta poikkeavien havaintojen vaikutusta analyysiin voidaan tällä menetelmällä lieventää. Z-score-standardoinnin matemaattinen esitys on seuraava:

$$x_{scaled} = \frac{x_i - \mu}{\sigma}, \tag{1}$$

jossa x_i on havainnon arvo, μ on havaintojen keskiarvo ja σ on havaintojen keskihajonta. (Normalization.)

Min-Max-skaalauksessa havainnot skaalataan arvojen 0 ja 1 välille (kaava 2). Siinä minimiarvo saa arvokseen 0, maksimiarvo saa arvokseen 1 ja muut arvot saavat arvot tältä väliltä. Min-Max-skaalauksen avulla muuttujien arvot saadaan muunnettua samaan skaalaan, mutta poikkeaviin havaintoihin ei voida vaikuttaa. Min-Max-skaalauksen matemaattinen esitys on seuraava:

$$x_{scaled} = \frac{x_i - x_{min}}{x_{min} - \mathrm{x}_{min}}, \tag{2}$$

jossa x_i on havainnon arvo, x_{min} on havaintojen minimiarvo ja x_{max} on havaintojen maksimiarvo. (Normalization.)

Poikkeavat havainnot voivat vaikuttaa klusteroinnin lopputulokseen tarpeettomasti, joten tällaiset havainnot tulee löytää ja käsitellä tarpeen mukaan. Poikkeavat havainnot voidaan etsiä ja kokonaan poistaa tai ne voidaan käsitellä sopivan menetelmän avulla. (Brownlee 2020.) Toisiin etäisyysmittoihin poikkeavilla havainnoilla on enemmän vaikutusta kuin toisiin. Jos poikkeavat havainnot halutaan pitää mukana analyysissa, voidaan myös käytettävä etäisyysmitta va-

lita niin, että poikkeavilla havainnoilla ei ole niin paljon vaikutusta lopputulokseen. Virheelliset havainnot voidaan kokonaan poistaa tai ne voidaan korvata esimerkiksi keskiarvolla. (Cowell & Victoria-Feser 1996.)

3.3.2 Läheisyysmitan valinta

Läheisyys- tai etäisyysmitan valinta on oleellinen osa klusterointiprosessia, koska sen avulla mitataan, kuinka samanlaisia tai erilaisia havainnot ovat. Esimerkkejä etäisyysmitoista, joita käytetään klusteroinnissa havaintojen etäisyyden mittaamiseen, ovat euklidinen etäisyys, Manhattan-etäisyys, kosinisamankaltaisuus ja Hamming-etäisyys.

Euklidisen etäisyyden matemaattinen määritelmä on:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{3}$$

jossa d on etäisyys ja x on piste, jonka etäisyys lasketaan pisteestä y. Euklidista etäisyyttä käytetään yleisesti jatkuville muuttujille kaksi- tai kolmeulotteisessa avaruudessa. Monidimensionaalisessa avaruudessa se on herkkä luvussa 3.3.1 mainitulle monidimensionaalisuuden kiroukselle. Jainin (2010) mukaan euklidinen etäisyys toimii hyvin, kun data-aineisto koostuu kompakteista ja eriytyneistä klustereista. Euklidista etäisyyttä käytettäessä muuttujat, joissa skaala on suurempi, painottuvat pienempiskaalaisia muuttujia enemmän. Data tulee siis standardoida ennen etäisyyksien mittaamista. (Jain 2010.)

Manhattan-etäisyys soveltuu euklidisen etäisyyden tapaan jatkuvien muuttujien etäisyyksien laskemiseen. Joidenkin tutkimusten mukaan se toimii euklidista etäisyyttä paremmin monidimensionaalisessa avaruudessa (Grootendorst 2021). Sen matemaattinen kaava on:

$$d = \sum_{i=1}^{n}|x_i - y_i| \tag{4}$$

Kosinisamankaltaisuusmittaa (kaava 5) käytetään usein euklidisen etäisyyden sijaan monidimensionaalisessa avaruudessa. Siinä kahden havainnon etäisyys

toisistaan on havaintojen paikkavektoreiden välisen kulman kosini. Sen matemaattinen määritelmä on:

$$d = \frac{\sum_{i=1}^{n} x_i \, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5}$$

Kosinisamankaltaisuus saa arvon -1 ja 1 väliltä. Siinä ei huomioida muuttujien arvojen suuruutta, vaan se perustuu siihen, miten muuttujat suhtautuvat toisiinsa. Tässä mittaustavassa ei ole vaikutusta sellaisilla muuttujilla, jotka saavat arvokseen 0, vaan ainoastaan sellaiset muuttujat, jotka saavat jonkin arvon, tekevät havainnot samankaltaisiksi. (Jiawei ym. 2012.)

Hamming-etäisyyttä käytetään luokkamuuttujien etäisyyksien laskennassa. Siinä etäisyyden arvo muodostuu saman arvon omaavien muuttujien lukumäärän perusteella. (Charu & Chandan 2013.)

3.3.3 Klusterointikriteerit

Klusterointiprosessin aikana määritellään minkälaista klusterointitulosta lähdetään hakemaan eli minkälaisten kriteerien perusteella tuloksen katsotaan olevan mielekäs. Klusteroinnissa löydetään yleensä aina jonkinlainen tulos, mutta tulee arvioida, löytyykö datasta todellista rakennetta, joka on mielekäs ratkaistavan tutkimusongelman kannalta vai onko datalla täysin satunnainen luonne. (Theodoris & Koutroumbas 2006.)

3.3.4 Klusterointialgoritmin valinta

Klusterointialgoritmeja ja niiden variantteja on olemassa satoja, mutta karkeasti klusterointialgoritmit voidaan jakaa hierarkkisiin, osittaviin, tiheyspohjaisiin ja ruudukkopohjaisiin algoritmeihin. Riippuu tutkimusongelmasta, klusterointikriteereistä, datan laadusta ja käytettävistä muista parametreista, mitä klusterointialgoritmia käytetään.

Hierarkkisissa algoritmeissa muodostetaan sisäkkäisiä klustereita, ja ne voivat olla joko kokoavia (agglomerative) tai jakavia (divisive). Kokoavissa algoritmeissa klusterit muodostetaan niin, että aluksi kaikki havainnot ovat omissa klustereissaan, joista ne yhdistetään kaikista lähimpänä olevien havaintojen kanssa. Tätä jatketaan, kunnes päästään hyvään lopputulokseen. Jakavissa algoritmeissa taas kaikki havainnot ovat aluksi yhdessä klusterissa, josta ne erotetaan pienempiin klustereihin. (Theodoris & Koutroumbas 2006.)

Osittavissa algoritmeissa, kuten yleisesti käytetyssä k-means-algortimissa ei ole hierarkkista rakennetta, vaan havainnot ositetaan klustereihin ennalta määritellyn klusterien lukumäärän perusteella (Jain 2010).

Klusterointimenetelmät voidaan jakaa myös koviin ja pehmeisiin menetelmiin. Pehmeissä menetelmissä havainnot voivat kuulua useampaan klusteriin yhtä aikaa, kun kovissa menetelmissä havainnot joko kuuluvat johonkin klusteriin tai eivät kuulu. (Kassambara 2017.)

3.3.5 Tulosten validointi

Klusteroinnin tulosten hyvyyden arvioinnissa voidaan hyödyntää erilaisia mittareita. Klusteroinnin validointimittarit jaetaan ulkoisiin ja sisäisiin mittareihin. Ulkoiset validointimittarit mittaavat klustereiden hyvyyttä perustuen ulkoiseen tietoon, kuten ennalta tiedossa oleviin luokkiin tai aikaisempaan tutkimukseen. Sisäisissä validointimittareissa arvioidaan klusterirakenteen hyvyyttä ilman ulkoista tietoa erilaisten laskennallisten kertoimien tai indeksien perusteella. (Charu & Chandan 2013.)

Kun klusteroinnin tavoitteena on löytää klusterit, joihin kuuluvat havainnot ovat mahdollisimman samankaltaisia keskenään ja mahdollisimman erilaisia muihin klustereihin kuuluvien havaintojen kanssa, perustuu klusteroinnin sisäinen validointi pääasiassa näiden kriteerien arvioimiseen. Sisäisen validoinnin yhteydessä puhutaan klusteroinnin tiiveydestä (compactness) ja erottuvuudesta (separation). Tiiveys kertoo, kuinka samanlaisia samaan klusteriin kuuluvat havain-

not ovat ja sitä voidaan mitata esimerkiksi samaan klusteriin kuuluvien havaintojen hajonnan perusteella ja/tai mittaamalla samaan klusteriin kuuluvien havaintojen etäisyydet toisistaan tai klusterikeskipisteestä. (Charu & Chandan 2013.)

Erottuvuus kertoo, kuinka erottuvia klusterit ovat toisista klustereista. Erottuvuutta voidaan analysoida esimerkiksi mittaamalla klusterikeskipisteiden etäisyydet toisistaan ja laskemalla eri klustereissa olevien havaintojen väliset minimi- tai maksimietäisyydet. Myös tiheyden mittaamiseen perustuvia mittareita käytetään erottuvuuden mittaamisessa. (Charu & Chandan 2013.)

Yleisesti käytetty klusteroinnin sisäisen validoinnin mittari on siluettikerroin (silhouette coefficient). Sen avulla voidaan arvioida sekä klusteroinnin tiiveyttä että erottuvuutta. (Charu & Chandan 2013.) Siluettikerroin mittaa, kuinka lähellä havainto on samaan klusteriin kuuluvia havaintoja suhteessa muihin klustereihin kuuluviin havaintoihin. Siluettikertoimen matemaattinen kaava yhden havainnon kertoimen laskemiselle on:

$$s = (b - a)/\max(a, b)\ , \tag{6}$$

jossa s on siluettikerroin, a on havainnon ja muiden samaan klusteriin kuuluvien havaintojen etäisyyksien keskiarvo ja b havainnon ja lähimmän klusterin havaintojen etäisyyksien keskiarvo. Siluettikerroin koko havaintojoukolle on kaikkien havaintojen siluettikertoimien keskiarvo. (Wei 2020.)

Siluettikerroin saa arvon väliltä [-1, 1]. Kun havainnon siluettikertoimen arvo lähestyy yhtä, tarkoittaa se, että havainto on kaukana muiden klustereiden havainnoista suhteessa oman klusterin havaintoihin ja klusteroinnin tulos on hyvä. Kun arvo on lähellä nollaa, on havainto kahden tai useamman klusterin välissä. Jos arvo on negatiivinen, havainto kuuluu väärään klusteriin. (Wei 2020.)

Eri klusterointimallien hyvyyttä voidaan vertailla myös Calinski-Harabasz-indeksin avulla, jossa klustereiden erottuvuutta verrataan klustereiden sisäiseen yhteneväisyyteen. Yhteneväisyyttä mitataan havaintopisteiden etäisyytenä klusterikeskipisteeseen ja erottuvuutta klusterikeskipisteiden etäisyytenä globaaliin

klusterikeskipisteeseen. Mitä suurempi arvo on, sitä parempi on klusteroinnin tulos. (Wei 2020.)

Davies-Bouldin-indeksissä käytetään klustereiden sisäistä hajontaa ja klustereiden välistä etäisyyttä kaikista samankaltaisimpien klusterien löytämiseksi. Mitä samankaltaisempia klusterit ovat, sitä suuremman arvon Davies-Bouldin-indeksi saa. Pienin mahdollinen arvo on 0. Mitä lähempänä 0:aa luku on, sitä parempi on tulos. Davies-Bouldin-indeksi on vanha, mutta paljon käytetty validointimittari. (Hassani 2016.)

3.4 K-means-algoritmi

Osittavista algoritmeista yksi tunnetuimmista ja käytetyimmistä on k-means-algoritmi. K-means-algoritmissa valitaan k painopistettä (centroid) satunnaisesti tai muulla halutulla tavalla edustamaan kutakin klusteria. Sitten lasketaan kaikkien havaintopisteiden etäisyydet painopisteisiin valitulla etäisyysmitalla ja osoitetaan kukin havainto sitä lähimpänä olevaan klusteriin. Tämän jälkeen lasketaan kuhunkin klusteriin kuuluvien havaintojen keskiarvo ja muutetaan klustereiden painopisteeksi laskettu keskiarvo. Tätä toistetaan, kunnes painopisteet ja havaintojen klusterit eivät enää muutu tai saavutetaan jokin muu valittu kriteeri. (Charu & Chandan 2013.)

K-means-klusterointialgoritmin tavoitteena on virheiden neliösumman (SSE) minimointi. Virheiden neliösumma kertoo kunkin klusterin sisäisestä hajonnasta ja pienenenee jokaisella iteraatiolla, kunnes saavuttaa lopulta lokaalin minimin. Klusterin virheiden neliösumma voidaan esittää matemaattisesti seuraavasti:

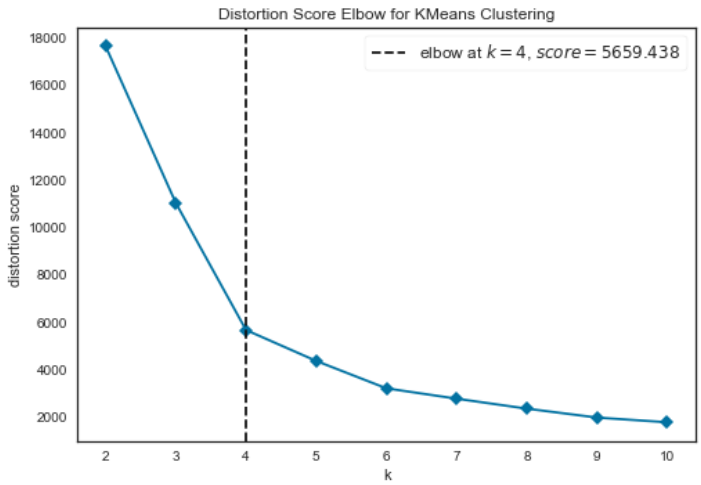
$$SSE(C) = \sum_{i=1}^{n}(x_i - c_k)^2, \tag{7}$$

jossa $SSE(C)$ on klusterin virheiden neliösumma, x_i on klusteriin kuuluvan havainnon arvo ja c_k on klusterin havaintojen keskiarvo. (Charu & Chandan 2013.)

K-means-algortimissa etäisyysmittana voidaan käyttää esimerkiksi euklidista etäisyyttä, Manhattan-etäisyyttä tai kosinisamankaltaisuutta, mutta yleisin käytetty etäisyysmitta k-means-algoritmin kanssa on euklidinen etäisyys. (Charu & Chandan 2013.)

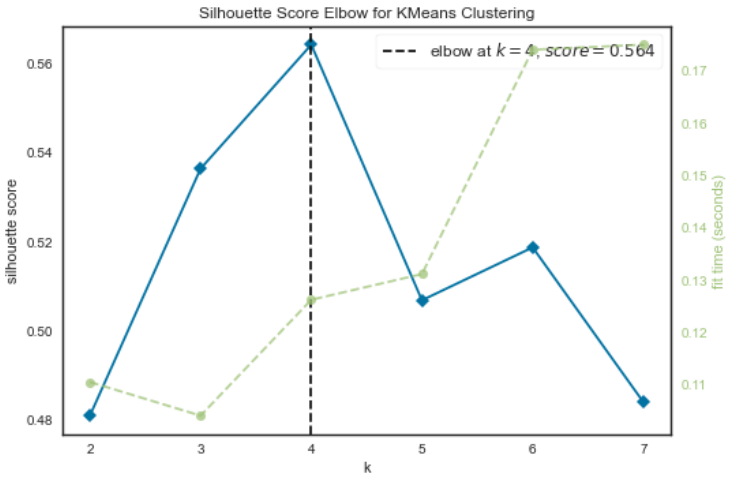
K-Means-klusteroinnin etuna on kirjallisuudessa raportoitu olevan sen helppo ymmärrettävyys, suhteellisen hyvä suoritusnopeus, soveltuvuus suurille data-aineistoille ja mahdollisuus vaikuttaa alkuasetelmaan, mikä nopeuttaa algoritmin läpivientiä (Yıldırım 2020). Algoritmin yhdeksi heikkoudeksi on kirjallisuudessa todettu se, että klustereiden lukumäärä k tulee määritellä ennakkoon. Usein se ei ole ennakkoon tiedossa, ja tiedon puute korvataan suorittamalla algoritmi useilla eri k:n arvoilla. Tämä ei kuitenkaan ole laskennallisesti kovin tehokasta, koska usein käsiteltävänä on suuria tietomääriä. (Charu & Chandan 2013.)

Klustereiden määrän arviointiin voidaan käyttää esimerkiksi kyynärpäämetodia. Siinä piirretään SSE-funktion kuvaaja eri k:n arvoilla ja tutkitaan, missä kohtaa kustannusfunktio ei enää merkittävästi pienene ja kuvaaja tekee selkeän koukkauksen. Tämä on todennäköisesti optimaalinen k:n arvo. Kuvassa 1 nähdään pystysuuntaisen katkoviivan kohdalla funktion kuvaajassa koukkaus, joten tässä tapauksessa optimaalinen k:n arvo olisi 4. (Scikit-learn a.)



Kuva 1. K:n arvon määrittäminen SSE-funktion kuvaajan avulla.

Kyynärpäämetodin lisäksi voidaan sopivaa klustereiden määrää arvioida validoinnin yhteydessä käsiteltyjen Calinski-Harabasz-indeksin, Davies-Bouldin-indeksin ja siluettikertoimen avulla. Arvot voidaan laskea etukäteen usealla eri k:n arvolla ja katsoa, mikä k:n arvo antaa parhaan tuloksen. (Scikit-learn a.) Kuten aiemmin mainittiin, ei tämä ole kovin tehokasta. Siluettikertoimen laskemisen tehokkuus on $O(n^2)$. Kuvassa 2 on esimerkki siluettikertoimen funktion kuvaajasta eri k:n arvoilla ja siihen kulunut aika.



Kuva 2. Siluettikertoimen kuvaaja k:n arvoilla 2, 3, 4, 5, 6 ja 7.

Kirjallisuudessa on todettu, että k-means-algoritmi on herkkä alkuasetelman suhteen. Lopputulokseen vaikuttavat satunnaisesti valitut klusterikeskipisteet ja havaintojen järjestys. Tämäkin ongelma ratkaistaan usein suorittamalla klusterointi useaan kertaan ja valitsemalla tuloksista paras. On myös kehitetty erilaisia alustusmenetelmiä, joilla voidaan vaikuttaa alkuasetelmaan. Näistä esimerkkinä on k-means++-algoritmi, jossa klusterikeskipisteet valitaan algoritmin suorituksen aikana perustuen todennäköisyyksiin. (Charu & Chandan 2013.)

K-means-klusteroinnissa muuttujien tulee olla vähintään välimatka-asteikollisia, jotta niiden keskiarvo voidaan laskea ja etäisyyttä mitata.

3.5 K-means-algoritmin variaatioita

K-means-algoritmista on kehitelty paljon erilaisia variaatioita, jotka soveltuvat erilaisiin data-aineistoihin ja joilla voidaan vaikuttaa edellisessä osiossa mainittuihin heikkouksiin. Tässä luvussa esitellään lyhyesti muutama näistä variaatioista: k-median-, k-medoids-, k-modes- ja k-prototypes-algortimit. (Charu & Chandan 2013.)

3.5.1 K-median

K-median-algoritmi perustuu k-means-algortmiin, mutta siinä klusterikeskipisteiksi valitaan havaintojen mediaani keskiarvon sijaan. Tässä algortimissa käytetään etäisyyksien mittaamisen useimmiten Manhattan-etäisyyttä. (Charu & Chandan 2013.)

K-median algoritmi toimii k-means-algoritmia tehokkaammin sellaisen datan klusterointiin, jossa on poikkeavia havaintoja, koska poikkeavat havainnot vaikuttavat keskiarvoon herkemmin kuin mediaaniin. Whelanin, Harrellin ja Wangin mukaan pitäisi 50 % havainnoista olla poikkeavia, että ne vaikuttaisivat oleellisesti mediaaniin. (Whelan ym. 2015.)

3.5.2 K-medoids

K-medoids-algoritmi on k-medians-algoritmin ohella k-means-algoritmia vastustuskykyisempi poikkeaville havainnoille. Siinä klusterikeskipisteiksi valitaan oikeat havaintopisteet ja ne päivitetään tutkimalla jokaisella iteraatiolla, miten kustannusfunktio käyttäytyy, jos klusterikeskipiste vaihdettaisiin joksikin toiseksi. (Charu & Chandan 2013.)

3.5.3 K-modes ja K-prototypes

K-means-algoritmi ei sovellu luokkamuuttujien klusterointiin SSE-funktion ja keskiarvon laskemisen vuoksi. Sen sijaan k-modes-algoritmia voidaan tällaisissa tapauksissa käyttää. Tässä algoritmissa käytetään aikaisemmin kuvattua

Hamming-etäisyyttä etäisyyden laskentaan, jonka perusteella havainnot osoitetaan tiettyyn klusteriin. (Charu & Chandan 2013.)

Kun k-means-algortimi soveltuu vain numeeriselle datalle ja k-modes kategoriselle datalle, voidaan k-prototypes-algoritmia käyttää data-aineistoon, jossa on sekä numeerisia että luokkamuuttujia. Eri muuttujia käsitellään algoritmissa niiden luonteelle sopivin menetelmin. (Aprilliant 2021.)

4 Analyysin toteutus

4.1 Analyysin lähtökohdat, tavoitteet ja rajaus

Analyysin tavoitteena oli selvittää, voidaanko UniSportin asiakastietojärjestelmään kertyneen datan perusteella muodostaa toiminnan kannalta hyödylliset asiakassegmentit klusterointimenetelmien avulla. Tavoitteena oli saada parempi ymmärrys siitä, miten asiakkaat käyttävät UniSportin palveluita, keitä palveluiden käyttäjät ovat ja miten tietyllä tavalla käyttäytyvät asiakkaat ovat jakautuneet. Ajatuksena oli, että yksittäisten muuttujien tarkastelun sijaan klusteroinnin avulla voitaisiin asiakkaiden käyttäytymisestä saada kokonaisvaltaisempi käsitys.

Mikäli segmentit saataisiin muodostettua, voitaisiin työtä hyödyntää hinnoittelun, palveluiden, viestinnän ja markkinoinnin suunnittelussa. Kun tiedetään, miten ja mitä palveluita asiakkaat käyttävät, voidaan se huomioida erilaisten kausikortti- tai jäsenyysmallien tarjonnassa. Esimerkiksi käytön perusteella voidaan pohtia, onko ryhmäliikunta- ja kuntosalikortit tarpeen eriyttää toisistaan. Segmentoinnin avulla voidaan myös löytää palveluita vähän käyttävät asiakkaat ja sitouttaa heitä palveluiden käyttöön markkinoinnin ja viestinnän keinoin sekä kohdentaa markkinointia tietyille asiakkaille tietyissä kanavissa. Tulosten pohjalta voidaan pohtia, miten mahdollisesti palveluiden karsiminen tai lisääminen vaikuttavat asiakkaiden sitoutumiseen. Lisäksi automatisoidun analyysimallin avulla asiakassegmentit on helppo jatkossa päivittää.

Data kerättiin asiakastietojärjestelmästä, johon on tallennettu asiakkaiden perustiedot sekä tiedot asiakkaiden ostoista ja palveluiden käytöstä. Opinnäytetyötä varten ei tehty erillistä asiakaskyselyä tai haastattelututkimusta, jonka avulla olisi saatu syvällistä tietoa asiakkaiden arvoista ja motiiveista, vaan tavoitteena oli tässä vaiheessa saada asiakastietojärjestelmän dataan pohjautuva käsitys asiakkaiden käyttäytymisestä, jonka pohjalta tutkimusta voidaan myöhemmin mahdollisesti jatkaa esimerkiksi asiakaskyselyiden muodossa.

Analyysi rajattiin koskemaan vain kausikorttiasiakkaiden käyttäytymistä ja kausikorttipalveluiden käyttöä. Kausikorttipalveluita ovat kuntosaliharjoittelu, ryhmäliikunta ja pallopelivuorot. UniSportilla ei ole käytössä jatkuvaa jäsenyysmallia, vaan asiakkaat voivat ostaa kausikortteja 12, 4 ja 1 kuukaudeksi kerrallaan.

4.2 Analyysissa käytetyt työkalut ja menetelmät

Opinnäytetyössä käytettävä data kerättiin tuotantokäytössä olevasta asiakastietojärjestelmästä, johon asiakastiedot kertyvät reaaliajassa. Järjestelmä tarjoaa admin-käyttäjälle käyttöliittymän, jonka avulla tietokannasta on mahdollista hakea raportteja erilaisin ryhmittelyin ja suodattimin csv-muodossa.

Data ladattiin asiakastietojärjestelmästä Microsoft Excel 365 -ohjelmistoon, jossa tarpeettomat henkilötiedot, kuten nimi- ja yhteystiedot poistettiin. Tämän jälkeen data tuotiin esikäsiteltäväksi ja analysoitavaksi web-pohjaiselle Jupyter Notebook -alustalle, jossa ohjelmointikielenä käytettiin Pythonia. Tätä varten tietokoneelle ladattiin Anaconda-ohjelmisto. Datamuunnoksissa ja mallinnuksessa käytettiin avoimen lähdekoodin Python-kirjastoja, kuten Scikit-learn-, Pandas-, NumPy-, Seaborn- ja Yellowbrick-kirjastoja.

Analyysi tehtiin k-means-klusterointialgoritmia sekä sen variaatioita hyödyntämällä. Lisäksi erilaisia etäisyysmittoja sekä alustus- ja standardointimetelmiä testattiin parhaan lopputuloksen löytämiseksi. Lopuksi eri menetelmien tuottamia tuloksia verrattiin keskenään.

4.3 Data-aineisto

Analyysia varten asiakastietojärjestelmästä ladattiin kolme eri raporttia, jotka yhdistettiin yhdeksi data-aineistoksi esikäsittelyvaiheen aikana. Analyysissa käytettiin palvelujen käytön dataa keväältä 2022 kahden kuukauden ajalta. Aineistoon valikoituivat seuraavat muuttujat:

- ikä
- asiakasryhmä
- kausikortin tyyppi
- pallopelivuorojen käyntitiheys
- kuntosalikäyntien tiheys
- ryhmäliikuntakäyntien tiheys
- asiakkuuden kesto
- aika, joka kulunut siitä, kun asiakas viimeksi käytti palveluita.

Seuraavissa aliluvuissa on kuvattu tarkemmin lopulliseen analyysiin valitut muuttujat, perusteet niiden valinnalle sekä aineiston esikäsittelyprosessi.

4.4 Muuttujien valinta

Analyysissa käytettävien muuttujien valintaan vaikutti tutkimusongelman lisäksi asiakastietojärjestelmästä saatavilla oleva tieto. Analyysin edetessä muuttujien tyypin ja jakauman perusteella osa muuttujista karsittiin, yhdistettiin tai muunnettiin sellaiseen muotoon, että ne soveltuivat suoritettavaan analyysiin.

Ikä valittiin yhdeksi segmentointikriteeriksi, koska kirjallisuudessa on todettu, että eri ikäisille asiakkaille toiset markkinointikanavat toimivat paremmin kuin toiset.

Asiakasryhmä sai tässä analyysissa arvokseen opiskelija, henkilökunta tai normaalihinta. Opiskelijoilla ja henkilökunnalla tarkoitetaan rahoittajayliopistojen (Helsingin yliopisto, Aalto yliopisto, Hanken, Taideyliopisto) opiskelijoita ja henkilökuntaa. Asiakasryhmän perusteella tehtävän segmentoinnin avulla voidaan eri asiakasryhmille kohdentaa markkinointia heidän käyttämissään kanavissa.

Ymmärtämällä eri asiakasryhmien käyttäytymistä voidaan myös palveluita ja hinnoittelua suunnitella asiakasryhmäkohtaisesti.

Asiakkuuden kestolla tarkoitetaan päivien lukumäärää siitä, kun asiakas on rekisteröitynyt UniSportin asiakkaaksi. Asiakkuuden kesto poimittiin mukaan analyysiin, jotta asiakkuuden vaiheesta saatiin tietoa. Uudet asiakkaat, jotka eivät välttämättä ole vielä kovin sitoutuneita palveluiden käyttöön, kaipaavat mahdollisesti erilaista viestintää kuin asiakkaat, jotka ovat jo pitkään olleet asiakkaina. On myös palveluiden suunnittelun kannalta hyvä olla tietoinen siitä, miten muutokset palvelutarjonnassa mahdollisesti vaikuttavat pitkäaikaisiin asiakkaisiin ja toisaalta taas uusiin asiakkaisiin.

Eri palveluiden käyntitiheys laskettiin jakamalla asiakkaiden käyntikerrat niiden päivien lukumäärällä, jolloin kausikortti oli tarkastelujakson aikana voimassa. Käyntitiheyden perusteella saatiin selville, mistä palveluista asiakkaat ovat kiinnostuneita ja kuinka sitoutuneesti he käyttävät palveluita. Vähemmän aktiivisten asiakkaiden aktivoimiseksi tai vain tiettyjä palveluja käyttäville asiakkaille voidaan kohdentaa markkinointia tai viestintää kohderyhmän piirteiden mukaan. Kun tiedetään, mitä palveluita asiakkaat käyttävät, voidaan arvioida, miten muutokset palvelutarjonnassa mahdollisesti vaikuttavat eri kohderyhmiin ja kuinka suurta osaa asiakaskunnasta ne todennäköisesti koskettavat.

Lisäksi analyysiin otettiin mukaan muuttuja, joka kertoo, milloin asiakas on viimeksi käyttänyt palveluita. Asiakas, joka on jo kauan aikaa sitten luopunut palveluiden käytöstä, tarvitsee mahdollisesti erilaista motivointia kuin asiakas, joka on vasta hiljattain käyttänyt palveluita.

4.5 Aineiston esikäsittely

Aineiston esikäsittelyvaiheessa muunnettiin data sopivaan muotoon, yhdistettiin lähdeaineistot analyysia varten yhdeksi aineistoksi ja käsiteltiin virheelliset ja poikkeavat havainnot.

4.5.1 Aineiston kokoaminen

Jotta halutut muuttujat saatiin koottua yhteen aineistoon analyysia varten, tuli eri aineistoissa olevia havaintoja käsitellä tarkoituksenmukaisesti ennen tietojen yhdistämistä. Tavoitteena oli tuottaa data-aineisto, jossa kukin havainto koostui yhdestä uniikista asiakkaasta ja asiakasta koskevista tiedoista. Tässä aliluvussa kuvataan pääpiirteissään, miten aineisto saatiin koottua lähdeaineiston perusteella.

Aineisto kerättiin kolmesta eri datalähteestä: pääsyoikeusraportilta, leimausraportilta ja asiakasraportilta. Kuvissa 3, 4 ja 5 nähdään pieni näyte kunkin lähdeaineiston sisällöstä alkutilanteessa. Raportit luettiin Jupyter Notebookissa, jossa pääsyoikeusraportilta poimittiin asiakkaiden tili-id:t, asiakasryhmä-id:t sekä pääsyoikeuksien tiedot. Pääsyoikeuksien tiedoista säilytettiin pääsyoikeuden tyyppi-id sekä alku- ja loppupäivän tiedot. Leimausraportilta poimittiin leimauspäivä, tili-id, palvelu-id ja leimausten määrä. Asiakasraportilta mukaan aineistoon tuli tili-id, syntymäpäivä ja asiakkaan luontipäivä.

	access_right_id	account_id	access_right_template_id	charging_timestamp	calculated_expiration_date	agreement_id
0	74448	62895	7	2020-07-27 08:05:49	2021-05-16	3
1	89282	78757	18	2020-10-16 10:00:30	2020-10-16	5
2	18327	2336	4	2019-12-18 19:26:43	2020-02-25	\N
3	87182	70215	2	2020-10-02 07:41:16	2020-10-21	\N
4	107448	101	21	2021-08-03 17:00:40	2021-08-03	248
...	...	...	...	...	...	...

Kuva 3. Ote pääsyoikeusraportin tiedoista lähtötilanteessa, kun ylimääräiset muuttujat oli suodatettu pois.

	day	account_id	service_id	quantity
0	2022-04-11	111926	14.0	1
1	2022-05-07	105578	3.0	1
2	2022-02-17	111799	2.0	1
3	2022-06-02	53570	2.0	1
4	2022-03-04	57057	2.0	1
...	...	...	...	...

Kuva 4. Ote leimausraportin tiedoista lähtötilanteessa, kun ylimääräiset muuttujat oli suodatettu pois.

	account	date_of_birth	created
0	112865	[illegible]	2021-09-01
1	77704	[illegible]	2019-12-01
2	94115	[illegible]	2019-12-01
3	112226	[illegible]	2021-09-01
4	101015	NaN	2020-06-01
...	...	...	...

Kuva 5. Ote asiakasraportin tiedoista lähtötilanteessa, kun ylimääräiset muuttujat oli suodatettu pois.

Pääsyoikeusraportti otettiin data-aineiston pohjaksi ja siihen otettiin mukaan vain sellaiset pääsyoikeudet, joiden tyyppi oli 12 kk:n, 4 kk:n tai 1 kk:n kausikortti ja jotka olivat olleet voimassa tarkastelujaksolla. Kultakin asiakkaalta otettiin mukaan viimeisin voimassa ollut kausikortti. Kausikortin alku- ja loppupäivän perusteella luotiin apumuuttujat ensimmäistä ja viimeistä päivää varten, jolloin kausikortti oli tarkastelujaksolla voimassa. Myös voimassaolopäivien lukumäärää varten luotiin uusi muuttuja ja aineistosta poistettiin sellaiset havainnot, joissa kausikortti oli ollut voimassa tarkastelujaksolla alle 14 vuorokautta. Myös UniSportin henkilökunnan tileihin liitetyt pääsyoikeudet ja sellaiset pääsyoikeudet, joissa kausikortin loppu- tai alkupäivä puuttui tai niiden katsottiin olevan asetettu virheellisesti, poistettiin. Asiakasryhmä-id ja pääsyoikeuden tyyppi-id

korvattiin tekstimuotoisilla arvoilla. Kuvan 6 datanäyte havainnollistaa pääsyoikeusdatan muotoa käsittelyn jälkeen.

	account_id	access_right	agreement	first_d	last_d	ar_valid
0	59894	12 kk	Opiskelijat	2022-04-01	2022-05-31	60
1	75011	12 kk	Normaalihinta	2022-04-04	2022-05-31	57
2	116530	12 kk	Opiskelijat	2022-04-01	2022-05-31	60
3	109961	12 kk	Opiskelijat	2022-04-01	2022-05-31	60
4	2832	12 kk	Opiskelijat	2022-04-01	2022-05-31	60
...	...	...	...	...	...	...

Kuva 6. Näyte pääsyoikeusaineistosta käsittelyn jälkeen.

Leimausraportilta mukaan aineistoon otettiin kunkin asiakkaan leimaukset, jotka oli tehty asiakkaan viimeisimmän kausikortin voimassaolon aikana. Leimaukset ryhmiteltiin tili-id:n mukaan ja kunkin palvelun leimausten lukumäärä tuotiin omiin sarakkeisiinsa. Viimeisimmän leimauksen päivämäärälle lisättiin myös oma sarakkeensa. Kuvassa 7 nähdään leimausaineiston muoto kuvatun käsittelyn jälkeen.

	account_id	last_seen	gym	group_training	ball_sports	quantity
0	104649	2022-05-31	13.0	1.0	NaN	1
1	113439	2022-05-29	14.0	3.0	NaN	1
2	82505	2022-05-18	3.0	NaN	NaN	1
3	118052	2022-05-31	14.0	5.0	NaN	1
4	60151	2022-05-30	NaN	11.0	NaN	1
...	...	...	...	...	...	...

Kuva 7. Näyte leimausaineistosta käsittelyn jälkeen.

Kun pääsyoikeudet ja leimaukset oli saatu oikeanlaiseen muotoon, yhdistettiin aineistot tili-id:n perusteella ja lisättiin asiakasraportilta aineistoon asiakkaan syntymäaika ja asiakkaan rekisteröitymispäivä. Syntymäaika puuttui 5 havainnosta. Nämä korvattiin tässä vaiheessa päivällä 1.1.1900. Palveluiden käytön

osalta ”Nan”-arvot korvattiin nollalla. Luontipäivä puuttui 26 havainnosta. Nämä havainnot poistettiin aineistosta kokonaan. Mikäli asiakas ei ollut käyttänyt palveluita tarkastelujaksolla ollenkaan, merkittiin viimeiseksi leimauspäiväksi tarkastelujakson ensimmäinen päivä.

Kun puuttuvat arvot oli käsitelty, laskettiin syntymäpäivän perusteella asiakkaan ikä ja viimeisen leimauspäivän avulla päivien lukumäärä siitä, kun asiakas oli käyttänyt palveluita edellisen kerran. Lisäksi palveluiden käyttökerrat jaettiin kausikortin voimassaolopäivien lukumäärällä ja kerrottiin 30:lla, jotta saatiin arvo, joka kertoo, kuinka monta kertaa keskimäärin asiakas käytti kausikorttiaan kuhunkin palveluun kuukaudessa kausikortin voimassaolon aikana. Aineistoon lisättiin myös muuttuja, joka kertoo kausikorttipalveluiden käyntitiheyden yhteensä siltä varalta, että se toimisi eri palveluiden käyntitiheyttä paremmin klusteroinnissa.

Edellä kuvattujen toimenpiteiden jälkeen aineistosta poistettiin tili-id, kausikortin ensimmäinen ja viimeinen päivä sekä tieto kausikortin voimassaoloajasta. Aineistosta poistettiin myös asiakkaat, joilla ei ollut kausikortti voimassa tarkastelujakson viimeisimpänä päivänä. Lopputuloksena saatiin kuvan 8 mukainen aineisto. Käsittelyn jälkeen aineistoon jäi yhteensä 9 muuttujaa ja 9831 havaintoa.

	access_right	agreement	gym	group_training	ball_sports	training_card_services	age	last_seen	created
0	12 kk	Opiskelijat	3.5	0.0	0.0	3.5	19.819178	43.0	119
1	12 kk	Normaalihinta	4.0	0.0	0.0	4.0	26.717808	5.0	912
3	12 kk	Henkilökunta	5.0	0.0	0.0	5.0	36.405479	2.0	364
4	12 kk	Opiskelijat	0.0	0.0	0.0	0.0	27.890411	0.0	912
5	12 kk	Opiskelijat	0.0	3.5	0.0	3.5	26.819178	21.0	912
...	...	...	...	...	...	...	...	...	...

Kuva 8. Klusterointia varten yhdistetty data-aineisto.

4.5.2 Muuttujien jakaumat ja poikkeavien havaintojen käsittely

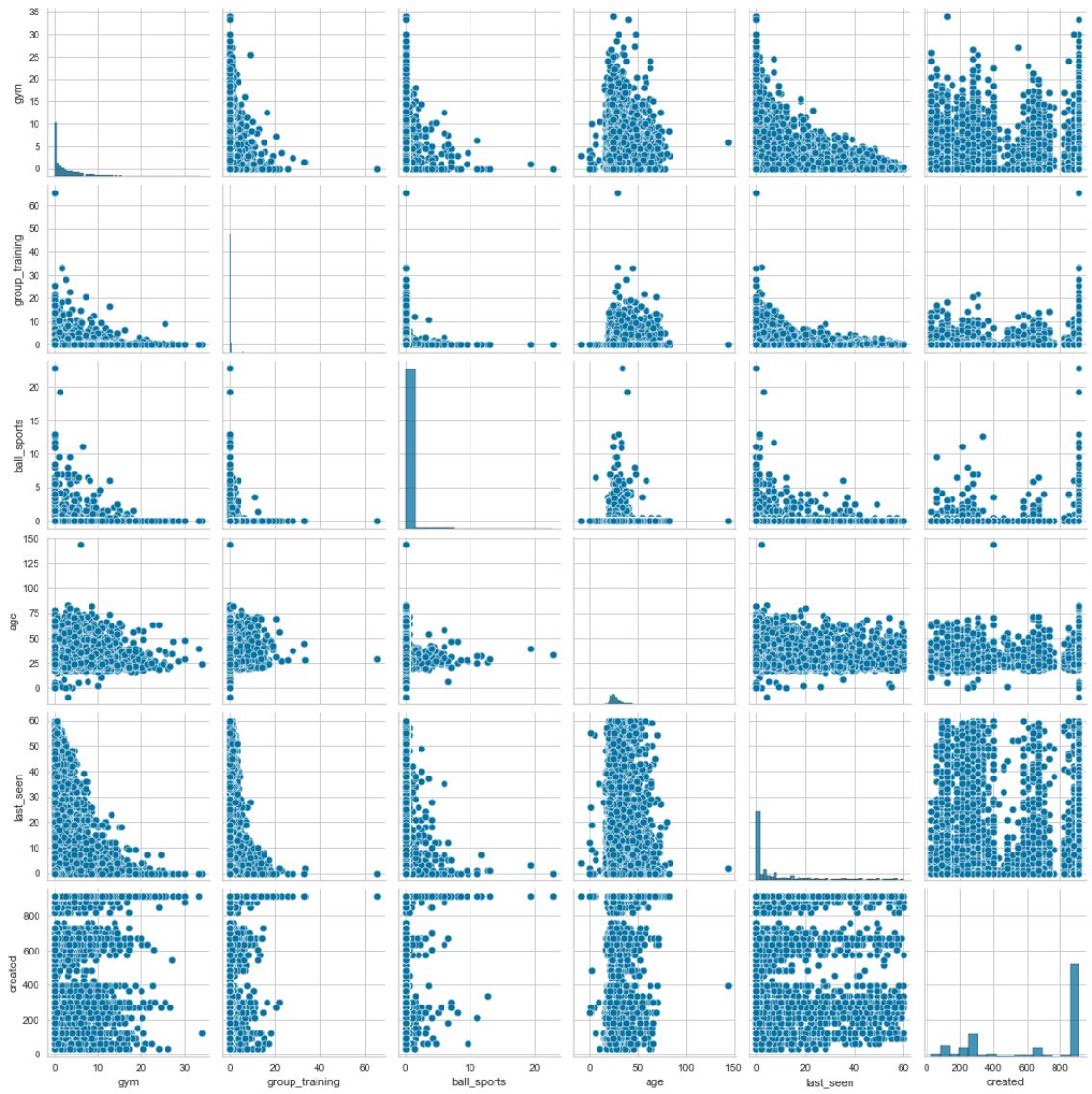
Kun lähdeaineisto oli saatu yhdistettyä yhdeksi data-aineistoksi klusterointia varten, tarkasteltiin muuttujien jakaumia ja tutkittiin, löytyykö aineistosta poik-

keavia havaintoja. Ennen muuttujien yksittäistä tarkastelua, tutkittiin numeeristen muuttujien perustunnuslukuja Pandas-kirjaston describe-funktion avulla tuotetusta yhteenvedosta. Funktion avulla saatiin luotua taulukko, joka näkyy kuvassa 9 ja josta nähdään muuttujien minimi- ja maksimiarvot, keskiarvo, keskihajonta ja kvantiilit.

	gym	group_training	ball_sports	training_card_services	age	last_seen	created
count	9831.000000	9831.000000	9831.000000	9831.000000	9831.000000	9831.000000	9831.000000
mean	3.149589	0.777331	0.067246	3.994166	28.963305	10.839182	664.061133
std	4.143026	2.190089	0.622560	4.430071	9.281515	15.538148	314.277344
min	0.000000	0.000000	0.000000	0.000000	-9.106849	0.000000	30.000000
25%	0.000000	0.000000	0.000000	0.500000	23.194521	0.000000	303.000000
50%	1.500000	0.000000	0.000000	2.500000	26.210959	3.000000	912.000000
75%	5.000000	0.500000	0.000000	6.000000	31.323288	15.000000	912.000000
max	34.000000	65.625000	22.777778	65.625000	143.967123	60.000000	912.000000

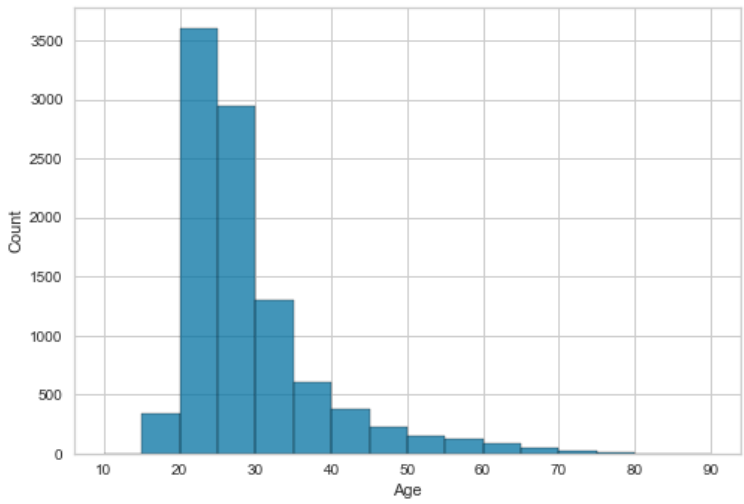
Kuva 9. Muuttujien tunnusluvut.

Tunnuslukujen tarkastelun lisäksi piirrettiin aineistosta hajontakuvio muuttujapareittain muuttujien suhteiden ja poikkeavien havaintojen havainnollistamiseksi (kuva 10).



Kuva 10. Hajontakuvio muuttujapareittain ennen poikkeavien havaintojen käsittelyä.

Kuvista 9 ja 10 nähdään, että ikä-muuttujan arvot ovat välillä -9-144. Koska UniSportilla alaikäraja on 12, pääteltiin tätä alempien arvojen olevan virheellisiä arvoja. Ne korvattiin ikämuuttujan mediaanilla. Samoin tehtiin myös yli 100 oleville arvoille. Muutosten jälkeen ikäjakauma oli kuvan 11 mukainen. Vaihteluväli oli 14-83 ja keskiarvo 29.



Kuva 11. Ikäjakauma.

Palvelujen käytön muuttujat: kuntosali, ryhmäliikunta ja pallopelivuorot kertovat, kuinka monta kertaa keskimäärin asiakas on käyttänyt kyseistä palvelua kuukaudessa. Hajontakuviosta ja tunnusluvuista nähdään, että muuttujista piirretyt histogrammit olivat voimakkaasti oikealle vinoutuneita. Erityisesti ryhmäliikunnan ja palloilun arvot sisälsivät paljon nollia ja muutamia paljon muuta aineistoa korkeampia arvoja. Nolla-arvot ja poikkeavat havainnot olivat kuitenkin tärkeitä analyysin kannalta, joten niitä ei haluttu kokonaan poistaa. Asia ratkaistiin tuomalla suurimpia arvoja hieman lähemmäs muuta aineistoa, jotta asteikko ei venynyt niiden vuoksi liiaksi. Tämä tehtiin Winsorizer-kirjaston capper-metodin avulla. Tämä tehtiin vain 0,1 prosentille arvoista, mikä auttoi tiivistämään asteikkoa, mutta arvot silti olivat mukana vaikuttamassa analyysiin. Käsittelyn jälkeen pallopelivuorot saivat arvoja 0:n ja 8:n väliltä, kuntosali 0:n ja 26:n väliltä ja ryhmäliikunta 0:n ja 19:n väliltä. Pallopelien keskiarvo oli 0,06, kuntosalin 3,1 ja ryhmäliikunnan 0,77. Muuttujien frekvenssit olivat kuvien 12, 13 ja 14 mukaiset. Kuvista nähdään, että muuttujat saivat paljon lähellä nollaa olevia arvoja.

ball_sports	f	%
[0, 1)	9659	98.25 %
[1, 5)	133	1.35 %
[5, 10)	39	0.40 %

Kuva 12. Pallopelivuorojen (käynnit/kk) frekvenssijakauma.

gym	f	%
[0, 1)	4110	41.81 %
[1, 5)	3233	32.89 %
[5, 10)	1657	16.85 %
[10, 15)	607	6.17 %
[15, 20)	177	1.80 %
[20, 25)	34	0.35 %
[25, 30)	13	0.13 %

Kuva 13. Kuntosalikäyntien (käynnit/kk) frekvenssijakauma.

group_training	f	%
[0, 1)	7890	80.26 %
[1, 5)	1455	14.80 %
[5, 10)	374	3.80 %
[10, 15)	83	0.84 %
[15, 20)	29	0.29 %

Kuva 14. Ryhmäliikuntakäyntien (käynnit/kk) frekvenssijakauma.

Asiakkuuden luontipäivä kertoo päivinä ajan, joka oli tarkastelujakson viimeisenä päivänä kulunut siitä, kun asiakas oli rekisteröitynyt asiakkaaksi. Kuvasta 10 nähdään, että asiakkaita oli rekisteröitynyt sykleissä. Asiakkuuden keston maksimiarvo oli 912. Tämä on ajankohta, jolloin nykyinen asiakastietojärjestelmä otettiin käyttöön eikä dataa tätä aikaisemmalta ajalta ole. Keskimäärin asiakkaat ovat rekisteröityneet 664 päivää sitten eli hieman alle 2 vuotta sitten. Kuvassa 15 on frekvenssitaulukko, josta nähdään, kuinka suuri osa asiakkaista on rekisteröitynyt alle vuosi sitten, 1-2 vuotta sitten ja yli kaksi vuotta sitten.

created	f	%
[0, 365)	3081	31.34 %
[365, 730)	1029	10.47 %
[730, 1000)	5721	58.19 %

Kuva 15. Luontipäivän frekvenssijakauma.

Viimeisintä käyntiä kuvaava ”last_seen”-muuttuja kertoo, kuinka kauan sitten asiakas on viimeksi käyttänyt palveluita. Tässä 0 tarkoittaa, että asiakas on käyttänyt palveluita tarkastelujakson viimeisenä päivänä ja suurempi arvo kertoo päivien lukumärän tarkastelujakson viimeisestä päivästä asiakkaan viimeisimpään käyntiin. Muuttuja saa maksimiarvonsa 60, kun asiakas on käynyt viimeksi tarkastelujakson ensimmäisenä päivänä tai ennen sitä. Keskimäärin asiakkaat ovat käyneet viimeksi 10 päivää sitten. Kuvassa 16 nähdään, miten asiakkaat jakautuvat ”last_seen”-muuttujan suhteen. Noin 60 % asiakkaista on käyttänyt palveluita viikon sisällä, 73 % kahden viikon sisällä ja 87% kuukauden sisällä.

last_seen	f	%
[0, 7)	5923	60.51 %
[7, 14)	1208	12.34 %
[14, 30)	1362	13.91 %
[30, 60)	1295	13.23 %

Kuva 16. Viimeisin käynti -muuttujan frekvenssijakauma.

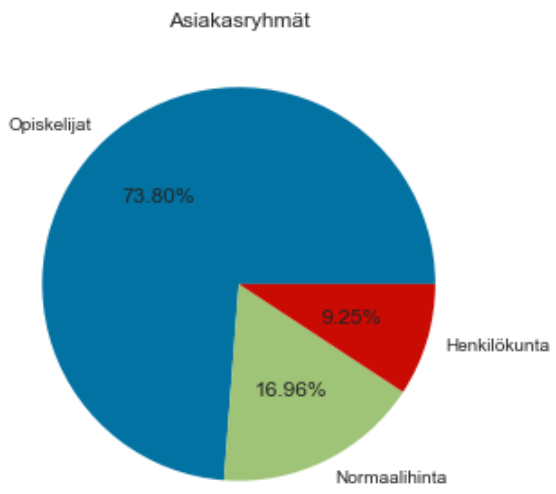
Kausikorttipalvelut-muuttujassa on yhdistetty pallopelien, kuntosalin ja ryhmäliikunnan käynnit eli se kertoo, kuinka paljon asiakas on käyttänyt kaikkia kausikorttipalveluita kuukaudessa yhteensä. Koska tässäkin muuttujassa oli arvoja, jotka poikkesivat huomattavasti muista arvoista, tehtiin sille samanlainen tiivistysoperaatio kuin yksittäisille kausikorttipalveluiden muuttujille. Operaation jälkeen muuttujan arvot vaihtelivat 0:n ja 29 välillä. Muuttujan keskiarvo oli 4,0. Kuvassa 17 on taulukoituna muuttujan frekvenssijakauma.

training_card_services	f	%
[0, 1)	2754	28.01 %
[1, 5)	3912	39.79 %
[5, 10)	2097	21.33 %
[10, 15)	765	7.78 %
[15, 20)	237	2.41 %
[20, 25)	45	0.46 %
[25, 30)	21	0.21 %

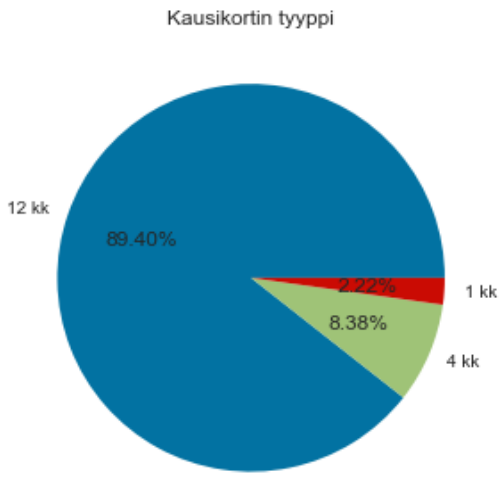
Kuva 17. Kausikorttipalveluiden käyntitiheyden (käynnit/kk) frekvenssijakauma.

Asiakasryhmä kertoo, onko asiakas rahoittajayliopiston opiskelija tai henkilökuntaa vai onko hän muu eli normaalihintainen asiakas. Kausikortin tyyppi kertoo kausikortin pituudesta. Kausikortti voi olla 1:n, 4:n tai 12 kuukauden pituinen.

Asiakasryhmä ja kausikortin tyyppi ovat luokkamuuttujia, joita ei ole tarkoitus käyttää varsinaisessa klusterianalyysissä. Niitä käytetään kuitenkin jälkianalyysissa näyttämään, onko mahdollisesti muodostuneiden ryhmien välillä eroavaisuuksia näiden muuttujien suhteen. Kuvista 18 ja 19 nähdään, miten kausikortin tyyppi ja asiakasryhmät ovat jakautuneet. Yli 70 % asiakkaista on opiskelijoita ja noin 90 % on 12 kuukauden kausikortti.



Kuva 18. Asiakasryhmien jakauma.



Kuva 19. Kausikortin tyypin jakauma.

Kuvaan 20 on koottu numeeristen muuttujien keskiarvot, keskihajonnat sekä minimi- ja maksimiarvot tässä osiossa tehtyjen muunnosten jälkeen. Siitä nähdään, että palveluiden käytön ja iän muuttujien vaihteluvälit ja muut tunnusluvut ovat hieman muuttuneet.

	group_training	gym	ball_sports	training_card_services	age	last_seen	created
count	9831.000000	9831.000000	9831.000000	9831.000000	9831.000000	9831.000000	9831.000000
mean	0.767140	3.146235	0.062347	3.986924	28.992690	10.839182	664.061133
std	2.036438	4.122563	0.515929	4.371775	9.140865	15.538148	314.277344
min	0.000000	0.000000	0.000000	0.000000	14.301370	0.000000	30.000000
25%	0.000000	0.000000	0.000000	0.500000	23.210959	0.000000	303.000000
50%	0.000000	1.500000	0.000000	2.500000	26.227397	3.000000	912.000000
75%	0.500000	5.000000	0.000000	6.000000	31.320548	15.000000	912.000000
max	19.048571	25.523182	8.085000	28.519474	82.761644	60.000000	912.000000

Kuva 20. Muuttujien tunnusluvut muutosten jälkeen.

4.6 Mallinnus

Kun datan rakenne oli selvillä ja virheelliset arvot käsitelty, siirryttiin mallintamaan data-aineistoa. Analyysi vietiin läpi eri muuttujayhdistelmiä, standardointimenetelmiä, etäisyysmittoja ja algoritmeja kokeilemalla.

Tässä vaiheessa lopullisesta analyysista päätettiin jättää pois ikä-muuttuja. Tämä olisi ollut mielenkiintoista pitää analyysissa mukana, mutta se jätettiin pois, koska muuttujia oli jo paljon, eikä haluttu, että syntyy ”dimensionaalisuuden kirous” -efektiä. Iän ei katsottu myöskään olevan niin käyttökelpoinen tutkimuksen tavoitteen kannalta kuin muiden muuttujien. Dimensionaalisuuden vähentämiseksi päätettiin myös jakaa aineisto kahteen osaan ja suorittaa klusterointi kahdella eri data-aineistolla. Toisessa aineistossa oli muuttujina kuntosali, ryhmäliikunta, pallopelivuorot ja viimeisin käynti. Toisessa aineistossa oli muuttujina kausikorttipalvelut, joka laskettiin palvelukohtaisista arvoista, asiakkuuden kesto ja viimeisin käynti -muuttuja. Selvyyden vuoksi kutsutaan ensimmäiseksi mainittua aineistoa data-aineisto 1:ksi ja toiseksi mainittua aineistoa data-aineisto 2:ksi.

4.6.1 Aineiston standardointi

Ennen algoritmin suorittamista aineisto standardoitiin. Standardointiin käytettiin Skicit-learn-kirjaston metodeja. Koska aineisto oli vinoutunut ja sisälsi jonkin verran poikkeavia havaintoja, kokeiltiin standardointiin Scikit-learn-kirjaston Power transformer ja Quantile transformer -metodeja. Power transformer muokkaa datan enemmän normaalijakauman muotoon ja minimoi vinouman (Scikit-learn b). Quatile transformer muokkaa jakaumaa myös normaalijakauman suuntaan ja hajauttaa kaikkein yleisimmät arvot vähentäen poikkeavien havaintojen vaikutusta (Scikit-learn c). Nämä eivät kuitenkaan tuottaneet mielekästä klusterointitulosta.

Edellä mainittujen standardointimenetelmien lisäksi aineiston starndardointiin kokeiltiin Min-Max-skaalausta sekä Z-score-standardointia. Mielekkäimmän tuloksen antoi Z-score-standardointi, jota päätettiin käyttää molempien aineistojen standardoimiseen.

Data-aineisto 1 koostui neljästä muuttujasta, joiden standardoiduista arvoista muodostettu tunnuslukutaulukko nähdään kuvassa 21. Tässä otsikko 0 on kuntosali, 1 ryhmäliikunta, 2 pallopeliryhmät ja 3 aika viimeisimmästä käynnistä. Kuvasta voidaan havaita, että pallopeliryhmien skaala on muita jonkin verran suurempi, mikä voi aiheuttaa sen, että pallopeliryhmät vaikuttavat dominoivasti klusteroinnin tulokseen.

	0	1	2	3
count	9.831000e+03	9.831000e+03	9.831000e+03	9.831000e+03
mean	4.183127e-16	-5.769862e-16	5.852132e-16	-1.360003e-15
std	1.000051e+00	1.000051e+00	1.000051e+00	1.000051e+00
min	-7.632132e-01	-3.767261e-01	-1.208506e-01	-6.976207e-01
25%	-7.632132e-01	-3.767261e-01	-1.208506e-01	-6.976207e-01
50%	-3.993434e-01	-3.767261e-01	-1.208506e-01	-5.045377e-01
75%	4.496861e-01	-1.311869e-01	-1.208506e-01	2.677944e-01
max	5.428197e+00	8.977619e+00	1.555070e+01	3.164040e+00

Kuva 21. Standardoidun data-aineisto 1:n tunnusluvut.

Data-aineisto 2 koostui kolmesta muuttujasta, joiden standardoitujen arvojen perusteella muodostettu tunnuslukutaulukko on kuvassa 22. Tässä 0 on viimeisin käynti, 1 on asiakkuuden kesto ja 3 kausikorttipalveluiden käyntitiheys.

	0	1	2
count	9.831000e+03	9.831000e+03	9.831000e+03
mean	-7.549765e-16	-1.531342e-17	-6.072291e-17
std	1.000051e+00	1.000051e+00	1.000051e+00
min	-6.976207e-01	-2.017624e+00	-9.120156e-01
25%	-6.976207e-01	-1.148920e+00	-7.976398e-01
50%	-5.045377e-01	7.889575e-01	-3.401365e-01
75%	2.677944e-01	7.889575e-01	4.604944e-01
max	3.164040e+00	7.889575e-01	5.611862e+00

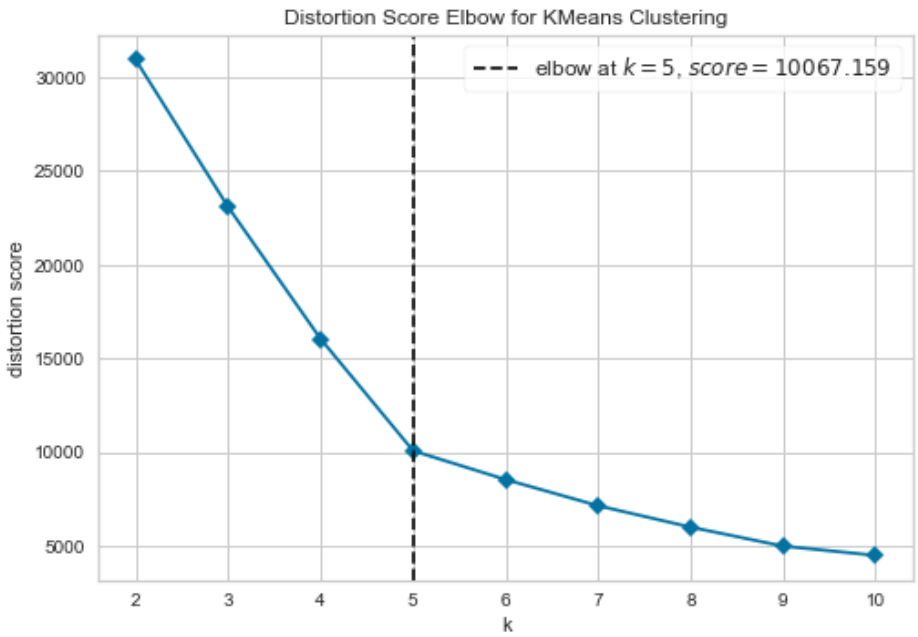
Kuva 22. Standardoidun data-aineisto 2:n tunnusluvut.

4.6.2 Algoritmin suoritus

Analyysin suorittamiseen käytettiin k-means- ja k-medoids-algoritmeja. Myös k-prototypes-algoritmia kokeiltiin niin, että luokkamuuttujat saatiin mukaan analyysiin, mutta heti ensimmäisistä kokeiluista nähtiin, että se ei tuota tulosta, joten sitä ei nähty merkitykselliseksi raportoida tässä yhteydessä suuremmin. Luokkamuuttujia päätettiin käyttää erikseen jälkianalyysissa. Analyysi suoritettiin aluksi aineistolle 1, jonka klusterointiprosessia käsitellään opinnäytetyössä seuraavaksi.

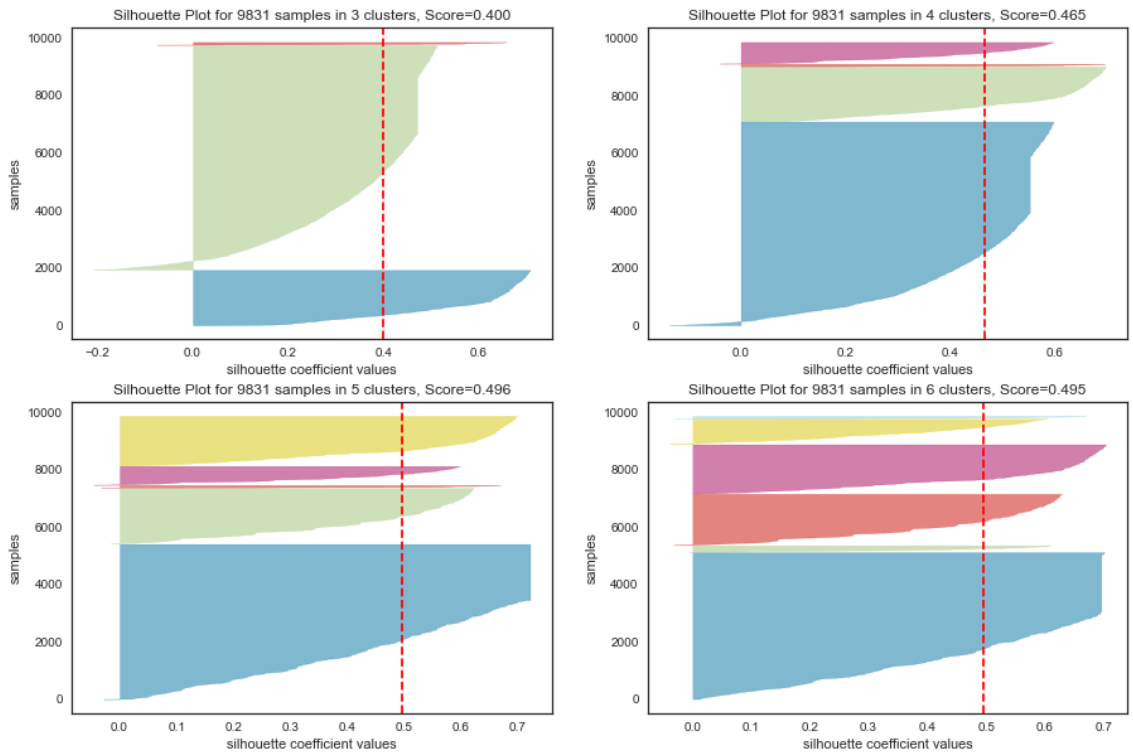
Ennen algoritmin suorittamista valittiin sopiva k:n arvo data-aineistolle. Ennakkotiedon pohjalta oletuksena oli, että klustereita muodostuisi 3-6. Jotta sopiva klustereiden määrä saatiin selville, katsottiin, mikä kyynärpäämetodin ja siluettikertoimen perusteella olisi paras k:n arvo.

Kuvassa 23 on piirretty virheiden neliösummafunktion kuvaaja, joka kuvaa klustereiden tiiveyttä eri k:n arvoilla. K:n arvon 5 kohdalla kuvaaja tekee kyynärpään näköisen koukkauksen eikä funktion arvo enää oleellisesti laske. Tämän analyysin mukaisesti sopiva klustereiden määrä olisi 5.



Kuva 23. SSE-funktion kuvaaja k:n arvoilla 2-10.

Kuvassa 24 on visualisoitu havaintojen siluettikertoimet, kun klusterointi suoritetaan k-means-algoritmilla ja k:n arvo on 3, 4, 5 tai 6. Kuvassa samaan klusteriin kuuluvat havainnot ovat saman värisiä. Vasemman laidan asteikon avulla nähdään, kuinka paljon havaintoja kuuluu kuhunkin klusteriin ja alalaidan asteikosta nähdään siluettikertoimen arvot. Pystysuuntainen katkoviiva kertoo, mikä on kaikkien havaintojen siluettikertoimien keskiarvo. Kuvasta nähdään, että siluettikertoimen keskiarvo on noin 0,5 koko havaintoaineistolle k:n arvolla viisi ja kuusi. Joidenkin havaintojen siluettikertoimet ovat alle nollan, mikä kertoo siitä, että kyseiset havainnot ovat siluettikertoimen tulkinnan mukaan luokiteltu väärään klusteriin. Kaikilla k:n arvoilla algoritmi muodostaa yhden hyvin pienen klusterin ja k:n arvolla kuusi kaksi pientä klusteria.



Kuva 24. Havaintojen siluettikertoimet.

K-means-algoritmi suoritettiin Scikit-learn KMeans-metodia käyttäen. Kuvassa 25 nähdään, mitkä parametrit valittiin k-means-algoritmin suoritukseen. Klustereiden lukumääräksi valikoitui lopulta viisi ja alustusmenetelmäksi k-means++. Scikit-learnin KMeans-metodissa ei ole mahdollisuutta erikseen valita etäisyysmittaa, vaan se on automaattisesti euklidinen etäisyys.

```
kmeans = cluster.KMeans(n_clusters=5, init='k-means++', n_init=10, max_iter=100, random_state=42)
kmeans.fit(df_scaled)
```

Kuva 25. K-means-algoritmin suoritus ja käytetyt parametrit.

K:n arvoksi valittiin viisi kuuden sijaan, koska kuudella klusterilla muodostui muuten samantapaiset klusterit kuin viidellä, mutta siinä ryhmäliikuntakävijät erotettiin useampaan klusteriin käyntikertojen perusteella ja syntyi kaksi hyvin pientä klusteria. Tämän ei katsottu olevan tarpeellista opinnäytetyön tavoittei-

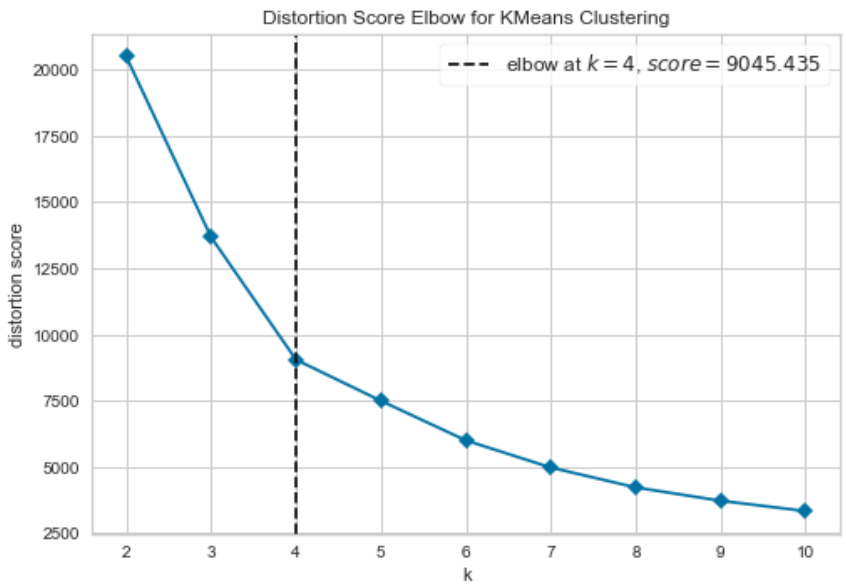
den kannalta. Myös SSE-funktion kuvaajan ja siluettikertoimen tarkastelun perusteella 5 oli sopiva k:n arvo. K-means-klusteroinnin tuloksiin palataan seuraavassa luvussa.

K-medoids-algoritmin suorittamiseen käytettiin Scikit-learn-kirjaston metodia KMedoids, jossa voidaan erikseen valita, mitä etäisyysmittaa käytetään. Analyysia kokeiltiin euklidista-, Manhattan- ja kosinisamankaltaisuusmittaa käyttäen. K:n arvona käytettiin samaa arvoa kuin k-means-algortimin kanssa. Kuvassa 26 nähdään parametrit, joilla jäljempänä kuvatut tulokset syntyivät.

```
kmedoids = KMedoids(n_clusters=5, init='k-medoids++', metric='cosine', max_iter=100, method='pam').fit(df_scaled)
```

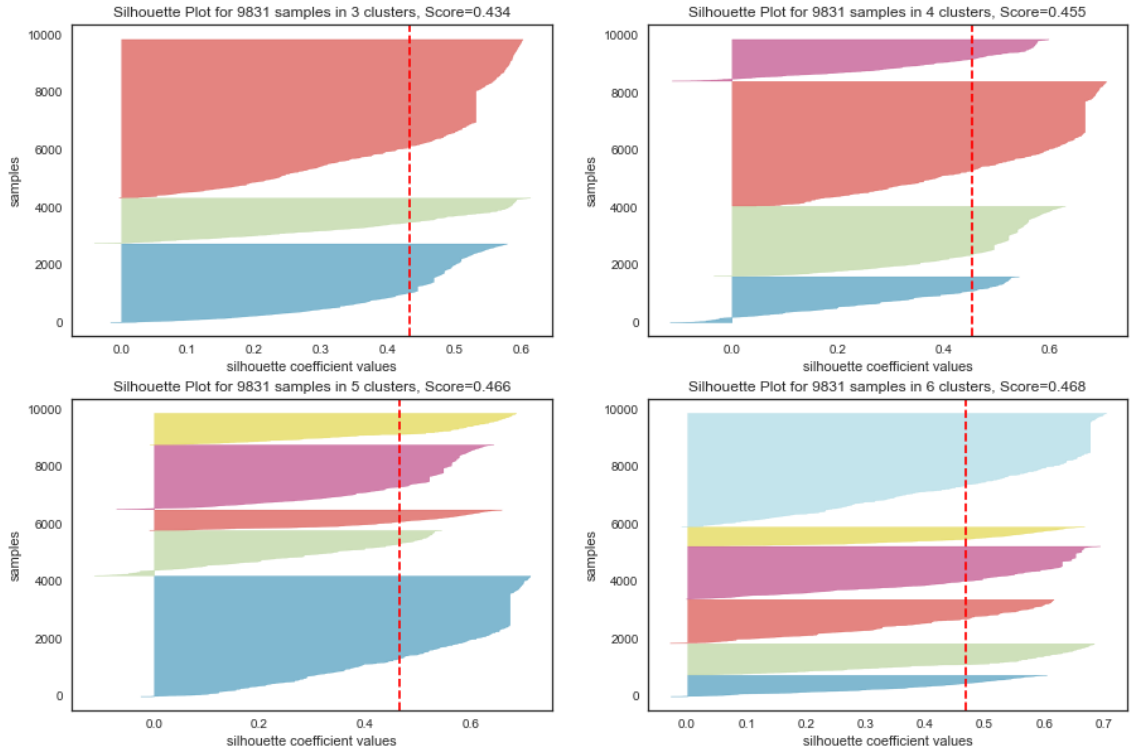
Kuva 26. K-medoids-algoritmin suorituksessa käytetyt parametrit

Data-aineisto 1:n tapaan aineistolle 2 etsittiin sopiva k:n arvo kyynärpäämetodia ja siluettikerrointa hyödyntäen. Kuvassa 27 kyynärpäämetodi osoittaa, että hyvä klustereiden määrä olisi 4. Tosin tämän jälkeen kustannusfunktio laskee vielä reilusti.



Kuva 27. SSE-funktion kuvaaja k:n arvoilla 2-10.

Kuvaan 28 piirretty kuvio siluettikertoimien arvoista näyttää, että siluettikertoimen keskiarvo olisi korkeimmillaan k:n arvoilla 5 tai 6. Analyysia kokeiltiin kummallakin k:n arvolla ja päädyttiin lopulta valitsemaan data-aineiston 2 klusterointiin k:n arvoksi 6.



Kuva 28. Data-aineisto 2:n siluettikertoimet k:n arvoilla 3, 4, 5 ja 6.

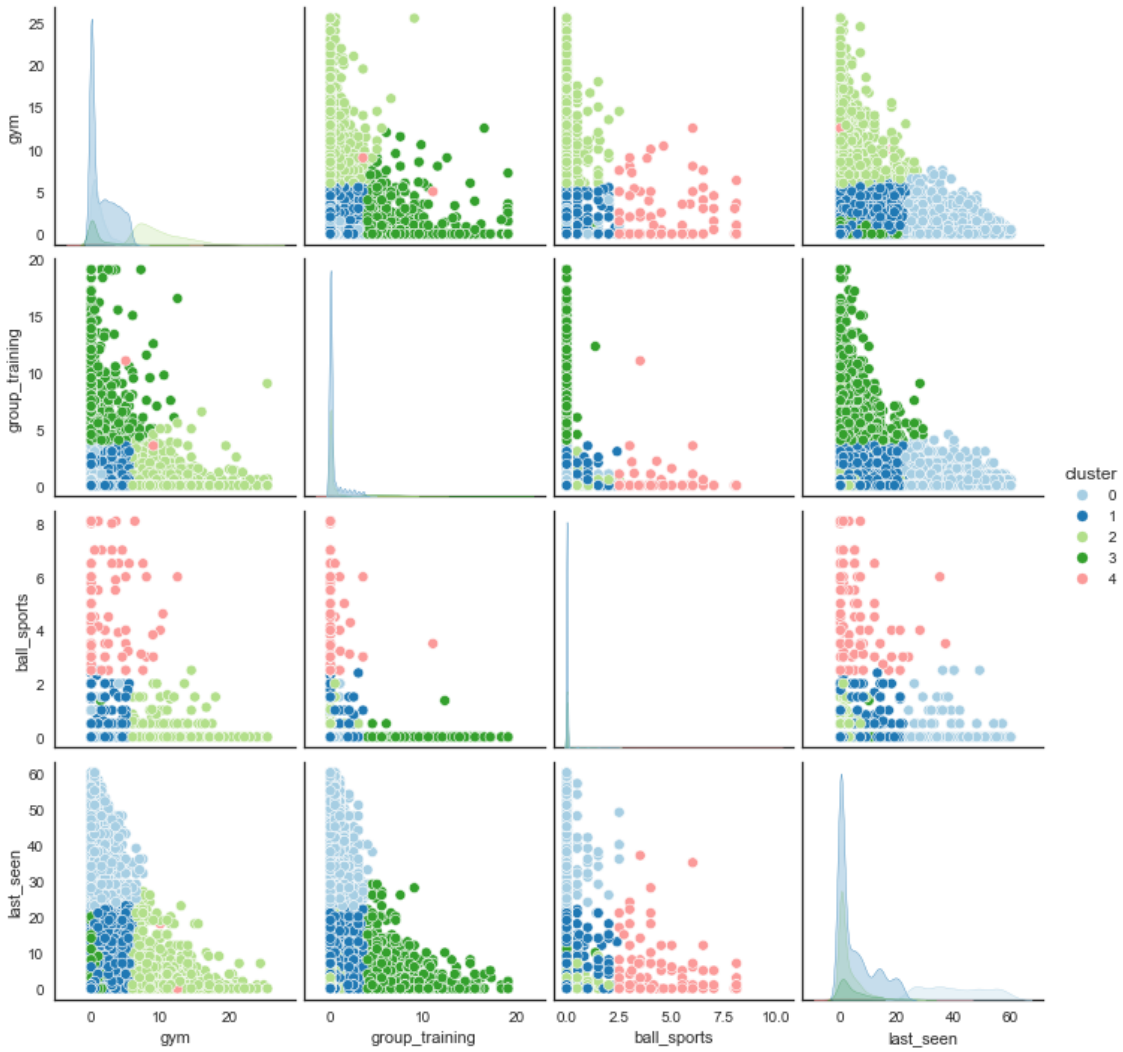
4.7 Tulokset

Data-aineistolle 1 saatiin k-means-algoritmista tuloksena kuvassa 29 näkyvät klusterikeskipisteet. Kuvasta nähdään kunkin klusterin havaintojen osuus koko data-aineistosta sekä eri muuttujien keskiarvot klusterien sisällä.

	f	%	gym	group_training	ball_sports	last_seen
0	1739	18 %	1.11	0.34	0.02	40.57
1	5409	55 %	1.59	0.35	0.02	5.03
2	1939	20 %	10.07	0.20	0.03	2.76
3	653	7 %	1.01	7.11	0.00	4.48
4	91	1 %	2.15	0.37	4.83	5.57

Kuva 29. Data-aineisto 1:n klusterikeskipisteet ja klustereiden frekvenssit.

Kuvassa 30 on hajontakuvio muuttujapareittain, jossa on eri värein kuvattu kuhunkin klusteriin kuuluvat havainnot.



Kuva 30. Data-aineiston 1 hajontakuvio muuttujapareittain ja klustereittain.

Kuvasta 29 nähdään, että klusteriin 0 kuuluu 18 % asiakkaista. Tähän klusteriin kuuluvat asiakkaat eivät juuri käytä palveluita ja ovat keskimäärin käyttäneet palveluita viimeksi yli 40 päivää sitten. Kuvaan 30 tämä klusteri on merkitty vaaleansinisellä. Kuvasta voidaan tulkita, että osa asiakkaista on käyttänyt palveluita aiemmin, mutta ovat lopettaneet palveluiden käytön tai jääneet tauolle.

Suurin muodostunut klusteri on klusteri 1, johon kuuluu 55 % havainnoista. Siinä kaikkien palveluiden käyttö on vähäistä. Ryhmäliikunnassa tähän klusteriin kuuluvat asiakkaat käyvät keskimäärin kerran kolmessa kuukaudessa ja kuntosalilla alle kaksi kertaa kuukaudessa. Kuvasta 30, jossa kyseinen ryhmä on merkitty tummansinisellä, voidaan tulkita, että osa näistä asiakkaista kuitenkin käy kuntosalilla ja/tai ryhmäliikunnassa vähän useammin, maksimissaan kerran viikossa. Tähän ryhmään kuuluvat asiakkaat ovat keskimäärin käyttäneet palveluita viimeksi viisi päivää sitten, joten he eivät todennäköisesti ole kokonaan lopettaneet palveluiden käyttöä, vaikka käyttävätkin niitä harvakseltaan.

Klusteriin 2 kuuluu 20 % havainnoista. Nämä asiakkaat käyttävät paljon kuntosalipalveluita, mutta eivät kovin paljon ryhmäliikuntaa tai pallopelivuoroja. Klusteri on merkitty vaaleanvihreällä kuvaan 30. Siitä nähdään, että jotkut asiakkaat käyvät myös muissa palveluissa, mutta pääsääntöisesti tähän ryhmään kuuluvat asiakkaat käyttävät kuntosalipalveluita.

Klusteri 3 on ryhmäliikkujien klusteri. Tähän klusteriin kuuluvat asiakkaat käyttävät ryhmäliikuntaa enemmän kuin muita palveluita ja käyvät palveluissa aktiivisesti - ryhmäliikunnassa keskimäärin 7 kertaa kuukaudessa ja kuntosalilla keskimäärin 1 kerran kuukaudessa. 7 % asiakkaista kuuluu tähän klusteriin.

Klusterissa 4 on havainnot, joissa pallopelivuorojen muuttujan korkea arvo korostuu. Tähän klusteriin kuuluvat asiakkaat käyvät pallopelivuorojen lisäksi myös kuntosalilla keskimäärin noin 2 kertaa kuukaudessa. Klusteriin kuuluu vain 1 % havainnoista. Havainnot on merkitty kuvaan 30 punaisella.

Kuvissa 31 ja 32 on taulukoituna eri asiakasryhmien ja kausikorttityyppien prosenttiosuudet klustereittain. Klusterien havaintojen jakaumat noudattelevat

melko hyvin kaikkien havaintojen jakaumia. Kuvasta 31 nähdään kuitenkin, että klusterissa 3, jossa ryhmäliikunta korostuu, on muita klustereita enemmän normaalihintaisia asiakkaita. Klusterissa 4, johon on jakautunut verrattain paljon pallopelivuoroilla kävijöitä, on enemmän henkilökuntaan kuuluvia asiakkaita kuin muissa klustereissa. Kuvasta 32 nähdään, että niissä klustereissa, joissa palvelujen käyttö on vähäistä (0 ja 1), on suhteessa enemmän 12 kk:n kausikorttilaisia kuin muissa klustereissa.

agreement	0	1	2	3	4	Total
Henkilökunta	9 %	9 %	8 %	13 %	15 %	9 %
Normaalihinta	15 %	16 %	17 %	26 %	18 %	17 %
Opiskelijat	76 %	74 %	75 %	61 %	67 %	74 %

Kuva 31 Data-aineisto 1:n asiakasryhmien osuudet klustereittain.

access_right	0	1	2	3	4	Total
1 kk	0 %	1 %	5 %	7 %	10 %	2 %
12 kk	91 %	92 %	84 %	81 %	85 %	89 %
4 kk	9 %	7 %	11 %	12 %	5 %	8 %

Kuva 32. Data-aineisto 1:n kausikorttityyppien osuudet klustereittain.

Data-aineisto 1:een sovellettiin k-means-algoritmin lisäksi k-medoids-algoritmia eri etäisyysmittoja käyttäen. Kuvassa 33 nähdään muodostuneet klusterikeskipisteet ja niiden frekvenssit, kun algoritmi suoritettiin kosinisamankaltaisuusmittaa käyttäen. Siitä nähdään, että algoritmi löytää samankaltaiset klusterikeskipisteet kuin k-means-algoritmi. Poikkeuksena k-means-algoritmiin on, että k-medoids-algoritmi ei erottele pallopelivuorolla kävijöitä omaksi klusterikseen, vaan sen sijaan kaksi klusteria kuntosalikävijöille, joissa toisessa havaintojen keskiarvo kuntosalimuuttujassa on 4 ja toisessa 9. Havainnot ovat jakautuneet

klustereihin k-medoids-algoritmin tuottamassa klusteroinnissa k-means-algoritmin tuottamaa tulosta tasaisemmin.

	f	%	gym	group_training	ball_sports	last_seen
0	1036	11 %	4.00	0.00	0.00	3.00
1	2409	25 %	1.18	0.00	0.00	33.00
2	2856	29 %	0.50	0.00	0.00	2.00
3	1296	13 %	0.50	4.00	0.00	7.00
4	2234	23 %	9.00	0.00	0.00	4.00

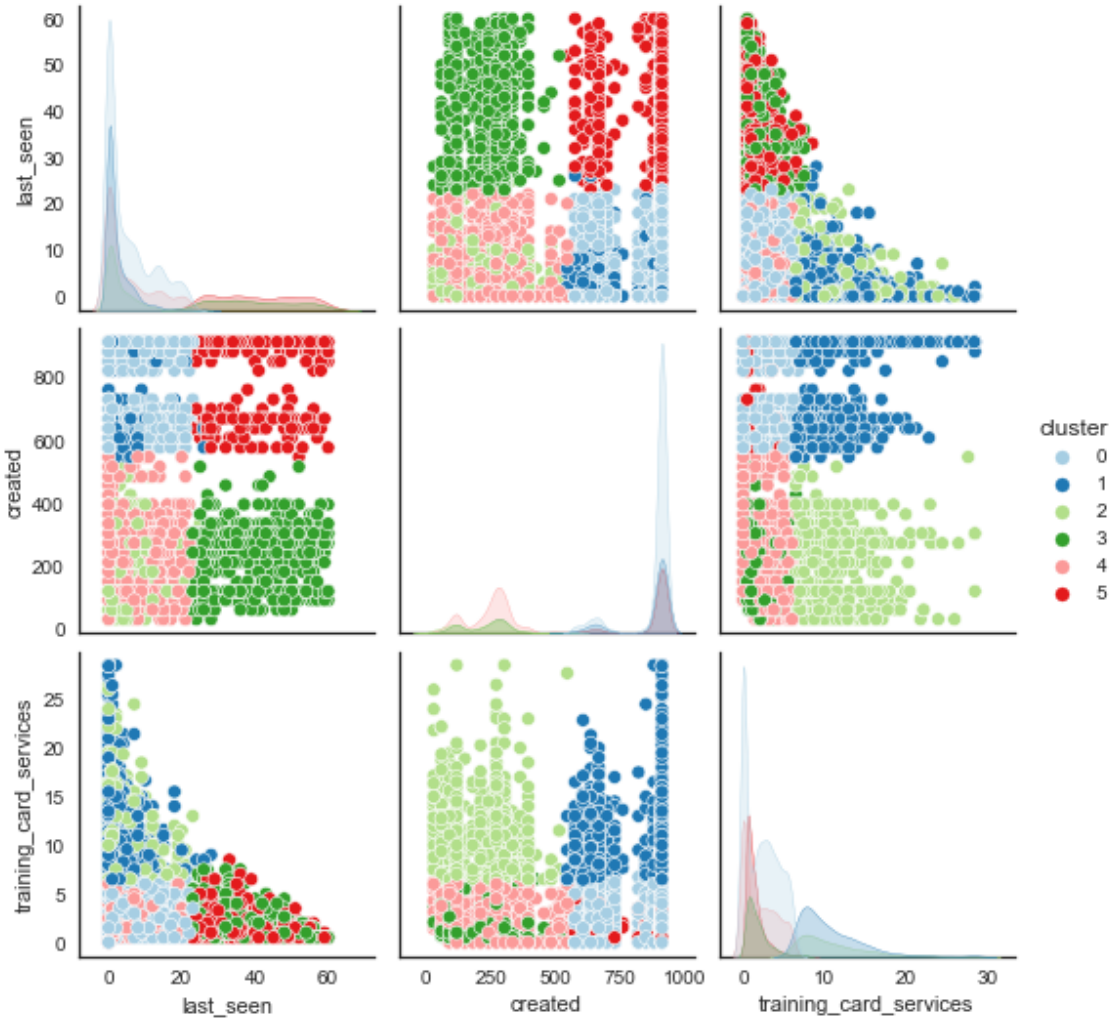
Kuva 33. Data-aineisto 1:n klusterikeskipisteet k-medoids-algoritmia käytettäessä.

Data-aineistolle 2 löytyi k-means-klusteroinnissa kuvan 34 mukaiset klusterikeskipisteet.

	f	%	last_seen	created	training_card_services
0	3959	40 %	5.14	878.57	2.23
1	1530	16 %	2.37	862.63	10.53
2	741	8 %	3.07	214.74	10.61
3	665	7 %	39.74	229.78	1.74
4	1845	19 %	5.05	248.24	2.03
5	1092	11 %	40.84	880.34	1.38

Kuva 34. Data-aineisto 2:n k-means-klusteroinnin klusterikeskipisteet ja klustereiden frekvenssit.

Kuvassa 35 on hajontakuvio muuttujapareittain, jossa on eri värein kuvattu kuhunkin klusteriin kuuluvat havainnot.



Kuva 35. Data-aineisto 2:n hajontakuvio muuttujapareittain ja klustereittain.

Kuvista 34 ja 35 nähdään, että klusteri 0, joka on merkitty hajontakuvioon vaaleansinisellä, on suurin klusteri. Siihen kuuluu 40 % asiakkaista. Nämä asiakkaat ovat pidempiaikaisia asiakkaita, jotka käyttävät palveluita verrattain vähän. He ovat kuitenkin käyneet palveluissa viimeisen kuukauden aikana.

Klusteriin 1, joka on merkitty hajontakuvioon tummansinisellä, kuuluvat asiakkaat, jotka käyttävät verrattain paljon palveluita ja ovat olleet asiakkaana jo yli kaksi vuotta. Heidän osuus kaikista asiakkaista on 16 %.

Klusteriin 2, joka on merkitty hajontakuvioon vaaleanvihreällä, kuuluvat asiakkaat, jotka käyttävät kausikorttipalveluita verrattain paljon. He ovat tulleet asiakkaiksi viimeisen vuoden aikana ja käyttäneet palveluita viimeisen kuukauden sisällä. Näiden asiakkaiden osuus kaikista asiakkaista on 8 %.

Klusteri 3 on merkitty hajontakuvioon tummanvihreällä. Tähän klusteriin kuuluvat asiakkaat ovat olleet asiakkaina alle vuoden eivätkä ole käyttäneet palveluita yli kuukauteen. Heidän osuus kaikista asiakkaista on 7 %.

Klusteri 4 on merkitty hajontakuvioon vaaleanpunaisella. He ovat kuten klusteri 0, mutta ovat uudempia asiakkaita. Heidän osuus kaikista asiakkaista on 19 %.

Klusteriin 5, johon kuuluvat havainnot on merkitty hajontakuvioon tummanpunaisella, kuuluvat asiakkaat, jotka eivät ole viimeisen kuukauden aikana käyttäneet palveluita ja ovat olleet asiakkaina yli kaksi vuotta. Heidän osuus asiakkaista on 11 %.

Kuvissa 36 ja 37 on taulukoituna eri asiakasryhmien ja kausikorttityyppien prosenttiosuudet klustereittain. Jakauma on melko tasainen. Klustereissa 2, 3 ja 4, joissa on uudempia asiakkaita, on suhteessa enemmän opiskelijoita kuin muissa klustereissa. Kuvan 37 mukaan klusterissa 2, jossa on uudempia asiakkaita ja jotka käyttävät palveluita verrattain paljon, on enemmän sellaisia asiakkaita, joilla on lyhyempi kuin 12 kuukauden kausikortti.

agreement	0	1	2	3	4	5	Total
Henkilökunta	12 %	10 %	6 %	6 %	5 %	11 %	10 %
Normaalihinta	18 %	22 %	12 %	11 %	13 %	18 %	18 %
Opiskelijat	70 %	68 %	81 %	84 %	82 %	71 %	72 %

Kuva 36. Data-aineisto 2:n asiakasryhmien osuudet klustereittain.

access_right	0	1	2	3	4	5	Total
1 kk	1 %	5 %	8 %	0 %	2 %	0 %	3 %
12 kk	93 %	88 %	72 %	86 %	89 %	93 %	89 %
4 kk	6 %	7 %	19 %	13 %	9 %	6 %	9 %

Kuva 37. Data-aineisto 2:n kausikorttityyppien osuudet klustereittain.

4.8 Arviointi

Aliluvussa 3.3.5 käsiteltiin klusteroinnin validointimenetelmiä, jotka voidaan jakaa ulkoisiin ja sisäisiin menetelmiin.

Sisäisissä menetelmissä klusteroinnin onnistumista arvioidaan erilaisilla mittareilla, jotka perustuvat klusteroinnin tiiveyden ja erottuvuuden mittaamiseen. Tässä opinnäytetyössä käytettiin mittareita, jotka mittaavat näitä molempia ulottuvuuksia: siluettikerrointa, Calinski-Harabasz-indeksiä ja Davies-Bouldin-indeksiä.

Data-aineisto 1 sai k-means-klusteroinnin tulosten osalta siluettikertoimen arvoksi 0,50, Davies-Bouldin-indeksin arvoksi 0,65 ja Calinski-Harabasz-indeksin arvoksi 7139. Data-ainesto 2 sai siluettikertoimen arvoiksi 0,46, Davies-Bouldin-indeksin arvoksi 0,75 ja Calinski-Harabasz-indeksin arvoksi 7695.

K-medoids-algortimin tuottamalle klusteroinnille laskettiin vain siluettikerroin, koska sen yhteydessä oli käytetty kosinisamankaltaisuutta etäisyysmittana eikä se toimi Calinski-Harabasz- tai Davies-Bouldin-indeksin kanssa. Siluettikertoimen arvoksi kosinisamankaltaisuusmittaa käyttäen saatiin 0,7.

Tuloksista nähdään, että klusterit muodostuivat, mutta ne eivät ole erityisen tiiviit ja toisistaan erottuvat. Klusterit ovat muodostuneet säännönmukaisesti kaikkien dimensioiden suhteen eikä visualisoinneissa näy vähemmän tiheää havaintopisteiden aluetta.

Kahden algoritmin vertailussa k-medoids-algoritmi toimi siluettikertoimen mukaan k-means-algoritmia paremmin. Kahden eri data-aineiston vertailussa data-aineistolle 1 tehty klusterointi sai hieman paremmat arvot siluettikertoimelle ja Davies-Bouldin-indeksille kuin aineistolle 2 tehty klusterointi. Calinski-Harabasz-indeksi oli data-aineiston 2 klusteroinnissa hieman korkeampi.

Ulkoisissa menetelmissä klusteroinnin hyvyyttä arvioidaan ulkoisen tiedon valossa. Tässä opinnäytetyössä aiempaa tietoa havaintojen oikeista luokista ei ollut eikä samanlaista tutkimusta ennestään ole tehty. Voidaan kuitenkin tutkia, löytyykö opinnäytetyössä muodostuneista segmenteistä jotain samankaltaisuutta aiemmissa tutkimuksissa kuvattuihin asiakasprofiileihin ja onko syntyneet klusterit käyttökelpoiset toiminnan kehittämisen kannalta.

Falardeaun ym. (2021) laatiman tutkimuksen mukaan 55 % asiakkaista on passiivisia osallistujia, jotka harrastavat liikuntaa epäsäännöllisesti. Myös UniSport x Frankly -loppumateriaalissa (2020) tunnistettiin aloittelevien liikkujien asiakasprofiili, johon kuuluvat asiakkaat liikkuvat satunnaisesti. Tämän opinnäytetyön analyysissa löytyi asiakkaiden ryhmä, jotka käyttivät UniSportin palveluita n. 2 kertaa kuukaudessa, mutta jotka vähäisistä käynneistä huolimatta olivat käyttäneet palveluita lähiaikoina. Näihin asiakkaisiin luokittui noin 50-60 % asiakkaista menetelmästä riippuen. Näihin asiakkaisiin saattaisi palvelujen käytön perusteella kuulua aiemmissa tutkimuksissa kuvattuja aloittelevia tai passiivisia liikkujia.

UniSport x Frankly -loppumateriaalissa (2020) kuvataan, että tavoitteelliset liikkujat liikkuvat usein ja monipuolisesti. Tämän opinnäyteyön analyysissa löytyi paljon palveluita käyttävien ryhmä, mutta siinä paljon palveluita käyttävät asiakkaat eriytyivät joko kuntosalipalveluiden tai ryhmäliikunnan suosijoiksi. Klusteriin, jossa palveluiden käyntitiheyden arvo on verrattain korkea ja jossa asiakkaaksi rekisteröitymisestä on kulunut pitkä aika, voisi kuulua Falardeaun ym. (2021) kuvailemia perinteisiin tukeutujia. Klusteriin, jossa ryhmäliikunta saa korkean arvon, saattaa kuulua UniSport x Franklyn materiaalissa kuvattuja ryhmässä liikkujia.

Analyysissa sellaisten asiakkaiden osuus, jotka eivät olleet yli kuukauteen käyttäneet palveluita oli noin 20 %. Näistä noin 40 % kuului klusteriin, jossa asiakkuus oli kestänyt alle vuoden. Vastaavanlainen tulos on saatu aiemmin lopettavien asiakkaiden tutkimuksesta (Unisport x Frankly -loppumateriaali 2020).

Tämän aiempiin tutkimuksiin perustuvan tarkastelun pohjalta voidaan sanoa, että analyysissa muodostuneista segmenteistä löytyy samanlaisia piirteitä ja jakaumia kuin opinnäytetyössä esitellyissä aiemmissa tutkimuksissa, mutta koska tutkimukset on toteutettu hyvin eri tavoin ja aiemmassa tutkimuksessa asiakasprofiilien luontiin on sisällytetty paljon enemmän erilaisia muuttujia kuin tässä tutkimuksessa ja ne on toteutettu hyvin eri tavalla, ei voida suoraan arvioida analyysin hyvyyttä näiden tutkimusten pohjalta. On myös hyvä huomata, että aiemmissa tässä opinnäytetyössä käsitellyissä tutkimuksissa on analysoitu asiakkaiden liikuntakäyttäytymistä yleensä ja tässä opinnäytetyössä tulokset pohjautuvat vain UniSportissa tapahtuneisiin käynteihin.

Vaikka klusterit eivät ole erityisen erottuvia ja tiiviitä, ovat ne UniSportin toiminnan kannalta mielekkäät. Ne ovat helposti ymmärrettävissä ja niiden pohjalta voidaan toimia. Toki on otettava huomioon, että klustereiden sisällä on hajontaa ja joidenkin asiakkaiden toiminta saattaa poiketa saman klusterin sisällä jonkin verran.

5 Yhteenveto

Opinnäytetyön tavoitteena oli selvittää, miten klusterointia voidaan hyödyntää UniSportin toiminnan kannalta hyödyllisten asiakassegmenttien muodostamisessa. Tavoitteena oli, että mikäli segmentit saadaan muodostettua, voidaan tuloksia hyödyntää palveluiden suunnittelussa, viestinnässä, markkinoinnissa sekä jatkotutkimuksissa. Tavoitteeksi asetettiin myös, että mikäli segmentit syntyvät, ne saataisiin jatkossa helposti päivitettyä.

Opinnäytetyössä perehdyttiin klusterointiin sekä asiakassegmentointiin markkinoinnin ja palvelumuotoilun näkökulmasta. Data-analyysin aineisto koottiin UniSportin asiakastietojärjestelmästä ja siihen sovellettiin k-means- ja k-medoids-

klusterointialgoritmia. Tuloksena saatiin kolmella eri tavalla muodostetut asiakassegmentit. Asiakassegmenttien muodostaminen perustui asiakkaiden palveluiden käytön aktiivisuuteen, käytettyihin palveluihin ja asiakkuuden kestoon.

Tutkimuksessa havaittiin, että muodostuneet asiakassegmentit eivät ole erityisen tiiviit ja toisistaan erottuvat, mutta ne ovat kuitenkin käyttökelpoiset. Segmenteissä oli havaittavissa samanlaisia piirteitä kuin aikaisemmissa tutkimuksissa on havaittu. Työssä saatiin myös luotua malli, jonka avulla segmenttejä on helppo jatkossa päivittää uudella data-aineistolla.

Tutkimusta rajoitti tutkimusajankohdan sijoittuminen koronapandemian ajalle. Vaikka palvelut olivatkin tarkastelujaksolla avoinna jo normaaliin tapaan, pandemia ja siihen liittyvät suositukset vaikuttivat ihmisten käyttäytymiseen. Tästä syystä tulokset eivät ole täysin sovellettavissa nykyhetkeen. Datan saatavuus oli rajoitettu myös siitä syystä, että nykyisin käytössä oleva asiakastietojärjestelmä on otettu käyttöön vasta noin kaksi vuotta ennen tutkimusajankohtaa, mistä johtuen kovin pitkältä ajalta ei saatu tietoa asiakkaiden käyttäytymisestä. Yhtenä rajoitteena oli se, että käytettävissä oli vain asiakastietojärjestelmän data, jossa on tiedot asiakkaiden ostoista ja palveluiden käytöstä, mutta ei esimerkiksi mielenkiinnonkohteista kuten verkkosivuanalytiikkatyökaluja käytettäessä on.

Opinnäytetyössä saatiin ymmärrys siitä, miten käytetyt klusterointimenetelmät soveltuvat tietyn data-aineiston klusterointiin ja asiakassegmenttien muodostamiseen. Aihe on kuitenkin laaja. Klusterointimenetelmiä on satoja ja asiakassegmentointia voidaan tehdä hyvin eri tavoin, joten tulosten yleistämistä koskemaan erilaisia tutkimusasetelmia, ei voida tämän opinnäytetyön perusteella tehdä. Tuloksia voidaan kuitenkin tällaisenaan hyödyntää UniSportin toiminnassa.

Työn tuloksena syntyneet asiakassegmentit lisäävät asiakasymmärrystä sekä auttavat palveluiden, markkinoinnin ja viestinnän suunnittelussa sekä jatkotutkimusten kohdentamisessa. Esimerkiksi klusteroinnin tuloksena syntynyttä segmenttiä, jossa asiakkaat eivät ole käyttäneet palveluita hetkeen, voidaan yrittää

houkutella takaisin palveluiden pariin kohdennetuilla markkinointiviesteillä tai heille voidaan lähettää kohdennettuja asiakaskyselyitä, joissa selvitetään syitä palveluiden käytön lopettamiselle.

Klusteroinnissa erottui omat segmentit asiakkaille, jotka eivät käytä palveluita säännöllisesti, mutta viimeisimmästä käynnistä ei ole pitkä aika sekä paljon palveluita käyttävät pitkäaikaiset kanta-asiakkaat. Molemmat ryhmät on tärkeää sitouttaa palveluiden käyttöön, mutta todennäköisesti heihin tehoaa erilainen viestintä. Segmenttien avulla sitouttaminen voidaan tehdä eri tavalla kohderyhmien tarpeet huomioiden.

Klusteroinnin tulosta voidaan edellisten esimerkkien lisäksi hyödyntää hinnoittelussa. Tuloksista nähdään, että kuntosali- ja ryhmäliikunta-asiakkaat ovat melko eriytyneet käyttämään vain toista palvelua ja he voisivat arvostaa hinnoittelumallia, jossa voi valita vain toisen kyseisistä palveluista.

Jatkossa opinnäytetyössä tehtyä tutkimusta voidaan laajentaa esimerkiksi yhdistämällä opinnäytetyön tuloksia asiakaskyselytutkimukseen. Analyysia voidaan myös käyttää pohjana uudelle analyysille, jossa voidaan käyttää uusia muuttujia ja klusterointialgoritmeja uudenlaisten mielenkiintoisten rakenteiden selvittämiseksi.

Lähteet

Agnes. 2022. Why target markets for gyms are crucial to success. Club Owners blog. Verkkoaineisto. PerfectGym. <https://www.perfectgym.com/en/blog/club-owners/why-target-markets-gyms-are-crucial-success-5-key-benefits>. Luettu 6.11.2022.

Ahonen, Laura ja Luoto, Sampo. 2015. Markkinointi boksin ulkopuolelta. Helsinki: Talentum Media Oy.

Alpaydin, Ethem ja Bach, Francis. 2014. Introduction to Machine Learning. MIT Press.

Aprilliant, Audhi. 2021. The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical). Verkkoaineisto. Towards Data Science <https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb> Luettu 1.11.2022.

Asiakasprofiili tehostaa viestintää ja markkinointia. Verkkoaineisto. Kanava.to. <https://kanava.to/asiakasprofiili>. Luettu 1.10.2022.

Baker, Kristen. Customer Segmentation: How to Effectively Segment Users & Clients. Verkkoaineisto. Hubspot. <https://blog.hubspot.com/service/customer-segmentation>. Luettu 15.9.2022.

Bergström, Seija ja Leppänen, Arja. 2021. Yrityksen asiakasmarkkinointi. Helsinki: Edita.

Brownlee, Jason. 2020. 4 Automatic Outlier Detection Algorithms in Python. Verkkoaineisto. Machine Learning Mastery. <https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/>. Luettu 6.11.2022.

Charu C. Aggarwal ja Chandan K. Reddy. 2013. Data Clustering: Algorithms and Applications. Chapman & Hall.

Collica, Randy. 2011. Customer Segmentation and Clustering Using SAS Enterprise Miner. SAS Institute.

Cooil, Bruce; Aksoy, Lerzan & Keiningham, Timothy. 2006. Approaches to Customer Segmentation. Journal of Relationship Marketing.

Cowell, Frank A. ja Victoria-Feser, Maria-Pia. 1996. Robustness Properties of Inequality Measures. The Econometric Society.

Customer Segmentation via Cluster Analysis. 2022. Verkkoaineisto. Optimove. <https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis> Luettu 2.11.2022.

Falardeau, Eric; Glynn, John & Ostromecka, Olga. 2021. Sweating for the fitness consumer. Verkkoaineisto. Mc Kinsey & Company. <https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/sweating-for-the-fitness-consumer> Luettu 1.11.2022.

Grootendorst, Maarten. 2021. Distance Measures in Data Science. Verkkoaineisto. Towards Data Science. <https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa> Luettu 5.11.2022.

Han, Jiawei; Kamber, Micheline ja Pei, Jian. 2012. Getting to know your data. Data Mining. The Morgan Kaufmann Series in Data Management Systems.

Hassani, Marwan. 2016. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. Vietnam Journal of Computer Science.

Huttunen, Kaisa. 2021. Asiakassegmentointi auttaa tuntemaan asiakkaasi. Verkkoaineisto. Zoner. <https://www.zoner.fi/digitaalinen-markkinointi/asiakassegmentointi>. Luettu 30.8.2022.

Jain, Anil K. 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. International Association for Pattern Recognition.

Jain, Anil K. ja Dubes, Richard C. 1988. Algorithms for Clustering Data. Michigan State University.

Karani, Dhruvil. 2022. A Practical Guide on K-Means Clustering. Verkkoaineisto. <https://towardsdatascience.com/a-practical-guide-on-k-means-clustering-ca3bef3c853d>. Luettu 5.10.2022.

Kassambara, Alboukadel. 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA.

Kotler, Philip ja Armstrong, Gary. 2018. Principles of marketing. Harlow: Pearson Education Limited.

Leskovec, Jure; Rajaraman, Anand & Ullman, Jeff. 2011. Mining of Massive Datasets. Cambridge University Press.

Normalization. Verkkoaineisto. Codecademy Team.<https://www.codecademy.com/article/normalization>. Luettu 6.11.2022.

Palveluiden analysointi syksy 2021. 2021. UniSport.

Scikit-learn a. Selecting the number of clusters with silhouette analysis on KMeans clustering. Verkkoaineisto. <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>. Luettu 2.9.2022.

Scikit-learn b. Power Transformer. Verkkoaineisto. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html>. Luettu 2.11.2022.

Scikit-learn c. QuantileTransformer. Verkkoaineisto. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.htm>. Luettu 2.11.2022.

Theodoridis, Sergios ja Koutroumbas, Konstantinos. 2006. Pattern Recognition. Elsevier Science & Technology.

Tuulaniemi, Juha. 2011. Palvelumuotoilu. Talentum Media Oy.

Unisport x Frankly -loppumateriaali 260820. 2020. UniSport.

Wei, Haitian. 2020. How to measure clustering performances when there are no ground truth?. Verkkoaineisto. Medium. <https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c>. Luettu 5.10.2022.

Whelan, Harrell ja Wang. 2015. Understanding the K-Medians Problem. International Conference on Scientific Computing.

Yıldırım, Soner. 2020. K-Means Clustering Explained. Verkkoaineisto. Medium. <https://towardsdatascience.com/k-means-clustering-explained-4528df86a120>. Luettu 1.10.2022.