Jan Ågren

LIIKETOIMINTATIETO

Opinnäytetyö

CENTRIA-AMMATTIKORKEAKOULU

Tietotekniikan koulutusohjelma

Tammikuu 2018

TIIVISTELMÄ OPINNÄYTETYÖSTÄ

Centria-ammattikorkea-koulu	**Aika** Tammikuu 2018	**Tekijä/tekijät** Jan Ågren
Koulutusohjelma Tietotekniikan koulutusohjelma		
Työn nimi LIIKETOIMINTATIETO		
Työn ohjaaja Sakari Männistö		**Sivumäärä** 42 + 3
Työelämäohjaaja Sakari Männistö		

Tämän opinnäytetyön tarkoitus oli tutkia liiketoimintatietoa ja sen eri aihealueita. Työ keskittyi eritoten tietovarastoinnin ja tiedonlouhinnan soveltamiseen liiketoimintaa tukevana toimintona.

Työn teoriaosuus käsittää liiketoimintatiedon kehittymisen, tietovarastoinnin perusteet ja eri arkkitehtuurit, tiedonlouhinnan prosessin ja sen yleisimmät tekniikat. Työn käytännönosuudessa visualisoidaan valittu tietojoukko tiedonlouhintaohjelmalla ja lopuksi luodaan ennuste käyttäen valittua tietojoukkoa analyysin perustana.

Asiasanat
Liiketoimintatieto, OLAP, Orange, tietovarasto, tiedonlouhinta.

Centria
UNIVERSITY OF APPLIED SCIENCES

ABSTRACT

Centria University of Applied Sciences	Date	Author/s
	January 2018	Jan Ågren

Degree programme Information Technology	
Name of thesis BUSINESS INTELLIGENCE	
Instructor Sakari Männistö	**Pages** 42 + 3
Supervisor Sakari Männistö	

The aim of this thesis was to research the field of business intelligence and its subjects. The main focus of this thesis was using data warehousing and data mining as a supportive function for running a business.

The theoretical part of the thesis was to explain business intelligence and its evolution, the basics and different architectures of data warehousing, the process of data mining and frequently used techniques of data mining. The practical part of the thesis was to visualize a selected dataset with selected data mining software and then create a prediction by using the selected dataset as the basis of the analysis.

Key words
Business intelligence, data mining, data warehouse, OLAP, Orange.

TIIVISTELMÄ
ABSTRACT
SISÄLLYS

KUVIOT

KUVAT

TAULUKOT

1 JOHDANTO

Hyvän liiketoiminnan harjoittajan voi tunnistaa hänen kyvystään käsitellä ja ymmärtää ympärillään tapahtuvia muutoksia ja niiden vaikutusta hänen liiketoimintaansa, kuten milloin myydä ja milloin ostaa. Tämänlainen ymmärtäminen syntyy yrittäjälle ajan myötä, mutta liiketoimintatiedon tarkoitus on tiedon keruulla ja sen analyysilla on tukea tällaista päätösprosessia ja opastaa datan käytössä liiketoiminnan tukena.

Tietotekniikan hyödyntämistä liiketoiminnan ja yritysten käyttötarkoituksiin on tapahtunut jo kauan aikaa, ja sen vaikutukset ovat näkyneet viimeisten vuosikymmenien aikana esimerkiksi lisääntyneenä automaationa ja palvelujen digitalisoitumisessa, mutta näistä saatua dataa on ollut vaikea käsitellä ja analysoida liiketoimintaa tukevaksi informaatioksi. Opinnäytetyön tarkoitus on avata tietovarastoinnin ja tiedonlouhinnan käyttöä liiketoiminnan tukena. Tietovarastoinnilla tarkoitetaan datan käsittelyä ja tallentamista erilaisiin tietokantoihin tai suunnitteluihin tietovarastoihin. Tiedonlouhinta on uusien mallien ja yhtäläisyyksien etsimistä ja esittämistä tietovarastossa olevasta datasta.

Työn käytännön osiossa käytetään osaa Yhdysvaltain vuoden 1994 väestönrekisteristä, josta tutkitaan, miten eri elämäntilanteet ja taustat vaikuttavat henkilön vuosituloihin, eritoten keskittyen korkeakoulututkinnon suorittaneiden tietoihin. Tämä tehdään ensimmäiseksi visualisoimalla tietojoukko ja lopuksi luomalla ennustukset, joiden pohjana käytetään erilaisia valvottuja ja ei-valvottuja tiedonlouhintatekniikoita. Työssä tutustutaan ja käytetään Ljubljanan yliopiston kehittämää ja ylläpitämää Orange-tiedonlouhintaohjelmaa, joka antaa yksinkertaisen ja selkeän tavan käydä läpi tiedonlouhinnan prosessin ja datan visualisoinnin.

2 LIIKETOIMINTATIETO

Liiketoimintatieto on sateenvarjotermi, joka yhdistää tietovarastot, tiedonlouhinnan ja analytiikan. Liiketoimintatiedon tarkoitus on tarjota käyttäjälle helpon ja vuorovaikutteisen pääsyn dataan, jonka avulla on mahdollista tehdä parempia analyysejä. Analysoimalla uutta ja vanhaa tietoa, tapahtumia ja suoritusta voidaan luoda pohja paremmalle päätöksenteolle. (Turban, Sharda, Delen & King 2011, 28-29.)

Termi liiketoimintatieto nousi esille ensimmäisen kerran Richard Millar Devensin kirjassa Cyclopaedia for Commercial and Business Anecdotes (1868, 210-211), jossa sitä käytettiin kuvailemaan taloudellista ja poliittista kokemusta tai tietoa, joilla pankkiiri Henry Furnese onnistui liiketoimissaan. (Heinze 2014.)

Vuonna 1958 IBM:n tutkija Hans Peter Luhn julkaisi artikkelin nimeltään ”A business intelligence system”, jossa hän kuvaili automaattista järjestelmää, jonka tarkoitus oli jakaa dataa yrityksen eri osastoille. Sama yhtiö vaikutti liiketoimintatiedon kehitykseen vahvasti kehittämällä kiintolevyn, joka tarjosi aivan uuden tavan tallettaa dataa. Kovalevyn kehityksen myötä alkoi myös ensimmäisten tietojärjestelmien kehitys, joita kutsuttiin päätöksentekoa tukeviksi järjestelmiksi (Decision support systems, DSS). (Heinze 2014.)

70-luvulla ilmestyi monia liiketoimintasovelluksia, joiden tarkoitus oli auttaa datan siirtämisessä tietokantoihin. Nämä tietokannat tarjosivat dataan ja raportteihin erittäin yksinkertaisen ja yksiulotteisen näkymän. Tietovarastojen tulon myötä 80-luvulla dataan saatiin kaivattua rakennetta ja järjestystä. (Joly 2016.)

90-luvulla liiketoimintatieto oli arkipäiväistä ja uusia työkaluja kehitettiin jatkuvasti. Työkalujen haaste oli niiden käytön hankaluus, sillä tavallinen käyttäjä ei päässyt käsiksi dataan ja joutui turvautumaan tietotekniikan asiantuntijoiden apuun saadakseen raportteja ja muuta dataa. Liiketoimintatiedon yleistyminen ja sen aseman kohoaminen liikemaailmassa saivat myyjät kehittämään käyttäjäystävällisempiä työkaluja, joiden avulla tavalliset käyttäjät pääsivät käsiksi tarvitsemaansa

dataan. Uusien työkalujen myötä oli vaikea taata datan laatua ja luotettavuutta, sillä kokeneiden analyytikoiden lisäksi nyt myös kokemattomat käyttäjät tekivät raportteja datasta. Tätä aikakautta kutsutaan BI 1.0:si ja sen aikana liiketoimintatiedon päätehtävät olivat datan ja raporttien luominen, datan järjestely ja sen visualisointi esitysmuotoon. (Heinze 2014.)

2000-luvun alussa teknologian kehittymisen seurauksena myös liiketoimintatieto edistyi. Tätä nykyistä kehitystä kutsutaan BI 2.0:ksi. Tietovarastot pystyivät toimimaan reaaliajassa, joka mahdollisti päätösten teon erittäin tuoreella datalla. Internetin kasvu toi mukanaan täysin uuden datan lähteen ja mahdollisuuden tutkia liiketoimintatiedon menetelmiä ja ohjelmia. Internet antoi myös alustan, jolla käyttäjät pystyivät jakamaan tietoa liiketoimintatiedon käyttökohteista ja käyttötavoista. Liikemaailman verkostoituminen loi yrityksille tarpeen reaaliajassa saatavalle datalle. Suurimmat syyt tälle tarpeelle olivat asiakkaiden halujen ymmärtäminen ja halu pysyä kilpailijoiden edellä. (Heinze 2014.)

3 TIETOVARASTOT

Tietovarasto on integroitu, aihekeskeinen ja pitkäkestoinen kokoelma dataa, jonka avulla voidaan tukea päätöksiä (Turban ym. 2011, 52–53). Tietovarasto pitää sisällään suuren joukon toisiinsa liittyviä toimintoja, joita käytetään tietovaraston suunnittelussa, toteutuksessa ja käytössä. Tietovarastossa oleva data voidaan luokitella kolmeen pääluokkaan: sisäinen data, ulkoinen data ja henkilökohtainen data. (Vercellis, 2009, 45–46.)

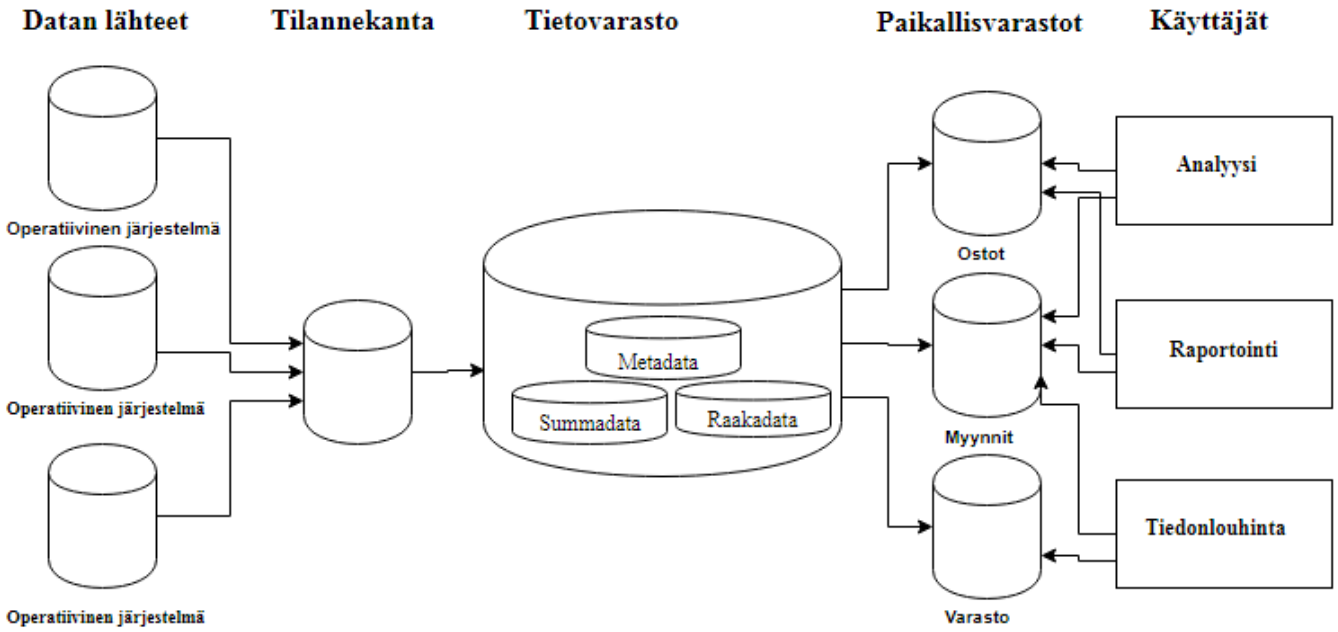
Sisäinen data tarkoittaa hallinnon, kirjanpidon, tuotannon ja logistiikan tietokannoissa ja operatiivisissa järjestelmissä esiintyvää dataa, kuten asiakastiedot, henkilöstön myynnit ja tuotteet. Yleisimmät lähteet sisäiselle datalle ovat yrityksen operatiiviset järjestelmät ja mahdolliset internet-pohjaiset järjestelmät. Operatiiviset järjestelmät keräävät dataa esimerkiksi tilauksista, varastoista, tuotannosta, ja asiakaspalvelusta. Internet-pohjaiset järjestelmät saavat tietonsa eri internet-sivuilta, mutta etenkin sivujen kautta tehdyistä tilauksista ja täytetyistä lomakkeista. (Vercellis 2009, 46.)

Ulkoisen datan tarkoitus on parantaa sisäisen datan laatua tarjoamalla selventävää dataa ja täten luoda laajempi näkemys yrityksen kokonaiskuvaan. Hyvä esimerkki tämän tyyppisestä datasta on markkinatutkimus ja paikkatietojärjestelmä (Geographical information systems, GIS), johon kuuluu erilaisia sovelluksia alueellisen datan hankintaan, organisointiin, varastointiin ja esitykseen tarkoituksena antaa laadukasta aihekohtaista ja aluekohtaista dataa. (Vercellis 2009, 46.)

Henkilökohtainen tieto on peräisin analyysien tekijöiden tai työntekijöiden tietokoneille tallentama data, kuten erilaiset laskentataulukot. Henkilökohtaisten datan integrointi sisäisen ja ulkoisen datan on yksi päämääristä tiedonhallintajärjestelmissä. (Vercellis 2009, 46.)

3.1 Arkkitehtuuri ja tyypit

Tietovaraston arkkitehtuuri koostuu kolmesta eri komponentista: tietovarasto, datan hankinta ja front-end ohjelmistot eli päätöksentekoa ja liiketoimintatiedon analyysia tukevat ohjelmistot. Tietovarasto-komponenttiin kuuluvat tietovarasto ja siihen kuuluvat paikallisvarastot, jonne data ladataan ja joka sisältää toiminnot datan kuvaamiseen, muokkaukseen ja tietenkin pääsyn dataan. (Vercellis, 2009, 51.)

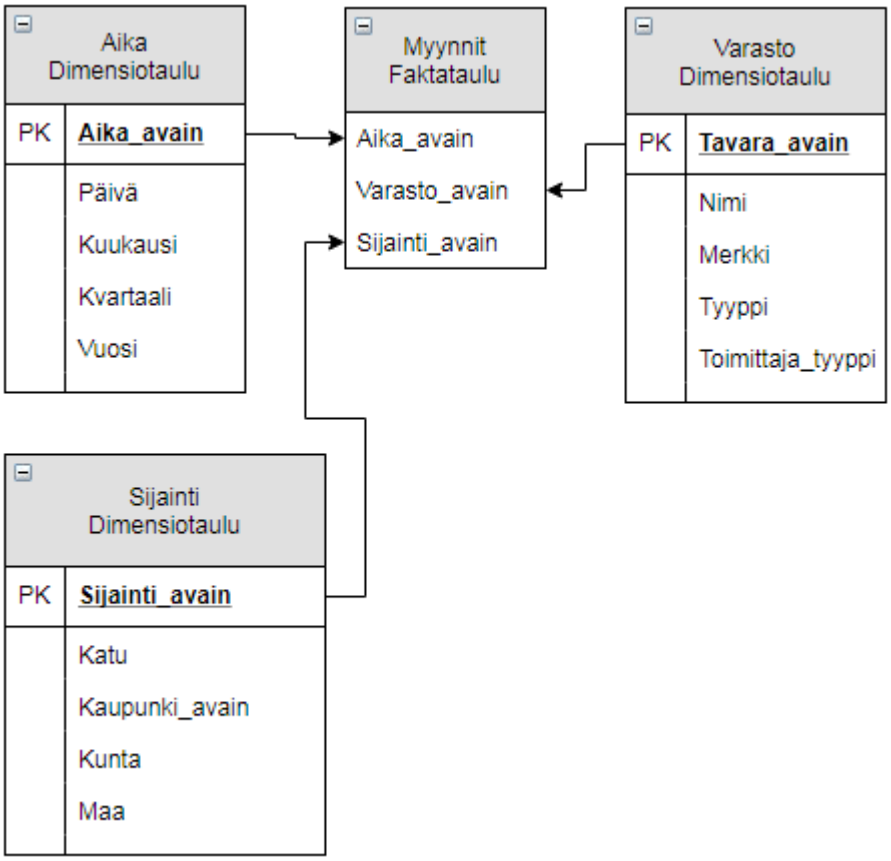


KUVIO 1. Keskitetyn tietovaraston arkkitehtuuri (Moore 2002)

Datan hankinta koostuu pääosin ohjelmista jotka hakevat, muokkaavat ja lataavat dataa tietovarastoon, kuten ETL-ohjelmat ja muut back end-ohjelmat. Liiketoimintatiedon analyysia ja päätöksentekoa tukevat ohjelmistot ovat ohjelmistoja joiden avulla analyytikot ja käyttäjät voivat analysoida ja visualisoida tietovarastossa olevaa dataa. (Vercellis, 2009, 51-52.)

3.1.1 Tietovaraston suunnittelu

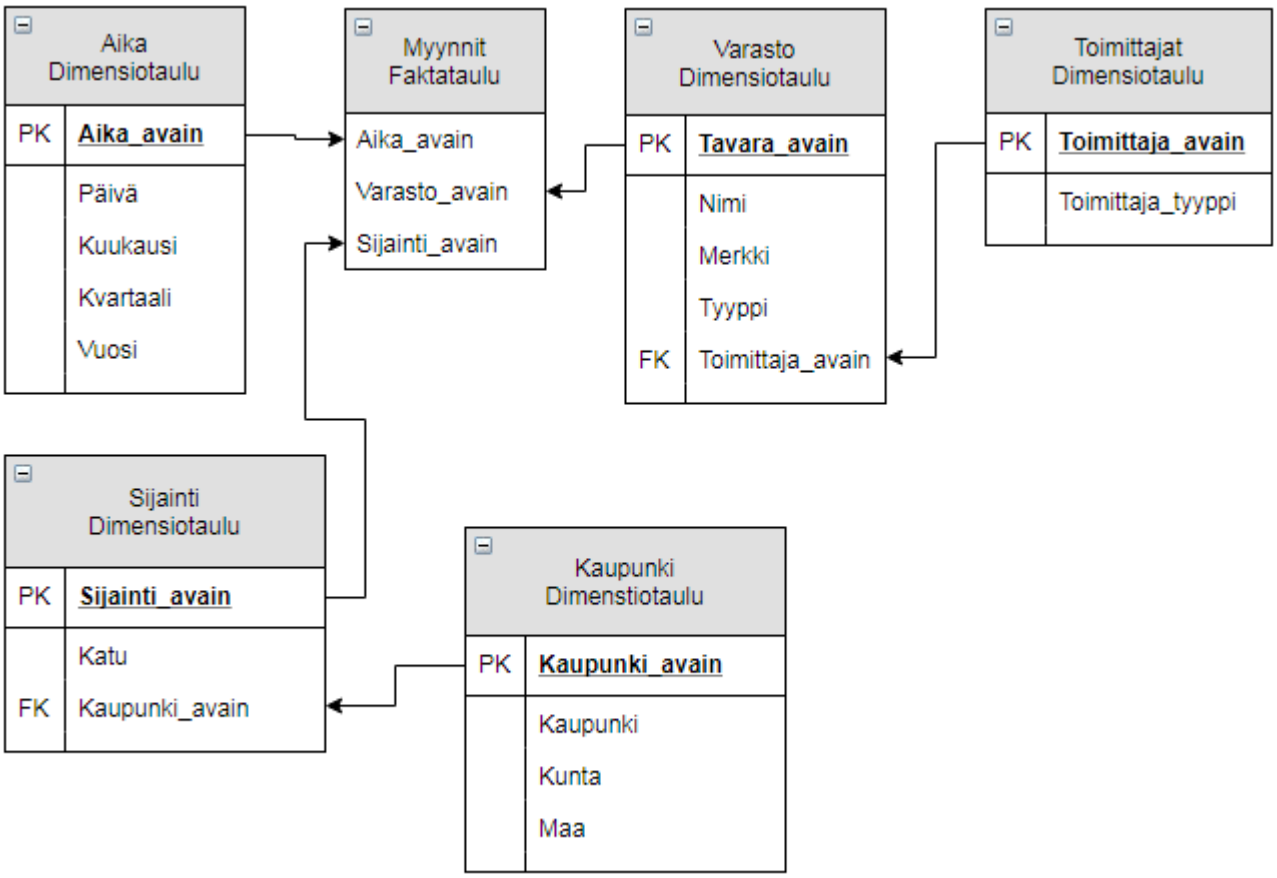
Tietovarastossa olevan datan havainnollistamiseen ja suunnitteluun käytetään yleensä tähti- tai lumihiutalemallia. Tähtimalli sopii hyvin numeerisen tai moniulotteisen datan mallintamiseen. Tähtimallin yleisin käyttökohde on paikallisvarastot, ja sen avulla simuloidaan moniulotteisuutta käyttäen relaatiotietokantaa. Tämän mallin tavoite on yksinkertaistaa kyselyiden ja raporttien luomista, tehdä datan ja taulujen lisäys helpoksi ja samalla tarjota hyvä suorituskyky. (Hovi, Ylinen & Kosinen 2001, 94.)



KUVIO 2. Tähtimalli (Tutorialspoint)

Tähtimallissa kyselyä tehtäessä dimensiotaulut yhdistetään faktatauluihin liitoksilla. Dimensiotaulut ovat tekstimuodossa olevia kuvauksia, jossa on yleensä useita eri kenttiä, joita kutsutaan attribuuteiksi tai ominaisuuksiksi. Näitä kenttiä käytetään yleensä hakuehtoina tai summausten ryh-

mittelykenttinä. Faktataulut ovat enemmänkin tapahtumatyyppisiä ja niiden tiedot ovat numeroarvoja, jotka yleensä summataan yhteen tai niistä lasketaan keskiarvo. Faktataulun perusavain on yleensä yhdistelmä dimensiotaulujen perusavaimia. (Hovi ym. 2001, 95–96.)



KUVIO 3. Lumihiutalemalli (Tutorialspoint)

Lumihiutalemalli on muokkaus tähtimallista, jossa tähtimallin ulottuvuudet on yksinkertaistettu muodostamalla lisää tauluja ja jakamalla data niihin. Monet pitävät puhdasta tähtimallia parempana kuin lumihiutalemallia, koska tähtimalli takaa paremman suorituskyvyn ja on yksinkertaisempi, mutta lumihiutalemallissa on joustavammat liitokset. (Hovi ym. 2001, 100.)

Tietovaraston rakenteen suunnittelussa voidaan toteuttaa kolmella eri ajattelutavalla: ”top-down” eli suomeksi ylhäältä alas, ”bottom-up” eli alhaalta ylös tai ”mixed” eli kahden edeltävän tavan

sekoitus. Top down -ajattelutavassa tarkastellaan tietovarastoa kokonaisuutena, jonka seurauksena se on järjestelmällisempi. Tämän ajattelutavan heikko puoli on pidempi kehitysaika, koska koko tietovarastoa suunnitellaan samanaikaisesti. Bottom up on enemmänkin prototyyppinen lähtökanta tietovaraston kehittämiseen, jossa luodaan prototyyppi tietovarastossa, jonka jälkeen sitä laajennetaan askeleittain. Projekti valmistuu nopeammin ja antaa parempia tuloksia, mutta tietovaraston kokonaiskuva jää kokonaan pois. Mixed-ajattelutavassa tietovarasto nähdään kokonaisuutena, mutta toteutetaan prototyyppisesti. Näistä kolmesta ajattelutavasta Mixed-ajattelutapaa pidetään parhaana, koska se antaa mahdollisuuden tehdä pieniä hallittuja muutoksia järjestelmään samalla ottaen huomioon koko järjestelmän. (Vercellis 2009, 52–53.)

.

3.1.2 Keskitetty tietovarasto

Keskitetyn tietovaraston idea on koota ja integroida monen liiketoiminta-alueen data yhteen isoon tietokantaan. Keskitetystä tietokannasta voi sitten muodostaa paikallisvarastoja raportointia ja kyselyitä varten. Puhtaimmillaan keskitettyyn tietovarastoon ei tehdä kyselyitä lainkaan, jonka seurauksena tietovarasto toimii operatiivisena tietojen yhdenmukaistamispaikkana ja suurien tietomäärien varastona. (Hovi, Ylinen & Koistinen, 2001, 67.)

3.1.3 Paikallisvarasto

Tietovarasto yhdistää koko yrityksen tietokannat, kun taas paikallisvarasto on pienempi ja keskittyy vain tiettyyn aihealueeseen tai osastoon. Paikallisvarasto voi olla joko riippuvainen tai itsenäinen. Riippuvainen paikallisvarasto luodaan suoraan tietovarastossa olevasta datasta ja on riippuvainen kyseisestä tietovarastosta. Kyseisen riippuvaisuuden takia tämän tyyppisen paikallisvaraston tieto on laadukasta ja varmistettua. Itsenäinen paikallisvarasto on yrityksen tietylle osastolle suunniteltu pieni tietovarasto joka ei ole yhdistetty suurempaan tietovarastoon. (Turban ym. 2011, 53.)

Paikallisvarasto ja tietovarasto ovat teknologialtaan samoja, mutta eroavat kokoluokassa. Tämän takia useimmat yritykset tehdä useita integroituja paikallisvarastoja yhden suuren paikallisvaraston sijasta, koska paikallisvarastot ovat yksinkertaisempi ja nopeampi rakentaa. (Vercellis, 2009, 49–50.)

3.1.4 Tilannekanta

Tilannekanta on tietovaraston tyyppi, jonka dataa voidaan päivittää vapaasti, kun taas tietovarastossa olevaa dataa ei voi muokata. Tilannekanta soveltuu lyhyen kantaman päätöksenteon tukemiseen ja tietovaraston pohjaksi. Tilannekanta kerää tietoa monista lähteistä ja antaa reaaliaikaisen ja integroidun näkymän nykyisestä ”haihtuvasta” datasta. (Turban ym. 2011, 53–54.)

Tilannekanta muistuttaa tietovarastokantaa, koska molemmat ovat integroituja ja data ladataan useasta eri perusjärjestelmästä, mutta tilannekantaan ei talleteta historiaa eikä se ole yleensä kyselyiden ja raporttien kohteena. Tilannekannasta on varsin helppo ladata tiedot tietovarastoon, varsinkin jos molemmat kannat ovat SQL-pohjaisia. (Hovi ym. 2001, 62.)

3.1.5 Relaatiotietokanta tietovarastona

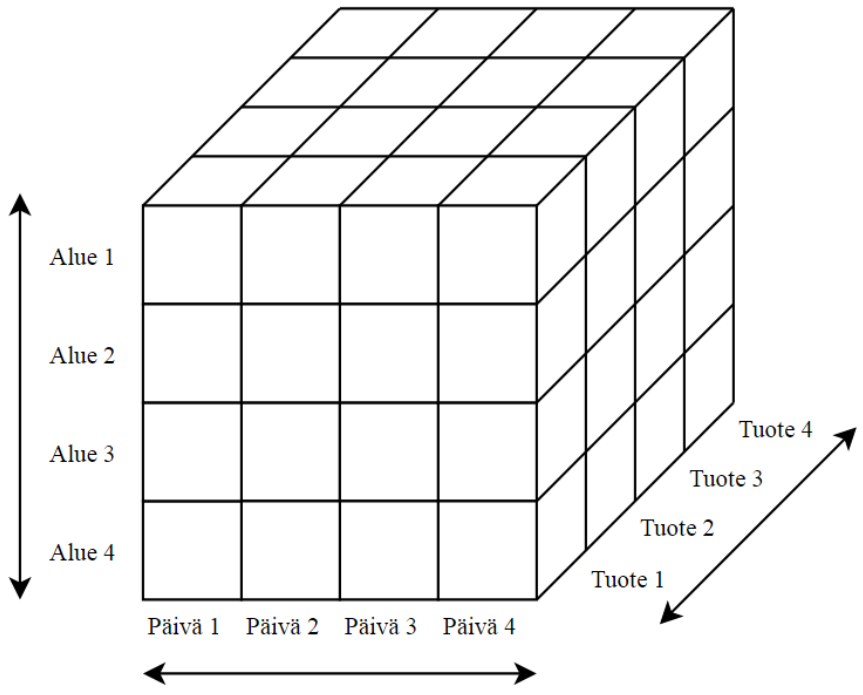
Relaatiotietokantoja käytetään yleensä operatiivisissa tietokannoissa, kuten tilijärjestelmissä ja varausjärjestelmissä, koska relaatiotietokantojen toimittajat ovat keskittyneet kehittämään tapahtumakäsittelyn suorituskykyä. Relaatiotietokantaa voidaan käyttää myös tietovarastona, vaikka siinä on tietovaraston kannalta turhia osia kuten erilaiset lukitukset, tapahtumaloki ja tapahtumien hallinta. Tietovarastot eivät tarvitse lukituksia, koska tietovarastokanta on suunniteltu vain lukua varten ja lataus tapahtuu yleensä kannan ollessa pois kyselykäytöstä.

Relaatiotietokannan käyttö tietovarastona on kuitenkin käytännöllistä, koska se antaa alustan hyvin suurille tietomassoille, siinä käytetty teknologia on toimivaa ja yleistä ja avoin SQL-rajapinta

takaa kehittyvän ympäristön ja hyvät edellytykset datan yhdistelyyn, jalostukseen ja summaamiseen. Huonot puolet ovat, ettei relaatiotietokanta tarjoa tukea moniulotteiselle käsittelylle, hakujen suorituskyky saattaa kärsiä ja siinä on turhia tapahtumankäsittelyominaisuuksia. (Hovi ym. 2011, 56–58.)

3.2 OLAP

OLAP (Online analytical processing) tai MOLAP (Multi-dimensional Online Analytical Processing) eli moniulotteiset kannat ovat niin sanottuja ”kuutiokantoja”, jotka on luotu moniulotteisen tiedon tehokkaaseen käyttöön. Tämän ominaisuutensa ansiosta OLAP-kannoissa on hyvät käsittelymahdollisuudet hierarkioille ja erilaisille luokituksille. Käyttöliittymältään OLAP-työkalut muistuttavat yleisiä taulukkolaskentaohjelmia. OLAP-kantojen suurin heikkous on rajoitukset ladattavan datan määrälle ja käyttäjien määrälle. (Hovi ym. 2001, 59.)



KUVIO 4. Esimerkki OLAP-kuutiosta (Chen 2012)

Moniulotteisten kantojen tarkastelua voidaan kutsua termillä ”porautuminen”. Termillä tarkoitetaan eri moniulotteisen kannan tasoille porautumista, kuten esimerkiksi käyttäjä voi porautua aluetasolle, sieltä piiritasolle ja sitten liiketasolle. Mitä alemmaksi käyttäjä porautuu, sitä tarkempia tietoja on saatavilla. Porautumisen vastakohta on karkeistaminen, jossa käyttäjä liikkuu tasolta ylöspäin. Tarkkailtavat ulottuvuudet vähenevät ja käyttäjälle aukenee isompi näkemys kannasta. (Hovi ym. 2011, 55.)

OLAP-kantoihin tieto ladataan usein tuotteiden omilla latausohjelmilla. Kantaan ei voi ladata kaikkia tauluja, joten yleensä tehdään useita kuutioita joka taas vaikeuttavat latausvaihdetta. Latauksen yhteydessä ohjelma laskee moniulotteisen kuution solmukohtiin valmiiksi summia, pakkaa tietoa ja indeksoi tietoja. Tämän käsittelyn ideana on tehdä kyselyistä ja porautumisesta mahdollisimman nopeaa. (Hovi ym. 2011, 59.)

OLAP-kannoista on tehty myös erilaisia yhdistelmiä muiden tietokantatekniikoiden kanssa, jotka kulkevat nimellä ROLAP ja HOLAP. ROLAP eli Relational OLAP on suoraan relaatiokantaan perustuva malli, jonka kantana on relaatiokanta mutta käyttäjälle tarjotaan moniulotteista käyttöliittymää. OLAP-käyttöliittymä yhdistetään relaatiokantaan väliohjelmiston avulla eli välikerroksen kautta relaatiokantaan syntyy joukko SQL-kyselyitä. ROLAP mahdollistaa suurien tietomassojen käsittelyn ja pystyy palvelemaan monta käyttäjää samanaikaisesti, mutta ROLAP-kannasta tiedot on luettava liitoksineen, joka johtaa hitaaseen vasteaikaan. Tämän ongelman selättämiseksi on ROLAP-tuotteet luovat tilapäisiä kuutioita muistiin, mikä nopeuttaa jatkokyselyitä. (Hovi ym. 2001, 60–61.)

HOLAP eli Hybrid OLAP käyttää niin OLAP- ja ROLAP-tekniikoita. HOLAP-kannan kuutiokantaan ladataan karkeammalla summatasolle olevat tiedot ja alimman tason tarkemmat tiedot tallennetaan relaatiokannan tauluun. Käyttäjälle tarjotaan moniulotteinen käyttöliittymä kannasta ja kun käyttäjä saapuu alimmalle kerrokselle, syntyy SQL-kysely, joka ohjautuu relaatiokannan tauluun. Tämä kysely lukee vain minimaalisen osan relaatiokannan taulusta, koska käyttäjä on jo porautuessaan kantaan rajoittanut hakuaan. (Hovi ym. 2011, 61.)

3.3 ETL

ETL eli Extract, Transformation, Load (suom. Yhdistele, muunna, lataa) on tietovaraston tietokäsittelyn ja liiketoimintatiedon hallinnan keskeisin vaihe, joka kattaa tiedon tuomisen yhdestä tai useammasta tietokannasta, tiedon muuntamisen tietovarastoon sopivaan muotoon ja tiedon lataamisen tietovarastoon. (Turban ym. 2011, 67.)

Datan yhdistely on ensimmäinen askel, jonka aikana data ladataan eri lähteistä. Tämä prosessi voidaan jakaa edeltävään lataukseen ja jälkilataukseen. Edeltävässä latauksessa kaikki saatava data ladataan tyhjään tietovarastoon ja jälkilatauksessa tietovarastoa päivitetään uuden tiedon tullessa saataville. Ladattava data valitaan tietovaraston tyypin ja analyytikoiden tarpeiden mukaan. (Vercellis 2009, 53.)

Datan siivous ja muuntaminen tulee yhdistelyn jälkeen. Muuntamisen tarkoitus on parantaa lähteistä saadun datan laatua korjaamalla epätarkkuuksia, ristiriitoja ja puuttuvia arvoja, kuten puuttuva data ja duplikaatit. Siivouksen aikana datasta korjataan useimmat toistuvat virheet käyttäen ennalta luotuja sääntöjä. Monessa tapauksessa käytetään luotuja sanakirjoja, jotka korvaavat väärät termit oikeilla riippuen termien välisistä yhtäläisyyksistä. Muunnoksen aikana tapahtuu myös lisämuunnoksia, kuten datan normalisointia, integrointia, eheyttämistä ja summatauluihin yhdistämistä. (Vercellis 2009, 53.)

Muuntamisen jälkeen tiedot ladataan tietovarastoon ja uudet tiedot lisätään vanhojen perään. Hyväksi todettu tekniikka latauksen nopeuttamiseksi on luoda erikseen työkanta, joka on SQL-kanta, minne tiedot luetaan ensiksi ja tehdään tarvittavat toimenpiteet kuten jalostus ja summaus. (Hovi ym. 2011, 82.)

3.4 Tiedon laatu

Tiedon integroiminen koostuu kolmesta prosessista, joiden tarkoitus on antaa analyysityökaluille pääsy dataan. Nämä kolme prosessia ovat dataan pääseminen, datan integroiminen ja muutosten tunnistaminen. (Turban ym. 2011, 65.)

Tiedon tarkistus, säilyttäminen ja laadun parantaminen ovat jatkuva prosessi tietovaraston suunnittelun ja päivittämisen kannalta. Datan laatuun vaikuttavat tekijät ovat tarkkuus, täydellisyys, yhtäläisyys, täsmällisyys, asiaankuuluvuus, tulkittavuus ja saatavuus. Tarkkuudella tarkoitetaan arvojen, nimien ja koodimerkintöjen oikeanlaista esitystä ja niiden arvojen esittämistä sopivissa rajoissa. Data ei saisi sisältää suuria määriä puuttuvia arvoja, mutta on hyvä pitää mielessä, että tiedonlouhinnassa käytetyt tekniikat pystyvät minimoimaan puuttuvien arvojen vaikutuksen, tätä kutsutaan datan täydellisyydeksi. Datan tulee olla integroitu yhtäläiseksi eri datalähteistä eli käytetyt merkinnät ja mittausyksiköt tulevat olla samoissa yksiköissä esimerkiksi valuutta tai paino. Tietovarastossa olevan datan pitäisi olla täsmällistä. Datan olisi hyvä olla liiketoiminnan kannalta asiaankuuluvaa, jotta se voisi antaa arvokasta tietoa analyyseille ja sitä seuraaville jälkianalyyseille. Tarpeeton ja toistuva data vähentää yhtäläisyyttä ja vie tallennustilaa, mutta dataa voidaan kopioida, jos vasteaika on huono etenkin monimutkaisissa kyseleissä. Datan tulee olla helposti saatavilla, ymmärrettävää ja oikein tulkittua analyytikon toimesta. (Vercellis 2009, 50–51.)

4 TIEDONLOUHINTA

Tiedonlouhinta on termi, jolla tarkoitetaan yhtäläisyyksien ja korrelaatioiden löytämistä isoista datamääristä. Teknisesti tiedonlouhinta on prosessi joka käyttää tilastollisia, matemaattisia ja koneoppimista hyödyllisen tiedon ja toimintamallien poimimiseen ja tunnistamiseen. (Turban ym. 2011, 157.)

Tiedonlouhinta pääero tilastotieteisiin ja OLAP-analyyseihin on analyysin luonne. Tiedonlouhinta luokitellaan aktiiviseksi, koska tiedonlouhinnassa käytetyt oppimismetodit pysyvät aktiivisessa roolissa analyyseissä luomalla uusia ennusteita ja tulkintoja jotka edustavat uutta tietoa. OLAP-analyysi ja tilastotieteet ovat passiivia, koska siinä käytetyt tekniikat ja työkalut ovat enemmänkin hypoteesien todistamista oikeaksi tai vääräksi. Tilastotieteissä luodaan hypoteesit, joille sitten etsitään todisteita pienestä otosjoukosta ja OLAP-analyysissä tekijällä on jo tietty käsitys, miten he joutuvat pohjaamaan tiedon poiminnan, raportoinnin ja kuvaamisen. Etenkin isojen tietomäärien analyysin onnistumisen kannalta on tärkeää käyttää aktiivisen luonteen omaavia malleja, koska isoista tietomassoista on vaikea muodostaa uusia ja merkityksellisiä hypoteeseja. (Vercellis 2009, 81.)

Tiedonlouhinnan toiminnot voidaan jakaa kahteen eri luokkaan, ennuste ja tulkinta, riippuen siitä mikä on analyysin päämäärä. Tulkinnan tarkoitus on löytää ja tunnistaa toistuvia yhtäläisyyksiä ja esittää ne määrättyjen sääntöjen ja kriteerien avulla. Sääntöjen ja kriteerien tarkoitus on tehdä analyysin tulkinnasta mahdollisimman helppoa käyttäjälle. Luodut säännöt tulisi olla alkuperäisiä ja asiaankuuluvia, koska niiden tarkoitus on löytää laadukasta tietoa louhittavasta järjestelmästä. Esimerkiksi myyntikampanjaa suunnitteleva tai uusia markkinarakoja etsivä jälleenmyyjä voi hyötyä louhimalla yhdistettyä tietokantaa asiakkaiden ostoprofiilista ja kanta-asiakaskortin omistajista ryhmitys-tekniikan avulla. (Vercellis 2009, 79.)

Ennusteen päämäärä on arvioida minkä arvon muuttuja ottaa tulevaisuudessa tai arvioida todennäköisyys tietylle tulevaisuuden tapahtumalle. Esimerkkinä puhelinoperaattori voi arvioida millä todennäköisyydellä tietty asiakas siirtyy toisen puhelinoperaattorin asiakkaaksi. Tämä voidaan

suorittaa esimerkiksi vertaamalla asiakkaan ikää, puhelujen kestoa, sopimuksen kestoa ja toisen operaattorin liittymiin soitettuihin puheluihin. (Vercellis 2009, 79.)

4.1 Tiedon luokitus

Data voidaan tiedonlouhinnassa luokitella kahteen pääluokkaan ja neljään alaluokkaan. Pääluokat ovat ehdoton ja numerollinen data. Ehdottoman tiedon alaluokat ovat luokiteltu ja järjestetty data. Numerollisen datan alaluokat ovat välimatkallinen ja suhteellinen datan. (Turban ym. 2011, 159.)

Ehdoton data tarkoittaa dataa joka voidaan luokitella tietyn tunnuksen alle, kuten sukupuoli, ikäryhmä ja etnisyys. Tietyt kategoriseen dataan kuuluvat tunnukset, kuten ikäryhmä, voidaan luokitella numeeriseksi dataksi, jos siinä oleva data on tarkka numero (esim. 25) eikä ryhmä (esim. 24-32). Kategorisen datan toinen nimitys on diskreetti eli jatkumaton data koska se toimii rajallisessa ulottuvuudessa ja siinä esitetyt arvot eivät ole matemaattisesti tarkkoja, mutta enemmänkin symboleja. (Turban ym. 2011, 159.)

Luokiteltu data voidaan esittää eri ryhmillä tai tunnuksilla, mutteivat ole tarkkoja arvoja, esimerkiksi henkilö voi olla (1) naimisissa (2) ei naimisissa tai (3) eronnut. Luokiteltu data voidaan esittää joko kahdella eri arvolla, esimerkiksi kyllä/ei, tai useammalla eri vaihtoehdolla esimerkiksi silmien väri. (Turban ym. 2011, 159.)

Järjestetty data tarkoittaa dataa joka voidaan järjestää ja esittää tietyillä arvoilla kuten esimerkiksi koulutus (ensimmäisen asteen, toisen ja kolmannen), ikäryhmä (lapsi, nuori, aikuinen, vanhus). Jotkut tiedonlouhinta algoritmit ottavat huomioon järjestetyn datan, jotta se voi rakentaa tarkemman mallinnuksen datasta. (Turban ym. 2011, 160.)

Numerollinen data voidaan esittää pelkästään numeroilla kuten vuositulot, ikä, lasten lukumäärä, pituus. Numerollinen data voi olla joko kokonaisluku tai desimaaliluku. Numerollinen data voi olla

loputonta eli se voi sisältää esimerkiksi loputtoman määrän desimaaleja. Välimatkallinen data tarkoittaa dataa, joka esitetään välimatka-asteikoilla, kuten lämpötilan mittaus. Suhteellinen data tarkoittaa dataa, jossa on nollapiste, jonka seurauksena arvojen keskiarvo ja suhteellisuus toisiinsa voidaan ilmoittaa tarkasti. Hyvä esimerkki suhteellisesta datasta on tulojen ilmoittaminen euroina tai yleisesti luonnontieteissä ilmenevät mittausyksiköt, kuten paino ja aika. (Turban ym. 2011, 160.)

4.2 Yleisiä tekniikoita

Tiedonlouhinnassa on monia eri metodeja tiettyjen toimintojen suorittamiseen. Moni näistä metodeista, kuten luokituspuu ja liitossäännöt on johdettu tietojenkäsittelytieteestä ja niistä käytetään termiä koneoppiminen tai tiedon löytäminen tietokannoista. Suurimalla osalla tämän tyyppisistä tekniikoista on empiirinen lähtökanta, mutta on olemassa tekniikoita, jotka voidaan luokitella monimuuttujatilastoihin, kuten regressio ja Bayesialainen luokitus. (Vercellis 2009, 79.)

Tekniikat voidaan jakaa joko valvottuihin tai valvomattomiin tekniikkoihin. Kun käyttäjä haluaa mallin, joka luo ennusteita syöttö- ja tuloarvoilla, sitä kutsutaan valvotuksi oppimiseksi. Valvotussa oppimisessa malli luodaan pienestä harjoitusosiosta, joka on yleensä pieni osa louhittavasta tietojoukosta. Tämän tarkoitus on kartoittaa arvojen toiminta ja luoda malli, joka pystyy ennustamaan tuloarvon syöttöarvon perusteella. Luokitus- ja regressiotekniikat ovat esimerkkejä valvotusta oppimisesta. Valvomaton oppiminen perustuu mallin luomiseen datan rakenteen tai datan jaon mukaan, jotta datasta voitaisiin oppia lisää. Toisin kuin valvotussa oppimisessa, valvomattomassa oppimisessa ei ole yhtä tiettyä vastausta, vaan tekniikat itsenäisesti luovat ja esittävät löydetyt rakenteet. Esimerkkejä tästä ovat ryhmitys- ja liitossäännöt-tekniikat. (Browniee 2016.)

4.2.1 Naiivi Bayesin luokitin

Naiivi bayesin luokitin perustuu klassiseen todennäköisyyslaskentaan, eli esimerkiksi, jos heitetään noppaa, todennäköisyys saada haluttu luku on 1/6 eli 17%. Tämä luokitin antaa mahdollisuuden yhdistää todennäköisyyden ja ehdollisen todennäköisyyden yhteen kaavaan, jonka avulla voidaan laskea jokaisen luokituksen todennäköisyys ja valita niistä suurimman todennäköisyyden omaava. Esimerkkinä voidaan käyttää taulukkoa tavarantoimituksen aikataulusta, jossa luokka kuvastaa sitä, kuinka ajoissa kuljetus oli.

TAULUKKO 1. Tavarantoimituksen aikataulu (Bramer 2013, 22–27.)

Päivä	Kausi	Tuuli	Sade	Luokka
Arkipäivä	Kesä	Ei	Rankkaa	myöhässä
Pyhä	Talvi	Ei	Ei	Ajoissa
Arkipäivä	Syksy	Kovaa	Rankkaa	Erittäin myöhässä
Pyhä	Kevät	Pientä	Kevyttä	Ajoissa
Arkipäivä	Kevät	Kovaa	Ei	Erittäin myöhässä
Pyhä	Syksy	Pientä	Ei	Ajoissa
Pyhä	Kesä	Ei	Kevyttä	Myöhässä
Arkipäivä	Talvi	Pientä	Rankkaa	Erittäin myöhässä
Arkipäivä	Syksy	Pientä	Ei	Ajoissa
Arkipäivä	Kevät	Kovaa	Kevyttä	Myöhässä

Sen sijaan, että laskemme todennäköisyyden sille, että kuljetus on ajoissa ja kausi syksy, laskemme ehdollisen todennäköisyyden sille, että kuljetus on ajoissa, kun kausi on syksy. Laskemme

tämän summaamalla merkinnät joissa luokka=”ajoissa” ja kausi=”syksy” ja jakamalla sen ”ajoissa”-merkintöjen kokonaissummalla, eli tässä tapauksessa 2/4 (Taulukko 1). Kun jokaisen merkinnän ehdollinen todennäköisyys lasketaan tällä tavalla, sitä voidaan soveltaa ennusteen tekoon seuraavan esimerkin mukaan.

TAULUKKO 2. Ennustettava tilanne (Bramer 2013, 22–27.)

Päivä	Kausi	Tuuli	Sade	Luokka
Arkipäivä	Kesä	Ei	Rankkaa	???

Vertaamalla taulukkoon voidaan olettaa, että kuljetus tulee myöhässä. Laskemalla ehdollisen todennäköisyyden jokaiselle mahdolliselle luokalle, taulukko näyttää tämänlaiselta.

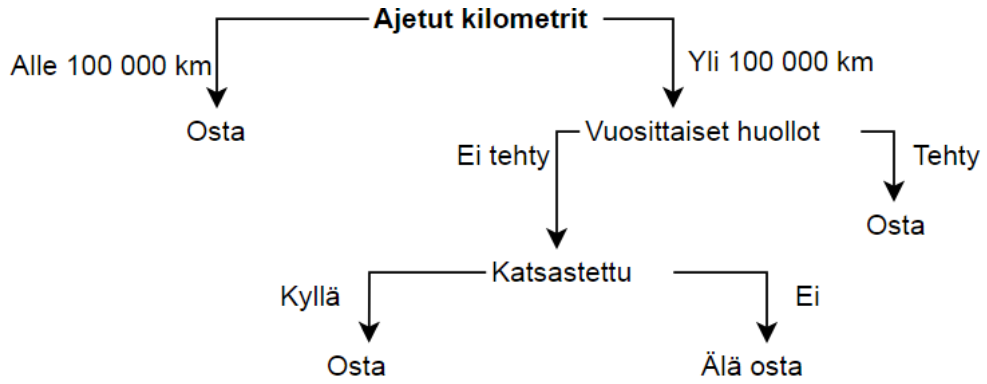
TAULUKKO 3. Lasketut todennäköisyydet ennustettavan tilanteen arvoille (Bramer 2013, 22–27).

Päivä = arkipäivä	Kausi = kesä	Tuuli = ei	Sade = rankkaa	Luokka
0.25	0	0.25	0	Ajoissa
0.66	0.66	0.66	0.33	myöhässä
1	0	0	0.66	Erittäin myöhässä

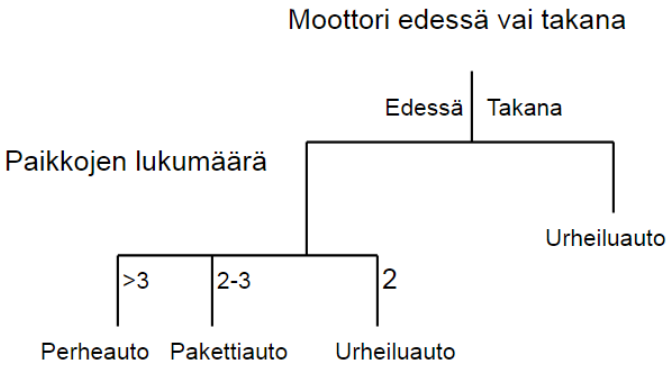
Taulukosta 3 voidaan lukea, että todennäköisin tapahtuma on kuljetuksen myöhästyminen. Ennuste ei ole kovin luotettava, sillä taulukon koko ja ominaisuusjakauma ovat erittäin pieniä, joka selittää 0% todennäköisyydet taulukossa. Tämä 0% todennäköisyyksien esiintyminen on yksi pääongelmista Bayesinn luokituksessa ja toinen pääongelma on, että tekniikka olettaa kaikkien ominaisuuksien olevan kategorista dataa. Tämä ongelma voidaan helposti korjata muuttamalla numeerinen data kategoriseksi. (Bramer 2013, 22–27.)

4.2.2 Päätös- ja luokituspuut

Päätöspuut (eng. Decision trees) perustuvat erilaisten kysymysten esittämiseen, joiden pohjalta luodaan sitten päätös. Yleensä päätöspuu aloitetaan esittämällä yksinkertainen kysymys, johon on kaksi tai useampi vastaus, jonka jälkeen puussa siirrytään alaspäin uuteen tarkentavaan kysymykseen. Päätöspuun tarkoitus on auttaa datan luokituksessa tai tunnistamisessa, jotta se voidaan kategorisoida. Tämän takia päätöspuu on hyödyllinen apuväline ennusteiden, valintakriteerien ja datan ja sen käyttökohteen valinnan apuna. Luokituspuut (eng. Classification trees) perustuvat päätöspuihin. Luokituspuiden tarkoitus on luokitella data eri luokkiin niiden eri ominaisuuksien perusteella, ja luokituspuita voidaan käyttää eri tekniikoiden tukena, kuten ryhmityksessä parantamaan suorituskykyä luokittelemalla attribuutit ensimmäiseksi luokituspuulla. (Brown 2012.)



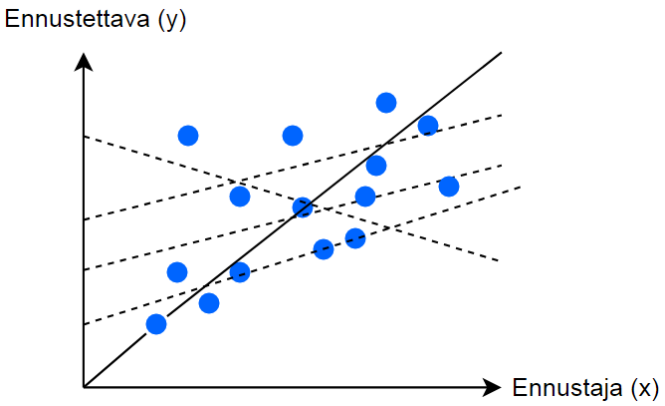
KUVIO 5. Esimerkki päätöspuusta (Brownlee 2016)



KUVIO 6. Esimerkki luokituspuusta (Frontline Systems Inc. 2017)

4.2.3 Lineaarinen regressio

Regressiotekniikoiden tarkoitus on luoda malli, joka kartoittaa muuttujan arvot suhteessa toiseen valittuun arvoon mahdollisimman pienellä virhemarginaalilla. Yksinkertaisin ja suosituin tekniikka on lineaarinen regressio, jossa on vain yksi ennuste-arvo ja yksi ennustaja-arvo. Näiden kahden muuttujan suhde voidaan kartoittaa yksinkertaisella x/y-asteikolla, jossa ennustettava on Y ja ennustaja X. Tämän jälkeen regressiomalli yrittää piirtää viivan, joka minimoi etäisyyden kaikkien merkintöjen välillä. Virhemarginaalin laskennasta yleisin käytettä tapa on neliöidä ennusteen ja oikean arvon ero. Tällä tavalla laskettuna kaukana viivasta olevat merkinnät liikkuvat viivaa kohti ja näin vähentäen virhemarginaalia. (Berson, Smith & Thearling.)



KUVIO 7. Lineaarinen regressio (Berson, Smith & Thearling)

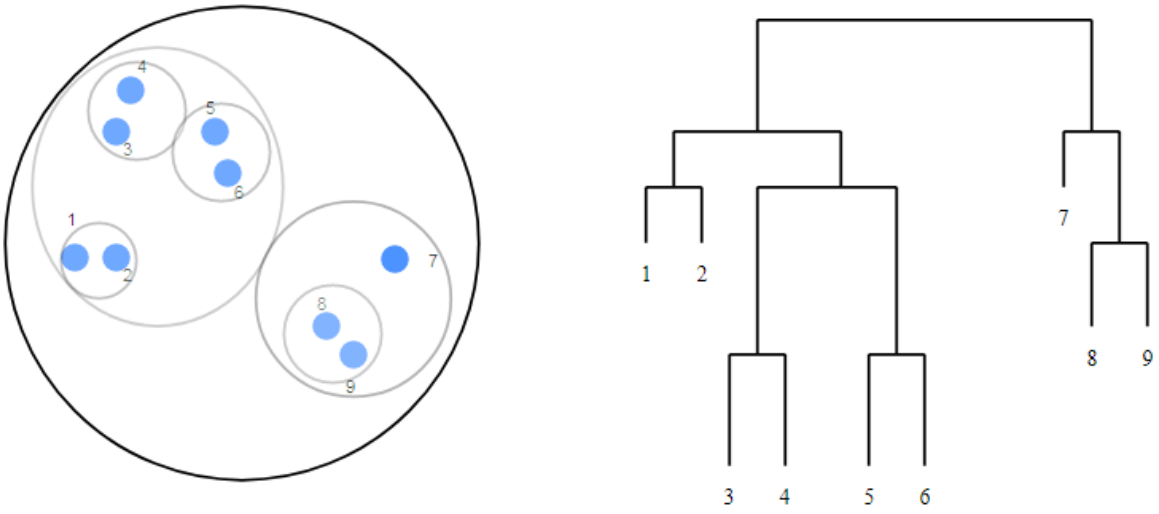
4.2.4 Liitossäännöt

Liitossäännöt (eng Association rules) on yksi tunnetuimmista tiedonlouhinnassa käytetyistä tekniikoista. Liitossääntöjen päämäärä on luoda korrelaatio kahden tai useamman, yleensä samaa tyyppiä olevan, objektin välillä toistuvien kuvioiden löytämiseksi. (Brown 2012.)

Esimerkkinä ostostapoja seuraamalla voidaan huomata, kun asiakas ostaa kahvia, hän ostaa myös maitoa. Tämän havainnon seurauksena voidaan olettaa, että kun asiakas ostaa kahvia, hän saattaa haluta myös maitoa tai jos hän ostaa maitoa, hän saattaa ostaa myös kahvia.

4.2.5 Ryhmitys

Ryhmitys on yhden tai useamman ominaisuuden käyttämistä pohjana korreloituvien objektien ryhmittämiseksi. Ryhmitys on hyödyllinen työkalu uuden datan tunnistamiseksi, koska sen avulla datan merkinnät saadaan korreloitumaan toistensa kanssa, jolloin huomataan mahdolliset yhdenmukaisuudet ja merkintöjen etäisyys toisistaan. On olemassa hierarkiatonta ja hierarkkisia ryhmitystekniikoita. (Brown 2012.)

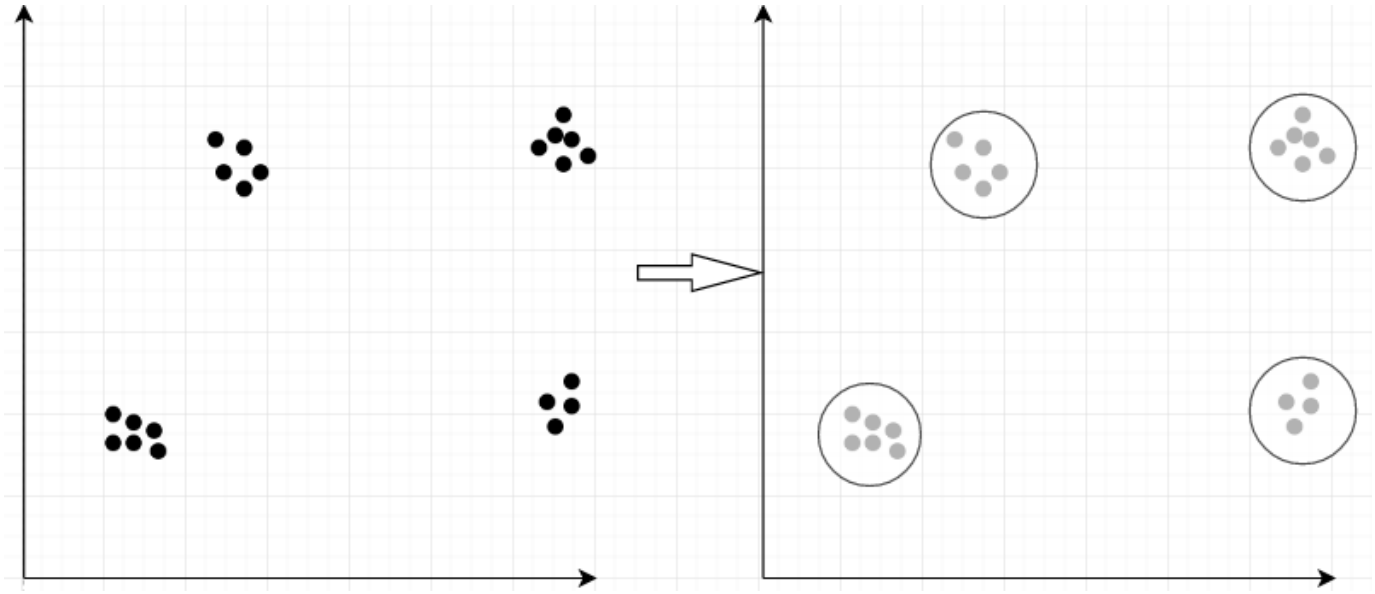


KUVIO 8. Hierarkkinen ryhmitys (Andale 2016)

Hierarkkinen tekniikka lajittelee merkinnöistä tehdyt ryhmityksen pienestä suurempaan. Tämä tehdään yhdistelemällä pienempiä ryhmiä keskenään, kunnes hierarkian yläosaan saadaan luotua yksi iso ryhmittymä, joka pitää sisällään kaikki datan merkinnät. Hierarkkisessa ryhmityksessä on olemassa kaksi eri lähestymistapaa, kasaava ja jakava. Kasaavassa tekniikassa ryhmittäminen aloitetaan hierarkian alaosasta, jossa jokaisella merkinnällä on oma henkilökohtainen ryhmittymä.

Nämä ryhmittymät liitetään pienestä suurempaan, kunnes saadaan luotua valmis hierarkia. Jakavassa tekniikassa prosessi aloitetaan hierarkian yläosan ryhmittymästä, joka pitää sisällään kaikki merkinnät. Ryhmittymiä jaetaan pienempiin osiin, kunnes se ei ole enää mahdollista. Hierarkkisen

tekniikan etu on ryhmittymien määrittäminen suoraan datasta eikä niiden lukumäärää voi määrittää ennalta kuten hierarkiattomassa tekniikassa, mutta katsottavien ryhmittymien määrää voi vähentää liikkumalla ylös tai alas hierarkiapuussa. (Berson, Smith &Thearling.)



KUVIO 9. Hierarkiaton ryhmitys (Polytechnic University of Milan)

Hierarkiaton tekniikka ryhmittää merkinnät mittaamalla etäisyyden merkintöjen välillä. Tämän tyyppiset tekniikat jaetaan yleensä sen mukaan, montako kertaa merkinnät luetaan datasta. Kerran lukevat metodit käyvät merkinnät läpi vain kerran, jonka jälkeen ne yrittävät luoda ryhmityksen datasta. Useasti lukevat eli kohdistavat metodit taas lukevat merkinnät useaan otteeseen ja siirtelevät merkintöjä ryhmästä toiseen luodakseen parempia ryhmiä. Tämä tekniikka antaa mahdollisuuden määrittää ryhmittymien määrän. Tämä saattaa tuottaa ongelmia, jos käyttäjä esimerkiksi yrittää luoda 10 ryhmää datasta, joka jakautuu 13 ryhmään. Jos ryhmitys suoritetaan näin, hierarkiaton tekniikka yrittää pakottaa ylimääräiset kolme ryhmää tehtyyn 10 ryhmään sen sijaan, että loisi datalle sopivat 13 ryhmää. (Berson, Smith & Thearling.)

4.3 Tiedonlouhinnan prosessi

Tiedonlouhinnan tavoite on luoda johtopäätöksiä vanhoista havainnoista ja sitten yleistää nämä havainnot suureen tietomäärään mahdollisemman tarkasti. Esimerkkejä tämän tyyppisistä kaavoista on lineaariset kaavat, ryhmitykseen pohjautuvat kaavat ja päättelysääntöihin pohjautuva lomakkeita.

Tiedonlouhinnan prosessi perustuu toistuvaan oppimiseen, jonka päätarkoitus on luoda uusia sääntöjä tietokannoissa olevista edellistä havainnoista ja sitten yleistää nämä havainnoit suureen tietomäärään mahdollisemman tarkasti. (Vercellis 2009, 78.)

Tiedonlouhinnan prosessin askeleet ovat tarkoituksen määrittäminen, datan keräys ja integrointi, lataus paikallisvarastoon, tutkiva analyysi, attribuuttien valinta, mallin kehittäminen ja tutkinta ja ennusteiden luominen siitä, jonka jälkeen prosessi palaa alkuun. Ensimmäinen askel on tarkoituksen määrittäminen, koska hyödyllisen tiedon löytämisen kannalta hyvin määritelty päämäärä takaa hyvän tiedon. Huonosti suunniteltu päämäärä voi johtaa ongelmiin tulevaisuuden jälkianalyysejä ajatellen. (Vercellis 2009, 84–85.)

Datan keräys ja integrointi voidaan aloittaa heti, kun analyysin tarkoitus on selvitetty. Data voi tulla monesta eri lähteestä, joten on suotavaa integroida data ja tarvittaessa lisätä siihen uusia kuvailevia muuttujia, kuten koordinaatteja tai asiakaslistalla. Ideaalisessa toteutuksessa data on jo ladattu tietovarastoihin OLAP-analyysejä ja päätöksentekoa varten, jolloin analyytikon tarvitsee vain valita attribuutit, jotka ovat sopivat tiedonlouhinnan tarkoituksen kannalta. Valmiiksi rakennetuissa tietovarastoissa on oma riskinsä, koska dataa on voitu yhdistellä tai koota yhteen muistin säästämiseksi sen verran, ettei datasta ole enää hyötyä jälkianalyyseissä. Esimerkiksi yritys on tallentanut asiakkaan kuitista kokonaissumman, muttei yksittäisiä esineitä, jonka seurauksena tulevaisuudessa on mahdoton suorittaa ostoskori-analyysiä. Tämän askeleen jälkeen voidaan siirtyä tutkivaan analyysiin. (Vercellis 2009, 87–88.)

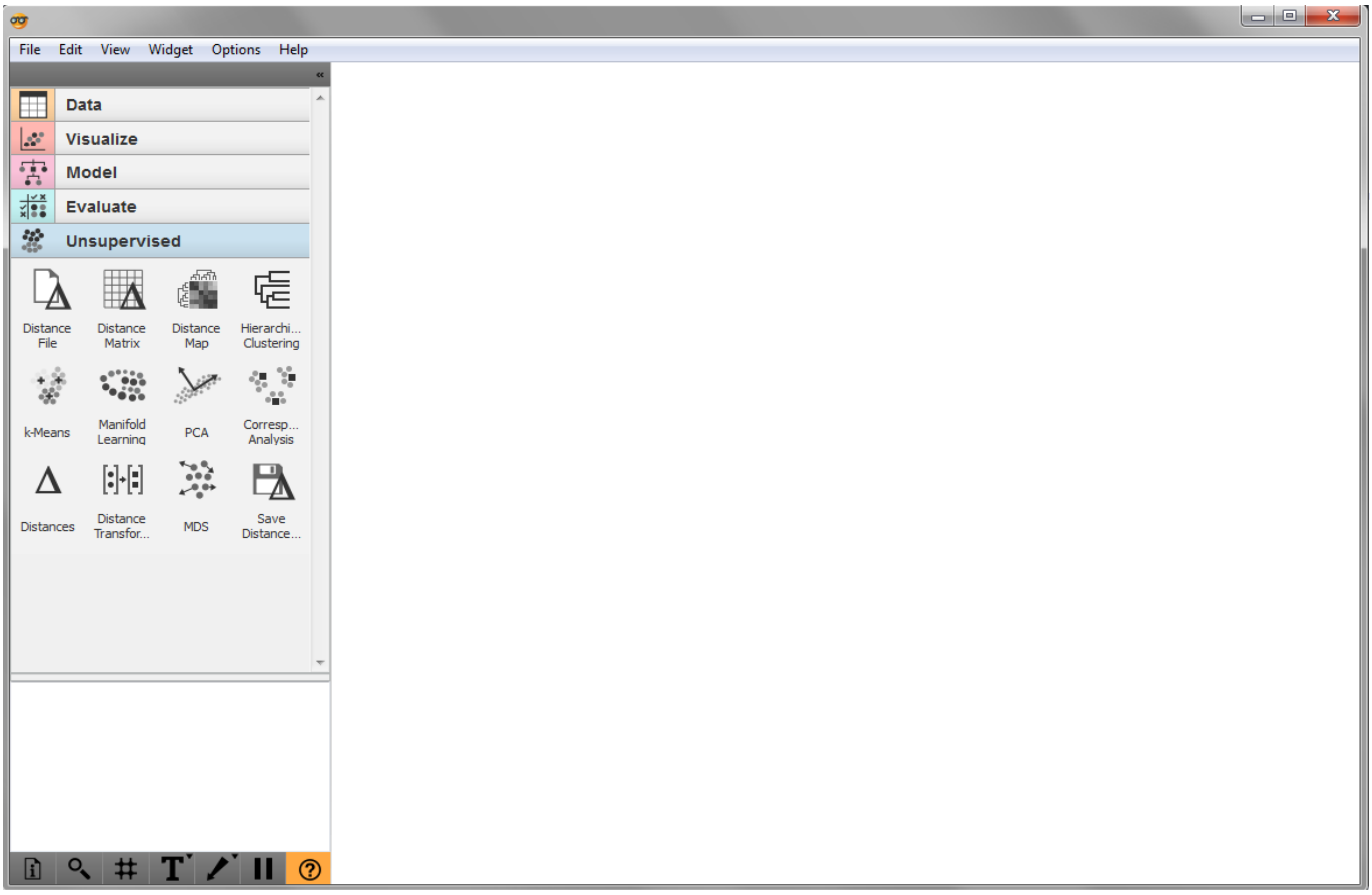
Tutkiva analyysin idea on suorittaa esianalyysi, jonka tarkoitus on perehtyä ja puhdistaa dataa. Yleensä dataa ladattaessa sitä puhdistetaan samalla eri toiminnoilla, kuten muuttujien normalisointi. Tiedonlouhinnan aikaista puhdistusta suoritetaan semanttisella eli merkityksellisellä tasolla. Tällä tarkoitetaan prosessia, jossa attribuuttien arvojen jakoa tutkitaan ja poikkeavat tai puuttuvat lukemat korostetaan, jonka jälkeen ne voidaan poistaa tarvittaessa analyysistä.

Tutkivan analyysin jälkeen suoritetaan attribuuttien valinta, jonka tarkoitus on muokata attribuutteja. Attribuuttien hyödyllisyyttä analyysin tarkoitukseen verrataan, hyödyttömät attribuutit poistetaan ja uusia tai muutettuja attribuutteja lisätään tietojoukkoon. Attribuuttien valinta ja tutkiva analyysi ovat kriittisimmät kohdat tiedonlouhinnan prosessissa, koska niillä on suuri vaikutus seuraaviin analyyseihin. (Vercellis 2009, 88.)

Attribuuttien valinta vaikuttaa myös seuraavaan askeleeseen eli mallin luomiseen ja varmentamiseen, koska attribuuttien rikastamalla tietojoukolla kehitetään tunnistus- ja ennustusmalleja. Ennen kuin malleja voidaan käyttää isoon määrään dataan, on mallia harjoitettava. Tämä suoritetaan jakamalla tietojoukko harjoittelu- ja testiosioon. Harjoitteluosio on pieni, mutta tilastollisesti merkittävä osio, jota käytetään tietyn oppimismallin löytämiseksi valittujen mallien joukosta. Testiosiolla määritetään luotujen mallien tarkkuus, jotta saadaan tunnistettua paras malli tulevaisuuden analyysejä varten. Prosessin lopuksi kehitysvaiheessa valittu malli otetaan käyttöön ja sitä sovelletaan analyysin tarkoituksen saavuttamiseen. Tulevaisuuden analyysien kannalta on järkevää sisällyttää käytetty malli myös päätöksentekoa tukeviin prosesseihin. (Vercellis 2009, 89.)

5 ORANGE

Orange on Ljubljanan yliopiston kehittämä koneoppimiseen ja tiedonlouhintaan käytetty ohjelma, joka perustuu enimmäkseen python-, C- ja C++-ohjelmointikieliin. Orange on kehitetty auttamaan piilotetun datan löytämiseen, tukemaan analyysin ymmärtämistä ja kommunikaatiota analyytikoiden ja asiantuntijoiden välillä. Orange käyttää datan visualisointiin monta eri tekniikkaa, kuten pistekaavioita korrelaatioiden havainnollistamiseen, laatikkokaaviota normaaleihin tilastollisiin toimintoihin ja histogrammia esimerkiksi normaalijakauman näyttämiseen. On myös mahdollista lisätä toimintoja lataamalla lisäosia Orangeen, joiden avulla käyttäjä voi mm. visualisoida tietoverkkoja. (Ljubljanan yliopisto.)



KUVA 1. Orange-ohjelman käyttöliittymä

Tiedonlouhinta on Orangessa toteutettu komponenttipohjaisesti. Tämä tarkoittaa, että tiedonlouhinnassa käytetyt tekniikat on ohjelmoitu eri komponentteihin, joista käytetään kutsumanimeä ”widget”, eli suomeksi sovelma. Tämän ansiosta käyttäjä tarvitsee vain löytää haluamansa tekniikoiden komponentti-ikonit ja lisätä nämä Orangen työtilaan, jossa käyttäjä voi aloittaa komponenttien yhdistelyn ja datan louhimisen. Nämä komponentit kommunikoivat keskenään ja muuttavat tuloksiaan muihin komponentteihin tehtyihin muutoksien perusteella, riippuen tuleeko data muutetusta komponentista vai ei. (Ljubljanan yliopisto.)

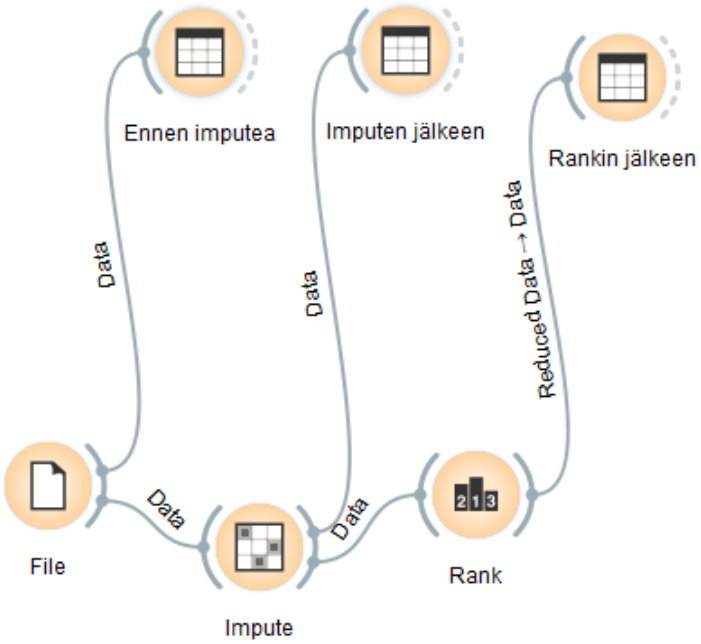
Visualisointi Orangessa on tehty interaktiiviseksi, jolla tarkoitetaan, että käyttäjä voi valita minkä tahansa merkinnän taulukosta tai kaaviosta ja Orange näyttää, mistä se data on peräisin. Kyseistä ominaisuutta voidaan myös soveltaa tutkivan datan analyysin (Explorative data analysis, EDA) suorittamiseen. Orange suoriutuu hyvin myös isojen datamäärien kanssa, koska se pystyy pisteyttämään luokat ja esittää käyttäjälle, millä luokalla datasta saadaan selkein ja kattavin. Orange mahdollistaa myös laadukkaiden raporttien luomisen nopeasti. (Ljubljanan yliopisto.)

6 TIETOJOUKON TUTKIMINEN ORANGEN AVULLA

Tässä luvussa käydään läpi väestönlaskentaan keskittyvän Adult-tietojoukon esikäsittely, sen osittainen visualisointi ja lopuksi ennusteen teko eri ominaisuuksien pohjalta. Käytetty tietojoukko on peräisin Kalifornian yliopiston koneoppimiseen keskittyvästä talletuskeskuksesta (Lichman, 2013). Tietojoukon koko on sen verran valtava, että käytän työssäni ohjelmassa jo valmiiksi olevaa pienempää osiota kyseisestä tietojoukosta, koska suuret datamäärät johtavat aina suureen määrään prosessoitavaa tietoa, joten on havainnollistamisen ja yksinkertaisuuden kannalta kannattavampi esittää pienempi osio tietojoukosta.

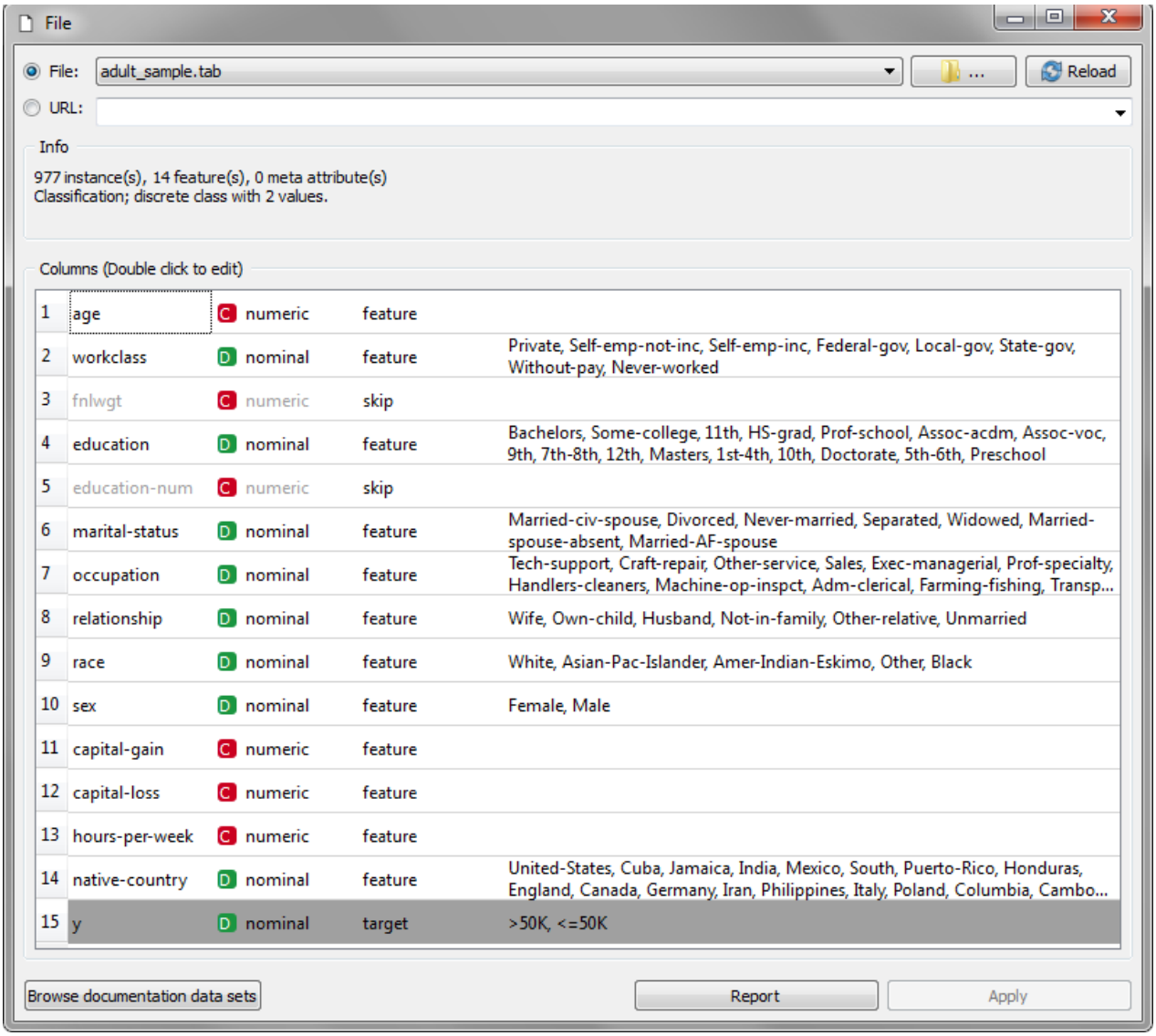
Tietojoukko koostuu 977 merkinnästä ja 15 eri ominaisuudesta, jotka kuvaavat henkilöiden taustoja ja sitä, paljonko he saavat palkkaa vuodessa. Attribuutit, niiden arvot ja kuvaukset löytyvät liitteistä (LIITE 1).

6.1 Esikäsittely ja Visualisointi



KUVA 2. Tietojoukon esikäsittely

Ensimmäiseksi tietojoukko haetaan File- eli tiedosto-toiminnon avulla (KUVA 2). Tämä toiminto myös näyttää meille tietojoukon eri sarakkeiden nimet, onko datan tyyppi nominaalinen vai diskreetti ja onko sarake ominaisuus, kohde vai metadataa eli kuvailevaa tai määrittävää tietoa tietojoukosta, kuten miten data on järjestetty ja miten voidaan hyödyntää sitä.

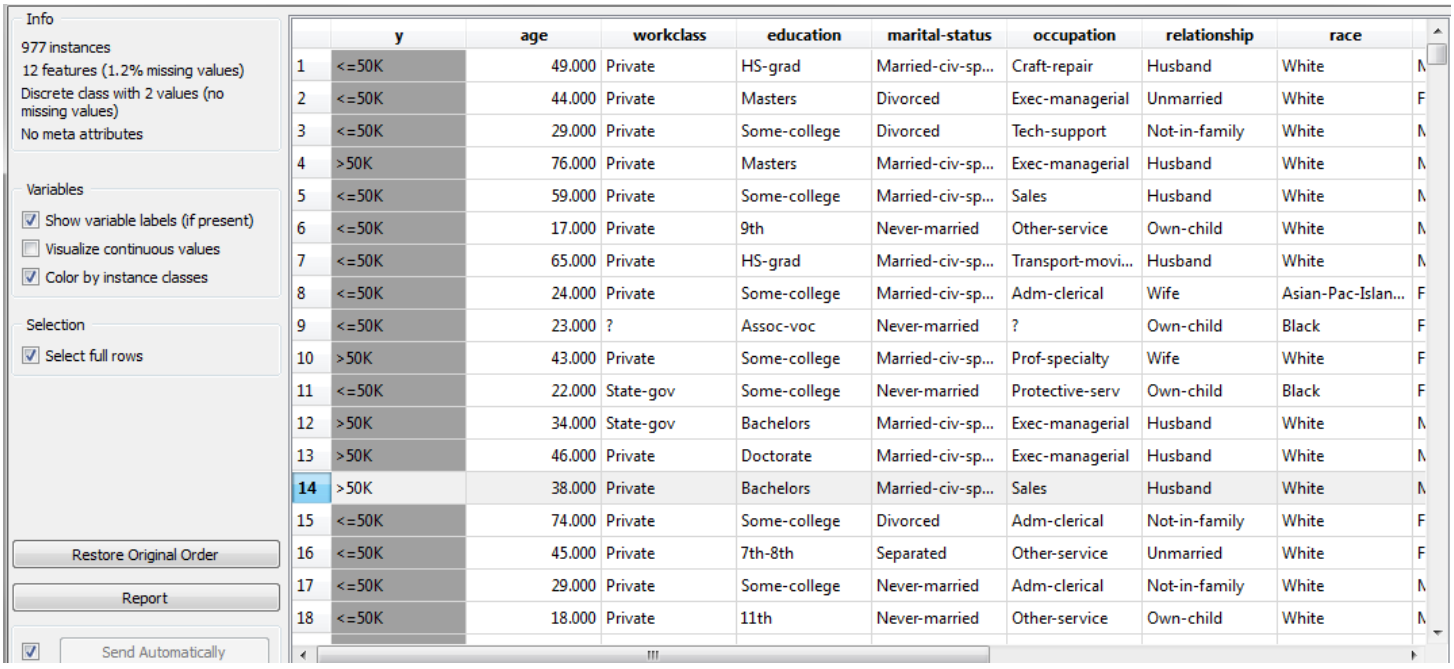


KUVA 3. Tietojoukon valinta Tiedosto-toiminnon avulla

Orange antaa mahdollisuuden muuttaa jokaisen kolumnin arvoja erikseen. Tämä tapahtuu klik-kaamalla muutettavaa arvoa ja kuten kuvassa nähdään, fnlwgt- ja educ-num-attribuuttien kohdalla

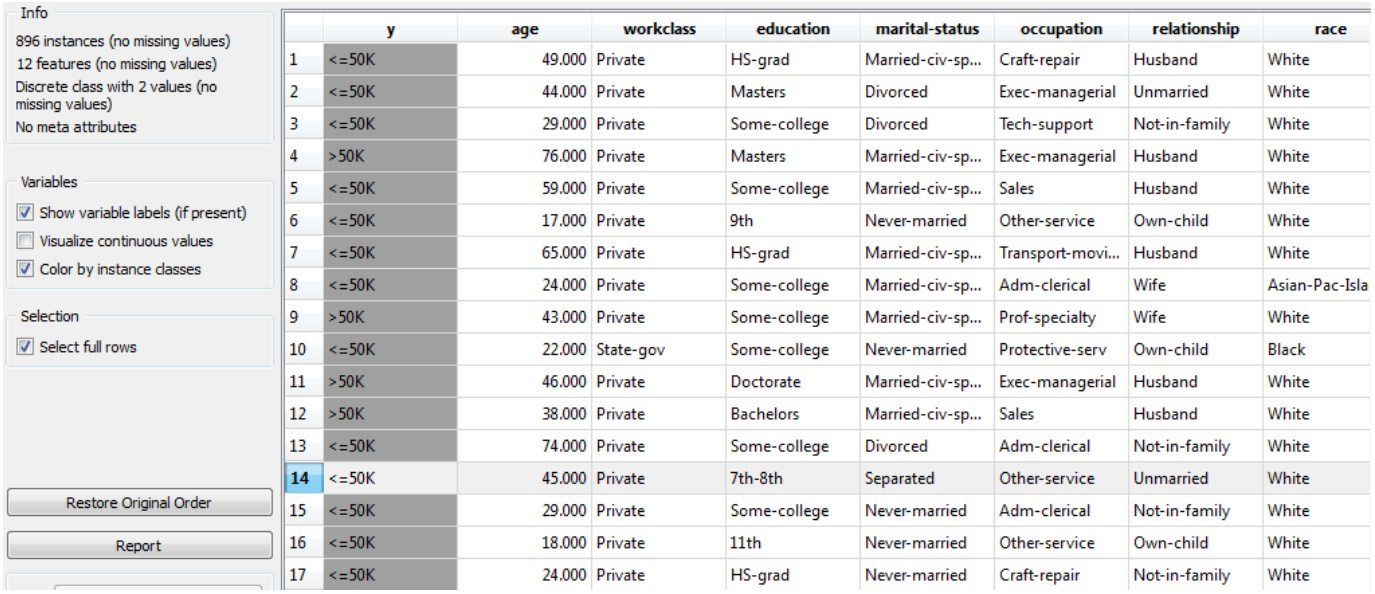
lukee ”skip” eli ohita. Tämä tarkoittaa, että nämä kaksi kolumnia poistetaan tietojoukosta kokonaan, koska nämä eivät ole analyysille tärkeitä attribuutteja ja saattavat pahimmillaan sekoittaa koko analyysin (KUVA 3).

Liittämällä file-toiminto table-toimintoon voidaan tietojoukkoa tarkastella taulukkomuodossa. Orange myös ilmoittaa, jos taulukossa on puuttuvia arvoja, joita tässä tietojoukossa on 1.2% (KUVA 4). Nämä arvoit voivat haitata analyysiä, joten on hyvä poistaa. Impute eli luku-toiminnon avulla voidaan hallita puuttuvia arvoja joko luomalla uudet arvot, poistamalla puuttuvat arvot, antamalla satunnaiset arvot tai luomalla arvot mallinnuksen pohjalta (Ljubljanan yliopisto, Widget catalog). Päätin poistaa kaikki puuttuvat arvot tietojoukosta ja kuten vertaamalla taulukkoja keskenään nähdään, tämä toiminto poisti taulukosta noin 81 merkintää (KUVA 4 ja KUVA 5).



	y	age	workclass	education	marital-status	occupation	relationship	race	
1	<=50K	49.000	Private	HS-grad	Married-civ-sp...	Craft-repair	Husband	White	M
2	<=50K	44.000	Private	Masters	Divorced	Exec-managerial	Unmarried	White	F
3	<=50K	29.000	Private	Some-college	Divorced	Tech-support	Not-in-family	White	M
4	>50K	76.000	Private	Masters	Married-civ-sp...	Exec-managerial	Husband	White	M
5	<=50K	59.000	Private	Some-college	Married-civ-sp...	Sales	Husband	White	M
6	<=50K	17.000	Private	9th	Never-married	Other-service	Own-child	White	M
7	<=50K	65.000	Private	HS-grad	Married-civ-sp...	Transport-movi...	Husband	White	M
8	<=50K	24.000	Private	Some-college	Married-civ-sp...	Adm-clerical	Wife	Asian-Pac-Islan...	F
9	<=50K	23.000	?	Assoc-voc	Never-married	?	Own-child	Black	F
10	>50K	43.000	Private	Some-college	Married-civ-sp...	Prof-specialty	Wife	White	F
11	<=50K	22.000	State-gov	Some-college	Never-married	Protective-serv	Own-child	Black	F
12	>50K	34.000	State-gov	Bachelors	Married-civ-sp...	Exec-managerial	Husband	White	M
13	>50K	46.000	Private	Doctorate	Married-civ-sp...	Exec-managerial	Husband	White	M
14	>50K	38.000	Private	Bachelors	Married-civ-sp...	Sales	Husband	White	M
15	<=50K	74.000	Private	Some-college	Divorced	Adm-clerical	Not-in-family	White	F
16	<=50K	45.000	Private	7th-8th	Separated	Other-service	Unmarried	White	F
17	<=50K	29.000	Private	Some-college	Never-married	Adm-clerical	Not-in-family	White	M
18	<=50K	18.000	Private	11th	Never-married	Other-service	Own-child	White	M

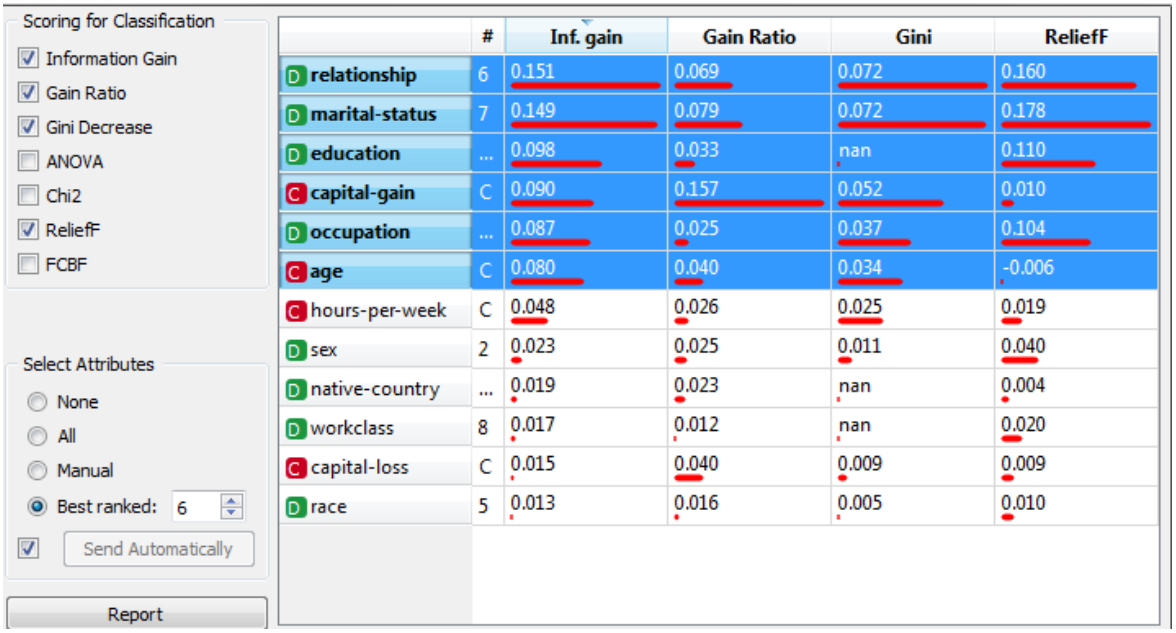
KUVA 4. Tietojoukko ennen puuttuvien arvojen poistamista



	y	age	workclass	education	marital-status	occupation	relationship	race
1	<=50K	49.000	Private	HS-grad	Married-civ-sp...	Craft-repair	Husband	White
2	<=50K	44.000	Private	Masters	Divorced	Exec-managerial	Unmarried	White
3	<=50K	29.000	Private	Some-college	Divorced	Tech-support	Not-in-family	White
4	>50K	76.000	Private	Masters	Married-civ-sp...	Exec-managerial	Husband	White
5	<=50K	59.000	Private	Some-college	Married-civ-sp...	Sales	Husband	White
6	<=50K	17.000	Private	9th	Never-married	Other-service	Own-child	White
7	<=50K	65.000	Private	HS-grad	Married-civ-sp...	Transport-movi...	Husband	White
8	<=50K	24.000	Private	Some-college	Married-civ-sp...	Adm-clerical	Wife	Asian-Pac-Isla
9	>50K	43.000	Private	Some-college	Married-civ-sp...	Prof-specialty	Wife	White
10	<=50K	22.000	State-gov	Some-college	Never-married	Protective-serv	Own-child	Black
11	>50K	46.000	Private	Doctorate	Married-civ-sp...	Exec-managerial	Husband	White
12	>50K	38.000	Private	Bachelors	Married-civ-sp...	Sales	Husband	White
13	<=50K	74.000	Private	Some-college	Divorced	Adm-clerical	Not-in-family	White
14	<=50K	45.000	Private	7th-8th	Separated	Other-service	Unmarried	White
15	<=50K	29.000	Private	Some-college	Never-married	Adm-clerical	Not-in-family	White
16	<=50K	18.000	Private	11th	Never-married	Other-service	Own-child	White
17	<=50K	24.000	Private	HS-grad	Never-married	Craft-repair	Not-in-family	White

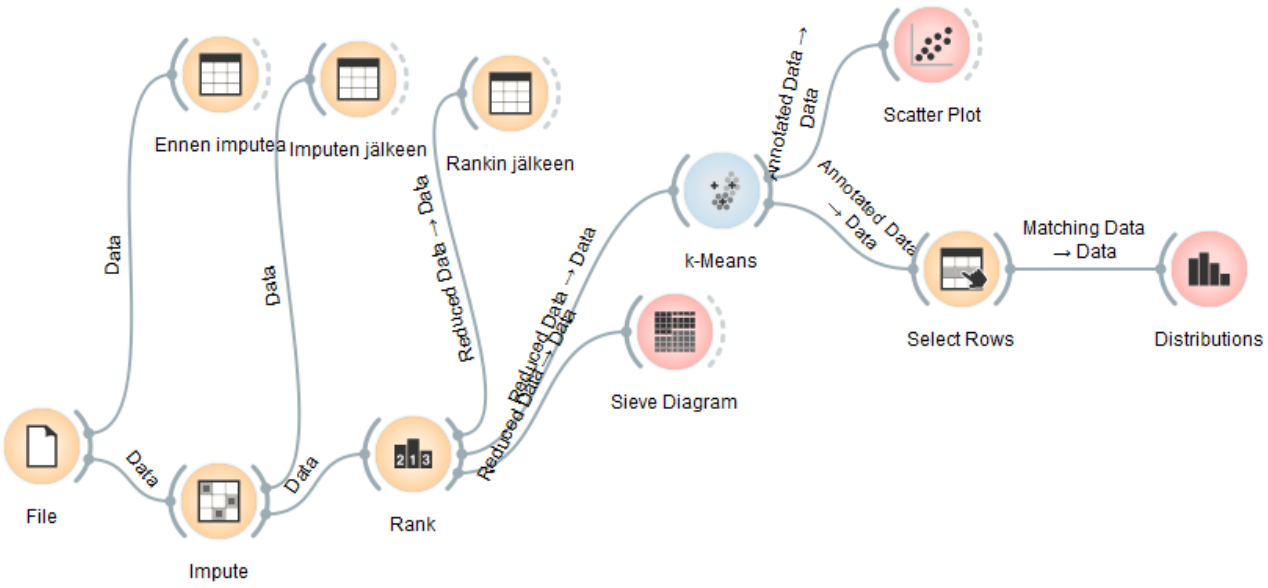
KUVA 5. Tietojoukko puuttuvien arvojen poistamisen jälkeen

Tyhjien arvojen poistamisen jälkeen päätin yksinkertaistaa analyysiä entisestään liittämällä mukaan Rank eli pisteytys-toiminnon. Tällä toiminnolla voidaan pisteyttää tietojoukossa olevat attribuutit eri muuttujien perusteella. Pisteytin attribuutit sen mukaan, että mistä attribuuteista vastaanotetaan eniten mahdollista dataa (inf. gain, information gain). Toiminnon jälkeen mahdolliset attribuutit laskivat kahdestatoista kuuteen. (KUVA 6.)



	#	Inf. gain	Gain Ratio	Gini	ReliefF
D relationship	6	0.151	0.069	0.072	0.160
D marital-status	7	0.149	0.079	0.072	0.178
D education	...	0.098	0.033	nan	0.110
C capital-gain	C	0.090	0.157	0.052	0.010
D occupation	...	0.087	0.025	0.037	0.104
C age	C	0.080	0.040	0.034	-0.006
C hours-per-week	C	0.048	0.026	0.025	0.019
D sex	2	0.023	0.025	0.011	0.040
D native-country	...	0.019	0.023	nan	0.004
D workclass	8	0.017	0.012	nan	0.020
C capital-loss	C	0.015	0.040	0.009	0.009
D race	5	0.013	0.016	0.005	0.010

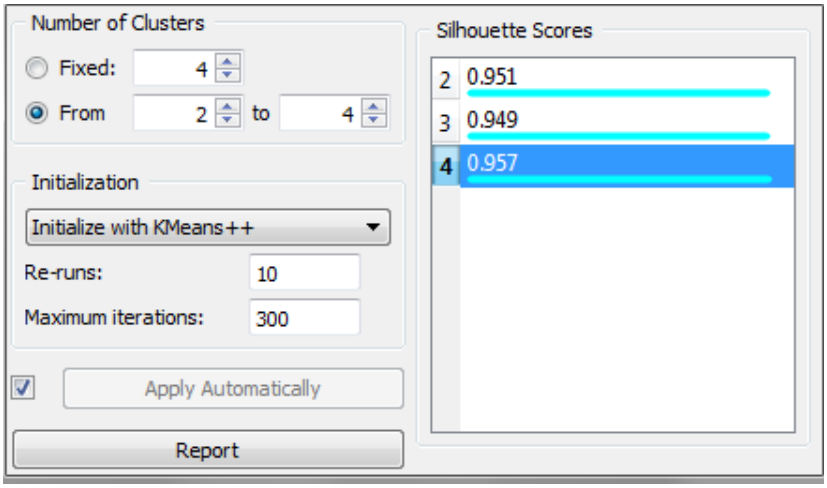
KUVA 6. Attribuuttien pisteytys



KUVA 7. Tietojoukon visualisointi

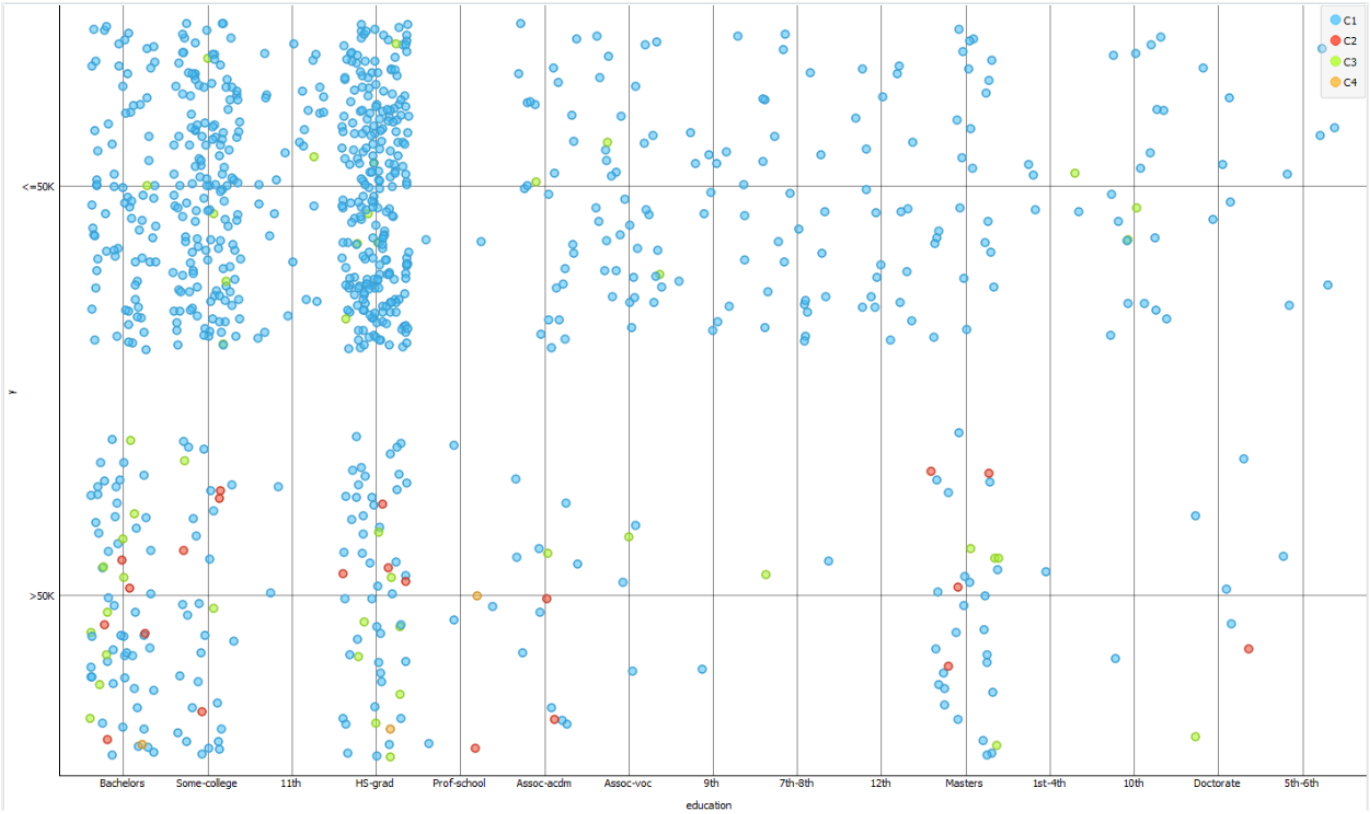
Kun datan esikäsittely on suoritettu, voidaan siirtyä visualisointiin. Merkintöjen määrän takia on hyvä käyttää toimintoja, jotka pystyvät näyttämään suuren määrän dataa siististi kerralla. Päätin tässä työssä käyttää k-means- ja sieve diagram eli seuladiagrammi-toimintoja tietojoukon visualisointiin (KUVA 7).

K-means-toiminto perustuu hierarkiattomaan ryhmitykseen, ja sitä voidaan käyttää eri visuaalisten toimintojen kanssa, kuten tässä tapauksessa käyttämääni scatter plot eli pistekaavio- ja distribution eli jakauma-toimintoa (Ljubljanan yliopisto, Widget catalog.). Ryhmitys pohjustettiin testaamalla, mikä ryhmien määrä soveltuu parhaiten tähän tietojoukkoon. Ohjelman mukaan paras tarkkuus saadaan 4 eri ryhmällä, jolloin arvioitu tarkkuus on noin 0.959 eli 96% (KUVA 8). Pistekaavio-toiminto näyttää ryhmityksen tulokset pistekaaviossa ja jakauma-toiminto näyttää ryhmityksen jakaumat. Lisäämällä jakauma-toimintoon select rows eli valitse rivit-toiminnon, voidaan tietojoukosta valita tietty arvoväli attribuuteille ja näin tutkia esimerkiksi vain ylemmän korkeakoulututkinnon suorittaneiden tietoja. (KUVA 10.)

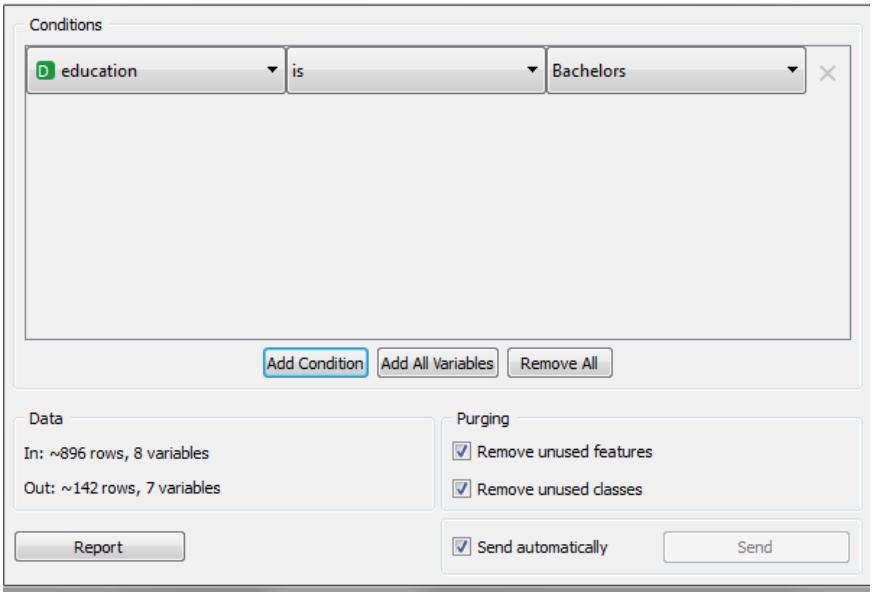


KUVA 8. Ryhmien lukumäärän pisteyttäminen

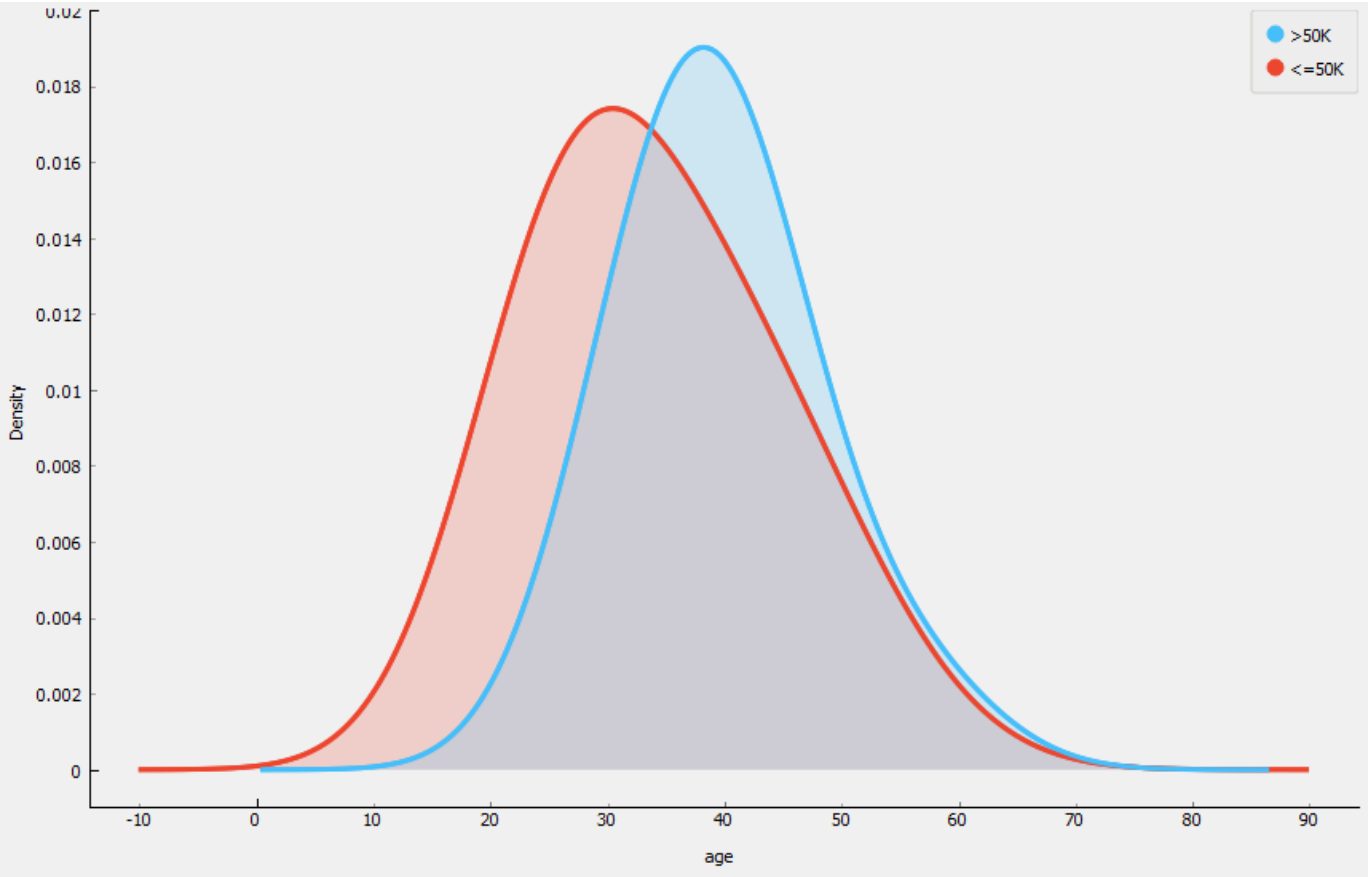
Sieve diagram-toiminto laskee ja vertaa kahden eri attribuutin esiintymistä tietojoukossa, tässä tapauksessa vuositulot ja koulutuksen taso. Orange suorittaa tämän ensiksi laskemalla odotetun esiintymän vuositulojen ja koulutuksen välillä; esimerkiksi montako merkintää löytyy, jossa vuositulot ovat yli 50k, kun koulutus on maisteri. Tämän jälkeen Orange laskee todellisen esiintymän tietojoukosta, vertaa sitä teoreettiseen esiintymään ja visualisoi sen. Visualisoinnissa erikokoiset neliöt ovat suhteellisia odotetun esiintymän kokoon, sen sisällä olevat neliöt merkitsevät todellisia esiintymiä ja väri merkitsee, onko todellinen esiintymä suurempi kuin odotettu, sininen jos on ja punainen jos ei ole. (Ljubljanan yliopisto, Widget catalog.)



KUVA 9. Ryhmityksestä saatu pistekaavio

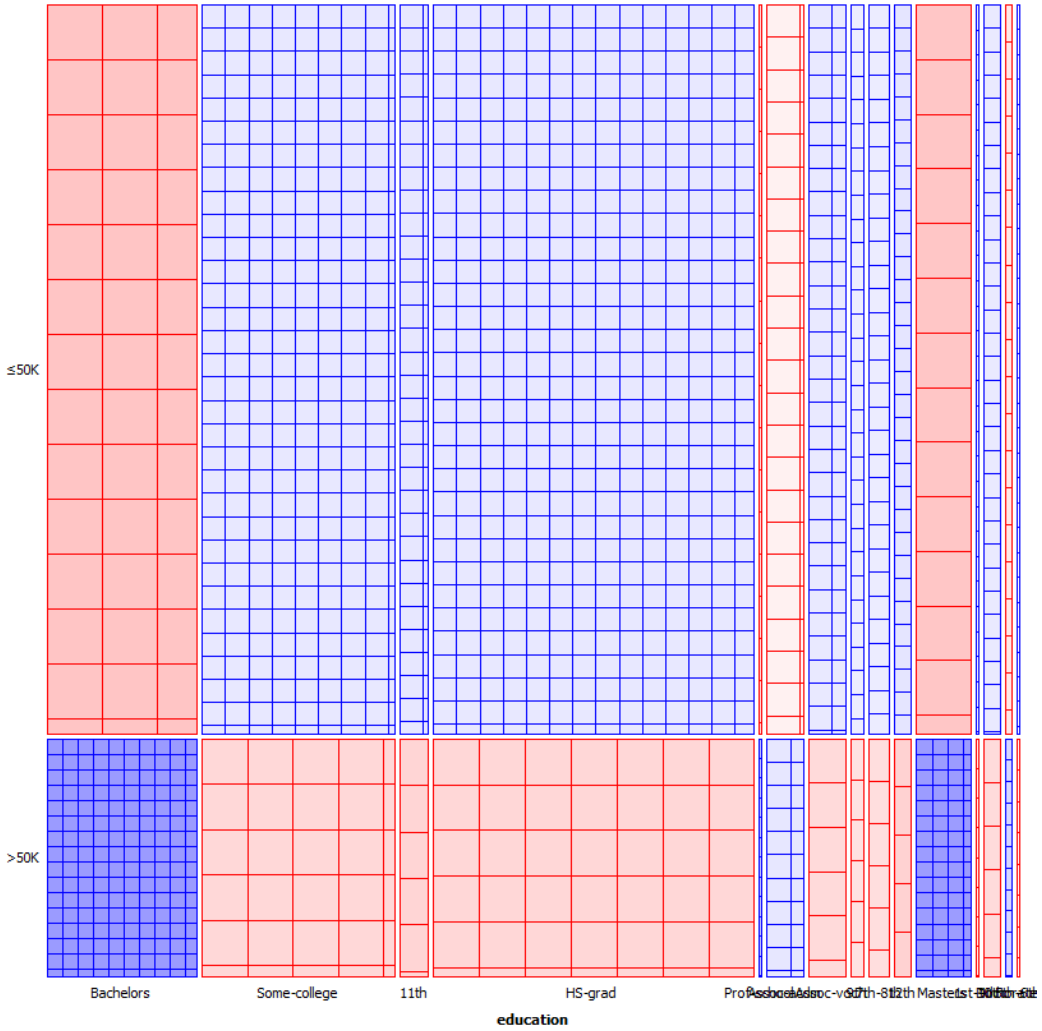


KUVA 10. Ehtojen asettaminen valitse rivit-toiminnolla



Kuva 11. Ehtojen asettamisesta johdettu ikäjakauma

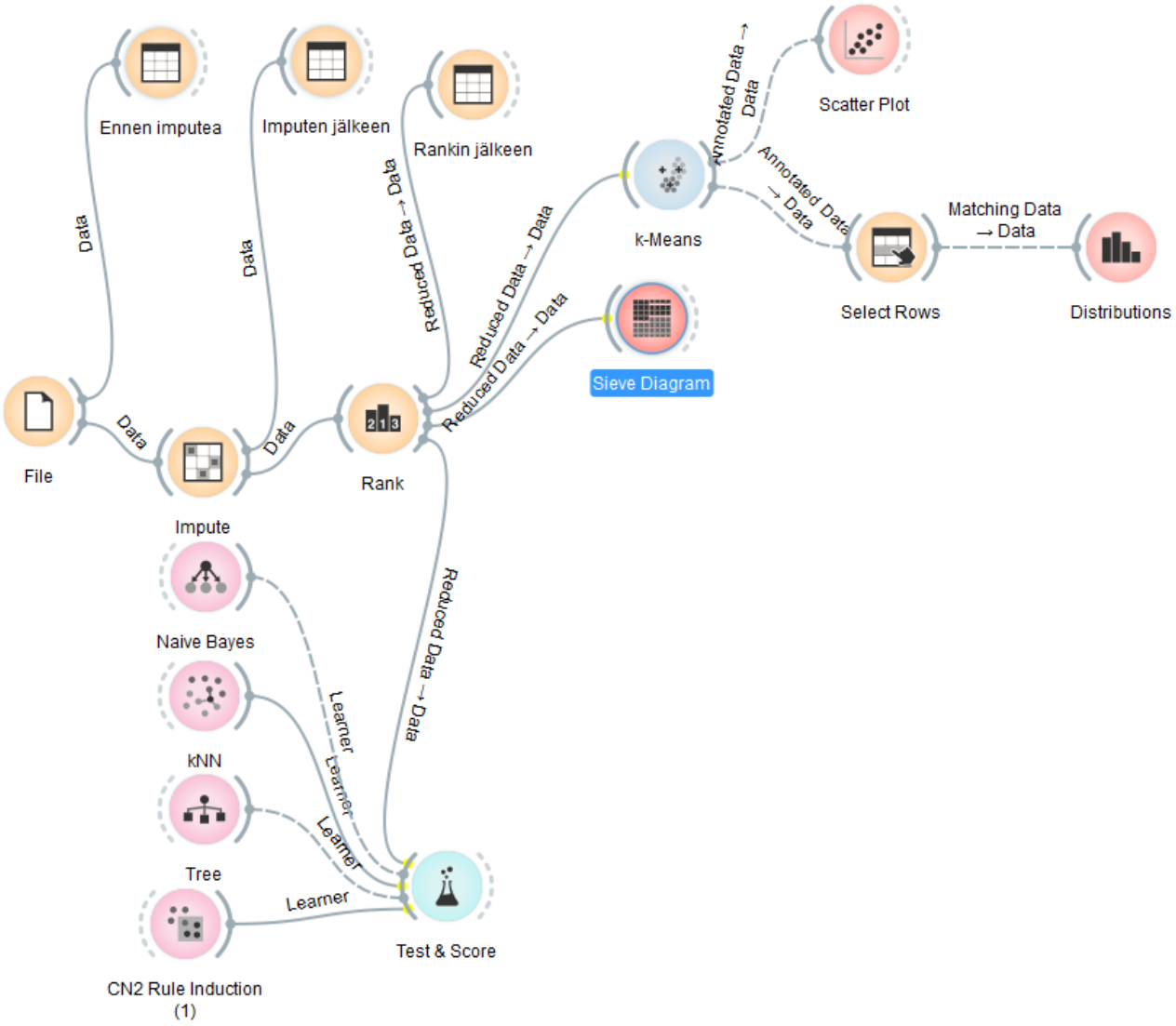
Kuten jakaumasta (KUVA 11) huomataan, yli 50 000 dollarin vuosituloihin pääsevät ovat todennäköisin iältään noin 35–40 ja alle 50 000 dollaria ansaitsevat ovat todennäköisin iältään noin 25-30. Seulasta taas huomataan, että yli 50 000 vuositulot ovat esiintymiltään korkeammat niillä, joiden koulutus on joko maisteri tai ylempi ammattikorkeakoulu ja alle 50 000 vuositulot esiintyvät eniten niillä, joiden koulutus on joko vain toiseen asteen koulutus tai ovat collegessa, mutta eivät ole vielä valmistuneet (KUVA 12).



Kuva 12. Seuladiagrammi tuloista ja koulutuksesta

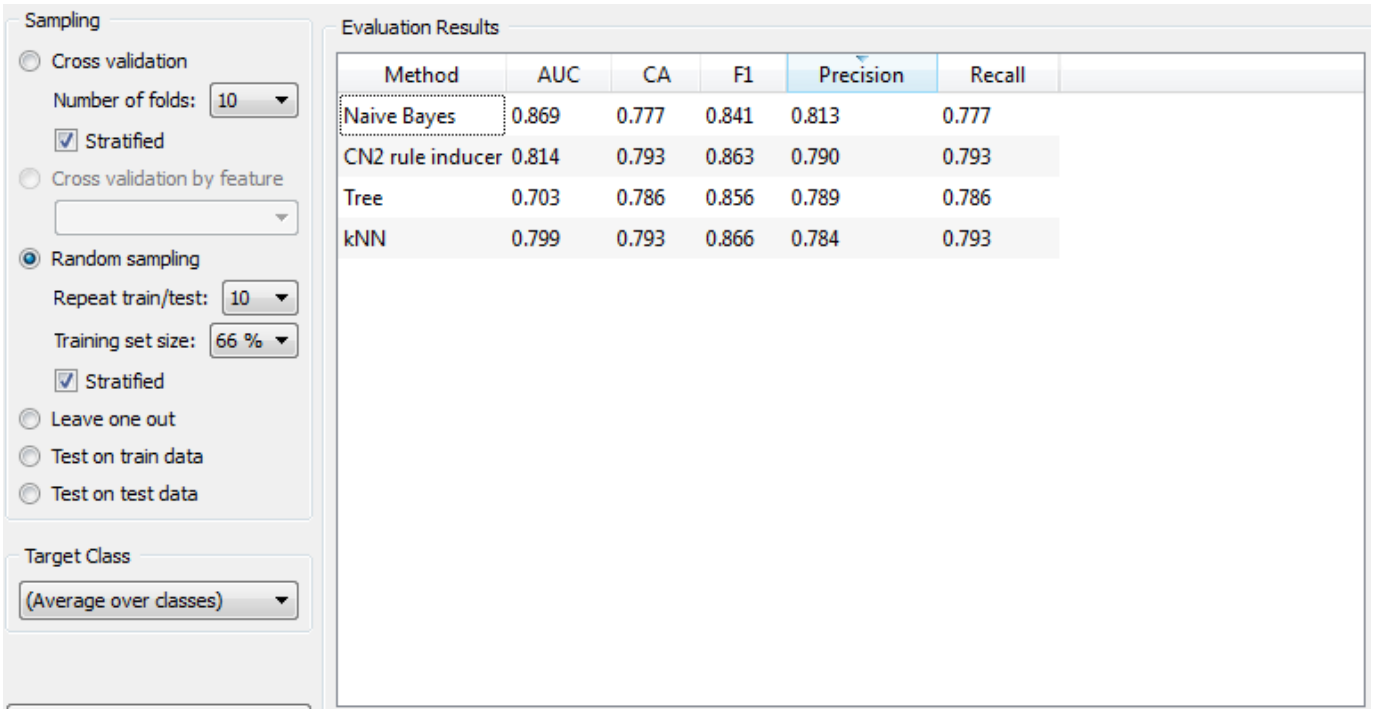
6.2 Ennusteen teko

Kun datan visualisointi on suoritettu, voidaan siirtyä ennusteen luomiseen. Ennusteen luomiseen käytetään yleensä erillistä opetuserää, joka sitten yleistetään koko dataan, mutta käytetyn tietojoukon koon takia tämä ei tässä tapauksessa ole tarpeen.



KUVA 13. Algoritmien pisteytys

Orangessa on mahdollisuus pisteyttää ja testata eri ennusteen luomiseen käytettyjä toimintoja käyttämällä Score & Test eli Pisteytä & testaa-toimintoa. Tämä toiminto näyttää kunkin ennuste-toiminnon mahdollisen tarkkuuden ennusteen luomisessa, josta voidaan valita sitten tarkimmat toiminnot. (Ljubljanan yliopisto, Widget catalog.)



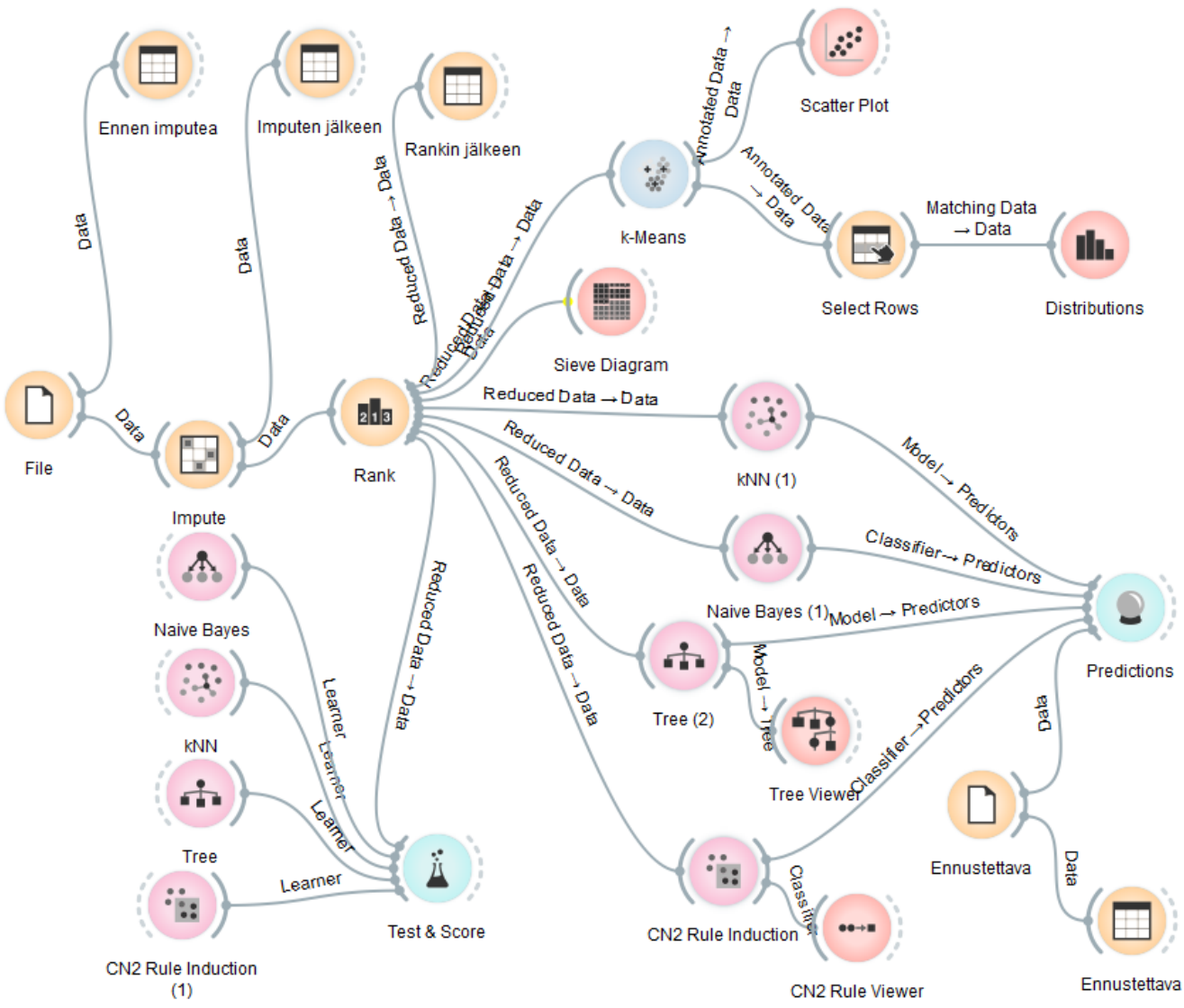
KUVA 14. Algoritmien pisteytyksen tulokset järjestetty tarkkuuden mukaan

Ennusteen tekoa varten luodaan uusi taulukko, jossa y-kolumni jätetään pois ja muut täytetään. Näin kerrotaan Orangelle, mitä attribuuttia haluamme ennustaa. Ennuste luodaan yhdistämällä Pisteytys-toiminnosta saatu data valittuihin toimintoihin, jotka sitten ennustavat vuositulot attribuuttien pohjalta.

	age	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	workclass	education	marital-status
1	34.000	Sales	Unmarried	White	Male	0.000	0.000	40.000	United-States	private	Bachelors	Divorced
2	45.000	Exec-managerial	Husband	Other	Male	1000.000	0.000	35.000	Ireland	federal-gov	Prof-school	Separated
3	20.000	Other-service	Unmarried	Black	Female	0.000	0.000	20.000	Ireland	without-pay	HS-grad	Never-married
4	23.000	Other-service	Not-in-family	Black	Male	0.000	0.000	30.000	Mexico	never-worked	Assoc-voc	Married-civ-sp...
5	60.000	Prof-specialty	Own-child	Other	Female	2300.000	500.000	40.000	United-States	state-gov	Some-college	Married-civ-sp...
6	55.000	Prof-specialty	Unmarried	Amer-Indian-Es...	Female	3000.000	1000.000	38.000	China	self-emp-inc	Masters	Widowed

KUVA 15. Taulukko ennustettavista

Taulukko viedään ohjelmaan käyttämällä Tiedosto-toimintoa, joka sitten yhdistetään valittuihin tiedonlouhinta-tekniikoihin, jonka jälkeen nämä yhdistetään ennustus-toimintoon, joka tekee ennustuksen kullakin tekniikalla ja kertoo todennäköisyydet sille, että henkilön vuositulot ovat joko yli tai alle 50 000.



KUVA 16. Ennustuksen luominen

Käytin ennustusten luomiseen ryhmitykseen kuuluvaa lähin naapuri-ryhmitystä (kNN, nearest neighbour), naiivi Bayesin luokitin-, luokituspuu- ja CN2-päättelysääntö-algoritmia. CN2-päättelysääntö-algoritmi on luokitusalgoritmi, joka luo joukon erilaisia ”jos”-ehtoja ja kertoo sitten todennäköisyydet näille ehdoille. (Ljubljanan yliopisto, Widget catalog.)

	CN2 rule inducer	kNN	Tree	Naive Bayes	age	occupation	relationship	race
1	0.20 : 0.80 → <=50K	0.00 : 1.00 → <=50K	0.25 : 0.75 → <=50K	0.23 : 0.77 → <=50K	34.000	Sales	Unmarried	White
2	0.20 : 0.80 → <=50K	0.00 : 1.00 → <=50K	0.25 : 0.75 → <=50K	0.34 : 0.66 → <=50K	45.000	Exec-managerial	Husband	Other
3	0.17 : 0.83 → <=50K	0.00 : 1.00 → <=50K	0.25 : 0.75 → <=50K	0.04 : 0.96 → <=50K	20.000	Other-service	Unmarried	Black
4	0.17 : 0.83 → <=50K	0.00 : 1.00 → <=50K	0.25 : 0.75 → <=50K	0.04 : 0.96 → <=50K	23.000	Other-service	Not-in-family	Black
5	0.20 : 0.80 → <=50K	0.00 : 1.00 → <=50K	0.25 : 0.75 → <=50K	0.25 : 0.75 → <=50K	60.000	Prof-specialty	Own-child	Other
6	0.14 : 0.86 → <=50K	0.60 : 0.40 → >50K	0.25 : 0.75 → <=50K	0.25 : 0.75 → <=50K	55.000	Prof-specialty	Unmarried	Amer-Indian-Es...

KUVA 17. Ennustuksen tulokset

Kuten ennustuksen tuloksista (KUVA 17) voidaan nähdä todennäköisyyksien eroavaisuudet eri algoritmien välillä. Esimerkiksi 0.2: 0.8 tarkoittaa, että 20% todennäköisyydellä vuositulot ovat yli 50 000 ja 80% todennäköisyydellä vuositulot ovat alle 50 000. Melkein kaikki algoritmit päätyivät samaan lopputulokseen, paitsi kNN- eli lähin naapuri-ryhmitys, joka ennusti yhden merkinnän tienaavan yli 50 000.

7 POHDINTA

Työn tarkoitus oli antaa lukijalle käsitys liiketoimintatiedosta sekä näyttää ja toteuttaa tiedonlouhinnan prosessi ja luoda ennustukset tiedonlouhintaan käytetyllä ohjelmalla. Käsitteenä liiketoimintatieto on erittäin laaja, ja ei vain rajoitu tässä työssä esitettyihin aiheisiin, mutta uskon että työni antaa ainakin alustavan käsityksen liiketoimintatietoon, jonka kautta sukellus sen syvempiin aiheisiin olisi helpompi ja selkeämpi lukijan kannalta.

Ennen tätä työtä minulla ei ollut kovin suurta tietoa aiheesta ja sen käytöstä, mutta tiesin jotakin koneoppimisesta ja tietovarastoinnista. Tätä työtä tehdessäni sain kokemusta kohtalaisesti jokaisesta liiketoimintatiedon osa-alueesta ja se syvensi mielenkiintoani koneoppimista ja tilastotieteitä kohtaan.

Opittuani enemmän aiheesta huomasin myös aiheen laajuuden ja työni lyhyyden. Liiketoimintatiedon jokaisesta aihe-alueesta voisi kirjoittaa oman kirjansa, mutta erityisesti jäi harmittamaan kirjoittaminen tiedonlouhinnasta. Kun käsittelin tiedonlouhinnassa käytettyjä tekniikoita, valitsin yleisimmät ja selkeimmät tekniikat, joten en kirjoittanut esimerkiksi keinotekoisista neuroverkostosta (eng. Artifical neural network), jonka eri muodoista, tekniikoista ja sovelluksista voisi itsessään tehdä kattavan opinnäytetyön. Käytännön osiossa mieleen jäi kaikki se avaamaton potentiaali, jonka olisin voinut Orangesta saada tutkimalla tarkemmin eri asetuksia ja vaihtoehtoja, mutta koen, että tämä olisi poikennut työni aiheesta ja samalla luonnut vaikeuksia selostaa jokaiset eri asetukset ymmärrettävästi.

LÄHTEET

Berson Alex, Smith Stephen, Thearling Kurt. An Overview of Data Mining Techniques. Saatavissa: http://www.thearling.com/text/dmtechniques/dmtechniques.htm. Viitattu 4.6.2017

Bramer Max. 2013. Principles of data mining 2nd edition. Springer-Verlag London.

Brown Martin. 2012. Data mining techniques. Saatavissa https://www.ibm.com/developerworks/library/ba-data-mining-techniques/**.** Viitattu 31.5.2017.

Brownlee, Jason. 2016. Supervised and unsupervised machine learning algorithms. Saatavilla: http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. Viitattu 15.6.2017

Chen. Danny H. 2012. OLAP Cubes in the SCSM Data Warehouse : Key Concepts. Saatavissa: https://blogs.technet.microsoft.com/servicemanager/2012/02/03/olap-cubes-in-the-scsm-data-warehouse-key-concepts/. Viitattu13.9.2017.

Devens, Richard Miller. 1865. Cyclopædia of commercial and business anecdotes. New York, London, D. Appleton and company. Saatavilla: https://archive.org/details/cyclopaediacomm00devegoog

Efraim Turban, Ramesh Sharda, Dursun Delen, David King. 2011. Business Intelligence: a managerial approach 2nd edition. Pearson Education Inc.

Frontline Systems Inc. 2017. Classification tree. Saatavilla: https://www.solver.com/classification-tree. Viitattu 21.9.2017.

Hovi Ari, Ylinen Jari ja Koistinen Heikki, 2001. Tietovarastot liiketoiminnan tukena. Talentum media /satku.fi

Lichman, M. 2013. Adult dataset. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Saatavilla: https://archive.ics.uci.edu/ml/datasets/adult

Ljubljanan yliopisto. Interactive visualization. Saatavilla: https://orange.biolab.si/features/interactive-data-visualization/. Viitattu 15.6.2017

Ljubljanan yliopisto. Visual programming. Saatavilla: https://orange.biolab.si/features/visual-programming/. Viitattu 15.6.2017

Ljubljanan yliopisto. Widget catalog. Saatavilla: https://orange.biolab.si/toolbox/

Moore, Valarie. Architecture of a Data Warehouse with a Staging Area and Data Marts. Saatavissa https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm. Viitattu 9.5.2017.

Vercellis Carlo. 2009. Business intelligence: Data mining and optimization for decision making. John Wiley & Sons Ltd.

Liitteet

TAULUKKO 4: Adult-tietojoukon attribuuttien kuvaukset ja arvot

Ominaisuus	**Kuvaus ja arvot**
y	Henkilön tulot vuodessa, joko alle 50 000 tai yli 50 000.
Age	Ikä. Tietojoukosta on poistettu kaikki alle 17-vuotiaat.
Workclass	Henkilön työnantaja: - Private (Yksityinen). - self-emp-not-inc (Itsenäistyöllistetty, ei yhtiötä). - self-emp-inc (Töissä omassa yhtiössä). - federal-gov (Liittovaltion hallitus). - local-gov (Kunnan hallitus). - State-gov (Valtion hallitus). - Without-pay(Palkaton). - Never-worked(Ei ole koskaan ollut töissä).
Fnlwgt	Kuvastaa arviota, että kuinka monta henkilöä yksi tietojoukon merkintä voi edustaa.
Education	Henkilön korkein koulutus: - Doctorate (Tohtori, korkeimman yliopiston tutkinto) - Masters (Maisteri), - Bachelors (Alempi korkeakoulututkinto). - Some-college (College, ei valmistunut). - Assoc-acdm (Academic associate’s degree. Akateeminen kahden vuoden koulutus yliopistossa tai collegessa). - Assoc-voc (Vocational associate’s degree. Ammattilinen kahden vuoden koulutus collegesssa tai yliopistossa). - Prof-school (Ylempi korkeakoulututkinto). - HS-grad (Toiseen asteen koulutus).

	- 1st-4th, 5th-6th, 7th-8th, 9th 10th, 11th, 12th (Vuodet koulussa, 1-6 on peruskoulu ja 6-12 toiseen asteen koulutus). - Preschool (esikoulu).
Education-num	Henkilön koulutus numeromuodossa.
Marital-status	Siviilisääty: - Married-civ-spouse (Siviiliavioliitto tai rekisteröity parisuhde). - Divorced (Eronnut). - Never-married (Ei naimisissa). - Separated (Asumusero). - Widowed (Leski). - Married-spouse-absent (Naimisissa, mutta puoliso asuu muualla esimerkiksi työn takia). - Married-AF-spouse (Naimisissa, puoliso puolustusvoimissa).
Occupation	Ammatti: - Tech-support (Tekninen tuki). - Craft-repair (Rakennus- ja korjauspalvelut). - Other-service (Muu palvelu). - Sales (Myynti). - Exec-managerial (Yrityksen hallituksessa esimerkiksi toimitusjohtaja). - Prof-specialty (Asiantuntija ja erityisasiantuntija). - Handlers-cleaners(Käsittely- ja siivouspalvelut). - Machine-op-inspct (Koneenkäyttäjä, konetarkastaja) - Adm-clerical (Virkailija). - Farming-fishing (Maanviljely - kalastus). - Transport-moving (Kuljetusala). - Priv-house-serv (Palvelija). - Protective-serv, (Poliisivoima). - Armed-Forces (Puolustusvoimat).

Relationship	Ihmissuhde / sukulaisuus: - Wife (Vaimo). - Own-child (Yksinhuoltaja). - Husband (Aviomies). - Not-in-family (Ei perhettä). - Other-relative (Muu tai sukulainen). - Unmarried (Naimaton)
Race	Etninen tausta: - Asian-Pac-Islander(Aasia tai Tyynenmeren saaret). - Black (Afrikan manner). - Amer-Indian-Eskimo (Amerikan intiaani, eskimo). - Other (Muu).
Sex	Sukupuoli: - Male (Mies). - Female (Nainen).
Capital-gain	Henkilön pääomatulot esimerkiksi osakkeista tai kiinteistöistä.
Capital-loss	Tappiot pääomatulosta. Henkilö on joutunut myymään omaisuuttaan osto-hintaa halvemmalla.
Hours-per-week	Montako tuntia viikossa henkilö työskentelee
Native-country	Henkilön synnyinmaa.