Ikä Tappaa Tieliikenteessä! - Osa 2 – Datan käsittely

Jatkamme tämän artikkelin edellisessä osassa mainitun Suomen Tieliikenneonnettomuudet 2005-2017 datan pyörittelyä, niin että voimme valituista muuttujista tehdä ennustemallin joka pyrkii ennustamaan todennäköisintä onnettomuusluokkaa.

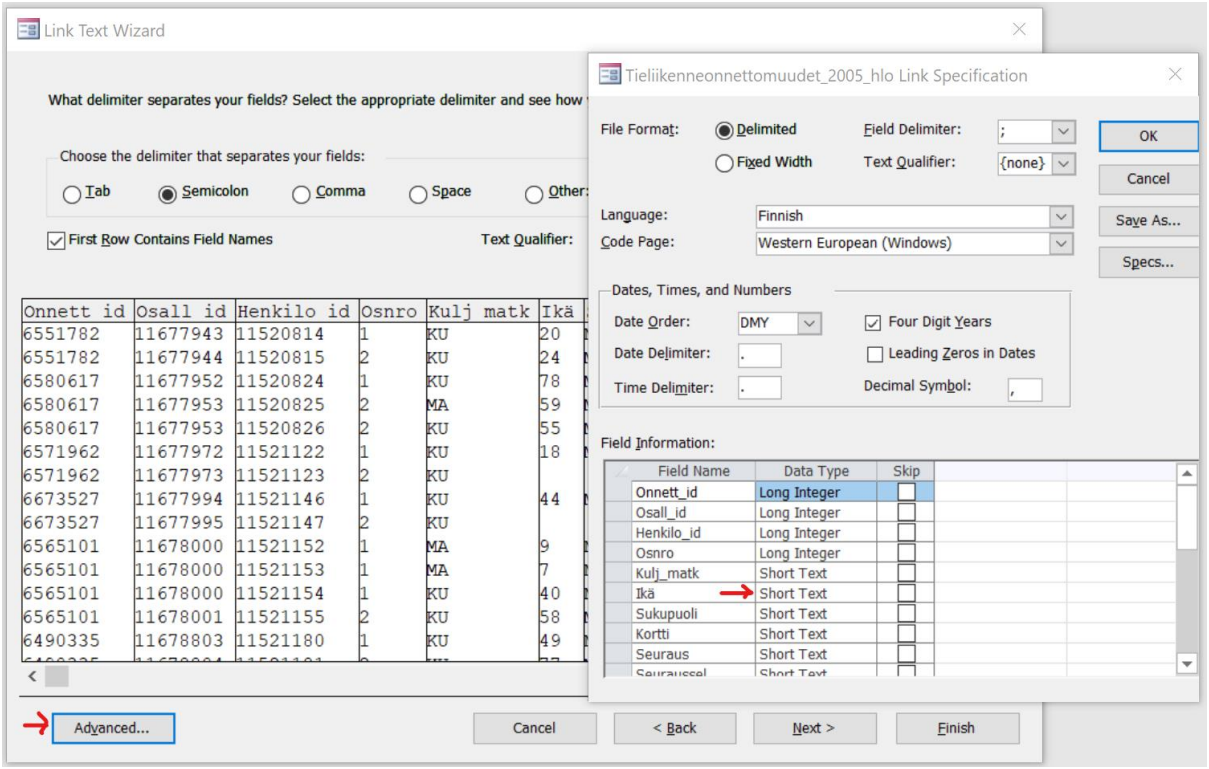
Väylävirasto on julkaissut tietoaineiston Tieliikenneonnettomuudet lisenssillä Creative Commons Attribution 4.0 International License. Tietoaineistossa on kullekin vuodelle kolme toisiinsa linkittyvää tiedostoa. Onnettomuudet-tiedostossa on onnettomuudet, joilla kullakin yksilöivä tunnus, ja massiivinen määrä taustatietoa itse onnettomuustilanteen tekijöistä. Henkilöt-tiedostossa on henkilötietoa, joilla kullakin yksilöivä tunnus ja se linkittyy onnettomuustunnuksella onnettomuudet-tiedostoon. Osalliset-tiedosto linkittyy sekä onnettomuustunnuksen että henkilötunnuksen kautta edellisiin tiedostoihin, tämä tiedosto kertoo lähinnä kunkin osallisen ajoneuvotiedon, sekä loukkaantumis/kuolemis tiedon, joka löytyy myös henkilöt-tiedostosta. Tietoaineiston tehokas käyttö vaatii jonkin verran relaatiotietokanta tietämystä, että voi yhdistää kunkin tiedoston sisältämän datan toisiinsa.

Koska data on vuosikohtaisissa tiedostoissa, aloittakaamme niiden yhdistely yhteen tiedostoon. Lisäksi data sisältää relaatioita ja rivejä on satojatuhansia, niin käytetään tässä tapauksessa ehkä vähiten käytettyä Microsoft Office paketin ohjelmistoa nimeltä Access, joka on pieni relaatiotietokantaohjelma. Käytössä oli Microsoft Office 365 versio 1912. Excel tukee maksimissaan miljoonan rivin taulua, mutta osoittaa hyytymisen merkkejä jo muutaman sadantuhannen rivin kohdalla, jos aikoo käsitellä ja muokata dataa isompana massana, eikä se sovellu relaatioiden käsittelyyn kovinkaan luontevasti. Accessin rajana on 2 GB Access-tiedoston koko, jota voi tarvittaessa kiertää linkittämällä useita tiedostoja ulkoisina datalähteinä.

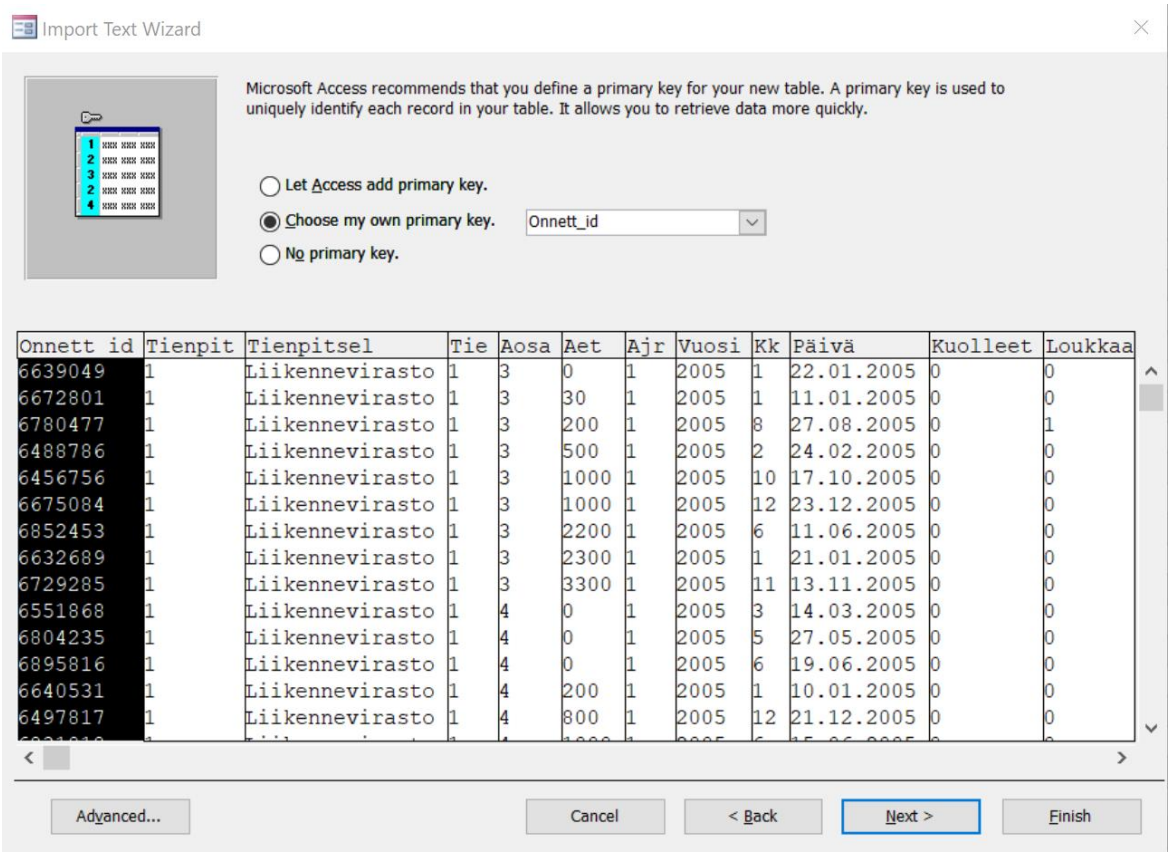
Huom! Kyseisen tietoaineiston tiedostojen yhdistely onnistuu millä tahansa tekstinkäsittelyohjelmallakin, sillä jokaisen vuoden tiedostossa on samat sarakkeet ja sarakenimet. Seuraava harjoite on lähinnä kertomaan kuinka Accessiin voi tuoda ulkoista dataa.

Seuraava on yleistasoinen ohje Access:in käyttöön, josta selviää suoritetut askeleet muttei yksityiskohtaisia ohjeita ohjelmiston käyttöön, jottei artikkelista paisuisi 100 sivuista opusta. Vaikka et olisi koskaan käyttänyt Access:ia, suosittelen kokeilemaan, siihen tutustuminen pitäisi onnistua ihan alla olevilla ohjeillakin.

Luodaan ensin tyhjä Access-tietokanta nimeltään "Tieliikenneonnettomuudet-staging.accdb" datakansion juureen, jonne tietoaineiston vuosikansiot ovat purettuna. Tämän jälkeen tuodaan kukin tiedosto Accessiin, External Data – New Data Source – From File – Text File. Avautuvasta dialogista valitaan "Import the source data into a new table in the current database", koska tässä tapauksessa data on staattista ja haluamme tarkistaa datan indeksoinnin (yksilöivät tunnukset) eheyden. Tuodaan data Accessin ehdottamilla datatyypeillä, koska haluamme vain yhdistää datan raakana yhteen tiedostoon ilman muutoksia. Näin vältämme sen ettei tässä vaiheessa karsita tietuita joiden sarakkeen datatyyppi vaihtelee soluttain, esim. numeroita ja kirjaimia.



Koska tietoaineistossa kussakin tiedostossa on kullakin tietueella yksilöivä tunnus (ID) jota voidaan käyttää indeksinä, aseta oma Primary Key. Onnettomuudet-tiedoston kohdalla valitse “Onnett_id”, Henkilöt-tiedoston kohdalla “Henkilo_id” ja Osalliset-tiedoston kohdalla “Osall_id”. Näin voimme varmistua että indeksointi on eheä ja ettemme vahingossa tuo samaa dataa kahdesti, sillä se ei onnistu jos Primary Key sarakkeesta löytyy samoja arvoja.

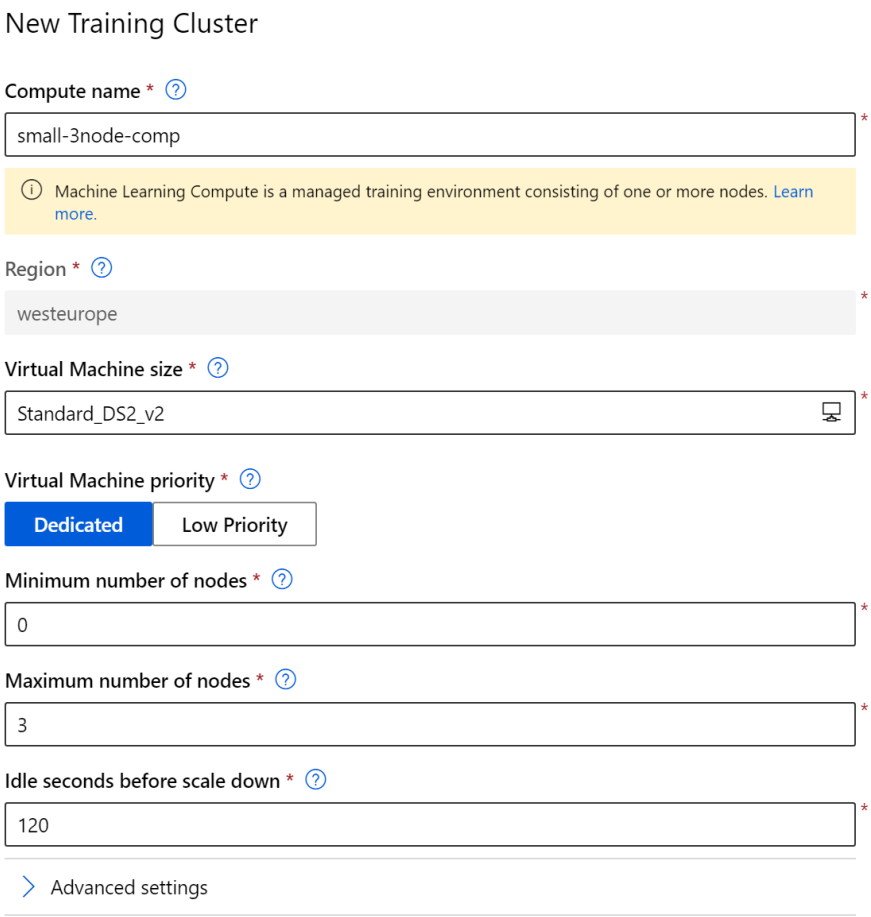


Seuraavan vuoden tiedostoa tuodessa valitaan “Append a copy of the records to the table” ja lisätään seuraavan vuoden data edellä luotuun tauluun. Toistetaan niin kauan että kunkin vuoden tiedostot ovat vastaavissa tauluissaan, jonka jälkeen Accessissa tulisi olla kolme erillistä tietokanta taulua Onnettomuuksille, Henkilöille ja Osallisille.

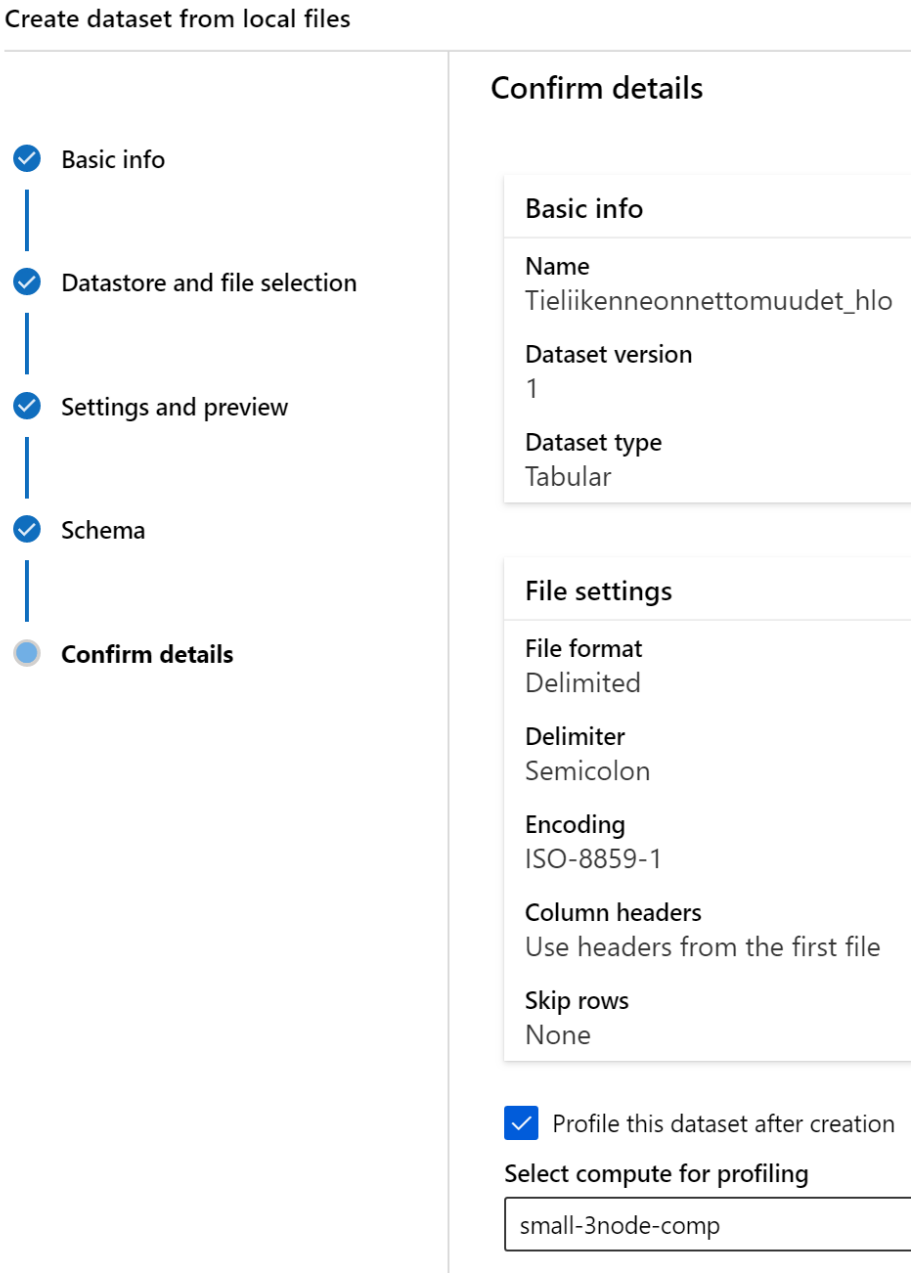
Joidenkin vuosien kohdalla Onnettomuus-tiedosto (2009,2010, 2012 ja 2016) antoi ilmoituksen tuontivirheestä muutaman rivin osalta, samalla luoden ImportErrors taulun, johtuen epäonnistuneesta tyyppikonversiosta joidenkin sarakkeiden kohdalla, josta ei kannata tässä tapauksessa välittää. Muutaman rivin puuttumisella ei käyttöskenaariossamme ole merkitystä.

Kun data on saatu tuotua tauluihin, onkin aika viedä ne tekstitiedostoiksi, joka tapahtuu kutakin taulua oikeaklikkaamalla – Export – Text File.

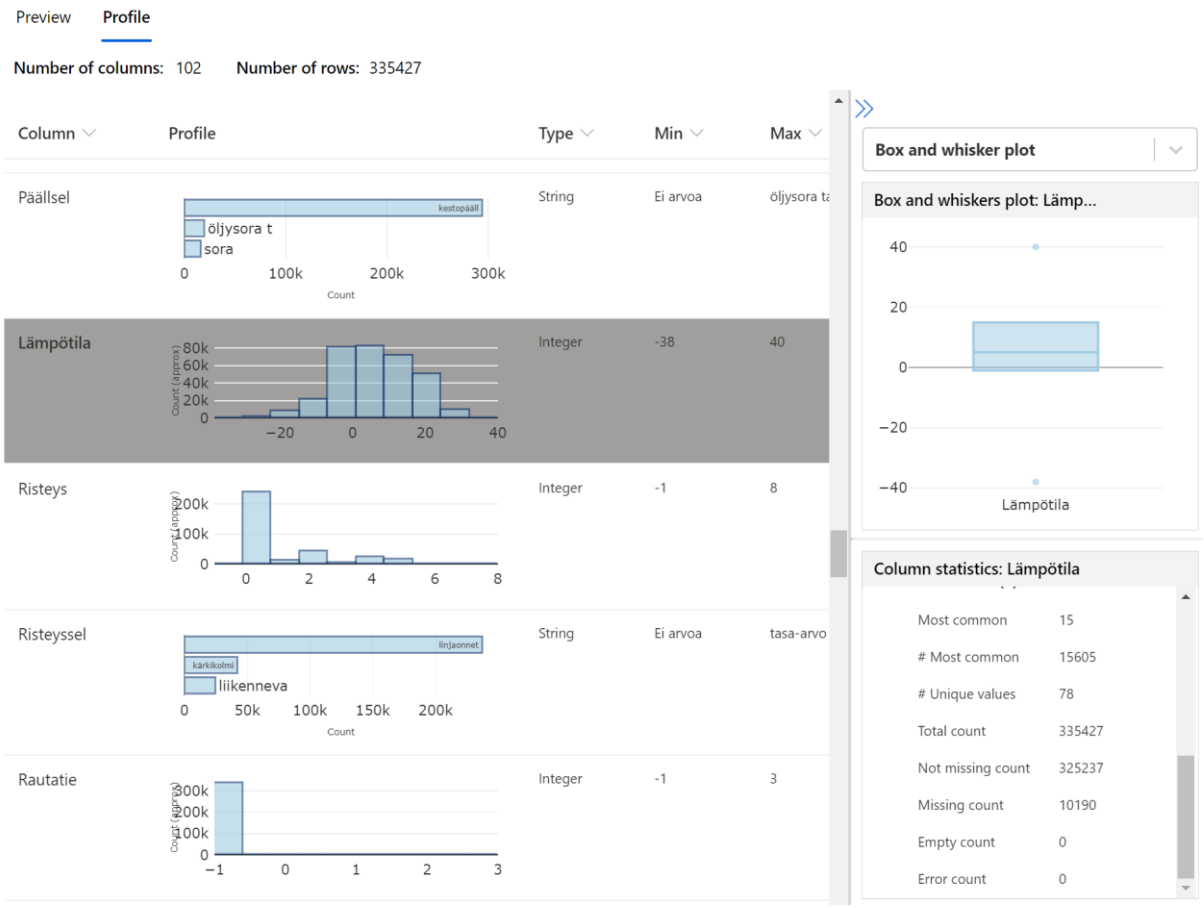
Nyt kun kaikki data on kolmessa tiedostossa, voimme analysoida datan ominaisuuksia. Tässä tapauksessa käytämme siihen Azure Machine Learning työtilasta (Preview:ssa kirjoituksen hetkellä) löytyvää tietoaineiston profilointi ominaisuutta (Datasets – Profile). Lisää Azure portaalissa tilaukseesi Azure Machine Learning Workspace, josta löytyy suora linkki työtilaportaaliisi. Datan profilointia varten Azure Machine Learning työtilassa tarvitsemme ensin palvelinklusterin (Compute – Training Cluster - New), kevyimmän mahdollisen, mieluusti kolme noodia kun kolme tiedostoa niin voi profiloida kaikki yhtä aikaa ja jota voimme käyttää jatkossa koneoppimismallien kouluttamiseen. Palvelinklusteri osaa sammuttaa noodit jos niitä ei käytetä, joten käyttökustannukset jäävät vähäisiksi vaikka klusteria ei poistaisikaan joka käytön jälkeen. Toki kannattaa seurailla käyttökustannuksia, sillä Preview-vaiheen alkupuolella klusteri aiheutti yllättäviä kuluja, ei ilmeisesti osannut sammutella noodeja.



Kun klusteri luotu, voidaan ladata tietoaineistot (Datasets – Create Dataset), valitse Confirm Details -sivulla "Profile this dataset after creation".



Kun tietoaineistot profiloitu, voi niitä tarkastella näkymästä Datasets – Explore – Profile. Profiili antaa monen näköistä statistiikkaa ja hyvän lähtökohdan arvioida mitkä muuttujat (feature) kannattaa ottaa lähempään tarkasteluun, jo siltä pohjalta ettei puuttuvia tai tyhjiä arvoja olisi suurin osa. Ko. tietoaineistossa käytetty puuttuvina arvoina mm. "Ei arvoa", "-1", "-", "NA", "XX", "Tieto puuttuu" ja "Ei tiedossa" tyyppisiä syötteitä.



Muuttujia etsiessä kannattaa kiinnittää huomiota datan määrän lisäksi siihen että se on tasaisesti jakautunutta eri arvojen välillä eikä suuria vääristymiä olisi, ts. yksi arvo ylipainottunut muiden suhteen. Muuttujia valitessa on myös pidettävä mielessä mitä niistä on ennustemallin käyttäjälle helposti saatavilla, kuten kuukausi tai lämpötila, ja mitä on melko mahdoton kyseisen käyttäjän selvittää, kuten esimerkiksi tien ylläpitäjä tai tien näkyvän osuuden pituus.

Seuraavaksi luomme uuden Access tietokannan, “Tieliikenneonnettomuudet.accdb”, jonne tuomme yhdistetyistä tietoaineistoista haluamamme sarakkeet. Tällä kertaa linkitämme sen ulkoiseen datalähteeseen (Link to the data source by creating a linked table) eli edellä luomiimme tekstitiedostoihin, siltä varalta että saisimme lisää tuoretta vuosidataa (kyseistä tietoaineistoa ei tosin enää päivitetä, dataa ei ole saatavilla ainakaan samassa muodossa ja linkitys ulkoiseen datalähteeseen luo myöhemmin selviävän haasteen datan jatkokäytölle). Toinen vaihtoehto olisi linkittää se suoraan edellä luotuun Access-tietokantaan (staging), jos sen datatyypit olisivat kunnossa, mitä ne eivät ole. Tällä kertaa kiinnitämme erityishuomion käytettyihin datatyypeihin emmekä tuo kaikkia sarakkeita.

Henkilöt-tiedostosta tuomme vain seuraavat sarakkeet, kaikki muut jätetään tuomatta (Skip):

- Onnett_id
- Osall_id
- Henkilö_id
- Osnro
- Kulj_matk
- Ikä

- Sukupuoli
- Seuraus
- Seuraussel

Samalla vaihdamme "Ikä" sarakkeen datatyypin numeroksi (Long Integer) ja tarkastamme että muissakin numerosarakkeissa datatyyppi on oikein.

Onnettomuus-tiedostosta tuomme vain seuraavat sarakkeet, kaikki muut jätetään tuomatta, datatyypin muutos suluissa:

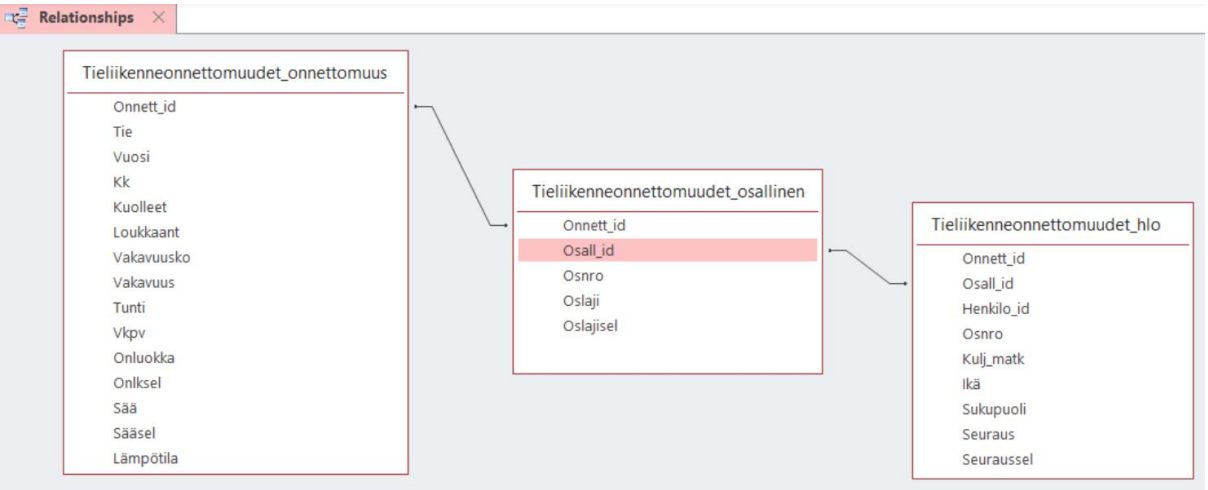
- Onnett_id
- Tie (Long Integer)
- Vuosi
- Kk
- Kuolleet
- Loukkaant
- Vakavuusko
- Vakavuus
- Tunti
- Vkpv
- Onluokka
- Onlksel
- Sää
- Sääsel
- Lämpötila (Long Integer)

Osallinen-tiedostosta tuomme vain seuraavat sarakkeet, kaikki muut jätetään tuomatta:

- Onnett_id
- Osall_id
- Osnro
- Oslaji
- Oslajisel

Nyt kun ulkoiset datalähteet on linkitetty, voimme luoda niiden väliset suhteet (Database Tools – Relationships), tuo relaatiotietokantojen mystisin ominaisuus jossa voi mennä niin monella tavoin mönkään. Luomalla vääränlaisia suhteita voimme lisätä tai rajata tietueiden määrää vaikka se ei olisikaan tarkoitus. Linkitys tapahtuu Relationships-näkymässä lisäämällä tahdotut taulut ja raahaamalla linkittävästä (yksilöivästä, yleensä ID) sarakkeesta toiseen, sekä määrittelemällä linkityssuhde (Join Type).

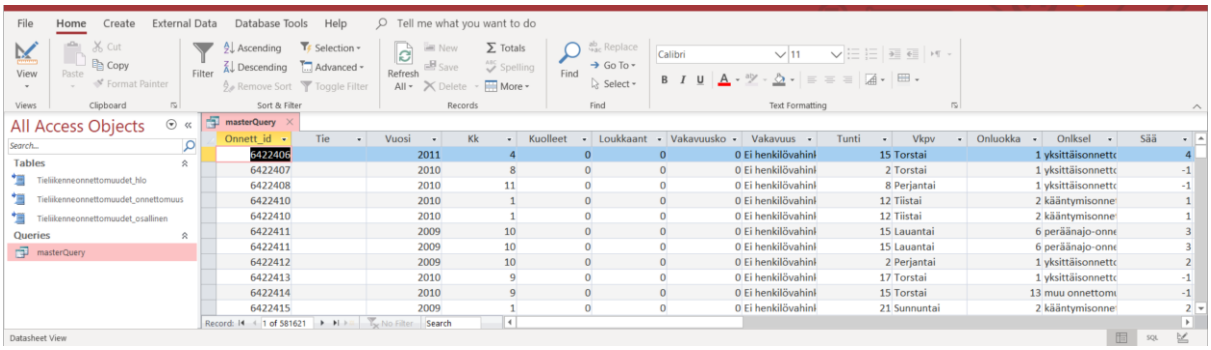
Linkitetään Onnettomuus-taulu Osallinen-tauluun, niin että kummankin yhteneväiset "Onnett_id" tietueet näytetään (Inner Join). Sen lisäksi linkitetään Osallinen-taulu Henkilöt-tauluun, niin että kummankin yhteneväiset "Osall_id" tietueet näytetään. Näin siksi että haluamme vain semmoista dataa joka löytyy kaikista tauluista ja vähennämme jatkossa tarveta poistaa tietueita joille ei löydykään arvoa kaikista sarakkeista. Tämän jälkeen kun luodaan kyselyitä (Query) on linkitys taulujen välillä oletuksena luotu eikä sitä tarvitse enää tehdä, muuttaa vain niissä tapauksissa kun halutaan jotain muuta lopputulemaa kuin oletuksena on tarjolla, esim. kaikki tietueet Onnettomuus-taulusta vaikka niille ei löytyisikään vastinetta muista tauluista (Left Join).



Seuraavaksi voimmekin alkaa tekemään tarkoituksenmukaisia kyselyitä (Query) tauluihin, Create – Query Wizard – Simple Query Wizard. Luodaan ensin yleisnäkymä kaikkeen ja nimetään se "masterQuery":ksi, eli lisää Onnettomuus-taulusta kaikki sarakkeet, Osallinen-taulusta kaikki paitsi "Onnett_id"-sarake ja Henkilöt-taulusta kaikki paitsi "Onnett_id"-, "Osall_id" ja "Osnro"-sarakkeet.

Seuraavat kyselyt tehdään "masterQuery":ä vasten. Tämä siksi että mahdollisten taulumuutosten tapahtuessa on helpompi korjata yksi kysely kuin kymmentä. Esimerkiksi itse huomasin myöhemmin, kun olin luonut jo useita kyselyitä, että Henkilöt-taulun "Ikä"-sarake ei ollutkaan numeerinen vaan tekstiä, jonka lajitteluominaisuudet ovat tyystin erilaiset kuin numeroilla (1,10,11,12,...2,20,21), eikä datatyypin vaihto suoraan taulun ominaisuuksista onnistunutkaan, joten jouduin tuomaan taulun datan uusiksi. Varovaisena toin taulun eri nimellä, että näin datatyypin vaihdoksen vaikutukset, ja loin vuorovaikutussuhteet uuteen tauluun kuten aikaisemmin. Tämän jälkeen vaihdoin vain master-kyselyssä kyseltävän taulun uuteen, koska sarakeotsikot pysyivät samana oli vaihdos hyvin yksinkertainen. Muihin kyselyihin ei tarvinnut tehdä muutoksia ja ne toimivat samoin tein. Ketjutetut kyselyt voivat vaikuttaa suorituskykyyn, mutta pienympäristössä erot ovat merkityksettömiä.

Mahtavaa! Olet tehnyt pohjatyöt ja Access-kantasi sekä "masterQuery" näyttää jotakuinkin tältä.



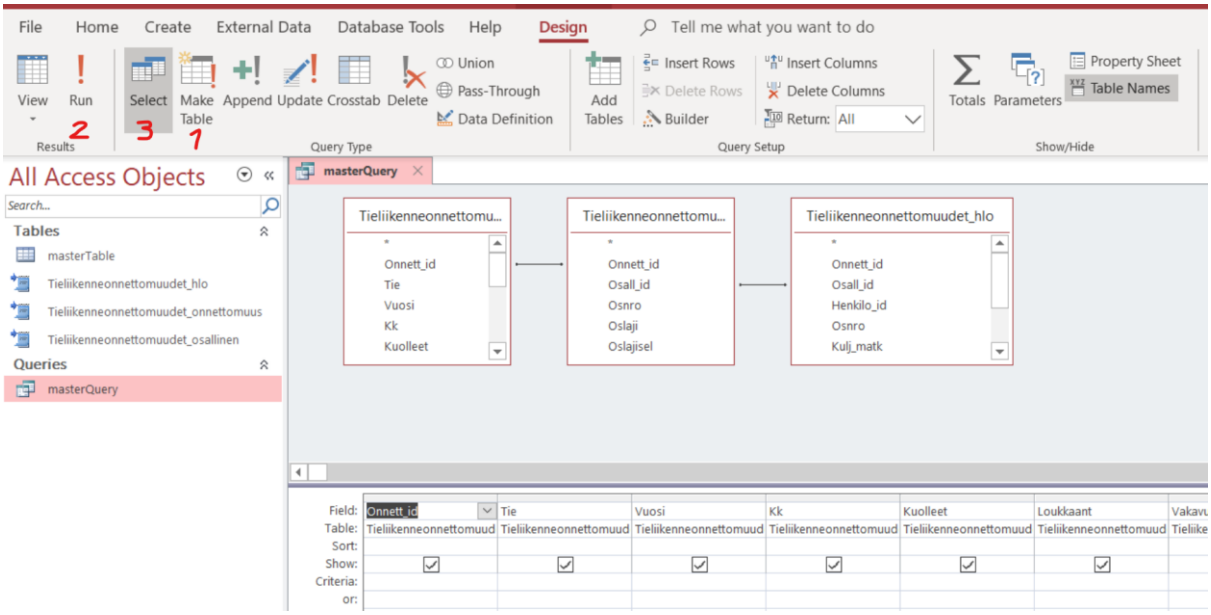
Jos haluat vilkaista miltä konepellin alla näyttää, vaihda näkymiä Datasheet View, SQL View ja Design View välillä esim. oikean alakulman pikanapeilla. Jos teet muutoksia ja haluat ne säilyttää, tallenna muutokset vasemman yläkulman Save-pikanapilla.

“MasterQuery”:n SQL lauseke näyttää tältä.

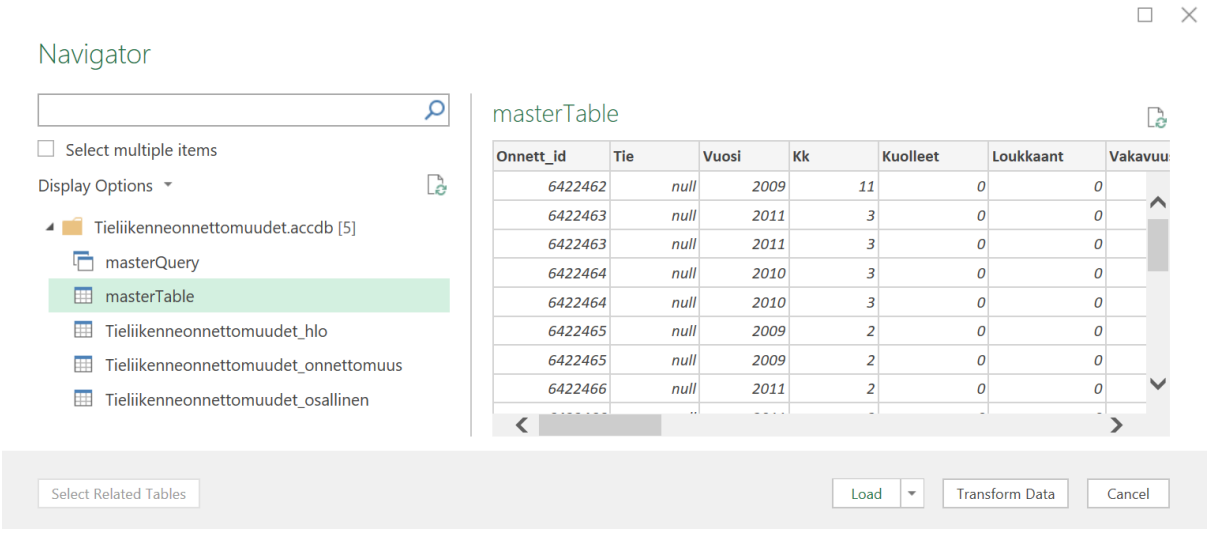
```
SELECT Tieliikenneonnettomuudet_onnettomuus.Onnett_id,
Tieliikenneonnettomuudet_onnettomuus.Tie,
Tieliikenneonnettomuudet_onnettomuus.Vuosi,
Tieliikenneonnettomuudet_onnettomuus.Kk,
Tieliikenneonnettomuudet_onnettomuus.Kuolleet,
Tieliikenneonnettomuudet_onnettomuus.Loukkaant,
Tieliikenneonnettomuudet_onnettomuus.Vakavuusko,
Tieliikenneonnettomuudet_onnettomuus.Vakavuus,
Tieliikenneonnettomuudet_onnettomuus.Tunti,
Tieliikenneonnettomuudet_onnettomuus.Vkpv,
Tieliikenneonnettomuudet_onnettomuus.Onluokka,
Tieliikenneonnettomuudet_onnettomuus.Onlksel,
Tieliikenneonnettomuudet_onnettomuus.Sää,
Tieliikenneonnettomuudet_onnettomuus.Sääsel,
Tieliikenneonnettomuudet_onnettomuus.Lämpötila,
Tieliikenneonnettomuudet_osallinen.Osall_id,
Tieliikenneonnettomuudet_osallinen.Osnro,
Tieliikenneonnettomuudet_osallinen.Oslaji,
Tieliikenneonnettomuudet_osallinen.Oslajisel,
Tieliikenneonnettomuudet_hlo.Henkilo_id,
Tieliikenneonnettomuudet_hlo.Kulj_matk, Tieliikenneonnettomuudet_hlo.Ikä,
Tieliikenneonnettomuudet_hlo.Sukupuoli,
Tieliikenneonnettomuudet_hlo.Seuraus,
Tieliikenneonnettomuudet_hlo.Seuraussel

FROM (Tieliikenneonnettomuudet_onnettomuus INNER JOIN
Tieliikenneonnettomuudet_osallinen ON
Tieliikenneonnettomuudet_onnettomuus.[Onnett_id] =
Tieliikenneonnettomuudet_osallinen.[Onnett_id]) INNER JOIN
Tieliikenneonnettomuudet_hlo ON Tieliikenneonnettomuudet_osallinen.[Osall_id]
= Tieliikenneonnettomuudet_hlo.[Osall_id];
```

Koska Access:ista puuttuu datan visualisointi, niin graafien tekemiseen Excel toimii vallan mainiosti isommillakin datamäärillä. Voisimme käyttää ulkoisena datalähteenä juuri luomaamme “masterQuery”:ä Access-tietokannassamme, jos se taasen EI käyttäisi ulkoiseen datalähteeseen linkitettyjä tauluja vaan data olisi tuota Access-kantaan. Jäljelle jää muutama vaihtoehto, “masterQuery”:n datan vieminen ulos tekstitiedostoksi tai Make Table Query. Valitaan jälkimmäinen.

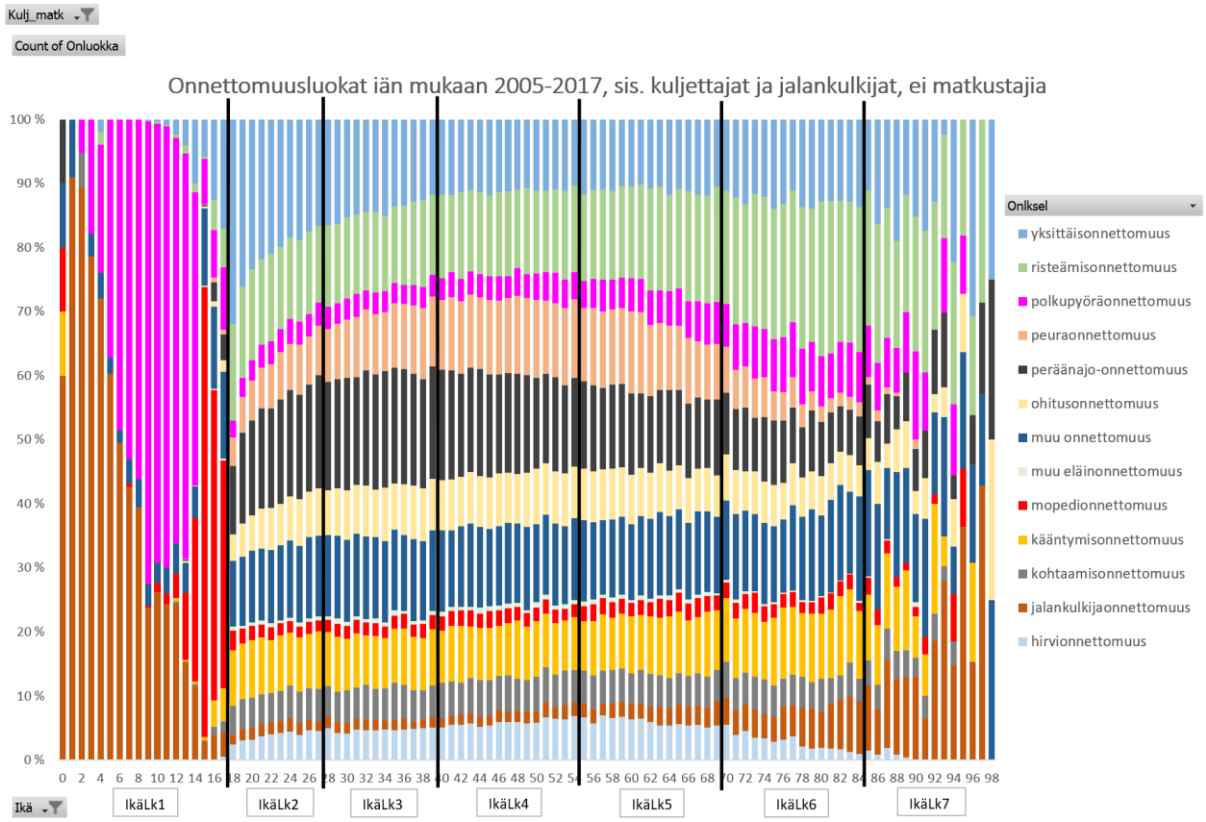
Mene “masterQuery”:n Design näkymään, jos kysely ei ole avoinna esim. oikeaklikkaamalla kyselyn nimeä – Design View - paina Make Table nappia (1) - anna taululle nimi “masterTable” – OK – paina Run nappia (2) ja odota, kun uusi taulu on muodostunut - vaihda “masterQuery”:n tyypiksi takaisin select query painamalla Select nappia (3).



Luo tyhjä Excel tiedosto "Tieliikenneonnettomuudet.xlsx", linkitä se Access-kannan "masterTable" tauluun, Data – Get Data – From Database – From Microsoft Access Database – etsi ja valitse Tieliikenneonnettomuudet Access-tiedosto – valitse "masterTable" - Load.



Luo näyttäviä graafeja, esimerkkejä voit katsella tämän artikkelisarjan ensimmäisestä osasta, ja tutki niiden kautta eri muuttujien vaikutusta onnettomuuksiin, jotta löydät vaihtoehtoisimmat valittavaksi ennustemalliin.



Kun käsitys muodostunut mitä muuttujia kannattaa ottaa mukaan ensimmäisiin ennustemallikokeiluihin, niin luodaan sen pohjalta kysely, jonka data viedään tekstitiedostoon. Käydään lävitse yhden mahdollisen muuttujavalikoiman luonti ja tällä kertaa ilman wizardia.

Access-tietokannassa, luo uusi Query, Create – Query Design – Select table dialogista vaihda Query näkymä ja valitse “masterQuery” - Add. Design näkymässä, kaksoisklikkaa lisätystä “masterQueary”:stä “Ikä”-sarake – poista valinta Show laatikosta – lisää Criteria riville ehto: >0.

Yllä olevasta kuvasta huomaat että iät on jaettu luokkiin tiettyjen onnettomuusluokkakuvioiden muutosten perusteella, varsinkin nuoruusvuodet ja vanhuus, välivuodet pyritti jakamaan tasaisesti. Luokaamme kyseiset ikäluokat kyselyyn suoraan “Ikä” sarakkeen viereen.

Kopioi seuraava Field riville “Ikä”-sarakkeen viereen, voit painaa Builder nappia kun oikea sarake Field riviltä valittuna niin helpompi nähdä kokonaisuus: IkäLk: IIf([Ikä] Between 0 And 17;1;IIf([Ikä] Between 18 And 27;2;IIf([Ikä] Between 28 And 39;3;IIf([Ikä] Between 40 And 54;4;IIf([Ikä] Between 55 And 69;5;IIf([Ikä] Between 70 And 84;6;IIf([Ikä]>84;7)))))))

Kaksoisklikkaa “Sää”-sarake – lisää Criteria riville ehto: >0. Kaksoisklikkaa “Kk”-sarake – lisää Criteria riville ehto: >0. Kaksoisklikkaa “Tunti”-sarake – lisää Criteria riville ehto: >0. Kaksoisklikkaa “Onluokka”-sarake – lisää Criteria riville ehto: >0 And <13.

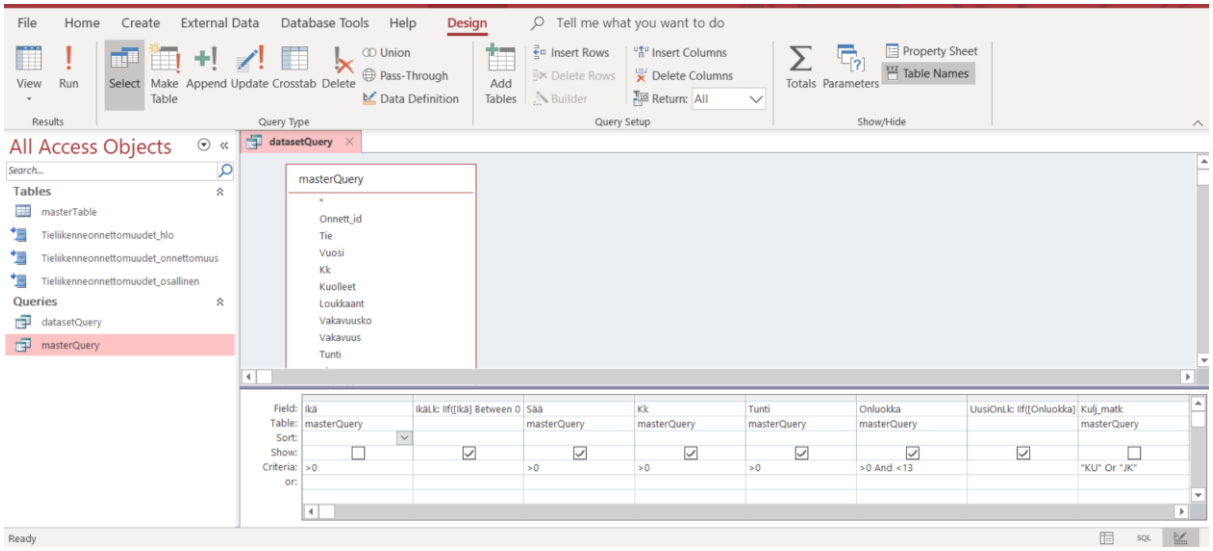
Koska onnettomuusluokat sisältävät samankaltaisuuksia ja niitä on syytä yhdistellä ennustetarkkuuden parantamiseksi, luomme valmiiksi “UusiOnLk”-sarakkeen perustellulla yhdistelysäännöllä (onnettomuusprofiilit samankaltaisia), jota voi myöhemmin hienosäätää esim. uuden onnettomuusluokkanumeroinnin suhteen (numerointi kasvanut isoksi useamman iterointikierroksen jälkeen).

Kopioi seuraava Field riville “Onluokka”-sarakkeen viereen, muista käyttää Builder näkymää: UusiOnLk:
IIf([Onluokka]=1;1;IIf([Onluokka]=2;44;IIf([Onluokka]=3;43;IIf([Onluokka]=4;44;IIf([Onluokka]=5;43;IIf([Onluokka]=6;6;IIf([Onluokka]=7;27;IIf([Onluokka]=8;27;IIf([Onluokka]=9;9;IIf([Onluokka]=10;30;IIf([Onluokka]=11;30;IIf([Onluokka]=12;30;IIf([Onluokka]=13;13;[Onluokka]))))))))))))

Kaksoisklikkaa “Kulj_matk”-saraketta, poista valinta Show laatikosta - lisää Criteria riville ehto: “KU” Or “JK”. Näin valitsemme pois onnettomuuteen useimmiten syyttömät matkustajat, vain kuljettajat ja jalankulkijat ovat mukana.

Paina Save nappia, anna nimeksi “datasetQuery”. Kun kysely tallennettu, paina Run nappia.

“DatasetQuery” näyttää Desing näkymässä tältä.



“DatasetQuery”:n SQL lauseke näyttää tältä.

SELECT IIf([Ikä] Between 0 And 17,1,IIf([Ikä] Between 18 And 27,2,IIf([Ikä] Between 28 And 39,3,IIf([Ikä] Between 40 And 54,4,IIf([Ikä] Between 55 And 69,5,IIf([Ikä] Between 70 And 84,6,IIf([Ikä]>84,7))))))) AS IkäLk, masterQuery.Sää, masterQuery.Kk, masterQuery.Tunti, masterQuery.Onluokka, IIf([Onluokka]=1,1,IIf([Onluokka]=2,44,IIf([Onluokka]=3,43,IIf([Onluokka]=4,44,IIf([Onluokka]=5,43,IIf([Onluokka]=6,6,IIf([Onluokka]=7,27,IIf([Onluokka]=8,27,IIf([Onluokka]=9,9,IIf([Onluokka]=10,30,IIf([Onluokka]=11,30,IIf([Onluokka]=12,30,IIf([Onluokka]=13,13,[Onluokka])))))))))))) AS UusiOnLk

FROM masterQuery

WHERE (((masterQuery.Ikä)>0) AND ((masterQuery.Sää)>0) AND ((masterQuery.Kk)>0) AND ((masterQuery.Tunti)>0) AND ((masterQuery.Onluokka)>0 And (masterQuery.Onluokka)<13) AND ((masterQuery.Kulj_matk)="KU" Or (masterQuery.Kulj_matk)="JK"));

Nyt ensimmäinen datasettisi on tekstitiedostoon vientiä vailla valmis, oikeaklikkaa “datasetQuery” – Export – Text File. Suosittelen että viet vain jomman kumman “Onluokka”- ja “UusiOnLk”-sarakkeista, poista Show valinta Design näkymässä toiselta – Save - Run. Ne ovat arvoja jota ennustemalli pyrkii ennustamaan ja on havaittu niiden ollessa yhdessä datasetissä aiheuttavan ongelmia ennustemallin pyytämiin muuttujiin vaikka kuinka olisi laitettu “Ignore” valintaa sarakkeeseen ennustemallia koulutettaessa.

Seuraavassa osassa pyrimmekin paneutumaan itse ennustemallin kouluttamiseen, miksi edellä kerrottu “UusiOnLk” oikein luotiin ja muita yksityiskohtia koneoppimismallista ja sen hyödyntämisestä.

Tämän LinkedIn artikkelisarjan osat:

Osa1: https://www.linkedin.com/pulse/ikä-tappaa-tieliikenteessä-osa-1-raakadata-marko-peltojoki

Osa2: https://www.linkedin.com/pulse/ikä-tappaa-tieliikenteessä-osa-2-datan-käsittely-marko-peltojoki

Osa3: https://www.linkedin.com/pulse/ikä-tappaa-tieliikenteessä-osa-3-ennustemalli-marko-peltojoki