# Datan valmistelu

CRISP-DM:n kolmas vaihe on datan valmistelu, vaiheen tarkoituksena on valita projektiin sopiva data, ennakko valmistella se mallinnusta varten, rakentaa lisää dataa tarvittaessa, ja yhdistää eri datat yhteen, jos tarpeellista.

## Valintojen perusteet

Kuvaa alle miksi valitut datat valittiin projektille käyttöön, sekä mahdolliset rajaukset, jos dataa rajattiin pois. Tarkoituksena on siis luoda säännöt miksi, kyseessä olevat datat valittiin projektin käyttöön.

|  |  |  |
| --- | --- | --- |
| Data | Perusteet | Rajaukset |
|  |  |  |

## Datan laadun parannukset

Kuvaa alle kaikki toimenpiteet mitä kullekin datasetille tehtiin sen laadun parantamiseksi. Arvioi myös tehtyjen toimenpiteiden vaikutusta datan sopivuuteen projektille.

|  |  |  |
| --- | --- | --- |
| Data | Toimenpiteet | Sopivuuden analyysi |
|  |  |  |

## Johdetut muuttujat

Jos dataan tehtiin uusia johdettuja muuttujia, kuvaa miten muuttujat ovat tehty ja miksi.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Uusi muuttuja | Miten johdettu | Miksi |
|  |  |  |  |

## Uudet rivit

Jos dataan tehtiin uusia rivejä, kuvaa miten rivit ovat tehty ja miksi.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Uusi Rivi | Miten muodostettu | Miksi |
|  |  |  |  |

## Yhdistetyt ja Aggregoidut datat

Kuvaa mahdolliset yhdistetyt ja aggregoidut datat, ja miten ne on tehty sekä miksi

### Yhdistetyt data

|  |  |  |
| --- | --- | --- |
| Lähdetiedot | Miten | Miksi |
|  |  |  |

### Aggrigoidut datat

|  |  |  |  |
| --- | --- | --- | --- |
| Aggrigoitu atribuutti/atribuutit | Uudet atribuutit | Miten | Miksi |
|  |  |  |  |