**Week 1 Introductie Big Data en Statistiek: Eindopdracht**

**Globaal projectplan voor het schatten van verkeersdrukte rondom een gemeente**

1. Informatiebehoefte

Een gemeente wil de verkeersdrukte op de snelwegen naar deze gemeente voorspellen. Doel daarvan is het verbeteren van de planning voor inzet van verkeersregelaars. In dit project wordt onderzocht of verkeersdrukte kan worden verklaard door het weer.

1. Benodigde data

Rijkswaterstaat verzamelt verkeersgegevens via verkeerslussen en stelt die als open data beschikbaar. Gegevens per uur van het weer, waaronder temperatuur, zon, bewolking en zicht, luchtdruk, wind en neerslag zijn te downloaden van de website van het KNMI. Neem data van de afgelopen twaalf maanden. Dan zijn de data recent en kan onderzocht worden of het seizoen invloed heeft op verkeersdrukte.

1. Verwerken data

Beschikbare data worden gecontroleerd op ontbrekende en foutieve waarden. Daartoe moeten indicatoren worden opgesteld. Breng extreme waarden en uitschieters van de data in kaart. Neem bij vreemde of opvallende waarden contact op met de broneigenaar.

1. Analyse data

Stel een regressiemodel op waarmee voor elke snelweg het aantal kilometers file per uur kan worden geschat op basis van gegevens over het weer. Toets dit model met een deel van de data dat niet is gebruikt voor het opstellen van het model. Stel vast welke afwijking acceptabel is.

1. Output

De uitkomsten van de analyse worden gepresenteerd in een onderzoeksrapport. In dit rapport worden aanbevelingen gedaan voor het gebruik van het model in de toekomst. Als de resultaten niet voldoen aan de informatiebehoefte wordt advies gegeven over verder onderzoek. Zo zou een conclusie kunnen zijn, dat niet alleen naar snelwegen moet worden gekeken, maar ook naar provinciale wegen. Het weer is wellicht niet voldoende voor een goede voorspelling en dan kan nagedacht worden over andere variabelen met bijbehorende databronnen.

1. Planning en benodigde capaciteit

In dit project is de informatiebehoefte duidelijk, waardoor de tijd voor het voortraject beperkt is. Een accountmanager, eventueel ondersteund door statistische experts, helpt de opdrachtgever hierbij. Een goed begin is het halve werk.

Het downloaden en verwerken van de benodigde data wordt gedaan door data analisten. Zij hebben programmeerervaring. Afhankelijk van de kwaliteit van de data ligt de benodigde tijd tussen enkele uren en enkele weken.

Daarna wordt de data gemodelleerd door statistische experts. Dit kost een paar dagen tot enkele weken, afhankelijk van de verbanden in de data. Sterke verbanden maken het makkelijker een goed model te vinden.

De output wordt meestal gemaakt door de modelmaker of een inhoudelijk expert en kost voor een niet al te ingewikkeld project een paar dagen tot een week.

Complexere projecten met veel betrokkenen worden begeleid door een projectleider.