KNMI data analyse en datavisualisatie Technical Report

Maurice Schaasberg - 11810866
Peter Heemskerk - 11988797
Jochem Holscher - 11007729
Pasqual Lunshof - xxxx

***Abstract*—Vanaf 1900 is er door de KNMI aan het weer gemeten. In dit project hebben we de door de KNMI beschikbaar gestelde data geanalyseerd en gevisualiseerd.**

I. Introductie

AL vanaf 1900 is er door KNMI data verzameld over het weer, eerst met een enkel weerstation, later uitgebreid.

In dit project is de werking onderzocht van verschillende technieken om data op te schonen en verbanden in beeld te brengen.

II. Literatuur

ER is gebruik gemaakt van college notities en aangeboden referentiemateriaal [1].

III. Methode

HIER wordt eerst ingegaan op de data, dan de wijze van opschonen en voorbewerken. Vervolgens wordt ingegaan op de wijze waarop verbanden zijn onderzocht en gepresenteerd.

A. Data

Voor dit onderzoek is gebruik gemaakt van de door KNMI beschikbaar gestelde data. [2]. Van deze dataset is alle data vanaf 1900 en voor alle weerstations geselecteerd.

	$vooropschonen$	$naopschonen$
weerstations	51	51
kenmerken	39	27
metingen	63986	63986

$Tabel 1$: Statistieken van KNMI dataset.

B. Methode

De KNMI data wordt in een aantal stappen bewerkt:

1) Data voorbewerken en opschonen
   De van de KNMI website gedownloadede data wordt als text file ingelezen, de data wordt opgeschoond en als csv bestand weggeschreven.
2) Exploratieve data analyse (EDA)
   Diverse univariate en multivariate verbanden zijn onderzocht.
3) visualisatie De interessante verbanden worden naar een web-site gepubliceerd voor presentatie.

C. Data opschonen en voorbewerken

De van de KNMI website gedownloadede data wordt als text file ingelezen. Vervolgens zijn de regels met beschrijvende gegevens verwijderd. De data wordt op een aantal punten geschoond. De kenmerken waarvan minder dan 60% gevuld zijn zijn verwijderd. De overblijvende kenmerken zijn:
STN - meetstation
YYYYMMDD - datum
DDVEC - vectorgemiddelde windrichting
FHVEC - vectorgemiddelde windsnelheid
FG - etmaal gemiddelde windsnelheid
FHX - hoogste uurgemiddelde windsnelheid
FHXH - uurvak meting FHX
FHN - laagste uurgemiddelde windsnelheid
FHNH - uurvak meting FHN
FXX - hoogste windstoot
FXXH - uurvak meting FXX
TG - etmaalgemiddelde temperatuur
TN - minimum temperatuur
TNH - uurvak meting TN
TX - maximum temperatuur
TXH - uurvak meting TX
SQ - zonneschijnduur berekend uit globale straling
SP - percentage van de langst mogelijke zonneschijnduur
PG - etmaalgemiddelde luchtdruk
PX - hoogste uurwaarde van de luchtdruk
PN - laagste uurwaarde van de luchtdruk
UG - etmaalgemiddelde relatieve vochtigheid
UX - maximale relatieve vochtigheid
UXH - uurvak meting UX
UN - minimale relatieve vochtigheid
UNH - uurvak meting UN
EV24 - referentie gewasverdamping (Makkink)

In een klein aantal tabellen kwamen waarden voor die niet passen. Deze waarden zijn vervangen door None. Het gaat om:

1) uurvakken met waarden boven 24 (parameters 'TNH' en 'TXH')

D. Exploratieve data analyse (EDA)

Op alle kenmerkingen is een univariate analyse uitgevoerd. Dit is met name nuttig om eventuele uitbijters en niet geldige waarden te achterhalen (zie Data opschonen). Vervolgens is een aantal bivariate analyses uitgevoerd om een aantal verbanden nader te onderzoeken. Deze worden in de presentatie inhoudelijk toegelicht. (onderzoeksvragen 1 en 2).

Als laatste is met behulp van meervoudige regressie een voorspelmodel gemaakt om het astronomische seizoen te voorspellen aan de hand van input parameters. Eerst wordt 1 input parameter geselecteerd met de beste fit, vervolgens wordt daarbij een tweede beste parameter toegevoegd. Mean squared error wordt als te minimaliseren cost functie gebruikt. (onderzoeksvraag 3).

E. Data Visualisatie

De resultaten worden voor gebruikers geimplementeerd. Hier is gebruik gemaakt van methode xxxx (college20/6)

F. Implementatie

De data is geanalyseerd met behulp van de python Pandas library, waarbij de data in DataFrame structuur wordt opgeslagen. Voor visualisatie is de python Bokeh library gebruikt. Het resultaat is als html in een browser gepresenteerd, middels Github-online.

IV. Experiment

HIERONDER zijn de resultaten en een discussie over deze resultaten beschreven.

A. Resultaten

Het bleek dat er in de loop van de tijd steeds meer metingen van weerstations. Hieronder een overzicht.

Daarnaast is een aantal interessante inzichten gebleken:

1) een breuk in meetmethode rond 1950 van luchtvochtigheid
2) een breuk in meetmethode van tijdstip vanaf 2005
3) temperatuur stijgt met 2 gr C. tussen 1900 en 2017
4) een aantal parameters kent een grote afhankelijkheid van windrichting.
5)

B. Discussion

Het is met de beschikbare methoden en technieken goed mogelijk gebleken een inzicht te geven in de door KNMI beschikbare gegevens. Hieronder een korte discussie over de methoden.

xxxx

Een inhoudelijke discussie over de resultaten:

1) stijging temperatuur
   Het lijkt gemakkelijk dit als klimatologische trend als gevolg van opwarming van de aarde te zien. We moeten voorzichtig zijn. Er kan ook heel goed een verandering in meetmethode achter zitten.
2) afhankelijkheid windrichting
   Past aardig bij onze algemene kennis van het weer
3) afhankelijkheid van luchtdruk
   Interessante afhankelijkheid die nader onderzoek vraagt. We kunnen ons niet wagen aan een inhoudelijke logica.

V. Verantwoording

Dit technisch rapport is onderdeel van in het voorjaar 2018 gegeven bachelor vak Data Analyse en Visualisatie door Gosia Migut binnen de studie Artificial Intelligence aan de Universiteit van Amsterdam.

References

[1] College notes and reference material (2018), *canvas,* .
[2] KNMI (2018), Klimatologie - Daggegevens van het weer in Nederland,