KNMI data analyse en datavisualisatie Technical Report

Maurice Schaasberg - 11810866
Peter Heemskerk - 11988797
Jochem Holscher - 11007729
Pasqual Lunshof - 11711019

***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. Tevens hebben we een voorspelmodel ontwikkeld dat op basis van inhoudelijke kenmerken een voorspelling doet van het seizoen van die meting.**

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 van deze door de KNMI beschikbare data gebruik gemaakt en is de werking onderzocht van verschillende technieken om data en verbanden in deze data in beeld te brengen.

Daarnaast is middels regressie een model gemaakt voor de voorspelling van de vier meteorologische seizoenen.

II. Literatuur

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

III. Methode

HIER wordt eerst ingegaan op de gebruikte data, 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. Voor verdere bewerking is de data eerst opgeschoond.

	$vooropschonen$	$naopschonen$
weerstations	51	51
kenmerken	41	27
metingen	794232	630353

$Tabel 1$ **:** Statistieken van KNMI dataset.

Het is ook goed om te beseffen dat het aantal meetstations dat meetresultaten in de loop van de jaren groeit.

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. Om te voorkomen dat conclusies worden getrokken op kenmerken met te weinig ingevulde gegevens zijn die kenmerken verwijderd die minder dan 60% gevuld met gegevens. De overblijvende kenmerken zijn opgesomd in tabel 2.

In een klein aantal tabellen (parameters 'TNH' en 'TXH') kwamen waarden voor die niet passen. Deze waarden zijn vervangen door NaN (Python Not a Number). Het gaat om uurvakken met waarden boven 24.

Ten behoeve van het seizoens voorspelmodel zijn de input parameters genormaliseerd naar waarden tussen 0 en 1. De windrichting (graden) is vertaald naar een cosinus en een sinus, om op deze manier een continue waarde te hebben.

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 als onderzoeksvraag 3 met behulp van meervoudige regressie een voorspelmodel gemaakt om het meteorologische seizoen te voorspellen aan de hand van input parameters. Het algoritme voegt telkens een parameter of een macht toe, of haalt een parameter weg. Op basis van een specifieke set parameters wordt met een trainingset met de in Numpy beschikbare Least Square Methode [3] de optimale weging van de dan gekozen set parameters vastgesteld. Dit gebeurt op een trainingset van 80% van de gegevens. Het algoritme stopt als een toevoeging minder bijdraagt dan een vooraf bepaalde waarde. Op deze wijze wordt iteratief de best passende set parameters, bijbehorende machten en de wegingen vast gesteld.

Deze regressieresultaten zijn getoets middels een validatie test set van 10% van de gegevens.

$afkorting$	$beschrijving$
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)

$Tabel 2:$ Beschrijving kenmerken.

E. Data Visualisatie

Ten behoeve van visualisatie zijn circa 30 grafieken gebruikt, waarbij gebruik gemaakt is van box plots, histogrammen, staafdiagramme, lijngrafieken, puntgrafieken.

Deze grafieken zijn in een web based demo omgeving opgenomen.

F. Implementatie

De data is geanalyseerd met behulp van de Python Pandas library, waarbij de data in DataFrame structuur wordt opgeslagen. Voor visualisatie zijn zowel de python MatPlotLib als de python Bokeh libraries gebruikt. Het resultaat is door middel van html-pagina's gepubliceerd middeld Github-online.

IV. Experiment

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

A. Resultaten

Bij een aantal gemeten grootheden bleek er een breuk in meetmethoden te bestaan. Rond 1950 zien we een plotselinge verschuiving in de luchtvochtigheid. Rond 2005 zien we een plotselinge verschuiving van 1 uur in het uur dat de maximale dagwind is gemeten.

We herkenden dat de gemiddelde temperatuur tussen 1900 en 2017 met 2gr C is gestegen.

Er bleek een aantal onderlinge verbanden tussen kenmerken. Met name opvallend was de afhankelijkheid van temperatuur en luchtvochtigheid van de windrichting, en de afhankelijkheid van temperatuur en luchtvochtigheid van de luchtdruk.

Het is mogelijk met ons regressiemodel het meteorologische seizoen te bepalen aan de hand van inhoudelijke parameters van een meting. Op basis van 10 testruns verkregen we een score van 67% goed voorspelde seizoenswaarde, dat aanzienlijk beter is dan de benchmark 25% bij een willekeurige keuze.

B. Discussie

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

Inhoudelijk is er over de resultaten het volgende toe te voegen:

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.

Het is een interessant resultaat dat we op basis van regressie van de inhoudelijke kenmerken in staat bleken het seizoen te voorspellen met een voor deze test prachtige score. Het is voor vervolgonderzoek om deze methode te vergelijken met andere methoden zoals clustering.

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,
[3] SciPi.org (2018), SciPi.org,