Igor Moens Verslag Klimaat

2023 BAMAJOR

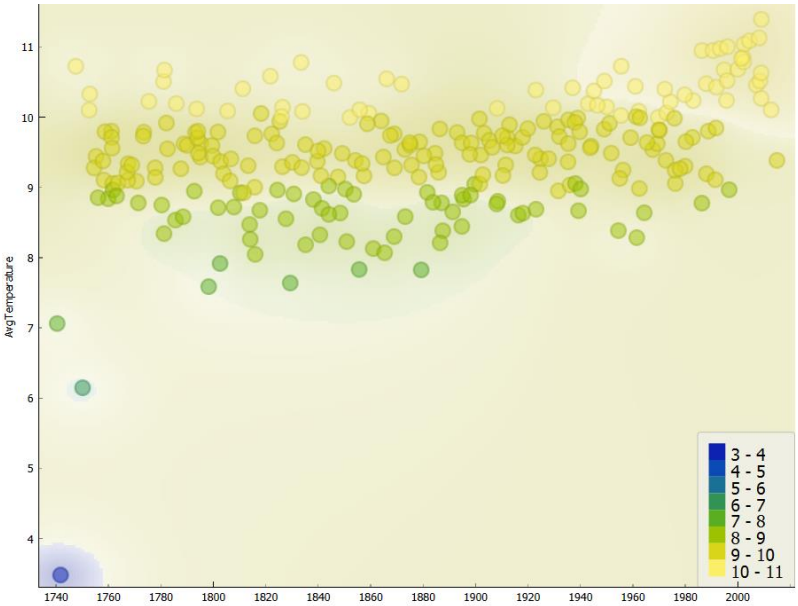
1. Intro

Het onderwerp voor dit verslag is Klimaat. Zo gaan we kijken naar de al veel besproken parameters, namelijk temperatuur, zeeniveau en de uitstoot.

2. Vragen

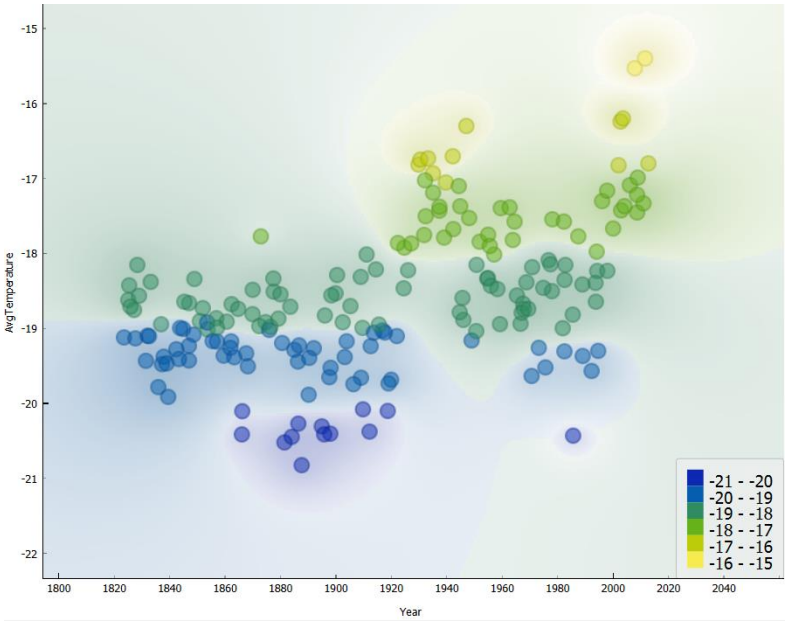
2.1 Als we naar de data kijken van Groenland en België is er dan een duidelijk verschil te zien?

We zullen starten bij België. Hieronder zie je de scatterplot van de gemiddelde temperatuur van België



Als we kijken naar de gemiddelde temperatuur van België in het begin zien we een paar uitschieters omdat deze data opgenomen werd met de eerste versies van de thermometer zoals beschreven in de eerste bron. Algemeen zien we een hele lichte stijging in temperatuur van het begin van de metingen naar het einde toe.

Nu gaan we kijken naar de scatterplot van Groenland.



Als we kijken naar de gemiddelde temperatuur van Groenland zien we dat de stijging van temperatuur sinds het begin van de metingen sterk is. We starten rond de -18 graden en eindigen rond de -16 graden.

Als we nu kijken naar het verschil tussen twee, zien we dat de stijging van de temperatuur in Groenland veel groter is dan die in België. We zouden hier kunnen uit kunnen vermoeden dat temperatuurstijging makkelijker gebeurt en waar te nemen is bij landen met gemiddelde temperaturen onder 0.

2.2 Welke conclusie kan je trekken over de regressiemodellen met betrekking tot het zeeniveau?

We hebben hiervoor het zeeniveau als target genomen. Hiervoor gaan we naar de performantie kijken naar verschillende regressiemodellen aangezien we hier te maken hebben met continue variabelen. We verdelen de data in 60 procent training, 20 procent validation en 20 procent test.

Model	MSE	RMSE	MAE	R2
Linear Regression	12.612	3.551	2.806	0.989
Random Forest	17.941	4.236	3.174	0.985
AdaBoost	17.963	4.238	3.232	0.985
Neural Network met lagen	522.395	22.856	19.003	0.565

We zien hier dat Linear Regression het beste scoort. Dit omdat de RMSE het laagste is, wat betekent dat dit is waar de error het kleinste is. We kunnen dit ook gaan bekijken in de ‘prediction widget’. Hier zien we dan ook dat linear regression altijd het de beste waarde voorspelt als je die vergelijkt

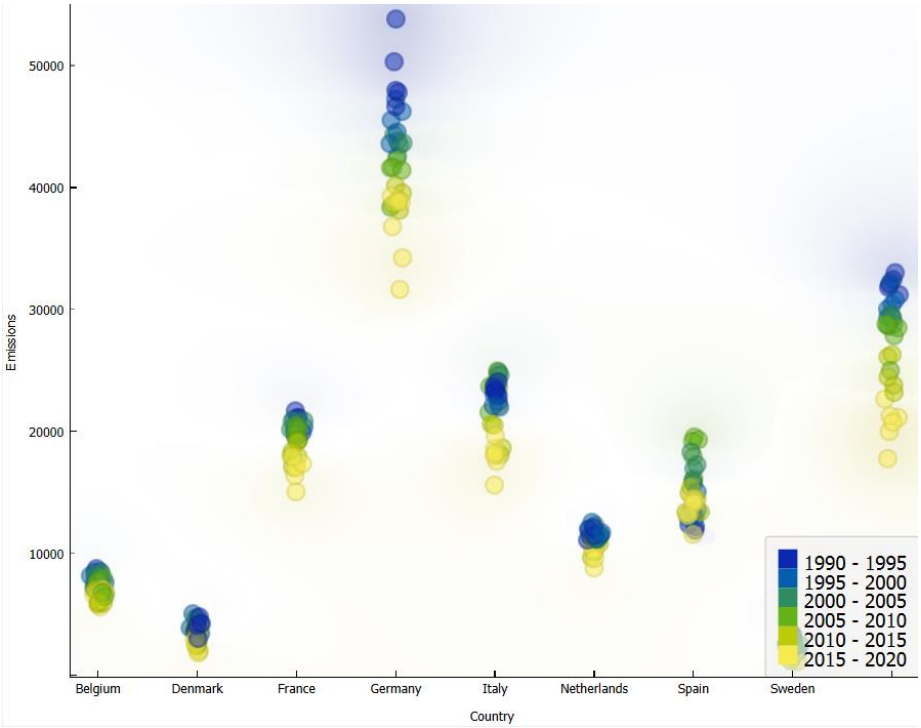
met de echte waarde.

SeaLevelShort	ral Network met l	Linear Regression	Random Forest	AdaBoost	Fold	Year
8.49	-26.5017	8.13701	7.85858	4.01	1	2016.0
-4.01	-28.7873	-2.9379	-2.41134	-3.16	1	2010.0
-15.83	-31.4416	-12.6357	-11.2651	-13.1	1	2007.0
-76.64	-60.3836	-77.6423	-79.8104	-79.37	1	2002.0
-125.82	-90.0111	-128.081	-118.856	-122.72	1	1994.0
-99.39	-70.6625	-99.108	-98.282	-98.06	1	1980.0
-73.91	-55.1597	-74.7002	-78.4209	-78.3	1	1973.0
-74.49	-53.7375	-72.673	-75.4677	-77.08	1	1971.0
-86.56	-56.2752	-80.3374	-81.4709	-79.37	1	1955.0

We kunnen wel zeggen dat om het zeeniveau accuraat te willen voorspellen dat we eigenlijk nog een kleinere error willen in sommige gevallen puur omdat er niet een heel groot uiteenlopend aantal meter in zeeniveau is.

2.3 Wat is je conclusie als je kijkt naar de uitstoot van landen over de jaren heen?

Als we kijken naar de uitstoot per land over de jaren heen zien we een heel positief iets.



We zien dat landen meer op hun uitstoot zijn beginnen letten met Duitsland en VK met de beste verbeteringen. Dit is natuurlijk ver van hoe het zou moeten zijn, maar we zien dat landen toch wel hun best doen om hun uitstoot te doen dalen, wat positief is.

2.4 Welke conclusie kan je trekken uit de classificatiemodellen met betrekking tot de uitstoot per land?

We hebben hiervoor een discrete waarde genomen als target namelijk het land. Hierdoor kunnen we classificatiemodellen gebruiken en gaan we de performantie er van bekijken. Ook hier passen we 60 procent training, 20 procent validatie en 20 procent test toe.

Model	AUC	CA	F1	Precision	Recall
kNN	0.937	0.836	0.831	0.850	0.836
Tree	0.877	0.782	0.767	0.787	0.782
Naive Bayes	0.881	0.455	0.407	0.402	0.455
Neural Network	0.886	0.418	0.409	0.485	0.418

Hieruit kunnen we afleiden dat kNN met in dit geval 4 buren het beste is omdat hier de hoogste classification accuracy heeft. We kunnen ook via de confusion matrix kijken waar de fouten gemaakt worden.

Predicted

	Belgium	Denmark	France	Germany	Italy	Netherlands	Spain	Sweden	United Kingdom (Convention)	Σ
Belgium	6	0	0	0	0	0	0	0	0	**6**
Denmark	2	4	0	0	0	0	0	0	0	**6**
France	0	0	6	0	0	0	0	0	0	**6**
Germany	0	0	0	6	0	0	0	0	0	**6**
Italy	0	0	3	0	2	0	0	0	1	**6**
Netherlands	0	0	0	0	0	6	1	0	0	**7**
Spain	0	0	0	0	0	1	5	0	0	**6**
Sweden	0	1	0	0	0	0	0	5	0	**6**
Convention)	0	0	0	2	1	0	0	0	3	**6**
Σ	**8**	**5**	**9**	**8**	**3**	**7**	**6**	**5**	**4**	**55**

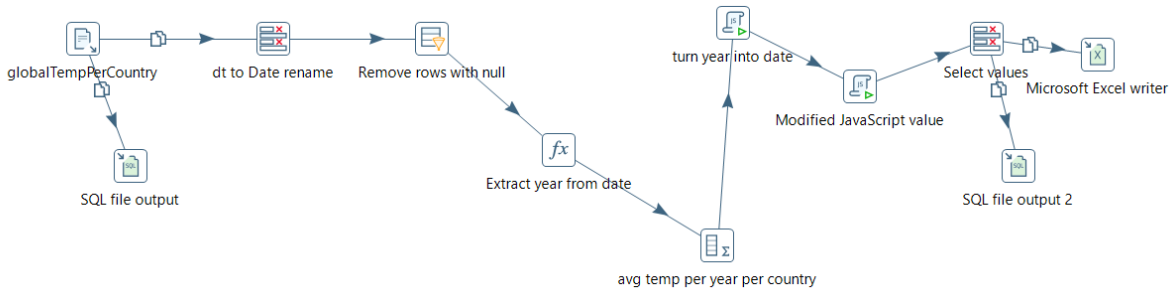
Hier kunnen we dus zien dat Frankrijk, Belgie en Duitsland de meeste foute landen voorspelt zijn. Nu kunnen we natuurlijk ook nadenken over waarom dit zo is. Eigenlijk is het redelijk logisch. Grote landen met dus meer uitstoot worden vaker verwisseld, analoog voor kleinere landen.

3. Stappen en tussentijdse resultaten.

3.1 Globale temperatuur per land

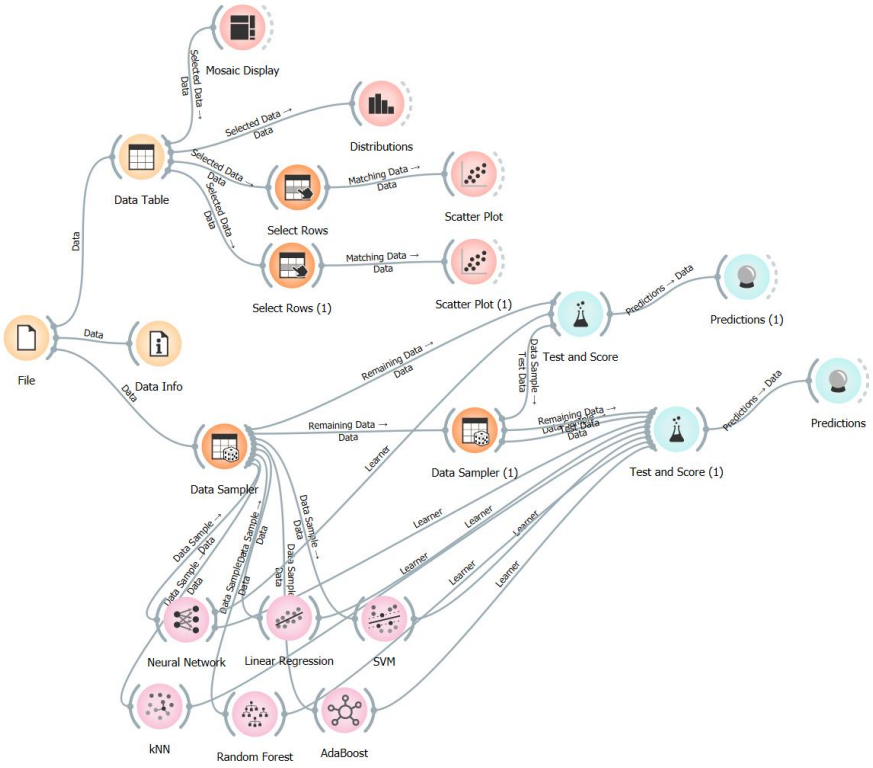
Ik ben gestart met data op te zoeken in ‘Kaggle’. Hier heb ik met de zoekterm klimaat een dataset gevonden met verschillende csv bestanden. Uit deze csv bestanden heb ik de globale temperatuur per land gepakt.

Na de csv op te slaan ben ik aan de slag gegaan in ‘Pentaho’ spoon. Pentaho is een tool om ETL (Extract, Transform, Load) mee uit te voeren. Hieronder ziet u mijn werkwijze die ik dan ook ga bespreken.



Ik ben dus gestart met de data in te laden via ‘CSV File Input’. In de volgende stap heb ik ‘Select values’ gebruikt om de namen van de variabelen te veranderen naar mooiere namen. Dan heb ik de rijen met null waardes verwijdered omdat deze niet nuttig zijn voor mijn berekening. De temperatuur per land was voor elke jaar ook elke maand. Ik vond dit een beetje overbodig dus heb ik het jaar gepakt via ‘Formule’. Omdat ik het jaar heb genomen kan ik nu een ‘Group by’ doen op het het land, dan het jaar en een average nemen van de temperaturen. Mijn twee volgende javascript stappen zijn gewoon om mijn jaartal een correct nummer te geven. Op het einde selecteer ik dan nog de values dat ik wil dat in mijn output file gestoken worden. Omdat onze data ook op de database moet komen hangt er bij de ruwe en de output data ook een ‘SQL output’ aan vast. Dit om een sql file te bekomen die ik kan laden in de database.

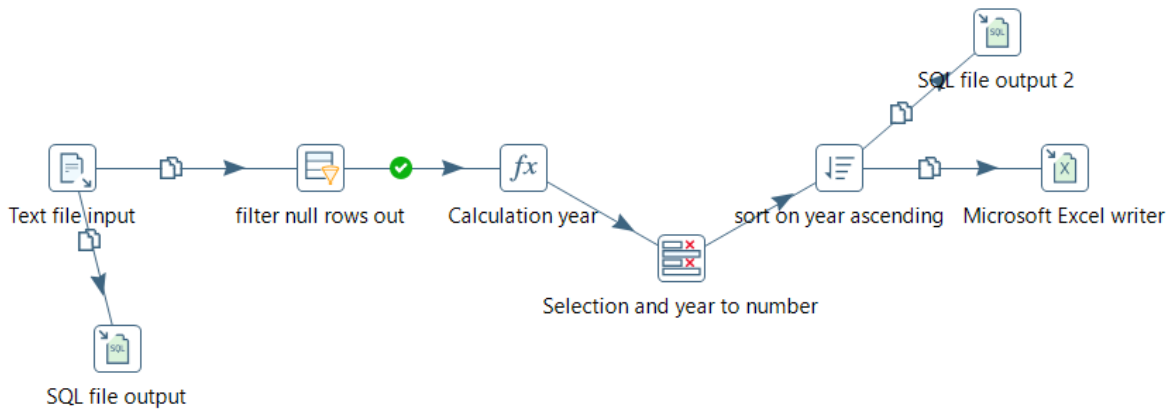
Daarna ben ik overgestapt naar het analyseren van de data via orange.



Hier ben ik eerst begonnen aan de verkenning van de data via mosaic display, scatter plots en distribution. Daarna heb ik een target value gekozen waarop ik mijn modellen wou maken. Ik moest regressie gebruiken omdat ik een continue variabele gekozen had als target. Hier heb ik dan ook mijn data eerst in 60 procent training gesplits en 40 procent validatie/test. En deze 40 procent heb ik dan nog eens door 2 gedeeld om 20 procent validatie en 20 procent test te bekomen. Dan heb ik naar de 'Test and Score' widget gekeken om de bepalen of er goede modellen waren of niet en welke dan de beste was.

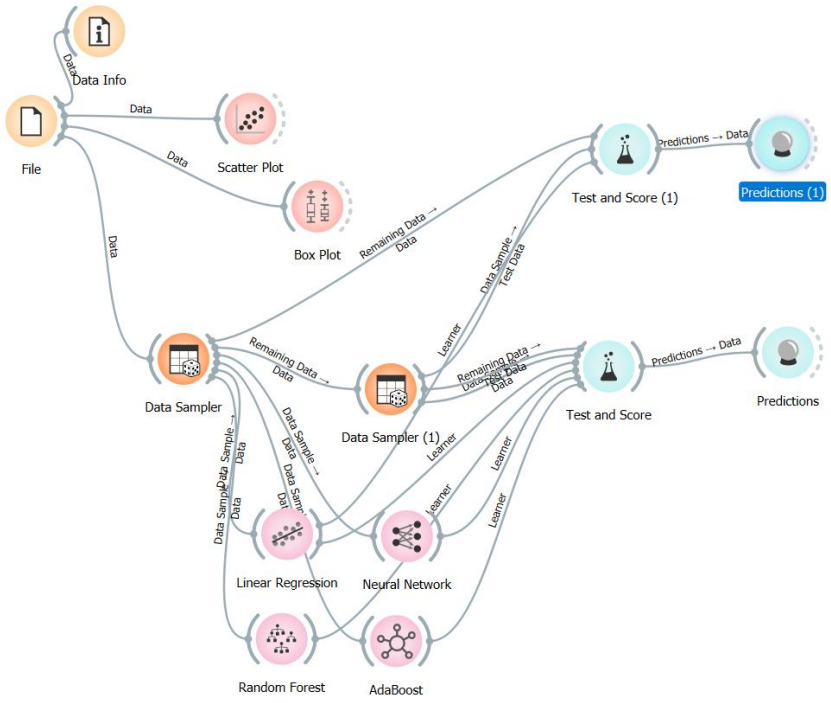
3.2 Zeeniveau

Deze data ben ik gaan zoeken op het internet en heb ik dus als txt bestand opgeslagen. Eerst heb ik de data in het tekst bestand moeten halen en dan nog een python script op moeten uitvoeren om de tabs naar ';' te veranderen want pentaho vond de tabs niet om de data te splitsen. Dan ben ik overgegaan naar Pentaho.



Eerst heb ik via de ‘Tekst file input’ mijn data ingeladen en gesplitst op ‘;’. Dan heb ik de null rows eruit gehaald zodat ik met propere data kon werken want maar één kolom bevatte dit. Dan heb ik weer het jaar berekent via ‘formula’. Dit moest omdat het jaar werd afgedrukt als het jaar vanaf het heden. Dus de laatste data van 2016 was 0, 2015 1 enzovoort. Dus heb ik dit naar de gewone datum omgezet. Dan heb ik deze datum accuraat gemaakt door decimale cijfers te verwijderen. Op het einde sorteer ik nog op jaar om zo eerste de oudste data te zien. Ook hier schrijf ik mijn output data weg in een excel bestand en heb ik ook de ruwe/output data in sql files gezet om op de databank te kunnen zetten.

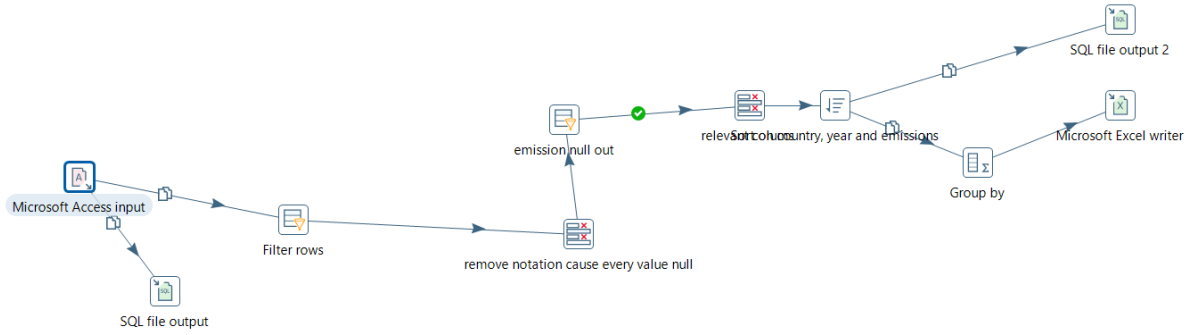
Hierna ben ik overgestapt naar Orange.



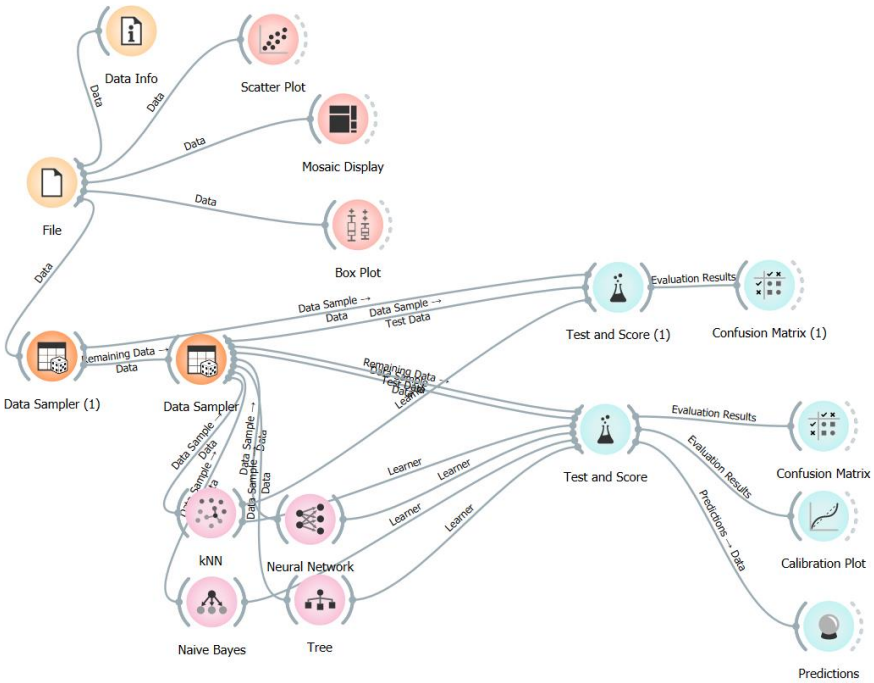
Hier ben ik ook weer begonnen met de exploratie met een scatterplot en een boxplot. Daarna heb ik ook weer de data in Training/Validatie/Test opgedeeld. Hier heb ik het waterniveau als target variable gepakt dus moesten we weer regressiemodellen gebruiken omdat die een continue variabele is. Op het einde heb ik dan ook weer ‘Test and Score’ gebruikt om te zien of ik goede modellen had en welke dan de beste was.

3.3 Uitstoot

Deze data ben ik ook gaan zoeken op het internet en ik kwam uit op een europese site met veel data over uitstoot. Dit omdat er veel conferenties zijn om de uitstoot te verlagen. Deze dataset was een accdb bestand (Microsoft Access). Met heel veel data in. Dan ben ik naar Pentaho gegaan om deze te verwerken.



Ik wou eerst de data van alle landen hebben maar kreeg op een gegeven moment een probleem de output data die over de Java stack ging of pentaho dat stopte met werken. Daarom heb ik ervoor gekozen om te filteren op bepaalde landen, namelijke België en omringende landen. Hier heb ik in de stap daarna een kolom met alleen maar nulwaarden eruit gehaald en alle uitstoot met null waarden er ook uitgehaald. Dan heb ik de relevante kolommen eruit gehaald met 'select Values' gevolgd door een sort op het land gevolgd door het jaar en dan de uitstoot. Hier had ik het probleem dat landen heel veel verschillende soorten uitstoot hadden en deze te veel waren voor orange dus heb ik voor de geselecteerde landen nog in een 'Group by' de gemiddelde uitstoor per land per jaar berekend en deze weggeschreven in een Excel bestand. Ook hier heb ik de ruwe en output data naar de database weggeschreven. Na dit ben ik verdergegaan in Orange.



Hier ben ik ook weer begonnen aan exploratie aan de hand van Scatter plot, mosaic displau en een box plot. Ook de data is weer verdeeld in Training/Validatie/Test. Wat verschilt van de andere twee analyses is dat ik hier country als target gekozen heb omdat dit een discrete variabele is en ik hier

dus classificatiemodellen kon gebruiken. Hier heb ik dus in ‘Test en Score’ gekeken naar CA om te zien welk model het beste presteerde om dan in de confusion matrix te kijken waar er dan wel nog fouten gemaakt werden.

4. Conclusie

Mijn conclusie van dit report is eerst en vooral dat er wel heel veel datasets te vinden zijn op het internet, maar dat er veel minder te vinden zijn die weldegelijk goed zijn. Het is ook spijtig dat pentaho het lastig heeft met heel grote hoeveelheid data omdat dit mijn werk wel verhinderd heeft. Ik heb na het analyseren van de data ook beseft dat het beter is om meer kolommen te hebben da degene die ik in elke analyse gebruikt heb omdat andere values interessante relaties kunnen toevoegen. Ik heb ook mijn pgadmin een lange periode aan de zijkant moeten laten draaien omdat het heel lang duurde voordat mijn data op de database gezet werd, vooral de ruwe data.

De conclusie over mijn onderwerp ‘Klimaat’ is dat de landen wel hun best doen om hun uitstoot naar beneden te brengen maar dat het nog lang niet genoeg is. Je ziet natuurlijk ook dat de gemiddelde temperatuur aan het stijgen is ookal stoten landen minder uit. Ik vermoed dat dit is omdat er natuurlijk een soort van ‘lag’ zit op het effect van de uitstoot op de temperatuur.

5. Bronnen

1. Deze bron bevat data van de temperatuur in landen in de wereld vanaf het jaar 1750. (csv)

   https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

2. Deze bron bevat het algemeen zeeniveau per jaar. (txt)

   https://www1.ncdc.noaa.gov/pub/data/paleo/contributions_by_author/spratt2016/spratt2016.txt

3. Deze data bevat per land in de EU plus het VK de uitstoot per jaar. (accdb)

   https://www.eea.europa.eu/data-and-maps/data/national-emissions-reported-to-the-unfccc-and-to-the-eu-greenhouse-gas-monitoring-mechanism-18