Det egna projektet, steg 1

Rikard Lang

Projektplan

Mitt problem

Det här var lite utav en utmaning måste jag erkänna, det finns så många intressanta dataset att titta på, vrida och vända, analysera och sedan låta en ML-modell se mönster som jag själv förmodligen inte skulle se lika lätt. Så att hitta data har inte varit något problem, det svåra har varit att hitta ett intressant problem där det också finns tillräckligt med data; för sig eller att jag kan göra en egen sammanställning.

Det mest intressanta dataset jag hittade var följande: https://riksarkivet.se/psidata/

Flyghaverier

Datum	Förband	F	F	F	F	N	N	H	C	E	A	C	F	C	C	F	C	Sammanfattning	H	K
1912-07-14	0	C	E	C	2	C	C	n	C	1	C	K	n	nej				Motorstörning i starten	A	F

flygvapenhaverier Det är en sammanställning över alla militära flyghaverier i Sverige under åren 1914 till 2007. För ett exempel på hur det kan se ut, se nedan.

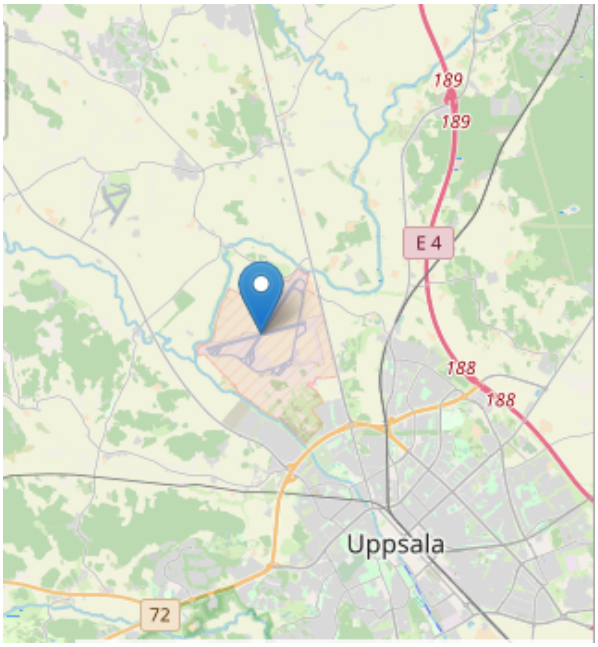
Problemet var att jag inte hittade något intressant problem som jag kunde överlåta till en maskininlärningsmodell. Intressant data men inget intressant problem.

Förutspå väder

Ett intressant och möjligen aktuellt problem är ju huruvida det kommer bli en vit jul i år, eller vilket år som helst i framtiden. Jag är ju novis när det kommer till väderdata men mitt arbete med den första inlämningsuppgiften lära mig en del om väderdata. Så jag bestämde mig för att mitt problem blir att förutse temperaturen och nederbörden baserat på historisk data.

Data

Jag bor 42 km nordost om Uppsala, varför jag specificerade det i kilometer och så exakt överlåter jag till läsaren att fundera på. Efter lite bläddrande på SMHIs tjänst för öppna data så hittade jag väderdata från en station som befinner sig på Ärna flygplats.



Alltså F16, det militära flygfältet som relativt nyligen återgått till att vara en aktiv flygflottilj. Anledningen till att jag valde denna väderstation är dels den relativt geografiska närheten till mig själv men att det är en flygflottilj gör ju det hela mer spännande (se datat om flyghaverier ovan) samt att till min förvåning så har stationen samlat in temperaturdata under en väldigt lång tid.

En närmare titt på datamängden - temperaturer

Den första posten är daterad 1944-07-01 och till en början är det tre mätningar per dygn.

Datum	Tid (UTC)	Lufttemperatur	Kvalitet
1944-07-01	07:00:00	16.2	G
1944-07-01	13:00:00	23.0	G
1944-07-01	18:00:00	20.2	G

Datum och tid angiven i UTC är trevligt. Sedan anges temperaturen i decimalform. Kvalitet har två olika märkningar: G står för ett kontrollerat värde medan Y står för ett misstänkt eller möjligen aggregerat värde. Y kan också betyda att det är grovt kontrollerat arkivdata eller okontrollerade realtidsdata.

1949-01-01	00:00:00	1.8	Y
1949-01-01	06:00:00	2.2	Y
1949-01-01	12:00:00	2.8	Y
1949-01-01	18:00:00	2.0	Y

1949-01-01 börjar man med att mäta fyra gånger per dygn. 1952 är det åtta mätningar per dygn och från och med 1962-01-01 så mäts temperaturen varje timme.

1962-01-01	00:00:00	0.8	Y
1962-01-01	01:00:00	0.4	Y
1962-01-01	02:00:00	-0.3	Y
1962-01-01	03:00:00	-1.1	Y
1962-01-01	04:00:00	-2.2	Y
1962-01-01	05:00:00	-2.3	Y
1962-01-01	06:00:00	-2.1	Y
1962-01-01	07:00:00	-1.1	Y
1962-01-01	08:00:00	-0.3	Y
1962-01-01	09:00:00	-0.6	Y
1962-01-01	10:00:00	-0.8	Y

Den sista posten i min datamängd är för idag. Nyare data kan laddas ned manuellt och läggas till i

2022-12-14	11:00:00	-10.2	G

träningsdata alternativt så kan ett öppet API användas för att maskinellt fylla på med träningsdata. Några reflektioner kring datat. Då datat sträcker sig över väldigt lång tid så kan det vara intressant att titta på hur medeltemperaturer och liknande förändrats över tid. Det är något att i åtanke när ML-modellen ska förutspå temperaturen, kommer så gammal data att påverka den negativt. Det får vi se sen när jag experimenterar i en Jupyter Notebook.

Den andra datamängden - nederbörd

2013-01-01	00:00:00	0.0	Y

Nederbörden har inte alls lika lång historik men som jag nämnde ovan så kanske det inte är helt relevant för modellen. Datamängden börjar 2013-01-01 och mäter mängden nederbörd per timme och den mäter i millimeter. I övrigt är datum, tid och kvalitet samma märkning som temperaturdatan.

2022-12-13	00:00:00	0.5	G
2022-12-13	01:00:00	0.6	G
2022-12-13	02:00:00	1.0	G
2022-12-13	03:00:00	2.0	G
2022-12-13	04:00:00	2.0	G
2022-12-13	05:00:00	1.0	G
2022-12-13	06:00:00	0.9	G
2022-12-13	07:00:00	0.9	G
2022-12-13	08:00:00	0.3	G

Här ser vi till exempel nederbördsdata 2022-12-13 mellan 00:00 till och med 08:00.

Det går att komplettera datat om stationen stödjer fler alternativa mätningar. De parametrar som finns i stort att välja på presenteras nedan.

Lufttemperatur (h)	Lufttemperatur (dygn)	Lufttemperatur (månad)
Lufttemperatur, min och max (12h)	Lufttemperatur, min och max (dygn)	Daggpunktstemperatur (h)
Nederbördsmängd (15 min)	Nederbördsmängd (h) ✔	Nederbördsmängd (dygn)
Nederbördsmängd (månad)	Nederbördsintensitet (15 min)	Nederbördsintensitet, max av medel (15 min)
Nederbördstyp (12h)	Nederbördstyp (dygn)	Snödjup och markytans tillstånd (dygn)
Relativ luftfuktighet (h)	Vindriktning och vindhastighet (h)	Vindhastighet, max av medel (h)
Byvind, max (h)	Total molnmängd (h)	Signifikanta moln (h)
Lägsta molnbas (h)	Lägsta molnbas, min (15 min)	Solskenstid (h)
Globalstrålning (h)	Långvågsstrålning (h)	Lufttryck (h)
Sikt (h)	Rådande väder (h)	

Plan framåt

Skapa en Jupyter Notebook där jag behandlar ovanstående data och möjligen kompletterar med fler datakällor om så behövs. Implementera och träna en ML-modell som kan förutspå vädret (temperatur och nederbörd). Detta är ett regressionsproblem och jag kommer initialt använda mig av supervised learning för att se hur väl modellen kan förutspå temperatur och nederbörd. Datat är redan uppmärkt korrekt. Jag kommer använda mig av Pandas, Numpy och Scikit-learn - kommer förmodligen att vilja visualisera lite data med seaborn också.