PROJECTE FINAL – CIÈNCIA DE DADES

IT ACADEMY – BARCELONA ACTIVA

Anna Masó Bagué

Novembre 2022

1. PRESENTACIÓ DEL CONJUNT DE DADES ESCOLLIT

He escollit un data set que aporta cinc anys de dades sobre el temps de Calcuta (Índia) entre els anys 2017 i 2021[1].

El meu interès des de petita per la geografia i la llicenciatura que tinc en la matèria, fan que tots els temes relacionats sempre m'interessin i busqui diferents punts des d'on estudiar-los.

El data set en concret, el vaig trobar interessant perquè té més informació de la normal respecte a les que es solen trobar (temperatura i precipitació), i a més era actual i d'una zona que en desconeixia aquest tipus d'informació, per tant era un bon data set per practicar el curs, per aprendre i per incrementar coneixements meteorològics d'un lloc que no tenia fins ara.

[1] Font del dataset: https://www.kaggle.com/datasets/kafkarps/five-years-weather-data-of-kolkata

2. CARACTERÍSTIQUES GENERALS

Ens trobem davant d'un data set compost per 1826 files i 25 columnes. Les columnes, que a partir d'ara les anomenarem variables, fan referència a les diferents variables meteorològiques d'entre els anys 2017 i el 2021 de la ciutat de Calcuta.

D'aquestes 25 variables, 6 són categòriques i la resta numèriques (totes float64). D'aquestes, n'eliminaré alguna que no conté dades, o que la dada sempre és la mateixa (per exemple, el lloc), però per altra banda, n'afegiré tres, ja que divideixo la data en dia, mes i any en columnes diferents.

```
RangeIndex: 1826 entries, 0 to 1825
Data columns (total 25 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Address              1826 non-null   object
 1   Date time            1826 non-null   object
 2   Minimum Temperature  1826 non-null   float64
 3   Maximum Temperature  1826 non-null   float64
 4   Temperature          1826 non-null   float64
 5   Dew Point            1826 non-null   float64
 6   Relative Humidity    1826 non-null   float64
 7   Heat Index           1538 non-null   float64
 8   Wind Speed           1826 non-null   float64
 9   Wind Gust            54 non-null     float64
 10  Wind Direction       1826 non-null   float64
 11  Wind Chill           13 non-null     float64
 12  Precipitation        1826 non-null   float64
 13  Precipitation Cover  1826 non-null   float64
 14  Snow Depth           1826 non-null   float64
 15  Visibility           1826 non-null   float64
 16  Cloud Cover          1826 non-null   float64
 17  Sea Level Pressure   1825 non-null   float64
 18  Weather Type         1825 non-null   object
 19  Latitude             1826 non-null   float64
 20  Longitude            1826 non-null   float64
 21  Resolved Address     1826 non-null   object
 22  Name                 1826 non-null   object
 23  Info                 0 non-null      float64
 24  Conditions           1826 non-null   object
dtypes: float64(19), object(6)
memory usage: 356.8+ KB
```

La importància de les variables meteorològiques per a les persones ve donada per l'afectació al nostre dia a dia, però moltes de les que no ens afecten tan directament són tant o més importants.

Al grup de els primeres, hi podem trobar la temperatura, la pluja o el vent, i en el grup de les segones podem trobar la cobertura de núvols, el punt de rosada o la direcció del vent. Per exemple, a més temperatura, més calor tenim i més afectació directa, o si fa vent, ens hem d'abrigar més, pot provocar caigudes d'arbres, etc. Però i la direcció del vent? Per exemple a casa sabem que perquè plogui durant 3 o 4 dies seguits i de forma abundant, la borrasca ha de venir de llevant (del mar), en canvi, si ve del nord xoca amb moltes muntanyes abans no arriba a nosaltres i l'afectació és menor o nul·la. Ser de família de pagesos t'ensenya moltes coses...

Amb tot això vull dir que hagués pogut fer un estudi molt més profund de les dades, però he preferit centrar-me en les més "conegudes" ja que per desconegut ja tenia Calcuta en general, i encara més la seva meteorologia.

3. DEFINICIÓ DE LES VARIABLES

Centrant-nos més en les variables del data set, com hem dit en tenim 6 de categòriques i la resta són numèriques.

De les categòriques he utilitzat bàsicament la de la data (Date_Time) i la de les condicions (Conditions). La primera m'ha servit per bàsicament per les línies temporals i la segona l'he trobat interessant per veure quin tipus de dia tenen més sovint (si ennuvolat, serè, plujós, etc).

De les variables numèriques, a part de fer histogrames i correlacions de totes (excepte la neu, que l'he eliminat perquè és un fenòmen excepcional en aquesta ciutat, per no dir nul), les que més he utilitzat han sigut les tres variables referents a la temperatura (mínima, màxima i mitjana) i la pluja. Són les més fàcils per començar a estudiar temes meteorològics d'un lloc desconegut i les més agraïdes a l'hora de treure'n resultats generals.

A la següent taula podem veure les dades bàsiques de les variables numèriques:

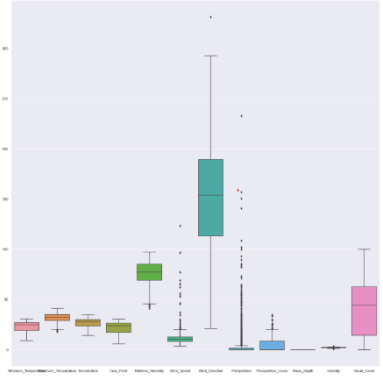
	Minimum Temperature	Maximum Temperature	Temperature	Dew Point	Relative Humidity	Heat Index	Wind Speed	Wind Gust	Wind Direction	Wind Chill	Precipitation	Precipitation Cover	Snow Depth	Visibility	Cloud Cover	Level Pressure	Latitude	Longitude	Info
count	1826.000000	1826.000000	1826.000000	1826.000000	1826.000000	1538.000000	1826.000000	54.000000	1826.000000	13.000000	1826.000000	1826.000000	1826.0	1826.000000	1826.000000	1825.000000	1.826000e+03	1.826000e+03	0.0
mean	22.611884	31.460022	26.628697	21.713527	76.602141	40.740182	11.097974	30.938889	151.183899	8.730769	4.650783	4.180986	0.0	2.019496	40.459693	1007.906411	2.257050e+01	8.837130e+01	NaN
std	5.300028	4.195294	4.473758	5.628949	10.552180	7.871842	6.294325	6.461772	51.267895	0.712255	13.994448	6.317213	0.0	0.316475	27.677867	6.050940	9.168512e-13	9.808176e-13	NaN
min	9.000000	18.000000	14.200000	5.800000	41.150000	26.300000	3.400000	20.800000	20.950000	7.600000	0.000000	0.000000	0.0	0.700000	0.000000	988.400000	2.257050e+01	8.837130e+01	NaN
25%	19.000000	29.000000	23.600000	17.100000	69.145000	34.100000	8.100000	26.725000	113.330000	7.900000	0.000000	0.000000	0.0	1.800000	14.325000	1003.100000	2.257050e+01	8.837130e+01	NaN
50%	24.800000	32.000000	28.200000	23.900000	77.410000	42.100000	10.300000	28.900000	153.810000	8.800000	0.000000	0.000000	0.0	2.100000	44.300000	1008.100000	2.257050e+01	8.837130e+01	NaN
75%	27.000000	35.000000	30.000000	26.500000	85.037500	46.700000	12.800000	34.400000	189.450000	9.100000	1.630000	8.330000	0.0	2.200000	62.600000	1013.100000	2.257050e+01	8.837130e+01	NaN
max	30.500000	41.000000	34.800000	30.100000	97.230000	61.200000	123.300000	55.300000	331.380000	9.600000	232.770000	34.780000	0.0	3.200000	100.000000	1019.700000	2.257050e+01	8.837130e+01	NaN

4. PRESENTACIÓ DELS OBJECTIUS:

Com he dit anteriorment, el meu objectiu principal era conèixer una mica més sobre el temps meteorològic de la ciutat de Calcuta, en aquest cas utilitzant les tècniques que he après durant el curs de ciència de dades.

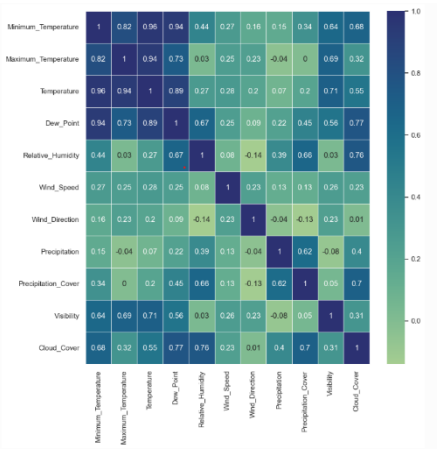
Així doncs:

- La informació bàsica del data set m'ha ajudat a entendre com era aquest i detectar amb quines variables volia treballar i amb quines no i preparar el data set per, a continuació, estudiar-lo més fàcilment.



- Fer diferents gràfics m'han permès visualitzar les dades i reafirmar que a

primer cop d'ull "val més una imatge que mil paraules". M'agrada molt per exemple el gràfic "Relació entre les variables meteorològiques i les condicions meteorològiques" i el boxplot per observar els outliers.



- Les taules de correlacions ajuden a interpretar molt ràpidament quines variables tenen correlacions més altes i quines altres van per lliure

- I tota la part dels algoritmes, m'ha permès veure com fer prediccions tant de variables numèriques com de categòriques, veure els percentatges d'error, estandarditzar les dades per a fer-les més mal·leables, convertir els algoritmes en plots (PCA 2 components), etc.

Per últim, vull deixar el gràfic següent, el qual a simple vista en podem treure dos raonaments, l'època monsònica coincideix amb els mesos de més temperatura (junt a octubre), però es podria percebre un lleuger canvi, amb més mesos de pluja, però amb pluja més intermitent però molt més abrupta, és a dir, una tendència de canvi a episodis més curts però més intensos de pluja.