

Plourà demà?

Predictió de pluja a Austràlia

Nil Farrés Soler



Abstract

This study explores the application of XGBClassifier for daily rainfall prediction in Australia, based on 10 years of meteorological observations (from 2007 to 2017). Using the AUPRC metric from the PR (Precision-Recall) curve as a key indicator, the model demonstrates an accuracy of 86%, with a notable ability to predict non-rainy days (precision of 88% and recall of 95%) and moderate performance for rainy days (precision of 75% and recall of 56%). The results provide an analysis of the importance of features in predicting the meteorological phenomenon, where wind direction, evaporation, humidity, and sunlight are the most influential variables for the model. The wind directions of southwest and west-northwest (dominant in Australia's wet season) at 3 pm appear to be related to increased humidity and decreased atmospheric pressure, two key indicators of the likelihood of rain the following day.

Keywords

XGBClassifier, rainfall prediction, Australia, meteorological observations, AUPRC metric, Precision-Recall (PR) curve, feature importance, wind direction, humidity, atmospheric pressure.

1. INTRODUCCIÓ

La capacitat de predir esdeveniments meteorològics ha sigut una fita persistent per a la ciència i la tecnologia, amb implicacions en l'agricultura, la seguretat i la planificació urbana. A Austràlia, on les condicions climàtiques extremes són freqüents, la precisió de les prediccions meteorològiques té una importància encara més gran. Aquest treball se centra en la predició de pluja a curt termini, una tasca complexa donat el sistema climàtic i els molts valors faltants que conté el dataset. El conjunt de dades recopila 10 anys d'observacions meteorològiques (del 2007 al 2017) mitjançant una àmplia gamma de variables (temperatura, precipitació, direcció del vent, humitat, pressió atmosfèrica, etc.), on els dies sense pluja (78%) superen els dies amb pluja (21%).

2. METODOLOGIA

Inicialment, s'ha realitzat una visualització de les dades, revelant un conjunt de dades notablement desequilibrat. Això ha motivat a fer un tractament meticulos en la fase de neteja i preprocessament. En aquesta etapa, s'han eliminat les files de les variables amb un percentatge molt baix de valors faltants (inferior al 3%, com és el cas de, per exemple, la variable *Rainfall*). També s'han creat tres columnes noves a partir de la columna *Date* (*Year*, *Month* i *Season*) tenint en compte que les estacions a Austràlia són diferents a les de l'hemicferi nord. Per enriquir l'anàlisi, s'han creat noves variables, com el promig de pluges de la setmana i del mes anterior, i la quantitat de pluja en els últims tres dies, oferint una perspectiva més completa i permetent la consideració de sèries temporals en el modelatge del conjunt de dades meteorològiques. Aquestes variables ajudaran al model a fer una millor predició. Finalment, s'ha utilitzat l'estrategia KNNImputer per a la imputació i, per a les variables categòriques, s'ha aplicat el mètode d'encoding Target Encoder. També s'ha realitzat una normalització de les dades mitjançant Standard Scaler.

Un cop netejat el dataset, s'ha seleccionat la mètrica més adequada per al model, triant l'AUPRC de la corba PR (Precision-Recall) per la seva alta sensibilitat en la detecció de casos positius. Després d'analitzar diferents models, incloent Random Forest, XGBClassifier, LGBMClassifier i AdaBoostClassifier (que apliquen gradient boosting), i KNeighborsClassifier, s'ha determinar que XGBClassifier ofereix el millor rendiment, establint-lo com l'eina principal per a les prediccions d'aquest treball.

3. RESULTATS I ANÀLISI

L'anàlisi de les dades netejades i preprocessades mitjançant XGBClassifier revela resultats significatius. El model aconsegueix una exactitud global del 86%, destacant-se en la predició de dies sense pluja amb una precisió del 88% i un recall del 95%. En comparació, la predició de dies amb pluja és més complexa, amb una precisió del 75% i un recall més baix del 56%.

	Precision	Recall	F1-Score	Support
False	0.88	0.95	0.91	20525
True	0.75	0.56	0.64	5815
Accuracy			0.86	26340
Macro avg	0.82	0.75	0.78	26340
Weighted avg	0.85	0.86	0.85	26340

Taula 1. Informe de classificació del model final

L'aplicació de l'AUPRC com a mètrica principal resulta en una puntuació de 0.75. Això reflecteix un equilibri adequat entre precisió i recall (especialment important en un conjunt de dades meteorològiques desequilibrat) i indica que el model és fiable en la detecció de la majoria de dies sense pluja, però encara hi ha marge de millora en la predicció precisa dels dies plujosos.

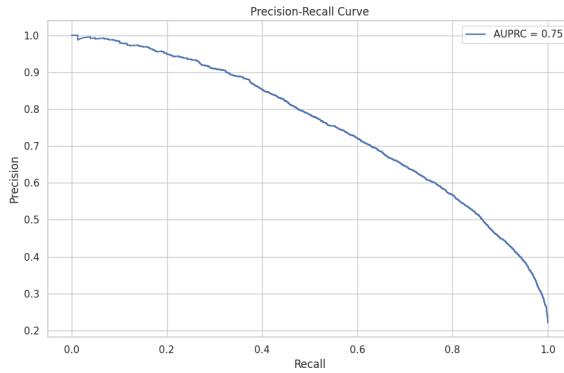


Fig. 1. Gràfic de la corba PR (Precision-Recall) amb un AUPRC de 0.75.

Podem observar que el model tendeix a tenir una alta confiança en les seves prediccions per a la classe negativa (la majoria de les probabilitats són baixes per la classe predita 0) i està menys segur en les seves prediccions de la classe positiva, amb les probabilitats esteses més uniformement.

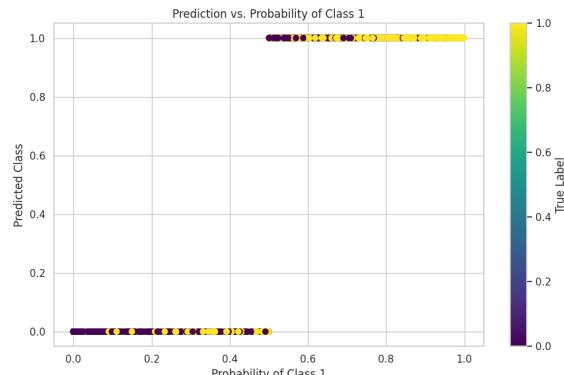


Fig. 2. Gràfic de probabilitat vs Classe predita del model final.

També s'ha realitzat una selecció d'un llindar òptim de 0.31, equilibrant precisió i sensibilitat. Aquest llindar acaba resultant en una millora substancial en la identificació de dies plujosos, amb una capacitat notable del model per mantenir baixos els falsos positius. La matriu de confusió amb aquest llindar mostra 19454 veritables negatius i 3228 veritables positius, amb els falsos negatius i falsos positius controlats a 2587 i 1071, respectivament. Aquests resultats confirmen l'eficàcia del model per predir dies no plujosos i destaquen l'oportunitat de millorar la detecció de dies plujosos.

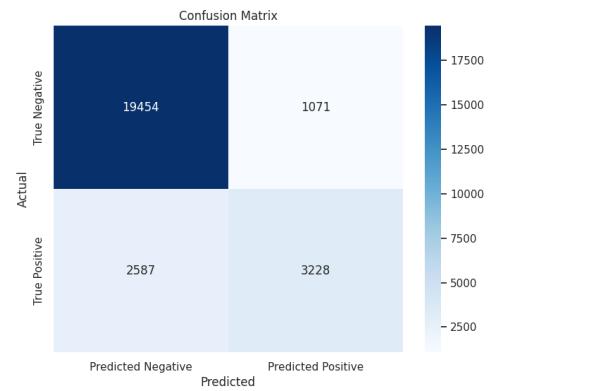


Fig. 3. Matriu de confusió del model final.

Addicionalment, l'anàlisi de la importància de les característiques ha revelat que la velocitat del vent a les 3PM (*WindDir3pm*) és el predictor més influent en el model, així com l'evaporació (*Evaporation*), la humitat a les 3PM (*Humidity3pm*), i la llum solar (*Sunshine*) són també de les més influents en la predicció de pluja per al dia següent.

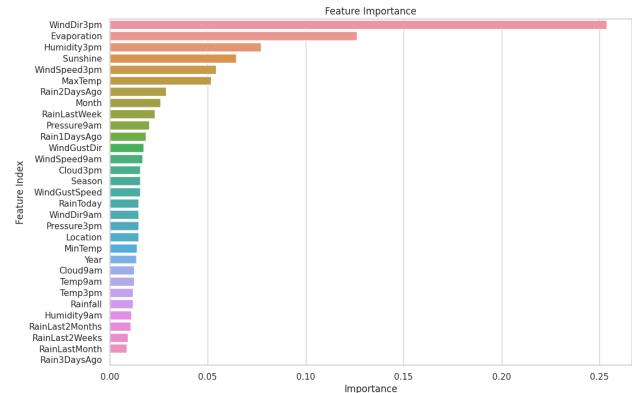


Fig. 4. Gràfic de la importància de les característiques.

Però per què precisament la velocitat del vent a les 3 de la tarda és la variable amb més importància pel model?

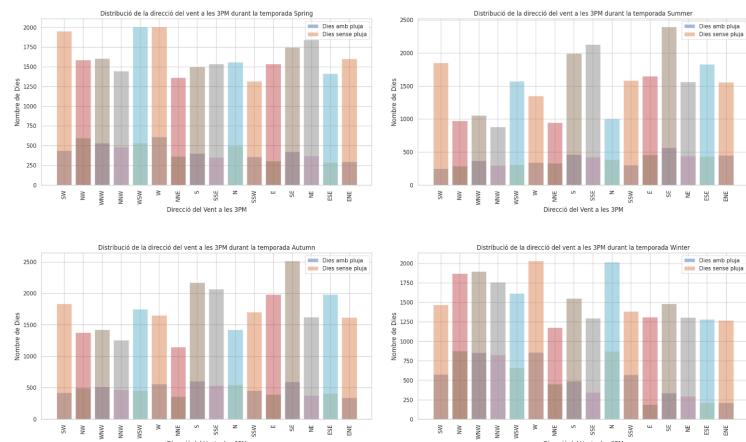


Fig. 5. Histogrames de distribució de *WindDir3pm* per a cada estació.

Els 4 histogrames apilats anteriors representen la distribució de la direcció del vent a les 3 PM (*WindDir3pm*) per a quatre estacions diferents: primavera, estiu, tardor i hivern. Cal tenir en compte que a Austràlia les estacions són diferents a les de l'hemisferi nord (l'estiu és de desembre a febrer, la tardor de març a maig, l'hivern de juny a agost i la primavera de setembre a novembre). Per a cada estació, els dies són separats en dos grups: els dies que van tenir pluja el dia següent (*Dies amb pluja*) i els dies que no (*Dies sense pluja*).

Es pot observar que cada estació mostra un patró únic, el que suggerix que la relació entre la direcció del vent i la precipitació pot variar segons la temporada. Durant la primavera hi ha algunes direccions del vent que sobresurten en els dies amb pluja, com NNE (nord-nord-est), W (oest), SW (sud-oest) i WNW (oest-nord-oest), que podrien estar associades amb un augment de la probabilitat de pluja. A l'estiu, en canvi, direccions del vent com E, SE i WNW també mostren una certa predominància en els dies amb pluja. A la tardor, S i SSE mostren un nombre més alt de dies amb pluja en comparació amb altres direccions i la distribució per als dies sense pluja és més uniforme entre les diferents direccions. Finalment, a l'hivern W i SW mostren una major freqüència en dies amb pluja. Hi ha una notable diferència en la freqüència de dies amb pluja per a la direcció NNE, que és molt més baixa en comparació amb altres estacions.

Cal destacar que, en el dataset utilitzat per aquest treball, l'estiu és l'època més plujosa en termes de volum total de pluja, mentre que l'hivern té el major nombre de dies plujosos. Això és consistent amb les estacions més plujoses a Austràlia, que tot i dependre de la regió, les pluges acostumen a ser molt més abundants a finals de primavera i durant l'estiu, que coincideix amb l'època humida.

També s'han realitzat diagrames de caixa amb cadascuna de les direccions del vent respecte la humitat i la pressió a les 3 de la tarda. Aquí es mostren els més rellevants.

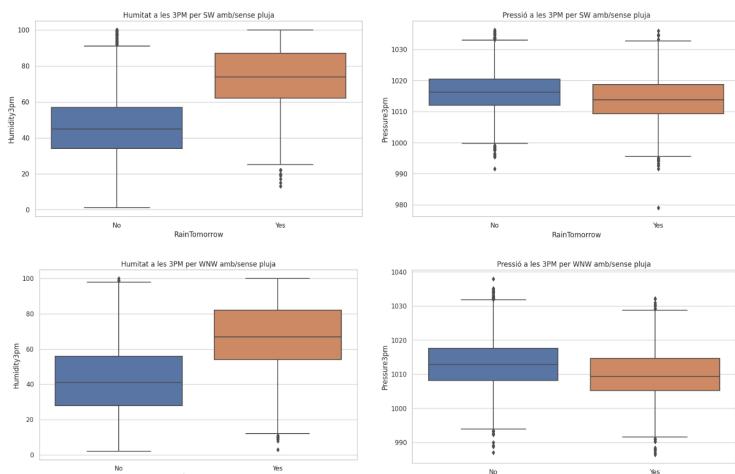


Fig. 6. Diagrames de caixa d'humitat i pressió per *WindDir3pm* rellevants: SW i WNW.

Pel que fa a la Humitat a les 3 de la tarda, les caixes per als dies amb pluja presenten valors més alts de humitat, indicant que una major humitat per la tarda sembla estar associada amb una major probabilitat de pluja el dia següent. A més, la dispersió (la

distància entre el quartil inferior i el superior) és lleugerament més gran en dies que segueixen amb pluja, especialment quan el vent és WNW.

Pel que fa a la Pressió a les 3 de la tarda, les caixes que representen la pressió en dies sense pluja mostren valors més alts de pressió en comparació amb els dies que segueixen amb pluja, indicant que una pressió més baixa per la tarda podria ser un indicador de pluja el dia següent. La dispersió de la pressió per a dies amb pluja és considerable, especialment per al vent WNW, suggerint una variabilitat més gran en la pressió en dies que segueixen amb pluja per aquesta direcció del vent.

Cal destacar que els vents del sud-oest (SW) i oest-nord-oest (WNW) a Austràlia soLEN estàr associats amb l'arribada de sistemes frontals que porten aire més fred i humit des de l'oceà, cosa que és consistent amb el model de predicció de pluja d'aquest treball. A més a més, aquests vents coincideixen en ser els que mostren una certa predominància en els dies amb pluja durant l'estiu, l'estació humida.

4. CONCLUSIONS

El model XGBClassifier ha demostrat un rendiment notable en la tasca de predicció de pluja, amb una exactitud general de l'86%. Això indica que en la majoria dels casos, el model pot distingir correctament entre dies de pluja i dies sense pluja.

El model tendeix a tenir un millor rendiment a l'hora de predir la classe negativa (dies sense pluja), com ho demostren la precisió del 88% i el recall del 95%. La classe positiva (dies amb pluja), en canvi, té una precisió del 75% i un recall més baix del 56%, resultant en una F1-Score del 64%. Això és probablement a causa de la gran quantitat de dies no plujosos del dataset (78%) enfront dels dies plujosos (21%).

La corba PR (Precision-Recall) i la seva AUPRC de 0.75 proporcionen una visió més matissada del rendiment del model, especialment en el context d'un conjunt de dades desequilibrat. La corba mostra la compensació entre precisió i recall a diferents líndars, ajudant a identificar un líndar òptim per a les prediccions del model.

L'anàlisi de la matriu de confusió confirma les tendències observades en les mètriques de rendiment, mostrant una quantitat significativa de falsos negatius, el que reafirma la dificultat del model per capturar tots els dies de pluja reals.

L'anàlisi de la importància de les característiques revela que La direcció del vent a les 3PM, l'evaporació, la humitat a les 3PM i la llum solar són les variables que tenen més influència en les decisions del model. Les direccions del vent SW i WNW a les 3PM semblen tenir relació amb l'augment de la humitat i la disminució de la pressió atmosfèrica, dos indicadors clau de la probabilitat de pluja al dia següent. De fet, coincideixen en ser els que mostren una certa predominància en els dies amb pluja durant l'estació humida a Austràlia (de finals de primavera a finals d'estiu).

5. BIBLIOGRAFIA

[1] Kaggle (Dataset utilitzat). Weather Dataset - Rattle Package.
Recuperat de
<https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package>.

[2] Bushman Tanks. 100 Years of Australian Rainfall. Recuperat de
<https://www.bushmantanks.com.au/100-years-of-australian-rainfall/>.

[3] Petheram, C., McMahon, T.A., Peel, M.C. Updated world map of the Köppen-Geiger climate classification. Hydrology and Earth System Sciences. Recuperat de
<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2007WR006373>.