Predicció de precipitacions: un estudi sobre la pluja a Austràlia

Youssef Cahouach Guella Ikhlaf

Abstract— Aquest treball presenta una anàlisi detallada d'un conjunt de dades meteorològiques amb l'objectiu de predir si plourà demà en diverses localitats d'Austràlia. Mitjançant un procés metòdic que inclou la neteja de dades, la transformació de variables, l'aplicació d'anàlisis de correlació i PCA, i finalment, l'entrenament i l'avaluació de models de classificació, hem aconseguit resultats prometedors. El model més potent identificat és XGBoost, que supera la Regressió Logística i KNN en totes les mètriques, assenyalant-se com el model més potent dels tres.

Keywords— Anàlisi de dades, Predicció meteorològica, Regressió Logística, KNearestNeighbours (KNN), XGBoost, Anàlisi de Components Principals (PCA), Correlació, Validació creuada StratifiedKFold, Optimització de paràmetres, Avaluació de models

1 Introducció

La predicció del temps ha estat sempre un element clau en la nostra vida quotidiana. Des de planificar les nostres activitats diàries fins a prendre decisions crítiques en sectors com l'agricultura, la construcció i fins i tot l'aviació, la capacitat de preveure les condicions meteorològiques té una gran importància.

La motivació d'aquest treball rau en la comprensió de com es realitzen aquestes prediccions en la vida real com es presenten a les notícies que veiem dia a dia.

El context d'aquest estudi és el 'dataset' "Rain in Australia" de Kaggle. Aquest 'dataset' proporciona un conjunt de dades exhaustives que recull observacions meteorològiques diàries de diverses localitzacions d'Austràlia, que es troba en un clima particularment variable.

En aquest treball, ens centrarem en l'anàlisi d'aquestes dades, amb l'objectiu de desenvolupar un model predictiu fiable per a la pluja.

Aquest conjunt de dades recopila observacions meteorològiques de diverses ciutats australianes, des del desembre de 2007 fins al juny de 2017. Inclou dades sobre les temperatures màximes i mínimes, així com la velocitat i direcció del vent a les 9 del matí i a les 3 de la tarda. A més, proporciona informació sobre la humitat i la pressió atmosfèrica en aquests mateixos intervals de temps. També es registren les temperatures i la cobertura de núvols a les 9 del matí i a les 3 de la tarda.

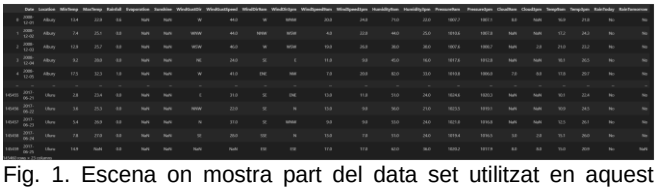

Fig. 1. Escena on mostra part del data set utilitzat en aquest projecte, on es pot observar les columnes i el tipus de dada.

Abans d'abordar l'anàlisi de les dades, és essencial comprendre la seva distribució en relació amb altres variables. Un exemple d'això és la comparació de les temperatures màximes i mínimes a cada ciutat d'Austràlia.

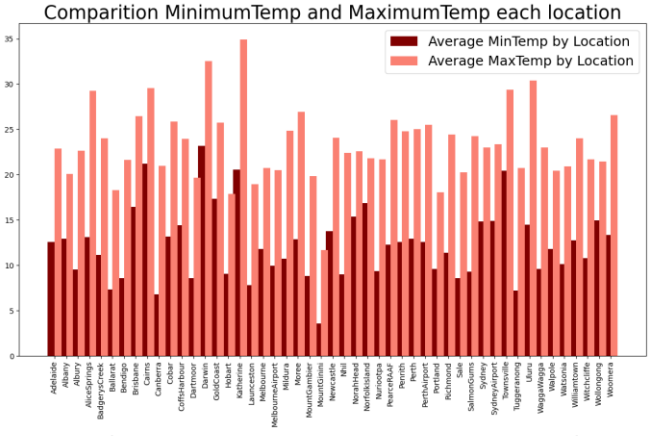

Fig. 2. Gràfica que mostra la temperatura mitjana màxima i mínima per a cada ciutat d'Austràlia. L'eix X representa les ciutats, mentre que l'eix Y representa les temperatures.

Podem observar que obtenim la temperatura més alta en el poble de Katherine arribant a una mitjana de 35 graus i una mínima temperatura de 20 graus, es a dir, no arriba a ser una zona freda a cap moment de l'any. La temperatura més baixa obtinguda s'obtè en el mont 'Mount Ginini' arribant als 4 graus. Per la resta de zones, podem veure que la temperatura va variant entre els 20 i els 30 graus.

Un altre aspecte important a considerar en l'anàlisi de les dades meteorològiques és la quantitat de pluja caiguda durant cada mes d'un any específic.

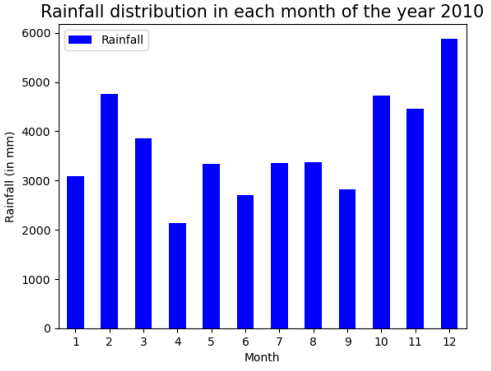

Fig. 3. Gràfica que representa la pluja caiguda, mesurada en mil·límetres, per a cada mes de l'any 2010. L'eix X indica el mes, mentre que l'eix Y mostra la quantitat de pluja mesurada en mil·límetres.

A partir d'aquesta gràfica, podem observar un patró interessant: les precipitacions comencen a augmentar a partir de setembre, coincidint amb la fi de l'estació d'estiu. Aquest augment continua fins a desembre, quan s'assoleix la quantitat màxima de pluja. En contrast, abril presenta la quantitat mínima de precipitacions.

La nostra variable objectiu en aquest estudi és 'RainTomorrow',

que pretenem predir. Per això, és crucial comprendre la distribució de les respostes 'Sí' i 'No' dins d'aquesta variable.

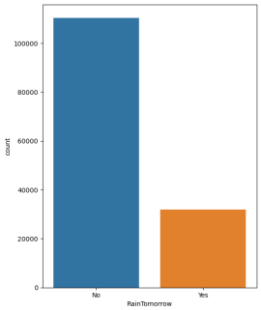

Fig. 4. Gràfica que representa la quantitat de respostes 'Sí' i 'No' per a 'RainTomorrow'.

Com podem observar, les dades no estan equilibrades. Això significa que haurem de procedir amb precaució a l'hora de dividir les dades en conjunts d'entrenament i de prova, per assegurar-nos que ambdós conjunts estiguin adequadament representats.

2 METODOLOGIA

Aquest conjunt de dades és de naturalesa temporal, la qual cosa significa que l'ordre cronològic és crucial per al seu tractament. En primer lloc, abans de comprovar i emplenear NaNs, hem de ordenar les dades en funció de la data amb format anymes-dia.

Un cop ordenades les dades, podem observar la quantitat de NaNs que hi ha al conjunt de dades.

Nombre de NaNs per columna:		Percentatge de NaNs per columna:	
Date	0	Date	0.000000
Location	0	Location	0.000000
MinTemp	1485	MinTemp	1.020899
MaxTemp	1261	MaxTemp	0.866905
Rainfall	3261	Rainfall	2.241853
Evaporation	62790	Evaporation	43.166506
Sunshine	69835	Sunshine	48.009762
WindGustDir	10326	WindGustDir	7.098859
WindGustSpeed	10263	WindGustSpeed	7.055548
WindDir9am	10566	WindDir9am	7.263853
WindDir3pm	4228	WindDir3pm	2.906641
WindSpeed9am	1767	WindSpeed9am	1.214767
WindSpeed3pm	3062	WindSpeed3pm	2.105046
Humidity9am	2654	Humidity9am	1.824557
Humidity3pm	4507	Humidity3pm	3.098446
Pressure9am	15065	Pressure9am	10.356799
Pressure3pm	15028	Pressure3pm	10.331363
Cloud9am	55888	Cloud9am	38.421559
Cloud3pm	59358	Cloud3pm	40.807095
Temp9am	1767	Temp9am	1.214767
Temp3pm	3609	Temp3pm	2.481094
RainToday	3261	RainToday	2.241853
RainTomorrow	3267	RainTomorrow	2.245978

Fig. 5. Representació del nombre i percentatge de NaN per a cada columna.

Observem que les columnes 'Evaporation', 'Sunshine', 'Cloud9am' i 'Cloud3pm' contenen més del 35% de valors NaN. Donada la gran quantitat de valors NaN, decidim eliminar les columnes 'Evaporation' i 'Sunshine'. No obstant això, decidim conservar les columnes 'Cloud', ja que la informació sobre la cobertura de núvols és rellevant per a la predicció de la pluja.

Per omplir els valors NaN, utilitzem diferents estratègies segons el tipus de columna. Per a les columnes numèriques, omplim els valors NaN utilitzant primerament el mètode 'backward fill', seguit del 'forward fill' i finalment la mitjana. Per a les columnes no numèriques, omplim els valors NaN amb el valor més freqüent, és a dir, el mode.

Després d'aplicar aquestes tècniques, encara queden alguns valors NaN. Això es deu principalment al fet que algunes localitzacions tenen poques o cap dada en determinades columnes. En aquests casos, decidim eliminar les files corresponents, ja que no és apropiat inventar dades per a aquestes localitzacions.

Amb un conjunt de dades lliure de valors NaN, podem procedir a la conversió de les variables categòriques a numèriques. Començant per les columnes 'RainToday' i 'RainTomorrow', les transformem en valors booleans 1s i 0s, corresponents a 'Yes' i 'No', respectivament.

Per a les columnes que indiquen la direcció del vent, utilitzem un codi personalitzat basat en les direccions cardinals mesurades en graus, tal com es mostra a la Figura 6.

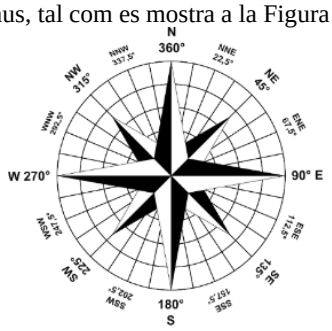

Fig. 6. Representació de direccions cardinals mesurada en graus.

Pel que fa a la data, que és única per a cada fila del conjunt de dades, utilitzem una tècnica basada en la funció sinusoidal per convertir cada data en un valor numèric.

A més, generem noves columnes com 'RainYesterday' i 'RainLastWeek', que indiquen si va ploure el dia anterior o la setmana passada, respectivament. Aquestes columnes poden millorar la precisió de les nostres prediccions. Si aquestes noves columnes generen valors NaN, optem per eliminar les files corresponents, ja que no és apropiat inferir les precipitacions de dies desconeguts.

Finalment, un cop totes les columnes estan tractades i creades, eliminem les columnes 'Date' i 'Location', que ja no són necessàries.

Ara tenim un conjunt de dades completament processat i preparat per a la divisió i l'entrenament. En aquest projecte, abans de procedir a aquesta etapa, prepararem dos conjunts de dades per a l'entrenament: un que ha estat simplement estandarditzat i un altre que ha passat per un anàlisi de correlació, una estandardització i un Anàlisi de Components Principals (PCA).

A continuació, analitzem la correlació entre les diferents columnes del nostre conjunt de dades utilitzant un mapa de calor, tal com es mostra a la Figura 7.

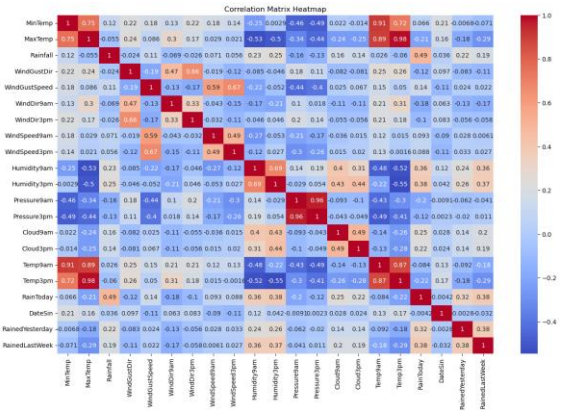

Fig. 7. Gràfica heatmap que representa correlacions entre les columnes.

Observem que hi ha una forta correlació entre algunes parelles de columnes, com ara 'Temp9am' i 'MinTemp', 'Temp9am' i 'MaxTemp', i 'Pressure3pm' i 'Pressure9am'. Per reduir la redundància, establim un llindar de correlació de 0.8 i eliminem les columnes 'Pressure3pm', 'Temp9am' i 'Temp3pm', reduint així el nostre conjunt de dades a 18 columnes.

A continuació, estandarditzem les dades del nostre conjunt de dades original i del conjunt de dades reduït per correlació utilitzant l'escala StandardScaler. La Figura 8 mostra la distribució de les dades per cada columna en ambdós conjunts de dades.

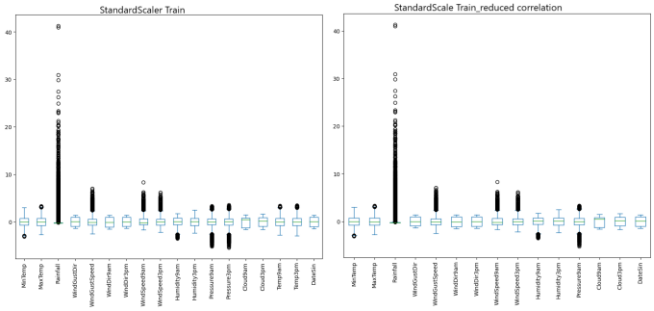

Fig. 8. Gràfica de tipus box que mostra la distribució de dades per cada columna de cada data set. Esquerra data set original i dreta data set tractat amb correlació.

Observem que, tot i que totes les dades estan estandarditzades de manera adequada, la columna 'RainFall' presenta una major variabilitat degut a la seva mesura en mil·límetres.

Un cop hem escalat les dades, apliquem l'Anàlisi de Components Principals (PCA) al conjunt de dades al qual hem aplicat el tractament de correlació. El nostre objectiu és assegurar que la variança explicada acumulada superi el 85% per obtenir resultats fiables. La Figura 9 mostra la variança acumulada en funció del nombre de components utilitzats.

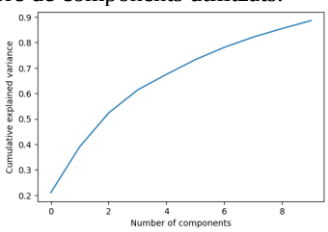

Fig. 9. Gràfica que mostra la variança cumulativa en funció del nombre de components utilitzats.

Observem que, a partir de 7 components, l'increment de la variança explicada es ralentitza. Quan utilitzem 10 components, la variança explicada acumulada supera el nostre objectiu del 85%, reduint així el nostre conjunt de dades a només 10 columnes.

3 EXPERIMENTS

Amb dos conjunts de dades preparats per a l'entrenament, el primer pas és seleccionar la mètrica que utilitzarem per avaluar el model durant la validació creuada, així com els classificadors que utilitzarem.

La mètrica que utilitzarem en aquest cas és la F1-Score de tipus macro. Aquesta mètrica és particularment útil quan es tracta de conjunts de dades desequilibrats, ja que proporciona una mesura més equitativa del rendiment del model que l'exactitud, que pot ser enganyosa en aquests casos.

Pel que fa als classificadors, utilitzarem tres: Regressió Logística, KNearestNeighbours (KNN) i XGBoost. La raó d'aquesta selecció és que ens permetrà comparar el rendiment d'un classificador relativament senzill, com és la Regressió Logística, amb el d'un classificador més complex, com és XGBoost, passant per un de complexitat intermèdia, com és KNN.

Per a la Regressió Logística, utilitzarem GridSearchCV per optimitzar els paràmetres del model. De manera similar, per a KNN i XGBoost, també utilitzarem GridSearchCV per trobar la millor combinació de paràmetres.

A més, utilitzarem l'estratègia de validació creuada StratifiedKFold amb 5 i 10 splits, que garanteix que cada subconjunt de dades conté aproximadament la mateixa proporció de mostres de cada classe objectiu com el conjunt de dades original.

Els paràmetres utilitzats per a cada classificador són els següents:

1. **Regressió Logística:**
   - penalty: Aquest paràmetre especifica la norma utilitzada en la penalització (regularització). Les opcions són 'l1', 'l2' i 'none'.
   - C: Aquest paràmetre és l'invers de la força de regularització. Els valors més petits indiquen una regularització més forta. Utilitzem un espai logarítmic entre -4 i 4.
   - solver: Aquest paràmetre especifica l'algorisme a utilitzar en el problema d'optimització. Les opcions són 'newton-cg', 'lbfgs' i 'liblinear'.
2. **KNearestNeighbours (KNN):**
   - n_neighbors: Aquest paràmetre indica el nombre de veïns a utilitzar per defecte per a les consultes dels k-veïns més propers. Utilitzem una llista de valors entre 5 i 20.
   - p: Aquest paràmetre és la potència del paràmetre Minkowski utilitzat per la mesura de distància.

Utilitzem 1 i 2, que corresponen a la distància de Manhattan i la distància euclidiana, respectivament.

- weights: Aquest paràmetre especifica la funció de pes utilitzada en la predicció. Les opcions són 'uniform', on tots els punts de la veïnatge tenen el mateix pes, i 'distance', on els punts més propers d'un punt de consulta tindran una influència més gran.

3. **XGBoost:**
   - learning_rate: Aquest paràmetre encogeix els pesos de les funcions de predicció en cada pas. Utilitzem 0.01, 0.1 i 0.2.
   - max_depth: Aquest paràmetre indica la profunditat màxima d'un arbre. Utilitzem 3, 5 i 10.
   - n_estimators: Aquest paràmetre especifica el nombre d'arbres a construir. Utilitzem 100, 200 i 500.
   - min_child_weight: Aquest paràmetre controla la complexitat del model. Els valors més alts redueixen el sobreajustament. Utilitzem 1, 3 i 5.

Finalment, una vegada que tenim els millors models per a cada classificador, calcularem la puntuació F1 mitjana utilitzant validació creuada. Aquesta puntuació ens proporcionarà una mesura robusta del rendiment dels nostres models.

4 RESULTATS

Amb els models preparats, procedim a l'entrenament amb el conjunt de dades d'entrenament i avaluem el seu rendiment. Els millors paràmetres trobats per GridSearch per a cada model són els següents:

Regressió Logística:

- StratifiedKfold 5: {'C': 0.23357214690901212, 'penalty': 'l2', 'solver': 'lbfgs'}
- StratifiedKfold 10: {'C': 0.03359818286283781, 'penalty': 'l2', 'solver': 'lbfgs'}
- StratifiedKfold 10 (conjunt de dades reduït): {'C': 0.0001, 'penalty': 'l2', 'solver': 'liblinear'}

KNN:

- StratifiedKfold 5: {'n_neighbors': 8, 'p': 1, 'weights': 'distance'}
- StratifiedKfold 10: {'n_neighbors': 8, 'p': 1, 'weights': 'distance'}
- StratifiedKfold 10 (conjunt de dades reduit): {'n_neighbors': 19, 'p': 2, 'weights': 'distance'}

XGBoost:

- StratifiedKfold 5: {'learning_rate': 0.2, 'max_depth': 10, 'min_child_weight': 1, 'n_estimators': 500}
- StratifiedKfold 10: {'learning_rate': 0.2, 'max_depth': 10, 'min_child_weight': 1, 'n_estimators': 500}
- StratifiedKfold 10 (conjunt de dades reduit): {'learning_rate': 0.2, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 500}

En primer lloc, examinem els resultats obtinguts quan utilitzem StratifiedKfold amb un valor de 5.

Train – CV = 5		
Model	F1-Score	
	Per Defecte	Millor paràmetres
Regressió Logística	0.74970679809	0.74995679744
KNN	0.73266295694	0.741963933729
XGBoost	0.76643084705	0.772752556181

Taula 1. Resultats obtinguts per crossvalidation amb cv=5.

Podem observar que, tot i la simplicitat de la Regressió Logística, aquesta ofereix un rendiment comparable al de KNN. No obstant això, XGBoost supera ambdós models.

A continuació, repetim el procés amb StratifiedKfold amb un valor de 10.

Train – CV = 10		
Model	F1-Score	
	Per Defecte	Millor paràmetres
Regressió Logística	0.74966722156244	0.749914586215
KNN	0.73438586866161	0.744864761043
XGBoost	0.76811363952401	0.775600887485

Taula 2. Resultats obtinguts per cross-validation amb cv=10.

En aquest cas, obtenim resultats similars als de l'experiment anterior, però amb una millora lleugera en el rendiment. En particular, XGBoost arriba a un F1-Score del 77%.

Finalment, apliquem el mateix procés al conjunt de dades reduït, que ha estat sotmès a un anàlisi de correlació i PCA.

Train_reduced – CV = 10		
Model	F1-Score	
	Per Defecte	Millor paràmetres
Regressió Logística	0.72449586303327	0.731658779505
KNN	0.71140631219274	0.720325228234
XGBoost	0.73279242636568	0.736043847877

Taula 3. Resultats obtinguts per cross-validation amb cv=10 en train_reduced.

En aquest cas, el rendiment és inferior en comparació amb els dos experiments anteriors. Per tant, descartem aquest conjunt de dades per a l'anàlisi final.

5 ANÀLISIS FINAL

Amb els millors paràmetres identificats, procedim a avaluar el rendiment dels nostres models utilitzant el conjunt de proves. En aquesta fase, també utilitzem la corba ROC per obtenir la mètrica de l'àrea sota la corba ROC (ROC AUC), i determinem els millors llindars de distància i Youden.

Regressió Logística			
Accuracy: 0.8535010030826442	Precision: 0.8015509950659305	Recall: 0.7280568600163834	F1-Score: 0.7543746950175318
ROC AUC: 0.874	Best Distance thr: 0.203	Best Youden thr: 0.203	

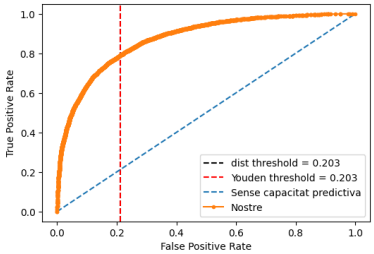

Fig. 10. Gràfica que mostra la corba ROC i els thresholds Youden i Distance.

Taula 4. Resultats de les mètriques del conjunt de test en RL.

KNN			
Accuracy: 0.8497333268092185	Precision: 0.7985719355135261	Recall: 0.7155572669910796	F1-Score: 0.7435651905957377
ROC AUC: 0.854	Best Distance thr: 0.229	Best Youden thr: 0.244	

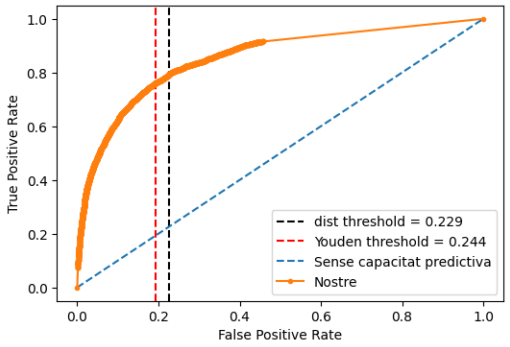

Fig. 11. Gràfica que mostra la corba ROC i els thresholds Youden i Distance.

Taula 5. Resultats de les mètriques del conjunt de test en KNN.

XGBoost			
Accuracy: 0.8637764838283	Precision: 0.814473196598	Recall: 0.7538899855	F1-Score: 0.7774820289
ROC AUC: 0.894	Best Distance thr: 0.076999999	Best Youden thr: 0.0769999995827	

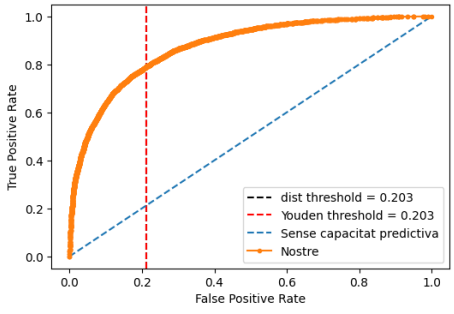

Fig. 12. Gràfica que mostra la corba ROC i els thresholds Youden i Distance.

Taula 6. Resultats de les mètriques del conjunt de test en XGB.

Observem que tots tres models ofereixen un rendiment respectable, amb una precisió superior a l'80% en tots els casos. No obstant això, XGBoost supera la Regressió Logística i KNN en totes les mètriques, assenyalant-se com el model més potent dels tres.

En aquest cas, hem utilitzat dos mètodes per determinar el millor llindar: el llindar de distància i el llindar de Youden. El llindar de distància es basa en la distància mínima a la cantonada superior esquerra del gràfic ROC, mentre que el llindar de Youden es basa en la màxima diferència entre la taxa de veritables positius i la taxa de falsos positius.

Per a la Regressió Logística, tant el llindar de distància com el de Youden són 0.203. Per a KNN, el llindar de distància és 0.229 i el de Youden és 0.244. Per a XGBoost, ambdós llindars són 0.077.

La Regressió Logística, tot i ser el model més senzill, ofereix un rendiment comparable al de KNN, el que demostra la seva eficàcia en aquesta tasca de classificació.

Pel que fa a KNN, tot i que el seu rendiment és lleugerament inferior al de la Regressió Logística i XGBoost, encara ofereix resultats útils, especialment tenint en compte la seva simplicitat relativa. Observem que tots tres models ofereixen un rendiment comparable en l'entrenament i la prova, indicant que estan generalitzant bé a noves dades. Així, si haguéssim de triar un sol model per a futures implementacions, XGBoost seria la nostra elecció degut al seu rendiment superior.

5 Conclusions

Aquest treball ha demostrat que, mitjançant un procés metòdic i l'ús de tècniques d'anàlisi de dades avançades, és possible predir amb precisió si plourà demà en diverses localitats d'Austràlia. Aquestes prediccions poden ser d'un gran valor en diversos àmbits, com ara l'agricultura, la planificació d'esdeveniments a l'aire lliure i la gestió de desastres naturals. A més, el treball destaca la importància de l'elecció del model adequat i de l'ajustament dels seus paràmetres per obtenir el millor rendiment possible.

Finalment, el treball demostra que, tot i la simplicitat relativa de models com la Regressió Logística i KNN, aquests poden oferir un rendiment comparable al de models més complexos com XGBoost, especialment quan s'optimitzen adequadament els seus paràmetres. No obstant això, XGBoost ha demostrat ser el model més potent dels tres, superant la Regressió Logística i KNN en totes les mètriques.