DataScientest

PRÉVISIONS MÉTÉOROLOGIQUES EN AUSTRALIE

Soutenance de **Lise Aujoulat** & **Geoffrey Foulon-Pinto**

Le 7 Janvier 2022

Supervision : Maxime Michel, Gaspard

Promotion DataScientest Bootcamp Octobre 2021

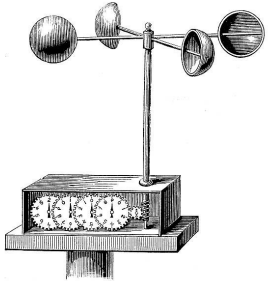
Parcours Data Scientist

Introduction

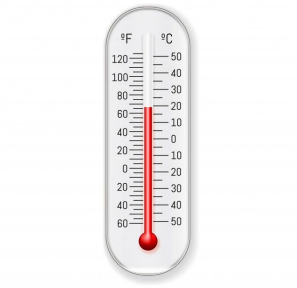
Thématique : prévisions météorologiques en Australie

- Météorologie : science antique, ayant évolué avec l'Homme
- Nombreux instruments de mesure

girouette



anémomètre



thermomètre

baromètre

Compréhension et prédiction du climat → **Apport de la data science**

Objectifs & méthodes

Réalisation d'un modèle de prédiction de la présence ou non de pluie à *J +1*

- Étude descriptive des données
- Traitement des valeurs manquantes
- Préprocessing
- Détermination du modèle de machine learning le plus adapté
- Modélisation en machine learning
- Essai d'amélioration du modèle
- Modélisation en deep learning

⇨ Objectif secondaire : étude de séries temporelles

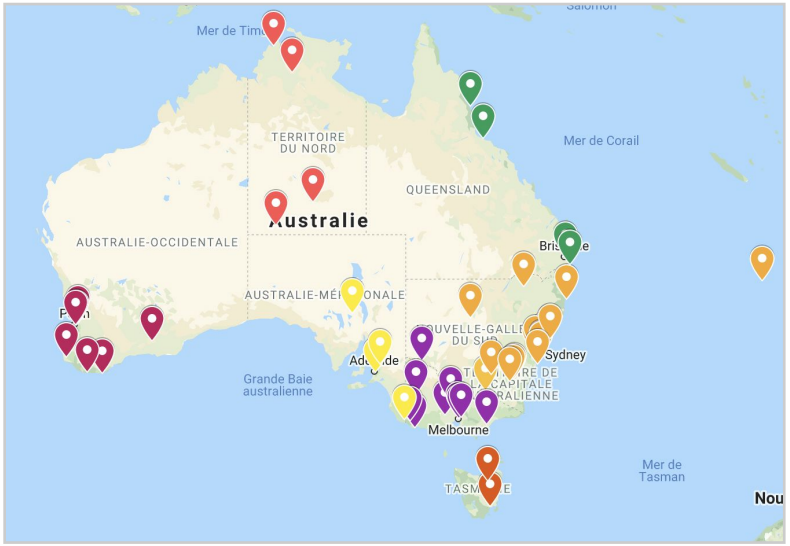
Base de données

Rain in Australia : predict next-day rain in Australia

Joe Young & Adam Young, www.kaggle.com

⟹ Origine des données : Bureau of Meteorology, Gouvernement Australien

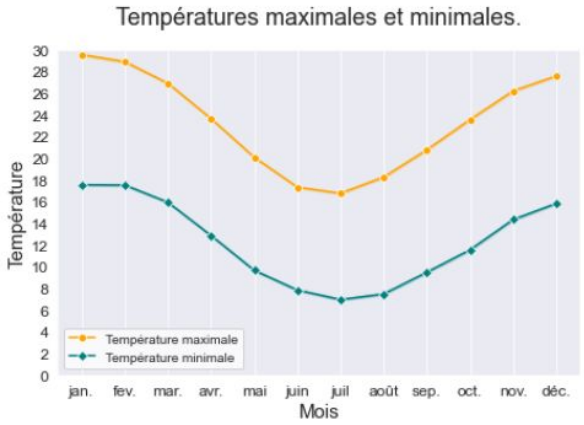
- 145 460 entrées provenant de 49 stations
- 22 variables explicatives
  - Températures
  - Humidité et précipitations
  - Ensoleillement, ennuagement
  - Pression atmosphérique
  - Direction et force du vent…



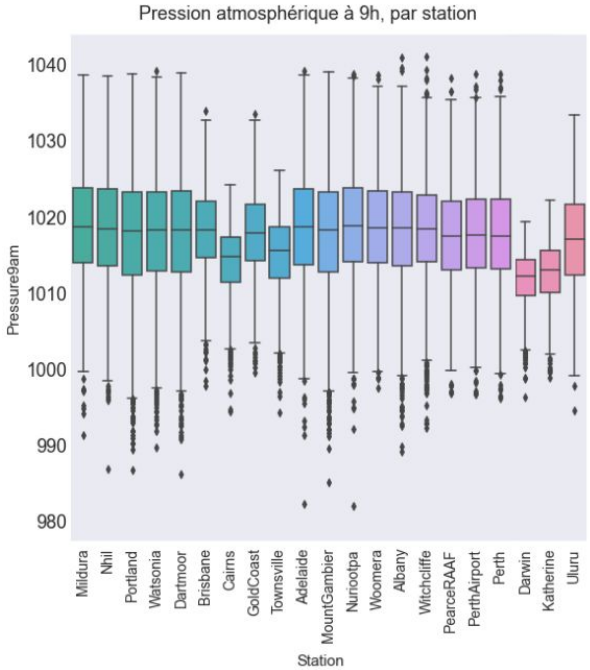
⟹ Variable à prédire : ***RainTomorrow***, booléen

Étude des données

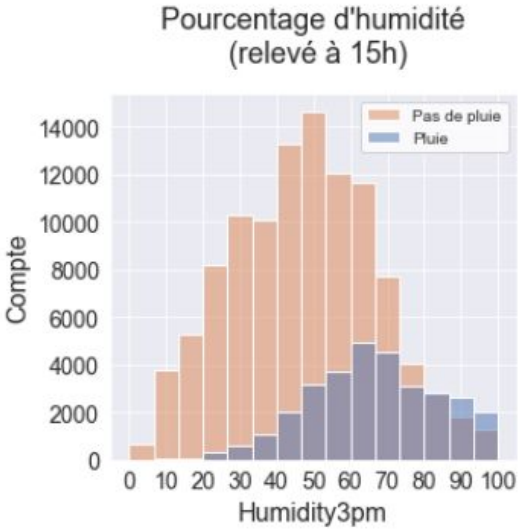
Saisonnalité



Influence géographique



Influence de la pluie



Modélisation machine learning

Plusieurs modèles testés après un cleaning simple :

Modèle Testé	Paramètres	Accuracy	Rappel (1 = jour de pluie)	Précision (1)	F1_score (1)
KNN	best param (*grid search*: k de 1 à 40, métriques 'minkowski', 'manhattan', 'chebyshev')	0.84	0.43	0.76	0.55
Decision Tree	criterion = 'entropy'	0.80	0.55	0.55	0.55
	criterion = 'gini	0.80	0.55	0.52	0.54
Random Forest	n_jobs = -1	**0.86**	0.54	0.77	**0.63**
Bagging	n_estimators = 10	0.85	0.52	0.74	0.61

Random Forest : **meilleures performances**, **combine arbre + bagging**, relativement **rapide**

Objectif: amélioration du **RF** avec focus sur la détection des **jours de pluie**
Importance de la stratégie au niveau du **preprocessing** +++

- Amélioration du preprocessing :
  → variables supprimées (*pression*, *évaporation*, *nuages*)
  → suppression ou remplacement des NaN par moyenne par station/année/mois
  → encoding des variables qualitatives
  → pas de standardisation / normalisation (même ordre de grandeur)

→ **84 % données conservées**

- Rééquilibrage des données : combinaison d'un *oversampling* suivi d'un *undersampling*

Accuracy	Rappel*	Précision*	f1-score*
0.83	0.63	0.61	**0.62**

** pour la classe 1 (jour de pluie)*

- Pistes explorées : GridSearch (paramètres), Feature Importance, Analyses des FN / VP (patterns, biais ?) → *cf Streamlit…*

→ *Objectif:* amélioration du modèle en **récupérant des variables** Importance des **connaissances métiers +++**

Essai d'amélioration du modèle

Révision du préprocessing

- limiter le nombre de variables supprimées
- regrouper les stations suivant leur localisation géographique

⇨ Conservation de **89 %** des données

Modélisation Random Forest

- réutilisation du code
- pas d'amélioration notoire

Modèle de deep learning

Layer (type)	Output Shape	Param #
Input (InputLayer)	[(None, 22)]	0
Dense_1 (Dense)	(None, 25)	575
Dense_2 (Dense)	(None, 50)	1300
Dense_3 (Dense)	(None, 2)	102

Total params: 1,977
Trainable params: 1,977
Non-trainable params: 0

Modèle à 2 couches denses

- Couches à 25 et 50 neurones, activation ReLu
- Couche de sortie à 2 neurones, activation Softmax

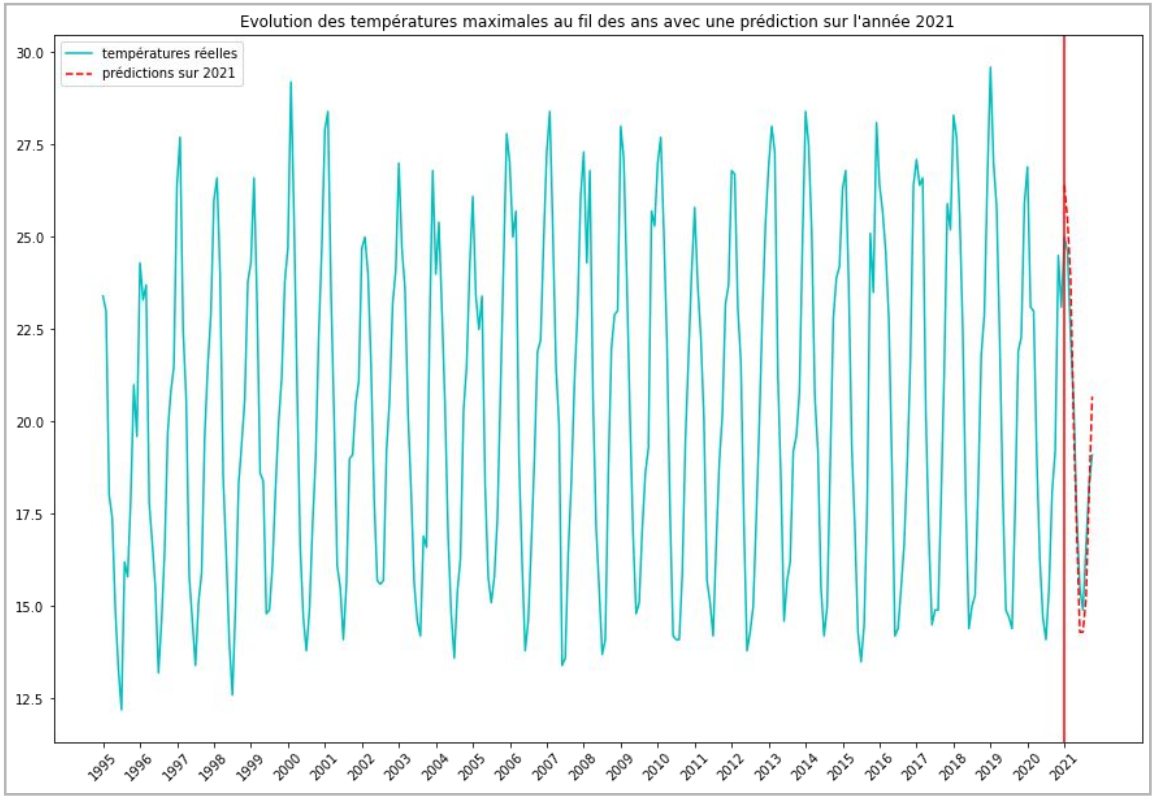
⇨ Précision de **85,6 %**

Essai d'amélioration des performances

- nombre de neurones
- nombre d'époch**s**
- taille des batchs
- fonction d'activation
- nombre de couches

Étude de série temporelle: Température maximale

```
model= sm.tsa.SARIMAX(series_fin, order=(1,1,1),seasonal_order=(0,1,1,12))
```

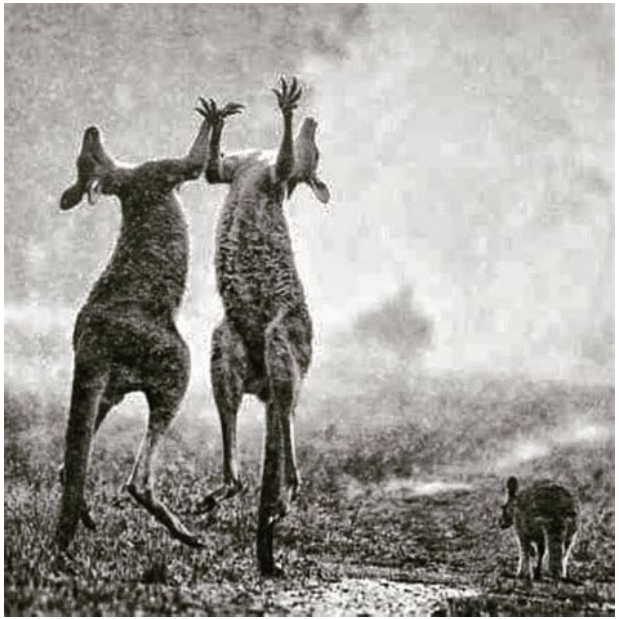


Qualité du Modèle

***Paramètres*: pvalue < 0.05**

***Ljung-Box / Jarque - Bera (résidus)*: pvalue > 0.05**

Démo Streamlit….



Conclusion / perspectives

Données et modèles

Importance des connaissances métiers dans la gestion des données (preprocessing)

RF: relativement correct, pistes d'améliorations ?
DL: moins performant (limite du modèle sur données tabulaires)
ST: intéressant à exploiter

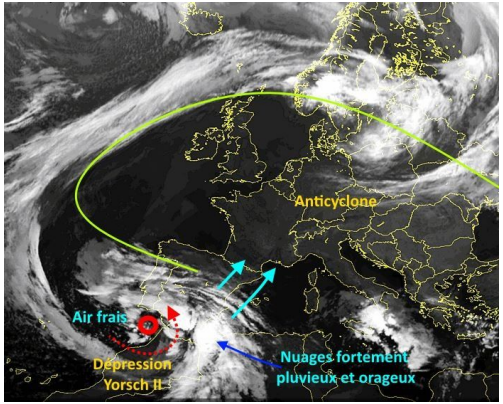
Applications métiers

Science / Société : études climatiques, préventions/stratégies

Applications touristiques

Perspectives

DL sur images satellites: prise en compte de phénomènes plus globaux (ex: *El Niño*)



source: meteo-sud-aveyron.over-blog.com