DataScientest • com

Rapport Technique d'évaluation

Projet : Prévisions météo en Australie

Promotion: AVRIL 2022

Réalisé par :

- Anne-Claire OGIERAIKHI
- Joseph CHARLES
- Olivier AMABLE
- Geneviève STEELE

Encadré par : Laurène BOUSKILA

GitHub : https://github.com/DataScientest-Studio/pyAusRainfall

Contexte (Joseph)

Contexte d'insertion du projet dans votre métier :

Ce projet porte sur l'étude d'une série chronologique regroupant environ 10 périodes d'observations météorologiques quotidiennes en Australie et couvre une gamme de techniques d'analyse, de modélisation et de prévision.

L'objectif principal est de prédire les précipitations du jour suivant l'observation à l'aide de la variable cible "RainTomorrow". Une étude des prévisions de température pour le lendemain de l'observation et des prévisions à plus long terme sont également réalisées en évaluant des modèles de séries chronologiques.

Le sujet proposé permet la mise en pratique des connaissances acquises pendant la formation sur la résolution d'un problème réel et le développement des compétences d'analyse et de prévision des données. Ce projet fil rouge offre la possibilité de collecter et d'explorer des données, d'en assurer la qualité, de les stocker dans des bases de données structurées, de développer des modèles prédictifs basés sur l'analyse statistique pour produire des outils décisionnels et l'opportunité d'apporter une première contribution à notre portfolio de science des données.

Du point de vue scientifique et technique :

Malgré les progrès fulgurants de la puissance de calcul, des observations disponibles et des méthodes numériques ces dernières années, la probabilité des précipitations reste l'un des principaux défis de la prévision météorologique. Pour faire de bonnes prévisions, les données atmosphériques ne suffisent pas car il faut regarder la Terre dans son ensemble, c'est-à-dire intégrer d'autres composants tels que les océans, les masses terrestres ou la cryosphère dans le modèle météorologique. De plus, la modélisation numérique de l'atmosphère doit faire face à plusieurs difficultés importantes :

- l'augmentation de la résolution des modèles et donc du nombre de calculs,
- l'augmentation des processus que l'on veut décrire telle que la chimie atmosphérique,
- l'augmentation du nombre de processeurs dans les supercalculateurs,
- l'augmentation de la consommation électrique des supercalculateurs,
- l'augmentation du nombre de lignes des programmes et donc de la difficulté de portabilité,
- l'augmentation de la masse des données à traiter et à stocker en sortie.

Les prévisions météorologiques pourraient bénéficier des algorithmes d'intelligence artificielle pour l'assimilation des données, car ils peuvent plus facilement intégrer et combiner des données provenant de différentes sources dans des modèles existants, améliorant ainsi le prétraitement des observations atmosphériques et physiques utilisées dans les modèles prédictifs.

L'apprentissage automatique pourrait également améliorer considérablement le paramétrage des équations traitant d'énormes masses de données physiques, ce qui augmenterait la résolution, la qualité et la vitesse d'exécution des modèles.

Actuellement, seuls les supercalculateurs de calcul haute performance actuels les plus puissants et les algorithmes hautement optimisés permettent aux prévisions d'être fiables et suffisamment rapides pour être utilisées efficacement. La puissance statistique des algorithmes d'apprentissage automatique devrait révolutionner toutes les prévisions météorologiques et offrir des opportunités pour une meilleure gestion des événements extrêmes.

DeepMind, un outil développé par Google, a récemment utilisé le modèle d'apprentissage en profondeur DGMR (Deep Generative Model of Radar) pour estimer non seulement l'emplacement des précipitations, mais également leur portée, leur durée, leur évolution et leur intensité de 5 à 90 minutes à l'avance.

Du point de vue économique :

L'utilisation de l'intelligence artificielle dans les modèles météorologiques représente un changement fondamental non seulement pour les météorologues, mais également pour plusieurs industries qui dépendent des aléas climatiques, telles que l'agriculture, les transports, la distribution, le textile, l'énergie, le tourisme, la construction et

l'événementiel. Les conditions climatiques affectent actuellement directement environ 70 % de l'économie mondiale, et on estime qu'environ 40 % des ventes françaises de biens de consommation seront affectées par le changement climatique.

Un autre effet des modèles d'intelligence artificielle est lié au paramétrage des événements physiques les plus difficiles à modéliser. En remplaçant la physique complexe par l'apprentissage automatique, les supercalculateurs pourraient fonctionner beaucoup plus rapidement, économiser de la puissance de calcul et consommer moins d'énergie.

Objectifs

Quels sont les principaux objectifs à atteindre ? Décrivez en quelques lignes. (Joseph)

Les principaux objectifs de cette étude sont d'appliquer les connaissances acquises au cours de la formation pour résoudre un problème réel et de se familiariser avec les outils et la méthodologie de gestion d'un projet en science des données.

Cela se traduit par les étapes les plus importantes du processus décisionnel de la science des données:

- l'exploration, la visualisation et l'analyse des données,
- la préparation du jeu de données,
- l'élaboration et l'évaluation de modèles prédictifs,
- l'interprétabilité des modèles de classification étudiés,
- le déploiement du modèle retenu via une interface graphique.

Pour chacun des membres du groupe, préciser le niveau d'expertise autour de la problématique adressée ? (Tous)

Joseph:

- Benchmarking d'applications scientifiques reposant sur PyTorch et TensorFlow sur des architectures de calcul standards et émergentes,
- Déploiement de modèles d'intelligence artificielle sur des plateformes HPC sans pour autant en comprendre vraiment l'anatomie ni le fonctionnement,
- Expérience professionnelle de 13 années dans le HPC dont 5 années au sein des communautés scientifiques des Sciences Atmosphériques et du Climat.

Geneviève:

- L'application des algorithmes de machine learning (principalement les arbres de décision boostés, les algorithmes de clustering et les réseaux de neurones) à des grands datasets pour la classification d'événements. Les événements en question étaient des désintégrations de particules, et les outils utilisés avaient été développés sur place, mais les techniques mathématiques restent les mêmes.

Olivier:

- Connaissance des algorithmes "classiques" de prévisions de ventes et en particulier les algorithmes de lissage exponentiel utilisés dans la plupart des APS (Advanced Planning System). Les APS sont les outils généralement utilisés dans la grande distribution pour prévoir les ventes.

Anne-Claire:

- Niveau débutant autour de la prédiction de météo

Êtes-vous entré en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents ? Si oui, détaillez l'apport de ces intéractions. (Tous)

Joseph:
Je ne suis entré en contact avec aucun expert métier pour affiner la problématique et les modèles sous-jacents. En revanche, je me suis documenté un peu sur le sujet via des blogs et vidéos sur internet.

Geneviève:

J'ai contacté une ancienne amie d'école qui a travaillé pour le Met Office (UK). Elle ne travaille plus là depuis plusieurs années mais à son époque, les études moins importantes utilisaient Fortran (!); et les plus gros frameworks de prédictions étaient construits sur mesure. La majorité des modèles donnant les prédictions étaient des réseaux de neurones.

Olivier:
Je ne suis pas entré en contact avec des experts métiers.

Anne-Claire:
Je n'ai contacté aucun expert métier pour m'aider au sein de ce projet.

Avez-vous connaissance d'un projet similaire au sein de votre entreprise, ou bien dans votre entourage ? Quel est son état d'avancement ? En quoi vous a-t-il aidé dans la réalisation de votre projet ? En quoi votre projet contribue-t-il à l'améliorer ? (Tous)

Joseph:
Ma dernière expérience professionnelle consistait à répondre aux appels d'offres Atos HPC-IA et à rédiger des rapports de benchmark pour de grands groupes industriels. J'ai pu constater une forte expansion ainsi qu'une intensification des demandes de solutions d'infrastructure d'IA à l'échelle mondiale.

Parmi mes derniers appels d'offres remportés figurent:

- le supercalculateur Européen "Leonardo" basé à CINECA en Italie, disposant d'environ 14,000 GPUs Ampere NVIDIA et de 10 ExaFlops de puissance AI FP16. Ce système contribue notamment à l'atténuation et à la gestion des risques liés aux conditions extrêmes, aux phénomènes naturels, aux tremblements de terre, aux tsunamis et aux événements volcaniques.
- les deux supercalculateurs "Belenos" et "Taranis" de Météo France, disposant d'un total de 300,000 cœurs de calcul CPU et de 21,48 PFlops de performance. Ces deux systèmes rendront les prévisions météorologiques plus fiables, feront des prédictions des aléas avec plus de précision, stimuleront les climats futurs plus finement et soutiendront les mesures d'adaptation au changement climatique.

Cette expérience m'a poussé à m'orienter vers une reconversion professionnelle en Data Science pour approfondir mes connaissances techniques et me familiariser avec les concepts clés de la visualisation, de l'analyse de données, du machine learning et des techniques de programmation en Python. Ce projet m'apporte une approche radicalement différente de la simulation numérique HPC dans le domaine des prévisions météorologiques.

Geneviève:
A part le projet de groupe, je ne travaille pas actuellement, et les personnes que je fréquente sont principalement d'autres parents d'élèves. Cependant, mon mari est Data Scientist, et j'ai déjà pu l'aider sur la partie théorique de certains algorithmes décisionnels.

Olivier:
Je ne sais pas si on peut parler ici de projet similaire mais la problématique de ce projet m'a fait penser à la difficulté de la prévision de la demande intermittente rencontrée parfois dans la grande distribution. En effet, il est parfois difficile de prédire une vente sachant que les ventes semblent être aléatoires et ne respecter aucune logique. Dans ces cas particuliers, le modèle "classique" utilisé est le modèle de demande intermittente de Croston (modèle à faible volume). Ce modèle étant moins performant que les modèles que nous avons utilisés lors de ce projet, nous n'avons pas creusé cette voie.

Anne-Claire:
Mes études universitaires m'ont permis d'avoir des connaissances théoriques afin de comprendre les différents types de régression utilisés lors de la modélisation des données. Lors d'un de mes précédents stages, j'ai eu l'occasion d'analyser des données provenant d'images satellites.

Data

Cadre (Anne Claire)

Quel(s) jeu(x) de donnée(s) avez vous utilisé pour atteindre les objectifs de votre projet ?

Le jeu de données utilisé contient environ 10 ans d'observations météorologiques quotidiennes provenant de différentes villes en Australie. Le jeu de données se compose de 23 variables :

- *Date* : Date d'observation.
- *Location* : Ville où se situe la station météorologique.
- *MinTemp* : Température minimale en degrés Celsius.
- *MaxTemp* : Température maximale en degrés Celsius.
- *Rainfall* : Quantité de pluie enregistrée pour la journée en millimètres.
- *Evaporation* : Niveau des bacs d'évaporation de classe A (en mm) dans les 24 heures jusqu'à 9h.
- *Sunshine* : Nombre d'heures d'ensoleillement dans la journée.
- *WindGustDir* : Direction des plus fortes rafales de vent dans les 24 heures jusqu'à minuit.
- *WindGustSpeed* : Vitesse en km/h des plus fortes rafales de vent dans les 24 heures jusqu'à minuit.
- *WindDir9am* : Direction du vent à 9h du matin.
- *WindDir3pm* : Direction du vent à 3h de l'après-midi.
- *WindSpeed9am* : Vitesse du vent en km/h moyennée sur 10 minutes avant 9h du matin.
- *WindSpeed3pm* : Vitesse du vent en km/h moyennée sur 10 minutes avant 3h de l'après-midi.
- *Humidity9am* : Humidité en pourcentage à 9h du matin.
- *Humidity3pm* : Humidité en pourcentage à 3h de l'après-midi.
- *Pressure9am* : Pression atmosphérique (hpa) au niveau de la mer à 9h.
- *Pressure3pm* : Pression atmosphérique (hpa) au niveau de la mer à 15h.
- *Cloud9am* : Opacité du ciel obscurci par les nuages à 9h mesuré en "oktas (0 signifie que le ciel est complètement éclairci tandis que 8 indique que le ciel est couvert.
- *Cloud3pm* : Opacité du ciel obscurci par les nuages à 15h mesuré en "oktas
- *Temp9am* : Température en degrés Celsius à 9h du matin.
- *Temp3pm* : Température en degrés Celsius à 3h de l'après-midi.
- *RainToday* : vaut 1 si les précipitations (en mm) dans les 24 heures avant 9h sont supérieur à 1 mm, 0 sinon.
- *RainTomorrow* : vaut 1 si les précipitations (en mm) du lendemain sont supérieur à 1 mm, 0 sinon

Ces données sont-elles disponibles librement ? Dans le cas contraire, qui est le propriétaire de la donnée ?

Ces données sont disponibles librement sur le site Kaggle à partir du lien suivant https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package. Il suffit de créer un compte Kaggle pour ensuite télécharger le fichier “weatherAUS.csv”.

Décrivez la volumétrie de votre jeu de données ?

Le jeu de données initial contient 23 variables comme indiqué précédemment ainsi que 145460 observations. Elles débutent au 01/12/2007 et s’arrêtent au 25/06/2017. On observe entre 2009 et 2016, une répartition annuelle des données assez équilibrée (autour de 16000 observations chaque année).

Le jeu de données final contient 14 variables ainsi que 140787 observations, ce qui correspond à une réduction significative de 3.2% des observations et de 39,1% des variables.

Pertinence (Anne Claire)

Avez-vous eu à nettoyer et à traiter les données ? Si oui, décrivez votre processus de traitement.

Dans un premier temps, nous avons supprimé les valeurs manquantes des variables *RainTomorrow* et *RainToday.* Nous avons ensuite encodé ses variables pour que la valeur “Non” soit remplacée par 0 et la valeur “Oui” soit remplacée par 1.
Nous avons utilisé la méthode KNN-Imputer() pour les variables numériques présentant la plus forte proportion de valeurs manquantes (*Sunshine, Evaporation, Cloud3pm, Cloud9am*). Concernant les variables quantitatives restantes, l’utilisation de l’interpolation avec la méthode ‘time’, nous a permis de remplacer les valeurs manquantes.

Pour finir, les valeurs manquantes des trois variables qualitatives *WindGustDir, WindDir9am, WindDir3pm* ont été respectivement remplacées par le mode.
Deux variables ont été créées; *Temp_Delta_MinMax* qui est la différence entre *MaxTemp* et *MinTemp* et *Humidity_Delta* qui est la différence entre les variables d'humidité.

Quelles variables vous semblent les plus pertinentes au regard de vos objectifs ?

Les variables qui nous semblent les plus pertinentes afin de prédire la pluie au lendemain sont :

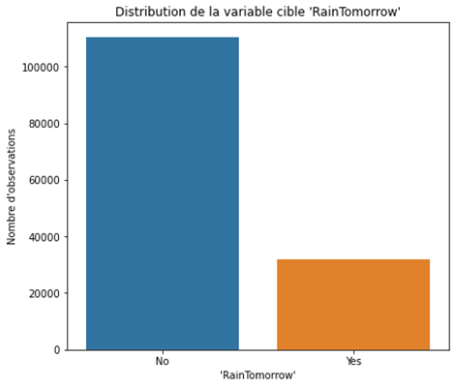
- *RainToday* car en général un épisode de fortes pluies est rarement isolé et s'accompagne donc de plusieurs autres épisodes dans le temps.
- *Humidity3pm et Humidity9am*, en effet, un taux d'humidité important au jour J peut indiquer que nous traversons plusieurs périodes pluvieuses.

Quelle est la variable cible ?

La variable cible est RainTomorrow. Elle signifie : a-t-il plu le jour suivant, oui ou non ? Cette colonne vaut "Oui" si la précipitation pour ce jour est de 1 mm ou plus et "Non" dans tous les autres cas.

Décrivez la distribution de ses valeurs ?

22.2% des observations correspondent à la classe minoritaire "Yes" tandis que 77.8% des observations correspondent à la classe majoritaire "No". Nous avons donc un jeu de données déséquilibré, en effet, la classe 'Yes' est minoritaire par rapport à la classe 'No'. Si aucune mesure n'est prise, les modèles auront donc du mal à identifier la classe minoritaire.



Avez-vous identifié des relations entre différentes variables ? Entre variables explicatives ? Et entre vos variables explicatives et la/les cible(s) ?

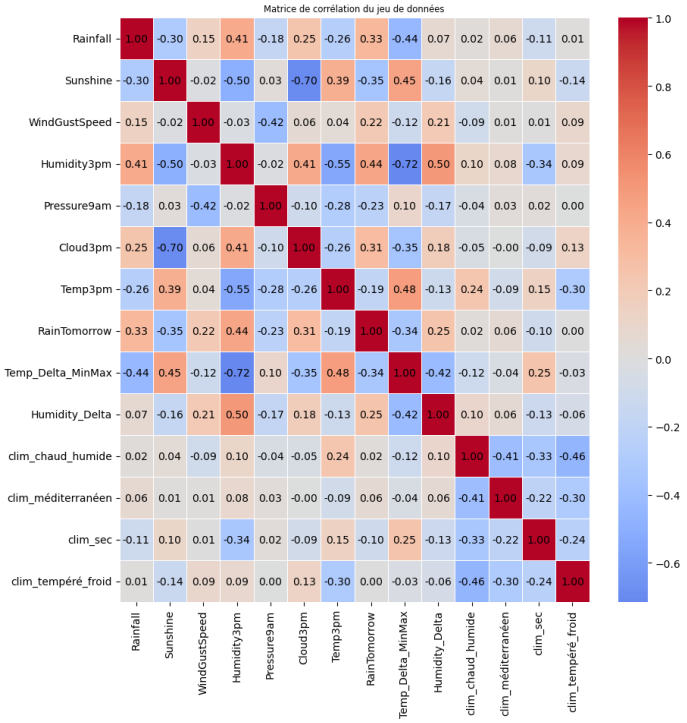
Dans un premier temps, nous avons étudié la corrélation entre la variable cible et les variables explicatives. Nous avons donc supprimé les variables les moins corrélés: *WindSpeed9am, WindSpeed3pm, MinTemp, Temp9am, Evaporation* du dataset. Les variables les plus corrélés à *RainTomorrow* sont Humidity3pm (0.438034), RainToday (0.313097), Temp_Delta_MinMax, (-0.336272) et Sunshine (-0.354955). Nous avons aussi décidé de la suppression de la date car ni l'année, le mois, la semaine ou la journée ne sont corrélées avec la variable cible (voir notebook concernant la data visualisation des données).

Concernant les corrélations identifiées entre les variables explicatives, nous obtenons les résultats suivants à partir de la matrice de corrélation :

- *Temp3pm* et *MaxTemp* sont très fortement corrélées (+0.97)
- *Pressure9am* et *Pressure3pm* sont très fortement corrélées (+0.92)
- *Humidity3pm* et Temp_Delta_MinMax sont très fortement corrélées (-0.72)
- *Sunshine* et *Cloud3pm* sont fortement corrélées (-0.70)
- *Sunshine* et *Cloud9am* sont fortement corrélées (-0.68)
- *Cloud9am* et *Cloud3pm* sont fortement corrélées (+0.62)
- *Humidity3pm* et *Temp3pm* sont fortement corrélées (-0.55)

Suite à ces constatations, nous avons supprimé les variables *Pressure3pm, Cloud9am, MaxTemp ainsi que* la variable *RainToday*, moins corrélée à la variable cible que *Rainfall.*

Suite à l'ajout de variables décrivant 4 catégories de climat en Australie, nous obtenons la matrice de corrélation finale suivante:



Quelles particularités de votre jeu de données pouvez-vous mettre en avant ?

De nombreuses particularités peuvent être relevées dans notre jeu de données:

- La différence de répartition des classes de la variable cible nous indique qu'un choix devra être fait afin de pallier cette problématique. Il existe plusieurs méthodes pour remédier à un déséquilibre des classes dans un dataset. Tout d'abord, il convient de vérifier que le déséquilibre du dataset soit représentatif de ce que l'on s'attend à trouver dans la réalité. Il est très important de choisir une métrique adaptée à ce type de problème, car la mesure de l'accuracy n'est pas suffisante. L'undersampling est utilisé lorsque l'on dispose d'un très grand nombre d'observations (>10K). Il s'agit ici simplement de retirer aléatoirement des instances de la classe majoritaire afin de rééquilibrer les proportions.
- Les variables qualitatives WindGustDir, WindDir9am et WindDir3pm (variables de direction du vent) ont des cardinalités élevées égales à 16. Nous avons fait le choix de les supprimer de notre dataset dans un premier temps. C'est un choix que nous avons fait de manière arbitraire.
- Nous avons rajouté des données concernant le climat des différentes villes australiennes. Après encodage de la variable *Clim_type*, nous obtenons les variables indicatrices clim_chaud_humide, clim_méditerranéen, clim_sec et clim_tempéré_froid.

Projet

Classification du problème (Joseph)

À quel type de problème de machine learning votre projet s'apparente-t-il ? (Classification, régression, clustering …)

Ce projet propose de résoudre un problème de classification binaire avec apprentissage supervisé, car il s'agit de prédire une variable quantitative cible à l'aide d'un ensemble de données étiquetées. Avec des entrées et des sorties étiquetées, le modèle peut mesurer sa précision et apprendre au fil du temps.

Pour l'élaboration de modèles prédictifs, nous avons eu recours à:

- des modèles de classification binaire avec apprentissage supervisé de type:
  - régression logistique ("Logistic Regression")
  - forêts aléatoires ("Random Forest")
  - séparateurs à vaste marge ("Support Vector Machine")
  - K-plus proches voisins ("K-Nearest Neighbors")
  - arbres de décision ("Decision Tree")
- des modèles de séries temporelles multivariées

À quelle tâche de machine learning votre projet s'apparente-t-il ? (détection de fraude, reconnaissance faciale, analyse de sentiment …) ?

Ce projet considère la classification de données déséquilibrées. Ceci est similaire à la détection d'anomalies car les échantillons d'apprentissage contiennent de fortes différences entre les classes à prédire. En effet, les classes "0" et "1" de la variable cible "RainTomorrow" représentent respectivement 78% et 22% de toutes les observations du jeu de données.

Pour réduire la redondance des informations fournies par un grand nombre d'individus majoritaires et pour mieux prédire les précipitations, nous avons utilisé et comparé des techniques de sous-échantillonnage telles que "RandomUnderSampler" et "ClusterCentroids", et de sur-échantillonnage telles que "RandomOverSampler" et "SMOTE". Cela nous a permis de rééquilibrer l'ensemble de données et de choisir l'approche la plus performante pour la métrique principale que nous avons choisie. La technique du sous-échantillonnage s'est imposée car nous disposions d'un très grand ensemble de données (145,460 observations) et cela impactait fortement la performance de certains algorithmes.

La répartition des classes de la variable cible "RainTomorrow" est la suivante:

- ensemble d'entraînement: {0: 76696, 1: 21854},
- ensemble d'entraînement ré-échantillonné: {0: 21854, 1: 21854}
- ensemble de test: {0: 32890, 1: 9347}

La classification pénalisée a été considérée car elle permet d'appliquer des coûts supplémentaires au modèle pour les erreurs de classification commises sur la classe minoritaire pendant l'entraînement. Ces pénalités sur les erreurs ont incité les modèles à accorder plus d'attention à la classe minoritaire.

L'argument "balanced", qui permet d'attribuer à chaque classe un poids inversement proportionnel à sa fréquence, a d'abord été testé puis une plage de différents ratios sur les classes a été considérée pour déterminer la configuration qui donnait les meilleurs résultats par rapport à nos métriques principales et secondaires. En complément, le seuil de probabilité à partir duquel les prédictions sont classées comme positives ou nulles a été ajusté d'une manière optimale afin de marquer plus distinctement les classes.

Le jeu de données est une série temporelle multivariée, car elle correspond à l'évolution temporelle de plusieurs variables explicatives. Puisqu'elle regroupe plusieurs séries univariées, elle permet la détection de corrélations entre plusieurs variables au fil du temps.

Quelle est la métrique de performance principale utilisée pour comparer vos modèles ?

Le choix d'une métrique d'évaluation appropriée pour comparer les performances de nos modèles de classification déséquilibrés et garantir leur qualité a été particulièrement difficile car nous nous intéressons dans ce projet aux prédictions de la classe minoritaire.

Lorsque les données ont une classe majoritaire forte, les algorithmes classiques sont souvent biaisés car leurs fonctions de perte tentent d'optimiser une métrique tout en ignorant la distribution des données et les classes minoritaires sont traitées comme des valeurs aberrantes par rapport à la classe majoritaire.

Les principales mesures de performance sont renseignées par la matrice de confusion, c'est-à-dire l'"accuracy", le "f1_score", la précision et le rappel. Les premières lignes concernent les évaluations par classe et les dernières lignes les évaluations moyennes.

Le **"f1_score"** permet de mesurer la capacité d'un modèle à prédire les individus positifs, tant en termes de précision (taux de prédictions positives correctes) qu'en termes de rappel (taux de positifs correctement prédits). Il est donc plus intéressant que l'"accuracy" dans le cas d'une situation de classes déséquilibrées, car le nombre de vrais négatifs n'est pas pris en compte dans son calcul.

Le "f1_micro" est une évaluation moyenne qui calcule le taux de prédictions correctes sur l'ensemble des observations, ce qui correspond tout simplement à l'"accuracy". Elle n'est donc pas adaptée car elle risque de refléter la surreprésentation de la classe majoritaire.

Le "f1_weighted" est une évaluation moyenne du "f1_score" par la pondération du "f1_score" pour chaque classe en fonction des distributions. Elle n'est donc pas adaptée à notre problème de classification car elle risque aussi de refléter la surreprésentation de la classe majoritaire.

Le **"f1_macro"** est la métrique d'évaluation qui nous intéresse le plus car elle calcule la moyenne arithmétique du "f1_score" de chaque classe. Cette métrique traite toutes les classes de la même manière, quelle que soit leur distribution. Puisqu'elle est adaptée à des situations de classes déséquilibrées, nous l'avons utilisée comme métrique de performance principale pour entraîner, évaluer et comparer nos modèles prédictifs.

Avez-vous utilisé d'autres métriques de performance qualitative ou quantitative) ? Si oui, détaillez.

D'autres métriques de performance ont été utilisées pour garantir au mieux la qualité de nos modèles prédictifs:

- **"balanced_accuracy":** cette métrique est la moyenne arithmétique de la sensibilité et de la spécificité. Elle est dérivée de l'"accuracy" standard qui a été ajustée pour mieux fonctionner sur des ensembles de données déséquilibrés. Elle calcule la précision moyenne pour chaque classe, au lieu de les combiner comme c'est le cas avec l'"accuracy" standard.
- **"geometric_mean":** cette métrique est la racine du produit de la sensibilité par classe. Cette mesure tente de maximiser l'"accuracy" sur chacune des classes tout en gardant ces "accuracy" équilibrées. Pour la classification binaire, elle est la racine carrée du produit de la sensibilité et de la spécificité.
- **"roc_auc":** cette métrique résume le compromis entre les taux de vrais positifs et les taux de faux positifs pour un modèle prédictif. Cette métrique peut être problématique avec des données déséquilibrées, car un petit nombre de prédictions correctes/incorrectes peut entraîner une grande modification du score.

Choix du modèle & Optimisation (Geneviève)

Quels algorithmes avez vous essayés ?

Nous avons essayé 5 modèles de machine learning durant la durée du projet:

- Régression Linéaire (lr),
- Arbres de Décision (dt),
- K-plus proches voisins (knn),
- SVM (svm),
- Forêt Aléatoire (rf).

Les meilleures performances obtenues correspondent aux modélisations des prévisions des précipitations avec étape de rééchantillonnage. Elles sont résumées dans le tableau suivant :

Final Test	f1_score	f1_macro	bal_acc	moy_geom	roc_auc
lr1	0.599413	0.719396	0.772318	0.772312	0.772318
lr2	0.599413	0.719396	0.772318	0.772312	0.772318
lr3	0.599388	0.719374	0.772302	0.772297	0.772302
rf0	0.621623	0.736925	0.787982	0.787981	0.787982
rf1	0.622077	0.736842	0.788919	0.788911	0.788919
rf2	0.621099	0.736475	0.787678	0.787677	0.787678
svm0	0.612058	0.729644	0.781015	0.781014	0.781015
svm1	0.611697	0.729226	0.780933	0.780932	0.780933
svm2	0.616739	0.732989	0.784738	0.784736	0.784738
knn	0.605684	0.720882	0.781044	0.780715	0.781044
dt0	0.585779	0.708628	0.762235	0.762228	0.762235
dt1	0.585861	0.708685	0.762304	0.762297	0.762304
dt2	0.578797	0.697870	0.761813	0.761201	0.761813

Pour chaque algorithme considéré (à l'exception du modèle KNN), les index 1,2 et 3 représentent les différentes itérations d'une étude comparative basée sur la pondération des classes, à savoir :

- 0 : correspond au paramétrage "class_weight = None" dans la définition du modèle,
- 1 : correspond au paramétrage "class_weight = 'balanced'" dans la définition du modèle,
- 2 : correspond au paramétrage "class_weight = {0:x, 1:1-x}" dans la définition du modèle,

De plus, le seuil de probabilité est adapté automatiquement dans toutes les modélisations afin de mieux distinguer les classes.

Décrivez celui / ceux que vous avez retenu et pourquoi ?

Bien que tous les modèles, lorsqu'on les entraîne sur des features sélectionnées des datasets bien nettoyés, donnent généralement des résultats décents, nous avons choisi un modèle de forêt aléatoire pour trois raisons principales:

- leur accuracy, f1, recall et précision étaient tous le plus haut de tous les modèles considérés.
- d'un point de vue pratique, le temps de calcul du modèle et les ressources CPU nécessaires étaient inférieures à l'autre candidat le plus proche : SVM. C'était donc plus facilement gérable pour des situations de calculs sur place, ou pour les services gratuits dans le cloud que nous avons utilisés, comme Google Collab (sans abonnement).
- l''interprétabilité des résultats de la forêt aléatoire est apparue comme supérieure à celle de l'autre modèle le plus proche, SVM. L'importance de chaque feature a été évaluée par deux méthodes. D'une part, la Détérioration Moyenne d'Impuretés, qui peut être affectée par le sur-entraînement du modèle, et être biaisée par les features qui présentent une haute cardinalité. D'autre part, par la permutation des features, qui inclut des calculs plus importants, et qui est influencée par les corrélations entre les variables. Les deux méthodes ont fait ressortir les colonnes Humidity3pm et Pressure9am comme étant parmi les features les plus importantes pour déterminer s'il va pleuvoir le lendemain.

Ces choix, prédits par les calculs du modèle mais aussi compréhensibles par les humains, se rapprochent aussi de certains proverbe concernant la météo: les oiseaux volent bas avant la pluie, car ils mangent les insectes qui volent bas eux aussi, car l'humidité de l'air se condense et les alourdit!

Qu'est ce qui a engendré une amélioration significative de vos performances ?

L'amélioration la plus significative des performances est venue d'une évaluation, d'une sélection et d'une manipulation très attentionnée des features. La structure du projet nous a amené à effectuer ces étapes tôt dans son déroulement. Après avoir testé les modèles, des modifications supplémentaires des features choisies --- ou l'exploration de nouvelles --- n'a apporté qu'une amélioration très limitée, voire pas du tout.

Tous les modèles ont aussi été testés avec une grille de paramètres, pour plusieurs paramètres pertinents. La sélection du meilleur ensemble de paramètres a amené une amélioration supplémentaire des performances.

Avez-vous analysé les erreurs de votre modèle ?

La forêt aléatoire a été testée avec une validation croisée, ce qui nous a permis d'établir que les erreurs étaient acceptables lors de la généralisation à un dataset indépendant.

Cela a-t-il contribué à son amélioration ? Si oui, décrivez.

Bien qu'il y ait quelques traces de sur-entraînement (légère différence de score entre le training set et le test set), les résultats de la validation croisée ont montré que l'erreur sur le score du dataset était minimale. Cela a prouvé qu'il n'y avait pas besoin d'aller plus loin.

Détaillez quelle a été votre contribution principale dans l'atteinte des objectifs du projet. (Tous)

Joseph:
Je pense avoir contribué à la réalisation des objectifs du projet à tous égards de manière inéquitable.

Geneviève:
Je pense avoir bien participé à toutes les étapes du projet (même si ce n'était pas toujours au moment idéal ni pour mes collègues ni pour ma famille!). En particulier, j'ai développé le travail initial de mes collègues en changeant la sélection des features, et en vérifiant les variations supplémentaires après l'étape de modélisation. J'ai aussi vérifié que le modèle était assez robuste pour donner des prédictions concernant d'autres variables que celle demandée.

Olivier:
J'ai travaillé sur la plupart des tâches mais ma contribution a été plus importante sur certaines tâches. Lors de la phase de visualisation des données, j'ai aidé à identifier l'ajout de variables supplémentaires pour améliorer les modèles. Lors de la phase de modélisation, je me suis concentré sur l'algorithme de KNN. J'ai effectué la classification des villes par type de climat. J'ai également creusé l'analyse des séries temporelles.

Anne-Claire:
J'ai principalement travaillé sur la représentation cartographique des données météorologiques, sur l'interprétation du modèle de régression logistique ainsi que sur une première ébauche du rapport final.

Description des travaux réalisés

Répartition de l'effort sur la durée et dans l'équipe (Olivier)

Morceler votre projet en un maximum de tâches unitaires. Produisez le diagramme de Gantt a posteriori en spécifiant qui s'est occupé de quelle tâche et à quelle moment. (joindre le diagramme en annexe du rapport)

Le diagramme de Gantt se trouve en annexe.

Bibliographie (Geneviève)

Sur quels éléments bibliographiques (articles de recherches, blog, livres, etc...) vous êtes vous appuyé pour réaliser votre projet ?

Technique:

- *https://scikit-learn.org/stable/*
- *https://pandas.pydata.org/docs/user_guide/index.html*
- *https://docs.python.org/3/*
- *https://seaborn.pydata.org/tutorial/introduction.html*
- *https://matplotlib.org/stable/gallery/index*

Spécifique à l'Australie:

- *http://www.bom.gov.au/climate/*
- *https://www.abcb.gov.au/resources/climate-zone-map*
- *https://www.kaggle.com/datasets/maryamalizadeh/worldcities-australia*

Théorie et études adjacentes:

- Interpretable Machine Learning - Christophe Molnar
- https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1974.tb00994.x
- https://www.sciencedirect.com/science/article/pii/S266682702100102X
- *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9099780/*
- *https://towardsdatascience.com/from-scratch-permutation-feature-importance-for-ml-interpretability-b60f7d5d1fe9*

Difficultés rencontrées lors du projet (Olivier)

Quel a été le principal verrou scientifique rencontré lors de ce projet ?

Notre principal verrou scientifique est lié à la théorie du chaos car, dans notre cas, il est difficile de prévoir un état dynamique sans avoir l'exhaustivité des conditions initiales. C'est un phénomène généralement illustré par l'effet papillon. La formulation exacte qui en est à l'origine fut exprimée par Edward Lorenz lors d'une conférence scientifique en 1972, dont le titre était : « Le battement d'ailes d'un papillon au Brésil peut-il provoquer une tornade au Texas ? ».
Afin de mieux définir notre système, il aurait été intéressant de connaître les conditions atmosphériques au large du pacifique, la présence ou non d'un anticyclone à proximité ou encore les données concernant les courants le long de la côte australienne.

Pour chacun des points suivants, si vous avez rencontré des difficultés, détaillez en quoi elle vous ont ralenti dans la mise en place de votre projet :

Prévisionnel: (tâches qui ont pris plus de temps que prévu etc …)

Parmis les tâches qui ont pris plus de temps que prévues, nous pouvons identifier les tâches suivantes :

Sélection des variables: Nous avons passé beaucoup de temps à sélectionner les variables. Etant donné que le temps de traitement des algorithmes était long, nous avons opté pour une approche simple qui consiste à ne sélectionner dans premier temps que les variables numériques. Nous avons ensuite appliqué les méthodes de sélection de variables apprises dans les cours.

Séries temporelles ARIMA: Nous avons perdu beaucoup de temps sur l'analyse des séries temporelles. Les résultats étaient peu probants et nous avons utilisé des contournements qui ne permettaient pas de répondre aux enjeux de notre projet.

Mise au propre des codes et rédaction sur GitHub: Nous avions commencé notre projet en partageant nos fichiers d'étude sur Google Collab. Nous avons par la suite transféré nos documents sur GitHub. La puissance de l'outil, nous a permis de mieux structurer nos fichiers. Un grand merci à Joseph qui a principalement œuvré à cette tâche. L'harmonisation des codes sur GitHub a pris un temps considérable car chaque modification de paramétrage ou

d'optimisation apportée sur un modèle se répercute à la fois sur tous les autres modèles de la partie modélisation, mais également sur tous les notebook de la partie interprétabilité. Cela demande donc une exécution complète de quasiment tous les notebooks à chaque modification de paramétrage pour obtenir une cohérence entre tous les résultats obtenus et une base de comparaison fiable entre les modèles.

Choix des métriques de performance: Nous avons commencé à évaluer les modèles en utilisant l'"accuracy" comme métrique de performance principale, puis nous avons réalisé que cette métrique ne convenait pas et nous avons privilégié le "F1_score". Ayant pris conscience des subtilités qu'il existe entre les différents variants de cette métrique, nous avons finalement opté pour le "F1_macro" comme métrique principale accompagné des métriques secondaires "balanced_accuracy", "geometric_mean" et "roc_auc".

Jeux de données : (acquisition, volumétrie, traitement, agrégation etc …)

Nous avions beaucoup de données manquantes dans notre jeux de données.
De plus, s'agissant d'une série temporelle, il nous manquait parfois des observations sur plusieurs jours.

Compétences techniques / théoriques : (timing d'acquisition des compétences, compétence non proposée en formation etc …)

Il aurait été intéressant d'aborder ce projet en ayant certaines bases sur les méthodes Agiles.

Rappel sur la méthode Agile :

4 valeurs fondamentales de la démarche :

1. L'équipe, soit des individus et des interactions, plutôt que des processus et des outils ;
2. L'application, c'est-à-dire des fonctionnalités opérationnelles plutôt que de la documentation exhaustive ;
3. La collaboration avec le client, plutôt que la contractualisation des relations ;
4. L'acceptation du changement, plutôt que le suivi d'un plan.

De ces valeurs découlent les 12 principes généraux suivants :

1. Satisfaire la clientèle en priorité
2. Accueillir favorablement les demandes de changement
3. Livrer le plus souvent possible des versions opérationnelles de l'application
4. Assurer une coopération permanente entre le client et l'équipe projet
5. Construire autour de personnes motivées
6. Privilégier la conversation en face-à-face
7. Mesurer l'avancement du projet en matière de fonctionnalité de l'application
8. Faire avancer le projet à un rythme soutenable et constant
9. Porter une attention continue à l'excellence technique et à la conception
10. Faire simple
11. Responsabiliser les équipes
12. Ajuster à intervalles réguliers son comportement et ses processus pour être plus efficace

Autre remarque : la MasterClass portant sur l'outil GitHub intervient un peu trop tard dans la formation. Nous aurions probablement gagné un peu de temps si nous avions utilisé GitHub dès le démarrage du projet.

Pertinence : (de l'approche, du modèle, des données etc ...)

L'approche du projet s'est faite de manière pragmatique (Visualisation des données, Modélisation, Finalisation, Démo). Il aurait été intéressant de découper ce projet en plusieurs sprints avec à la fin de chaque sprint une démo streamlite. Dans un contexte d'entreprise, cela aurait permis de délivrer plus rapidement de la valeur au projet en partageant les informations de manière plus régulière.

Nous avons sélectionné un modèle mais nous aurions pu aller plus loin (cf. Bilan & Suite du projet)

Une meilleure connaissance de la géographie propre à l'Australie nous aurait permis de mieux identifier les facteurs clés de prédiction de nos modèles.

IT : (puissance de stockage, puissance computationnelle, etc …)

Certains algorithmes de machine learning consomment plus de ressources que d'autres. Dans le cas de l'algorithme SVM, les temps de traitement dépassent plusieurs heures. L'utilisation d'une plateforme spécialement conçue pour la datascience (ex : databricks) nous aurait permis de réduire drastiquement nos temps de traitement.

Autres

Bilan & Suite du projet (Joseph)

En quoi votre projet a-t-il contribué à un accroissement de connaissance scientifique ?

De manière générale, ce projet a permis à chacun de prendre en main certains outils dédiés à l'analyse, un langage de programmation et des méthodes applicables aux problèmes de données. La manipulation des algorithmes étudiés est complexe et cela a nécessité d'aller continuellement se documenter pour approfondir les connaissances acquises au cours de cette formation.

Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non.

Exploration, visualisation et analyse des données

L'étude statistique exploratoire réalisée a permis de nettoyer les données en vérifiant les doublons et les valeurs aberrantes et en remplaçant les valeurs manquantes. Les premières observations sur la signification de chaque variable par rapport à notre variable cible ont été obtenues à l'aide d'une matrice de corrélation. Comme la matrice de corrélation obtenue après encodage des variables qualitatives n'était pas satisfaisante, les variables catégorielles ont été ignorées pour ne garder que les variables quantitatives. Après sélection des variables quantitatives en fonction de leur importance par rapport à la variable cible, la matrice de corrélation obtenue a mis en évidence les variables fortement corrélées entre elles. Plusieurs visualisations ont été également réalisées pour valoriser les influences et distributions les plus importantes.

Préparation du jeu de données

Cette étape s'appuie sur les constatations et conclusions obtenues lors de la première phase d'analyse exploratoire du jeu de données. Les variables catégorielles ont été écartées du jeu de données et les valeurs manquantes ont été remplacées en utilisant la méthode "KNN-Imputer" pour les variables présentant la plus forte proportion de valeurs manquantes et la méthode "interpolate" pour les variables présentant moins de 10% de valeurs manquantes. Les seules variables pour lesquelles les valeurs manquantes ont été supprimées sont "RainToday" et "RainTomorrow". Les variables les moins corrélées à la variable cible ont été supprimées du jeu de données et différentes tables de contingences ont été produites afin de sélectionner, parmi les variables très corrélées entre elles, celle qui présente l'importance la plus élevée par rapport à la variable cible. Lors de cette étape, nous avons créé deux nouvelles variables qui nous semblaient pertinentes: "Temp_Delta_MinMax" et "Humidity_Delta". De nouvelles variables spécifiant le type de climat australien selon 4 catégories ont également été créées. Enfin, la méthode "SelectKBest" a confirmé que toutes les variables que l'on considérait dans le cadre de cette étude étaient importantes à conserver. Différents jeux de données ont été générés et stockés afin de permettre l'accès à différentes configurations. Finalement, après avoir écarté les variables catégorielles, remplacé certaines variables trop fortement corrélées entre elles et ajouté de nouvelles variables explicatives, le jeu de données a subi une réduction significative de 3.2% des observations et de 39,1% des variables.

Elaboration et évaluation de modèles prédictifs

Une étude comparative de plusieurs modèles de classification binaire par apprentissage supervisé a été menée en se basant sur le choix de métriques d'évaluation appropriées aux jeux de données déséquilibrées. Cette étude compare différents modèles avec et sans méthode de rééchantillonnage. Chacune des approches établit les performances des métriques d'évaluation considérées ("f1_macro", "balanced accuracy", "geometric_mean" et "roc_auc") en sélectionnant le meilleur estimateur parmi une grille de paramètres. Les différents résultats obtenus

sont consolidés par validation croisée et accompagnés des courbes “ROC”, de “gain cumulé” et de “precision-rappel”. Une étude comparative basée sur la pondération des classes est effectuée pour tous les modèles (à l’exception du modèle KNN), et le seuil de probabilité est adapté dans toutes les modélisations pour mieux distinguer les classes.

Interprétabilité des modèles de classification étudiés

Pour chacun des modèles étudiés, nous avons proposé une approche d’interprétabilité en recourant à différentes techniques et visualisations qui viennent conforter le constat de départ sur l’importance relative des variables explicatives par rapport à notre variable cible. Cette étude nous apporte des informations complémentaires à celles obtenues lors de la phase de modélisation et qui nous aide à nous prononcer sur le choix d’un modèle final retenu.

Déploiement du modèle retenu via une interface graphique

<A VENIR>

S’ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s’inscrire ? Détaillez.

Ce modèle peut s’inscrire dans un processus de vigilance de phénomènes météorologiques dangereux pour le jour courant et le lendemain afin d’anticiper les événements, se préparer aux pluies intenses et mieux se coordonner.

Dans le cas contraire, quelles pistes d’amélioration suggérez-vous pour améliorer les performances de votre modèle ?

Il reste encore de nombreux points à explorer afin d’améliorer les performances des modèles prédictifs réalisés dans le cadre de ce projet:

- la collecte de davantage de données pour éventuellement rééquilibrer les deux classes de la variable cible à un degré variable,
- la relabellisation des données de la classe majoritaire en sous-classes pour obtenir un problème de classification multi-classes plus équilibré,
- l’utilisation des variables catégorielles ignorées dans le cadre de cette étude,
- l’utilisation de méthodes générant des sous-ensembles sous-échantillonnés, comme le Boosting ou le Bagging,
- l’utilisation d'autres méthodes de machine learning, comme l’”Anomaly Detection” ou l’”Active Learning”,
- l’utilisation de réseaux de neurones récurrents par des méthodes de Deep Learning avec "Keras",
- l’utilisation de "Spark ML" afin d'optimiser les performances des algorithmes de classification tels que "Logistic Regression", "Decision Tree", "Random Forest" et "SVM".

Annexes

Planning (Olivier)

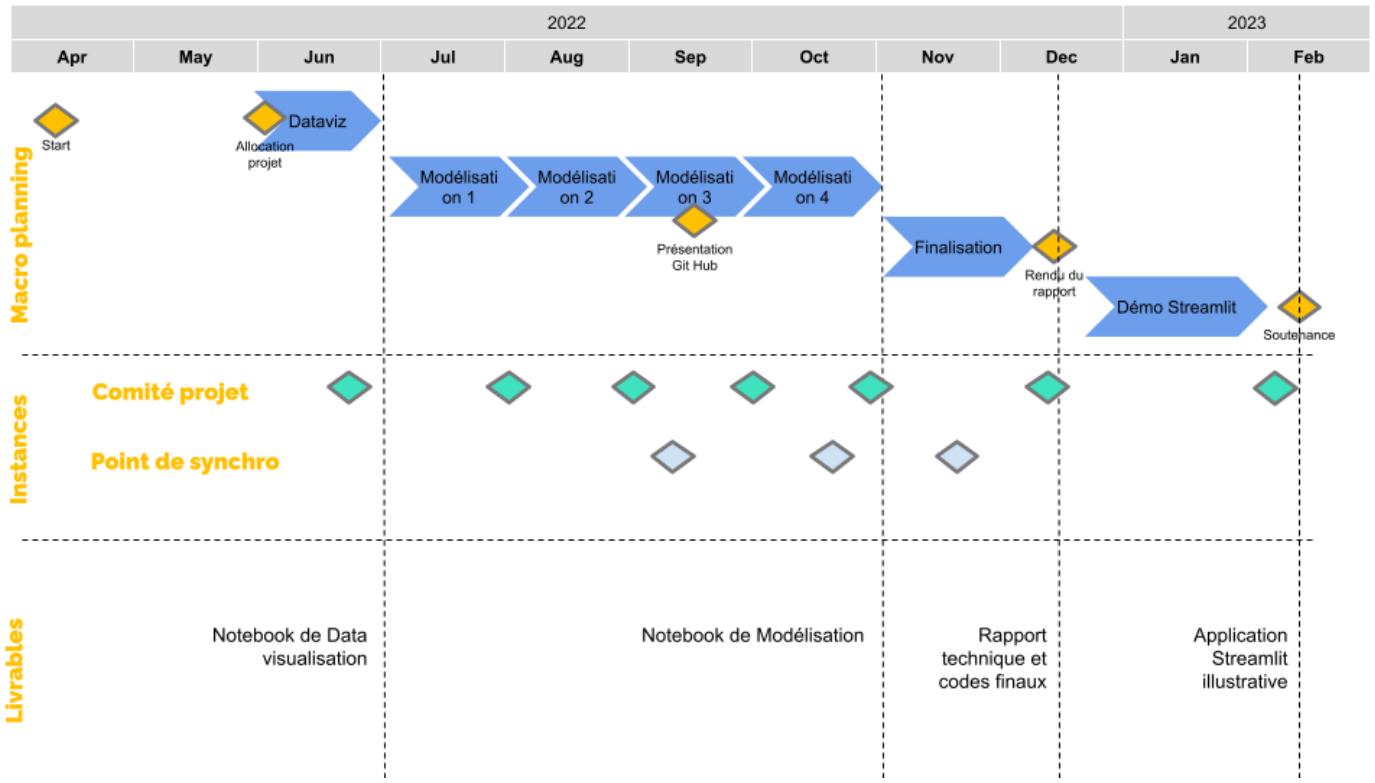


Diagramme de Gantt (Olivier)

Etape	Description	QUI ?	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb
DataViz	Préparation et nettoyage des données communes	Tous									
	Observation du jeu de données initial	Tous									
	Ajout de variables supplémentaires	O,G									
	Gestion des N/A	Tous									
	Encoding	Tous									
	Exporter le jeu de données	Tous									
	Etablir des axes d'intérêt pour les visualisations et les répartir	Tous									
	Corrélation	Tous									
	Représentation cartographique	AC									
	Moyenne annuelles des précipitations	Tous									
	Influence pour la prévision de pluie	Tous									
	Influence de certains critères (indépendamment du climat)	XXX									
	Influence de la pluie des jours précédents sur la pluie du lendemain	XXX									
	Influence des vents sur la pluie	G									
	Distribution des températures au cours de l'année suivant le climat	XXX									
Modélisation 1	Définir le type de problème ML	Tous									
	Définition de la variable cible	Tous									
	Preprocessing	Tous									
	Observer si la variable cible est équilibrée?	Tous									

Etape	Description	QUI ?	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb
	Explorer les techniques de rééchantillonnage et appliquer les algorithmes classiques:	Tous									
	SVM	J									
	KNN	O									
	DT	G									
	RF	G									
	LR	AC									
	Explorer les algorithmes de détection d'anomalies	N/A									
	Faire une recherche des paramètres optimaux avec GridSearchCV	Tous									
	Faire une étude comparative des algorithmes implémentées	J									
	Choisir le/les algorithmes qui vous offrent les meilleures performances	Tous									
Modélisation 2	Ajouter les coordonnées géographiques	AC									
	Sélection des variables avec le test du chi2	J									
	Extraction du jeu de données final	Tous									
	Tableau des performances des algo avec et sans rééquilibrage	O,J									
	Choisir une seconde stratégie	J									
	Classification des villes par clustering (faire le rapprochement avec le climat)	O									
	Deep Learning	N/A									
	Séries Temporelles	O,J									
Modélisation 3	Effectuer la même comparaison de performances sur la pluie à J+3 et J+7, et comparer	J									
	Explorer les stratégies Deep Learning ou séries temporelles	O,J									
	Tester différentes variables cibles tels que le vent ou la température	G,J									
Modélisation 4	Finaliser la modélisation	Tous									
	Derniers tests et modélisation	Tous									
	Séries temporelles ARIMA	O,J									
	Séries temporelles et climats	O,J									
	Deep Learning	N/A									
	Comparer les performances des modèles et stratégies	J									
	Etablir les forces et faiblesses des modèles	Tous									
	Juger de l'interprétabilité du modèle retenu	Tous									
	Établir les conclusions de la modélisation	Tous									
	Mettre au propre les codes sur github	J									
	Prépa support: rapport pdf ou word sur l'étape de modélisation	AC									
	Organiser le GitHub	J									
	Derniers tests et codes (J+3, J+7)	J									
	Comparer les performances des modèles et stratégies	J									
	Etablir les forces et faiblesses des modèles	Tous									
	Juger de l'interprétabilité du modèle retenu	Tous									
	Établir les conclusions de la modélisation	Tous									
	Mettre au propre les codes sur GitHub	J									
Finalisation	Rapport final:	Tous									
	Contexte (Joseph)	J									
	Objectifs (Joseph)	J									
	Cadre (Anne Claire)	AC									
	Pertinence (Anne Claire)	AC									
	Classification du problème (Joseph)	J									
	Choix du modèle & Optimisation (Geneviève)	G									

Etape	Description	QUI ?	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb
	Répartition de l'effort sur la durée et dans l'équipe (Olivier)	O									
	Bibliographie (Geneviève)	G									
	Difficultés rencontrées lors du projet (Olivier)	O									
	Bilan & Suite du projet (Joseph)	J									
	Diagramme de Gantt (Olivier)	O									
	Description des fichiers de code (Joseph)	J									

Description des répertoires et fichiers de code (Joseph)

01_Data_Visualization/

- **common_pyAusRainfall_dataviz.ipynb :**
  - exploration, visualisation et analyse du jeu de données original

02_Data_Preprocessing/

- **common_pyAusRainfall_features_selection.ipynb :**
  - étude standalone de sélection des variables quantitatives
- **common_pyAusRainfall_preprocessing.ipynb :**
  - préparation du jeu de données

03_Data_Modeling/

- **common_pyAusRainfall_modelisation_with_resampling.ipynb :**
  - étude de prévision de précipitations à J+1 par l'analyse de modèles de classification avec méthode de sous-échantillonnage
- **common_pyAusRainfall_modelisation_without_resampling.ipynb :**
  - étude de prévision de précipitations à J+1 par l'analyse de modèles de classification sans méthode de sous-échantillonnage
- **common_pyAusRainfall_temperature_forecasting.ipynb :**
  - étude standalone de prévision des températures par l'analyse du modèle de régression *"Gradient Boosting"*
- **common_pyAusRainfall_time_series_by_climate_type.ipynb :**
  - analyse temporelle générale de la pluviométrie en fonction du type de climat
- **common_pyAusRainfall_time_series_by_location.ipynb :**
  - analyse temporelle de la pluviométrie à J+1, J+3 et J+7 en fonction d'une ville

04_Model_Interpretability/

- **common_pyAusRainfall_DecisionTree.ipynb :**
  - interprétabilité du modèle de classification *"Decision Tree"*
- **common_pyAusRainfall_KNearestNeighbors.ipynb :**
  - interprétabilité du modèle de classification *"K-Nearest Neighbors"*
- **common_pyAusRainfall_LogisticRegression.ipynb :**
  - interprétabilité du modèle de classification *"Logistic Regression"*
- **common_pyAusRainfall_RandomForest.ipynb :**

        - interprétabilité du modèle de classification *"Random Forest"*

- **common_pyAusRainfall_SupportVectorMachines.ipynb :**
    - interprétabilité du modèle de classification *"Support Vector Machines"*

data/

- stockage des différents jeux de données utilisés

old/

- contenu obsolète

score/

- stockage des scores obtenus par les différents modèles

GUIDELINES :

- consignes, répartition du travail et échéances des différentes itérations du projet

NOTES :

- définition des variables, cadre méthodologique et suggestions d'améliorations