**Weather in Australia**

Forage de données

Baptiste MARTEL, Antoine DUTEYRAT, Joshua, Guillaume

# **Présentation des données**

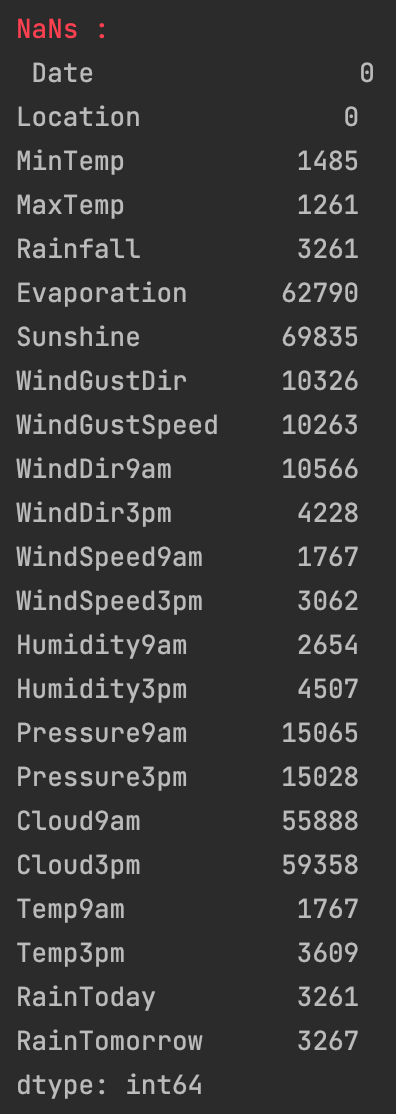
Pour mener à bien TP, nous avons choisi d’étudier un ensemble de données numériques basé sur le thème de la météorologie. Ces données, ci-dessous ont été collectées dans le but de prédire le temps du lendemain : pluvieux ou non, ce qui représente nos deux classes.

En enlevant notre attribut label du dataset, nous avons donc 22 variables distinctes qui nous aiderons à réaliser notre classification :

*Date, Location, MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, RainToday, RainTomorrow.*

Nous avons pu observer un total de 112 925 instances : 87 906 appartenant à la classe où il ne pleut pas le lendemain et 25 019 appartenant à l’autre classe. Cela crée un déséquilibre de classe.

# **Vérification et prétraitement des données**

Une image contenant table

Description générée automatiquement

On constate sur la capture d’écran de gauche que beaucoup de valeurs manquantes sont présentes dans le dataset de départ. De plus, il faut résoudre le problème d’équilibre.

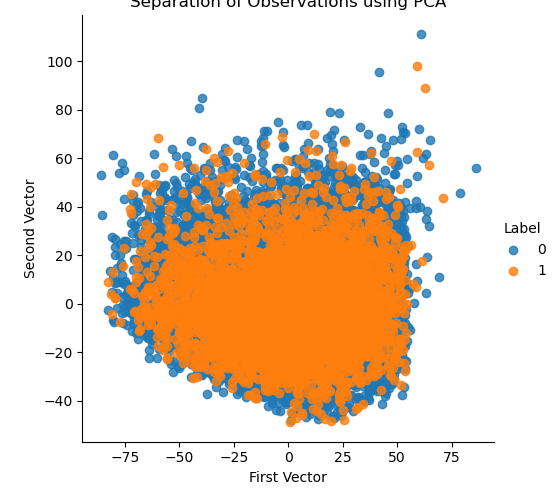
Nous avons donc effectué trois étapes pour arriver à un dataset exploitable :

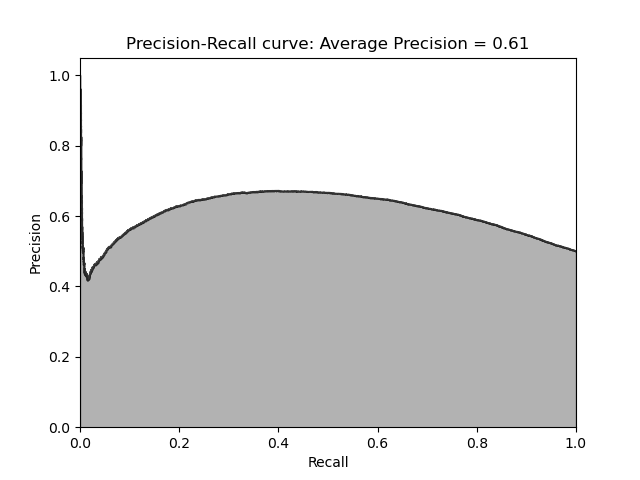
* Nous avons d’abord supprimé les attributs que nous trouvions inutiles et/ou qui comportaient trop de données manquantes.
* Nous avons ensuite encodé nos données, car certaines d’entre elles n’étaient pas numériques de base (ex : Wind Direction). Cette étape a globalement rendu les valeurs catégoriques, non-catégoriques.
* Enfin, nous avons supprimé certaines données appartenant à la classe majoritaire pour arriver à un ratio de 50:50.

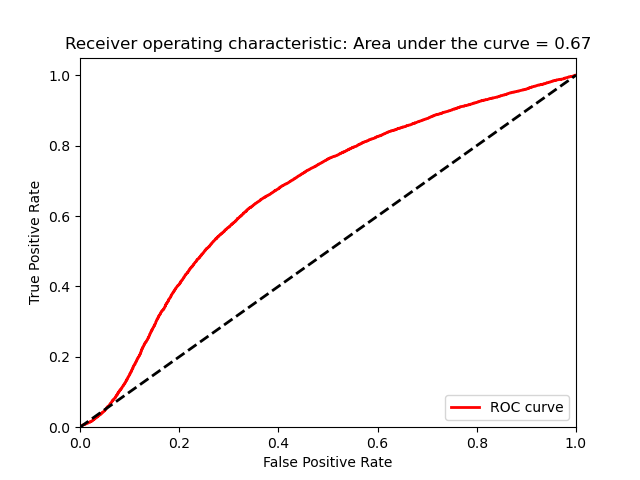
Après tout ce traitement, nous arrivons au dataset décrit dans la capture d’écran de droite.

Phase de Test :

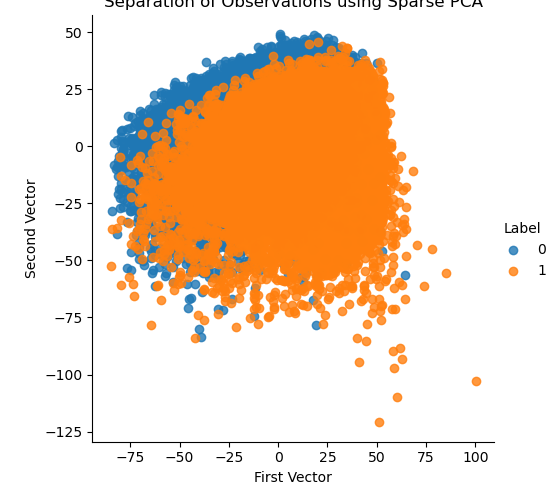
PCA :

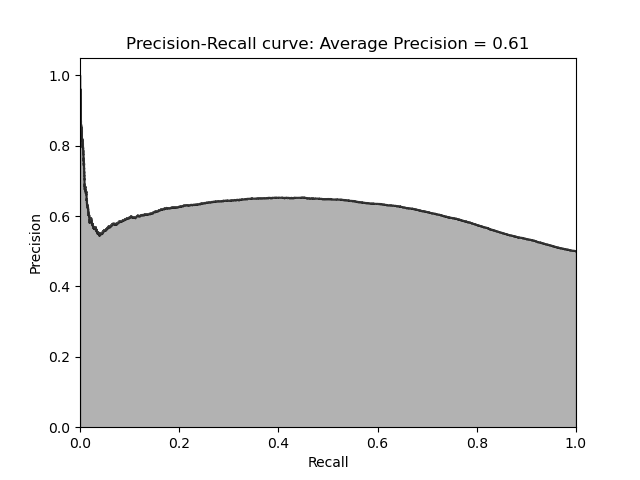


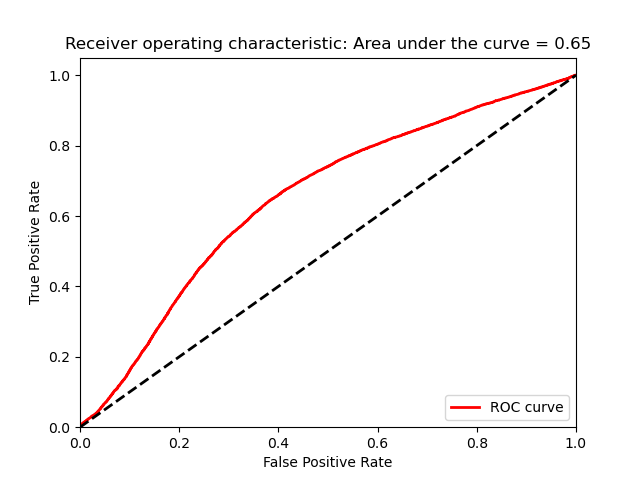




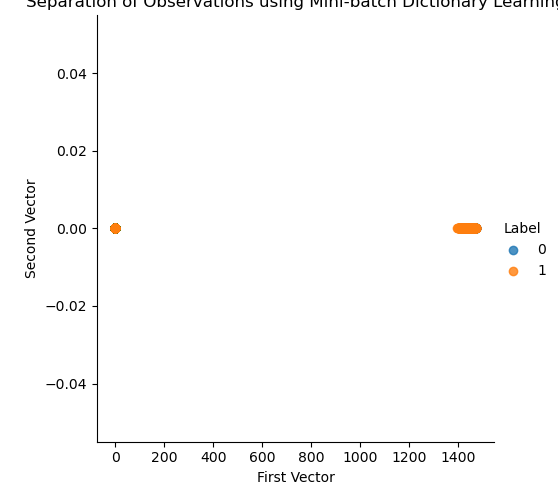
Sparse PCA :

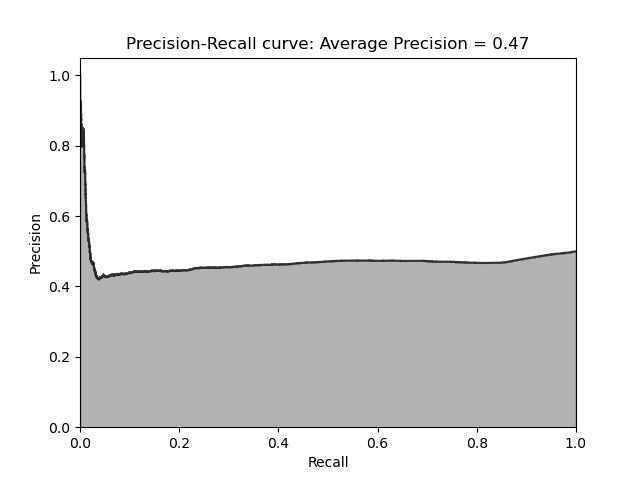


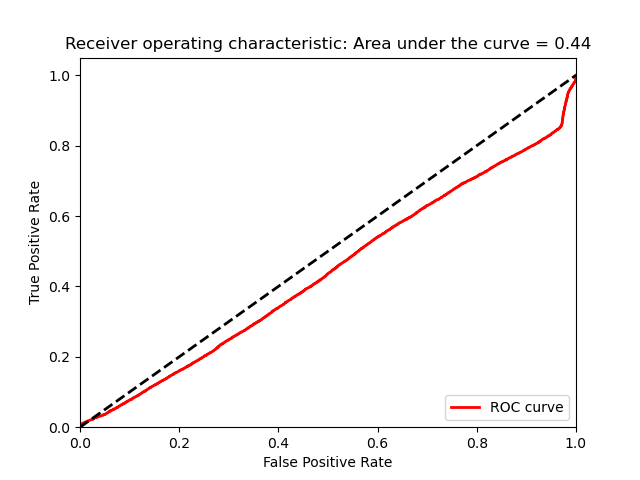




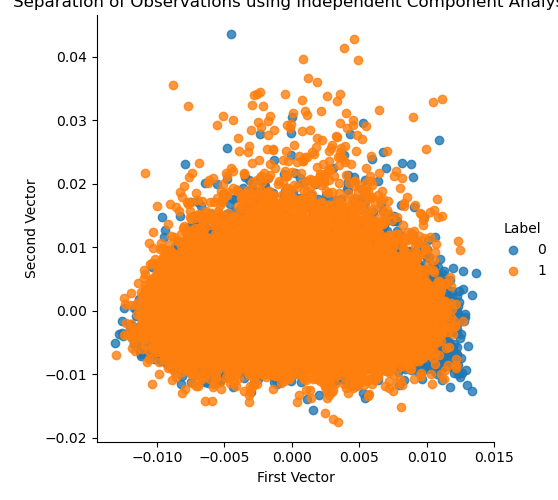
Dictionnaire :

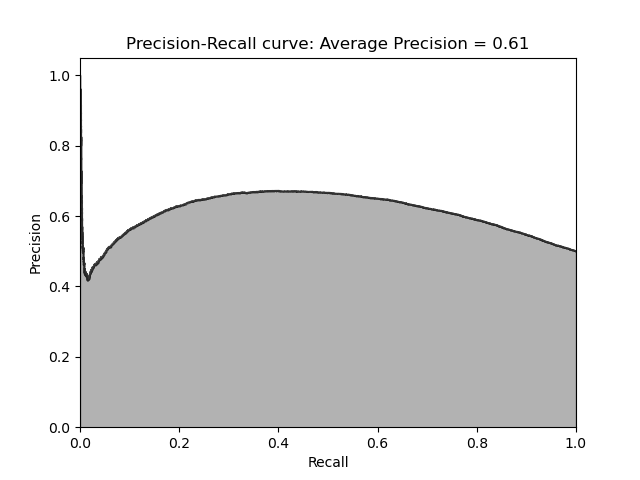


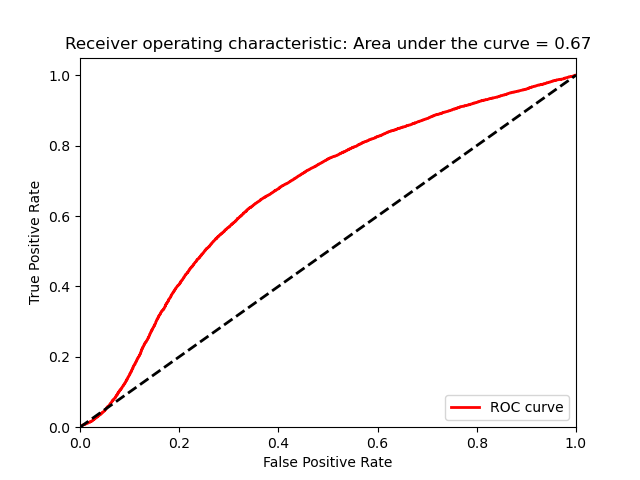




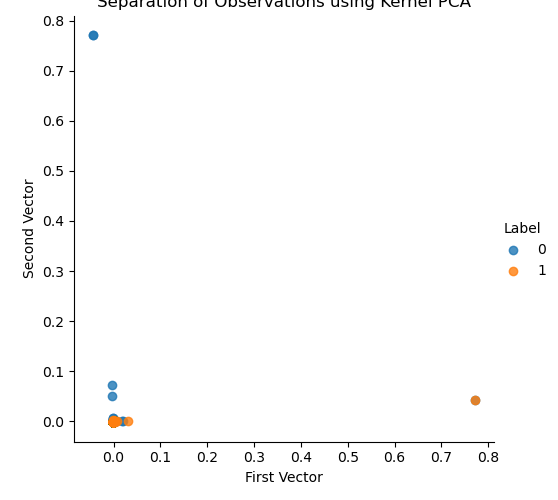
Fast-ICA :

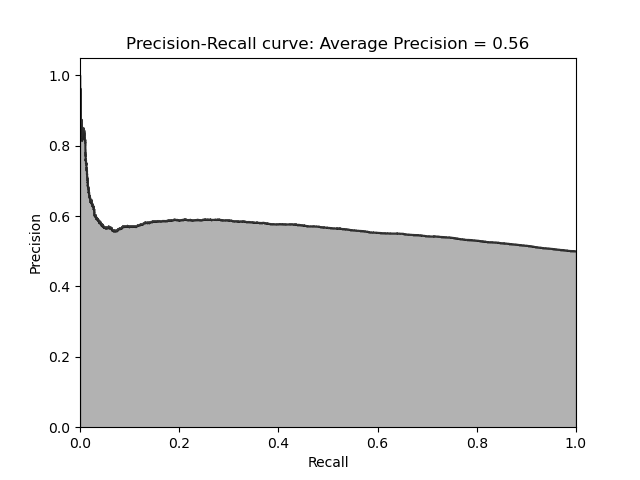


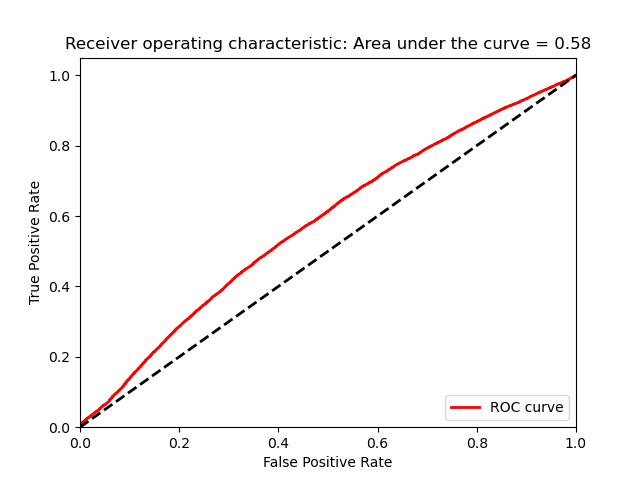




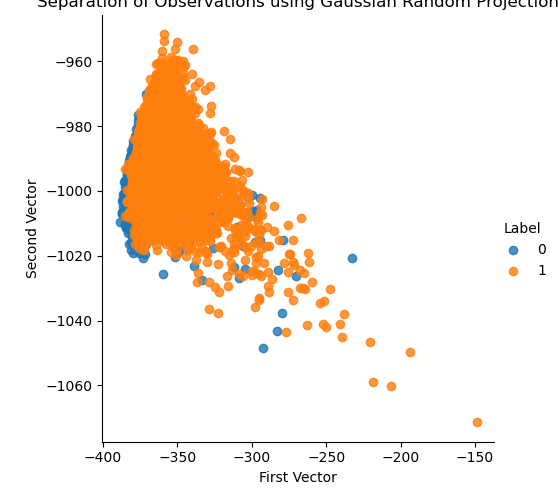
Kernel PCA :

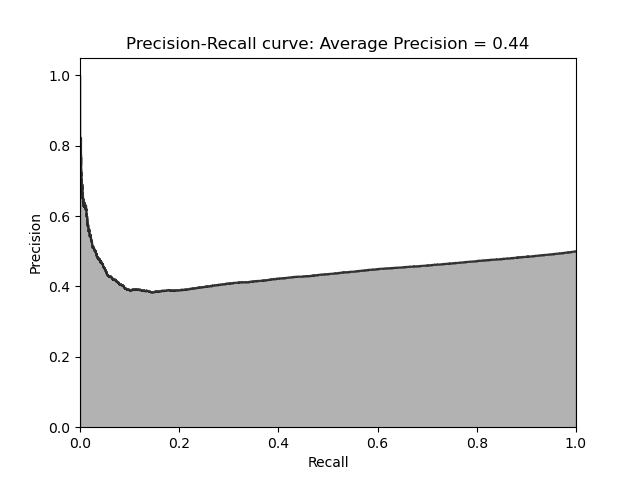




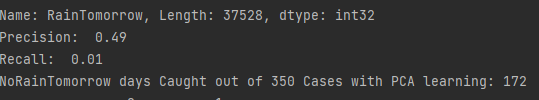


Gaussian-Random projection :





Les tests les plus prometteurs sont pour l’instant le PCA standard, Le sparse PCA et le fast ICA. Ils semblent être les plus précis. Bien sûr, dans le dataset présent, le rapport d’anomalie n’est pas aussi significatif car le data-set n’est pas débalancé, il n’y a donc pas de différence marquer dans la difficulté de reconstruction ce qui nous donne pour l’instant les résultats suivants et ce pour L’entièreté des algorithme utilisé:



Nous continuerons donc avec les trois plus prometteurs pour l’instant.