Prédiction de la variable « RainTomorrow » par les méthodes de classification.

Dans cette partie, nous utilisons toutes les données dans la base de données afin de prédire cette variable à l’aide de modèles de classification tels que Logistic Regression, Decision Tree, Random Forest et XGBoost.

Pour chaque modèle nous avons optimisé les hyperparamètres à l’aide de la bibliothèque Hyperopt, en cherchant à maximiser diverses métriques telles que l’accuracy, la précision, le recall, le score F1 et le ROC AUC.

Nous avons évalué chaque modèle optimisé en utilisant un ensemble de métriques significatives, notamment l’accuracy, la précision, le recall, le score F1 et le ROC AUC. Les résultats globaux démontrent que le modèle XGBoost surpasse les autres dans la prédiction des conditions météorologiques en Australie.

Nous avons également inclus des éléments visuels tels que des matrices de confusion, des graphiques de courbes ROC AUC, et des analyses des caractéristiques les plus importantes pour chaque modèle. Ces éléments permettent une compréhension approfondie de la performance de chaque modèle.

L’utilisation de SHAP (Shapley Additive exPlanations) pour évaluer la contribution de chaque variable explicative a été cruciale.

La section 1 compare les résultats des différents modèles en fonction de la métrique qu’on cherche à maximiser.

Les autres sections présentent des résultats détaillés pour chaque modèle.

# Comparaison des résultats des modèles

## Maximisation de l’accuracy

Les résultats présentés dans le tableau ci-dessous montrent les performances de quatre modèles différents en termes de cinq métriques d'évaluation, avec l'optimisation basée sur l’accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scores des modèles ayant meilleur accuracy** | | | | | |
|  | **accuracy** | **recall** | **precision** | **f1** | **auc** |
| LogisticRegression | 0.8461 | 0.5004 | 0.7278 | 0.5930 | 0.8694 |
| TreeDecision | 0.8389 | 0.4841 | 0.7049 | 0.5740 | 0.8456 |
| RandomForest | 0.8465 | 0.4761 | 0.7477 | 0.5818 | 0.8653 |
| **XGBoost** | **0.8526** | 0.5407 | 0.7315 | 0.6218 | 0.8844 |

* Accuracy : XGBoost a la meilleure accuracy parmi les quatre modèles, avec une valeur de 0.8526
* Recall : XGBoost a également le recall le plus élevé, ce qui suggère qu'il a une capacité supérieure à identifier les exemples positifs par rapport aux autres modèles.
* Precision : Random Forest a la precision la plus élevée, indiquant qu'il a moins de faux positifs par rapport aux autres modèles.
* F1-score : XGBoost a le F1-score le plus élevé, ce qui est une combinaison équilibrée de recall et precision.
* AUC : XGBoost a également la meilleure AUC, indiquant de bonnes performances globales pour la classification binaire.

Si l’objectif principal est d’optimiser l’accuracy, le modèle XGBoost semble être le meilleur choix

On peut observer dans le graphique ci-dessous que la ROC courbe du modèle XGBoost est au-dessus des ROC courbes des autres modèles. Cela suggère que le modèle XGBoost a une meilleure capacité à maintenir un taux élevé de vrais positifs tout en limitant le taux de faux positifs, même à différents seuils de classification.

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

## Maximisation de la précision

Les résultats présentés dans le tableau ci-dessous montrent les performances de quatre modèles différents en termes de cinq métriques d'évaluation, avec l'optimisation basée sur la précision.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scores des modèles ayant meilleur precision** | | | | | |
|  | **accuracy** | **recall** | **precision** | **f1** | **auc** |
| LogisticRegression | 0.8455 | 0.4894 | 0.7327 | 0.5869 | 0.8681 |
| TreeDecision | 0.8233 | 0.2987 | 0.7746 | 0.4311 | 0.7792 |
| **RandomForest** | 0.7830 | 0.0367 | **0.8897** | 0.0705 | 0.8315 |
| XGBoost | 0.8410 | 0.4144 | 0.7698 | 0.5388 | 0.8567 |

* Précision : La precision mesure la proportion d'exemples positifs parmi ceux que le modèle a identifiés comme positifs. Dans ce cas, Random Forest a la precision la plus élevée (0.8897), indiquant qu'il a une capacité à minimiser les faux positifs par rapport aux autres modèles.
* Accuracy : La précision seule ne donne pas une image complète de la performance du modèle, car elle ne prend pas en compte les faux négatifs. En termes d'accuracy, Linear Regression a la valeur la plus élevée (0.8455)
* Recall : Random Forest a le recall le plus faible (0.0367), ce qui signifie qu'il identifie très peu d'exemples positifs parmi tous les exemples réellement positifs.
* F1-score : XGBoost a un F1-score relativement équilibré, combinant recall et precision de manière équilibrée.
* AUC : Random Forest a la plus haute AUC dans ce cas (0.8315).

Si la précision est la métrique la plus importante, Random Forest pourrait être le choix préféré en se basant sur ces résultats spécifiques.

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

## Maximisation du recall

Les résultats présentés dans le tableau ci-dessous montrent les performances de quatre modèles différents en termes de cinq métriques d'évaluation, avec l'optimisation basée sur le recall.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scores des modèles ayant meilleur recall** | | | | | |
|  | **accuracy** | **recall** | **precision** | **f1** | **auc** |
| LogisticRegression | 0.7366 | **0.8464** | 0.4531 | 0.5902 | 0.8640 |
| TreeDecision | 0.8313 | 0.4944 | 0.6669 | 0.5679 | 0.8180 |
| RandomForest | 0.8397 | 0.5062 | 0.6956 | 0.5860 | 0.8469 |
| XGBoost | 0.8292 | 0.5547 | 0.6365 | 0.5928 | 0.8481 |

* Recall : Recall mesure la proportion d'exemples positifs réellement identifiés par le modèle. Linear Regression a le recall le plus élevé (0.8464), indiquant qu'il a la meilleure capacité parmi les modèles à capturer la majorité des exemples positifs. Cependant, il est important de noter que cette performance peut être associée à une baisse de précision.
* Precision : Linear Regression a la precision la plus faible (0.4531), indiquant qu'il a tendance à identifier un grand nombre de faux positifs.
* Accuracy : Random Forest a la valeur la plus élevée en termes d'accuracy (0.8397), mais cela peut être dû à une balance entre les performances en termes de faux positifs et de faux négatifs.
* F1-score : Le F1-score de Linear Regression est relativement équilibré, étant donné son recall élevé et sa precision basse.
* AUC : AUC mesure la capacité du modèle à discriminer entre les classes positives et négatives. XGBoost a la plus haute AUC dans ce cas (0.8481).

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

## Maximisation du F1-score

Les résultats présentés dans le tableau ci-dessous montrent les performances de quatre modèles différents en termes de cinq métriques d'évaluation, avec l'optimisation basée sur le F1-score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scores des modèles ayant meilleur f1** | | | | | |
|  | **accuracy** | **recall** | **precision** | **f1** | **auc** |
| LogisticRegression | 0.8460 | 0.5002 | 0.7277 | 0.5929 | 0.8695 |
| TreeDecision | 0.8409 | 0.4883 | 0.7114 | 0.5791 | 0.8498 |
| RandomForest | 0.8494 | 0.5049 | 0.7409 | 0.6006 | 0.8699 |
| **XGBoost** | 0.8554 | 0.5475 | 0.7397 | **0.6292** | 0.8852 |

* F1-score : Le modèle XGBoost affiche le F1-score le plus élevé (0.6292) parmi tous les modèles. Le F1-score est une métrique qui prend en compte à la fois la precision et le recall. Cela suggère que XGBoost atteint un équilibre optimal entre l'identification des vrais positifs (recall) et la minimisation des faux positifs (precision).
* Accuracy : XGBoost a également une accuracy élevée (0.8554), indiquant une classification correcte d'une grande proportion des exemples.
* Recall et Precision : XGBoost a des valeurs de recall et de precision compétitives par rapport aux autres modèles, ce qui renforce l'idée d'un équilibre entre la sensibilité aux vrais positifs et la limitation des faux positifs.
* AUC : XGBoost présente également la plus haute AUC (0.8852), ce qui confirme sa capacité à bien discriminer entre les classes positives et négatives.

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

## Maximisation de l’AUC

Les résultats présentés dans le tableau ci-dessous montrent les performances de quatre modèles différents en termes de cinq métriques d'évaluation, avec l'optimisation basée sur l’AUC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scores des modèles ayant meilleur auc** | | | | | |
|  | **accuracy** | **recall** | **precision** | **f1** | **auc** |
| LogisticRegression | 0.8460 | 0.5004 | 0.7273 | 0.5929 | 0.8695 |
| TreeDecision | 0.8399 | 0.4478 | 0.7341 | 0.5563 | 0.8493 |
| RandomForest | 0.8497 | 0.4980 | 0.7474 | 0.5978 | 0.8716 |
| **XGBoost** | 0.8538 | 0.5358 | 0.7400 | 0.6216 | **0.8844** |

* AUC : L'AUC (Area Under the Curve) est souvent utilisée comme métrique pour évaluer les modèles de classification binaire. Dans ce cas, XGBoost a la valeur d'AUC la plus élevée (0.8844), ce qui indique une performance globale solide pour la classification binaire. Cela suggère que le modèle XGBoost a une bonne capacité à discriminer entre les classes positives et négatives.
* Accuracy : XGBoost a également la meilleure accuracy (0.8538), ce qui signifie qu'il a correctement classé une grande proportion des exemples
* Recall : XGBoost a le recall le plus élevé, indiquant qu'il a une capacité supérieure à identifier les exemples positifs par rapport aux autres modèles.
* Precision : Random Forest a la precision la plus élevée, ce qui suggère qu'il a moins de faux positifs par rapport aux autres modèles.
* F1-score : XGBoost a le F1-score le plus élevé, ce qui est une combinaison équilibrée de recall et precision

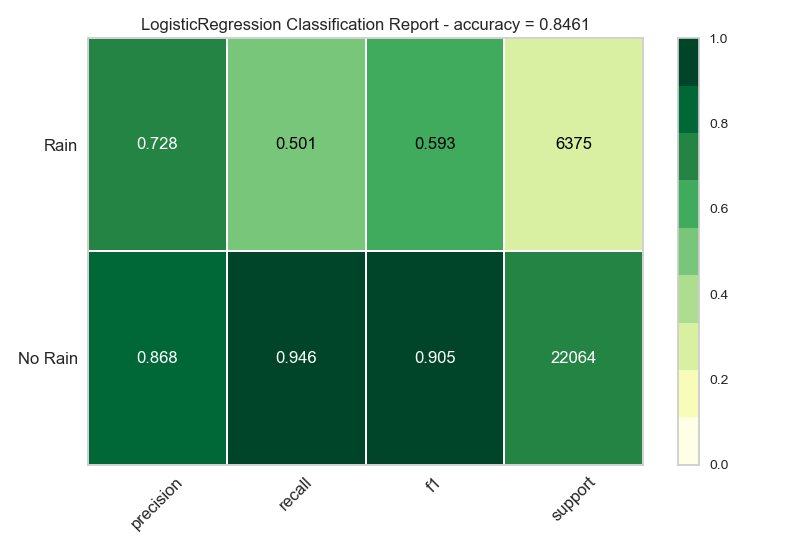
Les résultats suggèrent que XGBoost est le modèle qui offre la meilleure performance globale, en particulier en termes d'AUC, accuracy, recall, precision et F1-score.

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

# Modèle Logistic Regression

## Maximisation de l’accuracy



Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

## Maximisation de la précision

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

## Maximisation du recall

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

## Maximisation du score f1

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

## Maximisation de la ROC AUC

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

# Modèle Decision Tree

## Maximisation de l’accuracy

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran, Parallèle

Description générée automatiquement

## Maximisation de la précision

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

## Maximisation du recall

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

## Maximisation du score f1

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran, Parallèle

Description générée automatiquement

## Maximisation de la ROC AUC

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran, Parallèle

Description générée automatiquement

# Modèle Random Forest

## Maximisation de l’accuracy

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle

Description générée automatiquement

## Maximisation de la précision

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle

Description générée automatiquement

## Maximisation du recall

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle

Description générée automatiquement

## Maximisation du score f1

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

## Maximisation de la ROC AUC

Une image contenant texte, capture d’écran, carré, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

# Modèle XGBoost

## Maximisation de l’accuracy

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

## Maximisation de la précision

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, Rectangle

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

## Maximisation du recall

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

## Maximisation du score f1

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

## Maximisation de la ROC AUC

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquementUne image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, Parallèle, ligne

Description générée automatiquement

# SHAP pour le modèle XGBoost

## Mean SHAP plot

Pour déterminer quelles caractéristiques sont généralement les plus importantes pour les prédictions de notre modèle, nous pouvons utiliser un diagramme à barres des valeurs moyennes de SHAP pour toutes les observations. Prendre la moyenne des valeurs absolues garantit que les valeurs positives et négatives ne s’annulent pas.

On observe dans le graphique ci-dessous que la variable avec la valeur SHAP moyenne la plus élevée est Humidity3pm, ce qui indique qu’elle a l’impact le plus important sur les prédictions de notre modèle. Ces informations peuvent nous aider à comprendre quelles variables sont essentielles au processus décisionnel du modèle.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

## Beeswarm Plot

Le graphique Beeswarm est une visualisation utile pour examiner toutes les valeurs SHAP pour chaque entité. L'axe Y regroupe les valeurs SHAP par caractéristique, la couleur des points indiquant la valeur de la caractéristique correspondante. En règle générale, les points les plus rouges représentent des valeurs de caractéristiques plus élevées.

Le beeswarm plot peut aider à identifier les relations importantes entre les caractéristiques et les prédictions du modèle. Dans ce graphique, les caractéristiques sont classées selon leurs valeurs SHAP moyennes.

En examinant les valeurs SHAP dans le beeswarm plot ci-dessous, nous pouvons commencer à comprendre la nature des relations entre les variables et la pluie du lendement. Par exemple, pour Humidity3pm et WindGustSpeed, nous observons que les valeurs SHAP augmentent à mesure que la valeur de la fonctionnalité augmente. Cela suggère que des valeurs plus élevées de Humidity3pm et WindGustSpeed contribuent à la probabilité qu’il va pleuvoir demain plus élevée.

En revanche, pour la Pressure3pm et la Sunshine, nous remarquons la tendance inverse, où des valeurs de caractéristiques plus élevées conduisent à des valeurs SHAP plus faibles. Cette observation implique que des valeurs de Pressure3pm et de Sunshine plus élevées sont associées à la probabilité qu’il va pleuvoir demain plus faible.

Une image contenant texte, capture d’écran

Description générée automatiquement

## Dependence Plots

Pour mieux comprendre les relations entre les caractéristiques individuelles et leurs valeurs SHAP correspondantes, nous pouvons créer des tracés de dépendance. Un diagramme de dépendance est un nuage de points qui montre la relation entre la valeur SHAP et la valeur de caractéristique pour une seule caractéristique.

En analysant les diagrammes de dépendance, nous pouvons confirmer les observations faites dans le beeswarm plot. Par exemple, lorsque nous créons un diagramme de dépendance pour Humidity3pm et WindGustSpeed, nous observons une relation positive entre les valeurs de ces variables et les valeurs SHAP. En d’autres termes, des valeurs de ces deux variables plus élevées entraînent des prévisions de la pluie demain plus élevées.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Une image contenant texte, diagramme, Police, Tracé  Description générée automatiquement   |  | | --- | |  | |  |  |  |  |  |  | | | | | |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  | Une image contenant texte, diagramme, Tracé, ligne  Description générée automatiquement |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  | Une image contenant texte, diagramme, ligne, Tracé  Description générée automatiquement |  |  |  |  |
|  |  |  |  |  | Une image contenant texte, Tracé, ligne, diagramme  Description générée automatiquement |  |  |  |  |  |  |  |
| |  | | --- | |  | |  |  |  |  |  |  | | | | | | |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

## Waterfall Plot

Ce graphique nous aide à visualiser les valeurs SHAP de chaque échantillon dans nos données individuellement. Visualisons les valeurs SHAP du premier échantillon de test.

En ignorant les signes, l'ampleur de la valeur SHAP pour la Humidity, 0.9, est supérieure à celle des autres variables. Cela impliquait que Humidity3pm avait l'impact le plus significatif sur cette prédiction particulière.

Une image contenant texte, capture d’écran, nombre, diagramme

Description générée automatiquement

Tout comme nous avons visualisé les valeurs SHAP du premier échantillon, nous pouvons également visualiser les valeurs SHAP du deuxième échantillon de test.

Une image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquement