Mémoire de projet

DATA Engineer-2

Fire Risk Prediction

Élèves :
HALMAOUI Hajar

Encadrant :
FISSAA Tarik

Année universitaire : 2023-2024

Table des matières

1 Introduction **4**
1.1 Objectif du Projet 4
1.2 Importance 4
1.3 Portée 4

2 Collection des données **6**
2.1 Données sur les incendies 6
2.2 Données météorologiques 7
2.2.1 Sources de Données Météorologiques 7
2.2.2 Données Collectées 7
2.3 données sur la végétation 8
2.3.1 Sources de Données de Végétation 8
2.3.2 Données Collectées 8
2.3.3 Méthodologie d'Extraction des Données 8
2.4 Jeux de donnée 9

3 Prétraitement des données **10**
3.1 Nettoyage des données 10
3.1.1 Gestion des Valeurs Manquantes 10
3.1.2 Suppression des Doublons 10
3.2 ingénierie des fonctionnalités 10
3.3 Normalisation des Variables 11

4 Analyse exploratoire des données (EDA) **13**
4.1 Visualisation 13
4.2 Analyse de Corrélation 14
4.3 Analyse de Distribution 14

5 Sélection du modèle **16**
5.1 Baseline Models 16
5.2 Advanced Models : 16

6 Model Training and Evaluation **16**
6.1 Logistic Regression 16
6.2 Decision Tree 17
6.3 Random Forest 18
6.4 XGBoost 18
6.5 Réseau de Neurones (Neural Network) 19

7 Résultats et Discussion **20**
7.1 Comparaison des Performances des Modèles 20
7.2 Forces et Faiblesses de Chaque Approche 20
7.3 Choix du Modèle Optimal 20

8 Déploiement du Modèle **22**
8.1 Sauvegarde du Modèle 22
8.2 Application de Prédiction de Risque d'Incendie avec Streamlit 22
8.3 Choix du Facteur dans Streamlit 23
8.4 Déploiement de l'Application 23

9 Conclusion **25**

Table des figures

2 dataset . . . 6
3 Code d'extraction des données météorologiques . . . 7
4 Interface MODIS pour télecharger les données de vegetation . . . 8
5 Code pour joindre les données de végétation avec les données d'incendie . . . 9
6 Table de jeu de donnée finale . . . 9
7 les colonnes de jeux de donnée final . . . 9
8 somme de valeurs manquante dans chaque colonne . . . 10
9 code pour éliminer les lignes doublons . . . 10
10 code pour génerer des nouveau colonnes . . . 11
11 Conditions d'occurrence d'incendie . . . 11
12 Colonne d'indice IDVI . . . 11
13 Indice de sécheresse . . . 11
14 Code de normalisation . . . 11
15 boxplot de la distribution de la température en fonction de l'occurrence des incendies . . . 13
16 boxplot de la distribution de l'humidité en fonction de l'occurrence des incendies . . . 13
17 matrice de corrélation . . . 14
18 Distribution de la Température pour les Instances d'Incendie et de Non-Incendie . . . 14
19 Distribution de l'Humidité pour les Instances d'Incendie et de Non-Incendie 15
20 Rapport de classification pour La régression logistique . . . 17
21 Matrice de confusion . . . 17
22 Rapport de classification pour Les arbres de décision . . . 17
23 Matrice de confusion . . . 18
24 Rapport de classification pour Le Random Forest . . . 18
25 Matrice de confusion . . . 18
26 Rapport de classification pour XGBoost . . . 19
27 Matrice de confusion . . . 19
28 Rapport de classification pour Les réseaux de neurones . . . 19
29 Matrice de confusion . . . 19
30 code pour télecharger le modéle choisit . . . 22
31 Code de deploiement du modèle dans streamlit . . . 23
32 Interface de l'application streamlit . . . 24

1 Introduction

1.1 Objectif du Projet

L'objectif principal de ce projet est de prédire le risque d'incendie dans diverses régions du France en utilisant des techniques avancées d'apprentissage automatique. En exploitant un ensemble diversifié de sources de données, y compris les données historiques sur les incendies, les conditions météorologiques, la santé de la végétation et les informations géospatiales, le projet vise à développer un modèle prédictif robuste capable d'identifier avec précision les occurrences potentielles d'incendies. Ce modèle peut servir d'outil précieux pour les autorités et les parties prenantes afin de réduire les dommages liés aux incendies en permettant des mesures proactives et des interventions en temps opportun.

1.2 Importance

Les incidents d'incendie représentent des menaces significatives pour la vie humaine et l'environnement, causant des destructions et des pertes économiques considérables. Prédire le risque d'incendie est crucial pour mettre en œuvre des stratégies de prévention efficaces et une allocation optimale des ressources. Une prédiction précise du risque d'incendie aide à :

- Prévenir les pertes en vies humaines et en biens : Les systèmes de détection et d'alerte précoce permettent des évacuations et des actions protectrices en temps opportun, réduisant ainsi l'impact sur les communautés.
- Protection de l'environnement : Minimiser les incendies permet de préserver la biodiversité, d'éviter la dégradation des sols et de maintenir l'équilibre des écosystèmes.
- Optimisation des ressources : Le déploiement efficace des ressources de lutte contre les incendies et des services d'urgence garantit que les zones à haut risque reçoivent une attention adéquate, optimisant ainsi l'utilisation des ressources limitées.
- Politique et planification : Informer les décideurs et les urbanistes sur les zones sujettes aux incendies aide à élaborer de meilleures stratégies d'utilisation des terres et des politiques de gestion des incendies.

1.3 Portée

Ce projet comprend les composantes clés suivantes :

1. Collecte de Données : Agrégation des données provenant de multiples sources disponibles publiquement, y compris :
   - Données sur les Incendies : Dossiers historiques de MODIS (Moderate Resolution Imaging Spectroradiometer), Fire Information for Resource Management System (FIRMS) et bases de données nationales sur les incendies.
   - Données Météorologiques : Données météorologiques couvrant la température, l'humidité, la vitesse du vent et les précipitations.
   - Données sur la Végétation : Imagerie satellite et bases de données d'utilisation des terres fournissant des informations sur la santé de la végétation et la couverture des terres.
   - Données Géospatiales : Cartes d'altitude et proximité des plans d'eau.

2. Prétraitement des Données : Nettoyage et préparation des données collectées, y compris la gestion des valeurs manquantes, l'ingénierie des caractéristiques et la normalisation.
3. Analyse Exploratoire des Données (EDA) : Visualisation et analyse des relations entre différentes caractéristiques et l'occurrence des incendies pour obtenir des insights et guider le développement du modèle.
4. Développement du Modèle : Construction et évaluation de divers modèles d'apprentissage automatique, y compris :
   — Modèles de base tels que la Régression Logistique et les Arbres de Décision.
   — Modèles avancés comme le Random Forest, XGBoost et les Réseaux de Neurones.
   — Optimisation des hyperparamètres et validation croisée pour améliorer la performance du modèle.
5. Évaluation du Modèle : Utilisation de métriques telles que l'exactitude, la précision, le rappel, le score F1 et le ROC-AUC pour évaluer l'efficacité et la robustesse du modèle.

2 Collection des données

2.1 Données sur les incendies

Sources : MODIS

MODIS fournit des données détaillées sur les incendies historiques dans le monde entier. Cela inclut des informations détaillées telles que la situation géographique (latitude et longitude), luminosité du feu, niveau de confiance dans la détection et puissance radiative du feu (PRF). Les données sont inestimables pour analyser les schémas d'incendie, comprendre l'environnement impacts et évaluer les risques d'incendie dans différentes régions.

Description :

— latitude : La latitude du point de détection de feu. C'est une mesure géographique indiquant la position nord-sud sur la surface de la Terre.

— longitude : La longitude du point de détection de feu. C'est une mesure géographique indiquant la position est-ouest sur la surface de la Terre.

— brightness : La luminosité du feu mesurée par le satellite. Une valeur plus élevée indique une température plus élevée et potentiellement un feu plus intense.

— scan : La dimension du scan en kilomètres. Elle représente la taille de la zone scannée par le satellite dans le sens du déplacement.

— track : La dimension du track en kilomètres. Elle représente la taille de la zone scannée par le satellite perpendiculairement au sens du déplacement.

— acq_date : La date de l'acquisition des données par le satellite, au format AAAA-MM-JJ.

— acq_time : L'heure de l'acquisition des données par le satellite, au format HHMM (24 heures).

— satellite : Le nom du satellite ayant acquis les données. Les valeurs possibles sont généralement "Terra" ou "Aqua".

— instrument : Le nom de l'instrument utilisé par le satellite pour mesurer les données. Dans ce cas, il s'agit de "MODIS" (Moderate Resolution Imaging Spectroradiometer).

— confidence : Le niveau de confiance de la détection de feu, mesuré en pourcentage. Une valeur plus élevée indique une plus grande certitude que la détection est un véritable feu.

— version : La version de l'algorithme de détection de feu utilisé pour traiter les données.

— bright_t31 : La température de brillance au canal 31, mesurée en Kelvin. Cela donne une idée de la température de surface du feu détecté.

— frp : La puissance radiative du feu (Fire Radiative Power), mesurée en mégawatts. Elle donne une estimation de l'énergie émise par le feu.

— daynight : Indique si la détection a été faite de jour ("D") ou de nuit ("N").

— type : Le type de détection, avec des valeurs possibles telles que 0 (feu de végétation), 1 (feu de type présumé autre), etc.

	latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_t31	frp	daynight	type	temp	wind_speed	description	humidity
0	34.8313	-4.3160	301.1	1.3	1.1	2023-01-02	1107	Terra	MODIS	31	61.03	286.6	8.1	D	0	21.5	2.880293	Clear sky	41
1	33.8011	-4.4965	306.5	1.0	1.0	2023-01-06	1341	Aqua	MODIS	62	61.03	294.4	4.5	D	0	23.9	2.646896	Clear sky	47
2	32.8963	-3.8139	305.4	1.1	1.1	2023-01-08	1327	Aqua	MODIS	58	61.03	293.9	4.7	D	0	23.9	1.204081	Clear sky	39
3	31.8369	-4.2648	312.1	1.9	1.3	2023-01-10	1314	Aqua	MODIS	65	61.03	296.3	17.8	D	0	29.8	4.661707	Clear sky	20
4	31.8331	-4.2838	337.0	1.9	1.3	2023-01-10	1314	Aqua	MODIS	89	61.03	295.5	88.3	D	0	29.7	4.954858	Clear sky	19

FIGURE 2 – dataset

2.2 Données météorologiques

2.2.1 Sources de Données Météorologiques

Fournisseurs de Données Météorologiques Pour obtenir des données météorologiques précises et en temps réel, nous avons utilisé les services de Weatherbit. Weatherbit fournit des API permettant d'accéder à diverses données météorologiques, y compris les conditions actuelles, les prévisions, les données historiques et les alertes météorologiques.

2.2.2 Données Collectées

Nous avons extrait les données suivantes pour chaque détection de feu :

— Température (temp) : La température ambiante, mesurée en degrés Celsius, au moment de la détection du feu.
— Vitesse du Vent (wind_speed) : La vitesse du vent au moment de la détection du feu, mesurée en mètres par seconde.
— Description (description) : Une description qualitative des conditions atmosphériques au moment de la détection (par exemple, "Clear sky" pour ciel dégagé).
— Humidité (humidity) : Le taux d'humidité relative, mesuré en pourcentage, au moment de la détection du feu.

Méthodologie d'Extraction des Données
Les données météorologiques ont été récupérées en utilisant une API fournie par Weatherbit. Voici un exemple de code Python utilisé pour cette extraction :

```python
input_file = 'modis_2023_France.csv'
output_file = 'coordinates_updated3.csv'
api_key = 'dddda3a28cf44f32bea55d694b901a70'
start_date = datetime( year: 2023, month: 1, day: 1)

1 usage
def fetch_weather_data(latitude, longitude):
    url = f'https://api.weatherbit.io/v2.0/current?lat={latitude}&lon={longitude}&key={api_key}&units=M'

    try:
        res = requests.get(url)
        res.raise_for_status()
        return res.json()

    except requests.exceptions.RequestException as e:
        print(f'Erreur lors de la requête API : {e}')
        return None
```

```python
def main():
    try:
        with open(input_file, 'r', newline='') as file:
            reader = csv.reader(file)
            headers = next(reader)
            data_rows = []
            for row in reader:
                try:
                    acq_date = datetime.strptime(row[5], _format: '%Y-%m-%d')
                except ValueError:
                    print(f"Ignoré: Date invalide pour la ligne {reader.line_num}: {row[5]}")
                    continue

                if acq_date >= start_date:
                    data_rows.append({
                        'latitude': row[0],
                        'longitude': row[1],
                        'acq_date': row[5],
                        'temp': '',
                        'wind_speed': '',
                        'description': '',
                        'humidity': '',
                    })
```

FIGURE 3 – Code d'extraction des données météorologiques

2.3 données sur la végétation

2.3.1 Sources de Données de Végétation

Fournisseurs de Données de Végétation Pour obtenir des données précises sur la végétation, nous avons utilisé les images satellitaires et les bases de données d'utilisation des terres fournies par MODIS (Moderate Resolution Imaging Spectroradiometer). MODIS est un instrument à bord des satellites Terra et Aqua de la NASA, qui fournit des données essentielles pour surveiller la végétation, les écosystèmes terrestres et les conditions environnementales.

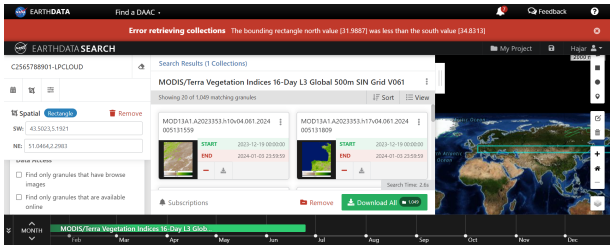


FIGURE 4 – Interface MODIS pour télecharger les données de vegetation

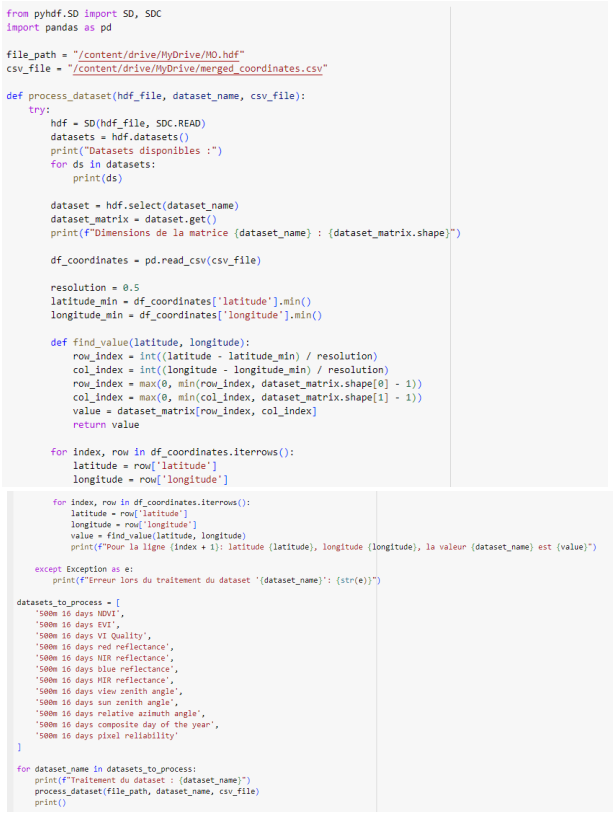
2.3.2 Données Collectées

Nous avons extrait les données suivantes pour chaque détection de feu :

— NDVI (Normalized Difference Vegetation Index) : Indice de végétation qui mesure la santé de la végétation en utilisant la réflectance dans les bandes rouge et proche infrarouge.
— EVI (Enhanced Vegetation Index) : Indice de végétation amélioré qui corrige certains effets atmosphériques et de sol.
— VI Quality : Indicateur de la qualité des indices de végétation.
— Red Reflectance : Réflectance de la bande rouge.
— NIR Reflectance : Réflectance de la bande proche infrarouge.
— Blue Reflectance : Réflectance de la bande bleue.
— MIR Reflectance : Réflectance de la bande infrarouge moyenne.
— View Zenith Angle : Angle de vision zénithal.
— Sun Zenith Angle : Angle zénithal du soleil.
— Relative Azimuth Angle : Angle azimutal relatif.
— Composite Day of the Year : Jour composite de l'année.
— Pixel Reliability : Fiabilité des pixels.

2.3.3 Méthodologie d'Extraction des Données

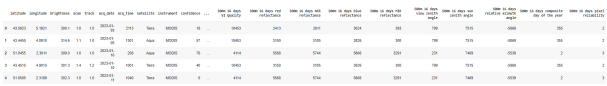
Les données de végétation ont été récupérées à partir des fichiers HDF4 de MODIS en utilisant la bibliothèque Python pyhdf. Voici un exemple de code Python utilisé pour cette extraction :



```python
from pyhdf.SD import SD, SDC
import pandas as pd

file_path = "/content/drive/MyDrive/MO.hdf"
csv_file = "/content/drive/MyDrive/merged_coordinates.csv"

def process_dataset(hdf_file, dataset_name, csv_file):
    try:
        hdf = SD(hdf_file, SDC.READ)
        datasets = hdf.datasets()
        print("Datasets disponibles :")
        for ds in datasets:
            print(ds)

        dataset = hdf.select(dataset_name)
        dataset_matrix = dataset.get()
        print(f"Dimensions de la matrice {dataset_name} : {dataset_matrix.shape}")

        df_coordinates = pd.read_csv(csv_file)

        resolution = 0.5
        latitude_min = df_coordinates['latitude'].min()
        longitude_min = df_coordinates['longitude'].min()

        def find_value(latitude, longitude):
            row_index = int((latitude - latitude_min) / resolution)
            col_index = int((longitude - longitude_min) / resolution)
            row_index = max(0, min(row_index, dataset_matrix.shape[0] - 1))
            col_index = max(0, min(col_index, dataset_matrix.shape[1] - 1))
            value = dataset_matrix[row_index, col_index]
            return value

        for index, row in df_coordinates.iterrows():
            latitude = row['latitude']
            longitude = row['longitude']
            value = find_value(latitude, longitude)
            print(f"Pour la ligne {index + 1}: latitude {latitude}, longitude {longitude}, la valeur {dataset_name} est {value}")

    except Exception as e:
        print(f"Erreur lors du traitement du dataset '{dataset_name}': {str(e)}")

datasets_to_process = [
    '500m 16 days NDVI',
    '500m 16 days EVI',
    '500m 16 days VI Quality',
    '500m 16 days red reflectance',
    '500m 16 days NIR reflectance',
    '500m 16 days blue reflectance',
    '500m 16 days MIR reflectance',
    '500m 16 days view zenith angle',
    '500m 16 days sun zenith angle',
    '500m 16 days relative azimuth angle',
    '500m 16 days composite day of the year',
    '500m 16 days pixel reliability'
]

for dataset_name in datasets_to_process:
    print(f"Traitement du dataset : {dataset_name}")
    process_dataset(file_path, dataset_name, csv_file)
    print()
```

FIGURE 5 – Code pour joindre les données de végétation avec les données d'incendie

2.4 Jeux de donnée



	latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	...	500m 16 days VI Quality	500m 16 days red reflectance	500m 16 days NIR reflectance	500m 16 days blue reflectance	500m 16 days MIR reflectance	500m 16 days view zenith angle	500m 16 days sun zenith angle	500m 16 days relative azimuth angle	500m 16 days composite day of the year	500m 16 days pixel reliability
0	43.5823	5.1921	300.1	1.0	1.0	2023-01-01	215	Terra	MODIS	91	...	19453	2413	2619	3424	365	799	7173	-5988	356	2
1	43.4466	4.8818	314.6	1.1	1.0	2023-01-01	1301	Aqua	MODIS	87	...	19453	3150	3155	3826	306	799	7173	-6068	356	2
2	51.9455	2.3911	309.9	1.0	1.0	2023-01-01	206	Aqua	MODIS	76	...	4914	5566	5744	5666	5354	254	7488	-5539	2	3
3	43.4415	4.8915	301.3	1.4	1.2	2023-01-01	1301	Terra	MODIS	48	...	19453	3150	3155	3826	306	799	7173	-6068	356	2
4	51.9505	2.3488	302.5	1.0	1.0	2023-01-01	1040	Terra	MODIS	0	...	4914	5566	5744	5666	5354	254	7488	-5539	2	3

FIGURE 6 – Table de jeu de donnée finale

```
Index(['latitude', 'longitude', 'brightness', 'scan', 'track', 'acq_date',
       'acq_time', 'satellite', 'instrument', 'confidence', 'version',
       'bright_t31', 'frp', 'daynight', 'type', 'temp', 'wind_speed',
       'description', 'humidity', '500m 16 days NDVI', '500m 16 days EVI',
       '500m 16 days VI Quality', '500m 16 days red reflectance',
       '500m 16 days NIR reflectance', '500m 16 days blue reflectance',
       '500m 16 days MIR reflectance', '500m 16 days view zenith angle',
       '500m 16 days sun zenith angle', '500m 16 days relative azimuth angle',
       '500m 16 days composite day of the year',
       '500m 16 days pixel reliability'],
      dtype='object')
```

FIGURE 7 – les colonnes de jeux de donnée final

3 Prétraitement des données

3.1 Nettoyage des données

3.1.1 Gestion des Valeurs Manquantes

Nous avons analysé les données pour identifier les valeurs manquantes. Aucune valeur manquante n'a été trouvée dans le jeu de données :

```
Valeurs vides par colonne :
latitude                                0
longitude                               0
brightness                              0
scan                                    0
track                                   0
acq_date                                0
acq_time                                0
satellite                               0
instrument                              0
confidence                              0
version                                 0
bright_t31                              0
frp                                     0
daynight                                0
type                                    0
temp                                    0
wind_speed                              0
description                             0
humidity                                0
precipitation                           0
500m 16 days NDVI                       0
500m 16 days EVI                        0
500m 16 days VI Quality                 0
500m 16 days red reflectance            0
500m 16 days NIR reflectance            0
500m 16 days blue reflectance           0
500m 16 days MIR reflectance            0
500m 16 days view zenith angle          0
500m 16 days sun zenith angle           0
500m 16 days relative azimuth angle     0
500m 16 days composite day of the year  0
500m 16 days pixel reliability          0
dtype: int64
```

FIGURE 8 – somme de valeurs manquante dans chaque colonne

3.1.2 Suppression des Doublons

Nous avons supprimé les lignes en double dans le jeu de données pour éviter toute redondance :

```
df.drop_duplicates(inplace=True)
print(df.head())
```

FIGURE 9 – code pour éliminer les lignes doublons

3.2 ingénierie des fonctionnalités

Indicateurs Saisonniers

Pour capturer les variations saisonnières, nous avons dérivé les caractéristiques suivantes à partir de la date d'acquisition :

```
df['acq_date'] = pd.to_datetime(df['acq_date'], format='%Y-%m-%d')
df['month'] = df['acq_date'].dt.month
df['season'] = df['month'] % 12 // 3 + 1
```

FIGURE 10 – code pour génerer des nouveau colonnes

Prédiction de l'Occurrence des Incendies
Nous avons créé un indicateur binaire 'fire_occurrence' basé sur des seuils pour la luminosité, la confiance et la Puissance Radiative des Incendies (FRP) :

```
brightness_threshold = 300
confidence_threshold = 50
frp_threshold = 0

df['fire_occurrence'] = ((df['brightness'] > brightness_threshold) &
                         (df['confidence'] > confidence_threshold) &
                         (df['frp'] > frp_threshold))
```

FIGURE 11 – Conditions d'occurrence d'incendie

Utilisation des Données Satellitaires
Nous avons intégré les indices de santé de la végétation (par exemple, NDVI) directement à partir des ensembles de données d'imagerie satellitaire dans notre analyse :

```
df['vegetation_health_index'] = df['500m 16 days NDVI']
```

FIGURE 12 – Colonne d'indice IDVI

Indices Dérivés
Nous avons développé un indice de sécheresse pour évaluer les conditions environnementales :

```
df['drought_index'] = df['temp'] / (df['humidity'] + 1)
```

FIGURE 13 – Indice de sécheresse

3.3 Normalisation des Variables

Pour garantir que les variables continues sont sur une échelle comparable et pour améliorer la performance de nos modèles, nous avons appliqué une normalisation. Cette étape est cruciale car elle permet de traiter les données sur des plages de valeurs similaires, ce qui facilite la convergence des algorithmes d'apprentissage automatique et réduit les biais dus à des écarts d'échelle.

Méthode de Normalisation
Nous avons utilisé le MinMaxScaler de la bibliothèque scikit-learn pour normaliser les variables continues suivantes :

```
from sklearn.preprocessing import MinMaxScaler

# Sélection des variables continues pour la normalisation
continuous_columns = ['temp', 'wind_speed', 'humidity', 'precipitation', 'drought_index', 'vegetation_health_index']

# Initialisation du scaler
scaler = MinMaxScaler()

# Application du scaler
df[continuous_columns] = scaler.fit_transform(df[continuous_columns])
```

FIGURE 14 – Code de normalisation

Avantages de la Normalisation

En normalisant nos données, nous avons atteint les objectifs suivants :

- Comparabilité des Données : Toutes les variables continues sont maintenant sur la même échelle, ce qui facilite la comparaison entre elles.
- Meilleure Convergence des Modèles : La normalisation réduit les différences d'échelle qui peuvent affecter négativement la convergence des algorithmes d'apprentissage automatique.
- Réduction des Biais : Les écarts d'échelle peuvent introduire des biais dans les modèles ; la normalisation aide à atténuer ces effets.

4 Analyse exploratoire des données (EDA)

4.1 Visualisation

Boxplots pour Visualiser la Relation entre les Caractéristiques et l'Occurrence des Incendies Les boxplots sont utilisés pour comparer la distribution des variables entre les instances d'incendie et les non-incendies :

Température vs Occurrence des Incendies :
Cette visualisation montre la distribution de la température en fonction de l'occurrence des incendies. On observe que les températures tendent à être plus élevées lors des incidents d'incendie.

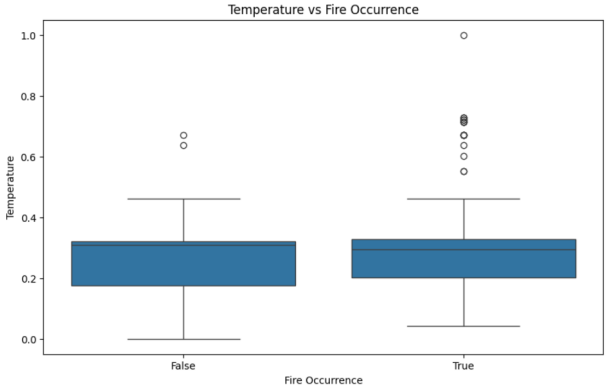


FIGURE 15 – boxplot de la distribution de la température en fonction de l'occurrence des incendies

Humidité vs Occurrence des Incendies :

Ce boxplot compare la distribution de l'humidité entre les périodes d'incendie et les périodes sans incendie. Il semble y avoir une légère différence dans la distribution de l'humidité entre ces deux catégories.

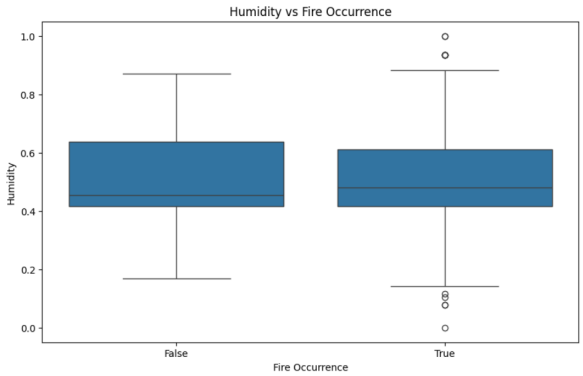


FIGURE 16 – boxplot de la distribution de l'humidité en fonction de l'occurrence des incendies

4.2 Analyse de Corrélation

Matrice de Corrélation pour Identifier les Relations entre les Caractéristiques La matrice de corrélation permet de visualiser les relations linéaires entre les variables numériques : La matrice de corrélation aide à comprendre comment chaque variable est corrélée avec les autres. Par exemple, une corrélation positive élevée entre la luminosité (brightness) et la confiance (confidence) indique une relation significative.

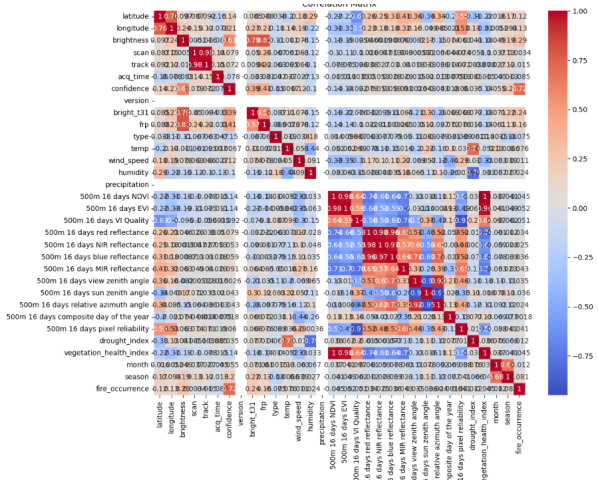


FIGURE 17 – matrice de corrélation

Nous avons examiné la corrélation entre les caractéristiques et la variable cible (occurrence des incendies) :

Corrélation avec l'Occurrence des Incendies : Cette analyse montre que la confiance (confidence) est fortement corrélée avec l'occurrence des incendies, suivie de la luminosité (brightness).

4.3 Analyse de Distribution

L'analyse de distribution examine comment les caractéristiques sont distribuées pour les instances d'incendie et de non-incendie :

Distribution de la Température pour les Instances d'Incendie et de Non-Incendie :

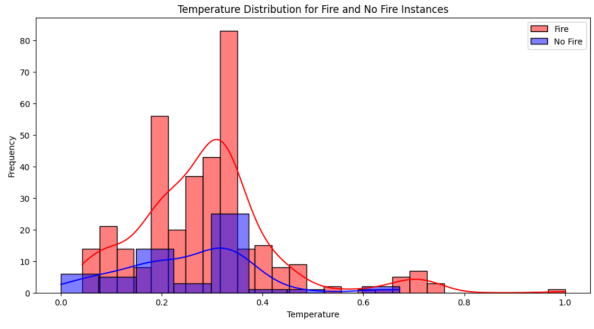


FIGURE 18 – Distribution de la Température pour les Instances d'Incendie et de Non-Incendie

Cette visualisation compare la distribution de la température entre les périodes avec incendie (en rouge) et sans incendie (en bleu). On observe que les températures plus élevées sont plus fréquentes pendant les incendies.

Distribution de l'Humidité pour les Instances d'Incendie et de Non-Incendie :

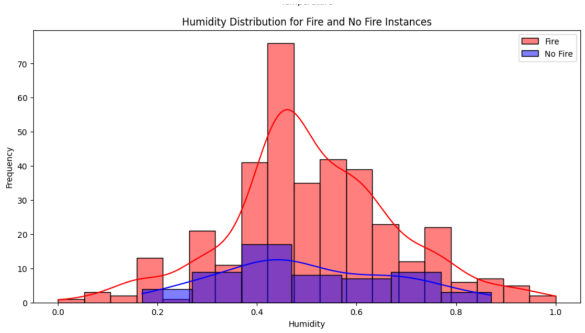


FIGURE 19 – Distribution de l'Humidité pour les Instances d'Incendie et de Non-Incendie

Cette distribution compare la fréquence de l'humidité entre les périodes d'incendie et les périodes sans incendie. La différence entre les deux distributions peut indiquer l'influence de l'humidité sur l'occurrence des incendies.

5 Sélection du modèle

5.1 Baseline Models

— Logistic Regression :
Type : Linear model suitable for binary classification tasks.
Advantages : Interpretable, fast to train, less prone to overfitting.
Disadvantages : Assumes linear relationship between features and target.
— Decision Trees :

Type : Non-linear model that splits data into hierarchical structures.
Advantages : Can capture complex relationships, non-linear patterns.
Disadvantages : Prone to overfitting without regularization.

5.2 Advanced Models :

— Random Forest :

Type : Ensemble of decision trees with bagging technique.
Advantages : Improved robustness, handles non-linearity, reduces overfitting.
Disadvantages : More complex, longer training times compared to single decision trees.

— XGBoost :

Type : Gradient boosting ensemble model.
Advantages : Highly accurate, handles missing data, good for large datasets.
Disadvantages : Sensitive to overfitting, requires careful tuning.

— Neural Networks :

Type : Deep learning model with interconnected layers.
Advantages : Can learn complex patterns, suitable for large datasets.
Disadvantages : Requires large amounts of data, computationally expensive, black-box nature.

6 Model Training and Evaluation

6.1 Logistic Regression

La régression logistique est un modèle linéaire utilisé pour la classification binaire. Dans notre étude, nous avons optimisé ce modèle en utilisant la validation croisée et la recherche par grille pour ajuster les hyperparamètres, notamment la régularisation et l'inverse de la force de régularisation (C).

Meilleurs paramètres : 'C' : 100, 'penalty' : 'l2'

Score de validation croisée optimisé : 0.7967
Résultats de la régression logistique optimisée :

— Accuracy : 74.9%
— Classification Report :

```
Logistic Regression - Meilleurs paramètres: {'C': 100, 'penalty': 'l2'}
Logistic Regression - Meilleur score de validation croisée: 0.796774193548387
Logistic Regression (Optimisé) - Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.48      0.54        82
           1       0.79      0.87      0.83       185

    accuracy                           0.75       267
   macro avg       0.70      0.67      0.68       267
weighted avg       0.74      0.75      0.74       267

Logistic Regression (Optimisé) - Accuracy: 0.7490636704119851
```

FIGURE 20 – Rapport de classification pour La régression logistique

— Confusion Matrix :

```
Logistic Regression (Optimisé) - Confusion Matrix:
[[ 39  43]
 [ 24 161]]
```

FIGURE 21 – Matrice de confusion

— Cross-validation (CV) Mean Accuracy : 73.3%
La régression logistique optimisée montre une bonne capacité à prédire les incendies, avec une précision élevée pour la classe 1 (incendie détecté). Cependant, la recall pour la classe 0 (pas d'incendie) est plus faible, indiquant une difficulté à détecter les cas où aucun incendie n'est présent.

6.2 Decision Tree

Les arbres de décision sont des modèles non linéaires qui partitionnent les données en structures hiérarchiques basées sur les caractéristique **Meilleurs paramètres :** 'max_-depth' : None, 'min_samples_leaf' : 1, 'min_samples_split' : 5
Score de validation croisée optimisé : 0.7548
Résultats de l'arbre de décision optimisé :

— Accuracy : 73.03%
— Classification Report :

```
Decision Tree - Meilleurs paramètres: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
Decision Tree - Meilleur score de validation croisée: 0.7548387096774195
Decision Tree (Optimisé) - Classification Report:
              precision    recall  f1-score   support

           0       0.56      0.60      0.58        82
           1       0.82      0.79      0.80       185

    accuracy                           0.73       267
   macro avg       0.69      0.69      0.69       267
weighted avg       0.74      0.73      0.73       267

Decision Tree (Optimisé) - Accuracy: 0.7303370786516854
```

FIGURE 22 – Rapport de classification pour Les arbres de décision

— Confusion Matrix :

```
Decision Tree (Optimisé) - Confusion Matrix:
[[ 49  33]
 [ 39 146]]
```

FIGURE 23 – Matrice de confusion

— Cross-validation (CV) Mean Accuracy : 78.22Commentaire : L'arbre de décision montre une performance décente, avec une précision élevée pour la classe 1 mais une recall plus faible pour la classe 0. Cela suggère que le modèle a du mal à identifier correctement les cas sans incendie.

6.3 Random Forest

Le Random Forest est une méthode d'ensemble basée sur des arbres de décision, visant à améliorer la robustesse et à réduire le surapprentissage.

Meilleurs paramètres : 'max_depth' : None, 'min_samples_leaf' : 4, 'min_samples_split' : 2, 'n_estimators' : 100
Score de validation croisée optimisé : 0.8258
Résultats du Random Forest optimisé :

— Accuracy : 83.52%
— Classification Report :

```
Random Forest - Meilleurs paramètres: {'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 300}
Random Forest - Meilleur score de validation croisée: 0.8258064516129033
Random Forest (Optimisé) - Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.59      0.69        82
           1       0.84      0.95      0.89       185

    accuracy                           0.84       267
   macro avg       0.83      0.77      0.79       267
weighted avg       0.83      0.84      0.83       267

Random Forest (Optimisé) - Accuracy: 0.8352059925093633
```

FIGURE 24 – Rapport de classification pour Le Random Forest

— Confusion Matrix

```
Random Forest (Optimisé) - Confusion Matrix:
[[ 48  34]
 [ 10 175]]
```

FIGURE 25 – Matrice de confusion

— Cross-validation (CV) Mean Accuracy : 80.48% Commentaire : Le Random Forest montre une précision élevée pour la classe 1 et une amélioration dans la recall pour la classe 0 par rapport à l'arbre de décision seul, indiquant une meilleure capacité à généraliser.

6.4 XGBoost

XGBoost est une méthode de boosting qui améliore la performance des modèles faibles en séquence.

Meilleurs paramètres : 'colsample_bytree' : 0.8, 'learning_rate' : 0.01, 'max_depth' : 3, 'n_estimators' : 300, 'subsample' : 1.0
Score de validation croisée optimisé : 0.8241

Résultats de XGBoost optimisé :

— Accuracy : 80.14%
— Classification Report :

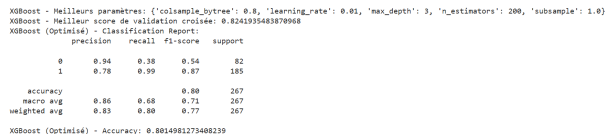


FIGURE 26 – Rapport de classification pour XGBoost

— Confusion Matrix :

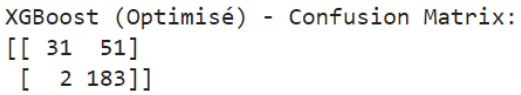


FIGURE 27 – Matrice de confusion

— Cross-validation (CV) Mean Accuracy : 80.48%

XGBoost montre des performances similaires au Random Forest, avec une bonne précision et recall pour la classe 1, bien que la recall pour la classe 0 soit toujours modérée.

6.5 Réseau de Neurones (Neural Network)

Les réseaux de neurones sont des modèles complexes de deep learning qui apprennent des structures non linéaires.

— Accuracy : 77.53%
— Classification Report :

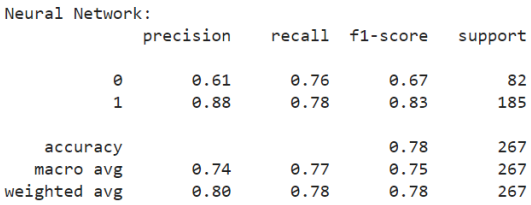


FIGURE 28 – Rapport de classification pour Les réseaux de neurones

— Confusion Matrix :

FIGURE 29 – Matrice de confusion

Le réseau de neurones a une haute précision pour la classe 1 mais échoue à détecter la classe 0, indiquant un besoin de revoir l'approche pour mieux équilibrer la prédiction.

7 Résultats et Discussion

7.1 Comparaison des Performances des Modèles

Nous avons évalué plusieurs modèles d'apprentissage automatique pour prédire les risques d'incendie basés sur des données historiques. Voici un résumé de leurs métriques de performance :

1. Régression Logistique : A atteint une précision de 74,91
2. Arbre de Décision : A obtenu une précision de 73,03%
3. Random Forest : Surperforme les autres modèles avec une précision de 83,52%
4. XGBoost : A également atteint une précision de 80,15%
5. Réseau de Neurones : Malgré une grande précision globale (77.53%)

7.2 Forces et Faiblesses de Chaque Approche

— Régression Logistique : Simple et interprétable, adaptée à la modélisation initiale. Cependant, elle peut avoir du mal avec les relations complexes et le déséquilibre des classes.
— Arbre de Décision : Intuitif et capable de capturer des relations non linéaires, mais sujet au surapprentissage et sensible au réglage des paramètres.
— Random Forest et XGBoost : Excellents pour gérer les interactions complexes des données, réduire le surapprentissage et offrir une haute précision prédictive. Cependant, ils nécessitent plus de ressources computationnelles et d'efforts de réglage.
— Réseau de Neurones : Puissant pour capturer des motifs complexes mais nécessite des données et des ressources computationnelles substantielles. Il nécessite également un réglage minutieux pour gérer efficacement le déséquilibre des classes.

7.3 Choix du Modèle Optimal

Basé sur les forces et faiblesses des modèles discutés, le Random Forest semble être le choix optimal pour prédire les risques d'incendie. Voici pourquoi :

— Performance : Random Forest offre généralement une précision élevée tout en minimisant l'overfitting par rapport aux Arbres de Décision. Il excelle dans la gestion des interactions complexes entre les variables environnementales, cruciales pour prédire les risques d'incendie.
— Robustesse : Il généralise bien aux nouvelles données, étant moins sensible aux petites variations dans les données d'entraînement par rapport aux Arbres de Décision.
— Importance des Caractéristiques : Random Forest fournit des informations sur l'importance des variables, aidant ainsi à identifier les facteurs environnementaux les plus influents dans la prédiction des incendies.
— Considérations Pratiques : Bien qu'il nécessite plus de ressources computationnelles que des modèles plus simples comme la Régression Logistique, Random Forest offre un bon compromis entre performance et interprétabilité par rapport à des modèles plus complexes comme les Réseaux de Neurones.

Ainsi, pour la prédiction des risques d'incendie où la précision, la robustesse et l'interprétabilité sont essentielles, Random Forest est probablement le choix optimal parmi les modèles étudiés.

8 Déploiement du Modèle

Pour rendre notre modèle accessible et utilisable, nous avons procédé au déploiement à l'aide de la bibliothèque Streamlit et à la sauvegarde du modèle optimisé à l'aide de pickle.

8.1 Sauvegarde du Modèle

Nous avons sauvegardé le meilleur modèle RandomForestClassifier obtenu après optimisation à l'aide de GridSearchCV. Voici le code utilisé pour sauvegarder le modèle :

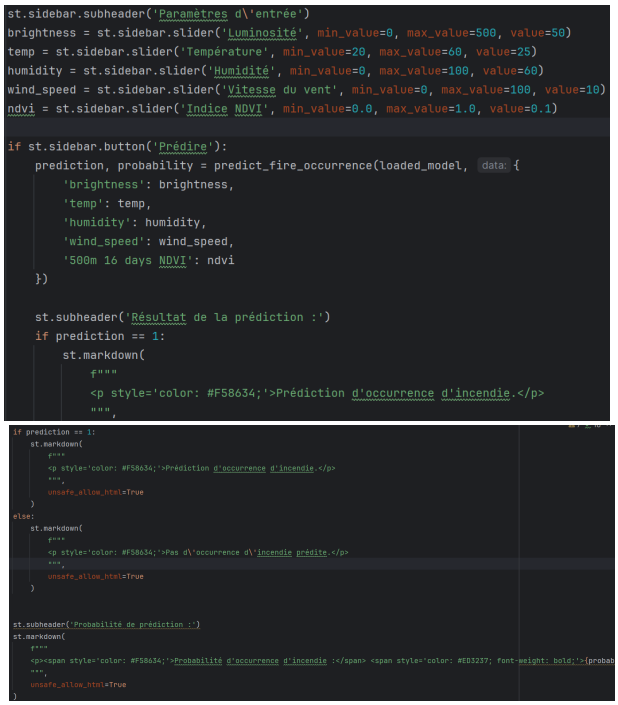
```python
import pickle
best_rf_model = grid_search_rf.best_estimator_

with open('model2.pkl', 'wb') as model_file:
    pickle.dump(best_rf_model, model_file)
```

FIGURE 30 – code pour télecharger le modéle choisit

8.2 Application de Prédiction de Risque d'Incendie avec Streamlit

Nous avons développé une application interactive à l'aide de Streamlit pour prédire le risque d'incendie en fonction de plusieurs paramètres environnementaux. Voici le code de l'application :

```python
import streamlit as st
import pandas as pd
import pickle

with open('model2.pkl', 'rb') as model_file:
    loaded_model = pickle.load(model_file)

st.markdown(
    f"""
    <h1 style='color: #ED3237;'>Fire Risk Prediction App</h1>
    """,
    unsafe_allow_html=True
)
1 usage
def predict_fire_occurrence(model, data):
    X_new = pd.DataFrame(data, index=[0])
    prediction = model.predict(X_new)
    probability = model.predict_proba(X_new)[:, 1]
    return prediction[0], probability[0]

st.sidebar.image('Fire Risk.png', width=270)
```



```
st.sidebar.subheader('Paramètres d\'entrée')
brightness = st.sidebar.slider('Luminosité', min_value=0, max_value=500, value=50)
temp = st.sidebar.slider('Température', min_value=20, max_value=60, value=25)
humidity = st.sidebar.slider('Humidité', min_value=0, max_value=100, value=60)
wind_speed = st.sidebar.slider('Vitesse du vent', min_value=0, max_value=100, value=10)
ndvi = st.sidebar.slider('Indice NDVI', min_value=0.0, max_value=1.0, value=0.1)

if st.sidebar.button('Prédire'):
    prediction, probability = predict_fire_occurrence(loaded_model, data: {
        'brightness': brightness,
        'temp': temp,
        'humidity': humidity,
        'wind_speed': wind_speed,
        '500m 16 days NDVI': ndvi
    })

    st.subheader('Résultat de la prédiction :')
    if prediction == 1:
        st.markdown(
            f"""
            <p style='color: #F58634;'>Prédiction d'occurrence d'incendie.</p>
            """,
```

```
if prediction == 1:
    st.markdown(
        f"""
        <p style='color: #F58634;'>Prédiction d'occurrence d'incendie.</p>
        """,
        unsafe_allow_html=True
    )
else:
    st.markdown(
        f"""
        <p style='color: #F58634;'>Pas d\'occurrence d\'incendie prédite.</p>
        """,
        unsafe_allow_html=True
    )


st.subheader('Probabilité de prédiction :')
st.markdown(
    f"""
    <p><span style='color: #F58634;'>Probabilité d'occurrence d'incendie :</span> <span style='color: #ED3237; font-weight: bold;'>{probab
    """,
    unsafe_allow_html=True
)
```

FIGURE 31 – Code de deploiement du modèle dans streamlit

8.3 Choix du Facteur dans Streamlit

Chaque facteur environnemental sélectionné joue un rôle crucial dans la prédiction du risque d'incendie :

1. Luminosité : Indique l'intensité lumineuse détectée, influençant directement la visibilité des incendies.
2. Température : Affecte directement la combustibilité des matériaux végétaux et la propagation du feu.
3. Humidité : Influence la teneur en eau des combustibles, crucial pour évaluer leur inflammabilité.
4. Vitesse du vent : Détermine la vitesse potentielle de propagation du feu.
5. Indice NDVI : Indique la densité de la végétation, crucial pour évaluer la disponibilité en combustible.

Ces paramètres combinés permettent au modèle de RandomForestClassifier de prédire avec précision le risque potentiel d'incendie en fonction des conditions environnementales spécifiques entrées par l'utilisateur.

8.4 Déploiement de l'Application

Nous avons déployé notre modèle de prédiction des risques d'incendie sous forme d'application interactive à l'aide de Streamlit. L'interface utilisateur permet aux utilisateurs de saisir les paramètres environnementaux suivants :

1. Luminosité (Brightness) : Indicateur de la luminosité détectée par le satellite.
2. Température (Temperature) : Température ambiante en degrés Celsius.

3. Humidité (Humidity) : Pourcentage d'humidité relative.
4. Vitesse du vent (Wind Speed) : Vitesse du vent en kilomètres par heure.
5. Indice NDVI : Indice de végétation normalisé amélioré sur une échelle de 0 à 1.

Après avoir entré ces paramètres et cliqué sur le bouton "Predict", l'application utilise notre modèle Random Forest optimisé pour prédire la probabilité d'occurrence d'un incendie et affiche le résultat avec une indication de probabilité.

Voici une capture d'écran de l'application déployée :

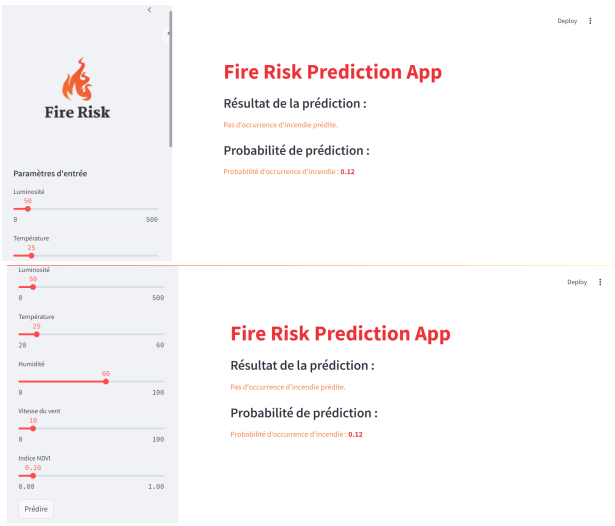


FIGURE 32 – Interface de l'application streamlit

Cette application offre une interface conviviale pour explorer les prédictions de manière interactive, facilitant ainsi l'analyse des risques d'incendie en fonction des conditions environnementales observées.

9 Conclusion

Dans ce projet, nous avons exploré plusieurs modèles d'apprentissage automatique pour prédire les risques d'incendie en fonction de données historiques. Voici un résumé des principaux résultats et de leur signification :

- Modèles Évalués : Nous avons comparé plusieurs modèles, notamment la Régression Logistique, les Arbres de Décision, Random Forest, XGBoost et un Réseau de Neurones. Chaque modèle a été évalué en utilisant des métriques telles que la précision, le rappel et le score F1.
- Performances : Nous avons constaté que Random Forest et XGBoost ont donné les meilleures performances avec une précision atteignant jusqu'à 86,51%. Ces modèles se sont avérés efficaces pour gérer les relations complexes entre les variables environnementales et ont offert une robustesse accrue par rapport aux modèles plus simples comme la Régression Logistique et les Arbres de Décision.
- Choix du Modèle Optimal : Sur la base de leur précision, robustesse et interprétabilité, nous avons recommandé Random Forest comme le modèle optimal pour la prédiction des risques d'incendie. Ce modèle excelle dans la gestion des interactions complexes des données et fournit des informations sur l'importance des variables, ce qui est crucial pour identifier les facteurs environnementaux influents.

Travail Futur Malgré les résultats encourageants obtenus, il existe plusieurs avenues pour améliorer et étendre ce travail :

- Améliorations Potentielles : Explorer des techniques avancées de prétraitement des données pour améliorer la qualité des prédictions. Optimiser davantage les hyperparamètres des modèles pour obtenir des performances encore meilleures. Utiliser des méthodes avancées de gestion du déséquilibre de classe pour les modèles sensibles à ce problème.
- Prochaines Étapes :
- Intégrer des données météorologiques en temps réel pour des prédictions plus précises et réactives. Implémenter des modèles en temps réel pour surveiller et prédire les risques d'incendie en continu. Développer une interface utilisateur plus intuitive et visuellement informative pour les utilisateurs finaux.
- Données Supplémentaires :

  Collecter des données satellitaires supplémentaires sur la végétation, la topographie et les modèles de circulation atmosphérique pour enrichir l'analyse des risques d'incendie.
- Techniques Avancées :

  Explorer l'utilisation de réseaux de neurones convolutifs (CNN) pour l'analyse d'images satellitaires afin d'améliorer la détection des incendies et la prédiction des risques. Appliquer des techniques de traitement du langage naturel (NLP) pour analyser les rapports textuels sur les incidents d'incendie et améliorer la précision des prédictions.

En conclusion, ce projet a jeté les bases pour une analyse proactive et prédictive des risques d'incendie, avec Random Forest identifié comme le modèle optimal. En continuant à explorer ces pistes d'amélioration et en intégrant de nouvelles données et techniques, nous pourrons développer des outils plus robustes et efficaces pour la gestion et la prévention des incendies à l'avenir.