Données climatiques de ECMWF pour la classification des activités infrasonores par machine learning

Préparé par **: BAHMAD Youssef & OUNIS Khalil**

M2 Data science : Santé, assurance, et finance

Année Universitaire : 2024-2025

Introduction

Les infrasons, qui sont des ondes acoustiques de très basse fréquence (inférieures à 20 Hz), proviennent de phénomènes naturels et d'activités humaines, tels que les mouvements de glace, les éruptions volcaniques et les explosions. En surveillance glaciaire, ces infrasons offrent un moyen efficace pour détecter et analyser, à distance, des événements comme les chutes de morceaux de glaciers (vêlage) et les écoulements d'eau de fonte. Contrairement aux techniques d'observation visuelle ou par satellite, l'utilisation des infrasons permet une surveillance passive et continue, précieuse pour suivre l'évolution des glaciers dans des régions difficiles d'accès, comme le Groenland**[1]**.

I Analyse et conception du projet

I.1 Objectif du projet

L'objectif principal de ce projet est de concevoir et mettre en place un système intégré basé sur AI qui permettra de différencier les niveaux infrasonores élevés causés par des processus dynamiques glaciaires d'ampleur, contre les activités infrasonores faibles dues à des processus glaciaires mineurs ou inexistants, en utilisant les techniques de machine learning. Le travail a été réalisé en binôme, avec un focus distinct sur deux aspects cruciaux : l'analyse et exploration approfondie des données, et l'utilisation de differents modèles de machine learning pour classifier les activités infrasonores.
Le projet vise à atteindre plusieurs objectifs clés :

- Choisir un seuil pour la discrétisation binaire des activités infrasonores (faible et élevée).
- Traiter et analyser les données de CEPMMT et signaux d'infrasons.
- Proposer et entraîner des modèles de ML.
- Mettre en œuvre le modèle de ML le plus adapté à notre problème.

I.2 Description du processus du projet

La conception d'un projet présente la chaîne des étapes entamées afin de réaliser ce dernier. Le processus de notre propre projet se résume dans la figure (1) ci-dessous.

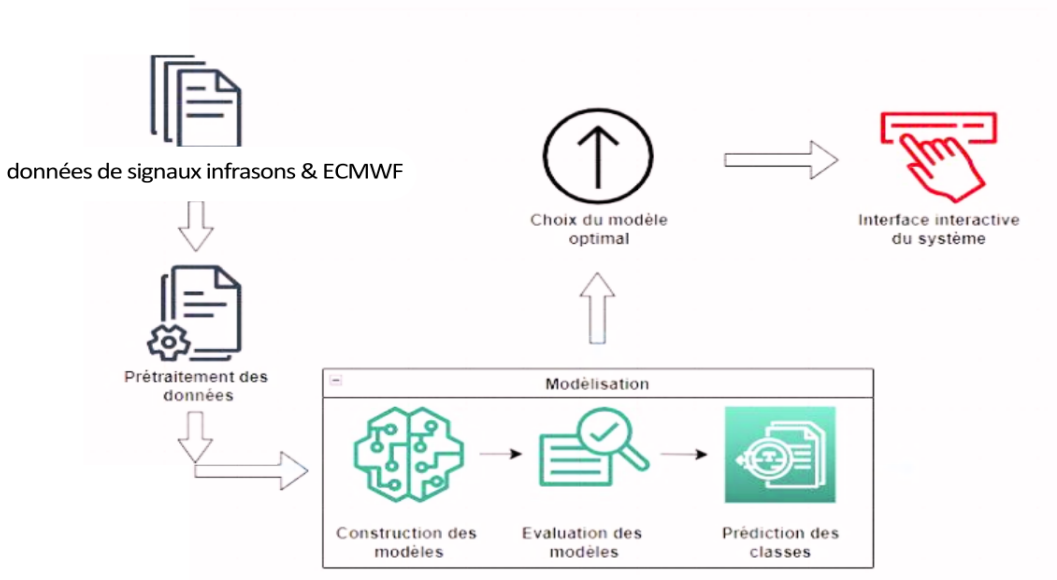


FIGURE 1 – Processus de classement des activités infrasonores.

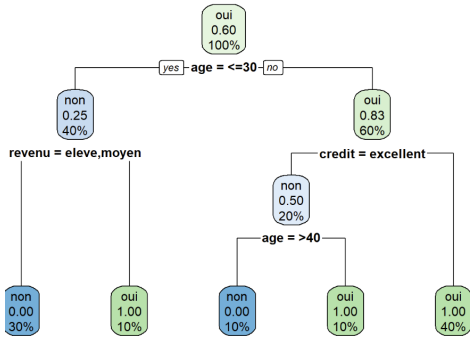
Nous allons travailler avec des données des signaux infrasons et avec les données environnementales associées(temperature, vitesse du vent...) fournies par ECMWF. Notre première étape consistera à nettoyer ces données, en éliminant tout ce qui pourrait affecter la réussite de l'apprentissage, tout en équilibrant les données pour obtenir des résultats précis.

Dans la phase de prétraitement des données, nous appliquerons diverses techniques pour bien préparer les données : feature engineering, transformation de probléme en un problème de classification binaire.

La phase de modélisation consistera à construire nos modèles d'apprentissage, que nous évaluerons ensuite en fonction de leur capacité à prédire les niveaux des infrasons(Faible ou elevée) . Nous choisirons ensuite le modèle d'apprentissage optimal pour résoudre notre problème de classification.

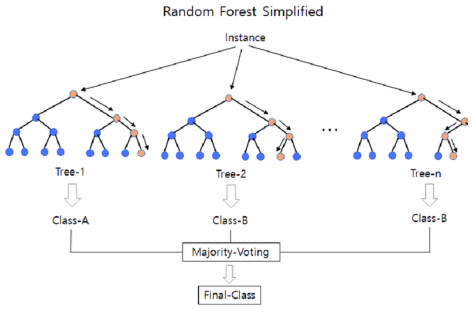
Enfin, nous mettrons en place une interface interactive pour faciliter les interactions entre les utilisateurs et le système.

I.3 Conception des modèles



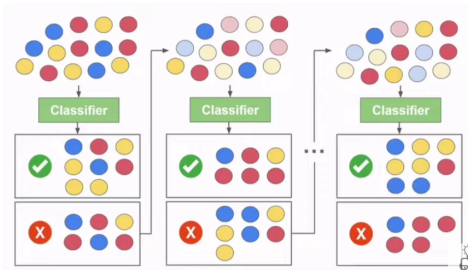
Arbre de décision

Un modéle de ML qui emploie une stratégie de division pour régner en effectuant une recherche gloutonne pour identifier les points de division optimaux au sein d'un arbre.



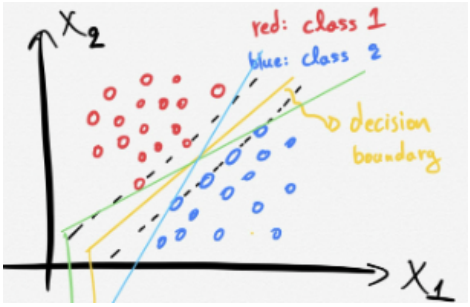
Forêt aléatoire

Un modèle qui consiste en une collection d'arbres de décision. L'algorithme sélectionne aléatoirement des échantillons des données de formation et des variables et crée un arbre de décision à partir de chaque échantillon.



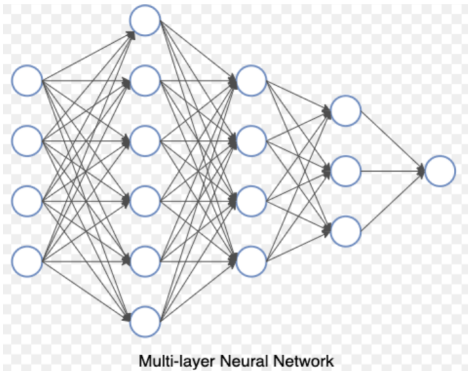
AdaBoost

L'idée principale de ce modèle est de donner plus de poids aux observations mal classées à chaque itération, pour que les modèles successifs se concentrent sur ces erreurs et améliorent progressivement la précision du modèle global.



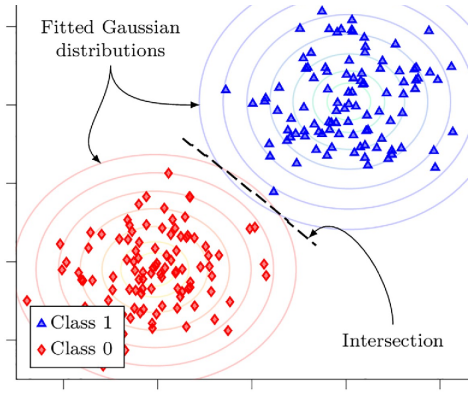
SVM

Un modèle qui consiste à séparer les données en classes en utilisant une frontière aussi **simple** que possible. Il sert ainsi à maximiser la distance, ou marge, entre les différents groupes de données, ainsi que la frontière qui les sépare.



réseaux de neurones

sont constituée de couches de nœuds, ou neurones artificiels : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque nœud se connecte à un autre et possède un poids et un seuil associés.



Gaussian Naïve Bayes

C'est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, appartenant à la famille des classifieurs linéaires.

II Modélisation et Exploration des données

II.1 Exploration des données

• **Présentation des données.**

	time	t2m	u10	v10	SST	SIC	r1_MAR	r2_MAR	r3_MAR	r4_MAR	r5_MAR	Y1
0	2013-01-01	-21.926931	-0.973994	3.149094	-1.690511	90.745710	0.034537	0.033345	0.0	0.0	0.0	0
1	2013-01-02	-23.696195	-6.502908	2.494894	-1.690511	88.502980	0.034527	0.033326	0.0	0.0	0.0	0
2	2013-01-03	-25.644027	-3.557411	1.025486	-1.689860	88.734091	0.034523	0.033321	0.0	0.0	0.0	0
3	2013-01-04	-23.566887	-1.888075	-3.486122	-1.690511	89.149576	0.034509	0.033306	0.0	0.0	0.0	0
4	2013-01-05	-22.897768	-2.748844	-3.491206	-1.689860	91.613955	0.034492	0.033290	0.0	0.0	0.0	0

FIGURE 2 – Présentation de la base de données.

Les données d'entrée se constituent de 11 variables qui décrivent les informations sur le climat :

— t2m : Température de 2 mètres au-dessous de la mer

— SST : Température à la surface de la mer

— u10, v10 : Vitesse du vent

— SIC : Concentration de la mer glacée

— r1_MAR, r2_MAR, r3_MAR, r4_MAR, r5_MAR : Débit d'eau liquide du Groenland simulé par les modèles climatiques régionaux dans 5 régions

Les données de sortie consistent à 4 variables qui représentent des enregistrements quantitatifs des infrasons. Dans notre étude, on va s'intéresser particulièrement à une seule sortie qui sera **Y1** qui prend des valeurs entre [0,443].

• **Analyses statistiques des données**

La description des variables est aussi nécessaire pour avoir une idée sur les ordres de grandeur des données si jamais on a besoin de normaliser. Ce tableau résume les différentes mesures apportées sur les données :

	time	t2m	u10	v10	SST	SIC	r1_MAR	r2_MAR	r3_MAR	r4_MAR	r5_MAR	Y1
count	2556	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000
mean	2016-07-01 12:00:00	-10.190040	0.139648	0.626351	-0.858922	73.267808	18.789275	11.522362	1.328740	4.382248	5.191286	3.525430
min	2013-01-01 00:00:00	-32.019122	-13.846656	-12.316128	-1.692462	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2014-10-01 18:00:00	-19.877500	-3.610168	-2.079422	-1.689860	70.000000	0.123798	0.121437	0.000000	0.000000	0.000000	0.000000
50%	2016-07-01 12:00:00	-9.596479	-0.187084	0.912780	-1.689860	84.601769	0.481870	0.481870	0.000000	0.000000	0.000000	0.000000
75%	2018-04-01 06:00:00	0.167824	3.807440	3.483397	-0.297573	90.362319	4.080174	3.961181	0.004413	0.010148	0.000150	0.000000
max	2019-12-31 00:00:00	7.840619	14.640913	12.811255	6.054536	99.500682	479.722174	281.673389	23.241791	115.876574	88.054318	433.000000
std	NaN	10.340583	5.013640	3.955417	1.446165	29.250724	47.703600	27.942124	3.393452	12.971518	13.406972	18.977537

FIGURE 3 – Description des données

• **Variation des valeurs moyennes mensuelles des infrasons Y1**

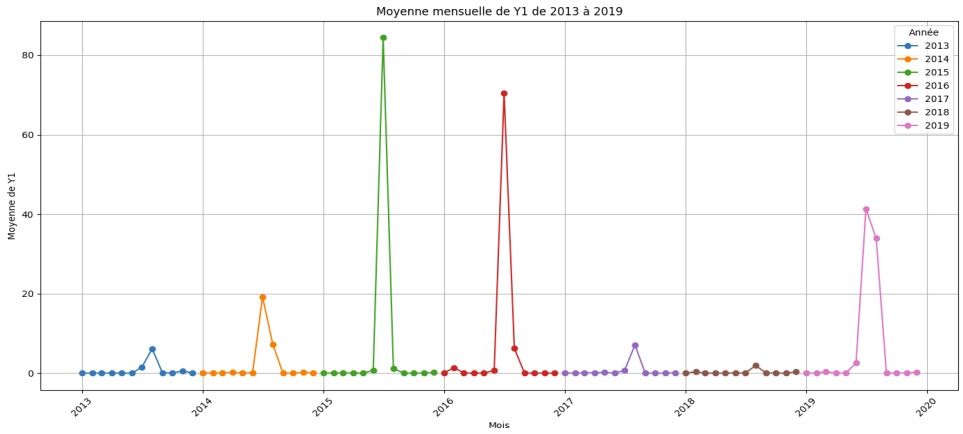


FIGURE 4 – Moyenne mensuelle de Y1 sur la plage des années 2013-2020

La visualisation des valeurs moyennes mensuelles de Y1 montrent un pique presque au niveau d'une seule période de l'année et on enregistre aussi des valeurs faibles presque quasiment nulles pour le reste de l'année.

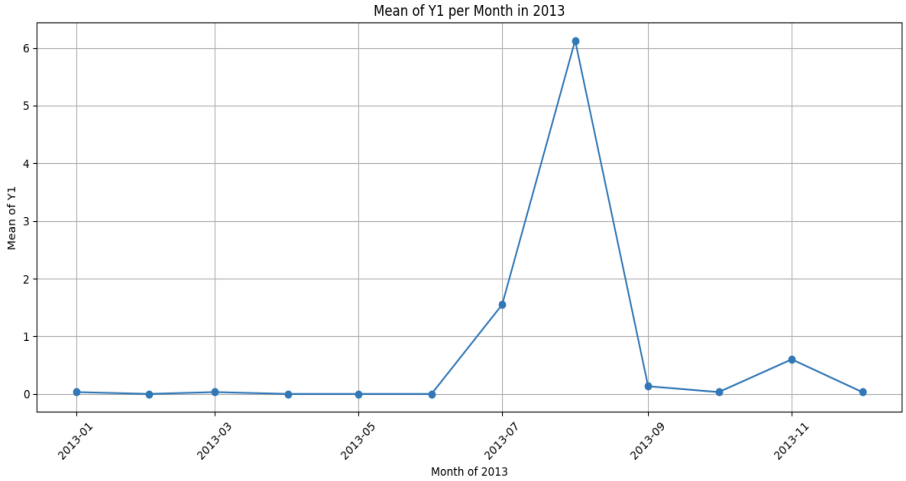


FIGURE 5 – Moyenne mensuelle de Y1 pour l'année 2013 (pareil aux autres années)

Sur les deux figures précédente, on remarque que les piques sont enregistrées dans la même période de l'année.Alors ces activités intenses sont enregistrées presque au mois de juin, juillet et août chaque année.

• **Matrice de corrélation**

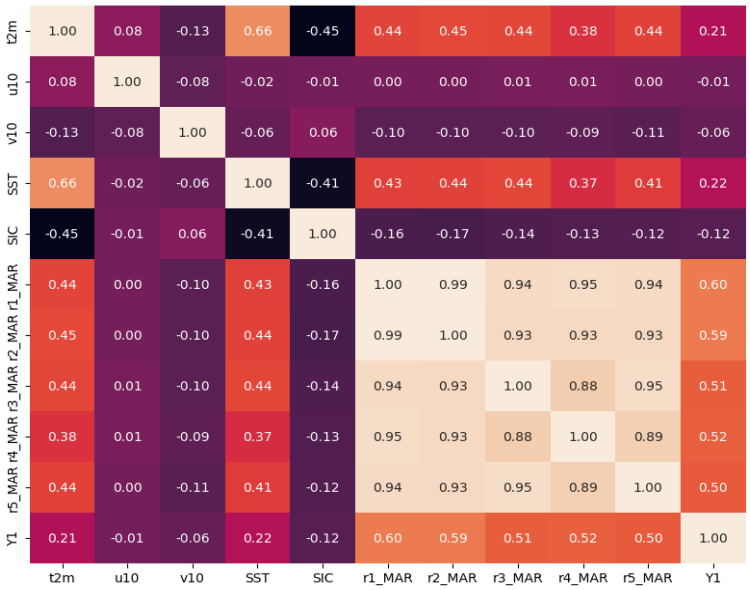


FIGURE 6 – matice de corrélation

• **Distribution de Y1**

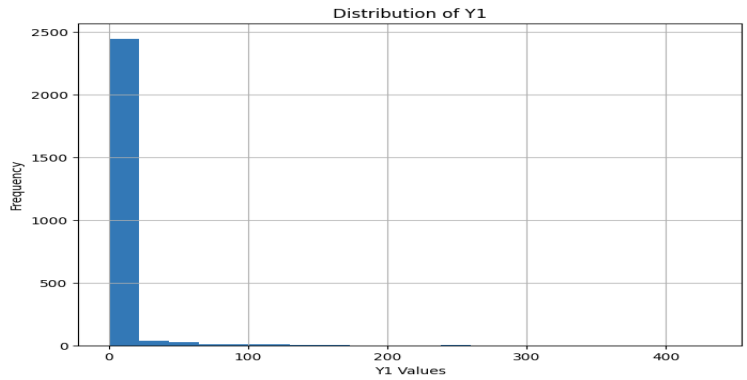


FIGURE 7 – Fréquence des valeurs de Y1 dans le dataset

• **Feature engineering (Pour la variable time)**

— Création d'une nouvelle variable binaire prenant en compte les événements saisonniers des infrasons pour les mois de juin, juillet et août.

$$\text{Variation\_intense}(m) = \begin{cases} 1 & \text{si } m \in \{6, 7, 8\} \\ 0 & \text{sinon} \end{cases}$$

où m représente le mois.

— Extraction des informations temporelles : le jour, le mois et l'année, à partir la variable time (et suppression de la variable time dans la dataset).

• **Choix du seuil de binarisation de Y1**

$$Y1_{\text{binaire}} = \begin{cases} 1 & \text{si } Y1 \geq \text{seuil} \\ 0 & \text{si } Y1 < \text{seuil} \end{cases}$$

Les valeurs de Y1 sont dans l'intervalle [0,433] et représentent des infrasons enregistrés. Les niveaux d'infrasons mesurés fluctuent en fonction de l'activité glaciaire. Par exemple : de faibles valeurs (comme 0, 1, ou 6) suggèrent une activité minimale ou normale sans événements glaciaires significatifs, comme des petites fontes de glace sans grand écoulement. Les valeurs moyennes (par exemple, entre 20 et 100) indiquent des niveaux modérés d'activité infrasonore, probablement associés à des décharges d'eau de fonte plus importantes ou à des petits événements de vêlage. Les valeurs élevées (au-delà de 100 et jusqu'à 433) correspondent à des événements plus intenses, tels que des épisodes de vêlage de blocs de glace ou de grandes quantités de fonte. De plus, on à essayer de choisir le seuil avec K-means, le resultat étais 57.

Justification du choix de seuil

Afin de séparer notre jeu de données en deux classes (0 et 1) et vu qu'on a 3 classes depuis l'interprétation physique des données, il est plus raisonnable d'opter pour un seuil de 60 qui est la valeur médiane de l'intervalle des valeurs moyennes. Dans ce cas, on va avoir 0 pour les valeurs qui appartiennent aux faibles valeurs des infrasons et la partie faible des valeurs moyennes ainsi que des 1 pour les valeurs élevées et la partie élevée des valeurs moyennes. Sans regarder les statistiques du jeu de données, cette répartition semble la plus correcte. Mais tenant compte de la présence de beaucoup de valeurs nulles, on a décidé de modifier le seuil à 20 Hz une valeur qui sépare les valeurs faibles et les valeurs moyennes/élevées.

II.2 Modélisation : Entrainement et évalution des modéles

• **Répartition des données**

Afin de mieux s'adapter aux proportions des données cibles déséquilibrées dans notre jeu de données, on a décidé d'entraîner les modèles en adoptant la technique du cross-validation qui va nous permettre d'évaluer les modèles de manière plus fiable et éviter le surapprentissage. Pour la division des données, on ne va pas se contenter d'utiliser un simple K-fold pour la division mais on va opter pour un Stratified K-fold qui va nous permettre de garder les mêmes pourcentage dans les différentes proportions des classes de la variable cible.

• **Analyse des résulats**

TABLE 1 – Benchmarking des modèles de classification - Mesures de Performance

Modèle	**Accuracy**	**F1 Score**	**Recall (Sensibilité)**	**Précision**
Decision Tree	0.96	0.53	0.47	0.61
Random Forest	0.97	0.63	0.63	0.64
AdaBoost Classifier	0.96	0.64	0.69	0.59
Neural Network	0.97	0.61	0.57	0.66
SVM	0.97	0.61	0.58	0.65
Naive Bayes	0.93	0.54	0.96	0.37

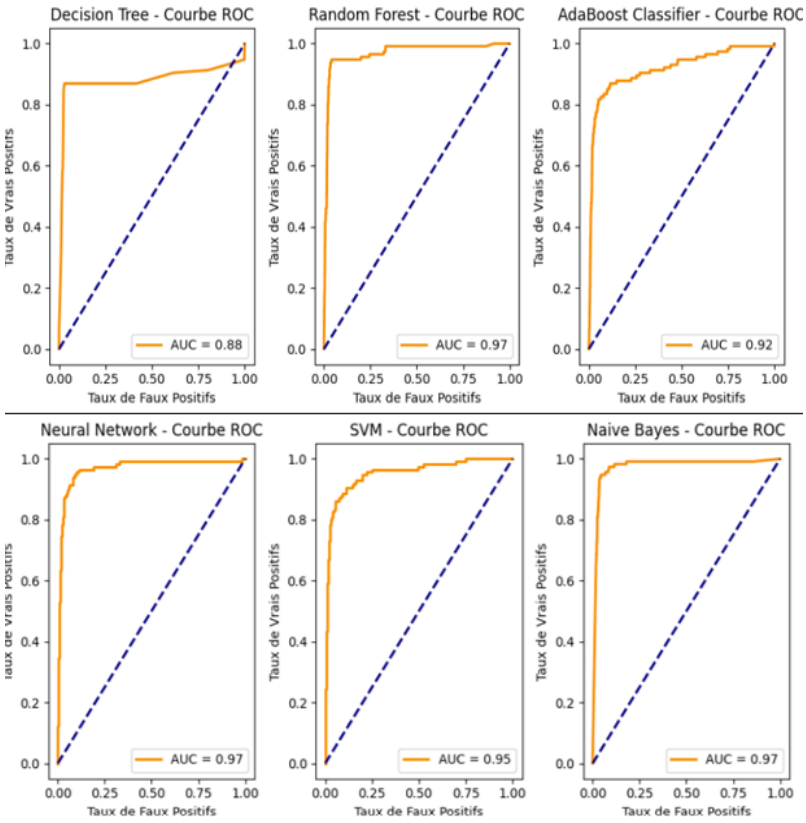


FIGURE 8 – Courbe ROC : capacité de diagnostique des classificateurs

Le tableau et les courbes ci-dessus présentent les performances de différents modèles de classification, évalués à l'aide de quatre métriques et courbe de ROC : Le Random Forest se distingue comme le meilleur modèle dans cet ensemble de résultats. Avec un bon équilibre entre toutes les métriques, il combine une haute Accuracy (0.97) avec un F1 Score élevé (0.63) et un bon compromis entre Recall et Précision. Cela en fait le choix optimal pour des tâches de classification où une bonne couverture des classes positives et une précision sont nécessaires.

Le Random Forest est donc recommandé comme le modèle le plus performant pour cette tâche.

• Interface interactive du système

L'objectif de l'interface interactive est de faciliter et de gérer les interactions entre l'utilisateur et le système. Nous avons développé cette interface en utilisant la bibliothèque Gradio, qui nous permet de créer rapidement des parties d'interface utilisateur faciles à utiliser et ajustables pour un modèle de Machine Learning. Pour nous, cette une interface qui va prendre comme entrée des données environnementales fournies par ECMWF associe et renvoyer la catégorie d'infrason : faible ou élevée.

Après l'exécution de la commande, nous avons obtenu comme résultat l'interface illustrée dans la figure ci-dessus :

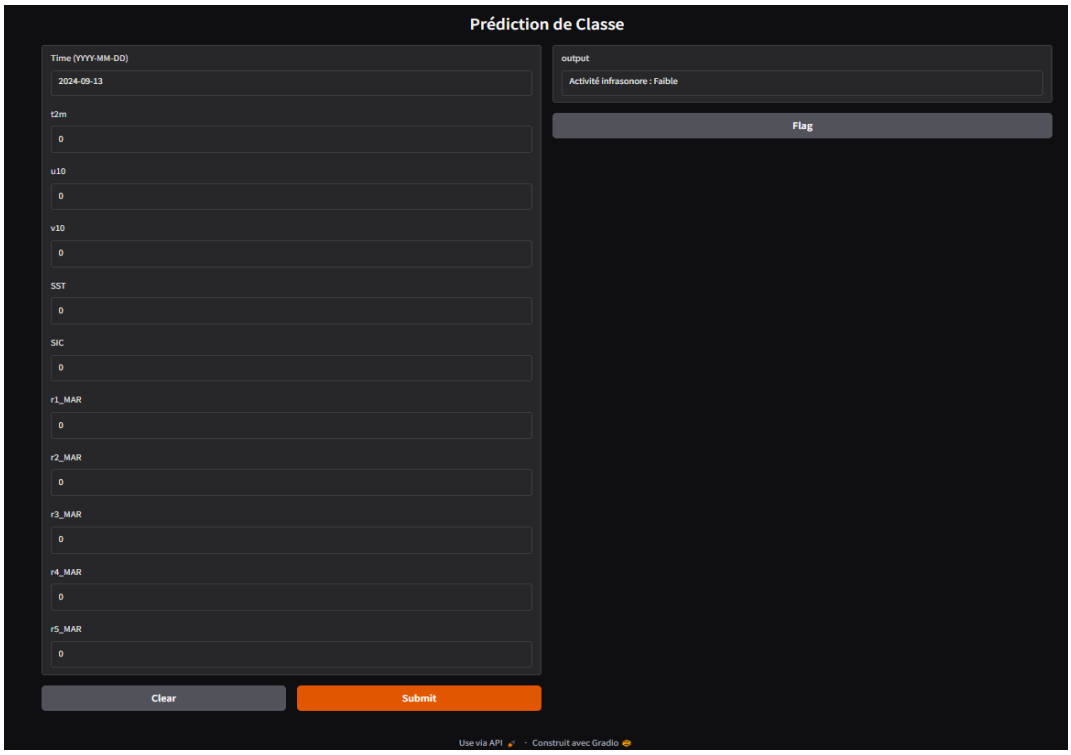


FIGURE 9 – Interface du système

Conclusion

L'analyse des données infrasonores pour classer l'activité glaciaire révèle à quel point les algorithmes de machine learning peuvent être efficaces pour détecter et prédire des phénomènes naturels complexes. Dans cette étude, plusieurs modèles de classification ont été mis à l'épreuve, notamment **Random Forest**, **AdaBoost**, **Réseaux de Neurones**, **SVM** et **Naive Bayes**. Chacun de ces modèles a ses propres atouts et limites lorsqu'il s'agit de saisir les subtilités des signaux infrasonores.

Cette recherche met en lumière l'importance de choisir le modèle de machine learning qui convient le mieux aux spécificités des données et aux objectifs visés. De plus, elle souligne l'importance de mettre en place des techniques de rééquilibrage des classes pour mieux représenter les événements rares qui apparaissent dans les données infrasonores. En tenant compte de ces éléments, nous pouvons améliorer notre compréhension des signaux et affiner nos capacités à analyser l'activité glaciaire.

Bibliographie

[1] Evers, L. G., & Smets, P. S. M. (2022). *Long-Term Infrasonic Monitoring of Land and Marine-Terminating Glaciers in Greenland. Geophysical Research Letters.*