UNIVERSITE ABDELMALEK ESSAADI
FACULTE DES SCIENCES ET TECHNIQUES DE TANGER
DEPARTEMENT GENIE INFORMATIQUE

DETECTION DU NIVEAU D'OZONE

Réaliser par :

- Ahmed Samady
- Marouan Daghmoumi
- Fahd Chibani

Encadrer par :

- Pr. M'hamed AIT KBIR

Année Universitaire : 2023/2024

Sigles et abréviations

Sigles	Significations
SVM	Support Vector Machines
CART	Classification and Regression Trees
DT	Decision Trees
PCA	Principal Component Analysis
CNN	Condensed Nearest Neighbour
SVSMOTE	Support Vector Machine Synthetic Minority Over-Sampling Technique
ROC Curve	Receiver Operating Characteristic Curve
AUC	Area under the ROC Curve

INTRODUCTION GÉNÉRALE

La surveillance des niveaux d'ozone est essentielle pour comprendre et atténuer les impacts de la pollution atmosphérique. Ce projet se concentre sur l'application de techniques d'apprentissage automatique, en particulier les machines à vecteurs de support (SVM) avec et sans l'utilisation de la réduction de dimension PCA (Principal Component Analysis), ainsi que les arbres de décision (DT). Notre objectif est d'évaluer l'efficacité de ces approches dans la classification des niveaux d'ozone à partir de données temporelles complexes. À travers cette étude, nous cherchons à évaluer l'efficacité de ces méthodes dans la résolution des défis liés à la surveillance de la qualité de l'air et à la protection de l'environnement.

CHAPITRE 1 :
PRESENTATION GÉNÉRALE

I. INTRODUCTION

La couche d'ozone, une composante cruciale de l'atmosphère terrestre, joue un rôle essentiel dans la protection contre les rayons ultraviolets nocifs du soleil. La surveillance et la prédiction des niveaux d'ozone revêtent donc une importance capitale pour la santé humaine et l'environnement. Dans cette étude, nous nous concentrons sur l'application de techniques d'apprentissage automatique pour prédire les niveaux d'ozone dans la région de **Houston**, **Galveston** et **Brazoria**, en nous appuyant sur un ensemble de données distincts : le pic d'ozone sur huit heures (**eighthr.data**), collectés entre 1998 et 2004.

Ces ensembles de données fournissent une perspective temporelle et multivariée des conditions météorologiques, notamment les mesures de température à différentes périodes de la journée, la vitesse du vent à divers moments, les niveaux d'humidité relative et d'autres paramètres météorologiques associés.

L'objectif principal de cette étude est d'explorer et de comparer deux approches d'apprentissage automatique : les machines à vecteurs de support (**SVM**) avec ou sans utilisation de l'Analyse en Composantes Principales (**PCA**), et les arbres de décision (**DT**), pour prédire les niveaux d'ozone dans cette région. Nous visons à évaluer la performance de ces techniques dans la classification des niveaux d'ozone à partir des données météorologiques disponibles.

II. DESCRIPTION DES DONNÉES

Les données utilisées dans cette étude représentent un ensemble multivarié et temporel de mesures des niveaux de pic d'ozone dans la région de **Houston**, **Galveston** et **Brazoria**, collectées entre 1998 et 2004. L'ensemble de données se concentre spécifiquement sur le pic d'ozone sur une période de huit heures.

- Les attributs commençant par T représentent les mesures de température à divers moments de la journée.
- Ceux débutant par WS indiquent la vitesse du vent à différentes périodes.
- Parmi les attributs, on trouve également des mesures de l'humidité relative, de la pression atmosphérique, des précipitations, ainsi que des indices météorologiques tels que le K-Index et les T-Totals.

L'ensemble de données est composé de 2535 instances et contient des valeurs réelles pour les différentes caractéristiques, avec la possibilité de présenter des valeurs manquantes.

Il est essentiel de noter que seuls les niveaux de pic d'ozone sur **huit heures** sont utilisés dans cette étude, excluant ainsi les données relatives au pic d'une heure. Cette sélection spécifique permet de se concentrer sur une période plus longue pour évaluer et prédire les niveaux d'ozone dans la région ciblée.

III. SYNTHÈSE DES TRAVAUX DE RECHERCHE SUR LE SUJET

L'article de Meng explore la prédiction des niveaux de pollution par l'ozone en utilisant des modèles d'apprentissage automatique. Voici un résumé de ses principaux points :

3.1. Données et Prétraitement:

Les valeurs manquantes sont gérées en utilisant une méthode de remplissage basée sur la moyenne des données voisines. La normalisation Z est appliquée pour équilibrer les attributs.

3.2. Modèles Utilisés:

Cinq modèles d'apprentissage automatique sont testés : Régression logistique, AdaBoost, Arbre de décision, Forêt aléatoire et Machine à Vecteurs de Support (SVM).

3.3. Résultats et Conclusions:

Le modèle SVM obtient le score de test le plus élevé de 0,949, suivi par Random Forests avec 0,942. La régression logistique obtient des scores de 0,859 en train et de 0,800 en test. L'article conclut que le SVM est le modèle le plus précis pour prédire la pollution par l'ozone dans cette région.

IV. PRETRAITEMENT DES DONNÉES

Le prétraitement des données est une étape cruciale pour garantir la qualité et la pertinence des résultats obtenus par les modèles d'apprentissage automatique. Dans le cadre de cette étude sur la prédiction des niveaux de pic d'ozone sur huit heures, plusieurs étapes de prétraitement ont été effectuées sur l'ensemble de données :

4.1. Gestion des valeurs manquantes :

Les données présentent des valeurs manquantes. Pour pallier ce problème, différentes approches ont été envisagées, notamment l'imputation des valeurs manquantes par la moyenne.

Figure 1: fichier eightht.data Avant Preprocessing

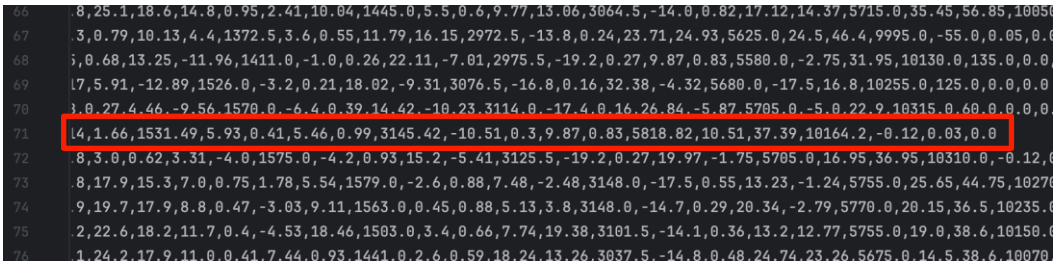

Figure 2: fichier eighthr.data Après Preprocessing

4.2. Eventuelle réduction de dimensionnalité :

Si nécessaire, une réduction de dimensionnalité a été envisagée pour réduire la complexité du modèle. L'Analyse en Composantes Principales (**PCA**) a été considérée comme une méthode pour réduire la dimension tout en conservant au maximum l'information pertinente.

Ces étapes de prétraitement visent à améliorer la qualité des données et à préparer l'ensemble de données pour l'entraînement des modèles d'apprentissage automatique. Le choix approprié des méthodes de prétraitement est fondamental pour assurer des résultats précis et fiables lors de la modélisation et de la prédiction des niveaux de pic d'ozone. Une fois ces étapes terminées, nous sauvegarderons nos nouvelles données dans un fichier distinct nommé (‘**0+1dataset.data**’).

V. MÉTHODOLOGIE

Dans cette étude visant à prédire les niveaux de pic d'ozone sur huit heures, deux techniques d'apprentissage automatique ont été considérées : les Machines à Vecteurs de Support (**SVM**) avec ou sans utilisation de l'Analyse en Composantes Principales (**PCA**), ainsi que les arbres de décision **CART** (Classification and Regression Trees).

5.1. Explication des Techniques

Support Vector Machines (SVM):

Cette méthode est utilisée pour la classification et la régression. Elle vise à trouver l'hyperplan optimal pour séparer les différentes classes en maximisant la marge entre les exemples d'entraînement. Lorsqu'elle est combinée avec la PCA, elle utilise la réduction de dimensionnalité pour améliorer la séparabilité des classes en projetant les données dans un espace de dimension inférieure.

Pour trouver les vecteurs de supports on calcule les α_i en le transformant en un problème d’optimisation convexe avec des contraintes sous l’équation suivants :

$$L(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \varphi(x_n, x_m)$$

Sous les contraintes suivantes :

$$\alpha_n \leq C\ , -\alpha_n\ \leq 0$$

Où $\varphi(x_n, x_m)$ une application de R^M vers H est le kernel qui est utiles pour transformer la dataset en une dimensionnalité plus grande que de la dataset original pour qu'on peut séparer la dataset par un hyperplan linéaire.

Les Kernels utiliser sont :

Kernel linéaire: $\varphi(x_1, x_2) = x_1^T . x_2$

Kernel Polynomial : $\varphi(x_1, x_2) = (x_1^T . x_2 + 1)^\gamma$

Kernel RBF: $\varphi(x_1, x_2) = e^{-\frac{(\|x_1 - x_2\|)^2}{2.nFeats.\gamma^2}}$

Kernel Sigmoïde: $\varphi(x_1, x_2) = \tanh(\gamma . x_1^T . x_2 + \gamma)$

Classification and Regression Tree:

Les arbres de décision sont des modèles utilisés pour la classification et la régression. Ils divisent l'ensemble de données en nœuds basés sur les caractéristiques pour atteindre des décisions. CART est une autre approche d'arbre de décision qui utilise l'indice d'impureté de Gini pour la classification et la chute de variance pour la régression, l'indice de Gini est donné par la formule suivante :

$$Gini = 1 - \sum_{i=1}^{K} p_i^2$$

Où p_i représente la probabilité d'apparition de la classe i dans l'ensemble de données.

Principal Component Analysis (PCA) :

Cette technique permet de simplifier la complexité des données tout en préservant au mieux leur structure et leurs relations essentielles. Elle est largement utilisée en statistique, en apprentissage automatique et dans divers domaines scientifiques pour explorer, visualiser et analyser les relations sous-jacentes entre les variables.

L'algorithme de PCA suit les étapes suivantes :

- **Normalisation des données :** Centrer les données pour avoir une moyenne nulle.

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{N} x_{ij}$$

- **Calcul de la matrice de covariance** : Mesure des relations linéaires entre les variables.

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y - \overline{Y})$$

- **Décomposition en valeurs propres** : Trouver les vecteurs propres (axes) et les valeurs propres (mesure de la variance) de la matrice de covariance.

$$CovMat = V \cdot D \cdot V^{-1}$$

- **Sélection des composantes principales** : Choix des axes qui capturent le plus de variance.
- **Projection des données** : Transformation des données vers les nouveaux axes principaux.

$$NewData = OriginalData \times EigenVectors$$

Over-Sampling :

Nous avons utilisé **SVMSMOTE** pour améliorer notre modèle. Cette méthode vise à rendre les données de la classe minoritaire plus équilibrées en introduisant des exemples synthétiques, en se concentrant sur les zones difficiles à classifier avec l'aide de **SVM**. Notre objectif est de renforcer la capacité du modèle à traiter la classe minoritaire (**Classe 1**) en identifiant les cas difficiles via **SVM** et en appliquant la technique **SMOTE** uniquement à ces exemples. Cela permet de réduire le biais potentiel introduit par le déséquilibre des données.

L'algorithme de **SVMSMOTE** suit les étapes suivantes :

- Calculer les vecteurs de support à l'aide de l'algorithme **SVM**.
- Choisir les vecteurs de support de la classe minoritaire.
- Trouver les k plus proches voisins des vecteurs de support de la classe minoritaire.
- Si le nombre de voisins majoritaires est inférieur à $k/2$, générer un nouvel échantillon par extrapolation, sinon générer un nouvel échantillon par interpolation.

Under-Sampling :

Nous avons utilisé **Condensed Nearest Neighbour (CNN)** pour améliorer notre modèle. Cette méthode vise à simplifier les données en éliminant les exemples redondants de la classe majoritaire, permettant ainsi de rendre l'ensemble de données plus équilibré et cela par préserver l'information importante tout en réduisant la complexité de l'ensemble de données. En identifiant les exemples de la classe majoritaire qui sont jugés non essentiels pour la représentation globale, cette méthode contribue à atténuer le déséquilibre entre les classes, améliorant ainsi la performance du modèle, notamment dans le cas où les classes sont désignées comme 1 et 0 dans notre contexte**.**

L'algorithme de **CNN** suit les étapes suivantes :

- Obtenir tous les échantillons minoritaires d'un ensemble.
- Ajouter un échantillon de la classe ciblée (classe à sous-échantillonner) dans et tous les autres échantillons de cette classe dans un ensemble.
- Parcourir l'ensemble, échantillon par échantillon, et classer chaque échantillon à l'aide d'une règle du plus proche voisin.
- Si l'échantillon est mal classé, ajoutez-le à sinon ne rien faire.
- Répéter jusqu'à ce qu'il n'y ait plus d'échantillons à ajouter.

CHAPITRE 2 : IMPLÉMENTATION DES SOLUTIONS

I. INTRODUCTION

Dans la phase d'expérimentation, l'ensemble de données des niveaux de pic d'ozone sur huit heures a été soumis à différentes techniques d'apprentissage automatique pour évaluer leur capacité à prédire ces niveaux dans la région de **Houston**, **Galveston** et **Brazoria**. Les techniques utilisées sont les **SVM** avec et sans **PCA**, ainsi que l'algorithme d'arbre de décision **CART**.

II. ÉTAPES DE L'EXPERIMENTATION

2.1. Préparation des données :

Les données ont été divisées en ensembles distincts pour l'entraînement et le test, avec une répartition de 80 % de la classe 0 et 90 % de la classe 1 pour l'entraînement, et le reste pour les tests. Pour la gestion des valeurs manquantes, nous avons choisi de les gérer en utilisant l'imputation par la moyenne pour garantir la cohérence des données, étant donné le déséquilibre entre les classes 0 et 1, avec 2374 observations pour la classe 0 et seulement 160 observations pour la classe 1.

Pour la partie **SVM**, la classe 0 a été remplacée par la classe -1 pour faciliter les calculs avec des α_i. Afin d'équilibrer les données déséquilibrées, deux approches ont été utilisées : l'Over-Sampling et l'Under-Sampling.

Under-Sampling : La méthode de l'Over-Sampling **CNN** a été appliquée dans la partie **SVM** pour réduire le nombre d'instances de la classe majoritaire. Dans ce cas précis, le compte des instances de la classe minoritaire (classe 1) est de 144, tandis que celui de la classe majoritaire (classe -1) est de 353.

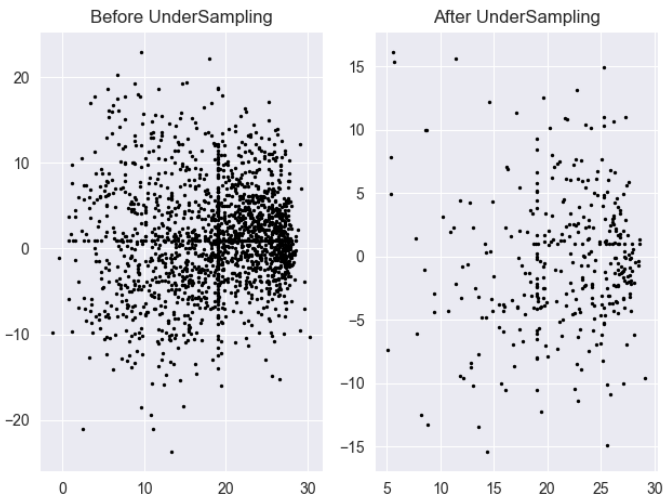

Figure 3 : classe 1 après et avant Under-Sampling.

Over-Sampling : La méthode de l'Under-Sampling **SVMSMOTE** a été appliquée dans la partie de l'arbre de décision **CART** pour augmenter le nombre d'instances de la classe minoritaire. Après cette

opération, le compte des instances pour la classe minoritaire (classe 1) est passé à 1519, et celui de la classe majoritaire (classe 0) est de 1899.

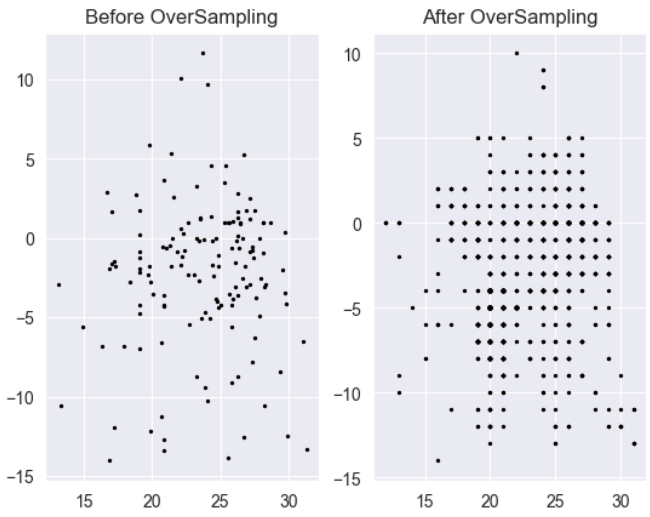

Figure 4 : classe 1 après et avant Over-Sampling.

Ces ajustements ont permis d'obtenir un équilibre relatif entre les deux classes, améliorant ainsi la capacité des modèles d'apprentissage automatique à prédire efficacement les niveaux de pollution par l'ozone, sans biais significatif envers une classe spécifique.

PCA : La méthode de **PCA** a été appliquée dans la partie **SVM** pour réduire la dimensionnalité du dataset. Nous avons sélectionné 7 attributs parmi les 72 attributs existants dans le dataset.

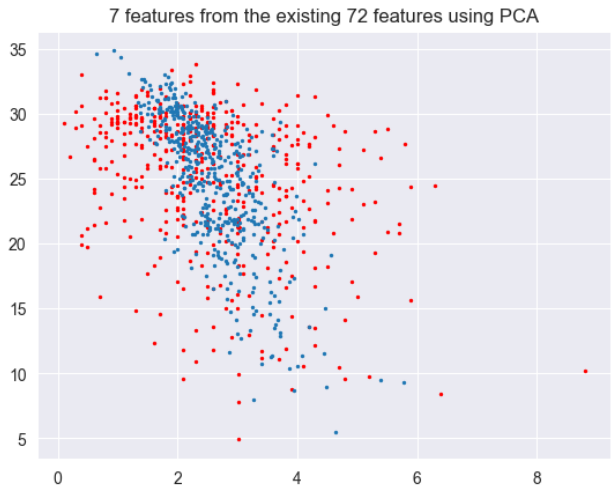

Figure 5 : la dispersion de dataset après PCA de l'attribut WSR11 en fonctions de l'attribut T9.

Ces ajustements ont permis d'obtenir un équilibre relatif entre les deux classes, améliorant ainsi la capacité des modèles d'apprentissage automatique à prédire efficacement les niveaux de pollution par l'ozone, sans biais significatif envers une classe spécifique.

2.2. Entraînement des modèles:

Les modèles **SVM** avec et sans **PCA** ont été entraînés sur l'ensemble d'entraînement pour prédire les niveaux de pic d'ozone. De même, l'arbre de décision **CART** a été construit en utilisant les caractéristiques météorologiques pour prédire ces niveaux.

SVM : Nous avons choisi les valeurs spécifiques de gamma=7 et C=13 pour nos implémentations de **SVM**, soit from scratch ou de la bibliothèque Sci-kit Learn, après avoir effectué plusieurs expérimentations et ajustements. Le paramètre gamma contrôle l'influence d'un exemple individuel. Un gamma plus élevé indique une influence plus spécifique sur le voisinage proche, adaptant potentiellement le modèle à des structures de données plus complexes. Quant au paramètre C, il régularise la classification incorrecte des exemples d'entraînement. Une valeur plus élevée de C permet au modèle de tolérer un plus grand nombre d'erreurs dans la classification, ce qui peut être bénéfique pour des ensembles de données où la frontière de décision est complexe et nécessite plus de flexibilité.

CART : Nous avons sélectionné des valeurs spécifiques comme max_depth=5 et min_samples_leaf=5 pour notre implémentation à partir de zéro après une série de tests visant à optimiser les performances du modèle **CART**. Le paramètre max_depth contrôle la profondeur maximale de l'arbre de décision. En fixant cette valeur à 5, nous limitons la croissance de l'arbre pour éviter un surajustement potentiel. Cela permet de créer un arbre qui capture les structures importantes des données sans devenir trop complexe et trop spécialisé pour l'ensemble d'entraînement spécifique. D'autre part, le paramètre min_samples_leaf détermine le nombre minimum d'échantillons requis pour constituer une feuille de l'arbre. En fixant cette valeur à 5, nous veillons à ce que les feuilles de l'arbre contiennent un nombre suffisant d'échantillons pour généraliser correctement sans être trop spécifiques aux données d'entraînement individuelles.

2.3. Évaluation des modèles :

Les performances des modèles ont été évaluées en utilisant différentes métriques telles que la précision, le **rappel**, le **F1-score**, les courbes **ROC** et les matrices de confusion. Ces mesures permettent de comprendre la capacité de chaque modèle à classifier correctement les niveaux de pic d'ozone.

2.4. Comparaison des performances:

Les résultats obtenus par les différentes techniques ont été comparés pour déterminer celle offrant les meilleures performances prédictives pour les niveaux de pic d'ozone sur huit heures dans cette région spécifique.

L'objectif principal est de sélectionner la méthode ou le modèle le plus performant pour la prédiction des niveaux de pic d'ozone, tout en évaluant la robustesse et la précision de chaque technique d'apprentissage automatique utilisée dans ce contexte spécifique. Cette comparaison permettra de déterminer quelle technique offre la meilleure prédiction pour ces données météorologiques complexe

CHAPITRE 3 : RÉSULTATS

I. PERFORMANCES DES MODÈLES

1.1. SVM from scratch sans PCA:

Pour le modèle de **SVM** from scratch sans **PCA** appliqué à l'ensemble de données a montré une performance globalement solide avec une précision globale de 91.45%. La matrice de confusion révèle une capacité élevée à prédire la classe majoritaire -1 (originalement classe 0), avec 438 prédictions correctes et seulement 37 faux positifs. Cependant, la prédiction de la classe minoritaire 1. est moins précise, avec seulement 11 prédictions correctes sur 16, mais un faible nombre de faux négatifs (seulement 5).

Le rapport de classification souligne que le modèle a une excellente précision pour la classe majoritaire 1 avec 99%, tandis que la précision pour la classe minoritaire 1 est beaucoup plus faible à 23%. Cependant, il est important de noter que le modèle a un bon rappel pour la classe minoritaire 1 à 69%, mais un rappel plus élevé pour la classe majoritaire à 92%.

L'exactitude générale de 91% indique une capacité globale du modèle à prédire correctement les deux classes, bien qu'il puisse y avoir une influence du déséquilibre des classes dans les données. **L'AUC** de 0.8 suggère que le modèle est raisonnablement capable de différencier entre les classes.

1.2. SVM from scratch avec PCA:

Pour ce modèle SVM construit à partir de zéro avec l'application de **PCA** sur la dataset, la performance est marquée par une précision globale de 94,7%. La matrice de confusion montre une forte capacité à prédire la classe majoritaire -1 avec 465 prédictions correctes et seulement 10 faux positifs. Cependant, le modèle n'a fait aucune prédiction correcte pour la classe minoritaire 1, avec 16 faux négatifs.

Le rapport de classification indique une précision élevée pour la classe majoritaire -1 à 97%, mais une précision de 0% pour la classe minoritaire 1. Le modèle a un bon rappel pour la classe majoritaire 98%, mais n'a aucun rappel pour la classe minoritaire 1.

L'accuracy globale de 94,7% est élevée, mais elle est influencée par la forte prévalence de la classe majoritaire dans les données. Cependant, l'AUC de 0,49 suggère que le modèle a du mal à discriminer entre les classes, montrant une performance très faible dans la prédiction de la classe minoritaire.

1.3. SVM de Sci-Kit Learn sans PCA :

Pour l'implémentation de Sci-kit learn du modèle **SVM** sans l'application de **PCA** sur la dataset, la précision globale est de 96,74%. La matrice de confusion montre que le modèle a correctement prédit tous les exemples de la classe majoritaire -1, avec 475 prédictions correctes et aucun faux positif. Cependant, de manière similaire au modèle précédent, il n'a fait aucune prédiction correcte pour la classe minoritaire 1, avec 16 faux négatifs.

Le rapport de classification indique une précision de 97% pour la classe majoritaire -1 et une précision de 0% pour la classe minoritaire 1. Le modèle a un rappel de 100% pour la classe majoritaire, mais aucun rappel pour la classe minoritaire.

L'accuracy globale de 96,74% est élevée, mais elle est biaisée par la forte prévalence de la classe majoritaire dans les données. Cependant, **l'AUC** de 0,50 indique que le modèle a une capacité très limitée à discriminer entre les classes, montrant une performance extrêmement faible dans la prédiction de la classe minoritaire.

1.4. SVM de Sci-Kit Learn avec PCA :

Les mêmes résultats ont été obtenus pour le modèle **SVM** de sci-kit learn avec l'application de **PCA**. Les performances de précision, de rappel et d'accuracy ainsi que la capacité limitée à prédire la classe minoritaire sont comparables à celles du l'implémentation de Sci-kit learn du modèle **SVM** sans l'application de **PCA**.

1.5. CART from scratch :

Pour ce modèle CART construit à partir de zéro, l'accuracy obtenue est de 90,63%. La matrice de confusion révèle une prédiction correcte de 435 exemples de la classe 0. Cependant, il y a eu 40 faux positifs pour cette classe. Pour la classe minoritaire 1, le modèle a prédit correctement 10 cas sur 16, avec 6 faux négatifs.

Le rapport de classification montre une précision élevée pour la classe 0 à 99%, mais une précision relativement faible de 20% pour la classe 1.0. Le modèle a un bon rappel pour la classe 0 à 92%, mais un rappel de 62% pour la classe 1.

L'accuracy globale de 90,63% montre une bonne performance globale du modèle, bien que cela puisse être influencé par le déséquilibre des classes dans les données. **L'AUC** de 0,77 suggère une capacité modérée du modèle à discriminer entre les classes.

1.6. CART de Sci-Kit Learn :

Pour le modèle CART construit à partir de zéro, l'accuracy atteint 91,45%. La matrice de confusion révèle que le modèle a correctement prédit 441 exemples de la classe 0. Cependant, il y a eu 34 faux positifs pour cette classe. Concernant la classe minoritaire 1, le modèle a correctement prédit 8 cas sur 16, mais avec 8 faux négatifs.

Le rapport de classification montre une précision élevée pour la classe majoritaire 0 à 98%, mais une précision relativement faible de 19% pour la classe minoritaire 1. Le modèle a un bon rappel pour la classe 0.0 à 93%, mais un rappel de 50% pour la classe 1.

L'accuracy globale de 91,45% indique une performance globalement solide du modèle, bien que cela puisse être influencé par le déséquilibre des classes dans les données. **L'AUC** de 0,71 suggère une capacité modérée du modèle à discriminer entre les classes.

II. COMPARAISON ET ANALYSE

1.1. SVM sans PCA vs SVM avec PCA :

Le **SVM** sans utiliser **PCA** a donné de meilleurs résultats que celui avec **PCA** car il semble que la réduction de dimension pour cette dataset n'améliore pas les performances. Les caractéristiques initiales semblent contenir des informations cruciales pour la classification, et en conservant toutes les caractéristiques, le modèle a pu capturer des schémas essentiels de manière plus efficace, conduisant à une meilleure performance par rapport à la réduction des caractéristiques via **PCA**.

1.2. SVM from scratch vs SVM de Sci-kit learn:

Le modelé de **SVM** from scratch a donné de meilleurs résultats que celui de Sci-Kit Learn, peut-être parce que la façon dont nous résolvons le problème avec la programmation quadratique pour les alphas en utilisant la bibliothèque **cvxopt** trouve des vecteurs de support plus pertinents. Cette approche spécifique pourrait permettre une sélection plus précise des vecteurs de support, améliorant ainsi la capacité du modèle à généraliser efficacement.

1.3. CART from scratch vs CART de Sci-kit Learn:

Le **CART** from scratch a donné de meilleurs résultats que celui de sci-kit learn, peut-être en raison de l'approche spécifique utilisée dans la construction manuelle du modèle. Il est possible que notre implémentation personnalisée ait ajusté les hyperparamètres ou la logique de manière plus adaptée aux données, offrant ainsi une meilleure capacité de généralisation par rapport à l'implémentation standard de sci-kit learn.

1.4. SVM vs CART :

Le **SVM** a donné des résultats légèrement supérieurs au **CART**, en particulier dans les implémentations from scratch. Cela peut s'expliquer par la capacité du **SVM** à mieux gérer les frontières de décision complexes dans les données, notamment lorsque les relations entre les caractéristiques sont non linéaires. Le **SVM** peut être plus adaptable à ces structures de données complexes, offrant ainsi des performances légèrement supérieures par rapport au **CART** dans ces scénarios.

III. CONCLUSIONS

Après avoir comparé différentes approches, il est clair que **PCA** n'a pas amélioré les performances dans notre cas. En conservant toutes les caractéristiques, le **SVM** sans **PCA** a surpassé celui avec **PCA**,

suggérant que la réduction de dimension n'a pas été bénéfique pour cette dataset. De plus, les implémentations personnalisées de **SVM** et de **CART** ont montré des avantages par rapport à leurs homologues de Sci-kit Learn, probablement en raison de méthodes plus adaptées à nos données.

Dans notre étude, le **SVM** a légèrement surpassé le **CART**, particulièrement dans les implémentations réalisées manuellement. Cette différence pourrait être attribuée à la capacité du SVM à gérer des frontières de décision plus complexes, notamment dans des relations non linéaires entre les caractéristiques. Ces résultats soulignent l'importance de choisir judicieusement les méthodes et les outils d'apprentissage automatique en fonction des caractéristiques spécifiques des données pour obtenir les performances optimales.

REFERENCES ET BIOGRAPHIE

- Zhang, K., Fan, W., & Yuan, X. (2008). Ozone Level Detection. UCI Machine Learning Repository. https://doi.org/10.24432/C5NG6W
- Under-sampling — version 0.11.0. (n.d.). https://imbalanced-learn.org/stable/under_sampling.html#condensed-nearest-neighbors
- Over-sampling — version 0.11.0. (n.d.). https://imbalanced-learn.org/stable/over_sampling.html
- The flowchart of the SVM-SMOTE algorithm. (n.d.-b). ResearchGate. https://www.researchgate.net/figure/The-flowchart-of-the-SVM-SMOTE-algorithm_fig3_358509219
- EzzatEsam. (n.d.). EzzatEsam/SVM-Implementation-Python-QuadraticProgramming. GitHub. https://github.com/EzzatEsam/SVM-Implementation-Python-QuadraticProgramming/blob/master/svm_quad.ipynb
- Meng, Z. (2019). Ground ozone level prediction using machine learning. Journal of Software Engineering and Applications, 12(10), 423–431. https://doi.org/10.4236/jsea.2019.1210026
- Rocklen. (2021, April 19). *Ozone level detection*. Kaggle.

  https://www.kaggle.com/code/rocklen/ozone-level-detection