Rapport du Projet

Sujet : *Risque de cancer cervical*

Encadrante : Prof. Célia da Costa Pereira

Étudiants : Binh Minh Tran, Serigne Diop, Ali Ben Said

IUT de Nice - Sophia Antipolis
24 mars 2025

Table des matières

1 Résumé

L'objectif de ce projet est de mettre en pratique des algorithmes de classification en apprentissage automatique dans le domaine médical, plus précisément pour la classification du cancer du col de l'utérus afin de prédire les risques. Nous avons appliqué quatre algorithmes de classification : SVM, arbre de décision, régression logistique et Naive Bayes. De plus, nous avons utilisé la méthode de rééchantillonnage SMOTE pour équilibrer les données, ce qui nous a permis de comparer l'efficacité des modèles avec et sans cette technique. Ce choix s'explique par le fait que les ensembles de données en santé sont souvent de petite taille, en raison des coûts et des contraintes liés à leur collecte.

2 Introduction

Le cancer du col de l'utérus est un grave problème de santé publique chez les femmes dans le monde. Cette maladie est globalement le deuxième cancer commun chez les femmes [1]. La réduction de la morbidité et de la mortalité du cancer du col passe par un dépistage précoce et un traitement rapide des lésions précancéreuses. Heureusement, cette maladie est évitable. La méthode de prévention ou de détection précoce reste ouverte et difficile. Il existe encore peu de recherches menées sur la détection du cancer du col de l'utérus basées sur les méthodes d'apprentissage automatique.

3 Méthodologie

3.1 Dataset

Dans ce projet, nous exploitons le jeu de données Sobar-72 [2], qui regroupe les réponses de 72 individus, dont 21 atteints du cancer du col de l'utérus (Ca Cervix) et 51 non atteints.

Selon le jeu de données, des théories des sciences sociales sont appliquées afin d'identifier les déterminants comportementaux liés au risque de cancer du col de l'utérus. L'évaluation des individus repose sur un questionnaire de 9 questions par variable. Trois comportements sont considérés comme les principaux facteurs de risque du cancer cervical : l'alimentation, le risque sexuel et l'hygiène personnelle. Ce jeu de données comprend 19 attributs représentant 8 variables liées aux comportements et à leurs déterminants, tels que la perception, l'intention, la motivation, la norme subjective, l'attitude, le soutien social et l'autonomisation.

3.2 Analyse Exploratoire des Données

Selon le boxplot des attributs entre les deux groupes (positif et négatif au Ca Cervix) (Fig.1), les scores des comportements d'hygiène personnelle dans le groupe négatif ont tendance à être plus élevés que dans le groupe positif. Les personnes négatives au Ca Cervix obtiennent des scores plus élevés dans des déterminants tels que l'automatisation, le soutien social, la perception et la norme, comparativement à l'autre groupe. En général, les scores des comportements et de leurs déterminants dans le groupe négatif sont décalés vers la droite, ce qui indique des valeurs plus élevées. De plus, Nous remarquons que l'échelle de la plupart des variables atteint un maximum de 15 points. Cependant, la variable "risque sexuel" a un maximum de 10 points, ce qui suggère

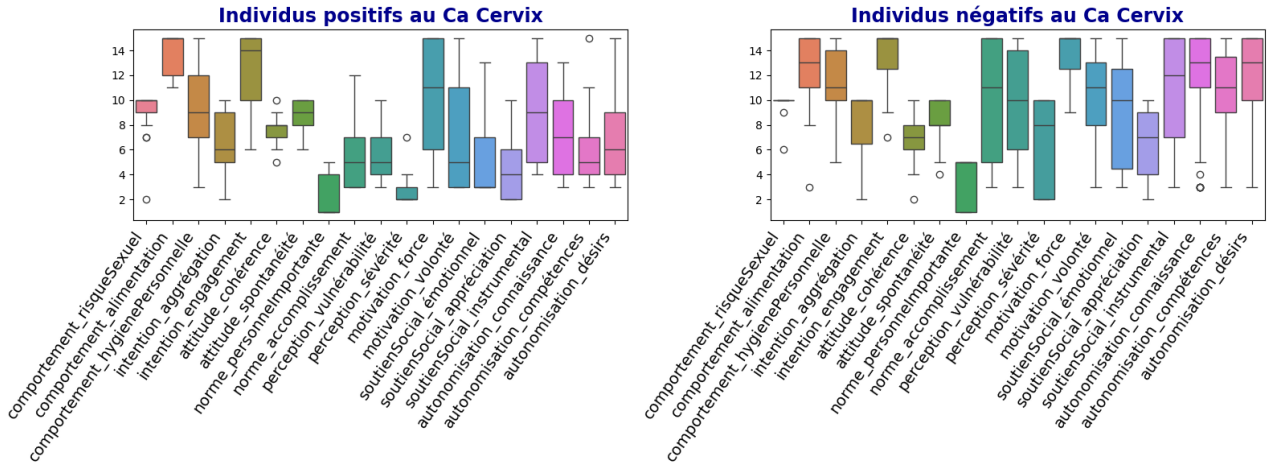


FIGURE 1 – Boxplot du ca cervix

soit qu'elle constitue une exception, soit qu'elle suit également une échelle sur 15 points, mais qu'aucun individu n'a dépassé le seuil de 10.

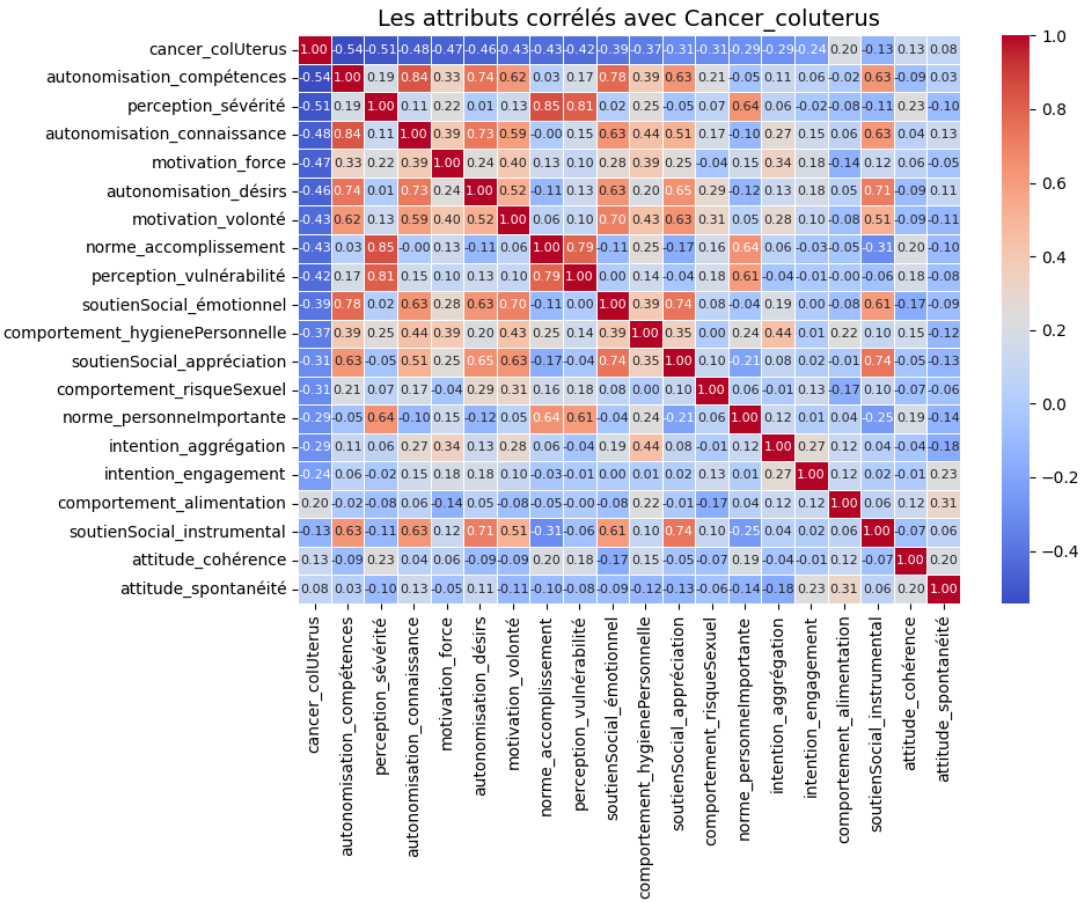


FIGURE 2 – Matrice de corrélation

Évidemment, les trois comportements principaux ne sont pas nécessairement corrélés entre eux (Fig.2). Toutefois, les déterminants 'autonomisation compétences' et 'autonomisation connaissances' présentent un coefficient de corrélation élevé, car ils sont souvent liés à l'acquisition de compétences et de connaissances dans des contextes similaires. De même, 'perception de vulnérabilité' et 'perception de la sévérité' montrent une forte corrélation, car elles concernent la manière dont une personne évalue son propre risque face à une situation donnée. Ces deux

variables, 'perception de vulnérabilité' et 'perception de la sévérité', sont également corrélées avec la norme d'accomplissement, car une perception élevée de la vulnérabilité et de la sévérité incite généralement les individus à adopter des comportements qui sont perçus comme socialement acceptables et conformes aux attentes des normes sociales, ce qui influence leur prise de décision.

3.3 Division du dataset

Nous choisissons de diviser les données de manière stratifiée afin de garantir que l'ensemble d'entraînement reflète fidèlement la distribution des deux catégories. Le découpage est réalisé selon un ratio de 60 % pour l'entraînement et 40 % pour le test, assurant ainsi que le test set dispose d'un volume suffisant de données car notre dataset n'est pas grand, notamment dans le cas où les classes sont déséquilibrées, évitant ainsi un accuracy artificiellement élevé pouvant atteindre 1.

Les 60 % de données d'entraînement seront ensuite utilisés pour effectuer une validation croisée, permettant d'optimiser les hyperparamètres du modèle. Une fois ces derniers déterminés, nous réentraînerons le modèle sur ces mêmes 60 % de données en utilisant les meilleurs hyperparamètres trouvés. Enfin, nous évaluerons la performance finale du modèle sur les 40 % restants (test set), qui n'auront jamais été utilisés durant l'entraînement ni l'optimisation, garantissant ainsi une estimation plus fiable de sa capacité de généralisation.

3.4 Équilibre des données

Ce jeu de données montre un déséquilibre entre les classes : 21 échantillons pour 'cervix' contre 51 pour 'non cervix', la classe minoritaire représentant environ la moitié de la classe majoritaire. Bien que ce déséquilibre soit modéré, il peut biaiser l'accuracy, par exemple si un modèle prédit toujours la classe majoritaire, obtenant ainsi une accuracy élevée sans détecter la classe minoritaire[3]. Donc, nous avons choisi de la mettre en œuvre afin de comparer les résultats.

Nous abordons le problème d'un dataset déséquilibré. Pour y remédier, nous appliquerons la méthode SMOTE (Synthetic Minority Oversampling Technique) [3], qui a prouvé son efficacité en améliorant l'AUC [4], un indicateur plus robuste dans ce contexte. Contrairement à une simple duplication des données, SMOTE génère des échantillons synthétiques en interpolant entre les échantillons de la catégorie minoritaire et leurs voisins proches.

3.5 Validation croisée (Cross-validation)

Pour évaluer la performance de notre modèle et affiner l'hyperparamètre, nous avons utilisé la technique de validation croisée à k plis (k-fold cross-validation). Dans cette méthode, nous divisons l'ensemble de données en k sous-ensembles de taille égale. À chaque pli, un sous-ensemble est utilisé comme ensemble de test, tandis que les $k-1$ autres sous-ensembles sont utilisés pour entraîner le modèle. Ainsi, chaque instance de l'ensemble d'entraînement est testée une fois. Enfin, nous évaluons la performance globale du modèle en calculant la moyenne des résultats obtenus lors des k plis.

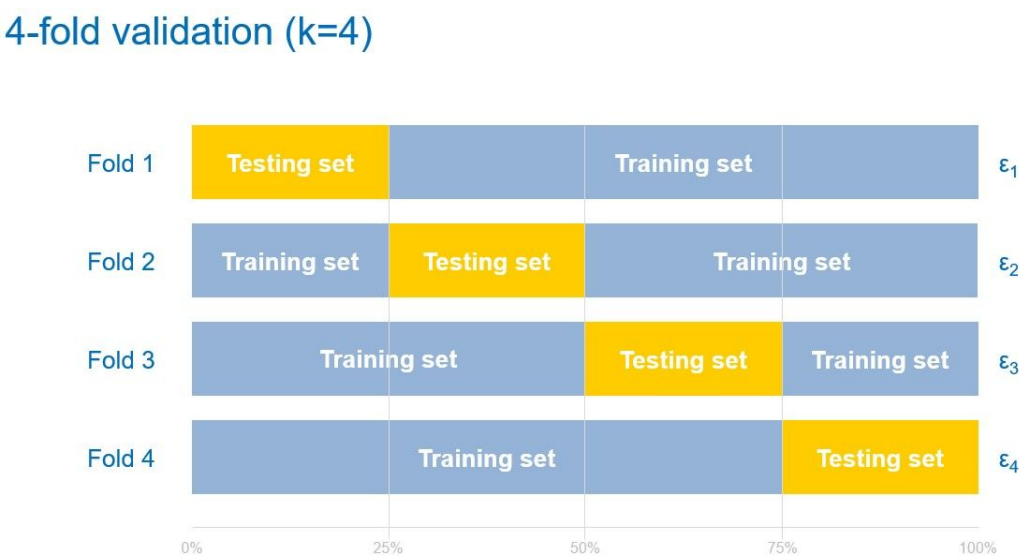


FIGURE 3 – K-fold CrossValidation

Il est recommandé de fixer k à des valeurs plus grandes comme 5, 10 ou 20, selon les règles empiriques (rules of thumb) [5]. Par exemple, si k = 2 (petit), cela signifie que 50% des données sont utilisées pour l'entraînement et 50% pour les tests, ce qui peut stimuler l'overfitting. À l'inverse, avec une valeur de k plus grande, le modèle peut s'entraîner sur un plus grand nombre d'exemples, ce qui permet d'obtenir des résultats plus stables et fiables.

3.6 Sélection du modèle

Support Vector Machine (SVM)

Les Machines à Vecteurs de Support (SVM) sont des algorithmes de classification très répandus, initialement conçus pour le binaire avant d'être adaptés à la régression et au classement [6]. Leur efficacité provient de leur capacité à maximiser la marge entre classes et à gérer des espaces de grande dimension, même lorsque les données ne sont pas linéairement séparables.
Dans ce projet, nous utilisons divers noyaux SVM pour entraîner un modèle sur les données du cancer du col de l'utérus et tester sa capacité de généralisation sur un nouveau jeu de données. Nous comparons un noyau linéaire et trois noyaux non linéaires, et évaluons l'effet de la méthode SMOTE pour équilibrer les données d'entraînement.

Naive Bayes

Les modèles Naive Bayes sont des algorithmes de classification probabilistes basés sur le théorème de Bayes, reposant sur l'hypothèse d'indépendance conditionnelle des variables. Initialement conçus pour des tâches simples, ils se sont révélés efficaces dans divers contextes, notamment grâce à leur rapidité et leur capacité à gérer des données de grande dimension. Dans ce projet, nous utilisons trois variantes de Naive Bayes pour entraîner un modèle sur les données du cancer du col de l'utérus et tester sa capacité de généralisation sur un nouveau jeu de données. Ces variantes sont GaussianNB, MultinomialNB et BernoulliNB, et nous évaluons l'effet de la méthode SMOTE pour équilibrer les données d'entraînement.

Regression Logistique

La régression logistique est un modèle de classification binaire largement utilisé pour estimer le risque de cancer du col de l'utérus en fonction de variables comportementales. Appliquée

à notre dataset, une première estimation a été réalisée avec les paramètres par défaut, suivie d'une optimisation via la validation croisée et l'ajustement des hyperparamètres. Cette approche permet d'améliorer la précision des prédictions et de mieux identifier les facteurs de risque. Dans la suite de notre étude, nous utiliserons le modèle optimisé pour affiner nos analyses et renforcer la fiabilité des résultats.

Abre de Décision

L'arbre de décision est une approche efficace pour estimer le risque de cancer du col de l'utérus grâce à sa capacité à gérer des données complexes et à capturer les interactions entre les variables. Son interprétabilité permet aux professionnels de santé de visualiser facilement les critères influençant le diagnostic, facilitant ainsi la prise de décision médicale. Appliqué à notre dataset, un premier modèle a été entraîné avant d'être optimisé par validation croisée et ajustement des hyperparamètres afin d'améliorer sa robustesse et sa précision. Cette optimisation a notamment consisté à comparer les critères d'impureté, tels que Gini et l'entropie, pour sélectionner celui offrant la meilleure séparation des classes et réduire le surajustement. Dans la suite de notre étude, nous utiliserons le modèle optimisé pour approfondir nos analyses et renforcer la fiabilité des prédictions

3.7 Optimisation des hyperparamètres

Afin de déterminer la meilleure valeur du hyperparamètre C pour chaque noyau/modèle, nous avons défini un ensemble de valeurs possibles pour C. Ensuite, nous avons entraîné le modèle pour chaque valeur de C.

En complément, nous avons appliqué la validation croisée K-Fold avec $K = 5$. Pour chaque valeur de C, nous avons entraîné K modèles distincts, un pour chaque pli de validation.

La précision moyenne pour chaque valeur de C a été calculée à partir des scores obtenus sur chaque ensemble de test dans la validation croisée, selon la formule suivante :

$$\bar{P}(C) = \frac{1}{K} \sum_{i=1}^{K} P_i(C)$$

où $P_i(C)$ est la précision du modèle entraîné avec la valeur C sur le **jeu de test** du $i^{\text{ème}}$ pli.

Nous sélectionnons ensuite la meilleure valeur de C en appliquant la règle suivante :

$$C^* = \arg\max_C \bar{P}(C)$$

Si plusieurs valeurs de C donnent la même précision moyenne, nous retenons la première rencontrée, car l'ordre n'a pas d'importance dans ce cas.

Dans ce projet, nous avons uniquement effectué le fine-tuning des hyperparamètres pour les modèles SVM, régression logistique et arbre de décision, car ces modèles possèdent des hyperparamètres importants à optimiser, contrairement au modèle Naive Bayes qui, bien qu'il ait des paramètres, n'en nécessite généralement pas une optimisation poussée.

4 Évaluation des modèles

4.1 SVM

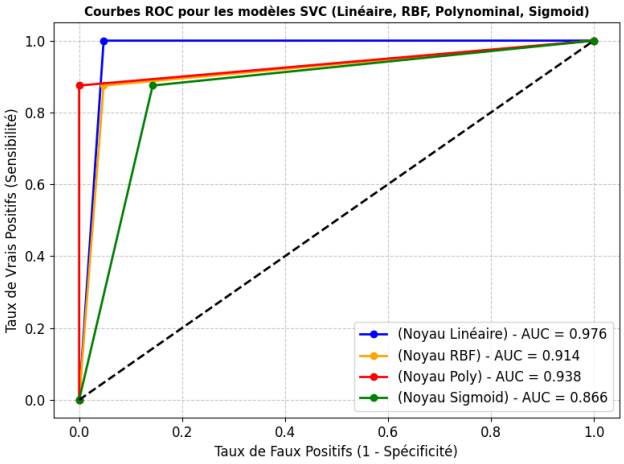


FIGURE 4 – ROC de SVM

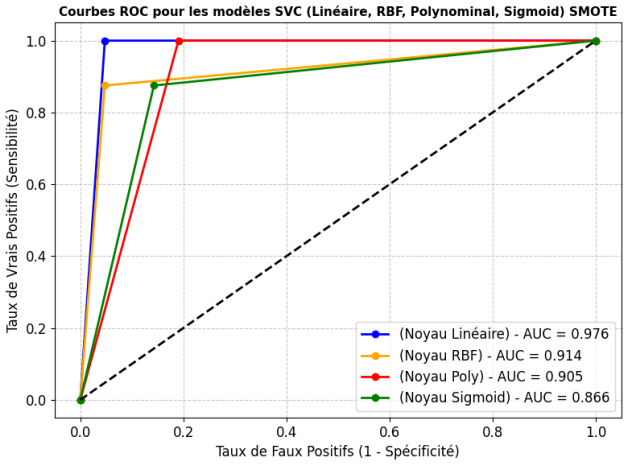


FIGURE 5 – ROC de SVM-SMOTE

Sur la base des courbes ROC-AUC obtenues pour les deux approches, avec et sans l'application de SMOTE (voir Fig.4 et Fig.5), il est difficile de tirer une conclusion définitive quant à une différence notable entre elles. En effet, les valeurs d'AUC semblent ne pas présenter d'écart significatif. Cependant, un point mérite attention : le noyau polynomial montre une variation marquée, avec une diminution de l'AUC après l'application de SMOTE. Ce résultat va à l'encontre des études précédentes qui montrent une amélioration positive de l'indice AUC lorsqu'on applique SMOTE [4].

Noyau	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUC**
Linéaire	**0.96552**	**0.94118**	0.88889	**1.0**	**0.97619**
RBF	0.93103	0.875	0.875	0.875	0.91369
Polynomiale	0.96552	0.93333	**1.0**	0.875	0.9375
Sigmoid	0.86207	0.77778	0.7	0.875	0.86607

TABLE 1 – Comparaison des noyaux de SVM sans SMOTE

Noyau	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUC**
Linéaire	**0.96552**	**0.94118**	**0.88889**	**1.0**	**0.97619**
RBF	0.93103	0.875	0.875	0.875	0.91369
Polynomiale	0.86207	0.800	0.66667	1.0	0.90476
Sigmoid	0.86207	0.77778	0.7	0.875	0.86607

TABLE 2 – Comparaison des noyaux de SVM avec SMOTE - Taille du ratio de rééchantillonnage de 100%

À partir des données présentées dans les tableaux (Tab.1 et Tab.2), on observe que les noyaux linéaire et RBF ne montrent aucune variation dans leurs métriques d'évaluation avant et après l'application de SMOTE. Cette stabilité est cohérente avec les découvertes de Shatnawi, Raed

(2012) [7], selon lesquelles un suréchantillonnage de la classe minoritaire par un facteur de 2 ou 3 n'entraîne pas de modification notable de l'AUC pour un modèle SVM, comme en témoignent les valeurs inchangées de l'AUC pour le noyau linéaire (0.97619) et RBF (0.91369). En revanche, le noyau polynomial subit une baisse de performance, avec une Accuracy passant de 0.96552 à 0.86207 et un F1-Score diminuant de 0.93333 à 0.800, bien que le Recall augmente de 0.875 à 1.0. Cette dégradation peut s'expliquer par le fait que SMOTE, en générant des échantillons synthétiques, modifie la frontière de décision, ce qui semble défavoriser la capacité de séparation du noyau polynomial, un effet moins marqué sur les noyaux linéaires. Par ailleurs, le noyau Sigmoid conserve des métriques identiques, suggérant une insensibilité à l'application de SMOTE. Si l'on conclut que ce jeu de données ne se prête pas pleinement à SMOTE lorsque l'objectif est d'optimiser l'Accuracy ou l'AUC, il convient toutefois de noter que l'amélioration du Recall pour le noyau polynomial pourrait être bénéfique selon les priorités de l'analyse.

Nous proposons le modèle SVM avec noyau linéaire (SMOTE/non SMOTE n'a pas d'importance car ils donnent les mêmes résultats) dans ce projet afin de le comparer avec d'autres modèles de classification.

4.2 Naive Bayes

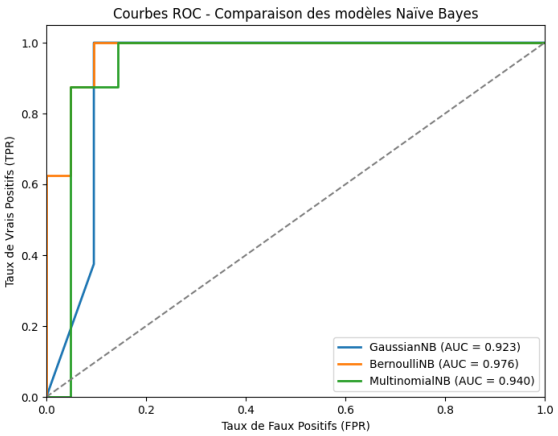


FIGURE 6 – ROC de NB

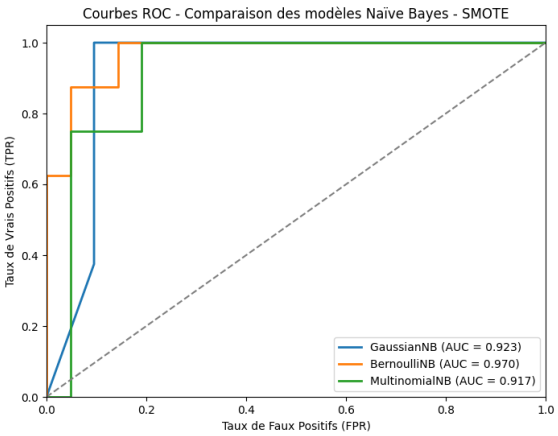


FIGURE 7 – ROC de NB-SMOTE

Sur la base des courbes ROC (Fig.6,Fig. 7), on peut observer que SMOTE n'a pas d'impact sur le modèle gaussien. Cependant, l'AUC des modèles Bernoulli et Multinomial diminue, en particulier celle du modèle Multinomial, qui chute de manière significative de 0,94 à 0,91. Cette baisse de l'AUC après l'application de SMOTE peut sembler contredire les conclusions sur les effets positifs de SMOTE sur l'AUC. Cela pourrait s'expliquer par un ensemble de données de petite taille et un suréchantillonnage insuffisant de la classe minoritaire.

Modèle	Accuracy	F1-Score	Precision	Recall	AUC
GaussianNB	0.89655	0.82353	0.77778	0.875	0.92262
BernoulliNB	0.93103	0.875	0.875	0.875	0.97024
MultinomialNB	0.79310	0.72727	0.57143	1.0	0.91667

TABLE 3 – Comparaison des modèles Naïve Bayes avec SMOTE

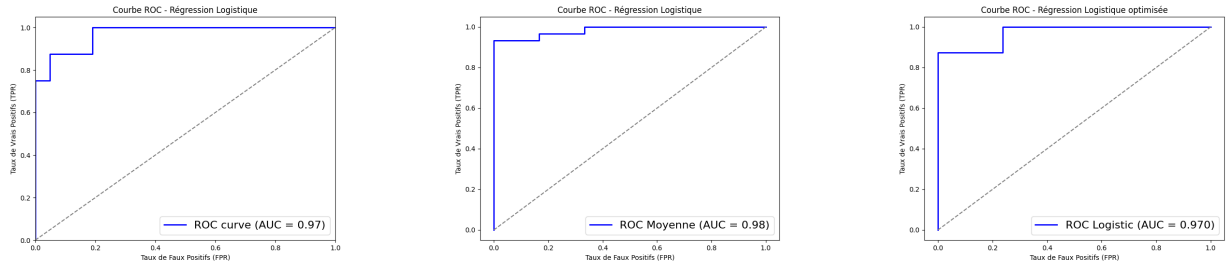
Modèle	Accuracy	F1-Score	Precision	Recall	AUC
GaussianNB	0.89655	0.82353	0.77778	0.875	0.92262
BernoulliNB	0.89655	0.84211	0.72727	1.0	0.97619
MultinomialNB	0.75862	0.36364	0.66667	0.25	0.94048

TABLE 4 – Comparaison des modèles Naïve Bayes sans SMOTE

D'après les tableaux (Tab. 4, Tab. 7), GaussianNB reste stable et performant (accuracy 0,89655, AUC 0,92262), sans impact notable de SMOTE, grâce à des données numériques conformes à l'hypothèse de normalité. BernoulliNB, adapté aux variables binaires, atteint un rappel parfait sans SMOTE, et avec SMOTE, il équilibre précision et rappel (accuracy 0,93103, F1-Score 0,875). MultinomialNB, conçu pour les comptages, montre des performances faibles sans SMOTE (F1-Score 0,36364, Recall 0,25), mais s'améliore avec SMOTE (F1-Score 0,72727, Recall 1), au détriment de la précision (0,57143), générant plus de faux positifs.

Nous décidons de proposer le meilleur modèle (BernoulliNB avec SMOTE) afin de le comparer avec d'autres algorithmes de classification.

4.3 Regression Logistique



(a) ROC de Logistique défault - SMOTE

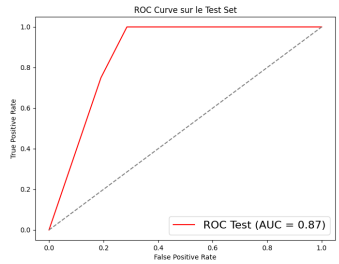
(b) ROC de Logistiquen Cross-Validation - SMOTE

(c) ROC de Logistique Optimisé - SMOTE

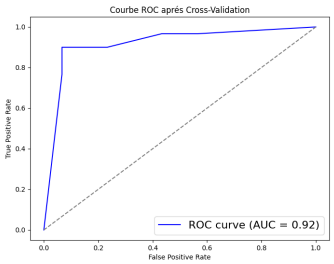
FIGURE 8 – Comparaison des courbes ROC des différents arbres de décision.

L'évaluation du modèle de régression logistique pour la détection du cancer du col de l'utérus montre une performance globalement élevée. Le modèle par défaut affiche une précision de 0,90 avec une AUC de 0,97, indiquant une bonne capacité à distinguer les cas positifs et négatifs. Toutefois, le rappel pour la classe 1 (cas positifs) est de 0,88, ce qui signifie que certains cas de cancer pourraient être manqués. L'optimisation par validation croisée améliore encore la précision moyenne à 0,95 et l'AUC à 0,98, renforçant ainsi la robustesse du modèle. Enfin, après un ajustement des hyperparamètres, le modèle optimisé conserve une précision globale de 0,90 avec un AUC de 0,97, mais réduit les faux positifs et faux négatifs, comme l'indique sa matrice de confusion (19 vrais négatifs, 7 vrais positifs). Ces résultats suggèrent que l'optimisation stabilise le modèle sans surajustement et assure une bonne généralisation sur de nouvelles données.

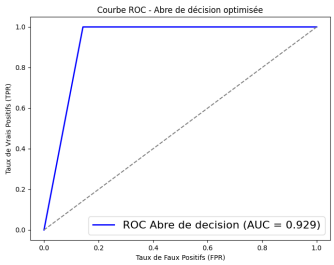
4.4 Abre de Décision



(a) ROC de l'Arbre de décision défault - SMOTE



(b) ROC de l'Arbre de décision CrossValidation - SMOTE



(c) ROC de l'Arbre de Décision Optimisé - SMOTE

FIGURE 9 – Comparaison des courbes ROC des différents arbres de décision.

L'évaluation du modèle d'arbre de décision pour la classification du risque de cancer du col de l'utérus montre une amélioration progressive des performances à travers l'optimisation. Le modèle par défaut atteint une précision de 0,79 avec une AUC de 0,87, indiquant une capacité correcte à différencier les cas positifs et négatifs, bien qu'avec une marge d'erreur notable. L'intégration de la validation croisée améliore la précision à 0,90 et l'AUC à 0,92, garantissant une meilleure robustesse du modèle. Après ajustement des hyperparamètres (gini comme critère d'impureté, une profondeur maximale de 5 et une sélection optimisée des variables), le modèle optimisé affiche une précision de 0,90 avec une AUC de 0,929, tout en réduisant les erreurs de classification. Sa matrice de confusion montre une bonne capacité à identifier correctement les cas positifs (rappel de 1,00 pour la classe 1), bien que la précision pour cette classe soit légèrement inférieure (0,73). Ces résultats suggèrent que l'optimisation améliore la stabilité et la généralisabilité du modèle, en réduisant les risques de surajustement et en assurant des prédictions plus fiables pour l'identification des cas de cancer.

4.5 Comparaison des modèles

Dans l'ensemble, nous avons choisi de comparer uniquement les performances obtenues avec SMOTE afin d'assurer une évaluation équitable. En effet, bien que Naïve Bayes et SVM aient été testés avec et sans SMOTE, les modèles de régression logistique et d'arbre de décision n'ont été entraînés qu'avec SMOTE. Ainsi, pour garantir une comparaison homogène, nous concentrons notre analyse sur le cas d'utilisation de SMOTE. Il apparaît que Naïve Bayes fonctionne très bien avec SMOTE, tandis que le SVM conserve ses performances élevées, ce qui confirme que la comparaison des quatre algorithmes dans ce contexte est pertinente.

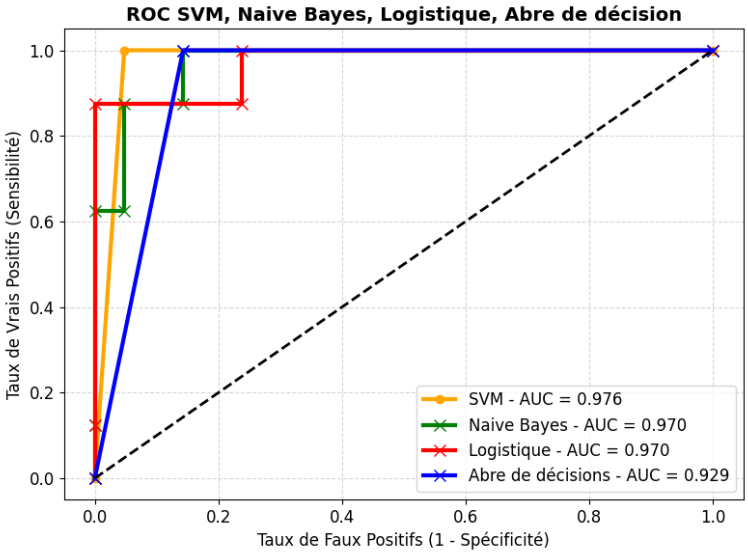


FIGURE 10 – ROC de SVM,NB,Logistique,Abre de décision - SMOTE

Modèle	Accuracy	F1-Score	Precision	Recall	AUC
SVM	0.96552	0.94118	0.88889	1.0	0.97619
Naïve Bayes	0.93103	0.875	0.875	0.875	0.97024
Régression Logistique	0.89655	0.82353	0.77778	0.875	0.97024
Arbre de Décision	0.89655	0.84211	0.72727	1.0	0.92857

TABLE 5 – Comparaison des modèles de classification

Selon les resultats (voir Fig.10, Tab 5), Nous constatons que le modèle SVM se démarque avec une accuracy de 0,96552, un F1-Score de 0,94118, un recall parfait et une AUC de 0,97619, ce qui indique une excellente capacité à discriminer les classes. Le Naïve Bayes suit avec une accuracy de 0,93103 et un équilibre satisfaisant entre précision, recall et F1-Score (tous à 0,875, AUC de 0,97024). La régression logistique et l'arbre de décision, tous deux avec une accuracy de 0,89655, présentent des F1-Scores respectifs de 0,82353 et 0,84211 ; l'arbre de décision affiche toutefois un recall parfait au prix d'une précision moindre (0,72727) et une AUC de 0,92857. En conclusion, le SVM s'avère être le modèle le plus performant dans notre contexte.

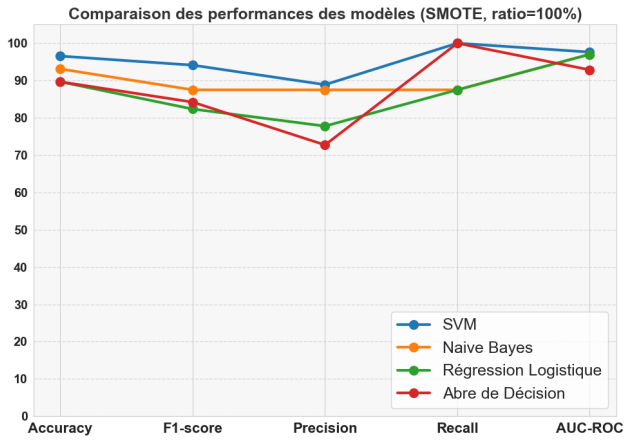


FIGURE 11 – Comparaison des performances des modèles (SMOTE, ratio=100%)

D'après le graphique (Fig. 11), l'écart d'accuracy entre les algorithmes est modéré. De plus, la régression logistique et l'arbre de décision ont une précision plus faible que le SVM et le Naïve Bayes. Cela indique que la régression logistique et l'arbre de décision génèrent davantage de faux positifs. De plus, l'arbre de décision montre un F1-Score faible et un recall élevé, ce qui signifie qu'il identifie bien les cas positifs, mais avec un nombre élevé de faux positifs, réduisant ainsi la précision.

5 Discussion et Perspectives

Enfin, nous avons constaté que la méthode SMOTE pourrait être plus efficace si les classes sont fortement déséquilibrées, avec un ratio de 90/10 sur l'ensemble des données, ou si la classe minoritaire représente moins de 25 % de la classe majoritaire. Dans notre cas spécifique, où la classe minoritaire représente près de la moitié de la classe majoritaire, l'utilisation de SMOTE ne s'est pas avérée très efficace, car plusieurs modèles n'ont montré aucun changement significatif dans les métriques avant et après l'application de SMOTE. De plus, selon Raed Shatnawi (2012)[7], augmenter la taille de l'échantillon de la classe minoritaire de 100 % ou 200 % (soit multiplier par 2 ou par 3) n'entraîne pas de différences notables. Dans notre projet, pour équilibrer les deux groupes, un rééchantillonnage de 100 % s'est révélé suffisant pour atteindre un équilibre des données, ce qui explique la stabilité des performances des modèles avant et après l'application de SMOTE. Cela souligne également les limites de l'utilisation de SMOTE sur un dataset de taille réduite. Cela souligne également un problème lié à la taille réduite du dataset lorsque l'on applique SMOTE.

De plus, pour choisir le meilleur modèle, il est important de prendre en compte la vitesse de calcul sur des jeux de données plus grands. Actuellement, avec les données SOBAR-72, ce critère n'est pas très significatif car l'entraînement des modèles prend seulement quelques secondes. Cependant, lorsque la taille des données augmente, les modèles plus complexes auront des temps de calcul plus longs.

En conclusion, nous recommandons le SVM pour ses performances exceptionnelles en précision et en AUC. Toutefois, dans le domaine médical, l'arbre de décision reste attractif grâce à son interprétabilité et sa capacité à détecter tous les cas positifs (rappel 1), ce qui est crucial pour la détection du cancer du col de l'utérus. Cependant, le SVM n'a pas donné de bons résultats dans notre cas spécifique.

Dans le cadre de ce projet, nous pouvons approfondir nos recherches en appliquant des techniques de sélection de variables, en augmentant la taille des données d'entraînement et en explorant des approches hybrides afin d'améliorer la généralisation des modèles. L'intégration de ces méthodes dans les systèmes cliniques pourrait permettre un diagnostic plus rapide et plus précis du cancer du col de l'utérus, contribuant ainsi à une détection précoce et à une meilleure prise en charge des patientes.

6 Annexe

Architecture du Projet

```
Cervical-Cancer-Risk-Classification/
|-- .gitattributes
|-- .gitignore
|-- Code/
|   |-- Comparer les modèles.ipynb
|   |-- NaiveBayes.ipynb
|   |-- Preparation.ipynb
|   |-- SVM.ipynb
|   |-- TreeDecision.ipynb
|   |-- Variables/
|   |   |-- roc_data_NB.pkl
|   |   |-- roc_data_log.pkl
|   |   |-- roc_data_svm.pkl
|   |   |-- roc_data_tree.pkl
|   |   |-- variables.pkl
|   |-- __pycache__/
|   |   |-- fonctionUtile.cpython-311.pyc
|   |-- fonctionUtile.py
|   |-- logistique_regression.ipynb
|-- Data/
|   |-- sobar-72.csv
|-- Image/
|   |-- ROC_4modeles.png
|   |-- ROC_NB.png
|   |-- ROC_NB_smote.png
|   |-- ROC_SVM.png
|   |-- ROC_SVM_SMOTE.png
|   |-- ROC_dt1.png
|   |-- ROC_dt2.png
|   |-- ROC_dt3.png
|   |-- ROC_log1.png
|   |-- ROC_log2.png
|   |-- ROC_log3.png
|   |-- ROC_opt.png
|   |-- boxplot_ca_cervix.png
|   |-- cropped-But-SD.png
|   |-- matrice_correlation.png
|   |-- perform_4modeles.png
|-- Rapport/
|   |-- Cervical_Rapport.pdf
|-- README.md
```

Références

[1] A. Saha et al. Awareness of cervical cancer among female students of premier colleges in kolkata, india. *Asian Pacific Journal of Cancer Prevention*, 11(4) :1085–1090, 2010.

[2] Sobar, R. Machmud, and Adi Wijaya. Behavior determinant based cervical cancer early detection with machine learning algorithm. *Advanced Science Letters*, 22 :3120–3123, 10 2016.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[4] Ahmed Jameel Mohammed, Masoud Muhammed Hassan, and Dler Hussein Kadir. Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(3) :3161–3172, 2020.

[5] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[6] Hwanjo Yu and Sungchul Kim. Svm tutorial : Classification, regression, and ranking. *Handbook of Natural Computing*, 01 2012.

[7] Raed Shatnawi. Improving software fault-prediction for imbalanced data. pages 54–59, 03 2012.