Rapport du Projet

Sujet : *Risque de cancer cervical*

Encadrante : Prof. Célia da Costa Pereira

Étudiant : Binh Minh Tran, Serigne Diop, Ali

IUT de Nice - Sophia Antipolis
21 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

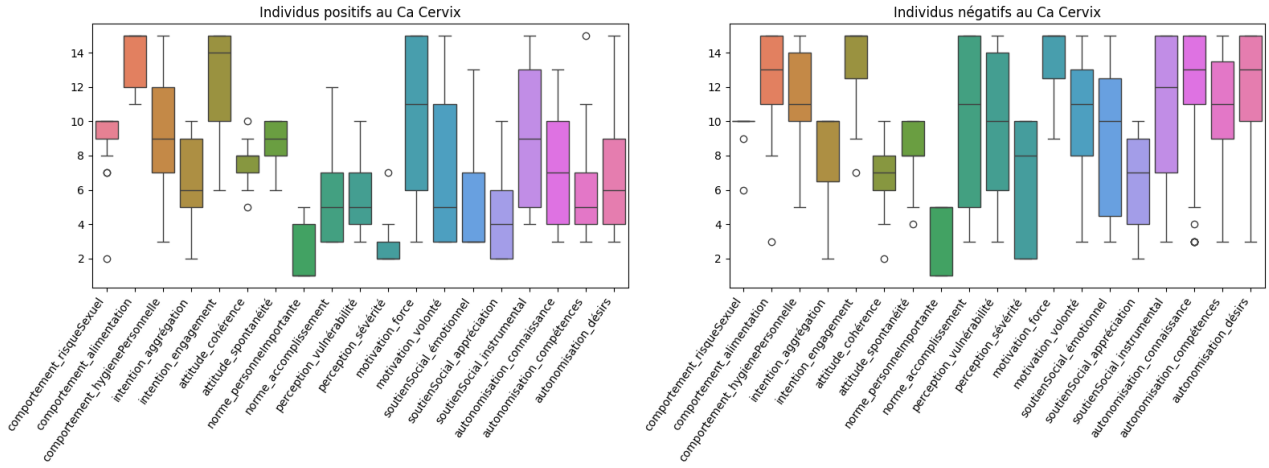

FIGURE 1 – Boxplot du ca cervix

leurs déterminants dans le groupe négatif sont décalés vers la droite, ce qui indique des valeurs plus élevées.

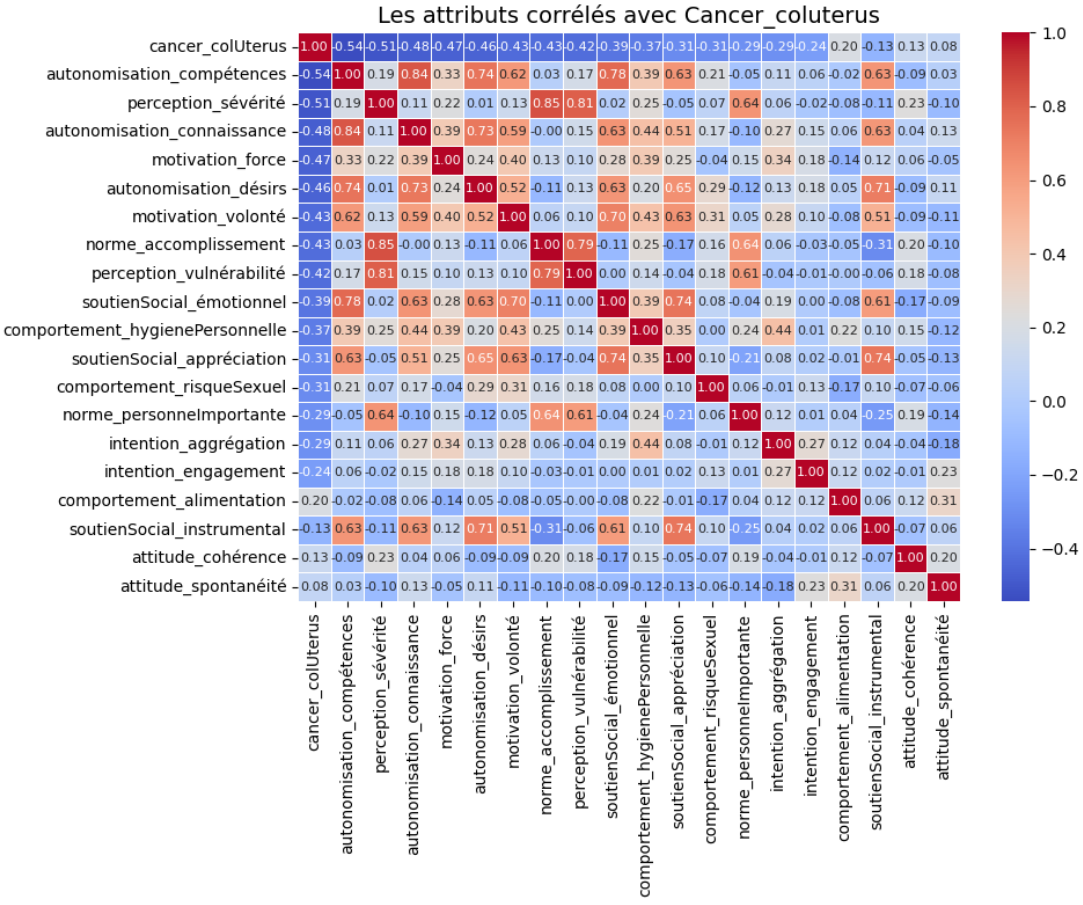

FIGURE 2 – Matrice de corrélation

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.

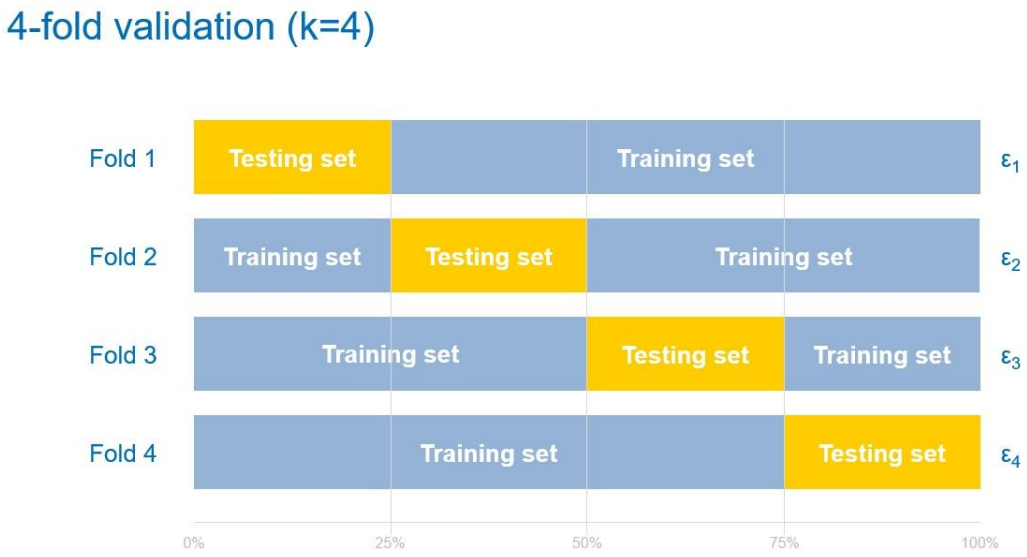

FIGURE 3 – K-fold CrossValidation

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

Nous allons commencer par une brève présentation de l'algorithme utilisé dans notre projet. Ensuite, nous expliquerons comment nous avons entraîné le modèle, en précisant les noyaux

utilisés et la manière dont nous avons comparé les résultats. En gros, nous détaillerons la méthode adoptée pour sélectionner le meilleur modèle. Consulter le SVM ci-dessus

Regression Logistique

ici , ácdasdasd

Abre de Décision

ici , ácdasdasd

3.7 Optimisation des hyperparamètres

Pour améliorer la performance du modèle, nous avons utilisé la méthode GridSearch de la bibliothèque scikit-learn pour trouver les meilleurs hyperparamètres. Plus précisément, afin de déterminer la meilleure valeur du hyperparamètre C pour chaque noyau, 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

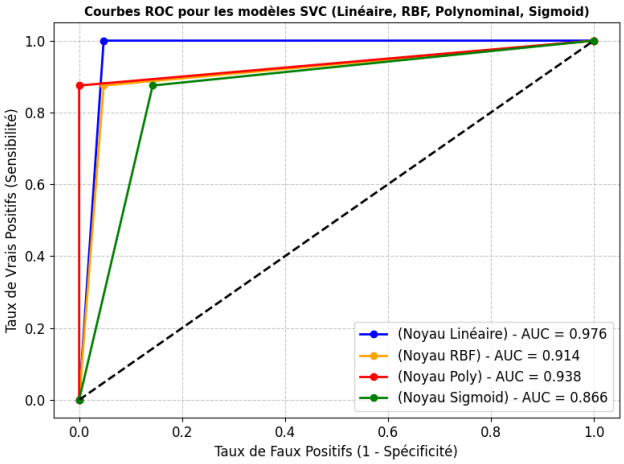

FIGURE 4 – ROC de SVM

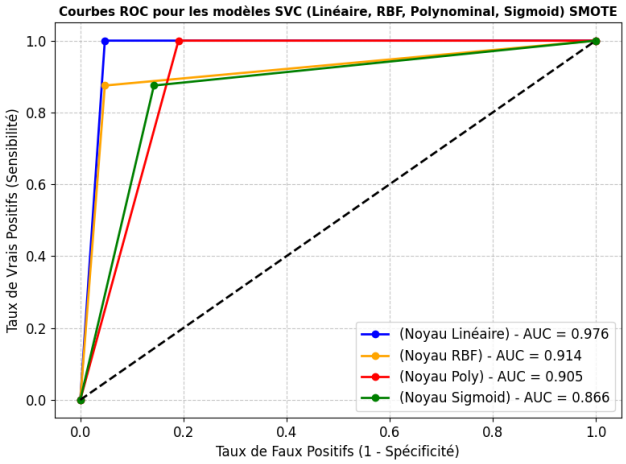

FIGURE 5 – ROC de SVM-SMOTE

4.1 SVM

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]. Nous pouvons donc conclure que l'application de SMOTE sur ce jeu de données ne semble pas efficace, car les classes ne sont pas très déséquilibrées et le jeu de données est encore trop petit, avec moins de 100 échantillons.

Modèle	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

Modèle	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 90%

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 dans ce projet afin de le comparer avec d'autres modèles de classification.

4.2 Naive Bayes

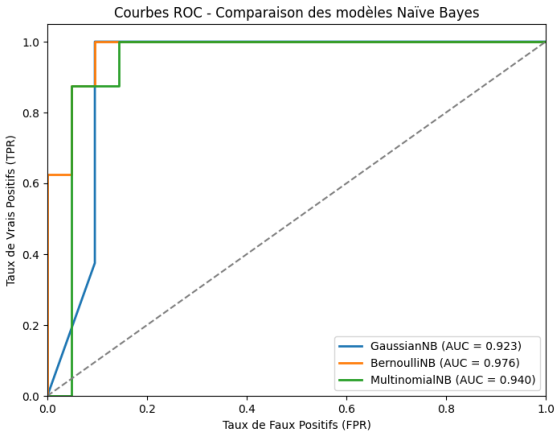

FIGURE 6 – ROC de NB

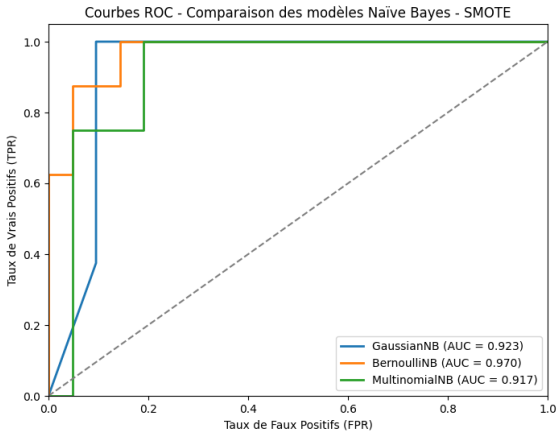

FIGURE 7 – ROC de NB-SMOTE

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

4.3 Regression Logistique

ici , ácdasdasd

4.4 Abre de Décision

ici , ácdasdasd

4.5 Comparer les modèles

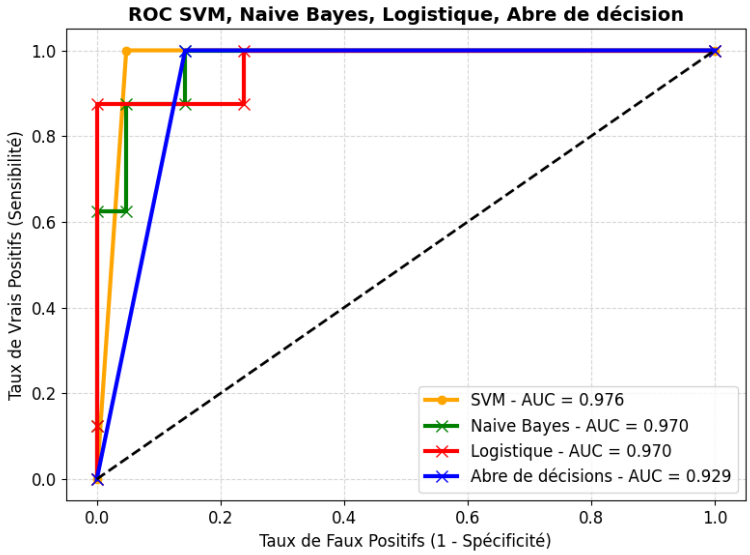

FIGURE 8 – ROC de SVM,NB,Logistique,Abre de décision

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

5 Discussion et Perspectives

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.