Rapport TP4

Pourquoi la stratification est-elle importante ici ?

- La stratification est importante pour maintenir les proportions des classes dans les ensembles d'entraînement et de test.

- Cela garantit une répartition équitable entre spam et ham.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1448
1	1.00	0.88	0.94	224
accuracy			0.98	1672
macro avg	0.99	0.94	0.96	1672
weighted avg	0.98	0.98	0.98	1672

Figure 1 : Rapport de classification du modèle SVM

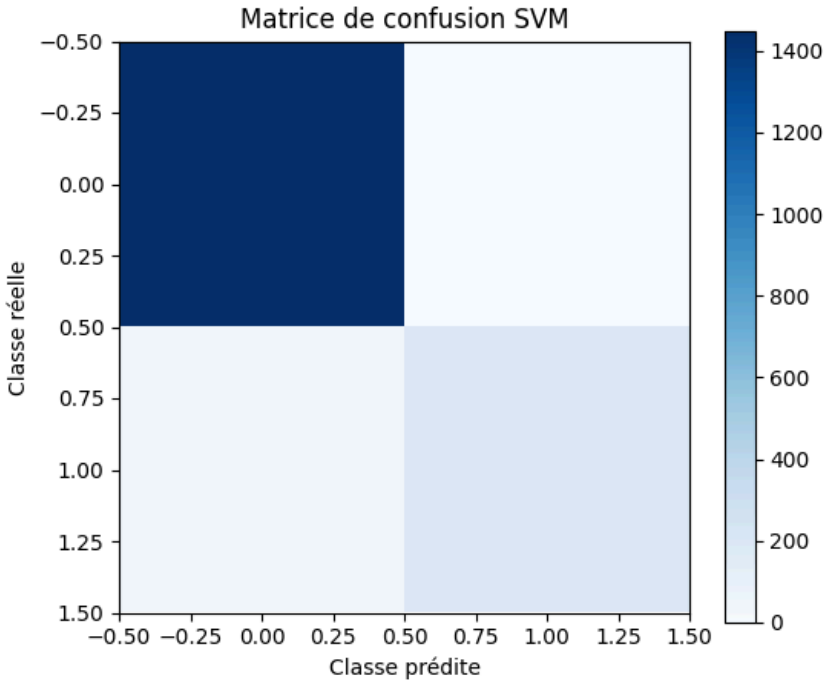


Figure 2 : Matrice de confusion du modèle SVM

Quelles erreurs sont les plus fréquentes ?

Les erreurs les plus fréquentes se trouvent généralement dans les faux négatifs (spam prédit comme ham), ce qui peut être problématique dans un filtre anti-spam.

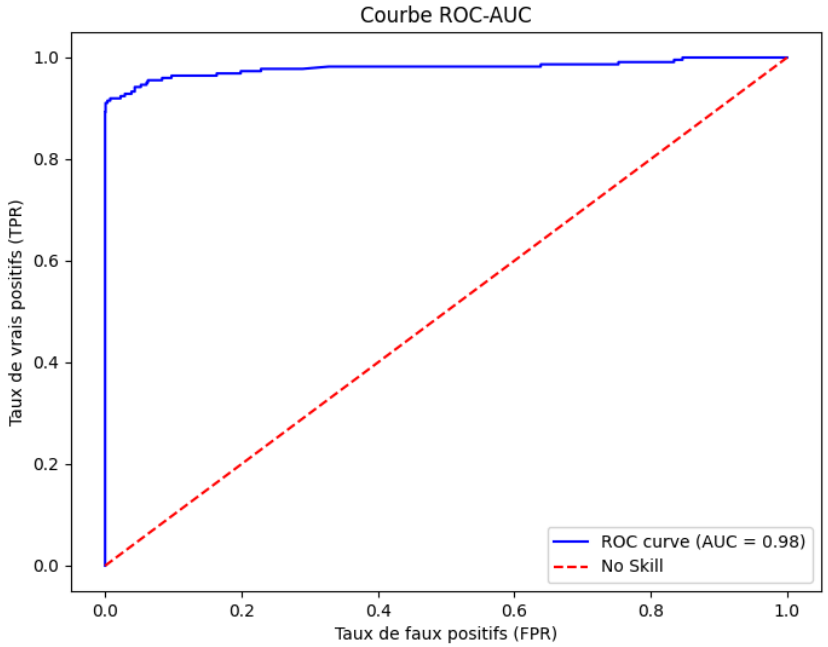
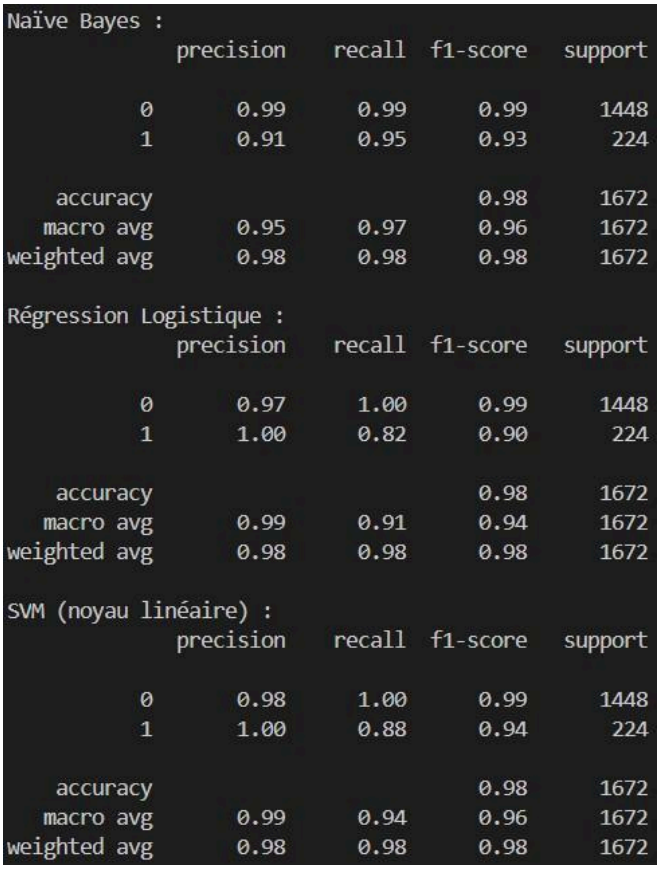


Figure 3 : Graphique de la courbe ROC-AUC du modèle SVM

Interprétation : Une AUC proche de 1 indique une bonne performance du modèle.



```
Naïve Bayes :
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1448
           1       0.91      0.95      0.93       224

    accuracy                           0.98      1672
   macro avg       0.95      0.97      0.96      1672
weighted avg       0.98      0.98      0.98      1672

Régression Logistique :
              precision    recall  f1-score   support

           0       0.97      1.00      0.99      1448
           1       1.00      0.82      0.90       224

    accuracy                           0.98      1672
   macro avg       0.99      0.91      0.94      1672
weighted avg       0.98      0.98      0.98      1672

SVM (noyau linéaire) :
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1448
           1       1.00      0.88      0.94       224

    accuracy                           0.98      1672
   macro avg       0.99      0.94      0.96      1672
weighted avg       0.98      0.98      0.98      1672
```

Figure 4 : Rapports de classification des modèles Naïve Bayes, Régression Logistique et SVM

Comparez leurs performances respectives.

1. Naïve Bayes :

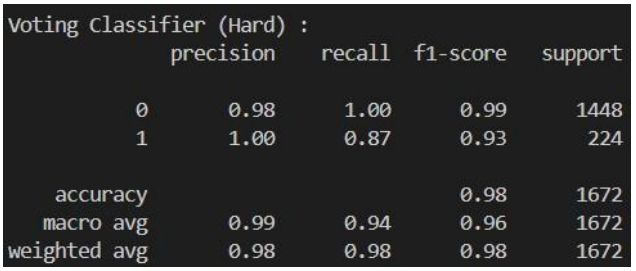
Naïve Bayes montre une bonne performance globale avec une précision pondérée de 98%. Cependant, il est légèrement moins performant en termes de rappel pour la classe des spams (95%), ce qui signifie qu'il rate encore quelques spams.

2. Régression Logistique :

La régression logistique est très précise pour prédire les spams (précision : 100%), mais elle a un rappel plus faible (82%). Cela signifie qu'elle classe certains spams comme des hams (faux négatifs). Le F1-score pour la classe des spams est légèrement inférieur à celui de Naïve Bayes.

3. SVM (noyau linéaire) :

Le modèle SVM a une performance équilibrée avec un rappel (88%) supérieur à celui de la régression logistique et une précision (100%) identique pour les spams. Le F1-score pour les spams (94%) est le plus élevé des trois modèles, ce qui est un choix judicieux pour cette tâche.



Voting Classifier (Hard) :

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1448
1	1.00	0.87	0.93	224
accuracy			0.98	1672
macro avg	0.99	0.94	0.96	1672
weighted avg	0.98	0.98	0.98	1672

Figure 5 : Rapport de classification du Voting Classifier avec un vote ‘hard’

1. Comparaison avec Naïve Bayes :

- Le Voting Classifier conserve une précision similaire (0.98), mais son rappel pour la classe spam diminue légèrement (0.87 contre 0.95 pour Naïve Bayes).

- Cependant, il améliore l'équilibre entre la précision et le rappel pour les spams, ce qui lui permet d'avoir un F1-score (0.93) similaire à celui de Naïve Bayes.

2. Comparaison avec Régression Logistique :

- Le Voting Classifier améliore nettement le rappel pour les spams (0.87 contre 0.82) tout en maintenant une précision élevée (1.00 pour la classe spam).

- Cela se traduit par un F1-score supérieur pour les spams (0.93 contre 0.90 pour la régression logistique).

3. Comparaison avec SVM :

- Les performances du Voting Classifier sont très proches de celles du SVM, avec une précision, un rappel, et un F1-score similaires.

- Cependant, en tant que modèle combiné, le Voting Classifier bénéficie d'une meilleure robustesse.

Voting Classifier (Soft) :

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1448
1	1.00	0.91	0.95	224
accuracy			0.99	1672
macro avg	0.99	0.95	0.97	1672
weighted avg	0.99	0.99	0.99	1672

Figure 6 : Rapport de classification du Voting Classifier avec un vote ‘soft’

- Rappel : Le vote ‘soft’ est plus performant, atteignant 0.91 contre 0.87 pour le vote ‘hard’. Cela signifie que le vote ‘soft’ détecte plus efficacement les spams, réduisant les faux négatifs.

- F1-Score : Le F1-score du vote ‘soft’ est également supérieur (0.95 contre 0.93), ce qui montre qu'il équilibre mieux précision et rappel.

- Accuracy : Le vote ‘soft’ atteint 0.99, légèrement supérieur au 0.98 du vote ‘hard’, ce qui confirme sa supériorité générale.

Donc, dans ce cas, le vote ‘soft’ est plus performant que le vote ‘hard'.

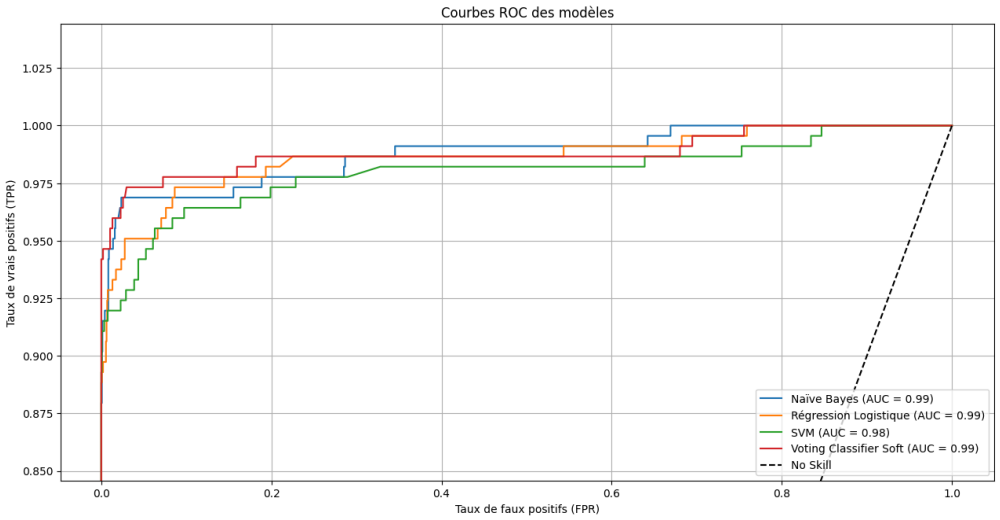


Figure 7 : Graphique des courbe ROC-AUC des modèles Naïve Bayes, Régression Logistique, SVM et le Voting Classifier

- Naïve Bayes (AUC = 0.99) : La courbe est légèrement inférieure aux meilleurs modèles par endroits, mais montre tout de même une bonne performance globale. Cela montre que Naïve Bayes est un bon classificateur pour ce dataset, bien que parfois il est dépassé par d'autres modèles.

- Régression Logistique (AUC = 0.99) : Ce modèle atteint presque la même performance que le Voting Classifier Soft. Sa courbe ROC est très proche de l'idéal, ce qui en fait l'un des meilleurs modèles.

- SVM (AUC = 0.98) : Bien que performant, sa courbe est légèrement en dessous des deux autres modèles individuels dans certaines zones, indiquant qu'il peut parfois manquer de flexibilité par rapport à la régression logistique.

- Voting Classifier (Soft) (AUC = 0.99) : Sa courbe se situe parmi les meilleures, avec une performance globale comparable à celle de la régression logistique.

En général, les trois modèles et le Voting Classifier Soft atteignent presque un AUC parfait (0.99 ou 0.98), ce qui démontre leur excellente capacité à différencier les spams des hams. Cependant, le Voting Classifier Soft est légèrement meilleur.

Pourquoi le mélange de modèles peut-il surpasser les performances des modèles individuels ?

Le mélange de modèles permet de surpasser les performances des modèles individuels en exploitant leur diversité, en réduisant à la fois le biais et la variance, et en atténuant les erreurs spécifiques grâce à l'agrégation de leurs prédictions.

Expliquez le principe des Gaussian Mixture Models (GMM) et leur rôle dans la classification.

Les Gaussian Mixture Models (GMM) sont des modèles probabilistes qui supposent que les données proviennent d'une combinaison de plusieurs distributions gaussiennes. Chaque composant de la combinaison correspond à une distribution gaussienne, et le modèle est entraîné pour estimer les paramètres (moyenne, covariance) de chaque composant.

Dans le cadre de la classification, chaque composant du modèle GMM peut être associé à une classe différente. Le GMM est utilisé ici pour modéliser les distributions des données dans chaque classe (spam et ham).

GMM :

	precision	recall	f1-score	support
0	0.87	1.00	0.93	1448
1	0.00	0.00	0.00	224
accuracy			0.87	1672
macro avg	0.43	0.50	0.46	1672
weighted avg	0.75	0.87	0.80	1672

Figure 8 : Rapport de classification du GMM

- Précision : La précision globale est de 0.87, ce qui semble bon, mais cela est principalement dû à la performance élevée pour la classe ham (prédictions correctes pour la classe majoritaire).

- Rappel : Le rappel de 1.00 pour ham signifie que tous les messages ham sont correctement classifiés, mais le modèle ne parvient pas à détecter les messages spam (rappel de 0.00 pour spam).

- F1-score : Le F1-score global de 0.80 est également élevé, mais cela reflète davantage la bonne performance pour ham, sans prendre en compte les échecs pour spam.

Donc, bien que la précision globale semble acceptable, le modèle échoue à identifier les messages spam, ce qui en fait une solution non fiable pour cette tâche de classification.

SVM:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1448
1	1.00	0.88	0.94	224
accuracy			0.98	1672
macro avg	0.99	0.94	0.96	1672
weighted avg	0.98	0.98	0.98	1672

Voting Classifier (Hard) :

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1448
1	1.00	0.87	0.93	224
accuracy			0.98	1672
macro avg	0.99	0.94	0.96	1672
weighted avg	0.98	0.98	0.98	1672

Voting Classifier (Soft) :

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1448
1	1.00	0.91	0.95	224
accuracy			0.99	1672
macro avg	0.99	0.95	0.97	1672
weighted avg	0.99	0.99	0.99	1672

GMM :

	precision	recall	f1-score	support
0	0.86	0.93	0.89	1448
1	0.01	0.00	0.01	224
accuracy			0.80	1672
macro avg	0.43	0.47	0.45	1672
weighted avg	0.74	0.80	0.77	1672

Figure 9 : Rapport de classification du SVM, Voting Classifier Hard, Voting Classifier Soft et GMM

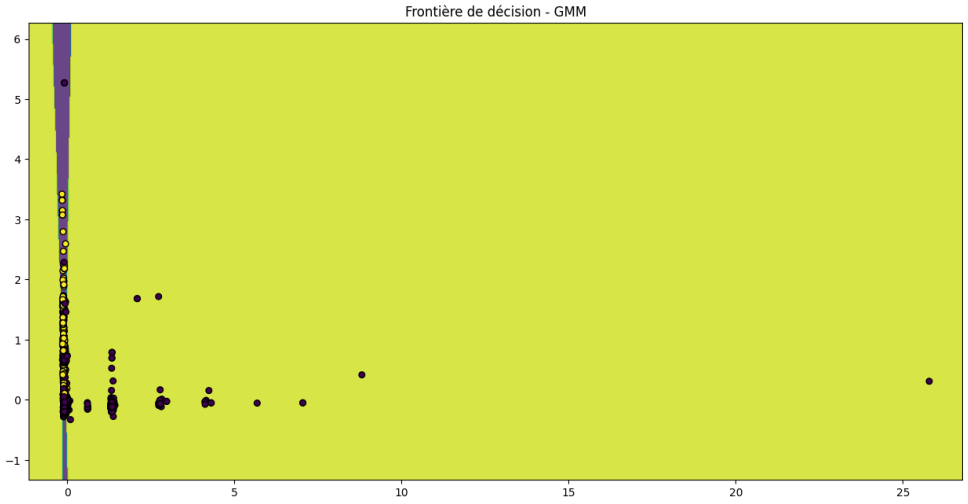


Figure 10 : Rapport de classification du GMM

La SVM crée une frontière linéaire qui sépare les classes de manière optimale, mais qui peut être trop rigide pour des données complexes.

La GMM peut générer des frontières non linéaires, car il modélise chaque classe comme une distribution gaussienne, ce qui lui permet de s'adapter plus facilement à des distributions complexes dans les données.

Ainsi, les frontières du GMM seront souvent plus flexibles et adaptées aux distributions des données, alors que celles de la SVM seront plus rigides et linéaires.

Quels sont les hyperparamètres principaux d’un SVM que vous pouvez optimiser ?

Les hyperparamètres principaux d'un SVM qu'on peut optimiser sont :
- C
- Kernel
- coef0
- gamma
- degree

```
Meilleurs paramètres SVM :
{'C': 0.1, 'kernel': 'linear'}
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1448
           1       1.00      0.87      0.93       224

    accuracy                           0.98      1672
   macro avg       0.99      0.94      0.96      1672
weighted avg       0.98      0.98      0.98      1672
```

Figure 11 : Les meilleurs paramètres trouvés avec une recherche par GridSearchCV et rapport de classification du SVM

```
Performance du Voting Classifier avec poids :
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1448
           1       1.00      0.91      0.95       224

    accuracy                           0.99      1672
   macro avg       0.99      0.96      0.97      1672
weighted avg       0.99      0.99      0.99      1672
```

Figure 12 : Rapport de classification de la performance du Voting Classifier avec les poids

```
Meilleurs paramètres pour le GMM :
{'covariance_type': 'spherical', 'n_components': 2}
Performances du modèle GMM optimisé :
              precision    recall  f1-score   support

           0       0.98      0.65      0.78      1448
           1       0.29      0.93      0.45       224

    accuracy                           0.69      1672
   macro avg       0.64      0.79      0.62      1672
weighted avg       0.89      0.69      0.74      1672
```

Figure 13 : Les meilleurs paramètres trouvés pour le GMM et rapport de classification du GMM optimisé