Faculté Des Sciences Economiques Et De Gestion De Nabeul

Master de recherche en Business Computing

Thème

**Exploration des Approches de Machine Learning pour la Classification Efficace des Emails**

Présenté Par :

Meriem Slama

Roudayna Bougatef

Responsable :

Dr. Fadoua BOUAFIF

Année Universitaire 2024 – 2025

Année Universitaire 2024 – 2025

Table des matieres

[Introduction 3](#_Toc164716064)

[I. Etat de l’art 4](#_Toc164716065)

[II. Analyse des méthodes choisis : 5](#_Toc164716066)

[1. Arbres de Décision : 5](#_Toc164716067)

[2. Réseaux Neuronaux Convolutionnels (CNN): 8](#_Toc164716068)

[3. K-means : 9](#_Toc164716069)

[4. Forêt d'Isolation : 11](#_Toc164716070)

[Conclusion 13](#_Toc164716071)

[Table de références 14](#_Toc164716072)

[Table des tableaux 15](#_Toc164716073)

[Table des figures 15](#_Toc164716074)

Introduction

La classification des emails est une tâche cruciale dans de nombreuses applications liées à la communication électronique, telles que la gestion des boîtes de réception, la détection de spams et la sécurité des données.

L'objectif principal de cette problématique est de développer des méthodes efficaces pour trier automatiquement les emails en différentes catégories, telles que les emails légitimes, les spams. Cette classification repose sur l'analyse et l'extraction de diverses caractéristiques des emails, telles que le contenu textuel, les en-têtes, les pièces jointes, et d'autres méta-données pertinentes.

L'enjeu majeur réside dans la capacité à discriminer avec précision entre les différents types d'emails, en minimisant les faux positifs (classification erronée d'un email légitime comme spam) et les faux négatifs (classification erronée d'un spam comme email légitime).

Pour relever ce défi, diverses approches d'apprentissage automatique, y compris les arbres de décision, les réseaux de neurones convolutionnels (CNN), les forêts aléatoires et les méthodes de clustering telles que K-Means, ont été largement explorées et utilisées avec succès. Ces approches offrent des techniques puissantes pour traiter cette problématique complexe et contribuent à améliorer l'efficacité et la sécurité des systèmes de gestion des emails.

# Etat de l’art

La classification des emails, une composante fondamentale de la gestion des communications électroniques, a fait l'objet de recherches approfondies dans la littérature académique et industrielle. Cette section présente un état de l'art des différentes méthodes et approches utilisées pour classifier les emails en fonction de leur contenu et de leurs caractéristiques.

L'article **[1]** examine la nécessité d'améliorer la classification de texte basée sur l'apprentissage automatique, malgré sa diversité d'applications. La littérature examinée met en lumière l'utilisation de divers modèles notamment les machines à vecteurs de support (SVM) **,** Naïve Bayes ,le KNN , l’arbre de decision , foret aleatoire , regression logistique ,les reseaux nouronaux (CNN/RNN)

La classification des e-mails à l'aide de l'algorithme K-means a été explorée dans divers articles de recherche. Mhaske-Dhamdhere et al. ont proposé une approche de classification des e-mails de phishing en temps réel utilisant l'algorithme K-means, permettant d'obtenir des taux de vrais positifs élevés pour les e-mails légitimes et de phishing **[2]**Sel et Hanbay ont également utilisé l'algorithme K-means pour la classification non supervisée des e-mails reçus sur la base des données traitées par Word2Vec, démontrant ainsi un regroupement réussi des e-mails avec un taux de réussite élevé **[3]**. En outre, Jáñez-Martino et al. ont introduit un ensemble de données multiclasses pour la classification des courriers indésirables, intégrant un clustering hiérarchique et des K-means pour améliorer la gestion des spams, démontrant ainsi l'efficacité de cette approche pour améliorer la précision de la classification des e-mails **[4].**

Tableau 1 -les approches de machine Learning des classifications des emails

|  |  |  |  |
| --- | --- | --- | --- |
| Approche | Caractéristiques | Avantages | Limitations |
| SVM | Efficace pour gérer des données de grande dimension et traiter des problèmes de classification linéaires et non linéaires. | Haute précision, adapté à la classification multi-classe. | Difficulté de choisir le noyau correct, temps d'entraînement long, occupation importante de la mémoire. |
| Naïve Bayes | Algorithme rapide et simple, performances acceptables avec de petits ensembles de données, adapté à la classification multi-classe. | Facile à implémenter. | Inadapté aux petits ensembles de données, performances médiocres lorsque les caractéristiques ne sont pas indépendantes. |
| kNN | Basé sur les voisins les plus proches, simplicité d'implémentation. | Pas de phase d'entraînement, adapté à diverses tâches. | Requiert plus d'espace de stockage, sensible aux erreurs. |
| Arbre de Décision | Construit une structure en forme d'arbre avec des branches pour classifier les échantillons. | Peut traiter des données numériques et catégoriques. | Instable, sensible aux variations mineures des données d'entraînement. |
| Random Forest (RF) | Ensemble d'arbres de décision, résultats finaux décidés par vote majoritaire. | Gestion efficace de milliers de caractéristiques, performances élevées avec données non linéaires. | Risque de surajustement. |
| Régression Logistique | Méthode statistique pour la classification binaire, extensible aux classifications multiclasses. | Bonne performance sur de petits ensembles de données. | Incapacité à classifier des variables continues. |
| Réseaux Neuronaux (CNN, RNN) | Captent des modèles complexes dans les données textuelles, adaptés à l'analyse des sentiments et à la catégorisation. | Capacité à traiter des données non linéaires. | Complexité de mise en œuvre, nécessité de grandes quantités de données pour l'entraînement. |
| K-means | Algorithme pour la classification non supervisée, utilisé pour le clustering des e-mails. | Efficace pour regrouper des e-mails similaires. | Sensible à la présence de valeurs aberrantes, nécessite de spécifier le nombre de clusters à l'avance. |

* Les auteurs notent que le SVM est le modèle le plus utilisé dans la littérature, suivi de Naïve Bayes et kNN. L'exactitude maximale de 98,88 a été obtenue par SVM dans l'ensemble de données 20Newsgroup. En outre, la classification des emails à l'aide de l'algorithme K-means a été explorée dans plusieurs études, démontrant son efficacité dans la détection de phishing et la classification non supervisée des emails.

# Analyse des méthodes choisis :

Pour la classification des emails, nous avons choisi d'utiliser les méthodes suivantes : les arbres de décision, les réseaux neuronaux convolutionnels (CNN), l'algorithme K-means et la forêt d'isolation. Voici une analyse détaillée de chaque méthode :

1. Arbres de Décision :

L'arbre de décision est un algorithme d'apprentissage supervisé largement utilisé pour la classification. Il fonctionne en créant une structure en forme d'arbre où chaque nœud interne représente une caractéristique, chaque branche représente une décision basée sur cette caractéristique, et chaque feuille représente une classe. Pour la classification des emails, les arbres de décision peuvent être efficaces car ils peuvent gérer à la fois des données catégorielles et numériques, ce qui est courant dans les données textuelles. Avec le jeu de données SpamAssassin, les arbres de décision peuvent être efficaces pour identifier les caractéristiques les plus discriminantes des emails spam et non-spam.

**Interprétation :**

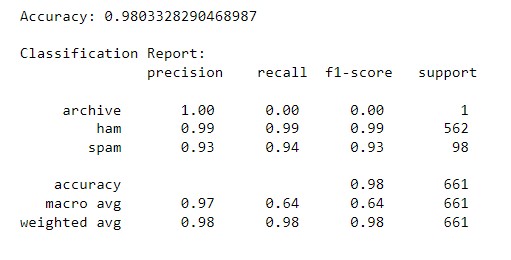


Figure 1 - résultat de l'approche de l'arbre de décision

Les résultats proviennent de l'évaluation d'un modèle d'arbre de décision pour la classification des e-mails. :

* L'exactitude globale du modèle est de 98.03%, avec une précision de 100% pour la classe "archive", 99% pour la classe "ham" (e-mails légitimes) et 93% pour la classe "spam" (e-mails indésirables).
* Le rappel est de 0% pour la classe "archive", 99% pour la classe "ham" et 94% pour la classe "spam".
* Le score F1, une moyenne harmonique de la précision et du rappel, n'est pas calculé pour la classe "archive".
* En conclusion, le modèle présente une performance élevée pour les classes "ham" et "spam", mais pourrait être amélioré pour détecter la classe "archive" et augmenter le rappel pour la classe "spam".

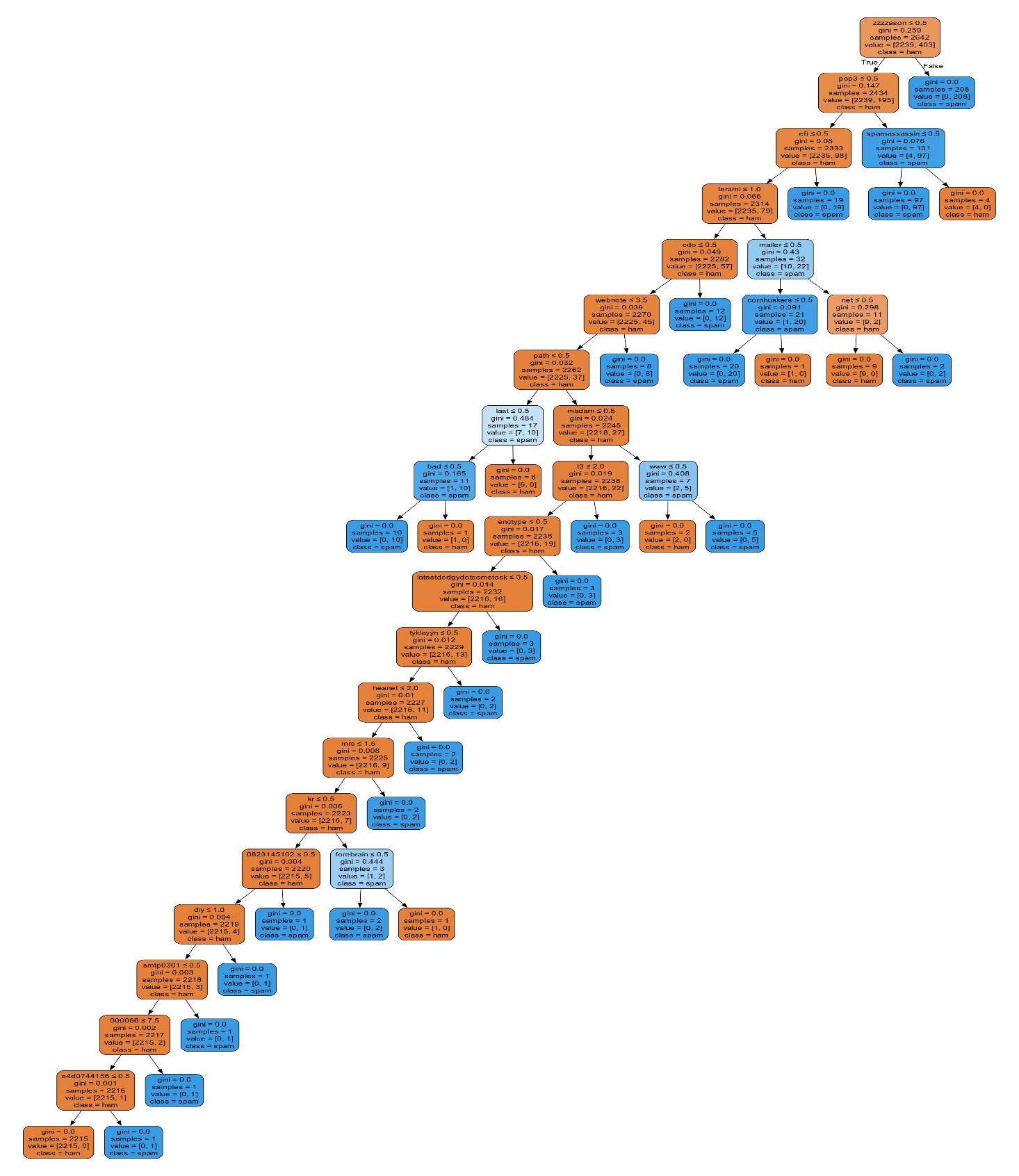


Figure 2- l'arbre de décision de modèle

1. Réseaux Neuronaux Convolutionnels (CNN):

- Les réseaux neuronaux convolutifs sont une classe de réseaux neuronaux profonds appliqués avec succès à des tâches de traitement de texte, y compris la classification des emails. Les CNN peuvent capturer des motifs complexes dans les données textuelles grâce à leurs couches de convolution, de pooling et de couches entièrement connectées. Ils sont capables d'apprendre des représentations de niveau élevé à partir des données brutes, ce qui en fait un choix prometteur pour la classification des emails. Avec SpamAssassin, les CNN pourraient être utilisés pour extraire des caractéristiques significatives des emails et améliorer les performances de classification.

**Interprétation :**

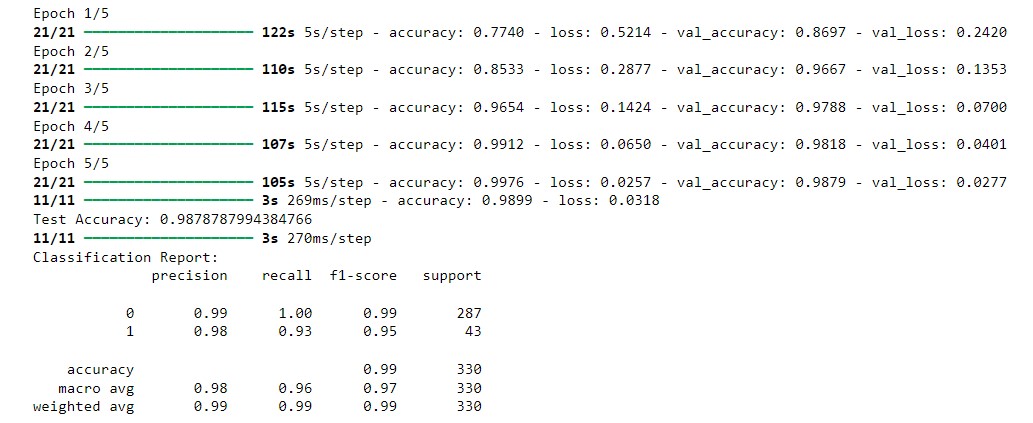


Figure 3- résultat de l'approche de reseaux de neurones convolutionneles CNN

* L'entraînement du modèle montre une amélioration progressive de la précision et de la perte (loss) sur les cinq epochs, avec une précision finale de 99.76% sur les données d'entraînement et une précision de validation de 98.79% sur l'ensemble de test.
* Le rapport de classification indique une bonne performance avec des scores élevés de précision, rappel et f1-score pour les deux classes, avec une légère baisse de rappel pour la classe 1.

1. K-means :

- K-means est un algorithme de regroupement non supervisé largement utilisé pour la segmentation des données. Dans le contexte de la classification des emails, K-means peut être utilisé pour regrouper les emails en clusters similaires en fonction de leurs caractéristiques. Cela peut être utile pour la détection d'anomalies, la segmentation de la boîte de réception ou même la classification des emails en fonction de leur contenu ou de leur expéditeur

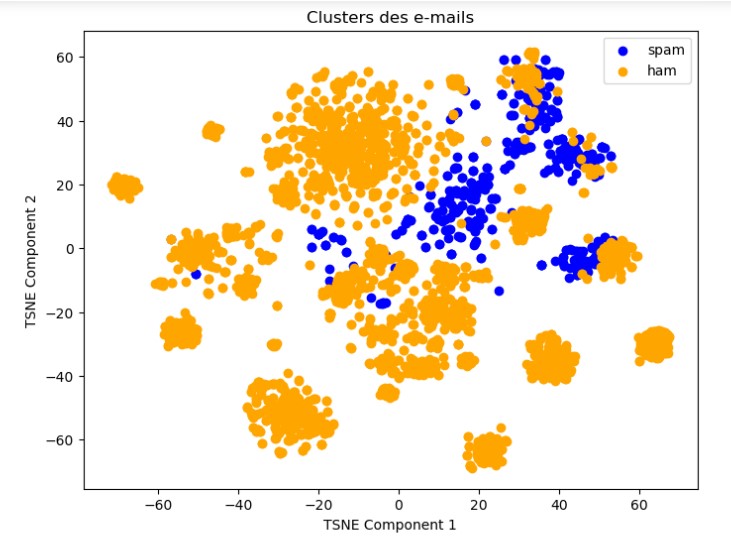
**Interprétation :**

Figure 4- les clusters des emails

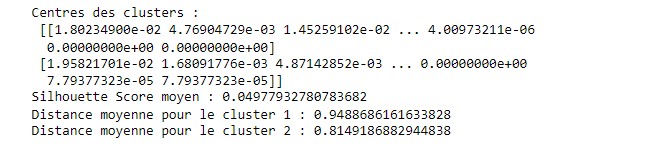


Figure 5 - résultat de l'approche de k-means

* Les "Centres des clusters" sont les vecteurs représentatifs de chaque cluster obtenu après l'application de l'algorithme de clustering. Chaque vecteur contient les valeurs des différentes caractéristiques pour chaque cluster.
* Le "Silhouette Score moyen" est une mesure de la cohésion et de la séparation des clusters. Un score proche de 1 indique une bonne séparation entre les clusters, tandis qu'un score proche de -1 indique un mauvais regroupement.
* Les "Distances moyennes pour le cluster 1" et "cluster 2" représentent la distance moyenne de chaque point dans chaque cluster par rapport au centre de ce cluster. Ces mesures permettent d'évaluer à quel point les points d'un cluster sont proches ou éloignés de leur centre respectif.

1. Forêt d'Isolation :

- La forêt d'isolation est un algorithme d'apprentissage non supervisé utilisé pour la détection d'anomalies. Dans le contexte de la classification des emails, la forêt d'isolation peut être utilisée pour identifier des emails anormaux ou suspects qui pourraient être des spam ou des emails malveillants. En isolant les anomalies dans le dataset, la forêt d'isolation peut aider à améliorer la précision globale du modèle de classification des emails.

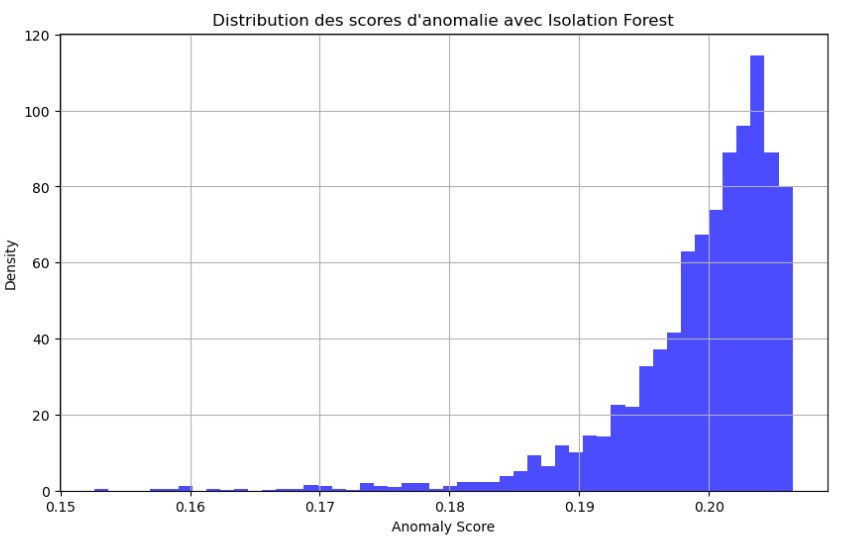
**Interprétation :**

Figure 6 - distribution des scores d'anomalie avec isolation forest

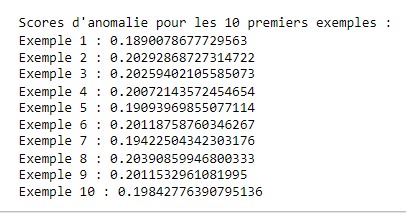


Figure 7 - résultat de l'approche de foret d'isolation

Ces valeurs représentent les scores d'anomalie attribués par le modèle Isolation Forest aux 10 premiers exemples de données. Un score d'anomalie plus élevé indique une probabilité plus élevée que l'exemple soit une anomalie ou un outlier dans le jeu de données.

* Les exemples 1, 2, 3, 4, 6, 8 et 9 ont des scores d'anomalie relativement élevés, indiquant qu'ils sont plus susceptibles d'être des anomalies.
* Les exemples 5, 7 et 10 ont des scores d'anomalie légèrement moins élevés, mais toujours significatifs, ce qui pourrait également indiquer une certaine anomalie potentielle, bien que moins prononcée que les autres.

Conclusion

En choisissant ces méthodes pour votre projet de classification des emails, vous exploitez une variété d'approches qui offrent des avantages et des inconvénients uniques. L'arbre de décision fournit une méthode simple et interprétable, les CNN offrent une puissance de modèle pour capturer des modèles complexes, K-means permet la segmentation et la détection d'anomalies, tandis que la forêt d'isolation se concentre spécifiquement sur la détection d'anomalies. En les combinant de manière appropriée et en les adaptant à votre ensemble de données SpamAssassin, vous pouvez construire un système robuste de classification des emails.

****Table de références****

|  |  |
| --- | --- |
| [1] | C. Z. E.-B. R. D. Ashokkumar Palanivinayagam, «Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review,» *Algorithms,* p. 28, 2023. |
| [2] | S. V. Vidya Mhaske-Dhamdhere, «A novel approach for phishing emails real time classifica-tion using k-means algorithm,» *International journal of engineering and technology,* p. 5, 2017. |
| [3] | E. F. ,. S. G.-M. ,. ı. V.-M. Francisco Ja´nez-Martino, «Classification of Spam Emails through Hierarchical,» *arXiv: Computation and Language,* p. 5, 2020. |
| [4] | D. H. Slhami Sel, «E-Mail Classification Using Natural Language Processing,» 2019. |

Table des tableaux

[Tableau 1 -les approches de machine Learning des classifications des emails 4](#_Toc164713995)

Table des figures

[Figure 1 - résultat de l'approche de l'arbre de décision 6](#_Toc164715240)

[Figure 2- l'arbre de décision de modèle 7](#_Toc164715241)

[Figure 3- résultat de l'approche de reseaux de neurones convolutionneles CNN 8](file:///C:\Users\pc\Desktop\mastere\semester%202\projets\Classification%20des%20emails.docx#_Toc164715242)

[Figure 4- les clusters des emails 9](file:///C:\Users\pc\Desktop\mastere\semester%202\projets\Classification%20des%20emails.docx#_Toc164715243)

[Figure 5 - résultat de l'approche de k-means 10](#_Toc164715244)

[Figure 6 - distribution des scores d'anomalie avec isolation forest 11](file:///C:\Users\pc\Desktop\mastere\semester%202\projets\Classification%20des%20emails.docx#_Toc164715245)

[Figure 7 - résultat de l'approche de foret d'isolation 12](file:///C:\Users\pc\Desktop\mastere\semester%202\projets\Classification%20des%20emails.docx#_Toc164715246)