École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

ENSIAS - RABAT

Rapport de projet

Détection des Annonces d'Emploi Frauduleuses : Approche par le Traitement du Langage Naturel et Machine Learning

Réalisé par :

MERZAQ Karim
KHATTABI Oussama

Encadré par :

Pr Y. TABII

Année universitaire : 2024-2025

Table des matières

Table des figures

Introduction générale

Avec l'essor des plateformes de recrutement en ligne, la détection d'annonces d'emploi frauduleuses est devenue essentielle pour protéger les chercheurs d'emploi. Ce projet vise à développer un modèle de machine learning capable d'identifier ces fraudes en s'appuyant sur l'analyse de données textuelles.

À partir d'un jeu de données Kaggle, nous appliquons des techniques de traitement du langage naturel (NLP) pour préparer et structurer les informations textuelles. Ensuite, des modèles de classification, comme la régression logistique et les arbres de décision, sont utilisés pour identifier les annonces suspectes. Ce projet met en avant les principales étapes et défis de la détection de fraude dans les annonces d'emploi en ligne.

Chapitre 1

Introduction

Le premier chapitre pose les bases du projet en introduisant le contexte et les objectifs. Pour un projet de détection de fausses annonces d'emploi, il est crucial de bien comprendre pourquoi ce problème est important et de définir précisément les objectifs que l'on souhaite atteindre.

1.1 Contexte du projet

Avec la montée en puissance des plateformes d'emploi en ligne, les fraudeurs exploitent de plus en plus ces plateformes pour diffuser de fausses offres d'emploi. Ces fausses annonces visent souvent à collecter des informations personnelles sensibles, à soutirer de l'argent aux candidats ou à attirer des personnes vers des services payants frauduleux. Les conséquences peuvent être graves pour les individus, touchant aussi bien leur vie privée que leurs finances.

Dans ce contexte, la détection automatisée de fausses annonces d'emploi devient une nécessité pour les plateformes d'emploi et les chercheurs d'emploi. Grâce aux avancées en intelligence artificielle et en traitement du langage naturel, il est désormais possible de concevoir des modèles capables de détecter des motifs de fraude dans les descriptions d'annonces.

1.2 Problèmatique

Les fausses annonces d'emploi peuvent être difficiles à détecter manuellement en raison de leur grand nombre et de leur sophistication croissante. Les escrocs développent des techniques de plus en plus élaborées pour rendre leurs annonces crédibles, en utilisant un langage proche de celui des annonces légitimes et en copiant parfois des éléments de véritables offres d'emploi.

La problématique de ce projet est donc : Comment concevoir un modèle efficace capable d'identifier automatiquement les fausses annonces d'emploi en ligne ? Ce modèle doit être capable de distinguer les caractéristiques spécifiques aux fausses annonces par rapport aux vraies annonces, malgré leur similarité apparente.

1.3 Objectifs du projet

Les objectifs principaux du projet sont :

- **Développer un modèle de détection :** Mettre en place un modèle de machine learning ou de deep learning capable d'analyser les annonces d'emploi et de détecter celles qui sont probablement frauduleuses.
- **Améliorer la précision de détection :** Optimiser le modèle pour qu'il puisse détecter les fausses annonces avec un haut niveau de précision, en minimisant les faux positifs et les faux négatifs.
- **Analyser les caractéristiques des fausses annonces :** Comprendre les attributs ou motifs spécifiques qui caractérisent les fausses annonces, afin d'améliorer la fiabilité du modèle.
- **Évaluer les performances du modèle :** Mettre en place un protocole d'évaluation qui permettra de mesurer l'efficacité du modèle, en utilisant des métriques de performance pertinentes.

1.4 Conclusion

Ce premier chapitre établit ainsi le cadre général du projet, en présentant le contexte, les défis, et les objectifs, et en donnant une vue d'ensemble de la structure du rapport.

Cette introduction permet au lecteur de comprendre l'importance du projet et de se familiariser avec la problématique à laquelle il répond.

Chapitre 2

Préparation des données

2.1 Introduction

Ce chapitre détaille la préparation des données pour l'analyse et la classification des annonces d'emploi. Nous commençons par présenter la source des données et les variables principales, suivies des étapes de prétraitement pour nettoyer et structurer les informations textuelles. Cette préparation permet de réduire le bruit dans les données et d'améliorer la qualité de l'analyse et de la classification des annonces frauduleuses.

2.2 Source de données

Le projet s'appuie sur un jeu de données public accessible sur Kaggle, une plateforme reconnue pour ses ressources en machine learning et en données. Ce jeu de données contient des informations détaillant des offres d'emploi, chaque offre étant identifiée comme réelle ou frauduleuse. Comme dans le monde réel, les annonces frauduleuses représentent une petite fraction de l'ensemble, ce qui correspond aux attentes.

2.2.1 Description du dataset

Notre jeu de données inclut un total de 17 880 enregistrements et comporte 18 variables descriptives.

	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	0	1	0	Other	Internship	NaN	NaN	Marketing	0
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	1	0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service	0
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	NaN	0	1	0	NaN	NaN	NaN	NaN	NaN	0
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	1	0	Full-time	Mid-Senior level	Bachelor's Degree	Computer Software	Sales	0
4	5	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care	Health Care Provider	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17875	17876	Account Director - Distribution	CA, ON, Toronto	Sales	NaN	Vend is looking for some awesome new talent to...	Just in case this is the first time you've vis...	To ace this role you:Will eat comprehensive St...	What can you expect from us? We have an open cu...	0	1	1	Full-time	Mid-Senior level	NaN	Computer Software	Sales	0
17876	17877	Payroll Accountant	US, PA, Philadelphia	Accounting	NaN	WebLinc is the e-commerce platform and service...	The Payroll Accountant will focus primarily on...	- B.A. or B.S. in Accounting- Desire to have l...	Health & WellnessMedical planPrescription ...	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Internet	Accounting/Auditing	0
17877	17878	Project Cost Control Staff Engineer - Cost Con...	US, TX, Houston	NaN	NaN	We Provide Full Time Permanent Positions for m...	Experienced Project Cost Control Staff Enginee...	At least 12 years professional experience.Abil...	NaN	0	0	0	Full-time	NaN	NaN	NaN	NaN	0
17878	17879	Graphic Designer	NG, LA, Lagos	NaN	NaN	NaN	Nemsia Studios is looking for an experienced v...	1. Must be fluent in the latest versions of Co...	Competitive salary (compensation will be based...	0	0	1	Contract	Not Applicable	Professional	Graphic Design	Design	0
17879	17880	Web Application Developers	NZ, N, Wellington	Engineering	NaN	Vend is looking for some awesome new talent to...	Who are we?Vend is an award winning web based ...	We want to hear from you if:You have an in-dep...	NaN	0	1	1	Full-time	Mid-Senior level	NaN	Computer Software	Engineering	0

17880 rows × 18 columns

FIGURE 2.1 – Dataset initiale

2.2.2 Description des variables

Notre jeu de données contient les variables suivantes :

- **job_id** : Identifiant unique pour chaque annonce d'emploi. Utilisé pour identifier individuellement chaque observation.
- **title** : Titre de l'offre d'emploi. Indique le poste proposé et peut fournir des indices sur le domaine ou la spécialité.
- **location** : Localisation géographique de l'emploi. Cette information peut aider à identifier des patterns régionaux ou localisés.
- **department** : Département de l'entreprise dans lequel le poste est proposé. Utile pour repérer des tendances dans des services spécifiques.
- **salary_range** : Échelle de salaire proposée pour le poste. Peut être utile pour identifier des fraudes, car des valeurs incohérentes pourraient être un signal.
- **company_profile** : Description de l'entreprise. Fournit des informations contextuelles qui peuvent aider à vérifier la légitimité d'une annonce.
- **description** : Description complète de l'offre d'emploi, souvent la plus riche en texte. Elle peut contenir des mots-clés importants pour classifier l'offre.
- **requirements** : Liste des exigences du poste. Peut inclure des compétences spécifiques, formations, et niveaux d'expérience, souvent cruciaux pour la détection de fraude.
- **benefits** : Avantages offerts aux employés pour le poste. Informations additionnelles qui peuvent renforcer l'analyse de la crédibilité.

— **telecommuting** : Indicateur binaire (0 ou 1) pour le télétravail, indiquant si le poste permet le travail à distance. Utile pour analyser les types de postes en fonction de leur flexibilité.

— **has_company_logo** : Indicateur binaire indiquant la présence d'un logo de l'entreprise dans l'annonce. Un logo peut être un gage de légitimité.

— **has_questions** : Indicateur binaire indiquant si des questions sont posées dans l'annonce, ce qui pourrait refléter une meilleure sélection des candidats.

— **employment_type** : Type d'emploi (temps plein, temps partiel, stage, etc.). Peut aider à segmenter les types d'emplois proposés.

— **required_experience** : Niveau d'expérience requis pour le poste. La présence ou l'absence de cette variable pourrait être significative pour identifier les fraudes.

— **required_education** : Niveau d'éducation requis pour le poste. Permet d'analyser les exigences du poste, qui peuvent être inconsistantes dans les annonces frauduleuses.

— **industry** : Secteur d'activité de l'entreprise. Cela peut fournir un contexte pertinent pour la classification des offres par domaine d'activité.

— **function** : Fonction ou rôle pour le poste. Aide à préciser le domaine d'intervention de l'offre.

— **fraudulent** : Étiquette cible indiquant si une annonce est frauduleuse ou non. Cette variable est essentielle pour la tâche de classification, car elle indique l'objectif de l'analyse.

```
job_id                   0
title                    0
location               346
department           11547
salary_range         15012
company_profile       3308
description              1
requirements          2696
benefits              7212
telecommuting            0
has_company_logo         0
has_questions            0
employment_type       3471
required_experience   7050
required_education    8105
industry              4903
function              6455
fraudulent               0
dtype: int64
```

FIGURE 2.2 – Nombre de valeur nulle par chaque caractèristique

2.3 Visualisation et exploration du dataset

Notre jeu de données couvre des enregistrements provenant de 83 pays différents. Parmi l'ensemble, 17 014 offres sont marquées comme réelles, tandis que 866 sont étiquetées comme frauduleuses.

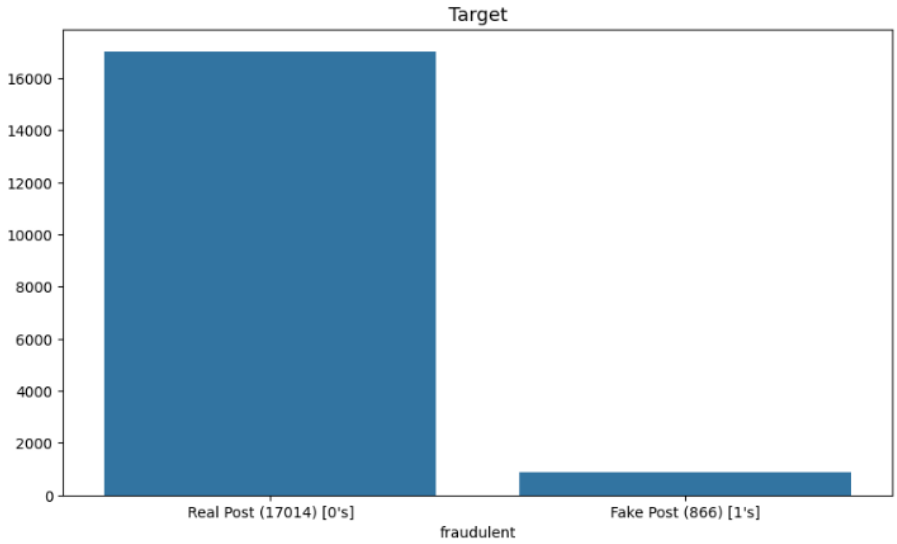

FIGURE 2.3 – Target count

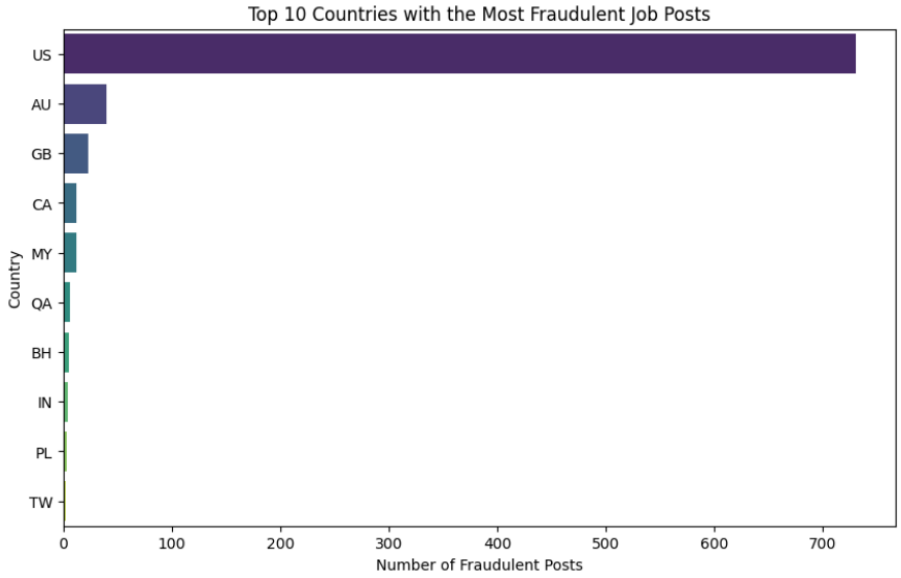

FIGURE 2.4 – Top 10 countries with the most fraudulent job posts

La figure 2.4 montre les dix pays avec le plus grand nombre d'offres d'emploi frauduleuses. Les États-Unis dominent avec un nombre élevé de cas, suivis à distance par l'Australie et le Royaume-Uni. Les autres pays affichent des chiffres nettement inférieurs, indiquant une forte concentration des fraudes aux États-Unis.

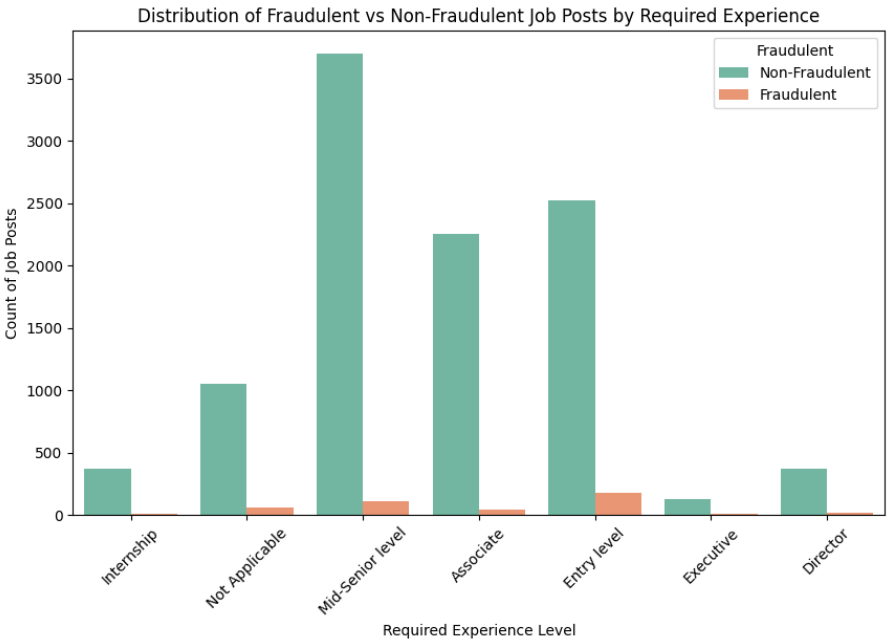

FIGURE 2.5 – Distribution of fraudulent vs non-fraudulent job

La figure 2.5 montre la distribution des offres d'emploi frauduleuses et non frauduleuses par niveau d'expérience requis. Les annonces non frauduleuses sont majoritairement pour des niveaux "Mid-Senior", "Associate", et "Entry", tandis que les annonces frauduleuses apparaissent à des niveaux d'expérience similaires mais en bien moindre nombre.

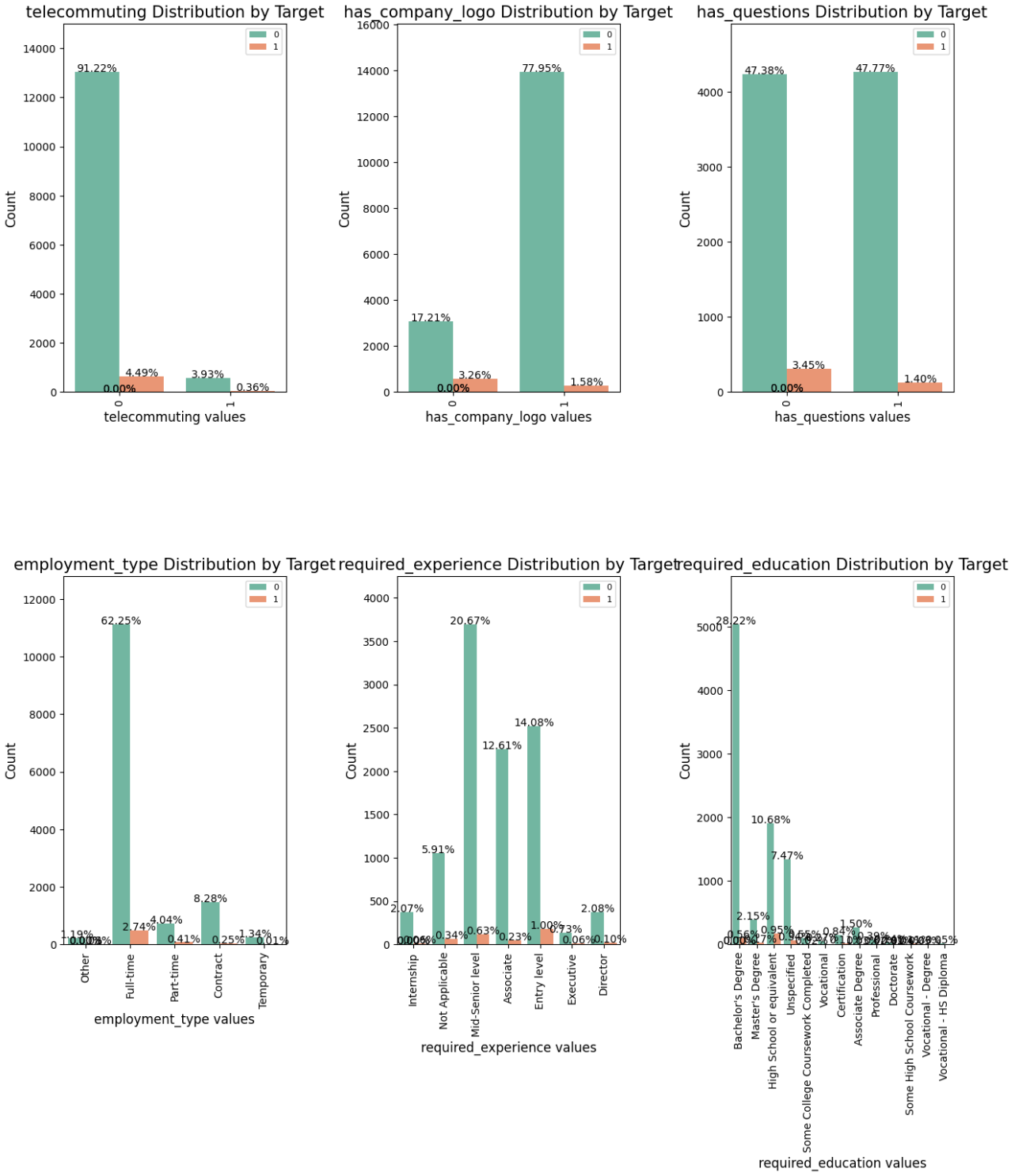

FIGURE 2.6 – Features description

La figure 2.6 montre la répartition des annonces selon des attributs comme le télétravail, le logo, les questions, le type d'emploi, l'expérience, et l'éducation, en distinguant les annonces frauduleuses et non frauduleuses. La plupart des annonces sont similaires sur ces attributs, avec une légère variation dans l'expérience et l'éducation requises.

2.3.1 Prétraitements

Parmi les 18 fonctionnalités de l'ensemble de données, toutes les colonnes, à l'exception de l'ID de travail, contiennent des valeurs textuelles ou binaires (Vrai ou Faux). La colonne ID de travail est une valeur entière, mais elle n'a pas de pertinence significative pour ce contexte de classification et est donc supprimée. Le prétraitement des données textuelles est une étape essentielle avant l'analyse, car des champs comme le titre, le profil de l'entreprise, la description, les exigences et les avantages contiennent du texte brut qui nécessite un nettoyage approfondi pour réduire le bruit.

Pour simplifier l'analyse, plusieurs champs textuels sont combinés en une seule colonne, englobant le titre, le $profil_entreprise, ladescription, lesexigences, lesavantages, laformation_requise, etl'in$ $suppressiondesmentions, deshashtags, desliens, etdelaponctuation.Enfin, lesvaleursnonnumriquessont$

2.3.1.1 Combinaison des Colonnes Textuelles

Les colonnes textuelles telles que `title`, `company_profile`, `description`, `requirements`, `benefits`, et `required_education` sont combinées en une seule colonne nommée `text`. Cela centralise les informations textuelles pour simplifier l'analyse et le traitement ultérieurs.

2.3.1.2 Suppression des Colonnes Inutiles

Certaines colonnes qui ne sont pas pertinentes pour l'analyse, comme `job_id`, `department`, `salary_range`, et `telecommuting`, sont supprimées. Cela permet de réduire la dimensionnalité des données et de conserver uniquement les informations essentielles pour l'analyse.

2.3.1.3 Nettoyage des Caractères Spéciaux

Le texte est nettoyé pour supprimer les caractères spéciaux (virgules, points-virgules, ponctuation, etc.) à l'aide de fonctions de substitution. Cela réduit le bruit dans les données textuelles et facilite leur traitement par les modèles de machine learning.

2.3.1.4 Suppression des Mentions et Hyperliens

Les mentions de type `@nom` et les hashtags `#` ainsi que les hyperliens (`https://`) sont retirés des textes pour éliminer les éléments propres aux réseaux sociaux qui n'apportent pas de valeur analytique.

2.3.1.5 Conversion des Types

Les colonnes contenant des valeurs non numériques sont converties en chaînes de caractères pour assurer la compatibilité avec les techniques de traitement textuel.

2.3.1.6 Tokenisation

La tokenisation est appliquée pour diviser chaque texte en une liste de mots, appelés *tokens*. Cela permet de traiter chaque mot individuellement et de préparer les textes pour des représentations numériques. La tokenisation est réalisée en divisant chaque texte par les espaces.

2.3.1.7 Vectorisation TF-IDF

Une transformation TF-IDF (Term Frequency-Inverse Document Frequency) est appliquée pour convertir le texte en une matrice de caractéristiques numériques. Cette méthode pondère chaque mot en fonction de sa fréquence dans le document et de sa rareté dans l'ensemble des documents, mettant ainsi en avant les mots les plus significatifs pour l'analyse.

2.4 Conclusion

Le prétraitement des données textuelles a permis de réduire le bruit et de standardiser les valeurs, en combinant les champs textuels et en appliquant un nettoyage pour supprimer les éléments non pertinents. Cette étape prépare efficacement les données pour la vectorisation et la modélisation, facilitant ainsi une classification plus précise des annonces frauduleuses.

Chapitre 3

Classification et Prediction

3.1 Introduction

Dans ce chapitre, nous nous sommes concentrés sur l'application d'algorithmes de machine learning pour résoudre un problème de classification, en particulier la détection des annonces frauduleuses.

3.2 Choix des modeles de classification

Nous avons sélectionné plusieurs modèles, chacun apportant des avantages spécifiques à la classification. Les modèles utilisés incluent :

3.2.1 Regression logistique

La régression logistique repose sur l'idée de transformer une sortie linéaire (comme dans la régression linéaire) en une probabilité comprise entre 0 et 1. Cela est réalisé à l'aide de la fonction sigmoïde (ou fonction logistique)

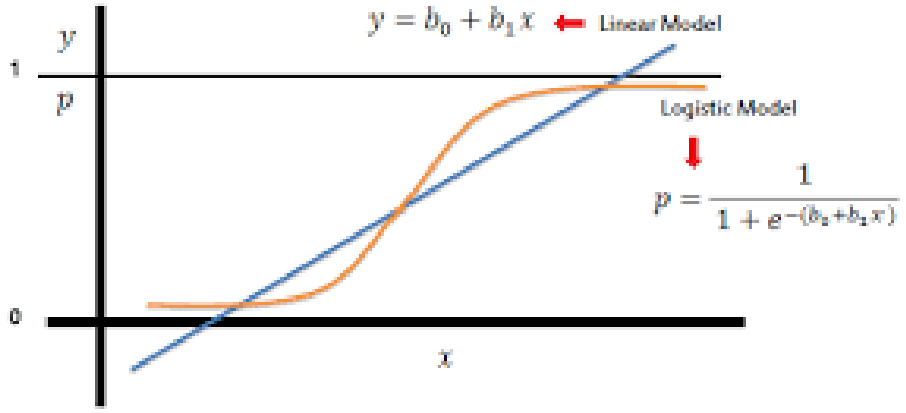

FIGURE 3.1 – principe de la regression logistique

3.2.2 Random Forest

Random Forest est un algorithme d'ensemble qui utilise une combinaison de nombreux arbres de décision pour effectuer des prédictions. Chaque arbre est entraîné sur un sous-ensemble aléatoire des données et des caractéristiques, ce qui améliore la diversité des arbres et réduit le risque de surapprentissage (overfitting). La décision finale est prise par un vote majoritaire des arbres (classification) ou par la moyenne des prédictions (régression).

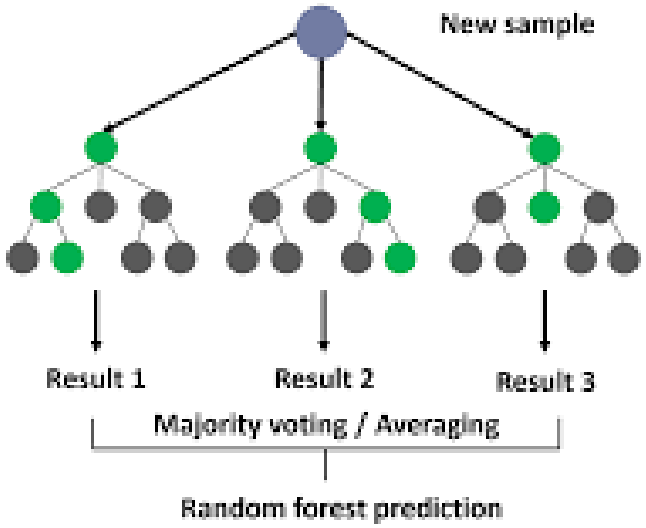

FIGURE 3.2 – principe du random forest

3.2.3 Support Vector Machine

Le SVM est un modèle qui crée une frontière de décision optimale entre les classes en maximisant la marge entre les points de données les plus proches des classes opposées, appelés vecteurs de support. Il peut utiliser différentes fonctions de noyau pour transformer les données et gérer les problèmes non linéaires, comme le noyau gaussien (RBF).

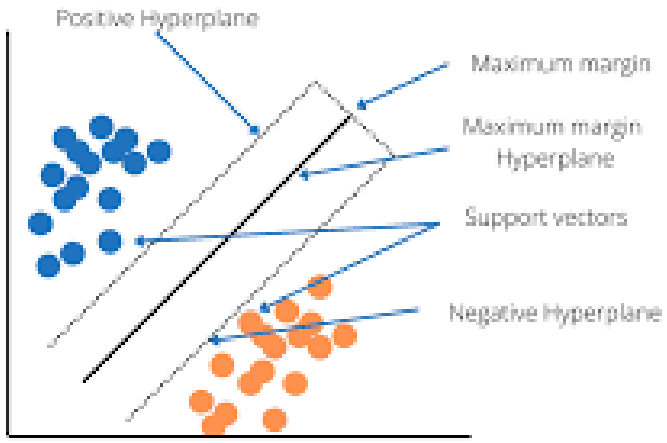

FIGURE 3.3 – principe de svm

3.2.4 Multinomial naive bayes

Basé sur le théorème de Bayes, il calcule la probabilité qu'un exemple appartienne à une classe donnée en multipliant les probabilités conditionnelles des caractéristiques sous l'hypothèse d'indépendance entre elles. Ce modèle suppose que les mots dans un document suivent une distribution multinomiale, ce qui est particulièrement adapté pour des données discrètes, comme celles rencontrées en analyse de texte.

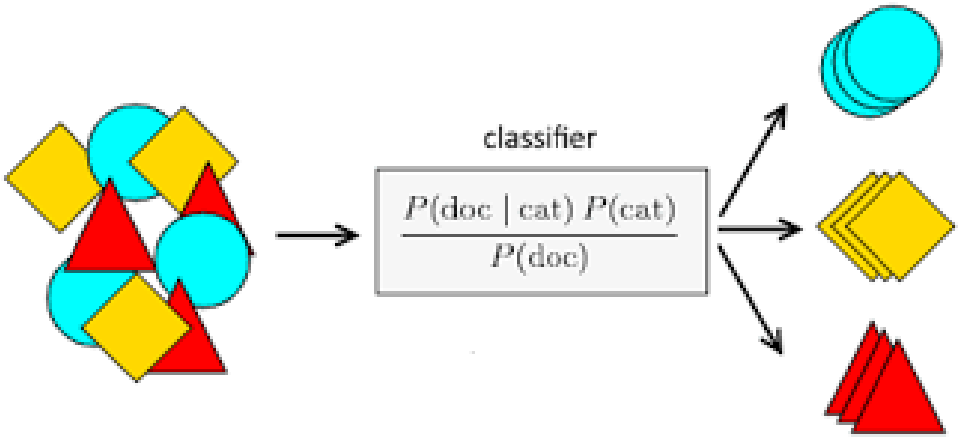

FIGURE 3.4 – principe du multinomial naive bayes

3.3 Metriques d'evaluation

3.3.1 Accuracy

L'Accuracy (ou précision globale) est une métrique utilisée pour évaluer la performance d'un modèle de classification. Elle mesure la proportion de prédictions correctes parmi toutes les prédictions effectuées. En d'autres termes, l'accuracy indique dans quelle mesure le modèle a correctement classifié les exemples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

FIGURE 3.5 – formule de l'accuracy

3.3.2 Recall

Le Recall (ou Sensibilité, aussi appelé True Positive Rate) est une métrique utilisée pour évaluer la capacité d'un modèle à identifier correctement les exemples appartenant à la classe positive. En d'autres termes, il mesure la proportion d'éléments positifs correctement identifiés par le modèle parmi tous les éléments réellement positifs.

FIGURE 3.6 – formule du recall

3.4 Évaluation des Modèles de Classification

3.4.1 Evaluation des modeles

Les modèles de régression logistique, SVM, forêt aléatoire et Naive Bayes Multinomial ont été entraînés et évalués sur les données. Après l'entraînement, leur performance a été mesurée à l'aide de différentes métriques, telles que l'accuracy, le recall et la précision, afin de comparer leur efficacité dans la tâche de classification. Ces évaluations ont permis de sélectionner le modèle le plus performant pour le problème traité.

Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	5104
1	1.00	0.37	0.54	260
accuracy			0.97	5364
macro avg	0.98	0.69	0.76	5364
weighted avg	0.97	0.97	0.96	5364

FIGURE 3.7 – classification report de logistique regression

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5104
1	1.00	0.64	0.78	260
accuracy			0.98	5364
macro avg	0.99	0.82	0.89	5364
weighted avg	0.98	0.98	0.98	5364

FIGURE 3.8 – classification report de random forest

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	3416
1	0.98	0.57	0.72	160
accuracy			0.98	3576
macro avg	0.98	0.78	0.85	3576
weighted avg	0.98	0.98	0.98	3576

FIGURE 3.9 – classification report de svm

```python
from sklearn.metrics import accuracy_score
y_pred = modelNB.predict(X_test)
# Calculate training accuracy
training_accuracy = accuracy_score(y_test, y_pred)
print(training_accuracy)
```

```
0.9552572706935123
```

FIGURE 3.10 – Multimodal naive bayes accuracy

3.4.2 Optimisation des modeles

Le GridSearch a été utilisé pour optimiser le recall et minimiser le nombre de faux négatifs (False Negatives). Cette méthode consiste à rechercher de manière exhaustive les meilleures combinaisons d'hyperparamètres pour un modèle donné, en évaluant différentes configurations sur un ensemble de validation. L'objectif principal de cette optimisation est d'améliorer la capacité du modèle à identifier correctement les cas positifs (maximiser le recall) , ce qui contribue à minimiser le nombre de faux négatifs. En ajustant des hyperparamètres clés comme le seuil de classification, les paramètres de régularisation ou les critères de décision, GridSearch permet d'améliorer la performance globale du modèle pour mieux équilibrer la détection des classes positives et négatives.

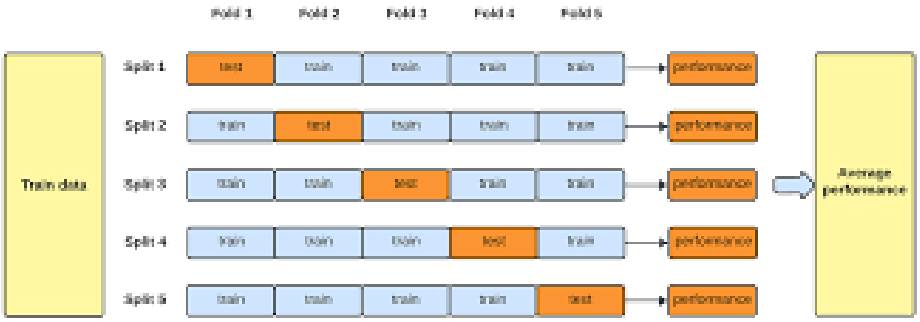

FIGURE 3.11 – Cross validation principe

Dans le cas du RandomForestClassifier, plusieurs hyperparamètres influencent directement la capacité du modèle à généraliser et à éviter le sur-apprentissage. Les paramètres choisis, tels que classweight='balanced', maxdepth=10, maxfeatures='sqrt', minsamplessplit=10, et nestimators=200, ont été ajustés pour offrir le meilleur compromis entre biais et variance. L'utilisation de classweight='balanced' permet de traiter les déséquilibres de classes, tandis que la limitation de la profondeur des arbres à 10 et la définition de minsamplessplit=10 contribuent à éviter un sur-apprentissage excessif. Le paramètre maxfeatures='sqrt' permet d'améliorer la diversité des arbres tout en réduisant la variance, et un grand nombre d'estimateurs (200 arbres) assure une meilleure stabilité du modèle. Après une recherche minutieuse de ces hyperparamètres à l'aide de méthodes telles que le GridSearch, ces configurations ont montré des performances optimales en termes de précision et de généralisation, offrant ainsi une solution robuste et efficace pour la tâche de classification.

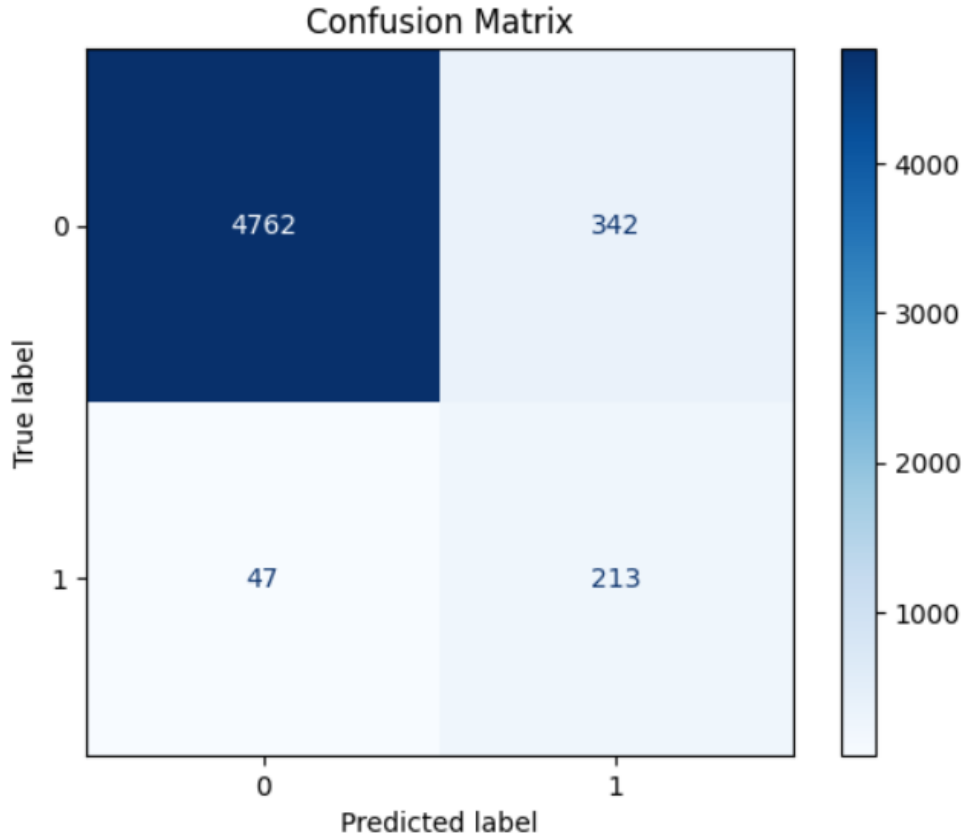

FIGURE 3.12 – matrice de confusion du random forest

Le modèle SVM (Support Vector Machine) avec les hyperparamètres sélectionnés, à savoir C=0.1, classweight='balanced', et kernel='linear', a été entraîné et optimisé pour améliorer sa capacité à généraliser sur les données. L'objectif principal était d'obtenir un bon compromis entre précision et recall, en particulier pour les classes minoritaires. Le paramètre C=0.1 a été choisi pour régulariser le modèle, lui permettant de maintenir une marge de séparation plus large entre les classes, même au prix de quelques erreurs de classification, afin de mieux généraliser sur de nouvelles données. En utilisant classweight='balanced', le modèle a été ajusté pour compenser les déséquilibres entre classes, ce qui est crucial pour maximiser le nombre de vrais positifs tout en réduisant les faux négatifs. Le noyau linéaire a été préféré en raison de sa simplicité et de son efficacité dans les cas où les données sont presque linéairement séparables.

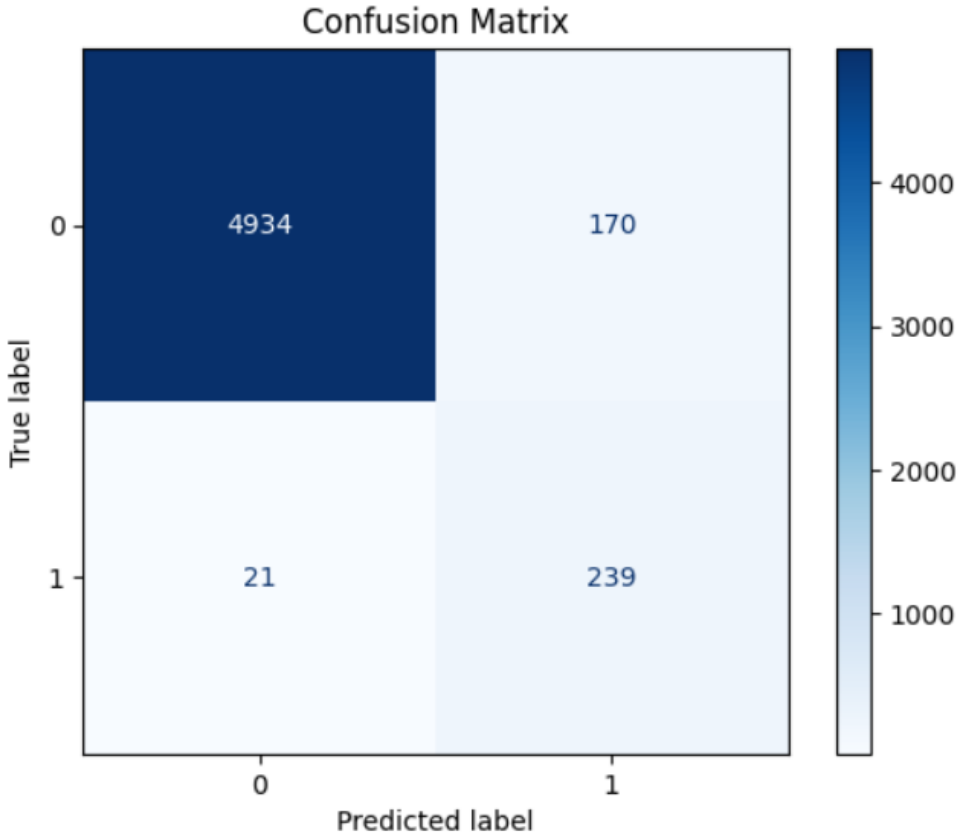

FIGURE 3.13 – matrice de confusion du svm

3.4.3 Selection du modele

Après l'évaluation de plusieurs modèles, c'est le Support Vector Machine (SVM) qui a présenté le nombre minimum de faux négatifs (False Negatives). En utilisant un noyau linéaire optimisé avec des hyperparamètres ajustés (C=0.1 et classweight='balanced'), le modèle SVM a démontré une capacité supérieure à identifier correctement les cas positifs, ce qui a permis de réduire au maximum le nombre de faux négatifs. Cette performance est particulièrement importante dans des contextes où il est crucial de capturer un maximum d'exemples positifs, même au prix d'un léger compromis sur d'autres métriques. Par rapport aux autres modèles testés, tels que la régression logistique, le Naive Bayes et le Random Forest, le SVM a donc été retenu comme le modèle le plus fiable pour minimiser les erreurs critiques liées à la non-détection de cas positifs.

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.97	0.98	5104
1	0.58	0.92	0.71	260
accuracy			0.96	5364
macro avg	0.79	0.94	0.85	5364
weighted avg	0.98	0.96	0.97	5364

FIGURE 3.14 – Classification report du svm apres optimisation

3.5 Conclusion

À l'issue de ce chapitre, nous avons pu identifier le SVM comme le modèle le plus performant pour la tâche de classification des annonces frauduleuses, notamment grâce à son recall élevé et sa capacité à minimiser les faux négatifs. Les techniques d'optimisation des hyperparamètres ont permis d'affiner les performances de chaque modèle, mais c'est le SVM, avec ses ajustements de class weight et paramètres de régularisation, qui s'est avéré le plus adapté à notre ensemble de données déséquilibré.

Conclusion générale

En conclusion, ce projet a permis de développer un modèle de détection des annonces d'emploi frauduleuses en appliquant des techniques de traitement du langage naturel et de machine learning. Les étapes, depuis la préparation des données jusqu'à l'évaluation des modèles, ont conduit à identifier des caractéristiques propres aux annonces frauduleuses et à construire un modèle performant.

Le modèle de Support Vector Machine (SVM) optimisé s'est distingué par sa capacité à minimiser les faux négatifs, crucial pour identifier efficacement les annonces frauduleuses. Ce projet met également en évidence l'importance du prétraitement des données textuelles et de l'optimisation des hyperparamètres pour garantir des performances optimales.

À l'avenir, il serait intéressant d'explorer des modèles plus avancés, comme les réseaux neuronaux ou les modèles de type Transformer, pour améliorer encore la précision. Intégrer d'autres sources de données, comme le comportement des utilisateurs, pourrait également enrichir la détection.

En somme, cette étude constitue une base prometteuse pour des applications de détection de fraude, avec un potentiel d'adaptation dans divers domaines touchés par la fraude en ligne.