



ECOLE CENTRALE CASABLANCA

ML OPERATIONNEL

Projet final : Resume screening

Membres :

DAHASSI Chaymae
EL MRABET Aïmane
DHIMEN Aymane
MEZIANY Imane
AIT BIHI Laila

Sous la supervision de :

M. Hicham Dakhli

16 décembre 2023

Table des matières

1	Scoping	2
2	Data	3
2.1	Data Definition	3
2.2	Data Quality	3
2.3	Data exploration/EDA	4
2.3.1	Aperçu des données	4
2.3.2	Visualisation des données	5
3	Model	7
3.1	Division des données	7
3.2	Construction du modèle :	7
3.3	Baseline metrics/KPIs	7
3.4	Prédiction et Évaluation	8
4	Deployment :	8
5	Monitoring	9

1 Scoping

Le recrutement de talents compétents constitue un défi universel pour les entreprises, une difficulté qui s'accroît particulièrement dans les entreprises intensives en main-d'œuvre, en phase de croissance, et confrontées à des taux élevés de rotation du personnel.

Les départements informatiques, confrontés à des marchés en expansion, ont fréquemment recours à des professionnels possédant une diversité de compétences techniques et une expertise pointue dans des domaines spécifiques. Le processus critique de sélection des meilleurs talents parmi de nombreux autres, également connu sous le nom de "Screening de CV", revêt une importance capitale.

L'intégration d'algorithmes de Machine Learning (ML) dans le processus de "Screening de CV" constitue une avancée majeure visant à enrichir considérablement les pratiques de recrutement des entreprises. Cette approche permet une évaluation rapide et efficace des candidats sans nécessiter un examen manuel de chaque CV, une pratique souvent irréalisable pour les grandes entreprises confrontées à des contraintes de temps.

Les bénéfices substantiels liés à cette approche incluent :

Efficacité accrue : En automatisant une partie du processus de tri des CV, les entreprises peuvent gagner en efficacité, accélérant ainsi le cycle de recrutement. Cela permet de traiter un grand volume de candidatures en un temps minimal.

Précision dans la sélection : Les algorithmes de ML peuvent être formés pour identifier rapidement les candidats correspondant le mieux aux critères spécifiques du poste. Cela contribue à une sélection plus précise et à une meilleure adéquation entre les compétences des candidats et les exigences du travail.

Réduction de la charge de travail manuel : En évitant l'examen manuel de chaque CV, les équipes de recrutement peuvent consacrer leur temps à des tâches plus stratégiques et à une interaction plus approfondie avec les candidats présélectionnés.

Minimisation des biais : En utilisant des algorithmes basés sur des critères objectifs, le processus de "Screening de CV" peut contribuer à réduire les biais humains potentiels, favorisant ainsi une évaluation plus équitable des candidats.

Adaptabilité continue : Les modèles de ML peuvent être ajustés et améliorés au fil du temps en fonction des retours d'expérience, assurant ainsi une adaptation constante aux besoins changeants de l'entreprise.

Gain financier : La réduction du temps nécessaire au processus de recrutement et l'amélioration de la qualité des embauches peuvent se traduire par des économies financières à long terme pour l'entreprise.

2 Data

2.1 Data Definition

A ce stade, notre objectif premier est de préciser les données destinées à l'entraînement et à l'évaluation de notre modèle.

La base de données utilisée a été téléchargée depuis [Kaggle](<https://www.kaggle.com/datasets/gauravkhandelwal/dataset/data>). Elle est structurée, adoptant un format tabulaire dans un fichier CSV. La structure est définie par deux colonnes principales : "Category" et "Resume". La colonne "Category" attribue une catégorie spécifique à chaque entrée, représentant le secteur industriel associé au CV du candidat, englobant des domaines tels que : Data Science, HR, Advocate, Arts, Web Designing, Mechanical Engineer, Sales, Health and Fitness, etc. D'autre part, la colonne "Resume" contient les données des CVs correspondants.

Cette organisation permet une gestion systématique des informations, facilitant une recherche, une analyse, et une compréhension claire des données. La nature structurée de cette base de données est cruciale pour la catégorisation des profils professionnels et offre un cadre organisé qui peut être exploité pour diverses analyses et applications dans les domaines des ressources humaines, du recrutement, et de l'analyse sectorielle.

2.2 Data Quality

Le prétraitement des données est une étape cruciale dans le processus d'analyse de données qui vise à préparer, nettoyer et organiser les données brutes avant leur utilisation dans des applications analytiques ou de modélisation. Cette phase comprend plusieurs étapes, telles que la gestion des valeurs manquantes, l'élimination du bruit, la normalisation, la conversion de formats, et d'autres techniques visant à rendre les données plus cohérentes, compréhensibles et adaptées à l'analyse. L'objectif principal du prétraitement est d'améliorer la qualité des données, de réduire les erreurs potentielles, et d'assurer la fiabilité des résultats obtenus lors de l'application de techniques d'exploration, de modélisation statistique, ou d'apprentissage automatique.

Dans ce sens, un ensemble complet d'étapes de prétraitement des données a été implémenté afin de s'assurer que les données textuelles extraites des CV sont de qualité et fiables :

- **Suppression des URL** : Toutes les URL ont été éliminées du texte du CV. Cette étape est cruciale pour éliminer les hyperliens, qui peuvent ne pas contribuer de manière significative à l'analyse et pourraient introduire du bruit.

- **Suppression de RT et CC** : Les occurrences de "RT" (Retweet) et "CC" ont été supprimées. Ce sont des artefacts courants issus du contenu des médias sociaux et ne sont pas pertinents dans le contexte de l'analyse du texte du CV.

- **Suppression des hashtags et mentions** : Les hashtags et mentions ont été

supprimés du texte. Cela améliore la propreté des données en éliminant des éléments centrés sur les médias sociaux.

- **Suppression de la ponctuation** : Divers signes de ponctuation ont été supprimés. Cela garantit un texte cohérent et standardisé, évitant les problèmes pouvant découler de la présence de ponctuations inutiles.

- **Suppression des caractères non ASCII** : Les caractères non ASCII, qui pourraient causer des incohérences d'encodage, ont été remplacés par des espaces. Cette étape favorise la compatibilité et l'uniformité des données.

- **Réduction des espaces blancs** : Les espaces blancs superflus à l'intérieur du texte ont été réduits. Cela contribue non seulement à une apparence plus propre, mais facilite également le maintien d'un format constant pour l'analyse.

En nettoyant et standardisant le texte, nous créons une base solide pour les analyses ultérieures, sans artefacts ni distractions inutiles.

2.3 Data exploration/EDA

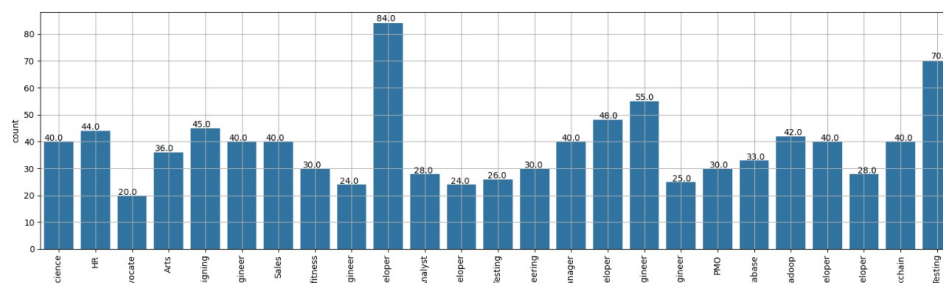
2.3.1 Aperçu des données

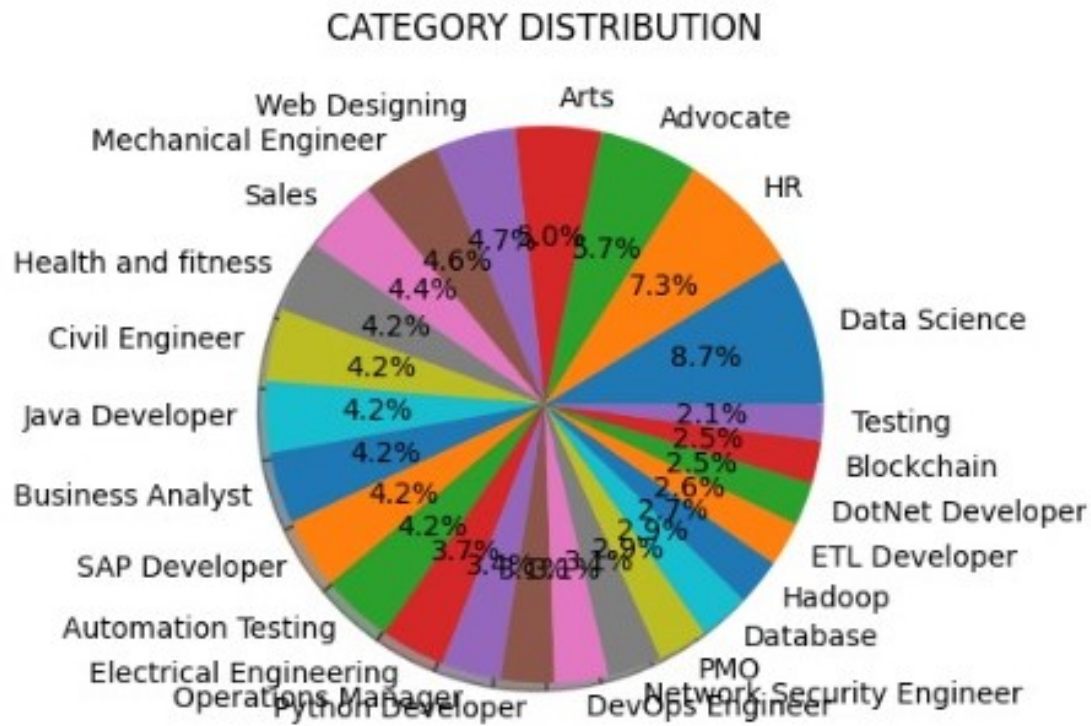
Dans cette phase, les données sont chargées à partir d'un fichier CSV à l'aide de la bibliothèque pandas. Une première exploration des données est réalisée en affichant les premières lignes du jeu de données à l'aide de la fonction `head()`. Cela permet d'avoir un aperçu rapide de la structure et du format des données.

A		B
1	ID	Resume_str
2	16852973	<p>HR ADMINISTRATOR/MARKETING ASSOCIATE</p> <p>HR ADMINISTRATOR Summary Dedicated Customer Service Manager with 15+ years of experience in Hospitality and Customer Service Management. Respected builder and leader of customer-focus strives to instill a shared, enthusiastic commitment to customer service. Highlights Focused on customer satisfaction Team management Marketing savvy Conflict resolution techniques Training and devel Skilled multi-tasker Client relations specialist Accomplishments Missouri DOT Supervisor Training Certification Certified by IHG in Customer Loyalty and Marketing by Segment Hilton Worldwide General Training Certification Accomplished Trainer for cross server hospitality systems such as Hilton OnQ, Micros Opera PMS, Fidelio OPERA Reservation System (ORS), Holidayx Completed courses and se customer service, sales strategies, inventory control, loss prevention, safety, time management, leadership and performance assessment. Experience HR Administrator/Marketing Associate</p> <p>HR Administrator Dec 2013 to Current Company Name - City, State Helps to develop policies, directs and coordinates activities such as employment, compensation, labor relations, benefits, training, a employee services. Prepares employee separation notices and related documentation Keeps records of benefits plans participation such as insurance and pension plan, personnel transactions such as h promotions, transfers, performance reviews, and terminations, and employee statistics for government reporting. Advises management in appropriate resolution of employee relations issues. Administers programs such as life, health, dental, insurance, pension plans, vacation, sick leave, leave of absence, and employee assistance. Marketing Associate Designed and created marketing collateral for sale meetings, trade shows and company executives. Managed the in-house advertising program consisting of print and media collateral pieces. Assisted in the complete design and launch of the company's 2 months. Created an official company page on Facebook to facilitate interaction with customers. Analyzed ratings and programming features of competitors to evaluate the effectiveness of marketing str Advanced Medical Claims Analyst Mar 2012 to Dec 2013 Company Name - City, State Reviewed medical bills for the accuracy of the treatments, tests, and hospital stays prior to sanctioning the claims to interpret the codes (ICD-9, CPT) and terminology commonly used in medical billing to fully understand the paperwork that is submitted by healthcare providers. Required to have organizational and an skills as well as computer skills, knowledge of medical terminology and procedures, statistics, billing standards, data analysis and laws regarding medical billing. Assistant General Manager Jun 2010 to C Company Name - City, State Performed duties including but not limited to, budgeting and financial management, accounting, human resources, payroll and purchasing. Established and maintained clo relationships with all departments of the hotel to ensure maximum operation, productivity, morale and guest service. Handled daily operations and reported directly to the corporate office. Hired and traine overall objectives and goals with an emphasis on high customer service. Marketing and Advertising, working on public relations with the media, government and local businesses and Chamber of Comme Executive Support / Marketing Assistant Jul 2007 to Jun 2010 Company Name - City, State Provided assistance to various department heads - Executive, Marketing, Customer Service, Human Resour Managed front-end operations to ensure friendly and efficient transactions. Ensured the swift resolution of customer issues to preserve customer loyalty while complying with company policies. Exemplifie second-to-none customer service delivery in all interactions with customers and potential clients. Reservation & Front Office Manager Jun 2004 to Jul 2007 Company Name - City, State Owner/ Partner to May 2004 Company Name - City, State Price Integrity Coordinator Aug 1999 to Dec 2001 Company Name - City, State Education N/A, Business Administration 1999 Jefferson College - City, St Business Administration Marketing / Advertising High School Diploma, College Prep. studies 1998 Sainte Genevieve Senior High - City, State Awarded American Shrubel Leadership Scholarship to Jef College Skills Accounting, ads, advertising, analytical skills, benefits, billing, budgeting, clients, Customer Service, data analysis, delivery, documentation, employee relations, financial management, gove relations, Human Resources, insurance, labor relations, layout, Marketing, marketing collateral, medical billing, medical terminology, office, organizational, payroll, performance reviews, personnel, policie presentations, public relations, purchasing, reporting, statistics, website.</p>
3	22323957	<p>HR SPECIALIST, US HR OPERATIONS Summary Versatile media professional with background in Communications, Marketing, Human Resources and Technology. Experience 09/2015 to Current HR 5 US HR Operations Company Name - City, State Managed communication regarding launch of Operations group, policy changes and system outages Designed standard work and job aids to create comprehensive training program for new employees and contractors Audited job postings for old, pending, on-hold and draft positions. Audited union hourly, non-union hourly and salary background chec drug screens Conducted monthly new hire benefits briefing to new employees across all business units Served as a link between HR Managers and vendors by handling questions and resolving system- issues Provide real-time process improvement feedback on key metrics and initiatives Successfully re-branded US HR Operations SharePoint site Business Unit project manager for RF/RFP on Backgro Check and Drug Screen vendor 01/2014 to 05/2015 IT, Marketing and Communications Co-op Company Name - City, State Posted new articles, changes and updates to corporate SharePoint site incl graphics and visual communications. Researched and drafted articles and feature stories to promote company activities and programs. Co-edited and developed content for quarterly published newsletter communication support for internal and external events. Collaborated with Communication team, media professionals and vendors to determine program needs for print materials, web design and digital communications. Entrusted to lead product, service and software launches for Digital Asset Management tool, Marketing Toolkit website and Executive Tradeshows Calendar. Created presentations for management and executive approval to ensure alignment with corporate guidelines and branding. Maintained the MySikorsky SharePoint site and provided timely solutions to mitigate issues. Created board and produced video for annual IT All Hands meeting. 10/2012 to 01/2014 Relationship Coordinator/Marketing Specialist Company Name - City, State Partnered with vendor to manage the in-hou advertising program consisting of print and media collateral pieces. Coordinated pre-show and post-show activities at trade shows. Managed marketing campaigns to generate new business and to supp and sales teams. Ordered marketing collateral for meetings, trade shows and advisors. Improved, administered and modified marketing programs to increase product awareness. Assisted in preparing int promotional publications, managed marketing material inventory and supervised distribution of publications to ensure high quality product output. Coordinated marketing materials including brochures, prc materials and products. Partnered with graphic designers to develop appropriate materials and branding for brochures. Used tracking and reporting systems for sales leads and appointments. 09/2009 to Assistant Head Teller Company Name - City, State Received an internal audit score of 100%. Performed daily and monthly audits of ATM machines and tellers. Educated customers on a variety of reta and available credit options. Consistently met or exceeded quarterly sales goals Promoted products and services to</p>

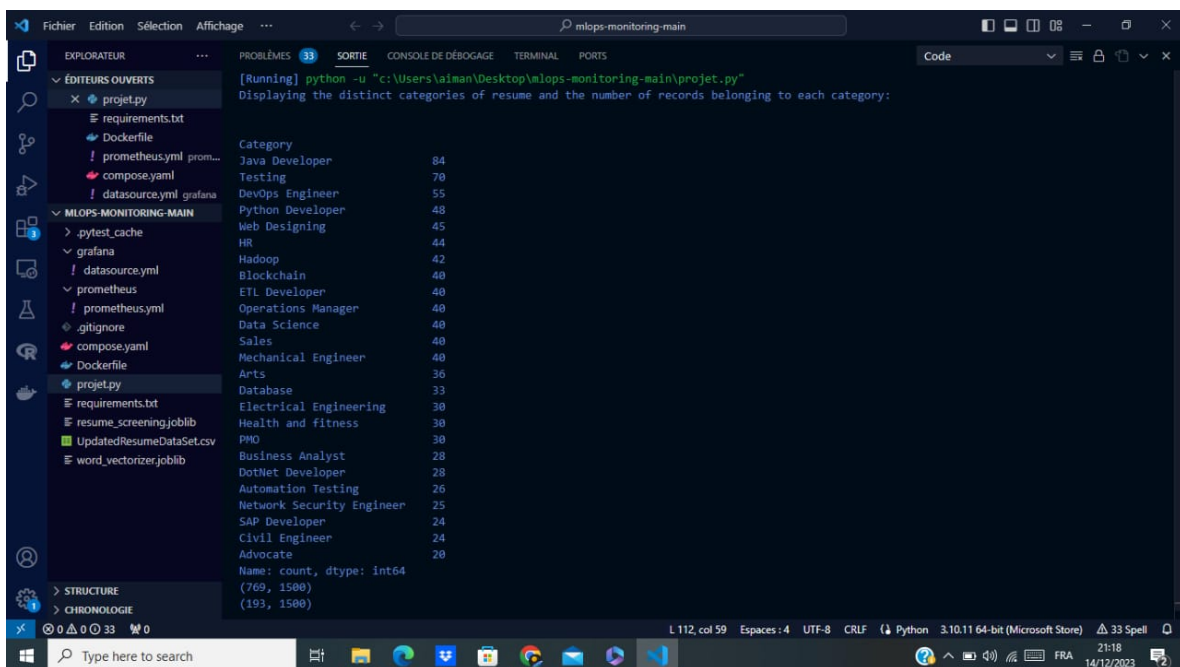
2.3.2 Visualisation des données

Une visualisation graphique est essentielle pour comprendre la distribution des catégories dans le jeu de données. Un graphique à barres, créé à l'aide de la bibliothèque seaborn, ainsi qu'un diagramme circulaire, permettent d'observer la fréquence de chaque catégorie, offrant ainsi une première impression visuelle de la distribution des données.





Nous avons révélé le nombre d'enregistrements appartenant à chaque catégorie unique dans la colonne 'Category' du DataFrame resumeDataSet. Cette information offre une compréhension détaillée de la répartition des données dans différentes catégories, fournissant ainsi un aperçu essentiel de la diversité des catégories de CV présentes dans le jeu de données.



3 Model

3.1 Division des données

Dans cette phase cruciale de construction du modèle, nous avons suivi un ensemble d'étapes méthodiques pour entraîner et évaluer notre modèle de classification basé sur les k plus proches voisins (KNeighborsClassifier). Tout d'abord, nous avons divisé nos données en ensembles d'entraînement et de test en utilisant la fonction `train_test_split`. Cette division, avec une taille de test de 20% et une randomisation fixée pour assurer la reproductibilité, a permis de préparer nos données pour l'entraînement et l'évaluation du modèle.

3.2 Construction du modèle :

En ce qui concerne la construction du modèle, nous avons opté pour une approche One-vs-Rest encapsulant l'algorithme k-NN.

Ce choix trouve sa pertinence dans sa simplicité conceptuelle et sa capacité à traiter efficacement des problèmes de classification. Le principe fondamental de cet algorithme repose sur la proximité des échantillons dans l'espace des caractéristiques : les individus similaires sont regroupés dans des catégories communes. Cette approche non paramétrique s'adapte bien à notre étude de Screening de CV, où la similarité entre les compétences et l'expérience des candidats peut jouer un rôle déterminant dans la classification. En outre, parmi ses avantages notables, on cite sa flexibilité pour gérer des données de nature textuelle, ce qui est particulièrement pertinent dans notre cas. De plus, ce modèle offre une interprétabilité accrue, facilitant la compréhension des résultats de classification.

Le modèle ainsi créé a été ensuite entraîné sur l'ensemble d'entraînement, composé des caractéristiques des mots (TF-IDF) extraites de nos CVs.

3.3 Baseline metrics/KPIs

1. La **précision** mesure le nombre de vrais positifs parmi tous les exemples positifs prédits par le modèle. C'est une mesure de l'exactitude des prédictions positives du modèle. Il s'agit du ratio des vrais positifs sur la somme des vrais positifs et les faux positifs.

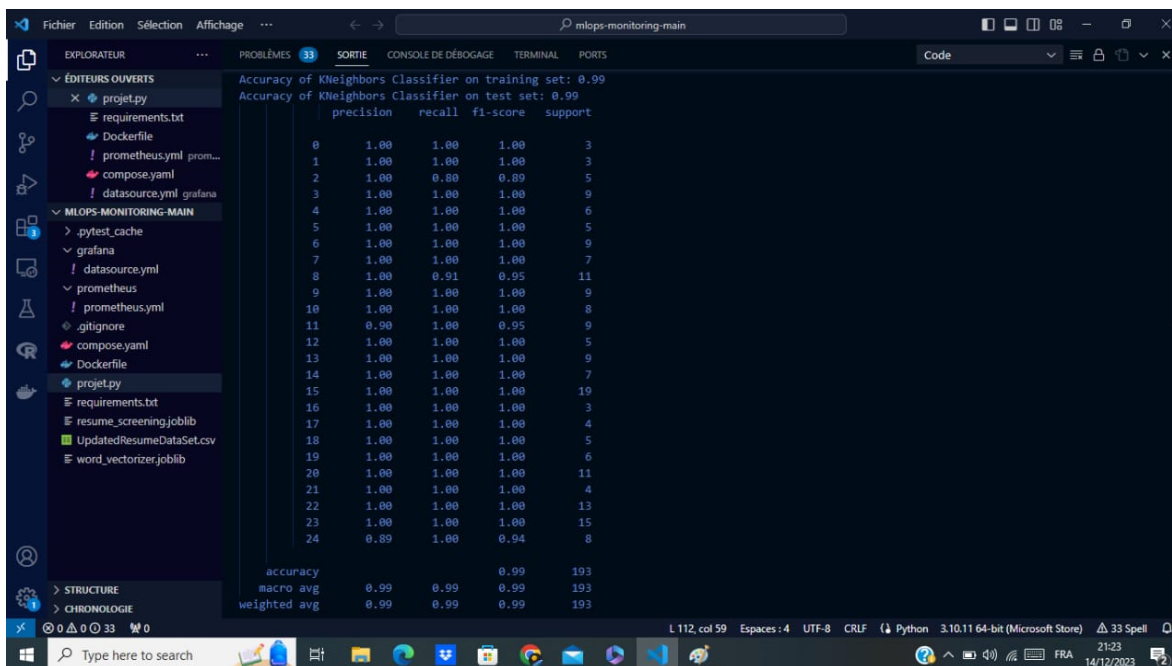
2. Le **Recall** évalue la capacité du modèle à identifier tous les exemples positifs. Il mesure la proportion d'exemples positifs réels que le modèle a correctement identifiés. Il s'agit du ratio des vrais positifs sur la somme des vrais positifs et les faux négatifs.

3. Le **F1-score** est une mesure qui combine la précision et le rappel en une seule valeur. Il est particulièrement utile lorsque les classes sont déséquilibrées.

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

3.4 Prédiction et Évaluation

Après l'entraînement, nous avons procédé à la prédiction sur l'ensemble de test pour évaluer la performance de notre modèle. Les résultats ont été analysés en termes de précision, aussi bien sur l'ensemble d'entraînement que sur l'ensemble de test. Ces métriques fournissent une indication précieuse de la capacité du modèle à généraliser à de nouvelles données.



The screenshot shows a VS Code editor with a Jupyter Notebook open. The notebook displays the accuracy of a KNeighborsClassifier on both training and test sets, followed by a detailed performance table.

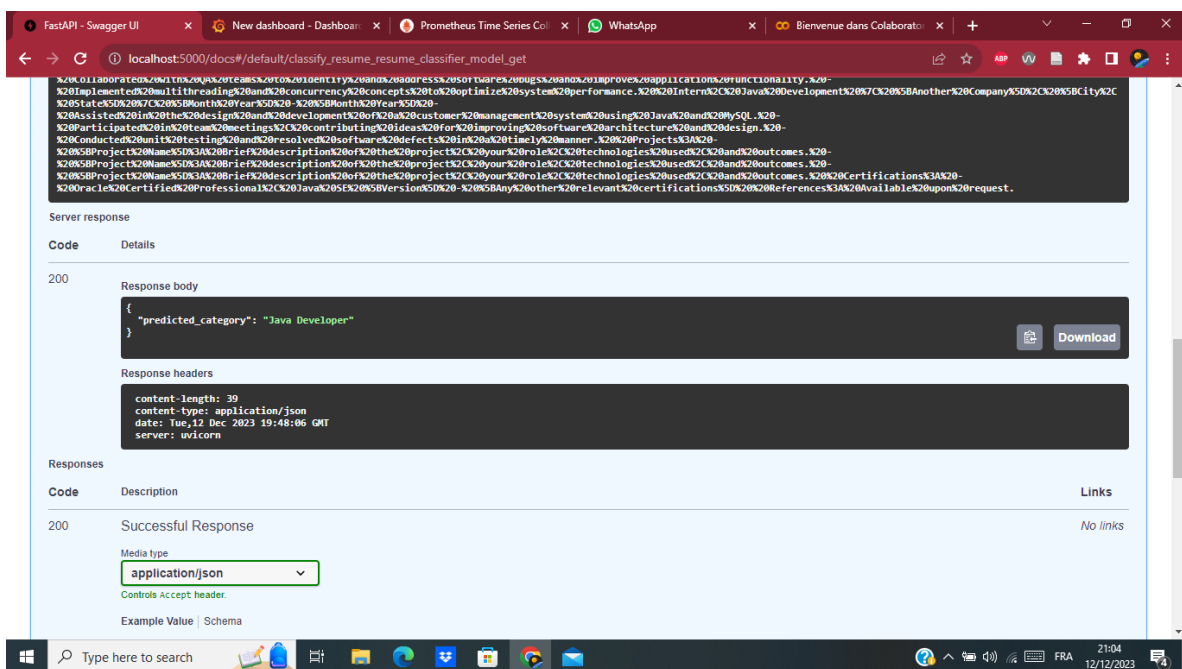
	precision	recall	f1-score	support
Accuracy of KNeighbors Classifier on training set: 0.99				
Accuracy of KNeighbors Classifier on test set: 0.99				
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	3
2	1.00	0.80	0.89	5
3	1.00	1.00	1.00	9
4	1.00	1.00	1.00	6
5	1.00	1.00	1.00	5
6	1.00	1.00	1.00	9
7	1.00	1.00	1.00	7
8	1.00	0.91	0.95	11
9	1.00	1.00	1.00	9
10	1.00	1.00	1.00	8
11	0.90	1.00	0.95	9
12	1.00	1.00	1.00	5
13	1.00	1.00	1.00	9
14	1.00	1.00	1.00	7
15	1.00	1.00	1.00	19
16	1.00	1.00	1.00	3
17	1.00	1.00	1.00	4
18	1.00	1.00	1.00	5
19	1.00	1.00	1.00	6
20	1.00	1.00	1.00	11
21	1.00	1.00	1.00	4
22	1.00	1.00	1.00	13
23	1.00	1.00	1.00	15
24	0.89	1.00	0.94	8
accuracy			0.99	193
macro avg	0.99	0.99	0.99	193
weighted avg	0.99	0.99	0.99	193

Enfin, conscient de l'importance de la sauvegarde des modèles entraînés, nous avons utilisé la bibliothèque joblib pour sauvegarder notre modèle de classification ainsi que le vectoriseur de mots utilisé dans le processus de prétraitement. Cette étape est importante pour permettre un déploiement facile et rapide du modèle dans des environnements de production.

4 Deployment :

Pour concrétiser notre modèle de Screening de CV, nous avons opté pour un déploiement en utilisant FastAPI, une bibliothèque Python dédiée à la création rapide d'API. Cette approche nous permet d'exposer notre modèle en tant que service Web, offrant une interface facilement accessible pour l'intégration avec d'autres applications. FastAPI excelle dans la prise en charge de la documentation automatique, facilitant ainsi la compréhension de l'API par les utilisateurs.

Ci-après un exemple d'utilisation de notre modèle où nous avons fourni au modèle un exemple de CV et nous a retourné la catégorie auquel appartient ce CV.



5 Monitoring

Afin de monitorer en temps réel les performances de notre modèle, nous avons intégré le tandem Prometheus/Grafana. Prometheus, un système open-source de surveillance et d'alerte, collecte les métriques pertinentes générées par notre API FastAPI. Ces métriques sont ensuite visualisées de manière graphique et intuitive grâce à Grafana, offrant une surveillance continue du comportement de notre modèle déployé. Cette combinaison de FastAPI, Prometheus, et Grafana constitue une infrastructure pour exposer notre modèle de machine learning, tout en fournissant des mécanismes de suivi essentiels pour garantir des performances optimales dans des environnements opérationnels.

Dans notre cas, nous avons décidé de suivre trois métriques, une qui mesure la durée du scrape et l'état du scrape (1 si réussi et 0 sinon), qui sont deux métriques prédéfinis dans Prometheus, et nous avons ajouté à notre API une métrique qui mesure le nombre des requêtes HTTP, et qui sert à suivre l'utilisation et le trafic de notre modèle, qui est une métrique importante à suivre lorsqu'un modèle de machine learning est déployé dans un environnement de production pour garantir que le modèle fonctionne de manière fiable, réactive et évolutive.

