Projet : Data Collection and Visualisation

Analyse des offres d'emploi tech en France

IMT Mines Alès – Département 2IA
Anas Seffraoui
Najoua Labriki
Salmane El hajouji
Sami Ait bella

Avril 2025

Table des matières

Introduction

Le marché de l'emploi dans le secteur tech et informatique connaît une forte croissance, notamment en France. Les entreprises recherchent des profils dotés de compétences spécifiques, mais ces exigences varient en fonction de plusieurs facteurs tels que la localisation géographique, le secteur d'activité, le type de contrat proposé ou encore le niveau de salaire.

Dans ce contexte, nous pensions qu'il serait pertinent de se demander si des éléments tels que le lieu de travail ou le salaire peuvent permettre d'anticiper les compétences informatiques attendues dans une offre d'emploi.

Afin de mieux comprendre cette problématique, nous avons formulé la question scientifique suivante :

"Peut-on estimer les compétences informatiques requises à partir du lieu de travail, du niveau de salaire et d'autres critères professionnels ?"

Ainsi, l'objectif principal du projet est de :

- Collecter un volume représentatif d'offres d'emploi tech en France via web scraping,
- Développer un tableau de bord interactif pour explorer les données collectées,
- Identifier les liens entre localisation, compétences, salaire et autres variables,
- Proposer des pistes pour une future analyse prédictive.

1 Méthodologie

1.1 Collecte des données

Outils et script de scraping

La collecte a été réalisée à l'aide d'un script Python dans un notebook Jupyter nommé `Data_scrapping_Hellowork_final.ipynb`, reposant sur les bibliothèques `requests`, `BeautifulSoup` et `pandas`.

Défis techniques rencontrés

L'un des principaux obstacles rencontrés concernait l'absence de standardisation des balises HTML sur HelloWork. Pour contourner cela, le script identifie dynamiquement des balises clés à l'aide de mots comme `missions` ou `profil`.

Données extraites

Nous avons extrait pour chaque offre :
- Titre du poste, entreprise, lieu, département,
- Secteur, type de contrat, télétravail,
- Salaire estimé, compétences, niveau d'études, années d'expérience.

1.2 Choix de la taille d'échantillon

Afin d'assurer la fiabilité des résultats, nous avons fondé notre stratégie de collecte sur une démarche d'échantillonnage statistique. En partant d'une estimation de 25 000 offres dans le domaine tech sur HelloWork, un calcul basé sur un niveau de confiance de 95% et une marge d'erreur de 5% donne une taille d'échantillon minimale de 381.

Pour anticiper les pertes après nettoyage (doublons, données incomplètes), nous avons fixé notre cible à 450 offres exploitables.

1.3 Prétraitement des données

Les principales étapes de préparation des données ont été :
- Suppression des doublons,
- Nettoyage des textes,

— Conversion des salaires en brut annuel,
— Extraction des compétences en listes,
— Export vers un fichier `.csv`.

2 Analyse exploratoire et visualisation

2.1 Outils utilisés

Un tableau de bord a été conçu avec `Shiny` et `flexdashboard` en R pour visualiser les résultats de manière dynamique et interactive.

2.2 Fonctionnalités du tableau de bord

Il permet d'explorer :

- Statistiques descriptives (nombre d'offres, salaires moyens, télétravail),
- Carte interactive des offres par département,
- Histogrammes et heatmaps des compétences par secteur,
- Répartition des contrats selon l'expérience,
- Corrélations entre salaire, diplôme, expérience.

2.3 Filtrage et recherche

Une section permet à l'utilisateur de filtrer dynamiquement les offres selon différents critères : type de contrat, localisation, compétences, secteur ou rémunération.

3 Pistes pour l'analyse prédictive

3.1 État actuel

Aucune modélisation prédictive n'a encore été mise en œuvre. L'analyse demeure descriptive. Cette section propose des pistes futures.

3.2 Objectif de modélisation

L'objectif serait de prédire les compétences attendues dans une offre à partir de :

- Lieu de travail,
- Type de contrat,
- Salaire,
- Expérience requise,
- Niveau d'études,
- Secteur d'activité.

3.3 Approche méthodologique

Un modèle de classification multilabel serait pertinent. Les algorithmes possibles sont :

- Random Forest,
- Gradient Boosting,
- Réseaux de neurones légers.

3.4 Préparation des données pour l'IA

Pour lancer cette modélisation :

- Encodage des variables qualitatives (One-Hot Encoding),
- Transformation des compétences en vecteurs multilabel,
- Normalisation des variables numériques,
- Validation croisée.

3.5 Intégration au tableau de bord

La modélisation pourrait à terme enrichir le dashboard, via un simulateur de profil. Un utilisateur saisirait ses caractéristiques et obtiendrait les compétences attendues correspondantes.

4 Conclusion

Ce projet nous a permis de :

- Comprendre les tendances du marché de l'emploi tech en France,
- Identifier les compétences les plus demandées,
- Visualiser les relations entre variables clés (expérience, salaire, diplôme, localisation),
- Poser les bases d'un futur modèle prédictif.

Il ouvre des perspectives pour des outils d'aide à la décision en orientation, recrutement ou reconversion. Une API ou une interface intelligente pourrait être envisagée pour recommander des compétences clés ou détecter des écarts de profil.