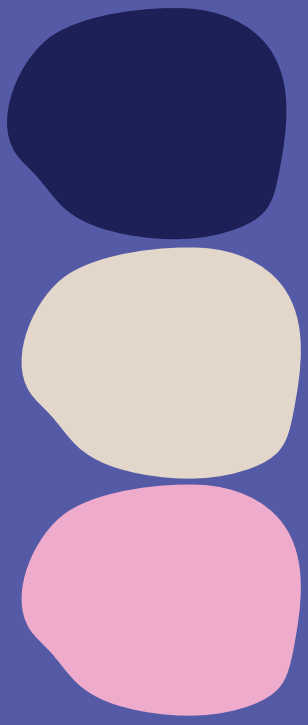
Projet I.A - Livrable Bibliographie

Présentation des références académiques
et techniques



Hichem CHERAFI, Tom NORMAND, Sulivan TULOUP

Février 2025

CESI

HUMANFORYOU

GROUPE 4

Sommaire

Introduction

3

Contexte du projet

L'entreprise pharmaceutique HumanForYou en Inde rencontre un gros problème avec ses employés. Chaque année 15 % d'entre eux quittent l'entreprise.

Cela pose des difficultés pour mener les projets à bien nuit à la réputation de l'entreprise et coûte cher en recrutement et formation de nouveaux talents.

Pour comprendre pourquoi ces départs ont lieu et trouver des solutions HumanForYou a demandé une analyse détaillée utilisant l'intelligence artificielle.

Plan de recherche

Le projet commence par récupérer et préparer les données. On nettoie les fichiers pour corriger les erreurs et les valeurs manquantes. Ensuite, on transforme les informations importantes pour les analyser. On explore les données pour comprendre les tendances et les liens entre les variables et les profils des départs.

Ensuite, on crée des modèles prédictifs. On choisit des méthodes comme la régression logistique ou les forêts aléatoires. On entraîne et teste ces modèles pour mesurer leur performance. Les résultats nous aident à identifier les facteurs clés et à proposer des solutions pour retenir les employés.

Enfin, on résume tout dans un rapport détaillé et on présente les conclusions à l'entreprise HumanForYou. Une présentation orale explique nos choix et les stratégies proposées. L'objectif est de fournir des solutions pratiques et adaptées aux besoins de l'entreprise.

Traitement des données

4

Sources des données

Les données du projet proviennent de fichiers CSV fournis par les ressources humaines de HumanForYou. Elles ont été anonymisées pour protéger les informations personnelles des employés, ce qui est essentiel pour respecter leur vie privée.

Ces données comprennent des détails sur les employés comme leur âge, salaire, ancienneté, niveau d'éducation, performances, satisfaction au travail et horaires de bureau.

Les fichiers principaux sont :

- general_data.csv : il contient des infos basiques comme l'âge, le salaire et si l'employé a quitté l'entreprise en 2016.
- manager_survey_data.csv : il montre les évaluations des managers sur l'implication et la performance des employés.
- employee_survey_data.csv : il donne des insights sur la qualité de vie au travail et le bien-être des employés.
- in_out_time.zip : ce dossier contient les horaires des employés, utiles pour analyser leur gestion du temps et comprendre les départs.

Traitement des données

4

Préparation des données

Le prétraitement des données est une étape clé pour rendre les données brutes utilisables. On commence par explorer les fichiers pour comprendre leur structure et repérer les valeurs manquantes ou incohérentes.

Gestion des valeurs manquantes

Le prétraitement des données est une étape importante pour obtenir des résultats fiables. Les données brutes contiennent souvent des erreurs qui peuvent fausser l'analyse. Il faut donc les nettoyer et les préparer avant de les utiliser.

Un problème fréquent est les valeurs manquantes. Elles peuvent venir d'erreurs de collecte ou d'informations non remplies. Il existe trois types de valeurs manquantes :

- MCAR : l'absence est totalement aléatoire.
- MAR : l'absence est liée à d'autres variables.
- NMAR : l'absence dépend de la valeur elle-même.

Pour gérer les valeurs manquantes, voici quelques approches simples :

- Utiliser des valeurs comme la moyenne, la médiane ou le mode pour les données catégoriques.
- Prédire la valeur manquante avec des modèles de régression basés sur d'autres variables.
- Prendre une valeur au hasard parmi celles disponibles.

Aucune méthode n'est parfaite. Souvent, il faut en combiner plusieurs pour réduire l'impact des données manquantes et rendre les analyses plus fiables.

Source : Comment traiter les données manquantes en Data Science (https://mrmint.fr/donnees-manquantes-data-science)

Traitement des données

4

Analyser des données avec la data exploration

L'exploration des données est une étape clé pour comprendre les tendances et les liens entre les variables.

Voici quelques méthodes utiles :

Compter les valeurs uniques permet de voir à quelle fréquence elles apparaissent, surtout pour les données catégorielles. Calculer la variance aide à comprendre comment les valeurs sont dispersées.

L'analyse de Pareto (80/20) identifie les variables les plus influentes. L'analyse de corrélation montre les relations entre les variables, souvent visualisées avec des matrices ou des heatmaps.
La clusterisation regroupe les données en segments similaires pour repérer des profils communs.
Enfin, détecter les valeurs aberrantes (outliers) est important pour éviter qu'elles ne faussent les résultats.

Une exploration bien menée est cruciale pour des analyses solides et des modèles plus précis

Les modèles d'IA

Les modèles d'intelligence artificielle analysent les données et font des prédictions. Il en existe plusieurs types adaptés à des besoins différents. Par exemple, la régression linéaire et logistique prédisent des nombres ou des catégories. D'autres modèles comme les machines à vecteurs de support ou le clustering classent et regroupent les données. Le choix du modèle dépend des données et de l'objectif.

Régression linéaire : Elle trouve une relation entre une variable à prédire et d'autres variables. Par exemple, prédire le prix d'une maison en fonction de sa taille et de son âge.
Source: https://fr.linedata.com/les-principaux-algorithmes-de-regression-pour-lapprentissage-supervise

Régression logistique : Elle sert à classer des données en deux catégories. Par exemple, déterminer si un e-mail est un spam ou non.
Source: https://developers.google.com/machine-learning/crash-course/logistic-regression

Perceptron : C'est un modèle simple qui sépare les données en deux groupes. Il est à la base des réseaux de neurones plus complexes.
Source: https://www.datacamp.com/fr/blog/machine-learning-models-explained

SVM (Machines à vecteurs de support) :

Les SVM sont des algorithmes de classification qui cherchent à tracer une frontière optimale entre les classes. Cette frontière est choisie pour maximiser la distance avec les points les plus proches de chaque classe, ce qui aide à mieux généraliser sur de nouvelles données. Les SVM peuvent aussi gérer des problèmes non linéaires en utilisant des techniques appelées noyaux, qui transforment les données pour les rendre séparables.
Source: https://www.datacamp.com/fr/blog/machine-learning-models-explained

K-Nearest Neighbors (KNN) :

Le KNN est un algorithme simple utilisé pour la classification ou la régression. Pour classer un nouveau point, il regarde les K points les plus proches et lui attribue la classe la plus fréquente parmi eux. Le choix de K et la manière de mesurer les distances sont importants pour sa performance. Cependant, il peut être lent sur de grandes bases de données car il doit calculer les distances entre tous les points.
Source: https://www.datacamp.com/fr/blog/machine-learning-models-explained

Les modèles d'IA

K-Means Clustering :

Le K-Means est un algorithme qui regroupe les données en K clusters. Il commence par placer K centres au hasard, puis assigne chaque point au centre le plus proche. Ensuite, il recalcule les centres en fonction des points assignés et répète le processus jusqu'à ce que les centres ne bougent plus. Il est souvent utilisé pour des tâches comme la segmentation de clients ou la compression d'images.
Source: https://www.datacamp.com/fr/blog/machine-learning-models-explained

Classifieur bayésien :

Les classifieurs bayésiens, comme le classifieur naïf bayésien, utilisent les probabilités pour prédire la classe d'une observation. Ils supposent que les caractéristiques sont indépendantes, ce qui simplifie les calculs. Même si cette hypothèse n'est pas toujours réaliste, ces classifieurs fonctionnent bien dans des applications comme la classification de textes ou la détection de spams.
Source: https://www.datacamp.com/fr/blog/machine-learning-models-explained

L'IA avec Python

Introduction

Cette partie explique comment utiliser Python pour créer et faire fonctionner un modèle d'intelligence artificielle. On y présente les étapes clés, de la préparation des données à l'application des algorithmes de machine learning. L'idée est de montrer comment Python aide à construire et améliorer des modèles fiables.

Préparation et manipulation des données

La qualité des données est cruciale en machine learning. Avant de lancer l'entraînement, il faut nettoyer et préparer les données pour éviter les erreurs et obtenir de meilleurs résultats. Python offre des outils comme pandas pour gérer les fichiers et numpy pour les calculs.

Pour importer les données, on utilise souvent pandas.read_csv() pour charger des fichiers CSV. Cela permet de les manipuler facilement. Ensuite, il faut traiter les valeurs manquantes. On peut soit supprimer les lignes incomplètes si ce n'est pas grave, soit remplir les trous avec des méthodes simples comme la moyenne.

Pour éviter que certaines données prennent trop d'importance, on normalise les valeurs avec StandardScaler de scikit-learn. Cela permet de tout mettre à la même échelle.

Enfin, il faut vérifier qu'il n'y a pas de doublons ou d'erreurs dans les données pour s'assurer que tout est prêt avant de lancer l'entraînement du modèle.

Sources :
Manipulation et nettoyage des données avec Python – OpenClassrooms
Introduction à pandas pour la data science – DataCamp

L’IA avec Python

Visualisation des données

Avant d'entraîner un modèle, il faut bien comprendre les données. On peut utiliser des outils comme Matplotlib et Seaborn en Python. Ces bibliothèques servent à créer des graphiques pour repérer des tendances, voir les liens entre les variables ou détecter des anomalies.

Avec Matplotlib, on peut tracer des histogrammes des courbes ou des nuages de points pour visualiser la répartition des données et repérer d’éventuelles valeurs extrêmes. Seaborn propose des graphiques plus avancés comme des heatmaps pour voir les corrélations ou des boxplots pour identifier les valeurs anormales.

Sources :
https://moncoachdata.com/blog/matplotlib-visualisation-de-donnees/
https://seaborn.pydata.org/

Séparation des données en ensembles d'entraînement et de test

Une fois que les données sont prêtes et nettoyées, il faut les diviser en deux groupes :

- Un ensemble d’entraînement qui sert à apprendre au modèle à faire des prédictions.
- Un ensemble de test qui permet de vérifier si le modèle fonctionne bien sur des données qu’il n’a jamais vues.

Pour faire cette séparation, on utilise la fonction train_test_split de la bibliothèque scikit-learn. Elle permet de choisir un pourcentage des données pour l'entraînement (souvent 80 %) et le reste pour le test (20 %). C'est une étape essentielle pour éviter que le modèle apprenne "par cœur" les données et ne soit pas capable de bien s’adapter à de nouvelles situations.

Source : Documentation scikit-learn – Train/Test Split

L'IA avec Python

Implémentation des modèles de classification

Cette étape est essentielle pour bien comprendre les données et choisir le bon modèle.
Ensuite, selon le problème à résoudre, on peut utiliser différents modèles de machine learning avec Scikit-learn.

Voici quelques exemples :

- Régression logistique : permet de classer des données en deux groupes
- Perceptron : un modèle simple qui s'améliore à chaque erreur
- SVM (Support Vector Machines) : efficace pour séparer des catégories avec des frontières optimisées
- KNN (k-Nearest Neighbors) : classe un élément en fonction de ses voisins les plus proches
- Classificateur bayésien : utilise des probabilités pour faire des prédictions

Chaque modèle a ses points forts et ses limites. Pour choisir le bon, on utilise des indicateurs comme la précision ou le F1-score qui permettent de voir lequel marche le mieux avec nos données.

Sources : https://scikit-learn.org/stable/

Conclusion

Ce projet nous a permis de mieux comprendre comment l'intelligence artificielle peut aider à analyser les raisons du départ des employés chez HumanForYou.

En suivant une approche méthodique, nous avons d'abord nettoyé et préparé les données avant d'appliquer différents modèles prédictifs à l'aide d'outils Python. Cette analyse nous a permis d'identifier des tendances et de proposer des pistes d'amélioration pour aider l'entreprise à mieux fidéliser ses employés.

L'apprentissage automatique s'est révélé être un outil puissant pour approfondir l'étude et obtenir des résultats concrets. Ce travail met en évidence l'importance de l'IA dans la gestion des ressources humaines et son potentiel pour améliorer la prise de décision en entreprise.

Bibliographie

- https://fr.linedata.com/les-principaux-algorithmes-de-regression-pour-lapprentissage-supervise
- https://developers.google.com/machine-learning/crash-course/logistic-regression
- https://www.datacamp.com/fr/blog/machine-learning-models-explained
- Introduction à pandas pour la data science – DataCamp
- Manipulation et nettoyage des données avec Python – OpenClassrooms
- https://moncoachdata.com/blog/matplotlib-visualisation-de-donnees/
- https://seaborn.pydata.org/
- Documentation officielle de scikit-learn
- https://scikit-learn.org/stable/