# Introduction

Expliquer brièvement le problème

# Analyse préliminaire

# Analyse exploratoire

Dans cette section, vous rendrez compte de votre analyse exploratoire. Elle doit être structurée en deux parties :

## Attributs qualitatifs

## Combien d’instances et combien de variables (attributs) y a-t-il ?

Il y a 10k instances avec chacune 11 attributs

## Quelle est la variable cible et est-elle quantitative ou qualitative ?

La variable cible est une variable qualitative qui prend deux valeurs

* 0 signifie que la personne est restée dans l’entreprise
* 1 signifie que la personne est partie de l’entreprise

## Les autres attributs sont-ils quantitatifs ou qualitatifs ?

Fournissez un tableau dans lequel vous indiquez le type de

Chaque attribut (quantitatif ou qualitatif) et la raison de ce choix. et la raison de ce choix)

|  |  |  |  |
| --- | --- | --- | --- |
| **Nom de Variable** | **Type** | **Valeurs** | **Raisons** |
| satisfaction\_level | Variable quantitative | Une note comprise entre 0 et 1 |  |
| last\_evaluation | Variable quantitative | Une note comprise entre 0 et 1 |  |
| number\_project | Variable quantitative |  |  |
| average\_monthly\_hours | Variable quantitative |  |  |
| time\_spend\_company | Variable quantitative |  |  |
| work\_accident | Variable qualitative | 0 signifie que la personne n’a pas eu d’accident  1 signifie que la personne a eu un accident du travail |  |
| promotion\_last\_5years | Variable qualitative | 0 signifie que la personne n’a pas eu de promotion dans les 5 dernières années  1 signifie que la personne a eu une promotion dans les 5 dernières années |  |
| department | Variable qualitative | Accounting, Hr, It, Management, Marketing, Management, product\_mng, RandD, Sales, Support,  technical |  |
| salary | Variable qualitative | High  Medium  low |  |

## Certaines variables doivent-elles être supprimées de l’analyse ? Pourquoi ?

Id est une variable d’identification et n’est ni quantitatif ni qualitatif

## Y a-t-il des données manquantes ?

Il n’y aucune instance avec des données manquantes

## Fournir une description de la variable cible et de sa distribution

P(Left = 0) = 76.19%

P(Left = 1) = 23.81%

Cela signifie qu’un quart des employés décide de quitter l’entreprise en fonctions des variables explicatives

A graph of a number of squares

Description automatically generated

## Pour chaque attribut qualitative f :

* Calculer la distribution de probabilité P(f)
* Calculer la probabilité conditionnelle de la variable cible, y,
* Compte tenu des valeurs des attributs, P(y|f).
* Les visualiser à l’aide des utils appropriées (barplots).

*Voir code*

## Analyser les probabilités conditionnelles

* Existe-t-il des attributs qualitatifs utiles pour distinguer les différentes classes?
* Si oui, sélectionnez deux ou trois attributs qui, selon vous, sont les plus prédictifs de l’attribut cible y.
* Expliquez clairement votre raisonnement, comment chacun de ces attributs affecte la valeur de y.
* Dans le rapport, incluez uniquement les résultats et la discussion pour les deux ou trois attributs les plus importants que vous avez sélectionnés.

### Niveau de salaire

Les employés à bas salaire ont un taux de départ de 30% ce qui est très significatif en comparaison des 20% et 6% des salaires medium et haut. Les écarts entre ces taux de départs est donc respectivement. 10% et 24%.

De plus la population des salaires bas et médium est presque équivalente environ 45% chacun.

Ceci permet de confirmer que le niveau de salaire est un indicateur important du taux de départ dans l’entreprise. Plus le salaire est bas plus les départs sont fréquent.

## En plus des deux ou trois attributs les plus importants, sélectionnez et décrivez également un attribut qui n’affecte pas la variable cible. Expliquez clairement pourquoi cet attribut n’affecte pas la cible.

### Work Accident

Les employé qui n’ont pas eu d’accident (work\_accident = 0) ont un taux de depart de 26.5% (P(left = 1 | work\_accident = 0) alors que ceux qui ont eu un accident (work\_accident = 1) ont un taux de depart bien plus bas à 8%. Ainsi la différence de taux départ entre ceux qui ont eu un accident et ceux qui n’en n’ont pas eu est d’environ 18%. Ce qui est un écart tout même important comparé aux autres variables.

Cela suggère que ne pas avoir d’accident est corrélé avec un taux de depart plus élevé. Cependant la majorité des employés (85.41%) n’ont jamais eu d’accident ce qui rend cet attribut moins significatif que d’autres pour expliquer les departs.

### Promotion dans les 5 dernières années

Les employés qui n’ont pas reçu de promotion depuis 5 ans (promotion\_last\_years\_5years = 0) ont un taux de départ de près d’un quart (24.2%) alors que ceux ayant reçu une promotion ont un taux de départ de seulement 5.7%. L’écart entre ces taux de départ est donc de prêts de 20%.

Ceci indique donc qu’un manque de promotion durant 5 ans est significativement corrélé aux départs des employé.

Cependant 97.8% des employés du datasets sont des employés n’ayant pas reçu de promotions depuis 5 ans. Ainsi ces employés sont peut-être sur représentés et donc cette variable peut fausser la prédiction de départ d’un employé.

### Département

Le taux de départ dans le différent département varie mais ce sont les départements R&D et ceux du management qui ont les taux de départ les plus bas (respectivement 14.1% et 15.5%)

En contraste la comptabilité et les ventes ont un taux de départ de 27.3% et 25.1%.

Cependant la sur représentation des employés de ventes (>25%), ainsi que des écarts relatifs limités (max 13%), font de cette variable un piètre indicateurs du départ des employés

### Sélectionnez un attribut qualitatif, f, de votre choix, idéalement avec un petit nombre de valeurs distinctes.

Ecrire une petite fonction nommée QualitativeAttrsTheory qui prend en entrée le jeu de données, l’index de l’attribut y et l’index de la variable cible et :

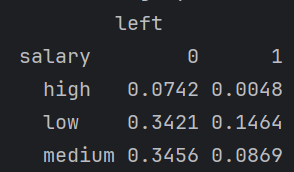
Établit la distribution conjointe P(f, y) où y est votre variable cible.

Utilise p(f, y) pour obtenir les distributions marginales p(f) et p(y).

Utilise p(f, y) et les distributions marginales p(f) et p(y) pour obtenir les distributions conditionnelles p(f|y) et p(y|f).

Pour les trois derniers points, expliquez clairement dans votre rapport comment vous passez de la distribution jointe à la distribution demandée en utilisant comme exemple les attributs que vous avez sélectionnés.

Tout d’abord, calcul la probabilités jointes des deux variable P(left, salary) ou left est la variable cible. Cette dernières nous donne un tableau dont la somme de toutes les valeurs égales à un.



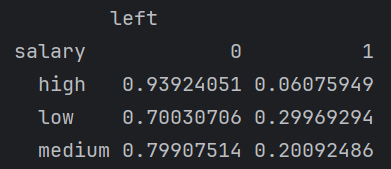
Par exemple dans le cas, du niveau de salaire, la probabilité qu’une personne avait un salaire bas et soit partie de l’entreprise est de 14,64% mais la probabilité qu’une personne ait un salaire bas ou medium et soit resté dans l’entreprise est d’environ 34% pour chacune des catéries de salaire,

Ensuite, on calcule les probabilités marginales de chacune des variables soit P(left) et P(salary). Si l’on observe le tableau des probabilité conjointe, on remarque que ces dernières peuvent être obtenu en sommant les valeurs de la bonne façon.

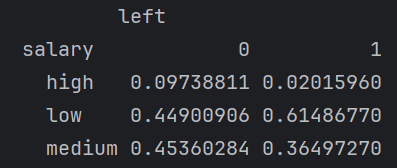
* Pour la variable cible left (en colonne dans le tableau des probabilités conjointes), si on somme l’ensemble des valeurs par colonnes on obtiendra P(left=0) = 76.19% et P(left=1)=23,81% indépendamment de P(salary)
* Pour la variable explicative salary (en ligne dans le tableau des probabilités conjointes), si on somme l’ensemble des valeurs par lignes on obtiendra P(salary=high)7,9%, P(salary=low)=48,85% et P(salary=medium)=43,2%5 indépendamment de P(left)

Enfin on calcule les distributions conditionnelles en utilisant les deux étapes précédentes.

* Si on divise chaque ligne de la probabilités conjointes par la probabilité marginale P(left) alors on obtient la probabilité conditionnelle P(target|salary). Cette probabilité représente la probabilité d’un depart d’un employé sachant sont niveau de salaire. Ainsi P(left=1|salary=high) = 6,07%



* Si on divise chaque colonne de la probabilités conjointes par la probabilité marginale P(salary) alors on obtient la probabilité conditionnelle P(salary|left). Cette probabilité représente la probabilité des différents niveau salaire sachant s’il est parti ou non de l’entreprise. Ainsi P(salary=high|left=1) = 2,07%



Utilisez les distributions que vous venez de calculer pour donner un exemple simple du théorème de Bayes.

P(Y|X) = P(X|Y)P(Y) / P(X) ou X est la variable cible et Y la variable explicative

Prenons par exemple

P(Y|X) = > P(salary=high|left=1) = 2,07%

P(X|Y) => P(left=1|salary=high) = 6,07%

P(X) => P(left=1)=23,81%

P(Y) => P(salary=high)7,9%

## Attributs quantitatifs

# Résumé

Résumé des principaux résultats et conclusions - factuel, clair et concis