Anthéa ABAH
Clément DROMER
Moïse NGOUALA
Maxim PEROLLE

MLOps

Analyse prédictive du salaire

Enseignant
Corentin VASSEUR

Master 2 SIAD (Parcours DS)
Année 2023 - 2024

Sommaire

Introduction

Le Centre d'études et de recherches sur les qualifications (CÉREQ) est un organisme public qui réalise des études dans les domaines du marché du travail, des qualifications et de la formation professionnelle.

Notre étude se base sur une enquête d'insertion de cet organisme. Celle-ci se concentre sur les jeunes sortis du système scolaire en 2013 et qui ont été interrogé trois ans après, en 2016, sur leur situation actuelle et leurs parcours passés.

Lors de cette enquête, 19 498 jeunes ont été interrogés. Parmi eux, 15 085 étaient en emploi au moment de l'enquête et 11 885 occupaient un emploi salarié à plein temps, il s'agit donc de notre population d'intérêt.

Pour des questions de confidentialité, nous n'avons pas accès à toute la base mais à une extraction aléatoire de 6 000 observations.

Grace à ces ressources, l'objectif est de mettre au point un modèle de Machine Learning prédictif de la capacité à obtenir un haut salaire pour un jeune arrivé sur le marché du travail depuis 3 ans. On considère qu'un jeune salarié touche un haut salaire s'il perçoit plus de 2 000 € net par mois.

Afin de répondre aux besoins de notre étude, nous allons dans un premier temps, procéder à une étude préalable des données afin de pourvoir repérer d'éventuelles données manquantes ou aberrantes, effectuer une description statistique des variables et étudier les liens existant entre celles-ci.

Par la suite, nous utiliserons divers algorithmes d'apprentissage automatique pour effectuer une modélisation sur notre jeu de données. Cette approche nous permettra d'obtenir des modèles prédictifs robustes basés sur les caractéristiques des individus.

Enfin, à partir de ces modélisations, nous sélectionnerons le modèle le plus performant qui correspondra aux attentes de l'étude.

Étude préalable des données

Description du fichier et vérification des variables

Liste et définition des variables

Nous disposons d'une base de données d'individus avec 6 000 observations qui recense :

SALPRSFIN ➔ Salaire mensuel net en fin de séquences (primes incluses) des salariés. (Il s'agit de la variable à expliquer)

Q1 ➔ Sexe de l'enquêté(e).

- Homme (1)
- Femme (2)

AGE13 ➔ âge de l'individu en 2013, à la sortie du système scolaire.

CFA ➔ Sortant de formation dispensée par voie d'apprentissage.

- Oui (1)
- Non (2)

CA9C ➔ Position professionnelle du père à la date de fin d'études.

- Ouvrier (1)
- Employé (2)
- Technicien, agent de maîtrise, VRP, profession intermédiaire (3)
- Cadre,ingénieur, profession libérale, professeur (4)
- Artisan, commerçant, chef d'entreprise (5)
- Agriculteur (6)
- NSP (7)
- Décédé (8)
- N'a jamais travaillé (9)

CA10C ➔ Position professionnelle de la mère à la date de fin d'études.

- Ouvrière (1)
- Employée (2)
- Technicienne, agent de maîtrise, VRP, profession intermédiaire (3)
- Cadre, ingénieure, profession libérale, professeur (4)
- Artisan, commerçant, chef d'entreprise (5)
- Agricultrice (6)
- NSP (7)
- Décédée(8)
- N'a jamais travaillée (9)

CA11 ➔ Niveau d'études du père.

- Sans diplôme, certificat d'études ou brevet des collèges (1)
- CAP BEP (2)
- Baccalauréat (3)
- Bac +2 (4)
- Bac +3, Bac +4 (5)
- Bac +5 ou plus (6)
- NSP (7)

CA12 ➔ Niveau d'études de la mère.

- Sans diplôme, certificat d'études ou brevet des collèges (1)
- CAP BEP (2)
- Baccalauréat (3)
- Bac +2 (4)
- Bac +3, Bac +4 (5)
- Bac +5 ou plus (6)
- NSP (7)

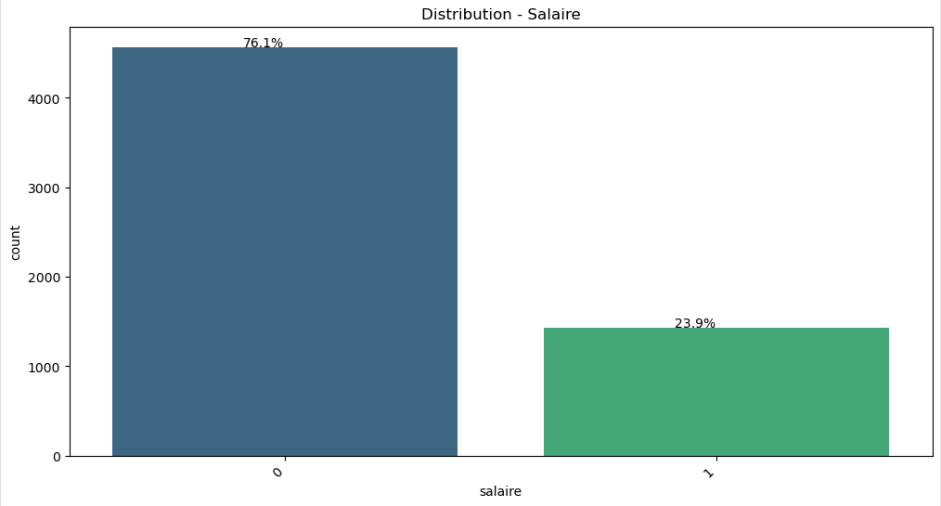
PHD ➔ Le plus haut diplôme obtenu en 2013

- 01 : NON DIPLOME
- 02I : CAP-BEP-MC INDUSTRIEL
- 02T : CAP-BEP-MC TERTIAIRE
- 03I : BAC PRO-BT-BP INDUSTRIEL
- 03T : BAC PRO-BT-BP TERTIAIRE
- 04I : BAC TECHNO INDUSTRIEL
- 04T : BAC TECHNO TERTIAIRE
- 05 : BAC GENERAL
- 06I : BTS-DUT INDUSTRIEL
- 06T : BTS-DUT TERTIAIRE
- 07I : AUTRE BAC+2 INDUSTRIEL
- 07T : AUTRE BAC+2 TERTIAIRE
- 08 : BAC+2/3 SANTE SOCIAL
- 09L : LICENCE PRO LSH GESTION DROIT
- 09M : LICENCE PRO MATHS SCIENCE TECHNIQUE
- 10L : L3 LSH GESTION DROIT
- 10M : L3 MATHS SCIENCE TECHNIQUE
- 11L : AUTRE BAC+3 LSH GESTION DROIT
- 11M : AUTRE BAC+3 MATHS SCIENCE TECHNIQUE
- 12L : BAC+4 LSH GESTION DROIT
- 12M : BAC+4 MATHS SCIENCE TECHNIQUE
- 13L : M2 LSH GESTION DROIT
- 13M : M2 MATHS SCIENCE TECHNIQUE
- 14L : AUTRE BAC+5 LSH GESTION DROIT
- 14M : AUTRE BAC+5 MATHS SCIENCE TECHNIQUE
- 15 : BAC+5 ECOLE DE COMMERCE
- 16 : INGENIEUR
- 17 : DOCTORAT SANTE
- 18L : DOCTORAT HORS SANTE LSH GESTION DROIT
- 18M : DOCTORAT HORS SANTE MATHS SCIENCE TECHNIQUE

Analyse de la distributions des variables

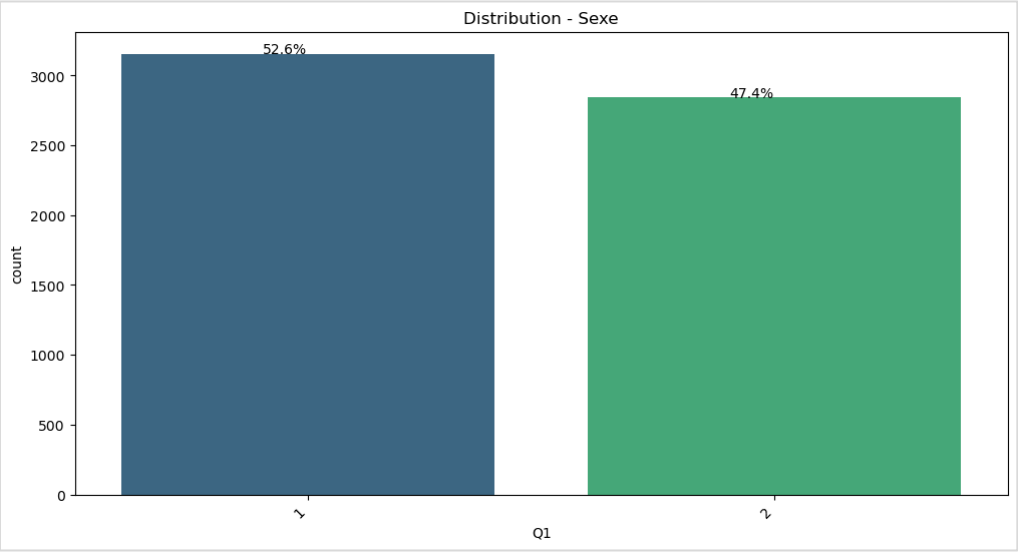
Pour toutes nos variables, nous avons représenté des graphiques pour visualiser nos données et étudier la répartition au sein de chaque variable, permettant ainsi d'avoir une meilleure compréhension de nos données.

Graphique 1 : Diagramme à barres représentant la répartition des individus en fonction du salaire



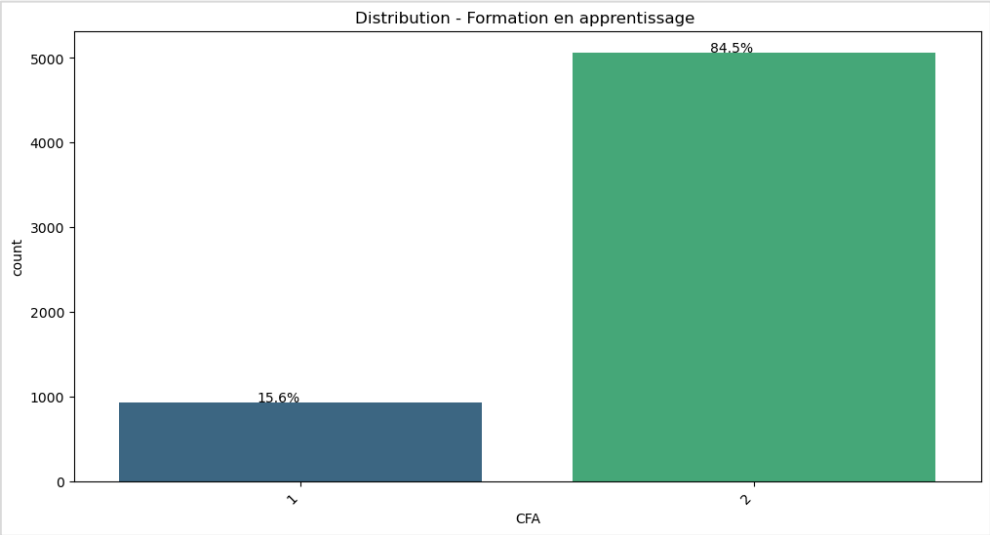
Sur ce graphique, on observe une distribution marquée avec 76,1% des individus percevant un salaire égal ou inférieur à 2000, tandis que 23,9% dépassent ce seuil. Cette répartition souligne une concentration importante de revenus dans la tranche inférieure.

Graphique 2 : Diagrammes à barres des individus la répartition des individus en fonction du sexe



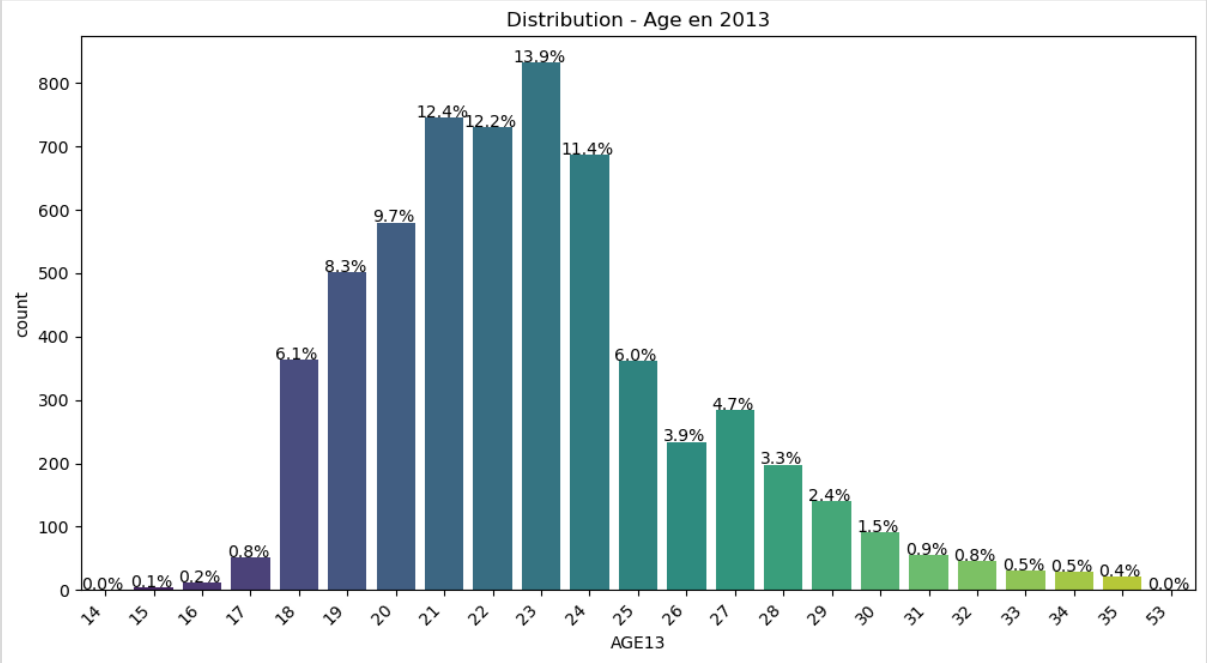
Nous constatons sur ce graphique que la très grande majorité des individus sont des Hommes (52,6%) tandis que les Femmes représentent 47,4%

Graphique 3 : Diagrammes à barres illustrant la répartition des individus selon le cursus en formation d'apprentissage



Nous constatons sur ce graphique qu'une grande partie des individus n'ont pas suivi de formation en apprentissage (84,5%) tandis que ceux ayant suivi une formation en apprentissage représentent 15,6%

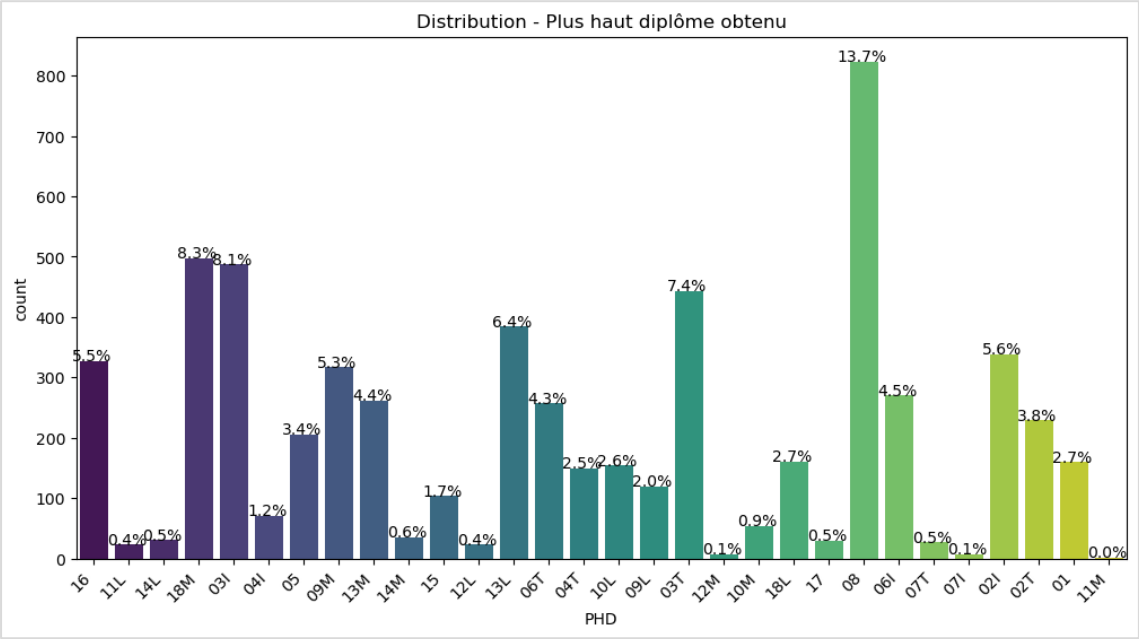
Graphique 4 : Diagramme à barre représentant la répartition des individus par rapport à l'âge



Sur ce graphique on constate une concentration importante des individus ayant entre 18 et 25 ans avec un pic marqué à 23 ans. Au-delà de cette âge, la fréquence des

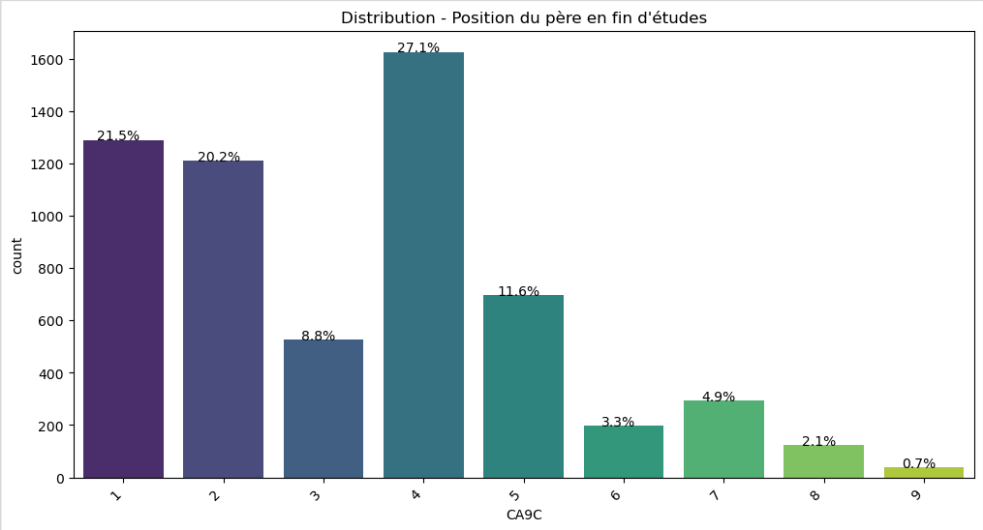
observations commence à décroître, suggérant une diminution du nombre d'individus dans ces tranches d'âge en sortie d'étude.

Graphique 5 : Diagramme à barres de la répartition des individus en fonction du plus haut diplôme obtenu



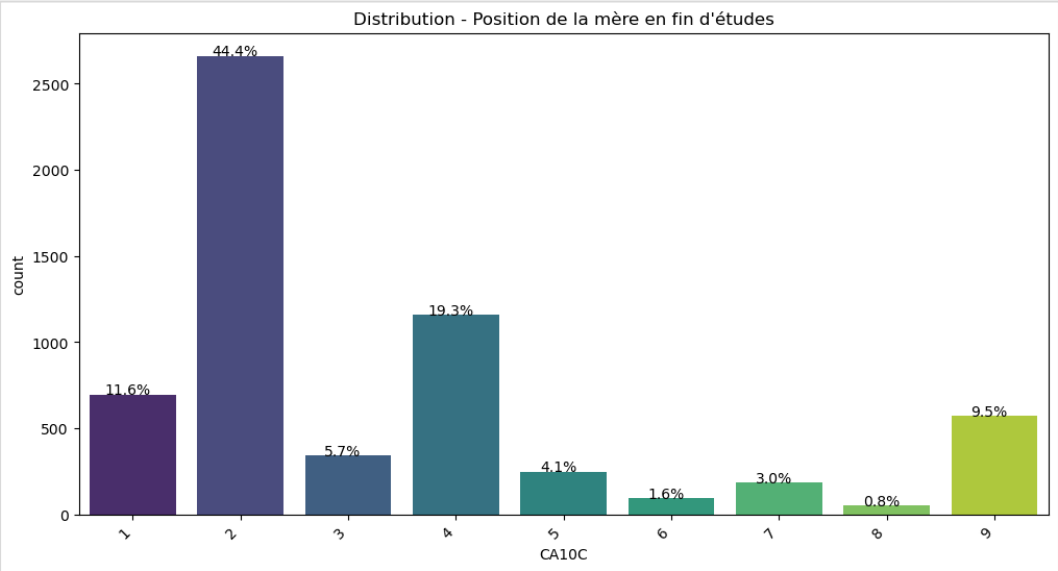
Sur ce graphique, on peut voir que la catégorie la plus représentée est *BAC+2/3 SANTE SOCIAL* (08) avec une fréquence de 13,7%. On observe également des concentrations importante dans les *catégories DOCTORAT HORS SANTE MATHS SCIENCE TECHNIQUE* (18M) et *BAC PRO-BT-BP INDUSTRIEL* (03I) avec 5,5%. D'autres niveaux d'éducation, tels que *BAC+4 MATHS SCIENCE TECHNIQUE* (12M) sont quant à eux moins présents. La présence de diverses catégories montre la diversité au niveau plus haut diplôme obtenu chez les individus.

Graphique 6 : Diagramme à barres de la répartition des individus en fonction de la position du père en fin d'études



On constate sur ce graphique que la catégorie 4 (*Cadre, ingénieur, profession libérale, professeur*) est la plus fréquente à 27,1%. En revanche, la catégorie 9 (N'a jamais travaillé) est la moins représentée, ne constituant que 0,7% des individus. Ces diverses catégories montrent les différents niveau d'études du pères pour les individus de notre échantillon de données.

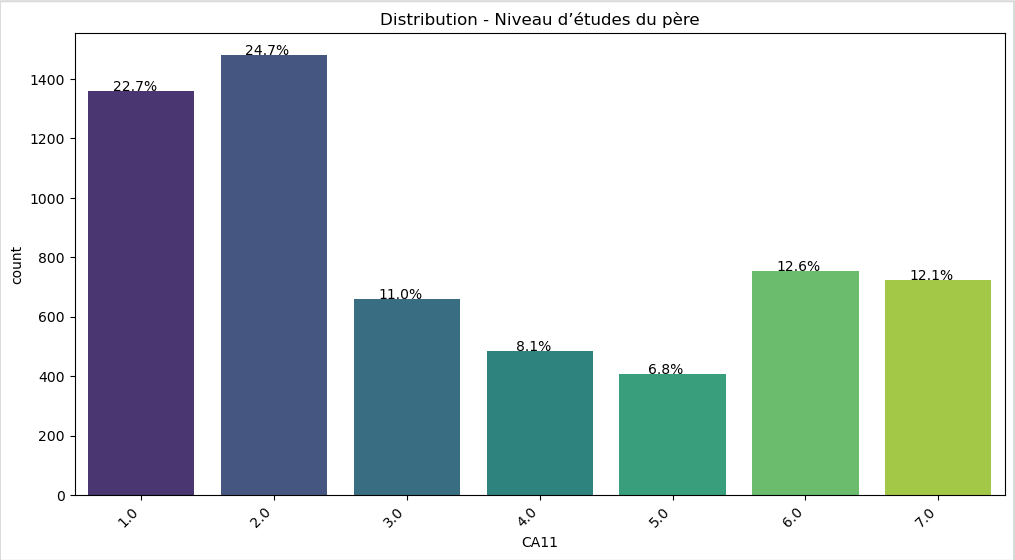
Graphique 7 : Diagramme à barres de la répartition des individus en fonction de la position du mère en fin d'études



Sur ce graphique, on peut voir que la catégorie 2 (Artisan, commerçant, chef d'entreprise) est la plus représentée avec 44,4%, soulignant une forte influence des

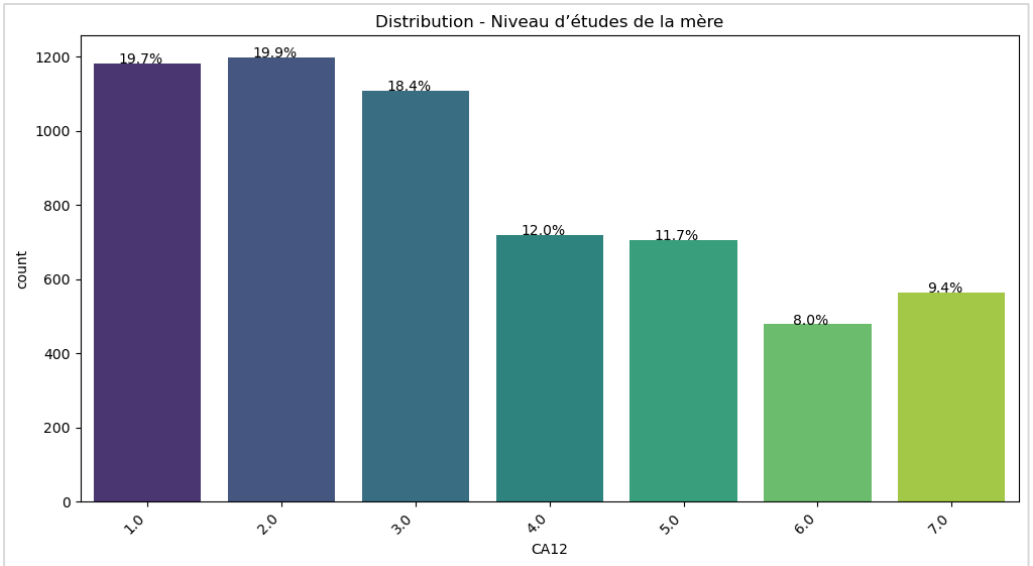
professions entrepreneuriales. En revanche, la catégorie 8 (Décédée) est la moins fréquente à 0,8%. Toutes ces différentes catégories suggèrent des contextes familiaux différents, susceptibles de jouer un rôle important dans les parcours éducatifs des individus.

Graphique 8 : Diagramme à barres de la répartition des individus en fonction du niveau d'études du père



On observe sur ce graphique que les catégories les plus représentées sont *CAP-BEP* (2) à 24,7% et l'absence de diplôme, certificat d'études ou brevet des collèges (1) à 22,7%. À contrario les catégories Bac +2 (4) et Bac +3, Bac +4 (5) sont les moins présentes.

Graphique 9 : Diagramme à barres de la répartition des individus en fonction du niveau d'études de la mère



Sur ce graphique, les catégories 1 (Sans diplôme, certificat d'études ou brevet des collèges) et 2 (CAP BEP) sont les plus fréquentes, représentant respectivement 19,7% et 19,9%. Les proportions relativement équilibrées observées dans les catégories 3 à 7 (de 8% à 18,4%) ce qui montre diversité des niveaux d'études chez la mère.

Les données manquantes

Nous avons parcouru notre base de données pour repérer les données manquantes. On en dénombre 174 :

Niveau d'étude du père (CA11) →124 données manquantes

Niveau d'étude de la mère (CA12) → 50 données manquantes

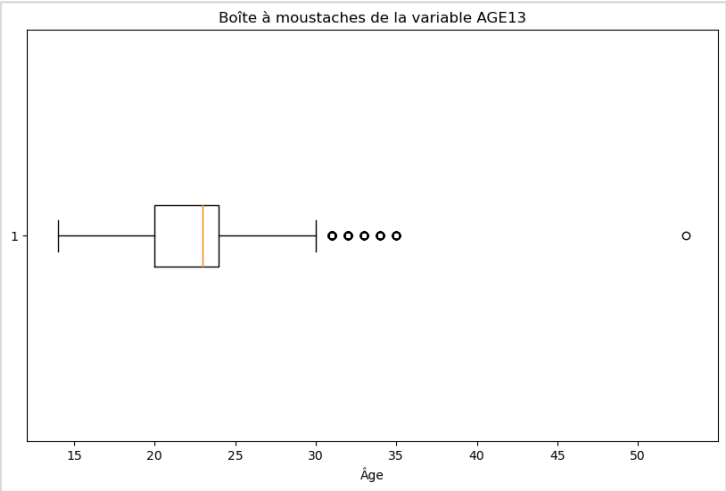
Nous avons décidé d'imputer ces valeurs manquantes avec le KNN (K-Nearest Neighbors) Imputer puisque que les observations similaires auront des valeurs similaires pour la variable à imputer. L'algorithme fonctionne en identifiant les k voisins les plus proches d'une observation avec des valeurs manquantes, en se basant sur les autres variables présentes. Ensuite, il utilise les valeurs observées dans ces voisins pour calculer une valeur imputée pour l'observation manquante (ici nous avons fixé k = 3).

Les données aberrantes

Pour détecter les valeurs aberrantes au niveau de l'âge nous avons utilisé une boite à moustache pour détecter les valeurs qui ne se trouvent pas dans cette boîte.

→ [`Q1 - 1.5 x écart interquartile ; Q3 + 1.5 x écart interquartile`] avec `Q1`, le premier quartile (25%), `Q3` le troisième quartile (75%) et `l'écart interquartile` qui correspond à la différence entre `Q3` et `Q1`.

Graphique 10 : Boîte à moustache de l'âge des individus en sortie d'études



Pour les valeurs qui ne se trouvent pas dans cette boite, nous avons décidé de les retirer de notre base car si l'on considère les durées des programmes d'études supérieures, la présence de telles valeurs pourrait indiquer une erreur de saisie ou de collecte des données.

Regroupement de modalités

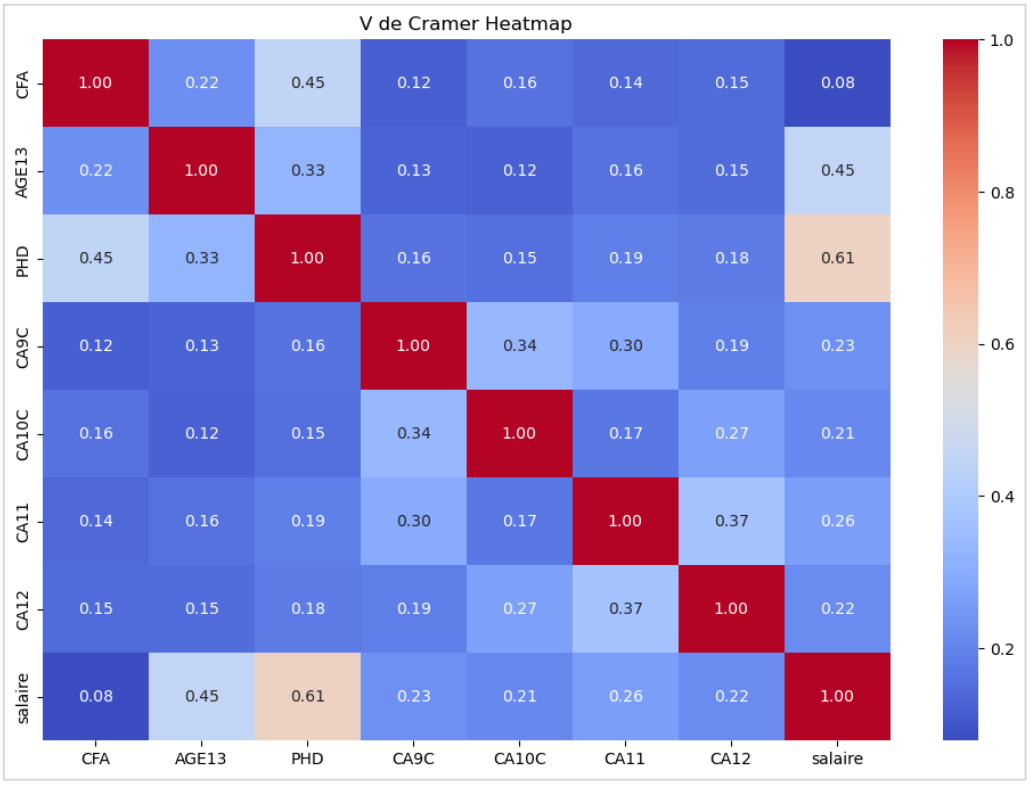
Nous avons effectué des regroupements au niveau de la variable PHD pour les diplômes similaires (selon le niveau) dans le but de simplifier et l'analyse des données.

Statistiques descriptives préalables à la modélisation

Analyse bivariée

Pour la suite de notre étude, nous avons effectué une analyse bivariée pour explorer les liens entre les différentes caractéristiques des individus.
Pour cela nous avons utilisé le V de Cramer pour mesurer le degré d'association entre les variables.

Graphique 11 : Matrice des V de Cramer entre les variables



Plusieurs variables sont fortement liées au salaire. On constate que le salaire et le niveau d'étude ont ainsi une forte influence, tout comme à égalité la position professionnelle de chaque parent.

On remarque aussi que les parents des individus ont un profil assez homogène avec plus de 30% de corrélation pour le niveau d'étude et la position professionnelle entre chaque parent.

Modélisation

Partitionnement du jeu de données

Tout d'abord, pour la modélisation, la division du jeu de données en ensembles d'entraînement (train) et de test est une étape primordiale dans le processus de construction et d'évaluation des modèles d'apprentissage automatique.

L'ensemble d'entraînement, est souvent constitué de 70 à 80% du jeu de données total. Il est utilisé pour ajuster les paramètres du modèle, permettant à celui-ci d'apprendre les relations sous-jacentes entre les variables explicatives et la variable cible.

L'ensemble de test, est généralement constitué des 20 à 30% restants du jeu de données. Il sert à évaluer la performance du modèle sur des données qui n'ont pas été utilisée lors de la phase d'apprentissage. Cela permet donc d'évaluer la capacité du modèle à généraliser (et que le modèle ne soit pas simplement « sur-ajusté » aux données d'entraînement).

Pour garantir que cette division soit représentative, nous utilisons une technique appelée stratification.

Lorsque nous divisons notre jeu de données en ensembles d'entraînement et de test, la stratification assure une répartition équilibrée des différentes classes de notre variable cible. Cela est essentiel, surtout lorsqu'il existe des classes sous représentées dans l'ensemble de données global. Ainsi, la stratification contribue à prévenir de tout biais potentiel et à assurer une évaluation plus fiable de la performance du modèle.

→ Pour notre modélisation, nous avons décidé de prendre 80% du jeu de données en apprentissage et 20 % en test.

Les métriques d'évaluation

L'évaluation des performances d'un modèle constitue une étape cruciale dans le processus de développement de tout algorithme d'apprentissage automatique. C'est à travers l'évaluation que nous mesurons la capacité d'un modèle à généraliser ses prédictions à de nouvelles données.

De plus cela permet d'ajuster ses paramètres, d'optimiser son efficacité, et d'assurer une utilisation fiable dans des situations réelles.

Des métriques permettent de comparer différents modèles et de choisir celle qui répond le mieux aux exigences spécifiques du problème.

Nous avons donc choisi les métriques suivantes :

Accuracy (exactitude)

- L'accuracy mesure le pourcentage de prédictions correctes parmi toutes les prédictions.
- Formule : (Vraies prédictions) / (Nombre total de prédictions)
- L'accuracy donne une vision globale de la performance du modèle.

Precision (précision)

- La précision mesure la proportion de vrais positifs parmi les prédictions positives du modèle.
- Formule : Vrais positifs / (Vrais positifs + Faux positifs)
- La précision est particulièrement utile lorsque le coût des faux positifs est élevé.

Recall (rappel)

- Le rappel mesure la proportion de vrais positifs parmi toutes les instances réellement positives.
- Formule : Vrais positifs / (Vrais positifs + Faux négatifs)
- Le rappel est important lorsque le coût des faux négatifs est élevé.

F1-Score (score F1)

- Le F1-score est la moyenne harmonique de la précision et du rappel. Il est utile lorsque l'on veut trouver un équilibre entre la précision et le rappel.
- Formule : 2 * (Précision * Rappel) / (Précision + Rappel)

Les modèles

La pipeline

Dans ce processus de modélisation, nous avons défini initialement plusieurs modèles, dont l'arbre de décision, les k-plus proches voisins (KNN), la régression logistique et la forêt aléatoire. Chacun de ces modèles est incorporé dans une pipeline, une séquence d'étapes de traitement des données et de modélisation.

La pipeline est construite différemment en fonction du modèle. Pour les modèles tels que KNN et la régression logistique, nous intégrons une étape de prétraitement, standardisation à l'aide de `StandardScaler()` afin d'avoir une distribution normale.

Le processus se poursuit avec la définition d'une grille d'hyperparamètres pour chaque modèle, spécifiant les combinaisons d'hyperparamètres à explorer lors de la recherche sur grille. Une fois la grille prête, nous lançons la recherche sur grille couplée à une validation croisée sur l'ensemble d'entraînement.

Chaque modèle est entraîné et évalué sur plusieurs configurations d'hyperparamètres et pour différentes partitions de l'ensemble d'entraînement, grâce à la pipeline qui assure le bon fonctionnement du processus. Une fois la recherche sur grille complétée, les meilleurs hyperparamètres sont identifiés pour chaque modèle.

Enfin, nous utilisons ces meilleurs hyperparamètres pour entraîner le modèle final sur l'ensemble d'entraînement et évaluons ses performances sur l'ensemble de tests et voir les résultats obtenus.

K-plus proches voisins (KNeighbors)

Le K-Nearest Neighbors (KNN) est un modèle d'apprentissage supervisé basé sur l'idée que des points de données similaires devraient avoir des étiquettes similaires. Ses points forts incluent sa simplicité et sa capacité à bien fonctionner sur des données non linéaires. Cependant, il peut être sensible aux valeurs aberrantes et nécessite souvent une mise à l'échelle pour des performances optimales.

Meilleurs paramètres pour K-Nearest Neighbors :

- Nombre de voisins (n_neighbors) : 7
- Poids utilisés pour la prédiction (model_weights) : uniform

Performance du modèle K-Nearest Neighbors sur l'ensemble de test :

- Accuracy : 80,42%

	Precision	Recall	F1-scoce	Support
0	0,85	0,90	0,88	871
1	0,60	0,48	0,53	263

Régression logistique (LogisticRegression)

La Régression Logistique est un modèle linéaire utilisé pour la classification. Son avantage majeur réside dans son interprétabilité et sa capacité à fournir des probabilités pour les différentes classes. Cependant, il suppose une relation linéaire entre les caractéristiques et la variable cible, ce qui peut être une limitation dans des cas plus complexes.

Meilleurs paramètres pour la Régression Logistique :

- Coefficient de de régularisation (model__C) : 0,1
- Pénalité (penalty) : l2 (Ridge Regression)

Performance du modèle de Régression Logistique sur l'ensemble de test :

- Accuracy : 81,57 %

	Precision	Recall	F1-scoce	Support
0	0,83	0,95	0,89	871
1	0,70	0,36	0,48	263

Arbre de décision (DecisionTree)

Le modèle d'arbre de décision est une technique d'apprentissage supervisé utilisée à des fins de classification et de régression. Ses points forts résident dans sa capacité à gérer des ensembles de données complexes et à fournir une interprétation facilement compréhensible. Cependant, il peut être sensible aux variations dans les données d'entraînement et peut être sujet au sur-apprentissage .

Meilleurs paramètres pour DecisionTree :

- Critère de division (model__criterion) : gini
- Profondeur maximale de l'arbre (model__max_depth) : 5

Performance du modèle DecisionTree sur l'ensemble de test :

- Accuracy : 85,78%

	Precision	Recall	F1-scoce	Support
0	0,88	0,94	0,91	871
1	0,73	0,56	0,63	263

Forêt aléatoire (RandomForest)

Le Random Forest est un modèle d'ensemble qui combine les prédictions de plusieurs arbres de décision pour améliorer la robustesse et la généralisation. Ses avantages incluent une réduction du sur-apprentissage par rapport à un seul arbre de décision et une meilleure performance sur des ensembles de données complexes. Cependant, il peut être plus complexe à interpréter.

Meilleurs paramètres pour RandomForest :

- Critère de division (model__criterion) : entropy
- Profondeur maximale de chaque arbre (model__max_depth) : 10
- Nombre d'arbres dans l'ensemble (model__n_estimators) : 100

Performance du modèle RandomForest sur l'ensemble de test :

- Accuracy : 84,3%

	Precision	Recall	F1-scoce	Support
0	0,88	0,92	0,90	871
1	0,69	0,60	0,64	263

Choix du meilleur modèle – Random Forest

Les raisons

Dans le cas présent, le modèle de Forêt Aléatoire (Random Forest) se démarque comme le meilleur choix pour plusieurs raisons :

- **Accuracy élevée** : Le Random Forest a démontré la plus haute précision parmi les modèles évalués, atteignant une valeur de 84,3 % sur l'ensemble de test. Cette précision élevée indique la capacité du modèle à classer correctement les individus dans les catégories cibles.
- **Réduction du sur apprentissage** : Le Random Forest, en tant que modèle d'ensemble, combine les prédictions de plusieurs arbres de décision. Cette approche contribue à réduire le risque de surapprentissage par rapport à un seul arbre de décision, ce qui signifie que le modèle est moins susceptible de s'adapter excessivement aux particularités des données d'entraînement et peut mieux généraliser à de nouvelles données.
- **Performance robuste** : Les hyperparamètres optimaux identifiés lors de la recherche sur grille, tels que le critère de division *entropy*, une profondeur maximale de chaque arbre de *10*, et *100* arbres dans l'ensemble, ont été déterminants pour la performance robuste du modèle sur l'ensemble de test.
- **Adaptabilité à la complexité des données** : Le Random Forest est efficace dans la gestion de jeux de données complexes en raison de sa capacité à agréger les décisions de multiples arbres. Cela fonctionne donc bien pour notre jeu de donnée (notamment les classes sous représentées)

Conclusion

Pour conclure, le but de notre étude consistait à réaliser une analyse prédictive des salaires.

Après avoir effectué l'exploration des données, permettant ainsi une meilleure compréhension de l'ensemble des individus, nous avons effectué la modélisation avec différents modèles d'apprentissage automatique.

Cela nous a permis d'identifier le meilleur modèle pour notre étude à savoir le *Random Forest* en raison de sa précision élevée de 84,3%, de sa capacité à réduire le surapprentissage grâce à son approche d'ensemble, de ses hyperparamètres optimisés, et de son adaptabilité à la complexité des données.