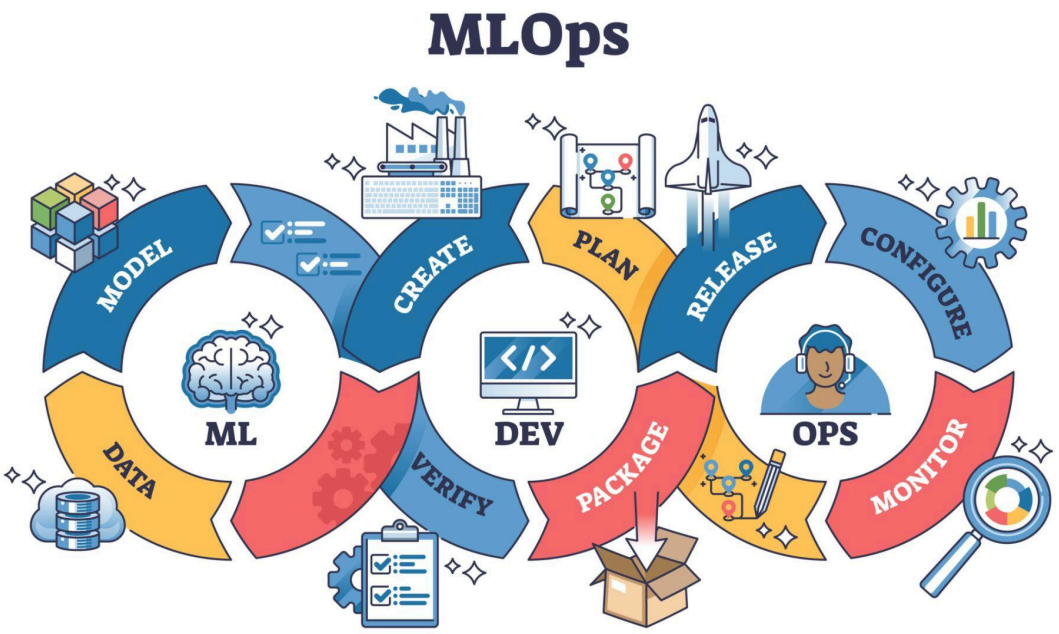
Sujet: Analyse exploratoire, modélisation prédictive et automatisation MLOps pour la prédiction du niveau de revenu



Rédigé par:

ADAM MOUSSA

SARA ABDI
MAMADOU DIALLO
DJAMILA BENBAHLOULI

2025-2026

Tables des matières

1. Introduction

La prédiction du niveau de revenu d'un individu à partir de caractéristiques socio-démographiques et professionnelles constitue un problème classique de la data science, largement utilisé pour illustrer les enjeux de la classification supervisée. Ce type de problématique présente un intérêt particulier, tant du point de vue méthodologique que pratique, en raison de la diversité des variables impliquées, de la présence de données hétérogènes et du déséquilibre fréquent des classes.

Dans ce projet, l'objectif est de prédire si le revenu annuel d'un individu est **inférieur ou égal à 50 000 dollars** ou **supérieur à ce seuil**, à partir d'un ensemble de variables décrivant son âge, son niveau d'éducation, sa situation professionnelle, son statut marital et d'autres caractéristiques socio-économiques. Les données utilisées proviennent du jeu de données *revenus.csv*, enrichi et transformé au fil des étapes de nettoyage et de préparation pour aboutir à des jeux de données adaptés à l'analyse et à la modélisation.

Au-delà de la simple construction d'un modèle prédictif, ce projet vise à mettre en œuvre une démarche complète de data science intégrant les principes du **MLOps**. Une attention particulière est ainsi portée à l'exploration des données, au traitement des valeurs manquantes et aberrantes, à la sélection des variables pertinentes ainsi qu'à la mise en place de pipelines reproductibles et automatisés.

Le présent rapport est structuré en plusieurs étapes. Il débute par une analyse exploratoire approfondie des données, permettant de comprendre leur structure et d'identifier les facteurs potentiellement explicatifs du niveau de revenu. Il se poursuit par une phase de pré-traitement visant à nettoyer et transformer les données. Enfin, les étapes de modélisation et d'automatisation sont présentées, avec pour objectif de construire un modèle fiable, interprétable et prêt à être intégré dans un pipeline opérationnel.

2. Étude préalable des données

Les données utilisées proviennent du fichier `revenus.csv`.
Le jeu de données contient **48 842 observations** et **15 variables**, réparties comme suit :

- **Variables numériques** :
  `age`, `fnlwgt`, `educational-num`, `capital-gain`, `capital-loss`, `hours-per-week`

- **Variables catégorielles** :
  `workclass`, `education`, `marital-status`, `occupation`, `relationship`, `race`, `gender`, `native-country`

- **Variable cible** :
  `income`

Ce jeu de données combine des informations sociales, éducatives et professionnelles, ce qui en fait un cas typique de problème de classification supervisée en data science.

3. Analyse des données manquantes

L'analyse de la qualité des données révèle la présence de valeurs manquantes codées par le caractère **"?"**, principalement dans les variables suivantes :

- `workclass` : 2 799 valeurs (5,73 %)

- `occupation` : 2 809 valeurs (5,75 %)

- `native-country` : 857 valeurs (1,75 %)

Au total, **6 465 valeurs manquantes** sont observées.
Ces valeurs manquantes concernent essentiellement des variables socio-professionnelles et ne semblent pas résulter d'un problème systémique de collecte. Elles ont été traitées lors de la phase de pré-traitement afin de conserver un maximum d'informations.

4. Analyse de la variable cible

La variable cible `income` est une variable catégorielle binaire indiquant si le revenu annuel d'un individu est inférieur ou égal à 50 000 dollars ou supérieur à ce seuil.

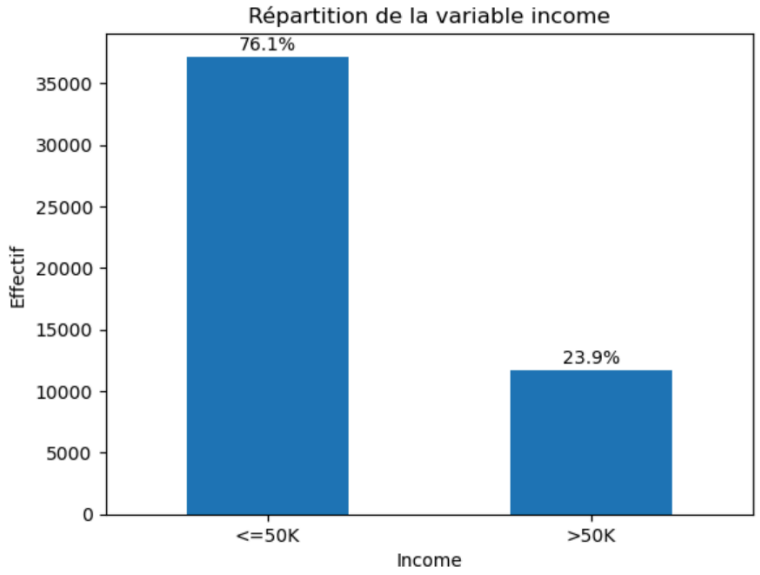


Figure 1 : Répartition des classes de la variable cible *income*

La variable cible *income* présente un déséquilibre de classes, avec une majorité d'individus percevant un revenu inférieur ou égal à 50K.

Ce déséquilibre devra être pris en compte lors de la phase de modélisation afin d'évaluer correctement les performances des modèles.

5. Analyse des variables numériques et valeurs aberrantes

Les statistiques descriptives mettent en évidence plusieurs caractéristiques importantes :

- **Âge** :
  moyenne ≈ 38,6 ans, médiane ≈ 37 ans, valeurs comprises entre 17 et 90 ans.
  La population est majoritairement composée d'individus en âge d'activité.

- **Heures travaillées par semaine** :
  moyenne ≈ 40,4 heures, médiane = 40 heures.
  Des valeurs extrêmes sont observées jusqu'à 99 heures.

- **Capital-gain et capital-loss** :
  distributions fortement asymétriques, avec une majorité de valeurs nulles et quelques valeurs très élevées.

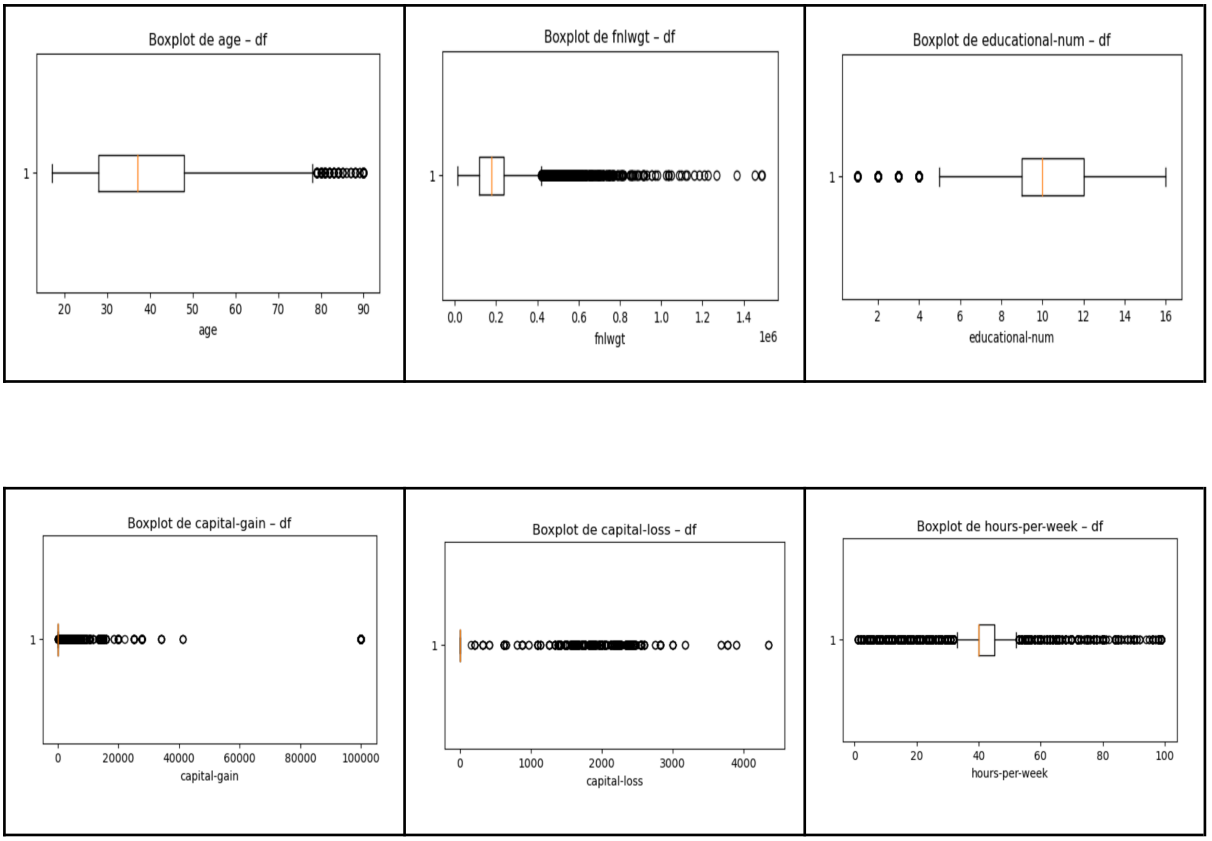


Figure 2 : Boxplots des principales variables numériques

L'analyse des statistiques descriptives générales met en évidence plusieurs éléments importants.
L'âge moyen des individus est d'environ 38,6 ans, avec une médiane proche (37 ans), ce qui indique une population relativement jeune et centrée autour de l'âge actif. Les valeurs minimales et maximales (17 à 90 ans) montrent toutefois une certaine hétérogénéité.

La variable *hours-per-week* présente une moyenne d'environ 40 heures, avec une médiane identique, traduisant un temps de travail hebdomadaire standard pour la majorité des individus. Néanmoins, des valeurs extrêmes sont observées, pouvant aller jusqu'à 99 heures par semaine.

Les variables *capital-gain* et *capital-loss* sont fortement asymétriques. La majorité des individus présente des valeurs nulles, tandis que quelques observations très élevées expliquent les valeurs maximales importantes. Ces variables nécessiteront une attention particulière lors du pré-traitement afin de limiter l'influence des valeurs extrêmes sur la modélisation.

Enfin, la variable *education-num* montre une dispersion modérée autour d'une moyenne proche de 10, traduisant des niveaux d'éducation relativement variés au sein de la population.

6. Analyse exploratoire des variables catégorielles

L'analyse des variables catégorielles met en évidence plusieurs déséquilibres structurels :

- **Genre** :
  66,85 % d'hommes contre 33,15 % de femmes.

- **Race** :
  forte dominance de la catégorie *White* (85,5 %).

- **Pays d'origine** :
  United-States représente près de 90 % des observations.

- **Niveau d'éducation** :
  concentration sur *HS-grad* et *Some-college*.

- **Occupation** :
  forte hétérogénéité, avec une dominance des catégories *White Collar*, *Blue Collar* et *Sales*.

Ces déséquilibres justifient des regroupements de modalités afin de limiter la sparsité et d'améliorer la robustesse des modèles.

7. Pré-traitement des données

Le pré-traitement a été réalisé à l'aide d'une fonction dédiée afin de garantir la reproductibilité du pipeline.

7.1 Nettoyage et harmonisation

- Suppression de la variable `fnlwgt`.

- Harmonisation des noms de colonnes (remplacement des tirets par des underscores).

- Remplacement des valeurs `"?"` par la modalité **"Non renseigné"**.

7.2 Transformation des variables

- **Âge** : discrétisation en classes d'âge.

- **Capital-gain / capital-loss** : transformation binaire (présence / absence).
- **Heures travaillées** : regroupement en trois catégories (*under*, *normal*, *over employed*).
- **Pays d'origine** : regroupement en *USA* / *Not USA*.

7.3 Regroupement des modalités

Plusieurs variables catégorielles ont été regroupées afin de réduire la cardinalité :

- `education` → niveaux d'éducation cohérents,
- `occupation` → familles de métiers,
- `workclass` → catégories institutionnelles,
- `marital_status` et `relationship` → statuts familiaux simplifiés,
- `race` → *White*, *Black*, *Other*.

À l'issue de cette étape, le jeu de données nettoyé contient **14 variables**.

8. Étude des relations entre variables

8.1 Matrice du V de Cramér

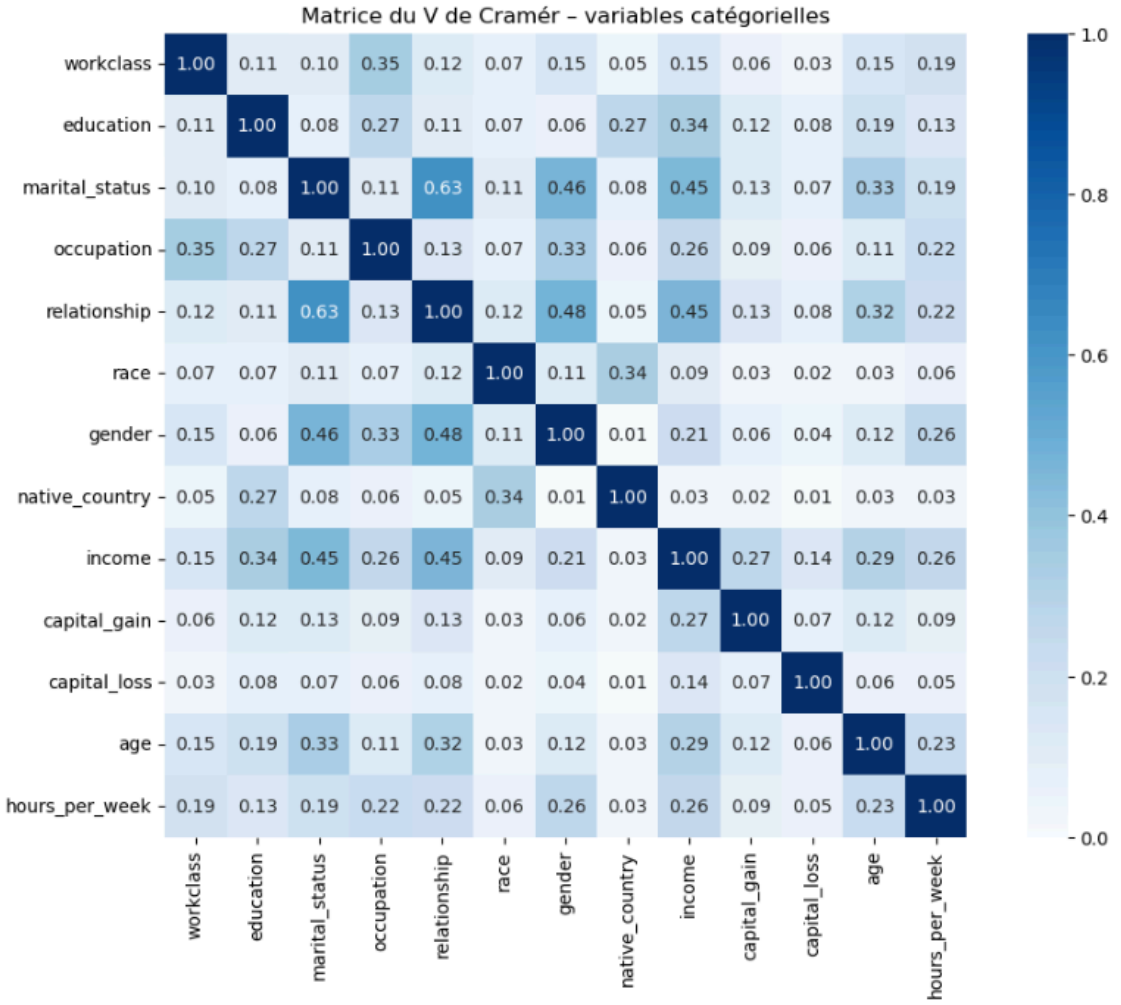


Figure 3 : Matrice du V de Cramér – variables catégorielles

L'analyse de la matrice du V de Cramér met en évidence plusieurs variables présentant une association notable avec la variable cible income, traduisant des liens structurels entre le niveau de revenu et certaines caractéristiques socio-démographiques et professionnelles.

- Fortement lié : Les variables marital_status et relationship apparaissent comme les plus fortement liées au revenu (V ≈ 0.45).
- Modérément lié : La variable education présente également une association significative avec income (V ≈ 0.34), traduisant le rôle central du capital humain dans la structuration des inégalités de revenus. L'âge (V ≈ 0.29) suggère un effet du cycle de vie professionnel, où l'accumulation d'expérience et l'ancienneté peuvent influencer le niveau de rémunération. Les variables hours_per_week (V ≈ 0.26) et occupation (V ≈ 0.26) indiquent que l'intensité du travail ainsi que le type d'activité professionnelle constituent également des déterminants importants du revenu. Enfin, la variable capital_gain (V ≈ 0.27) révèle l'impact des revenus du capital dans la distinction des niveaux de revenus, en particulier pour les individus appartenant aux catégories de revenus élevés.
- Faiblement lié : Les variables native_country (0.03), race (0,09), gender(0,21), workclass(0.15) et capital_loss (0.14) présentent des valeurs de V de Cramér faibles, traduisant une association limitée avec la variable income lorsqu'elles sont considérées isolément.

L'étude des dépendances entre variables explicatives met en évidence certaines redondances susceptibles d'influencer la construction du modèle.

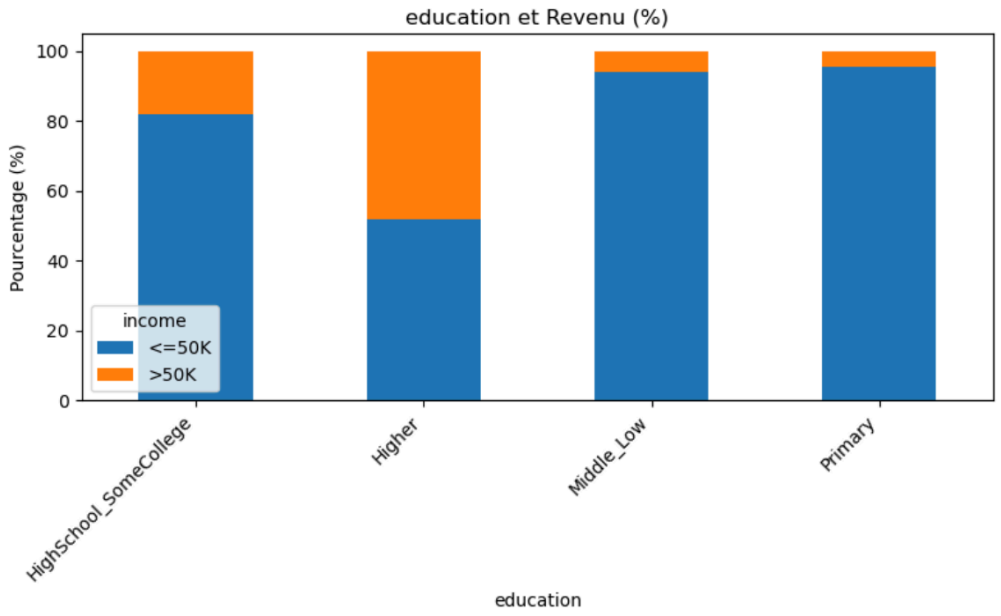
- La relation la plus marquée concerne les variables marital_status et relationship (V ≈ 0.63). Cette forte association indique que ces deux variables véhiculent une information très proche sur la situation familiale des individus.
- De fortes associations sont également observées entre relationship et gender (V ≈ 0.48), ainsi qu'entre marital_status et gender (V ≈ 0.46), traduisant des structures sociales genrées dans les rôles familiaux.
- Des liens modérés existent entre education et occupation (0.27), ainsi qu'entre occupation et workclass(0.35), ce qui est cohérent avec la structuration du marché du travail, sans pour autant constituer des redondances strictes. Ces dépendances doivent être prises en compte lors de la sélection finale des variables afin de limiter les effets de colinéarité.

À l'issue de cette analyse, les variables *education, age, hours_per_week, occupation, capital_gain et une seule variable décrivant la situation familiale : marital_status* apparaissent comme les plus pertinentes pour la suite de l'analyse bivariée et la phase de modélisation. Les variables présentant à la fois une faible association avec la variable cible et une redondance limitée pourront être écartées ou intégrées de manière secondaire selon les performances observées lors de la modélisation.

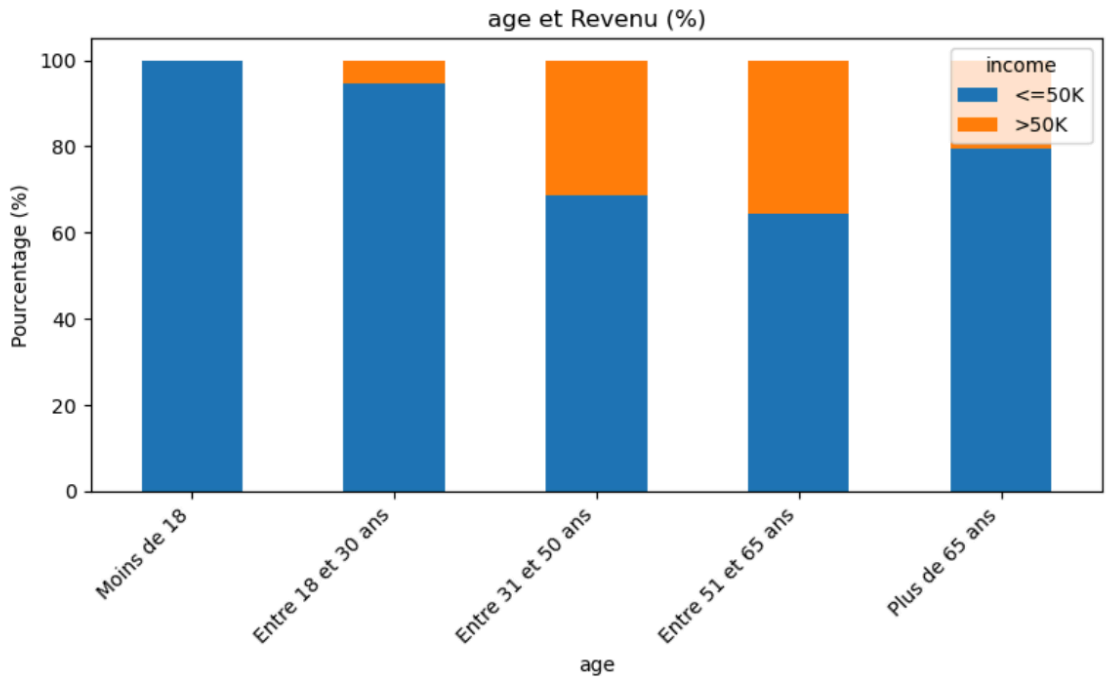
8.2 Analyses bivariées

Les analyses bivariées confirment les résultats précédents :

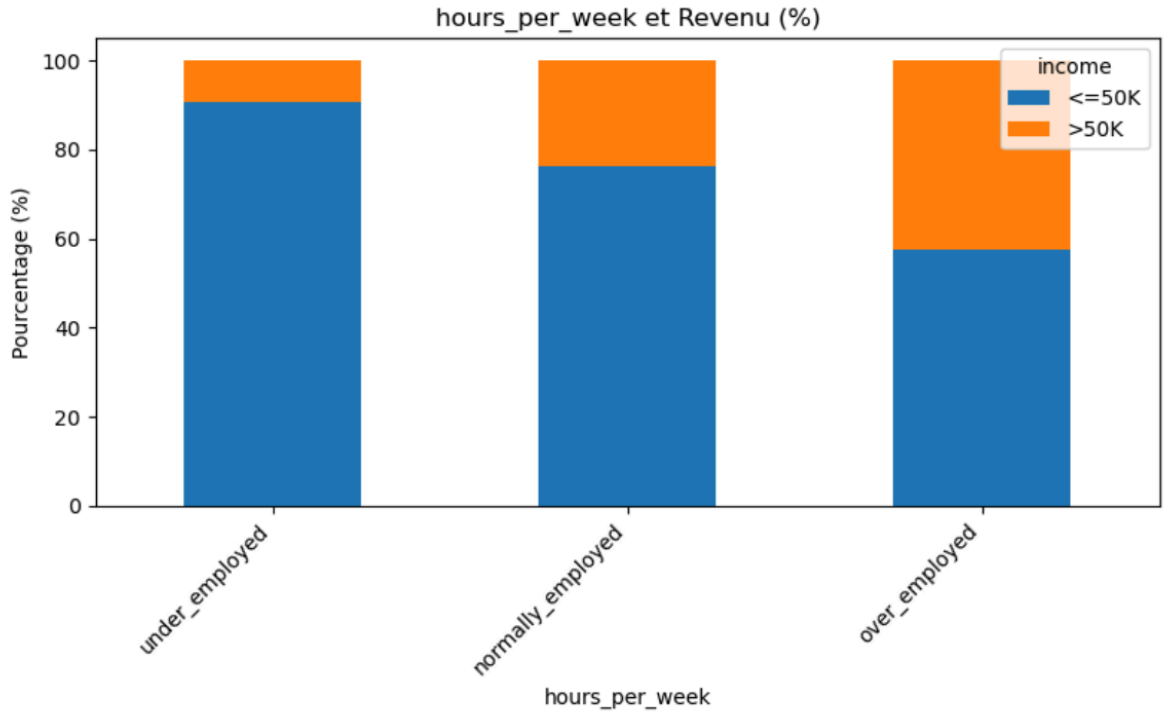
- **Éducation** : les individus de niveau *Higher* présentent une probabilité élevée de revenus >50K.



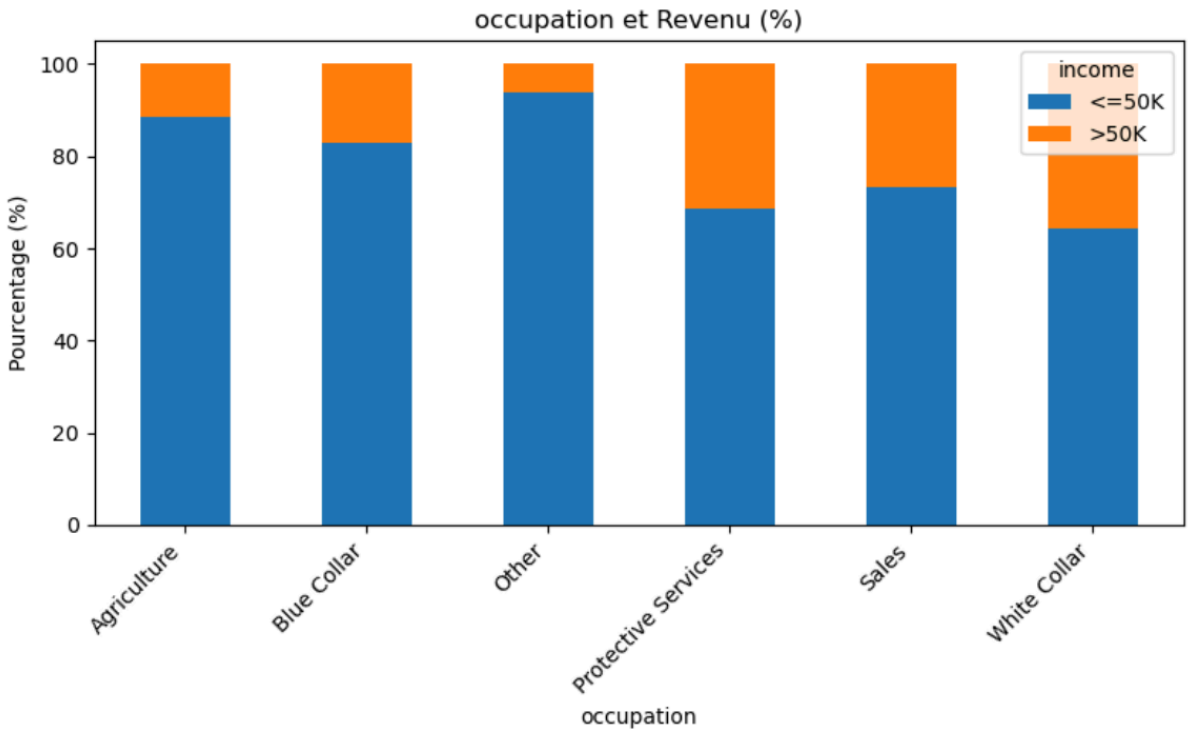
- **Âge** : la proportion de hauts revenus augmente fortement entre 30 et 65 ans.



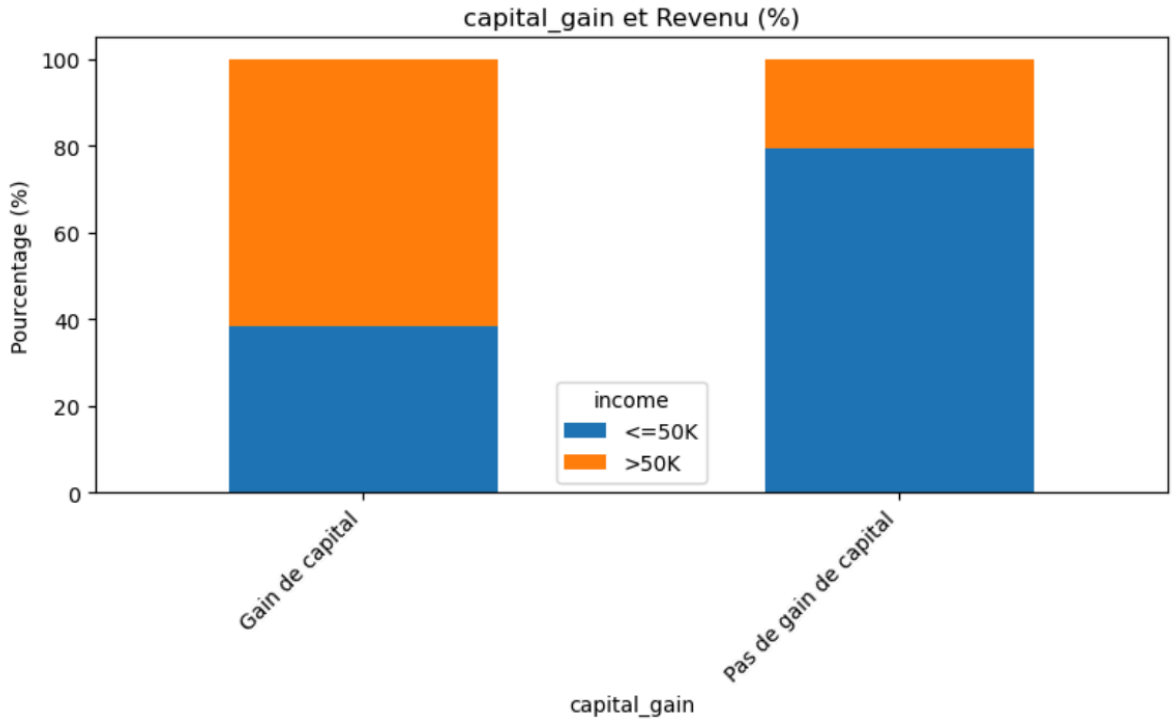
- **Heures travaillées** : les individus *over employed* présentent une proportion élevée de revenus >50K.



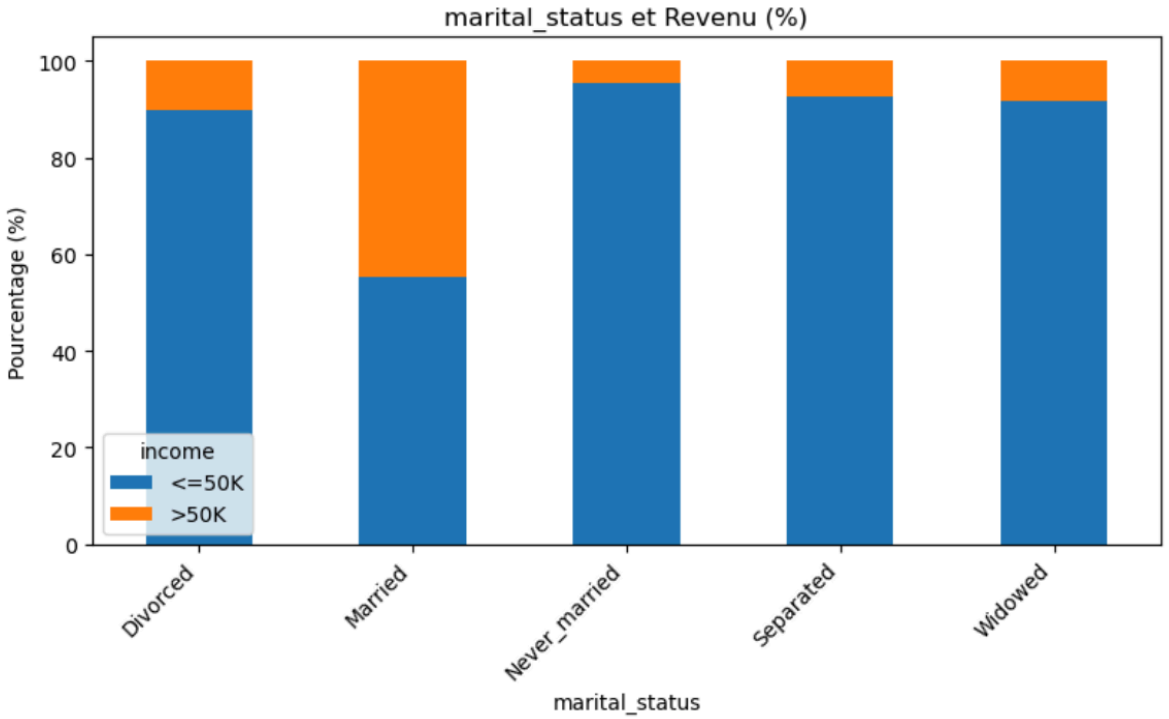
- **Occupation** : les catégories *White Collar* sont plus fréquemment associées à des revenus élevés.



- **Capital-gain** : la présence d'un gain de capital est fortement discriminante.



- **Statut marital** : les individus mariés présentent une proportion significativement plus élevée de revenus >50K.



Ces analyses bivariées confirment les résultats observés lors de l'étude de la matrice de dépendance. Les variables présentant des contrastes marqués dans la répartition du revenu seront retenues pour la phase de modélisation.

9. Modélisation

9.1 Objectif de la modélisation

À l'issue des phases d'exploration, de pré-traitement et d'analyse des relations entre variables, l'objectif est désormais de construire un modèle de classification supervisée permettant de prédire le niveau de revenu annuel d'un individu.

La variable cible considérée est **income**, avec deux modalités : **≤ 50K** et **> 50K**.
La modélisation vise à évaluer la capacité des variables explicatives sélectionnées à discriminer ces deux classes, tout en mettant en place un processus reproductible et automatisable.

9.2 Jeu de données de modélisation

La phase de modélisation repose sur le jeu de données **data_modeling.csv**, issu du pré-traitement.
Ce jeu de données contient uniquement les variables jugées les plus pertinentes à l'issue des analyses précédentes :

- education
- age
- hours_per_week
- occupation
- capital_gain
- marital_status

La réduction du nombre de variables permet de limiter la complexité du modèle, de réduire la redondance entre variables et d'améliorer la robustesse de l'apprentissage.

9.3 Séparation des données

Le jeu de données est séparé en :

- **80 %** de données d'entraînement,
- **20 %** de données de test.

La séparation est réalisée de manière **stratifiée** afin de conserver la proportion des classes de la variable cible, compte tenu du déséquilibre observé lors de l'analyse exploratoire.

9.4 Pipeline de modélisation

Afin de garantir la reproductibilité et d'éviter toute fuite d'information, l'ensemble du processus est intégré dans un **pipeline de machine learning**.

Ce pipeline comprend :

- un **transformateur de colonnes**,
- un **encodage One-Hot** des variables catégorielles,
- un **algorithme de classification**.

Cette approche garantit que les mêmes transformations sont appliquées de manière cohérente aux données d'entraînement et de test.

9.5 Modèle retenu

Plusieurs modèles ont été évalués. Le modèle retenu est un **Naive Bayes de type Bernoulli (BernoulliNB)**.

Ce choix s'explique par :

- la nature majoritairement catégorielle des variables,
- l'encodage binaire issu du One-Hot Encoding,
- la simplicité et la rapidité d'entraînement du modèle,
- sa robustesse dans des contextes de déséquilibre de classes.

9.6 Évaluation des performances

L'évaluation du modèle est réalisée sur le jeu de test à l'aide des métriques suivantes :

- accuracy,

- précision, rappel et F1-score par classe,
- F1-score macro, utilisé comme métrique principale.

Les résultats obtenus sont les suivants :

- **Accuracy : 83 %**
- **F1-score macro : 0,77**

Le modèle présente de très bonnes performances sur la classe majoritaire (≤ 50K) et une capacité satisfaisante à identifier les individus à revenus élevés (> 50K), malgré le déséquilibre initial des données.

9.7 Analyse de la matrice de confusion

L'analyse de la matrice de confusion montre :

- une forte proportion de prédictions correctes pour la classe ≤ 50K,
- une détection pertinente de la classe > 50K,
- des erreurs principalement liées à des individus à hauts revenus prédits comme appartenant à la classe ≤ 50K.

Ces résultats sont cohérents avec la distribution initiale des données.

10. Automation et démarche MLOps

10.1 Objectif de l’automatisation

Dans une logique MLOps, l’objectif de l’automatisation est d’assurer la **reproductibilité**, la **traçabilité** et la **robustesse** du processus de modélisation.

L’automatisation permet de relancer l’ensemble du pipeline de manière standardisée, sans intervention manuelle.

10.2 Pipeline automatisé

Un pipeline automatisé a été implémenté afin de regrouper :

- le chargement des données,
- l’entraînement du modèle,
- l’évaluation des performances,
- la sauvegarde des artefacts.

Ce pipeline peut être exécuté à partir d’un simple chemin vers le jeu de données.

10.3 Sauvegarde des artefacts

Les éléments suivants sont sauvegardés automatiquement :

- le modèle entraîné (sérialisé via *joblib*),
- les métriques d’évaluation.

Cette approche permet de conserver l’historique des modèles et facilite leur réutilisation ou leur déploiement.

10.4 Intérêt de la démarche MLOps

La démarche mise en place permet :

- une reproductibilité complète des expériences,
- une préparation à la mise en production,
- une base solide pour l’intégration future d’outils de CI/CD ou d’orchestration.

Conclusion

Ce projet a permis de mettre en œuvre l'ensemble des étapes clés d'un projet de data science appliqué à un problème réel de classification supervisée : la prédiction du niveau de revenu annuel d'un individu à partir de caractéristiques socio-démographiques et professionnelles. L'approche adoptée s'est appuyée sur une analyse rigoureuse des données, suivie d'un pré-traitement structuré, d'une phase de modélisation et enfin de la mise en place d'un pipeline automatisé dans une logique MLOps.

La phase d'exploration des données a mis en évidence la richesse et l'hétérogénéité du jeu de données, ainsi que la présence de déséquilibres importants, tant au niveau des classes de la variable cible que de certaines variables explicatives. L'analyse des distributions, des valeurs manquantes et des relations entre variables a permis de mieux comprendre les mécanismes sous-jacents aux données et d'orienter les choix méthodologiques pour la suite du projet.

Le pré-traitement des données a constitué une étape déterminante. Les opérations de nettoyage, de transformation et de regroupement des modalités ont permis de réduire la complexité du jeu de données tout en conservant l'information pertinente. Les analyses de dépendance et les études bivariées ont conduit à la sélection d'un ensemble restreint de variables explicatives, offrant un bon compromis entre pouvoir prédictif et robustesse du modèle.

La phase de modélisation a montré qu'il est possible d'obtenir des performances satisfaisantes, avec une accuracy d'environ 83 %, malgré le déséquilibre des classes. Le modèle retenu parvient à bien identifier les individus appartenant à la classe majoritaire et présente une capacité raisonnable à détecter les revenus élevés. Ces résultats confirment la pertinence des variables sélectionnées et soulignent l'importance d'utiliser des métriques adaptées, telles que le F1-score macro, pour évaluer correctement les performances.

Enfin, la mise en place d'un pipeline automatisé s'inscrit pleinement dans une démarche MLOps. L'automatisation de l'entraînement, de l'évaluation et de la sauvegarde des artefacts permet de garantir la reproductibilité, la traçabilité et la robustesse du processus de modélisation. Cette approche constitue une base solide pour une évolution future vers un déploiement opérationnel du modèle.

En perspective, plusieurs pistes d'amélioration peuvent être envisagées, notamment l'exploration de techniques de rééquilibrage des classes, l'optimisation des hyperparamètres, l'intégration de modèles plus complexes ou encore la mise en place d'un suivi des performances en production. Dans l'ensemble, ce projet illustre de manière concrète les enjeux et les bonnes pratiques d'un projet de data science moderne, intégrant à la fois les aspects analytiques et les principes du MLOps.