

RAPPORT SYNTHÉTIQUE

PROJET MACHINE LEARNING

Analyse et Prédition du Recrutement de Candidats

Modèle retenu : KNN sur dataset sans valeurs manquantes

Performance globale : F1-Score = 0.8682

Fiabilité : Validée par validation croisée

Réalisé par :

Fatoumata BAH
Ndeye Aïssatou CISSE
Papa Magatte DIOP
Ndoassan Armand DJEKONBE

Sous la supervision de :

Mme Fatou SALL

Janvier 2026

Table des matières

Résumé Exécutif	2
1 Contexte et Objectifs	3
1.1 Problématique	3
1.2 Objectifs du projet	3
1.3 Jeu de données	3
2 Méthodologie	3
2.1 Analyse Exploratoire des Données	3
2.2 Préparation des Données	4
2.3 Modélisation	4
2.4 Optimisation	5
3 Résultats	5
3.1 Top 3 des Meilleurs Modèles	5
3.2 Variables les Plus Importantes	6
3.3 Analyse des Erreurs	7
3.4 Synthèse des Résultats	8
4 Recommandations	9
4.1 Pour les Recruteurs	9
5 Conclusion	11
5.1 Synthèse des Contributions	11
5.2 Impact Attendu	12
5.3 Mot de Fin	12

Résumé Exécutif

Ce projet vise à développer un modèle prédictif capable d'identifier les candidats susceptibles d'être recrutés, à partir de leurs caractéristiques personnelles, académiques et professionnelles. L'approche méthodologique rigoureuse adoptée a permis de comparer plusieurs stratégies de prétraitement et algorithmes de classification, aboutissant à un modèle optimal avec d'excellentes performances.

- **Modèle final retenu :** K-Nearest Neighbors sur dataset nettoyé
- **Performance globale (F1-Score) :** 0.8682
- **Fiabilité :** Validée par validation croisée stratifiée

1 Contexte et Objectifs

1.1 Problématique

Dans un contexte de recrutement, il est crucial de pouvoir identifier rapidement les candidats ayant le plus de chances d'être embauchés, afin d'optimiser les ressources et d'améliorer l'efficacité du processus de sélection.

1.2 Objectifs du projet

- Analyser les facteurs influençant la décision d'embauche
- Construire un modèle prédictif fiable
- Comparer différentes approches de prétraitement et de modélisation
- Fournir des recommandations actionnables pour les recruteurs

1.3 Jeu de données

- **Source :** Données simulées de processus de recrutement
- **Volume :** 20 000 candidatures
- **Variables :** 11 features + 1 variable cible (embauche)
- **Types de variables :**
 - Démographiques : âge, sexe, couleur de cheveux
 - Académiques : diplôme, spécialité, note
 - Professionnelles : expérience, salaire, disponibilité
 - Temporelles : date de candidature

2 Méthodologie

2.1 Analyse Exploratoire des Données

2.1.1 Objectif

Comprendre la structure, la qualité et les relations dans les données avant toute transformation.

2.1.2 Analyses réalisées

- Distribution de la variable cible : déséquilibre de classes détecté
- Analyse univariée des variables numériques et catégorielles
- Étude des corrélations entre variables
- Identification des valeurs manquantes (< 1% par variable)
- Détection des valeurs aberrantes

2.1.3 Principaux constats

- Valeurs manquantes présentes sur toutes les features mais en faible quantité
- Valeurs aberrantes identifiées sur plusieurs variables numériques
- Corrélations attendues entre expérience, salaire et diplôme
- Déséquilibre modéré des classes

2.2 Préparation des Données

Pour maximiser les chances d'obtenir le meilleur modèle, nous avons testé 6 configurations différentes issues de la combinaison de :

2.2.1 Stratégies d'imputation

1. Suppression des valeurs manquantes

- Simple et sans biais d'imputation
- Perte acceptable de données

2. Imputation par médiane/mode

- Médiane pour les variables numériques
- Mode pour les variables catégorielles
- Méthode classique et rapide

3. Imputation par K-Nearest Neighbors

- Méthode sophistiquée préservant les relations entre variables
- k=5 voisins

2.2.2 Traitement des valeurs aberrantes

1. Conservation des valeurs aberrantes

- Préserve l'information complète
- Certaines valeurs aberrantes peuvent être des cas réels importants

2. Traitement par Winsorization

- Remplacement des valeurs extrêmes par les limites IQR
- Réduit le bruit sans perdre d'observations
- Appliqué uniquement si plus de 5% de valeurs aberrantes

Résultat : 6 datasets pour la modélisation (3 sans traitement des aberrantes, 3 avec traitement)

2.2.3 Autres prétraitements

- Encodage des variables catégorielles (One-Hot Encoding)
- Standardisation des variables numériques (StandardScaler)
- Gestion du déséquilibre de classes

2.3 Modélisation

2.3.1 Stratégie adoptée

Comparaison exhaustive de plusieurs algorithmes sur les 6 datasets.

2.3.2 Algorithmes testés

- Logistic Regression (baseline linéaire)
- Random Forest (ensemble learning)
- XGBoost / Gradient Boosting (boosting)
- Support Vector Machine (séparation optimale)
- K-Nearest Neighbors (apprentissage basé sur la proximité)

2.3.3 Métriques d'évaluation

- **Accuracy** : Taux de prédictions correctes global
- **Precision** : Proportion de candidats correctement identifiés comme embauchables parmi tous ceux prédits embauchables
- **Recall** : Proportion de candidats embauchables correctement identifiés parmi tous les candidats réellement embauchables
- **F1-Score** : Moyenne harmonique de Precision et Recall (métrique principale)
- **ROC-AUC** : Capacité de discrimination du modèle

2.3.4 Validation

- Split Train/Test : 80% entraînement / 20% test avec stratification
- Validation croisée : StratifiedKFold (5 folds) sur les 3 meilleurs modèles
- Analyse de l'overfitting : Comparaison des performances train vs validation

2.4 Optimisation

Recherche d'hyperparamètres via Grid Search / Randomized Search sur les meilleurs modèles, avec optimisation des paramètres clés de chaque algorithme et validation par cross-validation.

3 Résultats

3.1 Top 3 des Meilleurs Modèles

Le tableau suivant présente les performances des trois meilleurs modèles identifiés lors de la phase de comparaison exhaustive.

TABLE 1 – Comparaison des 3 meilleurs modèles

Rang	Modèle	Dataset	F1	Precision	Recall	Accuracy	AUC	Temps (s)
1	KNN	dropna	0.8682	0.9484	0.8005	0.9721	0.9286	0.005
2	Random Forest	dropna	0.6944	0.6119	0.8028	0.9191	0.9364	1.644
3	XGBoost	dropna	0.5104	0.3600	0.8761	0.8074	0.9142	0.475

Constat principal : Le dataset obtenu par suppression des valeurs manquantes a systématiquement donné les meilleures performances, ce qui indique que les valeurs manquantes n'apportaient pas d'information pertinente et pouvaient introduire du bruit.

3.1.1 Modèle retenu : K-Nearest Neighbors

Performances

- F1-Score : 0.8682 (excellent équilibre)
- Precision : 94.84% (très haute fiabilité des prédictions positives)
- Recall : 80.05% (bonne identification des candidats embauchables)
- Accuracy : 97.21% (excellente performance globale)
- ROC-AUC : 0.9286 (excellente capacité de discrimination)
- Temps d'entraînement : 0.005 secondes (très rapide)

Matrice de confusion

- Vrais Positifs : 349 candidats correctement identifiés comme embauchables
- Vrais Négatifs : 3350 rejets corrects
- Faux Positifs : 19 candidats sélectionnés à tort
- Faux Négatifs : 87 talents potentiellement ratés

Interprétation métier Le modèle identifie correctement 80% des candidats embauchables, avec une fiabilité exceptionnelle de 95% sur les candidats présélectionnés. Le taux de faux positifs est remarquablement bas, ce qui minimise la charge de travail liée aux entretiens de candidats non qualifiés. Le modèle est particulièrement adapté pour une utilisation en pré-sélection automatique.

Forces

- Très haute précision minimisant les erreurs sur les candidats sélectionnés
- Temps d'entraînement négligeable permettant un ré-entraînement fréquent
- Excellente accuracy globale
- Simplicité d'interprétation basée sur la proximité

Limites

- 20% de talents potentiellement ratés nécessitent une révision manuelle complémentaire
- Sensibilité au choix du nombre de voisins

3.1.2 Comparaison avec les alternatives

Random Forest Bien que présentant un recall légèrement meilleur, ce modèle génère un nombre trop élevé de faux positifs, avec une précision nettement inférieure. Cette configuration entraînerait une surcharge de travail pour les recruteurs avec de nombreux candidats non qualifiés à examiner.

XGBoost Malgré le meilleur recall identifiant 88% des candidats embauchables, la précision catastrophique rend ce modèle inutilisable en production. Le nombre très élevé de faux positifs surchargerait considérablement le processus de recrutement.

3.2 Variables les Plus Importantes

L'analyse par permutation importance sur le modèle KNN final révèle une hiérarchie claire des facteurs prédictifs, avec une concentration marquée de l'information sur les neuf premières variables qui représentent 99.7% de l'importance totale cumulée.

3.2.1 Les neuf variables décisives

TABLE 2: Importance des variables clés

Rang	Variable	Importance	Interprétation Métier
1	Ratio salaire/expérience	0.7100	Cohérence du profil entre prétentions salariales et niveau d'expérience
2	Interaction expérience-note	0.7097	Synergie entre performance académique et expérience validée sur le terrain

Suite à la page suivante

Table 2 – Suite de la page précédente

Rang	Variable	Importance	Interprétation Métier
3	Interaction diplôme-note	0.6440	Excellence académique : diplôme soutenu par de bonnes performances
4	Note académique	0.2958	Performance académique brute comme indicateur de compétence
5	Interaction diplôme-expérience	0.2395	Valorisation du diplôme par l'expérience professionnelle
6	Écart d'expérience	0.2279	Positionnement relatif par rapport aux exigences du poste
7	Âge	0.2206	Maturité professionnelle et équilibre expérience-potentiel
8	Score composite	0.1977	Vue d'ensemble synthétique du profil du candidat
9	Expérience professionnelle	0.0250	Années d'expérience considérées isolément

3.2.2 Insights clés

La synergie prime sur les critères isolés Les trois variables les plus importantes sont des interactions entre critères. Ce constat démontre que la cohérence entre différents aspects du profil est plus prédictive que chaque critère pris individuellement. Un candidat avec une excellente note académique mais sans expérience proportionnelle, ou avec des prétentions salariales incohérentes, sera moins valorisé qu'un profil équilibré.

L'expérience seule a un impact limité Paradoxalement, l'expérience professionnelle isolée n'occupe que le neuvième rang, alors qu'elle participe aux trois interactions majeures. Cela signifie que l'expérience ne crée de valeur que lorsqu'elle est mise en perspective avec la performance académique, le diplôme ou le salaire demandé.

Les variables démographiques ont un impact nul Les variables démographiques et catégorielles présentent des importances proches de zéro. Le modèle ne discrimine pas sur des critères subjectifs ou protégés (sexe, apparence physique), ce qui garantit une équité du processus de sélection. Les décisions reposent uniquement sur des critères objectifs de compétence et de cohérence de profil.

L'effet temporel est négligeable Les variables temporelles ont des importances marginales, ce qui contredit les fluctuations saisonnières observées lors de l'analyse exploratoire. Cela signifie que le modèle capture la qualité intrinsèque des candidats, indépendamment de leur période de candidature. Les variations temporelles observées résultent de fluctuations de la qualité des candidatures plutôt que de biais temporels dans les décisions.

L'analyse temporelle révèle un paradoxe significatif : alors que le volume de candidatures reste stable, le taux d'embauche fluctue fortement, et ces deux métriques évoluent de manière inverse. Cette relation contre-intuitive suggère que la qualité des candidatures prime sur leur quantité, les périodes de fort afflux étant potentiellement associées à une dilution des profils pertinents.

3.3 Analyse des Erreurs

3.3.1 Répartition des erreurs

La matrice de confusion du modèle KNN révèle la distribution suivante :

- Vrais Positifs : 349 candidats correctement identifiés comme embauchables
- Vrais Négatifs : 3350 rejets corrects
- Faux Positifs : 19 candidats sélectionnés à tort
- Faux Négatifs : 87 talents ratés par le modèle

3.3.2 Analyse des erreurs critiques

Faux Négatifs - Risque de perte de talents Les 87 faux négatifs représentent des candidats embauchables rejetés par le modèle. L'impact principal est la perte d'opportunités de recrutement. Pour mitiger ce risque, une révision manuelle des candidats avec des scores de probabilité dans une zone intermédiaire est recommandée.

Faux Positifs - Risque faible Les 19 faux positifs correspondent à des candidats non embauchables sélectionnés par erreur. L'impact est minimal car la charge de travail supplémentaire reste très faible. Cette configuration permet parfois de découvrir des profils atypiques intéressants.

Cas d'erreurs typiques Les faux négatifs concernent souvent des profils avec des parcours atypiques ou des valeurs limites sur plusieurs critères. Les faux positifs, très rares, correspondent généralement à des candidats ayant de bonnes caractéristiques individuelles mais manquant d'adéquation globale.

3.3.3 Équilibre Precision-Recall

Le modèle KNN priviliege la precision au détriment du recall. C'est un choix conservateur qui minimise les faux positifs et convient parfaitement à un usage en pré-sélection, acceptant de rater une proportion de talents qui pourra être compensée par une révision manuelle des cas limites.

3.4 Synthèse des Résultats

3.4.1 Justification du choix

1. Meilleur F1-Score parmi tous les modèles testés, indiquant un équilibre optimal
2. Precision exceptionnelle minimisant les erreurs sur les candidats sélectionnés
3. Rapidité d'entraînement adaptée à un ré-entraînement fréquent
4. Simplicité d'interprétation facilitant l'appropriation par les équipes métier

3.4.2 Performance en contexte métier

Sur un volume annuel de candidatures, le modèle permet :

- Un gain de temps considérable en écartant automatiquement la majorité des candidats non qualifiés
- Une identification efficace des bons candidats
- Un nombre minimal de faux positifs à filtrer manuellement
- Une perte acceptable de talents compensable par révision des cas limites

Le modèle atteint un niveau de performance opérationnel pour une mise en production.

4 Recommandations

4.1 Pour les Recruteurs

4.1.1 Priorisation des Critères d'Évaluation

Sur la base des variables critiques identifiées, nous recommandons une grille d'évaluation structurée en trois niveaux de priorité.

Niveau 1 : Critères de cohérence Les trois premières variables démontrent l'importance fondamentale de la cohérence du profil.

1. Ratio Salaire/Expérience

- Établir une fourchette de référence par niveau d'expérience
- Identifier les écarts significatifs nécessitant justification
- Utiliser comme premier filtre de pertinence du profil

2. Synergie Expérience-Performance Académique

- Pondérer la performance académique par l'expérience validée
- Valoriser les profils équilibrés
- Définir un seuil minimal pour présélection automatique

3. Excellence Académique Validée

- Valoriser les excellents résultats dans des filières exigeantes
- Alerter sur les incohérences entre diplôme et performance
- Privilégier la cohérence diplôme-résultats

Niveau 2 : Critères de validation Ces variables permettent de valider et affiner l'évaluation initiale.

1. Performance Académique Brute

- Fixer un seuil minimal selon le type de poste
- Relativiser selon l'ancienneté du candidat
- Compléter par l'analyse des interactions

2. Valorisation du Diplôme par l'Expérience

- Vérifier l'adéquation diplôme-parcours professionnel
- Identifier les potentiels inexploités ou sur-qualifications
- Analyser la progression de carrière

3. Positionnement Relatif

- Évaluer l'adéquation avec les exigences du poste
- Identifier les sous-qualifications ou sur-qualifications
- Ajuster selon le contexte spécifique

4. Maturité Professionnelle

- Utiliser uniquement comme proxy de maturité
- Ne jamais mentionner comme critère éliminatoire
- Vérifier la cohérence avec l'expérience déclarée

Niveau 3 : Vue d'ensemble

1. Score Composite

- Calculer un score global agrégeant tous les critères
- Utiliser pour le classement et la priorisation
- Définir des seuils de décision

2. Expérience Brute

- Vérifier le minimum requis par le poste
- Toujours coupler avec d'autres critères
- Ne pas considérer isolément

4.1.2 Optimisation du Processus selon l'Analyse Temporelle

L'analyse temporelle révèle un paradoxe entre volume et qualité des candidatures.

Constats clés

- Volume de candidatures relativement stable
- Taux d'embauche variant significativement
- Corrélation négative entre volume et qualité

Recommandations opérationnelles

1. Stratégie de sourcing qualitatif

- Privilégier le ciblage précis plutôt que l'attraction massive
- Concentrer les efforts sur les canaux à forte valeur ajoutée
- Mesurer systématiquement le taux de conversion par source

2. Gestion des pics de candidatures

- Identifier les périodes à risque via l'historique
- Renforcer les critères de présélection lors des pics
- Maintenir un flux constant de candidatures qualifiées

3. Utilisation prédictive du modèle

- Scorer automatiquement les candidatures dès réception
- Prioriser les profils à forte probabilité
- Examiner manuellement la zone de probabilité intermédiaire
- Automatiser les rejets pour les probabilités très faibles

4. Amélioration continue

- Tracer systématiquement les décisions finales
- Analyser les écarts entre prédictions et réalité
- Identifier les profils atypiques pour enrichissement futur
- Mettre à jour régulièrement le modèle

4.1.3 Vigilance Éthique et Conformité

Points positifs du modèle actuel

- Absence d'impact des variables démographiques
- Décisions basées sur des critères objectifs de compétence
- Pas de biais temporel ou géographique détecté

Risques à surveiller

1. Biais indirect potentiel

- Documenter l'usage des variables sensibles
- Former les équipes aux bonnes pratiques
- Auditer régulièrement les décisions

2. Conformité réglementaire

- Informer les candidats de l'usage d'un modèle algorithmique
- Expliquer les critères de décision sur demande
- Permettre une révision manuelle si sollicitée

3. Audit régulier

- Fréquence semestrielle recommandée
- Vérifier l'absence de discrimination
- Analyser les écarts par segment
- Valider la stabilité des variables non-discriminantes

Documentation légale

- Tenir un registre des traitements algorithmiques (Article 30 RGPD)
- Nommer un référent IA/éthique chargé du suivi
- Préparer une notice explicative pour les candidats
- Établir une procédure de recours pour les candidats
- Maintenir une traçabilité complète des décisions automatisées

5 Conclusion

5.1 Synthèse des Contributions

Ce projet a permis de développer un système de prédiction de recrutement performant basé sur l'apprentissage automatique, avec des résultats opérationnels validés.

5.1.1 Apports principaux

1. Méthodologie rigoureuse

- Comparaison exhaustive de 6 stratégies de prétraitement
- Évaluation de 5 algorithmes de classification
- Validation croisée systématique
- Analyse approfondie des performances

2. Modèle performant

- F1-Score de 0.8682 démontrant un excellent équilibre
- Precision de 94.84% minimisant les faux positifs
- Rapidité d'exécution adaptée à la production
- Stabilité validée par cross-validation

3. Insights actionnables

- Identification des 9 variables critiques
- Importance primordiale de la cohérence du profil

- Rôle central des interactions entre critères
- Absence de biais démographiques détectables

4. Recommandations opérationnelles

- Grille d'évaluation structurée en 3 niveaux
- Stratégie de déploiement progressive
- Plan de maintenance et d'amélioration continue
- Cadre éthique et de conformité

5.2 Impact Attendu

5.2.1 Pour les recruteurs

- Gain de temps significatif sur le tri initial
- Meilleure priorisation des candidatures
- Réduction de la charge cognitive
- Focus sur l'évaluation humaine à forte valeur ajoutée

5.2.2 Pour l'organisation

- Amélioration de la qualité du recrutement
- Réduction du temps de traitement
- Optimisation des ressources RH
- Standardisation des pratiques
- Mesure objective de la performance

5.2.3 Pour les candidats

- Délai de réponse réduit
- Processus plus équitable basé sur critères objectifs
- Feedback constructif possible
- Transparence des critères d'évaluation

5.3 Mot de Fin

Ce projet démontre le potentiel de l'intelligence artificielle pour assister et améliorer les processus de recrutement. Le modèle développé offre des performances remarquables tout en maintenant des standards éthiques élevés.

La clé du succès réside dans l'équilibre entre automatisation et expertise humaine. Le modèle doit être perçu comme un outil d'aide à la décision, non comme un substitut au jugement des recruteurs. L'intervention humaine reste essentielle pour capturer les nuances et le potentiel des candidats que les algorithmes ne peuvent pleinement appréhender.

Le déploiement progressif, accompagné d'une amélioration continue basée sur les retours terrain et les données de performance post-embauche, permettra d'affiner constamment le système et de maximiser sa valeur ajoutée.

Nous sommes convaincus que cette approche data-driven, combinée à l'expertise métier, contribuera significativement à l'optimisation du processus de recrutement et à l'identification des meilleurs talents.