

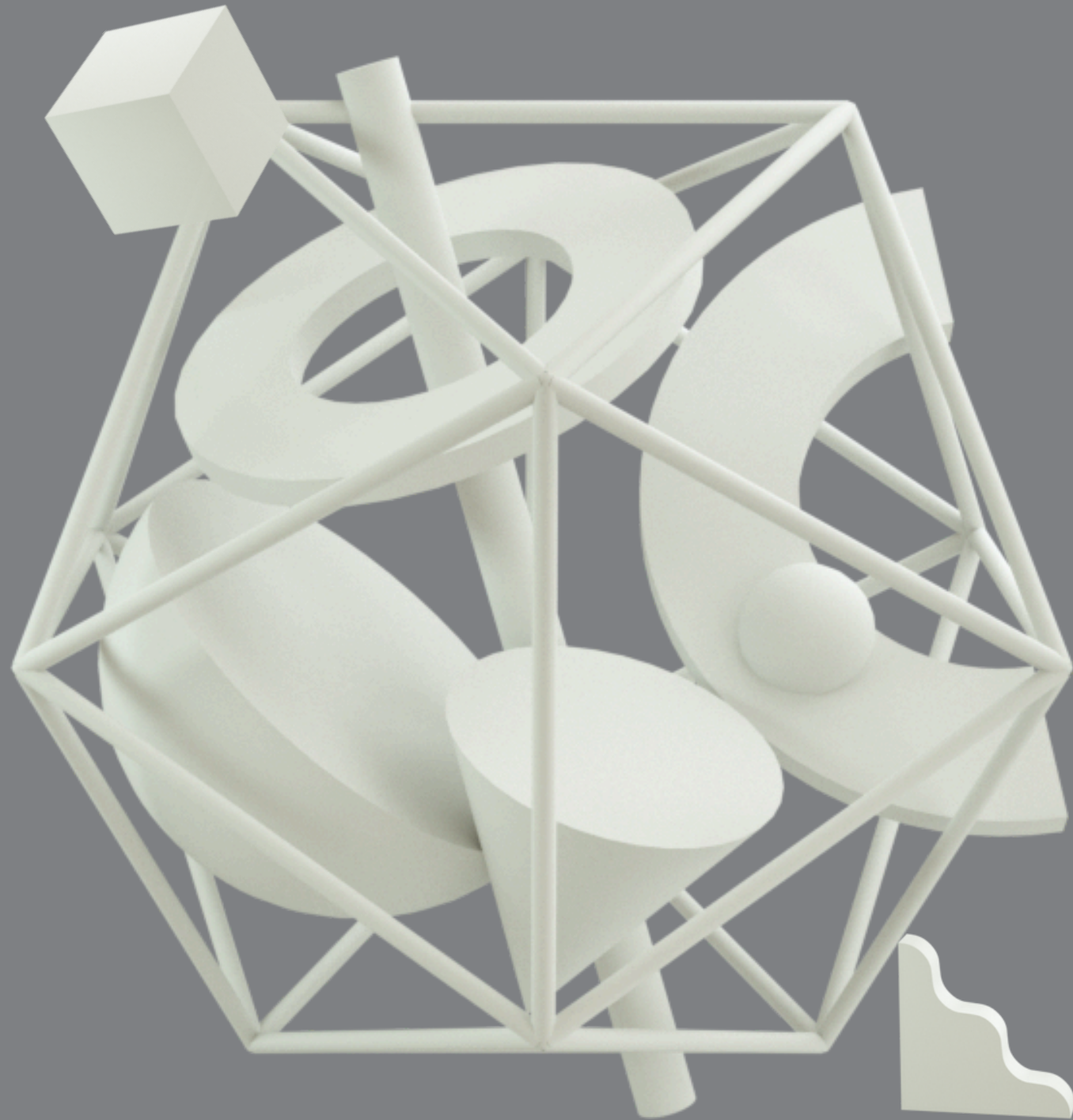
The background is a light gray with various 3D geometric shapes scattered across it. These include a small cube in the top left, a large sphere in the top left, a wavy line in the top right, a cone in the top right, a large cylinder in the center, a sphere in the center, a cone in the bottom left, a large cylinder in the bottom left, a sphere in the bottom left, a large cylinder in the bottom right, a sphere in the bottom right, and a small cube in the bottom right.

Fiabilité des données dans

# L'Internet of Things

(dans la santé)

AIT HAMOUDA Gaya  
CHARFI Sirine  
ISSAAD Célia



# Sommaire

I. Introduction du projet

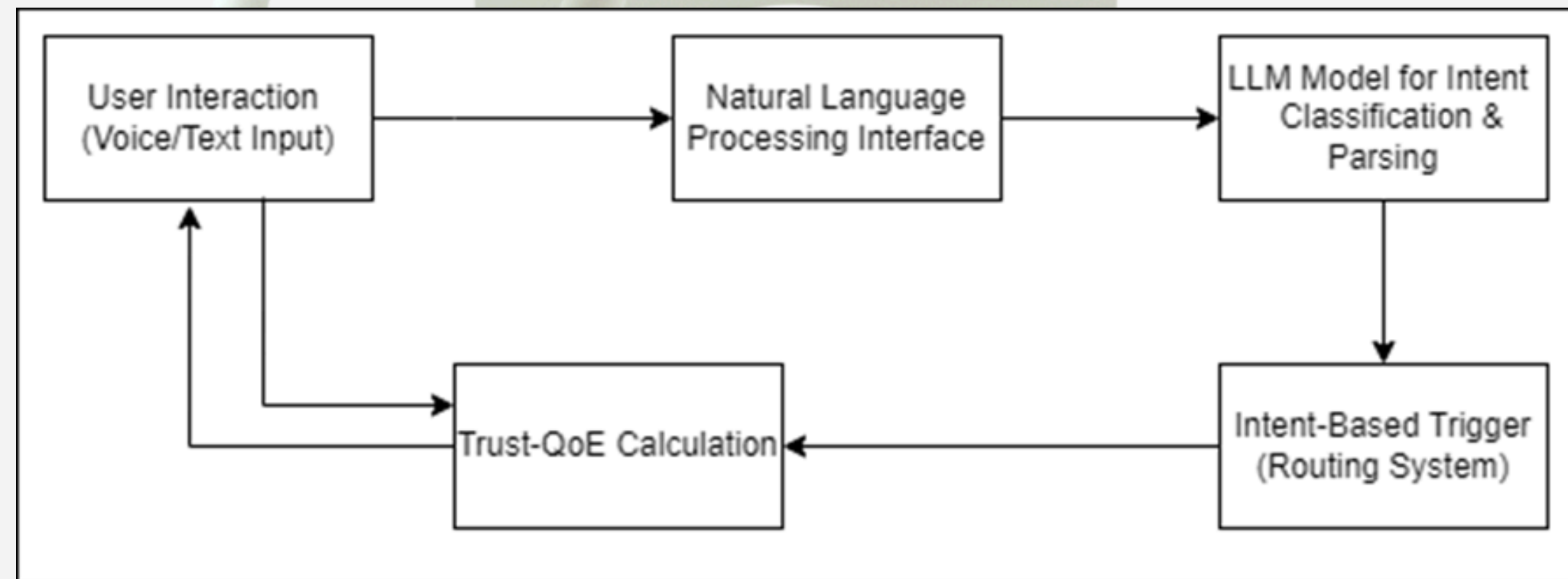
II. Présentation des 5 méthodes de ML utilisées

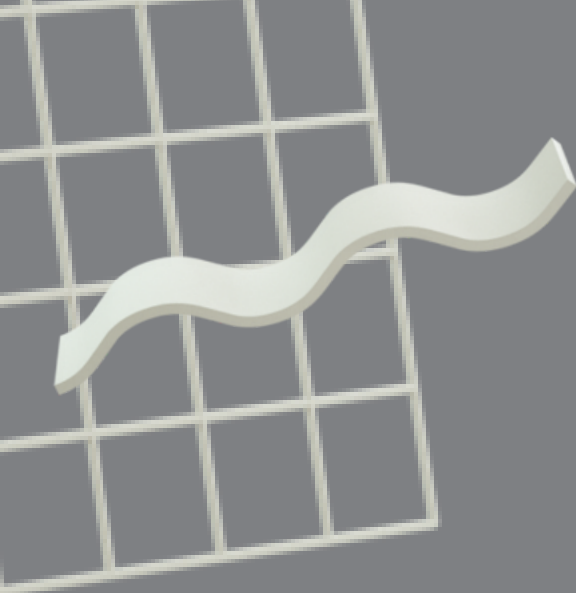
III. Perspectives

IV. Conclusion

# I. Introduction

- **Objectif** : Assurer la fiabilité et la pertinence des réponses des appareils IoT basées sur l'intention utilisateur.
- **Compréhension des Intentions** : Utiliser le NLP et un modèle LLM pour analyser et comprendre les besoins des utilisateurs.
- **Réponse Adaptée** : Déclencher automatiquement l'action appropriée selon l'intention détectée, améliorant l'expérience utilisateur.
- **Confiance Mesurée** : Évaluer et affiner la qualité des réponses pour construire un système de confiance évolutif.
- **Plan de Production** : Développement d'un prototype (POC) en conditions réelles pour tester et valider le concept.





## **II. Travail effectué**



# 5 modèles de Machine Learning

A decorative background featuring a 3D cube in the upper left and a stylized, light-colored figure in the center. The figure appears to be a person with arms raised, possibly in a celebratory or dancing pose, rendered in a simple, rounded style.

- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Logistic Regression
- Support Vector Machine



# Decision Tree

L'arbre de décision est un algorithme d'apprentissage supervisé qui utilise une structure arborescente pour prendre des décisions en fonction des caractéristiques des données. Il peut être utilisé pour les tâches de classification et de régression.

- Fonctionnement:

Nœud racine : L'arbre commence par une question (condition) sur une caractéristique du jeu de données.

Nœuds internes : Chaque nœud représente une condition basée sur une caractéristique spécifique, divisant les données en sous-groupes.

Feuilles : Les nœuds finaux de l'arbre (feuilles) contiennent la prédiction finale ou la classe.

Critère de division : L'arbre utilise des mesures comme l'entropie, le gini ou l'erreur quadratique pour décider la meilleure façon de diviser les données à chaque nœud.

- Avantages :

Interprétable : Facile à comprendre et à visualiser, ce qui rend les décisions transparentes.

Rapide et efficace : Adapté aux petits jeux de données, avec une exécution rapide.

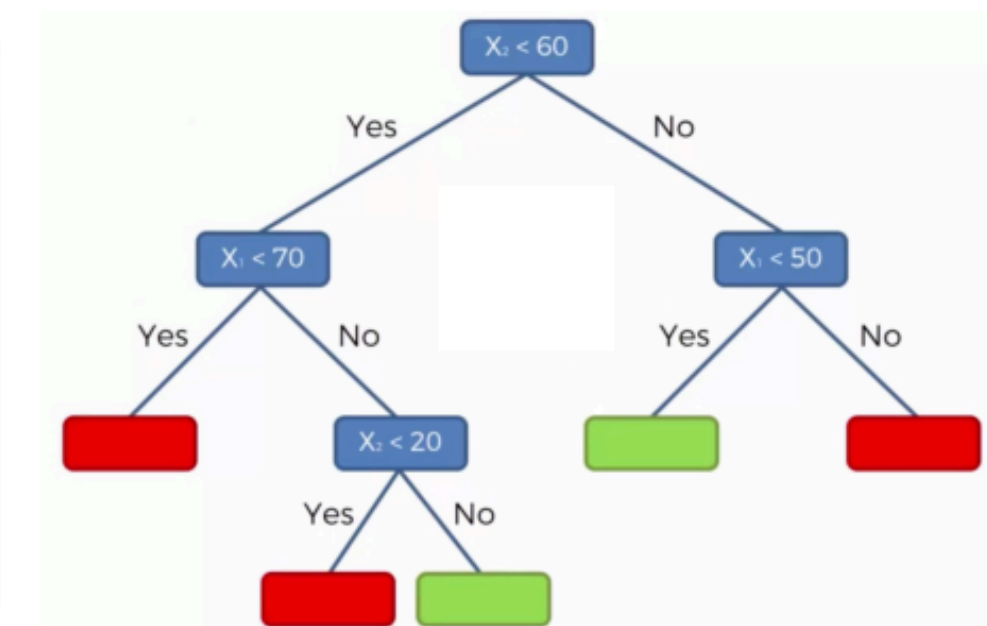
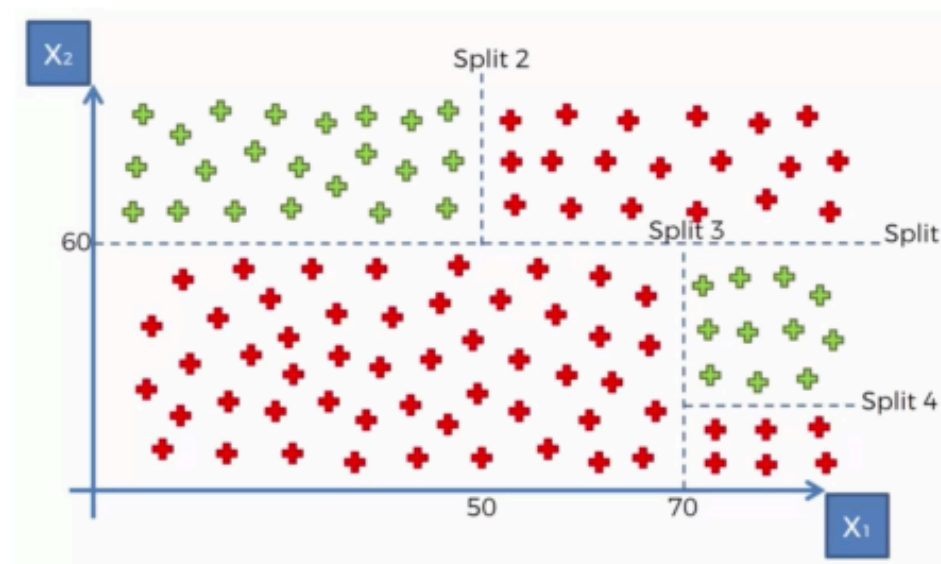
Pas de prétraitement nécessaire : Gère les données numériques et catégorielles sans transformation.

- Inconvénients :

Surapprentissage : Peut être sensible au surapprentissage si l'arbre est trop profond.

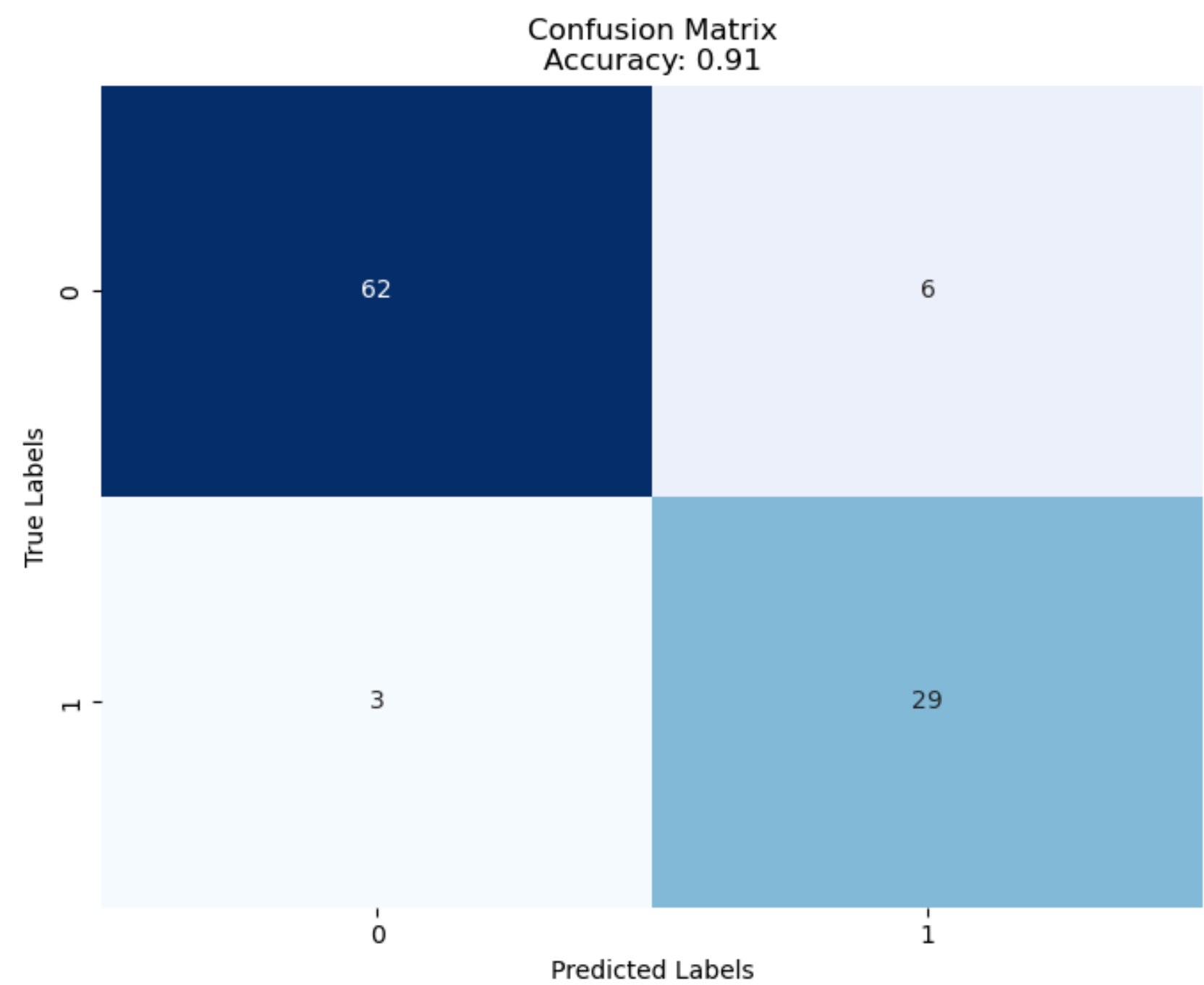
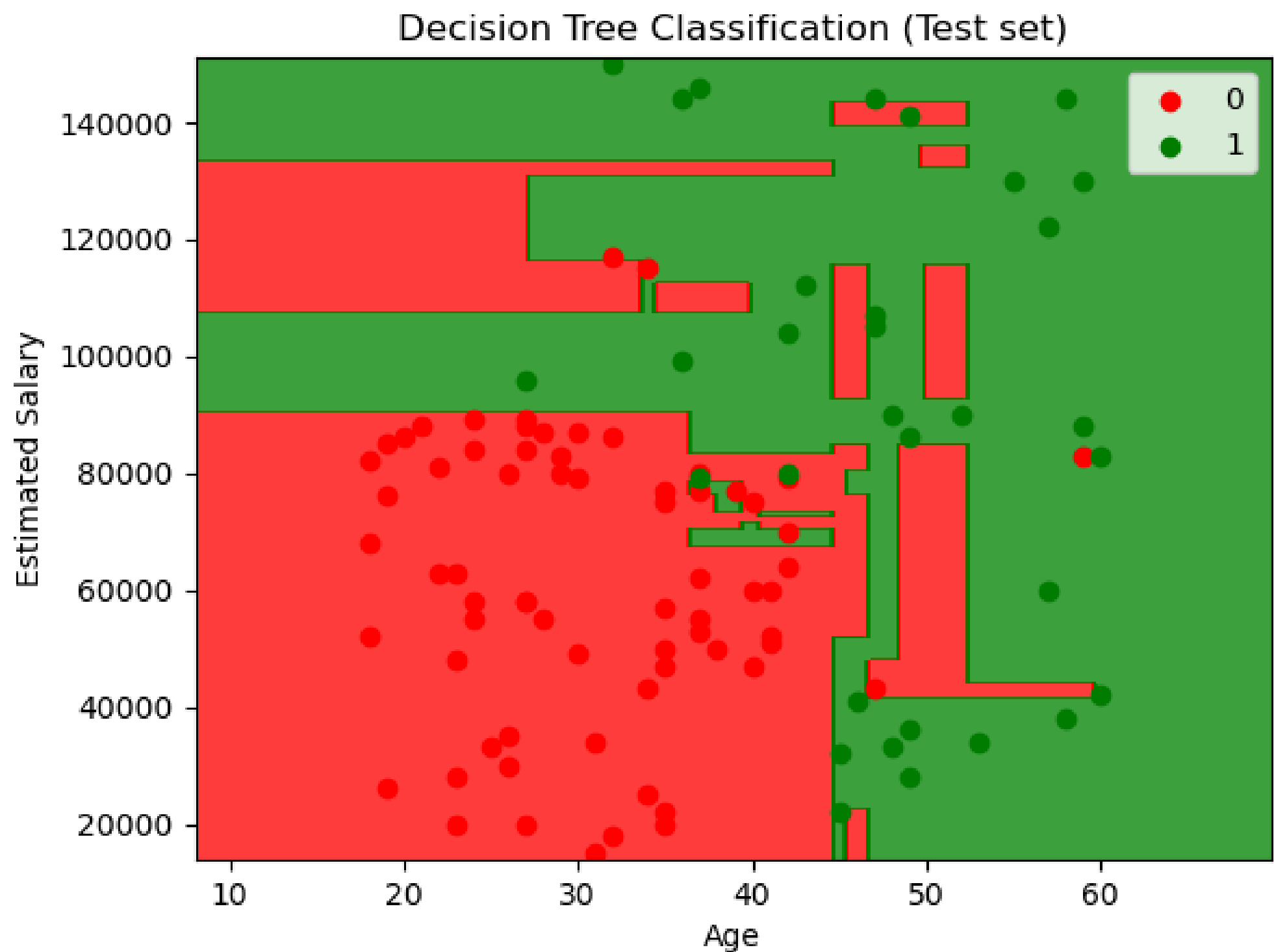
Précision limitée : Moins performant sur des jeux de données complexes par rapport à d'autres méthodes comme le Random Forest.

- Cas d'utilisation : Classification de clients, diagnostic médical, prise de décision dans les processus industriels, etc.



# Decision Tree

Résultats obtenus :



# Random Forest

Le Random Forest est une méthode d'apprentissage par ensemble qui construit plusieurs arbres de décision et combine leurs résultats pour produire une prédiction finale. Il est utilisé pour les tâches de classification et de régression.

- Fonctionnement:

Étape 1 : L'algorithme crée plusieurs arbres de décision en utilisant différents sous-ensembles des données d'entraînement, sélectionnés par bootstrap (échantillonnage aléatoire avec remise).

Étape 2 : Pour chaque division d'un arbre, un sous-ensemble aléatoire de caractéristiques est sélectionné, ce qui réduit la corrélation entre les arbres.

Étape 3 : Le modèle fait des prédictions en agrégeant les résultats de tous les arbres :

- Classification : Prend la majorité des votes entre les arbres.
- Régression : Fait la moyenne des prédictions de tous les arbres.

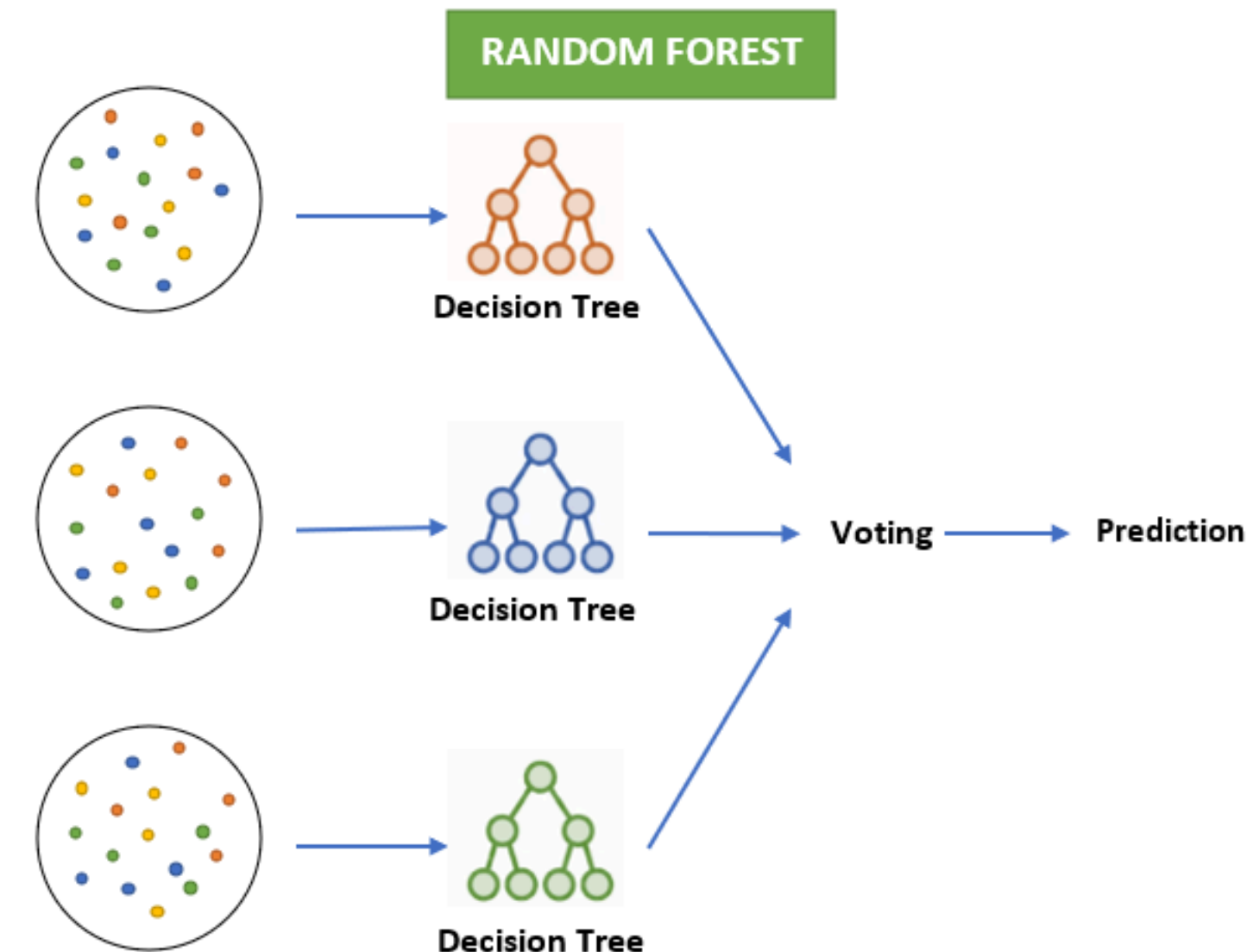
- Avantages :

Haute précision : Réduit le surapprentissage par rapport à un arbre de décision unique.

Robustesse : Gère de grands jeux de données et tolère les valeurs manquantes.

Importance des caractéristiques : Identifie les caractéristiques les plus influentes du jeu de données.

- Cas d'utilisation : Détection de fraude, diagnostic médical, systèmes de recommandation, etc.

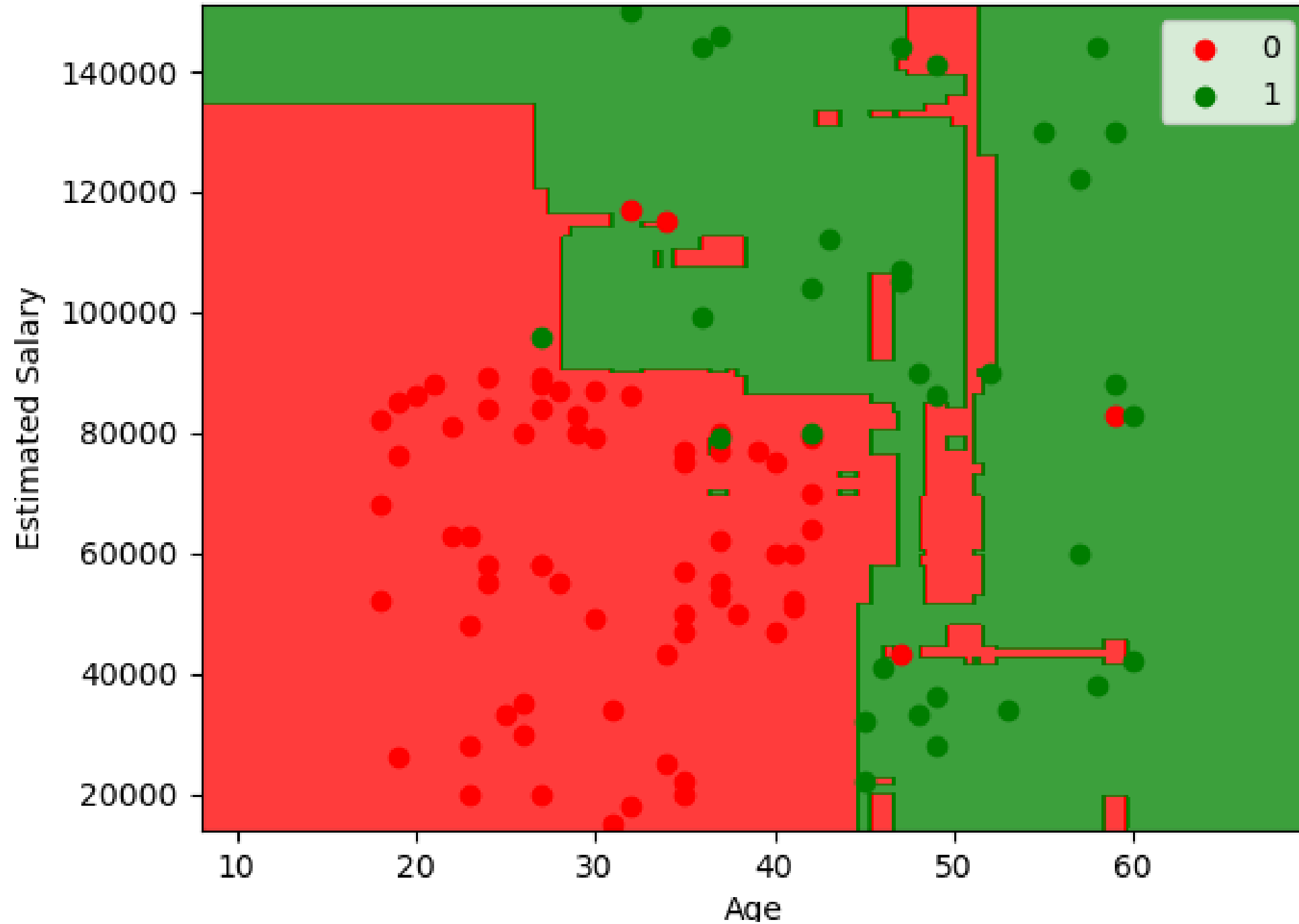




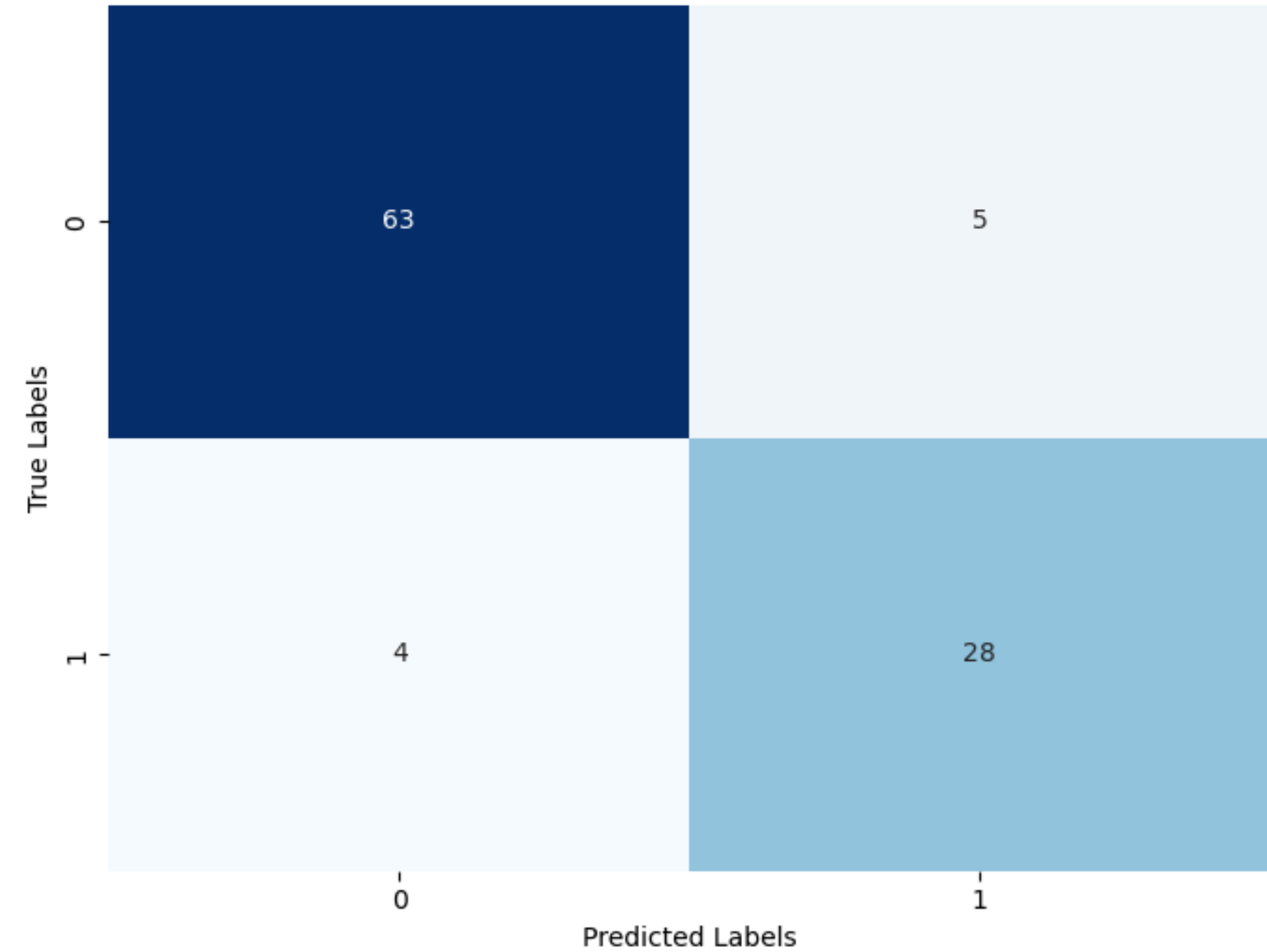
# Random Forest

Résultats obtenus :

Random Forest Classification (Test set)



Confusion Matrix  
Accuracy: 0.91



# K-Nearest Neighbors (KNN)

Algorithme de classification non paramétrique qui classe un point de données en fonction des classes de ses k voisins les plus proches dans l'espace de caractéristiques. Il est également utilisé pour les tâches de régression en prenant la moyenne des valeurs des voisins.

- Fonctionnement :

Étape 1 Calcul de la distance : Pour chaque nouvel échantillon de données, calcul de la distance entre cet échantillon et chaque observation dans le jeu de données d'entraînement.

Étape 2 Choix des Voisins (k) : sélection des k points (nombre de voisins) les plus proches: valeur de k faible -> modèle sensible au bruit, valeur de k élevée -> modèle moins flexible

Étape 3 Prédiction de la classe: détermine la classe majoritaire parmi les k voisins (le nouvel échantillon est assigné à cette classe) et calcule la moyenne des valeurs des k voisins pour obtenir la prédiction (=régression).

- Avantages :

Simplicité et Intuitivité : KNN est facile à comprendre et à implémenter. Le concept de classification par proximité est intuitif.

Flexibilité : KNN ne fait aucune hypothèse sur la distribution des données, ce qui le rend adapté aux distributions non linéaires.

Polyvalence : Fonctionne pour la classification et la régression, ce qui le rend applicable à divers types de problèmes.

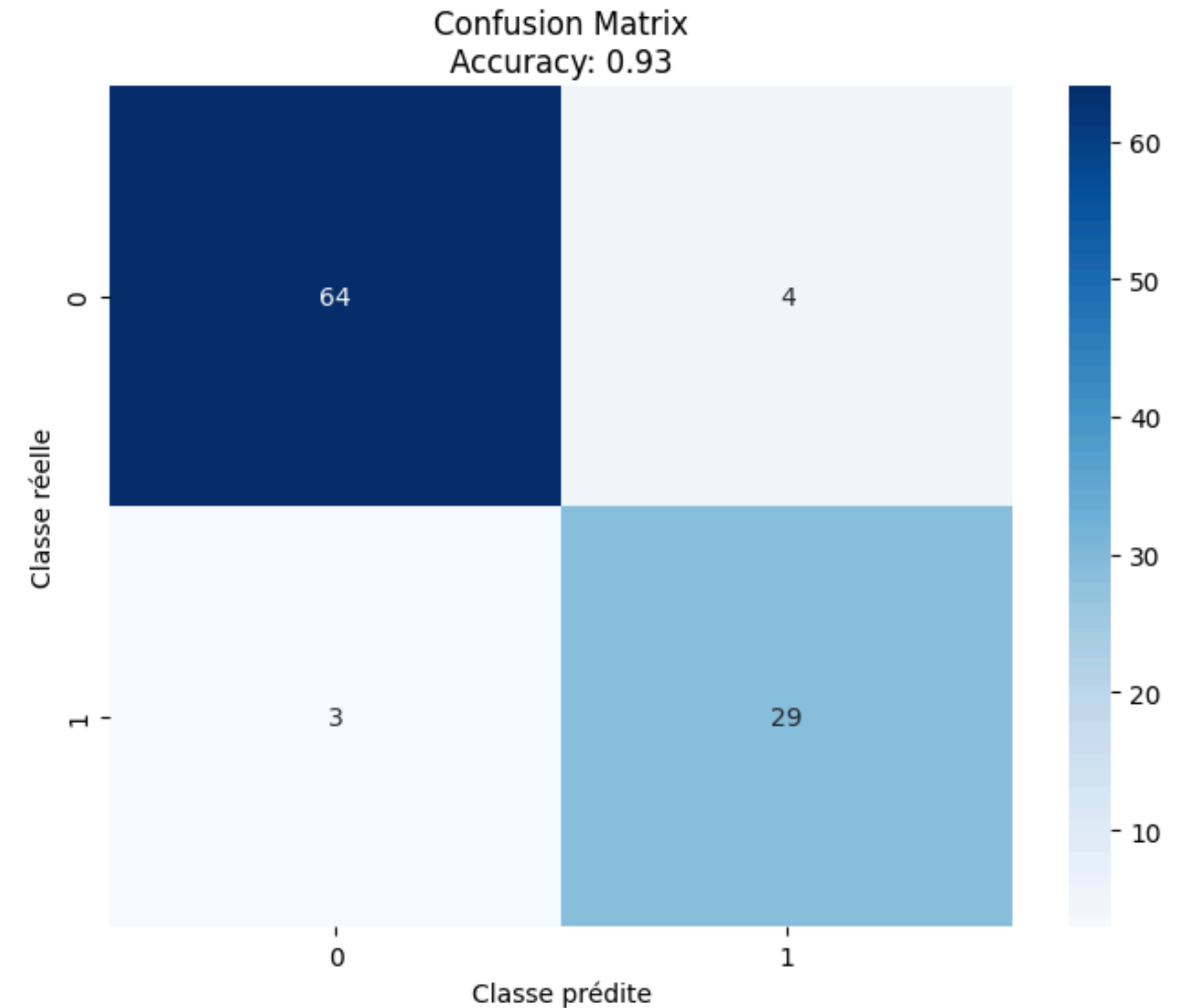
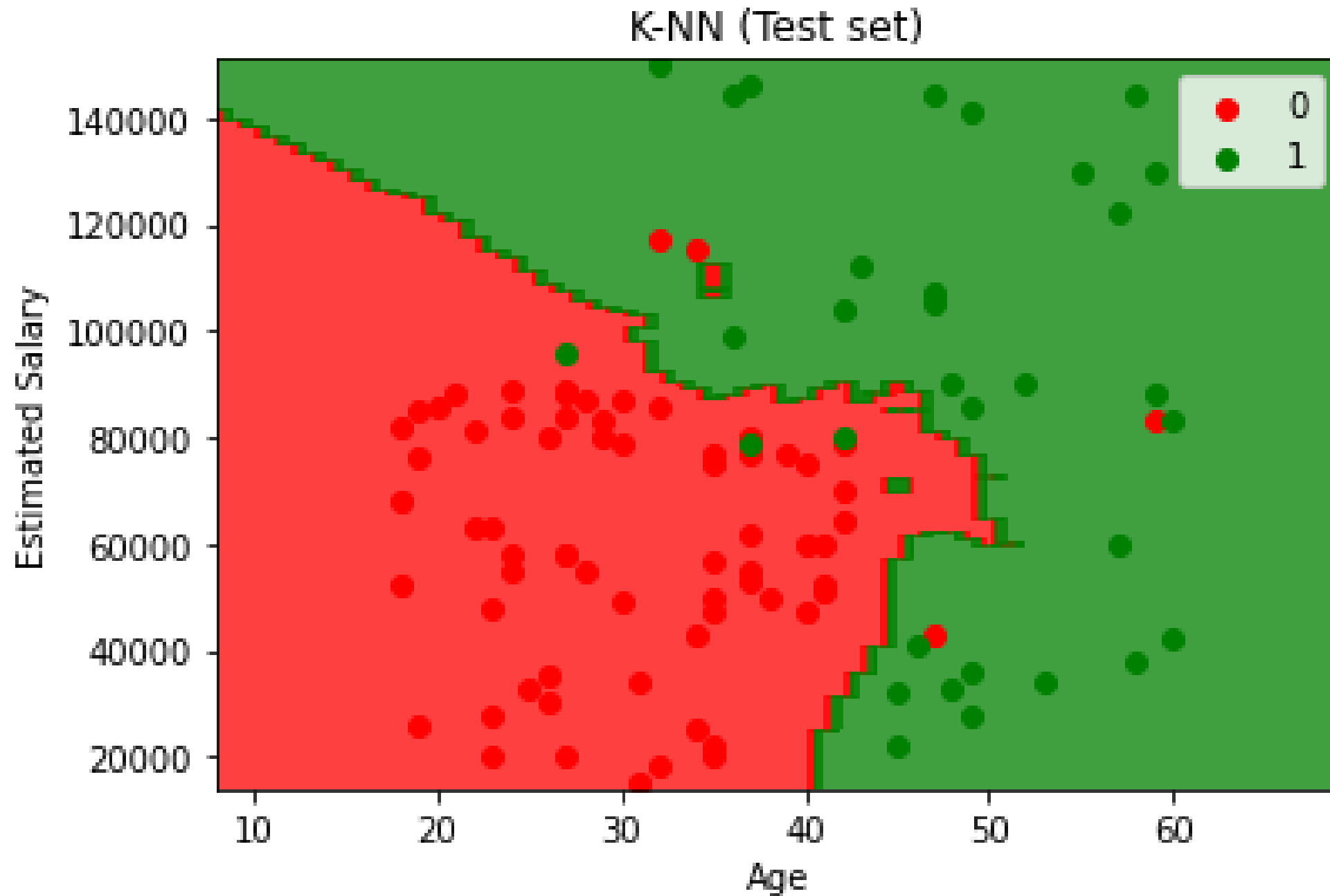
- Inconvénients :

Dépend fortement du choix de k et du type de mesure de distance utilisée. Un mauvais choix peut affecter les résultats  
Les valeurs aberrantes peuvent fausser les résultats, surtout si k est faible.

- Cas d'utilisation: reconnaissance d'Images, systèmes de Recommandation, prédiction de Diagnostics Médicaux

# K-Nearest Neighbors (KNN)

Résultats obtenus :



# Logistic Regression

**Méthode d'apprentissage supervisé** utilisée principalement pour des tâches de classification binaire, où elle prédit la probabilité qu'une observation appartienne à une classe ou à une autre.

## Fonctionnement :

La régression logistique **utilise une fonction sigmoïde pour transformer les valeurs continues en probabilités**, allant de 0 à 1. **Le modèle apprend des coefficients pour chaque caractéristique d'entrée**, qui définissent l'influence de chaque caractéristique sur la probabilité de la classe cible.

Une fois la probabilité calculée, un **seuil (par défaut 0,5)** est appliqué pour classer les observations. Si la probabilité dépasse ce seuil, l'observation est assignée à la classe "1" ; sinon, à la classe "0".

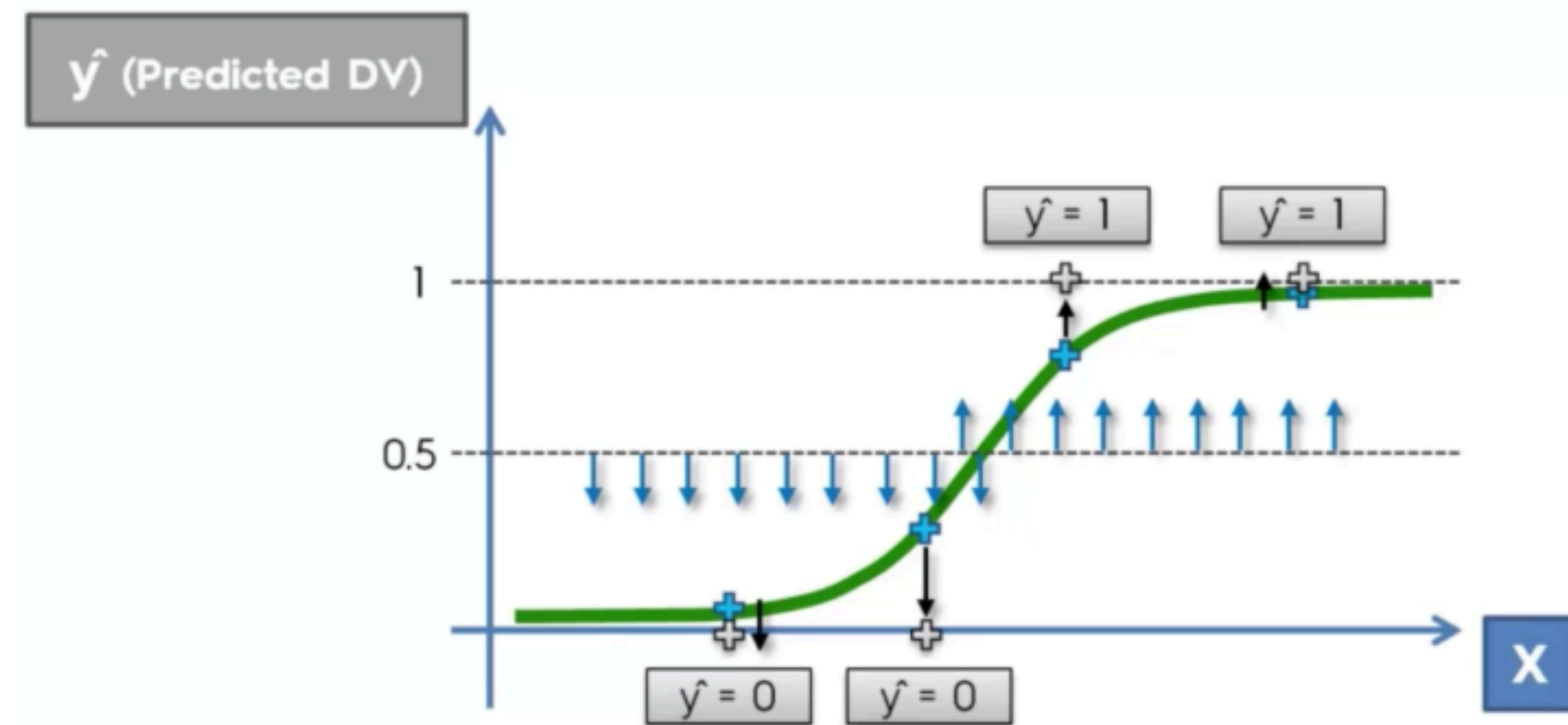
Le modèle ajuste les coefficients en utilisant une méthode de maximisation de vraisemblance, **cherchant à réduire l'écart entre les prédictions et les valeurs réelles**.

## Avantages :

- Les coefficients permettent d'interpréter l'impact de chaque caractéristique sur le résultat final.
- Fournit des probabilités prédictives, utiles pour prendre des décisions informées.

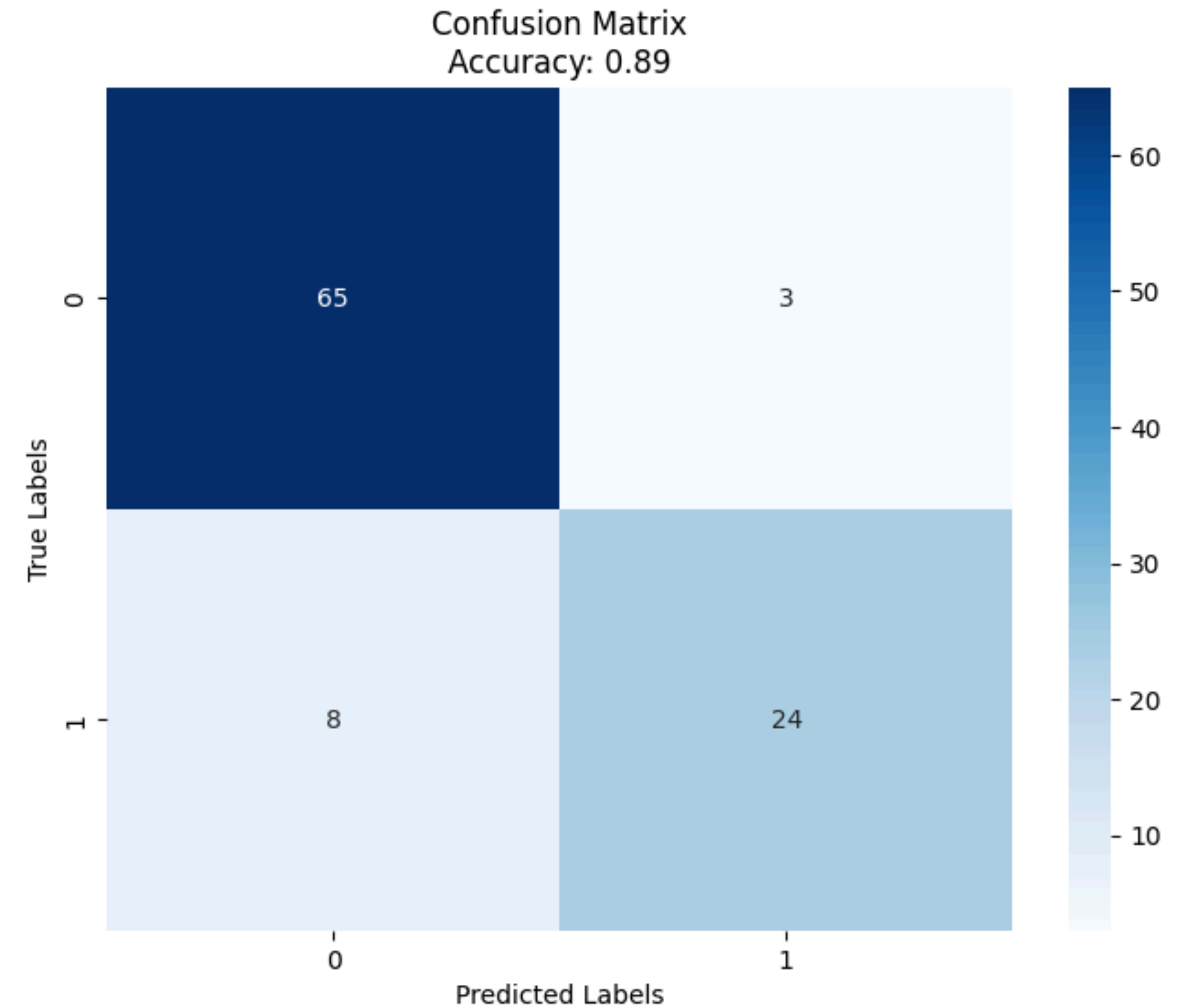
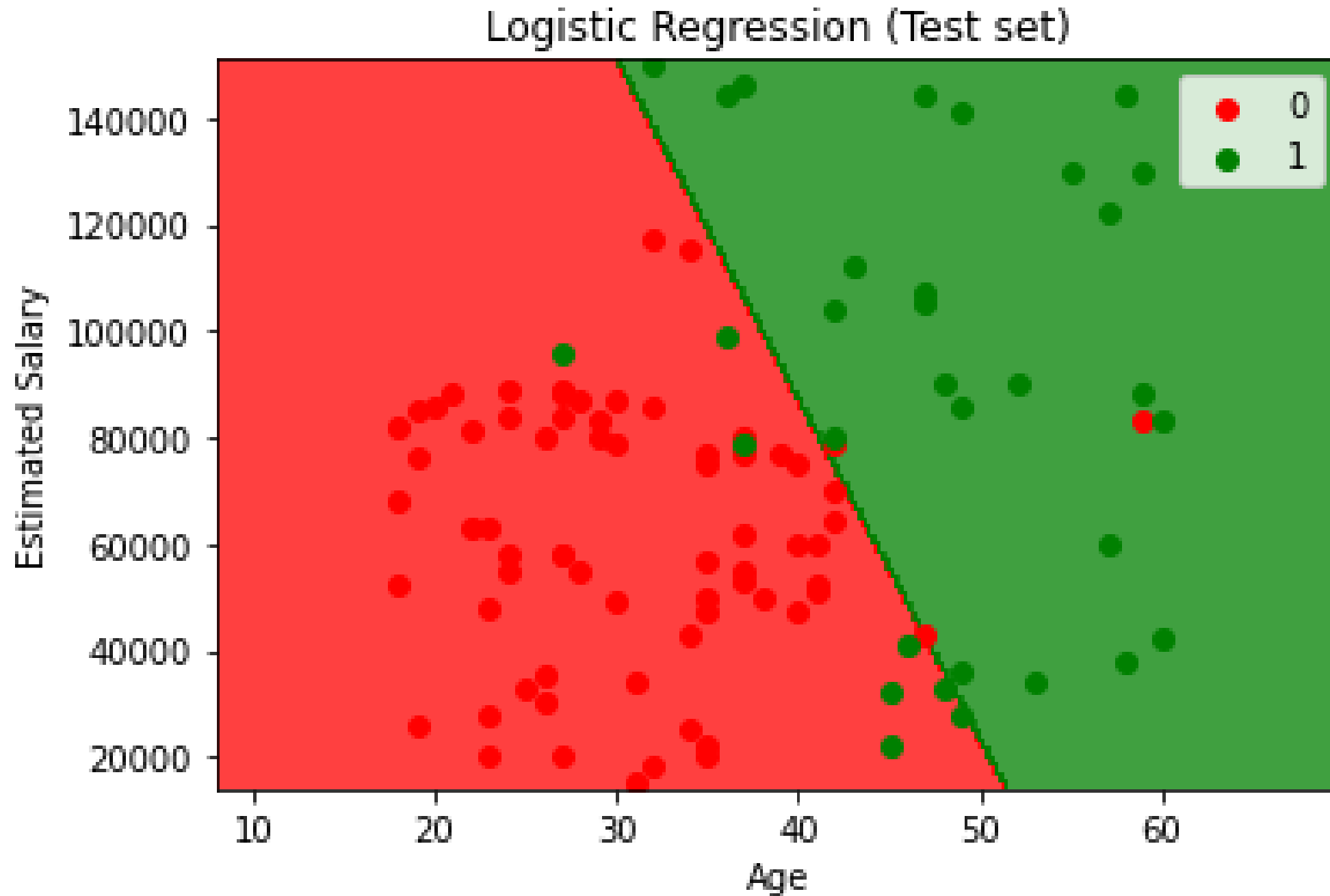
## Inconvénients :

- Linéarité : Supposant une relation linéaire entre les caractéristiques et la probabilité de l'issue, elle peut être moins performante pour des données très complexes ou non linéaires.
- Sensibilité : Les valeurs extrêmes dans les données peuvent influencer significativement les coefficients.



# Logistic Regression

Résultats obtenus :





# Support Vector Machine

**Algorithme d'apprentissage supervisé** utilisé pour la classification et la régression, bien qu'il soit surtout connu pour ses performances dans les tâches de classification.

## Fonctionnement :

**Le SVM cherche à créer un hyperplan qui sépare les classes de données de la manière la plus nette possible.** Dans le cas de données en 2D, cet hyperplan est une ligne qui sépare les classes, mais en dimensions plus élevées, il peut s'agir d'un plan ou d'une hypersurface.

**Le SVM utilise uniquement certains points de données proches de la frontière** entre les classes, appelés vecteurs de support, **pour déterminer l'hyperplan optimal.** Ces points sont importants pour maximiser la marge de séparation entre les classes.

L'objectif du SVM est de **maximiser la distance entre les vecteurs de support et l'hyperplan**, ce qui contribue à une meilleure généralisation du modèle.

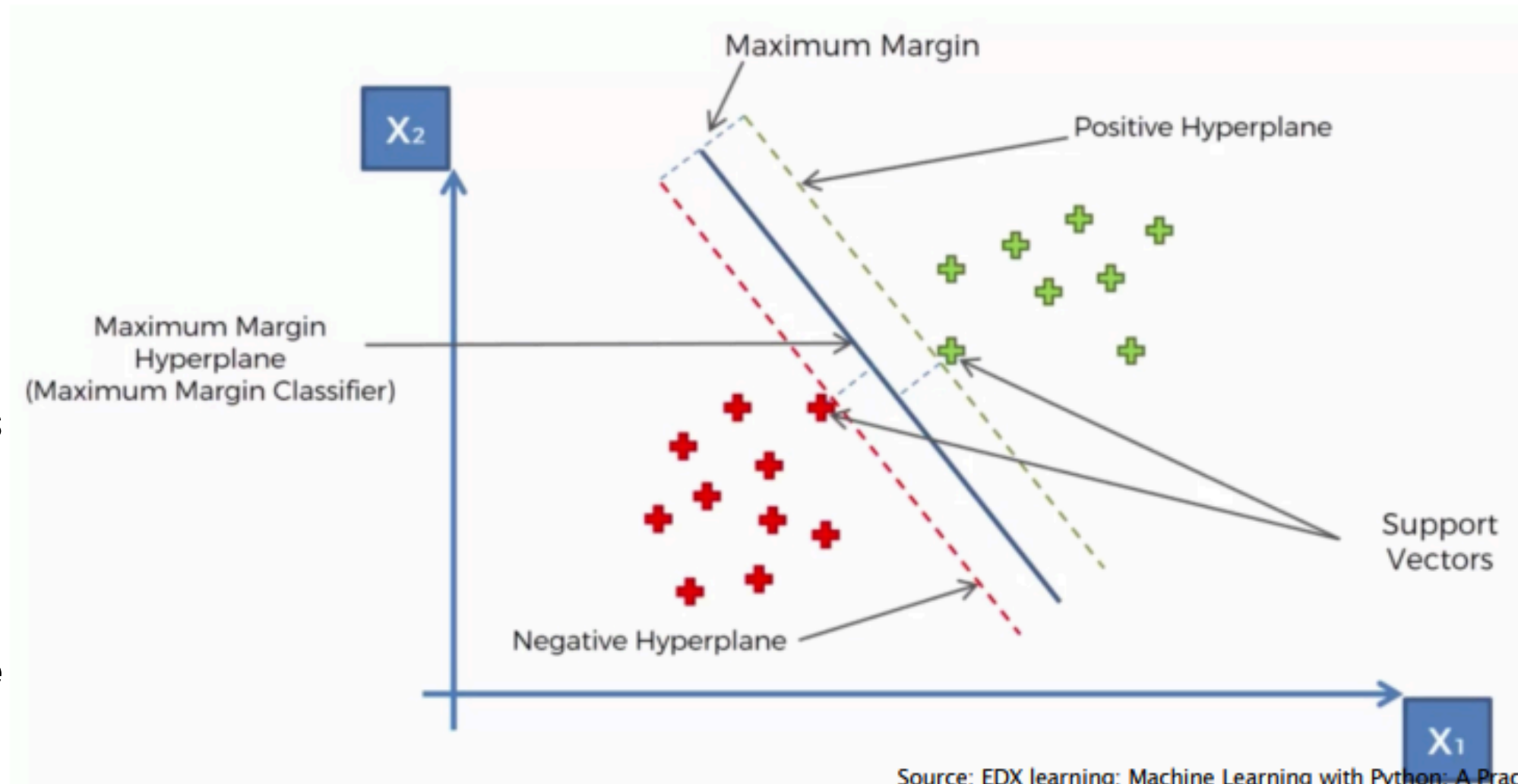
**Lorsque les données ne sont pas linéairement séparables dans un espace de dimension supérieure**, on inclut le linéaire, le gaussien (RBF) et le polynôme.

## Avantages :

- Polyvalent : Peut être utilisé avec différents kernels pour s'adapter à des données linéaires ou non linéaires.

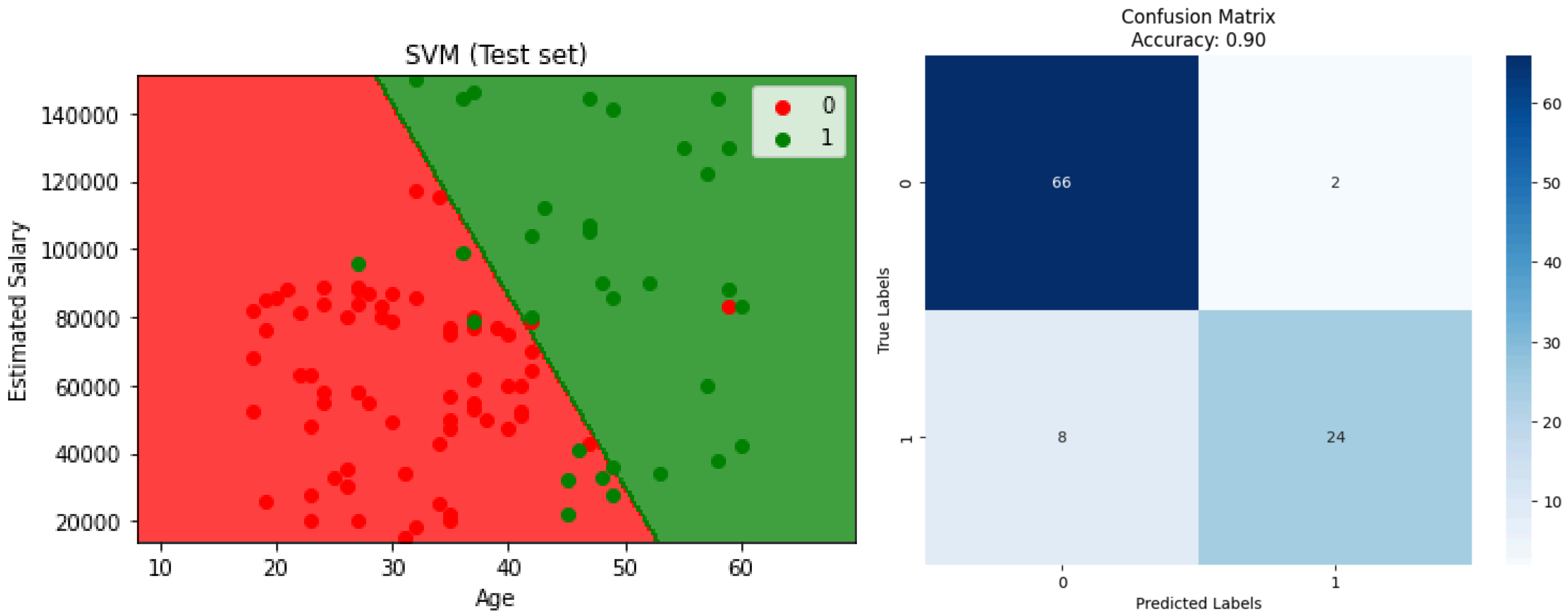
## Inconvénients :

- Choix du Kernel : La sélection du kernel et des paramètres peut être complexe et nécessite souvent des ajustements manuels.



# Support Vector Machine

Résultats obtenus :

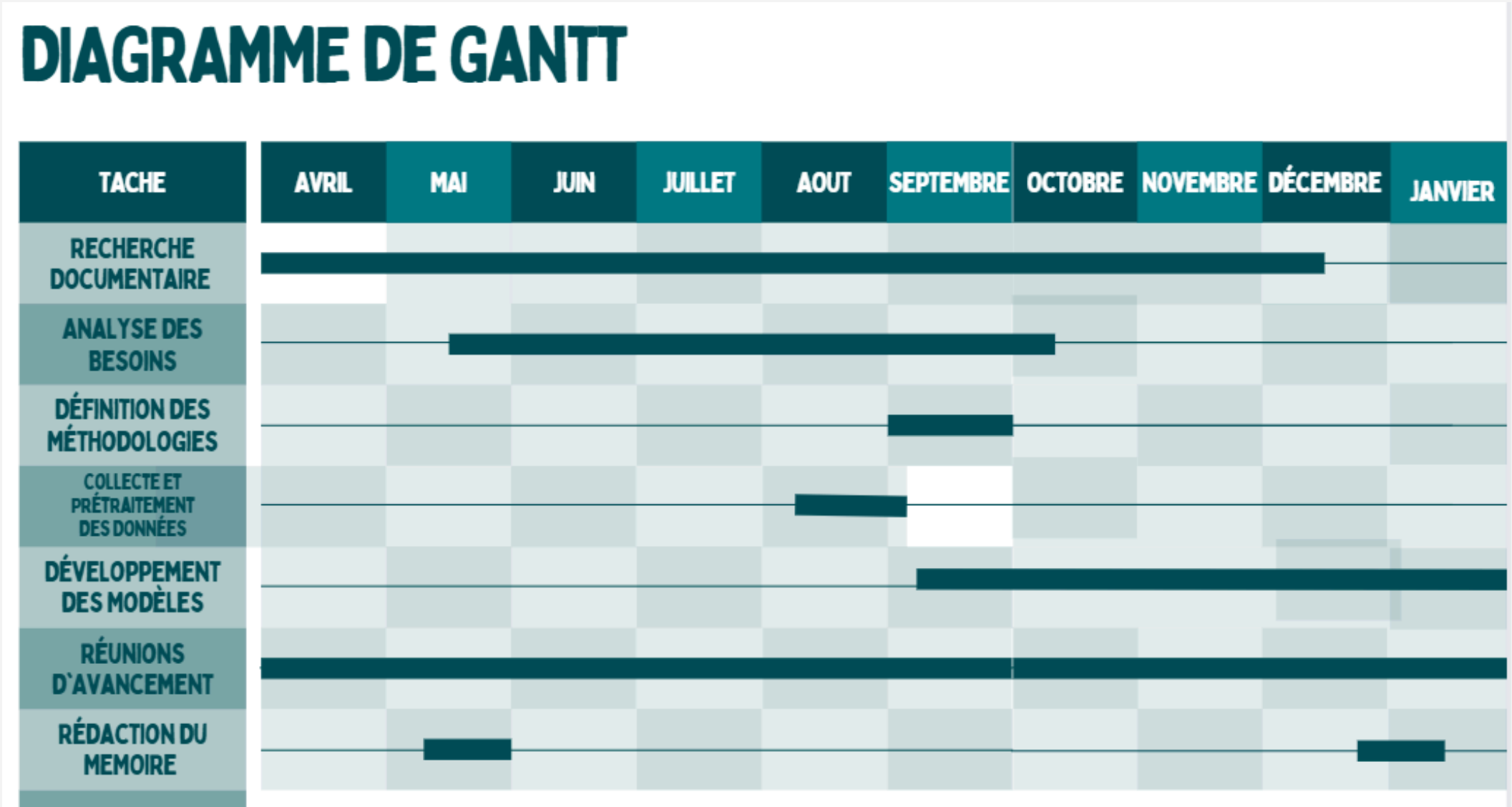




**Les perspectives**

# Nos perspectives

- Construction d'une l'Interface de Traitement du Langage Naturel (NLP) avec Modèle LLM
- Construction d'un Système de Routage Basé sur les Intentions
- Intégration de la Confiance et de la Qualité de l'Expérience (trust-QoE)
- Développer la Preuve de Concept en Conditions Réelles






# Conclusion

Notre projet démontre l'importance de la fiabilité des données dans les dispositifs médicaux, en utilisant cinq modèles de machine learning pour valider leur précision et pertinence. Chaque modèle a apporté une vision complémentaire de la classification, essentielle dans un contexte où la qualité des informations est cruciale pour la prise de décision médicale.

L'intégration de techniques NLP pour analyser les intentions utilisateur permettra également d'améliorer l'expérience. La prochaine étape sera le test en conditions réelles pour évaluer la performance et renforcer la sécurité des données médicales connectées.







**Merci pour votre  
attention !**

