



Devoir Machine Learning :

Prédiction de la présence d'une maladie à partir de symptômes

1. Contexte :

Un centre médical souhaite automatiser la prédiction de maladies probables à partir des symptômes rapportés par les patients
(ex. : *itching, skin_rash, nodal_skin_eruptions*, etc.).

🎯 **Objectif :** classifier automatiquement chaque patient comme **malade** ou **sain**
→ pour aider au diagnostic et accélérer le triage.

2. Préparation des données :

✓ Nettoyage

- Suppression des espaces
- Gestion des valeurs manquantes

✓ Encodage

Transformation des symptômes/maladies en valeurs numériques avec **LabelEncoder**.

✓ Séparation des données

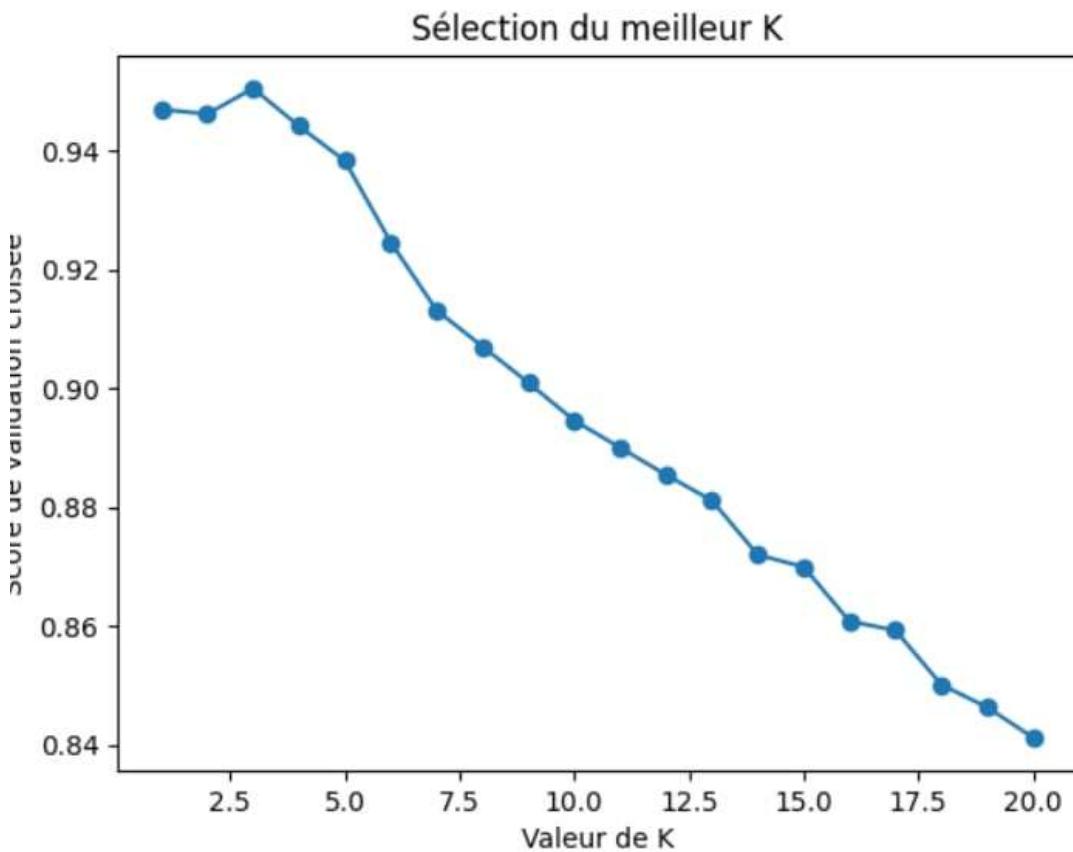
80 % pour l'apprentissage — 20 % pour le test.

3. Modèles testés :

Modèle	Description
KNN	Recherche du "voisin le plus proche". Normalisation nécessaire.
Naïve Bayes	Très rapide, adapté aux données catégorielles.

4. Sélection du meilleur K (KNN) :

Le meilleur K trouvé est : **K = 3**



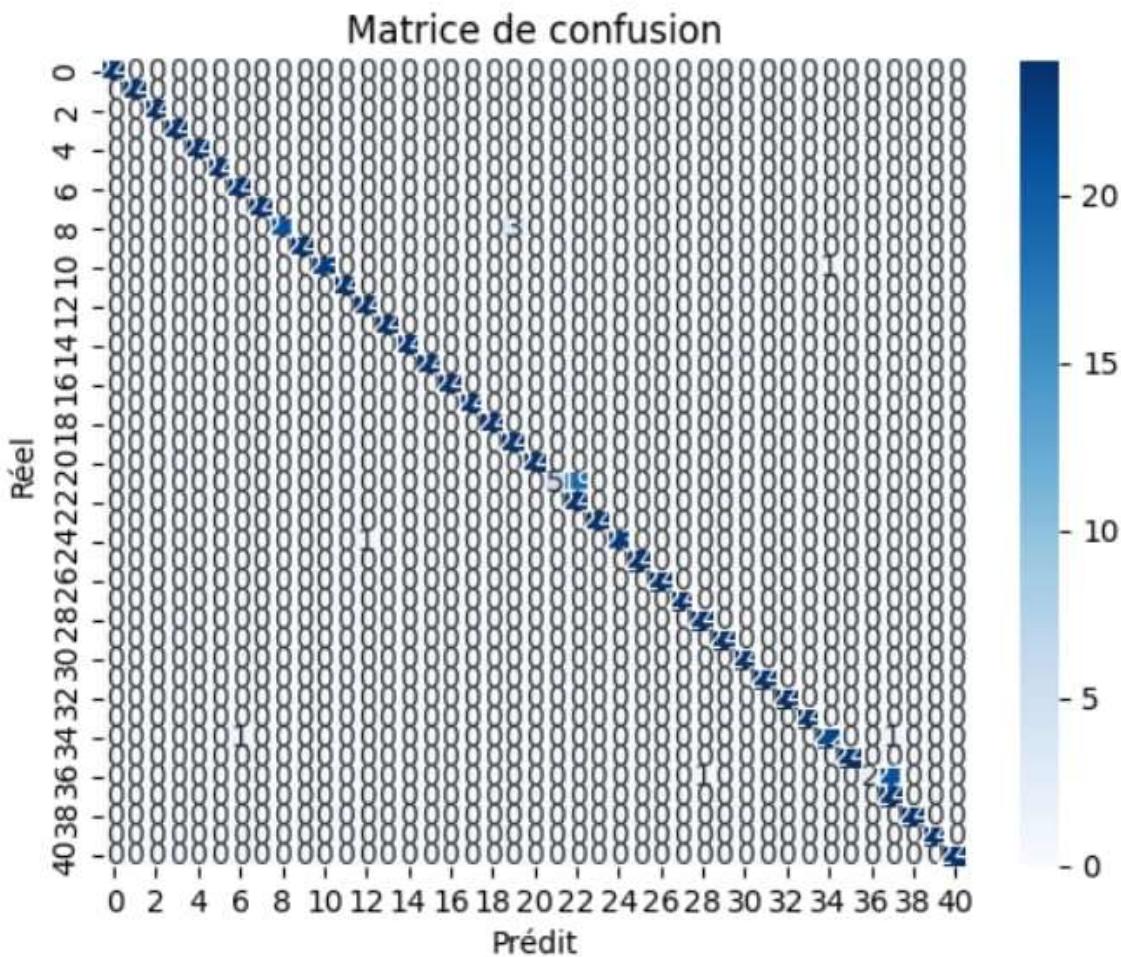
meilleur K trouvé : 3

Pourquoi K=3 ?

→ la précision en validation croisée est maximale pour 3 voisins.

Code : Normalisation + KNN + Score

5. Matrice de confusion (KNN) :



Interprétation

- Beaucoup d'éléments dans la diagonale → **bonnes classifications**
- Les erreurs concernent des maladies rares ou similaires

6. Modèle Naïve Bayes :

Interprétation

- Très rapide
- Précision légèrement inférieure à KNN (~88–90 %)

7. Prédiction sur nouveaux patients :

DataFrame des nouveaux cas : Prédiction nouveaux patients

Interprétation

Les deux modèles prédisent une **infection fongique** pour les deux patients.

8. Variables les plus discriminantes :

Symptôme	Importance
itching	élevée
skin_rash	moyenne
dischromic_patches	moyenne
nodal_skin_eruptions	faible

- itching et skin_rash influencent le plus la classification.

Comparaison KNN vs Naive Bayes

Critère	KNN	Naive Bayes
Précision test	★ 95%	88–90%
Rapidité	Moyenne	★ Très rapide
Déploiement	Moyen	★ Facile
Robustesse petites données	Moyenne	★ Bonne
Complexité	Moyenne	★ Faible

Résultat des nouvelles prédictions :

KNN :

	Symptom_1	Symptom_2	Symptom_3
0	0	0	0
1	0		0

Naive Baises :

➊ Prédictions :

Patient 1 → 3

Patient 2 → 3

9. Conclusion : méthode la plus pertinente :

KNN est le meilleur modèle pour cette étude :

- Précision très élevée
- Très bon sur données de symptômes
- Capte bien la similarité entre patients

Naive Bayes reste utile

Mais moins précis → adapté pour un modèle léger et très rapide.