

**UNIVERSITE DU BURUNDI**

Campus KIRIRI

FSI – BAC IV – TIC

**Travail Pratique** : Détection du Cancer du Côlon à partir de Données Génomiques

**Cours** : Machine Learning

Par:

* BAKIZE BLESS ALLEGRE
* BASHEMEZE RICHARD
* BENIMANA SOSTHENE
* RUNYENYERI DIAMANT
* CIZA JOELLE
* NZIZA CHRISTA BELLA
* IGIRANEZA CLERY
* IRAKOZE JEAN DE DIEU
* KANEZA GRETTA
* IRAKOZE PACIFIQUE

# Objectif

L'objectif de ce projet est de développer un modèle de machine learning capable de **prédire l'état tissulaire** (normal ou tumoral) à partir de l'expression d'un unique gène, sélectionné parmi plusieurs, en utilisant un ensemble de données génomiques.

# Méthodologie

## Chargement des Données

* Les données sont chargées à partir du fichier CSV colon\_cancer.csv.
* Le fichier contient des mesures d'expression génique pour différents échantillons.
* Le label cible (tissue\_status) est transformé en valeurs numériques :
  + 0 pour "normal",
  + 1 pour "tumoral".
* Les données manquantes sont supprimées pour éviter toute perturbation lors de l'entraînement.

## Sélection du Meilleur Gène

* Pour chaque gène, un **test t de Student** est réalisé pour comparer les niveaux d'expression entre tissus normaux et tumoraux.
* La **p-value** est utilisée pour évaluer la significativité de la différence d'expression.
* Le gène avec la **p-value la plus faible** est sélectionné comme étant le plus discriminant.

## Préparation des Données

* Seule l'expression du **gène sélectionné** est utilisée comme caractéristique (feature) pour la classification.
* Les données sont **normalisées** avec StandardScaler afin de standardiser la moyenne et l'écart-type.

## Séparation des Données

* Les données sont divisées en un **ensemble d'entraînement** (80%) et un **ensemble de test** (20%).

## Entraînement et Évaluation des Modèles

* Cinq algorithmes de machine learning sont testés :
  + Arbre de Décision
  + Forêt Aléatoire
  + Régression Logistique
  + Machine à Vecteurs de Support (SVM)
  + k-Plus Proches Voisins (KNN)
* Chaque modèle est évalué par **validation croisée (5-fold)** sur l'ensemble d'entraînement.
* Le modèle avec la **meilleure précision moyenne** est sélectionné pour l'entraînement final.

## Entraînement Final et Test

* Le meilleur modèle est **réentraîné** sur l'ensemble d'entraînement complet.
* Sa **précision** est mesurée sur l'ensemble de test.

## Sauvegarde des Résultats

* Le modèle entraîné et l'outil de normalisation (StandardScaler) sont sauvegardés au format .pkl via joblib.
* Les fichiers sont mis à disposition pour téléchargement via Google Colab.

# Déploiement du Modèle

## Backend - Django

Le modèle est déployé dans une application Django :

* **Chargement du modèle et du scaler** au démarrage de la vue PredictCancerView grâce à joblib.
* **Route POST /api/predict/** : permet d'envoyer une ou plusieurs entrées pour obtenir des prédictions sur :
  + Le statut tumoral (Normal ou Tumoral)
  + Le niveau de confiance (probabilité de la prédiction)
* **Route GET /api/predictions/** : permet de **lister l'historique** des prédictions réalisées, enregistrées en base de données.

## Interface Frontend

Une interface utilisateur a été développée pour interagir avec l'API :

* **Formulaire dynamique** :
  + Permet d'ajouter manuellement plusieurs patients et leurs valeurs d'expression du gène (UGP2).
  + Vérification de la validité des entrées avant envoi.
* **Prédiction** :
  + Envoie les entrées au backend via la route /api/predict/.
  + Affiche immédiatement les résultats (statut tumoral + confiance) dans un tableau.
* **Historique** :
  + Permet d'afficher l'**historique complet** des prédictions enregistrées en base.
  + Utilisation d'une boîte de dialogue (Dialog) pour afficher les prédictions passées.

# Résultats

* **Meilleur gène sélectionné** : UGP2
* **Modèle choisi** : Logistic Regression avec une précision de 0.9735
* Le modèle est disponible **en ligne** via <https://tp-machine-learning-groupe-2.onrender.com/>
* **Historique conservé** : toutes les prédictions sont sauvegardées pour consultation ultérieure.

Ces résultats démontrent que même en utilisant l'expression d'un **seul gène**, il est possible d'obtenir une **bonne capacité de discrimination** entre tissus normaux et tumoraux.

# Conclusion

Ce projet montre qu'une **analyse statistique simple** combinée à **des modèles de machine learning classiques** peut donner des résultats intéressants en détection précoce du cancer à partir de données biologiques.  
Il constitue une bonne base pour des travaux plus avancés en bioinformatique et en classification supervisée.