**Bujumbura International University**

**Bac2 GL/BST**

**Cours Big Data**

**Projet Examen: Medical Insurance Charges**

**Travail académique présenté par :**  
MUKIZA Chris Alick (BIU-EED-11-059-2023)

MATONDO LENZO DANIEL (BIU-EED-11-121-2023)

KEZAKIMANA KENTHA (BIU-EED-11-xxx-2023)

**Encadré par :**  
Dr Thierry MUGENZI

**Année académique :**  
2024 – 2025

**1. Introduction et Méthodologie**

**1.1 Contexte et Objectifs**

Ce projet pédagogique applique un workflow complet de data science sur des données réelles d'assurance santé. Le dataset de 1,338 patients avec 7 variables permet de résoudre trois problèmes de machine learning distincts.

**Objectifs :**

* Prédire les coûts d'assurance (régression)
* Classifier le statut tabagique (classification binaire)
* Identifier la région d'origine (classification multiclasse)

**1.2 Méthodologie**

Approche structurée en 5 phases :

1. Analyse exploratoire et prétraitement
2. Modélisation par régression
3. Classification binaire
4. Classification multiclasse
5. Synthèse et présentation

**1.3 Variables d'étude**

**Indépendantes :** âge, sexe, BMI, enfants, région  
**Dépendantes :** charges (régression), smoker (binaire), région (multiclasse)

**2. Analyse Exploratoire et Résultats**

**2.1 Découvertes Clés**

**Impact du tabagisme :**

* Coût moyen fumeurs : 32,050$ vs non-fumeurs : 8,434$
* **Multiplicateur : 3.8×** pour les fumeurs

**Distribution :**

* 20.5% fumeurs, 79.5% non-fumeurs
* Répartition équilibrée par genre et région

**2.2 Performances des Modèles**

**RÉGRESSION (Prédiction coûts) :**

| Modèle | R² | RMSE |
| --- | --- | --- |
| **Régression Linéaire** | **0.164** | **11,107$** |
| Arbre de Décision | -0.807 | 22,415$ |
| Random Forest | -0.003 | 12,115$ |

**CLASSIFICATION BINAIRE (Fumeur) :**

| Modèle | Accuracy | ROC AUC |
| --- | --- | --- |
| **Random Forest** | **0.873** | **0.850** |
| Régression Logistique | 0.794 | 0.822 |
| SVM | 0.794 | 0.822 |

**CLASSIFICATION MULTICLASSE (Région) :**

| Modèle | Accuracy |
| --- | --- |
| **Random Forest** | **0.250** |
| KNN | 0.216 |
| Arbre de Décision | 0.205 |

**3. Discussion et Analyse**

**3.1 Insights Techniques**

**Sur-apprentissage :**

* Modèles complexes (Arbre, Random Forest) sur-apprennent sévèrement
* Régression linéaire plus stable mais performance modeste (R²=0.164)

**Gestion déséquilibre :**

* class\_weight='balanced' essentiel pour éviter la prédiction systématique de la classe majoritaire
* Stratification préservant la distribution des classes

**3.2 Insights Business**

**Priorité tabagisme :**

* Facteur de risque n°1 (impact ×3.8 sur coûts)
* **Recommandation :** Programmes de prévention prioritaires

**Hiérarchie variables :**

1. Statut tabagique
2. Âge
3. BMI

**3.3 Défis Rencontrés**

**Limitations données :**

* Variables insuffisantes pour prédiction région
* Effectif limité pour certains sous-groupes
* Relations potentiellement non-linéaires

**4. Conclusion et Perspectives**

**4.1 Bilan**

**Réussites :**

* Pipeline ML complet et fonctionnel
* Insights business actionnables
* Gestion efficace des défis techniques

**Performances :**

* Bonne discrimination fumeur/non-fumeur (AUC=0.85)
* Performance modeste mais réaliste en régression
* Difficulté prédiction région (proche du hasard)

**4.2 Recommandations**

**Améliorations données :**

* Ajout d'antécédents médicaux
* Variables de style de vie
* Données socio-économiques

**Approches techniques :**

* Feature engineering (catégories BMI, tranches âge)
* Optimisation hyperparamètres
* Techniques ensemble learning avancées

**4.3 Conclusion**

Ce projet démontre l'application réussie d'un workflow data science complet. Malgré les limitations du dataset, l'étude génère des insights business valorisables et établit une base solide pour des améliorations futures.

**Valeur principale :** Démonstration que la simplicité algorithmique peut surpasser des approches complexes face au risque de sur-apprentissage.