



Ecole Supérieure Polytechnique (ESP)

Département Statistique et Ingénierie des Données (SID)

Projet Machine Learning

Aliou Diallo

1 Introduction

Dans cette étude, nous allons développer un modèle de machine learning supervisé basé sur la régression logistique, afin de prédire la probabilité qu'un ménage soit classé comme pauvre ou non. Cette prédiction sera réalisée à partir d'un ensemble de variables socio-économiques et de conditions de vie des ménages. La régression logistique est une méthode bien adaptée pour les problèmes où la variable dépendante est binaire, ici la variable "pauvre" (Variable dépendant) qui distingue les ménages pauvres des non-pauvres. Le modèle s'appuiera sur une série de facteurs socio-économiques tels que le revenu, la taille du ménage, le niveau d'éducation du chef de ménage, le statut professionnel, l'accès aux services de base (comme l'électricité, l'eau potable, et l'accès à la santé), ainsi que des informations géographiques (comme la région, la commune, et la wilaya) et des conditions de vie (par exemple, les dépenses en alimentation, logement, transport, et santé). L'objectif est de comprendre comment ces facteurs interagissent pour influencer le statut de pauvreté des ménages. Une fois le modèle construit, il pourra non seulement prédire les ménages à risque de pauvreté, mais aussi offrir des informations essentielles sur les variables ayant le plus grand impact sur la pauvreté. Cela pourrait aider à orienter les politiques publiques et les programmes de soutien aux ménages les plus vulnérables.

2 Variables Indépendante

La base de données utilisée est celle de l'Enquête Permanente sur les Conditions de Vie des Ménages (EPCVM). Les variables sont :

Variable	Description
Milieu	Milieu de résidence
Taille	Taille du ménage
Sitmatcm	Situation matrimoniale du chef de ménage
Nivcm	Niveau d'éducation du chef de ménage
Depsante	Dépenses sur la santé
Logement	Type de logement
Transp	Mode de transport

Table 1: Description des variables étudiées