Prédiction de l'Admission Universitaire

Mathématiques Pour La Sciences Des Données

Présenté par :

- Manal Rhoni Aref
- Souhaila Benaouate

Introduction & Objectif

Contexte

- L'admission en Master est un processus hautement compétitif.
- Les étudiants et les universités ont besoin d'outils pour évaluer objectivement les dossiers.

Objectif du Projet

- Développer un modèle prédictif capable de classer automatiquement un candidat comme **"Admis"** ou **"Rejeté"**.
- Identifier les critères académiques les plus déterminants (CGPA, GRE, etc.).

Notre Approche

1. **Exploration :** Analyse et nettoyage des données.
2. **Prétraitement :** Normalisation et équilibrage des classes.
3. **Modélisation :** Comparaison entre **Régression Logistique** et **Arbre de Décision**.

Présentation du Dataset

Source des Données

- Fichier : `Admission_Prediction.csv`
- Taille : 500 candidats (Lignes) x 8 variables (Colonnes).

Les Variables (Features)

- **Scores Académiques :** GRE Score (sur 340), TOEFL Score (sur 120), CGPA (sur 10).
- **Dossier Qualitatif :** University Rating (1-5), SOP (Statement of Purpose), LOR (Lettre de Recommandation).
- **Recherche :** Expérience en recherche (0 ou 1).

La Cible (Target)

- **Variable d'origine :** `Chance of Admit` (Probabilité continue entre 0 et 1).
- **Transformation :** Conversion en **Classe Binaire** avec un seuil de **0.75**.
  - `>= 0.75` : **Admis (1)**
  - `< 0.75` : **Rejeté (0)**

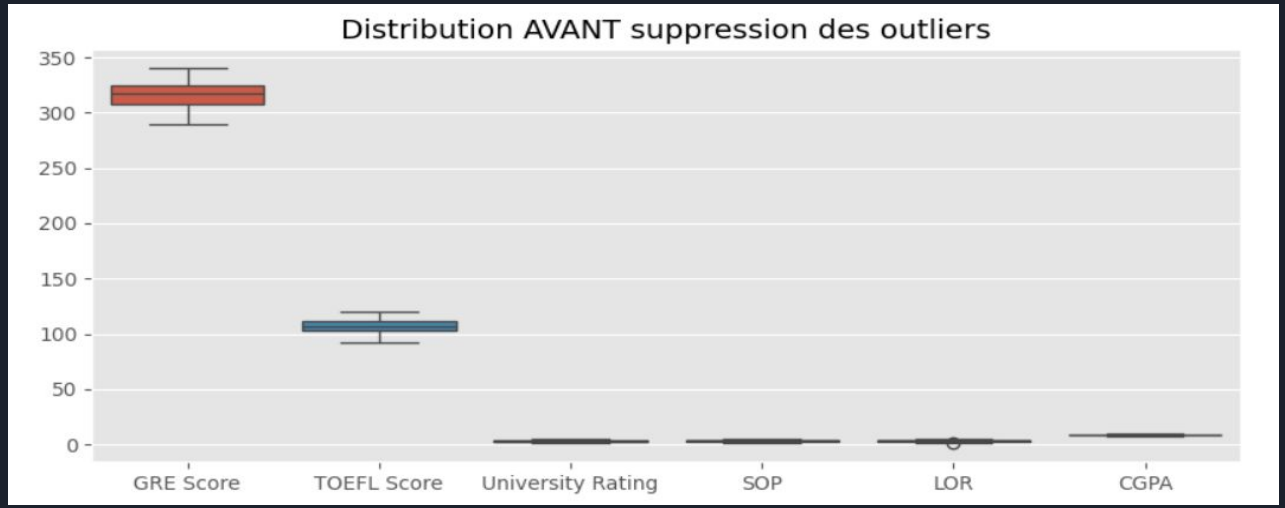
Analyse Exploratoire & Nettoyage

1. Valeurs Manquantes

- Analyse : Vérification de la complétude du dataset.
- Résultat : Aucune valeur manquante détectée (dataset complet).

2. Traitement des Outliers (Valeurs Aberrantes)

- **Méthode :** Utilisation de l'intervalle interquartile (**IQR**).
- **Pourquoi ?** Les valeurs extrêmes faussent la Régression Logistique.
- **Action :** Suppression automatique des lignes aberrantes pour garantir un modèle robuste.



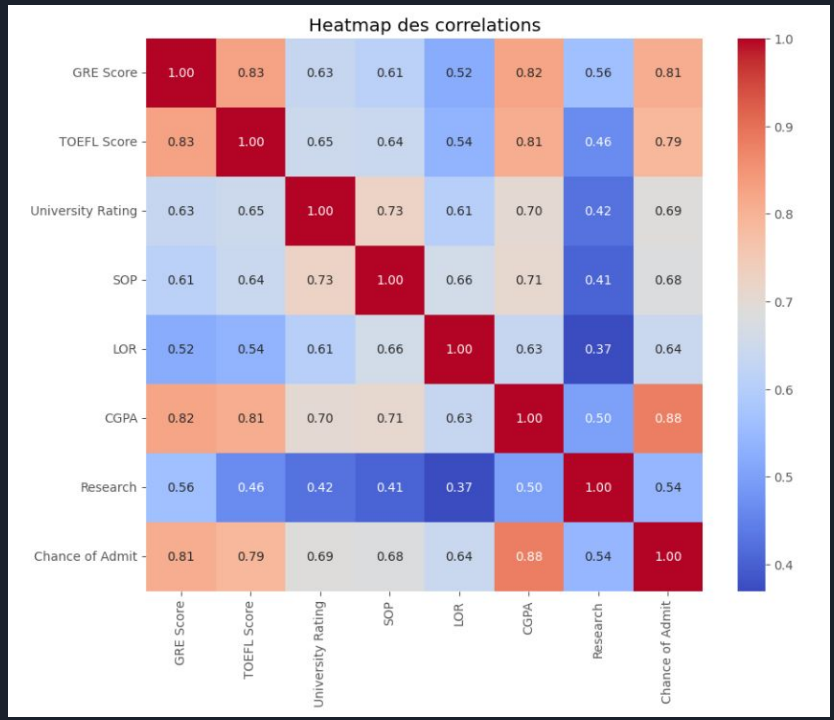
Corrélations des Variables

Analyse de la Heatmap

- Nous avons analysé les liens entre les critères d'admission.

Observations Clés

1. **Forte corrélation :** Le **CGPA** (Moyenne générale) et le **GRE Score** sont très fortement liés à l'admission.
2. **Impact de la Recherche :** Avoir une expérience de recherche augmente significativement les chances.
3. **Indépendance :** Les lettres de recommandation (LOR) sont moins corrélées aux scores bruts, apportant une information complémentaire.



Prétraitement Avancé

1. Normalisation (StandardScaler)

- **Problème :** Le GRE va jusqu'à 340, le CGPA jusqu'à 10.
- **Solution :** Mise à l'échelle (Moyenne = 0, Écart-type = 1).
- **Nécessité :** Indispensable pour la convergence de la Régression Logistique.

2. Gestion du Déséquilibre (Oversampling)

- **Constat :** Il y a naturellement moins d'admis que de rejetés (ou inversement selon le seuil).
- **Technique utilisée : Random Oversampling** (Sur-échantillonnage aléatoire).
- **Principe :** Dupliquer aléatoirement des exemples de la classe minoritaire pour que le modèle apprenne équitablement les deux cas.
- **Précaution :** Appliqué uniquement sur le jeu d'entraînement (Train Set) pour éviter toute fuite d'information (Data Leakage).

Modélisation & Algorithmes

Nous avons entraîné et comparé deux algorithmes majeurs vus en cours :

Modèle A : Régression Logistique

- **Type :** Modèle Linéaire.
- **Pourquoi ?** Simple, rapide et très efficace quand les relations sont proportionnelles (ex: plus la note monte, plus la chance monte).

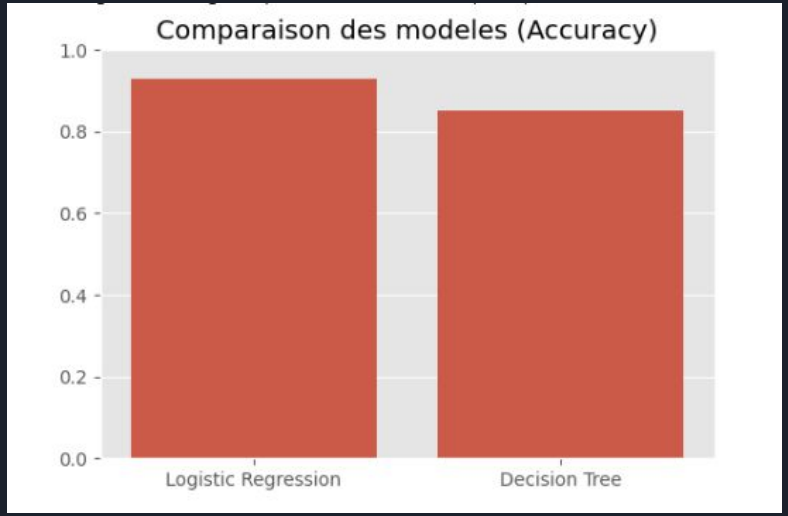
Modèle B : Arbre de Décision

- **Type :** Modèle Non-Linéaire.
- **Paramètres :** Critère `Gini`, Profondeur maximale limitée pour éviter le sur-apprentissage.
- **Pourquoi ?** Capable de capturer des règles de décision complexes ("Si CGPA > 9 et Recherche = 1 Alors...").

Résultats & Comparaison

Tableau de Performance (sur Test Set)

Modèle	Accuracy (Exactitude)	Précision
Régression Logistique	**~93%**	Élevée
Arbre de Décision	~85%	Moyenne



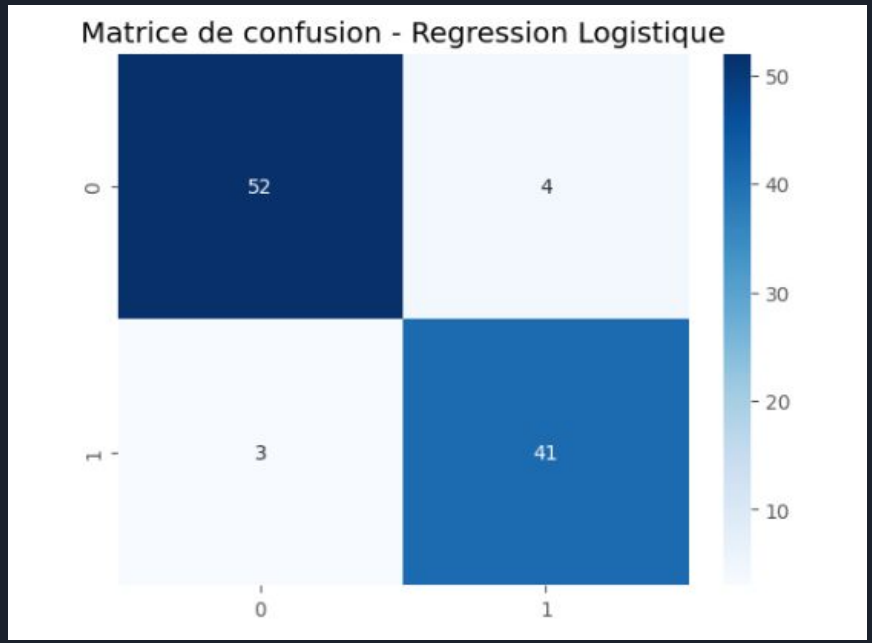
Analyse

- La **Régression Logistique** obtient les meilleurs résultats avec une accuracy de 93%.
- Cela confirme que le processus d'admission suit une logique majoritairement **linéaire** (les notes sont le facteur dominant).

Analyse de la Matrice de Confusion

Focus sur la performance de notre meilleur modèle (Régression Logistique).

- **Vrais Positifs (TP) :** Candidats admis correctement détectés.
- **Faux Négatifs (FN) :** Candidats admis que le modèle a ratés.
- **Résultat :** Le modèle présente un excellent équilibre, ce qui signifie qu'il est fiable pour les deux classes.



Conclusion & Perspectives

Bilan du Projet

- Objectif atteint avec une précision supérieure à **90%** (Régression Logistique).
- La **suppression des outliers** et l'**oversampling** ont été des étapes décisives pour la performance.
- Le **CGPA** est confirmé comme le critère n°1 pour l'admission.

Perspectives

- Élargir le dataset (actuellement 500 lignes) pour plus de robustesse.
- Tester des modèles d'ensemble (Random Forest) pour améliorer encore le score de l'Arbre de Décision.
- Explorer des techniques de rééchantillonnage plus avancées (comme SMOTE).