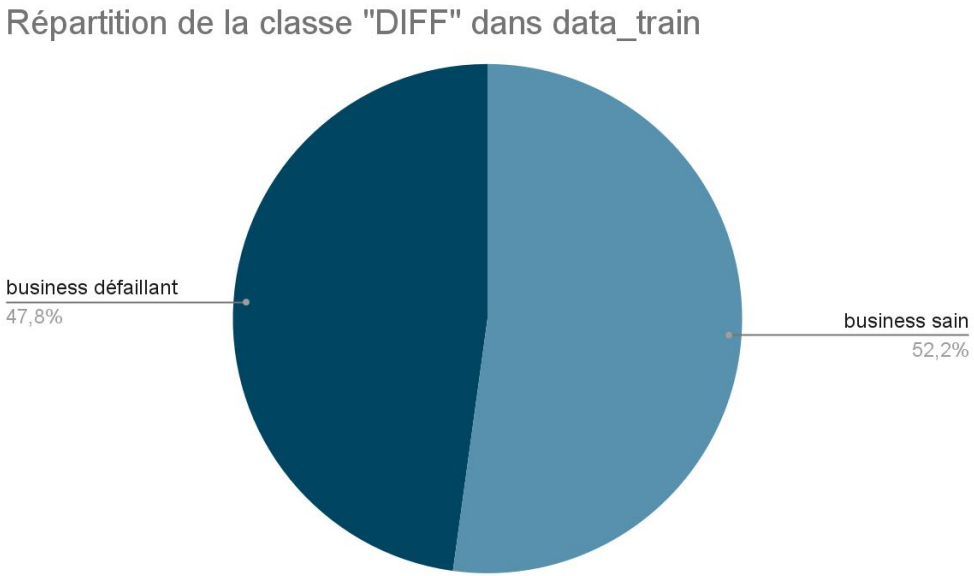
Projet Classification : Exploitations agricoles

Maghraoui Nawfal et Yassine Ait Said

Description de la base de données et Problématique

- R2 : capitaux propres / capitaux permanents,
- R14 : dette à long et moyen terme / produit brut,
- R17 : frais financiers / dette totale,
- R32 : (excédent brut d'exploitation - frais financiers) / produit brut
- DIFF : la variable difficulté de paiement (0=sain et 1=défaillant)

❖ Prédiction de la variable DIFF pour une base de donnée (R2,R14,R17,R32) en entraînant un modèle basé sur les algorithmes de classification.



Approche :

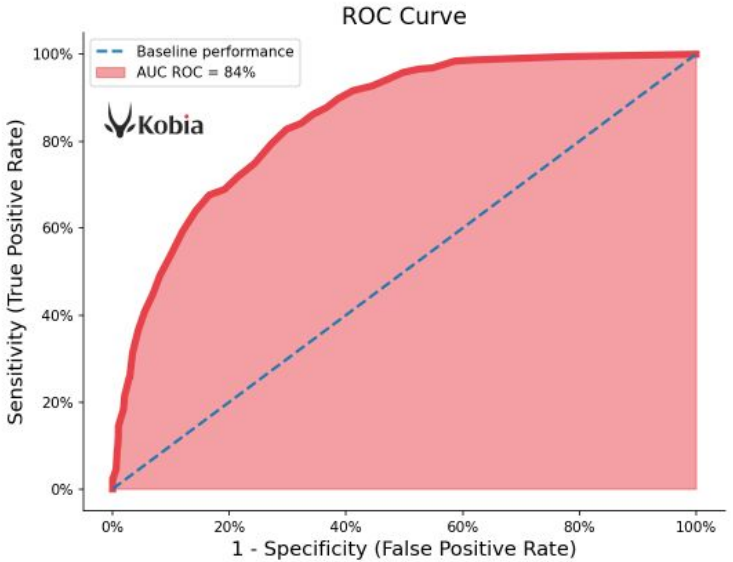
- Comparaison des différents modèles de classification en utilisant la cross-validation et les courbes ROC et l'AUC.
- Choix de l'algorithme adapté
- Prédiction de la variable DIFF

Méthodes de validation :

- DATA SPLIT
- Leave One Out Cross-Validation(LOOCV)
- K-FOLD Cross-Validation

Outils de validation

- Matrice de confusion
- Visualisation de la courbe de ROC
- AUC
- Error rate
- Indice de Gini

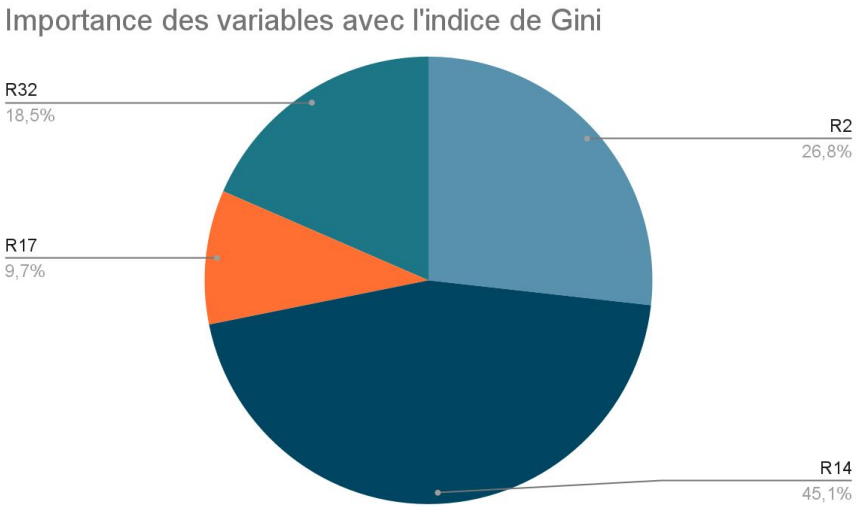


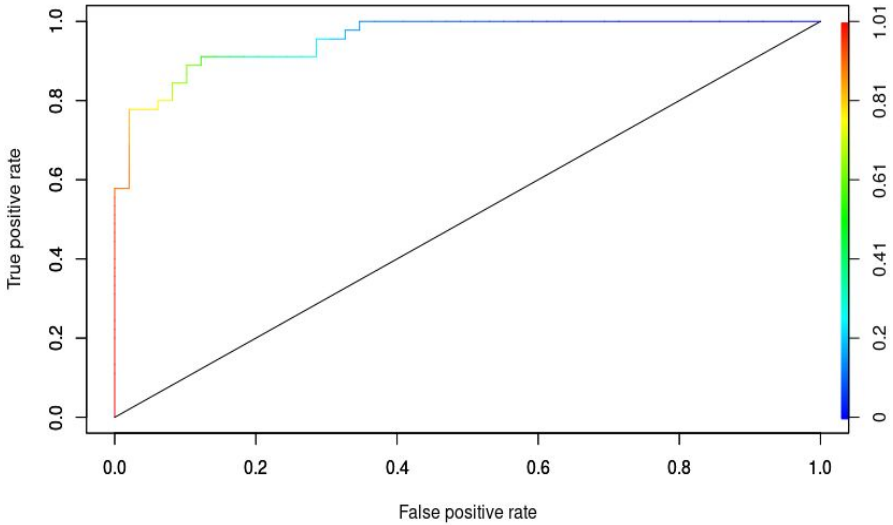
True Positive (TP)	**False Positive (FP)**
False Negative (FN)	**True Negative (TN)**

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Résultats : Courbes ROC et seuils





- R2 : capitaux propres / capitaux permanents,
- R14 : dette à long et moyen terme / produit brut,
- R17 : frais financiers / dette totale,
- R32 : (excédent brut d'exploitation - frais financiers) / produit brut

modèle : Random Forest

Résultats : Approches basées sur un modèle.

Modèle	erreur moyenne	AUC moyen
Régression logistique	0.129	0.94
LDA	0.145	0.94
QDA	0.131	0.942
Bayes naïf	0.145	0.938

Résultats : Approches de type prototype.

Modèle	erreur moyenne	AUC moyen
KNN	0.132	0.934
D.TREE	0.148	0.89
R.FOREST	0.127	0.932

Choix du modèle et Prédiction

- Dans ce cas d'étude la régression logistique semble la plus appropriée pour prédire la variable DIFF.
- Entraîner ce modèle sur 95 % de données et le tester sur les 47 observations qui restent.
- Erreur = 0.042 & AUC = 0.964

Conclusion

Notre étude nous a amené à considérer que la régression logistique est la plus adéquate pour un problème de classification.