Analyse de données d'agréage

Identification de facteurs pouvant expliquer l'apparition de défauts dans les cargaisons d'ananas

Membres du groupe :

Julien Kossi KOWOU
Prince Foli ACOUETEY

Superviseur :

Madame Anne GEGOUT-PETIT

Sommaire

1 Introduction

La société SIIM (Société Internationale d'Importation), spécialisée dans la commercialisation des fruits, dans une démarche qualité, prévoit d'identifier les facteurs/conditions pouvant influer sur la qualité de ses cargaisons d'ananas.

Les caractéristiques des ananas avant leur exportation, les conditions de transport des ananas et l'état des ananas après leur exportation sont regroupés sous le nom de **données d'agréage**. Ces données, après correction, peuvent être schématisées par une matrice de 7468 lignes (7468 observations/cargaisons d'ananas) et 29 colonnes (29 variables) ayant 10.62% de valeurs manquantes (valeurs désignées « NA »).

Une première analyse exploratoire des données d'agréage permet d'identifier les caractéristiques des ananas avant leur exportation et les conditions de transport des ananas comme l'espace possible de variables explicatives. Quant aux informations sur l'état des ananas après exportation, elles représentent l'espace des variables pouvant servir de mesure(s) de la qualité des cargaisons d'ananas (variable à expliquer). Deux mesures de qualité se distinguent :

- *Ratio.Ananas.Defaut* : le ratio d'ananas avec défauts (compris entre 0 et 1)
- *Fruit.Quality* : un score des ananas issu de tests de qualité (valeurs entières entre 1 et 8).

Sur la base de ces informations, une méthodologie fut définie afin d'atteindre l'objectif général.

2 Gestion des valeurs manquantes

Nous calculons la proportion de valeurs manquantes par variable au sein des données d'agréage :

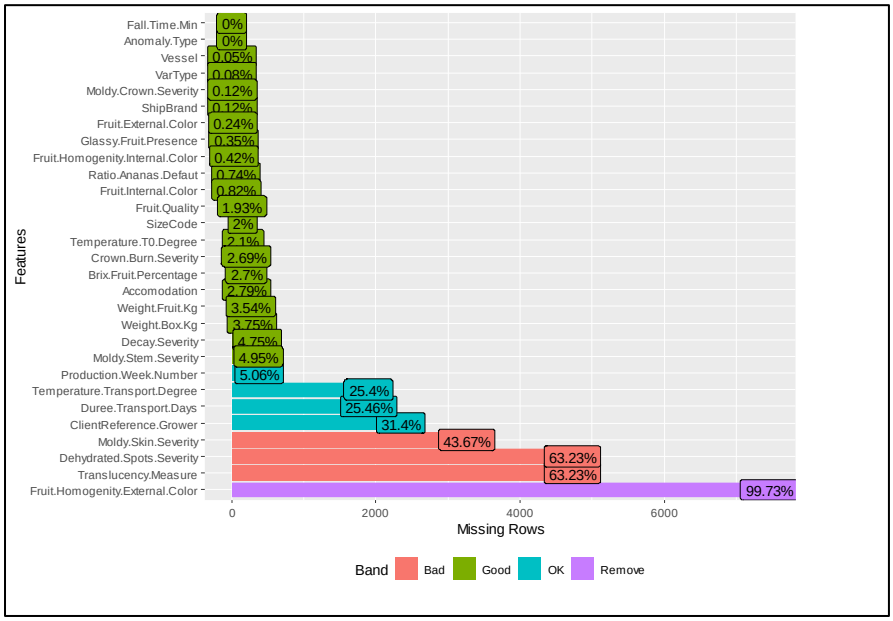

Figure 1 : Proportion de valeurs manquantes par variable

99.73% de valeurs sont manquantes chez la variable *Fruit.Homogenity.External.Color* (soit moins de 30 observations sans valeur manquante) et la variable *Decay.Severity* présente une seule modalité : ces deux variables sont supprimées.

Les données d'agréage passent de 29 à 27 variables (19 variables qualitatives et 8 variables quantitatives) sur lesquelles sont réalisés trois types d'imputations de valeurs manquantes :

- Une imputation par analyse factorielle mixte des données (AFMD) pénalisée

Une sélection discrète sur l'espace des combinaisons possibles des composantes principales retient la combinaison des 2 premières comme celle fournissant une meilleure estimation des valeurs manquantes des données d'agréage pour un MSEP (Mean Squared Error of Prediction) en-dessous de 10^{-6}.

- Une imputation par l'algorithme de la forêt aléatoire

10 arbres ont servi à la construction de la forêt et chacune des valeurs manquantes est imputée 25 fois : 5 imputations possibles des valeurs manquantes des données d'agréage sont fournies. La moyenne des imputations est retenue pour les variables quantitatives et le mode pour les variables qualitatives.

- Bagging des imputations par analyse factorielle mixte des données (AFMD) pénalisée et par forêt aléatoire

Pour les variables quantitatives, est affectée à chaque valeur manquante la moyenne de ses estimations issues de deux imputations précédentes ; quant aux variables qualitatives, un tirage aléatoire permet de choisir entre les deux estimations issues des deux imputations.

3 Analyse des distributions des variables

Après le bagging des imputations, une analyse des distributions des variables est faite afin de corriger toute inconsistance/valeur aberrante pouvant provenir des imputations (notamment, une valeur de ratio d'ananas avec défaut supérieure à 1 fut détectée et remplacée par 0.99).

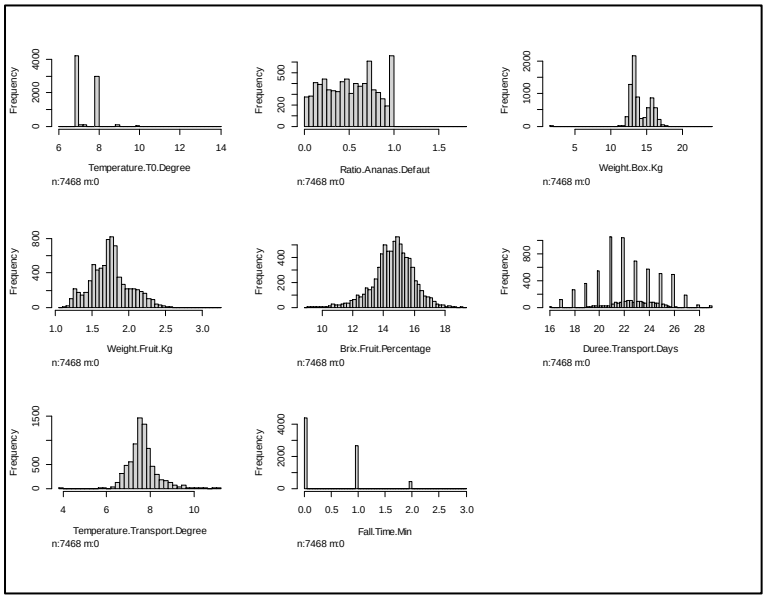

Figure 2 : Distribution des variables quantitatives après imputation

Les distributions des variables quantitatives sont particulières et ne peuvent aucunement être assimilées à de possibles gaussiennes. Quant aux variables qualitatives, elles présentent des déséquilibres d'effectifs au sein de leurs modalités. Ci-dessous la répartition des modalités de la variable Fruit.Quality :

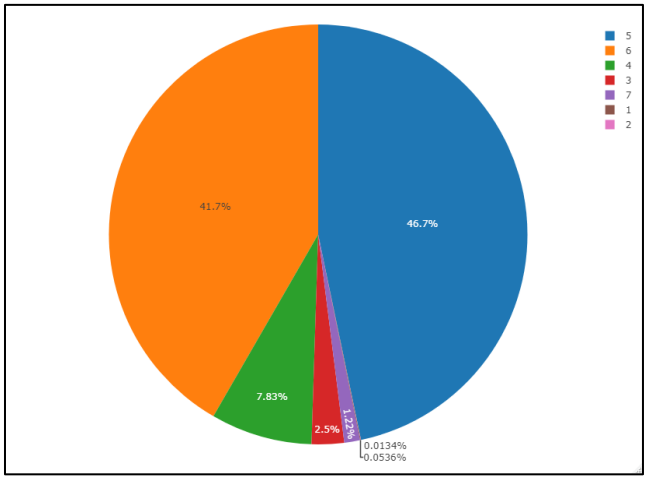

Figure 3 : Répartition des modalités de la variable Fruit.Quality

Le score affecté aux cargaisons d'ananas à l'issue des tests de qualité est résumé par la variable *Fruit.Quality*. Seules des scores d'au-moins 5 sont avantageux pour les cargaisons d'ananas de la SIIM.

4 Clustering des observations

Les données d'agréage, après gestion des valeurs manquantes, ne contiennent que sept (7) variables pouvant donner une mesure de la qualité des cargaisons d'ananas. Toutes ces 7 variables sont utilisées pour estimer des clusters d'observations ; et une nouvelle mesure de qualité est définie uniquement à partir des variables *Fruit.Quality* et *Ratio.Ananas.Defaut*.

4.1 Définition d'une nouvelle mesure de qualité

Les cargaisons d'ananas ayant un score en-dessous de 5, ainsi que celles ayant un ratio d'ananas avec défauts trop élevé sont celles que la SIIM cherche à éviter.

Le score (en-dessous / au-dessus de 5) permet de définir la variable « *Overall.Quality.Level1* » affectant :

- La catégorie « Not Acceptable » aux cargaisons d'ananas avec un score en-dessous de 5
- La catégorie « Acceptable » aux cargaisons d'ananas avec un score au-dessus de 5

La catégorie « Acceptable » en fonction de la valeur du ratio d'ananas avec défauts, définit la variable « *Overall.Quality.Level2* » scindée en deux (2) :

- La catégorie « Average » pour les cargaisons ayant un ratio d'ananas avec défaut supérieur à 0.2
- La catégorie « Good » pour les cargaisons ayant un ratio d'ananas avec défaut inférieur à 0.2

La Figure 4 illustre la proportion de cargaisons selon chacune des modalités de la variable *Overall.Quality.Level1*.

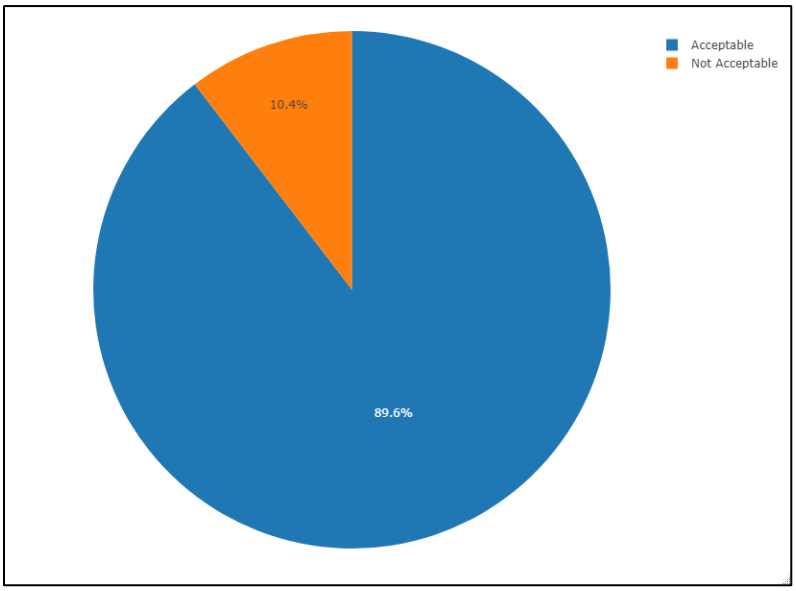

Figure 4 : Répartition des modalités de la variable Overall.Quality.Level1

La Figure 5 illustre la proportion de cargaisons selon chacune des modalités de la variable *Overall.Quality.Level2*.

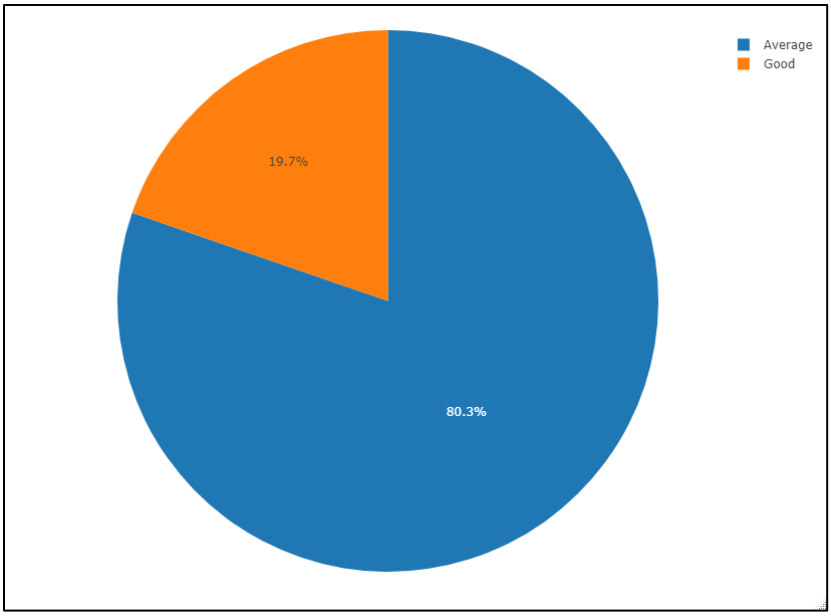

Figure 5 : Répartition des modalités de la variable Overall.Quality.Level2

4.2 Clustering hiérarchique basée sur les variables de mesure de qualité

L'AFMD permet de résumer l'information issue des 7 variables de mesure de qualité en 25 composantes principales. La Figure 6 illustre la part de variance expliquée par les 10 premières composantes principales.

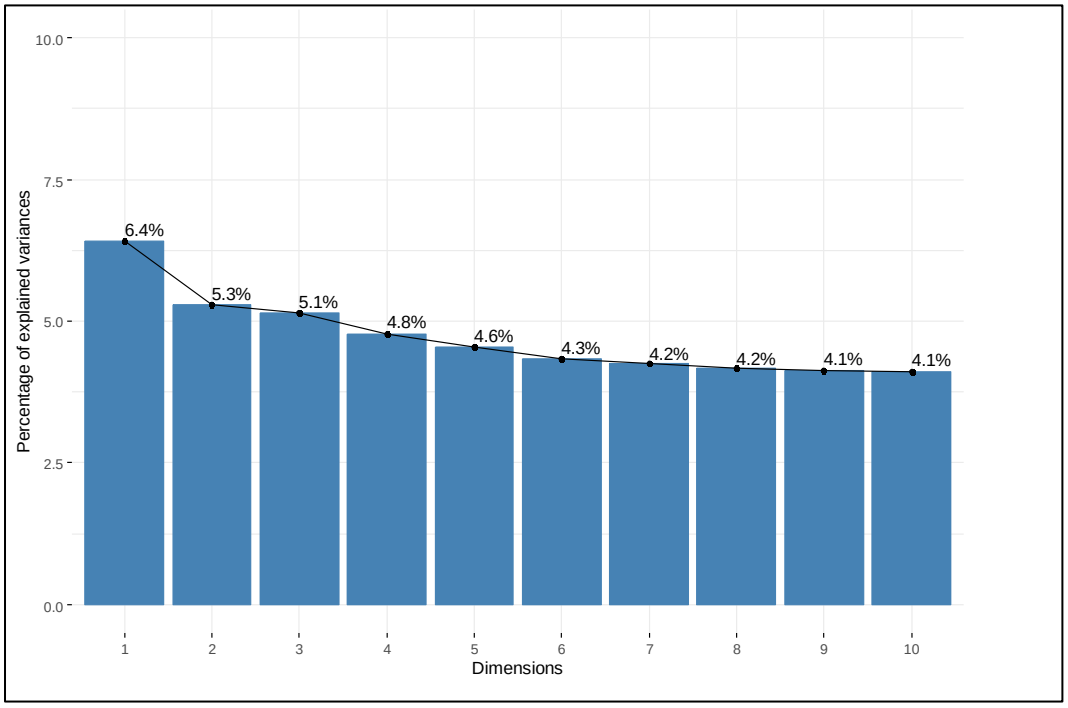

Figure 6 : Part de variance expliquée par les 10 premières composantes principales de l'AFMD

Les 25 composantes principales sont utilisées pour déterminer dans un premier temps le nombre optimal de clusters pouvant définir les possibles catégories d'observations (en utilisant la méthode de la

silhouette). Dans un second temps, ce nombre optimal de clusters est utilisé pour faire une classification hiérarchique des observations.

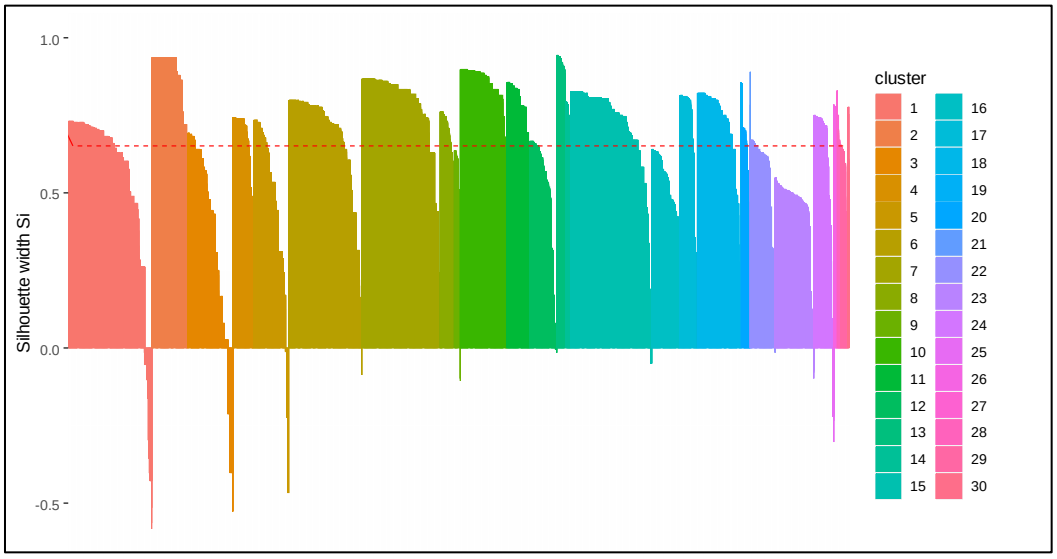

Figure 7 : Silhouettes des 30 clusters d'observations

Les valeurs de silhouette négatives indiquent la présence d'individus mal classés dans des clusters. Sur les 7468 observations, 168 ont été mal classées. Pour ces observations mal classées, les meilleurs clusters desquels ils sont proches (en termes de norme L1) sont déterminés, puis ils y sont affectés.

Les clusters obtenus ont des tailles qui varient entre 750 et 3 observations. La Figure 8 montre la disparité des clusters obtenus en termes de nombre d'observations.

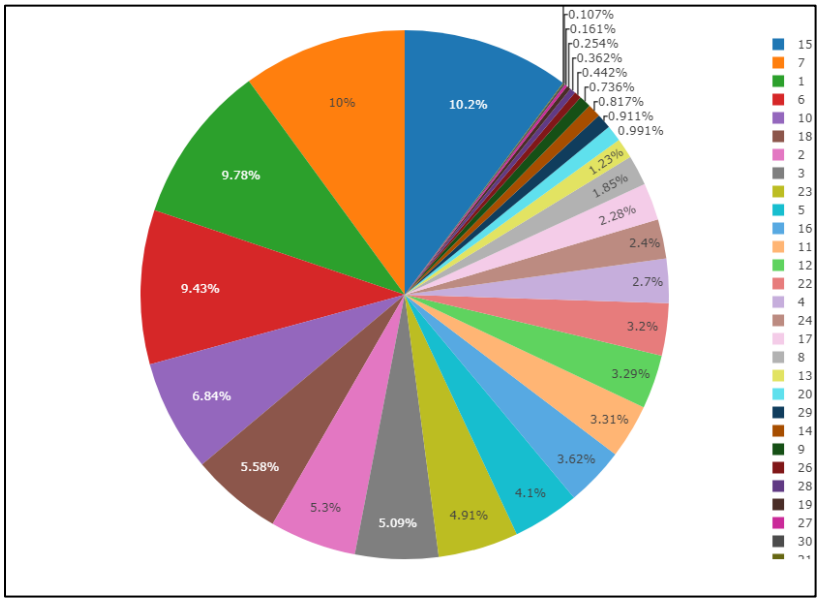

Figure 8 : Taille des clusters issus de la classification hiérarchique

5 Arbres de décisions et régressions logistiques par cluster

Les données d'agréage sont séparées en échantillon d'apprentissage ($^2/_3$ des cargaisons, soit 4980 observations) et en échantillon de test (2488 observations).

5.1 Arbres de décision

Un sur-échantillonnage est effectué sur l'ensemble d'apprentissage afin de mieux gérer tout déséquilibre entre les modalités de la variable *Overall.Quality.Level1* qui pourrait affecter l'estimation des paramètres de l'arbre de décision.

L'échantillon d'apprentissage est séparé en deux :

- L'une des moitiés reste inchangée
- Et un sur-échantillonnage est appliqué sur l'autre moitié afin d'augmenter les proportions de la modalité minoritaire « Not Acceptable »

La répartition des modalités de la variable *Overall.Quality.level1* au sein du nouvel ensemble d'apprentissage est illustrée par la Figure 9 :

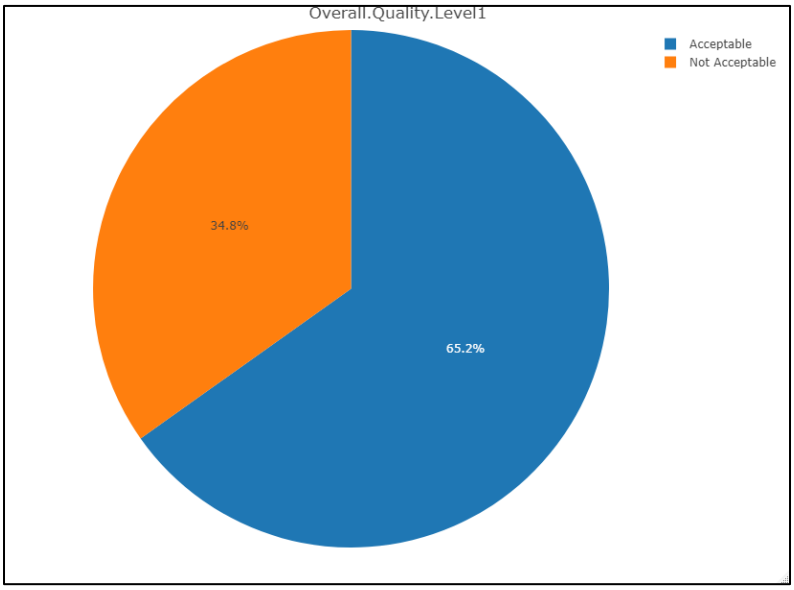

Figure 9 : Répartition des modalités de la variable Overall.Quality.Level1 au sein de l'échantillon d'apprentissage

5.1.1 Arbre de décision basée sur les p-valeurs ajustées de Bonferroni

A chaque nœud un test statistique de niveau de confiance 95%) (soit un test de Chi-deux ou de Fisher, soit un test de Mann-Whitney est effectué entre la variable *Overall.Quality.Level1* et chacune des variables explicatives.

Les p-valeurs issues de ces tests sont ajustées par la méthode de Bonferroni ; puis la variable du test statistique ayant la plus faible p-valeur est choisi pour réaliser la séparation des observations.

Un nœud ne subira une division que si le nombre d'observations qui y sont présents est supérieur à 30 et s'il y existe une variable dont la p-valeur ajustée du test statistique entre elle et la variable *Overall.Quality.Level1* est inférieure à 0.05.

Dans chaque feuille, la valeur prédite de la variable *Overall.Quality.Level1* pour les observations qui y sont présentes est celle de la classe majoritaire.

La profondeur de l'arbre construit est fixée à 3 et les variables ayant plus de 31 modalités (*Vessel*, *Production.Week.Number*) ne sont pas considérées (dû à des problèmes complexités d'algorithmes et de temps de calcul).

La Figure 10 présente l'arbre obtenu :

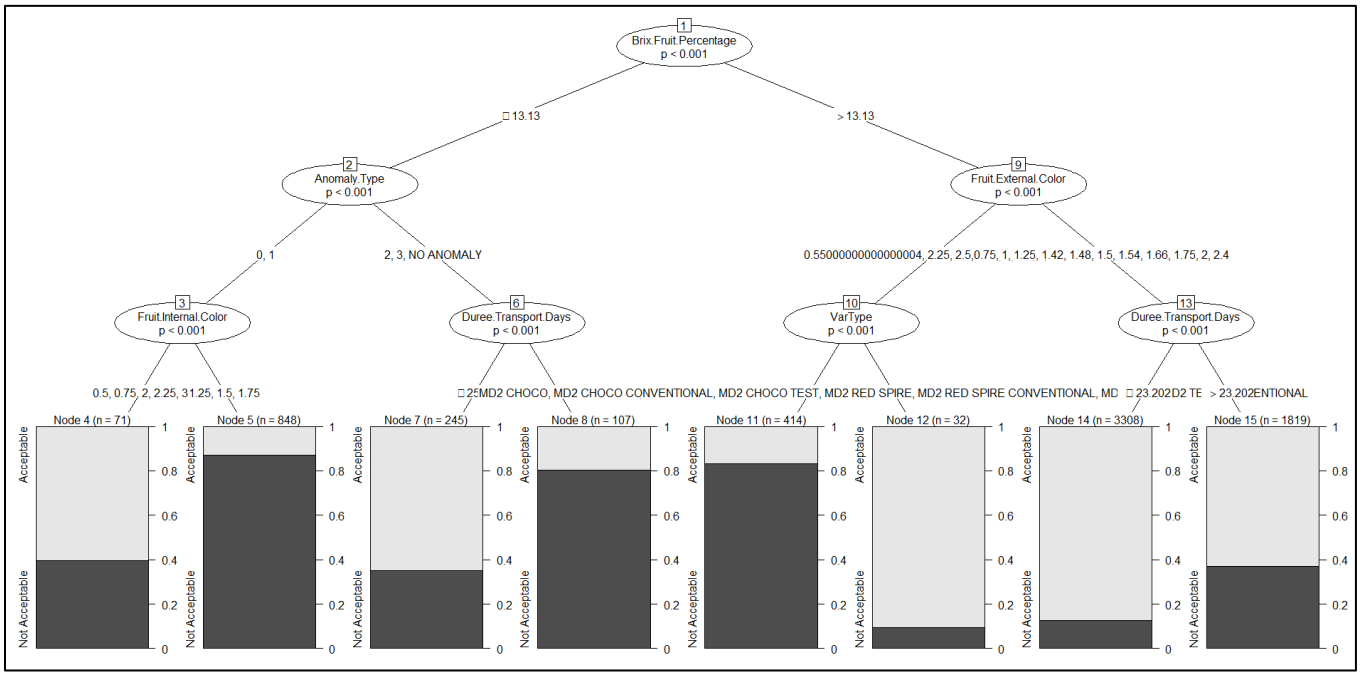

Figure 10 : Arbre de décision basée sur les p-valeurs ajustées de Bonferroni

Sachant que 4 observations n'ont pu être classées :

- *Sensibilité* : 95.15% des observations ayant la modalité « Acceptable » ont pu effectivement être classées comme « Acceptable »
- *Spécificité* : 54.47% des observations ayant la modalité « Not Acceptable » ont pu effectivement être classées comme « Not Acceptable »

5.1.2 Arbre de décision CART

Un arbre de décision CART de profondeur 3, est construit en utilisant l'indice d'impureté de Gini. Une feuille ne sera considérée comme un nœud lorsque le nombre d'observations qui y sont présentes excède 30.

Après élagage, la Figure 11 représente l'arbre de décision CART retenu :

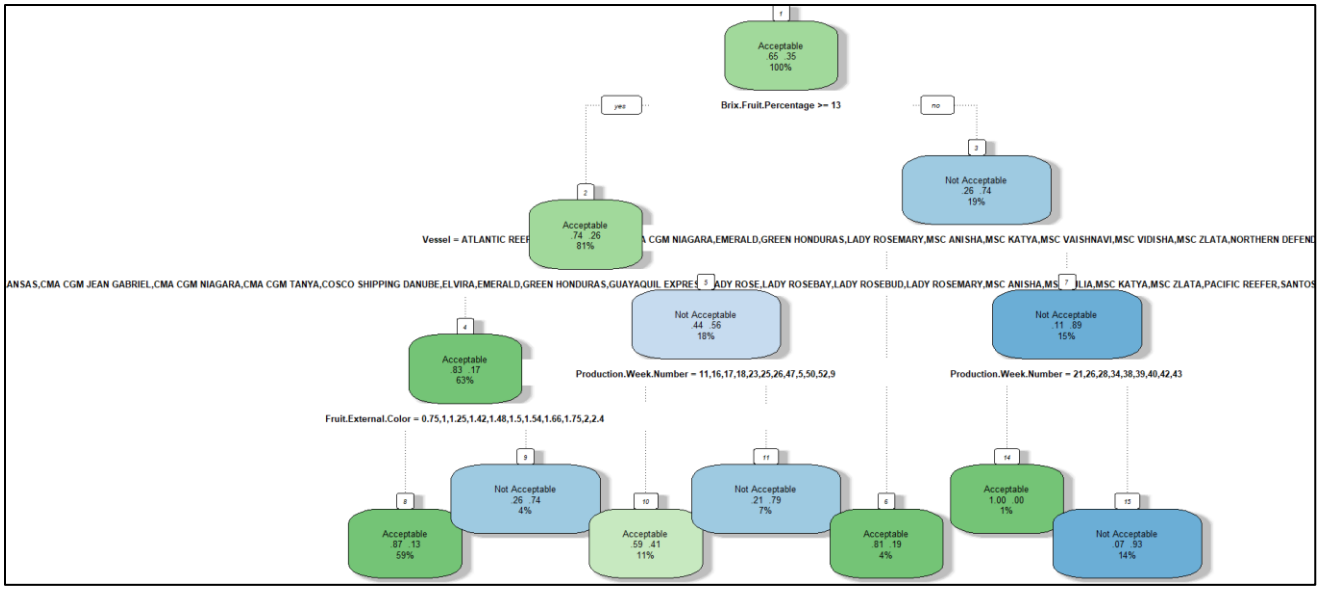

Figure 11 : Arbre de décision CART

Les variables les plus discriminantes selon l'arbre de décision CART sont la variable *Vessel*, puis la variable *Brix.Fruit.Percentage*.

- *Sensibilité* : 94.04% des observations ayant la modalité « Acceptable » ont pu effectivement être classées comme « Acceptable »
- *Spécificité* : 61.78 % des observations ayant la modalité « Not Acceptable » ont pu effectivement être classées comme « Not Acceptable »

L'arbre de décision CART permet de mieux prédire la modalité « Not Acceptable » de la variable *Overall.Quality.Level1*, tandis que l'arbre de décision basée sur les p-valeurs ajustées de Bonferroni est faiblement mieux spécifié pour la prédiction de la modalité « Acceptable » de la variable *Overall.Quality.Level1.*

Un arbre mixte tel que pour une observation donnée, l'arbre de décision basée sur les p-valeurs ajustées de Bonferroni prédit la probabilité qu'elle ait la modalité « Acceptable » de la variable Overall.Quality.Level1, tandis que l'arbre CART prédit la probabilité qu'elle ait la modalité « Not Acceptable » fut évalué. A chaque observation, selon la plus grande des deux probabilités une modalité adéquate de la variable Overall.Quality.Level1 sera affectée. Cependant, ses performances en termes de spécificité et de sensibilité étaient plus basses que celles ces deux arbres de base.

5.2 Régressions logistiques par cluster

Pour chaque cluster composé d'au moins 30 observations ayant la modalité « Acceptable » de la variable *Overall.Quality.Level1* et les deux modalités de la variable *Overall.Quality.Level2*, un modèle de régression logistique est estimé et optimisé par une approche backward de sélection de variables.

Quant aux clusters composés de moins de 30 observations ayant la modalité « Acceptable » de la variable *Overall.Quality.Level1* et uniquement une des modalités de la variable *Overall.Quality.Level2*, ils sont regroupés et un unique modèle de régression logistique est estimé et optimisé par une approche backward de sélection de variables.

Trois (3) régressions logistiques sont obtenues. La prédiction pour une observation sera la modalité la fréquente prédite par les trois modèles.

Pour un seuil de coupure de 0.5 entre les modalités « Good » et « Average », le modèle issu du regroupement de clusters a une valeur d'AUC de 0.5 ; tandis que la valeur de l'AUC des deux autres vaut 0.4614.

L'analyse des courbes ROC de ces deux modèles permet de conclure sur la non-nécessité d'améliorer la spécificité (la spécificité étant définie par rapport à la modalité « Good ») afin d'augmenter les valeurs d'AUC. En effet, classer une cargaison comme « Average » alors qu'elle possède des qualités « Good » n'est pas aussi préjudiciable que l'inverse.

Le modèle obtenu sur le regroupement de clusters est meilleur en termes d'AUC : il est retenu et peut être utilisé afin de prédire efficacement la catégorie « Average » d'une cargaison pouvant avoir un score au-dessus de 5.

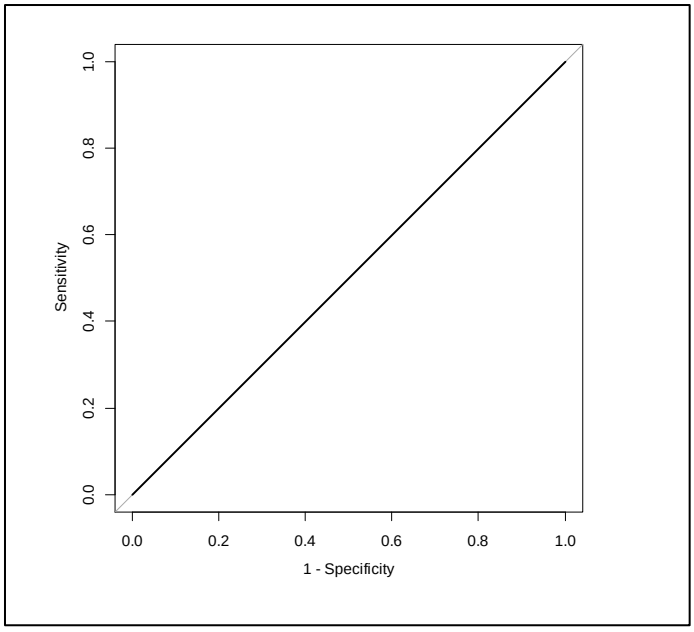

Figure 12 : Courbe ROC du modèle de régression logistique optimal

6 Conclusion

Les arbres de décision ont permis d'identifier les combinaisons possibles de facteurs pouvant influer négativement sur la qualité des cargaisons d'ananas

Le taux de sucre, la durée des transports, la semaine de production, les couleurs internes et externes des fruits sont des facteurs influant sur la qualité des ananas.

Les caractéristiques communes identifiées aux cargaisons de mauvaise qualité sont :

- Un taux de sucre inférieur à 13% et une couleur interne en dessous de 3
- Un taux de sucre inférieur à 13% et une durée de transport de plus de 25 jours
- Un taux de sucre inférieur à 13% et une variété d'ananas MD2 CHOCO, MD2 CONVENTIONNAL, MD2 CHOCO TEST, MD2 RED SPIRE, MD2 RED SPIRE CONVENTIONNAL
- Un taux de sucre inférieur à 13% et des semaines de production 21, 26, 28, 34, 38, 39, 40, 40, 42, 43, 11, 16, 17, 18, 23, 25, 26, 47, 5, 50, 52., 9