Haute Ecole d'Ingénierie et de Gestion du Canton de Vaud
University of Applied Sciences Western Switzerland

Amélioration de la productivité de cultures tropicales par des méthodes d'apprentissage automatique

Caractériser et prédire la qualité des cafés colombiens

Thibault Schowing
Travail de Bachelor
2 mai 2017

Superviseur : Prof. Carlos Andrès Peña
Expert : Sylvain Delerce
Expert : Daniel Jimenez

Remerciements

Merci

Résumé

Résumé

Table des matières

1 Introduction

1.1 Question de recherche

A partir de données sur le climat, la qualité du sol et les pratiques culturales, est-il possible d'expliquer et de prédire les différents traits de la qualité en bouche des cafés du département de Risaralda ?

1.2 Problématique

Le sujet de ce Travail de Bachelor a été proposé par le « *Centro Internacional de Agricultura Tropical* » (CIAT) qui travaille dans le but d'améliorer la productivité et la gestion de l'agriculture en zone tropicale, et dont les bureaux se trouvent à Cali, en Colombie.

À 200 kilomètres de Cali, le comité des caféiculteurs de Risaralda souhaite pouvoir expliquer les différents traits de la qualité en bouche des cafés produits dans les différents secteurs de leur département. La filière café colombienne est en effet en concurrence avec d'autres pays exportateurs sur le marché international, et un des avantages comparatifs de la Colombie est que ses terroirs produisent des cafés de qualité et de caractères affirmés. Il est donc stratégique pour la fédération des caféiculteurs de Colombie d'être en mesure de faire valoir ces spécificités pour aller chercher la valeur ajoutée associée aux produits démarqués du lot.

Ce projet a pour but de trouver des méthodes de modélisation afin d'identifier les caractéristiques du café spécifiques à chaque secteur de la région en se basant sur des analyses gustatives, des données climatiques et géographiques, et d'autres données de pratiques culturales.

Dans un premier temps, l'objectif est de catégoriser les cafés en tentant de trouver des tendances gustatives par rapport aux conditions de culture. Dans un second temps, il faudra pouvoir prédire la qualité en bouche des cafés par rapport aux conditions environnementales.

Le but de cette collaboration sur le long terme est de permettre au département de Risaralda de mettre en valeur la diversité de ses cafés, principalement à des fins de promotion auprès des acheteurs.

1.3 Données

Les données gustatives sont très relatives aux sens et à la perception de chaque goutteur. Cependant, la SCAA, *Speciality Coffee Association of America*, dispose d'un système de notation basé sur des hypothèses communautaires reconnues ce qui permet d'avoir une certaine régularité dans les données de dégustations. Les cafés sont notés sur 100 points répartis sur plusieurs critères : parfum/arôme, saveur, arrière-goût, acidité, corps, équilibre, douceur, clean-cup (absence de défauts marqués), uniformité et évaluation personnelle du testeur. Chacun de ces critères est noté sur 10 mais aussi par des termes qualitatifs. Par exemple, la saveur, c'est-à-dire la combinaison de l'odeur et du goût, la première impression qu'on a en goûtant le café, peut être notée 7/10 et "Caramel".

Les données climatiques comprennent les températures maximales, minimales et moyennes, la variation de température pendant la journée (DTR) et les quantités de précipitations. Les moyennes de ces mesures ont été calculée pour chaque mois et extrapolées sur une grande partie du territoire (à partir de stations météorologiques), permettant ainsi d'accéder aux mesures selon l'emplacement désiré à environ 500 mètres près.

En prenant par exemple les données de température maximale pour le mois de janvier 2011, en affectant pour chaque valeur une couleur, nous pouvons visualiser les données sous la forme d'une image comme sur la figure 1.1.

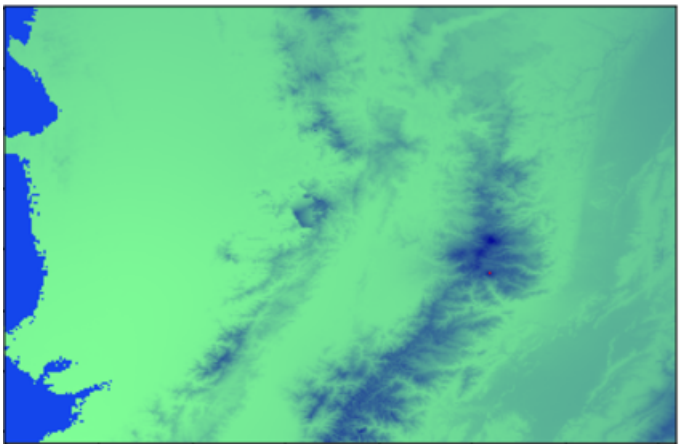


FIGURE 1.1 – Mise sous forme graphique du tableau des température maximales pour le mois de janvier 2011

Les données de qualité de sol sont subdivisées en profils. Chaque profil est séparé en une ou plusieurs couches d'une certaine profondeur dont sont renseignées les caractéristiques comme le pH, la gravimétrie ou encore le taux de matière organique.

Le système SICA , pour *Sistema de Información Cafetera*,

2 Processus de modélisation

2.1 Régression et classification

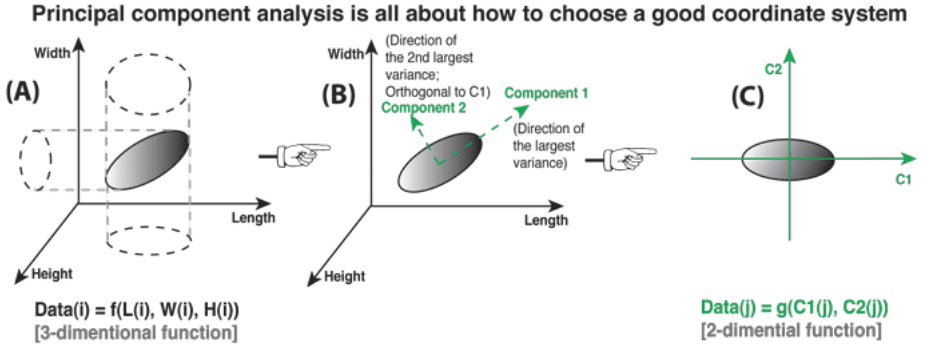
La régression est un ensemble de méthodes statistique utilisées afin d'estimer les relations entre des variables, plus précisément entre variables indépendantes (entrées) et dépendantes (sorties). Afin d'estimer les relations entre les variables, des *features* peuvent être extraites comme une moyenne ou la PCA (section 2.1.1) par exemple.

La classification, définit simplement le fait de mettre un objet dans une certaine classe ou une autre. Elle appartient à l'apprentissage non-supervisé qui consiste en méthodes permettant de diviser un groupe hétérogène de données en sous-groupes de données considérées comme étant les plus proches.

2.1.1 Principal Component Analysis (PCA)

La PCA, pour Analyse en Composante Principale en français, est une méthode qui consiste à transformer un jeu de variables corrélées en nouvelles variables dé-corrélées les unes des autres. Ces nouvelles variables sont appelées composantes principales et permettent de rendre l'information moins redondante. Pour faire plus simple, l'utilité de la Composante Principale est de réduire le nombre de variables tout en gardant un maximum d'information. La figure 2.1 montre une représentation graphique de la composante principale.

FIGURE 2.1 – Description de l'Analyse en Composante Principale. (A) Description d'un objet simple de manière compliquée (trois dimensions pour par exemple une ellipse en papier) (B) Trouver des nouvelles variables (axes de coordonnées) orthogonaux l'un à l'autre qui pointent dans les directions de la plus grande variance (C) Utiliser les nouvelles variables (axes) pour décrire l'objet d'une manière plus simple.



2.1.2 Partial Least Square (PLS)

PLS, originalement pour *Partial Least Squares regression* puis plus récemment pour *Projection to Latent Structures* est une méthode qui combine des propriétés de la PCA ainsi que de multiples régressions linéaires. Au lieu de trouver un hyperplan de la variance maximale, entre les variables dépendantes et indépendantes, cette méthode va trouver un modèle de régression linéaire en projetant les variables indépendantes et dépendantes dans un nouvel espace. Ce sont les variables latentes. Cette méthode est particulièrement utile lorsqu'il est nécessaire de prédire un jeu de variables dépendantes à partir d'un très grand jeu de variables indépendantes.

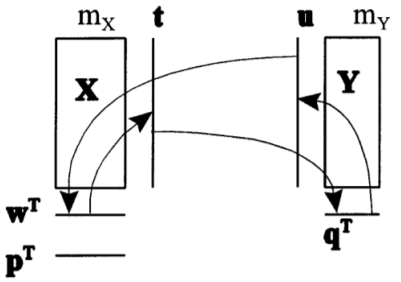


FIGURE 2.2 – Méthode PLS. X est représenté par son score t et Y par u. Une première estimation de U est multipliée à travers X pour obtenir une aproximation du poid ω_t. Le poid est normalisé pour être de longueur 1 et remultiplié à travers X pour produire t. A partir de t et de Y, le poid q^T est obtenu ce qui donne un nouveau vecteur u. Cette opération est répétée jusqu'à la convergence de t. [4]

2.1.3 Multi Block PLS

La PLS multi block est une extension de la méthode PLS qui sépare les variables indépendantes en plusieurs blocks afin de leur donner une plus grande interpértabilité et plus d'informations sur la structure générale des données. Dans le cadre de ce projet, on peut imaginer séparer les données climatiques des données de sol par exemple. L'exécution est très similaire à la méthode PLS classique.

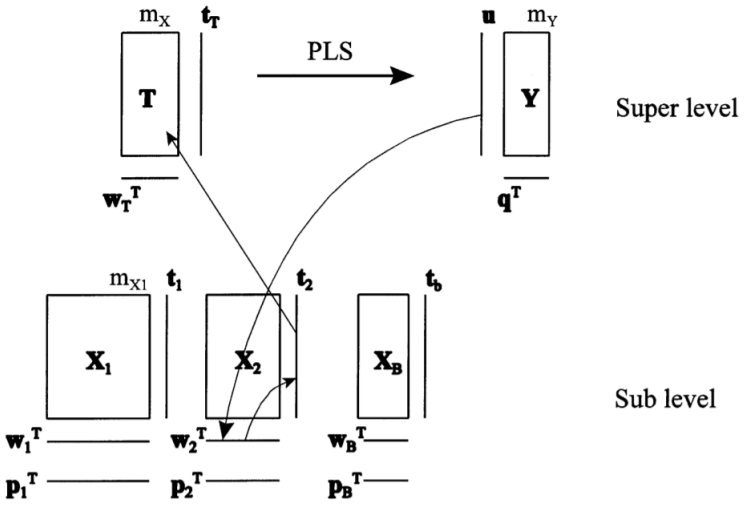


FIGURE 2.3 – Méthode MBPLS. Un score de départ u est régressé sur tous les blocs X_b pour donner les poids variables du bloc w_b^T Les poids des variables de blocs sont normalisés à la longueur un et multipliés par les blocs pour donner les scores de blocs t_b. Les scores de blocs sont combinés dans le super bloc T. Un cycle PLS entre T et Y est effectué pour donner le poids superieur W_T^T, qui est également normalisé à la longueur un, et le super score t_T. L'opération est répétée jusqu'à la convergence de t_T. [4]

2.2 Apprentissage supervisé

Le but de l'apprentissage supervisé est d'expliquer des sorties (outputs) à partir d'entrées (inputs). Des règles sont calculées à partir de données d'apprentissage selon différents modèles. Par la suite, le modèle est utilisé pour catégoriser des nouvelles données.

2.2.1 Random Forest

La méthode Random Forest, ou *forêts d'arbres décisionnels* en français, fait partie des méthodes ensemblistes [1], qui utilisent la combinaison de plusieurs modèles de base, d'apprentissage automatique. Elle combine les concepts de sous-espaces aléatoires et de bagging.

Le bagging, ou *bootstrap agregation*, consiste à sous-échantillonner (ou ré-échantillonner au hasard avec doublons) le training set et de faire générer à l'algorithme voulu un modèle pour chaque sous-échantillon. On utilise le bagging pour réduire la variance de la fonction de prédiction estimée. Le bagging semble bien fonctionner pour les procédures avec une grande variance et un petit biais, comme les arbres de décision. [2]

Random Forest effectue donc un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents [3].

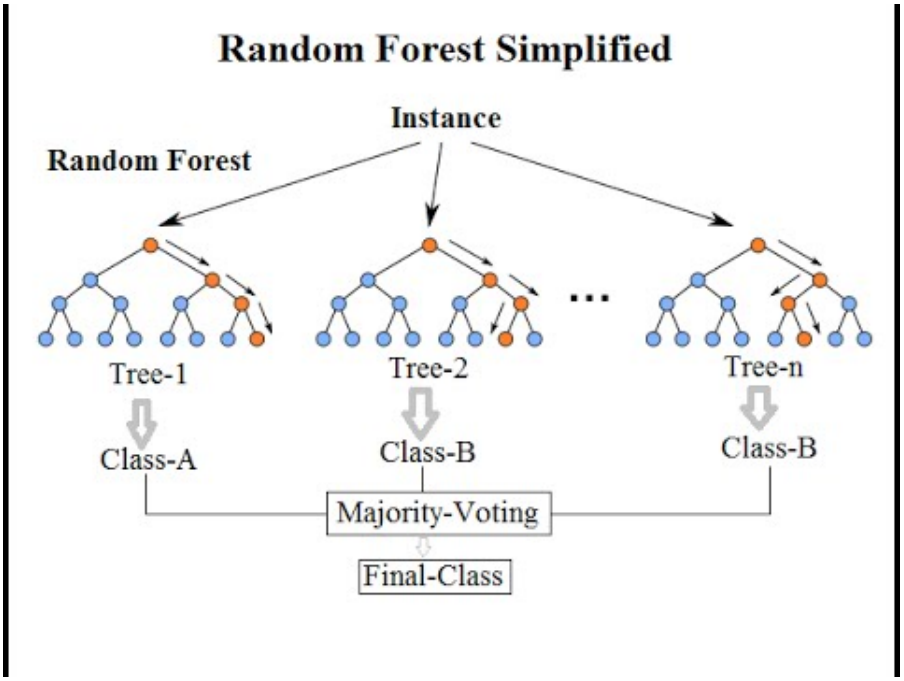


FIGURE 2.4 – Schéma simple du fonctionnement de Random Forest. Source : *https ://www.youtube.com/watch ?v=ajTc5y3OqSQ*

2.3 Apprentissage non-supervisé

Contrairement à l'apprentissage supervisé, l'apprentissage non-supervisé tente de trouver des groupes dans des données hétérogènes. Le but est d'extraire des connaissances à partir de ces données.

2.3.1 SOM

Les différentes classes gustatives d'un café peuvent être considérées comme des entrées afin vérifier s'il est possible de regrouper différents cafés qui se distingueraient. Afin de trouver les différentes catégories de café, nous testerons les capacités de l'algorithme SOM (pour Self Organizing Map ou Cartes Auto Adaptatives en français) qui utilise un réseau de neurones pour étudier la répartition des données dans un espace de grande dimension.

Un bel exemple de SOM est celui de la carte de la pauvreté mondiale réalisé par le *Department of Computer Science and Engineering* de l'université *Helsinki University of Technology.*

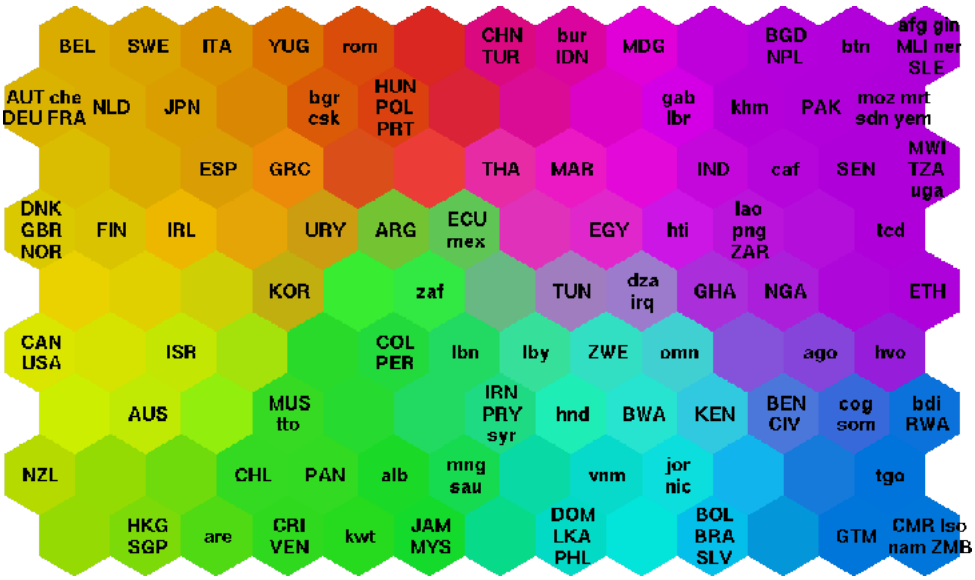


FIGURE 2.5 – Pays organisés en SOM d'après des indicateurs de pauvreté. Source : *http ://www.cis.hut.fi/research/som-research/worldmap.html*

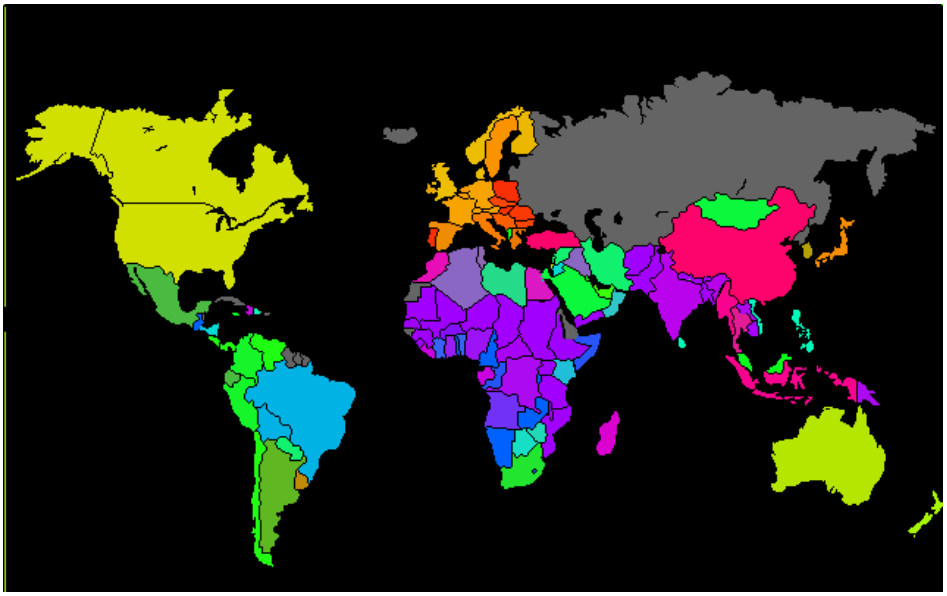


FIGURE 2.6 – Pays correspondants à la carte SOM de la figure 2.5
Source : *http ://www.cis.hut.fi/research/som-research/worldmap.html*

2.3.2 Classification multi-classes

Dans ce projet, le but est de trouver différentes classes de café en trouvant différents points pour les différencier. Les classes de cafés ne seront pas binaires (bon / pas bon, amer ou pas) mais seront composées de multiples classes différentes. Les différents traits du goût (acidité, corps,...) et les différentes classes dans ces catégories (Acide malique ou acide phosphorique pour décrire l'acidité par exemple) composent le café comme différents mots composent le sujet d'un texte.

Les algorithmes comme KNN, Logistic Regression et Naive Bayes sont naturellement multi-classes alors que d'autres comme les perceptrons, SVM ou le boosting sont essentiellement des classificateurs binaires. Il est tout de fois possible de combiner ces différents classificateurs binaires, avec des méthodes comme *One Against All* ou *All-Pairs* afin de les rendre multi-classe.

Bibliographie

[1] Issam El Alaoui. Les méthodes ensemblistes pour algorithmes de machine learning. *Octo blog*, 2014.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference and prediction*. Springer, 2 edition, 2009.

[3] Leo Breiman Statistics and Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[4] Johan A. Westerhuis, Theodora Kourti, and John F. MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12(5) :301–321, 1998.