Règlement du « Crop Data Challenge 2018 - Prédiction des pertes de rendement du blé en France »

Objectif

Le rendement agricole annuel d'une culture représente la quantité de produits récoltée par unité de surface une année donnée. Dans le cas du blé d'hiver, le rendement correspond à la quantité récoltée de grains, souvent exprimée en tonnes par hectare. Le rendement dépend des caractéristiques de la région où le blé est cultivé et des conditions climatiques dans cette région (températures, rayonnements, précipitations etc.). La valeur du rendement est susceptible de varier fortement entre régions et entre années. Par exemple, le rendement pourra être anormalement faible une année présentant un déficit hydrique important à un stade clé du développement de la culture ou, au contraire, très élevé une année présentant des conditions climatiques optimales tout au long de la saison.

Il est important de prédire précisément le rendement avant la récolte. En France, celle-ci a généralement lieu en juillet pour le blé. Des prédictions fiables avant la récolte offrent la possibilité aux opérateurs économiques régionaux de planifier leur récolte, de gérer leurs stocks et d'optimiser leurs contrats (achats et ventes de grains). Les prédictions de rendement constituent également une information stratégique utilisée par les acteurs opérants sur les marchés internationaux. Des prédictions de récoltes abondantes ou, au contraire, des prédictions de pertes importantes, peuvent fortement impacter les cours des marchés agricoles mondiaux.

L'objectif de ce challenge est de développer des outils permettant de classer aussi précisément que possible les valeurs de rendement du blé en France.

Qui peut participer ?

Ce défi ouvert à la fois aux professionnels et aux étudiant.e.s. Tout.e étudiant.e inscrit.e dans une formation universitaire ou dans une école d'ingénieur française peut participer. Les stagiaires inscrits en master et les doctorants sont bienvenus. Tout professionnel peut participer. Un classement différent sera établi pour les professionnels et les étudiants.

Données

Le(a) participant.e dispose de plusieurs fichiers accessibles depuis le site du data challenge :

- **Un fichier de données « entrainement »** pour le blé (TrainingDataSet_Wheat) incluant (i) une variable binaire indiquant l'occurrence d'une perte sévère de rendement ou l'absence de perte sévère pour différents départements français et pour 39 années tirées au hasard et (ii) les valeurs correspondantes de variables climatiques pour les mêmes départements et les mêmes années qui pourront être utilisées pour prédire les rendements.
- **Un fichier « test »** blé (TestDataSet_Wheat_blind) incluant les valeurs des variables climatiques pour les mêmes départements mais pour 19 années non incluses dans les fichiers « entrainement ». Ce fichier n'inclut pas les valeurs de la variable indiquant l'occurrence de perte sévère de rendement. Ces valeurs ont été retirées et seront

utilisées par l'organisateur pour évaluer les performances des classements fournis par les participants.

Règles

Pour le challenge blé, le(a) participant.e développe une méthode de classification des pertes de rendement annuel du blé (ex : régression , forêt aléatoire, réseau de neurone) à l'échelle des départements français à partir du fichier « entrainement » blé en utilisant **le logiciel R (https://cran.r-project.org/)** ou **python**. Le résultat de la méthode doit correspondre à un **indicateur continu de risque de perte** (ex : probabilité de perte ou autre indice).

Une fois sa méthode au point, le(a) participant.e **calcule son indicateur de risque de perte pour toutes les situations (départements*années) du fichier « test »** blé. Il/elle dépose un fichier sur le site du data challenge incluant les valeurs de l'indicateur **dans le même ordre que les situations du fichier « test ».** La précision de ces prédictions sera évaluée par les organisateurs en calculant **l'aire sous la courbe ROC (AUC)** à partir des pertes observées**.**

Format des réponses au challenge

Le(a) candidat.e devra envoyer les éléments suivants à l'adresse cland_fcy@lsce.ipsl.fr :

- Un document au format .pdf décrivant de manière détaillée la méthode utilisée (package R/python utilisé, nature du modèle, variables d'entrée, algorithmes utilisés etc.),
- Un fichier au format .txt (séparateur tabulation) incluant les valeurs de l'indicateur de risque de perte dans le même ordre que les situations incluses dans le fichier test.
- Un fichier incluant le code R ou Python suffisamment documentés pour que le jury puisse utiliser la méthode de prédiction.

Le document décrivant la méthode devra s'intituler :
`ble_document_nomCandidat.pdf`.

Le fichier incluant les prédictions devra être un fichier intitulé :
`ble_prediction_nomCandidat.txt` incluant les prédictions d'anomalie de rendement.

Sélection du vainqueur

Seules les candidatures comportant un document descriptif détaillé et les codes R/Python seront considérés.

Parmi ces candidatures, deux classements seront établis, un pour les étudiants, un pour les professionnels. Chaque classement sera établi à partir des AUC calculés avec les jeux de données « test ». Le vainqueur sera celui ou celle ayant obtenu la valeur de AUC la plus forte pour le blé.

Le 1er du classement « étudiant » recevra un prix de 500 euros. Le 1er du classement « professionnel » recevra un prix offert par le Réseau Mixte Technologique « Modélisation et analyse de données » (http://www.modelia.org/moodle/).

Les meilleur.e.s candidat.e.s seront invité.e.s à venir présenter leurs méthodes de prédiction lors d'un séminaire de restitution le 7 décembre.

Dates clés

- Dates limites pour déposer le fichier des prédictions : 16 novembre 2018
- Publication des résultats : 7 décembre 2018
- Séminaire de restitution : 7 décembre 2018

Liste des variables

Class : variable égale à 1 en cas de perte sévère de rendement de blé et à zéro sinon. Il s'agit de la variable cible à prédire.

year_harvest : année (anonyme) de récolte (1 à 58)

NUMD : numéro (anonyme) indiquant le département (de 1 à 94).

La valeur de la variable Class doit être prédite uniquement à l'aide des variables suivantes (ou d'une partie de ces variables) :

- ETP_1… ETP_12 : Evapotranspiration potentielle moyenne mensuelle par année et par département (1= janvier, 12=décembre)
- PR_1… PR_12 : Précipitation cumulée mensuelle par année et par département (1= janvier, 12=décembre)
- RV_1… RV_12 : Rayonnement moyen mensuel par année et par département (1= janvier, 12=décembre)
- SeqPR1…SeqPR12 : Nombre de jours de pluie mensuel par année et par département (1= janvier, 12=décembre)
- Tn_1…Tn_12 : Température minimale journalière moyenne mensuelle par année et par département (1= janvier, 12=décembre)
- Tx_1…Tx_12 : Température maximale journalière moyenne mensuelle par année et par département (1= janvier, 12=décembre)
- Tn17.1_1 … Tn17.1_12 : Nombre de jours où la température minimale journalière est inférieure à -17 degrés C pour chaque mois par année et par département (1= janvier, 12=décembre)
- Tx010_1 … Tx010_12 : Nombre de jours où la température maximale journalière est comprise entre zéro et 10 degrés C pour chaque mois par année et par département (1= janvier, 12=décembre)
- Tx34_1… Tx34_12 : Nombre de jours où la température maximale journalière est supérieure à 34 degrés C pour chaque mois par année et par département (1= janvier, 12=décembre)

Important : Le blé d'hiver est semé à l'automne et est habituellement récolté en juillet. Les valeurs des variables climatiques pour les mois 9 à 12 des fichiers « entrainement » et « test » correspondent aux valeurs obtenues l'année qui précède l'année de récolte. Les valeurs des variables climatiques pour les mois 1 à 6 correspondent aux valeurs obtenues l'année de récolte. Toutes ces valeurs sont disponibles avant juillet et peuvent donc être utilisées directement pour prédire le rendement avant la récolte. Les valeurs des variables climatiques des mois 7 et 8 sont absentes.