Une image contenant texte, carte de visite, Police, logo

Description générée automatiquement

|  |  |
| --- | --- |
| Image illustrant une route sinueuse et des arbres  Rapport dataviz  Challenge-Edition 2023 | Résumé  Comprendre les facteurs qui peuvent améliorer la production agricole potentielle par hectare ou par tête d'animal, en dehors de toute aide au pays de Basque.  Ferol TATANG FOMEKON  Élève ingénieure en Big Data et IA. |

Table des matières

[Introduction 2](#_Toc138033970)

[I. Analyse Global 3](#_Toc138033971)

[4](#_Toc138033972)

[4](#_Toc138033973)

[II. Nettoyage des données 4](#_Toc138033974)

[III. Exploration des données 6](#_Toc138033975)

[A. Analyse des variables de types numériques 6](#_Toc138033976)

[B. Analyse des variables de types catégoriels 9](#_Toc138033977)

[IV. Prédiction de la production agricole 12](#_Toc138033978)

[A. Définition du pipeline 12](#_Toc138033979)

[B. Evaluation de la prédiction 14](#_Toc138033980)

[Conclusion 15](#_Toc138033981)

[Licence et version 16](#_Toc138033982)

[Utilisation de logiciels 16](#_Toc138033983)

# 

# Introduction

L'histoire de l'agriculture dans le Pays Basque entre 1970 et 2020 est marquée par une évolution de l'agriculture traditionnelle vers des exploitations plus modernes et spécialisées. Depuis lors, le Pays Basque a subi d'importants recensements sur son agriculture par l'INSEE (Institut National de la Statistique et des Études Économiques). De ce fait, le Pays Basque a créé son portail de données ouvertes sur l'agriculture nommé "ZABAL Open Data Agriculture". Ainsi, le Pays Basque pourrait désormais se poser des questions sur ces données et y trouver une solution afin de les comparer à celles de l'INSEE. Partant de cette optique, il nous a été soumis des données très fiables issues de ZABAL auxquelles nous nous sommes posé des questions très pertinentes, à savoir : quels sont les facteurs qui expliquent au mieux la production agricole ? Peut-on prédire la production agricole au fil du temps ? Ainsi, tout au long de ce document, nous allons y apporter des réponses aux différentes questions.

# Analyse Global

Une analyse globale des données consiste à observer le comportement des données de façon brute, sans y apporter de changement. Ainsi, l'observation globale des données nous montre que le jeu de données soumis à notre étude ne possède aucune ligne dupliquée et compte exactement 2561 enregistrements et 23 colonnes, dont il semblerait que nous ayons 7 colonnes de types catégoriels et le reste de types numériques (figure 1). Par ailleurs, l'observation de la description de ces données semble nous révéler qu'entre 1972 et 2020, le Pays Basque a compté exactement 486 productions agricoles (pbs) avec un effectif à temps plein de 640 (etp). Cependant, il semblerait aussi que nous ayons 882 unités de gros bétail (ugb) pour une superficie agricole utilisée de 563 hectares (sau\_ha) (figure 2). Toutes ces remarques nous poussent à penser à l'existence de valeurs aberrantes dans nos données.

Une image contenant texte, menu, capture d’écran, document

Description générée automatiquement

Une image contenant texte, Police, blanc, algèbre

Description générée automatiquement

Figure 1

# 

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

# 

Figure 2

# Nettoyage des données

Au vu du paragraphe précédent, nous avons constaté que nos données contiennent des valeurs manquantes. Raison de plus pour les nettoyer afin d'apporter une meilleure analyse sur la production agricole au pays basque. De ce fait, nous avons estimé que nous ne pouvons pas nous permettre de supprimer les valeurs manquantes, sinon nous allons perdre suffisamment d'informations. C'est pourquoi nous avons remplacé certaines valeurs manquantes numériques par leur moyenne, leur maximum et leur médiane. La moyenne est une mesure statistique qui représente la valeur centrale ou typique d'un ensemble de données, tandis que le quantile est une mesure statistique utilisée pour diviser un ensemble de données en sous-groupes égaux ou inégaux en termes de taille. La figure 3 nous montre le choix de la médiane pour deux variables. Cependant, pour la variable "âge", nous nous sommes intéressés à la moyenne des valeurs extrêmes, puis nous avons remplacé ces valeurs manquantes par la moyenne globale. Par ailleurs, concernant les coordonnées géographiques, après avoir observé la position exacte de ces points sur le globe terrestre (figure 4 et 5), nous avons estimé qu'il serait judicieux de s'intéresser à la longitude et à la latitude. Ensuite, nous avons remplacé ces valeurs manquantes par leurs moyennes.

Une image contenant texte, capture d’écran, nombre, ligne

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Figure 3

Une image contenant carte, croquis, texte

Description générée automatiquement

Figure 4

Une image contenant carte, atlas, Monde, texte

Description générée automatiquement

Figure 5

# Exploration des données

Après nettoyage des données, nous pouvons faire une première description et constater que nous avons bien 2561 productions agricoles avec une moyenne de 6626,09 productions agricoles au total entre 1970 et 2020 au Pays de Basque (figure 6). Par ailleurs, une observation globale montre que l'agriculteur le plus jeune aurait presque 13 ans et le plus âgé 85 ans. La figure 6 montre que les données nous parlent désormais de manière beaucoup plus cohérente et commencent à expliquer la production agricole au Pays Basque.

Une image contenant texte, Police, nombre, ligne

Description générée automatiquement

Figure 6

## Analyse des variables de types numériques

La corrélation est un indicateur de la force et de la direction de la relation linéaire entre les variables. Mais pour effectuer une corrélation, nous devons nous assurer que les variables suivent une distribution normale. Ainsi, en observant la figure 7, il semblerait que les données suivent une distribution normale. De ce fait, nous avons donc effectué un test de Shapiro-Wilk sur les valeurs numériques. Nous obtenons une statistique de test de 0.33 et une p-valeur de 0.0. La valeur p obtenue est très faible (inférieure à 0,05), ce qui suggère un rejet de l'hypothèse nulle selon laquelle les données suivent une distribution normale. Cependant, nos données sont assez grandes, donc nous ne pouvons pas nous focaliser uniquement sur le test de Shapiro-Wilk pour rejeter l'hypothèse de la distribution normale.

Par la suite, nous avons effectué une matrice de corrélation sur nos données (figure 8), ce qui nous a permis de visualiser la corrélation linéaire entre plusieurs variables et la production agricole (pbs) au Pays Basque. Ainsi, la figure 9 nous montre davantage cette corrélation. Cependant, d'autres variables semblent ne pas expliquer significativement la production agricole par hectare ou par tête d'animal au Pays Basque, comme nous pouvons l'observer dans la figure 10.

Une image contenant texte, diagramme, ligne, Police

Description générée automatiquement

Figure 7

Une image contenant texte, capture d’écran, diagramme, Parallèle

Description générée automatiquement

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquement

Figure 8

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

Figure 9

Une image contenant texte, capture d’écran, affichage, diagramme

Description générée automatiquement

Figure 10

En définitive de cette rubrique, il en ressort que le temps plein (ETP), l'unité de référence permettant de calculer les besoins nutritionnels ou alimentaires de chaque type d'animal d'élevage (UGB), la superficie agricole utilisée en hectares, les valeurs et les années expliquent au mieux la production standard agricole (PBS) au Pays Basque entre 1970 et 2020.

## Analyse des variables de types catégoriels

La production agricole standard au pays de Basque dépend aussi de certains facteurs non quantitatifs comme le type d’agriculture, l’échelle, la catégorie et les communes. Une premier observation de la production en fonction des catégories figure 11 semble nous apporter aucune information, raison pour lesquelles nous avons pousser l’analyse un peu plus loin en observant la surface agricole utilise en fonction de l’échelle entre 1970 et 2020 figure 12, nous montre pertinemment que depuis 1970 jusqu'en 2020, le pôle de Bidache a presque toujours eu, une superficie agricole utiliser plus grande que les autres pôles soit 39.74 ha ensuite vient les pôles Amikuze (38,65 ha), Soule Xiberoa (38.44 ha).

Par ailleurs, la figure 13 nous montre que le pôle de Garazi Biagorri a eu la plus grande production de l'histoire du pays de basque en 2010 avec une production potentielle par hectare ou par tête d'animale de plus de 40 milles. Nous observons aussi une croissance pour les autres pôles qui ont pu atteindre en 2020 plus de 20 milles productions. Cependant, nous pouvons affirmer que lorsqu'on regroupe tous les pôles de du pays de basque nous constatons qu'ils ont eu une production de plus de 166 milles en 2010 et une production en baisse de plus de 133 milles en 2020.

En ce qui concerne la figure 14 qui représente la production agricole en fonction des types, il en ressort de façon globale que les types par production ont connu une baisse au fil des années soit la plus importante production de plus de 161 milles en 2010 contre plus de 105 mille productions en 2020. Cependant, les types d'agriculture tels que l'exploitation individuelle, ouvin ou caprin, petite et les moyennes ont une valeur de production potentielle par hectare ou par tête d'animal très importante que les autres types. Par ailleurs, nous pouvons soupçonner qu'il serait judicieux d'investir dans les types d'exploitation individuelle, ouvin ou caprin, petites et moyennes, malgré une baisse très considérables de la production en 10 ans.

Tout comme les analyses précédentes la figure 15, nous constatons que la production est en baisse au fil des années. Cependant, nous pouvons dire que les catégories tels que le statut d'exploitation, les Orientations technico-économiques de l'exploitation (Otex) et la taille d'exploitation expliquer significativement la production agricole.

À la vue de la figure 16, Il semblerait que globalement les communes n’expliquent pas significativement la production agricole au fil des années.

En définitive de cette sous partie, nous disons avec certitude que les variables comme l’échelles, les types, les catégories et les communes explique la production agricole standard.

Une image contenant texte, capture d’écran, affichage, nombre

Description générée automatiquement

Figure 11

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

Figure 12

Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Figure 13

Une image contenant texte, capture d’écran, Parallèle, nombre

Description générée automatiquement

Figure 14

Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Figure 15

Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Figure 16

Par la suite nous allons prédire la production agricole en fonction de toutes les autres variables.

# Prédiction de la production agricole

## Définition du pipeline

Un pipeline est un enchaînement d'opérations qu'on code en enchainant des pairs (clé, valeur). Ainsi, notre pipeline est utilisé pour prétraiter et entraîner un modèle de régression forestière à l'aide de caractéristiques numériques et catégorielles figure 17.

Voici un résumé du pipeline :

1. La première étape du pipeline est integer\_transformer. Elle utilise deux étapes :
   1. imputer : Cela remplace les valeurs manquantes des caractéristiques numériques par la valeur la plus fréquente de chaque colonne.
   2. scaler : Il standardise les caractéristiques numériques en les centrant sur zéro et en les mettant à l'échelle pour avoir une variance unitaire.
2. La deuxième étape du pipeline est categorical\_transformer. Elle utilise deux étapes :
   1. imputer : Cela remplace les valeurs manquantes des caractéristiques catégorielles par la valeur la plus fréquente de chaque colonne.
   2. onehot : Il transforme les caractéristiques catégorielles en variables binaires (one-hot encoding), en créant de nouvelles colonnes pour chaque catégorie unique.
3. La troisième étape du pipeline est preprocessor. Elle utilise ColumnTransformer pour appliquer les transformations spécifiques à chaque type de caractéristique :
   1. Pour les caractéristiques numériques, elle utilise integer\_transformer.
   2. Pour les caractéristiques catégorielles, elle utilise categorical\_transformer.
4. La quatrième étape du pipeline est base. Elle utilise deux étapes :
   1. preprocessor : Cela applique les transformations définies précédemment aux données d'entrée.
   2. regressor : Cela utilise un modèle de régression forestière pour effectuer la prédiction finale.

En résumé, ce pipeline prétraite les caractéristiques numériques et catégorielles, remplace les valeurs manquantes, effectue le codage one-hot pour les caractéristiques catégorielles, standardise les caractéristiques numériques et utilise un modèle de régression forestière pour effectuer la prédiction.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Figure 17

Ensuite nous avons sérialiser le modèle base en le nommant model\_rf. Qui nous a donné un coefficient de détermination de 89,17% en d’autres termes, cela signifie que le modèle de régression forestier utilisé pour effectuer des prédictions à une bonne capacité à expliquer la variation de la variable dépendante, car environ 89,17 % de la variance de la variable dépendante est expliquée par le modèle.

Par ailleurs, nous nous sommes intéressés a observé (figure 18) les coefficients de déterminations de la régression linéaire et du modèle Ridge, nous avons obtenu respectivement 54,2% pour le modèle linéaire et 57,1% pour le modèle Ridge.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Figure 18

En résumer de cette partie, il ressort que la production agricole est prédite a 89,17 % par ses différents facteurs.

## Evaluation de la prédiction

La figure 19 nous montre bien que les points qui ont été prédits sont très comparables aux données de test.

s

Une image contenant capture d’écran, ligne, Tracé, texte

Description générée automatiquement

Figure 19

# Conclusion

En définitive, cet exposé avait pour but de comprendre les facteurs qui peuvent améliorer la production agricole potentielle par hectare ou par tête d'animal, en dehors de toute aide, à partir des données fournies par ZABAL open data agriculture. Ces données sont constituées de fiches territoriales fournies par le recensement agricole entre 1970 et 2020 pour le Pays Basque. Il en ressort que nous avons tout d'abord nettoyé les données en début d'étude, puis effectué une analyse exploratoire, réalisé une prédiction et enfin évalué les résultats. De ce fait, nous pouvons donc dire que la production agricole par hectare ou par tête d'animal au Pays Basque s'explique en fonction des variables soumises à notre étude et est mieux expliquée en particulier par la superficie agricole utilisée, le temps plein effectué, l'unité de gros bétail, le nombre d'exploitations et le temps. Par ailleurs, nous avons réalisé une prédiction et nous sommes rendu compte que dans les années suivantes, la production agricole serait bonne à 89,17%.

# Licence et version

• pandas version: 1.5.3 (Open source)

• numpy version: 1.22.4 (Open source)

• matplotlib version: 3.7.1 (Open source)

• seaborn version: 0.12.2 (Open source)

• plotly version: 5.13.1 (Open source)

• mpl\_toolkits.basemap version: 1.3.7 (Open source)

• folium version: 0.14.0 (Open source)

• pickle version: 4.0 (Open source)

• scikit-learn version: 1.2.2 (Open source)

• sklearn.linear\_model version: sklearn (Open source)

• sklearn.ensemble version: sklearn (Open source)

• sklearn.linear\_model version: sklearn (Open source)

• sklearn.linear\_model version: sklearn (Open source)

• sklearn.linear\_model version: sklearn (Open source)

• sklearn.impute version: sklearn (Open source)

• sklearn.preprocessing version: sklearn (Open source)

• sklearn.preprocessing version: sklearn (Open source)

• sklearn.preprocessing version: sklearn (Open source)

# Utilisation de logiciels

Voici les rôles de chacun des logiciels utilisés :

1. import pandas as pd: Importe la bibliothèque Pandas qui est utilisée pour la manipulation et l'analyse des données.
2. import numpy as np: Importe la bibliothèque NumPy qui est utilisée pour effectuer des calculs numériques efficaces et pour manipuler des tableaux multidimensionnels.
3. import matplotlib.pyplot as plt: Importe la bibliothèque Matplotlib, qui est utilisée pour créer des graphiques et des visualisations en utilisant une syntaxe similaire à celle de MATLAB.
4. import seaborn as sns: Importe la bibliothèque Seaborn, qui est utilisée pour créer des graphiques statistiques attrayants et informatifs.
5. import plotly.express as px: Importe la bibliothèque Plotly Express, qui facilite la création de graphiques interactifs tels que des graphiques à dispersion, des histogrammes, des boîtes à moustaches, etc.
6. import plotly.graph\_objects as go: Importe la bibliothèque Plotly Graph Objects, qui offre plus de contrôle sur la création de graphiques interactifs en utilisant une syntaxe orientée objet.
7. from mpl\_toolkits.basemap import Basemap: Importe la classe Basemap de la bibliothèque mpl\_toolkits.basemap, qui est utilisée pour créer des cartes et des visualisations géographiques.
8. import folium: Importe la bibliothèque Folium, qui est utilisée pour créer des cartes interactives et des visualisations géospatiales.
9. from sklearn.model\_selection import train\_test\_split: Importe la fonction train\_test\_split de la bibliothèque scikit-learn (sklearn), qui est utilisée pour diviser les données en ensembles d'entraînement et de test.
10. from sklearn.compose import ColumnTransformer: Importe la classe ColumnTransformer de scikit-learn, qui est utilisée pour appliquer différentes transformations sur les colonnes spécifiées d'un tableau de données.
11. from sklearn.pipeline import Pipeline: Importe la classe Pipeline de scikit-learn, qui est utilisée pour définir des pipelines d'apprentissage automatique, qui sont des séquences ordonnées d'estimateurs (transformations et modèles) à appliquer sur les données.
12. from sklearn.impute import SimpleImputer: Importe la classe SimpleImputer de scikit-learn, qui est utilisée pour remplacer les valeurs manquantes dans un tableau de données par des valeurs spécifiées.
13. from sklearn.preprocessing import OneHotEncoder, StandardScaler: Importe les classes OneHotEncoder et StandardScaler de scikit-learn, qui sont utilisées respectivement pour encoder les variables catégorielles et standardiser les variables numériques.
14. from sklearn.ensemble import RandomForestRegressor: Importe la classe RandomForestRegressor de scikit-learn, qui est utilisée pour entraîner des modèles de régression à base d'arbres de décision aléatoires.
15. from sklearn.linear\_model import LinearRegression, Ridge, Lasso, BayesianRidge: Importe les classes LinearRegression, Ridge, Lasso et BayesianRidge de scikit-learn, qui sont utilisées pour entraîner des modèles de régression linéaire avec différentes régularisations.
16. from sklearn.preprocessing import PolynomialFeatures: Importe la classe PolynomialFeatures de scikit-learn