Sistema de predicción de producción de cultivos

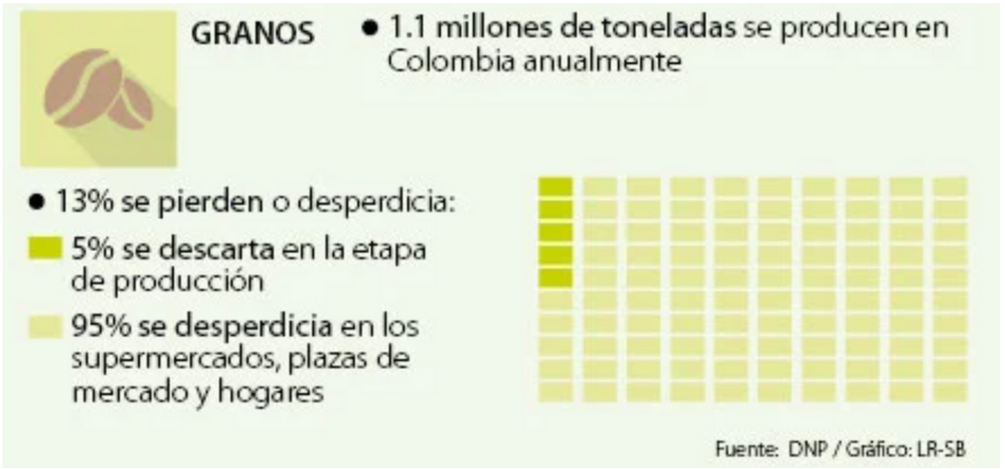
OSCAR FERNANDO BOSIGAS PUERTO
YEIMY ANDREA CANO M
DAVID POLANIA MEJIA

Entendimiento del negocio

EN EL SECTOR AGRÍCOLA SE PIERDEN 6 MILLONES DE TONELADAS DE ALIMENTOS AL AÑO

60% de la comida que se consume el país proviene del sector agrícola

A corte de mayo de 2022 las ventas al exterior de este sector sumaron US$5.055 millones



https://www.agronegocios.co/agricultura/en-el-sector-agricola-se-pierden-6-millones-de-toneladas-de-alimentos-al-ano-2706145#:~:text=La%20p%C3%A9rdida%20de%20alimentos%20es,58%25%20de%20los%20alimentos%20disponibles.

OBJETIVO GENERAL

Crear un sistema que permita predecir la producción de un determinado cultivo de acuerdo con características más importantes, mediante la implementación de técnicas de Data Science.

OBJETIVOS ESPECÍFICOS

- Implementar una metodología CRISP-DM
- Entrenar y encontrar el modelo de Machine Learning que mejor prediga la Producción (t).

Agricultor

Demographics

Guillermo, 40 años, se dedicó al tema de la agricultura desde hace poco, vive de su finca y lo que produce en ella.

Behaviors & Habits

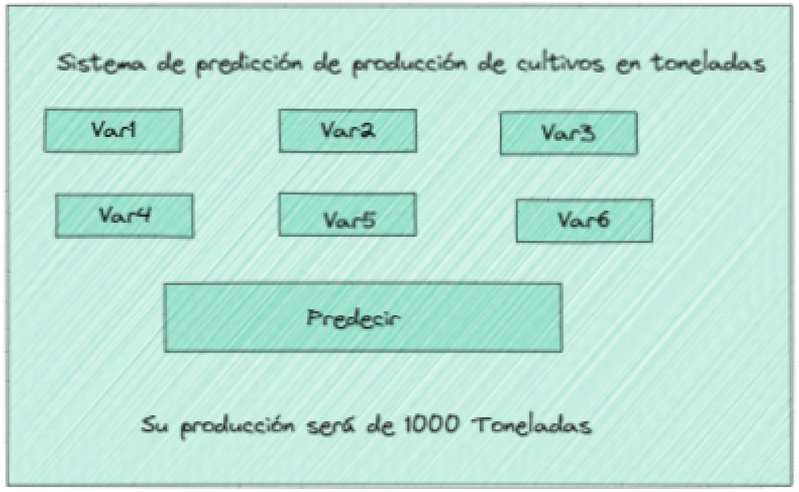
Es una hombre trabajador y perseverante, se levanta todas las mañanas muy temprano para cuidar de sus cosechas.
Le gusta compartir tiempo con su familia en la finca.

Pain Points & Frustrations

Guillermo ha tenido inconvenientes para conocer las cantidades que debe sembrar para tener una cosecha rentable, además, aún no conoce cuál es el mejor cultivo que se siembra en donde está.

Needs & Goals

- Le gustaría contar con un servicio que le permita conocer cuales cultivos producen más en el lugar donde se encuentra
- Le gustaría conocer las cantidades que debería sembrar para tener una cosecha productiva



DATOS

Dataset: Base Agrícola de Cultivos tomado de datos abiertos del gobierno de Colombia

- Tiene un tamaño de 206,068 registros con 17 columnas.

Nombre de la columna	Descripción
COD DEPARTAMENTO	Código del departamento, según lo establecido por el DANE
DEPARTAMENTO	Departamento Colombiano
COD MUNICIPIO	Código del municipio, según lo establecido por el DANE
MUNICIPIO	Municipio Colombiano
GRUPO DE CULTIVO	Categoría del cultivo
SUBGRUPO DE CULTIVO	Tipo de cultivo según categoría
CULTIVO	Nombre del cultivo
DESAGRAGACION REGIONAL Y/O SISTEMA PRODUCTIVO	Nombre genérico del cultivo
AÑO	Año de producción
PERIODO	Periodo médico, siendo A los primeros 6 meses y B los últim
ÁREA SEMBRADA (ha)	Área sembrada en hectáreas
ÁREA COSECHADA (ha)	Área cosechada en hectáreas
PRODUCCIÓN (t)	Tiempo de producción
RENDIMIENTO (t/ha)	Rendimiento de la cosecha
ESTADO FISICO PRODUCCION	Estado del producto
NOMBRE CIENTIFICO	Nombre científico del cultivo
CICLO DE CULTIVO	Ciclo del cultivo en el país

Preparación de datos

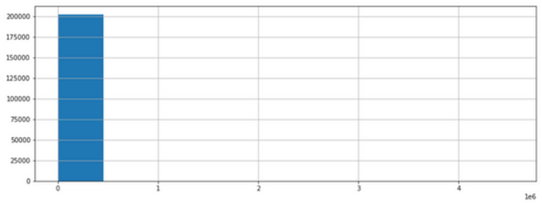
Variables de entrada

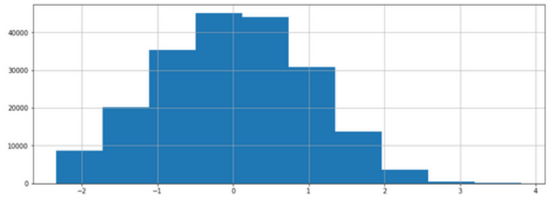
- DEPARTAMENTO
- MUNICIPIO
- GRUPO DE CULTIVO
- PERIODO
- ESTADO FISICO PRODUCCION
- CICLO DE CULTIVO

Transformación de los datos

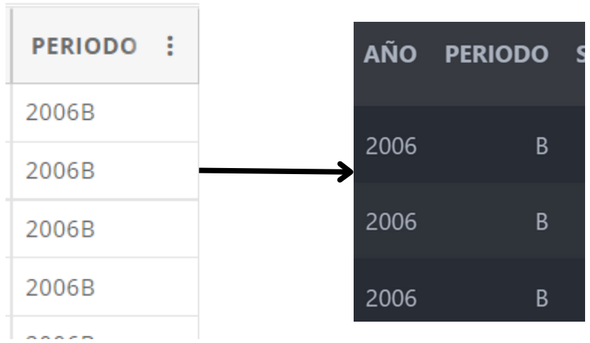
1. Eliminación de valores en 0 en la variable objetivo
2. Extracción de información de columna Periodo que contiene los valores de el año y el periodo en dos columnas distintas.
3. Conversión de variables categóricas, mediante One Hot encoding. Debido a que tenemos muchas variables categóricas resultantes
4. Min Max Scaler. Para estandarizar los datos
5. Transformación Logarítmica

Producción





Periodo



Modelado

Modelo	Error datos entrenamiento	Error datos evaluación	R-2 datos entrenamiento	R-2 datos evaluación
Regresión Lineal	5867.559	6247.714	0.45272	0.45971
Ridge	5790.458	6170.881	0.4526	0.45884
Lasso	5602.839	5975.188	0.45189	0.45818
SVM	-	-	-	-
Random Forest	158.209	483.062	0.99855	0.98918
Red Neuronal	3338.669	3707.196	0.39302	0.37172
Random Forest – Grid Search		-	-	-
Random Forest Selección manual de Hiperparámetros 1	2607.9341478175024	2839.0088205477714	0.80407	0.81717
Random Forest Selección manual de Hiperparámetros 2	2962.017009461782	3331.731134240868	0.31266	0.30847

Tabla 1. Comparación de evaluación de modelos

Evaluación

	Mean Absolute Error	R2 Score
Datos de entrenamiento (train)	0.0702	0.98751
Datos de prueba (test)	0.1807	0.92368

	departamento	municipio	grupo_cultivo	year	period	area	estado_produccion	ciclo_cultivo	Producción (t)	Y_predict	Y	Y2
0	CAQUETA	SAN JOSE DEL FRAGUA	FRUTALES	2013	C	7	FRUTO FRESCO	PERMANENTE	-0.576625	-0.586833	45.0	44.097193
1	ANTIOQUIA	CALDAS	TUBERCULOS Y PLATANOS	2012	A	8	TUBERCULO FRESCO	TRANSITORIO	0.046518	0.006185	160.0	147.065943
2	VALLE DEL CAUCA	LA CUMBRE	TUBERCULOS Y PLATANOS	2008	C	108	FRUTO FRESCO	PERMANENTE	0.542553	1.023464	465.0	1389.506006
3	BOYACA	DUITAMA	HORTALIZAS	2017	A	300	HORTALIZA FRESCA	TRANSITORIO	1.747958	1.709629	8260.0	7483.066646
4	CAUCA	CALOTO	LEGUMINOSAS	2006	B	3	LEGUMINOSA FRESCA	TRANSITORIO	-1.396000	-1.691238	9.0	4.945505
5	HUILA	TIMANA	CEREALES	2018	A	120	GRANO SECO	TRANSITORIO	-0.003921	0.132176	144.0	191.588295
6	ANTIOQUIA	SANTUARIO	LEGUMINOSAS	2017	A	420	GRANO SECO	TRANSITORIO	0.648040	0.645008	588.0	584.023284
7	VALLE DEL CAUCA	BUGA	CEREALES	2006	B	260	GRANO SECO	TRANSITORIO	0.995014	0.956622	1300.0	1188.736425
8	META	CUMARAL	OTROS PERMANENTES	2016	C	26	CAFE VERDE EQUIVALENTE	PERMANENTE	-0.874952	-0.880927	25.0	24.709247
9	NARIÑO	ALBAN	CEREALES	2017	B	20	GRANO SECO	TRANSITORIO	-0.874952	-0.693282	25.0	35.722959

-Áreas menores a 30: 111.80
-Áreas entre 30-150: 86.47
-Áreas entre 150 – 1200: 69.73
-Áreas mayores a 1200: 50.64

Departamento	diferencia promedio de predicción vs original %
AMAZONAS	187.101402
GUAVIARE	150.744445
VAUPES	107.377363
ATLANTICO	92.447821
CAUCA	85.446929

Tabla 2. Top de promedio de porcentaje de error por departamento

Despliegue

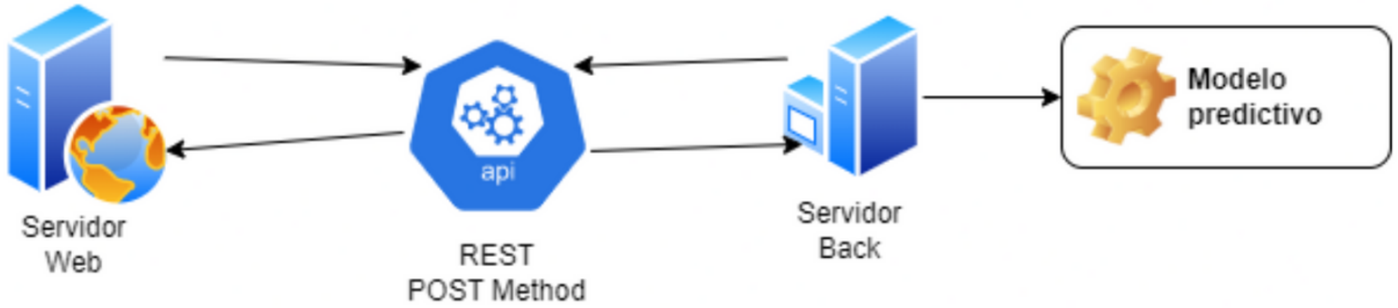
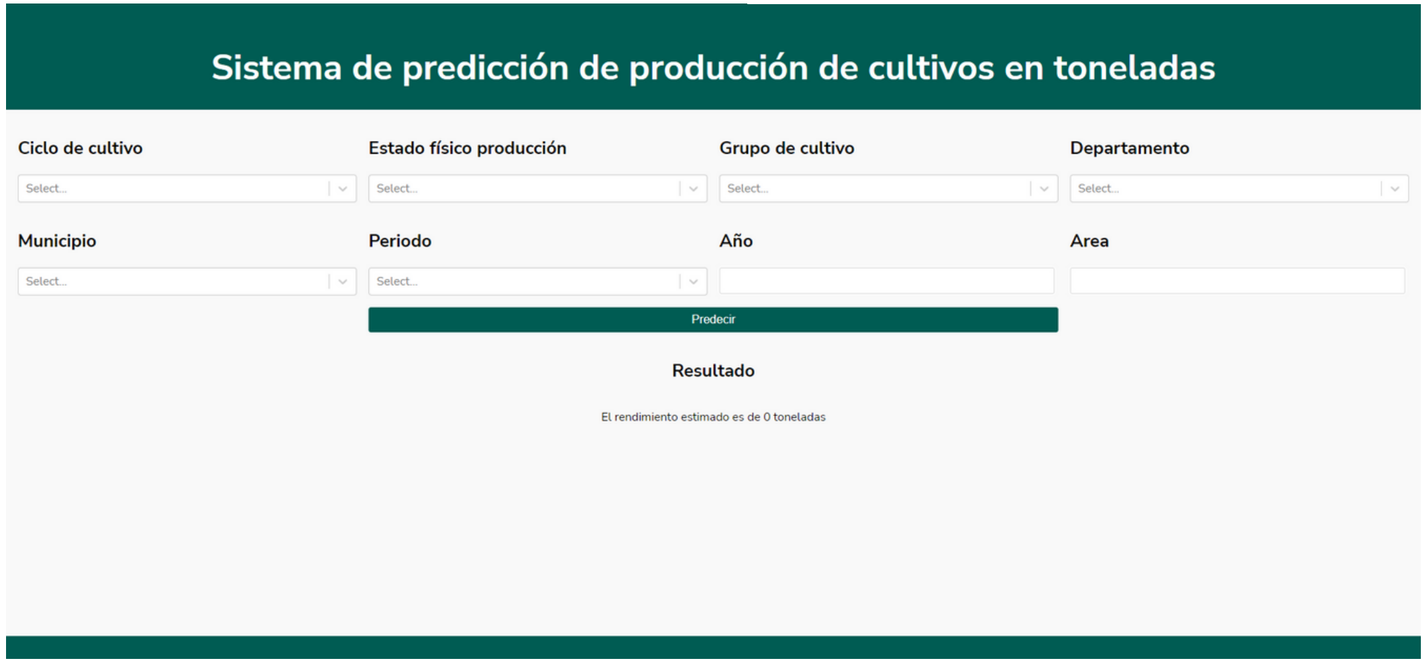


Figura 4. Diagrama de Arquitectura



Conclusiones

¿Qué condiciones considera que deberían tener los datos para obtener mejores resultados?

Por una parte, contar con datos de los años más recientes dado que solo se tiene información del 2018. Se identificó que departamentos con menos registros fueron los que en promedio tuvieron más errores en la predicción.

Además, sería bueno tener datos adicionales que también contengan información de las frutas y otros cultivos.

¿El modelo obtenido es suficiente para soportar la necesidad u oportunidad de negocio identificada?

- El modelo obtenido no es suficiente, dado que vemos que el error absoluto puede es más alto de 33 (valor objetivo era de 30) pero consideramos que es un primer punto de apoyo para aquellos departamentos en los que tiene mejor rendimiento: Cundinamarca, Caldas, Magdalena, Boyacá, Cesar, Guainía, Antioquia, Valle del Cauca, Huila, San Andrés y Providencia, Aruaca, Norte de Santander