Contextualización

DATASET

- Se utiliza el dataset: "Base Agrícola Cultivos" tomado de datos abiertos del gobierno de Colombia

OBJETIVO GENERAL

- Predecir la producción que puede tener una cosecha de un determinado cultivo por departamento y periodo del año. 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).

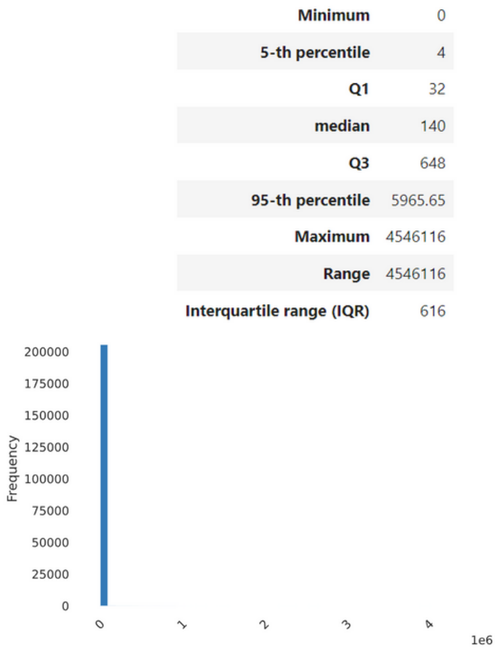
Preparación de datos

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



Diccionario de datos

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
DESAGREGACION 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 últimos
Área Sembrada (ha)	Área sembrada en hectáreas
Área Cosechada (ha)	Área cosechada en hectáreas
Producción (t)	Producción en toneladas
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

Estrategia de validación y selección del modelo

Métricas de rendimiento

Para calcular el Error en las predicciones del modelo se escogió el error absoluto medio o MAE para poder tener las mismas unidades que se tienen de la variable objetivo. A demás se utiliza R2 para saber en qué porcentaje la variable objetivo se puede explicar por las variables de entrada

Aceptación del modelo

Conociendo los percentiles de la variable objetivo: 90P=2655, 50P=35 y 25P=35. Lo ideal es tener errores en la misma magnitud no superiores a 35. Respecto al R2, se espera tener valores superiores al 70%.

Modelos Candidatos

Al ser un escenario de regresión, los modelos candidatos fueron:

- Random Forest
- Regresión Lineal (Con y sin regularización)
- Regresión Polinomial
- Red neuronal (Perceptrón simple)
- SVM

Construcción del modelo

Modelos entrenados

Al ser un escenario de regresión, los modelos candidatos fueron:

- Random Forest
- Regresión Lineal (Con y sin regularización)
- Regresión Polinomial
- Red neuronal (Perceptrón simple)
- SVM

Se utilizó 80% de los datos para entrenamiento y 20% de test, para todos los modelos.

Selección de hiperparámetros

Modelo	Hiperparámetros	Comentario
Red neuronal	hidden_layer_sizes=200 alpha=0.001	
Random Forest * (Search Grid)	'n_estimators':[100,150,200,500], 'max_depth': [None,1,2], 'min_samples_split':[1,2,4],	No funcionó
Manual 1 random forest:	(n_estimators=150, max_depth=2, min_samples_split=2)	
Manual 2 Random Forest	(n_estimators=200, max_depth=2, min_samples_split=0.1)	

Evaluación del modelo

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 de comparación de evaluación de modelos

Mejor modelo

Como se puede observar, los primeros modelos como las regresiones lineales tenían un error significativo, aunque tanto Ridge como Lasso bajaron un poco el error aun así no eran viables. De igual manera, la red neuronal obtuvo errores muy altos. Random Forest obtuvo mejores resultados que los demás, por lo cual se optó por una búsqueda de hiperparametro

Conclusiones

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

Muchas de las variables observadas, se encontraban desbalanceadas.

Además se evidenció que hay outliers que son demasiado grandes, sin embargo, no los hemos eliminado, ya que después de realizar la búsqueda correspondiente, corresponden a datos reales, por lo cual no pueden ser eliminados, pero sí alteran drásticamente el análisis de la variable

¿Cuáles son las mayores dificultades que se han tenido en el proyecto?

- Debido al gran volumen de datos, los modelos tardaron mucho tiempo en entrenar.
- Los modelos obtenidos han tenido un margen de error muy alto, esto puede ser causado por el desbalanceo de los datos ya mencionado.
- La variable de predicción tiene un rango demasiado alto, lo cual hace que sea más dificil para el modelo entrenar sobre los datos.

¿Qué estrategias se plantean para mitigarlas?

- Implementar estrategias de balanceo de datos.
- Eliminación de outliers.
- Separación de los valores altos de los bajos en la variable objetivo, entrenamiento individual de los datasets resultantes y finalmente realizar un ensembre que de el valor real.

¿El mejor modelo obtenido hasta el momento es suficiente para soportar el problema u oportunidad de negocio identificada?

Nuestro mejor modelo, es Random Forest, sin embargo, tiene un margen de error alto, por fuera de los parámetros estáblecido,

¿Cómo se usará este modelo dentro del producto o solución que se construirá?

- Elección de los departamentos más productivos para determinado tipo de cultivos.
- Para obtener X cantidad de producción se debe sembrar Y cantidad de area de Z determinado producto.
- Selección del mejor momento del año, para cultivar ciertos cultivos.