PROYECTO FINAL MINE-4101: CIENCIA DE DATOS APLICADA - SEGUNDA ENTREGA.

Presentado por:

Oscar Fernando Bosigas Puerto – 202220008 - o.bosigas

Yeimy Andrea Cano M – 202213304 - y.cano

David Polania Mejia - 202213328 - d.polaniam

Universidad de los Andes

Bogotá D.C., Colombia.

04/11/2022

Maestría en Ingeniería de la Información - MINE

Contenido

Introducción.

El presente trabajo busca presentar los avances del proyecto final de la asignatura ciencia de datos aplicada. Para la segunda entrega se tiene como objetivo finalizar actividades de entendimiento de datos, preparación de estos para construcción de un modelo basado en Machine Learning (ML). Y realizar el entrenamiento de los primeros modelos.

Para el desarrollo de este, se utilizó la metodología CRISP-DM, en donde se siguieron las etapas de metodologías para proyectos de ciencias de datos, se inicia con el entendimiento del negocio y objetivos, con un repaso de entendimiento de datos que nos permite ser más claros en las decisiones del proceso de preparación para la construcción del modelo de ML.

Enlace al repositorio.

https://github.com/OscarBosigas/Proyecto-Final-Ciencia-de-Datos-Aplicada/

1. Entendimiento del negocio

Contextualización del problema

La agricultura es una actividad económica relevante para el gobierno colombiano y dadas las ventajas geográficas y naturales con las que cuenta el país, es un enfoque priorizado en el desarrollo productivo. En el caso de la agricultura colombiana, se requiere el desarrollo de una serie de sistemas de información que permitan que entidades territoriales y los mismos agricultores tomen decisiones y se informen sobre la actividad. Este proyecto propone un Sistema de Información predictiva para la agricultura en Colombia, donde mediante la utilización de técnicas avanzadas de Data Science y Machine Learnig se pueda crear un modelo que permita predecir el comportamiento de los cultivos en los siguientes años igualmente, presentando esto visualmente en una herramienta que nos permita crear un dashboard de como los

cultivos se han comportado en años anteriores y su tendencia en el futuro, con el fin de identificar oportunidades y datos que ayuden a la toma de decisiones en el país.

a. Objetivos del proyecto

i. Objetivo General

- Predecir la producción en toneladas, que puede tener la cosecha de un determinado cultivo por departamento y periodo del año en que se siembra, mediante la implementación de técnicas de Data Science.

ii. Objetivos Específicos

- Implementar una metodología CRISP-DM.
- Entrenar y encontrar el modelo de Machine Learning que mejor prediga la variable objetivo Producción (t).
- Encontrar las variables más significativas para el modelo a la hora de predecir la producción (t).

2. Entendimiento de los datos

a. Información del dataset

El dataset proviene del sitio de datos abiertos del gobierno colombiano. Dicho dataset se puede descargar en formato csv, del cual se tiene una copia en el repositorio del proyecto para que el notebook pueda tomar los datos desde el repositorio.

El dataset tiene un tamaño de 206,068 registros con 17 columnas.

b. Diccionario de datos

El diccionario de datos que se encuentra en la misma fuente indica:

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

De acuerdo con las necesidades del negocio se toma la columna “Producción (t)” como variable objetivo a predecir, la cual es de tipo numérico y por lo tanto este un problema de regresión.

c. Exploración de datos:

Se puede observar que hay datos faltantes para algunas columnas:

- MUNICIPIO (1 dato faltante)
- Rendimiento (3433 datos faltante)
- NOMBRE CIENTIFICO (2857 datos faltantes)

Se puede observar que para todos los valores de 'Rendimiento (t/ha)' en nulos, los valores de 'Área Cosechada(ha)' y ' Producción (t)' están en 0. Los 3433 registros corresponden al 1.67% de los datos. También se observa alta correlación entre Rendimiento y la variable objetivo, esto de acuerdo con que la primera es consecuencia de la segunda.

Se hace la revisión de los valores en 0 para la variable objetivo (Producción) y se encontraron 3807 registros, que corresponden al 1.85% de los datos. De los cuales 164 registros (0.08%) son datos en los que la columna cosecha también está en 0, y aunque para los demás si se tiene un valor válido en área de sembrada no se considera prudente realizar algún tipo de imputación sobre los valores objetivos en 0, y dado que no es un porcentaje muy alto se prefiere eliminar estos registros.

Respecto a la búsqueda de datos duplicados, no se encontró ninguno.

Se encontró alta cardinalidad para las columnas: SUBGRUPO DE CULTIVO (120 diferentes categorías), CULTIVO (223 diferentes categorías), DESAGREGACIÓN REGIONAL Y/O SISTEMA PRODUCTIVO (271 diferentes categorías) y NOMBRE CIENTIFICO (214 diferentes categorías). En la revisión de datos en el notebook se puede establecer que la columna GRUPO DE CULTIVO solo tiene 13 categorías y que puede agrupar las anteriores, por lo cual con esta se suple las demás de alta cardinalidad.

No se encontraron problemas de calidad en el caso de los nombres de los departamentos. Ni en los casos de los nombres CICLO DE CULTIVO o ESTADO FISICO PRODUCCION.

En el caso de la columna GRUPO DE CULTIVO, las categorías más frecuentes contienen alrededor del 95% de los datos.

3. Preparación de los datos

a. Limpieza de datos

En el caso del registro faltante para el caso de “MUNICIPIO”, se hace el ajuste manualmente, teniendo en cuenta que se tiene el código del municipio “27077” que corresponde a “BAJO BAUDO”.

Dado que “Rendimiento” es una columna que es resultado de la variable objetivo 'Producción (t), de antemano conocemos que no la vamos a requerir, así como tal no hay imputación para estos valores nulos, sino que se elimina la columna.

Como se comentó en la exploración de datos, en los casos de variable objetivo iguales a 0 se eliminan dado que no se considera apropiado modificar la columna a predecir.

A causa de la alta correlación entre la columna “Área cosechada” y la “Área Sembrada” (0.97), se decide eliminar la columna, y dentro de datos de entrada solo usar “Área Sembrada”, que para casos futuros es el valor que si se puede conocer de cuánto se va a sembrar.

Como las columnas “CÓD. DEP”. y “DEPARTAMENTO” corresponden a lo mismo, se decide dejar “DEPARTAMENTO”, dado que al ser un código el modelo no lo tome como número ordinal, sino que posteriormente a departamento se haga una transformación One-Hot-Encoder. Se aplica el mismo razonamiento para el caso de “CÓD. MUN.” Y “MUNICIPIO”.

Como se comentó en la sección de exploración, dado la alta cardinalidad se eliminan las columnas: SUBGRUPO DE CULTIVO, CULTIVO, DESAGREGACIÓN REGIONAL Y/O SISTEMA PRODUCTIVO y NOMBRE CIENTIFICO.

b. Formatear datos

En el caso de la variable “periodo”, se observa que el valor corresponde al año y este puede o no estar concatenado a un carácter que hace referencia al periodo del año. De acuerdo con el entendimiento de datos, por ejemplo:

2007A: Es el periodo A del año 2007 (Primer semestre)
2007B: Es el periodo B del año 2007 (Segundo semestre)
2007: El periodo es para todo el año.

Por lo que se hace una transformación para tener solo 3 posibles valores: A, B o C en el caso de todo el año.

De esta manera el set de datos queda con 202261 registros y 9 columnas: Departamento, Municipio, Grupo de cultivo, Año, Periodo, Área sembrada, Estado físico producción, Ciclo del cultivo y Producción (que corresponde a nuestra variable objetivo)

	DEPARTAMENTO	MUNICIPIO	GRUPO DE CULTIVO	AÑO	PERIODO	Área Sembrada (ha)	Producción (t)	ESTADO FISICO PRODUCCION	CICLO DE CULTIVO
0	BOYACA	BUSBANZA	HORTALIZAS	2006	B	2	1	FRUTO FRESCO	TRANSITORIO
1	CUNDINAMARCA	SOACHA	HORTALIZAS	2006	B	82	1440	FRUTO FRESCO	TRANSITORIO
2	CUNDINAMARCA	COTA	HORTALIZAS	2006	B	2	26	FRUTO FRESCO	TRANSITORIO
3	NORTE DE SANTANDER	LOS PATIOS	HORTALIZAS	2006	B	3	48	FRUTO FRESCO	TRANSITORIO
4	NORTE DE SANTANDER	PAMPLONA	HORTALIZAS	2006	B	1	5	FRUTO FRESCO	TRANSITORIO

Fig. Imagen de muestra de los primeros datos del datase, luego de un primer preprocesamiento de datos.

Dado que, de las 9 columnas, se tienen 3 de tipo numéricas y 6 categóricas no ordinales, se procede a hacer una transformación de tipo One-Hot-Encoder, que nos

deja como resultado un total de 1094 columnas (la mayoría a causa de la columna municipios, pero que se prefiere mantener por interés del negocio).

Dado que se conoce que se obtienen mejores modelos con datos escalados, procedemos a aplicar la función MinMaxScaler para poder normalizar todos nuestros datos, a excepción de la columna "Producción (t)" que corresponde a columna objetivo de predicción.

Se identifican valores outliers superiores para área sembrada (75 percentil está en 156 mientras hay valores máximos en 47,403), y también en la variable objetivo (muy posiblemente como resultado de contar con áreas tan altas). Para los valore outliers a veces se presenta una única instancia la cual contiene valores muy grandes. Un ejemplo de ello, es la caña de azúcar. En primera instancia no se quiere intervenir dichos valores dada la naturaleza de variedad del problema, se prefiere revisar el comportamiento y métricas del problema y si es el caso tomar decisiones posteriormente.

4. Modelado

Para hacer el modelamiento, se hace una separación de datos: 80% de los datos son para entrenamiento y un 20% de los datos para Evaluación del modelo.

5. Estrategia de validación y selección de modelo:

a. Selección de técnica de modelado

Se utilizaron modelos de aprendizaje supervisado como: regresión lineal, regresión lineal regularizada (Ridge, Lasso), SVM, Random Forest y red neuronal**.** Primero se ejecutaron unas versiones base de estos, para luego determinar si se necesitaba alguna agregación, como el caso del hiperparametro para el modelo de Random Forest. Algunos modelos como la regresión polinomial no fueron considerados debido a la cantidad de columnas generadas al realizar el One-Hot-Encoding, sin embargo, se planea realizar una reducción de dimensionalidad PCA para posteriormente probar este modelo y los anteriormente mencionados.

b. Definición de métricas de aceptación del modelo

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.

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%.

c. Entrenamiento del modelo

Se realiza el enteramiento usando los algoritmos descriptos anteriormente, seguidos de la evaluación con datos de entrenamiento y luego con datos de evaluación.

En el caso de regresión al comparar las métricas de error MAE, se observa que el error pasa de 5867 a 6247, lo cual si muestra que existe un sobreajuste en el modelo. Por lo que se aplican las dos técnicas de regularización para ver con cual se obtiene un mejor resultado.

Por otra parte, se plateó entrenar un modelo con SVM, pero los tiempos de ejecución eran muy altos y no se logró obtener resultados. Se implementaron 3 modelos modelo Random Forest, uno utilizando Search Grid, el cual no funcionó, y los otros dos se pasaron manualmente los hiperparámetros y se obtuvieron mejores resultados. Finalmente, se entrenó un modelo de Red Neuronal.

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)	

6. Evaluación

a. Resumen de resultados

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

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 cumplen con los estándares establecidos. Por otra parte, la red neuronal obtuvo errores muy altos. Finalmente, Random Forest obtuvo mejores resultados que los demás, por lo cual se optó por realizar una búsqueda de hiperparametros.

b. Selección del modelo

Desafortunadamente, no se logró entrenar un modelo que cumpliera con los estándares establecidos, sin embargo, el modelo que obtuvo mejores resultados es Random Forest sin hiperparámetros.

7. Conclusiones

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

- Clases balanceadas: Se identificó que muchas de las variables observadas, se encontraban desbalanceadas, lo cual dificulta la capacidad de predicción del modelo.

- Además, se evidenció que algunos de los outliers produzcan clases muy dispersas, sin embargo, se tomó la decisión de eliminados, ya que se preguntó al dueño del dataset, pero al no tener respuesta, se realizó una pequeña investigación, en donde se concluye que estos valores sí pueden ser factibles.

b. ¿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 objetivo tiene un rango demasiado alto, lo cual hace que sea más difícil para el modelo entrenarse.

c. ¿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 del valor real.

d. ¿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 establecido,

e. ¿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 área de Z -- determinado producto.

- Selección del mejor momento del año, para cultivar ciertos cultivos.