

Proyecto Aprendizaje Automático

Jhonnatan Alexander Pinilla Bolaños

Juan Gabriel Cortes Villamil

Dayler Haver Oviedo Capera

Camilo Andres Barbosa Molina

Facultad de Ciencias Naturales e Ingeniería

Universidad Jorge Tadeo Lozano

009841 – Aprendizaje Automático

Olmer Garcia Bedoya

Mayo 21 2022



Tabla de Contenido

Aprendizaje autónomo con XGBoost y Sector agrícola Colombiano	3
Introducción	3
Justificación	4
Metodología	6
Entendimiento de datos	6
Preparación de los datos	8
¿Modelo seleccionado y porque?	10
Resultados	10
Discusión	11
Conclusión	11
Enlaces	11
Referencias	12

Listado de Figuras

Fig 1. Visualización general de los datos del Dataset importado	6
Fig 2. Matriz de correlación de variables	7
Fig 3. Participación de los diferentes cultivos dentro del Dataset	8
Fig 4. Participación de los diferentes grupos de cultivos dentro del Dataset	9
Fig 5. Participación de los departamentos dentro del Dataset	9
Fig 6. Proceso para convertir variables categóricas a numéricas	10
Fig 7. Configuración de hiper parámetros del modelo	11

Aprendizaje autónomo con XGBoost y Sector agrícola Colombiano

Jhonnatan Pinilla, Camilo Barbosa, Juan Cortes, Dayler Oviedo

Universidad de Bogota Jorge Tadeo Lozano Carrera, 4 22-61 Bogotá, Colombia
{jhonnatana.pinillab, camiloa.barbosam, juang.cortesv, daylerh.oviedoc}@utadeo.edu

Resumen. En el presente documento se presentará el proyecto de aprendizaje automático, que tiene como fin obtener información importante que sirva como base para procesos de toma de decisiones sobre un contexto específico, todo esto por medio del tratamiento de datos con modelos de Machine learning. Para este proyecto es necesario el uso del lenguaje de programación de Python junto a la técnica de machine learning de XGBoost regresión, que nos ayudará a procesar un Dataset sobre [evaluaciones agrícolas municipales de Colombia](#)[3], tomada del repositorio nacional de Datasets, y con esto realizar un análisis descriptivo que determine la producción de cultivos de ciertos departamento de Colombia de acuerdo a características específicas de estas áreas

Palabras clave: Sector agrícola, XGBoost, Machine learning

Introducción

La baja productividad agrícola de Colombia limita la competitividad del país. Son necesarias una serie de acciones estructurales para acelerar las mejoras de la productividad y de la competitividad, y para facilitar el desarrollo rural, las prácticas sostenibles a largo plazo y la internacionalización de los productos producidos. En especial, el sector de la agricultura está sujeto a una gama amplia de políticas, así como a instrumentos específicos que generan incentivos y desincentivos para la productividad agrícola cuando se encuentran entre sí.

Por lo cual, una estrategia para incrementar la productividad agrícola requiere de un análisis integral de los instrumentos, modelos algorítmicos y a su vez de su interacción y relación con la productividad de la producción.

Este documento tiene como objetivo principal realizar un análisis, diagnóstico y predicción global de la productividad de la producción de cultivos cosechados del sector agrícola en municipios de Colombia, esto mediante la utilización de un Dataset proporcionado por el repositorio nacional de Datasets denominados datos abiertos, el cual nos brinda información en el sector de agricultura y desarrollo rural, como por ejemplo sobre el tipo de cultivo, el área sembrada, área cosechada, producción, rendimiento y el nombre de los municipios en donde se cosechan estos cultivos y con base en ello, presentar una propuesta concreta de hoja de ruta para aumentar la productividad del sector agrícola, y su impacto en los encadenamientos productivos, la sostenibilidad y la internacionalización.

Justificación

Este modelo predictivo dotará al sector agrícola de herramientas capaces de leer y manipular datos (características del cultivo y aspectos específicos). De esta manera, se puede obtener información cuantitativa temprana sobre el estado de los cultivos para implementar intervenciones efectivas, eficientes y específicas.

Este modelo de predicción del sector agrícola proporciona información útil para apoyar la toma de decisiones, como:

- Fenología del cultivo (etapa de desarrollo del cultivo o etapa de maduración del fruto)
- Riesgo de infestación por patógenos o plagas específicas en cada etapa del ciclo de cultivo
- Requerimientos de nutrientes de los cultivos
- Requisitos de cultivo

Los modelos predictivos son un apoyo invaluable para quienes trabajan en la agricultura. Son un elemento muy importante porque simplifican el trabajo y aumentan la eficacia en la toma de decisiones en el ámbito de la gestión del riego y la fertilización.

Para este tipo de problemas existen algunas soluciones ya en el mercado como lo son RawData[[1](#)] un modelo de predicción de cosecha validado en 32.606 parcelas, este modelo maneja cuatro (4) puntos importantes:

- **Predicción de volumen:** Visualiza los kg de cosecha y planifica tu comercialización
- **Predicción de maduración:** Conoce con antelación la maduración de tu cosecha.

- **Control y registro de visitas:** Conteo de aforos, plagas, enfermedades, puntos de control y mucho más.
- **Visión de negocio:** Centraliza toda la información y ten visión de negocio 360°

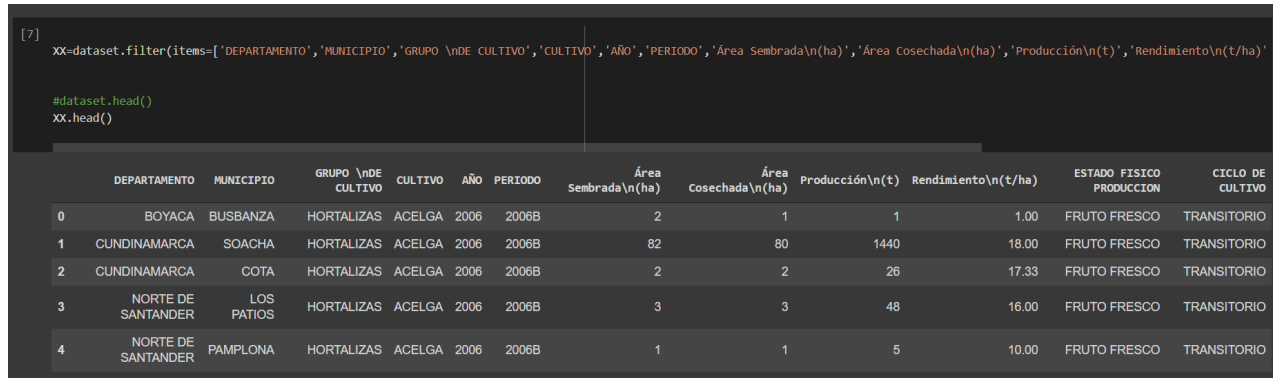
Igualmente están plataformas como AgriTech creada por Agricolus[[2](#)] simplifica y mejora el trabajo en el campo de los agricultores y operadores del sector para conseguir una sostenibilidad tanto económica como medioambiental. La web de Agricolus y la app monitorizan tus campos y te dan información en tiempo real de lo que sucede en tus cultivos, enfocando siempre en seis (6) puntos importantes:

- **Mapeo de campo:** Los sistemas GIS permiten mapear parcelas y georreferenciar toda la información
- **Imágenes de satélite:** Las imágenes satelitales permiten un seguimiento remoto eficaz de los cultivos
- **Modelos de predicción:** Los modelos de predicción suponen un valioso apoyo para la prevención de plagas y enfermedades
- **Soporte a las decisiones – DSS:** Los Sistemas de Soporte a la Decisión (DSS) asesoran a los agricultores sobre las acciones más adecuadas a realizar
- **Sensores:** Los sensores para la agricultura permiten detectar datos muy importantes sobre la salud de las plantas
- **Agricultura de precisión:** La agricultura de precisión permite llevar a cabo intervenciones agronómicas específicas gracias a herramientas innovadoras

Metodología

Entendimiento de datos

Para el trabajo se hace necesario el uso de un Dataset, el cual fue extraído de la página datos.gov (repositorio nacional Colombiano), se visualizan los datos en [google colab](https://colab.research.google.com/)



```
[7]
xx=dataset.filter(items=['DEPARTAMENTO', 'MUNICIPIO', 'GRUPO \nde CULTIVO', 'CULTIVO', 'AÑO', 'PERIODO', 'Área Sembrada\n(ha)', 'Área Cosechada\n(ha)', 'Producción\n(t)', 'Rendimiento\n(t/ha)']

#dataset.head()
xx.head()
```

	DEPARTAMENTO	MUNICIPIO	GRUPO \nde CULTIVO	CULTIVO	AÑO	PERIODO	Área Sembrada\n(ha)	Área Cosechada\n(ha)	Producción\n(t)	Rendimiento\n(t/ha)	ESTADO FISICO PRODUCCION	CICLO DE CULTIVO
0	BOYACA	BUSBANZA	HORTALIZAS	ACELGA	2006	2006B	2	1	1	1.00	FRUTO FRESCO	TRANSITORIO
1	CUNDINAMARCA	SOACHA	HORTALIZAS	ACELGA	2006	2006B	82	80	1440	18.00	FRUTO FRESCO	TRANSITORIO
2	CUNDINAMARCA	COTA	HORTALIZAS	ACELGA	2006	2006B	2	2	26	17.33	FRUTO FRESCO	TRANSITORIO
3	NORTE DE SANTANDER	LOS PATIOS	HORTALIZAS	ACELGA	2006	2006B	3	3	48	16.00	FRUTO FRESCO	TRANSITORIO
4	NORTE DE SANTANDER	PAMPLONA	HORTALIZAS	ACELGA	2006	2006B	1	1	5	10.00	FRUTO FRESCO	TRANSITORIO

Fig 1. Visualización general de los datos del Dataset importado

Este Dataset contiene algunos datos faltantes (na), por lo cual se eliminaron estos datos nulos. Además con la ayuda de la biblioteca de Python, Sweetviz, podemos tener un mejor análisis de las variables con las que cuenta este Dataset facilitando la selección de las variables de entrada y variable de salida que vamos a tomar para realizar el proyecto.

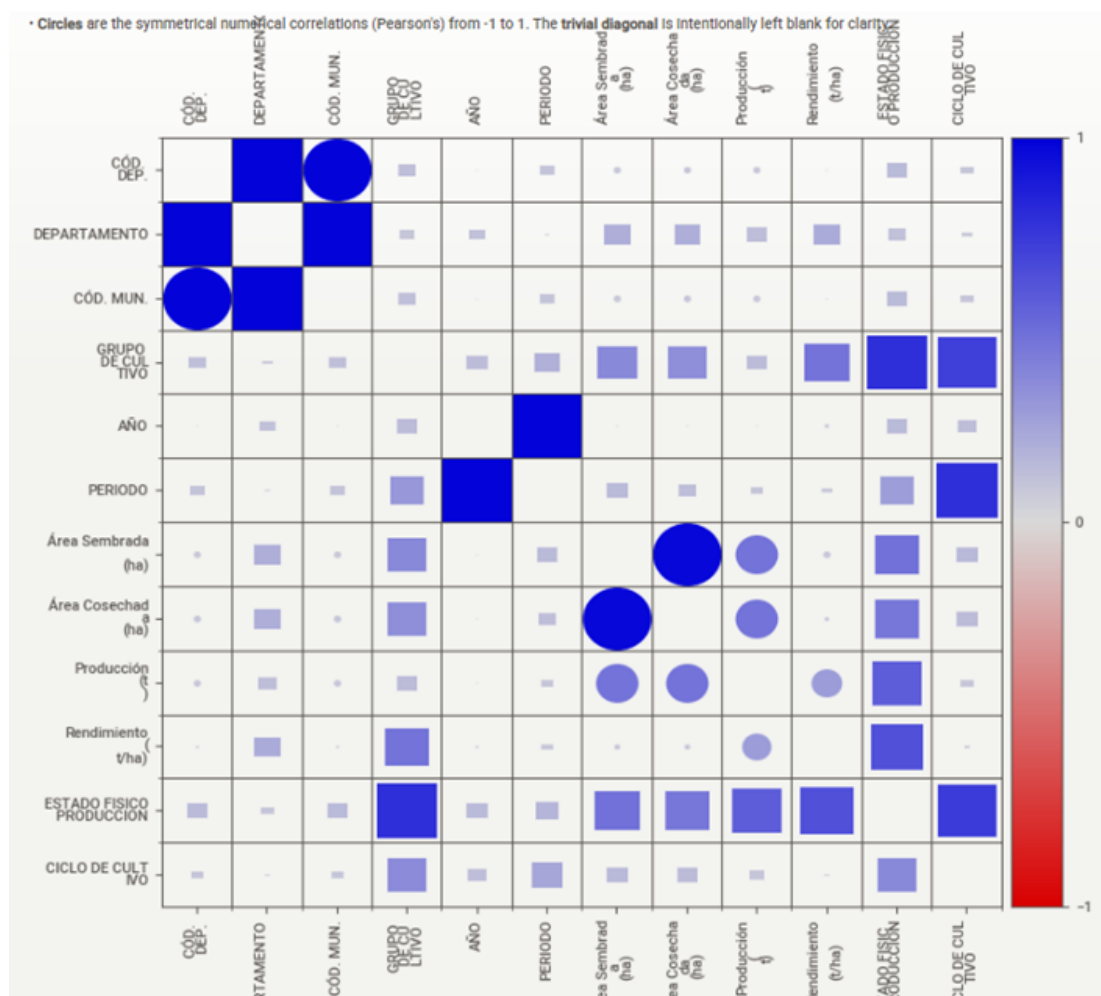


Fig 2. Matriz de correlación de variables

Como se puede observar en la anterior, se unifican todas las variables del Dataset, en donde los círculos representan correlaciones de variables numéricas y los cuadrados correlaciones categóricas y numéricas, también el tamaño de estas figuras nos dicen algo, si son grandes es que tienen una mayor correlación, por ejemplo se puede observar que la variable “área sembrada” se correlaciona con la variable “área cosechada”, este tipo de análisis nos ayudó a prescindir de la variable “área cosechada” ya que es algo redundante, de igual forma esto sucede con la variable “departamento” y la variable “código de departamento”. De las dieciséis columnas que ofrece este Dataset algunas de las columnas escogidas en este punto pueden ser excluidas con el desarrollo de este proyecto ya que quizás no aporten de una manera positiva al mismo.

Preparación de los datos

Se comienza aplicándole al Dataset la función “Dataset = Dataset.dropna()” que lo que hace es eliminar los valores Nan (no un número) y Nat (no un tiempo) es decir valores nulos, después de esto con ayuda de Sweetviz descartamos algunas variables que evidenciamos que pueden generar una sobrestimación en los datos. Esto no quiere decir que los datos sean atípicos, lo que sucede es que hay zonas en donde se cultivan proporciones muy grandes con respecto a la mayoría de los datos.

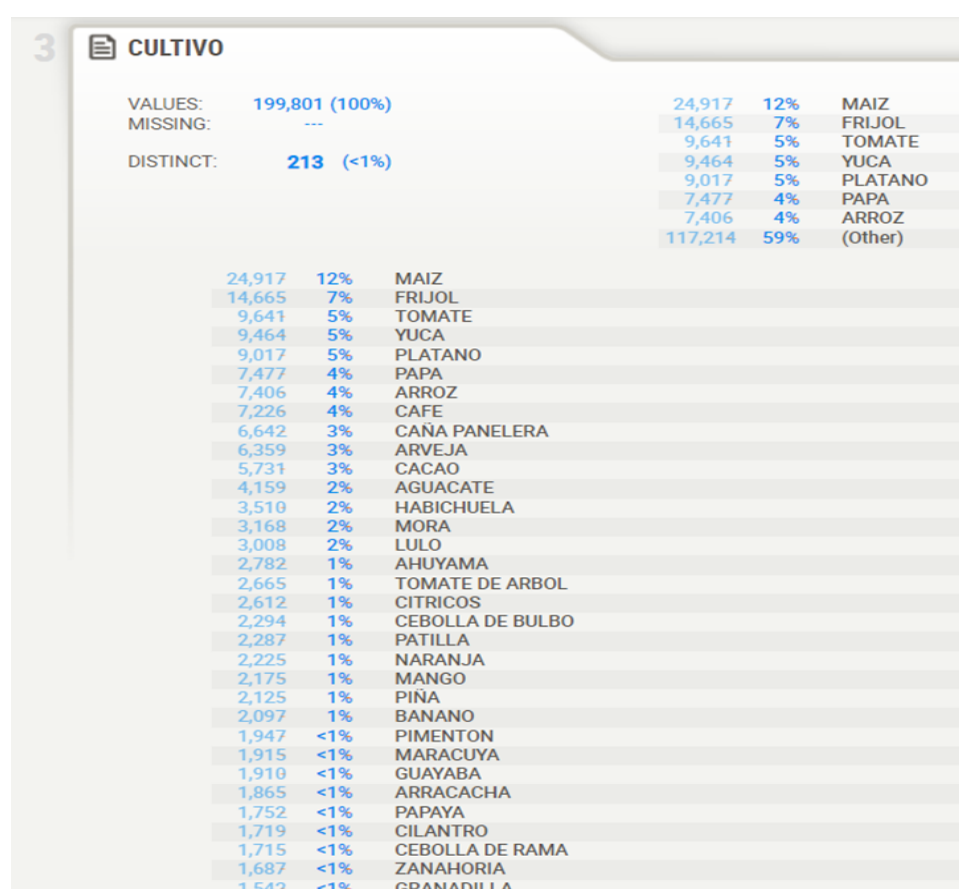


Fig 3. Participación de los diferentes cultivos dentro del Dataset

Dado que existen parcelas pequeñas entre una a cuatro hectáreas y pocos terrenos grandes que manejan hectáreas por arriba de las sesenta hectáreas de sembradío se pueden evidenciar datos fuera del rango como por ejemplo fibras, hongos y forestales.

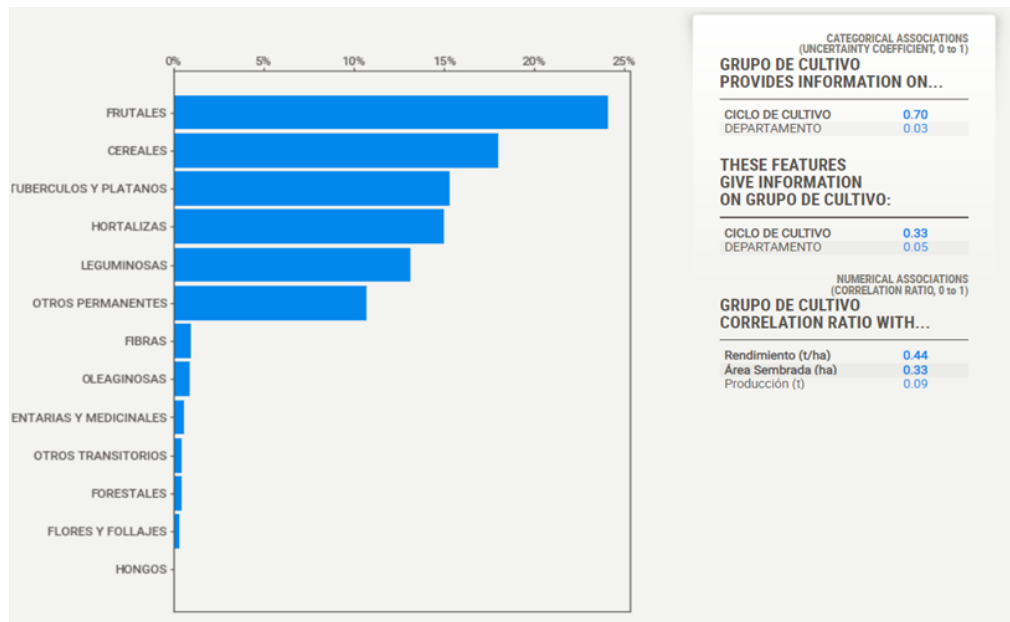


Fig 4. Participación de los diferentes grupos de cultivos dentro del Dataset

De acuerdo con este análisis se opta por quitar estos registros que no tienen una gran participación en el Dataset y que se salen del rango en las siguientes variables “departamento”, “cultivo” y “grupo de cultivo”, como se muestra en la siguiente gráfica, en donde seleccionamos los departamentos de Boyacá al Meta

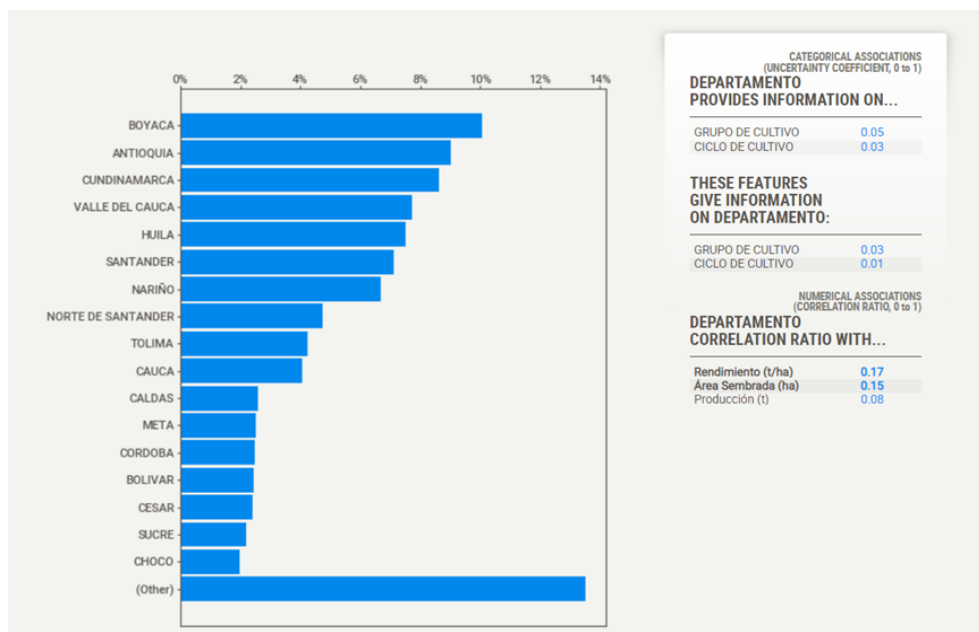


Fig 5. Participación de los departamentos dentro del Dataset

También con algunas variables se tiene que realizar el proceso de convertirlas de categóricas a numéricas con la siguiente función.

```
[ ] from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import OneHotEncoder

    encoder = OneHotEncoder(categories='auto',
                             drop='first', # devuelve k-1, usa drop=false para devolver k dummies
                             sparse=False)

    encoder.fit(X.fillna('Missing'))

    OneHotEncoder(drop='first', sparse=False)
```

Fig 6. Proceso para convertir variables categóricas a numéricas

¿Modelo seleccionado y porque?

Extreme Gradient Boosting (XGBoost) es el modelo que se ha seleccionado ya que es una biblioteca de código abierto que proporciona una implementación eficiente y efectiva del algoritmo de aumento de gradiente.

XGBoost es un componente clave para obtener soluciones para una variedad de problemas en las competencias de aprendizaje automático. Especialmente problemas de modelos predictivos de regresión implican la predicción de un valor numérico. XGBoost se puede utilizar directamente para modelado predictivo de regresión.[5]

Se seleccionó este tipo de modelo por su gran capacidad de velocidad de ejecución y el rendimiento del modelo. XGBoost domina los conjuntos de datos estructurados o tabulares sobre problemas de modelado predictivo de clasificación y regresión.

Resultados

Si no se llega a parametrizar correctamente el XGBoost este llega a producir sobre ajustes, entonces cuando se da un resultado en donde el error cuadrático medio que es una de las métricas da alrededor del 99.9 % y la desviación estándar es del 0.01% se entiende que el modelo está sobre ajustado. A esto se le da solución creando el objeto regresor del modelo XGBoost en donde colocando unos parámetros con nombre y valor estos quedan remplazados, en este caso se ajusta la tasa de aprendizaje (Learning_rate) a 0.05 ya que si el valor es muy alto el modelo aprende muy rápido y tiende a sobre ajustarse por ende utilizamos intervalos pequeños, número de estimadores igual a 40, que corresponden a qué

cantidad de árboles se va a modelar para que genere la predicción y finalmente max_depth nos da la profundidad de los árboles en este caso 3 grados de profundidad.

```
[ ] from xgboost import XGBRegressor  
  
regresor = XGBRegressor(learning_rate = 0.05, n_estimators=40, max_depth= 3)  
regresor
```

Fig 7. Configuración de hiper parámetros del modelo

Discusión

Este proyecto se relaciona directamente con la agricultura campesina, ya que implementando este modelo entrenado se puede llegar a mejorar la toma de decisiones por parte del campesino al momento de invertir en un terreno, o querer expandir su comercio. Este modelo puede ser implementado para crear una interfaz o un aplicativo en donde se puedan ingresar algunas características como el área a sembrar, el departamento y el tipo de cultivo que se quiera y con esto dar una visualización de la producción que se podría llegar a obtener.

Conclusión

Se puede concluir que este tipo de modelos constituyen una herramienta muy útil para poder desarrollar una agricultura eficiente y poder llegar a usar eficientemente el recurso del suelo, teniendo en cuenta la productividad deseada por los agricultores, este modelo se puede convertir en una herramienta que facilite y mejore la toma de decisiones al momento de invertir en un proyecto agrícola.

Enlaces

- [Ejercicio Colab](#)
- [Repositorio GitHub](#)

Referencias

1. RawData. (2022, 27 abril). *Predicciones*. Recuperado 19 de mayo de 2022, de <https://agrawdata.com/predicciones/>
2. Sgargi, C. (2022b, mayo 6). Agricolus - La plataforma para la agricultura de precisión. Agricolus. Recuperado 19 de mayo de 2022, de <https://www.agricolus.com/es/>
3. Unidad de Planificación Rural Agropecuaria - UPRA. (2021). Evaluaciones Agropecuarias Municipales – EVA. 2019 - 2020 [Data set]. Recuperado 19 de mayo de 2022, de <https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Evaluaciones-Agropecuarias-Municipales-EVA-2019-20/p5fp-pay3>
4. Calvo, L. A. (2020, mayo). *Estrategia de prediccion en procesos biologicos del campo agricola con datos limitados: casos de aplicacion en cafe y banano*. Doctorado en Ciencias Naturales para el Desarrollo Enfoque en Tecnologías Electrónicas Aplicadas. https://repositoriotec.tec.ac.cr/bitstream/handle/2238/11454/TFG_Alexander_Calvo.pdf?sequence=1&isAllowed=y
5. García-Arteaga, J. J., Zambrano-Zambrano, J. J., Alcivar-Cevallos, R., & Zambrano-Romero, W. D. (2020). Predicción del rendimiento de cultivos agrícolas usando aprendizaje automático. *Revista Arbitrada Interdisciplinaria Koinonía*, 5(2), 144. <https://doi.org/10.35381/r.k.v5i2.1013>
6. D. (2021, 13 marzo). Extreme Gradient Boosting (XGBoost) es una biblioteca de código abierto que proporciona. Top Big Data. Recuperado 19 de mayo de 2022, de <https://topbigdata.es/XGBoost-para-regresion/>