Introducción a la Ciencia de Datos con Python

Edwin Olivar Montes, Sandra Cotúa Barrera, Estefani Cárdenas Ramírez

Ciencias Agrarias
Ingeniería Agropecuaria
Universidad de Antioquia
Caucasia, Colombia

edwin.olivar@udea.edu.co
sandra.cotuab@udea.edu.co
estefani.cardenas@udea.edu.co

Abstract – Este documento contiene un análisis descriptivo y estadístico de la base de datos Evaluaciones Agrícolas Municipales en Colombia años 2006-2017, obtenida de la página de Datos Abiertos de Colombia. Se buscó, a través de diferentes modelos predictivos, obtener valores prospectivos de la variable Rendimiento (t/ha) haciendo uso del lenguaje de programación de Python. Las métricas de desempeño de los dos (2) modelos usados evidenciaron que el conjunto de variables y su disposición eran insuficientes para la predicción de la variable.

Introducción

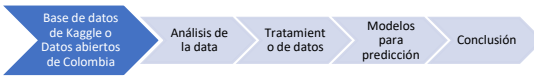
En la Universidad de Antioquia, Seccional Bajo Cauca, se viene fomentando la enseñanza y aprendizaje de la Ciencia de Datos. Para iniciar en este campo o afianzar los conocimientos se nos habilitó el curso de ***Introducción a la Ciencia de Datos*** con uso del lenguaje de programación Python y el entorno virtual de Colaboratory debido a que, en el conjunto de roles y actividades que conforman la ciencia de datos, obtener los conocimientos básicos es de gran importancia. Realizar un análisis de los elementos que conforma la data y el estado en que se encuentran es de gran importancia, esto nos dará una idea las acciones de limpieza, tratamiento, determinar las herramientas (librerías, funciones) y hasta el tipo de modelo a usar.

El ejercicio empieza con reconocer el formato o estructura de la base de datos, ya sea Excel, CSV entre otros, pasando por determinar la cantidad de variables y la naturaleza de dato al que pertenecen sus entradas o filas (flotante, entero, booleano o un carácter). Además, se recomienda identificar y tratar los datos ausentes.

Nuestro objetivo con el análisis de la base de datos de las Evaluaciones Agropecuarias Municipales desde el año 2006 hasta 2017, será conocer la variación en el ***área sembrada, cosechada, producción*** y como influencian la variable objetivo de ***rendimiento***. Luego, usar el mejor modelo que prediga el rendimiento de futuras producciones.

Materiales y Métodos

El algoritmo usado para el trabajo fue el siguiente:



a) Base de Datos Evaluaciones Agrícolas Municipales 2006 – 2017 de la pagina Datos Abiertos de Colombia.

La base de datos usada para el tratamiento y análisis fue las Evaluaciones Agropecuarias Municipales desde el año 2006 hasta 2017 en formato CSV conformada por 17 columnas, 206.068 filas. Con 0,17% de datos nulos. Las siguientes son las variables monitoreadas a través de la función *value_counts ()*:

Variables

- CÓD. \nDEP: o código de departamento que es asignado por el sistema de codificación para Departamentos y Municipios del territorio nacional (DANE). Tipo de dato: Entero.
- DEPARTAMENTO: Designa cada uno de los 32 departamento colombianos. Tipo de dato: Objeto.
- CÓD. MUN: o código de municipio que es asignado por el sistema de codificación para Departamentos y Municipios (DANE. Tipo de dato: Objeto
- MUNICIPIO: Designa 1018 municipios de Colombia. Tipo de dato: objeto.
- GRUPO \nDE CULTIVO: o grupo de cultivos. Con 13 grupos de cultivos. Tipo de dato: objeto
- SUBGRUPO \nDE CULTIVO: o subgrupo de cultivos. Con 120 subgrupos. Tipo de dato: objeto
- CULTIVO: Con 223 cultivos. Tipo de dato: objeto.
- DESAGREGACIÓN REGIONAL Y/O SISTEMA PRODUCTIVO: Con 271 entradas, en este se agrupa la categoría de cultivos por región y sistema productivo
- AÑO: Años es los que se recopiló la información que reposa en esta base de datos. Desde 2006 hasta 2017. Tipo entero.
- PERIODO: Corresponde a Periodo A para el primer semestre y periodo B para el segundo semestre. Tipo entero.
- Área Sembrada\n(ha): Indica el área en la que se estableció los tipos de cultivos. Tipo entero.
- Área Cosechada\n(ha): Hace referencia al área que, una vez sembrada, fue cosechada. Tipo entero.
- Producción\n(t): Muestra las toneladas producidas sobre el área cosechada por cultivo en el tiempo referenciado. Tipo entero.
- Rendimiento\n(t/ha): Rendimiento obtenido por la división entre el área producida y el área cosechada. Tipo entero.
- ESTADO FISICO PRODUCCION: Estado en el que fue recolectado el producto. 23 entradas o categorías. Tipo entero.
- NOMBRE \nCIENTIFICO: Nombre científico del cultivo.

Con 214 entradas o categorías. Tipo entero.

- CICLO DE CULTIVO: Indica cuántos de los cultivos registrados es transitorio, permanente o anual. Tipo entero.

b) Herramientas del Python

Este lenguaje de programación posee una serie de bibliotecas con sus correspondientes funciones, necesarias para el tratamiento de los datos y construcción de modelos predictivos.

c) Análisis de la Data

Se orientó hacía la identificación de la densidad, distribución y sesgo del conjunto de datos. Este último refleja una tendencia hacia la derecha con poca variabilidad y con una frecuencia de datos concentrada en las primeras clases*. Se procedió al tratamiento de datos al realizar imputación de datos nulos, normalización y/o estandarización de los datos.

Resultados y Análisis

Posterior a la limpieza y organización de los datos (eliminación de variables 13 e incorporación de la mediana en datos nulos 41%) se analizó la distribución de los datos a través de gráficas que permitían su visualización:

Para la variable objetivo:

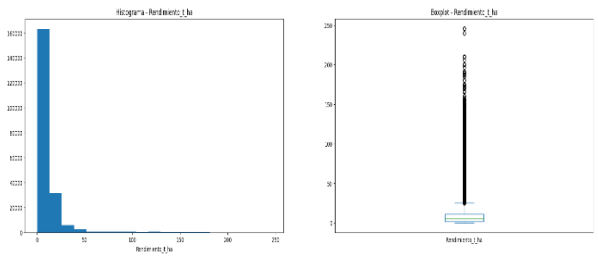


Ilustración 1Histograma_boxplot_Rendimiento

Para las variables independientes:

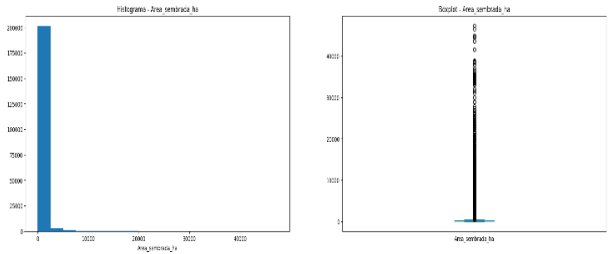


Ilustración 2 Histograma_boxplot_Area_sembrada_ha

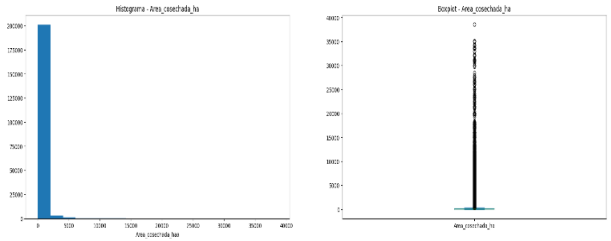


Ilustración 3 Histograma_boxplot_Área_cosechada_ha

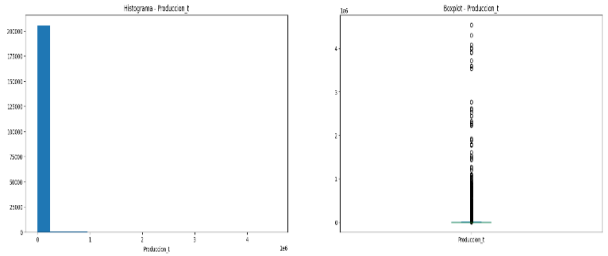


Ilustración 4 Histograma_boxplot_Producción_t

Para el conjunto de ilustraciones (1, 2, 3, 4) los datos, para cada vector, presentan poca variabilidad lo que aumenta

*Se usó la regla de Sturges para obtener el número de clases en los histogramas obteniendo 19. Formula (1+ 3,33*log10(n) donde n es el número de entradas de las variables en nuestro caso 206068.

colinealidad entre independientes y, en este caso, poca correlación con la variable objetivo (ilustración 5), concentrándose en las primeras clases, con sesgo hacía la derecha (ilustración 6) y el 12,6% aproximado de datos atípicos para variables independientes (26000) y 4,8% para la variable objetivo (10000).

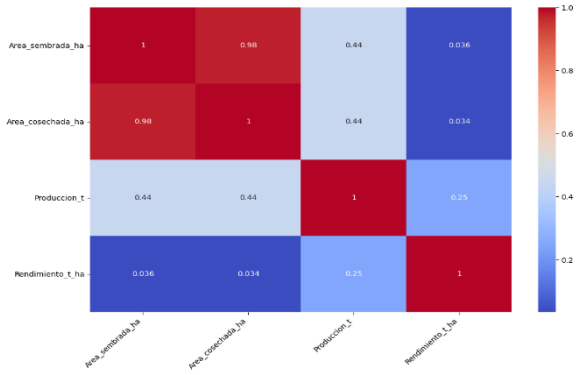


Ilustración 5Correlación de variables

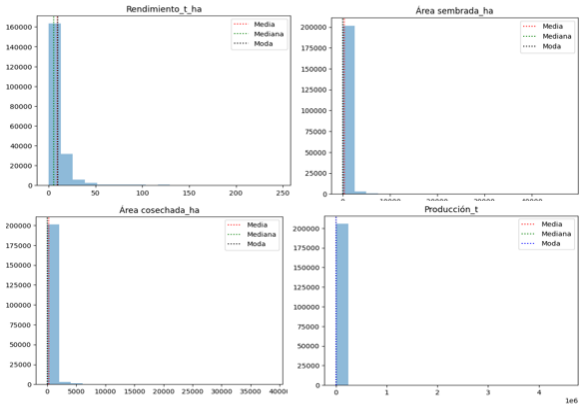


Ilustración 6 Estadísticos cola hacia la derecha

La imputación a través del cambio de datos atípicos por la mediana, solo cambió el sesgo del conjunto de variables esta vez hacía la izquierda a excepción del vector rendimiento (ilustración 7).

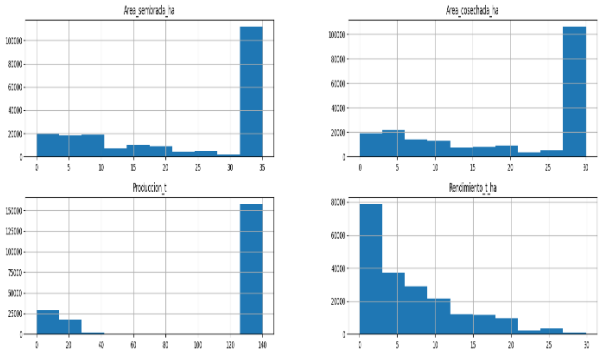
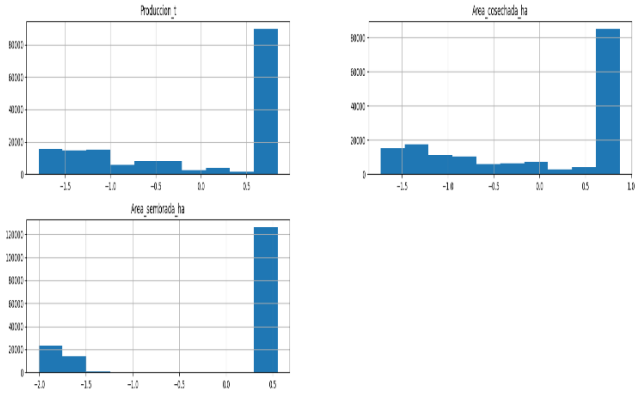


Ilustración 7Histograma con datos imputados

El último tratamiento fue el de normalización, en el que se buscó ajustar las variables en una escala común. (centrados en 0 con desviación estándar de 1).



Los modelos usados para la predicción fueron el Modelo Regresión Lineal Múltiple (RLM) y El Árbol de decisión de Regresión (ADR).

Antes del tratamiento de los datos el modelo de RLM arrojó lo siguientes parámetros:

Coeficientes	**-3.62976565e-05** **-1.39864267e-03** **9.50830336e-05**
Intercepto	9.332883078651136
Accuracy	6.88%
MSE	202.98

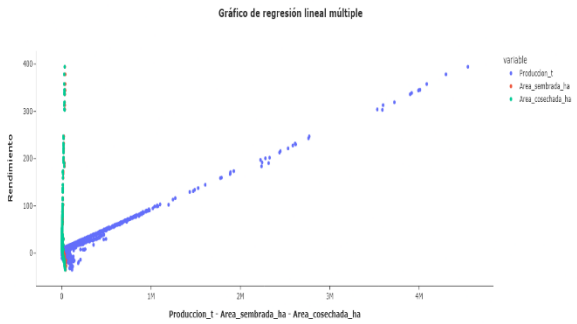


Ilustración 8 RLM de predicción

Dos variables con relación inversa y una directa e intercepto positivo. La precisión del modelo fue baja, siete por ciento (7%), y su Error Cuadrático Medio bastante alto, MSE de 203.

Luego del tratamiento de los datos, el modelo de RLM presentó valores muy similares para los conjuntos de datos Train y Test, así:

Coeficientes	**0.57884899** **-3.19976552** **3.31034777**
Intercepto	6.697098572911616
Accuracy	19.03%
MSE	30.13

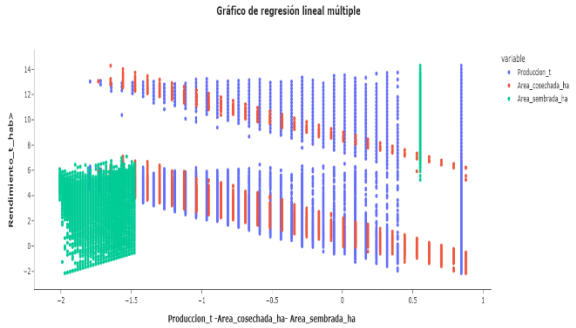


Ilustración 9 RLM de predicción luego de tratamiento de datos

Se nota, más claramente, la relación lineal entre las variables, que esta vez se relaciona una de manera inversa y dos de forma directa.

El modelo de ADR arrojó las siguientes medidas muy similares para los conjuntos de Train y Test:

```
DecisionTreeRegressor
DecisionTreeRegressor(max_depth=6, max_leaf_nodes=10, random_state=0)
```

Ilustración 10 ADR características

Accuracy	**23 %**
MSE	28

La precisión del modelo ADR mejoró con respecto al modelo de RLM en, aproximadamente, 4 puntos porcentuales. Mientras que el Error Cuadrático Medio bajó en 2 puntos porcentuales.

Conclusiones

Después de haber realizado este procedimiento de modelación podemos concluir que:

- El carácter de las variables nos brinda información verás de la realidad que representan. Es así que, con respecto al área sembrada en Colombia se nota una gran desproporción entre los cuartiles 3 y 4, donde el 75% de las áreas sembradas son menores a 151 ha y en el máximo de esta variable podemos observar registros de hasta 47403 ha. Queda claro que el 75% del área sostiene la producción Nacional agrícola con áreas menores a 151 ha.
- Con este modelo de árbol de decisión de Regresión La precisión del modelo y el error cuadrático medio mejoraron muy poco, sigue sin ser confiable, pasando de 7 a 23%. Para el conjunto datos de entrenamiento y de testeo el desempeño del modelo es muy similar. Se concluye que el conjunto de variables y su disposición actual son insuficientes para la predicción de la variable Rendimiento.
- Se recomienda usar otro tipo de modelos como lo son las redes neuronales.

Referencias

Bibliografía

Gobierno de Colombia. (24 de Noviembre de 2019). *Datos Abiertos de Colombia*. Obtenido de Datos Abiertos de Colombia: https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Evaluaciones-Agropecuarias-Municipales-EVA/2pnw-mmge

Universidad de Antioquia. (01 de Marzo de 2023). *Curso Introducción a la*

Ciencia de Datos. Obtenido de Curso Introducción a la Ciencia de Datos: https://classroom.google.com/c/NTQzMDM2NTcyNTE0