Un Proceso de Big Data para la Predicción de Rendimiento y Producción de Cultivos*

Stefano Fabi[1] and Agustina Buccella[1][0000−0002−8516−7453]

GIISCO Research Group
Departamento de Ingeniería de Sistemas - Facultad de Informática
Universidad Nacional del Comahue
Neuquen, Argentina
{stefano.fabi,agustina.buccella}@fi.uncoma.edu.ar

Abstract. En Argentina, la agricultura desarrolla un papel crucial para la economía, especialmente en provincias con alta dependencia de cultivos como maíz, girasol, trigo y soja. Este trabajo describe la aplicación de un proceso Big Data que toma un gran conjunto de datos y se enfoca en la predicción del rendimiento y la producción agrícola utilizando datos oficiales del sector junto con la temperatura y los distintos tipos de suelos de cada provincia.

Keywords: Análisis de datos · Redes neuronales · Predicción de cultivos

1 Introducción

Al manejar grandes volúmenes de datos es interesante poder identificar posibles relaciones o patrones para incrementar el valor de los mismos y hacer que sean más útiles para la toma de decisiones.

El objetivo del trabajo es presentar un caso de estudio que aplica un proceso de Big Data [1,3] con un enfoque ETL [7]. Los datos, que provienen desde diferentes fuentes como datos de siembra y cosecha, temperatura y tipos de suelos de la Argentina, se preparan e integran para la creación de modelos predictivos que permitan analizar el rendimiento y la producción de los cultivos de girasol, maíz, trigo y soja.

2 Caso de Estudio

En este trabajo aplicamos el proceso de Big Data presentado en [1] el cual consta de 4 pasos que se describen a continuación.

1. Evaluación del caso de negocio. Los objetivos definidos en este caso de estudio son: (1) predecir si el rendimiento de un cultivo superará la media anual,

* Este trabajo es presentado en el marco de la materia electiva "Almacenamiento y Análisis para Big Data" perteneciente al quinto año de la Carrera "Licenciatura en Sistemas de Información" de la Facultad de Informática de la Universidad Nacional del Comahue. Cursada el año 2024 cuya Profesora a cargo es Agustina Buccella

medida en kg/ha, lo que permitirá identificar áreas con alto o bajo rendimiento agrícola; y (2) predecir si la producción total de un cultivo en una provincia será mayor o menor que la media anual, medido en toneladas.

2. Identificación y recolección de datos. Los tres conjuntos de datos con los que trabajamos son: (1) *dataset de cultivos* que contiene 153.889 registros con la superficie de siembra y cosecha en hectáreas, lugar, rendimiento en kg/ha y producción en toneladas de campañas agrícolas para diversos cultivos y departamentos provinciales desde el año 1969 hasta 2019, provenientes de datos oficiales publicados por el Ministerio de Agricultura y Pesca [5]; (2) *dataset de clima* que proporciona mediciones diarias de la temperatura del suelo promedio en grados celcius (℃) desde 1855 hasta 2013, proporcionada por [2]; y (3) *dataset de suelos* donde utilizamos ChatGPT [6], para identificar los tipos de suelos por provincia según su color en una imagen provista por el INTA [4] generando datos de provincia y tipos de suelos.

3. Preparación de datos. Llevamos a cabo varios procedimientos para transformar los conjuntos de datos. Para realizar el trabajo utilizamos las librerías de Python ***Pandas***[1] y ***NumPy***[2]. La mayor parte del trabajo fue realizado con el *dataset de cultivos* en el cual eliminamos columnas innecesarias, tratamos las inconsistencias, rellenamos datos faltantes con diferentes técnicas, reemplazamos nulos y agrupamos el conjunto de datos por año, provincia y cultivo para ir a la misma granularidad que los otros conjuntos de datos. Para el *dataset de clima* filtramos los registros por el país "Argentina" y agrupamos por año y por promedio anual de la temperatura del suelo. En el *dataset de suelos*, no fue necesario realizar ninguna preparación previa.

Aquí también realizamos tareas de **integración** para unir los tres conjuntos de datos mediante los años y las provincias. El conjunto de datos finalmente integrado contiene 9.710 registros de cultivos con el año de inicio y fin de la campaña, la provincia, el tipo de cultivo, la superficie de siembra y cosecha, la producción, el rendimiento, la temperatura y los 8 tipos de suelos. El mismo fue almacenado en ***Apache HDFS***[3] con el nombre *dataset siembra y suelos final*.

4. Análisis de datos: El procesamiento utilizado fue por lotes y los dividimos en cuatro pasos (Figura 1):

1. **Análisis de correlación de las variables.** En base a nuestras pruebas, nos dimos cuenta que unir todos los cultivos en un único mapa de correlación no tenía sentido debido a que cada cultivo tiene sus propias características. Entonces seleccionamos sólo los cultivos más representativos (girasol, maíz, trigo y soja) y provincias clave (Buenos Aires, Córdoba, Santa Fe, Entre Ríos, La Pampa, Chaco, Santiago del Estero y San Luis) para crear mapas de correlación de las variables entre sí, como mostramos en la Figura 2.
   Se puede observar, para todos los cultivos, correlaciones fuertes del 95% entre la superficie cosechada y la producción. Sin embargo, en otros casos

[1] `https://pandas.pydata.org/docs/`
[2] `https://numpy.org/`
[3] `https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html`

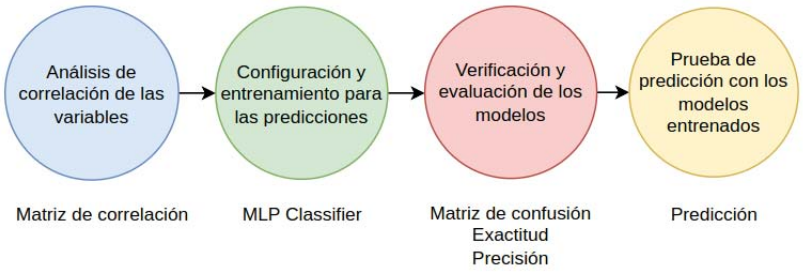

Fig. 1. Pasos principales del proceso de predicción de cultivos

como para el cultivo de girasol (Figura 2a), la correlación entre la superficie sembrada y el rendimiento es significativamente más baja, específicamente de sólo el 30%. También se observa una correlación prácticamente nula entre las distintas variables y la temperatura. Los tipos de suelos que más se destacaron positivamente son "molisoles" y "vertisoles" (Figura 2a y d), en contraste con los otros tipos donde fue nula.

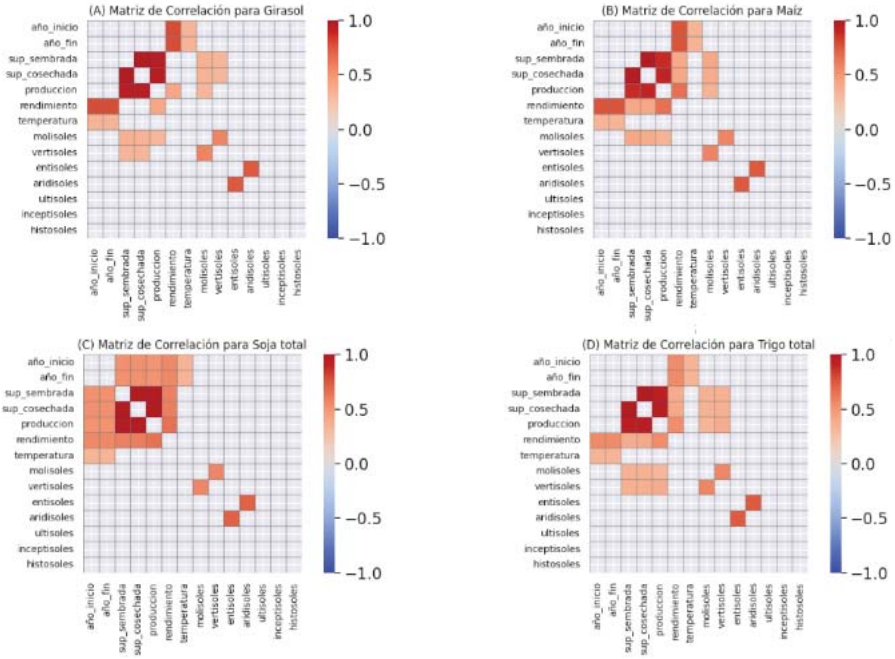

Fig. 2. Matriz de correlación para (a) girasol, (b) maíz, (c) soja y (d) trigo

2. **Configuración y entrenamiento para las predicciones.** Para generar los modelos predictivos, creamos y entrenamos una red neuronal utilizando las clases ***MLPClassifier***[4] y las métricas de rendimiento de la librería ***scikit-learn***[5]. Utilizamos como entrada el conjunto de datos *dataset siembra y suelos final* con todas las variables menos el año de inicio y fin. También quitamos los tipos de suelos quedando únicamente "molisoles" y "vertisoles" ya que como describimos previamente, fueron los de mayor correlación positiva. Para ambos objetivos realizamos cuatro configuraciones creando diferentes modelos predictivos. En cada una de ellas realizamos algunas variaciones con el fin de observar y comparar los diversos resultados obtenidos y analizar el comportamiento de los modelos. Por ejemplo, se varió el número de capas ocultas para permitir al modelo aprender representaciones más complejas, el número de iteraciones para asegurar que se disponga del tiempo suficiente para aprender, etc. Las diferentes configuraciones fueron creadas en base a varios artículos de la Web indicando mejores u optimizaciones de cada algoritmo o función de activación. Así, los modelos creados fueron:
   - Modelo 1: 1 capa con 100 neuronas y 200 iteraciones. Usa el algoritmo "adam" y la función de activación "relu".
   - Modelo 2: Igual al Modelo 1, pero con un máximo de 500 iteraciones.
   - Modelo 3: 2 capas, misma configuración del Modelo 2. La segunda capa tiene 50 neuronas.
   - Modelo 4: Igual al Modelo 3, pero usa el algoritmo "sgd" y la función de activación "tanh".
3. **Verificación y evaluación de los modelos.** Para el ***primer objetivo***, y a pesar de que la configuración más simple era la del modelo 1 y la más compleja la del modelo 4, encontramos que la mejor configuración fue la intermedia del modelo 3. La misma arrojó una exactitud y precisión del 75%. Esto subraya la importancia de evitar tanto el sobreajuste como el subajuste. Sin embargo, la matriz de confusión de este modelo (Figura 3a) reveló que todavía hay dificultades para clasificar correctamente los rendimientos agrícolas.
   Para el ***segundo objetivo*** cada modelo que fue aumentando la complejidad obtuvo un mejor resultado que el anterior, pasando de una precisión del 93% en el modelo 1 hasta llegar a una precisión del 95% en el modelo 4. La matriz de confusión del modelo 4 (Figura 3b) revela los buenos resultados obtenidos.
4. **Prueba de predicción con los modelos entrenados.** Para ambos objetivos nos realizamos la misma pregunta: ¿que pasará con el cultivo "Soja total" en la provincia de "Buenos Aires", donde se sembrarán 900.000 hectáreas y se cosechará el 100% de lo cultivado, con una temperatura de suelo promedio de 18°C?. Sin embargo, las predicciones son diferentes ya que se utilizan distintos clasificadores: para el ***primer objetivo*** el modelo 3 predijo con

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html`
[5] `https://scikit-learn.org`

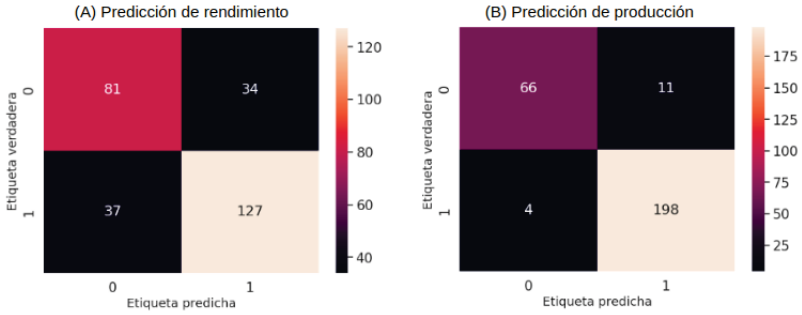

Fig. 3. (A) Matriz de confusión para el modelo 3 del primer objetivo y (B) Matriz de confusión para el modelo 4 del segundo objetivo

un 75% de exactitud que el rendimiento será mayor a la media anual, es decir, superior a 2.041 kg por hectárea; mientras que el ***segundo objetivo*** el modelo 4 predijo con un 95% de exactitud que la producción será mayor a la media anual de 1.394.642 toneladas.

3 Conclusiones y Trabajo Futuro

En este trabajo presentamos un caso de estudio orientado a la predicción del rendimiento y la producción de cultivos considerando datos obtenidos de diferentes fuentes. Se puede continuar este caso, tanto variando los modelos y sus configuraciones como agregando más información que permita mejorar las predicciones. Por último, las fuentes analizadas y los programas creados para realizar este trabajo están disponibles en GitHub[6].

References

1. Bahga, A., Madisetti, V.: Big Data Science & Analytics: A Hands-On Approach. VPT, 1st edn. (2016)
2. Earth, B.: Cambio climático: Datos de la temperatura superficial de la tierra (2017), `https://berkeleyearth.org/data/`
3. Erl, T., Khattak, W., Buhler, P.: Big Data Fundamentals: Concepts, Drivers Techniques. Prentice Hall Press, 1st edn. (2016)
4. Instituto Nacional de Tecnología Agropecuaria (INTA): Mapa de órdenes de suelos de la argentina según soil taxonomy (2014), `https://www.suelos.org.ar/sitio/mapa-de-ordenes-de-suelos-de-la-argentina-segun-soil-taxonomy-68-kb/`
5. Ministerio de Agricultura, Ganadería y Pesca de Argentina: Estimaciones agrícolas (2019), `https://datos.magyp.gob.ar/dataset/estimaciones-agricolas`
6. OpenAI: Chatgpt (versión del 15 de julio) [modelo de lenguaje de gran tamaño] (2023), `https://chat.openai.com/chat`
7. Vaisman, A., Zimnyi, E.: Data Warehouse Systems: Design and Implementation. Springer Publishing Company, Incorporated, 1st edn. (2016)

[6] https://github.com/IngSisFAI/BigDataFabi2024/