Estudio de Producción de Cultivos en la República Argentina

Facundo Hernández y Santiago Rubio

Cátedra de Ciencia de Datos – UTN FRBA – Cluster AI

Abstract

El trabajo desarrollado a continuación se realizó en base a un dataset publicado por el Ministerio de Agroindustria de la República Argentina que reúne datos sobre la producción de diversos cultivos en el país, indicando para cada muestra la provincia y el departamento en la que se realizó la producción, el nombre del cultivo, las hectáreas sembradas, las hectáreas cosechadas, las toneladas producidas y el rendimiento de la cosecha para el año medido.

El objetivo del estudio es descubrir el comportamiento de los cultivos en el país.

Durante la investigación se realizó en primera instancia un análisis exploratorio de datos para consecuentemente entrenar dos modelos, uno de clasificación basado en el algoritmo K nearest neighbors (Knn) y uno de clustering basado en el método K means.

Keywords

Agricultura, Argentina, producción, cultivos.

1 Introducción

El objetivo del estudio es descubrir el comportamiento de los cultivos en el país. Para esto se realizó un análisis exploratorio de datos para los cultivos predominantes en las distintas provincias a lo largo del tiempo mediante la visualización de la correlación que tiene la producción de los mismos.

Luego de esto, se realizaron dos modelos. El primero permite clasificar una muestra de cultivos que se ingrese, la cual el modelo asigna a una provincia. El segundo consiste en un clustering de la producción de los distintos cultivos en cada provincia, generando grupos según el rendimiento de los mismos.

Se analizaron 136051 samples con una dimensionalidad de 12 features.

2 Análisis Exploratorio de Datos (EDA)

El objetivo de este análisis es el de encontrar la naturaleza de los datos, sabiendo qué influencia tienen entre ellos mismos, y qué provincias y departamentos eran para los que más datos se tenían.

La primera acción a tomar fue revisar la presencia de NaNs en el dataset y obtener las distintas labels de las columnas. Para el primer caso, el resultado fue negativo para la presencia de NaNs, sin embargo, en cuanto a las labels de las columnas, tuvimos que cambiar una ya que se encontraba escrita con un espacio al final que resultaba incómodo a la hora de referenciarla.

La estructura de las labels del dataset quedó de la siguiente manera:

id_provincia	id_campaña
provincia	campaña
id_departamento	sup_sembrada
departamento	sup_cosechada
id_cultivo	produccion
cultivo	rendimiento

Teniendo en mente los objetivos propuestos para el EDA, el paso inicial fue graficar la frecuencia de aparición de las provincias y los departamentos en el dataset. Para la primera visualización se realizó un countplot tomando como eje x del mismo a la columna "provincia" (figura 1), mientras que para llevar a cabo la segunda (figura 2) se tuvo que recurrir a aplicar un

GroupBy al DataFrame, agrupando el id_departamento en función de la provincia y el nombre del departamento. Se recurrió a realizar esto frente al hecho de que hay departamentos de distintas provincias que tienen el mismo nombre, como por ejemplo 9 de Julio o 25 de Mayo.

Figura 1: *Cantidad de apariciones en el dataset de cada provincia*

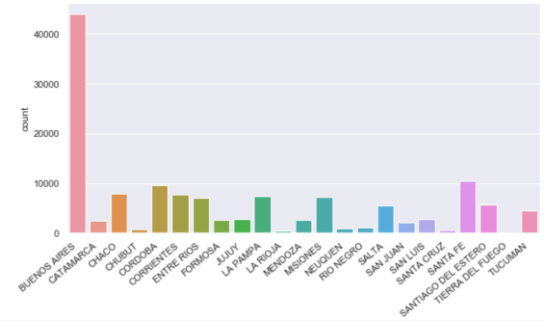
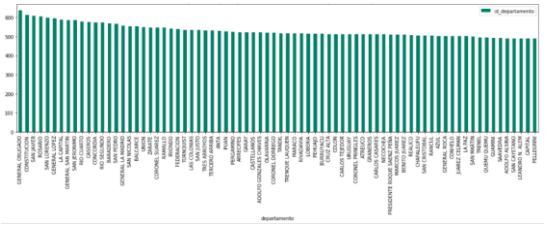


Figura 2: *Cantidad de apariciones por departamento en el datasetde los primeros 70 departamentos*



Como segunda acción, y para profundizar el estudio, se decidió analizar por provincia, la variación de los cultivos en el tiempo y la correlación que había entre los niveles de producción de estos. Para poder llevar a cabo esta investigación se tuvo que hacer una reestructuración de los datos con el fin de generar un DataFrame por provincia con los datos de producción de los distintos cultivos, lo cual fue realizado de la siguiente manera:

Se generaron dos diccionarios, el primer contaba con una entrada por cada provincia del país y dentro de esta entrada se generó un DataFrame para cada cultivo presente en el dataset, sea este cultivado en la provincia pertinente o no. Paso seguido se colapsó, por provincia, a todos los DataFrames existentes en la entrada de esta, generando un solo cuadro de datos que tenía como índice a los años distintos años de medición y como columnas a la producción de cada cultivo para esa provincia en cada año.

Al llevar a cabo esta operación, se generó un problema que era la existencia de NaNs en cada uno de estos DataFrames como resultado de que no todas las provincias tienen todos los cultivos ni hay datos de los cultivos para todos los años en todas las provincias. Para solucionarlo, se procedió a eliminar todas las plantaciones que tuvieran NaNs, ya que estos estaban originados en el no cultivo de estas en la provincia o en la no medición del nivel de producción de estos durante todos los años abarcados por el DataFrame.

Es a partir de esto que se decidió que el análisis de correlación en la producción de cultivos de una provincia y la distribución del nivel de producción de los mismos se iba a realizar sólo para aquellos que fueran los principalmente producidos en esa región, debido a que era para los cuales contábamos con datos completos.

En las figuras 3, 4 y 5 se pueden visualizar un pairplot, un heatmap de pairwise correlation y un boxplot, respectivamente, para las plantaciones propias de la provincia de Corrientes.

Figura 3: *Pairplot de cultivos cosechados en la provincia de Corrientes.*

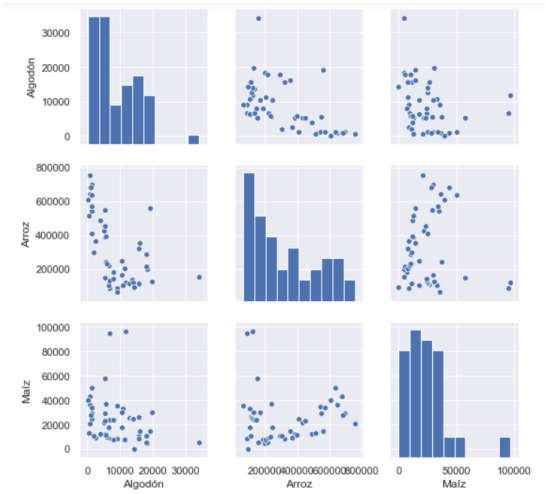


Figura 4: *Heatmap de pairwise corerlation de los cultivos cosechados en la provincia de Corrientes.*

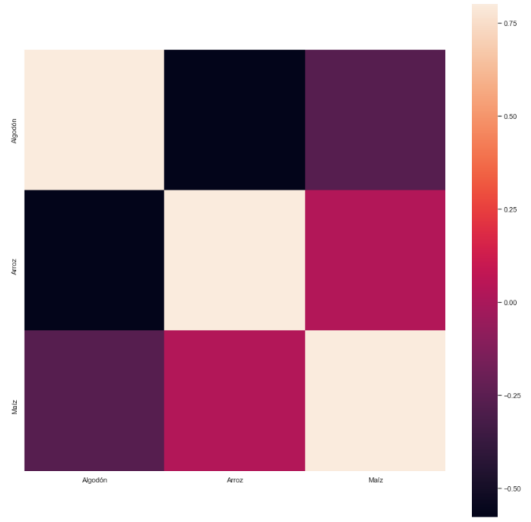
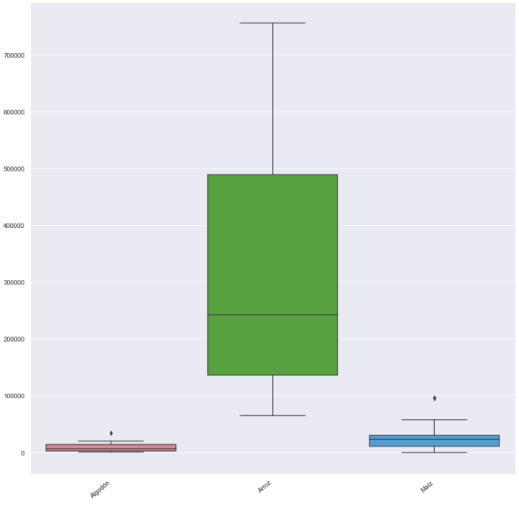


Figura 5: *Boxplot de las toneladas producidas de los cultivos cosechados en la provincia de Corrientes*



En este ejemplo, se puede entender gráficamente que, por ejemplo, el cultivo más producido en corrientes es el arroz y que hay una correlación casi nula entre este cultivo y la producción de maíz en la provincia, probablemente a razón de que las áreas de sembrado son distintas.

Para facilitar la visualización de estas relaciones para todas las provincias, se generó una función con nombre functions.analisis, la cual muestra en pantalla estos gráficos para la provincia que se quiera investigar, mediante la indicación del nombre de la misma como parámetro de la función.

3 Modelización

Luego de haber ejecutado el EDA se procedió al entrenamiento de dos modelos con los datos de trabajo. El primer modelo es de clasificación y aplica el algoritmo K Nearest Neighbors para realizar el etiquetado de las distintas muestras, mientras que para la segunda modelización se realizó un clustering basado en el método K means, con el objetivo de encontrar 5 categorías distintas de rendimiento de los cultivos en función de las provincias en las que eran sembrados.

3.1 Modelo de clasificación con KNN

3.1.1 Estructuración de los datos

Como paso inicial para este objetivo, se realizó un proceso similar al que se realizó durante el EDA. Se generó un diccionario que tiene una entrada por cada cultivo y cuyas columnas son la producción de ese cultivo por provincia según la campaña (año). Esto vuelve a generar una tabla con NaNs debido a que no todos los cultivos se producen en todas las provincias y a que hay cultivos que no tienen datos para todos los años que abarca el dataset.

Luego de haber generado el diccionario, se lo transformó en un data set cuyas column labels son los id de cultivos y se volvió a iterar con ayuda del DataFrame "cultivos", para renombrar todas las columnas, reemplazando el id de cultivo con el nombre del cultivo con el que se corresponde.

Acto seguido se procedió a procesar los NaNs presentes pero con una diferencia respecto a lo efectuado en el EDA. En este caso se optó por reemplazar a todos los nulls por cero debido a que todas las muestras y todas las columnas los tenían (debido a las razones explicadas anteriormente), por lo que eliminarlos no era una opción y reemplazarlos por el promedio tampoco, ya que al ser cada cultivo una columna, de esta manera se hubiera llenado a distintas provincias con un mismo promedio de producción, lo cual hubiera sido erróneo.

Si bien sabemos que esto produjo un poco de ruido en el modelo, concluimos que era la mejor opción que se podía tomar.

Como última acción de este paso del proceso, se separó al DataFrame en etiquetas y

variables dependientes y se aplicó un LabelEncoder a las primeras ya que eran objetos del tipo string y se necesitaba que los mismos fueran números para la realización del modelo.

3.1.2 Feature Engineering

El primer paso en esta etapa fue separar a nuestros datos en un set de train y otro de test (70/30) para luego escalar nuestras variables mediante un StandardScaler que fue ajustado a nuestras features del train set y que luego fue aplicado a estas y a las variables del test set.

3.1.3 Armado del modelo y resultados

Habiendo realizado todos los pasos previos necesarios para poder generar nuestro clasificador, se continuó con la definición del mismo. Con este fin, se llevó a cabo un Grid Search Cross Validation para encontrar los parámetros que obtuvieran los mejores resultados. Los valores de los parámetros que se pasaron para iterar fueron:

- n_neighbors: [1, 2, 3, 4, 5, 6, 7]
- weights: [uniform, distance]
- p: [1, 2]

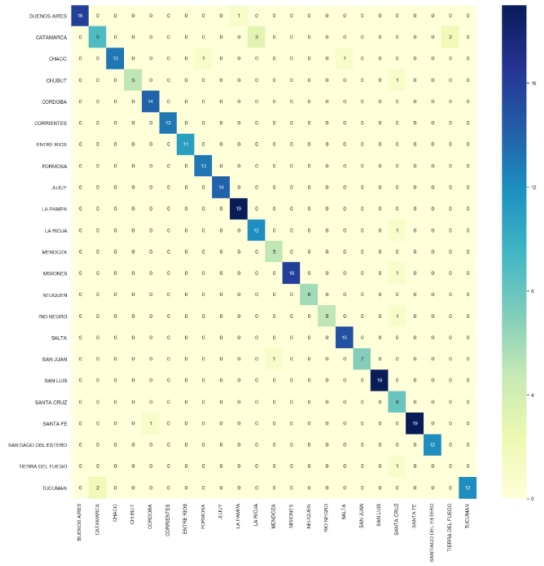
De los que obtuvimos como valores a usar:

- n_neighbors = 1
- weights = uniform
- p = 1

Con esta combinación de parámetros, se llegó a un modelo que clasificó con 96% de accuracy el train set y que predijo con un accuracy del 94% el test set, siendo esto un resultado que encontramos como muy satisfactorio.

Para poder visualizar la predicción que realizó nuestro modelo del test set, realizamos una Confusion Matrix y la graficamos mediante un heatmap. A la hora de realizar la misma, realizamos un inverse_transform de nuestras etiquetas para mostrar el nombre de las provincias en vez del valor numérico que se había asignado a cada una.

Figura 6: *Heatmap de la matriz de confusión generada a partir de la predicción de nuestro modelo respecto a las muestras del test.*



Algo que resalta a simple vista es que ninguna muestra se clasificó como perteneciente a la provincia de Tierra del Fuego, esto es así ya que de nuestro data set original, de 136000 samples, sólo nueve pertenecían a esta provincia.

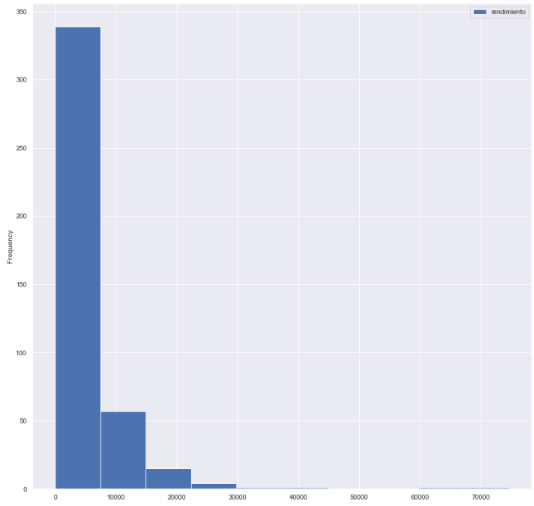
Por otro lado, también se puede notar que la mayoría de errores de clasificación se dan con provincias que se encuentran juntas geográficamente salvo por dos muestras que fueron clasificadas como pertenecientes a la provincia de Santa Cruz cuando en realidad pertenecían a las provincias de Misiones y La Rioja.

3.2 Modelo de clusterización con K means

3.2.1 Estructuración de los datos

Antes de comenzar con la realización del modelo se agrupó en un nuevo DataFrame al rendimiento de los cultivos, haciendo un promedio del mismo por cultivo y por provincia y acto seguido se buscó visualizar con un histograma la cantidad de samples que compartían un mismo rango de rendimiento.

Figura 7: *Distribución del rendimiento de las muestras en kilogramos por hectárea cosechada.*



Se puede notar que la amplia mayoría de las muestras (más de 300 sobre las 419 analizadas) tienen un rendimiento menor a 10 toneladas por hectárea cosechada.

3.2.2 Feature Engineering

Si revisamos mediante un df.head() la estructura de nuestro DataFrame generado, la misma es:

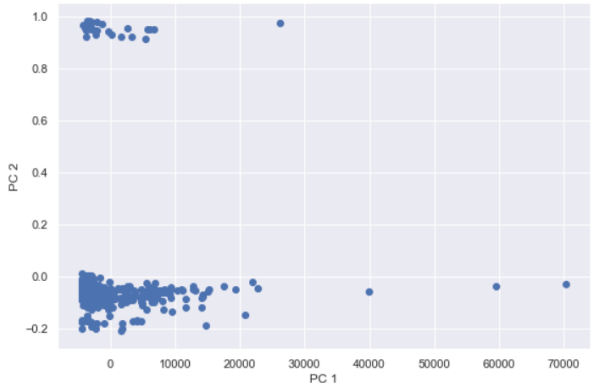
	provincia	cultivo	rendimiento
0	BUENOS AIRES	Ajo	3978.088748
1	BUENOS AIRES	Alpiste	954.178904
2	BUENOS AIRES	Avena	1504.611014
3	BUENOS AIRES	Cebada cervecera	2414.116224
4	BUENOS AIRES	Cebada forrajera	625.169038

Es entonces que resultó necesaria la generación de dos grupos de dummies distintos, uno para la columna provincia y otro para la columna de cultivo, con el fin de transformar los objetos de tipo string en objetos numéricos, quedando así un DataFrame de trabajo de 419 filas por 57 columnas.

Una aclaración pertinente en este análisis es que se decidió no escalar a los valores del rendimiento ya se trata de la variable principal y al escalarla iba a perder importancia en comparación con el valor de los dummies.

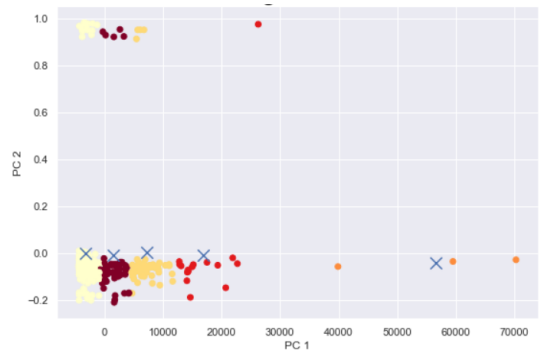
Como acto seguido, se recurrió a la ejecución de un PCA para poder graficar las muestras en dos dimensiones.

Figura 8: *Distribución del rendimiento de las muestras en kilogramos por hectárea cosechada.*



3.3.3Armado del modelo y resultados

Como tercer etapa de la modelización se definió el modelo como un clusterizador por K means con 5 clusters, para el cual se obtuvo un Índice de Silhoouette de 0,697. El gráfico de los clusters y el centroide de los mismos está a continuación:



Para poder obtener conclusiones de esto, se unió la lista de labels que generó el K means al DataFrame original que fue generado al inicio de este proceso de modelización, obteniendo la siguiente estructura:

	provincia	cultivo	rendimiento	Cluster
0	BUENOS AIRES	Ajo	3978.088748	4
1	BUENOS AIRES	Alpiste	954.178904	0
2	BUENOS AIRES	Avena	1504.611014	0
3	BUENOS AIRES	Cebada cervecera	2414.116224	0
4	BUENOS AIRES	Cebada forrajera	625.169038	0

Con la cual se generó una función llamada functions.clusters que lista todas las combinaciones de cultivo/provincia que entraron en los distintos clusters según su rendimiento. Gracias a esto se puedo determinar qué intervalo productivo definía cada cluster:

- Rendimiento alto = Cluster 3
- Rendimiento medio alto = Cluster 1
- Rendimiento medio = Cluster 2
- Rendimiento medio bajo = Cluster 4
- Rendimiento bajo = Cluster 0

Siendo los valores límite de rendimiento de cada cluster:

- Rendimiento alto: Más de 44229 Kg/ha
- Rendimiento medio alto: Entre 30606 Kg/ha y 17125 Kg/ha
- Rendimiento medio = Entre 15989 Kg/ha y 8978 Kg/ha
- Rendimiento medio bajo: Entre 8503 Kg/ha y 3537 Kg/ha
- Rendimiento bajo: Menos de 3436 Kg/ha

Como resultado de esta modelización se puede definir que el cultivo con mayor rendimiento es la caña de azúcar, la cual tiene mayores rendimientos en las provincias de Jujuy, Salta y Tucumán (Rendimiento alto) respecto de las provincias de Santa Fé, Formosa, Corrientes y Chaco (Rendimiento medio alto).

4 Conclusiones y observaciones

Tanto de la primera etapa del trabajo (EDA) como de la segunda (Modelización) se han podido obtener definiciones relevantes acerca del comportamiento de la siembra y la cosecha de los diferentes cultivos producidos en el país

Con el trabajo explorativo de datos realizado se han conseguido observaciones gráficas de la existencia de correlaciones entre los distintos cultivos que se producen en una misma provincia mientras que con la modelización se encontró información valiosa respecto a que cultivos presentan mejor rendimiento en cada locación geográfica.

En cuanto al data set en sí, el mismo resulta muy interesante debido a la información que presenta, sin embargo, está incompleto, por lo que el análisis realizado sobre el mismo es de la misma naturaleza y podría ser mucho más fructuoso. Esperamos que muestras nuevas sigan siendo añadidas de manera constante y completa a esta base, debido al potencial investigativo que presenta la misma.

Por otro lado, si bien no fue algo que realizamos, sería interesante poder relacionar este set de datos con otro que presente información sobre hechos económicos o climáticos que pudieran explicar la variación en el tiempo y en las distintas provincias del nivel de producción de los cultivos y el rendimiento de los mismos.

5 Referencias

[1] Trevor, Hastie, Tibshirani Robert, y Friedman JH. "The elements of statistical learning: data mining, inference, and prediction." pág. 459-474. (2009).

[2] Jacob T. VanderPlas. "Python Data Science Handbook: Essential Tools for Working with Data." (2016)

[3] Pandas: powerful Python data analysis toolkit. "User Guide" URL https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html

[4] Documentation of scikit-learn 0.21.3. "User Guide" URL https://scikit-learn.org/stable/user_guide.html

[5] Seaborn: statistical data visualization. "Official seaborn tutorial" URL https://seaborn.pydata.org/tutorial.html