Predicción de rinde de cultivos basado en condiciones y prácticas agropecuarias

El siguiente trabajo está dirigido a productores agropecuarios de cualquier escala. Particularmente a aquellos que se dediquen a la siembra de los siguientes cultivos:

* Batata
* Mandioca
* Trigo
* Cebada
* Banana
* Oleaginosas

Se ofrecerá un modelo que permita estimar el rinde de cierto cultivo en función de la caída de lluvia anual y el uso de pesticidas y fertilizantes. El rinde se tomará como razón de unidad de producción por área sembrada. La hipótesis que guiará el análisis es que esas tres variables, que de aquí en adelante se llamarán “variables de interés”, aumentan el rinde de forma directamente proporcional.

## Dataset utilizado

Se usó el dataset “Agricultural Crop Yield in Indian States Dataset”, publicado por Akshat Gupta et al.[[1]](#footnote-2). El mismo contiene datos de producción de más de cincuenta cultivos en la India entre los años 1997 y 2020. La información está agregada por año y por estado y describe el área sembrada, la producción en toneladas métricas, la caída de lluvia anual, el uso de pesticidas en kilos, el uso de fertilizantes, también en kilos; y el rinde como razón entre la producción y el área sembrada.

Si bien esta información está tomada de la India, la idea es aplicarlo a la Argentina. Por ello se seleccionaron algunos cultivos que crecen en el país

## Objetivo

Los datos se procesarán en python, principalmente utilizando la librería pandas. A su vez, se implementará un modelo de regresión lineal con scikit-learn. El objetivo será generar un modelo para cada tipo de cultivo seleccionado que pueda recibir los parámetros de entrada y devolver el rinde teórico.

## Análisis de datos exploratorio

El dataset cuenta con 19689 registros y no cuenta con valores nulos ni repetidos. Cada uno tiene información sobre la producción de cierto cultivo en cierto estado de la India para un determinado año. También detalla la producción total, el área de siembra, el rinde, que es la razón entre estas últimas; la cantidad de fertilizantes y pesticidas utilizados y, por último, la caída de lluvia en ese año.

Una vez validada la consistencia de los datos, sobrevino el desafío de encontrar alguna correlación entre las variables de análisis y el rinde en los cultivos de interés. Para ello se utilizó la función correlacion\_alta que retorna los cultivos para los que al menos una de las variables de interés tiene una correlación mayor a +/- 0.5 reposando en el método DataFrame.corr. Así se obtuvieron dichos cultivos y de ellos se hizo una selección de los que crecen en Argentina.

Por último hubo que determinar el nivel de agregación. Como hay una entrada por trinomio cultivo-año-estado, podía hacerse de tres formas:

* Agrupada por año y sumando las variables de interés por estado
* Agrupada por estado y sumando las variables de interés por año
* Desagregada sin agrupar

Las formas agrupadas reducen sensiblemente el tamaño de la muestra mientras que la desagregada no. Se concluyó que mantener el nivel original del dataset era la mejor aproximación ya no distorsiona el peso de cada variable de interés. Se generó un archivo .csv para cada cultivo.

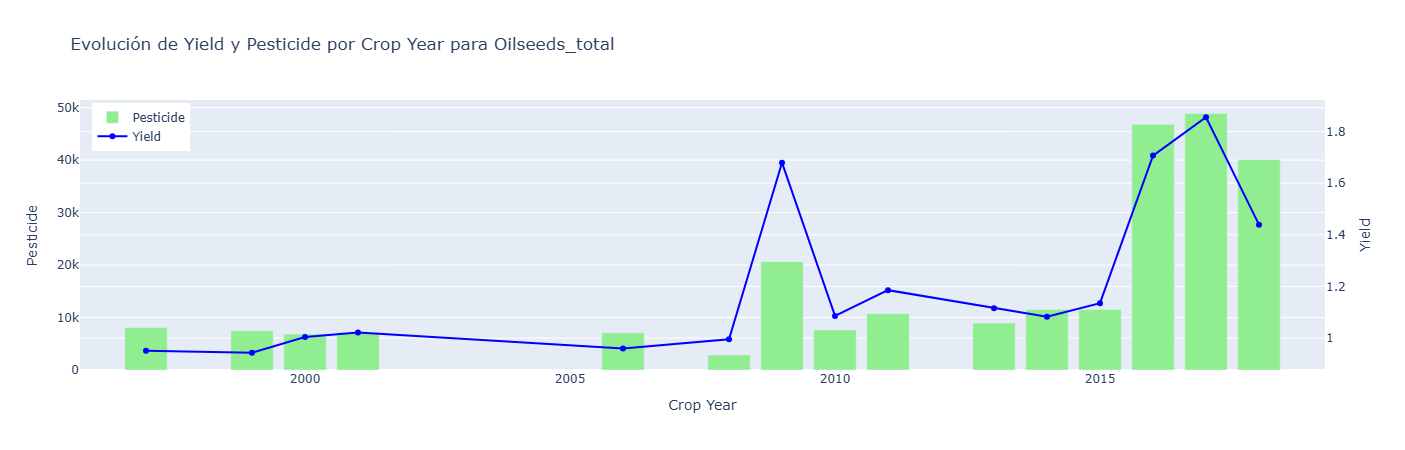
Una vez obtenidos los datasets de trabajo, se prosiguió a analizar las relaciones entre las variables de estudio con el rinde. Para eso se hizo un gráfico, esta vez sí agrupado por año para sintetizar, que relaciona una de las variables de interés con el rinde. A partir de la función rinde\_por\_columna\_interactivo se toma un dataset de trabajo, una variable de interés, y el nombre del cultivo. Así se determinó que las relaciones son mayormente lineales en todos los casos. Sin embargo cabe destacar que para el arroz y la cebada se observa menos dispersión mientras que se observan con alta dispersión para la batata, la mandioca y las oleaginosas.

Se adjuntan algunos de gráficos para el arroz y las oleaginosas a modo de ejemplo. Los gráficos interactivos están disponibles en trabajoFinal/cropYield.ipynb

### Arroz



### Oleaginosas



## Modelado

Una vez obtenidos los datasets se implementa un pipeline que utiliza StandardScaler para escalar los valores de las variables de interés y alimentar los modelos. Se va a utilizar GridSearchCV y Pipeline para seleccionar los mejores hiperparámetros para cada modelo y se van a comparar entre modelos a través de la métrica raíz del error mínimo cuadrado (RMSE).

Los modelos para ajustar van a ser regresión lineal Ridge, random forest, y red neuronal multicapa. Para cada cultivo se va a correr el pipeline y se elegirá el mejor modelo basado en RMSE con sus mejores hiperparámetros. Los resultados son

|  |  |  |  |
| --- | --- | --- | --- |
| **Cultivo** | **Ridge** | **RandomForest** | **MLP** |
| Cebada | 0.76 | 0.64 | 0.63 |
| Oleaginosas | 0.05 | 0.06 | 0.13 |
| Arroz | 0.66 | 0.58 | 0.63 |
| Batata | 1.40 | 0.68 | 1.22 |
| Mandioca | 0.37 | 0.16 | 0.69 |

El desarrollo del código está disponible en trabajoFinal/modeling.ipynb

## Conclusi**ón**

Fue posible crear modelos con un nivel de error aceptable en una primera instancia y lograr predicciones específicas para cada cultivo. Contraintuitivamente, los cultivos que tenían distribuciones más irregulares, como las oleaginosas o la mandioca, tuvieron menos error en la predicción. Para el futuro podría probarse aplicar un análisis no solo de rinde sino de rinde en el tiempo y/o por unidad geográfica. Este sería un buen desafío si tuviéramos información de magnitud más pequeña. Idealmente hasta llegar a la parcela específica.

También podría extenderse este análisis a otros cultivos y automatizar el procesamiento para actualizar los modelos cada cierto tiempo

1. Disponible en <https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset> [↑](#footnote-ref-2)