UADE
UNA GRAN UNIVERSIDAD

# *Modelo de Estimación de Producción de Soja*

*Buzarquis, Maia - 1142683*
*Calp, Diego - 1143421*
*Marzocca, Tomas Raul - 1088865*
*Tejada Borges, Alejandro - 1132294*

# *Agenda*

- Introducción
- Descripción del dataset
- Limpieza y Preparación de Datos
- Modelos y Performance
- PCA
- Conclusiones

# *Producción de Soja*

- El **ciclo** de crecimiento de soja es de 5/7 meses, se produce en el verano, desde septiembre hasta junio del próximo año.
- **Producción** de la última campaña (2020/2021): 43,5 millones de t
- **Superficie** sembrada 2020/2021:17 millones de ha
- **Uso** de la soja: aceite de soja, alimentación animal, biodiesel.

# *Dataset*

- Registro de **productores agropecuarios** para el cultivo de soja desde 2004-5 hasta 2020-1 con 259 variables y más de 120.000 registros.
- Luego de la limpieza y preparación de datos nos quedamos con 43 variables.
- Variable Target: producción

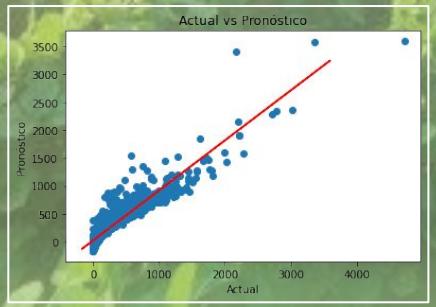
# *Machine Learning Pipeline*

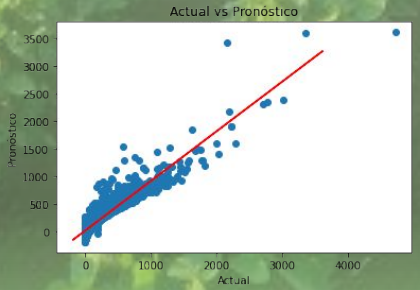
# *Limpieza y Preparación de Datos*

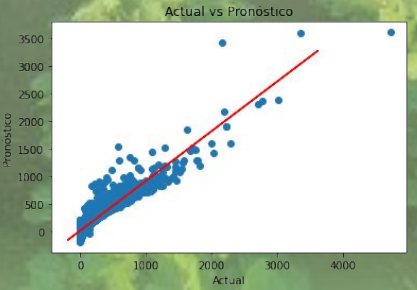
- Variables categóricas a numéricas
- Fechas a períodos en días
- Tratamiento de Nulos
- Transformaciones
- Variable Fenómeno Enos (Niño o Niña)
- Escalado de Variables

# *Modelo: Regresión Lineal Múltiple*

| | Nulos = 0 | Nulos = Media | Nulos = Mediana |
|---|---|---|---|
| $R^2$ | 0,898 | 0,901 | 0,903 |
| MAE | 35,92 | 33,95 | 33,92 |
| MSE | 3901,38 | 3730,99 | 3676,14 |
| RMSE | 62,46 | 61,08 | 60,63 |

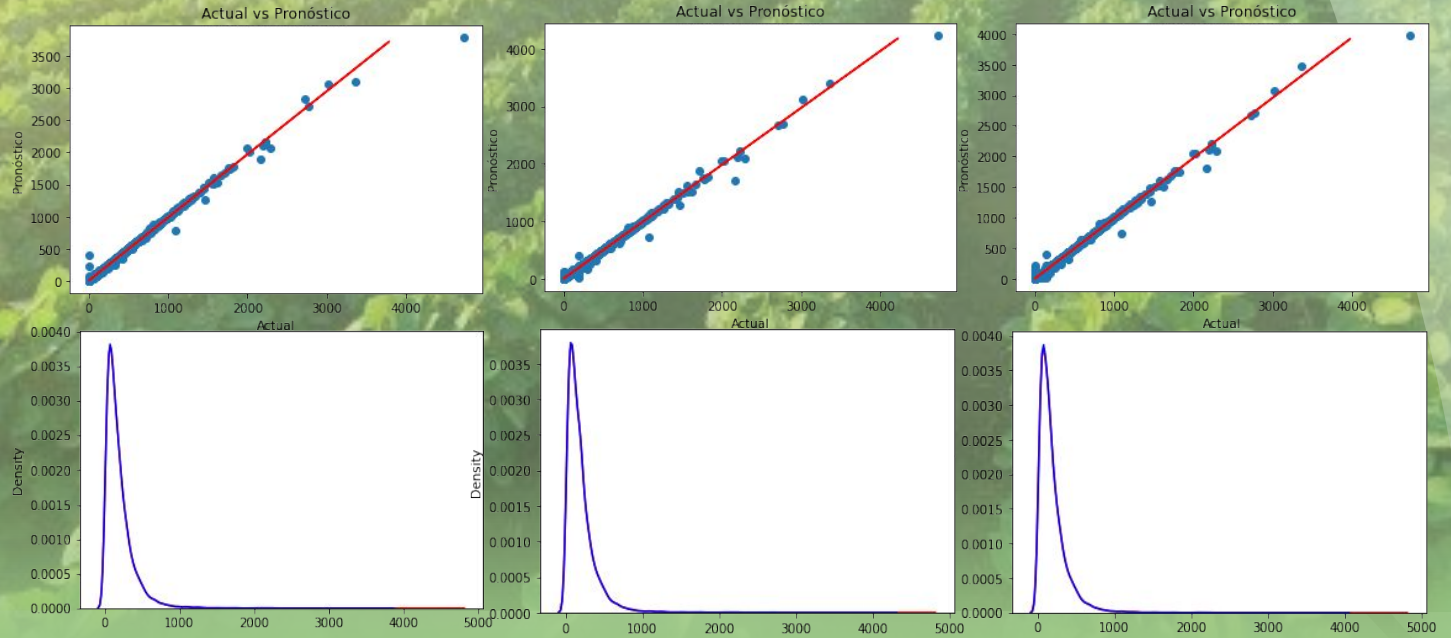


# *Modelo: Random Forest*

| | Nulos = 0 | Nulos = Media | Nulos = Mediana |
|---|---|---|---|
| $R^2$ | 0,99 | 0,99 | 0,99 |



# *Validaciones*

| Cross Validation | Nulos = 0 | Nulos = media | Nulos = mediana |
|---|---|---|---|
| Regresión Múltiple | 0,885 | 0,886 | 0,892 |
| Random Forest | 0,983 | 0,983 | 0,981 |
| Ridge Regression | 0,901 | 0,905 | 0,906 |
| Árbol de Decisión | 0,963 | 0,969 | 0,969 |

# *PCA - Principal Component Analysis*

- PCA(0.95) = 36 componentes
- Nulos = mediana
- Standard Scaler

| | Regresión | Random Forest |
|---|---|---|
| $R^2$ | 0,86 | 0,98 |

| Cross Validation | Nulos = mediana |
|---|---|
| Regresión Múltiple | 0,892 |
| Random Forest | 0,981 |
| Ridge Regression | 0,906 |
| Árbol de Decisión | 0,969 |

# *Conclusiones*

- La Regresión Lineal Múltiple tiene buenos ajustes pero errores altos.
- Esto se repitió con otros modelos y en las validaciones cruzadas.
- Aún eliminando variables relacionadas (por ej. adversidades del cultivo), se mantuvo el nivel de ajuste y errores.
- Ocurrió lo mismo con selección de variables por PCA.

# *Próximos Pasos*

- Métodos de reemplazo de nulos más elaborados: MICE o KNN.
- Hacer un análisis exhaustivo de Multicolinealidad.
- Evaluar nuevos métodos de selección de las variables independientes.
- Simulación seleccionando para test un rango del negocio, por ej. una campaña.
- Otros modelos de pronóstico.
- Probar métodos de Auto ML.

UADE
UNA GRAN UNIVERSIDAD

# *Muchas gracias!*

*Buzarquis, Maia - 1142683*
*Calp, Diego - 1143421*
*Marzocca, Tomas Raul - 1088865*
*Tejada Borges, Alejandro - 1132294*

# *Conversión de fechas a días*