

21 de julio 2025

# COSECHANDO DATOS

Estimaciones Agrícolas en Argentina

## Integrantes Grupo 15:

Andrea Rodriguez  
Belen Massuco

# Índice de Contenidos

**01** Introducción

**02** Objetivos

**03** Metodología de análisis

**04** Análisis exploratorio

**05** Modelo de aprendizaje supervisado

**06** Modelo de aprendizaje no-supervisado

**07** Resultados

**08** Conclusiones

# INTRODUCCIÓN

Seleccionamos un dataset de estadísticas agrícolas históricas de Argentina, proveniente del portal oficial del *Ministerio de Agricultura*.

Incluye registros desde el ciclo agrícola 1969/1970 hasta 2022/2023, con información organizada por provincia y departamento para **42 cultivos**.

Las variables principales son:

**provincia\_nombre**: Nombre de la provincia

**departamento\_nombre**: Nombre del departamento

**cultivo**: Nombre del cultivo

**ciclo**: Año de la campaña

**sup\_sembrada**: Cantidad de superficie sembrada en hectareas

**sup\_cosechada**: Cantidad de superficie cosechada en hectareas

**produccion**: Cantidad de produccion en toneladas

**rendimiento**: Cantidad de rendimiento en kilos por hectarea

# OBJETIVO DEL PROYECTO

Analizar la evolución de la producción agrícola en distintas regiones del país a lo largo del tiempo. Nos enfocamos en identificar tendencias en los rendimientos, variaciones en la superficie sembrada y cosechada, y diferencias regionales por cultivo.

Como objetivo específico nos proponemos estimar la producción de los principales cultivos del país.

# METODOLOGIA DE ANALISIS

## ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Comprendemos la estructura de los datos, identificando relaciones y anomalías.

## MODELADO SUPERVISADO Y NO SUPERVISADO

Aplicamos algoritmos avanzados para descubrir patrones y hacer predicciones.

## OPTIMIZACIÓN DE HIPERPARÁMETROS

Afinamos los modelos para maximizar la precisión y robustez de las predicciones.

# EXPLORACION DE DATOS

Comenzamos explorando la estructura general del dataset: contiene más de 150.000 registros y 12 features originales con información sobre cultivos, superficie y producción a nivel departamental.

## Primero

Identificamos que algunas columnas clave estaban almacenadas como texto y contenían valores no numéricos como "SD". Se convirtieron a valores float.

Los registros con valores SD se eliminaron, ya que los intentos de imputación generaron valores atípicos e inconsistentes.

## Segundo

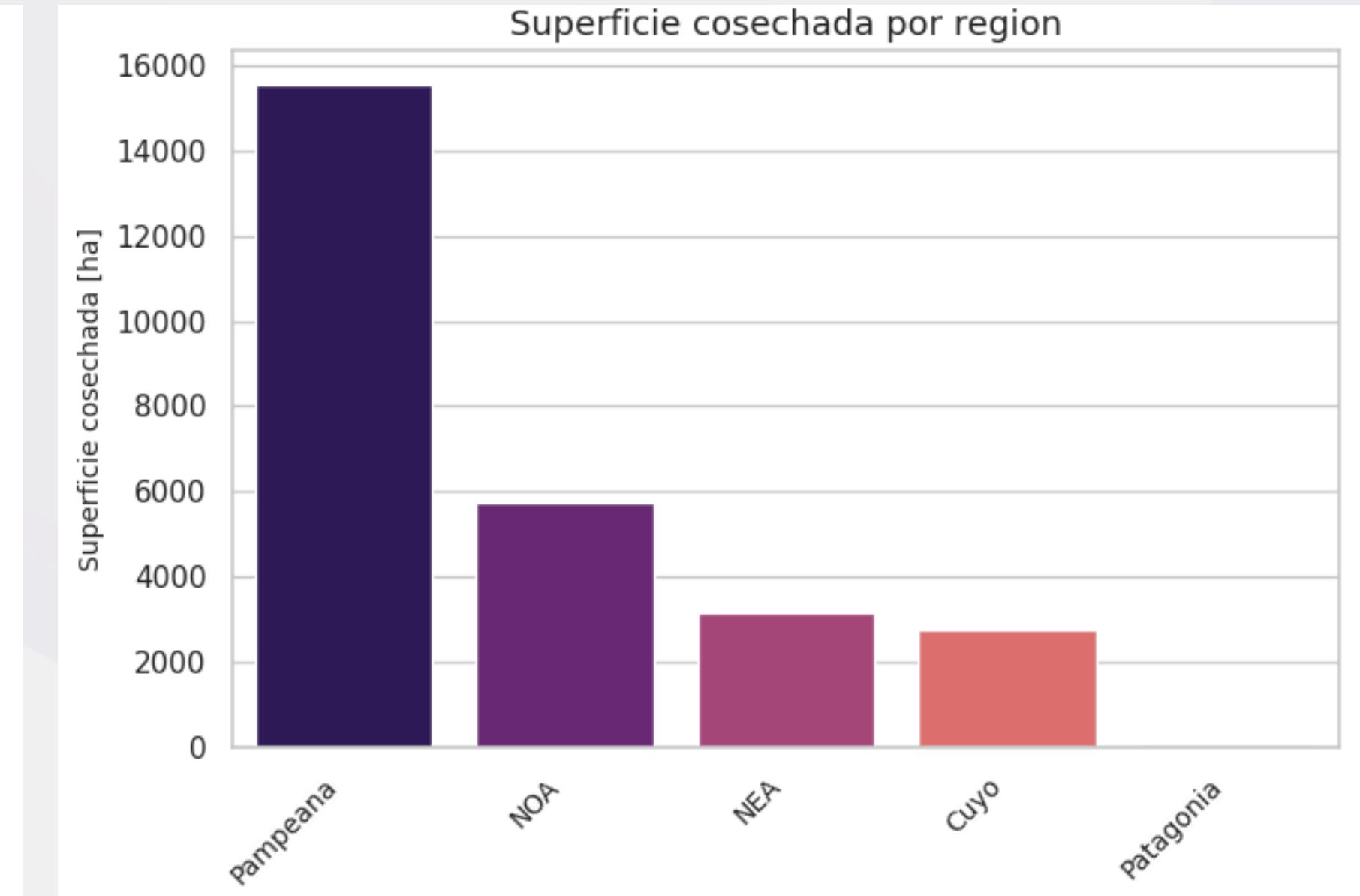
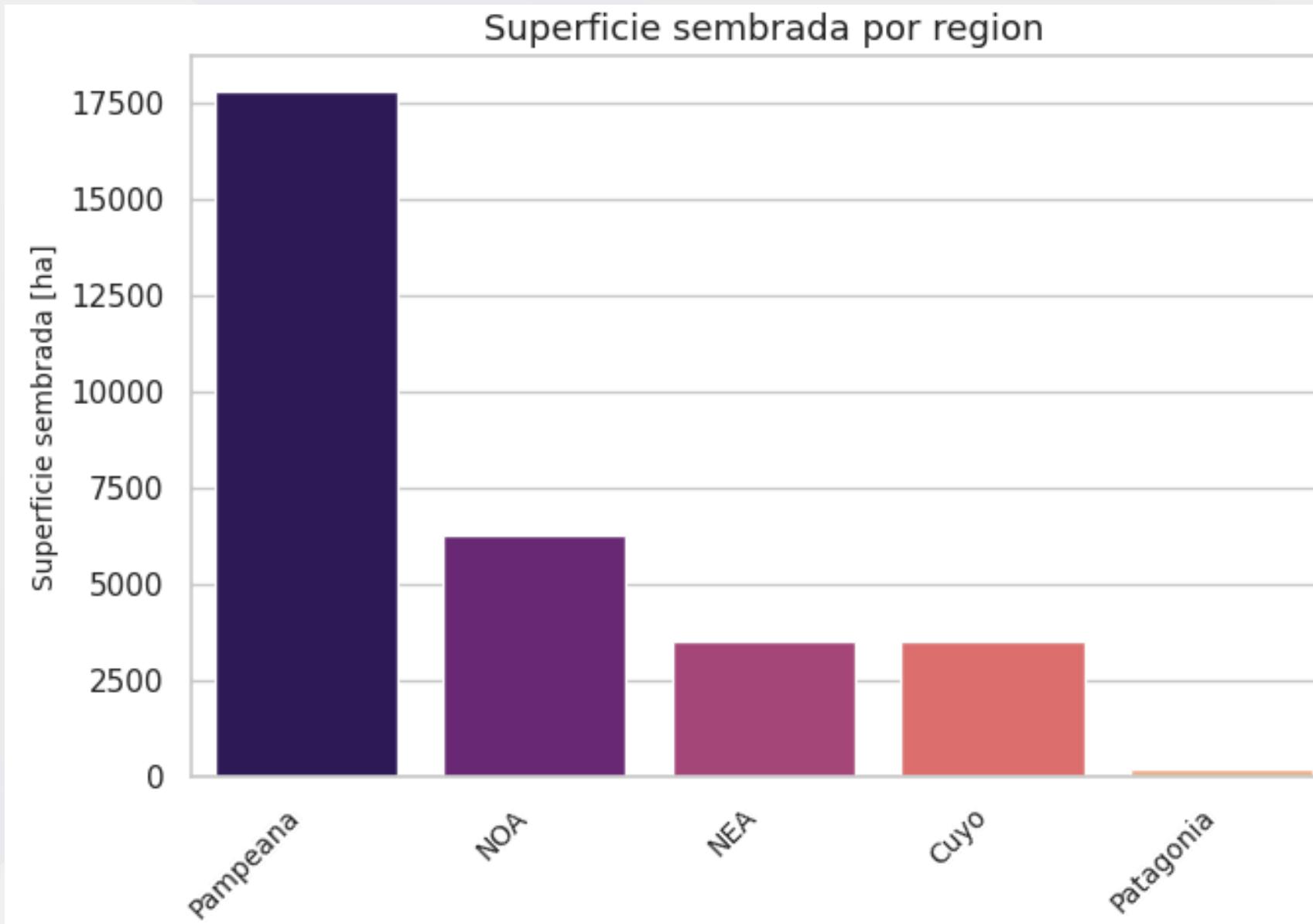
Realizamos limpieza de datos faltantes, renombramos features y creamos nuevas:

- Usamos un diccionario para agrupar provincias por región.
- Tasa de cosecha

## Tercero

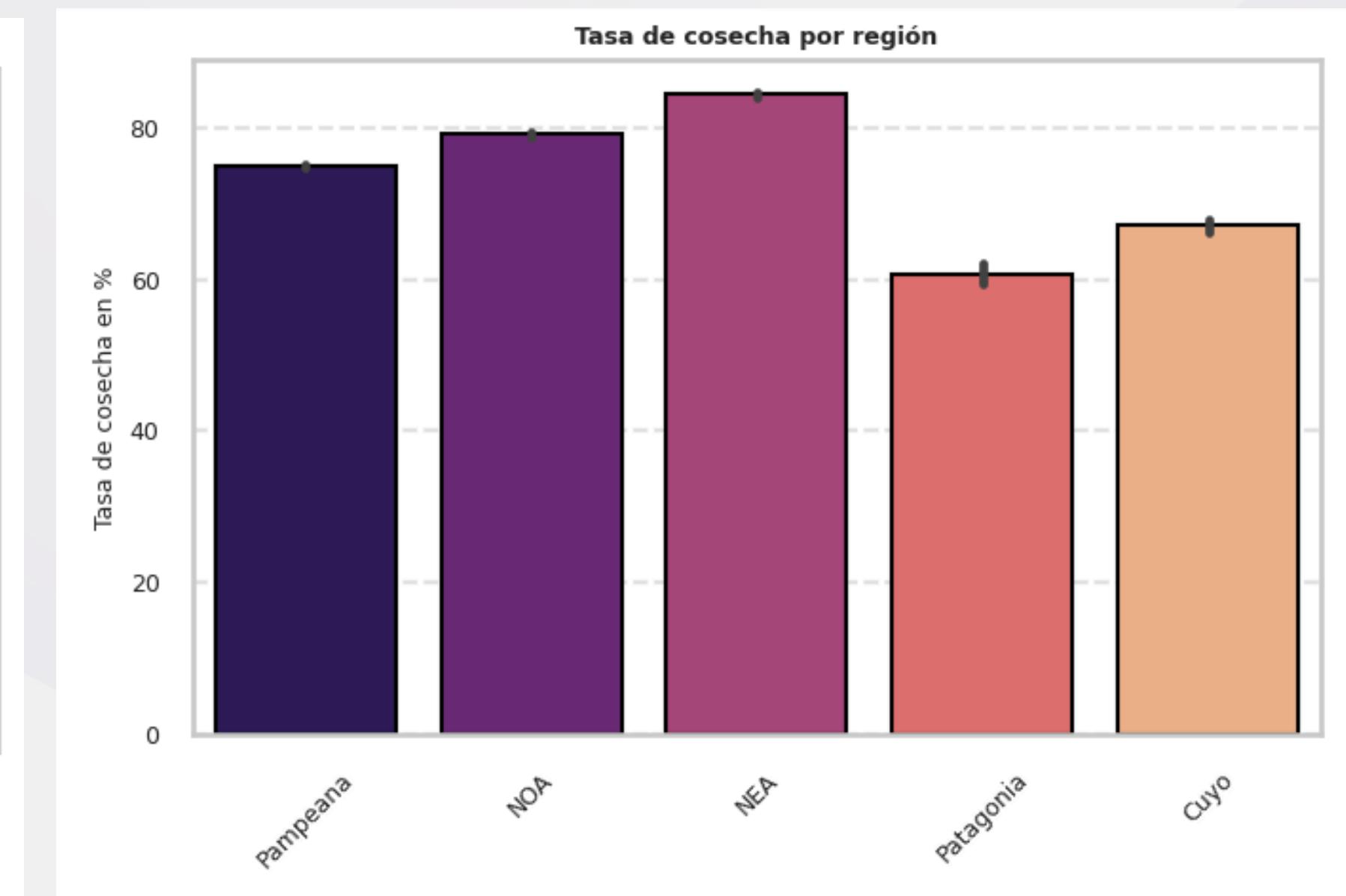
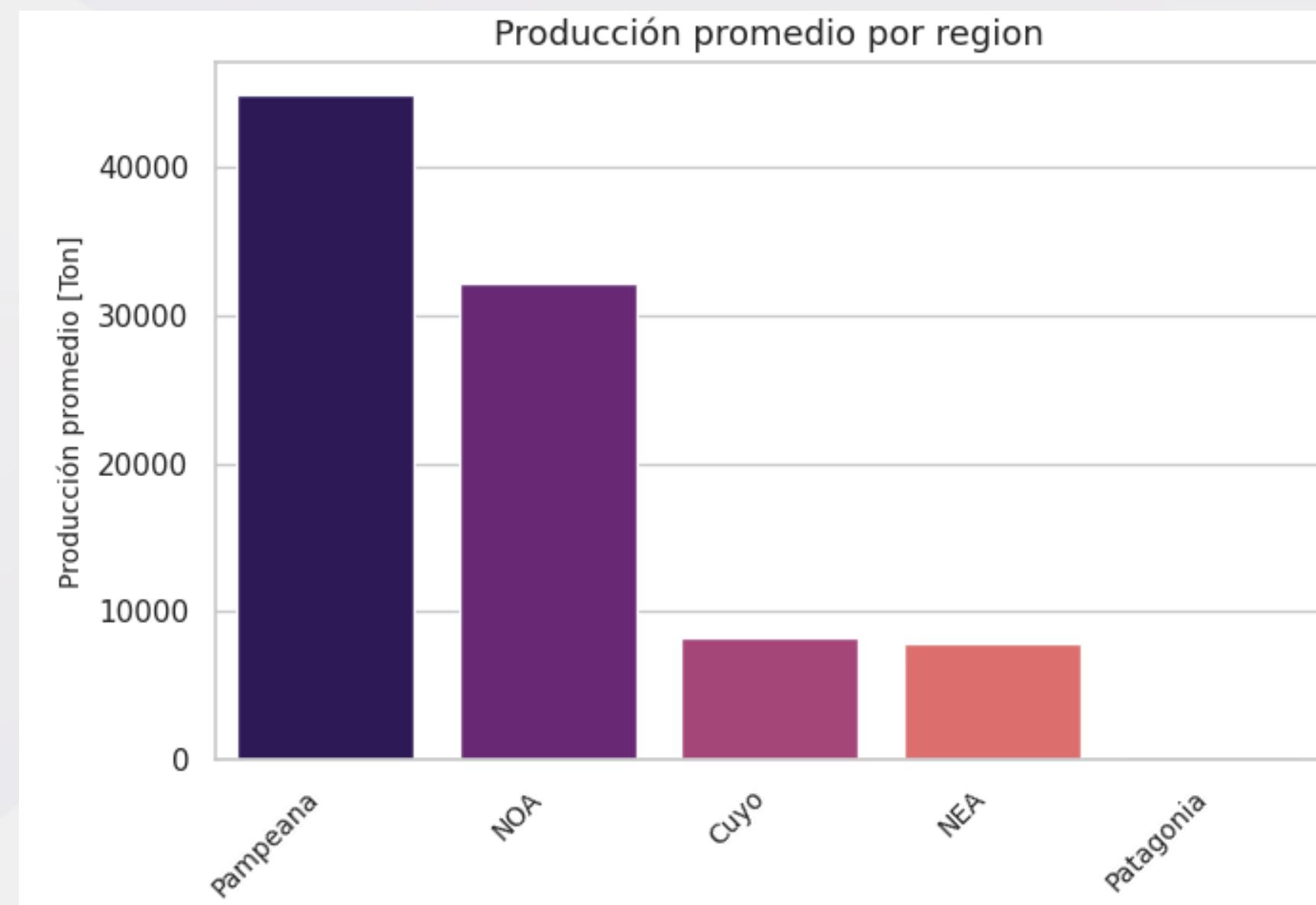
Se crearon distintos gráficos para visualizar el comportamiento y la distribución de los valores y features

## • COMPARACIÓN POR REGIÓN: SUPERFICIE SEMBRADA VS. COSECHADA



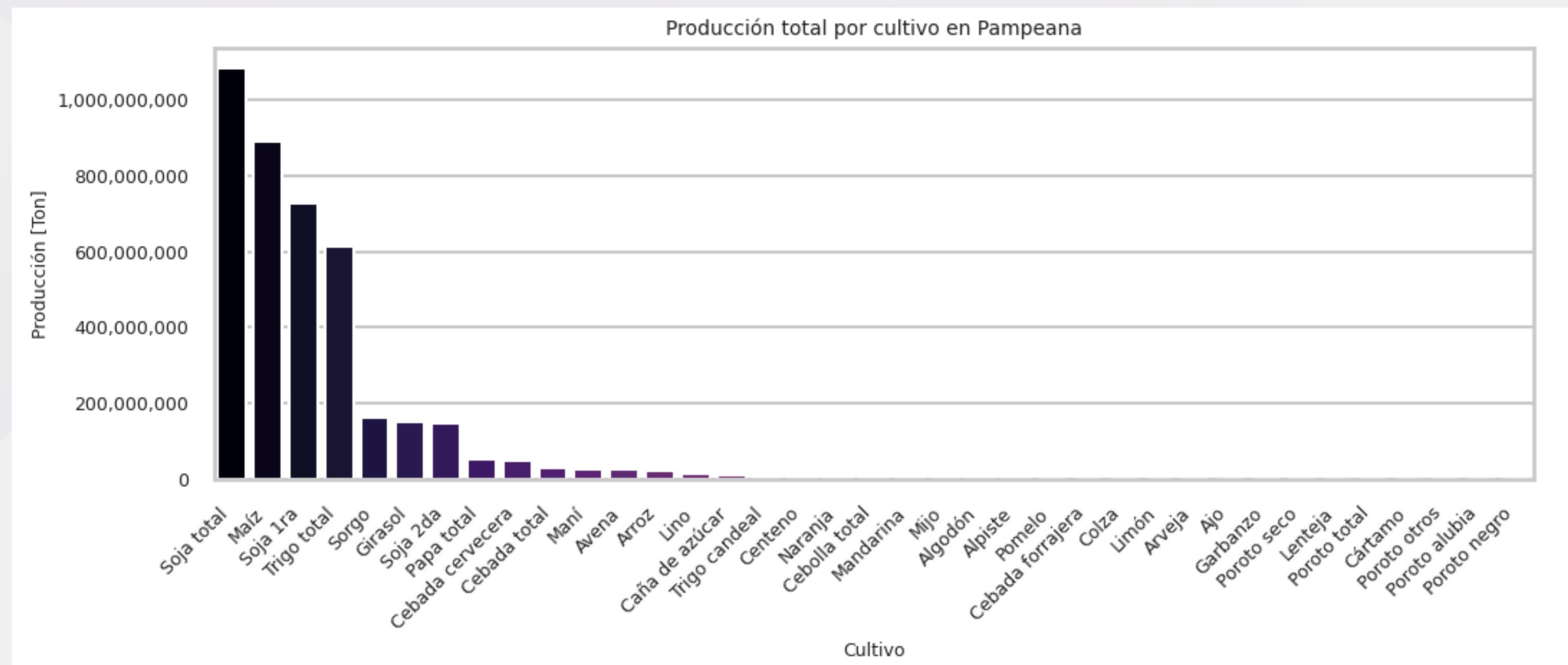
**La distribución regional muestra una concentración fuertemente desigual de la actividad agrícola, con un dominio absoluto de la región pampeana.** Este patrón refleja la realidad productiva del país y puede influir directamente en decisiones sobre políticas públicas, inversión y logística agrícola.

## • PRODUCCIÓN Y EFICIENCIA AGRÍCOLA POR REGIÓN



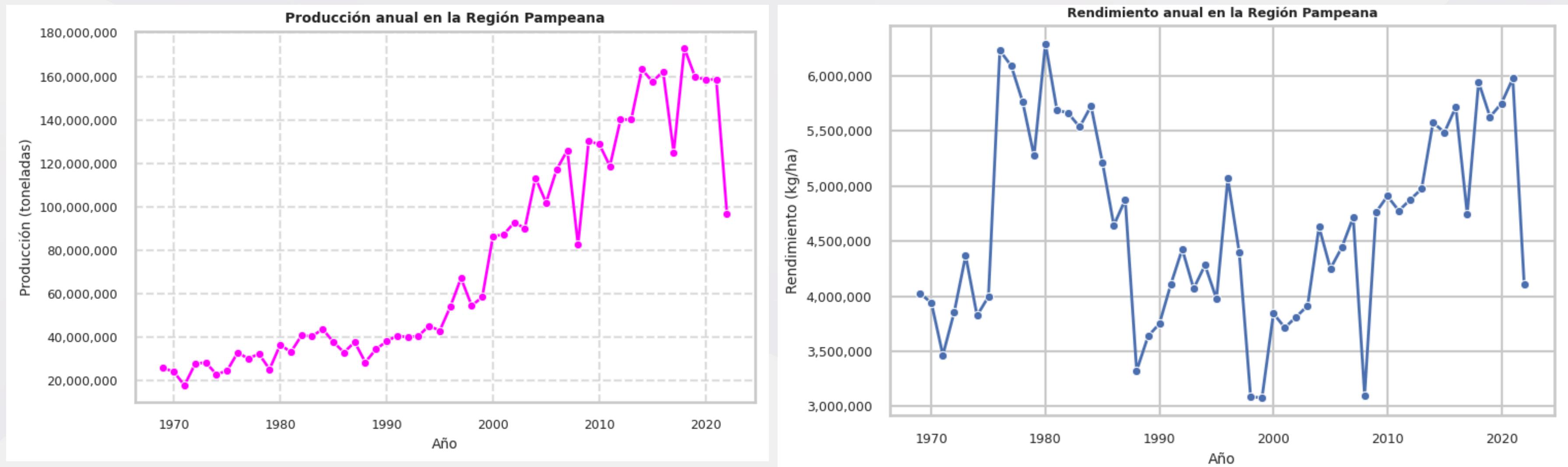
La región pampeana lidera ampliamente en producción promedio, seguida por el NOA. Sin embargo, al analizar la tasa de cosecha, se observa que NEA alcanza el mayor porcentaje, superando incluso a regiones con mayor volumen productivo.

Esto sugiere que algunas regiones, aunque tengan menor superficie o producción total, logran una mayor eficiencia en la cosecha. La Patagonia y Cuyo, en cambio, muestran bajos niveles de producción y tasas de cosecha más reducidas, reflejando limitaciones estructurales o ambientales que podrían abordarse con mejoras tecnológicas o de infraestructura.



Soja, maíz y trigo concentran la mayor parte de la producción pampeana.

# EVOLUCIÓN HISTÓRICA DE LA PRODUCCIÓN Y EL RENDIMIENTO EN LA REGIÓN PAMPEANA



La producción total en la región pampeana creció fuertemente desde los años 90, mientras que el rendimiento muestra una mejora sostenida a pesar de su variabilidad anual.

# APRENDIZAJE SUPERVISADO

## 01. Se eliminan features

Para el análisis se usaron:

- 'region'
- 'cultivo'
- 'sup\_sembrada\_ha'
- 'sup\_cosechada\_ha'
- 'produccion\_ton'
- 'rendimiento\_kg/ha'
- 'tasa\_cosecha'
- 'anio'



## 02. Elección cultivos

Con el objetivo de enfocar el análisis en los cultivos de mayor impacto productivo a nivel nacional, se procedió a identificar aquellos con la mayor producción total acumulada:

- Soja
- Maiz
- Trigo

## 03. Transformación

Transformación de variables categóricas para "cultivo" y "region":

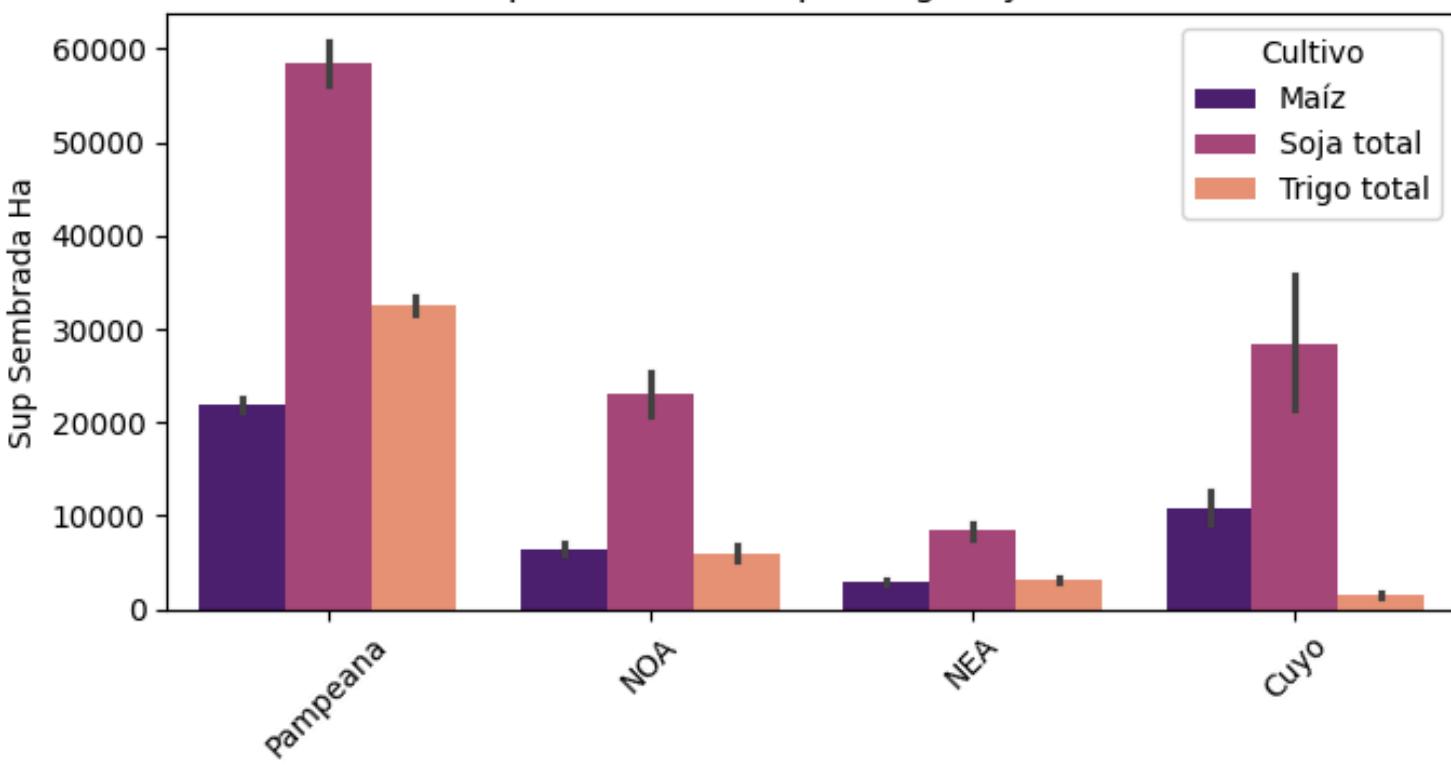
- OneHotEncoder

## 04. Implementación del modelo

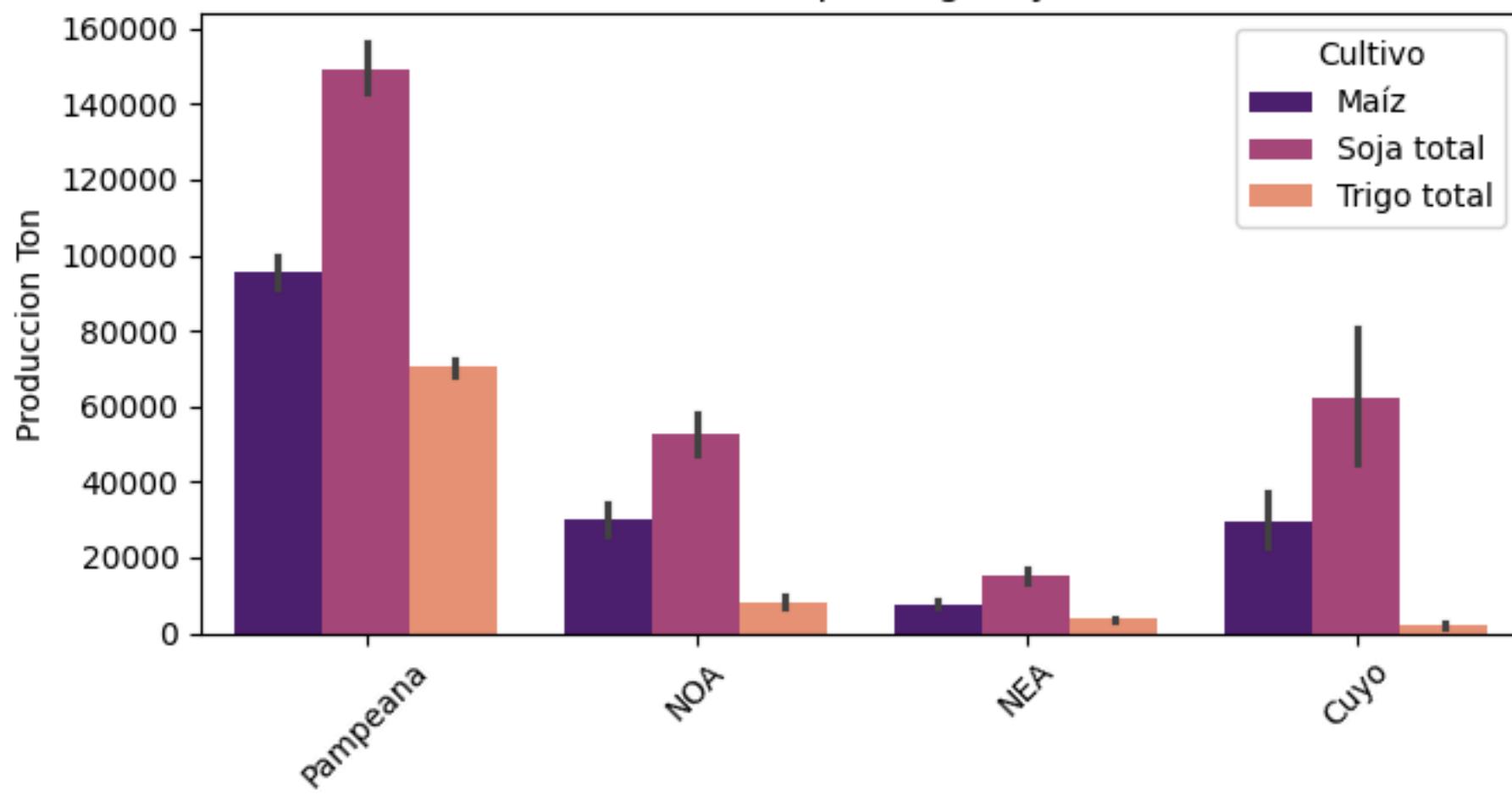
Se implementaron los siguientes modelos:

- Regresión lineal
- Random forest regressor

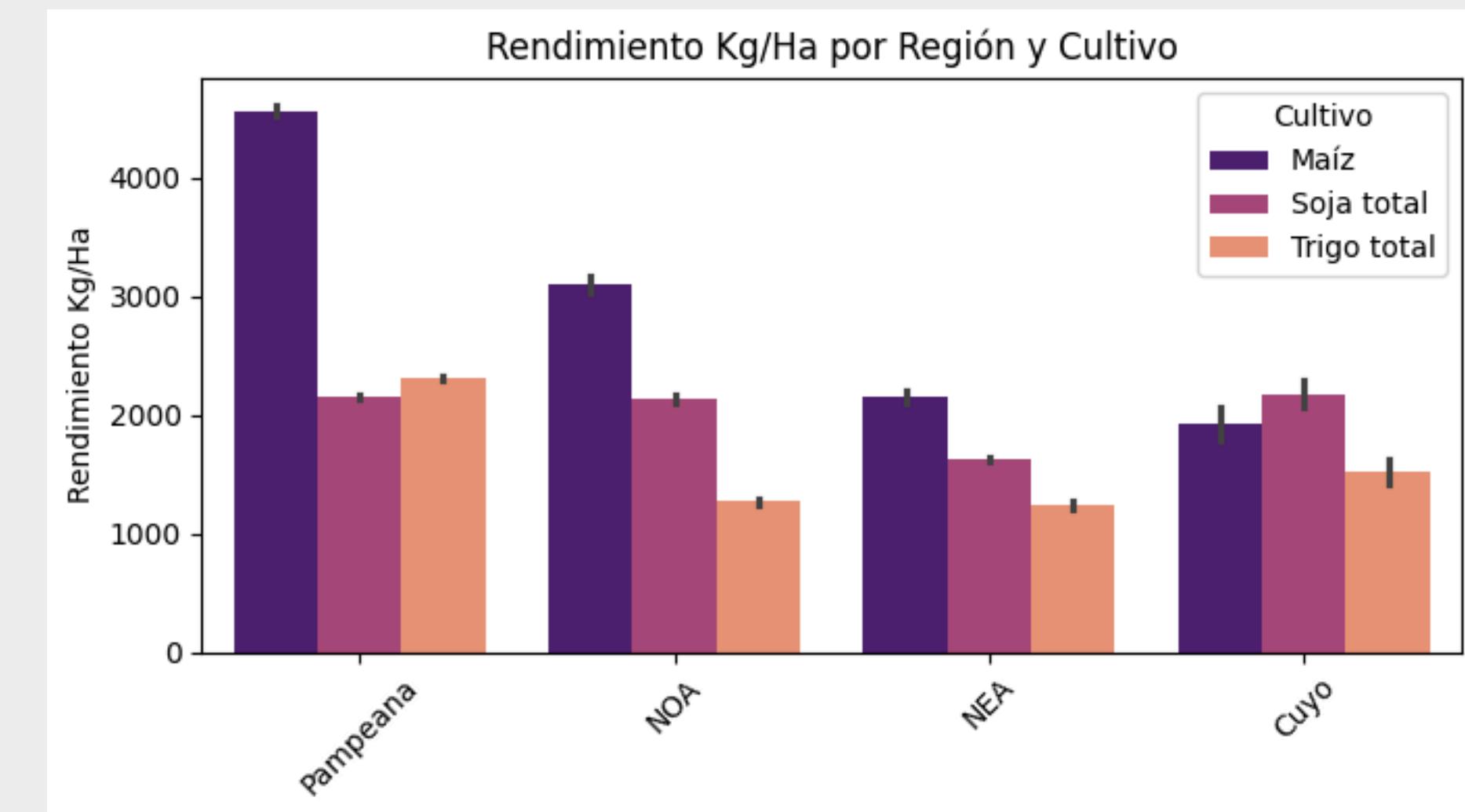
Sup Sembrada Ha por Región y Cultivo



Producción Ton por Región y Cultivo



Rendimiento Kg/Ha por Región y Cultivo



# RESULTADOS

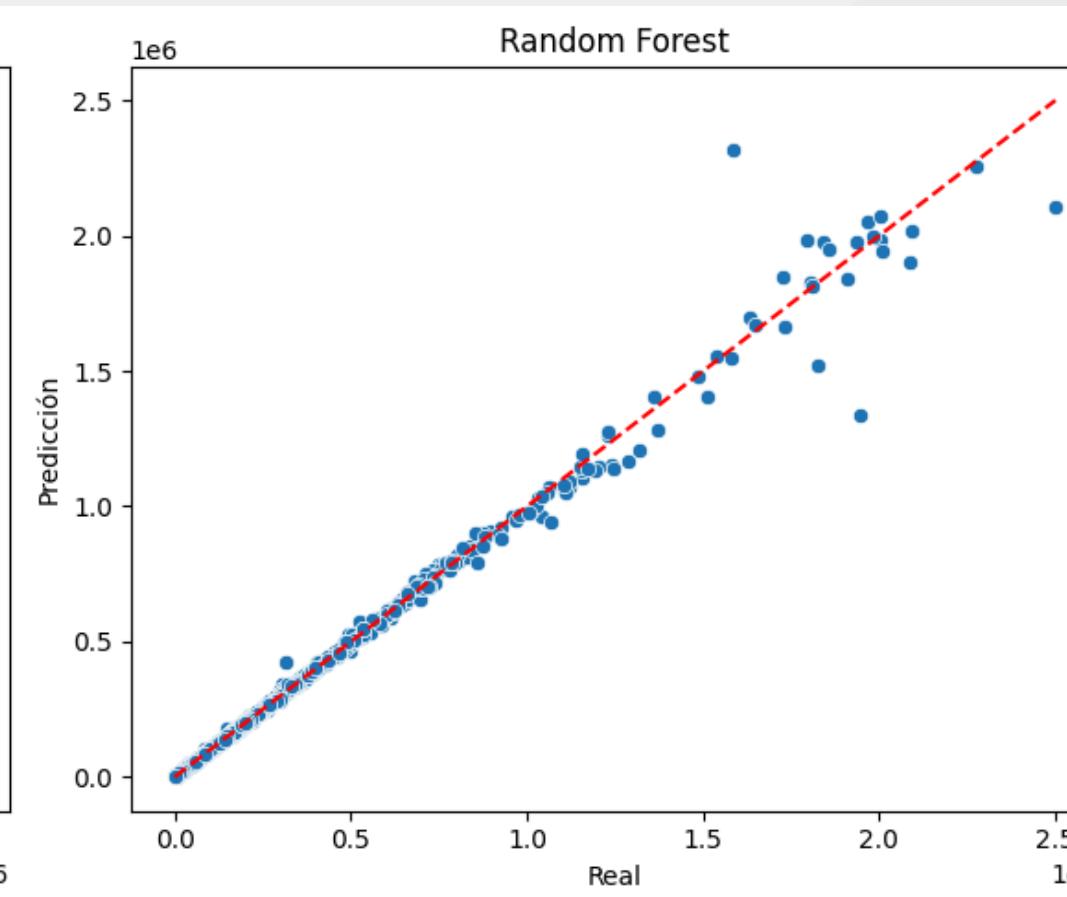
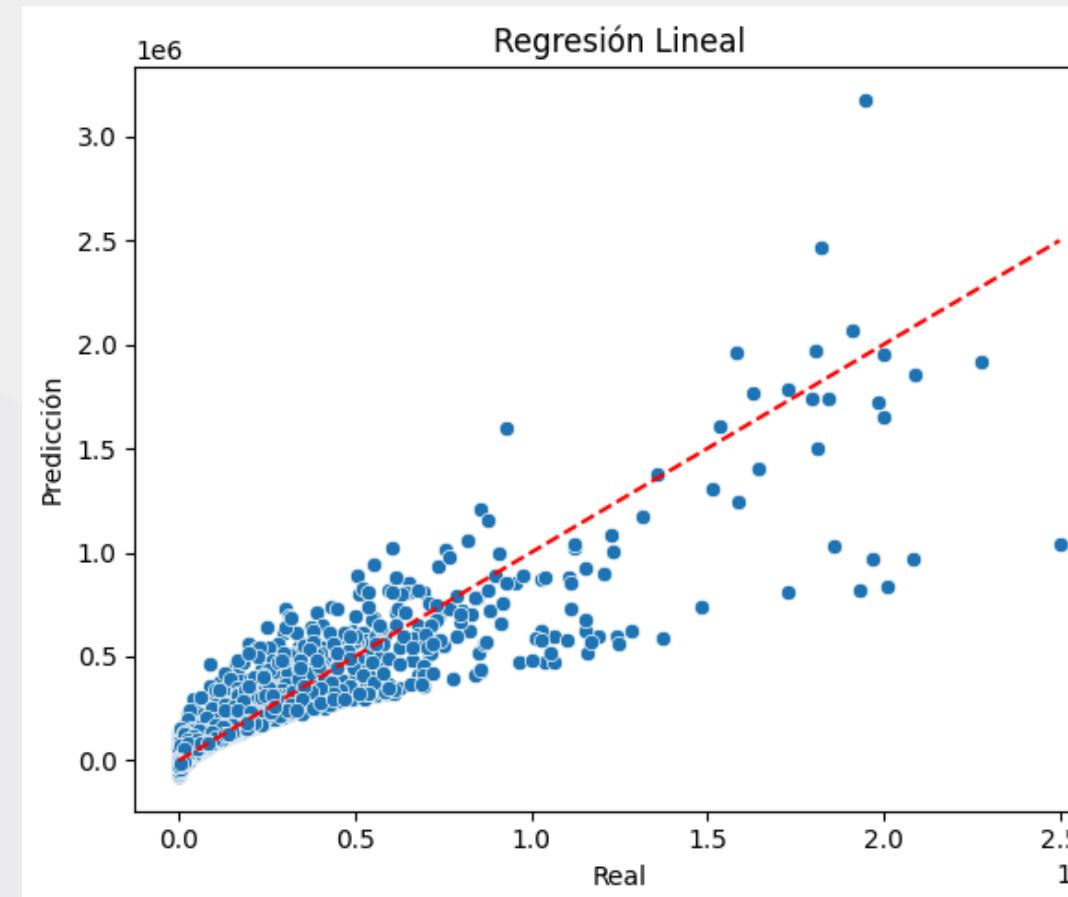
## MODELO UTILIZANDO LOS 3 CULTIVOS CON MAYOR PRODUCCION

Para ambos modelos se evaluaron las siguientes métricas:

- Error medio absoluto
- Error cuadrado medio
- R2

### REGRESION LINEAL

MAE: 32551.07  
RMSE: 69073.10  
 $R^2$ : 0.84



### RANDOM FOREST REGRESSOR

MAE: 1278.67  
RMSE: 13573.66  
 $R^2$ : 0.99

# RESULTADOS

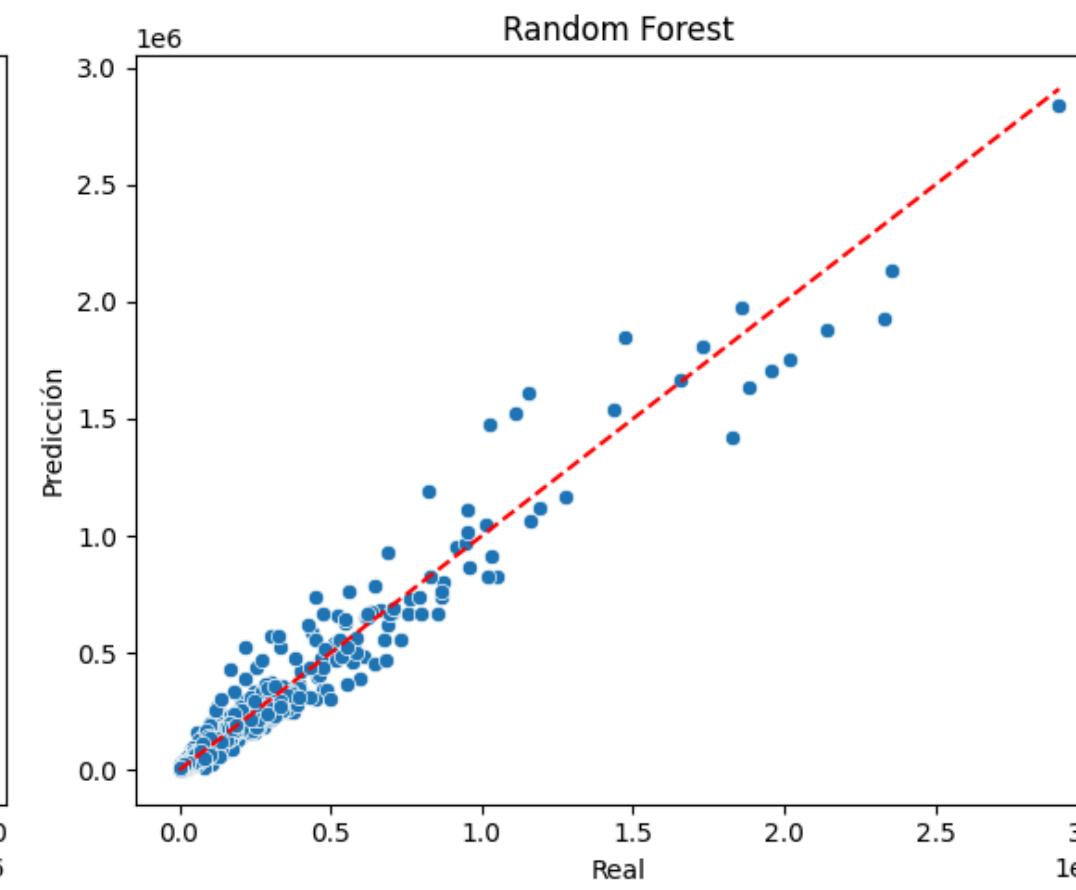
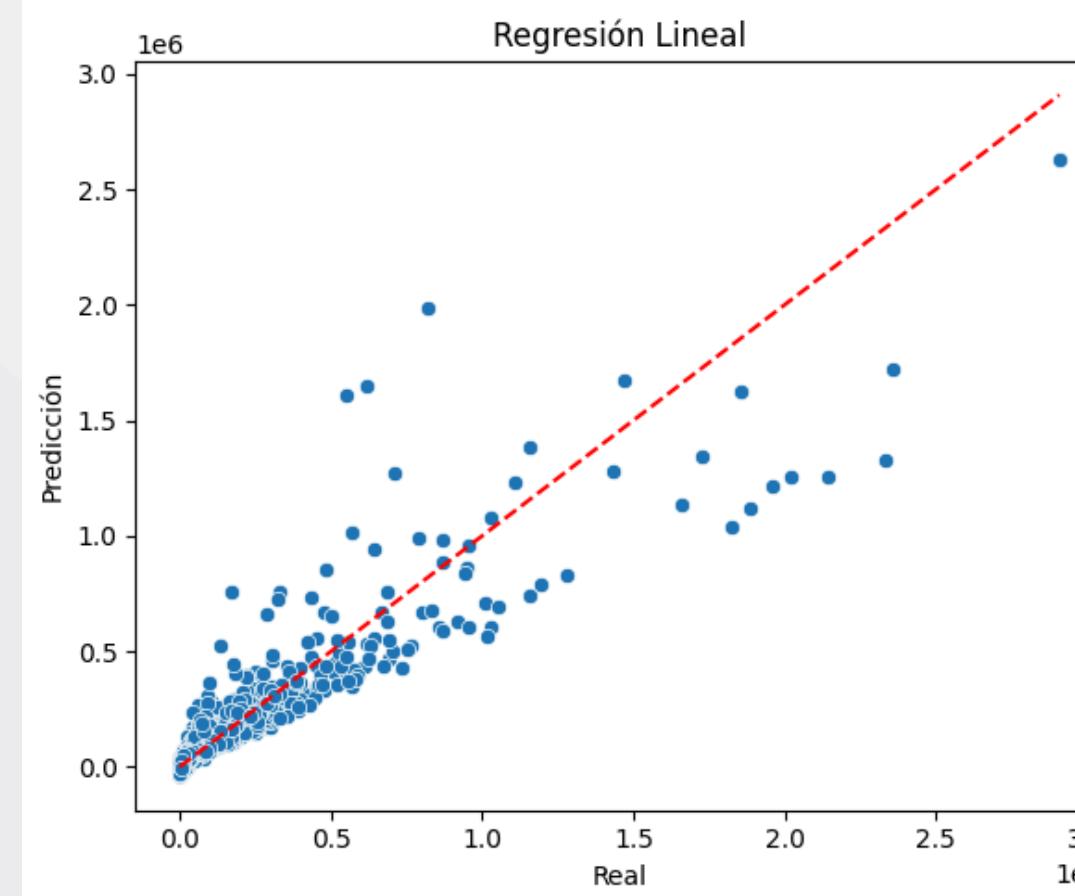
## MODELO UTILIZANDO EL CULTIVO MAÍZ

Para ambos modelos se evaluaron las siguientes métricas:

- Error medio absoluto
- Error cuadrado medio
- R<sup>2</sup>

### REGRESION LINEAL

MAE: 27990.84  
R<sup>2</sup>: 0.84



### RANDOM FOREST REGRESSOR

MAE: 11255.85  
R<sup>2</sup>: 0.96

# RESULTADOS

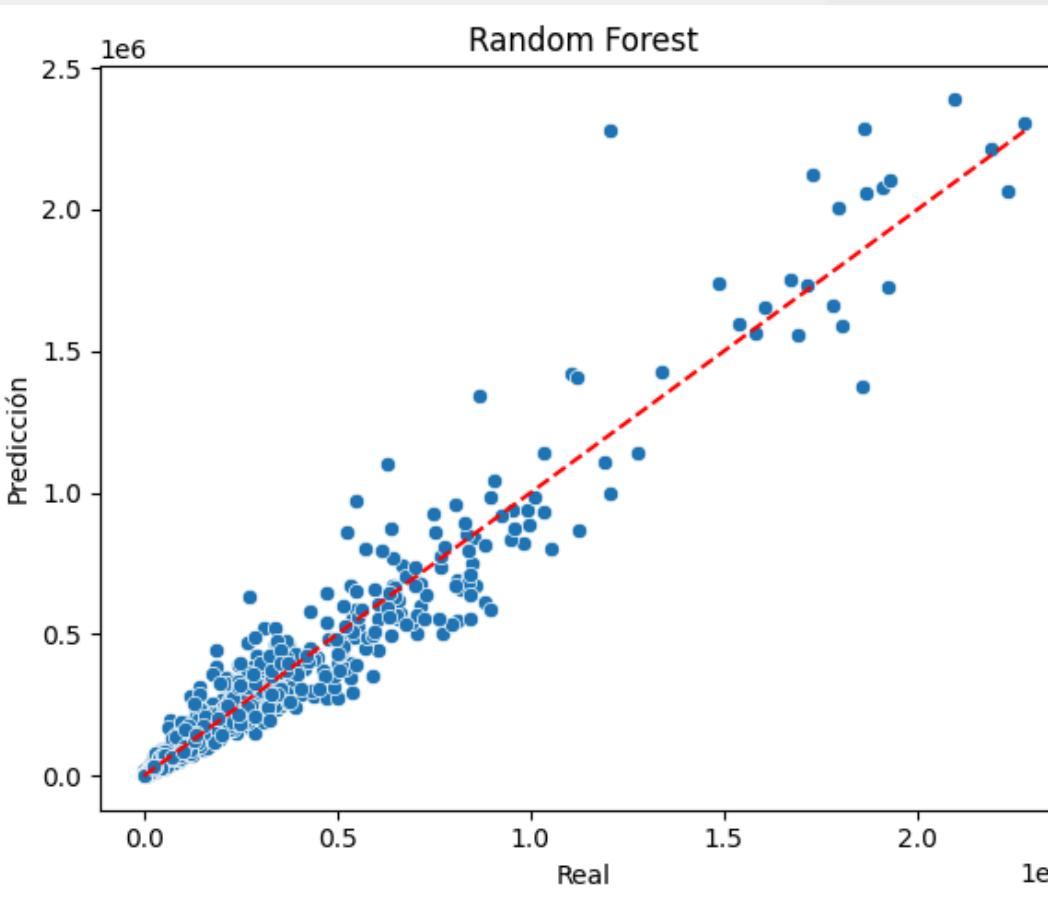
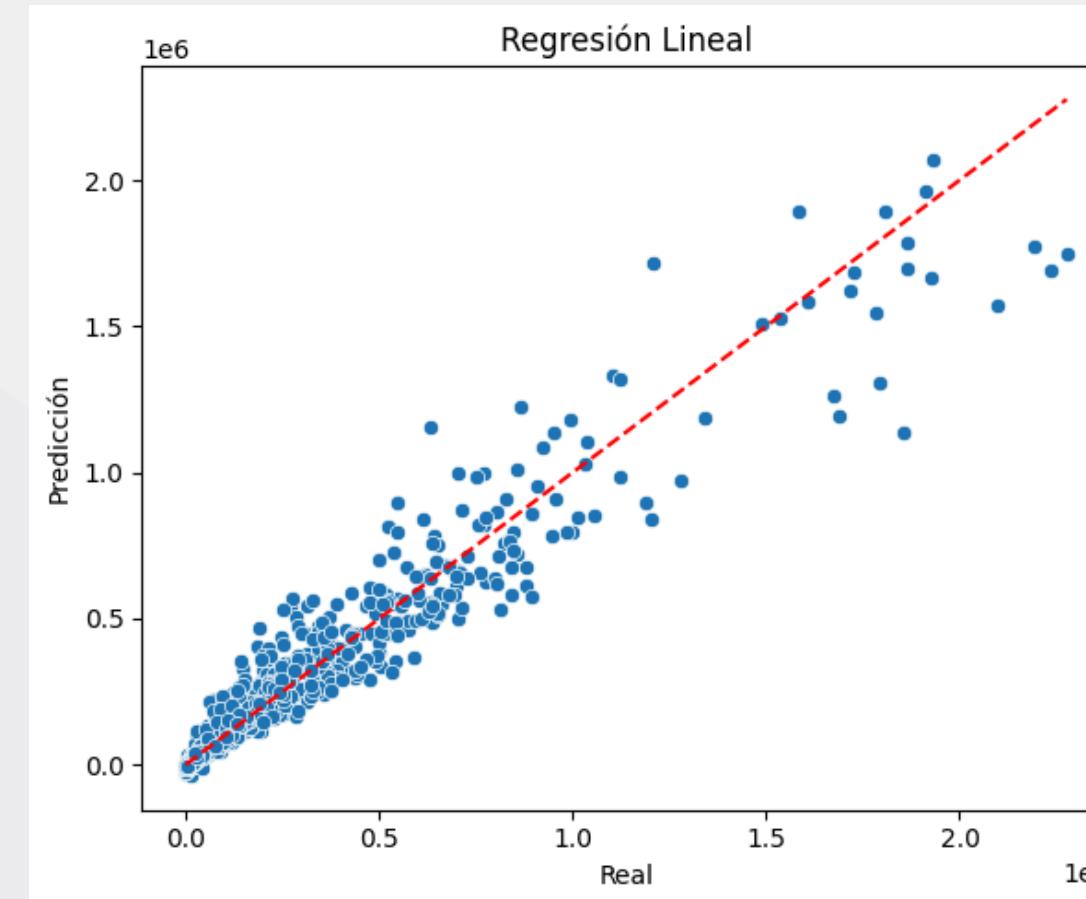
## MODELO UTILIZANDO EL CULTIVO SOJA

Para ambos modelos se evaluaron las siguientes métricas:

- Error medio absoluto
- Error cuadrado medio
- R<sup>2</sup>

### REGRESION LINEAL

MAE: 26247.54  
R<sup>2</sup>: 0.94



### RANDOM FOREST REGRESSOR

MAE: 19815.99  
R<sup>2</sup>: 0.95

# RESULTADOS

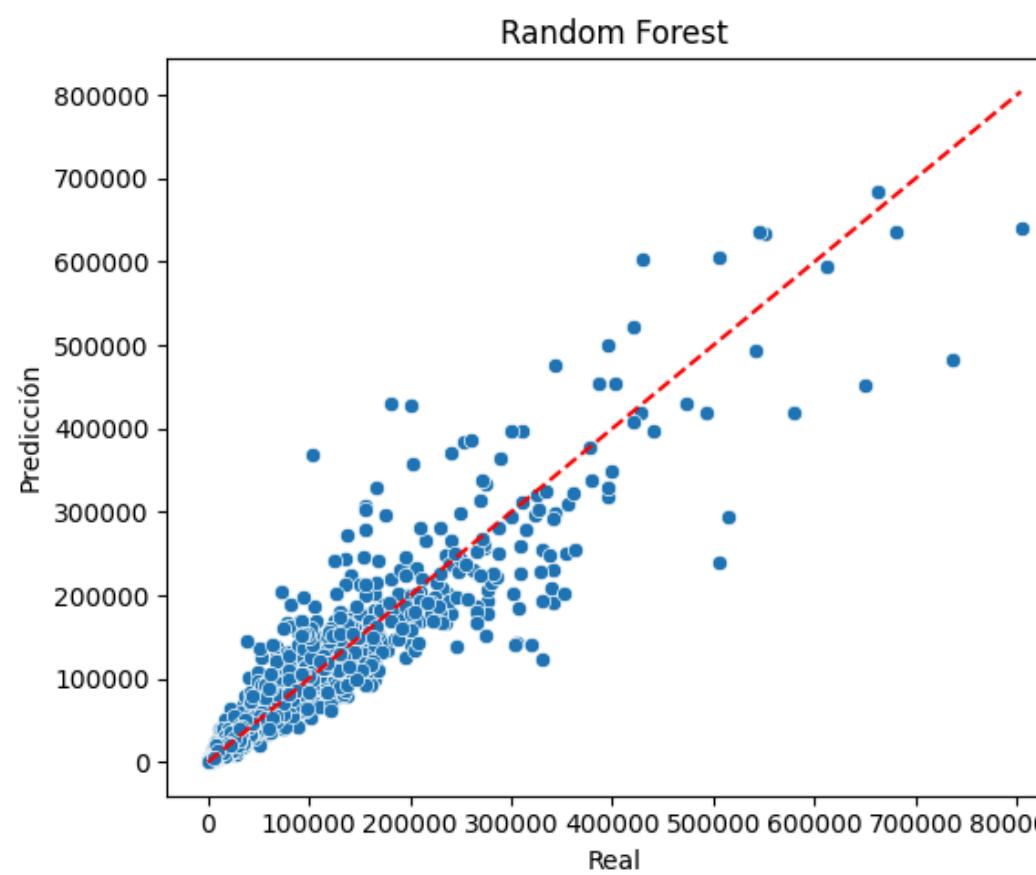
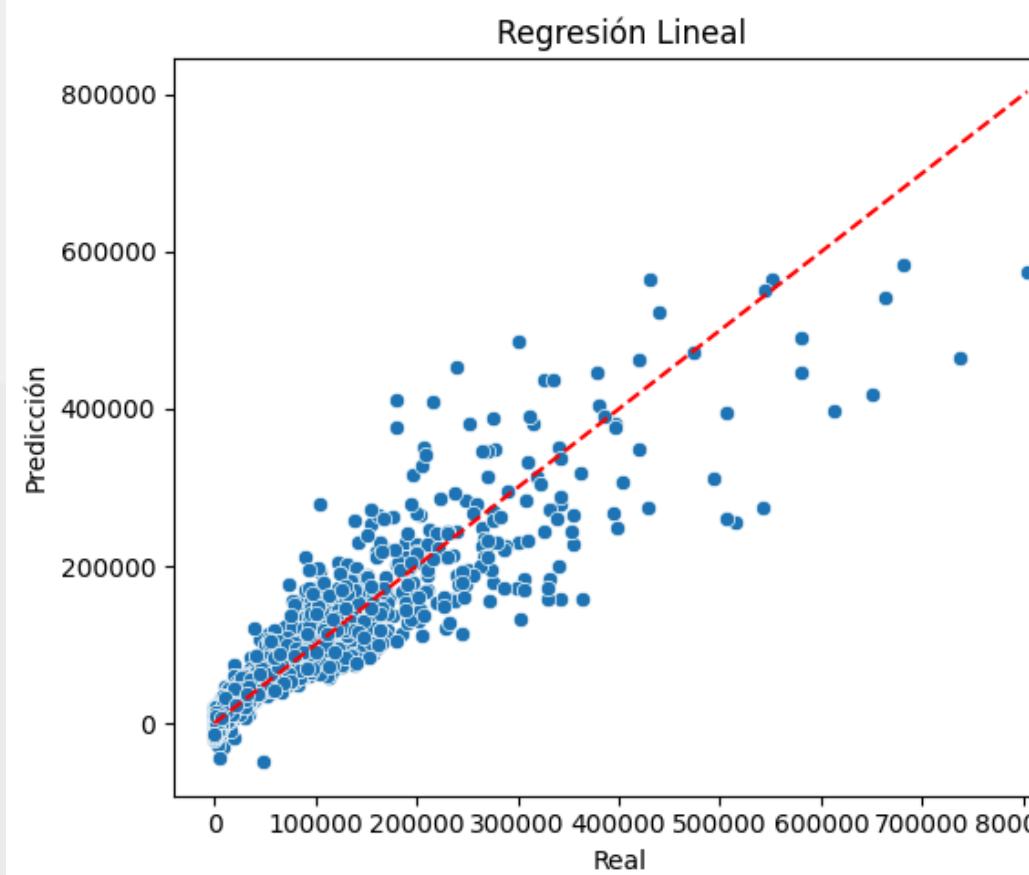
## MODELO UTILIZANDO EL CULTIVO TRIGO

Para ambos modelos se evaluaron las siguientes métricas:

- Error medio absoluto
- Error cuadrado medio
- R2

### REGRESION LINEAL

MAE: 17655.14  
 $R^2$ : 0.86



### RANDOM FOREST REGRESSOR

MAE: 11720.09  
 $R^2$ : 0.89

# HIPERPARÁMETROS

Se evaluaron los hiperparámetros para encontrar el mejor modelo resultando:

## MODELO UTILIZANDO LOS 3 CULTIVOS

Mejor configuración encontrada:  
{'n\_estimators': 200,  
'min\_samples\_split': 2,  
'min\_samples\_leaf': 1,  
'max\_features': None,  
'max\_depth': None}

### Métricas:

- MAE: 1149.99
- R<sup>2</sup>: 0.995

## MODELO PARA MAÍZ

Mejor configuración encontrada:  
{'n\_estimators': 500,  
'min\_samples\_split': 2,  
'min\_samples\_leaf': 1,  
'max\_features': None,  
'max\_depth': 20}

### Métricas:

- MAE: 11188.91
- R<sup>2</sup>: 0.96

## MODELO PARA SOJA

Mejor configuración encontrada:  
{'n\_estimators': 100,  
'min\_samples\_split': 2,  
'min\_samples\_leaf': 2,  
'max\_features': 'sqrt',  
'max\_depth': 20}

### Métricas:

- MAE: 19250.19
- R<sup>2</sup>: 0.95

## MODELO PARA TRIGO

Mejor configuración encontrada:  
{'n\_estimators': 300,  
'min\_samples\_split': 10,  
'min\_samples\_leaf': 1,  
'max\_features': 'log2',  
'max\_depth': 30}

### Métricas:

- MAE: 11217.47
- R<sup>2</sup>: 0.90

# OVERFITTING

## REGULARIZAR EL MODELO

Se pueden ajustar aún más los hiperparámetros y aplicar técnicas simples de regularización como max\_depth o min\_samples\_leaf y verificar si el rendimiento se mantiene sin sobreajuste.

## VALIDACIÓN CRUZADA

Usa todo el dataset para entrenamiento y validación

Si el modelo rinde muy bien en entrenamiento pero mal en validación cruzada, hay overfitting.

No depende de una sola partición de test

```
[95] from sklearn.model_selection import cross_val_score
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.metrics import make_scorer, r2_score

    # Random Forest con hiperparámetros iniciales
    rf_cv = RandomForestRegressor(
        n_estimators=100,
        max_depth=15,
        min_samples_leaf=5,
        min_samples_split=10,
        max_features='sqrt',
        random_state=42,
        n_jobs=-1
    )

    # Usamos R2 como métrica, en 5 folds
    scores = cross_val_score(rf_cv, X, y, cv=5, scoring='r2')
    print("R2 en cada fold:", scores)
    print(f"R2 promedio: {scores.mean():.3f}")
    print(f"Desviación estandar: {scores.std():.3f}")

→ R2 en cada fold: [0.99040755 0.9888277 0.98981825 0.98701928 0.94223494]
   R2 promedio: 0.980
   Desviación estandar: 0.019
```

# APRENDIZAJE NO SUPERVISADO



## 01. Features relevantes

Para el modelado se usaron:

- 'sup\_sembrada\_ha'
- 'produccion\_ton'
- 'rendimiento\_kg/ha'

El análisis de correlación mostró que 'rendimiento\_kg/ha' tiene baja correlación con las otras dos, lo cual indica que podría aportar información complementaria útil para la segmentación.



## 02. Transformación

Transformación de variables categóricas:

- LabelEncoder



## 03. PCA

Análisis de componentes principales y su aporte a la varianza



## 04. Implementación del modelo

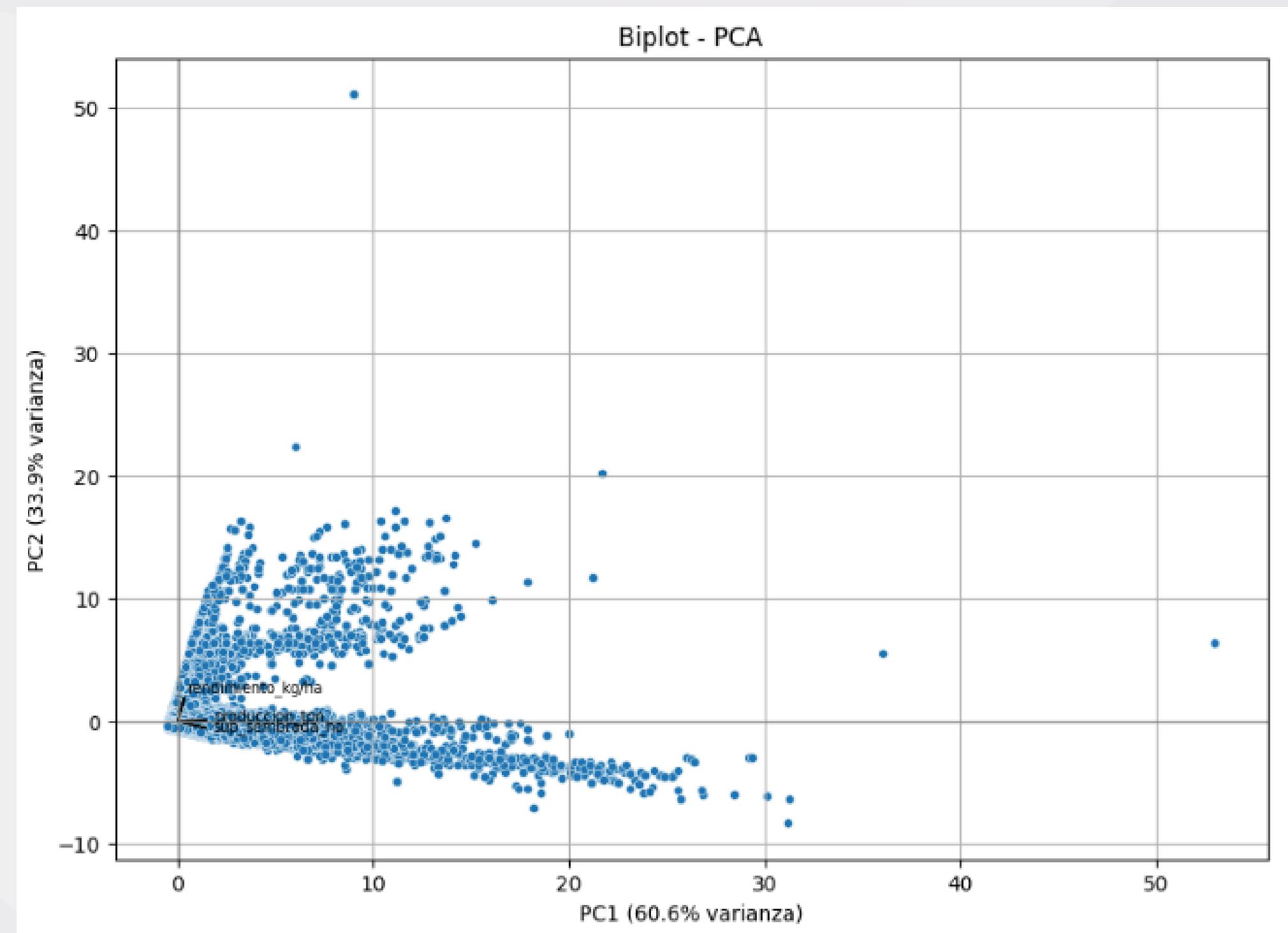
Se implementaron los siguientes modelos:

- K-Means
- DBSCAN

# PCA

	PC1	PC2
sup_sembrada_ha	0.681424	-0.280756
produccion_ton	0.710231	0.030670
rendimiento_kg/ha	0.176726	0.959289

- PC1 explica el 60.56% de la varianza
- PC2 explica el 33.87% de la varianza

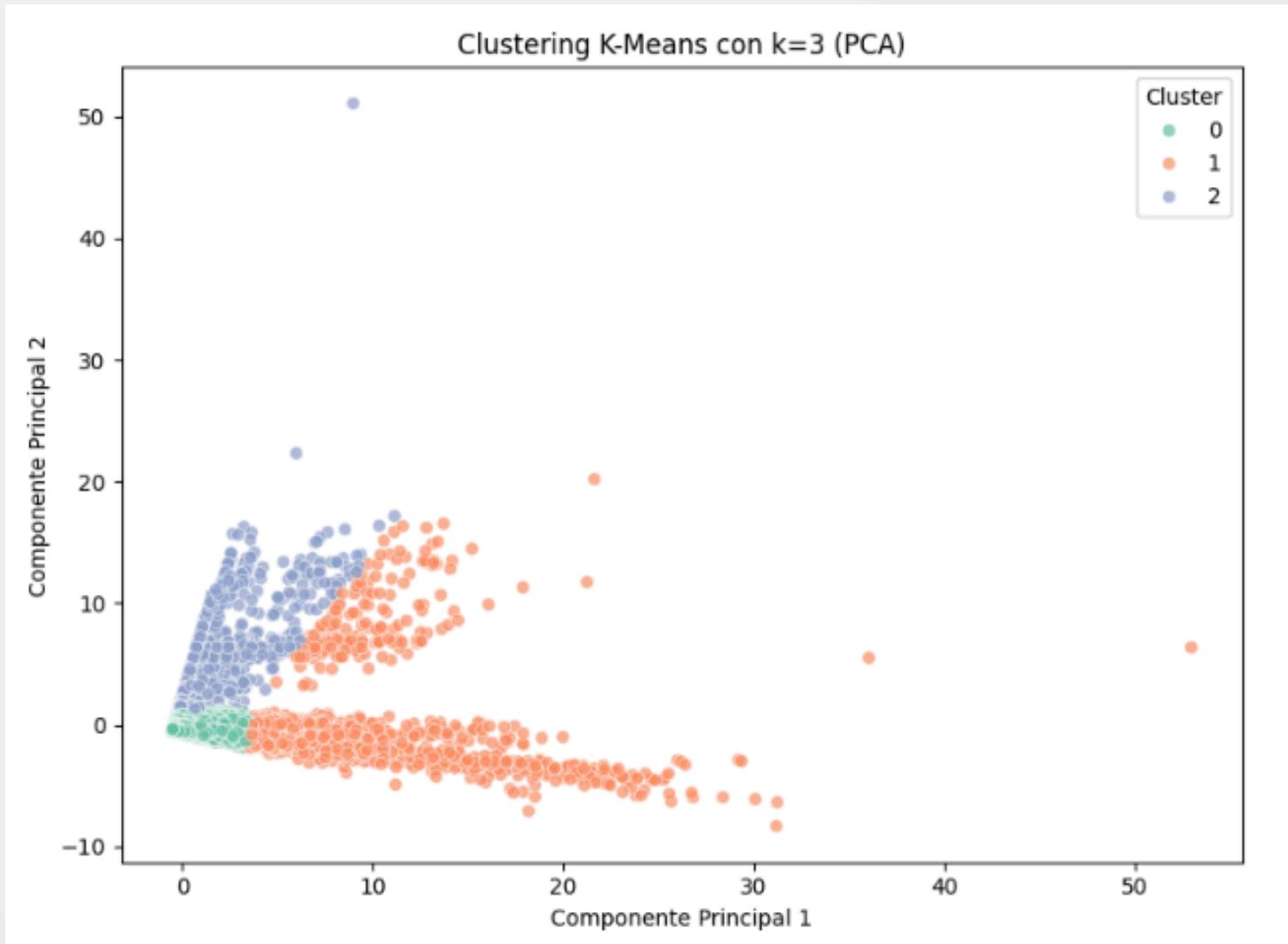


# RESULTADOS

La reducción de dimensionalidad con PCA permitió una visualización efectiva de los grupos en 2D.

## K-MEANS

Se utilizó el método del codo y el silhouette score para definir la cantidad de clusters adecuada.

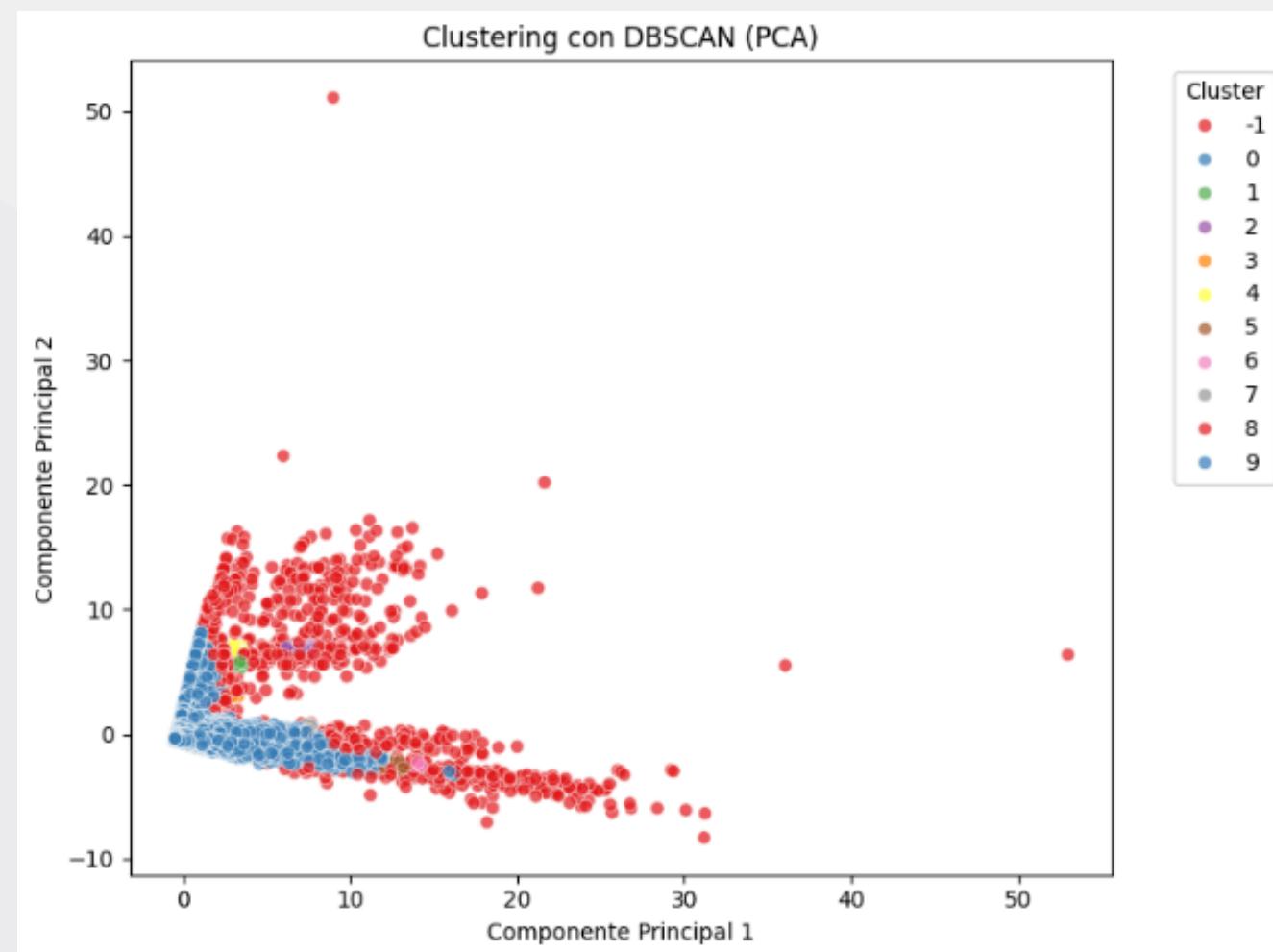


KMeans forzó la segmentación sin una separación clara de comportamientos productivos.

# RESULTADOS

La reducción de dimensionalidad con PCA permitió una visualización efectiva de los grupos en 2D.

## DBSCAN



Etiquetas únicas encontradas: [-1 0 1 2 3 4 5 6 7 8 9]  
Cantidad de clusters (sin contar ruido): 10

dbSCAN\_cluster

-1	779
0	152957
1	10
2	40
3	9
4	11
5	28
6	8
7	12
8	12
9	9

DBSCAN detectó 10 clusters.  
No se encontraron agrupamientos coherentes ni interpretables.

# CONCLUSIONES

En conclusión, los modelos supervisados fueron claramente los más efectivos para predecir la producción.

Las variables disponibles permiten construir buenos modelos regresivos. Los métodos no supervisados no lograron identificar patrones útiles en este caso, pero sirvieron para contrastar y validar el enfoque.

**MUCHAS  
GRACIAS**