Análisis de historiales agrícolas para predicción, agrupación y clasificación de cosechas

Samsung Innovation Campus

Joshua David Salcedo Monroy, Jesús Antonio Torres Contreras, Luis Angel Lozano Reyes, Gregorio Salazar Solís

Abstrac - En este estudio se explora la aplicación de técnicas de machine learning en la agricultura para poder predecir y clasificar variables claves relacionadas con el rendimiento de los cultivos basados en las características más importantes para la producción agrícola. Se utilizan principalmente tres enfoques: Regresión lineal múltiple para estimar la producción en función del fertilizante y la lluvia, Random Forest para la clasificación binaria considerando factores como el fertilizante y la irrigación y K-Means para agrupar cultivos en función de cada característica de la base de datos. Los resultados muestran la efectividad de estos métodos en la toma de decisiones agrícolas y su aplicabilidad en la optimización de recursos.

I. INTRODUCCIÓN

En la agricultura moderna, la capacidad de predecir el crecimiento de los cultivos y maximizar su rendimiento es fundamental para garantizar la seguridad alimentaria y la rentabilidad del sector. Factores como el clima, el tipo de riego, la calidad del suelo y el uso de fertilizantes influyen directamente en la productividad, por lo que contar con herramientas que permitan analizar estos datos y generar proyecciones precisas se ha convertido en una necesidad crítica. En este contexto, el machine learning (ML) emerge como una tecnología clave para desarrollar modelos predictivos que ayuden a los agricultores y agroindustrias a tomar decisiones basadas en datos.

Este reporte explora la aplicación de tres técnicas fundamentales de ML en el ámbito agrícola:

a) Modelos de regresión: Para predecir variables continuas como el rendimiento en toneladas según condiciones ambientales y de manejo del cultivo.
b) Algoritmos de clasificación: Que permiten categorizar cultivos según su potencial de crecimiento o susceptibilidad a enfermedades en función de parámetros como humedad, temperatura y tipo de fertilizante.
c) Métodos de agrupamiento (clustering): Útiles para segmentar zonas de cultivo con características similares y optimizar recursos como el riego o la aplicación de nutrientes.

El objetivo principal es demostrar cómo estas técnicas pueden predecir el crecimiento de los cultivos y estimar las ganancias esperadas, permitiendo una planificación más eficiente. Para ello, se analizarán conjuntos de datos reales que incluyen las siguientes variables:

- Lluvia
- Temperatura
- Uso de fertilizante
- Riego
- Días para la cosecha
- Rendimiento en toneladas por hectárea
- Región
- Tipo de suelo
- Cultivo
- Condición del clima

Los resultados obtenidos no solo ayudarán a maximizar la producción, sino también a reducir costos operativos y minimizar el impacto ambiental mediante un uso más inteligente de los recursos. Este estudio busca ser un referente para la implementación de agricultura de precisión, donde el machine learning se convierte en un aliado estratégico para enfrentar los desafíos de la producción agrícola en un escenario de cambio climático y creciente demanda de alimentos.

Revisión de la literatura

Explorando los antecedentes de proyectos con esta misma temática, se encontró que existen diferentes publicaciones que abordan este mismo tema, aunque cada uno con diferentes enfoques, lo que ayudó a encontrar las metodologías óptimas para la realización del proyecto

1. Implementation of Prediction of Crop Using SVM Algorithm por Arya Phadnis, Shivam Panchal, Rajat Jadhav, Buddhi Rajdeep, Deepak Patil (2023): En él se explora la aplicación del algoritmo de Máquinas de Soporte Vectorial para la predicción de cultivos, con el objetivo de optimizar la gestión agrícola y maximizar los rendimientos. Se destaca que la predicción de cultivos, basada en datos históricos como condiciones climáticas, calidad del suelo y rendimientos previos, es fundamental para tomar decisiones informadas sobre siembra, cosecha y manejo de riesgos como enfermedades o plagas. Se propone un

sistema que utiliza SVM para clasificar y predecir el rendimiento de cultivos en función de las características del suelo y otros factores ambientales.

2. Nuevo sistema de predicción de cosecha de cereales de castilla y león por Alberto Gutiérrez, Ignacio Villarino, David A. Nafría, Nieves Garrido, Inmaculada Abia, Miriam Fernández y Lorenzo Rodríguez: El artículo presenta un sistema para estimar los rendimientos de trigo y cebada de secano en Castilla y León, una región clave en la producción cerealista de España. Este sistema utiliza el modelo agronómico AquaCrop, calibrado para la región, integrando datos meteorológicos reales, predicciones a 10 días y escenarios climáticos históricos de 30 años.

3. Crop Prediction Using Machine Learning de Rao et al. (2022): El artículo aborda la predicción de cultivos mediante el uso de algoritmos de aprendizaje supervisado, con el objetivo de asistir a los agricultores en la selección de cultivos óptimos según las condiciones climáticas y los nutrientes del suelo. Se comparan tres modelos: K-Nearest Neighbor (KNN), Decision Tree y Random Forest, aplicados a un conjunto de datos de Kaggle con 22 variedades de cultivos y siete características

4. Inteligencia Artificial para la Predicción de Cosechas de Saturdays.AI (2022): El artículo describe un proyecto que emplea inteligencia artificial para predecir rendimientos de cosechas de maíz en Illinois, Estados Unidos, ante la falta de datos normalizados en España, inicialmente considerada como región de estudio. Se destaca la importancia de la agricultura en Illinois, un estado con clima inestable y alta producción de maíz y soja, afectada por eventos climáticos extremos. El enfoque se centra en la limpieza y gestión de datos históricos, utilizando series temporales de temperatura a 2 metros del suelo y el índice de vegetación EVI, procesadas mediante diferenciación regular e interpolación para alinearlas.

5. Crop Prediction Model Using Machine Learning Algorithms de Elbasi et al. (2023): Él explora el uso de algoritmos de aprendizaje automático para optimizar la predicción de cultivos en la agricultura moderna, integrando datos de sensores IoT en tiempo real. Los autores evalúan 15 algoritmos, como Bayes Net, Naïve Bayes Classifier, Random Forest y Multilayer Perception, utilizando un conjunto de datos de Kaggle con 2200 registros y 22 etiquetas de cultivos, basado en siete características.

6. Crop Prediction using Machine Learning Approaches de Nischitha et al. (2020): El publicado en el International Journal of Engineering Research & Technology, presenta un sistema basado en algoritmos de aprendizaje automático para recomendar cultivos óptimos en India, donde la agricultura es predominante pero enfrenta desafíos como la falta de rotación de cultivos y el uso inadecuado de fertilizantes. Los autores diseñaron un modelo que utiliza Support Vector Machine (SVM) con kernel de función de base radial para predecir la precipitación anual a partir de datos históricos, y Decision Tree para recomendar cultivos basándose en parámetros como temperatura, humedad, pH del suelo y la precipitación predicha.

II. PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

El problema que se busca resolver es la dificultad para predecir y optimizar el rendimiento de los cultivos en la agricultura moderna, debido a la influencia de múltiples factores variables como el clima, el riego, la calidad del suelo y el uso de fertilizantes. Esta falta de precisión en el análisis y proyección de datos limita la capacidad de los agricultores y agroindustrias para tomar decisiones efectivas, lo que puede resultar en menor productividad, mayores costos operativos y un impacto ambiental negativo por el uso ineficiente de recursos.

Justificación

La implementación de técnicas de machine learning en la agricultura se justifica por la necesidad urgente de enfrentar los desafíos que afectan el rendimiento de los cultivos y la seguridad alimentaria en un contexto de cambio climático y creciente demanda de alimentos. Factores como el clima, el riego, la calidad del suelo y el uso de fertilizantes impactan directamente la productividad agrícola, y su análisis preciso mediante herramientas avanzadas es esencial para optimizar recursos y mejorar la toma de decisiones. Según la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO), la población mundial alcanzará los 8,500 millones de personas en 2030, lo que implicará un aumento del 60% en la demanda global de alimentos. Este incremento, combinado con los efectos del cambio climático, pone en riesgo la capacidad de los sistemas agrícolas tradicionales para satisfacer las necesidades alimentarias.

El cambio climático agrava esta situación al alterar patrones climáticos y reducir el rendimiento de cultivos clave. Por ejemplo, un estudio publicado en Scielo ("Cambio climático y agricultura: una revisión de la literatura con énfasis en América Latina") estima que, de continuar las tendencias actuales, el rendimiento del maíz podría disminuir hasta un 10% para 2055 debido a fluctuaciones en precipitaciones y temperaturas. Asimismo, la agricultura contribuye significativamente a las emisiones globales de gases de efecto invernadero, responsables del calentamiento global. Según la FAO, el sector agrícola es responsable de aproximadamente el 17% de estas emisiones, a las que se suma entre un 7% y un 14% derivado de la deforestación y cambios en el uso del suelo. Este doble papel de la agricultura, como víctima y causante del cambio climático, subraya la importancia de adoptar soluciones innovadoras.

El uso de machine learning, como los modelos de regresión lineal múltiple, Random Forest y K-Means aplicados en este estudio, permite predecir variables críticas, clasificar condiciones de producción y agrupar cultivos según sus

características, ofreciendo una base científica para maximizar la producción y minimizar el impacto ambiental. Estas herramientas no solo mejoran la rentabilidad y la seguridad alimentaria, sino que también promueven una agricultura de precisión sostenible. Por

ejemplo, la reducción del uso ineficiente de fertilizantes y agua mediante estas tecnologías puede disminuir los costos operativos y las emisiones asociadas, alineándose con los objetivos de desarrollo sostenible. En un mundo donde el 25-30% de los alimentos producidos se pierde o desperdicia (según datos de la Secretaría de Medio Ambiente y Recursos Naturales de México), optimizar los recursos agrícolas mediante inteligencia artificial se convierte en una estrategia clave para garantizar un futuro alimentario viable y respetuoso con el medio ambiente.

III. MÉTODOLOGIA

Algoritmo de regresión

Objetivo: El objetivo de usar el modelo de la regresión lineal múltiple es modelar la relación entre las features independientes (específicamente en este caso se utilizaron Rainfall_mm, Fertilizer_Used e Irrigation_Used) y una variable dependiente (la cual fue Yield_tons_per_hectare). La finalidad del modelo es dar una predicción de la producción agrícola en función de las variables que anteriormente comentamos, aplicando el método de descenso de gradiente para estimar los parámetros del modelo.

Etapas de la implementación:

Carga y preprocesamiento de los datos: Se cargó el dataset desde un archivo CSV utilizando pandas.

Inspección de los datos: Se inspeccionaron las primeras filas de los datos con df.head(10), lo cual es útil para revisar la estructura del dataset y obtener una visión general de las primeras observaciones.

Normalización de Datos: Variables Independientes: Para asegurar que las diferentes escalas de las variables no afectarán el modelo, las variables independientes fueron normalizadas utilizando MinMaxScaler, lo cual se ajusta al rango [0, 1].

Variable Dependiente: La variable dependiente Yield_tons_per_hectare fue normalizada utilizando StandardScaler, asegurando que tenga una media de 0 y una desviación estándar de 1. Esto mejora la eficiencia y estabilidad del algoritmo de descenso de gradiente.

Definición del Modelo de Regresión Lineal: Se definió la clase RegresionLinealMultiple, que implementa el algoritmo de regresión lineal mediante descenso de gradiente. El modelo comienza con coeficientes aleatorios y ajusta estos coeficientes iterativamente, minimizando la función de costo, que en este caso es el Error Cuadrático Medio (MSE). La optimización se realiza con un aprendizaje basado en el cálculo del gradiente de la función de pérdida.

Entrenamiento del Modelo:

Se entrenó el modelo usando el conjunto de entrenamiento. En cada iteración, el modelo ajustó los parámetros (coeficientes) para reducir el error, es decir, el MSE. Se utilizó un número de iteraciones de 10,000 y una tasa de aprendizaje de 0.05, elegidos tras un proceso de experimentación.

Evaluación del Modelo: El modelo fue evaluado tanto en los datos de entrenamiento como en los de prueba. El MSE se utilizó como métrica de rendimiento para determinar qué tan bien se estaba ajustando el modelo a los datos. Se implementó validación cruzada K-Fold con 5 particiones, lo que permite evaluar la capacidad de generalización del modelo al dividir los datos en múltiples subconjuntos y realizar un entrenamiento y evaluación cruzada. Finalmente, se graficó la evolución del MSE durante las iteraciones para observar cómo el modelo convergía y la relación entre los errores de entrenamiento y prueba.

Visualización: Se presentó una visualización 3D de los datos de prueba y la superficie de regresión ajustada, mostrando la relación entre las variables independientes y la variable dependiente. La superficie de regresión muestra cómo el modelo predice el rendimiento de la cosecha en función de las características normalizadas de lluvia, fertilizantes y riego.

Algoritmo de clasificación

Para el apartado de clasificación se presentó el modelo de Random Forest, esto es gracias que utilizamos una cantidad considerable de datos siendo este modelo más útil cuando se trabajan con muchos datos de clasificación, un punto importante a mencionar es que esta clasificación es una clasificación binaria, dándonos como resultados posibles, True o False, esto es gracias a que vamos a intentar clasificar si un cultivo utilizó irrigación y por otra parte fertilizante de manera separada.

A continuación, mostraremos una tabla comparativa la cual ejemplifica los mejores algoritmos de clasificación:

Algoritmo	Precisión	Velocidad	Desventaja	Ventaja
Árbol de decisión	Media	Alta	Propenso a overfitting con datos ruidosos	Replicabilidad clara, maneja características categóricas
Random forest	Alta	Media	Menos interpretable que un solo árbol	Reduce overfitting, robusto con datos desbalanceados
Naive bayes	Media	Media	Asume independencia entre variables (poco realista)	Rápido, funciona bien con pocos datos
MLP	Alta	Media	Requiere muchos datos y	Captura relaciones no

			potencia computacional	lineales complejas
XGBoosting	Alta	Baja	Requiere ajuste de hiperparámetros	Alta precisión, maneja datos desbalanceados

Dentro de nuestro DataSet no contamos con ruido o datos atípicos, además de esto un punto importante a tratar es que este DataSet contiene 1 millón de datos por lo que para una computadora común y corriente son cantidades de procesamiento.

altos y aunque podemos hacer una minimización en la toma de datos, tomando solo una muestra de los datos podemos perdernos de información importante que más adelante podamos utilizar. Necesitamos un algoritmo que no tenga un costo de procesamiento alto, que tenga una precisión alta y que funcione con muchos datos por lo que el algoritmo de Random Forest es la mejor opción según las circunstancias que tenemos en esta problemática.

Algoritmo de clústers

Según la temática de nuestro proyecto tomamos en cuenta las capacidades de procesamiento que tenemos, tipos de datos a procesar y el resultado que queremos obtener del algoritmo. Para este caso utilizamos el algoritmo de k-Means ya que este nos permite trabajar con una precisión decente según los tipos de datos que tenemos que agrupar y aunque a nivel de complejidad este algoritmo no es muy robusto este destaca en ser de procesamiento rápido aun con cantidades de datos grandes.

Algoritmo	**Precisión**	**Velocidad**	**Desventaja**	**Ventaja**
K-Means	Media	Rápida	- Requiere especificar *k*. - Sensible a outliers e inicialización	- Simple y rápido. - Escala bien con grandes datasets.
DBSCAN	Alta	Moderada	- Difícil elegir parámetros - Lento con datos de alta dimensión.	- No requiere *k* inicial. - Detecta clusters de formas arbitrarias.
Hierárquico	Media	Lenta	Computacionalmente costoso. - Sensible a ruido.	- No requiere *k* inicial. Visualización intuitiva
GMM	Alta	Moderada	- Asume distribución gaussiana. - Lento con muchos datos.	- Modela clusters no esféricos.

K-Means nos permite modificar el parámetro de k y aunque al inicio no podemos garantizar la eficacia de los valores impuestos podemos irlos modificando hasta el punto de encontrar la mejor precisión del modelo según el MSE. Todo esto aunado a que es simple de explicar los resultados de este algoritmo lo hace uno de los mejores para nuestro caso.

IV. RESULTADOS

Algoritmo de regresión

En este algoritmo implementamos la regresión lineal múltiple donde se hicieron diversas pruebas, entre las 2 más resaltadas tenemos en la que usamos como variables independientes a 'Rainfall_mm' y 'Fertilizer_Used' ya que al hacer nuestra correlación resultó en que estas eran de las 2 más significativas para predecir que tanto rendimiento de toneladas por hectárea nos iba a dar cada cultivo (como se ve en el siguiente mapa de calor).

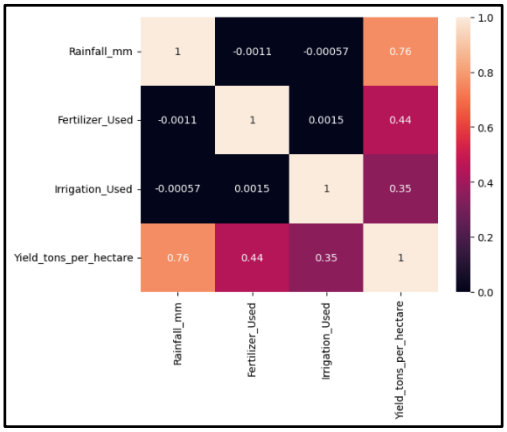

Fig. 1 HeatMap de las variables elegidas

Como variable dependiente, como se explico, pues usamos la de 'Yield_tons_per_hectare'. Entonces al hacer distintas pruebas con estos datos (cambiando solamente el "test_size" ya que al hacer distintas pruebas observe que el mejor "learning_rate" era de 0.05 y el mejor número de iteraciones era de 10000) obtuvimos los siguientes resultados:

%Prueba	**MSE entrenamiento**	**MSE prueba**	**MSE Cross-validation**	**Varianza**
20%	0.21914	0.21976	0.21926	1.00199
30%	0.21935	0.21954	0.21941	1.00354
40%	0.21924	0.21949	0.21934	1.00186

Al analizar los resultados, se observa que la distribución 80-20 produce los mejores resultados, aunque la diferencia con otras configuraciones es mínima. Un aspecto destacado es la proximidad entre los valores del MSE en el conjunto de entrenamiento y el conjunto de prueba, lo que indica que el modelo no presenta ni overfitting ni underfitting. Además, la validación cruzada muestra valores de MSE similares al de la prueba, lo que sugiere que el modelo mantiene su estabilidad cuando se evalúa en diferentes subconjuntos de datos. También es relevante que el MSE de prueba es considerablemente menor que la varianza, lo

que respalda la capacidad predictiva del modelo y su funcionalidad.

Aunque estos resultados son prometedores, se considera que aún pueden mejorarse. Por ello, se decidió incluir una variable adicional, 'Irrigation_Used'. Aunque esta variable no mostró una correlación tan alta como las dos variables independientes previamente utilizadas (como se puede observar en el mapa de calor a continuación), se espera que su inclusión aporte valor adicional al modelo.

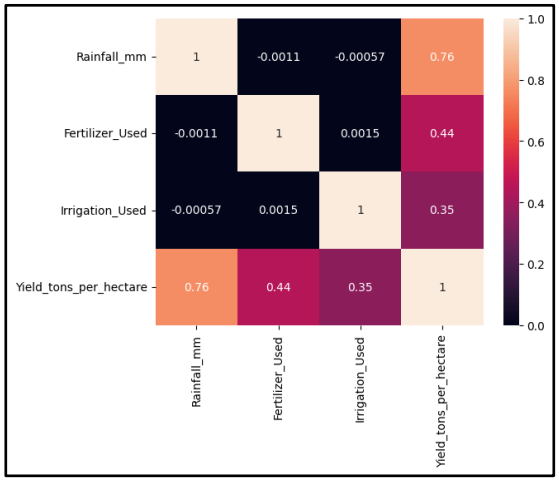

Fig. 2

Igual tenía la suficiente relevancia como para hacer un cambio significativo. Ahora sí, los resultados con las 3 variables independientes son los siguientes:

%Prueba	MSE entrenamiento	MSE prueba	MSE Cross-validation	Varianza
20%	0.09423	0.09432	0.09425	1.00199
30%	0.09428	0.09436	0.09431	1.00354
40%	0.09428	0.09428	0.09428	1.00186

De manera similar a la tabla anterior, se observa que la partición 80-20 continúa arrojando los mejores resultados, aunque la diferencia con otras distribuciones sigue siendo marginal. Sin embargo, en este caso se destaca una mejora general en los indicadores de desempeño. Al realizar las mismas comparaciones consideradas en el análisis previo, se confirma que el modelo mantiene un equilibrio adecuado, evitando tanto el sobreajuste (overfitting) como el subajuste (underfitting). Asimismo, los valores obtenidos en la validación cruzada son consistentes con los del conjunto de prueba, lo cual indica que el modelo presenta una buena estabilidad al ser evaluado en distintos subconjuntos de datos. Finalmente, el hecho de que el error cuadrático medio (MSE) de prueba se mantenga por debajo de la varianza respalda que el modelo posee capacidad predictiva y resulta funcional en su aplicación.

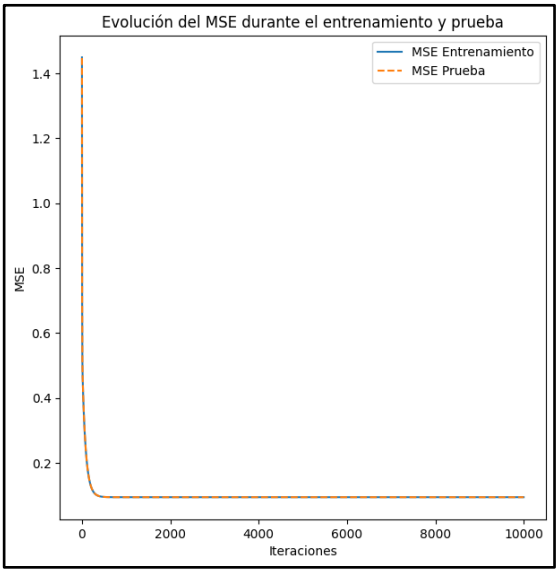

Fig. 3 Evolución del MSE

Tal como se mencionó anteriormente, la gráfica del error cuadrático medio (MSE) respalda los hallazgos previos, ya que al observar ambas curvas —correspondientes al conjunto de entrenamiento y al de prueba— se evidencia una tendencia estable. Esta estabilidad en las líneas indica que el modelo no presenta señales de sobreajuste (overfitting) ni subajuste (underfitting), lo que refuerza la solidez de su desempeño general.

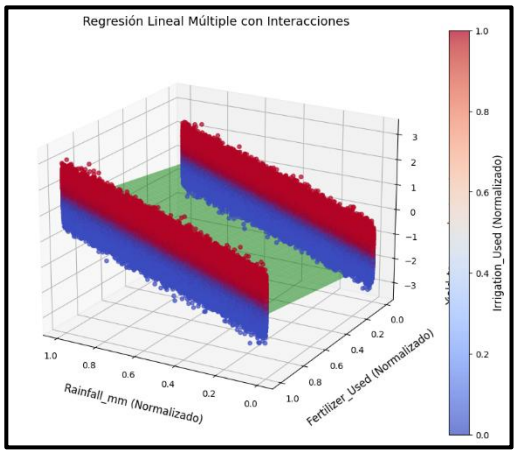

Fig. 4 Representación 3D del modelo

También se incluye una segunda gráfica que permite visualizar claramente la relación positiva entre la cantidad de lluvia, el uso de fertilizante y la aplicación de riego, mostrando cómo la producción aumenta significativamente cuando estos tres factores se combinan de manera adecuada.

Conclusiones

En conclusión, a partir de las pruebas realizadas y el análisis de los resultados obtenidos, se puede deducir que cada cultivo alcanza su máximo potencial productivo

cuando se cumplen ciertas condiciones clave. En primer lugar, la variable más determinante es la cantidad de lluvia, la cual influye de manera decisiva en el rendimiento. En segundo lugar, la aplicación de fertilizante contribuye significativamente al desarrollo óptimo del cultivo. Finalmente, aunque en menor medida, el riego constante también tiene un impacto positivo al asegurar que las plantas dispongan de suficiente agua durante todo el ciclo productivo. Por tanto, una combinación adecuada de estos tres factores —lluvia, fertilización y riego— es fundamental para maximizar la producción agrícola.

Algoritmo de clasificación

En el apartado de clasificación, se obtuvo una precisión del 81 % tanto en la predicción del uso de fertilizante como en la del uso de irrigación. A continuación, se presenta el código utilizado para alcanzar estos resultados.

Se llevaron a cabo diversas pruebas con el objetivo de identificar la configuración óptima para el entrenamiento del algoritmo Random Forest, ajustando parámetros clave como el tamaño del conjunto de validación, el número de árboles (n_estimators) y la profundidad máxima del modelo (max_depth). Estas pruebas permitieron evaluar el impacto de cada parámetro en el desempeño del clasificador, optimizando así su capacidad para predecir correctamente el uso de insumos agrícolas.

Num. Validación	Num. árboles	Profundidad max.	Precisión
30%	30	10	81%
40%	30	10	81%
20%	30	10	81%

Al modificar la proporción del conjunto de validación, se observó que los valores de precisión, *recall* y F1-score se mantuvieron prácticamente constantes. Esto sugiere que la cantidad de datos destinados a la validación no constituye un factor determinante en la mejora o deterioro del rendimiento del algoritmo. Por lo tanto, dentro del rango evaluado, el modelo muestra una robustez considerable frente a variaciones en la partición de los datos, lo cual es indicativo de una buena generalización.

Num. Validación	Num. árboles	Profundidad max.	Precisión
30%	60	10	81%
30%	30	10	81%
30%	2	10	80%

En cuanto al número de árboles que conforman el modelo de Random Forest, se realizaron pruebas para identificar el equilibrio entre precisión y costo computacional. Los resultados indican que, si bien un modelo con más de dos árboles mejora el rendimiento, no es necesario un número excesivo para alcanzar una buena precisión. En este caso, se observó que un rango de 3 a 5 árboles ofrece el mejor compromiso entre desempeño y eficiencia computacional, proporcionando resultados óptimos sin incurrir en un costo innecesario de recursos.

Num. Validación	Num. árboles	Profundidad max.	Precisión
30%	5	40	78%
30%	5	10	81%
30%	5	2	66%

En este último apartado, se observó que la profundidad de los árboles sí generó un impacto significativo en la precisión del modelo. Específicamente, se detectó que, al incrementar la profundidad, la precisión tiende a disminuir ligeramente, posiblemente debido al sobreajuste. Sin embargo, al reducir excesivamente la profundidad, la precisión disminuye de manera drástica, lo que indica que el modelo pierde capacidad para capturar patrones relevantes en los datos.

Conclusiones

A partir de los experimentos realizados, se puede concluir que la mejor configuración del algoritmo Random Forest para clasificar si un cultivo utiliza fertilizante o irrigación es la siguiente:

- Porcentaje de validación: 30 % (variable con poca influencia en la precisión del modelo).
- Número de árboles: 5 (equilibrio óptimo entre rendimiento y costo computacional).
- Profundidad máxima: 10 (permite capturar patrones relevantes sin caer en sobreajuste).

Esta configuración proporciona un modelo con buena capacidad predictiva y eficiencia computacional adecuada para las tareas de clasificación planteadas.

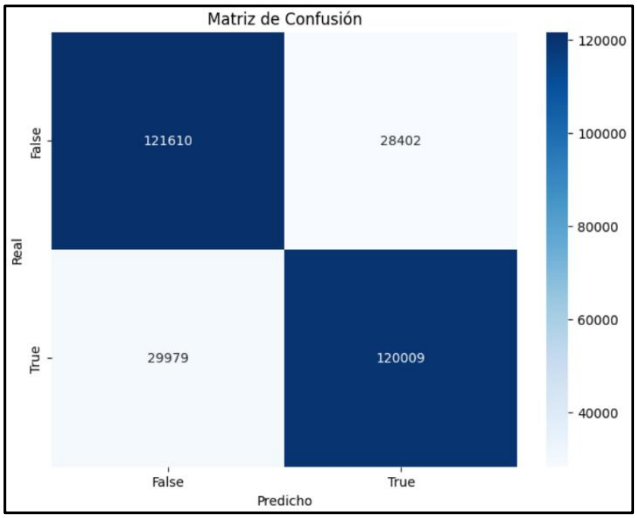

Fig. 5 matriz de confusión

	precision	recall	f1-score	support
False	0.80	0.81	0.81	150012
True	0.81	0.80	0.80	149988
accuracy			0.81	300000
macro avg	0.81	0.81	0.81	300000
weighted avg	0.81	0.81	0.81	300000

Fig.6 Métricas de rendimiento

Algoritmo de Clústeres

Para esta sección, correspondiente a un algoritmo no supervisado, se implementó el método de K-means con el objetivo de agrupar los datos en clústeres representativos. Una característica fundamental de este algoritmo es la necesidad de definir previamente la cantidad de clústeres (*K*) que se desea utilizar. Para determinar un valor óptimo de *K*, se aplicó el conocido método del codo (*elbow method*).

Este método consiste en graficar la suma de los errores cuadráticos dentro de los clústeres (WCSS, por sus siglas en inglés) para distintos valores de *K*. Al observar la gráfica resultante, se identifica el punto en el que la disminución del WCSS comienza a estabilizarse, formando un "codo". Este punto indica el número óptimo de clústeres, ya que a partir de ahí agregar más clústeres no mejora significativamente la compactación del modelo. La siguiente gráfica muestra los resultados obtenidos tras entrenar el modelo:

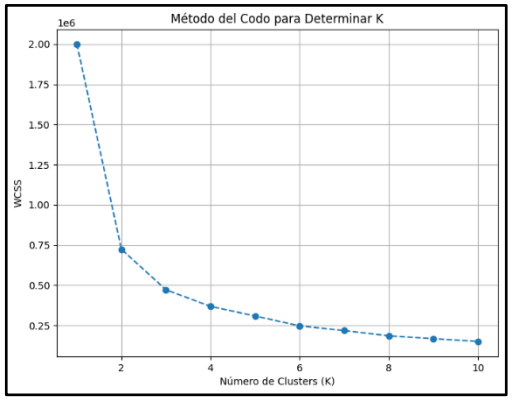

Fig. 7 Gráfica del método del codo

El método del codo se basa en identificar el punto en el que la disminución del WCSS (*Within-Cluster Sum of Squares*) comienza a desacelerarse de forma significativa. Este punto marca el número óptimo de clústeres, ya que a partir de ahí, agregar más clústeres no mejora sustancialmente la calidad del agrupamiento.

Con base en la gráfica obtenida, se determinó que el valor óptimo de *K* es 4, ya que es en este punto donde se forma el "codo" característico. A continuación, se aplicó el algoritmo K-Means utilizando como variables de entrada 'Rainfall_mm' y 'Yield_tons_per_hectare'. Con estos datos, se construyó la gráfica de clústeres, donde cada grupo se representa con un color distinto y su respectivo centroide está marcado, como se muestra a continuación:

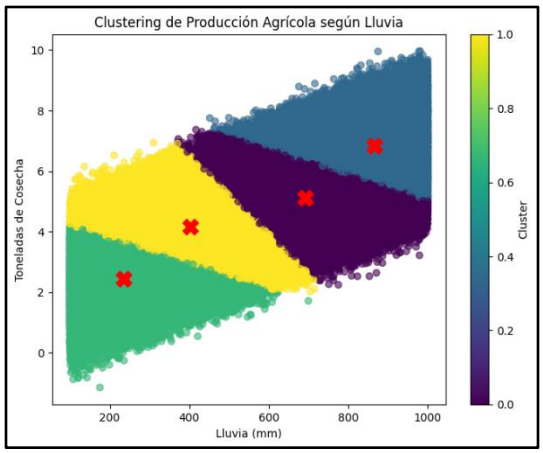

Fig. 8 Clustering

Al observar la representación gráfica inicial del clustering, notamos que los resultados no eran completamente satisfactorios en términos de separación y estructura de los clústeres. Por esta razón, decidimos mejorar el preprocesamiento de los datos implementando dos técnicas clave:

1. ColumnTransformer: Esta herramienta nos permitió aplicar transformaciones específicas a determinadas columnas del conjunto de datos, como la estandarización de variables numéricas, sin afectar al resto de la estructura del dataset.

2. PCA (Análisis de Componentes Principales): Utilizamos PCA con el objetivo de reducir la dimensionalidad del conjunto de datos a dos componentes principales, permitiendo así una visualización más clara en 2D del agrupamiento generado por K-means. Esta técnica nos permite preservar la mayor cantidad de varianza posible de los datos originales, facilitando una interpretación más significativa del clustering.

Una vez transformados y reducidos los datos, se volvió a aplicar el método del codo, ya que la estructura de los datos ha cambiado y, con ello, también puede variar el número óptimo de clústeres. La nueva gráfica del WCSS en función del número de clústeres se muestra a continuación:

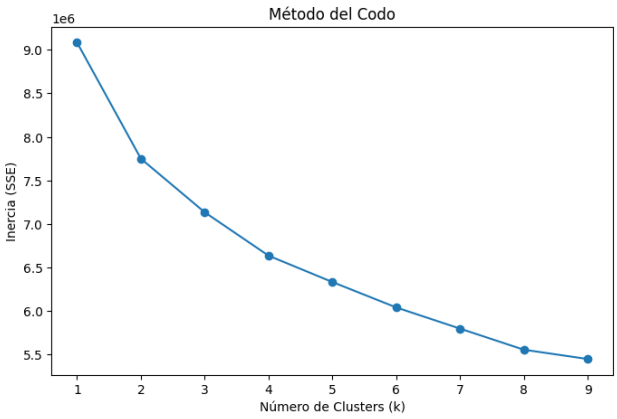

Fig.9 Método del codo actualizado

Ahora en esta gráfica observamos que tiene diferencias claras con la primera gráfica, pero en este caso podemos deducir que, al igual que la anterior gráfica, también vamos a usar 4 clusters.

Gracias a estos procesos, logramos crear nuestra gráfica final de clustering:

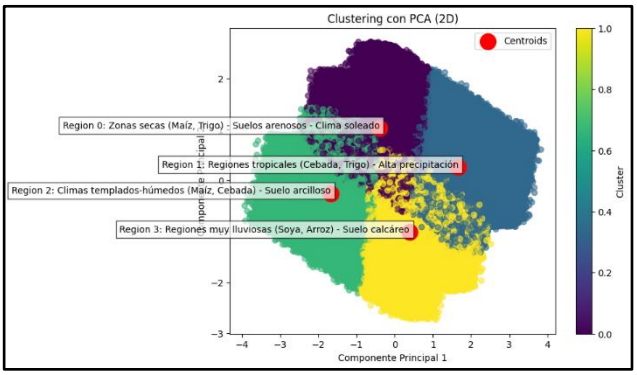

Fig.10 Clustering con PCA

Conclusiones

Al analizar la gráfica final, podemos concluir que el algoritmo de K-Means funcionó de manera efectiva. Los centroides de los clústeres están claramente posicionados, y las regiones de cada clúster son fácilmente distinguibles, lo que indica que el modelo logró agrupar los datos de manera coherente. Cada clúster refleja las relaciones subyacentes dentro de los datos.

Por ejemplo, en las regiones 0 y 2 (como se observa en la gráfica a continuación), podemos identificar que ambas presentan baja precipitación. Sin embargo, en el clúster 0 hay algunos valores atípicos con precipitaciones más altas, mientras que en el clúster 2 los valores de precipitación son más homogéneos, sin tanta variabilidad. Por otro lado, los clústeres 1 y 3 comparten la característica de tener alta precipitación, lo que sugiere que estos grupos de cultivos experimentan condiciones climáticas similares.

Este análisis nos ayuda a entender cómo las distintas variables, como la precipitación, pueden influir en la clasificación y agrupamiento de los cultivos, proporcionando información valiosa para tomar decisiones en la gestión agrícola.

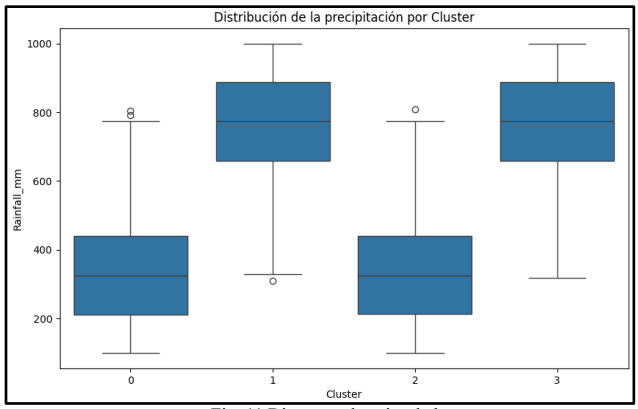

Fig. 11 Diagrama de cajas de la precipitacion

A partir de la interpretación detallada de las características principales por región, podemos obtener una comprensión más clara de las divisiones de los clústeres generados:

1. Interpretación de Tipos de Suelo por Clúster:
    - Región 0: Predomina la presencia de Loam y Sandy, suelos con buen drenaje que permiten una rápida evacuación del agua, lo cual es compatible con su baja precipitación.
    - Región 1: Tiene una mayor presencia de Peaty, un suelo con alta capacidad de retención de agua, lo cual concuerda con su alta precipitación.
    - Región 2: Se encuentra dominada por Clay, suelos que también retienen agua pero pueden ser difíciles para ciertas plantas debido a su alta densidad.
    - Región 3: Se caracteriza por Chalky y Sandy, suelos que suelen ser más secos y permiten menos retención de agua.

Conclusión: Las regiones 1 y 2 representan suelos con mayor retención de agua, lo que las hace más adecuadas para cultivos que requieren humedad constante, mientras que las regiones 0 y 3 están asociadas con suelos más secos o con mejor drenaje, lo que las hace más adecuadas para cultivos resistentes a la sequía.

2. Interpretación de Tipos de Cultivo por Clúster:
    - Región 0: Predomina el Maíz y el Trigo, que son cultivos que pueden adaptarse a suelos con drenaje moderado.

- Región 1: Se observa una mayor cantidad de Cebada y Trigo, cultivos que requieren más agua, lo que es coherente con su mayor precipitación.
- Región 2: Se encuentran cultivos como Cebada (Barley) y Maíz, lo que sugiere una combinación de climas templados y secos.
- Región 3: Hay una predominancia de Soya y Maíz, cultivos que toleran suelos más secos.

Conclusión: Las regiones 1 y 2 representan áreas de cultivo con una mayor necesidad de agua, mientras que las regiones 0 y 3 son más adecuadas para cultivos que pueden resistir períodos de sequía.

3. Interpretación de Condiciones Climáticas por Clúster:
   - Región 0: Experimenta más días soleados, lo que puede explicar su baja precipitación.
   - Región 1: Tiene una combinación de días soleados y lluviosos, indicando un clima tropical.
   - Región 2: Se caracteriza por más días nublados y lluviosos, lo que sugiere un clima templado-húmedo.
   - Región 3: Se observa un mayor número de días lluviosos, lo que indica un clima con precipitación constante.

Conclusión: Las regiones 1 y 2 tienen condiciones climáticas más húmedas y variables, favoreciendo los cultivos que requieren alta humedad, mientras que las regiones 0 y 3 tienen un clima más seco o soleado, lo que favorece cultivos adaptados a condiciones más áridas.

En resumen, la interpretación de los clústeres generados nos ha permitido segmentar las áreas agrícolas de acuerdo con sus condiciones climáticas, de suelo y de cultivo. Esto proporciona una visión más clara sobre las necesidades hídricas y de tipo de suelo de las diferentes regiones, lo que puede ser utilizado para tomar decisiones más informadas sobre la planificación agrícola y la gestión de los recursos en función de las condiciones específicas de cada zona.

Región	Clima principal	Suelo más común	Principales cultivos	Precipitación
0	Seco, soleado	Arenoso (Sandy), Franco (Loam)	Maíz y trigo	Baja
1	Tropical, alta humedad	Turba (Peaty), Franco (Loam)	Cebada	Alta
2	Templado-Húmedo, Nublado	Arcilloso (Clay), Arenoso (Sandy)	Maíz y cebada	Media
3	Lluvioso	Calcáreo (Chalky), Arenoso (Sandy)	Soya y arroz	Muy alta

V. DISCUSIÓN Y CONCLUSIONES

Para concluir, es importante resaltar varios aspectos clave que deben tenerse en cuenta para mejorar la precisión de las predicciones agrícolas en el futuro:

Uno de los aspectos fundamentales es la necesidad de contar con una base de datos lo más completa y diversa posible. Es esencial incluir información sobre distintos tipos de cultivos en una variedad de entornos y condiciones, considerando también factores como el riego y el uso de fertilizantes. Durante este proyecto, la dificultad principal fue la falta de bases de datos confiables, ya que la base utilizada presentaba inconsistencias. Esto nos lleva a reconocer que la fiabilidad de los datos impacta directamente en la calidad de los resultados obtenidos.

Otro desafío relevante fue la gestión de grandes volúmenes de datos. La base de datos utilizada en este estudio contenía una gran cantidad de información, lo que generó dificultades a la hora de ejecutar los algoritmos debido a los recursos limitados de procesamiento. A pesar de que herramientas como Google Colab son útiles para manejar grandes conjuntos de datos, la ejecución de los algoritmos fue un proceso lento, con tiempos de espera considerables. Este desafío fue aún mayor debido a la necesidad de aplicar técnicas de reducción de dimensionalidad para optimizar el procesamiento.

En cuanto a la fiabilidad de los resultados, aunque los modelos generaron información valiosa sobre el comportamiento de los cultivos bajo distintas condiciones, no se puede garantizar que los algoritmos sean completamente eficientes y confiables para la toma de decisiones agrícolas. Esto se debe a la falta de certeza sobre la fiabilidad de los datos utilizados para entrenar los modelos, lo que afecta la validez de las predicciones.

Asimismo, se identificó una dificultad significativa en la recolección y estandarización de datos agrícolas a nivel global. Si bien algunos países cuentan con bases de datos más completas, en otros, como México, la información

disponible es escasa, desactualizada, dispersa y difícil de acceder. Esta falta de estandarización y organización en los datos dificulta la creación de bases de datos sólidas que puedan utilizarse para entrenar modelos predictivos de manera efectiva.

A pesar de estos retos, si se pudiera acceder a datos más específicos, confiables y actualizados, y contar con mayores recursos de procesamiento, sería posible desarrollar modelos más precisos y confiables. Esto permitiría a los agricultores tomar decisiones más informadas y mejorar la eficiencia de sus prácticas agrícolas, beneficiando así al sector agrícola en su conjunto.

En resumen, aunque los resultados obtenidos hasta ahora proporcionan una visión útil sobre los comportamientos de los cultivos, la precisión y confiabilidad de los modelos dependen en gran medida de la calidad de los datos utilizados. Superar las limitaciones actuales en cuanto a la recolección de datos y optimizar los recursos de procesamiento disponibles permitirá desarrollar modelos más robustos que podrían transformar la gestión agrícola, contribuyendo a una agricultura más eficiente y sostenible.

VI. REFERENCIAS

[1]K. Kirasich, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," *Data Science Review*, 2018.

[2]Z. Keita, "Clasificación en machine learning: Introducción," *DataCamp*, 2024.

[3]A. Phadnis, S. Panchal, R. Jadhav, R. Rajdeep, and D. Patil, "Implementation of prediction of crop using SVM algorithm," *International Journal for Research in Applied Science & Engineering Technology*, vol. 11, no. 5, pp. 3812–3816, 2023.

[4]A. Gutiérrez, I. Villarino, D. A. Nafría, N. Garrido, I. Abio, M. Fernández, and L. Rodríguez, "Nuevo sistema de predicción de cosecha de cereales de Castilla y León," Instituto Tecnológico Agrario de Castilla y León y Delegación de AEMET en Castilla y León.

[5]M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012033, 2022, doi: 10.1088/1742-6596/2161/1/012033.

[6]Saturdays.AI, "Inteligencia artificial para la predicción de cosechas," 9 de marzo de 2022.

[7]E. Elbasi et al., "Crop prediction model using machine learning algorithms," *Applied Sciences*, vol. 13, no. 16, p. 9288, 2023, doi: 10.3390/app13169288.

[8]K. Nischitha, D. Vishwakarma, A. Ashwini, N. Mahendra, and M. R. Manjuraju, "Crop prediction using machine learning approaches," *International Journal of Engineering Research & Technology*, vol. 9, no. 8, pp. 23–26, 2020.

[9]Organización de las Naciones Unidas para la Alimentación y la Agricultura, *The future of food and agriculture: Trends and challenges*, Roma: FAO, 2018.

[10]P. Ramírez and C. Villegas, "Cambio climático y agricultura: Una revisión de la literatura con énfasis en América Latina," *Revista de Estudios Latinoamericanos*, vol. 12, no. 2, pp. 45–67, 2020.

[11]Secretaría de Medio Ambiente y Recursos Naturales, "Día Internacional de la Concientización sobre la Pérdida y el Desperdicio de Alimentos," 28 de septiembre de 2022. Enlace

[12]Programa de las Naciones Unidas para el Medio Ambiente, *Food Waste Index Report 2021*, Nairobi: PNUMA, 2021.