Desarrollo de modelos de machine learning para la predicción en la recomendación de cultivos mediante datos medioambientales y físico-químicos para distintas regiones de la India

Ana Inés Ortega[1]

[1] *IT Academy, Data Science, Barcelona Activa. Barcelona, España.*

Resumen

El sector agrícola de la India es fundamental en su economía, aunque aún enfrenta desafíos como la baja productividad, el cambio climático y la falta de acceso a la financiación. La adopción de tecnologías como la inteligencia artificial, el big data y el machine learning pueden abordar estos desafíos y mejorar la eficiencia y productividad del sector. En este sentido, se busca crear un modelo de machine learning basado en datos recopilados de múltiples zonas de la India que corresponden a características medioambientales y físico-químicas del suelo, para predecir el cultivo óptimo a implantar. El presente trabajo se enfoca en crear este modelo, el cual procesa los datos ingresados mediante la aplicación de modelos de aprendizaje supervisado y no supervisado.

Palabras Claves: *Clasificación Multiclase, Clustering, Machine Learning, Aprendizaje Supervisado, Aprendizaje No Supervisado, Algoritmos, Predicciones, DataSet, Acurracy, F1, F1 (macro), Cultivos, Factores ambientales, Factores físico - químcos del suelo.*

1. Introducción

El sector agrícola de la India desempeña un papel fundamental en la economía del país, contribuyendo con el 18 % del producto interno bruto (PIB) y el 40 % del producto interno neto (PND) rural total [1]. Sin embargo, aún enfrenta una serie de desafíos, como la baja productividad, el cambio climático y la falta de acceso a la financiación. A su vez, la creciente densidad poblacional mundial y la necesidad de producir una mayor cantidad de alimentos han llevado a la adopción de tecnologías como la inteligencia artificial, el big data y el machine learning en la agricultura. Estas tecnologías tienen el potencial de abordar los desafíos del sector y mejorar su eficiencia y productividad [2].
En este sentido, el presente trabajo tiene como objetivo crear un modelo de machine learning basado en datos recopilados de múltiples zonas de la India que corresponden a características medioambientales y físico-químicas del suelo, para predecir el cultivo óptimo a implantar. El análisis de grandes cantidades de datos permitirá a los productores tomar mejores decisiones y mejorar la eficiencia del sector agrícola, lo que tendrá un impacto positivo en la economía en general.

2. Antecedentes

En el campo de la agricultura de precisión, la adquisición de datos del ciclo del cultivo, la evaluación del estado del suelo y la monitorización de los factores medioambientales son de gran importancia para la detección de enfermedades de las plantas, la predicción de fechas de siembras y cosecha, entre otras. Para lograr estos objetivos, se requieren datos e información de alta precisión y métodos eficaces para realizar una evaluación correcta del estado del cultivo y del suelo. Esto implica tener en cuenta variables ambientales, elementos del suelo y pH.
Praguer (2019) realiza una investigación aplicada basada en la modelación de cultivos y ganado, lo que permite obtener información a partir de la utilización de modelos de regresión y máquinas de vectores de soporte. Esto permite comprender las diferentes condiciones de adaptación tanto del sembradío como de la ubicación del ganado. Estos comportamientos se ven influenciados por factores medio ambientales y pronósticos estacionales que pueden ser analizados mediante lenguajes usados en inteligencia artificial como python y R[3].
En el presente trabajo, se busca desarrollar modelos de machine learning para la predicción en la re-

comendación de cultivos en diferentes zonas de la India.
Se procesará el ingreso de datos ambientales y físico – químicos para predecir si el suelo donde se va a realizar la siembra es el indicado para un determinado cultivo mediante la aplicación de modelos de aprendizaje supervisado y no supervisado, con el fin de contribuir al agricultor para que ejecute de manera correcta el manejo de su cultivo a mediano y largo plazo.
En conclusión, la aplicación de la inteligencia artificial en el campo agrícola para la predicción de siembras y recomendación de cultivos es un tema en constante evolución y desarrollo. Los trabajos previos en este campo han utilizado técnicas de adquisición de datos, simulación de cultivos, modelación y aprendizaje automático para lograr resultados prometedores en la predicción de rendimientos, la selección de cultivos y la gestión de la fertilización y el riego. Sin embargo, todavía existen desafíos importantes a superar, como la falta de datos precisos y completos, la necesidad de técnicas más sofisticadas para la adquisición y procesamiento de datos, y la integración efectiva de los sistemas de inteligencia artificial con los sistemas de información agrícola existentes. A medida que se avanza en la investigación y el desarrollo en este campo, es probable que se logren avances significativos en la mejora de la eficiencia y la sostenibilidad de la agricultura.

3. Metodología

3.1. DataSet

Se obtuvo el conjunto de datos (Dataset) a través de la página web "www.kaggle.com"[4], el cual contiene información sobre factores medioambientales y físico-químicos del suelo de diferentes regiones en la India, en las cuales se han implantado ciertos cultivos. El Dataset consta de un total de 2200 muestras y 8 columnas.

3.2. Preparación y análisis de datos

Se identificaron 8 características (columnas) diferentes dentro del DataSet. El conjunto de datos estaba limpio, sin valores nulos ni caracteres especiales (NaN). Posteriormente, se corrigió el tipo de dato (Dtype) del DataSet, ya que los valores de la columna Çultivo"se encontraban como objeto, se los transformó a formato categórico. Como el target del DataSet no requería transformaciones, no se realizaron cambios.

3.3. Transformación de datos

Luego del análisis realizado a cada columna (ver fig. 1), se decidió normalizar todo el conjunto de datos en función de lo observado. Para realizar la transformación se decidió utilizar pipeline con el fin de automatizar el flujo de trabajo en el desarrollo de modelos. Se los transformó de la siguiente manera:

- Columnas con outliers: RobustScaler
- Columnas sin outliers: MinMaxScaler

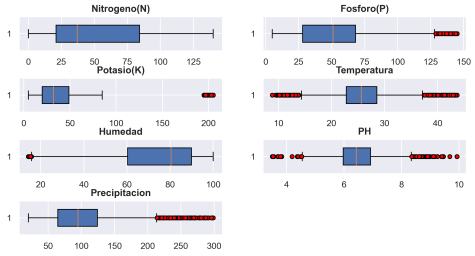

Figura 1: Análisis de las características de cada columna de datos.

3.4. Creación de modelos

Para predecir el cultivo óptimo a implantar, se implementaron modelos de aprendizaje supervisado y no supervisado. Los modelos de aprendizaje supervisado utilizados fueron algoritmos de clasificación, en este caso en particular se trató de una Clasificación Multiclase, mientras que los modelos de aprendizaje no supervisado fueron algoritmos de Clustering para buscar patrones. Debido a la variedad de algoritmos de machine learning dentro de la librería Scikit-Learn [5], se seleccionaron algunos modelos en función del tipo de problema, los datos y recursos disponibles, así como del rendimiento deseado. Finalmente, se eligieron los siguientes modelos, en función de las mejores métricas (F1 macro) obtenidas:

- Gaussian Naive Bayes (GaussianNB).
- Random Forest Classifier.
- Gradient Boosting Classifier.

Para cada modelo seleccionado, se determinaron diferentes hiperparámetros, con los cuales se realizó un GridSearch para ajustar los modelos y evaluar su puntaje, con el fin de obtener un F1 macro, que es una combinación armónica de la precisión y la exhaustividad (recall), así como un Accuracy, que es la proporción de instancias clasificadas correctamente sobre el total de instancias, lo más cercano a 1 posible.
Posteriormente, se llevó a cabo una validación cruzada (Cross Validation) con el total de los datos,

y finalmente se ajustó el modelo para obtener las métricas de análisis planteadas (F1 macro y Accuracy).
Finalmente, se realizó un análisis de Clustering utilizando diferentes métodos de evaluación, incluyendo el método de la Silueta y del Codo, para identificar patrones en todo el conjunto de datos. Además, se aplicó PCA para reducir la dimensionalidad de los datos y mejorar la precisión del Clustering.

3.5. Resultados

La Tabla 1 muestra los resultados de cada modelo junto con la transformación de datos correspondiente y los mejores hiperparámetros de cada uno. Se observa que los algoritmos Gaussian Naive Bayes (GaussianNB) y Random Forest Classifier alcanzaron las mejores métricas utilizando los siguientes hiperparámetros: 'priors': None, 'var smoothing': 1e-09 para GaussianNB y 'criterion': 'gini', 'max depth': None, 'class weight': None, 'max features': 'auto', 'min samples leaf': 1, 'min samples split': 5, 'n estimators': 50, 'random state': 42 para Random Forest Classifier. Estos modelos ofrecen un Acurracy de 1 y un F1 macro de 0.995.

MODELO	F1 MACRO	ACURRACY
GaussianNB	0.995	1.00
RandomForest Classifier	0.995	1.00
Gradient Boosting Classifier	0.993	0.991

Cuadro 1: Métricas finales de los modelos elegidos

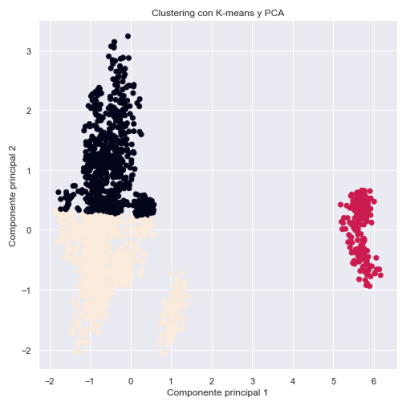

Figura 2: Resultado clusters K-means y PCA

Además, los resultados del análisis de Clustering revelaron la presencia de diferentes grupos en los datos. Tanto el método de la Silueta como el del Codo sugieren que el número óptimo de clusters es de 3 (ver Fig.2) También se identificaron como características más importantes, que contribuyen a la formación de cada cluster, el fósforo y el pH del suelo (ver Fig. 3).

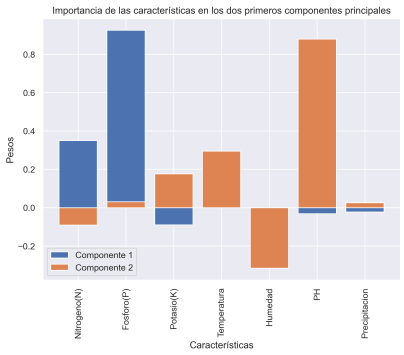

Figura 3: Importancia de las características en los componentes principales.

En conjunto, estos hallazgos proporcionan una comprensión más profunda de la relación entre los factores medioambientales y físico-químicos en los cultivos de la región. Estos resultados pueden ser útiles para mejorar la producción y la eficiencia de los cultivos en el futuro.

4. Conclusiones

En resumen, el sector agrícola de la India enfrenta varios desafíos, pero la adopción de tecnologías como la Inteligencia Artificial, el Big Data y el Machine Learning pueden mejorar su eficiencia y productividad. En este estudio, se utilizó un conjunto de datos que contenía factores medioambientales y físico-químicos de 22 cultivos diferentes para crear un modelo de machine learning capaz de predecir el cultivo óptimo a implantar en una determinada región.
La metodología empleada incluyó la identificación y transformación de características, así como la creación de modelos de aprendizaje supervisado y no supervisado. Los resultados obtenidos fueron prometedores y sugieren que el modelo propuesto puede ser una herramienta útil para los productores agrícolas al tomar decisiones informadas sobre la elección de cultivos.
Si bien existen desafíos importantes a superar

en la implementación de la inteligencia artificial en la agricultura, como la falta de datos precisos y completos y la necesidad de técnicas más sofisticadas para la adquisición y procesamiento de datos, es probable que se logren avances significativos en la mejora de la eficiencia y la sostenibilidad de la agricultura a medida que se avanza en la investigación y el desarrollo en este campo.

En definitiva, la aplicación de la inteligencia artificial en el sector agrícola tiene el potencial de transformar la forma en que se produce y maneja la agricultura en la India y en todo el mundo, contribuyendo a mejorar la eficiencia, productividad y sostenibilidad del sector. Por lo tanto, se sugiere continuar con la investigación buscando nuevas características que aporten valor a la misma y sean de utilidad práctica para los agricultores y los productores buscando optimizar su producción y el manejo de sus cultivos para maximizar su rendimiento.

Referencias

[1] Invest India. Agriculture & Forestry. Recuperado de https://www.investindia.gov.in/es-es/sector/agriculture-forestry

[2] Express Computer. Indian Agriculture Goes Hi-tech with new technologies like AI, ML and IoT. Recuperado de https://www.expresscomputer.in/features/indian-agriculture-goes-hi-tech-with-new-technologies-like-ai-ml-and-iot/45432/

[3] Coppiano Marín, A. D., & Herrera Vargas, C. J. (2022). Desarrollo de aplicativo web basado en máquinas de vectores de soporte(SVM) de aprendizaje supervisado para la predicción en la recomendación de cultivos mediante datos ambientales para fincas agroecológicas del cantón La Maná, provincia del Cotopaxi. Trabajo de titulación, Universidad Técnica de Cotopaxi, Extensión La Maná.

[4] Kaggle. Agricultural Crop Dataset. Recuperado de https://www.kaggle.com/datasets/agriinnovate/agricultural-crop-dataset

[5] F. P. et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, no. 1, pp. 2825–2830, 2011.