APLICACIÓN DEL MACHINE LEARNING EN AGRICULTURA DE PRECISIÓN
APPLICATION OF MACHINE LEARNING IN PRECISION AGRICULTURE

Article *in* Revista CINTEX · December 2020

DOI: 10.33131/24222208.356

CITATION
1

READS
446

2 authors, including:

Carlos Alejandro Ramirez Gomez
Servicio Nacional de Aprendizaje SENA
11 PUBLICATIONS **19** CITATIONS

SEE PROFILE

Revista Cintex | Vol25(2) | pp. 14-27 | julio-dic. | 2020 | ISSN: 2422-2208 | Medellín–Colombia

APLICACIÓN DEL MACHINE LEARNING EN AGRICULTURA DE PRECISIÓN

APPLICATION OF MACHINE LEARNING IN PRECISION AGRICULTURE

Carlos Alejandro Ramírez Gómez
MsC, Ingeniero Electricista, Centro de Automatización Industrial, SENA, Regional Caldas. Instructor Sennova. Grupo de investigación Electrónica, Automatización y Energías Renovables EAYER, Manizales, Colombia
alramiezgo@sena.edu.co

(Recibido el 02-11-2020. Aprobado el 03-12-2020)

Resumen – El presente artículo propone un modelo de Machine Learning para predecir el estado de la cosecha a partir de información de consumo de pesticidas y otras variables del cultivo, para lo cual se sigue la metodología de machine Learning, la cual consiste en cuatro pasos que son: Preprocesamiento y análisis de la información, separación de los datos de entrenamiento, test y validación; selección de los modelos, y evaluación de los hiperparámetros del modelo a partir de una métrica. Para eso se proponen cinco modelos de clasificación, para la evaluación se toma como métrica Accuracy score, como resultado final se obtienen los valores de los hiperparámetros correspondientes a los Modelos y se selecciona uno de ellos.

Palabras clave: Machine Learning, Agricultura de precisión, KNN, Métrica de precisión, árbol de decisión.

Abstract – This article proposes a Machine Learning model to predict the state of the harvest from information on the consumption of pesticides and other crop variables. A machine learning methodology is followed, which consists of four steps. At first, a stage of preprocessing and analysis of information, and separation of training, test, and validation data. The final stages include the selection of models and evaluation of hyperparameters of the model from a metric. For this, five classification models are proposed, and the accuracy score is taken as a metric for evaluation. As a result, the hyperparameters for every model are obtained, and the best-performing model is selected.

Keywords: Machine Learning, Smart Agriculture, KNN, Accuracy Score, Decision Tree.

1 INTRODUCCIÓN

La agricultura tradicional se ve afectada por factores ambientales externos, los cuales reducen el rendimiento y la calidad de los cultivos [1], uno los factores que acentúa el impacto es el cambio climático. Los análisis indican que para el 2050 es probable que se presenten aumentos significativos de la temperatura, precipitación más errática y mayor prevalencia de plagas y enfermedades [2]. En Colombia el impacto seria significativo debido a que el sector agropecuario en Colombia es responsable de más de una décima parte del PIB del país y fuente de empleo para más de una quinta parte de su población. [3]

En el caso de las plagas y enfermedades va en aumento, y es probable que la situación empeore con la agudización del cambio climático [4]. Los cultivos actualmente afectados incluyen las musáceas (bananos, plátanos) en áreas por encima de los 500 msnm, el café en áreas por encima de los 1500 msnm [5], la papa en áreas por debajo de los 2500 msnm, así como el cacao, el maíz y la yuca.[6], lo cual genera un uso intensivo de pesticidas que puede representar altos costos económicos para los pequeños productores y costos a largo plazo para el agroecosistema [6].

Con el uso de los pesticidas se debe tener mucho cuidado ya que la dosis correcta permite hacer un control de las plagas y enfermedades, pero si se agrega más de lo requerido, se puede estropear toda la cosecha. Como una solución a este problema ha tomado fuerza la implementación de agricultura inteligente, la cual busca que la agricultura sea más eficiente y eficaz con la ayuda de algoritmos de alta precisión, para lo cual usa Machine Learning [7]. Este ha surgido, junto con tecnologías de big data y computación de alto rendimiento, para crear nuevas oportunidades para desentrañar, cuantificar y comprender los procesos intensivos de datos en entornos operativos agrícolas.[6].

Citar como:
C. Ramírez. "APLICACIÓN DEL MACHINE LEARNING EN AGRICULTURA DE PRECISIÓN" Revista CINTEX, Vol. 25(2), pp. 14-27. 2020.

En este artículo se propone un modelo de machine Learning para predecir el estado de la cosecha a partir de información del consumo de pesticidas y otras variables del cultivo, para lo cual el artículo está dividido en las siguientes sesiones: Estado del arte, Marco teórico, Metodología, conclusiones y bibliografía.

2 ESTADO DEL ARTE

La agricultura inteligente pretende brindar una herramienta al agricultor para poder procesar la información del cultivo y con esta predecir el comportamiento que va a tener la cosecha, esto le permitirá mejorar la toma de decisiones haciendo que la agricultura sea más eficiente y eficaz. La agricultura inteligente se ha venido desarrollando en los últimos años, y con el fin de dar un panorama de avance de esta nueva área de conocimiento, se realizó una revisión en la cual no solo se busca mostrar cuales son las tendencias en el área, si no también ver los referentes teóricos.

Para realizar la revisión se definió una ecuación de “Machine Learning” and “Precision agriculture”, con la cual se realizó una búsqueda en la base de datos de Scopus. Dicha búsqueda arrojó una cantidad de 1812 documentos, con los cuales se lograron hacer los siguientes análisis: número de documentos por año, número de publicaciones por país, número de publicaciones por institución y número de publicaciones por Autor.

En la figura 1 se puede observar el número de publicaciones en los últimos 15 años, de la cual es posible observar que el área temática se encuentra en crecimiento, en especial los últimos seis años, lo cual refleja la importancia de la agricultura inteligente.

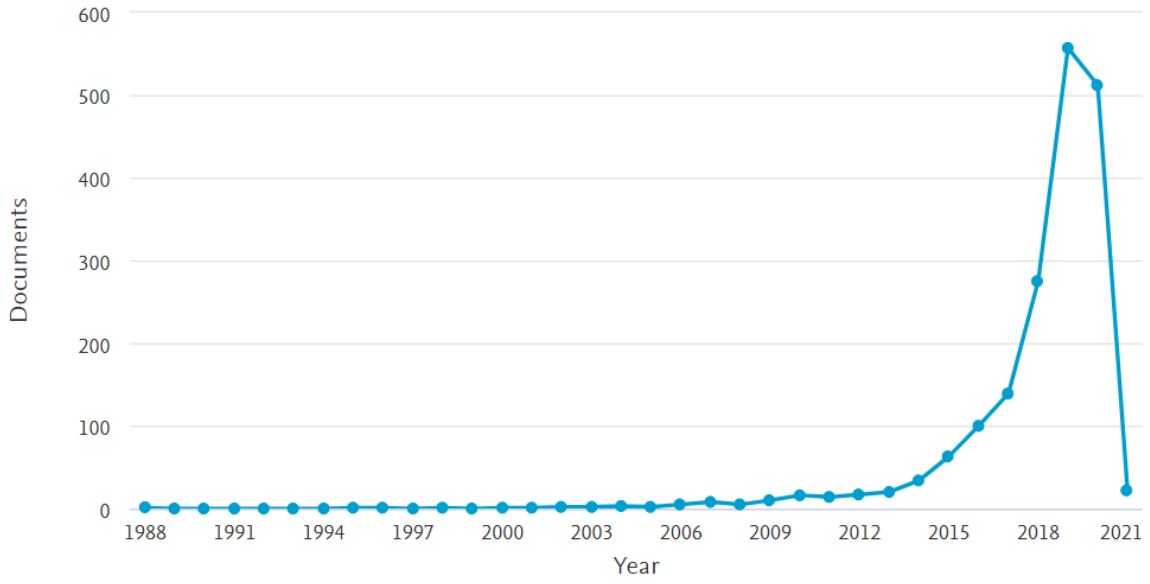

Fig. 1. Número de publicaciones por año para las temáticas de “Machine Learning” and “Precision agriculture” en Scopus

En la figura 2 se puede ver cuáles son los países que más han publicado en el área y se convierten en referentes teóricos, los países más relevantes son: India, Estados Unidos, China y Australia.

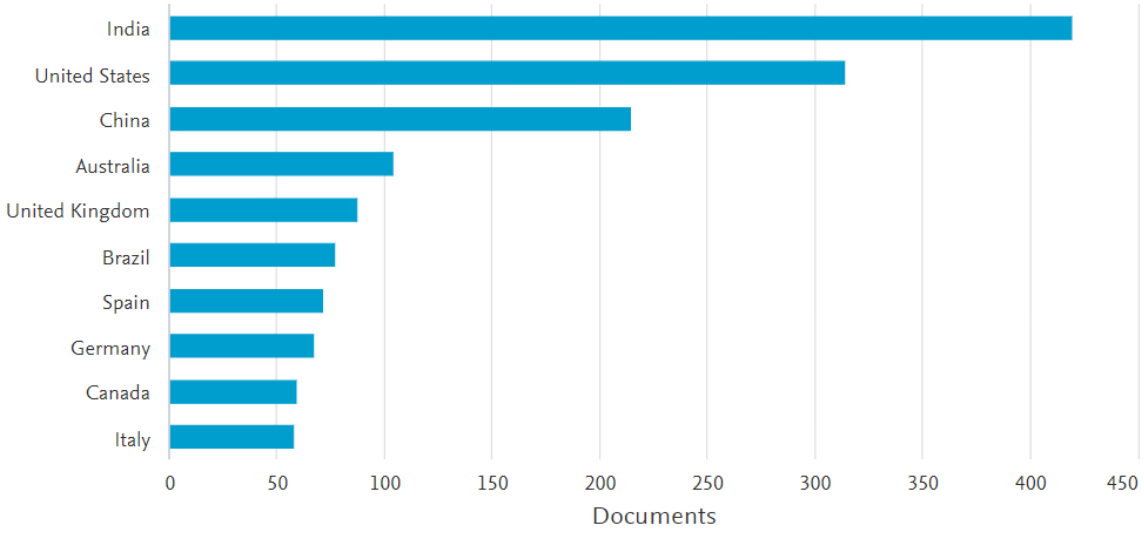

Fig. 2. Número de publicaciones por país para las temáticas de “Machine Learning” and “Precision agriculture” en Scopus

En la figura 3, se puede ver cuáles son las instituciones más importantes en el área, entre las que se destacan: Chinese Academy of Sciences de China, Universidade de Sao Paulo – USP de Brazil, Wageningen University & Research de Nueva Zelanda, University of Southern Queensland de Australia y Consejo Superior de Investigaciones Científicas de España

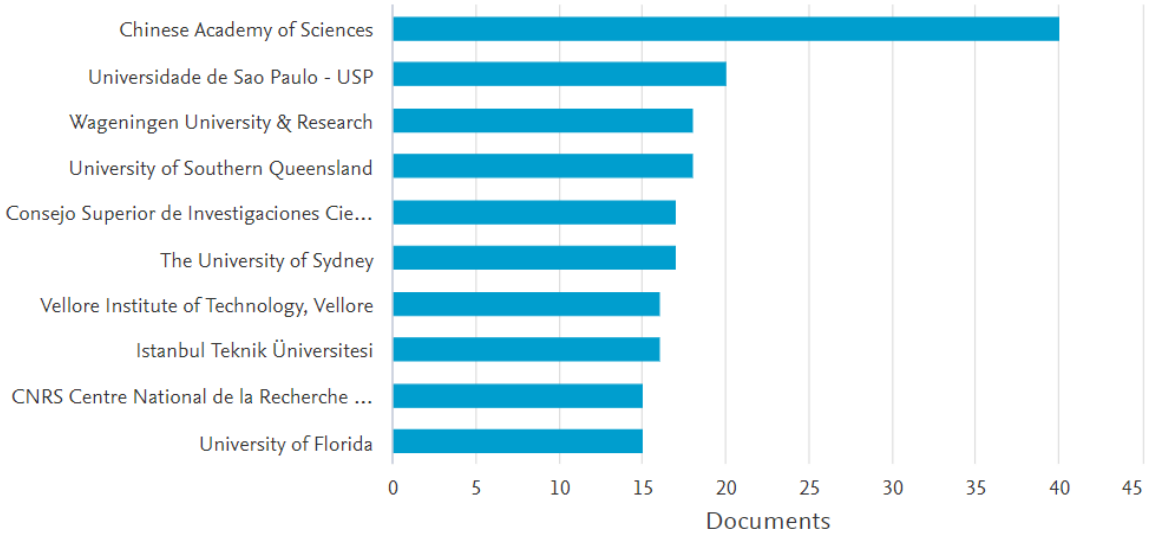

Fig. 3. Número de publicaciones por institución para las temáticas de “Machine Learning” and “Precision agriculture” en Scopus

En la figura 4 se puede ver cuáles son los autores más representativos en el área, entre los cuales se destacan: Deo, Ravinesh C, de University of Southern Queensland, Toowoomba, Australia; Corrales, Juan Carlos, de Universidad del Cauca, Popayan, Colombia; Yalçin, Hülya, de Istanbul Teknik Üniversitesi, Istanbul, Turkey; y Ampatzidis, Yiannis G, University of Florida Institute of Food and Agricultural Sciences, Gainesville, Estados Unidos

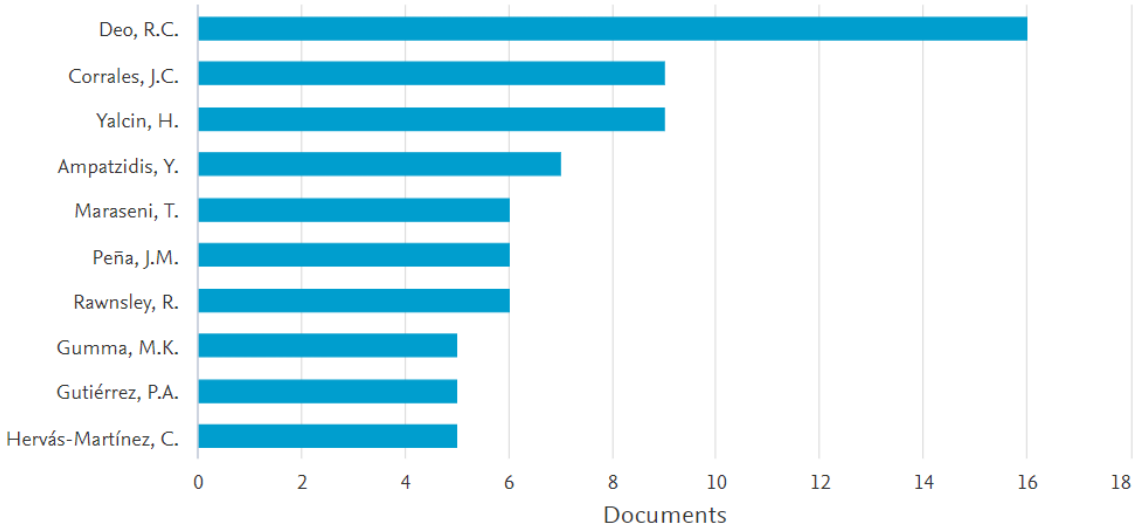

Fig. 4. Número de publicaciones por autor para las temáticas de “Machine Learning” and “Precision agriculture” en Scopus

Adicionalmente con la ayuda de Vosviewer se identificaron las temáticas en las cuales se está desarrollando el área de agricultura inteligente, a través de la construcción de una red de palabras claves, la cual se presenta en la figura 5.

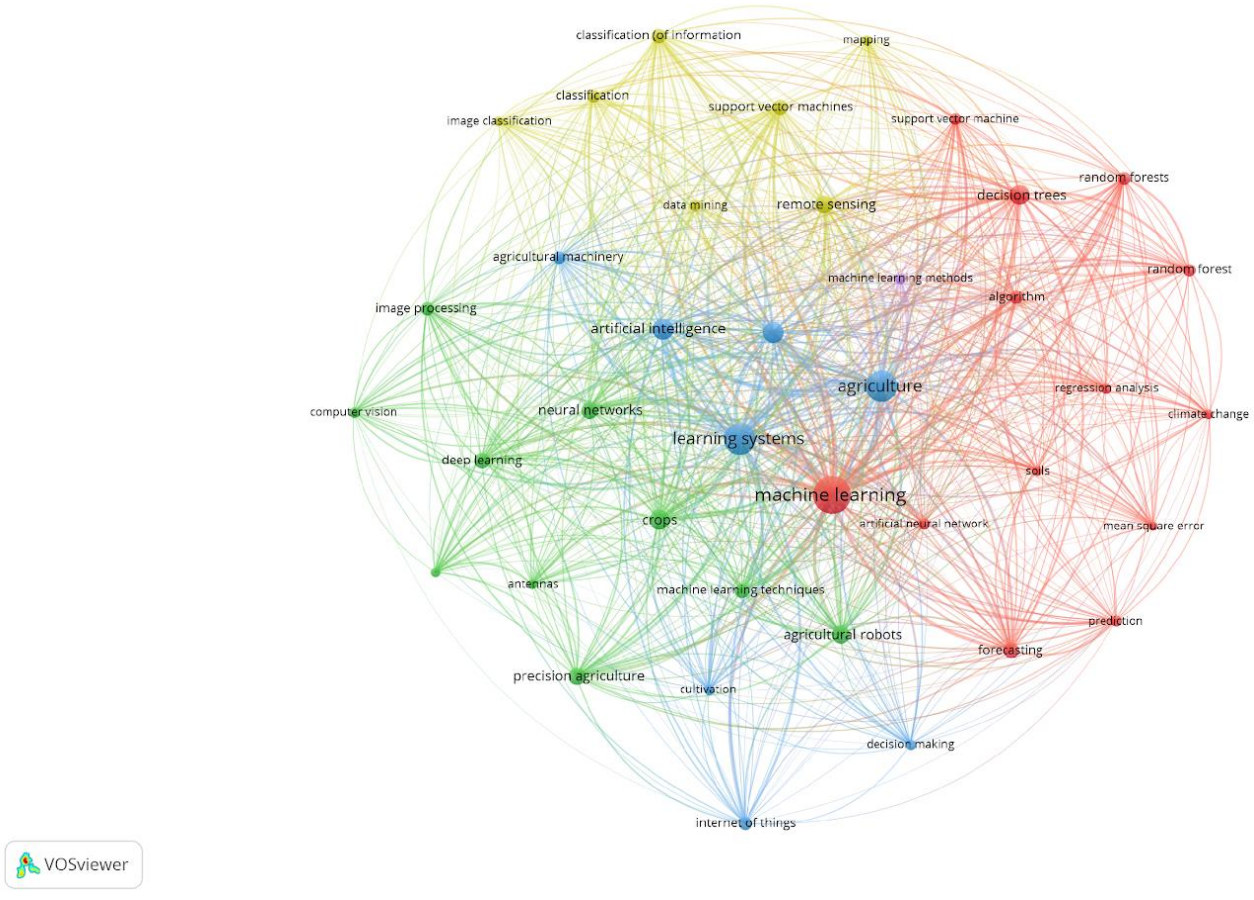

Fig. 5. Red de palabras clave de Agricultura inteligente.

Entre las palabras encontradas se destacan varios términos que son útiles a la hora de conocer más información sobre las áreas relacionadas: Agricultura robots, unmanned aerial vehicles, image processing, computer vision; modelos de machine learning: classification (of information),Decision trees, support vector machines, Regression analysis, Random Forest, Support vector machine, forecasting, Neural networks, Data mining y Artificial intelligence adicionalmente se destacan medidas de desempeño de los modelos Mean square error [8]. En la figura 5 se puede ver que las palabras se pueden dividir cuatro clústeres, los cuales se presentan en la tabla 1.

TABLA 1
CLÚSTER DE TÉRMINOS RELACIONADOS CON AGRICULTURA INTELIGENTE

Clúster uno	**Clúster dos**	**Clúster tres**	**Clúster cuatro**
algorithm	agricultural robots	agricultural machinery	classification
artificial neural network	antennas	agriculture	classification (of information)
climate change	computer vision	artificial intelligence	data mining
decision trees	crops	cultivation	image classification
forecasting	deep learning	decision making	mapping
mean square error	image processing	internet of things	remote sensing
prediction	machine learning techniques	learning algorithms	support vector machines
random forest	neural networks	learning systems	
regression analysis	precision agriculture		
soils	unmanned aerial vehicles (uav)		

3 MARCO TEORICO

El Machine Learning en la agricultura de precisión busca encontrar patrones a partir de la información histórica de las diferentes variables del cultivo, con el fin de mejorar toma de decisiones que permitan garantizar una mayor productividad y calidad de la cosecha [1]. Los algoritmos de clasificación buscan predecir los posibles estados a partir de información [9] como se muestra en la figura 6. Estos algoritmos buscan patrones que permitan hacer la clasificación [10].

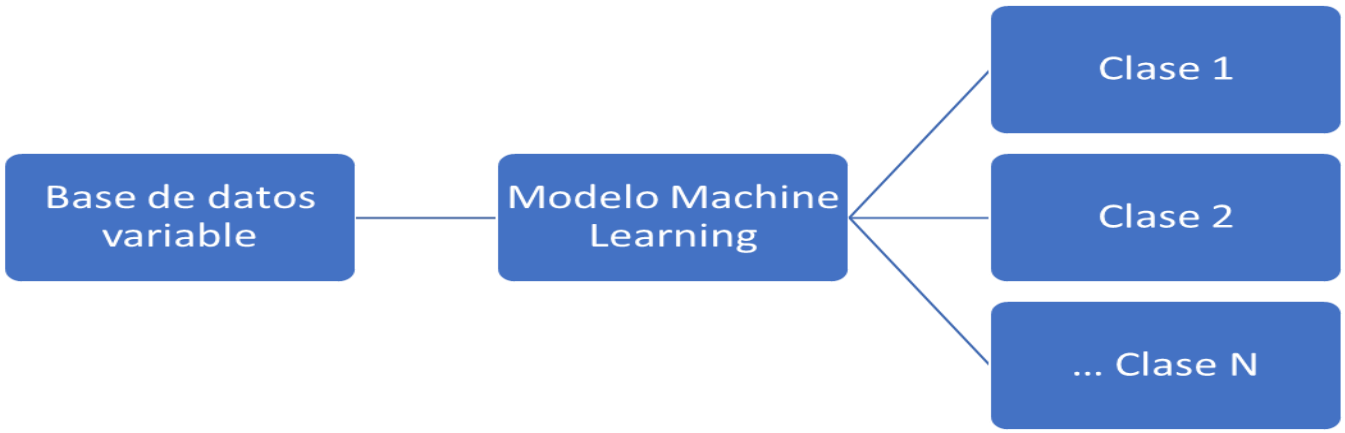

Fig. 6. Funcionamiento de algoritmo de clasificación

Con el fin de procesar la información se han utilizado diferentes modelos clasificación, como son: Decisión Tree, Naïve Bayes, Support Vector Machine (SVM) y KNN [11].

3.1 Decisión Tree

El árbol de decisión consiste la formulación de un número de preguntas condicionales, de tal forma que para cada pregunta se dividan los datos hasta lograr definir claramente a que categoría pertenecen; el número de preguntas se conoce como el hiperparámetro de este modelo. En la figura 7 se presenta un ejemplo de árbol de decisión [11].

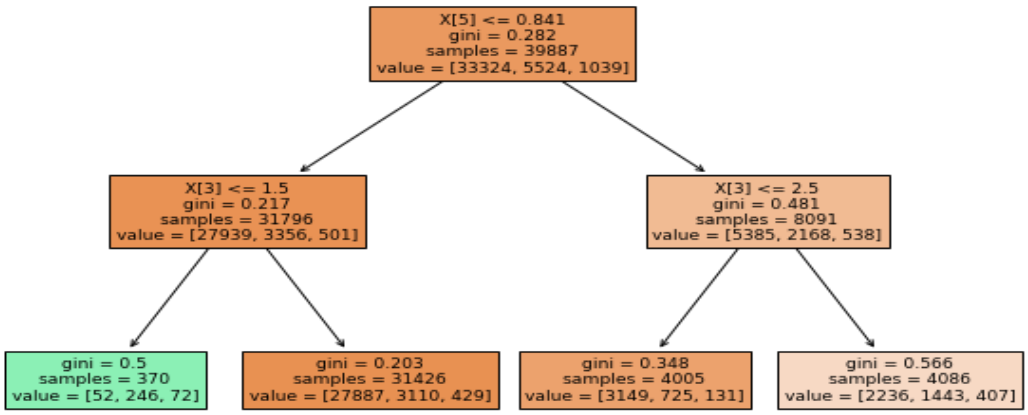

Fig. 7. Ilustración de un clasificador por árbol de decisión.

En la figura 7 se puede ver el árbol de decisión en donde la primera pregunta es $X[5] \leq 0.841 X[5] \leq 0.841$; si la respuesta es si pasa al lado derecho y si la respuesta en no al izquierdo, así para cada pregunta divide los datos y se generan los subconjuntos en los cuales se mide la probabilidad de pertenecer una categoría con el denominado índice de Gini: a medida que se realizan más preguntas se aumenta la probabilidad de pertenecer a una categoría. Para el árbol de decisiones utilizaremos la librería *sklearn DecisionTreeClassifier* [11], [12].

3.2 Clasificador KNN

El modelo de vecinos cercanos parte del concepto de vecindad, en el cual define la categoría a partir del número K de vecinos más cercanos [13], por esta razón el valor K se conoce como el hiperparámetro de este modelo. En la figura 11 se muestra un ejemplo de este clasificador.

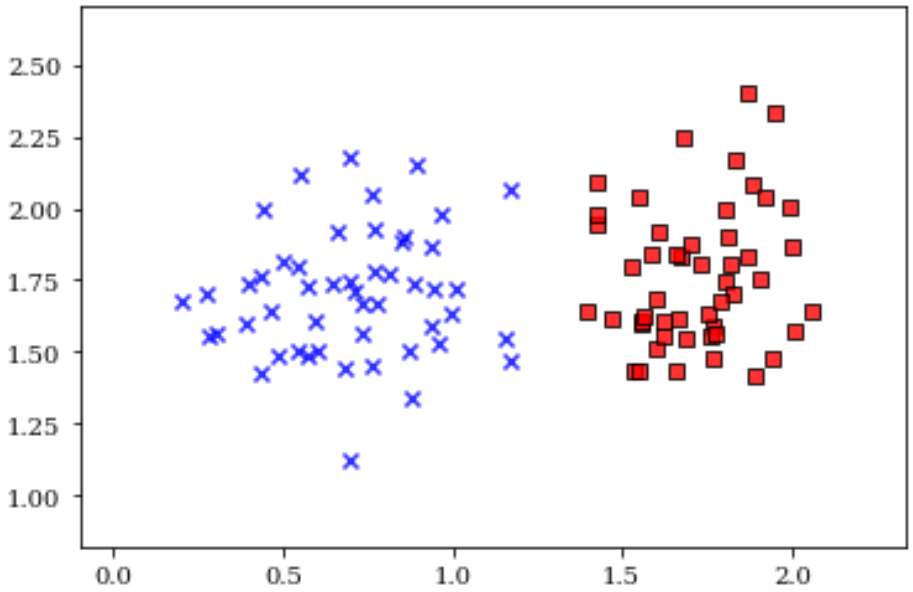

Fig. 8. Ilustración de un clasificador KNN

De la figura 8, se elegimos un punto de la gráfica donde encuentre que hay 3 vecinos de color rojo, por mayoría el punto debería pertenecer a la categoría de cuadros rojos. Para número de vecinos cercanos KNN utilizaremos la librería de *sklearn KNeighborsClassifier* [11].

3.3 Naive Bayes

El método Naive Bayes es un método de clasificación, el cual genera una hipótesis para el conjunto de clases de salida en función de los atributos de entrada, su enfoque es probabilístico, al determinar la probabilidad de una clase de acuerdo con los atributos de entrada [11]. Para Naive Bayes se utilizará la librería de *sklearn.naive_bayes.MultinomialNB* [11][14].

3.4 Máquina de Soporte Vectorial

La máquina de soporte vectorial es un método de clasificación en el cual se define un vector de separación entre las diferentes clases, el cual debe cumplir la máxima distancia posible a los elementos de cada clase. Se usa particularmente en dominios ruidosos y complejos. Para Support Vector Machine se usará la librería *sklearn.svm. Svc* [11], [15].

4 METODOLOGIA

4.1 Datos

La mayoría de las personas no aprecian el trabajo del agricultor, pero este trabajo es una verdadera prueba de resistencia y determinación. Una vez sembradas las semillas, trabaja días y noches para asegurarse de que cultiva una buena cosecha al final de la temporada. Existe varios factores que pueden afectar el éxito de la cosecha como son: la disponibilidad de agua, la fertilidad del suelo, la protección de los cultivos de los roedores, el uso oportuno de pesticidas y otros productos químicos [16]. Si bien muchos de estos factores son difíciles de controlar, la cantidad y frecuencia de los pesticidas es algo que el agricultor puede controlar.

Los pesticidas también son especiales, porque mientras protegen el cultivo con la dosis correcta, pero, si agregas más de lo requerido, pueden estropear toda la cosecha [17]. Un alto nivel de pesticida puede considerar el cultivo muerto / inadecuado para el consumo entre muchos resultados. Estos datos se basan en cultivos cosechados por varios agricultores al final de la temporada de cosecha. Para simplificar el problema, se puede suponer que todos los demás factores como las variaciones en las técnicas de cultivo han sido controlados. Es necesario determinar el resultado de la temporada de cosecha, es decir, si el cultivo estaría sano (vivo), dañado por pesticidas o dañado por otras razones.

4.2 Análisis exploratorio de los datos

Los datos están compuestos por 10 variables los cuales se presentan en la tabla 2, para los cuales se dispone de 88854 datos. De las variables se tiene una identificación ID, 5 variables categóricas que son: Tipo de cultivo, tipo de suelo, tipo de uso de pesticidas, estación y categoría de daños a los cultivos. Adicionalmente, se cuenta con 4 variables numéricas que son: número estimado de insectos por metro cuadrado, número de dosis por semana de pesticida, número de semanas que se utilizó pesticida y número de semanas sin uso de pesticidas.

TABLA 2 Variable de la base de datos

Columna	Descripción	Valores
ID	Identificación	ID
Estimated_Insects_Count	número estimado de insectos por metro cuadrado	Numérico
Crop_Type	Tipo de cultivo	Categorías (0 ,1)
Soil_Type	tipo de suelo	Categorías (0,1)
Pesticide_Use_Category	tipo de uso de pesticidas	Categorías (1: nunca, 2: previamente usado, 3: usado actualmente)
Number_Doses_Week	número de dosis por semana de pesticida	Numérico
Number_Weeks_Used	número de semanas que se utilizó pesticida	Numérico
Number_Weeks_Quit	número de semanas sin uso de pesticidas	Numérico
Season	estación	categoría (1,2,3)
Crop_Damage	categoría de daños a los cultivos	Categorías (0: cultivo saludable, 1: daños por otras causas, 2: daños por el uso de pesticidas)

4.3 Procesamiento de la información

Al revisar la información se identifica que las variables “Number_Weeks_Used” tabla 3, tiene un total de 79858 datos no nulos, lo que equivale a tener 9000 datos nulos que en porcentaje son 10 % de los registros. Para no perder la información de las demás variables que si cuentan con el registro, se procede a sustituir los datos nulos faltantes [18].

TABLA 3
RESUMEN DE LOS DATOS

Columna	Número de datos no nulos	Tipo de datos
ID	88858	object
Estimated_Insects_Count	88858	int64
Crop_Type	88858	int64
Soil_Type	88858	int64
Pesticide_Use_Category	88858	int64
Number_Doses_Week	88858	int64
Number_Weeks_Used	79858	float64
Number_Weeks_Quit	88858	int64
Season	88858	int64
Crop_Damage	88858	int64

Para este proceso se utilizan tres posibilidades que son:

- Reemplazar por la media: se remplaza los datos faltantes por la media
- Rellenar hacia adelante: se remplaza los datos faltantes por los datos adelante (series de tiempo)
- Rellenar hacia atrás: se remplaza los datos faltantes por los datos atrás (series de tiempo)

Con el fin de identificar si el promedio pudiera ser un valor adecuado para realizar la sustitución de los datos faltantes, se procedió a realizar los estadísticos de la variable “Number_Weeks_Used”, los cuales se presentan en la tabla 4. Adicionalmente, se grafica un histograma con los valores de la variable, la cual se puede ver en la figura 9 y en la cual se ubicó aproximadamente el promedio con una línea roja.

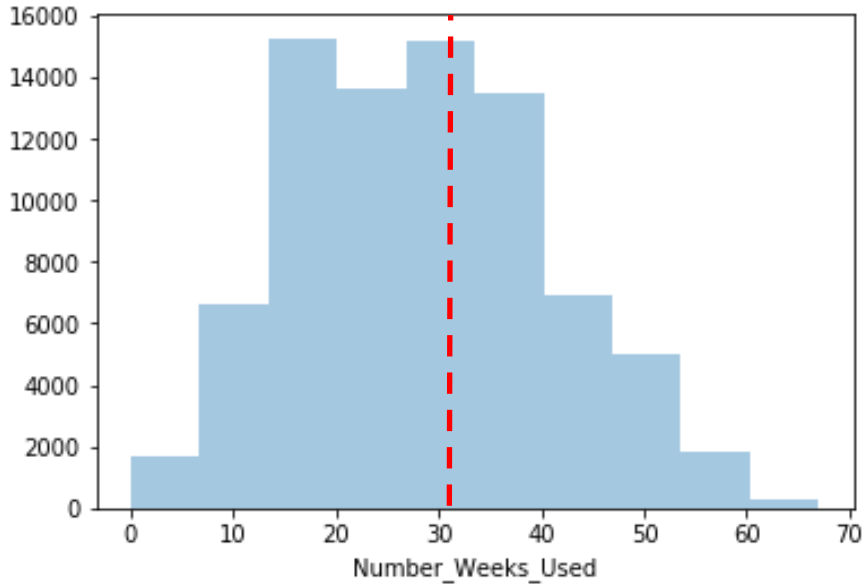

Fig. 9. Histograma del Número de semanas que se utilizó pesticida.

En la figura 9 se puede ver que los datos de la variable tienen una distribución normal, con una concentración de datos alrededor de la media y, adicionalmente, que los percentiles de 25 y 75 se encuentran cercanos a la media. Por este motivo se tomó como decisión utilizar la media para sustituir los datos faltantes debido a que es el valor más representativo.

TABLA 4
DATOS RELACIONADOS CON EL NÚMERO DE SEMANAS QUE SE UTILIZÓ PESTICIDA

Número de datos	79858
promedio	28.623970
std	12.391881
mínimo	0
Q 25%	20
Q 50%	28
Q 75%	37
Máximo	67

TABLA 5
DESCRIPCIÓN DE LOS DATOS

Variable	**Número de datos**	**promedio**	**std**	**mínimo**	**Q 25%**	**Q 50%**	**Q 75%**	**Máximo**
Estimated_Insects_Count	88858	1399.012210	849.048781	150.000000	731	1212	1898	4097
Crop_Type	88858	0.284375	0.451119	0	0	0	1	1
Soil_Type	88858	0.458417	0.498271	0	0	0	1	1
Pesticide_Use_Category	88858	2.264186	0.461772	1	2	2	3	3
Number_Doses_Week	88858	25.849952	15.554428	0	15	20	40	95
Number_Weeks_Used	88858	28.623970	11.747567	0	20	20,623	36	67
Number_Weeks_Quit	88858	9.589986	9.900631	0	0	7	16	50
Season	88858	1.896959	0.701322	1	1	2	2	3
Crop_Damage	88858	0.190562	0.454215	0	0	0	0	2

Luego de ajustar los datos faltantes es necesario estandarización de los datos de las variables, esto se debe a que hay variables que tiene una escala mayor que las demás, lo cual puede producir que estas cuenten con un mayor peso relativo a la hora de entrenar el modelo incurriendo así en un error en la predicción [19]. En nuestro caso las variables "Estimated_Insects_Count", "Number_Doses_Week", "Number_Weeks_Used" y "Number_Weeks_Quit", requieren un ajuste de escala, para lo cual se utilizó el comando estándar escalar sklearn, después de normalizar los datos.

TABLA 6
ESTANDARIZACION DE LOS DATOS

Variable	**Número de datos**	**promedio**	**std**	**mínimo**	**Q 25%**	**Q 50%**	**Q 75%**	**Máximo**
Estimated_Insects_Count	88858	6.712815e-16	1	-1.4711	-0.7867	-0.2202	0.58770	3.177677
Number_Doses_Week	88858	4.375397e-17	1	-1.66191	-0.6975	-0.3761	0.90971	4.445708
Number_Weeks_Used	88858	4.375397e-17	1	-2.43660	-0.73411	3.024229e-16	0.62788	3.266740
Number_Weeks_Quit	88858	3.900862e-15	1	-9.68629	-0.96862	-0.26159	0.64743	4.081582

4.4 Análisis de la información

En nuestro modelo de Machine Learning la variable a predecir es “daños a los cultivos”, la cual es una variable categórica y puede tomar los siguientes valores (0: cultivo saludable, 1: daños por otras causas, 2: daños por el uso de pesticidas). Con el fin de conocer la interacción de nuestra variable de salida, se realizaron una serie de gráficos que la relacionan con las demás variables. En la figura 10 se puede observar que la probabilidad de tener un cultivo sano disminuye a medida que se aumenta la cantidad de insectos; sin embargo, también se observa que los cultivos podrían mantenerse saludables manteniendo el número de insectos en un rango aceptable entre las 300 y 800 unidades por metro cuadrado sin incrementar el uso de pesticidas, debido a que estos también afectan de manera significativa la calidad de los cultivos.

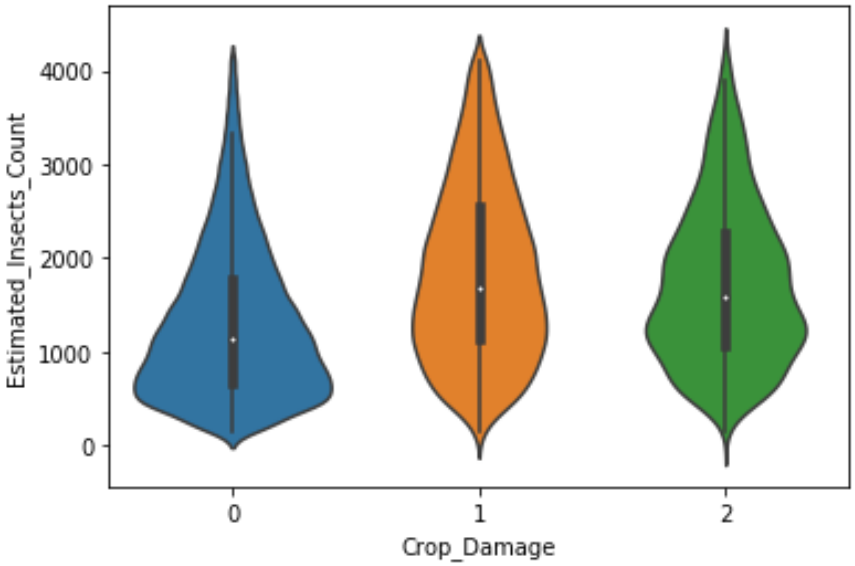

Fig. 10. Daños a los cultivos vs Número estimado de insectos por metro cuadrado.

En la figura 11 se observa que a menor número de semanas de uso de pesticida aumenta la posibilidad de tener un cultivo sano, y a mayor número de semana usando pesticida mayor posibilidad de dañar el cultivo por exceso de pesticidas.

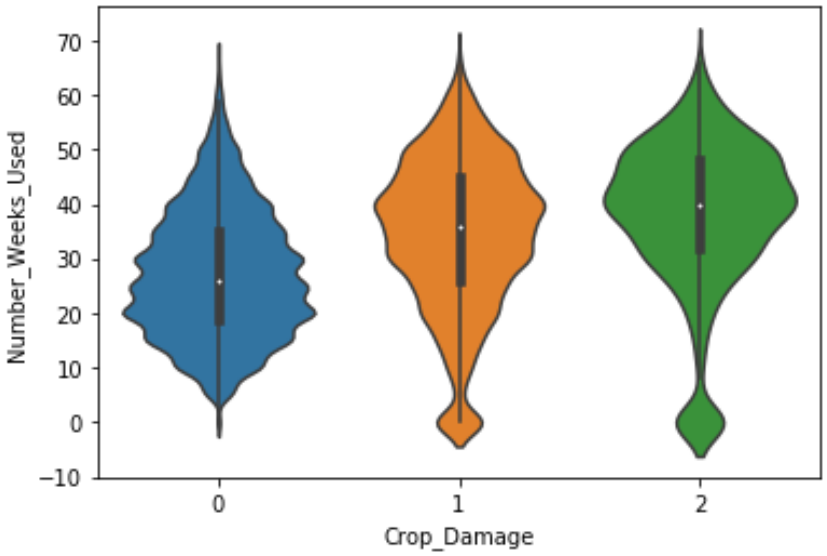

Fig. 11. Daños a los cultivos vs Número de semanas usando pesticidas.

En la figura 12 se puede ver que cuando se deja de usar pesticida, la probabilidad de pérdida del cultivo es alta, con respecto a la posibilidad de un cultivo sano y o con daños por otras causas. Según el análisis realizado, el uso de pesticida es recomendado entre la segunda y la veinteava semana, debido a que el uso prolongado del pesticida también incrementa los daños al cultivo.

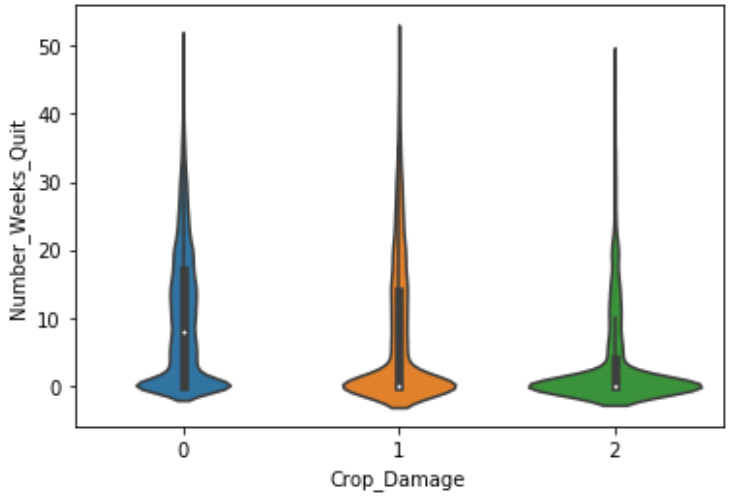

Fig. 12. Daños a los cultivos vs Número de semanas sin uso de pesticidas**.**

4.5 Selección de Modelo

De acuerdo con la variable de salida, que es el estado del cultivo, una variable categórica que puede tomar solo tres valores (0: cultivo saludable, 1: daños por otras causas, 2: daños por el uso de pesticidas), nuestro modelo de Machine Learning debe ser de categorización. De acuerdo con la revisión de literatura, se propone implementar los algoritmos de: KNN Número de vecinos cercanos (KNeighborsClassifier), árbol de decisiones (DecisionTreeClassifier), SVM- Máquina de Soporte Vectorial y Naive Bayes. Para la selección de modelo se tomará como métrica el "accuracy score" o medida de precisión.

Para realizar el entrenamiento y la validación se realizó la división de los datos con la librería "train_test_split" la cual permite dividir los datos en entrenamiento, test y validación, de tal forma que los subconjuntos de datos conserven la misma proporción de datos originales y de esta forma no afectar la predicción. Para el entrenamiento se empleó el comando. "fit" y para la predicción. "predict".

TABLA 6
COMPARACIÓN DE LOS MODELOS

Modelo	Accuracy train	Accuracy test
DecisionTreeClassifier	0.8403239150600447	0.8406372474169085
KNeighborsClassifier	0.8421540852909469	0.8410516982676306
SVM	0.8354601749943591	0.8410516982676306
Naive Bayes	0.8354601749943591	0.8354590096848997

De acuerdo con la tabla 6 los modelos que tiene un mejor desempeño son KNN y Decision tree, para estos modelos se realizará un análisis de sensibilidad para seleccionar los hiperparámetros.

4.6 Selección de hiperparámetros

Para la selección del hiperparámetro de cada uno de los modelos se tomó como métrica el "accuracy score".

4.6.1 KNN

Para el modelo KNN el hiperparámetro es el número de vecinos; para este se tomaron los siguientes valores de vecinos: (20,25,30,40,45,50,60,70,80,90,100,110,120,130). La gráfica de los datos se presenta en la figura 13.

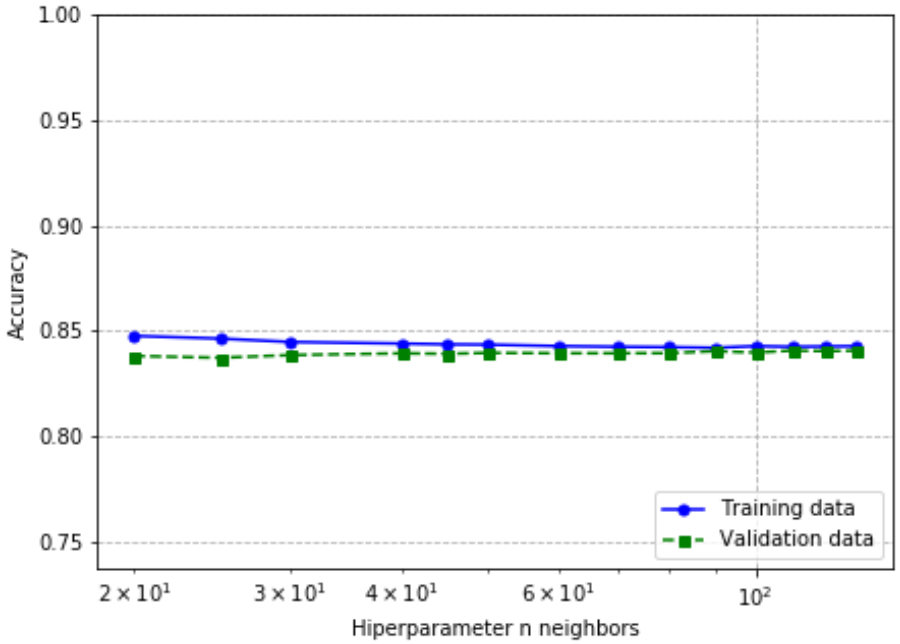

Fig. 13. Gráfica de los Hiperparámetros de KNN.

Con el modelo de Knn se puede ver que el accuracy alcanzado por el modelo, con 200 vecinos, es de 0.8410516982676306 en validación y para el entrenamiento es de 0.841960604672469.

4.6.2 Decision Tree

Para el modelo árbol de decisiones se tomaron los siguientes valores de profundidad (1,2,3, 4,5, 6,7, 8, 10, 12, 14, 16,18,20). Al graficar los datos se obtiene la figura 14.

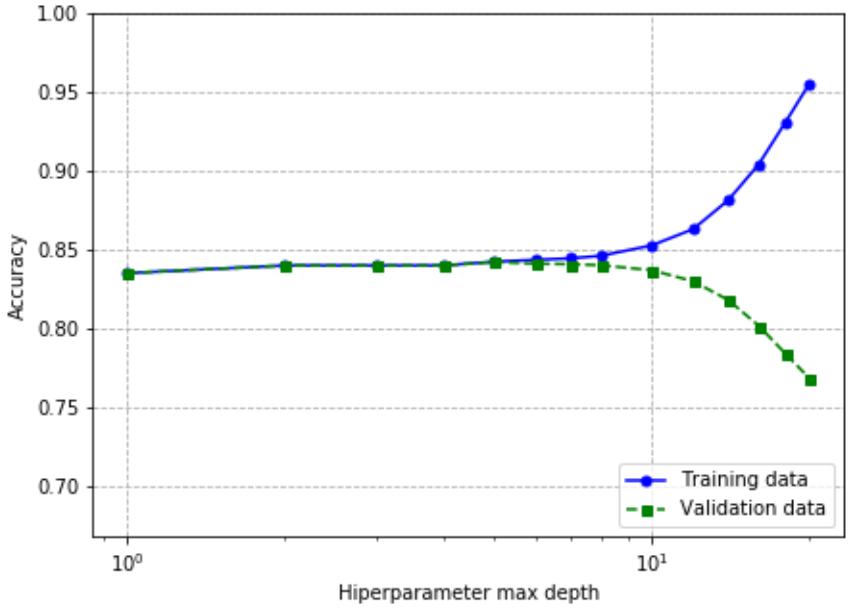

Fig. 14. Gráfica de búsqueda de Hiperparámetros de DecisionTreeClassifier

Con el modelo de árbol de decisión, desde una profundidad de 6, se tiene un accuracy de validación de 0.8425521756922657 y entrenamiento es 0.8426805726176448. A partir de esta profundidad se cae en el problema

del sobre ajuste, ya que el entrenamiento tiene un mejor ajuste que la validación, lo que se entiende como que el modelo se está aprendiendo los resultados y no tiene una buena predicción.
Al comparar los modelos KNN y Decision-tree se puede ver que este último tiene un mejor desempeño con un costo computacional menor, por lo que el modelo seleccionado es el Decision-tree. Adicionalmente, se realizará el análisis de un modelo de ensamble Boosting. El Boosting engloba a una familia de algoritmos cuya idea general es tomar modelos sencillos como los árboles de decisión y mejorar sus predicciones de manera secuencial. Para mejorar esas predicciones, el algoritmo entrena cada modelo secuencialmente con todos los datos en cada iteración y para cada nuevo modelo se le da más peso a los datos que no fueron bien clasificados o cuyo error en regresión sea más alto. Finalmente, la predicción va a ser una suma ponderada de todos los clasificadores base o en el caso de clasificación por votación Es similar a bagging en el sentido de que usa varios modelos base; sin embargo, el entrenamiento es secuencial y dependiente, es decir, el modelo en la iteración actual depende de las predicciones en la iteración anterior. Para implementar el modelo de ensamble Boosting se utilizó la librería *sklearn.svm ensemble.AdaBoostClassifier* [20]. Para el modelo Boosting se utilizó un ajuste de tres modelos Decision Tree con una profundidad de 2, donde se obtiene el accuracy score de la Tabla 7.

TABLA 6
COMPARACIÓN DE LOS MODELOS

Modelo	Accuracy train	Accuracy test
AdaBoostClassifier	0.8403239150600447	0.8402673577956623

Al analizar los resultados obtenidos el mejor modelo es Boosting ya que es el que permite un mejor ajuste tanto para el entrenamiento como la validación

5 CONCLUSIONES

En este trabajo se propusieron cinco modelos de Machine Learning para predecir el estado de un cultivo en la cosecha a partir de la información recolectada sobre el uso de pesticidas y otras variables.
Para la primera comparación se tiene los modelos de KNN, Decision Tree, SVM y Naive Bayes, al realizar el ajuste con “accuracy score” se obtiene que los dos mejores modelos son KNN y Decision Tree por lo cual se procedió hacer un mejor ajuste de sus parámetros a través de un análisis de sensibilidad. Como resultado, los parámetros fueron de 200vecinos para KNN, mientras que para árbol de decisión el parámetro de profundidad es de 5.
Al comparar los dos métodos el árbol de decisión presenta un mejor desempeño ya que logra obtener un mejor ajuste, adicionalmente, lo logra con un coste computacional significativamente menor por lo cual se puede decir que es el mejor modelo para la predicción durante esta primera ronda de comparación.
Por último, se propuso un modelo de ensamble tipo Boosting a partir de modelos Decision Tree con el cual se logró hacer una mejor aproximación con la suma varios modelos más simples dando como resultado el modelo seleccionado para la predicción del estado de la cosecha.

REFERENCIAS

[1] R. Katarya, A. Raturi, A. Mehndiratta, y A. Thapper, «Impact of Machine Learning Techniques in Precision Agriculture», *Proc. 3rd Int. Conf. Emerg. Technol. Comput. Eng. Mach. Learn. Internet Things ICETCE 2020*, n.º February, pp. 18-23, 2020, doi: 10.1109/ICETCE48199.2020.9091741.
[2] U. L. C. Baldos, K. O. Fuglie, y T. W. Hertel, «The research cost of adapting agriculture to climate change: A global analysis to 2050», *Agric. Econ.*, vol. 51, n.º 2, pp. 207-220, mar. 2020, doi: 10.1111/agec.12550.
[3] C. A. Amaya Corredor, C. H. Contreras, N. Pedroza Rosas, y R. S. Cáceres Quintero, «Estrategias de adaptación y mitigación al cambio climático de las Unidades Tecnológicas de Santander», *Rev. CINTEX*, vol. 22, n.º 2, pp. 89-109, dic. 2017, doi: https://doi.org/10.33131/24222208.301.
[4] J. J. Castro-Maldonado, J. A. Patiño-Murillo, A. E. Florian-Villa, y O. E. Guadrón-Guerrero, «Application of computer vision and low-cost artificial intelligence for the identification of phytopathogenic factors in the agro-industry sector», *J. Phys. Conf. Ser.*, vol. 1126, p. 012022, nov. 2018, doi: 10.1088/1742-6596/1126/1/012022.
[5] F. A. González, J. J. Gómez, y D. F. Amaya, «Multispectral image processing in coffee and cocoa crops», *Rev. CINTEX*, vol. 22, n.º 2, pp. 51-67, dic. 2017, doi: 10.33131/24222208.294.
[6] C. Lau, A. Jarvis, y J. Ramirez, «Agricultura Colombiana: Adaptación al Cambio Climático», *CIAT - Cent. Int. Agric. Trop.*, vol. 1, p. 4, 2011.

[7] A. Chlingaryan, S. Sukkarieh, y B. Whelan, «Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review», *Comput. Electron. Agric.*, vol. 151, pp. 61-69, ago. 2018, doi: 10.1016/j.compag.2018.05.012.
[8] J. Patiño, J. D. López, y J. Espinosa, «Sensitivity Analysis of Frequency Regulation Parameters in Power Systems with Wind Generation», en *Advanced Control and Optimization Paradigms for Wind Energy Systems*, vol. 172, R.-E. Precup, T. Kamal, y S. Zulqadar Hassan, Eds. New York, NY: Springer New York, 2019, pp. 67-87. doi: 10.1007/978-981-13-5995-8_3.
[9] M. I. Ardila Marín, W. Orozco Murillo, J. Galeano Echeverri, y A. M. Medina Escobar, «Desarrollo de software para la gestión del mantenimiento en los laboratorios de la I.U. Pascual Bravo», *Rev. CINTEX*, vol. 23, n.º 1, pp. 43-50, oct. 2018.
[10] R. Shirsath, N. Khadke, D. More, P. Patil, y H. Patil, «Agriculture decision support system using data mining», *Proc. 2017 Int. Conf. Intell. Comput. Control I2C2 2017*, vol. 2018-January, pp. 1-5, 2018, doi: 10.1109/I2C2.2017.8321888.
[11] S. Umadevi y K. S. J. Marseline, «A survey on data mining classification algorithms», *Proc. IEEE Int. Conf. Signal Process. Commun. ICSPC 2017*, vol. 2018-January, n.º July, pp. 264-268, 2018, doi: 10.1109/CSPC.2017.8305851.
[12] Scikit-learn, «Scikit-learn user guide - Release 0.23.2», 2020.
[13] C. A. Duarte-Salazar, A. E. Castro-Ospina, M. A. Becerra, y E. Delgado-Trejos, «Speckle Noise Reduction in Ultrasound Images for Improving the Metrological Evaluation of Biomedical Applications: An Overview», *IEEE Access*, vol. 8, pp. 15983-15999, 2020, doi: 10.1109/ACCESS.2020.2967178.
[14] Scikit-learn, «sklearn.naive_bayes», 2020.
[15] C. V. Classification *et al.*, «sklearn.svm», 2020. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[16] V. Correchel, R. de O. Custodio Filho, F. Correa Veloso dos Santos, y I. Colares de Freitas, «Suelos Con Alta Susceptibilidad A La Degradación Ocupados Por Beneficiarios De Políticas De Reforma Agraria En Brasil», *Rev. CINTEX*, vol. 21, n.º 2, pp. 113-131, dic. 2016.
[17] G. Brookes y P. Barfoot, «Environmental impacts of genetically modified (GM) crop use 1996-2016: Impacts on pesticide use and carbon emissions», *GM Crops Food*, vol. 9, n.º 3, pp. 109-139, jul. 2018, doi: 10.1080/21645698.2018.1476792.
[18] F. Hoyos Gómez, J. D. Betancur Gómez, D. Osorio Patiño, y J. G. Ardila Marín, «Construcción de curvas de factor de concentración de esfuerzos por medio de simulaciones», *Rev. CINTEX*, vol. 21, n.º 1, jun. 2016, Accedido: dic. 19, 2019. [En línea]. Disponible en: https://revistas.pascualbravo.edu.co/index.php/cintex/article/view/8
[19] Y.-W. Chung, B. Khaki, T. Li, C. Chu, y R. Gadh, «Ensemble machine learning-based algorithm for electric vehicle user behavior prediction», *Appl. Energy*, vol. 254, p. 113732, nov. 2019, doi: 10.1016/j.apenergy.2019.113732.
[20] Scikit-learn, «sklearn.ensemble.AdaBoostClassifier», *Scikit-learn*, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html