Detección deforestación

Nicolás Penagos, Nicolás Colmenares, Bryan Guapacha, Juan Pablo Sanin
Inteligencia Artificial I, Universidad Icesi
Cali, Colombia
nicolas.penagosm98@gmail.com
nicoescol0818@gmail.com
juanpablosaninb@gmail.com
bryanicesi0822@gmail.com

Palabras clave— Deforestación, machine learning, k-means, regresión logística, imágenes satelitales, amazonas, sostenibilidad, bosques, medio ambiente.

Abstract— Every year, the levels of deforestation in the world's forests increase dramatically affecting environmental conditions and, in the future, the sustainability of life on earth. This is why this project intends to contribute to the situation by providing a satellite detection system for deforestation areas using machine learning techniques. This project has been developed in an academic context under the crisp dm methodology. As a result, a web application was developed based on a neural network and a linear regression model allowing the detection of satellite images that contain deforestation zones and also the prediction of deforestation probabilities based on data such as date and time and the geographical location

I. INTRODUCCIÓN

Los bosques son de vital importancia para la sostenibilidad y la preservación de la vida en la tierra, pues estos albergan la biodiversidad de los ecosistemas y además, juegan un papel fundamental en la producción de oxígeno y la reducción de las emisiones de dióxido de carbono. Sin embargo, la creciente explotación de grandes coberturas arbóreas por desmedidas actividades extractivas como la agricultura a gran escala o la ganadería extensiva, ponen cada vez más en peligro a estos ecosistemas y con ello, al equilibrio ambiental del planeta.

De acuerdo con un estudio realizado por la Universidad de Maryland, tan solo en el año 2020, 12.2 millones de hectáreas de bosques y selvas tropicales fueron destruidas [1]. Adicionalmente, este estudio también ha demostrado como Brasil se corona como el país con la mayor cantidad de pérdidas de bosques primarios del planeta y del mismo modo, otros países de la región como Bolivia, Perú o Colombia también se encuentran entre los diez países con mayores niveles de deforestación.

Por todo lo anterior, existe una gran necesidad de desarrollar soluciones tecnológicas que contribuyan a detectar y cuantificar las zonas geográficas que están sufriendo procesos de deforestación para ejecutar planes de acción que contribuyan a garantizar la protección ambiental del planeta.

II. MARCO TEÓRICO

- **Machine Learning**: Machine learning es “una forma de la IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita.” [4]. Un modelo de machine learning es “la salida de información que se genera cuando entrena su algoritmo de machine learning con datos” Entre las diferentes categorías de machine learning existen: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje de refuerzo y el deep learning.

- **Regresión Lineal:** es un algoritmo de aprendizaje supervisado el cual utiliza el análisis de la regresión lineal para predecir el valor de una variable según el valor de otra. La variable que desea predecir se denomina variable dependiente. La variable que se está utilizando para predecir el valor de la otra variable se denomina variable independiente”.[4]

- **Redes neuronales:** son técnicas de deep learning que intentan procesar datos de una forma similar a la que funciona el cerebro humano y “funcionan propagando entradas, ponderaciones y sesgos hacia adelante. Sin embargo, es en el proceso inverso de propagación hacia atrás donde la red aprende realmente al determinar los cambios exactos que se deben aplicar a las ponderaciones y sesgos para producir un resultado exacto.”[5]

III. METODOLOGÍA

Este trabajo fue desarrollado utilizando la metodología CRISP-DM. Por tanto, el proceso de investigación inicia con una etapa de entendimiento del negocio en la cual se definieron los objetivos iniciales del negocio y las siguientes preguntas de investigación:

1. .¿Cuáles son las características relevantes de imágenes satelitales que contengan áreas deforestadas y no deforestadas?

2. ¿Qué áreas geográficas presentan mayores niveles de deforestación?

3. ¿Cómo determinar si hay o no deforestación en una determinada zona geográfica?

4. ¿Cómo podemos predecir la evolución de la deforestación a partir de los datos actuales de una zona geográfica?

Adicionalmente, se definió como criterio de éxito lograr un modelo que fuera capaz de determinar si una imagen aleatoria presentaba procesos de deforestación con una precisión cercana a 80%. Por su

parte, se definió un plazo de desarrollo de 3 meses y además se realizó un análisis de riesgos y limitaciones entre otras actividades que se detallan en el documento anexo a este reporte.

Las siguientes etapas abordadas fueron la de recolección y entendimiento de los datos. Para esto se llevó a cabo una búsqueda intensiva de datos sobre deforestación en el Amazonas en la plataforma Kaggle (https://www.kaggle.com/). Se obtuvieron, por un lado datos recolectados en forma de tablas de varios datasets encontrados en la plataforma kaggle. Estos datos contienen variables tales como temporalidad, áreas de deforestación, cantidad de incendios desatados, el cambio neto del área forestal, hectáreas de arborizadas pérdidas, entre otros.

Por otro lado, se trabajará con un dataset de cerca 40.000 imágenes satelitales de la amazonía, en los cuales el modelo estará orientado al procesamiento de imágenes con técnicas de reconocimiento de redes neuronales para detectar los posibles casos de deforestación.

Una vez finalizado el proceso de recolección y exploración de datos. Se procedió a realizar la preparación de los datos y el análisis para la solución desde dos perspectivas: imágenes y datos.

Para el primer caso, con el dataset "Understanding the Amazon from Space" se contaba con 40478 imágenes de 256x256 pixeles (abarcando un área de 221.7 acres) etiquetadas con un total de 17 tags.

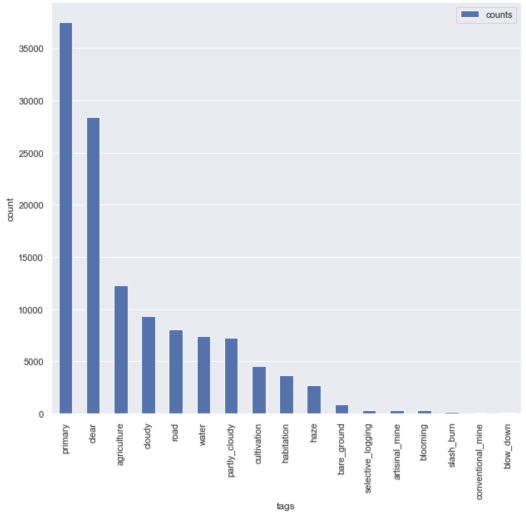

Fig. 1 Distribución de las etiquetas del dataset

A continuación, se procedió a convertir estas etiquetas de una cadena de texto que contenía todas las etiquetas separadas por espacio, a una tabla de categorías binaria en donde cada etiqueta cuenta con sus respectiva columna en donde el valor 0, significa la ausencia de la categoría y el valor 1 la presencia de esta.

	image_name	tags
0	train_0	haze primary
1	train_1	agriculture clear primary water
2	train_2	clear primary

	image_name	primary	clear	agriculture	cloudy	road	water	p
0	train_0	1	0	0	0	0	0	
1	train_1	1	1	1	0	0	1	
2	train_2	1	1	0	0	0	0	
3	train_3	1	1	0	0	0	0	

Fig. 2 Tratamiento de conversión a categorías en el dataset

Para realizar el análisis de estas etiquetas, se tuvieron en cuenta trabajos similares como el de André Ferreria [2], en el cual, dado a la variada de distribución de diferentes tags relacionados y no relacionados con situaciones de deforestación, se realiza una agrupación binaria en dos categorías finales: deforestación y no deforestación:

TABLA 1.
AGRUPACIÓN DE ETIQUETAS

No deforestación	*"bare_ground","blooming","blow_dow","clear","cloudy","habitation","haze","partly_cloudy","primary","water", "agriculture", "cultivation"*
Deforestación	*"artisinal_mine","conventional_mine","road","selective_logging","slash_burn"*

Por tanto, cada imagen que contenga una o más etiquetas del grupo de deforestación va a ser clasificada como deforestada y solo las que no contengan ninguna etiqueta de deforestación serán categorizadas como no deforestadas.

A continuación, las imágenes son reescaladas a una resolución de 75x75 pixeles divididas en las dos carpetas *deforestation* y *no deforestation* con base en el criterio de clasificación presentado anteriormente. Llegados a este punto y dadas las características del problema (reconocimiento de imágenes), se decide entrenar modelos de redes neuronales para la clasificación de imágenes los cuales se encuentran en el apartado siguiente.

Para el segundo caso, los datos, trabajamos con el dataset "Brazilian Amazon Rainforest Degradation 1999-2019" donde se contaba con 16 filas y 11 columnas que representaban el área deforestada anual en kilómetros cuadrados por regiones del 2004 al 2019. También incluía otro dataset con los puntos de incendio en el Amazonas del 1999 al 2019 compuesto de 2104 filas y 6 columnas.

En el dataset separaban el Amazonas de Brasil en 10 regiones, parte del trabajo de ajuste del dataset fue renombrar las columnas para facilitar la visualización de los datos. Las regiones se pueden ver en la siguiente tabla:

TABLA 2.
REGIONES DEL AMAZONAS

Regiones	*'Acre', 'Amazonas', 'Amapa', 'Maranhao', 'Mato Grosso', 'Para', 'Rondonia', 'Roraima', 'Tocantins'*

Luego de explorar el dataset pudimos determinar las regiones con mayor deforestación e incendios con apoyo en un trabajo del Salma B [3]. Obtuvimos el siguiente gráfico:

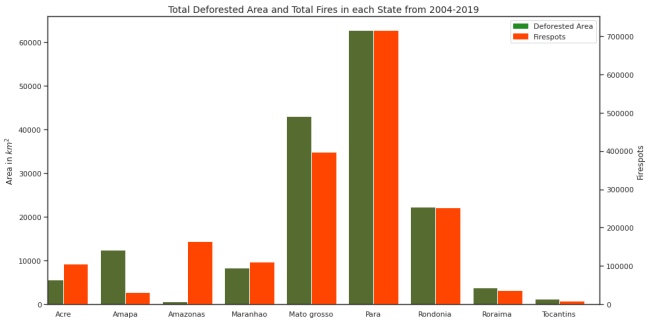

Fig. 3 Deforestación vs Incendios por zonas

Por consecuente, determinamos que existía una relación entre la cantidad de incendios y el área deforestada, planteamos desarrollar un modelo de regresión lineal que se encuentra en la siguiente sección. Además, se busca establecer una relación entre el año y el área deforestada utilizando una regresión lineal.

IV. RESULTADOS Y ANÁLISIS

Posteriormente, se llevó a cabo las etapas de modelado y evaluación. Para la construcción de los modelos, en primera instancia para las imágenes, se llevó a cabo un particionamiento del dataset con 80% (32384) para pruebas y un 20% (8095) para test. Posteriormente se entrenaron evaluaron a traves de graficas de precisión los siguientes modelos:

1. Modelo de tensor flow sin convolución

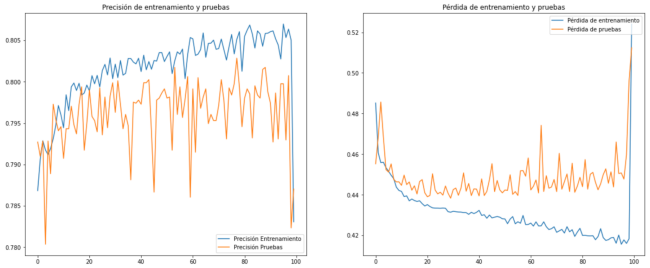

Fig. 4 Gráficas de precisión modelo de tensor flow sin convolución.

2. Modelo convolucional.

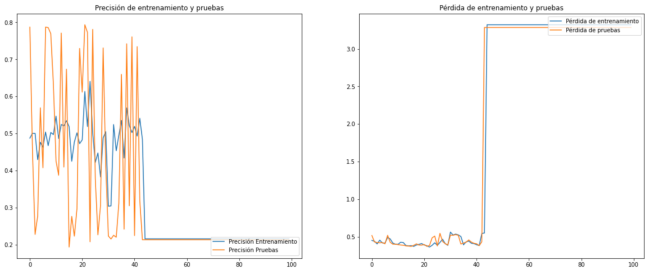

Fig. 5 Gráficas de precisión modelo convolucional

3. Modelos convolucionales con drop out y batch size de (8 a 100).
   a. 8

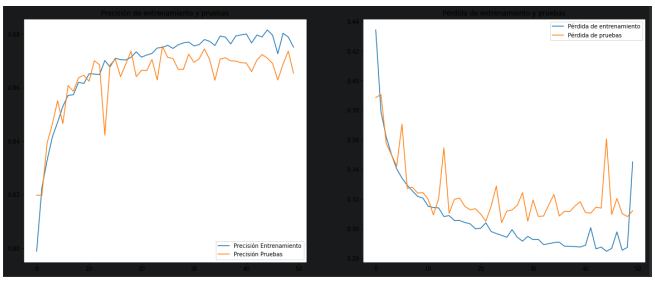

Fig. 6 Gráficas de precisión modelo convolucional batch size 8

b. 16

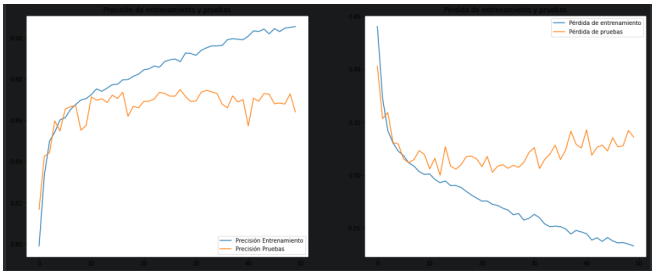

Fig. 7 Gráficas de precisión modelo convolucional batch size 16

c. 32

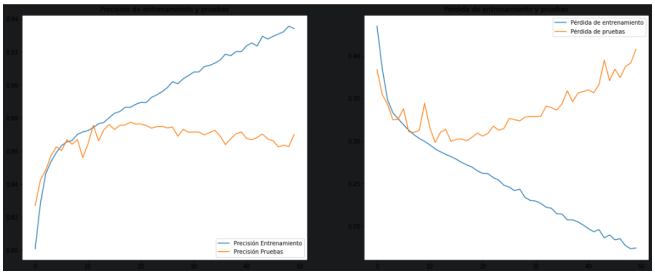

Fig. 8 Gráficas de precisión modelo convolucional batch size 32

d. 64

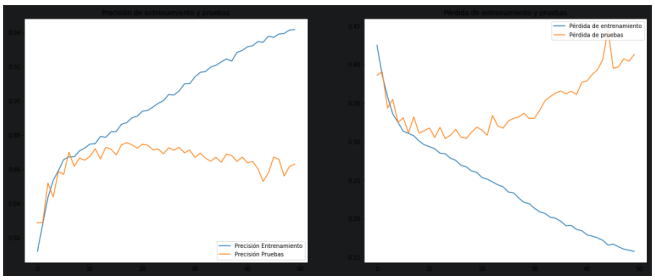

Fig. 8 Gráficas de precisión modelo convolucional batch size 64

e. 100

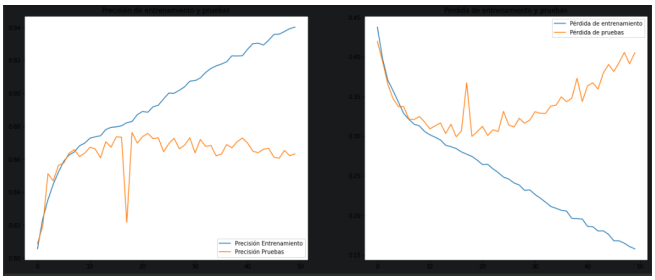

Fig. 9 Gráficas de precisión modelo convolucional batch size 100

La mejor arquitectura es usando redes convolucionales y dropout. Además de que sin importar el batch todos tienen 87.xx% de accuracy, la diferencia es en qué época llegan a la mejor versión, pareciendo que un batch de 32 es la mejor opción.

Por otro lado, para el set de datos del amazonas se desarrollaron tres modelos de regresión lineal:

En primer lugar, se tomaron los años y área deforestada por región. Los resultados al correr los modelos fueron se pueden observar a continuación.

TABLA 3.
R² y RMSE Regresión lineal año, deforestación por región

Región	R²	RMSE (km²)
Acre	0.0047	166.6048
Amazonas	0.1152	273.8208
Amapa	0.2862	19.2705
Maranhao	0.6332	185.7088
Mato Grosso	0.4371	2135.9187
Para	0.6047	1184.5386
Rondonia	0.2980	765.4515
Roraima	0.0009	145.7232
Tocantins	0.5403	41.5157

Ningún R² logró superar 0.65 además no se puede establecer una correlación fuerte entre el año y el área deforestada por zona.

En segundo lugar, se implementó un modelo de regresión lineal con el área total deforestada y el año. En este caso se obtuvo un R² de 0.5403, que nos lleva a pensar en otras alternativas para estimar la deforestación en el tiempo.

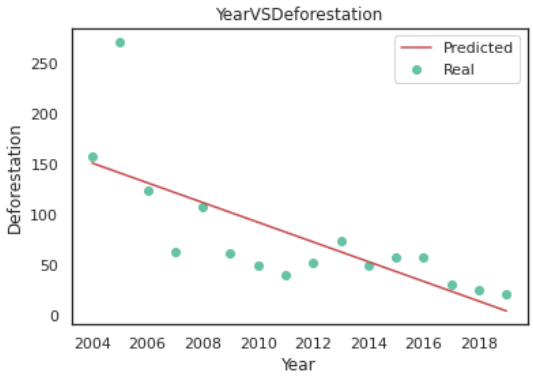

Fig. 10 Regresión lineal área deforestación total por año

Luego de identificar que las métricas no fueron las deseadas se decide proceder a explorar con diferentes modelos para buscar resultados mejores. Entre los modelos a implementar se decide utilizar la regresió lasso, regresión ridge, árboles de decisión y random forest. En la tabla a continuación se presentan los resultados.

TABLA 4.
RMSE (km²) año, deforestación por región por modelos

Región	Lasso	Ridge	DT	RF
Brasil	4220.1	4365.8	27.3	16.4
Acre	183.7	179.5	236.5	44.2
Amazonas	188.1	308.7	466.6	82.8
Amapa	13.9	16.9	9.4	6.3
Maranhao	156.6	143.4	43.3	70.3
Mato Grosso	2223.8	7039.5	91.5	585.2
Para	1221.4	1125.9	1006.5	422.8
Rondonia	832.1	785.32	220.6	132.7
Roraima	140.4	139.25	232.7	54.5
Tocantins	35.42	37.4	22.3	18.4

DT: Árbol de decisión, RF: Random Forest

Después de haber realizado este trabajo para los modelos se decide implementar el random forest al ver que obtiene los mejores resultados en cuanto a RMSE para casi todos los modelos. Las únicas excepciones fueron las regiones de Maranhao y Mato Grosso, y se procedió a implementar el random forest para todas las regiones.

Finalmente, se planteó un último modelo donde trabajamos con la cantidad de puntos de incendio y el área total deforestada. Se obtuvo un R² de 0.6952 y RMSE de 3267 km² que nos permite decir que hay espacio para mejorar este modelo.

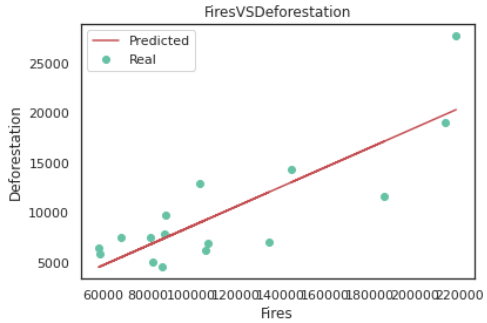

Fig. 11 Regresión lineal área deforestación total por año

Al identificar este espacio de mejora se decidió explorar la posibilidad de ampliar el dataset de 16 registros de deforestación anual a 192 registros usando los incendios por mes para segmentar el área deforestada anual.

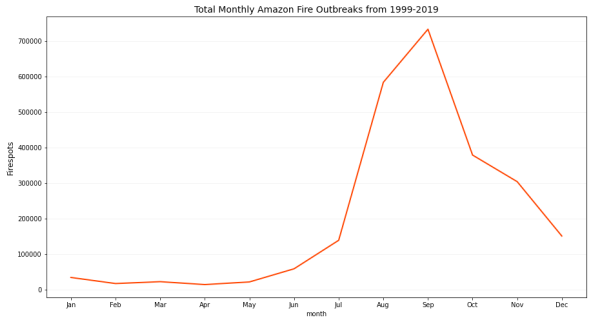

Fig. 12 Total de incendios por mes

Observando la figura 12, los incendios se presentan en mayor cantidad en la segunda mitad del año, así que se generaron datos a partir de la fórmula *(incendios mes/incendios año) * total deforestación anual.* De esta manera se aumentaron los registros y volvimos a generar un nuevo modelo de regresión lineal.

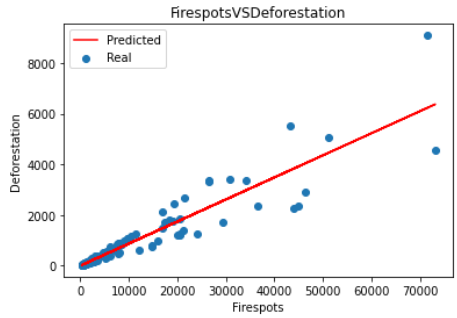

Fig. 13 Regresión lineal área deforestación e incendios

Para el nuevo modelo con mayor cantidad de datos se obtuvo una mejora esperada, se manejo la partición de datos en entrenamiento validación y prueba. En la tabla 4 a continuación se presentan los resultados obtenidos.

TABLA 5.
Resultados Regresión Lineal

	R^2	RMSE (km^2)
Entrenamiento	0.8524	516.1171
Validación	0.9468	276.6133
Prueba	0.9394	257.6945

Podemos observar que si se obtuvo una mejora significativa al aumentar la cantidad de registros, se evidencia un aumento en el R^2 y una disminución en el RSME en el conjunto de entrenamiento. El modelo presenta un buen ajuste en los conjuntos de validación y prueba al mejorar las métricas de la misma manera.

Como último paso, una vez terminados los modelos tanto en su entrenamiento como en su evaluación, se procedió a desarrollar el despliegue en una aplicación web utilizando el framework Dash para realizar una interfaz gráfica con la cual los usuarios finales pudieran interactuar con los modelos y obtener las respectivas predicciones. La aplicación se compone de las siguientes tres pestañas:

Fig. 14 Pestaña de detección a partir de imagen.

Aquí, los usuarios podrán cargar una imagen en formato .jpg o .png y obtener como resultado la clasificación de sí la imagen presenta o no deforestación.

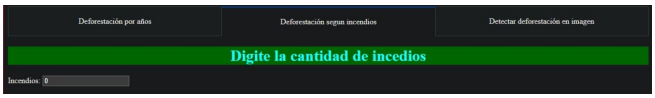

Fig. 15 Pestaña predicción área deforestada a partir de incendios.

En este apartado, los usuarios podrán obtener la predicción de deforestación con base al número de incendios en un área determinada.

Fig. 16 Pestaña regresión a partir de zona y año.

Finalmente, los usuarios también tendrán la posibilidad de obtener la respectiva predicción sobre deforestación ingresando una zona y año especifico.

V. CONCLUSIONES Y OPORTUNIDADES DE MEJORA

Se obtiene un modelo con una precisión aceptable de 87% en comparación a una línea base aleatoria de 38.8%. Esta línea base se obtuvo a partir del conjunto de pruebas en donde 12566 de 32384 son imágenes deforestadas. Sin embargo, no se descarta que para futuras iteraciones esta exactitud pueda mejorar llegando a valores de entre 90% y 95%.

Usar la convolución y el drop out es de suma importancia, permitiendo un mejor aprendizaje. También vemos que usando 3 capas de convolución es suficiente, agregar más empeora el aprendizaje.

La mayoría de algoritmos después de 25 épocas de entrenamiento termina sobre entrenando el modelo.

Para los modelos de regresión lineal obtuvimos resultados de R^2 para las regiones bajos y se nos dificulta establecer una relación por región anual para hacer predicciones. En cuanto a los totales anuales también podemos decir lo mismo. Así que podemos mejorar la predicción en el tiempo implementando otros modelos ARIMA y SARIMA que cuentan con mejor desempeño cuando de temporalidad se habla. Al implementar más modelos logramos reducir el RMSE de 7 de 9 regiones lo cuál fue un resultado positivo para la posterior implementación en el despliegue.

Por otro lado, al trabajar con los incendios llegamos a un R^2 aceptable con posibilidades de mejorar. Al tener pocos registros (16 años) fue difícil establecer este modelo. Sin embargo, al aumentar la cantidad de registros de 16 a 192 utilizando el

porcentaje de la cantidad de puntos de incendios por mes se logra ver una mejora esperada. En últimas instancias podemos afirmar que el modelo desarrollado a partir del dataset generado tiene métricas exitosas dentro de nuestro marco de trabajo. Generar mayor cantidad de registros acorde a los meses fue una decisión acertada y se evidencia en los resultados obtenidos. Podemos decir que manejamos los datos de manera correcta para aumentar las métricas, aprendimos de primera mano que el volumen de datos juega un rol importante a la hora de entrenar nuestros modelos.

Para cerrar, aún hay espacio para mejorar, siempre está presente en cualquier trabajo, dentro de nuestro proyecto se podría expandir a otras regiones del amazonas y del planeta tierra. También buscar vincular las imágenes a los datos de incendios o área deforestada deja camino para elevar el nivel del proyecto. Explorar diferentes modelos externos al curso como ARIMA o SARIMA para las predicciones en el tiempo o técnicas más avanzadas de computer vision aportarían a este objetivo.

Ahora bien, en cuanto a la resolución de nuestras preguntas de investigación se les pudieron dar las siguientes respuestas con base al trabajo realizado:

1. ¿Cuáles son las características relevantes de imágenes satelitales que contengan áreas deforestadas y no deforestadas?

La respuesta a esta pregunta va a depender mucho del tipo de dataset que esté siendo utilizado. Para el caso abordado en esta solución, con el dataset de imágenes satelitales, cada imagen tiene un tag asociado y el grupo de etiquetas asociadas a deforestación se comprende por:
"artisinal_mine","conventional_mine","road","selective_logging","slash_burn"
Ahora bien, los algoritmos de redes neuronales son los encargados de con base en los tags y los valores de los píxeles de las imágenes, llevar a cabo el aprendizaje para hacer las predicciones.

2. ¿Qué áreas geográficas presentan mayores niveles de deforestación?

Con base a los hallazgos de nuestra investigación, se encontró que la zona amazónica es la región del planeta que presenta los mayores índices de deforestación, y enfocándonos en Brasil, las zonas que presentan mayores de deforestación son las regiones de Pará, Mato Grosso y Rondonia tal como se puede apreciar en el siguiente gráfico.

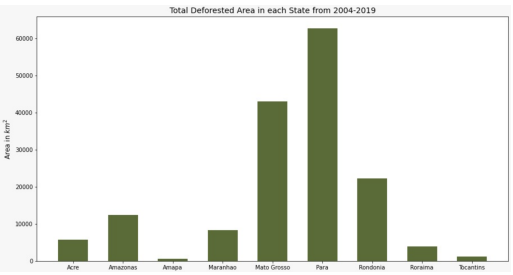

Fig. 17 Totales de áreas deforestadas en brasil de 2004 a 2019

3. ¿Cómo determinar si hay o no deforestación en una determinada zona geográfica?

Con base a los datos recolectados en nuestra investigación, se puede afirmar que es posible determinar si una zona ha sufrido un proceso de deforestación haciendo predicciones con base a variables como, el número de incendios, la ubicación geográfica de la zona o si se presenta minería artesanal entre otros factores.

5. ¿Cómo podemos predecir la evolución de la deforestación a partir de los datos actuales de una zona geográfica?

Al ser esto un problema netamente de predicción (no de clasificación, ni recomendación, etc) uno de los posibles acercamientos es utilizando modelos de regresión, pues a partir de estas técnicas es posible hacer estimaciones a partir del modelamiento de variables que ya se tienen (como las mencionadas anteriormente) para descubrir las relaciones entre las variables dependientes y las predictoras y modelar una función matemática que puede ser más o menos compleja para hacer estas estimaciones.

VI. REFERENCIAS

[1] M. Weisse y L. Goldman. "La Destrucción de los Bosques Primarios Aumentó un 12 % de 2019 a 2020". Global Forest Watch. https://www.globalforestwatch.org/blog/es/data-and-research/datos-globales-de-perdida-de-cobertura-arborea-2020/ (accedido el 22 de octubre de 2022).

[2] André Ferreira, Detecting deforestation from satellite images, a full stack deep learning project (2021), Towards Data Science

[3] Salma B, Brazilian Amazon Rainforest Degradation Analysis (2021)

[4] IBM, ¿Qué es el Machine Learning? (s.f.) https://www.ibm.com/co-es/analytics/machine-learning

[5] Frank L. ¿Cómo aprenden las redes neuronales? (2019). https://learn.microsoft.com/es-es/archive/msdn-magazine/2019/april/artificially-intelligent-how-do-neural-networks-learn