Agricultura Inteligente: Predicción de Rendimiento Mediante Aprendizaje Profundo

Juan Joseph Mora Estrada

Cc. 1193329412

Bootcamp de Inteligencia Artificial

Julio de 2025

I. Resumen

El presente trabajo propone un modelo predictivo basado en redes neuronales profundas para estimar el rendimiento de cultivos agrícolas en el departamento del Huila, Colombia. Con base en registros históricos provenientes de fuentes como el Ministerio de Agricultura y Desarrollo Rural y variables climáticas extraídas de la plataforma NASA POWER, se construyó una base de datos que combina información agrícola, geográfica y meteorológica entre los años 2006 y 2023. Se desarrolló un sistema de red neuronal multitarea que primero estima las condiciones climáticas en función del municipio, cultivo, periodo y área sembrada, y posteriormente predice el rendimiento esperado en toneladas por hectárea.

El modelo fue entrenado sobre datos normalizados y codificados, logrando métricas de rendimiento destacables, con un coeficiente de determinación (R^2) de 0.92, un error cuadrático medio (RMSE) de 1.4 y un error absoluto medio (MAE) de 0.91. Se construyó además una interfaz interactiva en Dash y Plotly que permite a los usuarios seleccionar parámetros y obtener predicciones acompañadas de visualizaciones como series temporales, mapas de ubicación y comparaciones históricas por cultivo.

Este enfoque demuestra el potencial del aprendizaje profundo para apoyar la toma de decisiones en el sector agrícola, especialmente en contextos donde el cambio climático y la escasez de recursos dificultan la planificación eficiente. Finalmente, se contemplan lineamientos éticos relacionados con la recolección, el uso responsable de los datos y las implicaciones sociales de la predicción agrícola automatizada.

Palabras clave:

Agricultura, redes neuronales, predicción de rendimiento, Huila, Dash, aprendizaje automático, variables climáticas.

II. Introducción

La agricultura constituye una de las principales actividades económicas del departamento del Huila, Colombia, siendo fuente fundamental de ingresos, empleo y seguridad alimentaria. No obstante, este sector enfrenta desafíos significativos derivados de la variabilidad climática, el uso ineficiente de los recursos y la limitada implementación de tecnologías avanzadas para la toma de decisiones. En un escenario donde el cambio climático afecta progresivamente los patrones de producción agrícola, la necesidad de herramientas predictivas y automatizadas resulta esencial para optimizar el rendimiento y mitigar riesgos.

Gracias al acceso a grandes volúmenes de datos y al avance de la inteligencia artificial (IA), hoy es posible construir modelos capaces de anticipar el comportamiento de variables agrícolas y climáticas con un alto grado de precisión. En este trabajo se propone un sistema de predicción de rendimiento agrícola basado en redes neuronales profundas, diseñado para estimar el rendimiento en toneladas por hectárea a partir de variables como el municipio, el cultivo, el periodo del año y el área sembrada.

El modelo se entrena utilizando información histórica del rendimiento de cultivos recopilada por el Ministerio de Agricultura y Desarrollo Rural, así como datos climáticos obtenidos mediante la plataforma NASA POWER. La propuesta contempla además una interfaz desarrollada en Dash, la cual permite a los usuarios interactuar con el sistema, realizar predicciones y visualizar resultados mediante gráficos dinámicos y mapas geográficos.

Este informe detalla cada una de las etapas del proyecto, incluyendo la exploración y limpieza de datos, el diseño del modelo, su entrenamiento, evaluación, y una reflexión ética sobre el uso de datos en contextos agrícolas.

III. Metodología

El desarrollo del sistema de predicción de rendimiento agrícola se estructuró en varias etapas, desde la recolección y depuración de datos hasta la construcción, entrenamiento y evaluación del modelo. A continuación, se describe el procedimiento seguido:

1. Recolección de datos

Se integraron dos fuentes principales de datos:

1. Datos agrícolas: obtenidos del Ministerio de Agricultura y Desarrollo Rural, los cuales incluyen información sobre cultivos, áreas sembradas y cosechadas, producción y rendimiento (t/ha) para varios municipios del Huila, un dataset venía desde el año 2006 hasta el 2018 y el otro desde el año 2019 hasta el 2023, organizados en periodos "a" y "b". Se juntaron ambos y su contenido venia de la siguiente manera:

	Código Dane departamento	Departamento	Código Dane municipio	Municipio	Grupo cultivo	Subgrupo	Cultivo	Desagregación cultivo	Año	Periodo	Área sembrada	Área cosechada	Producción	Rendimiento
157815	66	RISARALDA	66001	PEREIRA	HORTALIZAS	PEPINO	PEPINO COHOMBRO	PEPINO COHOMBRO	2006	2006B	2.0	2.0	16.0	8.0
157814	41	HUILA	41807	TIMANA	HORTALIZAS	PEPINO	PEPINO COHOMBRO	PEPINO COHOMBRO	2006	2006B	2.0	2.0	22.0	11.0
157813	41	HUILA	41020	ALGECIRAS	HORTALIZAS	PEPINO	PEPINO COHOMBRO	PEPINO COHOMBRO	2006	2006B	2.0	2.0	40.0	20.0
157812	95	GUAVIARE	95200	MIRAFLORES	HORTALIZAS	PEPINO	PEPINO COHOMBRO	PEPINO COHOMBRO	2006	2006B	2.0	2.0	26.0	13.2
157806	17	CALDAS	17486	NEIRA	HORTALIZAS	PEPINO	PEPINO COHOMBRO	PEPINO COHOMBRO	2006	2006B	4.0	4.0	102.0	25.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41931	52	Nariño	52320	Guaitarilla	Cereales	Cereales	Maíz	Maíz blanco tradicional	2023	2023B	400.0	0.0	0.0	0.0

Fig

1. Datos de rendimiento

2. Datos climáticos: extraídos de la plataforma NASA POWER, que proporciona variables como temperatura, humedad relativa, humedad del suelo, radiación solar, velocidad del viento y precipitación. El proceso para la adquisición de datos fue así:

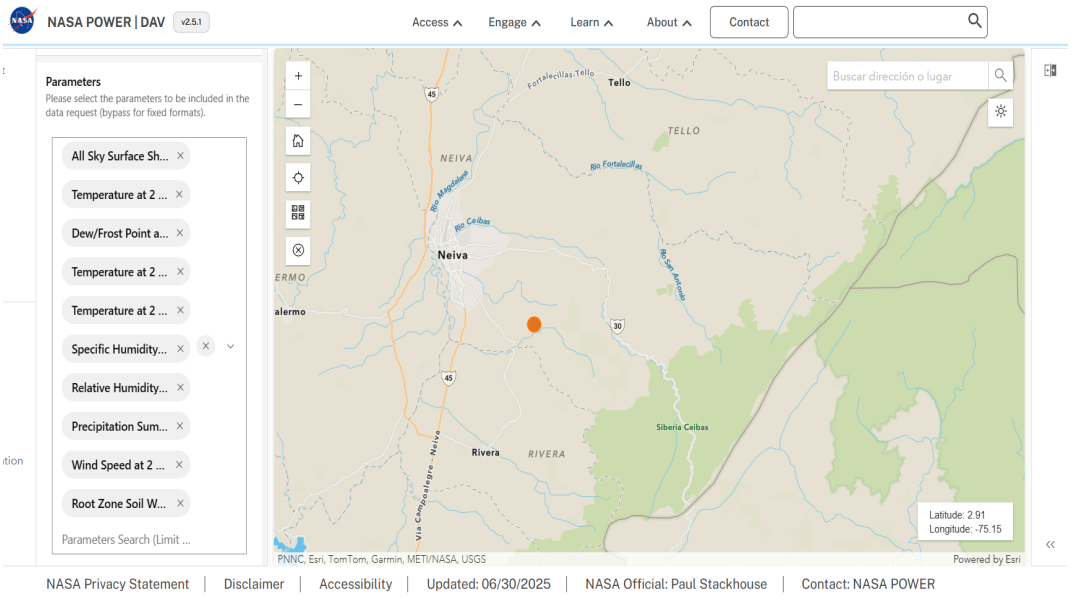

Fig 2. Interfaz de la NASA

Luego de ingresar a la página de la NASA, se seleccionan las coordenadas de los 5 municipios escogidos con las variables que se deseen trabajar y se descargan en formato CSV, estos vienen de la siguiente manera:

index	PARAMETER,YEAR,JAN,FEB,MAR,APR,MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC,ANN
0	ALLSKY_SFC_SW_DWN,2006,16.63,17.86,15.24,16.17,16.99,15.94,16.02,16.76,17.25,15.93,16.42,16.31,16.45
1	ALLSKY_SFC_SW_DWN,2007,18.16,20.45,15.53,16.79,16.61,15.41,18.17,15.24,16.36,16.72,16.75,15.32,16.77
2	ALLSKY_SFC_SW_DWN,2008,16.19,16.82,16.59,16.95,16.3,16.0,16.3,16.9,17.13,16.5,15.24,16.13,16.42
3	ALLSKY_SFC_SW_DWN,2009,15.29,17.28,14.93,15.4,17.19,16.35,16.26,16.44,19.12,17.74,16.89,18.57,16.78
4	ALLSKY_SFC_SW_DWN,2010,19.79,17.56,16.82,16.77,17.06,16.47,16.46,15.43,16.11,16.22,15.2,15.37,16.6
5	ALLSKY_SFC_SW_DWN,2011,19.26,15.33,16.25,16.41,15.97,16.74,16.41,18.0,17.13,16.45,16.21,16.43,16.73
6	ALLSKY_SFC_SW_DWN,2012,16.66,18.04,16.05,15.6,17.0,17.05,16.0,16.04,15.58,16.51,16.5,16.57,16.46
7	ALLSKY_SFC_SW_DWN,2013,18.24,15.49,15.85,17.09,15.19,17.97,16.86,15.87,17.49,17.81,15.15,16.0,16.59
8	ALLSKY_SFC_SW_DWN,2014,17.16,16.74,15.56,17.03,16.97,16.31,16.71,15.95,17.81,16.09,16.67,17.78,16.73
9	ALLSKY_SFC_SW_DWN,2015,18.1,17.24,16.33,16.5,16.29,15.15,15.84,15.32,18.15,16.63,16.01,18.5,16.67
10	ALLSKY_SFC_SW_DWN,2016,18.76,16.18,15.56,15.98,16.51,16.29,16.06,16.12,17.36,17.43,15.73,16.4,16.54
11	ALLSKY_SFC_SW_DWN,2017,15.51,18.57,16.25,17.65,17.79,16.65,16.0,18.05,17.93,15.86,16.82,16.72,16.97
12	ALLSKY_SFC_SW_DWN,2018,16.81,15.45,17.3,14.94,16.12,16.05,15.88,16.65,17.57,16.87,16.06,18.08,16.49
13	ALLSKY_SFC_SW_DWN,2019,16.78,16.88,14.37,15.98,16.45,15.83,16.76,15.51,16.61,16.45,15.88,16.03,16.12
14	ALLSKY_SFC_SW_DWN,2020,18.3,18.84,16.14,15.91,15.52,16.35,16.32,16.21,17.87,17.15,15.48,16.91,16.74
15	ALLSKY_SFC_SW_DWN,2021,17.7,15.47,14.24,16.32,17.11,16.31,16.11,16.26,17.96,17.44,16.51,15.8,16.44
16	ALLSKY_SFC_SW_DWN,2022,16.01,14.97,15.52,15.68,16.3,16.59,16.06,17.04,17.39,16.77,16.01,16.97,16.29
17	ALLSKY_SFC_SW_DWN,2023,16.38,18.31,14.43,15.76,17.22,17.7,15.74,17.88,18.21,17.52,16.36,16.66,16.84
18	ALLSKY_SFC_SW_DWN,2024,20.87,17.29,15.74,17.79,17.59,16.76,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0
19	ALLSKY_SFC_SW_DWN,2025,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0,-999.0
20	GWETROOT,2006,0.65,0.61,0.62,0.65,0.62,0.6,0.56,0.53,0.53,0.55,0.62,0.62,0.6
21	GWETROOT,2007,0.58,0.54,0.55,0.63,0.61,0.57,0.54,0.53,0.53,0.6,0.62,0.61,0.58
22	GWETROOT,2008,0.6,0.63,0.71,0.69,0.71,0.72,0.67,0.61,0.61,0.64,0.68,0.72,0.67
23	GWETROOT,2009,0.64,0.62,0.64,0.72,0.63,0.57,0.54,0.54,0.52,0.54,0.54,0.57,0.59
24	GWETROOT,2010,0.54,0.52,0.53,0.61,0.7,0.71,0.65,0.56,0.55,0.6,0.75,0.81,0.63

Show 25 per page

Fig 3. Formato de datos NASA

2. Preprocesamiento y limpieza

Se realizaron los siguientes pasos de limpieza y transformación en ambos datasets:

Dataset de rendimiento:

1. Normalizar el texto: Al unir los datasets de rendimiento que venían en diferentes grupos de años, sus datos no venían uniformes, un dataset contiene mayúsculas y tildes y el otro no, por lo tanto se realizó una normalización de su contenido para una mayor facilidad de manejo.

2. Eliminación de columnas innecesarias al no aportar información adicional al cultivo como lo eran Grupo cultivo y Desagregación cultivo

3. Filtrado de municipios con mayor cantidad de datos para asegurar robustez en el entrenamiento: Se realizó un value_count() para ver cual era el departamento con mayor cantidad de datos y que municipios contenía:

index	Departamento	Municipio	count
0	Huila	Garzon	947
1	Valle Del Cauca	Roldanillo	929
2	Huila	Gigante	926
3	Valle Del Cauca	Yumbo	914
4	Huila	Neiva	893
5	Huila	La Plata	892
6	Huila	Algeciras	889

Fig 4. Municipios mas representativos

4. Exclusión de cultivos con baja frecuencia de registro: Existían cultivos con muy baja cantidad de información por año, lo cual iba a entorpecer el modelo al no contar con suficiente información, por tal razón los cultivos con menos de 10 datos fueron borrados:

index	Departamento	Municipio	Cultivo_final
300	Huila	La Plata	Limon (Limon Tahiti)
301	Huila	La Plata	Limon (Limon Pajarito)
302	Huila	Garzon	Mandarina (Mandarina Arrayana Y/o Oneco)
303	Huila	Garzon	Naranja (Naranja Valencia)
304	Huila	Gigante	Limon (Limon Pajarito)
305	Huila	Gigante	Naranja (Naranja Valencia)
306	Huila	Neiva	Pepino Cohombro
307	Huila	Gigante	Limon (Limon Tahiti)
308	Huila	La Plata	Mandarina (Mandarina Arrayana Y/o Oneco)
309	Huila	La Plata	Papa (Papa Criolla)
310	Huila	Garzon	Limon (Limon Pajarito)
311	Huila	La Plata	Naranja (Naranja Valencia)
312	Huila	Neiva	Naranja (Naranja Valencia)
313	Huila	Neiva	Mandarina (Mandarina Arrayana Y/o Oneco)
314	Huila	Algeciras	Limon (Limon Pajarito)
315	Huila	Neiva	Gulupa
316	Huila	Garzon	Limon (Limon Tahiti)
317	Huila	Algeciras	Mandarina (Mandarina Arrayana Y/o Oneco)

Fig 5. Cultivos con menos datos

5. Eliminación de registros nulos o incompletos: Antes de realizar los filtros por municipio y cultivo la cantidad de datos nulos era considerable (3.433 datos), pero luego de los filtros los nulos se redujeron (12 datos) y se tomó la decisión de eliminarlos

6. Asignar periodos a años no especificados: Existían datos en donde el año no venía especificado por periodo, entonces se le asignó uno de forma aleatoria por año, bien sea periodo “a” o “b”, de no hacerlo, se iban a generar datos nulos al unirlo con el dataset climático:

icipio	Municipio	Año	Periodo	Área sembrada	Área cosec
41396	La Plata	2007	2007	4.0	
41306	Gigante	2007	2007	4.0	
41020	Algeciras	2007	2007	9.0	
41020	Algeciras	2007	2007	4152.0	3
41306	Gigante	2007	2007	4412.0	4
...	...	...	...	...	
41298	Garzon	2007	2007	65.0	
41020	Algeciras	2007	2007	83.0	
41396	La Plata	2007	2007	11.0	
41306	Gigante	2007	2007	18.0	
41001	Neiva	2007	2007	20.0	

Fig 6. Años sin periodo

Dataset de condiciones climáticas:

7. Reconstrucción de formato: Como se mostró en la fig 3, el formato en que se descargan los datos climáticos no era el adecuado para trabajar con ellos. Por lo tanto, se reestructuró dejando las variables como columnas, añadiendo una nueva columna de municipio y realizando un promedio del primer semestre para periodo “a” y el segundo para periodo “b” con el fin de poderlo unir con el dataset de rendimiento. Finalmente se unieron los datos de los 5 municipios y quedó de la siguiente manera:

PARAMETER	Año_periodo	Municipio	ALLSKY_SFC_SW_DWN	GWETROOT	PRECTOTCORR_SUM	QV2M	RH2M	T2M	T2MDEW	T2M_MAX	T2M_MIN	WS2M
0	2006a	Neiva	16.471667	0.625000	100.278333	12.940000	78.856667	19.496667	15.250000	26.981667	13.005000	1.506667
1	2006b	Neiva	16.448333	0.568333	68.656667	12.115000	71.780000	20.240000	14.178333	28.950000	13.223333	1.671667
2	2007a	Neiva	17.158333	0.580000	66.708333	12.506667	73.821667	20.258333	14.671667	28.681667	13.596667	1.630000
3	2007b	Neiva	16.426667	0.571667	73.280000	12.040000	71.608333	20.151667	14.090000	29.438333	13.500000	1.601667
4	2008a	Neiva	16.475000	0.676667	148.408333	12.938333	80.961667	18.971667	15.243333	25.966667	12.856667	1.406667
...	...	...	...	...	...	...	...	...	...	...	...	...
195	2023b	La Plata	16.013333	0.690000	81.998333	11.773333	84.123333	15.416667	12.423333	22.715000	9.695000	1.923333

Fig 7. Formato de la NASA organizado

8. Se renombraron las columnas para un mejor entendimiento de los datos.

9. Ajuste a valores más reales: El mapa en la página de la NASA manejaba una resolución o ventana en donde los datos descargados en zonas cercanas tendrían el mismo contenido. Por lo tanto, se tomaron los datos en coordenadas un poco más alejadas del municipio objetivo, con el fin de que el contenido de los datos fuera distinto en cada uno de ellos, pero esto conllevaba hacerle un ajuste a las variables con valores más reales de cada municipio.

10. Datos erróneos: El dataset de clima se descargó desde 2006 al año actual 2025, pero los datos de 2024 y 2025 al ser un poco recientes venían erróneos con valores negativos como -999, al inicio se pensó realizar una imputación, pero como el dataset de rendimiento iba hasta 2023 esos últimos 2 años igual se perderán, por lo tanto se dejaron asi.

Dataset final:

11. Se unieron ambos datasets ya limpios y procesados, mediante la columna de Año_periodo y Municipio, quedando así:

	Código Dane departamento	Departamento	Código Dane municipio	Municipio	Año_periodo	Área sembrada	Área cosechada	Producción	Rendimiento	Cultivo_final	Radiacion_solar	Humedad_suelo	Precipitacion	Humedad_especifica	Hum
0	41	Huila	41020	Algeciras	2006b	2.0	2.0	40.00	20.00	Pepino Cohombro	16.20	0.67	107.56	14.81	
1	41	Huila	41298	Garzon	2006b	40.0	40.0	91.00	2.28	Arveja	16.67	0.60	129.03	17.60	
2	41	Huila	41001	Neiva	2006b	80.0	80.0	140.00	1.75	Arveja	16.82	0.65	67.36	12.13	
3	41	Huila	41306	Gigante	2006b	100.0	100.0	193.00	1.93	Arveja	15.28	0.62	108.67	14.41	
4	41	Huila	41020	Algeciras	2006b	120.0	120.0	273.00	2.28	Arveja	15.84	0.67	103.37	15.04	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3796	41	Huila	41306	Gigante	2023a	13.5	13.5	132.30	9.80	Pitahaya	15.61	0.61	126.76	15.60	
3797	41	Huila	41306	Gigante	2023a	25.5	25.3	442.75	17.50	Tomate	14.94	0.67	129.64	15.75	
3798	41	Huila	41306	Gigante	2023b	8.0	8.0	137.60	17.20	Tomate	15.54	0.44	73.99	16.03	

Fig 8. Dataset final

12. Transformación de la variable categórica Cultivo_final mediante codificación One-Hot.

13. Conversión de la variable Año_periodo mediante LabelEncoder para preservar su naturaleza ordinal (2006a = 0, 2006b = 1...).
14. Conversión a logaritmo de las variables Área sembrada, Área cosechada, Producción y Rendimiento.

15. Normalización de variables numéricas (Área sembrada, código DANE, periodo) con StandardScaler para mejorar el desempeño del modelo.

3. Construcción del modelo

En la fase inicial del desarrollo del proyecto se implementó un modelo de regresión utilizando Random Forest Regressor de la biblioteca scikit-learn. En un primer intento, el modelo fue entrenado con las variables Área sembrada, Cultivo, Año_periodo, así como Producción y Área cosechada como entradas, y el Rendimiento agrícola como variable de salida.

Sin embargo, este enfoque presentó varios inconvenientes:

1. Presencia de valores atípicos: Al trabajar con las variables sin transformar, los rendimientos extremos afectaban negativamente el desempeño del modelo, generando predicciones inestables.

2. Fuga de información (data leakage): Al incluir Producción y Área cosechada como variables predictoras, se incorporaba información que en la práctica solo está disponible luego de la cosecha, lo que hacía inviable el uso del modelo para predicción futura.

3. Uso de logaritmos: Se aplicó una transformación logarítmica a las variables de entrada y salida para estabilizar la varianza y reducir el impacto de los outliers. Esta mejora permitió obtener mejores métricas sobre los datos históricos ($R^2 \approx 0.83$ y RMSE ≈ 0.27); sin embargo, el modelo continuaba dependiendo de variables futuras, por lo que persistía el problema de data leakage.

Finalmente, se probó un modelo de Random Forest sin las variables Producción y Área cosechada, pero con logaritmo aplicado al rendimiento. Esta versión, aunque válida desde el punto de vista de predicción futura, mostró un rendimiento significativamente inferior, ya que el modelo perdió capacidad predictiva al no contar con suficientes variables explicativas.

Dadas estas limitaciones, se optó por una segunda estrategia basada en redes neuronales multitarea, que ofrecieran mayor capacidad para modelar relaciones no lineales y permitieran predecir primero las condiciones climáticas esperadas y, con base en ellas, estimar el rendimiento. Esta arquitectura demostró ser más adecuada para escenarios de predicción futura, incluso cuando no se cuenta con variables como Producción o Área cosechada.

3.1 Arquitectura final: red neuronal multitarea

Se implementó una red neuronal debido a la no linealidad de las variables, con dos salidas:

- **Salida 1:** Predicción de 10 variables climáticas (Temperatura, Precipitación, Humedad relativa, entre otras).
- **Salida 2:** Predicción del rendimiento agrícola (variable continua en t/ha), a partir de las salidas climáticas.

Características del modelo:

- Entradas: 47 variables, incluyendo:

    Código DANE del municipio (numérica)

    Año_periodo codificado con LabelEncoder

    Área sembrada

    Cultivo codificado con OneHotEncoder

- Preprocesamiento: escalamiento mediante StandardScaler aplicado a las variables numéricas.

Estructura de la red neuronal:

- Capa de entrada con 47 neuronas.
- Bloque climático:

    Capa densa con 256 neuronas (ReLU)

    Capa densa con 128 neuronas (ReLU)

    Salida climática: 10 neuronas lineales (una por variable climática)

- Bloque de rendimiento:

    Toma como entrada la salida del bloque climático

    Capa densa de 64 neuronas (ReLU)

    Capa densa de 32 neuronas (ReLU)

    Salida final: 1 neurona lineal (rendimiento en toneladas por hectárea)

Entrenamiento:

- Función de pérdida: Error cuadrático medio (MSE) para ambas salidas.
- Ponderación de pérdidas: 0.2 para clima y 0.8 para rendimiento, enfocando el entrenamiento en la predicción final.
- Optimización: Adam con batch_size=64 y early stopping con paciencia de 10 épocas.
- División del dataset: 80% entrenamiento, 20% prueba.

4. Evaluación

El conjunto de datos se dividió en entrenamiento y prueba con una proporción 80/20. Se utilizaron métricas como el coeficiente de determinación (R^2), el error cuadrático medio (RMSE) y el error absoluto medio (MAE) para evaluar el rendimiento del modelo. Además, se generaron gráficos de dispersión entre valores reales y predichos para analizar visualmente la precisión de las predicciones.

5. Interfaz de usuario

Finalmente, se construyó una aplicación interactiva con Dash (Plotly), que permite a los usuarios ingresar información sobre el municipio, cultivo, periodo y área sembrada. La app retorna la predicción del rendimiento agrícola, comparaciones con el promedio histórico del cultivo y representaciones visuales como líneas de tiempo y mapas de ubicación.

IV. Resultados

La implementación del modelo de red neuronal multitarea permitió abordar con mayor precisión el problema de predicción del rendimiento agrícola en escenarios futuros, eliminando la dependencia de variables no disponibles antes de la cosecha como la producción o el área cosechada.

El modelo fue entrenado utilizando como entrada las variables Código DANE del municipio, Año_periodo (codificado), Área sembrada (en escala real), y Cultivo (mediante codificación one-hot). A partir de esta información, el modelo primero estima las condiciones climáticas esperadas (temperatura, humedad, precipitación, entre otras) y luego utiliza dichas predicciones como entrada para calcular el rendimiento agrícola esperado (en toneladas por hectárea).

Los principales resultados obtenidos fueron:

- R^2 (coeficiente de determinación): 0.92
- RMSE (Raíz del error cuadrático medio): 1.38 t/ha
- MAE (Error absoluto medio): 0.91 t/ha

Estas métricas indican un desempeño muy adecuado, especialmente considerando que el modelo no utiliza variables críticas como producción ni área cosechada para evitar el "Data leakage". Además, el enfoque multitarea permitió capturar mejor la dinámica climática que influye en el rendimiento.

A continuación, se presenta el resultado de los modelos que muestran la relación entre los valores reales y predichos del rendimiento agrícola. Una buena distribución alrededor de la diagonal indica un alto grado de precisión:

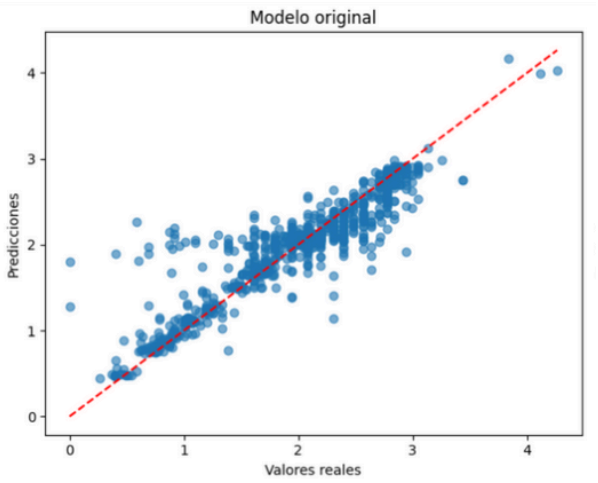

Fig 9. Predicción vs Real (Random Forest)

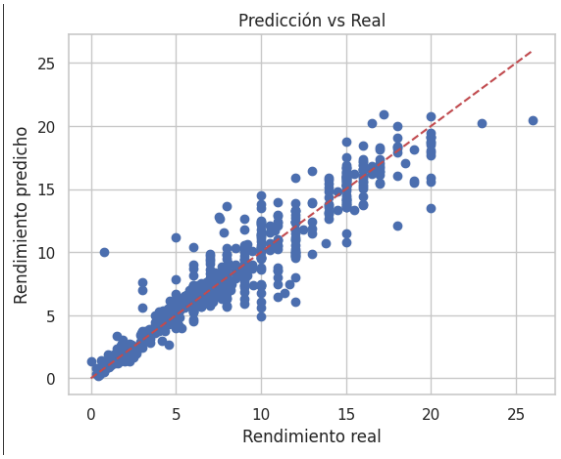

Fig 10. Predicción vs Real (Redes neuronales)

En comparación con el modelo previo basado en Random Forest, la red neuronal mostró una mejora en capacidad predictiva, mayor estabilidad frente a variaciones de entrada, y mejor adaptación a escenarios futuros. Además, permite un flujo de inferencia más lógico: del contexto (año, cultivo, área sembrada, ubicación) → al clima esperado → al rendimiento estimado. Aquí se muestra una tabla comparativa con el Random Forest y las redes neuronales:

Modelo	R^2	RMSE	MAE	Observaciones
Random Forest (sin log)	0.85	0.45	0.24	Afectado por valores atípicos (outliers)
Random Forest (con log) + prod/cosecha	0.989	0.27	0.16	reciso, pero dependiente de variables no disponibles a futuro
Random Forest (sin prod/cosecha)	0.83	0.27	0.16	Menor precisión al remover variables clave
Red neuronal multitarea (final)	0.92	1.38	0.91	Apta para predicción futura, sin data leakage

Interfaz de usuario

Como producto final, se construyó una interfaz web interactiva con Dash, que permite al usuario seleccionar el municipio, el cultivo, el período (ej. 2026a), y el área sembrada, para obtener la predicción estimada del rendimiento en toneladas por hectárea. La aplicación además incluye gráficos de comparación con el promedio histórico, evolución temporal y localización geográfica del municipio. Esta herramienta está diseñada para ser utilizada por agricultores, entidades gubernamentales y técnicos del sector agropecuario.

Fig 11. Interfaz de usuario con Dash

V. Ética en el análisis de datos

El desarrollo del presente proyecto consideró principios éticos fundamentales a lo largo de todas las etapas del proceso, desde la recolección de los datos hasta el uso del modelo de predicción.

Privacidad y anonimato

La base de datos utilizada no contiene información sensible ni de carácter personal. Todos los datos provienen de fuentes públicas como Datos Abiertos Colombia y la NASA POWER Project, asegurando el cumplimiento de principios de transparencia y acceso libre.

Consentimiento y uso de fuentes

Los conjuntos de datos utilizados están disponibles bajo licencias abiertas, lo que permite su uso para fines educativos, de investigación y desarrollo. No se violaron derechos de autor ni restricciones de propiedad intelectual en el uso de estas fuentes.

Transparencia del modelo

El modelo de red neuronal fue documentado en detalle, desde la preprocesamiento de los datos hasta la interpretación de sus salidas. Además, se evitó el uso de variables que implicarán data leakage, asegurando que las predicciones fueran realistas y reproducibles en contextos futuros.

Equidad y no discriminación

No se utilizaron variables que pudieran generar sesgos sociales o económicos. El modelo fue diseñado para ser aplicado en diversos municipios y cultivos, sin favorecer a ninguna región o tipo de producción en particular.

Limitaciones éticas identificadas

Si bien el modelo busca brindar apoyo a la toma de decisiones en el sector agrícola, no debe ser usado como única fuente de verdad. Existen incertidumbres en las proyecciones climáticas y limitaciones por falta de variables como nutrientes del suelo o presencia de plagas, que pueden afectar la precisión del modelo.

VI. Conclusiones y trabajo futuro

Conclusiones clave

El presente proyecto logró desarrollar un sistema de predicción de rendimiento agrícola que integra datos climáticos y de producción agrícola utilizando inteligencia artificial, específicamente redes neuronales multitarea. A través del análisis exploratorio, la depuración de datos y la construcción de modelos, se concluye que:

- Las redes neuronales multitarea superaron al modelo inicial de Random Forest en la capacidad de generalización hacia períodos futuros y condiciones climáticas no vistas.

- La exclusión de variables como Producción y Área cosechada, aunque disminuyó la precisión, fue fundamental para evitar data leakage y garantizar la validez del modelo en escenarios reales de predicción anticipada.

- El modelo logró un R^2 superior al 0.92, indicando un alto nivel de ajuste en los datos históricos cuando se entrena con datos depurados y normalizados.

- El sistema fue implementado en una interfaz visual con Dash, permitiendo una interacción amigable con usuarios no técnicos.

Trabajo futuro

Para fortalecer y escalar esta solución, se propone:

- Incorporar nuevas variables como fertilización, calidad del suelo, presencia de plagas o prácticas agrícolas que puedan mejorar la precisión del modelo.

- Implementar una red neuronal recurrente o modelos basados en series temporales (como LSTM) que puedan capturar mejor las tendencias históricas.

- Ampliar la cobertura territorial del sistema, incluyendo más municipios, departamentos y cultivos para lograr una herramienta nacional escalable.

- Integrar APIs meteorológicas en tiempo real para obtener datos climáticos más recientes y fortalecer la predicción futura.

- Realizar pruebas piloto con agricultores o entidades del sector agropecuario que puedan aportar retroalimentación desde la práctica.

VII. Referencias

NASA POWER. (2024). *NASA Langley Research Center – POWER Data Viewer*. https://power.larc.nasa.gov/data-access-viewer/

Ministerio de Agricultura y Desarrollo Rural. (s.f.). *Evaluaciones Agropecuarias Municipales EVA (2007-2018)* [Conjunto de datos]. Datos Abiertos Colombia. https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Evaluaciones-Agropecuarias-Municipales-EVA/2pnw-mmge

Ministerio de Agricultura y Desarrollo Rural. (s.f.). *Evaluaciones Agropecuarias Municipales EVA 2019-2021* [Conjunto de datos]. Datos Abiertos Colombia. https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Evaluaciones-Agropecuarias-Municipales-EVA-2019-20/uejq-wxrr

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283. https://www.tensorflow.org/

Chollet, F. et al. (2015). *Keras* [Software]. GitHub. https://keras.io/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://scikit-learn.org/

Plotly Technologies Inc. (2015). *Dash: A productive Python framework for building web applications*. https://dash.plotly.com/

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://matplotlib.org/

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://seaborn.pydata.org/