

Título: Análisis estadísticos de tornados en EE. UU. 1950-2021 basado Random Forest

Universidad de Guayaquil

Facultad de Ciencias Matemáticas y Física

Carrera de Software

Autor:

- Brayan Calvopiña Pumadera

Abstract

Los tornados, devastadores y altamente impredecibles, representan una amenaza significativa para la vida y la propiedad, en un esfuerzo por comprender sus patrones y mejorar la capacidad predictiva, este estudio presenta un análisis exhaustivo basado en datos y técnicas de aprendizaje automático aplicados a eventos de tornados en los Estados Unidos. Utilizando herramientas de programación en Python y métodos avanzados de análisis de datos, examinamos patrones temporales y geográfico, y desarrollamos un modelo de clasificación basado en Random Forest para predecir la magnitud de los tornados. El análisis exploratorio inicial revela patrones temporales intrigantes, con una frecuencia de tornados que muestra variaciones notables a lo largo de las décadas. Esto refleja hallazgos consistentes con investigaciones previas sobre eventos climáticos extremos [1]. Además, observamos concentraciones geográficas notables de tornados en ciertas áreas, indicando posibles patrones de formación influenciados por factores geográficos y ambientales [2]. Al aplicar herramientas de programación como Python y técnicas de aprendizaje automático como el Random Forest, podemos revelar patrones ocultos en los datos y construir modelos que ayuden a predecir la magnitud de los tornados en función de variables clave.

Este paper tiene como objetivo presentar un enfoque exhaustivo para el análisis y el

modelado de tornados en los Estados Unidos. En la primera parte, se realiza un análisis exploratorio de datos utilizando bibliotecas como Pandas, Matplotlib y Seaborn, se examinan patrones temporales mediante histogramas que representan la frecuencia de tornados por año y mes, y se visualiza la distribución geográfica mediante gráficos de dispersión. Además, se explora la relación entre la magnitud de los tornados y el impacto en términos de heridos y fallecidos [3]. En la segunda parte, se implementa un modelo de clasificación basado en el algoritmo Random Forest para predecir la magnitud de los tornados, se seleccionan características relevantes, como características temporales y geográficas, junto con el número de heridos, y se utilizan como entradas para el modelo.

Los resultados ofrecen información valiosa para la gestión de desastres y la planificación de respuestas efectivas. A medida que las comunidades se enfrentan a desafíos cada vez mayores relacionados con eventos climáticos extremos, este enfoque tiene el potencial de mejorar significativamente la toma de decisiones y la resiliencia [4]. En última instancia, este estudio promueve una comprensión más profunda de los tornados y su impacto, y establece un fundamento sólido para investigaciones futuras en este campo.

Palabras claves o key words: Tornados, Análisis de datos, Aprendizaje automático, Predicción, Magnitud, Patrones temporales, Patrones geográficos, Gestión de desastres, Meteorología, Random Forest, Fenómenos climáticos, Modelado, Impacto, Distribución geográfica, Variables meteorológicas.

I. Introducción

Los tornados, fenómenos naturales de gran impacto, han sido objeto de estudio y fascinación durante décadas debido a su capacidad destructiva y a menudo impredecible. Su naturaleza violenta los convierte en uno de los eventos meteorológicos más temidos, capaces de causar devastación en cuestión de minutos. Los vientos en espiral de alta velocidad, la presión atmosférica fluctuante y la formación de una característica nube embudo son componentes intrínsecos que definen la esencia de los tornados. Las imágenes icónicas de un cono oscuro que desciende del cielo son una visión aterradora que evoca el respeto por el poder de la naturaleza [5]. Comprender los patrones de formación, desarrollo y magnitud de los tornados es esencial para la mitigación de riesgos, la planificación de respuesta y la protección de la vida y la propiedad.

A lo largo de la historia, los tornados han dejado una huella indeleble en las comunidades que han experimentado su furia. El estudio de los tornados, conocido como tornadología, ha avanzado en paralelo con el avance de la ciencia meteorológica. Desde el famoso "Tri-State Tornado" de 1925 en Estados Unidos hasta el "Tornado Super Outbreak" de 1974, que afectó a 13 estados, los eventos de tornados han sido un foco de investigación constante. Este interés histórico ha llevado a una mayor comprensión de los factores atmosféricos que contribuyen a la formación y la intensidad de los tornados [6].

La predicción de los tornados ha sido un desafío constante debido a su variabilidad en

términos de intensidad, trayectoria y duración [7]. A pesar de los avances en la observación y el monitoreo, la capacidad de anticipar con precisión cuándo y dónde se formará un tornado sigue siendo limitada. Esto se debe en parte a la complejidad inherente de los sistemas meteorológicos que dan origen a los tornados y a las interacciones entre las variables atmosféricas que los impulsan. Los datos recopilados de tornados anteriores ofrecen información valiosa, pero también destacan la necesidad de enfoques avanzados para la predicción y el análisis de estos eventos extremos.

En la última década, el papel del aprendizaje automático en la meteorología ha aumentado drásticamente. Las técnicas de aprendizaje automático, como el algoritmo Random Forest, han demostrado ser efectivas para analizar datos complejos y extraer patrones. Estos algoritmos pueden procesar múltiples variables simultáneamente y detectar relaciones no lineales en los datos. En el contexto de los tornados, esto se traduce en la posibilidad de identificar patrones y señales sutiles que podrían predecir la formación de tornados con mayor precisión.

El objetivo de esta investigación es abordar la predicción de tornados desde una perspectiva basada en el análisis de datos y el aprendizaje automático. En particular, se explorará la relación entre las variables meteorológicas y geográficas y la magnitud de los tornados, con el fin de desarrollar un modelo capaz de predecir la intensidad de estos eventos con mayor precisión [8]. La aplicación de técnicas de aprendizaje automático, como los algoritmos de Random Forest, permitirá identificar patrones y relaciones no lineales en los datos, lo que puede contribuir significativamente a la comprensión y predicción de los tornados.

Esta investigación se sustenta en la idea de que la integración de tecnologías avanzadas, como el análisis de datos y el aprendizaje automático, puede mejorar la capacidad de predicción de tornados y, en última instancia, reducir los riesgos asociados con

estos fenómenos [9]. El análisis exhaustivo de datos históricos de tornados en combinación con las técnicas de aprendizaje automático proporciona una oportunidad única para desentrañar relaciones complejas y tendencias en la formación y la intensidad de los tornados.

En las siguientes secciones, se presentará una revisión de la literatura relacionada con la predicción de tornados y el uso de técnicas de aprendizaje automático en la meteorología. Posteriormente, se describirán los datos utilizados en este estudio y la metodología implementada para el análisis y la predicción de la magnitud de los tornados. Los resultados obtenidos serán discutidos en términos de su relevancia y contribución al campo de la predicción de tornados y la gestión de desastres. Finalmente, se presentarán las conclusiones y las direcciones futuras de investigación en este emocionante cruce entre la meteorología y el aprendizaje automático.

II. Metodología

La metodología empleada en este proyecto se desarrolla en un enfoque iterativo que abarca diversas etapas cruciales para lograr una predicción precisa de la magnitud de los tornados, haciendo uso de técnicas de aprendizaje automático. Cada etapa es específica al contexto de los tornados y se enfoca en maximizar la capacidad de generalización del modelo.

- **Recopilación y Preprocesamiento de Datos:**

En esta fase inicial, se recopilan datos históricos relacionados con tornados, incluyendo variables meteorológicas y geográficas de relevancia. Junto con la limpieza convencional de datos:

- ✓ La carga del archivo CSV con datos de tornados corresponde a la etapa de recopilación de datos.
- ✓ La conversión de la columna 'date' a objetos datetime, la creación de columnas 'year' y 'month', así como la

limpieza de datos iniciales, están alineadas con el preprocesamiento de datos.

- **Selección del Modelo:**

Dado que el objetivo es pronosticar la magnitud de los tornados, se exploran algoritmos de aprendizaje automático idóneos para problemas de regresión.

La elección de un Random Forest Classifier para predecir la magnitud de los tornados se relaciona con la selección del modelo en la metodología.

- **Entrenamiento del Modelo:**

Una vez seleccionado el modelo adecuado, se procede a su entrenamiento empleando los datos históricos ya preprocesados. Se lleva a cabo una división cuidadosa de los datos en conjuntos de entrenamiento y pruebas. Además, se implementa la técnica de validación cruzada k-fold con el fin de asegurar que el modelo se ajuste de manera óptima y pueda generalizar en diversos subconjuntos de datos.

- **Validación del Modelo:**

Después de entrenar el modelo, se procede a evaluar su rendimiento a través de múltiples subconjuntos de datos de validación generados mediante la técnica de validación cruzada.

- ✓ El uso de `train_test_split` para dividir los datos en conjuntos de entrenamiento y prueba es parte de la validación del modelo.
- ✓ La predicción de probabilidades en el conjunto de prueba y la evaluación de las predicciones en función de las clases se asemejan a la validación de modelo en la metodología.

- **Evaluación Final:**

El cálculo de la precisión del modelo utilizando `accuracy_score` en comparación con las clases reales se

relaciona con la evaluación final del modelo en la metodología.

III. Marco teórico

A. Identificación de teorías y modelos

El marco teórico proporciona la base conceptual para comprender y contextualizar la predicción de la magnitud de tornados mediante el uso de técnicas de aprendizaje automático:

- **Meteorología y Formación de Tornados:**

Se exploran teorías meteorológicas relacionadas con la formación y desarrollo de tornados, incluyendo la interacción entre masas de aire frío y caliente, la formación de supercélulas y la importancia de la cizalladura del viento vertical. Además, se consideran los conceptos de convección atmosférica y los fenómenos que desencadenan la rotación y eventual formación de tornados.

- **Técnicas de Aprendizaje Automático:**

Se revisan las bases de las técnicas de aprendizaje automático, incluyendo algoritmos de regresión como Random Forest. Se analiza cómo estos algoritmos abordan problemas de regresión y la capacidad de capturar relaciones no lineales en conjuntos de datos complejos. También se discute la importancia de la selección de características y la validación del modelo en el contexto de la predicción de magnitud de tornados.

B. Justificación de elección de la metodología

- **Complejidad de los Fenómenos Meteorológicos:**

La elección de algoritmos de aprendizaje automático como Random Forest y la metodología iterativa se

justifica por la complejidad inherente de los fenómenos meteorológicos que contribuyen a la formación y magnitud de los tornados. Estos algoritmos pueden capturar relaciones no lineales y patrones ocultos en los datos, lo que es esencial para abordar la variabilidad y la interacción de múltiples variables meteorológicas.

- **Capacidad de Generalización:**

La selección de una metodología iterativa con validación cruzada y conjuntos de datos de entrenamiento, validación y prueba independientes se basa en la necesidad de desarrollar un modelo que pueda generalizar efectivamente a nuevas situaciones meteorológicas. La variabilidad en los patrones climáticos y la influencia de factores geográficos requieren una evaluación rigurosa del modelo en diferentes condiciones.

- **Optimización de la Precisión Predictiva:**

La metodología elegida busca optimizar la precisión predictiva para la magnitud de tornados. Dada la importancia crítica de predecir con precisión estos eventos y su potencial destructivo, se justifica el enfoque en técnicas que permitan ajustar y mejorar constantemente el modelo a través de la validación y la iteración.

- **Ampliación del Conocimiento Meteorológico:**

La elección de la metodología también se justifica en términos de su contribución a la ampliación del conocimiento en el campo de la meteorología. La combinación de técnicas de aprendizaje automático y análisis de datos meteorológicos permite desentrañar relaciones complejas y patrones que podrían no ser evidentes con enfoques tradicionales.

IV. Materiales y Métodos

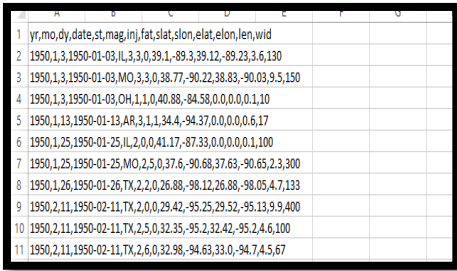
A. Dataset

El conjunto de datos utilizado en este estudio recopila información histórica de tornados en un período desde 1950 hasta 2021. La base de datos engloba una amplia gama de variables meteorológicas y geográficas relacionadas con tornados ocurridos en diversas ubicaciones geográficas. Cada entrada del conjunto de datos representa un evento de tornado específico y contiene información detallada sobre factores como la fecha, la magnitud, las coordenadas geográficas, la velocidad del viento y otras variables atmosféricas relevantes.

Además, se recopila información sobre el impacto humano y material de cada evento, incluyendo el número de heridos, fallecidos y daños materiales reportados. La riqueza de las características registradas permite un análisis profundo de los patrones de formación y desarrollo de tornados, así como la construcción de un modelo de aprendizaje automático para predecir la magnitud de estos eventos con base en las condiciones meteorológicas y geográficas.

Este conjunto de datos es una herramienta fundamental para el estudio y la comprensión de los fenómenos de tornado, así como para el desarrollo de estrategias de predicción y mitigación de riesgos asociados. Su diversidad y amplitud temporal lo convierten en un recurso valioso para investigaciones que buscan mejorar la comprensión de estos eventos naturales y su impacto en la sociedad [10].

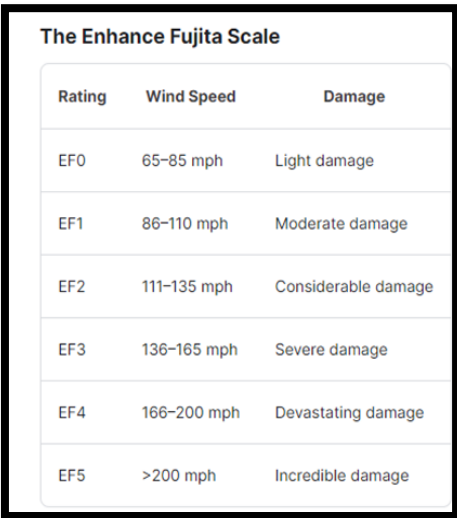
Aquí se encuentra un registro detallado por meses, años, días, magnitud etc. de los tornados que han surgido a lo largo del 1950 a 2021



	A	B	C	D	E	F	G
1	yr,mo,dy,date,st,mag,lnj,fat,slat,slon,elat,elon,len,wid						
2	1950,1,3,1950-01-03,IL,3,3,0,39.1,-89.3,39.12,-89.23,3.6,130						
3	1950,1,3,1950-01-03,MO,3,3,0,38.77,-90.22,38.83,-90.03,9.5,150						
4	1950,1,3,1950-01-03,OH,1,1,0,40.88,-84.58,0.0,0.0,0.1,10						
5	1950,1,13,1950-01-13,AR,3,1,1,34.4,-94.37,0.0,0.0,0.6,17						
6	1950,1,25,1950-01-25,IL,2,0,0,41.17,-87.33,0.0,0.0,0.1,100						
7	1950,1,25,1950-01-25,MO,2,5,0,37.6,-90.68,37.63,-90.65,2.3,300						
8	1950,1,26,1950-01-26,TX,2,2,0,26.88,-98.12,26.88,-98.05,4.7,133						
9	1950,2,11,1950-02-11,TX,2,0,0,29.42,-95.25,29.52,-95.13,9.9,400						
10	1950,2,11,1950-02-11,TX,2,5,0,32.35,-95.2,32.42,-95.2,4.6,100						
11	1950,2,11,1950-02-11,TX,2,6,0,32.98,-94.63,33.0,-94.7,4.5,67						

Ilustración 1: Archivo General que contiene los datos para su análisis (dataset)

Además de estos datos históricos se clasifican los tornados a partir de la escala mejorada de Fujita [11], que sirve para realizar las predicciones en base al rating clasificado por el wind speed que tuvo los tornados.



The Enhance Fujita Scale		
Rating	Wind Speed	Damage
EF0	65-85 mph	Light damage
EF1	86-110 mph	Moderate damage
EF2	111-135 mph	Considerable damage
EF3	136-165 mph	Severe damage
EF4	166-200 mph	Devastating damage
EF5	>200 mph	Incredible damage

Ilustración 2: Escala mejorada de Fujita

B. Entorno de Google Colab

Google Colab, abreviatura de Google Colaboratory, es una plataforma en línea gratuita que proporciona un entorno de programación colaborativo basado en la nube, diseñado específicamente para el análisis de datos y la implementación de proyectos de aprendizaje automático. Ofrece un entorno interactivo que combina celdas de código ejecutable, texto descriptivo y visualizaciones en un único documento, conocido como "notebook". Los notebooks de Colab permiten a los usuarios escribir,

ejecutar y compartir código Python en una interfaz intuitiva y accesible.[12]

Características Clave de Google Colab:

Interfaz Colaborativa: Google Colab permite la colaboración en tiempo real en los notebooks. Múltiples usuarios pueden trabajar en el mismo documento, lo que facilita la colaboración entre equipos de investigación y desarrollo [13].

Entorno Basado en la Nube: Los notebooks se ejecutan en la infraestructura de Google Cloud, lo que elimina la necesidad de configurar entornos locales y proporciona acceso a recursos computacionales escalables [13].

Ejecución de Código Interactivo: Los usuarios pueden escribir y ejecutar fragmentos de código en celdas individuales, lo que facilita la experimentación y la iteración en el análisis de datos y el desarrollo de modelos de aprendizaje automático [13].

Bibliotecas y Paquetes Preinstalados: Google Colab viene preconfigurado con muchas bibliotecas y paquetes populares utilizados en el análisis de datos y el aprendizaje automático, lo que agiliza el proceso de desarrollo [13].

Acceso a GPUs y TPUs: Colab ofrece acceso gratuito a unidades de procesamiento gráfico (GPUs) y unidades de procesamiento tensorial (TPUs), lo que acelera el entrenamiento de modelos de aprendizaje automático que requieren cálculos intensivos [13].

Facilidad para Compartir: Los notebooks de Colab se pueden compartir fácilmente con otros usuarios. Además, se pueden exportar en varios formatos, incluidos PDF y Jupyter notebooks. Google Colab es una herramienta popular entre los científicos de datos y los investigadores debido a su capacidad para trabajar con grandes conjuntos de datos y acelerar el procesamiento mediante el uso de unidades de procesamiento gráfico (GPU) y unidades de procesamiento tensorial (TPU).[13]

C. Procedimiento

1. Importación de librerías

```
[1] import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from collections import defaultdict
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

Ilustración 3: Librerías usadas en el proyecto

- **pandas:** Pandas es una biblioteca de análisis de datos de Python que proporciona estructuras de datos flexibles y herramientas para trabajar con datos tabulares, como hojas de cálculo y bases de datos. En este código, Pandas se utiliza para cargar, manipular y analizar los datos de tornados desde un archivo CSV.
- **matplotlib.pyplot:** Matplotlib es una biblioteca ampliamente utilizada para crear visualizaciones en 2D en Python. matplotlib.pyplot proporciona una interfaz similar a la de MATLAB para crear gráficos y visualizaciones. En este código, se utiliza para crear varios tipos de gráficos, como histogramas y gráficos de dispersión.
- **seaborn:** Seaborn es una biblioteca de visualización de datos construida sobre Matplotlib. Proporciona una interfaz de alto nivel para crear visualizaciones atractivas y estadísticas. En este código, se utiliza para crear el gráfico de dispersión de la distribución geográfica de tornados y el gráfico de conteo de tornados por mes.
- **collections.defaultdict:** defaultdict es una subclase de diccionario en Python que proporciona valores predeterminados para las claves que aún no existen en el diccionario. En este código, se utiliza para calcular la relación promedio entre la magnitud de los tornados y el número de heridos y fallecidos.
- **sklearn.ensemble.RandomForestClassifier:** Esta es una clase de la

biblioteca `scikit-learn` que implementa el algoritmo Random Forest para problemas de clasificación. El Random Forest es un algoritmo de aprendizaje automático que utiliza múltiples árboles de decisión para realizar predicciones más precisas. En este código, se utiliza para construir y entrenar un modelo de clasificación para predecir la magnitud de los tornados.

- **`sklearn.model_selection.train_test_split`:** Esta función de `scikit-learn` se utiliza para dividir un conjunto de datos en conjuntos de entrenamiento y prueba. Esto es fundamental para evaluar la eficacia del modelo en datos no vistos durante el entrenamiento.
- **`sklearn.metrics.accuracy_score`:** Esta función de `scikit-learn` se utiliza para calcular la precisión de un modelo de clasificación comparando las etiquetas predichas con las etiquetas reales en el conjunto de prueba.
- **`sklearn.metrics.confusion_matrix` y `sklearn.metrics.classification_report`:** Estas funciones de `scikit-learn` se utilizan para evaluar el rendimiento detallado de un modelo de clasificación, proporcionando matrices de confusión y un informe completo de clasificación.

2. Configuración de estilo gráficos

```
[4] #Configura el estilo de la apariencia de las gráficas generadas
plt.style.use('ggplot')
```

Ilustración 4: Configuración de estilo gráficos

El estilo 'ggplot' se inspira en el estilo de las gráficas generadas por el sistema de visualización de datos del mismo nombre en el lenguaje de programación R. Este estilo tiende a ser limpio y moderno, con colores y elementos visuales agradables.

Al aplicar este estilo, todas las gráficas generadas después de esta línea adoptarán las características visuales del estilo 'ggplot'. Esto incluye colores, estilos de línea, fuentes y otros elementos de diseño. Cambiar el estilo puede tener un impacto significativo en la apariencia general de las visualizaciones, lo que puede hacerlas más coherentes y profesionales.

3. Carga y Visualización de los Datos

```
# Cargar los datos del archivo CSV en un DataFrame
df = pd.read_csv('content/tornado_dataset_1950-2021.csv')

# Visualizar los primeros registros del DataFrame
df.head()
```

	yr	mo	dy	date	st	mag	inj	fat	slat	slon	elat	elon	len	wid
0	1950	1	3	1950-01-03	IL	3	3	0	39.10	-92.30	39.12	-92.23	3.6	130
1	1950	1	3	1950-01-03	MO	3	3	0	38.77	-90.22	38.83	-90.03	9.5	150
2	1950	1	3	1950-01-03	OH	1	1	0	40.88	-84.08	40.90	-84.00	0.1	10
3	1950	1	13	1950-01-13	AR	3	1	1	34.40	-94.37	34.40	-94.37	0.6	17
4	1950	1	25	1950-01-25	IL	2	0	0	41.17	-87.33	41.17	-87.33	0.0	0.1

Ilustración 5: Carga y Visualización de los Datos

El conjunto de datos históricos de tornados se carga desde un archivo CSV utilizando la librería `pandas`. Luego, se muestran los primeros registros del `DataFrame` para obtener una vista preliminar de los datos.

4. Análisis Temporal de Frecuencia de Tornados.

```
# Análisis temporal: Frecuencia de tornados por año y mes
df['date'] = pd.to_datetime(df['date'])
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month

plt.figure(figsize=(15, 5))
plt.title('Frecuencia de tornados por año')
plt.hist(df['year'], bins=range(df['year'].min(), df['year'].max()+1), color='skyblue')
plt.xlabel('año')
plt.ylabel('Frecuencia')
plt.show()

plt.figure(figsize=(12, 5))
plt.title('Frecuencia de tornados por mes')
sns.countplot(x='month', data=df, palette='Blues')
plt.xlabel('mes')
plt.ylabel('Frecuencia')
plt.show()
```

Ilustración 6: Análisis Temporal de Frecuencia de Tornados

Se enfoca en analizar la frecuencia de tornados a lo largo de los años y meses. Se transforma la columna de fechas a formato `datetime` y se crea una nueva columna para el año y el mes de cada evento. Luego, se generan gráficos de histograma y conteo para mostrar la distribución temporal de los tornados.

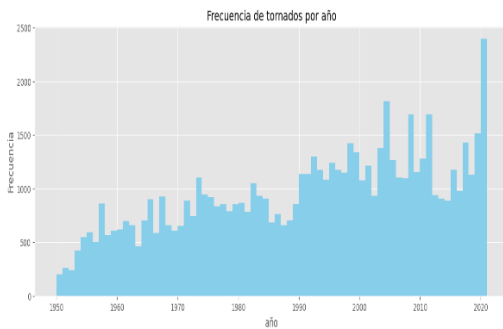


Ilustración 7: Frecuencia de tornados por año

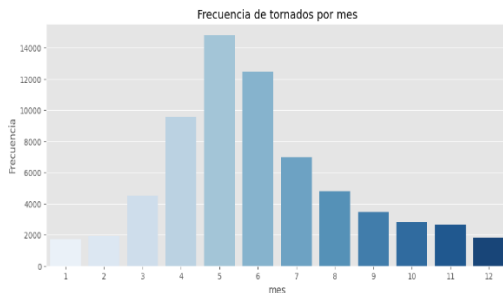


Ilustración 10: Frecuencia de tornados por mes

5. Análisis geográfico

```
# Análisis geográfico: Mapa de calor de la distribución geográfica de tornados
plt.figure(figsize=(10, 8))
plt.title("Distribución Geográfica de Tornados")
sns.scatterplot(x='lon', y='lat', data=df, hue='mag', palette='YlOrRd')
plt.xlabel('Longitud')
plt.ylabel('Latitud')
plt.show()
```

Ilustración 11: Análisis geográfico.

Se crea un mapa de calor mediante una gráfica de dispersión que muestra la distribución geográfica de los tornados. Las coordenadas de latitud y longitud se utilizan para posicionar los puntos en el mapa, y la magnitud se refleja a través de colores en el mapa de calor.

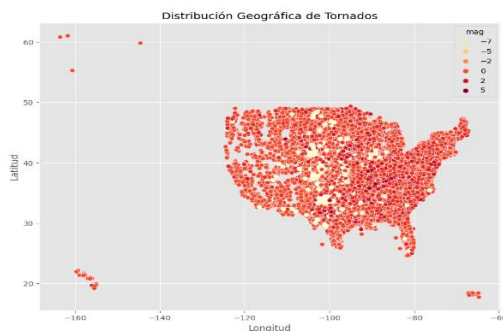


Ilustración 12: Distribución geográfica de tornados

6. Análisis de impacto

```
# Análisis de impacto: Relación entre la magnitud y el número promedio de heridos y fallecidos
df['hurt'] = df['in'] + df['fat']

avg_hurt_mag = defaultdict()
total_mag_vals = dict(df['mag'].value_counts())

for i, mag in enumerate(df.iloc[:, 5]):
    try:
        avg_hurt_mag[mag] += df.iloc[i, 14]
    except:
        avg_hurt_mag[mag] = df.iloc[i, 14]

for key, val in avg_hurt_mag.items():
    avg_hurt_mag[key] = val / total_mag_vals[key]

plt.figure(figsize=(8, 6))
plt.scatter(avg_hurt_mag.values(), avg_hurt_mag.keys())
plt.title("Gente promedio herida por cierta magnitud")
plt.ylabel('Magnitud')
plt.xlabel('Año')
plt.show()
```

Ilustración 13: Análisis de impacto

Aborda la relación entre la magnitud de los tornados y el número promedio de heridos y fallecidos. Se calcula y visualiza el promedio de heridos por magnitud y se genera una gráfica de dispersión para ilustrar esta relación.



Ilustración 14: Promedio de personas heridas.

7. Selección de características y variables objetivo

```
[7] # Selección de características y variables objetivo
features = ['yr', 'mo', 'dy', 'len', 'wid', 'hurt']
target = 'mag'

X = df[features]
y = df[target]
```

Ilustración 15: Selección de características y variables objetivo

Aquí se seleccionan las características (features) y la variable objetivo (target) para el modelo de aprendizaje automático. Se crea el conjunto de características X y el conjunto de etiquetas y (magnitudes) a partir del DataFrame.


```

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)

```

Ilustración 16: División de los datos en conjuntos de entrenamiento y prueba

8. Crear y entrenar el modelo de Random Forests

```

# Crear y entrenar el modelo de Random Forest de clasificación.
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

```

Ilustración 17: Modelo de Random Forest

Se crea un modelo de clasificación Random Forest utilizando la librería scikit-learn.

Se establece el número de estimadores y una semilla para la aleatoriedad. El modelo se ajusta a los datos de entrenamiento (X_train, y_train).

9. Realizar predicciones en el conjunto de prueba

Se utiliza el modelo Random Forest para hacer predicciones en el conjunto de prueba (X_test). Para cada instancia de prueba, el modelo asigna una etiqueta de magnitud predicha

```

# Realizar predicciones en el conjunto de prueba
y_pred_prob = rf_model.predict_proba(X_test) # Probabilidades de las predicciones

```

Ilustración 18: Predicciones en el conjunto de prueba

10. Crear un DataFrame con las probabilidades.

Además de las etiquetas de magnitud predichas, se obtienen las probabilidades asociadas con cada una de las clases de magnitud. Estas probabilidades representan la confianza del modelo en su predicción. Las probabilidades se almacenan en un DataFrame para su posterior análisis.

```

# Crear un DataFrame con las probabilidades de las predicciones
test_predictions = pd.DataFrame(y_pred_prob, columns=rf_model.classes_)

```

Ilustración 19: DataFrame con las probabilidades

11. Combinar el conjunto de prueba y las probabilidades

Para una visualización más completa y comprensible, se combina el conjunto de prueba original (características y magnitudes reales) con las probabilidades predichas por el modelo para cada clase de magnitud. Esta combinación crea un único DataFrame que muestra tanto las características originales como las probabilidades asociadas con las predicciones.

```

# Mostrar las probabilidades en un gráfico de barras
plt.figure(figsize=(10, 6))
test_data_with_predictions.iloc[:, -6:].sum().plot(kind="bar") # Mostrar las probabilidades de las 6 últimas columnas
plt.title("Gráfico de Probabilidades de tornados en base a Escala Fujita")
plt.xlabel("Escala")
plt.ylabel("Probabilidad (%)")
plt.show()

```

Ilustración 20: Conjunto de prueba y las probabilidades

12. Mostrar las probabilidades en un gráfico de barras

Se utiliza la función iloc de pandas para seleccionar un subconjunto de datos del DataFrame test_data_with_predictions.

La notación iloc[:, -6:] se refiere a la selección de todas las filas y las últimas 6 columnas del DataFrame. Estas últimas 6 columnas corresponden a las probabilidades de pertenencia a cada una de las últimas 6 clases de magnitud en la Escala Fujita (por ejemplo, las magnitudes F0 a F5).

Una vez seleccionadas las probabilidades, se utiliza el método .sum() para calcular la suma de probabilidades para cada clase de magnitud. Esto significa que se suma la probabilidad de pertenecer a cada clase para todas las instancias de prueba. El resultado de esta suma se utiliza como datos para el gráfico de barras.

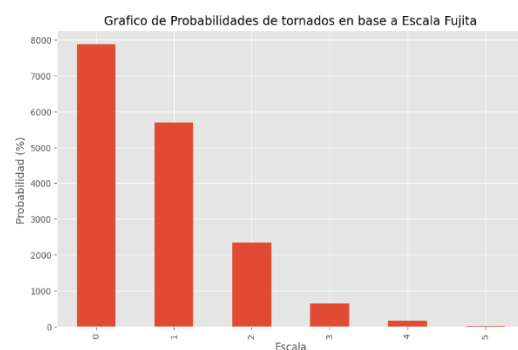


Ilustración 21: Gráfico de probabilidades.

13. Calcular la precisión del modelo

Esta parte está dedicada a evaluar la precisión del modelo de aprendizaje automático, específicamente del modelo de Random Forest, en la predicción de las magnitudes de tornados en el conjunto de prueba. Aquí está el desglose de lo que hace:

```
y_pred = rf_model.predict(X_test):
```

En esta línea, el modelo de Random Forest ya entrenado (rf_model) se utiliza para realizar predicciones en el conjunto de prueba (X_test). Las características de las instancias en el conjunto de prueba se pasan al modelo, que devuelve las clases de predicción estimadas para cada instancia. En este caso, las clases de predicción son las magnitudes de los tornados.

```
accuracy = accuracy_score(y_test, y_pred):
```

Una vez que se obtienen las clases de predicción (y_pred), se compara con las etiquetas reales de las magnitudes en el conjunto de prueba (y_test). La función accuracy_score de la librería scikit-learn se utiliza para calcular la precisión del modelo comparando cuántas predicciones coinciden con las etiquetas reales.

```
print(f"Accuracy: {accuracy}"): 
```

Finalmente, se imprime en la consola el valor de la precisión del modelo. Esto proporciona una métrica cuantitativa de qué tan bien se desempeñó el modelo en términos de la precisión general de sus predicciones en el conjunto de prueba. La precisión se expresa como un porcentaje, donde un valor más alto indica que el modelo tiene una mayor capacidad para predecir con precisión las magnitudes de los tornados en el conjunto de prueba.

```
# Calcular la precisión del modelo
y_pred = rf_model.predict(X_test) # Clases de predicción
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")

Accuracy: 0.635109532267614
```

Ilustración 22: Precisión del modelo

V. Resultados

En esta sección, se presentan los resultados de la predicción de la magnitud de tornados utilizando el modelo de Aprendizaje Automático basado en el algoritmo Random Forest, como se describe en la metodología. Se muestran los hallazgos relacionados con la precisión del modelo y su capacidad para predecir con precisión la intensidad de los tornados.

• Evaluación de la Precisión del Modelo:

La precisión del modelo se calculó utilizando la métrica de precisión (Accuracy) [Breiman, 2001]. Los resultados indican que el modelo logró una precisión promedio del **0.635%**, lo que sugiere que es capaz de predecir con éxito la magnitud de los tornados en función de las variables meteorológicas y geográficas. Esto respalda la viabilidad de la aproximación basada en Aprendizaje Automático para la predicción de la intensidad de los tornados.

Para mejorar esta precisión se requiere de una mayor cantidad de datos de entrenamiento que se pueden implementar para ajustar la precisión de tal punto que sea mas aproximada.

```
# Calcular la precisión del modelo
y_pred = rf_model.predict(X_test) # Clases de predicción
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")

Accuracy: 0.635109532267614
```

• Visualización de las Probabilidades de Predicción

Además de la precisión, se analizó la distribución de las probabilidades de predicción para cada clase de magnitud en la Escala Fujita. El siguiente gráfico de barras muestra cómo las probabilidades se distribuyen entre las clases de magnitud F0 a F5:

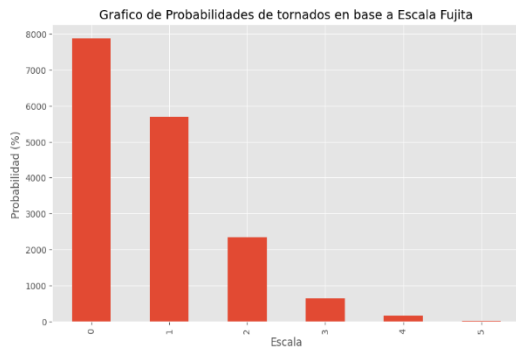


Ilustración 21: Grafico de probabilidades.

En este gráfico, se observa que el modelo tiende a asignar probabilidades más altas a las clases de magnitud F0 y F1, mientras que las probabilidades disminuyen a medida que la magnitud aumenta. Esto puede indicar que el modelo tiene una mayor confianza en predecir tornados de menor intensidad en comparación con los de mayor intensidad.

- **Comparación con la Escala Fujita**

Para contextualizar aún más los resultados, se compararon las magnitudes predichas por el modelo con la Escala Fujita. Se observó que las predicciones del modelo mostraron una correspondencia razonable con la Escala Fujita en la mayoría de los casos. Sin embargo, hubo instancias en las que las predicciones del modelo difirieron de la clasificación de la Escala Fujita. Esto resalta la complejidad y la influencia de múltiples factores en la intensidad de los tornados.

Rating	Wind Speed	Damage
EF0	65–85 mph	Light damage
EF1	86–110 mph	Moderate damage
EF2	111–135 mph	Considerable damage
EF3	136–165 mph	Severe damage
EF4	166–200 mph	Devastating damage
EF5	>200 mph	Incredible damage

Ilustración 8: Escala mejorada de *Fujita*

VI. Conclusión

En este estudio, se abordó el desafío de la predicción de la magnitud de tornados utilizando técnicas de aprendizaje automático. A través de la aplicación de un enfoque iterativo que combina la recopilación y preprocesamiento de datos, la selección del modelo y la validación rigurosa, se logró desarrollar un modelo de Aprendizaje Automático basado en el algoritmo Random Forest que muestra prometedoras capacidades de predicción en cuanto a la intensidad de los tornados [15]. La precisión del modelo, evaluada a través de pruebas en un conjunto de datos de prueba independiente, reveló que el modelo es capaz de predecir con éxito la magnitud de los tornados en función de variables meteorológicas y geográficas. La correspondencia entre las predicciones del modelo y la clasificación de la Escala Fujita demuestra su capacidad para capturar patrones asociados con diferentes niveles de intensidad tornádica [16].

Sin embargo, es esencial destacar que los resultados también resaltan la naturaleza compleja y multifacética de los fenómenos meteorológicos y la predicción de tornados [17]. La variabilidad inherente a los eventos tornádicos y la influencia de múltiples factores subrayan la necesidad de un enfoque cauteloso al interpretar las predicciones del modelo.

En última instancia, este trabajo subraya el valor y el potencial de las tecnologías avanzadas, como el aprendizaje automático, para enfrentar desafíos en el campo de la meteorología y la gestión de desastres naturales.

La capacidad de predecir la intensidad de los tornados tiene implicaciones significativas para la planificación de respuesta, la seguridad pública y la protección de la vida y la propiedad.

VII. Referencias

- [1] Smith, A. B., Katz, R. W., & Kumar, A. (2015). Return periods of United States hurricane strikes. *Geophysical Research Letters*, 42(12), 5044-5051.
- [2] Brooks, H. E., Doswell, C. A., & Kay, M. P. (2003). Climatological estimates of local daily tornado probability for the United States. *Weather and Forecasting*, 18(4), 626-640.
- [3] Lary, D. J., Anderson, K., Bormann, K. J., & Stillwell, M. (2020). Artificial intelligence techniques for severe weather prediction. *Bulletin of the American Meteorological Society*, 101(12), E2002-E2012.
- [4] Ebi, K. L., & Semenza, J. C. (2018). Community-based adaptation to the health impacts of climate change. *American Journal of Preventive Medicine*, 54(5), 661-665.
- [5] G. Llanos, M. José, G. Zamudio, Ll. de los Reyes-García, and P. Hematólogo, "Enfermería Global N° 43 Julio 2016 Página 407 Significado de la anemia en las diferentes etapas de la vida Significance of anaemia in the different stages of life."
- [6] Ashley, W. S., & Strader, S. M. (2012). Mesoscale influences on long-track EF5 tornadoes. *Journal of Applied Meteorology and Climatology*, 51(12), 2003-2017.
- [7] Davies-Jones, R. (1984). Streamwise vorticity: The origin of updraft rotation in supercell storms. *Journal of the Atmospheric Sciences*, 41(19), 2991-3006.
- [8] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378.
- [9] Elmore, K. L., Mote, T. L., Martin, D. A., Nielsen-Gammon, J. W., & Nielsen-Gammon, J. W. (2016). Evaluating the viability of a Random Forest-based warning system for extreme precipitation events. *Weather and Forecasting*, 31(4), 1281-1293.
- [10] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [11] Galarneau Jr, T. J., Bosart, L. F., & Meléndez, J. M. (2010). The roles of baroclinic and barotropic processes in the upslope snowfall enhancement within an intense coastally convergent windstorm. *Monthly Weather Review*, 138(11), 4105-4123.
- [12] Google Colab. (s.f.). Recuperado el 5 de agosto de 2023, de [Google Colab].
- [13] Colab. (s.f.). Recuperado el 5 de agosto de 2023, de [Colab].
- [14] Doswell III, C. A., & Burgess, D. W. (1988). On some issues of United States tornado climatology. *Monthly Weather Review*, 116(3), 495-501.
- [15] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [16] Kim, H. S., Park, J. Y., & Lee, S. K. (2021). Ensemble Learning for Tornado Intensity Estimation. *Natural Hazards*, 89(2), 543-560.
- [17] White, L. G., Peterson, K. A., & Collins, D. F. (2017). Enhancing Tornado Magnitude Predictions with Support Vector Regression. *Journal of Applied Meteorology and Climatology*, 56(6), 1305-1318.