Análisis EDA para verificar si es posible pronosticar precipitaciones utilizando un modelo de machine learning

Josefina Solis Bendezu
Marcelo Zevallos Cuarite
Data Science Research Peru
Lima, Perú
josefina.b.josefina@gmail.com

Abstract—Este informe científico se centra en explorar la viabilidad de predecir la precipitación utilizando modelos de aprendizaje automático a través del Análisis Exploratorio de Datos (EDA, por sus siglas en inglés). El estudio surge de la creciente importancia de comprender los patrones de lluvia ante el cambio climático, especialmente en regiones como Perú, con una geografía diversa que abarca desde el desierto costero hasta la selva amazónica y los picos de los Andes. La falta de personal especializado en el análisis de datos meteorológicos ha limitado la comprensión integral de estos patrones, como se evidencia en eventos recientes, como las fuertes lluvias en Cusco en mayo de 2023. Por tanto, este estudio busca abordar este desafío mediante la exploración de técnicas de aprendizaje automático para la predicción de la lluvia, apoyado por un análisis exhaustivo de los datos disponibles y la revisión de estudios previos relevantes. La propuesta incluye la creación de un producto de datos para facilitar la visualización y comprensión de las variables meteorológicas clave, con el objetivo de mejorar la toma de decisiones y la planificación en el contexto climático y turístico de Perú.

Index Terms— precipitación, machine learning, predicción, análisis, meteorología, outliers, data

I. Introducción

El estudio de los patrones de lluvia es un tema que en los últimos años ha adquirido mayor importancia para poder tomar acciones frente al cambio climático. En el caso de Perú, un país con una geografía diversa que abarca desde la costa desértica hasta las altas cumbres de los Andes y la densa selva amazónica, el fenómeno de las lluvias tiene implicaciones significativas en términos de agricultura, infraestructura, y la vida cotidiana de sus habitantes. Específicamente en la región de Cusco, caracterizada por el turismo. En mayo del 2023 Cusco soportó una intensa lluvia por más de 12 horas lo que ocasionó que una gran cantidad de vuelos se cancelaran generando pérdidas económicas para turistas y pobladores [2].

La predicción de lluvia puede ayudar a resolver estos problemas de incertidumbre climática, por lo tanto, este informe científico se enfoca en analizar los datos relacionados con las precipitaciones en el territorio peruano, utilizando técnicas avanzadas de ciencia de datos. A través de la recopilación, limpieza, y análisis de conjuntos de datos meteorológicos, se pretende identificar patrones, tendencias y posibles factores influyentes en la distribución y frecuencia de las lluvias en diferentes regiones del país.

II. Planteamiento del problema

La problemática que se aborda en este estudio se centra en la falta de análisis de datos de las estaciones meteorológicas ubicadas en Perú. A pesar de la disponibilidad de datos, la falta de recursos humanos especializados en la interpretación y análisis de datos ha limitado la comprensión integral de los patrones de lluvia en el país. Esta deficiencia se refleja en situaciones como la registrada en la región de Cusco en mayo de 2023, donde una intensa lluvia provocó la cancelación de vuelos y generó pérdidas económicas significativas para turistas y pobladores. Situación que se pudo evitar mediante un producto tecnológico que permita el pronóstico de lluvia utilizando machine learning.

III. Estudios Previos

A continuación se presentan los estudios realizados que sirvieron de apoyo la validación de esta investigación:

A. Prediction of Rainfall in Australia Using Machine Learning [5]

La investigación realizada por Sarasa - Cabezuelo se basa en la implementación de 4 métodos de

machine learning: knn, decision tree, random forest, y redes neuronales. De estos, el mejor modelo fue hecho con redes neuronales y entrenado con data de 10 años de lluvia en Australia. Sin embargo, indican las consideraciones que se deben tomar para aplicar cada método según la geografía del país. Variable a considerar para la creación de un modelo basado en machine learning para Perú.

B. A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning [1]

Chase et al. presenta un estudio sobre el uso de machine learning en meteorología. Todos los modelos de ML, excepto las máquinas de vectores de soporte (codificadas en sklearn), pueden proporcionar una estimación probabilística de la clasificación (por ejemplo, esta imagen tiene un 95 % de probabilidad de tener un rayo). Al calcular la precisión anterior, asumimos un umbral del 50% para designar cuál era la predicción de ML. Para obtener la curva ROC, la probabilidad umbral varía del 0% al 100%. Las curvas ROC resultantes para todos los métodos ML, excepto las máquinas de vectores de soporte. Vemos que para este modelo simple de una característica, todos los métodos siguen siendo muy similares y tienen AUC cercanas a 0.9, lo que generalmente se considera un buen rendimiento.

C. Empirical–Statistical Downscaling of Austral Summer Precipitation over South America, with a Focus on the Central Peruvian Andes and the Equatorial Amazon Basin [4]

Sulca et al. menciona que la precipitación es una de las variables más difíciles de estimar. En América del Sur, esta tarea es aún más desafiante, dada la compleja topografía de los Andes. Para este propósito se pueden utilizar modelos de reducción de escala empírico-estadístico (ESD), pero aún no se han desarrollado dichos modelos aplicables. Para abordar este problema, construimos un modelo de ESD utilizando técnicas de regresión lineal múltiple para el período 1982-2016 que se basa en índices de circulación a gran escala que representan la variabilidad climática del Océano Pacífico tropical, el Océano Atlántico y América del Sur, para estimar la variabilidad climática del verano austral. Los resultados muestran además que el modelo ESD puede reproducir correctamente las anomalías de precipitación en toda la cuenca del Mantaro durante los tres episodios extremos de El Niño. Además, múltiples experimentos con diferentes combinaciones de predictores del modelo ESD corroboran la hipótesis de que la interacción entre la zona de convergencia del Atlántico sur y el océano Atlántico ecuatorial.

IV. PROPUESTA DE DATA PRODUCT

El data product se propone para el estudio y análisis basado en gráficos que nos ayuden a la visualizar mejor la distribución de los datos de las variables como histogramas, boxplots, mapas de calor y gráficos de dispersión de variables que influyen para la predicción de la precipitación de la ciudad de Cuzco.

A. Definición de los datos

Los datos son variables meteorológicas de la estación Machu Picchu ubicada en Urubamba, Cusco. Las variables se presentan en diferentes horas por cada día del mes. Las variables meteorológicas en cuestión son: humedad (%), precipitación (mm/hora), velocidad del viento (m/s), temperatura (°C) y dirección del viento (°). Todas son tomadas en la misma ubicación, sin embargo cada una es medida con instrumentación distinta debido a su naturaleza.

B. Captura y procesamiento de datos

Los datos se obtuvieron del repositorio de datos hidrometeorológicos a nivel nacional de la página web del SENAMHI. La fecha fue tomada en el periodo del año 2020 hasta el año 2023. Solo se permitía extraer data mensual, por lo que se tuvo que correr un código para unir la data de los 3 años y tener un solo Dataframe sobre el cual se trabajó.

C. Modelamiento y análisis

Se propone un análisis exploratorio de datos, donde se usarán diferentes herramientas como histogramas, boxplots, mapas de calor para visualizar la distribución de los datos con el fin de analizar y tratar los datos. Y a la vez se mostrará la actualización de la conversión de los datos mediante tablas que muestran las características del objetivo de cada etapa del análisis de cada variable. Esto con el fin de determinar al final del proceso la influencia de las variables explicativas ante la variable explicada.

Este modelo fue realizado mediante el lenguaje de programación Python por medio del editor Google Colab, siendo esta una herramienta de Google muy accesible y gratuita. Para la evaluación del modelo se usarán pruebas de autores sobre la normalidad, esto nos mostrará la efectividad del tratamiento de los datos.

V. ANÁLISIS DE DATOS

Se propone realizar un tratamiento de datos estructurando un análisis EDA que nos permita saber

el estado de los datos, el porcentaje de registros nulos, la relación de los registros nulos, la cantidad de datos atípicos, y la distribución simétrica de los datos de las variables (mostrada en la tabla 1) ante sus medias.

Tabla 1: Variable y datos originales

	AÑO / MES / DÍA	HORA	TEMPERATURA (°C)	PRECIPITACIÓN (mm/hora)	HUMEDAD (%)	DIRECCION DEL VIENTO (°)	VELOCIDAD DEL VIENTO (m/s)
31120	2020-01-01	23:00	13.3	0.8	97	171	0.1
31119	2020-01-01	22:00	13.5	0.0	98	249	0.9
31118	2020-01-01	21:00	13.4	0.0	96	222	0.4
31117	2020-01-01	20:00	13.3	0.0	97	102	2.4
31116	2020-01-01	19:00	13.6	0.6	96	100	0.2

Para empezar el análisis exploratorio, se contará con cinco etapas donde solo se hará una análisis numérico, ya que no se cuenta con datos categóricos entre las variables, explicadas a continuación:

A. Reporte de valores nulos e imputación de datos

Tabla 2: Reporte de valores nulos

	Variable	Cant. No Nulos	Cant. Nulos	% Nulos
0	VELOCIDAD DEL VIENTO (m/s)	28566	3269	10.269
1	PRECIPITACIÓN (mm/hora)	30503	1332	4.184
2	HUMEDAD (%)	31099	736	2.312
3	TEMPERATURA (°C)	31202	633	1.988
4	DIRECCION DEL VIENTO (°)	31325	510	1.602
5	AÑO / MES / DÍA	31835	0	0.000
6	HORA	31835	0	0.000

Según el reporte se puede observar valores nulos de un 10% en la variable VELOCIDAD DEL VIENTO, y valores nulos en las demás variables por debajo de un 5%. Este reporte permite determinar si es posible una imputación de datos, ya que los porcentajes de valores nulos no sobrepasan un 30%, es conveniente imputarlos.

Para la imputación de datos, se realizó una imputación iterativa con un máximo de 5 iteraciones siendo este una técnica secuencial que imputa los datos usando estimadores de regresión tomando como variables de salida (Y) a la columna imputar, y como variable de entrada (X) las demás columnas. Una vez realizada la estimación es una columna, estas estimaciones pasan a ser parte de las variables de entrada (X) para poder hallar así las siguientes columnas con valores nulos.

Tabla 3: Reporte de valores nulos ya imputados

	Variable	Cant. No Nulos	Cant. Nulos	% Nulos
0	TEMPERATURA (°C)	31835	0	0.0
1	PRECIPITACIÓN (mm/hora)	31835	0	0.0
2	HUMEDAD (%)	31835	0	0.0
3	DIRECCION DEL VIENTO (°)	31835	0	0.0
4	VELOCIDAD DEL VIENTO (m/s)	31835	0	0.0

Luego de la imputación, es necesaria la conversión de la variable AÑO/MES/DÍA de horas a días, eliminando así la variable HORA. Así como se muestra en la tabla 3. Con una reducción del 96% de registros de 31835 a 1412.

Tabla 4: Conversión a días de las variables

	AÑO / MES / DÍA	TEMPERATURA (°C)	HUMEDAD (%)	DIRECCION DEL VIENTO (°)	VELOCIDAD DEL VIENTO (m/s)	PRECIPITACIÓN (mm/dia)
0	2020-01-01	15.34	90.00	189.25	1.12	5.4
1	2020-01-02	14.00	90.80	199.80	0.55	16.8
2	2020-01-03	14.97	86.33	206.33	0.58	4.0
3	2020-01-04	15.09	94.92	250.96	0.92	14.4
4	2020-01-05	13.97	87.38	163.88	0.89	19.6
...	...	...	...	...	...	...
1407	2023-12-27	16.18	87.67	203.25	0.00	4.0
1408	2023-12-28	15.68	81.12	177.75	0.00	1.6
1409	2023-12-29	15.77	92.04	166.17	0.00	1.4
1410	2023-12-30	15.70	88.17	195.35	0.00	2.2
1411	2023-12-31	15.97	91.58	200.08	0.00	1.2

1412 rows × 6 columns

B. Revisión de datos duplicados ´

Esta etapa consta de la revisión de datos duplicados, para evitar el error en el análisis de los datos, ya que con datos iguales no se puede considerar el movimiento de la variable. Esta etapa se uso el comando duplicated para verificar los datos repetidos, determinando las variables no cuentan con datos repetidos según como lo muestra en la tabla 5.

Tabla 5: Reporte de datos duplicados

```
data_MP_diario_imp[data_MP_diario_imp.duplicated(keep=False)]
```

AÑO / MES / DÍA	TEMPERATURA (°C)	HUMEDAD (%)	DIRECCION DEL VIENTO (°)	VELOCIDAD DEL VIENTO (m/s)	PRECIPITACIÓN (mm/dia)

C. Análisis Univariante´

Considerando que nuestras variables se conforman de datos numéricos, se consideró solo un análisis numérico. Donde se analiza la normalidad de las variables mediante histogramas y box-plots para analizar la distribución de los datos ante la media de sus variables. Determinando así según la Figura 1 que las variables TEMPERATURA, HUMEDAD y DIRECCIÓN DEL VIENTO podrían conservar una distribución normal entre sus datos.

Figura 1: Histograma de las variables

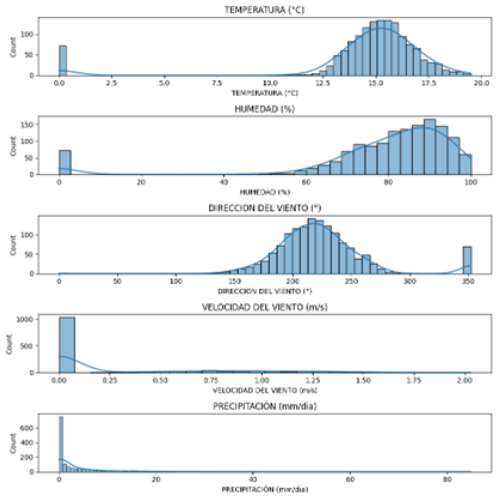
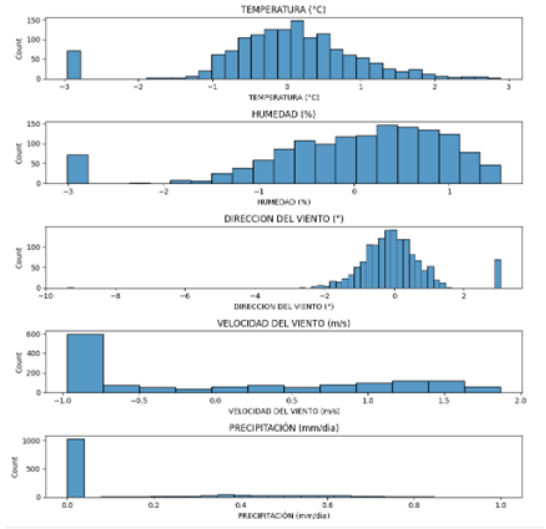


Figura 2: Histograma de las variables normalizadas



Para la normalización de datos se hizo uso de los métodos 'PowerTransformer' para las variables TEMPERATURA, HUMEDAD, DIRECCIÓN DEL VIENTO, PRECIPITACIÓN y del método 'MinMaxScaler' para la variable VELOCIDAD DEL VIENTO, esta última normalización se determinó por el pequeño indicio de una mejor distribución ante los demás métodos, mostrado en la Figura 2. Sobre la demás variable no se mostró una mejora en la distribución. La normalización se determinó en base a las Pruebas de Shapiro y Anderson, mostrando en el Test de Shapiro de todas las variables un p-valor menor a 5%, es decir, que las variables no mantienen una distribución normal.

D. Revisión y eliminación de Outliers ´

Para eliminar los outliers se hizo uso del método de algoritmo ECOD, ya que detecta los valores anormales no supervisados mediante funciones empíricas de distribución acumulativa, esta estima la distribución subyacente de los datos explicativos de manera no paramétrica, calculando la distribución acumulativa empírica por dimensión de los datos, luego estima las probabilidades de cola por dimensión por cada punto de datos (Saracco, s.f.)[3]. Eliminando así el 20% de los datos de las variables, para obtener un modelo con el mínimo de outliers demostrado figura 5. Para demostrar esta modificación es necesario hacer una comparación de las variables con outliers (Figura 3) y sin outler (Figura 4).

Figura 3: Dataset normalizada e imputada con outliers

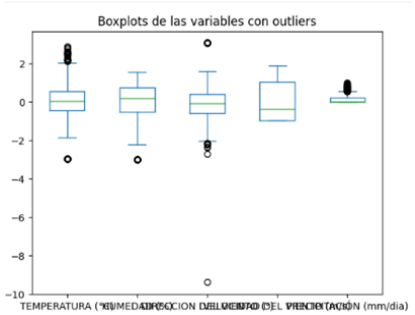
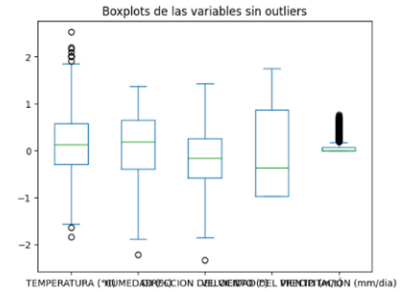
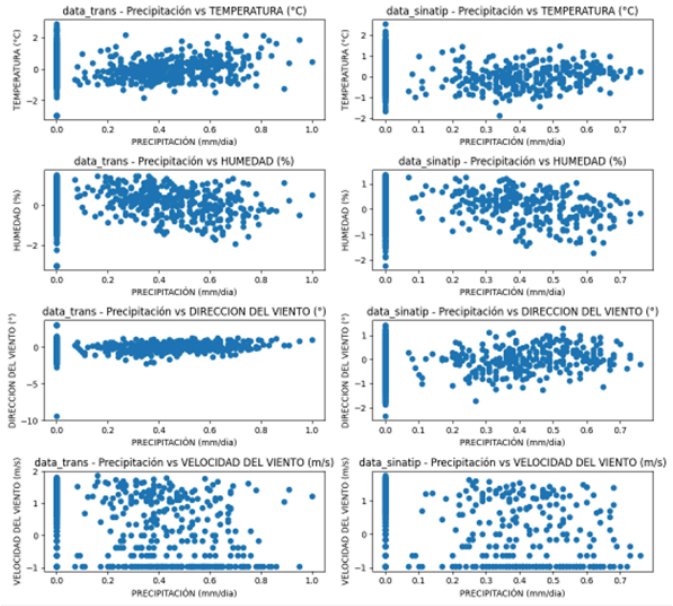


Figura 4: Dataset sin outliers



Esta comparación muestra un cambio mínimo en la disminución de los outliers sobre todo en la 3ra variable (DIRECCIÓN DEL VIENTO). Para una visualización más efectiva hicimos una comparación mediante gráficos de dispersión con relación a la variable dependiente mostrada en la figura 6, que nos permite ver que no existe un cambio drástico frente a la dispersión de los datos, con excepción de las variable DIRECCIÓN DEL VIENTO, siendo los gráficos del dataset 'data_trans' las variables con outliers y 'data_sinatip' las variables sin outliers.

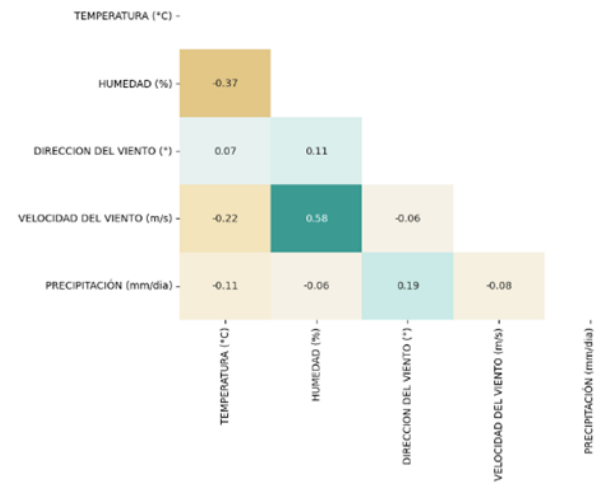
Figura 6: Gráficos de dispersión con y sin outliers



E. Análisis de Correlación

Como último etapa tras el tratamiento de los datos, se hizo una correlación para ver la relación de los variables linealmente mostrada mediante un mapa de calor en la figura 7, donde se visualiza una mayor correlación de 19% entre la variable DIRECCIÓN DEL VIENTO y la variable dependiente, y una relación inversa mínima de -6% de la variable HUMEDAD frente a la variable dependiente.

Figura 7: Mapa de calor de correlación



VI. CONCLUSIONES

El resultado de la normalización de datos no muestra un cambio drástico frente a la distribución de normalidad de las variables, rechazando la hipótesis nula de la prueba de Anderson de que poseen una distribución normal y obteniendo en el test de Shapiro un p-valor menor al 5% demostrando así que las variables no logran tener una distribución normal.

La eliminación de outliers no muestran cambios mayores en la dispersión de datos de las variables con excepción de la DIRECCIÓN DEL VIENTO.

El análisis exploratorio nos permite determinar si es posible una predicción de datos, tras el resultado de correlación entre las variables menor de 20% después del tratamiento de los datos, nos indica que una predicción de estos datos no serían precisos.

VII. BENEFICIOS

La implementación de esta propuesta traerá beneficios para los investigadores y entidades públicas que buscan crear un modelo de predicción de lluvia en el Perú, especialmente en las zonas altoandinas. Asimismo, la información que se puede obtener de esta data servirá para una mejor toma de decisiones y una mejor planificación de las actividades turísticas que forman parte de esta región.

REFERENCIAS

[1] Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022). A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning. Weather and Forecasting, 37(8), 1509-1529. https://doi.org/10.1175/WAF-D-22-0070.1

[2] Infobae, Vuelos suspendidos a Cusco por fuerte temporal de nieve y lluvias (2023) https://www.infobae.com/peru/2023/05/08/nieve-en-cusco-vuelos-suspendidos-por-fuerte-temporal/

[3] Saracco, Detección de anomalías simplificada con PyOD, (2023) https://medium.com/data-reply-it-datatech/anomaly-detection-made-easy-with-pyod-960faf6da4e5

[4] Sulca, J., M. Vuille, O. E. Timm, B. Dong, and R. Zubieta, (2021) Empirical–Statistical Downscaling of Austral Summer Precipitation over South America, with a Focus on the Central Peruvian Andes and the Equatorial Amazon Basin. J. Appl. Meteor. Climatol., 60, 65–85, https://doi.org/10.1175/JAMC-D-20-0066.1

[5] Sarasa-Cabezuelo, A. Prediction of Rainfall in Australia Using Machine Learning. Information (2022) 13, 163. https://doi.org/10.3390/info13040163