
Evaluación de Modelos de Machine Learning para Pronóstico de Precipitación de Corto Plazo en Guatemala

Isaac Solórzano Quintana



UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ciencias y Humanidades



**Evaluación de Modelos de Machine Learning para
Pronóstico de Precipitación de Corto Plazo en
Guatemala**

Trabajo de graduación en modalidad de tesis presentado por
Isaac Solórzano Quintana
para optar al grado académico de Licenciado en Física

Guatemala
2025

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ciencias y Humanidades



**Evaluación de Modelos de Machine Learning para
Pronóstico de Precipitación de Corto Plazo en
Guatemala**

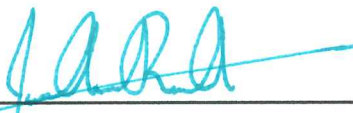
Trabajo de graduación en modalidad de tesis presentado por
Isaac Solórzano Quintana
para optar al grado académico de Licenciado en Física

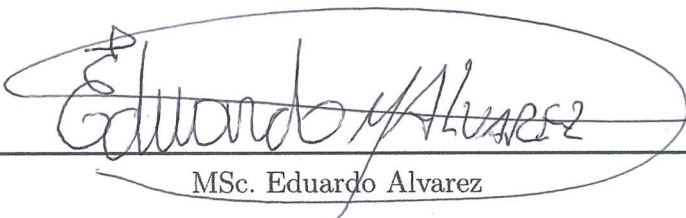
Guatemala
2025


Vo.Bo.:

(f) 
Lic. Jorge Rodríguez

Tribunal Examinador:

(f) 
Lic. Jorge Rodríguez

(f) 
MSc. Eduardo Alvarez

(f) 
Lic. Luis Tun

Fecha de aprobación: Guatemala, 4 de junio de 2025.

Agradecimientos

A Dios, por permitirme llegar hasta aquí.

Al Departamento de Investigación y Servicios Meteorológicos de INSIVUMEH, por abrirme las puertas y proporcionar todos los datos necesarios para el desarrollo de esta investigación.

A mi asesor y amigo, Lic. Jorge Rodríguez, quien desde el primer día manifestó su interés por el proyecto. Su orientación y apoyo constante han sido fundamentales para el desarrollo y culminación exitosa de este trabajo.

A la Fundación Juan Bautista Gutiérrez, por creer en mí y en mi sueño.

A mis padres, Ing. Ramiro Solórzano y Lcda. Milvia de Solórzano, por su amor incondicional e invaluable enseñanzas a lo largo de todos estos años. Sin ellos, no estaría hoy donde estoy.

A todo aquel que confió en mí y me dio una oportunidad. ¡Muchas gracias!

Agradecimientos	III
Lista de figuras	VII
Lista de cuadros	IX
Glosario	X
Resumen	XI
1. Introducción	1
2. Objetivos	2
2.1. Objetivo general	2
2.2. Objetivos específicos	2
3. Justificación	3
4. Marco teórico	4
4.1. Patrones y tendencias de la precipitación	4
4.1.1. Influencia de la temperatura	4
4.1.2. Influencia de la humedad relativa	4
4.1.3. Influencia de la presión atmosférica	5
4.2. Dinámica regional de la precipitación	5
4.3. Modelos numéricos de predicción meteorológica	7
4.4. Introducción a machine learning	9
4.4.1. Tipos de algoritmos de machine learning	10
4.5. Algoritmos de regresión	11
4.5.1. Evaluación de algoritmos de regresión	12
4.6. Algoritmos de clasificación	13
4.6.1. Evaluación de algoritmos de clasificación	14
4.7. Algoritmos de machine learning	15
4.7.1. Regresión lineal	15
4.7.2. Regresión polinomial	16
4.7.3. Regresión logística	17
4.7.4. Bosque aleatorio	18
4.7.5. XGBoost	19

5. Antecedentes	21
6. Metodología	25
6.1. Especificaciones de <i>hardware</i> y <i>software</i>	25
6.2. Definiciones	26
6.3. Procedimiento experimental	27
6.3.1. Variables objetivo	27
6.3.2. Variación de parámetros	28
6.3.3. Preprocesamiento de datos	30
6.3.4. Evaluación de modelos y comparación con WRF y NextGen	31
6.4. Diagrama de flujo procedimental	33
7. Resultados	34
7.1. Modelos de regresión	34
7.2. Modelos de clasificación de cotas percentiles	38
7.3. Modelos de clasificación de lluvia binaria	44
8. Discusión y análisis de resultados	50
8.1. Análisis de rendimiento	50
8.2. Análisis de variación de parámetros	51
8.3. Aplicabilidad de los modelos	53
8.4. Interpretación de métricas de evaluación	54
8.5. Discusión general	54
9. Conclusiones	56
10.Recomendaciones	57
11.Bibliografía	59
12.Anexos	62

Lista de figuras

4.1. Región Mar Intra-Americano	6
4.2. Representación discreta de un campo continuo	8
4.3. Ejemplo de evolución de la precipitación en un modelo NWP	8
4.4. Diferencia entre machine learning y programación tradicional	9
4.5. Ejemplo de modelo de regresión	11
4.6. Ejemplo de modelo de clasificación	13
4.7. Ejemplo de árbol de decisión	18
5.1. Mapa de estaciones convencionales de INSIVUMEH	21
5.2. Representación gráfica de los archivos de datos <i>raster</i> proporcionados por el modelo WRF-BMJ GFS	22
6.1. Esquema de la estructura temporal aplicada en el análisis predictivo	26
6.2. Representación gráfica del funcionamiento de los modelos de machine learning	27
6.3. Diagrama de flujo procedimental	33
7.1. Rendimiento de modelos de regresión durante todo el año	35
7.2. Rendimiento de modelos de regresión durante la época lluviosa	35
7.3. Rendimiento de modelos de regresión durante la época seca	36
7.4. MAE promedio de modelos de regresión según algoritmos	36
7.5. MAE promedio de modelos de regresión según número de árboles	37
7.6. MAE promedio de modelos de regresión según combinación de variables	37
7.7. MAE promedio de modelos de regresión según número de días en el rango de observación	37
7.8. MAE promedio de modelos de regresión según tipo de entrenamiento	38
7.9. Rendimiento de modelos de clasificación con cotas percentiles durante todo el año	39
7.10. Rendimiento de modelos de clasificación con cotas percentiles durante la época lluviosa	40
7.11. Rendimiento de modelos de clasificación con cotas percentiles durante la época seca	41
7.12. Exactitud promedio de modelos de clasificación con cotas percentiles según algoritmos	42
7.13. Exactitud promedio de modelos de clasificación con cotas percentiles según número de árboles	43
7.14. Exactitud promedio de modelos de clasificación con cotas percentiles según combinación de variables	43
7.15. Exactitud promedio de modelos de clasificación con cotas percentiles según número de días en el rango de observación	43
7.16. Exactitud promedio de modelos de clasificación con cotas percentiles según tipo de entrenamiento	44
7.17. Rendimiento de modelos de clasificación de lluvia binaria durante todo el año	45
7.18. Rendimiento de modelos de clasificación de lluvia binaria durante la época lluviosa	46

7.19. Rendimiento de modelos de clasificación de lluvia binaria durante la época seca . . .	47
7.20. Exactitud promedio de modelos de clasificación de lluvia binaria según algoritmos .	48
7.21. Exactitud promedio de modelos de clasificación de lluvia binaria según número de árboles	48
7.22. Exactitud promedio de modelos de clasificación de lluvia binaria según combinación de variables	49
7.23. Exactitud promedio de modelos de clasificación de lluvia binaria según número de días en el rango de observación	49
7.24. Exactitud promedio de modelos de clasificación de lluvia binaria según tipo de entrenamiento	49
B.1. Distribución de clases según variable categórica	63

Lista de cuadros

4.1. Cantidades importantes para evaluación de algoritmos de clasificación	14
4.2. Métricas de evaluación de algoritmos de clasificación	14
4.3. Ejemplo de matriz de confusión	15
6.1. Interpretación de la variable objetivo para los modelos de clasificación con cotas percentiles	28
6.2. Algoritmos de machine learning a implementar	29
6.3. Combinaciones de variables meteorológicas	29
6.4. Cantidad y proporción de valores faltantes en la base de datos original	30
7.1. Rendimiento del modelo WRF-BMJ-GFS para regresión	36
7.2. Rendimiento del modelo WRF-BMJ-GFS para clasificación con cotas percentiles	42
7.3. Rendimiento del modelo NextGen para clasificación con cotas percentiles	42
7.4. Rendimiento del modelo WRF-BMJ-GFS para clasificación de lluvia binaria	48
8.1. Cuadro comparativo entre los resultados de machine learning (ML) y WRF	51
D.1. MAE para modelos de regresión evaluados con datos de todo el año	65
D.2. MAE para modelos de regresión evaluados con datos de la época lluviosa	66
D.3. MAE para modelos de regresión evaluados con datos de la época seca	67
D.4. MdAE para modelos de regresión evaluados con datos de todo el año	68
D.5. MdAE para modelos de regresión evaluados con datos de la época lluviosa	69
D.6. MdAE para modelos de regresión evaluados con datos de la época seca	70
E.1. Exactitud para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año	72
E.2. Exactitud para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa	73
E.3. Exactitud para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca	74
E.4. Precisión de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año	75
E.5. Precisión de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa	76
E.6. Precisión de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca	77
E.7. Sensibilidad de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año	78

E.8. Sensibilidad de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa	79
E.9. Sensibilidad de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca	80
E.10. Especificidad de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año	81
E.11. Especificidad de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa	82
E.12. Especificidad de la categoría <i>NN</i> para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca	83
F.1. Exactitud para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año	85
F.2. Exactitud para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa	86
F.3. Exactitud para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca	87
F.4. Precisión para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año	88
F.5. Precisión para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa	89
F.6. Precisión para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca	90
F.7. Sensibilidad para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año	91
F.8. Sensibilidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa	92
F.9. Sensibilidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca	93
F.10. Especificidad para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año	94
F.11. Especificidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa	95
F.12. Especificidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca	96
G.1. Tiempos de entrenamiento para los modelos de regresión	98
G.2. Tiempos de entrenamiento para los modelos de clasificación con cotas percentiles . .	99
G.3. Tiempos de entrenamiento para los modelos de clasificación de lluvia binaria	100

<i>Ensemble de modelos:</i>	Conjunto de modelos que se combinan para estimar una variable, permitiendo capturar en mayor medida la incertidumbre y mejorar las predicciones.
Clima:	Se refiere al estado promedio de las condiciones atmosféricas en un sitio particular durante un período largo de tiempo.
Climatología:	Es el estudio científico del clima.
Datos <i>raster</i>:	Es un formato de datos geospaciales donde el dominio se representa mediante una cuadrícula en la que cada celda contiene un valor que representa información.
Dominio geográfico:	Región del espacio terrestre sobre la cual se recopilan, analizan o representan datos atmosféricos.
Meteorología:	Ciencia que estudia el comportamiento del tiempo atmosférico, es decir, las condiciones atmosféricas en el corto plazo.
Método <i>kriging</i>:	Método geoestadístico de interpolación que estima valores desconocidos a partir de la correlación espacial entre puntos de datos conocidos.
Orografía:	Es el estudio y descripción de las montañas y, por extensión, de todas las formas de relieve terrestre.
Subsidencia:	Se refiere al descenso de una capa de aire sobre una zona extensa. Genera condiciones atmosféricas estables.
Tiempo:	Es el estado de la atmósfera en un lugar y momento determinado.
Vientos alisios:	Vientos que soplan constantemente desde los trópicos hacia el ecuador. Son impulsados por la rotación terrestre y la circulación atmosférica.

Esta investigación se centra en la evaluación de diferentes modelos de machine learning para el pronóstico de la precipitación en Guatemala con un día de anticipación. Se desarrollan modelos de clasificación de lluvia binaria para predecir la ocurrencia de lluvia y modelos basados en percentiles para estimar su intensidad en comparación con registros pasados. Además, se construyen modelos de regresión para pronosticar la anomalía absoluta de precipitación respecto del promedio climatológico. El entrenamiento de los modelos se realiza con la base de datos meteorológicos proporcionada por el Instituto Nacional de Sismología, Vulcanología, Meteorología e Hidrología (INSIVUMEH).

Cada modelo es sometido a tres evaluaciones: con datos de la época seca, de la época lluviosa y del año completo (evaluación general). El rendimiento de los modelos de machine learning es comparado con el del modelo numérico *Weather Research and Forecasting* (WRF), utilizado actualmente por INSIVUMEH en el pronóstico de precipitación diaria.

Los resultados muestran que la exactitud general se encuentra entre el 26 % y el 65 % para los modelos de percentiles y entre el 55 % y el 75 % para los modelos de lluvia binaria. Los modelos de regresión presentaron un MAE entre 5.14 mm y 6.50 mm al ser evaluados con datos de todo el año. El 23 % de los modelos de lluvia binaria y el 96 % de los modelos de regresión superaron los resultados de evaluación general de WRF, mientras que los modelos de percentiles obtuvieron resultados por debajo del modelo de referencia en esta misma evaluación. Las mejores combinaciones de parámetros para los modelos de clasificación resultaron ser *bosque aleatorio* y *XGBoost* con 100 árboles de decisión, y para los modelos de regresión fueron la *regresión lineal múltiple* y la *regresión polinomial*. No se observó un cambio relevante al implementar las diferentes combinaciones de variables y número de días en el rango de observación que se utilizan como predictores. En general, el rendimiento durante la época seca fue mejor en comparación con la época lluviosa, debido a la diferencia de la variabilidad de precipitación entre ambas épocas.

Debido a su ubicación en una región marítima tropical, Centroamérica presenta variaciones de temperatura generalmente pequeñas, y la precipitación es, con diferencia, el fenómeno meteorológico más importante (Portillo et al., 2024). La distribución de la precipitación está influenciada por una amplia gama de factores locales y globales, tanto de corto como de largo plazo, lo que genera una alta variabilidad y complejidad a lo largo del año. Esta variabilidad, combinada con las condiciones socioeconómicas de la región, hace que Centroamérica sea particularmente vulnerable a desastres naturales como inundaciones y deslaves.

El Departamento de Investigación y Servicios Meteorológicos de INSIVUMEH ha implementado diferentes modelos numéricos de predicción meteorológica (NWP, por sus siglas en inglés). Estos modelos hacen uso de los principios físicos que rigen la dinámica de la atmósfera para plantear y resolver numéricamente una serie de ecuaciones diferenciales parciales. El dominio geográfico es dividido en celdas. Conociendo el estado inicial en cada celda, se pueden resolver las ecuaciones para conocer la evolución del estado a lo largo del tiempo. Desde el año 2000, la exactitud de estos modelos ha mejorado gracias a una mayor disponibilidad de datos meteorológicos; sin embargo, a menudo requieren de gran poder de cómputo. (Waqas et al., 2024).

Recientemente, diferentes estudios han demostrado que la inteligencia artificial es una herramienta prometedora para mejorar, y potencialmente reemplazar a los modelos numéricos. A diferencia de los NWP, los modelos de inteligencia artificial no tienen conocimiento acerca de las leyes físicas involucradas, más bien se centran en aprender patrones complejos a partir de grandes conjuntos de datos (Waqas et al., 2024). En concreto, la subcategoría de la inteligencia artificial más importante para la ciencia atmosférica es el machine learning, siendo los algoritmos de tipo supervisado los que más se repiten en las publicaciones recientes (Bochenek y Ustrnul, 2022). Algoritmos como la regresión logística, bosque aleatorio y XGBoost han sido utilizados en el pronóstico de precipitación, a menudo superando el desempeño de los modelos tradicionales (Elbasheer et al., 2023; Grazzini et al., 2024; Ko, C-M et al., 2020).

A partir de este contexto, se desarrolla la presente investigación. Esta tiene como propósito, medir el desempeño de diversos modelos de machine learning en el pronóstico de precipitación a corto plazo en Guatemala. Para ello, se desarrollan múltiples modelos con diferentes combinaciones de parámetros, a fin de identificar una configuración óptima. Se realiza una comparación entre el rendimiento de estos modelos y los que utiliza actualmente INSIVUMEH, estableciendo así una referencia que permita valorar el potencial del machine learning frente a los métodos de pronóstico convencionales.

2.1. Objetivo general

Evaluar el desempeño de diferentes modelos de machine learning en el pronóstico de precipitación diaria en Guatemala con un día de antelación.

2.2. Objetivos específicos

- Evaluar el desarrollo de modelos de clasificación para la predicción cualitativa de la precipitación.
- Evaluar el desempeño de modelos de regresión en el pronóstico de un valor de precipitación esperado.
- Determinar qué modelo proporciona los mejores resultados.
- Comparar el desempeño de los distintos modelos durante la época seca y la época lluviosa.
- Comparar las predicciones obtenidas mediante machine learning con las proporcionadas por el modelo WRF que utiliza INSIVUMEH.

La República de Guatemala se sitúa en la posición 44 de los países con mayor riesgo de desastres en el mundo (WorldRiskIndex, 2024). De acuerdo con CONRED (2024), durante el período 2019-2023 se registraron en Guatemala un total de 12,796 incidentes ocasionados por lluvias. De hecho, solo durante 2023, la temporada de lluvias afectó a más de 4.5 millones de personas (CONRED, 2024), lo que representa el 30 % de la población total del país, de acuerdo con el Censo Nacional de Población y Vivienda realizado por el Instituto Nacional de Estadística (INE) (2019). Ante esta problemática, surge la necesidad de contar con un adecuado sistema de pronóstico de precipitación, que permita conocer con anticipación la ocurrencia de fenómenos potencialmente peligrosos, evitando así, pérdidas materiales, pero sobre todo, humanas.

La precipitación, y el tiempo atmosférico en general, se caracterizan por ser fenómenos altamente complejos y sensibles a condiciones iniciales. Esto significa que dos sistemas que difieren mínimamente en su estado inicial, pueden desembocar en estados completamente distintos al cabo de un tiempo. Esta naturaleza caótica dificulta la aplicación de los modelos de pronóstico tradicionales, ya que pequeños errores en la medición de los valores de entrada pueden amplificarse rápidamente. Así mismo, los métodos numéricos empleados suelen introducir errores que se acumulan con el tiempo, y reducir esta incertidumbre a menudo implica un mayor uso de recursos computacionales. En consecuencia, resulta fundamental explorar métodos alternativos que permitan mitigar estas limitaciones y mejorar la calidad de los pronósticos.

El uso del machine learning en el pronóstico meteorológico ha demostrado tener un rendimiento notable, superando en muchos casos la exactitud, eficiencia y adaptabilidad de los modelos numéricos. A pesar de que ya existen múltiples estudios sobre este tema, sigue siendo un área de investigación relativamente reciente, con múltiples interrogantes aún por resolver. Además, no se ha identificado ningún estudio publicado que evalúe la aplicación de estos modelos específicamente en Guatemala. Por lo tanto, esta investigación adquiere mayor relevancia al introducir en el país un enfoque innovador con gran potencial para mejorar los pronósticos de precipitación actuales.

4.1. Patrones y tendencias de la precipitación

La precipitación es un fenómeno meteorológico sumamente complejo, caracterizado por su alta variabilidad no solo en el tiempo, sino en el espacio. Entre los factores que intervienen directamente en la precipitación se encuentra la orografía del terreno, la circulación de corrientes de convección locales, la humedad relativa y la evaporación de agua, entre otros (Mesa y Peñaranda, 2015). A continuación, se explica de forma cualitativa, y a grandes rasgos, el impacto de tres variables meteorológicas en la precipitación.

4.1.1. Influencia de la temperatura

Cuando el aire está en contacto con un cuerpo líquido de agua, algunas moléculas de agua saltan constantemente del líquido al gas mediante evaporación, mientras que otras pasan del gas al agua mediante el proceso de condensación. Inicialmente, la tasa de evaporación es superior a la de condensación, por lo que el aire se va cargando de vapor de agua hasta que se llega a un equilibrio dinámico en el que se dice que el aire está saturado; a partir de este momento, las moléculas se condensan y evaporan a la misma velocidad (McIveen, 1992).

La presión de vapor de agua aumenta con la temperatura. Esto quiere decir que a mayor temperatura, el aire presenta mayor capacidad de almacenar humedad, lo que favorece la formación de nubes, creación de tormentas y, en general, una mayor probabilidad de precipitación. Esta relación se explica en que al incrementar la temperatura se aumenta la energía interna de las moléculas, por lo que éstas tienen una mayor facilidad de desprenderse de la capa más superficial del líquido (McIveen, 1992).

4.1.2. Influencia de la humedad relativa

La humedad relativa se define como el vapor de agua presente en el aire en comparación con el vapor máximo que el aire puede soportar. Una humedad relativa del 0 % indica que no hay presencia de agua en el aire, mientras que un valor del 100 % es indicador de que el aire está completamente saturado (McIveen, 1992).

Un aire saturado es aquel que ha llegado a su máxima capacidad de absorber vapor de agua, por lo que toda la humedad adicional es condensada en agua líquida. En este sentido, una mayor humedad relativa aumenta las probabilidades de precipitación, ya que si se sigue suministrando vapor de agua, eventualmente se llegará a un valor del 100 % y se comenzará a formar agua líquida. Es importante mencionar que la precipitación no ocurre inmediatamente después de que el agua comienza a condensarse, ya que las corrientes de convección presentes en la atmósfera empujan constantemente las pequeñas gotas de agua hacia arriba, impidiendo la precipitación. Para que las gotas caigan, es necesario que alcancen un tamaño suficientemente grande como para caer por su propio peso. Esto se logra mediante el agrupamiento por colisión de múltiples gotas, o simplemente mediante más condensación de agua (Wallace y Hobbs, 2006).

4.1.3. Influencia de la presión atmosférica

La presión influye directamente en el movimiento de las corrientes de vapor de agua en la atmósfera. El aire que se encuentra en contacto con la superficie terrestre tiende a calentarse debido a la energía proporcionada por los rayos solares; al calentarse, su densidad es menor en comparación con las capas más altas, por lo que tiende a ascender, llevando consigo la humedad absorbida mediante evaporación. Este proceso se conoce como convección. Cuando este aire caliente llega a las capas superiores de la atmósfera, se expande debido a la disminución en la presión atmosférica. Este aumento de volumen implica un descenso de temperatura mediante un proceso conocido como enfriamiento adiabático, por lo que el aire también se enfría. Al bajar la temperatura, el aire pierde capacidad de retención de vapor de agua, por lo que parte de la humedad se condensa, y se comienzan a formar pequeñas gotas de agua. La precipitación se ve influenciada por la pérdida de presión atmosférica que se experimenta con el aumento de altitud, lo que favorece el enfriamiento del aire en las capas superiores, y la consecuente condensación del exceso de vapor de agua (McIveen, 1992).

Cuando las capas de aire caliente ascienden por convección, se produce un vacío relativo de aire al nivel del suelo, lo que genera un sistema de baja presión. En este sentido, y en términos generales, un valor de presión atmosférica bajo (en comparación con el registro histórico) suele estar asociado a altas precipitaciones, pues es un fuerte indicador de que el proceso de convección ha estado presente en la atmósfera durante algún tiempo transportando humedad hacia las capas más altas de aire (McIveen, 1992).

4.2. Dinámica regional de la precipitación

La precipitación en Centroamérica presenta altas variaciones locales y estacionales, las cuales se atribuyen en gran parte a la dinámica de la atmósfera centroamericana y su interacción con la topografía local. A nivel general, la distribución de precipitación de la región presenta dos máximos: uno en junio, y un segundo (y más largo) en septiembre. En medio de estos dos máximos se presenta un mínimo local en el período julio-agosto conocido como *canícula*. La temporada seca ocurre durante el invierno e inicio de la primavera del hemisferio norte, que comprende aproximadamente el período noviembre-abril, y suele ser más intensa en la costa pacífica en comparación con la atlántica (Alfaro, 2002).

La ubicación geográfica de Centroamérica juega un papel importante en la variabilidad climática de la región. Se encuentra rodeada por dos grandes masas de agua: el Océano Pacífico Tropical Este a un lado, y el Mar Caribe al otro; esta región junto con el Golfo de México recibe el nombre de Mar Intra-Americano (IAS, por sus siglas en inglés), cuyo mapa se muestra en la Figura 4.1. El IAS es una región sensible a sistemas dinámicos que ocurren en su vecindad, tanto de carácter regional como de largo alcance. Algunos ejemplos son: los fuertes vientos provenientes del este, un gran cinturón de calor latente, la temperatura superficial del mar, así como la precipitación (Maldonado, 2016).



Figura 4.1: Región Mar Intra-Americano

Fuente: Maldonado (2016)

Uno de los fenómenos que tiene mayor impacto en la variabilidad de lluvias de la región es el anticiclón NASH (del inglés *North Atlantic Subtropical High*). Se trata de un anticiclón subtropical semipermanente dominado por dos patrones espaciales: el primero consiste en una región de alta presión que ocurre en la cuenca del Atlántico durante el verano boreal, mientras que el segundo incluye múltiples anticiclones en el este de Norteamérica y noroeste de África (Davis et al., 1997). El NASH ha demostrado inducir grandes corrientes de viento en el Caribe provenientes del este, así como el descenso de masas de aire seco mediante un proceso conocido como *subsistencia* (Martínez et al., 2019).

Otro fenómeno de gran relevancia en la región es la Zona de Convergencia Intertropical (ITCZ), ubicada en torno al ecuador en el rango $\pm 10^\circ$ de latitud. En esta zona se produce el encuentro de los vientos alisios del norte y del sur de la circulación de Hadley; dando lugar a la convergencia superficial de aire húmedo, su ascenso y, en consecuencia, a la precipitación. Se estima que en esta región se produce un tercio de la precipitación total de la Tierra (Wolf et al., 2021).

Por otro lado, las corrientes de convección, principalmente en la región del Caribe, se ven influenciadas por el Reservorio de Agua Caliente del Atlántico (AWP). Este hace referencia a áreas de agua con temperatura superficial superior a los 28.5°C distribuidas en el Golfo de México, el Mar Caribe y el Atlántico Norte Tropical Occidental. El AWP está ausente durante el invierno boreal, comienza a surgir en torno a junio y alcanza su tamaño máximo en septiembre. La precipitación que ocurre en el período de agosto a septiembre en las regiones del Caribe, Centroamérica y este de Suramérica se debe principalmente al tamaño del AWP (Martínez et al., 2019; Wang y Enfield, 2006).

El NASH, el ITCZ y el AWP son fenómenos de largo alcance que se combinan para dar lugar a las lluvias que caracterizan el verano boreal en Centroamérica y el Caribe. Sin embargo, durante el invierno, los fuertes vientos, la subsistencia, una baja temperatura superficial del mar y la migración hacia el sur de la ITCZ provocan una ausencia de lluvias, por lo que el invierno boreal es una época seca (Martínez et al., 2019).

Durante la temporada lluviosa, los fenómenos de largo alcance anteriormente mencionados se combinan con otros fenómenos de alcance regional y mecanismos de retroalimentación. Uno de estos fenómenos regionales es el Chorro de Bajo Nivel del Caribe (CLLJ), principal responsable del transporte de humedad proveniente del Mar Caribe, y a cuya variabilidad se atribuye, en gran medida, el fenómeno de la canícula. Otro factor que también influye son las Ondas del Este Africanas, perturbaciones de escala sinóptica que se propagan a través del Atlántico en el período junio-octubre, y que son fuente de precipitaciones, principalmente en la región del Caribe, además de ser responsables

de la mitad de ciclones tropicales y cerca del 85 % de los huracanes que se generan en la región (Martinez et al., 2019).

4.3. Modelos numéricos de predicción meteorológica

Debido al avance exponencial de la computación, en los últimos años, los modelos numéricos de predicción meteorológica (NWP, del inglés *Numerical Weather Prediction*) se han convertido en la herramienta más importante para el pronóstico meteorológico, siendo utilizados por la mayoría de países alrededor del mundo (Pu y Kalnay, 2018).

El eje central de los NWP's consiste en resolver una serie de ecuaciones diferenciales parciales que gobiernan el comportamiento y evolución de la atmósfera. El sistema consiste en un total de 7 ecuaciones con 7 incógnitas, las cuales corresponden a las siguientes leyes físicas: conservación de momento (3 ecuaciones, una por cada componente de la velocidad), ecuación de continuidad o conservación de masa, ecuación de estado para un gas ideal, la primera ley de la termodinámica o conservación de la energía, y una ecuación para la conservación de masa de agua (Pu y Kalnay, 2018). Estas leyes se muestran, en el mismo orden, en el sistema de ecuaciones 4.1.

$$\left\{ \begin{array}{l} \frac{d\vec{V}}{dt} = -\alpha\vec{\nabla}p - \vec{\nabla}\phi + \vec{F} - 2\vec{\Omega} \times \vec{V} \\ \frac{\partial\rho}{\partial t} = -\vec{\nabla} \cdot (\rho\vec{V}) \\ p\alpha = RT \\ Q = C_P \frac{dT}{dt} - \alpha \frac{dp}{dt} \\ \frac{\partial\rho q}{\partial t} = -\vec{\nabla} \cdot (\rho\vec{V}q) + \rho(E - C) \end{array} \right. \quad (4.1)$$

Donde:

- \vec{V} es la velocidad del aire
- t es el tiempo
- ρ es la densidad
- $\alpha = 1/\rho$ es el volumen específico
- p es la presión
- T es la temperatura
- ϕ es la energía potencial gravitatoria
- q es la razón de vapor de agua presente en la atmósfera (kg de agua por cada kg de aire)
- Q es el calor
- E es la evaporación
- C es la condensación
- R es la constante de gas ideal
- \vec{F} es la fuerza de fricción
- $\vec{\Omega}$ es la velocidad angular de rotación de la Tierra (con lo que el término $-2\vec{\Omega} \times \vec{V}$ es la aceleración debida a la fuerza de Coriolis)

Las ecuaciones del sistema 4.1 se conocen como *ecuaciones primitivas*, y conforman un conjunto de ecuaciones universales para los NWP's. Generalmente, cada modelo trabaja con una versión simplificada de las ecuaciones al asumir ciertas características en el sistema; esto permite tener una aplicación única para las diferentes regiones del mundo, así como enfocarse en procesos atmosféricos específicos (Knievel, 2006; Pu y Kalnay, 2018).

En los NWP's, el estado de la atmósfera en cualquier instante se representa por medio de una cuadrícula tridimensional de puntos que cubre todo el dominio espacial. La distancia entre dos puntos adyacentes corresponde a la resolución del modelo. Cada punto contiene el valor de las diferentes variables meteorológicas asociadas a dicha ubicación espacial. Este proceso se ejemplifica en la Figura 4.2 para un caso bidimensional. En los últimos años, los avances en términos de capacidad computacional han permitido incrementar la resolución de los modelos desde algunos cientos de kilómetros en la década de 1950, (Collins et al., 2013) hasta modelos actuales con casi 1 km de resolución (Waqas et al., 2024).

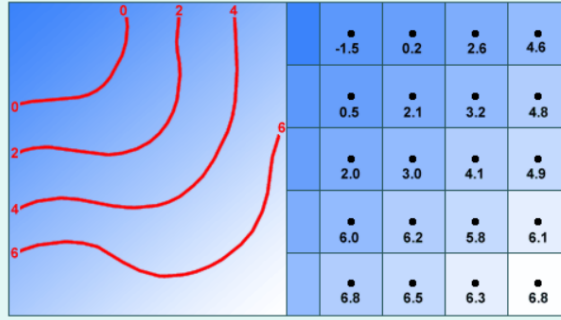


Figura 4.2: Representación discreta de un campo continuo
Fuente: The COMET Program (2015)

Los NWP's funcionan mediante la discretización del tiempo en intervalos de tamaño Δt . Indicando el estado inicial de la atmósfera en el tiempo t_0 , así como las condiciones de frontera de la región bajo análisis, se puede construir el estado del sistema en un tiempo $t_1 = t_0 + \Delta t$; este estado se puede utilizar nuevamente como estado inicial para construir el siguiente estado $t_2 = t_1 + \Delta t$. Siguiendo este proceso iterativo, el modelo es capaz de simular la evolución temporal de la atmósfera. De forma matemática, y a grandes rasgos, el método consiste en determinar la función $F(\psi, t)$ que define la evolución del sistema como se muestra en la ecuación 4.2, y que depende del tiempo t y del estado actual ψ . A partir de esta función se puede encontrar el estado en el siguiente paso de tiempo mediante la ecuación 4.3 (Pu y Kalnay, 2018). En la Figura 4.3 se muestra una representación gráfica de este proceso para un caso hipotético de precipitación.

$$F(\psi, t) = \frac{\partial \psi}{\partial t} \quad (4.2)$$

$$\psi|_{t+\Delta t} = \psi|_t + F(\psi|_t, t)\Delta t \quad (4.3)$$

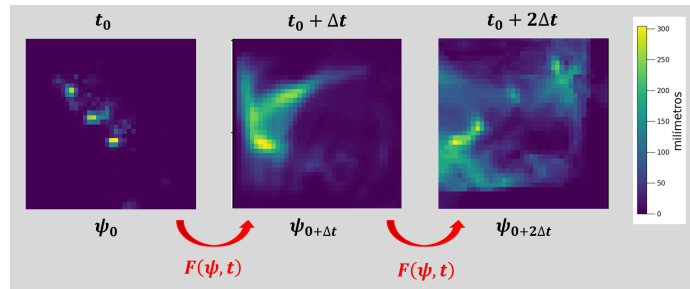


Figura 4.3: Ejemplo de evolución de la precipitación en un modelo NWP
Fuente: elaboración propia

4.4. Introducción a machine learning

La programación tradicional se fundamenta, a grandes rasgos, en dos pasos: primero, se crea un diseño detallado del paso a paso que debe seguir el algoritmo para resolver la tarea, y segundo, dicho algoritmo es implementado en una computadora mediante un lenguaje de programación. Esta forma de resolver tareas ha funcionado bastante bien prácticamente desde los inicios de la computación; sin embargo, carece de un problema fundamental que limita su implementación en muchos problemas de la vida cotidiana, y es que en ocasiones resulta difícil establecer una lista de pasos que defina la realización exitosa de una tarea. Por ejemplo, si se desea programar un algoritmo para clasificar un correo como spam o no spam, no basta con un conjunto de instrucciones objetivas, ya que lo que para una persona es spam, para la otra puede no serlo. Es aquí donde la inteligencia artificial, y en concreto, el machine learning, surge como una solución, pues éste se fundamenta en el uso de algoritmos y técnicas para automatizar soluciones a problemas complejos que son difíciles de implementar mediante la programación convencional (Rebala et al., 2019).

La inteligencia artificial se basa en replicar el comportamiento humano, y traducirlo en términos computacionales para que un ordenador pueda comportarse de manera similar. Las tres funciones básicas de la inteligencia humana son recordar, adaptarse y generalizar: la persona recuerda que cuando actuó de cierta forma ante determinada situación, el resultado fue positivo, por lo que lo más seguro es que el comportamiento se repita. De manera similar, un modelo de inteligencia artificial recuerda que cuando vio ciertos datos y proporcionó cierta respuesta, ésta estuvo correcta, de manera que lo más seguro es que lo vuelva a repetir, mientras que si la respuesta fue incorrecta, intentará dar nuevas respuestas hasta dar con la correcta, momento en el que aprenderá cómo reaccionar ante la situación. Mediante la simulación artificial de la inteligencia humana, se logra que un programa recuerde, se adapte y generalice sobre una gran cantidad de datos con el fin de responder ante los estímulos que se le proporcionan (Marsland, 2015).

El machine learning (ML) combina la estadística, la inteligencia artificial y la informática; consiste en la programación de computadoras para optimizar el rendimiento de un modelo matemático que hace uso de ciertos parámetros para transformar los datos de entrada en la salida deseada (Alpaydin, 2014). Los algoritmos de machine learning son capaces de extraer información a partir de datos; información que, para un humano, sería muy complicado obtener. No necesitan una programación explícita sobre las reglas que rigen el fenómeno, pues son capaces de aprender y descifrar dichas reglas a partir de los datos (Rebala et al., 2019).

En la Figura 4.4 se resume la gran diferencia entre el machine learning y la programación tradicional: mientras que la programación tradicional requiere el ingreso de un programa (serie de instrucciones) para transformar los datos de entrada en la salida, el machine learning analiza las entradas y las salidas para generar el programa que rige la relación entre estas dos.

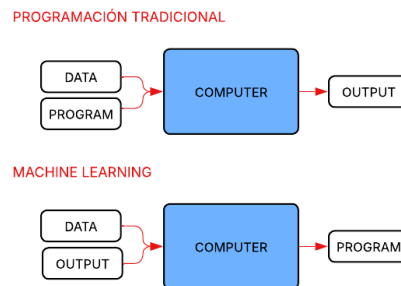


Figura 4.4: Diferencia entre machine learning y programación tradicional

Fuente: elaboración propia

El machine learning tiene múltiples aplicaciones en tareas que requieren la toma de decisiones basadas en la experiencia, y para las cuales no existe una serie de pasos a seguir que se apliquen en todas las situaciones por igual. Algunos ejemplos son: traducir un texto escrito a mano, determinar si un tumor es maligno o benigno, identificar los temas que se tratan en un blog de internet, o incluso segmentar los clientes de una empresa en grupos con preferencias similares (Müller y Guido, 2017).

4.4.1. Tipos de algoritmos de machine learning

En función de la estrategia utilizada para optimizar el modelo de machine learning, estos pueden dividirse, a muy grandes rasgos, en dos categorías principales:

Algoritmos supervisados

El aprendizaje supervisado ocurre cuando a partir de los datos de entrada \vec{X} y una variable de salida Y se utiliza un algoritmo para descifrar la función f que mapea la entrada en la salida:

$$Y = f(\vec{X}) \quad (4.4)$$

El objetivo es aproximar la función de mapeo lo suficientemente bien de tal forma que cuando se presenten nuevos valores de entrada \vec{X} , el modelo sea capaz de predecir correctamente la variable de salida Y (Brownlee, 2016).

En los algoritmos supervisados, el programa es alimentado con un conjunto de *datos de entrenamiento*. Estos datos contienen múltiples instancias del fenómeno, y son *etiquetados*, es decir, no solo presentan los valores de entrada para cada instancia, sino también el valor correcto de la variable de salida. No obstante, si se conociera el valor de respuesta para todos los posibles valores de entrada, el machine learning perdería sentido, porque simplemente bastaría con referirse a la base de datos para conocer el valor deseado. El machine learning presenta la ventaja de ser generalizable, en el sentido de que, a partir de los datos de entrenamiento, el modelo deberá ser capaz de producir salidas correctas para datos de entrada que no estuvieron presentes durante el entrenamiento (Marsland, 2015).

Los algoritmos supervisados resuelven las tareas de una manera indirecta: primero, se genera un modelo basado en el entrenamiento a partir de los datos etiquetados, para posteriormente hacer uso del modelo para predecir la variable de salida a partir de un nuevo conjunto de datos de entrada (Rebala et al., 2019). Se llama aprendizaje supervisado porque el proceso que sigue el algoritmo a partir de los datos de entrenamiento es análogo a un profesor supervisando el proceso de aprendizaje (Brownlee, 2016).

Los algoritmos supervisados pueden ser de regresión, si la variable de salida Y es continua, por ejemplo el precio de una casa o el salario de una persona. Por el contrario, si la salida Y es una variable categórica se dice que el algoritmo es de clasificación, por ejemplo al predecir la especie de una planta o la marca de un vehículo (Marsland, 2015). Más adelante se abordará en detalle esta distinción de algoritmos, así como algunos ejemplos de los mismos.

Algoritmos no supervisados

En el aprendizaje supervisado, es necesario conocer el valor de salida asociado a cada instancia de la base de datos; sin embargo, en ocasiones, puede resultar muy complicado obtener dicho valor, ya sea porque alguien debe ingresarlo manualmente y la base de datos es demasiado grande, o

porque simplemente no se conoce la respuesta correcta. En dichos casos, hacer uso del aprendizaje no supervisado suele ser una buena idea, donde, si bien el modelo no será capaz de predecir la variable de salida, sí se podrán hacer tareas relevantes como agrupar instancias que tienen características en común, o encontrar reglas de asociación entre los datos (Marsland, 2015).

En este tipo de aprendizaje no se cuenta con la variable de salida, sino únicamente con los datos de entrada \vec{X} . El propósito es modelar la estructura subyacente o la distribución de los datos con el fin de encontrar patrones de comportamiento ocultos a simple vista. Se conocen como no supervisados porque, a diferencia del aprendizaje supervisado, no se cuenta con respuestas correctas, por lo que no hay un "profesor" que guíe el aprendizaje, sino que se deja que el algoritmo trabaje por su cuenta en el análisis de datos (Brownlee, 2016).

Los datos de entrada generalmente presentan una estructura donde ciertos patrones se repiten más que otros; una estrategia para encontrar estos patrones, y que se utiliza mucho en el aprendizaje no supervisado, es el agrupamiento (en inglés, *clustering*). El objetivo del agrupamiento es dividir el conjunto de datos en grupos que compartan características similares. Un ejemplo es la segmentación de clientes, donde una empresa busca clasificar a sus compradores por sus características en común, con el fin de diseñar estrategias específicas para cada perfil de comprador (Alpaydin, 2014).

4.5. Algoritmos de regresión

El análisis de regresión consiste en un tipo de aprendizaje supervisado que se basa en la predicción de variables continuas, es decir, variables que pueden tomar cualquier valor dentro de un rango determinado. Algunos ejemplos son el precio justo de una acción, el comportamiento de la deuda estudiantil en los próximos años, así como el pronóstico de ganancias de un negocio. Se proporcionan las variables predictoras y una variable de respuesta continua; el objetivo es encontrar la relación entre dichas variables que permita predecir la respuesta correcta (Raschka et al., 2022; Watt et al., 2020).

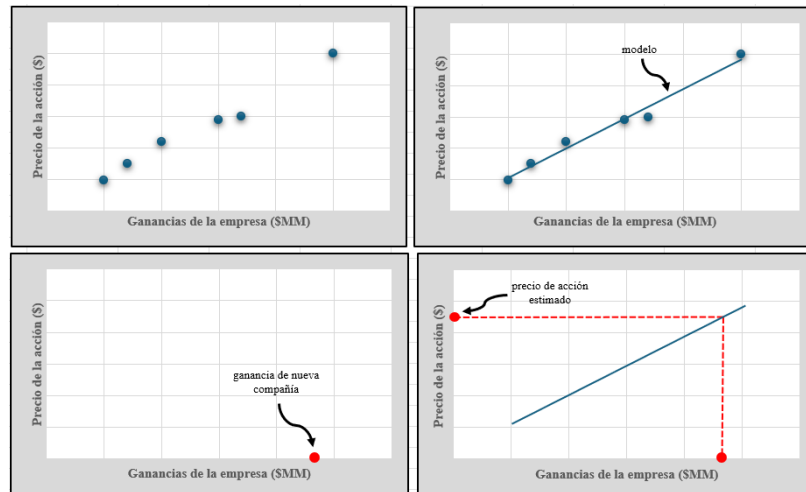


Figura 4.5: Ejemplo de modelo de regresión

Fuente: elaboración propia

Tomando el ejemplo proporcionado por Watt et al. (2020), la regresión se puede aplicar para determinar el precio justo de la acción de una compañía que está a punto de salir a bolsa. Lo primero que se debe hacer es recopilar el conjunto de datos de entrenamiento con compañías cuyo precio de acción sea conocido. Después, se deben seleccionar las características que sean relevantes

para la variable objetivo, en este caso, y por simplicidad, se tomará únicamente la ganancia neta de la empresa. Ahora, para hallar la dependencia que existe entre la ganancia de la empresa y el precio de su acción, se procede a entrenar un modelo de regresión con los datos de entrenamiento. Este proceso se muestra en la parte superior de la Figura 4.5 para un conjunto hipotético de datos. Una vez el modelo ha sido entrenado, el precio justo de la acción de una nueva empresa puede ser determinado con facilidad ingresando su ganancia neta como variable de entrada del modelo, tal como se observa en la parte inferior de la Figura 4.5.

4.5.1. Evaluación de algoritmos de regresión

Existen múltiples métricas que se pueden utilizar para evaluar el rendimiento de algoritmos de regresión. Cada métrica evalúa el modelo de una forma particular, por lo que es importante conocer el funcionamiento interno de cada una y así saber qué información proporcionan. Para los fines que competen el presente trabajo, se presentan las siguientes dos métricas:

- *Media del Error Absoluto (MAE)*
- *Mediana del Error Absoluto (MdAE)*

El MAE, como su nombre indica, corresponde al valor esperado del error absoluto, y se calcula como se indica en la ecuación 4.5. El MdAE, por su parte, se calcula como la mediana del conjunto de errores absolutos entre los valores reales y los predichos por el modelo; su cálculo se muestra en la ecuación 4.6.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.5)$$

$$MdAE = \text{mediana}(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \dots, |y_n - \hat{y}_n|) \quad (4.6)$$

La media y la mediana son ambas medidas de tendencia central. La media es una medida de magnitud que indica la ubicación central de los datos; la mediana, por su parte, es una medida de posición que define el valor ubicado en la posición de en medio cuando los datos están ordenados de forma ascendente (Anderson et al., 2012).

Se observa que para el cálculo del MAE y MdAE se aplica valor absoluto a las diferencias entre el valor real y el valor predicho. De no realizar esto, habría tanto errores positivos como negativos que tenderían a anularse mutuamente para el cálculo de la media; y si bien el cálculo de la mediana se ubicaría en el medio de la distribución, ésta no proporcionaría valor estadístico sobre la calidad de las mediciones. El cálculo del error absoluto permite que ambas métricas representen adecuadamente la distribución de los errores (Marroquín, 2023). Por ejemplo, para los errores $(-0.5, 1, -0.25, 0.45, 0.55)$ la media es 0.25 y la mediana 0.45, sin embargo, el signo del error no es relevante en términos de cuantificar el rendimiento del modelo, ya que éste simplemente indica si la predicción fue superior o inferior al valor real; al aplicar valor absoluto y repetir el cálculo se tiene una media de 0.55 y una mediana de 0.50

Para distribuciones con pocos valores, el MdAE presenta la ventaja de ser bastante robusto frente a valores atípicos, a diferencia del MAE (Bonnin, 2017). Para el caso de la media, el error introducido por el dato atípico se distribuye de forma homogénea entre todas las observaciones debido al cálculo del promedio, por lo que si el número de predicciones es pequeño el impacto será grande, mientras que si hay muchas predicciones el impacto será menor. En el caso de la mediana, esta simplemente se desplaza hacia el siguiente error más grande de la distribución, por lo que el impacto del valor atípico

en esta métrica depende más de la distribución y no del número total de observaciones. Siguiendo con el ejemplo del párrafo anterior, observe que al añadir un valor atípico a la distribución de errores, por ejemplo 3.85, y al tratarse de una distribución con pocos errores, la media experimenta un gran cambio al duplicarse desde 0.55 a 1.10; la mediana por su parte pasa de ser 0.50 a 0.525.

4.6. Algoritmos de clasificación

Los métodos de clasificación son un tipo de aprendizaje supervisado en el que la variable que se busca predecir es de tipo categórica, es decir, variables cuyo valor está limitado a un número finito de valores discretos (también llamados clases o categorías). Un modelo que cumple con esta característica recibe el nombre de *clasificador*, y sus aplicaciones incluyen la detección de rostros humanos en fotografías, el análisis del sentimiento de un cliente al comprar un producto, el diagnóstico de enfermedades a partir del historial del paciente, entre muchos otros (Watt et al., 2020).

Un clasificador puede ser de tipo *binario* o *multiclase*. Un clasificador binario identifica la variable de respuesta como perteneciente a una de dos posibles categorías; por ejemplo, la clasificación de un correo como spam o no spam, o el pronóstico meteorológico de si lloverá o no. Por el contrario, en un clasificador multiclase, el número de categorías entre las que puede ser clasificada la variable es superior a dos; un ejemplo de esto es la clasificación de un correo como de carácter personal, académico, recreativo o spam, o distinguir la marca de un carro en base a sus características físicas (Rebala et al., 2019).

Una forma común de entender la clasificación, y que utilizan muchos modelos de este tipo, es en términos de la optimización de parámetros para hallar la frontera de decisión que mejor separe las clases de datos presentes en el conjunto de entrenamiento. Esta frontera de decisión, también llamada hiperplano, es una línea si el vector de entrada del modelo tiene dos dimensiones, un plano si tiene tres, o, en el caso más general, una superficie de dimensión $n - 1$ donde n es la dimensión del vector de entrada (Watt et al., 2020).

Observe como ejemplo la clasificación de la Figura 4.6, tomada del libro de Watt et al. (2020). Se utilizan dos características como predictores, por lo que el vector de entrada tiene dos dimensiones y la frontera de decisión es simplemente una línea. Una vez entrenado el modelo, la categoría de una nueva observación se puede predecir fácilmente determinando de qué lado de la frontera se ubica la instancia.

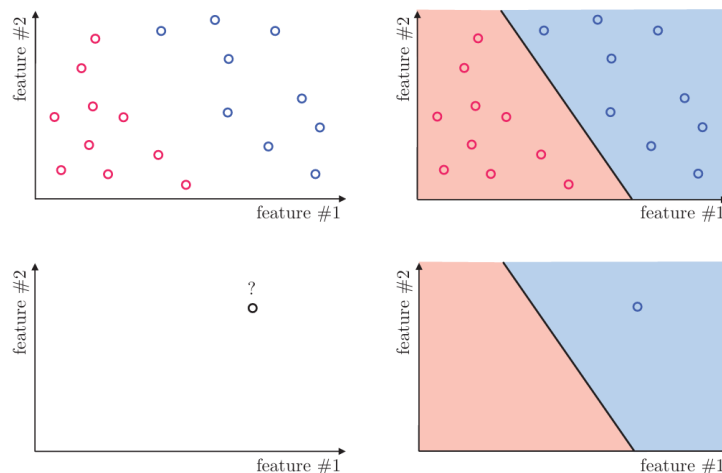


Figura 4.6: Ejemplo de modelo de clasificación

Fuente: Watt et al. (2020)

4.6.1. Evaluación de algoritmos de clasificación

Antes de definir las métricas de evaluación para algoritmos de clasificación, es importante introducir el concepto de *matriz de confusión*. La matriz de confusión, denotada \mathbf{M} , es un método que permite resumir los resultados de algoritmos de clasificación durante la fase de evaluación. El elemento de la i – esima fila y k – esima columna de la matriz de confusión (M_{ik}) corresponde al número de instancias pertenecientes a la categoría i , pero que el modelo clasificó como pertenecientes a la categoría k . A partir de esta matriz, se definen 4 cantidades importantes que se muestran en el Cuadro 4.1. Nótese que para un clasificador multiclase, estas cantidades presentan un valor diferente para cada una de las posibles categorías de la variable objetivo; además, para calcular el Verdadero Negativo (VN) se hace uso de las otras tres cantidades. El número total de predicciones realizadas por el modelo se denota con la letra n (Kroese et al., 2023).

Cantidad	No. de instancias que fueron...	Cálculo categoría j
Verdadero Positivo (VP)	correctamente clasificadas como pertenecientes	M_{jj}
Falso Positivo (FP)	incorrectamente clasificadas como pertenecientes	$\sum_{k \neq j} M_{kj}$
Falso Negativo (FN)	incorrectamente clasificadas como no pertenecientes	$\sum_{k \neq j} M_{jk}$
Verdadero Negativo (VN)	correctamente clasificadas como no pertenecientes	$n - VP_j - FP_j - FN_j$

Cuadro 4.1: Cantidades importantes para evaluación de algoritmos de clasificación

Estas cantidades permiten definir las métricas de evaluación más importantes para los algoritmos de clasificación, cuya descripción y cálculo, de acuerdo con Kroese et al. (2023), se muestra en el Cuadro 4.2. Notar que la exactitud no es distinta para cada categoría, sino que es una métrica general del modelo, mientras que las demás métricas sí difieren en función de la categoría que esté bajo análisis.

Métrica	Proporción de instancias...	Cálculo
Exactitud	predichas correctamente	$\frac{\sum_j M_{jj}}{n}$
Precisión	clasificadas como categoría j y que sí pertenecen a j	$\frac{VP_j}{VP_j + FP_j}$
Sensibilidad	de la categoría j y que fueron clasificadas como tal	$\frac{VP_j}{VP_j + FN_j}$
Especificidad	no pertenecientes a la categoría j y que fueron clasificadas como tal	$\frac{VN_j}{FP_j + VN_j}$

Cuadro 4.2: Métricas de evaluación de algoritmos de clasificación

Para ejemplificar lo anterior, considere un sistema automatizado de identificación facial para permitir o denegar el acceso de una persona a archivos militares secretos. Este es un ejemplo de clasificación binaria; suponga que la matriz de confusión del clasificador es la que se muestra en el Cuadro 4.3. En este caso, y considerando la autorización como resultado positivo y la no autorización como negativo, se tienen las cantidades:

$$n = 100550; VP = 100; FP = 50; FN = 400; VN = 100000 \quad (4.7)$$

De esta manera, se calcula una exactitud del 99.55%, precisión del 33.33%, sensibilidad del 20.00% y una especificidad del 99.95%

Real	Predicho	
	autorizado	no autorizado
autorizado	100	400
no autorizado	50	100000

Cuadro 4.3: Ejemplo de matriz de confusión

La elección de la métrica a utilizar depende de la aplicación del modelo de clasificación (Kroese et al., 2023). Se observa que el ejemplo anterior presenta una exactitud bastante alta; sin embargo, el sistema dio acceso a 50 personas no autorizadas, lo que es inadmisibile para un sistema de protección de archivos secretos. En este caso, probablemente sea más inteligente evaluar el rendimiento según su precisión, y no tanto por su exactitud.

También existen casos en donde lo más importante es la sensibilidad. Considere un modelo encargado de indicar si una persona tiene cáncer o no; el caso más crítico en esta situación es el de personas que tienen la enfermedad pero que no son diagnosticadas correctamente, debido a que estas personas no recibirán tratamiento y la enfermedad seguirá avanzando. En estos casos, es recomendable utilizar un modelo que clasifique correctamente la mayor proporción de casos positivos de cáncer, por lo que sería correcto utilizar un modelo con alta sensibilidad.

4.7. Algoritmos de machine learning

4.7.1. Regresión lineal

El método de regresión lineal, como su nombre indica, asume que la relación entre las variables de entrada y la variable de salida es lineal. La variable objetivo se expresa como una combinación lineal de los parámetros de entrada, en la que los coeficientes reciben el nombre de *pesos*, ya que indican la influencia que tiene cada parámetro en modificar el valor de la variable objetivo. La regresión lineal puede ser *simple*, si únicamente existe una variable de entrada, o *múltiple* si el número de predictores es mayor que uno (Brownlee, 2016).

En la ecuación 4.8 se muestra el modelo general de una regresión lineal múltiple, donde la predicción y_p se obtiene a partir de n variables de entrada. Se agrega un peso adicional w_0 para proporcionar al modelo un grado de libertad más (moverse hacia arriba o hacia abajo en un plano bidimensional) que recibe el nombre de *intercepto* (Brownlee, 2016). La ecuación 4.8 se debe cumplir para cada una de las m instancias del conjunto de entrenamiento; esto se resume en el sistema matricial de la ecuación 4.9, donde:

- \mathbf{X} es la matriz de dimensión $(m, n + 1)$, una fila por cada instancia, y una columna por cada variable de entrada. La primer columna contiene valores de 1 para no alterar el intercepto.
- $\vec{\mathbf{W}}$ es el vector columna de $n + 1$ filas que contiene los pesos desde w_0 hasta w_n .
- $\vec{\mathbf{Y}}_p$ es el vector columna de m filas; contiene las predicciones para cada una de las m instancias.

$$y_p = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n \quad (4.8)$$

$$\vec{\mathbf{Y}}_p = \mathbf{X}\vec{\mathbf{W}} \quad (4.9)$$

El proceso de entrenamiento tiene por objetivo determinar el valor ideal de los pesos w 's que minimize el error introducido por el modelo al realizar predicciones. Uno de los métodos más comunes para encontrar el valor de los pesos se conoce como *mínimos cuadrados*. Este consiste en minimizar el cuadrado de la diferencia entre el valor real y el valor predicho por el modelo. Este valor se calcula para cada instancia y se suman los resultados. El proceso de elevar al cuadrado se realiza para evitar que diferencias negativas y positivas se cancelen mutuamente (Rebala et al., 2019). El error total, en términos matriciales, se muestra en la ecuación 4.10, donde $\vec{\mathbf{Y}}$ contiene los valores reales y $\mathbf{X}\vec{\mathbf{W}}$ corresponde a las predicciones. Nótese que si bien el cálculo involucra matrices, el resultado es un escalar.

$$J(\vec{\mathbf{W}}) = (\vec{\mathbf{Y}} - \mathbf{X}\vec{\mathbf{W}})^T (\vec{\mathbf{Y}} - \mathbf{X}\vec{\mathbf{W}}) \quad (4.10)$$

Para minimizar esta cantidad, se hace uso de cálculo diferencial matricial para igualar a cero la derivada $J'(\vec{\mathbf{W}})$, con lo que, mediante operaciones matriciales, se obtiene que el vector $\vec{\mathbf{W}}$ que minimiza el error de mínimos cuadrados, y que por lo tanto mejor se ajusta al conjunto de entrenamiento, es el que se muestra en la ecuación 4.11

$$\vec{\mathbf{W}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{\mathbf{Y}} \quad (4.11)$$

La regresión lineal presenta la ventaja de ser sumamente fácil de interpretar y consumir pocos recursos computacionales; sin embargo, presenta la desventaja de ser poco útil para modelar relaciones complejas y no lineales (Harrington, 2012). Es importante considerar que el cálculo de la matriz inversa de $\mathbf{X}^T \mathbf{X}$ en la ecuación 4.11 puede tornarse complejo y consumir bastantes recursos, especialmente para casos con muchas variables de entrada, o incluso la inversa podría simplemente no existir. En estos casos se debe recurrir al método de *descenso del gradiente* (Rebala et al., 2019).

4.7.2. Regresión polinomial

En ciertos casos modelar una relación como una regresión lineal múltiple no es suficiente, ya que puede darse el caso de que la salida esté influenciada por el cuadrado de una variable de entrada, o que el producto de dos entradas sea más relevante que sus valores separados en términos de predecir la variable de interés. En estos casos se introduce una generalización de la regresión lineal múltiple en la que se aumenta el espacio de características mediante la inclusión de combinaciones (potencias) de las variables de entrada. Si se consideran las potencias introducidas simplemente como otras variables, que al igual que las variables originales tienen un peso asociado en la combinación lineal, entonces este caso se reduce a una regresión lineal múltiple, por lo que la deducción realizada en la subsección anterior es completamente válida para describir el entrenamiento de un modelo de regresión polinomial (Rebala et al., 2019).

Para ejemplificar lo anterior, considere que se toman las entradas x_1 y x_2 para predecir una variable de salida y . Ahora bien, por alguna razón, se presume que el valor de y depende en gran medida de potencias de grado 2 de las variables x_1 y x_2 , por lo que se puede plantear la relación como se muestra en la ecuación 4.12.

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 \quad (4.12)$$

A primera instancia, este podría verse como un caso completamente diferente a lo visto previamente; no obstante, realizando las sustituciones $x_3 = x_1 x_2$, $x_4 = x_1^2$ y $x_5 = x_2^2$, la expresión se transforma en la ecuación 4.13, que claramente es una regresión lineal múltiple. El entrenamiento

consiste en calcular las nuevas variables para cada instancia de la base de datos y utilizar la ecuación 4.11 para encontrar los pesos óptimos.

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 \quad (4.13)$$

La regresión polinomial presenta una mayor capacidad de modelar relaciones no lineales, al considerar la influencia de nuevas variables construidas a partir de potencias de las variables originales. Es importante mencionar que el número de pesos incrementa drásticamente a medida que aumenta el grado del polinomio, lo que puede ocasionar un alto consumo de recursos dependiendo de las herramientas que se utilicen (Rebala et al., 2019).

4.7.3. Regresión logística

La regresión logística es uno de los algoritmos de clasificación más utilizados y sumamente fácil de implementar en situaciones donde las clases son linealmente separables, es decir, clases que se pueden dividir de forma sencilla mediante una frontera de decisión. Es un algoritmo probabilístico, en el sentido de que, además de la clase predicha, también se proporciona la probabilidad de que el pronóstico esté correcto. Se puede aplicar tanto a clasificación binaria como multiclase (Raschka et al., 2022).

Antes de definir el funcionamiento del algoritmo, es importante introducir el concepto de *odds*. Los odds se pueden escribir como $\frac{p}{1-p}$, donde p es la probabilidad de clase positiva. Por ejemplo, en un modelo de detección de una enfermedad, p sería la probabilidad de que el paciente sí esté enfermo. La función *logit* se define como el logaritmo natural de los odds. Esta función mapea valores en el intervalo $(0, 1)$ a todo el rango de números reales. En la regresión logística, se asume que existe una relación lineal entre la función logit y las entradas del modelo; dicha relación se representa en la ecuación 4.14

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = w_0 + w_1x_1 + \dots + w_nx_n \quad (4.14)$$

La inversa de la función logit, llamada *función logística* o *sigmoide*, define la probabilidad de clase positiva en términos de las variables de entradas. Esta función tiene forma de letra S, y mapea el conjunto de los números reales en el intervalo $(0, 1)$; su forma se presenta en la ecuación 4.15 donde $z = w_0 + w_1x_1 + \dots + w_nx_n$.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (4.15)$$

Nótese que esta función tiende a 1 cuando $z \rightarrow \infty$, y tiende a 0 cuando $z \rightarrow -\infty$. La función logística se utiliza como función de activación: si $\sigma(z) < 0.5$ se predice la clase negativa, mientras que si $\sigma(z) > 0.5$ se predice la clase positiva. Si $\sigma(z) = 0.5$, se entiende que ambas clases presentan la misma probabilidad.

Sabiendo que la variable de respuesta sigue una distribución de Bernoulli, se puede afirmar que la probabilidad de que la predicción de la i -ésima instancia sea correcta está dada por la ecuación 4.16, donde $y^{(i)}$ es la clase real con valor 0 o 1. Nótese que esta probabilidad se reduce a $\sigma(z^{(i)})$ si $y^{(i)} = 1$, y a $1 - \sigma(z^{(i)})$ si $y^{(i)} = 0$.

$$\text{Probabilidad de predicción correcta} = (\sigma(z^{(i)}))^{y^{(i)}} (1 - \sigma(z^{(i)}))^{1-y^{(i)}} \quad (4.16)$$

Ahora bien, asumiendo que todas las predicciones son independientes, se puede construir la función *likelihood* como la probabilidad de que todas las predicciones sean correctas. Esta función se calcula simplemente como el producto de las probabilidades de cada instancia, y su forma se muestra en la ecuación 4.17. El proceso de entrenamiento del modelo de regresión logística consiste en maximizar esta función mediante el método de ascenso del gradiente, logrando así determinar los pesos w 's que definen la mejor frontera de decisión entre las clases (Raschka et al., 2022).

$$\mathcal{L} = \prod_{i=1}^n (\sigma(z^{(i)})^{y^{(i)}} (1 - \sigma(z^{(i)}))^{1-y^{(i)}}) \quad (4.17)$$

La regresión logística presenta la ventaja de ser sencilla de implementar y eficiente en cuanto a consumo de recursos computacionales, además de ser bastante fácil de interpretar: la predicción está ligada al valor de la función sigmoide. En adición, se presenta no solo la clase predicha, sino también la probabilidad de que el pronóstico sea correcto. Sin embargo, en ocasiones puede ser propensa a sobreajuste, es decir, que el modelo se alinee bastante bien con los datos de entrenamiento pero que no funcione adecuadamente para realizar pronósticos de instancias nuevas. Además, la regresión logística funciona bajo la premisa de que las clases son (al menos en buena medida) linealmente separables, por lo que si las clases están superpuestas en el espacio de características, el rendimiento del modelo podría verse afectado (Harrington, 2012).

4.7.4. Bosque aleatorio

Antes de definir el algoritmo de bosque aleatorio, es importante introducir el concepto de *árbol de decisión*. Un árbol de decisión es un algoritmo muy útil que se puede utilizar tanto para tareas de clasificación como de regresión. El árbol cuenta con nodos que representan decisiones deterministas basadas en las variables de entrada. Las ramas que parten de un nodo representan los diferentes caminos que se pueden seguir en función de la decisión tomada. Al final de cada rama se puede encontrar un nuevo nodo con otra decisión a tomar, o una hoja que contiene el valor de la predicción del modelo (Rebala et al., 2019). En la Figura 4.7 se muestra un ejemplo de árbol de decisión para determinar si una persona es apta, o no, para un préstamo.

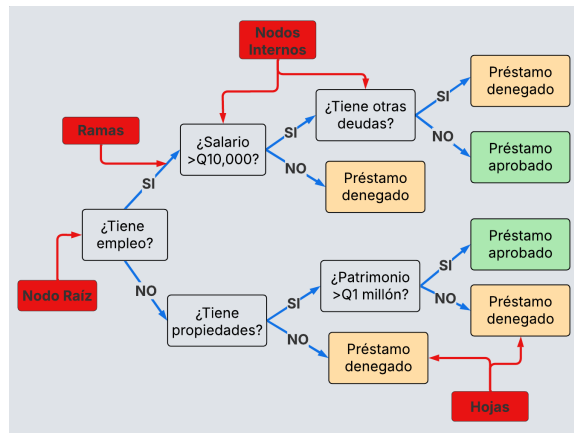


Figura 4.7: Ejemplo de árbol de decisión

Fuente: elaboración propia

Para construir el árbol durante la fase de entrenamiento se utiliza el algoritmo CART (del inglés *Classification and Regression Tree*). Este comienza en el nodo inicial, y mediante la optimización

de una función de costo, se determina la mejor partición del conjunto de características, con lo que se tiene el primer nodo del árbol. A continuación, para cada una de las ramas que se desprenden del nodo recién construido, se repite el proceso para construir la siguiente capa de nodos. Este proceso se repite constantemente hasta completar el árbol. Para evitar el crecimiento deliberado del árbol, se recomienda indicar un *criterio de parada*, el cual suele ser el número mínimo de instancias del conjunto de entrenamiento que debe haber en cada subdivisión del espacio de características. La función de costo utilizada para determinar la mejor partición se calcula como la suma de las ponderaciones relativas de una métrica a ambos lados de la partición. La métrica utilizada es escogida por el usuario; algunas de las más conocidas son el índice de Gini o la entropía para casos de clasificación, así como el error cuadrático medio para casos de regresión (Brownlee, 2016).

El algoritmo de árbol de decisión es bastante intuitivo, en el sentido de que es fácil conocer la lógica utilizada para obtener la predicción. No obstante, presenta la desventaja de ser susceptible al sobreajuste, pues leves cambios en el conjunto de entrenamiento pueden llevar a un árbol completamente distinto. El bosque aleatorio resuelve este problema, al combinar la predicción de varios árboles de decisión para producir un resultado más exacto y libre de sesgo (Kroese et al., 2023).

El bosque aleatorio es un algoritmo de *ensemble* que hace uso de muchos árboles de decisión. Se utiliza el concepto de *bagging* (también conocido como *bootstrap aggregation*) para generar diferentes subconjuntos de entrenamiento a partir del conjunto original; cada subconjunto se obtiene mediante *muestreo por reemplazo*, lo que significa que algunas instancias de la base de datos pueden repetirse, mientras que otras pueden quedar fuera del subconjunto. Cada árbol de decisión es entrenado con un subconjunto diferente. La predicción final se realiza por agregación de las predicciones individuales de cada árbol: se toma la clase que más se repite para el caso de clasificación, o el promedio de las predicciones para regresión (Rebala et al., 2019).

El bosque aleatorio presenta múltiples ventajas, como el hecho de que el tiempo necesario para entrenar el modelo es menor en comparación con otros algoritmos, lo que lo hace sumamente útil para casos donde el conjunto de entrenamiento es muy grande. Además, suele generar predicciones bastante exactas, debido a que reduce la tendencia del árbol de decisión al sobreajuste al combinar la predicción de múltiples árboles (Doshi et al., 2022).

4.7.5. XGBoost

El algoritmo XGBoost (del inglés *extreme gradient boosting*) es muy similar al bosque aleatorio, pues ambos hacen uso de múltiples árboles de decisión para generar una predicción más exacta. Sin embargo, mientras que el bosque aleatorio utiliza el *bagging* para entrenar cada árbol con un subconjunto diferente y luego combinar la predicción de cada uno para producir la respuesta final, XGBoost utiliza un concepto diferente conocido como *impulso del gradiente* (en inglés, *gradient boosting*).

El impulso del gradiente consiste en la construcción secuencial de múltiples modelos, donde cada modelo intenta corregir los errores producidos por los modelos anteriores. La idea es empezar con un modelo simple g_0 generado a partir del conjunto de entrenamiento, y convertirlo en un modelo mejorado $g_1 = g_0 + h_1$, donde h_1 es el modelo que surge a partir del entrenamiento de los datos de entrada, pero tomando como variable objetivo el residuo del modelo g_0 . Así, h_1 predice el error que tendrá g_0 ; al tomar en cuenta este error, el nuevo modelo g_1 corregirá el error del modelo inicial. Este proceso se realiza varias veces, donde al nuevo modelo siempre se le agrega una corrección que corresponde al error del último modelo. El modelo final consiste en la agregación de todas las correcciones al modelo original, y se muestra en la ecuación 4.18 (Kroese et al., 2023).

$$g(\vec{x}) = g_0(\vec{x}) + \sum_{b=1}^B h_b(\vec{x}) \tag{4.18}$$

El algoritmo XGBoost emplea el método de impulso del gradiente para construir de forma iterativa un modelo que implemente árboles de decisión. Un árbol es capaz de producir buenas predicciones únicamente en cierto rango de los valores de entrada; mediante la corrección en serie de los errores de cada árbol, XGBoost es capaz de generar un modelo que se acerque más a la realidad. Por lo general, los árboles utilizados son bastante simples, con una profundidad en el rango de 1 a 5 (recordar que la profundidad es el número de ramas que hay entre el nodo inicial y la hoja más lejana), lo que dota al algoritmo de una gran eficiencia en términos de memoria de almacenamiento y le permite realizar predicciones de forma más rápida (Müller y Guido, 2017).

Los algoritmos que emplean árboles de decisión potenciados con impulso del gradiente, como XGBoost, son a menudo los algoritmos ganadores en el ámbito del machine learning, y son ampliamente utilizados por la industria para un sinnúmero de aplicaciones. Si bien suelen ser más sensibles a los parámetros de entrada, proporcionan una mayor exactitud si estos son escogidos adecuadamente (Müller y Guido, 2017).

El Departamento de Investigación y Servicios Meteorológicos de INSIVUMEH cuenta con un total de 64 estaciones convencionales distribuidas a lo largo del territorio guatemalteco, cuya ubicación se muestra en la Figura 5.1. Cada estación cuenta con la instrumentación necesaria para el registro de diversas variables meteorológicas. Entre las variables que se miden están la precipitación, la temperatura máxima, media y mínima, la velocidad y dirección del viento, la humedad relativa, entre otras. Las variables son registradas diariamente por un observador climático en tres horarios diferentes: a las 7:00, 13:00 y 18:00 hrs. Posteriormente, los datos son enviados por medio de una aplicación móvil para su análisis y almacenamiento (Argueta, 2024).

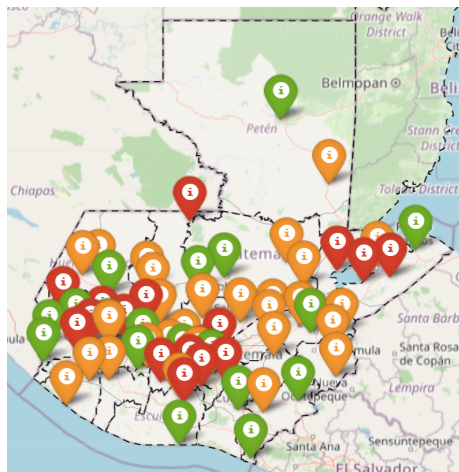


Figura 5.1: Mapa de estaciones convencionales de INSIVUMEH

Fuente: Argueta (2024)

INSIVUMEH colabora con la iniciativa ENACTS¹ del programa ACToday de la Universidad de Columbia (Turner, 2021). ENACTS es una iniciativa que mejora la disponibilidad, acceso y uso de información climática mediante la generación de bases de datos históricos de alta resolución (Dinku et al., 2017). Mediante este programa, INSIVUMEH ha sido capaz de obtener un registro histórico de

¹*Enhancing National Climate Services*

precipitación, así como de temperatura máxima, media y mínima en todos los puntos del territorio nacional donde no se cuenta con una estación meteorológica. Los datos se obtienen mediante el ajuste de información satelital con los datos de la red de estaciones meteorológicas de INSIVUMEH (INSIVUMEH, 2023). Para la generación de datos de precipitación se utilizan los datos satelitales CHIRPS, administrados por el Climate Hazards Group de la UCSB²; mientras que para los valores de temperatura se hace uso de los datos MERRA2 proporcionados por la NASA. Mediante este método, se ha logrado construir una base de datos históricos con valores diarios y mensuales para todo el dominio geográfico de Guatemala con una resolución de 0.05°.

En la actualidad, los pronósticos meteorológicos de INSIVUMEH son realizados utilizando el modelo numérico WRF; en concreto, se hace uso del esquema Betts-Miller-Janjic (BMJ), el cual representa los mecanismos de convección de la atmósfera mediante perfiles termodinámicos de equilibrio cuasi-estático determinados empíricamente (Vaidya y Singh, 1999). Como datos de entrada del modelo, INSIVUMEH utiliza dos fuentes: los datos proporcionados por el modelo global GFS (del inglés *Global Forecast System*) y los del Modelo de Mesoescala de América del Norte (NAM). Los primeros generan pronósticos con un horizonte temporal máximo de 15 días, mientras que los segundos son aptos para pronósticos que no superen los 3 días. El modelo proporciona un archivo de datos *raster* en formato *GeoTIFF*, que contiene el valor predicho para cada punto del dominio geográfico con una resolución de 8 km; en la Figura 5.2 se observa un ejemplo de representación gráfica de un pronóstico de WRF proporcionado por Ixcamparij (2025). Mediante este método, INSIVUMEH realiza pronósticos diarios de precipitación, velocidad y dirección del viento, así como de temperatura media y otras variables meteorológicas.

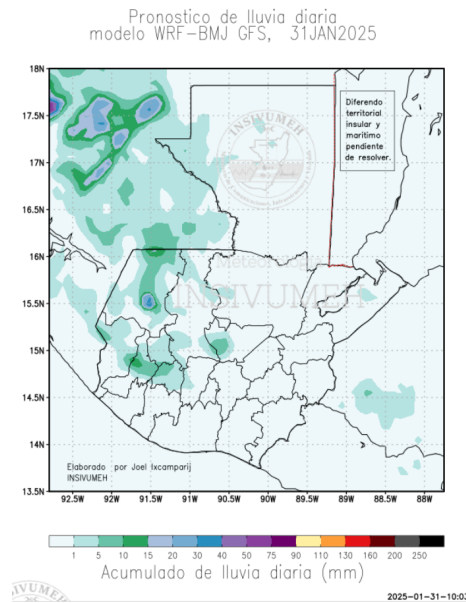


Figura 5.2: Representación gráfica de los archivos de datos *raster* proporcionados por el modelo WRF-BMJ GFS

Fuente: Ixcamparij (2025)

NextGen es un modelo de predicción meteorológica y climatológica capaz de producir pronósticos en intervalos que van desde semanas hasta años. Fue desarrollado y publicado en 2019 por el Instituto Internacional de Investigación para el Clima y la Sociedad de la Universidad de Columbia (IRI, por sus siglas en inglés), tras más de 25 años de investigación. Este modelo se basa en el uso de múltiples Modelos Generales de Circulación (GCM), los cuales hacen uso de las ecuaciones de Navier-

² *University of California, Santa Barbara*

Stokes para simular las corrientes de circulación de la atmósfera. NextGen utiliza un análisis de correlación canónica para calibrar los pronósticos de los diferentes GCM's; posteriormente, los GCM's se combinan, cada uno con diferente peso, para dar lugar a un *ensemble* de modelos, el cual se utiliza para obtener los pronósticos finales (Acharya et al., 2021; IRI, 2020). Actualmente, INSIVUMEH hace uso del modelo NextGen para la obtención de pronósticos estacionales; las variables que se predicen son la precipitación acumulada, así como la temperatura máxima, media y mínima a lo largo de cada período estacional (INSIVUMEH, 2025).

En los últimos años, el machine learning, como parte de la inteligencia artificial, se ha empezado a utilizar en el pronóstico de variables meteorológicas. Algunos de los modelos más importantes y ya consolidados son:

- *GraphCast*: Desarrollado por colaboradores de Google Deepmind, es un modelo basado en machine learning y redes neuronales gráficas (GNN) capaz de predecir cientos de variables meteorológicas a lo largo de un horizonte temporal de 10 días con una resolución de 0.25° en menos de un minuto. Se toman como parámetros de entrada el estado del tiempo hace 6 horas y el estado actual, para predecir el estado en las próximas 6 horas; este proceso se puede realizar de forma iterativa para realizar pronósticos en intervalos de 6 horas hasta completar el máximo de 10 días. En cada ubicación geográfica, se predicen cinco variables al nivel del suelo, como la precipitación total y la temperatura, así como seis variables atmosféricas en cada una de las 37 altitudes definidas. (Lam et al., 2023)
- *FourCastNet*: Liderado por Nvidia, es un modelo global de pronóstico meteorológico basado en datos, que hace uso del deep learning y las redes neuronales para predecir variables como la precipitación y la velocidad del viento. Posee una resolución de 0.25° y realiza pronósticos de una semana en menos de dos segundos. Ha demostrado tener mayor exactitud que modelos numéricos, como el Sistema de Pronóstico Integrado (IFS) del Centro Europeo de Previsiones Meteorológicas a Plazo Medio (ECMWF) para variables de pequeña escala, incluyendo la precipitación. (Pathak et al., 2022)
- *ClimaX*: Promovido por Microsoft y la UCLA, es un modelo de deep learning basado en redes neuronales profundas (DNN). Se realiza una fase de preentrenamiento donde se utiliza aprendizaje supervisado, para después implementar una fase de entrenamiento final en la que se pueden implementar variables no vistas en el preentrenamiento, y la cual depende del tipo de pronóstico deseado. Esto dota al modelo de mayor flexibilidad y lo generaliza para poder ser aplicado en una variedad de contextos. (Nguyen et al., 2023)

Estos modelos han sido patrocinados por grandes empresas, y se concentran principalmente en el uso del deep learning y las redes neuronales para identificar, generalizar y predecir patrones de comportamiento atmosférico. Los parámetros de entrada de estos modelos suelen ser mallas bidimensionales de datos.

Hacia comienzos de 2020, científicos de Corea del Sur desarrollaron un modelo de machine learning basado en XGBoost para corregir los pronósticos de precipitación proporcionados por la Administración Meteorológica Coreana (KMA) y desarrollar así un modelo de pronóstico hidrológico. Se demostró que, luego de implementar XGBoost, los pronósticos proporcionaron un mejor rendimiento en términos de media del error absoluto (MAE), lo que en palabras de los autores indica claramente que la técnica de machine learning es capaz de mejorar el rendimiento de las previsiones de precipitación (Ko, C-M et al., 2020).

Posteriormente, en abril de 2023, Elbasheer et al. (2023) utilizaron diferentes modelos de machine learning para intentar mejorar los pronósticos estacionales y subestacionales del ECMWF, para lo cual realizaron una clasificación de la lluvia en función del registro histórico (abajo de lo normal, cerca de lo normal y arriba de lo normal). Se utilizaron como predictores la precipitación total y la temperatura a 2 m del suelo; la métrica de rendimiento escogida fue la RPSS (del inglés *ranked*

probability skill score). Se encontró que la regresión logística y las redes neuronales con perceptrón multicapa (ANN-MLP) superaron el rendimiento del ECMWF.

En abril de 2024, científicos desarrollaron un sistema especializado en el reconocimiento de condiciones favorables para la formación de eventos de precipitación extrema en el norte y centro de Italia, conocido como MalCoX, el cual se basa en una secuencia de bosques aleatorios para clasificar los eventos según su intensidad (Grazzini et al., 2024). Por su parte, Nunno et al. (2022) emplearon una extensión del árbol de decisión tradicional, conocido como M5P, así como el modelo de vectores de soporte de regresión para predecir la precipitación en dos regiones al norte de Bangladés; incluso probaron un modelo híbrido, el cual resultó ser mejor, con coeficientes R^2 de 0.87 y 0.92 para cada región, respectivamente.

En este capítulo se desarrolla el planteamiento teórico, así como el procedimiento para desarrollar y evaluar los distintos modelos que se estudiarán en esta investigación.

6.1. Especificaciones de *hardware* y *software*

Para el desarrollo de la presente investigación, se utilizó un computador portátil marca Lenovo con las siguientes características:

- Sistema operativo: Windows 10 de 64 bits
- Procesador: Intel Core i7-10750H, de 6 núcleos y 12 hilos
- Memoria RAM: 16 GB
- Almacenamiento: 512 GB SSD
- Tarjeta gráfica: Nvidia GeForce GTX 1650 (dedicada)
 - VRAM: 4GB GDDR5
 - Arquitectura: Turing
 - Núcleos: 896 CUDA
 - Frecuencia base: 1485 MHz
 - Frecuencia boost: 1665 MHz

El código fue programado en lenguaje *Python* a través de la interfaz *Jupyter Notebook*. Para acceder a dicha interfaz, se instaló la versión 25.1.0 de la plataforma *Anaconda*. Durante el desarrollo de los modelos, se hizo uso de las siguientes bibliotecas:

- pandas
- pykrige
- xgboost
- joblib
- numpy
- scikit-learn
- time

6.2. Definiciones

A continuación, se define una serie de conceptos fundamentales para la correcta comprensión de la metodología aplicada en la elaboración de la investigación. Para un mejor entendimiento de los primeros 3, puede referirse a la Figura 6.1.

- *Fecha objetivo*: Es la fecha cuya precipitación se desea pronosticar. Usualmente, se utilizará el día siguiente al día en que se realiza el análisis, de esta manera, se obtendrá un pronóstico antes que el día ocurra. Sin embargo no se limita al día posterior al análisis, ya que pueden realizarse pronósticos en retrospectiva para días anteriores. La fecha objetivo no puede ser posterior al día siguiente del análisis, ya que no se contaría con los datos necesarios para realizar un pronóstico de corto plazo.
- *Fecha de referencia*: Se denomina como fecha de referencia, el día anterior a la fecha objetivo. Esta fecha se utilizará como referencia para hacer alusión al pronóstico realizado.
- *Rango de observación*: Es el intervalo de tres, cinco o siete días consecutivos, inmediatamente anteriores a la fecha de referencia. Las variables meteorológicas en cada uno de estos días son tomadas como variables predictoras de los modelos de machine learning.
- *Climatología*: Corresponde al período de años 1991-2020. Se utiliza como intervalo de referencia para comparar la precipitación de la fecha objetivo con el comportamiento histórico registrado en dicha fecha.
- *Anomalía de precipitación*: Es la cantidad que surge de restar la precipitación esperada menos el promedio de la Climatología. Puede expresarse en términos absolutos (*mm*) o porcentuales (%). Permite identificar desviaciones con respecto al comportamiento habitual de la lluvia.
- *Época lluviosa*: Se define como la temporada del año comprendida entre **mayo** y **octubre**.
- *Época seca*: Intervalo estacional que abarca desde el mes de **noviembre** hasta **abril**.

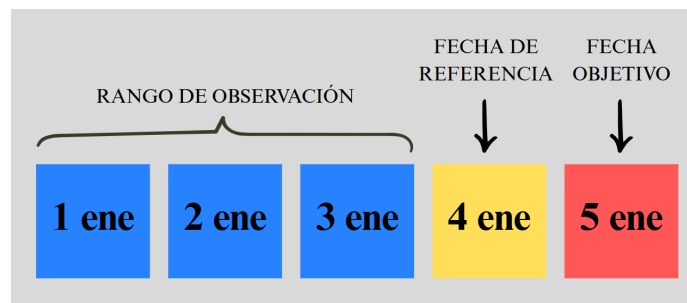


Figura 6.1: Esquema de la estructura temporal aplicada en el análisis predictivo

Las variables meteorológicas de la fecha de referencia no se incluyen como predictores del modelo. Esta decisión permite que los modelos puedan predecir la precipitación esperada para el día siguiente cuando la fecha de referencia coincide con el día en el que se realiza el pronóstico. En estos casos, las variables de la fecha de referencia son desconocidas al momento de realizar la predicción, por lo que es inviable incluirlas como variables de entrada del modelo.

6.3. Procedimiento experimental

Se desarrollan diversos modelos de machine learning que permiten realizar pronósticos de precipitación en Guatemala en la fecha objetivo. Los modelos son entrenados a partir de la base de datos meteorológicos proporcionada por el Departamento de Investigación y Servicios Meteorológicos de INSIVUMEH. Los modelos funcionan tomando como predictores los valores diarios de diferentes variables meteorológicas en el rango de observación, así como el número de mes (del 1 al 12) correspondiente a la fecha de referencia y las coordenadas geográficas (latitud y longitud) del punto donde se desea realizar el pronóstico. En la Figura 6.2 se representa este comportamiento, para un rango de observación de m días y n variables meteorológicas.

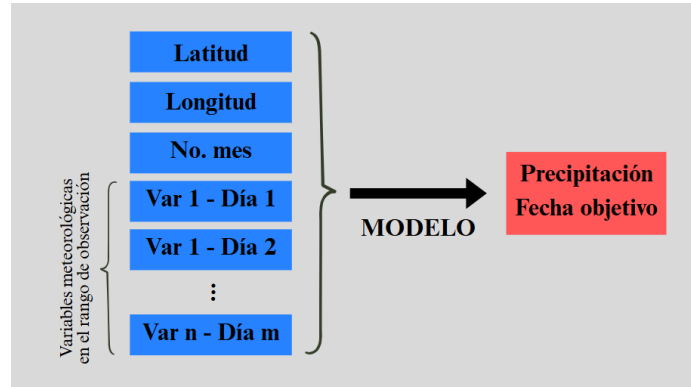


Figura 6.2: Representación gráfica del funcionamiento de los modelos de machine learning

Los modelos de machine learning desarrollados son guardados en archivos de formato *pkl*, los cuales almacenan los pesos y parámetros entrenados. De esta forma, se garantiza la permanencia en el tiempo y facilidad de uso de los modelos.

6.3.1. Variables objetivo

Con el objetivo de evaluar la utilidad del machine learning en el pronóstico meteorológico, se desarrollarán modelos para predecir tres variables objetivo: una de regresión y dos de clasificación. Cada uno de los múltiples modelos estará especializado en el pronóstico de una de las tres variables. Estas variables son:

1. Regresión (Variable Continua): Se pronostica la anomalía absoluta de precipitación del día siguiente respecto al promedio de la precipitación durante la Climatología.
2. Cotas Percentiles (Variable Categórica): Se pronostica el comportamiento cualitativo de la precipitación en comparación con el registro histórico. Para ello, se utilizan los percentiles del conjunto de precipitaciones de la Climatología. De acuerdo con el intervalo donde se encuentre la precipitación, así será su interpretación cualitativa. Ver Cuadro 6.1.
3. Lluvia Binaria (Variable Categórica): Se realiza una clasificación binaria de la precipitación del día siguiente como igual a $0mm$ (ausencia de lluvia), o mayor a $0mm$ (presencia de lluvia). En esencia, se trata de un clasificador que simplemente indica si lloverá o no, independientemente de la intensidad o del valor puntual de precipitación.

Intervalo de precipitación	Interpretación
$[0, P(1/3)]$	Abajo de lo normal (<i>BN</i>)
$(P(1/3), P(2/3))$	Dentro de lo normal (<i>NN</i>)
$[P(2/3), \infty)$	Arriba de lo normal (<i>AN</i>)

Cuadro 6.1: Interpretación de la variable objetivo para los modelos de clasificación con cotas percentiles

El promedio y los percentiles de la Climatología se calculan a partir del conjunto de 30 valores de precipitación diaria en el día y punto geográfico de interés, de cada año en el período 1991-2020. Por ejemplo, para la precipitación del 1 de enero en las coordenadas (*lat, lon*), se calcula el promedio y los percentiles del conjunto de 30 valores de precipitación del 1 de enero en dicha ubicación para cada año. La única excepción es el 29 de febrero, cuyos valores son determinados en base al conjunto de 8 valores de precipitación, correspondientes únicamente a los años bisiestos del período.

Es importante notar que los valores de la Climatología son diferentes para cada una de las coordenadas geográficas y para cada uno de los 365 días del año. Por lo tanto, un mismo valor de anomalía se corresponde con un valor de precipitación diferente en función de las coordenadas y día del año respectivo (la precipitación es la suma de la anomalía y el promedio de la Climatología). De igual forma, un mismo valor de precipitación puede ser clasificado como *BN*, *NN* o *AN*, dependiendo de los percentiles asociados.

6.3.2. Variación de parámetros

Al momento de implementar modelos de machine learning de pronóstico de precipitación, existe una diversidad de factores que pueden ser modificados. Sin embargo, conocer con anticipación los parámetros óptimos para potenciar el rendimiento del modelo puede llegar a ser complicado. Por esta razón, el trabajo propuesto incluirá la variación de diferentes parámetros al momento de desarrollar los modelos, con el objetivo de determinar los ajustes que proporcionan un mejor rendimiento. Es importante mencionar que se desarrolló un modelo para cada una de todas las posibles combinaciones de parámetros. Los parámetros a modificar son:

- Algoritmo de machine learning
- Número de árboles (para RF y XGB)
- Combinación de variables predictoras
- Número de días en el rango de observación
- Tipo de entrenamiento

Con respecto a los algoritmos de machine learning, en el Cuadro 6.2 se muestran los algoritmos a utilizar en función de la naturaleza de la variable objetivo (regresión o clasificación). En el caso de clasificación, se implementarán los algoritmos indicados para cada una de las dos variables objetivo mencionadas en la sección anterior.

Tipo de variable objetivo	Algoritmos a implementar
Regresión	Regresión lineal múltiple Regresión polinomial (grado 2) Bosque aleatorio XGBoost
Clasificación	Regresión logística Bosque aleatorio XGBoost

Cuadro 6.2: Algoritmos de machine learning a implementar

Es importante mencionar que para el caso de regresión polinomial se escogió únicamente trabajar con grado 2 debido a que para grados superiores la computadora utilizada presentó problemas de memoria (el número de pesos como función del grado sigue una relación polinómica no lineal). Además, en las pocas pruebas que sí se pudieron realizar para grado 3, se observó que, en general, el tamaño de los pesos es menor en los términos de mayor grado, lo que evidencia una disminución en la relevancia de los términos de orden superior.

Todos los algoritmos fueron desarrollados mediante el uso de la librería Scikit-Learn, con excepción de XGBoost, para el cual se utilizó la librería del mismo nombre. Así mismo, todos los algoritmos fueron implementados con sus hiperparámetros por defecto. El único hiperparámetro modificado fue el número de árboles para los modelos de bosque aleatorio y XGBoost. Los valores utilizados fueron 10, 20, 50 y 100 árboles de decisión.

Con respecto a los predictores del modelo, se implementaron diferentes combinaciones de variables meteorológicas. También se varió el número de días del rango de observación. Se desarrollaron tres combinaciones de variables: *básica*, *intermedia* y *avanzada*, cuya composición de variables se muestra en el Cuadro 6.3. Para la combinación *básica* se realizaron modelos con 3, 5 y 7 días de observación, con el objetivo de ver la evolución del rendimiento en términos de la cantidad de días. Por su parte, para las combinaciones *intermedia* y *avanzada* únicamente se utilizaron 5 días de observación, con el objetivo de comparar el rendimiento según la combinación, fijando el número de días en 5.

Variable meteorológica	Dimensional	Combinación		
		Básica	Intermedia	Avanzada
Precipitación	<i>mm</i>	✓	✓	✓
Temperatura media	°C	✓	✓	✓
Temperatura máxima	°C	✓	✓	✓
Temperatura mínima	°C	✓	✓	✓
Humedad relativa	%	✓	✓	✓
Velocidad del viento	<i>km/h</i>	✗	✓	✓
Dirección del viento ¹	°	✗	✓	✓
Presión atmosférica	<i>mmHg</i>	✗	✗	✓

Cuadro 6.3: Combinaciones de variables meteorológicas

El último parámetro que se varió fue el tipo de entrenamiento al que son sometidos los modelos. Se trabajó con entrenamiento general (a partir de datos de todo el año), entrenamiento con datos únicamente de la época lluviosa (período mayo-octubre) y entrenamiento con datos exclusivos de la época seca (período noviembre-abril). La separación de datos por épocas se realiza en base a la fecha de referencia. El entrenamiento selectivo se realiza con el objetivo de verificar si hay alguna mejora potencial en función del tipo de entrenamiento suministrado.

¹La dirección del viento se mide en grados sexagesimales, a partir del norte (0°) y en sentido horario. Hace referencia a la dirección desde donde sopla el viento.

6.3.3. Preprocesamiento de datos

La base de datos proporcionada por INSIVUMEH consiste en un archivo en formato *xlsx* que contiene registros diarios de diferentes variables para un total de 47 estaciones meteorológicas distribuidas a lo largo del territorio nacional. La base de datos cuenta con una columna para el nombre de la estación, otra para la fecha, y una columna para cada una de las variables meteorológicas. Los registros comienzan a partir del 1 de enero de 1980, y los valores más recientes corresponden al 31 de mayo de 2023.

La base de datos presenta múltiples entradas vacías, incluso existen días para los que algunas estaciones no cuentan con ningún registro (i.e. no existe una fila en la base de datos). En el Cuadro 6.4 se indica la distribución de valores faltantes para cada una de las variables meteorológicas. Para solucionar esto, primero se agregan las filas correspondientes a los pares fecha-estación ausentes, y a continuación, se procede a rellenar los valores vacíos. El proceso seguido para las diferentes variables es el siguiente:

- *Precipitación, temperatura media, máxima y mínima*: Se ingresa el valor proporcionado por la base de datos de ENACTS, la cual es proporcionada por INSIVUMEH en archivos de formato NetCDF. Se busca el valor asociado a la fecha y las coordenadas geográficas de la estación meteorológica.
- *Humedad relativa, velocidad y dirección del viento*: Se realiza una interpolación de la distribución geográfica de la variable en el día de valor desconocido. Para realizar la interpolación se utiliza la latitud y longitud de las estaciones cuyo valor es conocido para dicho día. Se utiliza un modelo de variograma lineal de *kriging*, en el que la variable cambia linealmente con la distancia entre los puntos de la red. Una vez realizada la interpolación, el valor desconocido se obtiene al evaluar la latitud y longitud de la estación en el modelo resultante.
- *Presión atmosférica*: Se realiza una regresión lineal por mínimos cuadrados para el comportamiento de la presión atmosférica como función de la altitud en el día de valor desconocido. Se ajusta el modelo al conjunto de altitudes de las estaciones meteorológicas con valor de presión conocido para ese día. Al tener la regresión, se ingresa la altitud de la estación para conocer el valor de presión asociado.

Variable meteorológica	Valores faltantes	
	Cantidad	Proporción (%)
Precipitación	113590	15.24
Temperatura media	145664	19.54
Temperatura máxima	124420	16.69
Temperatura mínima	125502	16.84
Humedad relativa	231449	31.06
Velocidad del viento	225313	30.23
Dirección del viento	232584	31.21
Presión atmosférica	564892	75.80

Cuadro 6.4: Cantidad y proporción de valores faltantes en la base de datos original

La base de datos original no cuenta con las variables objetivo de interés, por lo que es necesario construirlas. Para ello, primero se calcula el promedio y los percentiles de la Climatología; esto se realiza para cada uno de los días del año en cada una de las estaciones meteorológicas (coordenadas geográficas). Los resultados son guardados en un archivo *xlsx* con el fin de garantizar su permanencia en el tiempo. Mediante el uso de estos promedios y percentiles, se procede a construir las variables objetivo para cada una de las entradas de la base de datos de acuerdo con lo mencionado en el apartado *Variables objetivo*. Al terminar el proceso, se aplica un desfase de un día hacia arriba para

garantizar que en cada fila las variables objetivo correspondan a la precipitación del día posterior a la fecha de referencia.

En muchos casos ambos percentiles presentaron un valor de 0 mm , de hecho esto ocurrió en casi el 43 % de las ocasiones. Esto se debe a que en la época seca, existen días en los que consistentemente la precipitación ha sido nula a lo largo de la Climatología para algunas estaciones, por lo que ambos percentiles se corresponden con un valor de 0 mm . Nótese que en estos casos una precipitación de 0 mm sería clasificada, simultáneamente, como BN y AN de acuerdo con el Cuadro 6.1, lo que carece de sentido. En estos casos, un registro de 0 mm se clasifica como NN , ya que este valor es el 'normal' climatológico, mientras que registros mayores a 0 mm se clasifican como AN .

Una vez construidas las variables objetivo para cada una de las entradas de la base de datos, se procede a agregar columnas correspondientes a los valores de las variables meteorológicas en cada uno de los días del rango de observación, las cuales serán utilizadas como predictores. Para ello, se duplica la columna original de la variable y se le aplica un desfase equivalente al número de días que separa cada día de observación de la fecha de referencia.

A continuación, se agrega una columna para la latitud y otra para la longitud geográfica de la estación meteorológica. También se agrega una columna con el número de mes (del 1 al 12) de la fecha de referencia. Se eliminan las columnas que contienen el nombre de la estación y la fecha de referencia, así como los valores de las variables meteorológicas en este día.

Seguidamente, se realiza un proceso de barajado en el que el orden de las filas de la base de datos se reestructura de una forma aleatoria. Nótese que la base de datos resultante contiene entradas de todos los meses del año. Esta base de datos se separa en función del número de mes para obtener dos bases de datos adicionales: un conjunto de datos exclusivos de la época lluviosa y otro con datos exclusivos de la época seca. Finalmente, las tres bases de datos (general, época lluviosa y época seca) son divididas en un 80 % para datos de entrenamiento y un 20 % para datos de evaluación. Los conjuntos resultantes son exportados en archivos *xlsx* para su posterior uso en el desarrollo de los modelos.

6.3.4. Evaluación de modelos y comparación con WRF y NextGen

Cada uno de los modelos desarrollados es evaluado con el fin de cuantificar la calidad de sus pronósticos. Para esto, se le ingresa al modelo las variables predictoras del conjunto de evaluación, de manera que éste proporciona un pronóstico para cada una de las instancias del conjunto. A partir de los pronósticos y los valores reales, se calculan diferentes métricas que permiten cuantificar el rendimiento del modelo. Para los modelos de regresión se calcula el MAE y el MdAE, mientras que para los modelos de clasificación se calcula la exactitud, precisión, sensibilidad y especificidad. Este proceso de evaluación se realiza con datos generales (de todos los meses del año), con datos de la época lluviosa, y con datos de la época seca; esto se realiza con el fin de evaluar el rendimiento a lo largo de diferentes períodos de tiempo.

Con respecto a los modelos de regresión, puede darse el caso de que la anomalía pronosticada por el modelo sea menor que el negativo del promedio de la Climatología. En estos casos, el pronóstico corresponde a un evento de precipitación negativa, lo que es imposible en la vida real. Esta incongruencia se soluciona mediante el reemplazo del pronóstico original por una anomalía igual al negativo del promedio de la Climatología, ya que el valor de precipitación asociado en este caso es de 0 mm (que es el valor físicamente posible más cercano al pronóstico original). Por ejemplo, considere que el promedio de la precipitación durante la Climatología es de 5 mm y que el modelo pronostica una anomalía de -7 mm , esto indica que la precipitación esperada es de $5\text{ mm} - 7\text{ mm} = -2\text{ mm}$, lo que no tiene interpretación física; en este caso se reemplaza el pronóstico por una anomalía de -5 mm , con lo que la precipitación esperada es de 0 mm .

Es importante mencionar que para el caso de los modelos de cotas percentiles, dado que se trata de un clasificador multiclase, existe una precisión, sensibilidad y especificidad diferentes para cada una de las tres categorías. Se decidió calcular estas métricas únicamente para la categoría *NN* por dos motivos principales: primero, para no saturar el estudio con demasiados resultados que lo desvíen de su objetivo fundamental, y segundo, porque evaluar la categoría *NN* permite conocer el rendimiento del modelo en términos de diferenciar correctamente un fenómeno de precipitación normal de uno no normal, lo que resulta útil en el pronóstico de tormentas y sequías.

El rendimiento de los modelos de machine learning es comparado con el del modelo WRF-BMJ GFS, utilizado actualmente por INSIVUMEH para pronósticos meteorológicos de corto plazo. Para cada fecha, se trabaja con el *raster* correspondiente al pronóstico de precipitación de la fecha objetivo, y que no emplea como datos de entrada las variables en la fecha de referencia (esto último para comparar escenarios con condiciones equivalentes). Se toma la precipitación del *raster* correspondiente a la ubicación geográfica de cada una de las 47 estaciones meteorológicas. A partir de este valor se construyen las tres variables objetivo mediante el uso del promedio y los percentiles de la Climatología calculados durante el preprocesamiento. Para obtener los valores reales, se consulta en los datos de INSIVUMEH la precipitación registrada en cada una de las estaciones para ese día, y, nuevamente, se construyen las variables objetivo a partir del promedio y los percentiles. Este proceso se repite para 7 días de cada mes en el período Nov 2023 - Oct 2024, abarcando por completo una época seca y una época lluviosa. Finalmente, habiendo calculado los pronósticos y los valores reales, se procede a calcular las métricas de evaluación para cada una de las tres variables objetivo. Al igual que con los modelos de machine learning, las métricas de WRF se calculan para datos generales, para datos solo de la época lluviosa y para datos exclusivos de la época seca.

Los modelos de machine learning también son comparados con el modelo NextGen. Dado que los pronósticos de NextGen son estacionales, y no diarios (como es el objetivo de este estudio), la única variable objetivo que se evalúa es la clasificación con cotas percentiles, cuya interpretación es la misma del Cuadro 6.1, con la diferencia de que en lugar de trabajar con precipitaciones diarias se consideran precipitaciones acumuladas. Para ello, primero es necesario determinar los percentiles de la Climatología, los cuales se calculan para cada período estacional y para cada estación meteorológica a partir del conjunto de 30 valores de precipitación acumulada durante el período 1991-2020. Los períodos estacionales son MJJ (may-jul), ASO (ago-oct) y DEFM (dic-mar). A continuación, se sigue un procedimiento similar al de WRF: a partir del *raster* se construye la categoría pronosticada, y la categoría real se construye a partir de los registros de precipitación de INSIVUMEH. La evaluación de NextGen toma en cuenta todos los pronósticos en el período ASO 2022 - ASO 2024. Una vez calculada la categoría pronosticada y la categoría real, se procede a calcular las métricas de evaluación de algoritmos de clasificación. Se realiza una evaluación con datos generales, una con datos de la época lluviosa (períodos MJJ y ASO) y una con datos de la época seca (que incluye únicamente el período DEFM).

6.4. Diagrama de flujo procedimental

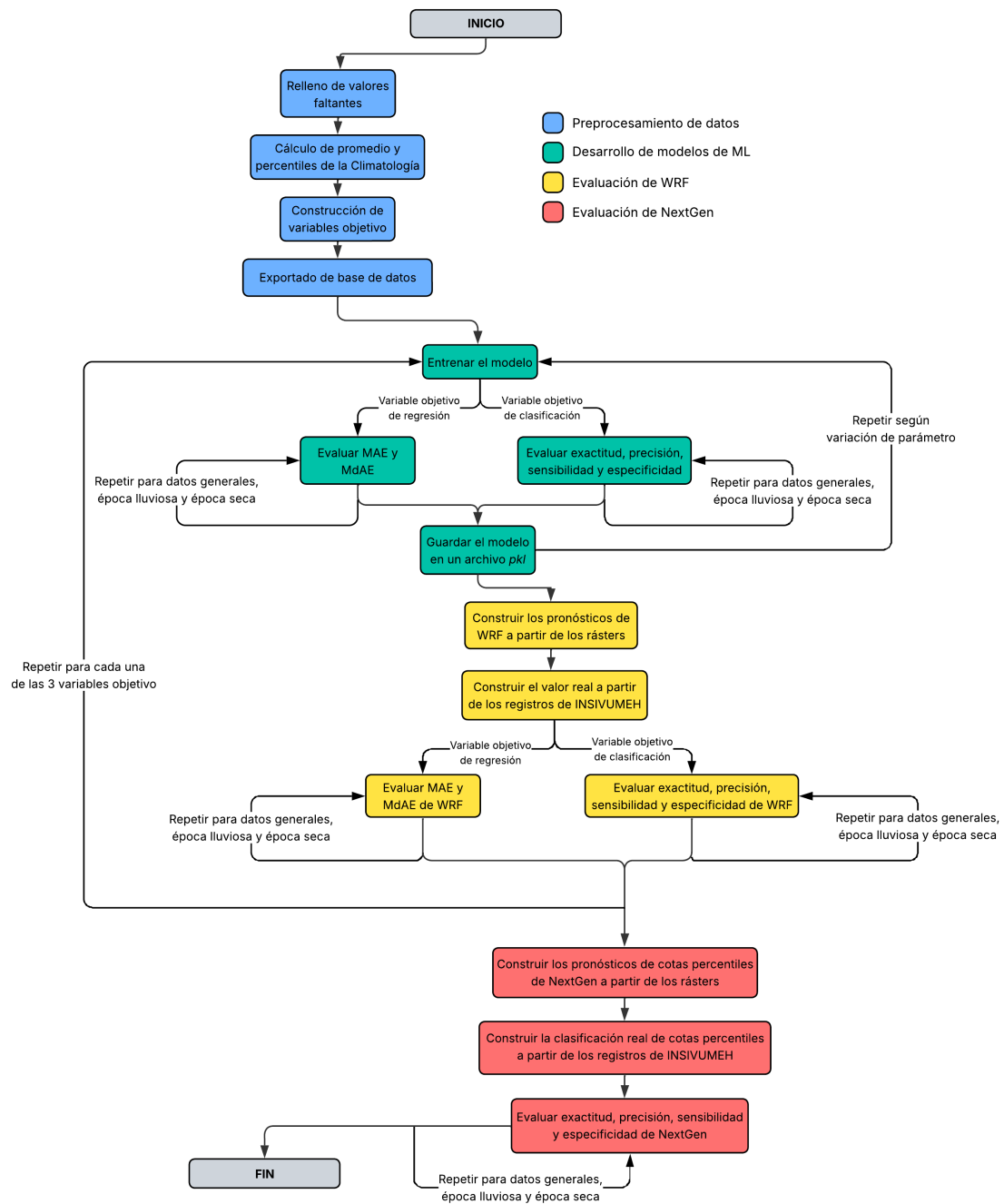


Figura 6.3: Diagrama de flujo procedimental

En la sección de *Anexos* se encuentran los valores puntuales de rendimiento de los modelos, así como su tiempo de entrenamiento. También se incluye el enlace a un repositorio de *GitHub*, que permite acceder al código de programación desarrollado.

Cada una de las figuras presentadas en esta sección permite acceder a su versión interactiva, ya sea escaneando el código *QR* incluido, o presionando directamente sobre la figura. En dicha versión, es posible realizar múltiples acciones, como hacer *zoom* al contenido o descargar la imagen en formato *png*. En la sección de *Anexos* se incluye el enlace a un video explicativo sobre el uso de esta versión interactiva; presione **aquí** para verlo.

7.1. Modelos de regresión

En el diagrama de dispersión de la Figura 7.1 se resume el rendimiento de los modelos de regresión evaluados con datos de todo el año. En el eje vertical se muestra el MdAE, mientras que el eje horizontal corresponde al MAE. Cada modelo de machine learning es representado con una figura geométrica dentro del gráfico. Los factores como el color de relleno, el tipo de borde, así como el tamaño y la forma de la figura geométrica se utilizan para identificar los diferentes parámetros del modelo. Con el fin de no saturar demasiado el gráfico, se presentan únicamente modelos con 100 árboles de decisión. Las líneas azules corresponden al rendimiento del modelo WRF evaluado con datos de todo el año. De forma análoga, en las figuras 7.2 y 7.3 se muestra el rendimiento de los modelos de regresión evaluados con datos de la época lluviosa y con datos de la época seca, respectivamente. Por su parte, en el Cuadro 7.1 se muestran las métricas de rendimiento de WRF.

En las figuras de la 7.4 a la 7.8 se muestran gráficas de barras que indican el promedio de MAE de los modelos de regresión según la variación de cada parámetro (algoritmo, número de árboles, combinación de variables, número de días en el rango de observación y tipo de entrenamiento). Estos valores son calculados al promediar el MAE de todos los modelos que comparten el parámetro indicado. Por ejemplo, la primera barra (de izquierda a derecha) de la Figura 7.4 es el promedio de MAE durante todo el año de los modelos de regresión que hacen uso del algoritmo de regresión lineal múltiple. En cada una de estas figuras se presentan líneas horizontales que corresponden al

MAE del modelo WRF en cada uno de los tres tipos de evaluación.

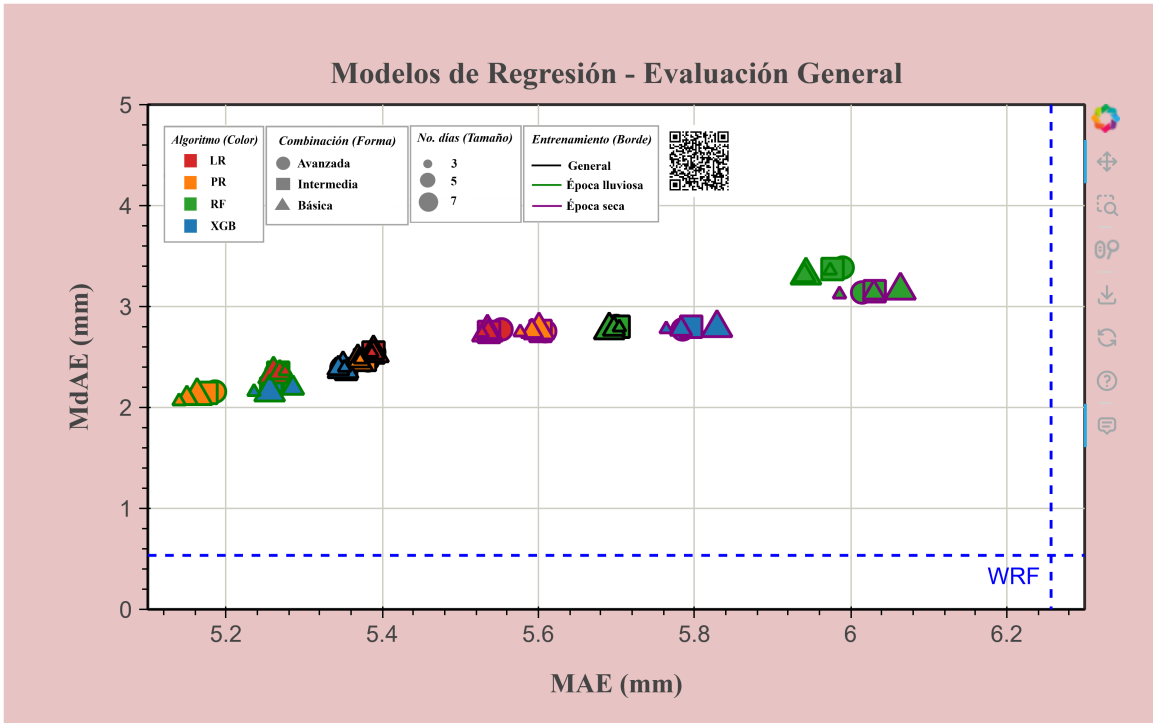


Figura 7.1: Rendimiento de modelos de regresión durante todo el año

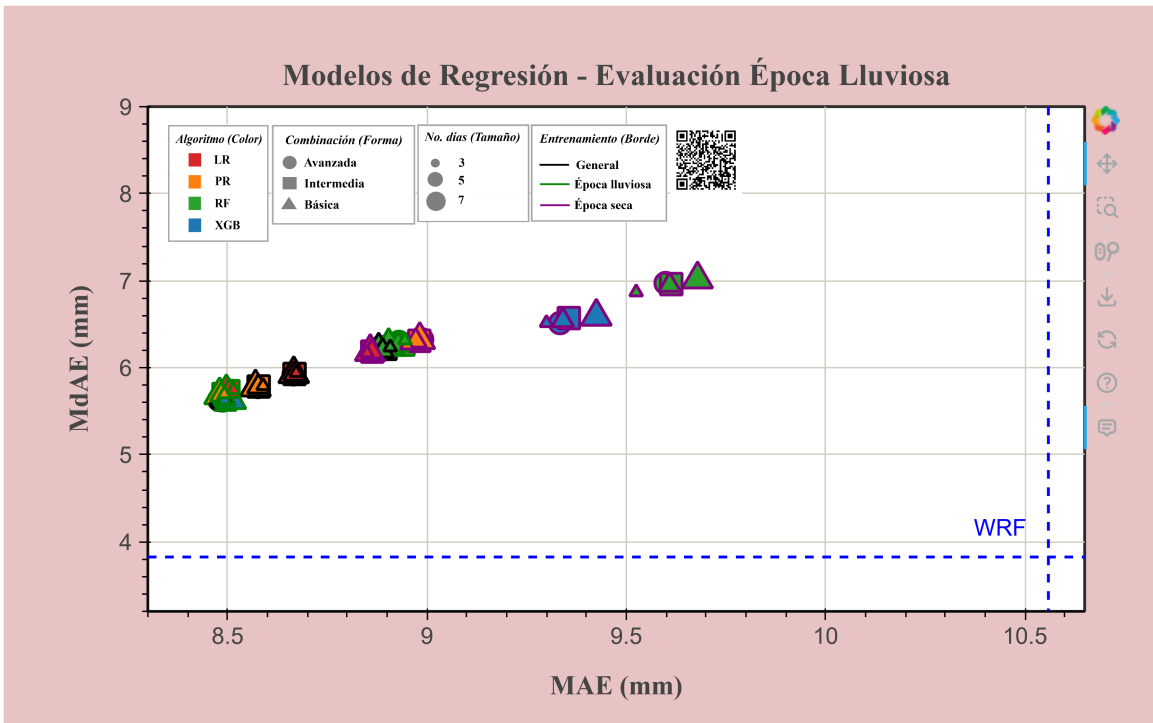


Figura 7.2: Rendimiento de modelos de regresión durante la época lluviosa

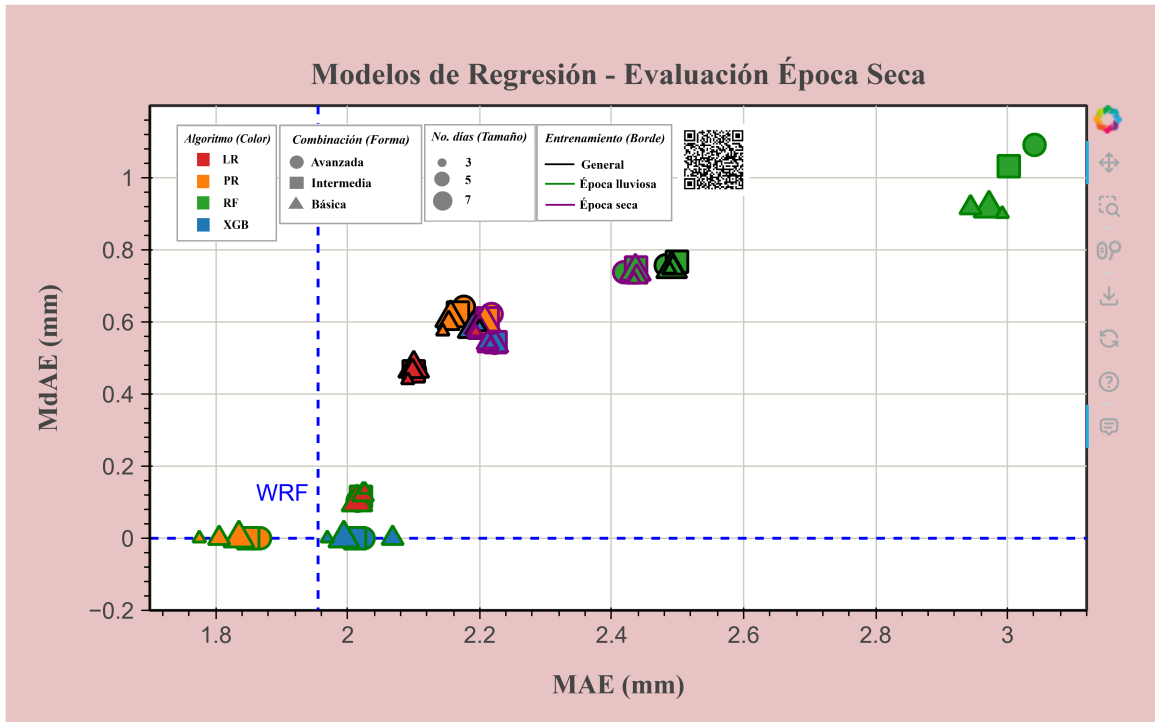


Figura 7.3: Rendimiento de modelos de regresión durante la época seca

Evaluación	MAE (mm)	MdAE (mm)
General	6.257	0.535
Época Lluviosa	10.559	3.825
Época Seca	1.955	0.000

Cuadro 7.1: Rendimiento del modelo WRF-BMJ-GFS para regresión

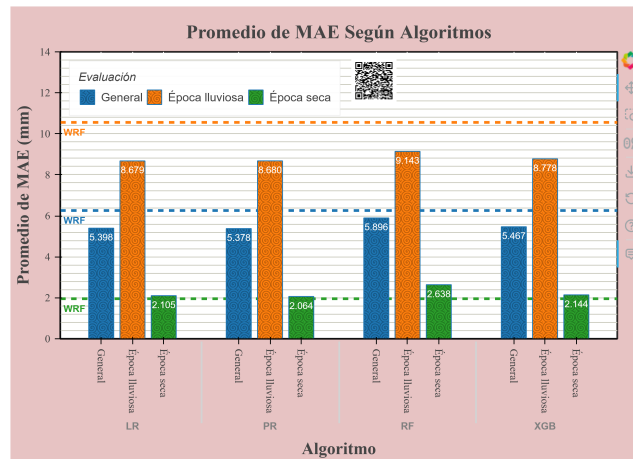


Figura 7.4: MAE promedio de modelos de regresión según algoritmos

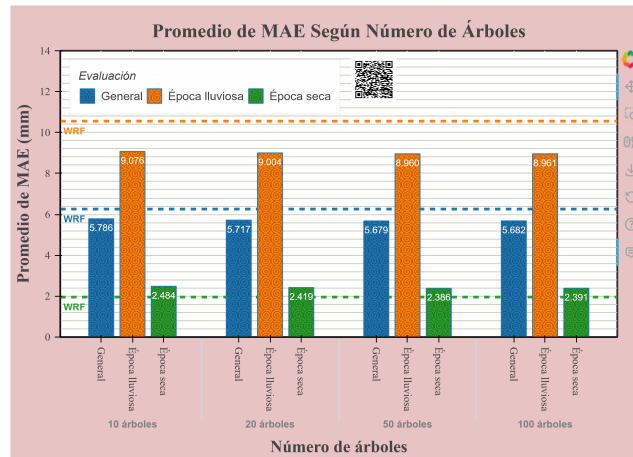


Figura 7.5: MAE promedio de modelos de regresión según número de árboles

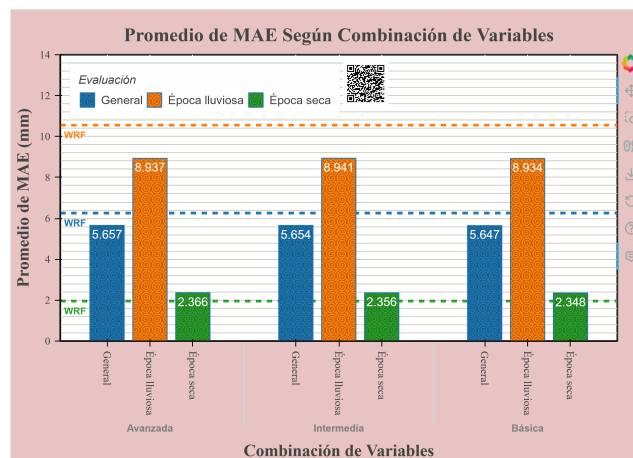


Figura 7.6: MAE promedio de modelos de regresión según combinación de variables

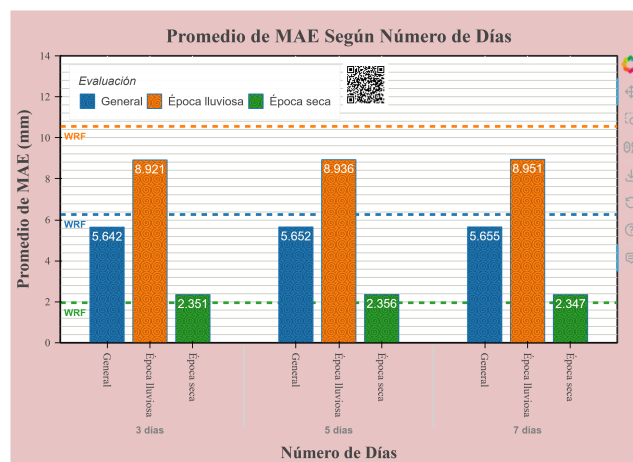


Figura 7.7: MAE promedio de modelos de regresión según número de días en el rango de observación

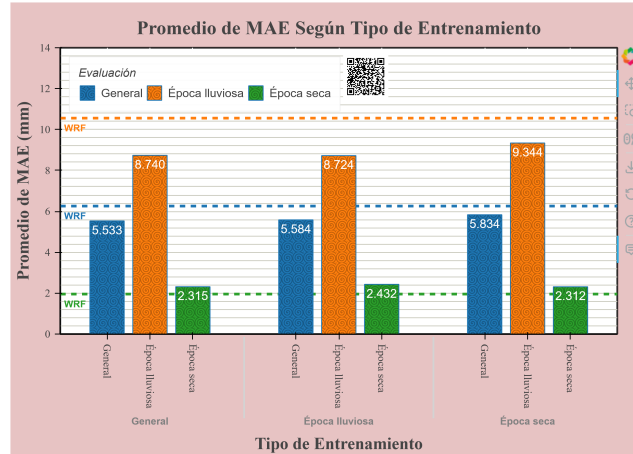


Figura 7.8: MAE promedio de modelos de regresión según tipo de entrenamiento

7.2. Modelos de clasificación de cotas percentiles

En la Figura 7.9 se presentan tres diagramas de dispersión que resumen la evaluación con datos de todo el año de los modelos de clasificación con cotas percentiles. Los tres diagramas comparten la misma escala horizontal: la exactitud del modelo. Cada diagrama muestra una métrica diferente en su escala vertical (precisión, sensibilidad y especificidad) que corresponde a la evaluación de la categoría *NN*. Los modelos de machine learning son representados mediante figuras geométricas, cuyo borde, tamaño, forma y color de relleno son diferentes en función de los parámetros utilizados. Con el fin de no saturar el gráfico, se presentan únicamente modelos con 100 árboles de decisión. De forma similar, en las figuras 7.10 y 7.11 se muestra el rendimiento de los modelos evaluados con datos de la época lluviosa y con datos de la época seca, respectivamente. Las líneas punteadas azules y rojas corresponden al rendimiento de los modelos WRF y NextGen, respectivamente, cuyos valores se muestran en los cuadros 7.2 y 7.3.

En las figuras de la 7.12 a la 7.16 se muestran gráficas de barras que indican el promedio de exactitud de los modelos de cotas percentiles según la variación de los diferentes parámetros (algoritmo, número de árboles, combinación de variables, número de días en el rango de observación y tipo de entrenamiento). Estos valores se calculan promediando la exactitud de todos los modelos que comparten el parámetro indicado. A manera de ejemplo, el valor de la primera barra de la Figura 7.12 es el promedio de la exactitud durante todo el año de los modelos de cotas percentiles que utilizan el algoritmo de regresión logística. Las rectas horizontales de estas figuras indican la exactitud de los modelos WRF y NextGen para cada tipo de evaluación.

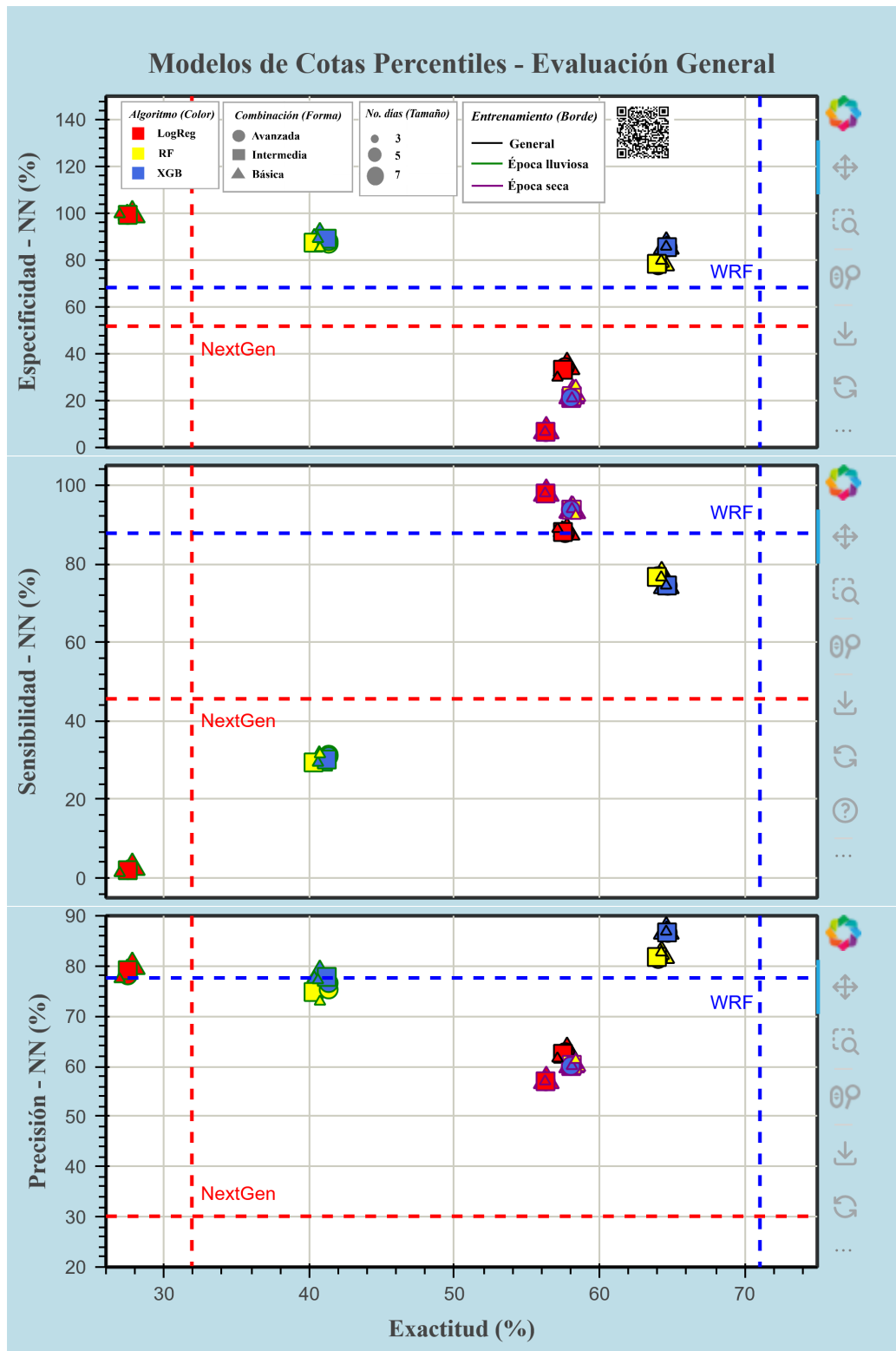


Figura 7.9: Rendimiento de modelos de clasificación con cotas percentiles durante todo el año

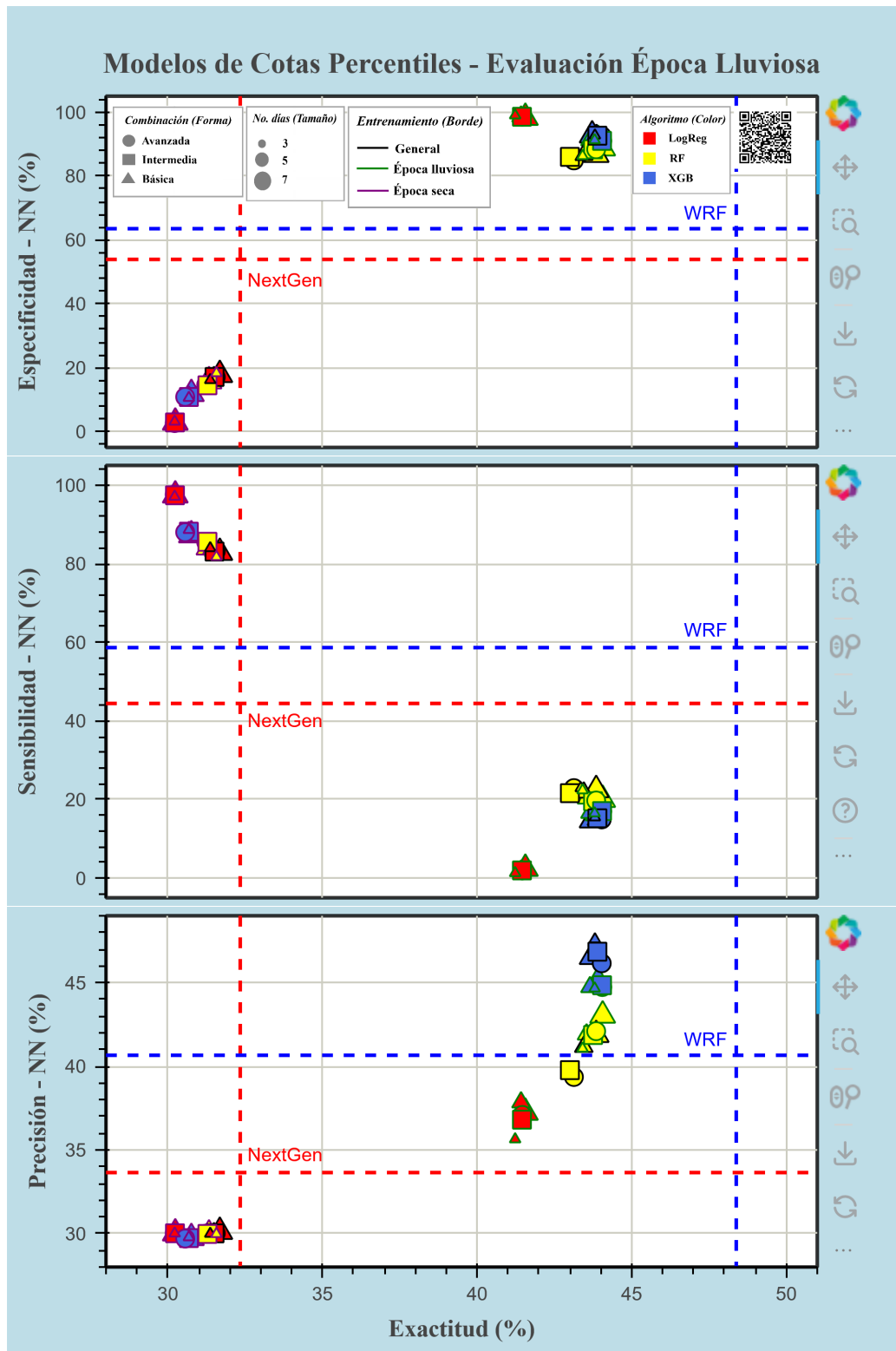


Figura 7.10: Rendimiento de modelos de clasificación con cotas percentiles durante la época lluviosa

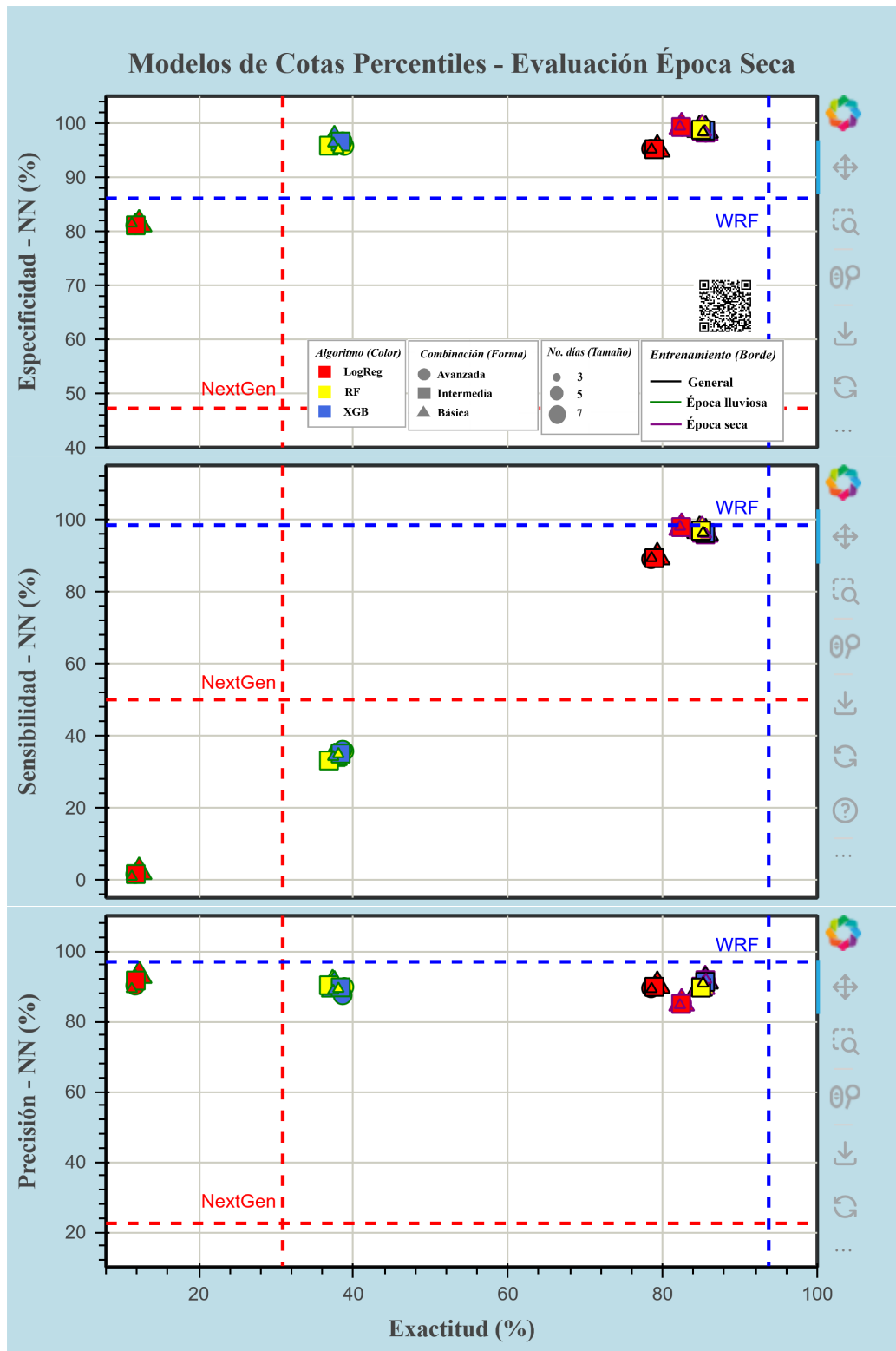


Figura 7.11: Rendimiento de modelos de clasificación con cotas percentiles durante la época seca

Tipo Evaluación	Exactitud (%)	Categoría <i>NN</i>		
		Precisión (%)	Sensibilidad (%)	Especificidad (%)
General	71.05	77.62	87.76	68.25
Época Lluviosa	48.38	40.66	58.64	63.51
Época Seca	93.72	96.88	98.44	86.10

Cuadro 7.2: Rendimiento del modelo WRF-BMJ-GFS para clasificación con cotas percentiles

Tipo Evaluación	Exactitud (%)	Categoría <i>NN</i>		
		Precisión (%)	Sensibilidad (%)	Especificidad (%)
General	31.91	30.13	45.63	51.77
Época Lluviosa	32.34	33.64	44.44	53.90
Época Seca	30.85	22.45	50.00	47.22

Cuadro 7.3: Rendimiento del modelo NextGen para clasificación con cotas percentiles

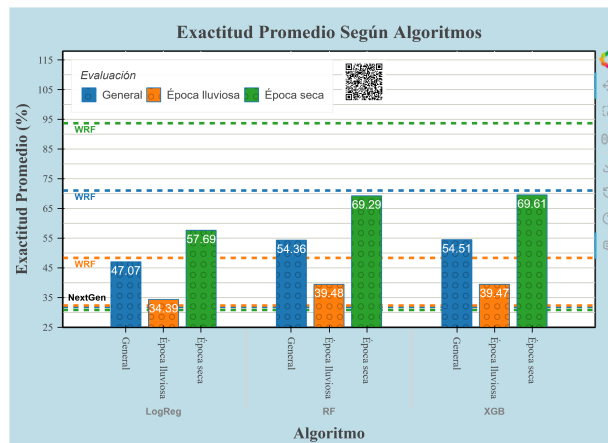


Figura 7.12: Exactitud promedio de modelos de clasificación con cotas percentiles según algoritmos

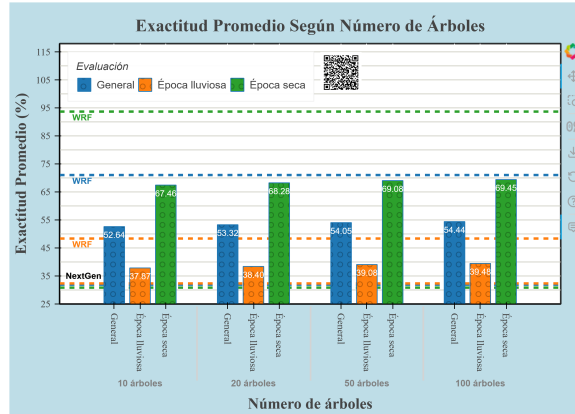


Figura 7.13: Exactitud promedio de modelos de clasificación con cotas percentiles según número de árboles

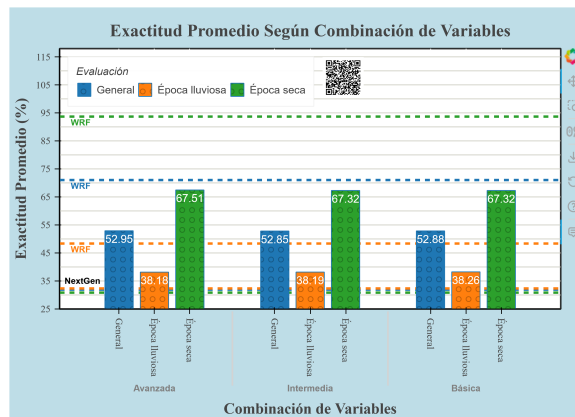


Figura 7.14: Exactitud promedio de modelos de clasificación con cotas percentiles según combinación de variables

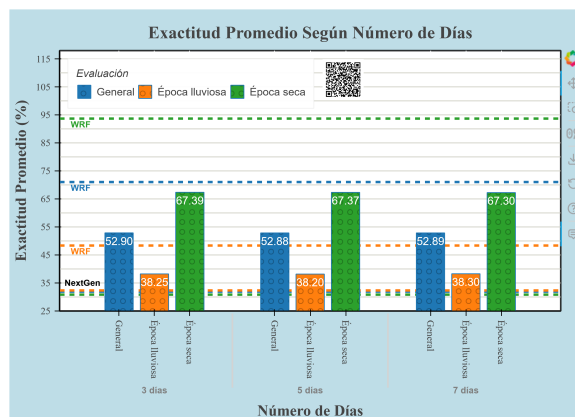


Figura 7.15: Exactitud promedio de modelos de clasificación con cotas percentiles según número de días en el rango de observación

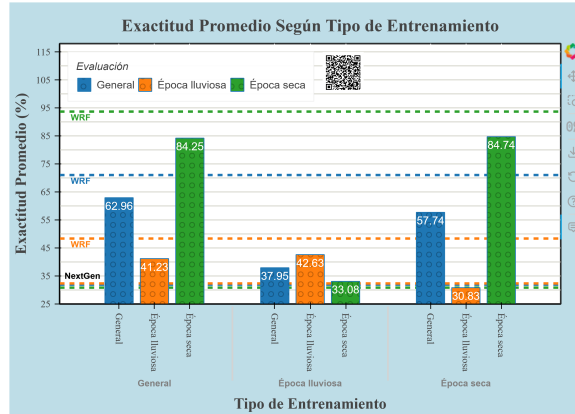


Figura 7.16: Exactitud promedio de modelos de clasificación con cotas percentiles según tipo de entrenamiento

7.3. Modelos de clasificación de lluvia binaria

En las figuras 7.17, 7.18 y 7.19 se resume el rendimiento de los modelos de clasificación de lluvia binaria durante todo el año, la época lluviosa y la época seca, respectivamente. Cada una de las figuras presenta tres diagramas de dispersión cuya escala horizontal es la exactitud del modelo. El eje vertical corresponde a una métrica diferente en cada diagrama (precisión, sensibilidad y especificidad). Los modelos de machine learning son representados con una figura geométrica, cuyo tamaño, forma, borde y color de relleno indican la variación de parámetros utilizada, de acuerdo con la leyenda de la figura. Con el objetivo de no saturar demasiado el gráfico, solamente se presentan modelos con 100 árboles de decisión. Las líneas punteadas de color azul indican las métricas de rendimiento del modelo WRF, cuyos valores se presentan en el Cuadro 7.4.

En las figuras de la 7.20 a la 7.24 se muestran gráficas de barras con el promedio de exactitud de los modelos de lluvia binaria según la variación de los diferentes parámetros (algoritmo, número de árboles, combinación de variables, número de días en el rango de observación y tipo de entrenamiento). Estos valores se calculan promediando la exactitud de todos los modelos que comparten el parámetro indicado. Por ejemplo, la última barra de la Figura 7.23 es el promedio de la exactitud durante la época seca de los modelos de lluvia binaria que utilizan 7 días como rango de observación. Las líneas horizontales de estas figuras indican la exactitud del modelo WRF.

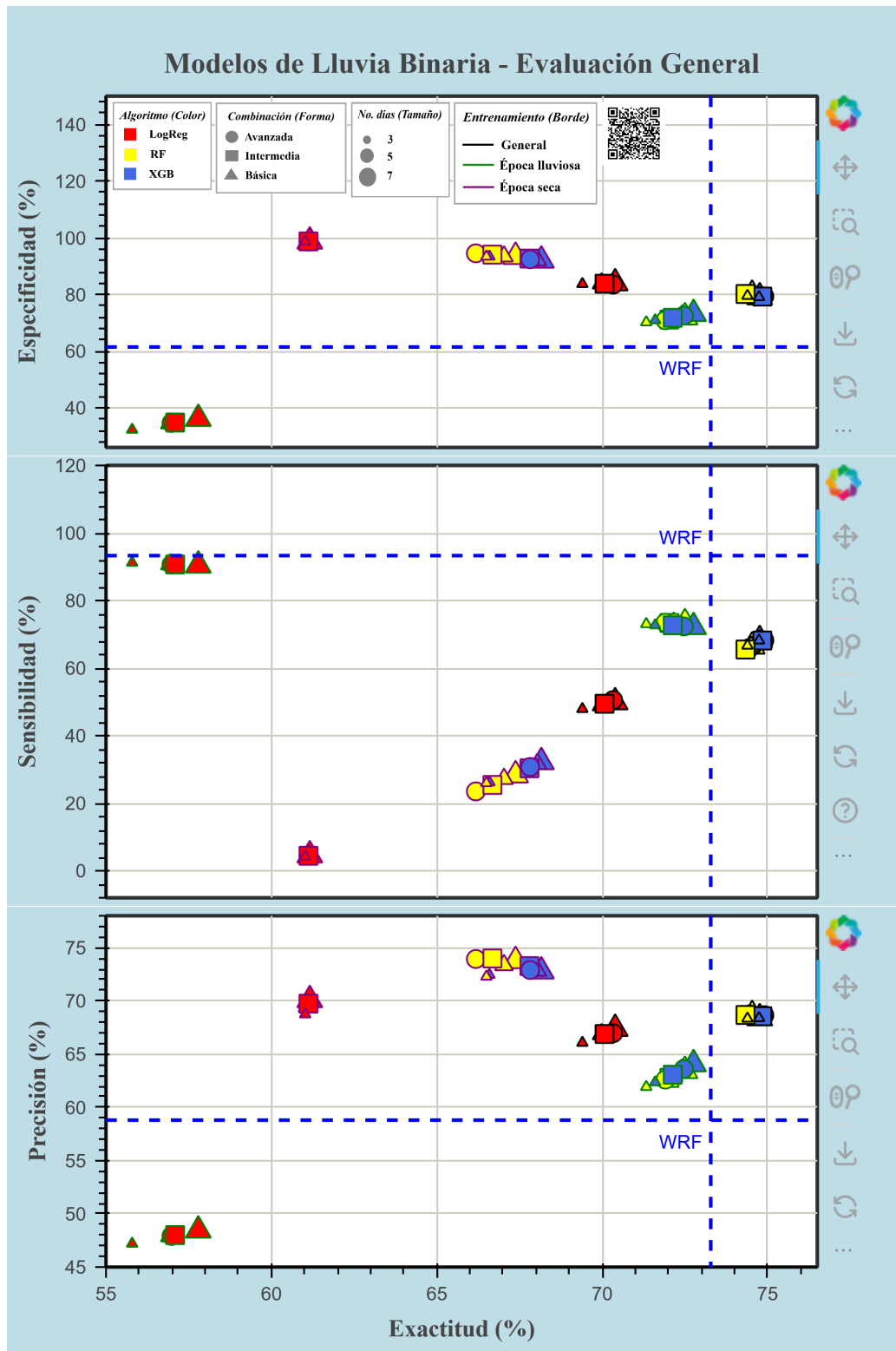


Figura 7.17: Rendimiento de modelos de clasificación de lluvia binaria durante todo el año

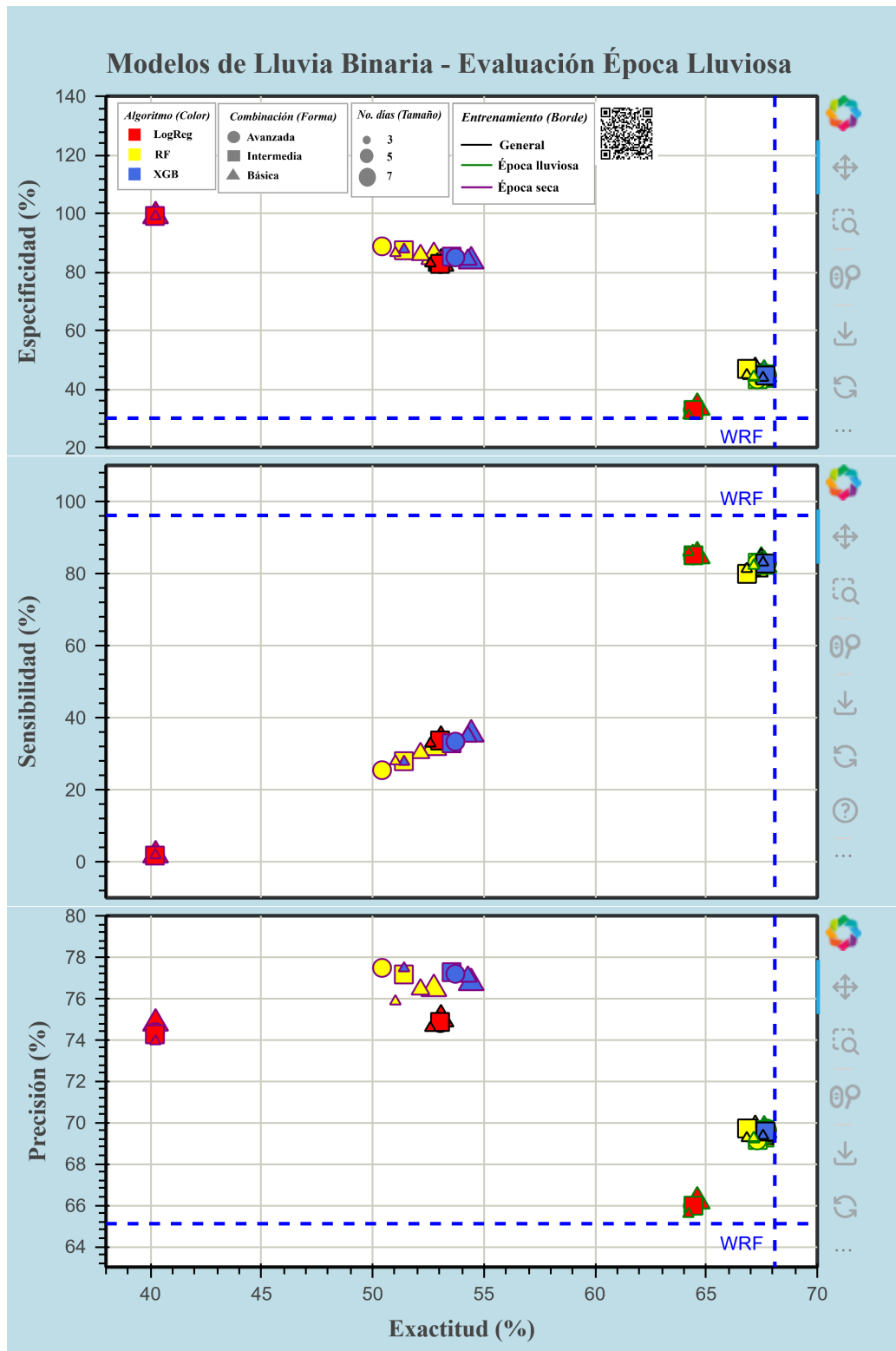


Figura 7.18: Rendimiento de modelos de clasificación de lluvia binaria durante la época lluviosa

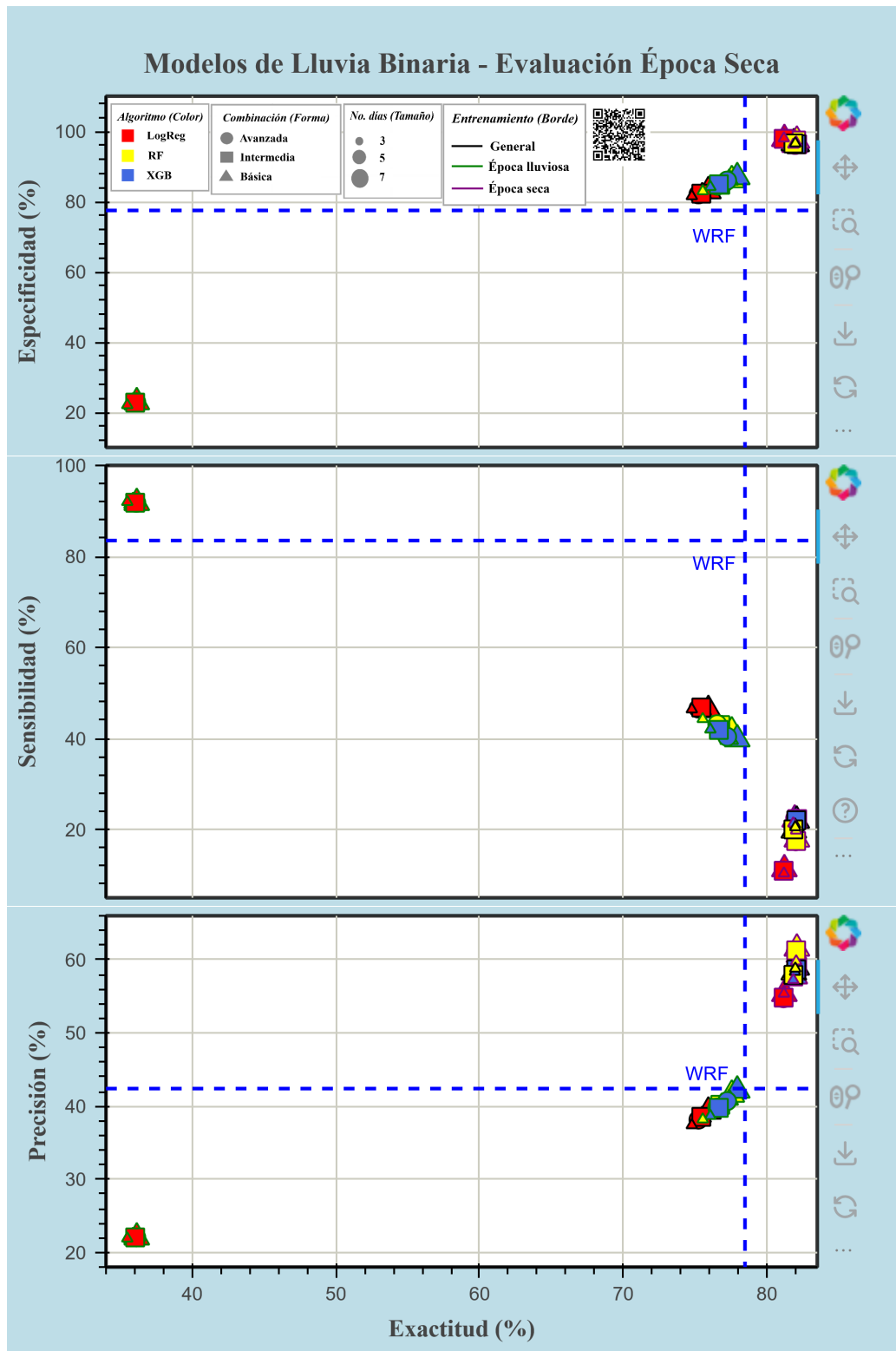


Figura 7.19: Rendimiento de modelos de clasificación de lluvia binaria durante la época seca

Evaluación	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
General	73.28	58.80	93.31	61.47
Época Lluviosa	68.09	65.10	96.13	29.99
Época Seca	78.47	42.39	83.49	77.47

Cuadro 7.4: Rendimiento del modelo WRF-BMJ-GFS para clasificación de lluvia binaria

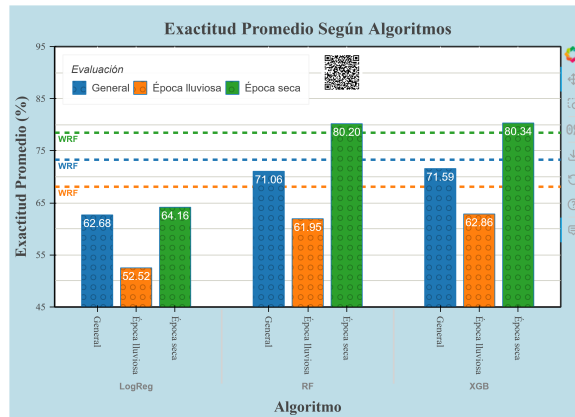


Figura 7.20: Exactitud promedio de modelos de clasificación de lluvia binaria según algoritmos

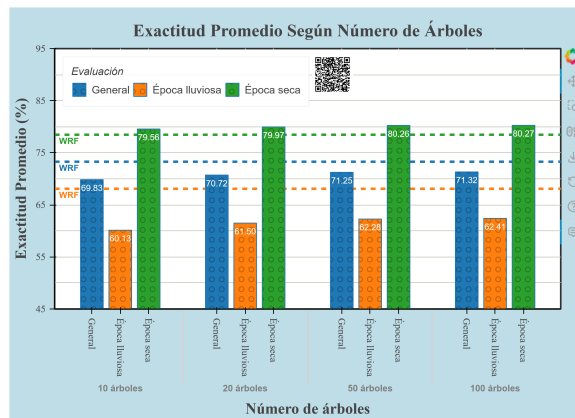


Figura 7.21: Exactitud promedio de modelos de clasificación de lluvia binaria según número de árboles

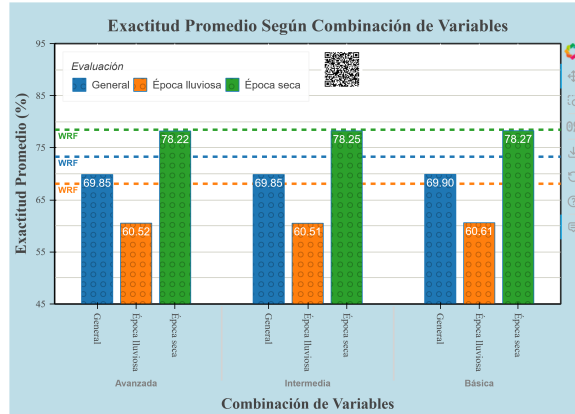


Figura 7.22: Exactitud promedio de modelos de clasificación de lluvia binaria según combinación de variables

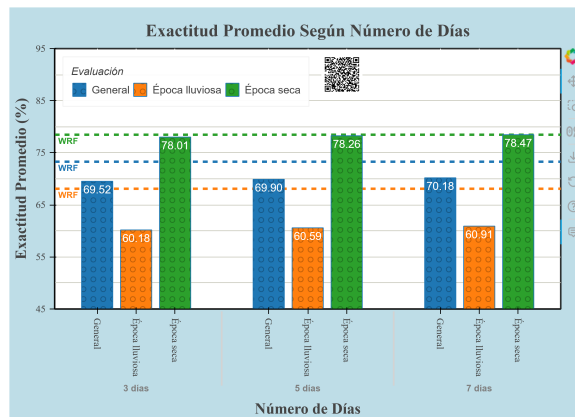


Figura 7.23: Exactitud promedio de modelos de clasificación de lluvia binaria según número de días en el rango de observación

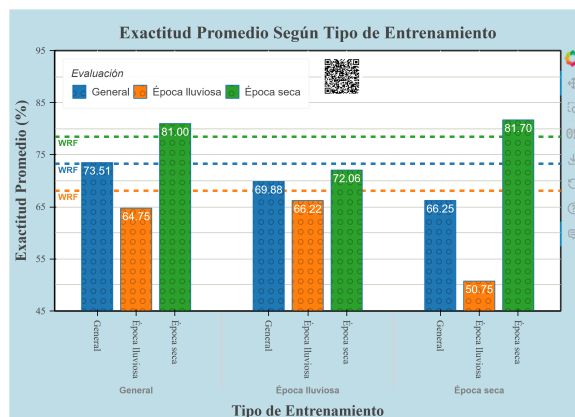


Figura 7.24: Exactitud promedio de modelos de clasificación de lluvia binaria según tipo de entrenamiento

8.1. Análisis de rendimiento

Al comparar el rendimiento de los modelos de regresión, en las figuras 7.1, 7.2 y 7.3, se observa que los modelos de machine learning tienden a presentar un menor valor de MAE (con excepción de la evaluación en época seca), y un valor igual o superior de MdAE, en comparación con el modelo WRF. Pese a tener una mediana de error alta, el promedio del error se reduce en los modelos de machine learning. Esto es un indicativo de que las mejores predicciones, aunque son minoría, tienen errores asociados muy bajos. En cambio, el modelo WRF, a pesar de tener un MAE más alto, mantiene una mediana baja, lo que implica que al menos la mitad de sus predicciones tienen errores pequeños y que solo algunos casos extremos elevan el promedio. Por lo tanto, WRF resulta conveniente cuando la prioridad es obtener errores pequeños en la mayoría de ocasiones, pero en pronósticos puntuales, los modelos de machine learning obtienen resultados más cercanos a la realidad.

Ninguno de los modelos de machine learning para clasificación de cotas percentiles logró superar la exactitud del modelo WRF en ninguno de los tres tipos de evaluación (Ver figuras 7.9, 7.10 y 7.11), por lo que en términos de clasificar correctamente la mayor cantidad posible de instancias, el modelo ganador es WRF. Con respecto a las otras métricas de evaluación, existieron algunos modelos que sí tuvieron un mejor rendimiento que WRF. Por otro lado, todos los modelos (con excepción de *regresión logística* entrenada con época lluviosa) presentaron una mayor exactitud que NextGen durante la evaluación general; esto tiene sentido, ya que los pronósticos de NextGen no solo se realizan con un mayor tiempo de anticipación, sino que además consideran la precipitación durante un mayor rango temporal (son estacionales), por lo que evidentemente la calidad de sus pronósticos tiende a ser menor. Aún así, al considerar los tres tipos de evaluación, se observa que NextGen logró vencer a varios modelos de machine learning en términos de precisión, sensibilidad y especificidad, por lo que a pesar de que, en general, NextGen clasifica incorrectamente una mayor proporción de instancias, en ocasiones es más consistente en sus resultados.

Al observar los resultados de modelos de clasificación de lluvia binaria en las figuras 7.17, 7.18 y 7.19, se tiene que WRF presentó un rendimiento notable en términos de exactitud durante la evaluación de época lluviosa; durante la evaluación general y la época seca existieron modelos de machine learning que sí superaron la exactitud de WRF. En cuanto a precisión y especificidad, el

rendimiento del machine learning se hizo notar, superando en muchas ocasiones a WRF. Se observa que ningún modelo logró superar la sensibilidad de WRF, por lo que este modelo es ideal para minimizar casos en los que el pronóstico indicó que no llovería, pero en realidad sí llovió (Falsos Negativos). Disminuir este tipo de errores es sumamente importante, ya que puede darse el caso de que la lluvia sea intensa, y al no conocer de antemano que lloverá, no se podrán tomar las acciones necesarias para minimizar los daños.

Todo lo discutido hasta este punto se resume en el Cuadro 8.1, el cual realiza una comparación entre el machine learning y el modelo WRF en función de la métrica en la que cada uno obtuvo un rendimiento superior al otro, así como la utilidad particular de cada modelo. En este sentido, resulta evidente que ambos pueden complementarse mutuamente, utilizando cada uno en los contextos donde su desempeño sea más favorable.

Variable objetivo	Modelo	Mejor Métrica	Útil para...	Notas
Regresión	ML	MAE	Obtener un error mínimo en pronósticos puntuales	Por sí solo, el modelo no determina cuándo arroja un resultado con error mínimo
	WRF	MdAE	Obtener consistentemente un error medianamente bajo	
Cotas Percentiles	ML			No se observó una tendencia que se cumpliera para todos los modelos
	WRF	Exactitud	Clasificar correctamente la mayor cantidad de veces posible	
Lluvia Binaria	ML	Precisión y Especificidad	Evitar la detección errónea de lluvias inexistentes	
	WRF	Sensibilidad	Detectar eventos aislados de lluvia durante época seca	

Cuadro 8.1: Cuadro comparativo entre los resultados de machine learning (ML) y WRF

En general, se observó que el rendimiento durante la época seca fue mejor en comparación con el rendimiento durante la época lluviosa. Esto se verifica al observar (en los diagramas de dispersión) que el rango de variación de cada una de las métricas es más favorable durante la época seca. El motivo de esto es la alta variabilidad de la precipitación durante la época lluviosa, que a menudo presenta fenómenos extremos y cambios abruptos cuyos patrones son más difíciles de generalizar. En consecuencia, los modelos presentan mayor facilidad para pronosticar instancias de la época seca. Este comportamiento se observó principalmente en los modelos de regresión y en los modelos de cotas percentiles. Por el contrario, en los modelos de lluvia binaria esta tendencia se vio disminuida a tal punto de ser opuesta en términos de precisión. Una posible explicación para esto es que, de las tres variables objetivo, la lluvia binaria es la más simple, por lo que la diferencia de complejidad entre las dos épocas disminuye y el modelo enfrenta menos dificultades para realizar pronósticos durante la época lluviosa. Finalmente, la evaluación con datos de todo el año estuvo en un punto intermedio (mejor que la época lluviosa, pero peor que la época seca), lo que tiene sentido dado que este período surge de combinar ambas épocas.

8.2. Análisis de variación de parámetros

En la Figura 7.4 se observa que el rendimiento promedio en términos de MAE fue muy similar para todos los algoritmos, siendo la *regresión lineal múltiple* y la *regresión polinomial* ligeramente

mejores al tener un menor valor de MAE. En cuanto a los modelos de clasificación, se observa en las figuras 7.12 y 7.20 que los mejores algoritmos en términos de exactitud fueron *bosque aleatorio* y *XGBoost*, los cuales superaron por un amplio margen el rendimiento de la *regresión logística*.

En cuanto al número de árboles, se tiene en la Figura 7.5 que el número óptimo para los modelos de regresión es de 50 árboles de decisión; mientras que para los modelos de clasificación es de 100 árboles (figuras 7.13 y 7.21). Sin embargo, es importante mencionar que la diferencia de rendimiento entre usar un modelo de 10 árboles y uno de 50 o 100 es relativamente pequeña (entre un 1% y un 4% de incremento relativo en la métrica de evaluación), por lo que utilizar el caso más simple (10 árboles) podría justificarse si el tiempo de entrenamiento aumenta rápidamente con el número de árboles (como es el caso del *bosque aleatorio*).

Se observa que para los modelos de clasificación, el rendimiento aumenta a medida que se incrementa el número de árboles, por lo que no se descarta que para modelos con más de 100 árboles el rendimiento pueda ser aún mejor. No obstante, también se observa que la mejora de rendimiento es cada vez menor a medida que el número de árboles crece (por ejemplo, para lluvia binaria, pasar de 10 a 20 árboles mejora la exactitud en un 0.41 – 1.37% dependiendo del tipo de evaluación, mientras que pasar de 50 a 100 árboles solo la incrementa en un 0.01 – 0.13%). Por lo tanto, aunque el rendimiento siguiera aumentando para más de 100 árboles, probablemente este no sería muy diferente de los resultados ya obtenidos.

En las figuras 7.8, 7.16 y 7.24 se observa que no existe un "mejor tipo de entrenamiento", sino que, dependiendo de la época del año donde se evalúe el modelo, será mejor un tipo de entrenamiento distinto. De hecho, el rendimiento es máximo cuando el entrenamiento coincide con el período de evaluación (durante todo el año es mejor utilizar modelos con entrenamiento general, durante la época lluviosa es mejor el entrenamiento con época lluviosa, y lo mismo para la época seca). Esto es lógico, ya que evidentemente, el modelo se ajustará mejor a datos que pertenezcan al mismo período de entrenamiento. Este comportamiento sugiere que para maximizar el rendimiento se debe utilizar un modelo entrenado con época lluviosa para pronosticar durante la época lluviosa, y uno entrenado con época seca para pronosticar durante la época seca. Si solo se puede entrenar un modelo, se recomienda hacerlo con datos de todo el año, ya que este tipo de entrenamiento presentó el máximo rendimiento general. Finalmente, es importante notar que la diferencia según el tipo de entrenamiento es menor en los modelos de regresión, por lo que lo anterior se aplica principalmente a los modelos de clasificación.

En ninguna de las tres variables objetivo se observó una diferencia relevante al variar la combinación de variables predictoras, pues en las figuras 7.6, 7.14 y 7.22 se observa como el cambio relativo en el rendimiento al variar la combinación de variables es siempre inferior al 1%, con lo que se garantiza que la combinación de variables tiene un impacto mínimo en el rendimiento del modelo. Una tendencia similar se observó en cuanto al número de días que componen el rango de variación, pues en las figuras 7.7, 7.15 y 7.23 se tiene que el rendimiento es prácticamente el mismo sin importar cuántos días se consideren. Este comportamiento se evidencia claramente en los diagramas de dispersión, donde se observa que las figuras geométricas se mueven en función del algoritmo y el tipo de entrenamiento; sin embargo, al variar la combinación y el número de días de observación, estas no se mueven, sino que se agrupan, indicando que no hay un cambio considerable en el rendimiento. Este agrupamiento ocurre no solo en el eje horizontal, sino también en el eje vertical, lo que sugiere que tampoco hay una variación importante en términos de MdAE (para regresión) ni de precisión, sensibilidad y especificidad (para clasificación).

Se observó que en la mayoría de ocasiones el peso que el modelo le asigna al último día del rango de observación es superior al peso de los demás días. En algunos casos esta diferencia es de hasta un orden de magnitud, aunque en la mayoría de ocasiones se mantiene en el mismo orden. Esto ocurrió principalmente con las variables de precipitación y humedad relativa. De esta forma, se evidencia que el modelo es capaz de reconocer que el día más reciente es el que tiene mayor impacto en la precipitación. Con respecto a los demás días del rango, no se pudo observar una tendencia en

cuanto a los pesos que el modelo les asigna. Esto explica por qué el número de días en el rango de observación no tuvo un impacto relevante en el rendimiento del modelo, pues dado que el último día es el que concentra la mayor cantidad de información, los pronósticos no se ven afectados, ya que todos los modelos incluyen a este día como un predictor.

Por su parte, en algunos modelos se observó una tendencia a asignar un menor peso a los valores de dirección, velocidad del viento, y en algunos casos, también de presión atmosférica en comparación con los pesos de las demás variables (principalmente la precipitación, cuyos pesos fueron los más altos). Si bien esta diferencia de pesos es menos intensa a la que se presentó entre los días del rango de observación, sigue siendo uno de los motivos por los que la combinación de variables no impactó en el rendimiento de los modelos: todas las combinaciones contienen las variables más relevantes. En este sentido, se considera que la mejor combinación es la *básica*, ya que deja de lado las variables menos importantes, minimizando así el número de variables sin afectar el rendimiento.

Los modelos fueron desarrollados utilizando los hiperparámetros definidos por defecto en cada uno de los algoritmos de machine learning. El único hiperparámetro que se varió fue el número de árboles de decisión en el caso de los algoritmos de *bosque aleatorio* y *XGBoost*. Sin embargo, es probable que los valores por defecto no sean los que maximicen el rendimiento del modelo, por lo que podrían existir combinaciones de hiperparámetros para las que el rendimiento es aún mejor que cualquiera de los modelos presentados en este trabajo.

El análisis sobre el rendimiento de los modelos en función de la variación de parámetros se realiza en términos de MAE para los modelos de regresión, y de exactitud para los modelos de clasificación, por lo que las conclusiones pueden ser diferentes si se utiliza una métrica distinta. El motivo de utilizar estas métricas se fundamenta en que cuantifican de forma general el rendimiento (el MAE promedia el error absoluto, y la exactitud mide la proporción de instancias correctamente clasificadas). Nótese además, que se utiliza el promedio de la métrica como indicador del rendimiento del parámetro, esto se realiza para tener una perspectiva global del desempeño sin tener que entrar a analizar cada modelo individualmente.

8.3. Aplicabilidad de los modelos

De acuerdo con el entrenamiento de los modelos, estos no toman en cuenta el comportamiento de la atmósfera en los alrededores, sino que se limitan a recibir como parámetros de entrada las variables meteorológicas en las mismas coordenadas geográficas que se busca pronosticar. Esto es una fuente de error, ya que se omite por completo la interacción que existe entre las diferentes partes de la atmósfera. Por ejemplo, considere que se registra un fuerte ingreso de humedad procedente del Océano Pacífico; evidentemente la probabilidad de lluvia en la capital aumenta, sin embargo esto no es detectado por el modelo, que solo toma en cuenta información de la capital al realizar el pronóstico. Este desconocimiento por parte del modelo sobre lo que ocurre en los alrededores se vuelve más importante a medida que el pronóstico se realiza con un mayor tiempo de anticipación, ya que existe una mayor ventana temporal para la interacción meteorológica. No obstante, dado que los modelos implementados realizan pronósticos con solamente un día de antelación, se puede considerar que el impacto en el rendimiento se ve reducido.

Los modelos pueden ser utilizados para realizar pronósticos en cualquier punto geográfico del país. Para tener una mejor interpretación de los resultados (principalmente la anomalía de precipitación y la clasificación con percentiles) se deberá conocer el promedio y los percentiles de la Climatología en el punto deseado, para lo cual se puede utilizar la base de datos de ENACTS. Es importante recordar que los modelos fueron entrenados y evaluados únicamente con datos de las estaciones meteorológicas de INSIVUMEH, que cubren un número limitado de puntos, por lo que realizar pronósticos para cualquier ubicación del país puede tener un rendimiento menor al presentado en la sección de Resultados.

El hecho de tener modelos para tres variables objetivo diferentes tiene la ventaja de poder combinar diferentes pronósticos para obtener un resultado más robusto. Por ejemplo, se puede comenzar prediciendo si lloverá o no; después, si el modelo indica que sí lloverá, se pueden utilizar los modelos de regresión para conocer el valor puntual; y finalmente se puede utilizar cotas percentiles para saber la intensidad de dicha lluvia. La detección de inconsistencias puede ser un indicativo de que el pronóstico no es confiable. Algunos ejemplos de inconsistencias son: que la regresión prediga un valor fuera del intervalo pronosticado por cotas percentiles, o que mientras la lluvia binaria indica que no lloverá, la regresión produce un valor de precipitación positivo. En estos casos, se recomienda complementar con modelos fuera del machine learning.

8.4. Interpretación de métricas de evaluación

Se recomienda el uso de modelos con bajo MdAE durante la época lluviosa y modelos con bajo MAE durante la época seca. Durante la época lluviosa, la precipitación es sumamente variable, lo que provoca que algunos días el modelo cometa errores muy grandes. Dado que el MAE es un promedio, estos errores le afectan mucho, haciendo que la evaluación del modelo se vea distorsionada por unos pocos días. Por lo tanto, durante la época lluviosa es mejor emplear el MdAE, que permite evaluar cuán bueno es el modelo en la mayoría de días (50%) sin que los valores atípicos perjudiquen la métrica. Por otro lado, durante la época seca, la distribución suele ser más homogénea, por lo que los errores suelen ser pequeños. En este caso, el MAE no se ve afectado por valores atípicos, por lo que se considera una buena métrica para evaluar el rendimiento promedio del modelo.

En cuanto a las métricas de los modelos de clasificación, si bien la exactitud cuantifica el rendimiento en términos generales, las demás métricas también proporcionan información relevante. De acuerdo con su definición, una alta sensibilidad minimiza los falsos negativos (FN), esto es útil para evitar casos en los que el modelo no detecta correctamente un evento potencialmente peligroso, cuyas consecuencias se habrían podido evitar al conocer de antemano lo que iba a ocurrir. Por su parte, los modelos con baja precisión y especificidad minimizan los falsos positivos (FP), lo que evita el desperdicio de recursos en el resguardo de la población frente a fenómenos que, si bien el modelo predijo que ocurrirían, finalmente no sucedieron. A medida que se desarrollen modelos con un mayor tiempo de anticipación, las ventajas de minimizar los FN y los FP serán aún más relevantes, ya que se tendrá un mayor tiempo para actuar en función del pronóstico arrojado por el modelo, y por lo tanto el costo de realizar una mala predicción aumentará.

8.5. Discusión general

Aunque inicialmente se consideró incluir el algoritmo de *máquinas de vectores de soporte* (SVM) y *k vecinos más cercanos* (KNN), estos resultaron ser demasiado costosos computacionalmente, por lo que finalmente no fueron implementados. Al intentar desarrollar un modelo de SVM con combinación avanzada y 5 días de observación, pasaron más de 24 horas sin que terminara la fase de entrenamiento. Por su parte, KNN requiere demasiado tiempo durante la fase de evaluación, ya que para clasificar un nuevo punto se debe calcular la distancia hacia todos los datos del conjunto de entrenamiento. Existen dos formas de solucionar este problema: (1) aumentar el poder de cómputo, lo que requiere de cierto poder adquisitivo, o (2) disminuir el tamaño de la base de datos de entrenamiento, por ejemplo, considerando un período de tiempo más corto.

Otra modificación que se realizó fue no desarrollar modelos de clasificación con cotas fijas, como inicialmente se había planteado, debido al desbalanceo de clases. Estos modelos clasificarían, en términos absolutos, la intensidad de precipitación en 4 categorías (leve, moderada leve, moderada intensa, intensa). Sin embargo, luego de construir la clasificación para cada entrada de la base de datos, se observó una sobrerepresentación de categorías. Inicialmente se utilizaron las cotas 10 mm,

20 mm y 30 mm, pero se observó que el 86 % de los datos pertenecían a la categoría *leve*. Se intentó disminuir el valor de las cotas a 2 mm, 4 mm, 8 mm y 16 mm (introduciendo una categoría adicional), pero aún así la mayoría de los datos siguieron perteneciendo a la categoría *leve* con un 72 % de la totalidad. En la Figura B.1 se observa la distribución de clases en la base de datos para cada una de las variables objetivo de tipo categóricas. Se observa que, a diferencia de las cotas fijas, las cotas percentiles y la lluvia binaria tienen un mayor balanceo de clases. La sobrerrepresentación es negativa porque genera modelos sesgados a predecir siempre la clase mayoritaria, lo que si bien puede resultar en una alta exactitud, no permite que el modelo detecte correctamente fenómenos de las demás categorías. Aún si se consiguiera un buen balanceo de clases con cotas fijas, es preferible la clasificación con cotas percentiles, ya que esta última facilita la detección de fenómenos anómalos en comparación con el registro histórico al utilizar intervalos específicos para cada estación y día del año.

1. En general, el estudio realizado permitió evaluar satisfactoriamente el desempeño de diferentes modelos de machine learning en el pronóstico de precipitación en Guatemala con un día de anticipación.
2. Se evaluaron modelos de clasificación con cotas percentiles para la predicción cualitativa de la intensidad de precipitación en comparación con la Climatología, obteniendo una exactitud general entre el 26 % y el 65 %. También se evaluaron modelos de clasificación de lluvia binaria en el pronóstico de presencia o ausencia de precipitación, presentando entre el 55 % y el 75 % de exactitud general.
3. Se evaluó el rendimiento de modelos de regresión en el pronóstico de la anomalía absoluta de precipitación respecto del promedio de la Climatología, presentando valores de MAE de 5.14 mm a 6.50 mm, y valores de MdAE entre 2.06 mm y 3.50 mm al ser evaluados con datos de todo el año.
4. En términos de MAE promedio, los mejores algoritmos de regresión son *regresión lineal múltiple* y *regresión polinomial*. Los algoritmos de clasificación con mayor exactitud promedio son *bosque aleatorio* y *XGBoost* con 100 árboles de decisión. No se observó un cambio relevante al implementar las diferentes combinaciones de variables predictoras y número de días que componen el rango de observación. Los resultados de la evaluación son susceptibles a la época del año utilizada durante el entrenamiento.
5. En general, se presentó un mejor rendimiento en las evaluaciones realizadas con datos de la época seca. Esto se debe a la diferencia en la distribución de la precipitación, la cual suele presentar más variabilidad durante la época lluviosa, mostrando fenómenos extremos y cambios abruptos cuyos patrones son más difíciles de generalizar. Cabe destacar que esta tendencia también se observó en el modelo WRF. En el caso de NextGen, si bien el comportamiento fue inverso, la diferencia de rendimiento entre épocas es mucho menor a la presentada por WRF y los modelos desarrollados.
6. El 96 % de los modelos de regresión y el 23 % de los modelos de lluvia binaria tuvieron una evaluación general superior o igual a la del modelo WRF en términos de MAE y exactitud, respectivamente, lo que evidencia el potencial del machine learning como complemento a los modelos que actualmente utiliza INSIVUMEH.

Al realizar pronósticos utilizando los modelos entrenados en este estudio, se recomienda:

1. Obtener los valores de Climatología y percentiles a partir de los registros históricos de la estación meteorológica en la ubicación a pronosticar. En caso de no contar con registros históricos en dicha ubicación, se sugiere calcular los valores a partir de la base de datos de ENACTS. Si bien estas cantidades no son indispensables para la ejecución del modelo, permiten realizar una mejor interpretación de los resultados obtenidos.
2. Contrastar entre sí los resultados de los tres tipos de pronóstico y sus respectivas variables objetivo, con el fin de realizar un análisis más íntegro que ayude a determinar el nivel de confiabilidad de los resultados.

Para futuras investigaciones, se proponen las siguientes sugerencias:

1. Dado que los modelos de clasificación con mayor exactitud fueron *bosque aleatorio* y *XGBoost* con 100 árboles de decisión, se recomienda variar los demás hiperparámetros de estos algoritmos para determinar su configuración óptima. Esto permitirá mejorar aún más el rendimiento.
2. Se sugiere construir modelos que tomen en cuenta el comportamiento de las variables meteorológicas, no solo en el punto exacto a pronosticar, sino también en los alrededores del lugar. Esto es especialmente importante si en el futuro se desarrollan modelos de pronóstico con varios días de anticipación, ya que en estos casos las regiones cercanas tienen más tiempo para interactuar con el punto de interés.
3. Para hacer factible la implementación de otros algoritmos como KNN y SVM, se recomienda reducir el tamaño de la base de datos. También se sugiere explorar el uso de algoritmos de deep learning.
4. Se recomienda que la selección de variables predictoras se realice en base a un criterio definido. Por ejemplo, calculando el coeficiente de correlación de Pearson entre la precipitación de la fecha objetivo y cada una de las variables disponibles. De esta forma, se asegura que los predictores estén fuertemente relacionados con la variable a predecir, lo que facilita la identificación de patrones por parte del modelo.

5. Explorar la posibilidad de entrenar modelos utilizando únicamente la precipitación como variable predictora, o una combinación de precipitación y valores de temperatura. Dado que estas variables son las más fáciles de medir, sería relativamente accesible para cualquier usuario obtener pronósticos de precipitación.

- Acharya, N., Ehsan, M., Admasu, A., Teshome, A., y Chen, K. (2021). On the next generation (NextGen) seasonal prediction system to enhance climate services over Ethiopia. *Climate Services* 24(5):100272. <https://doi.org/10.1016/j.cliser.2021.100272>
- Alfaro, E. (2002). Some Characteristics of the Precipitation Annual Cycle in Central America And their Relationships with its Surrounding Tropical Oceans. *Universidad de Costa Rica. Centro de Investigaciones Geofísicas*.
- Alpaydin, E. (2014). *Introduction to Machine Learning* (3.^a ed.). Massachusetts Institute of Technology.
- Anderson, D., Sweeney, D., y Williams, T. (2012). *Estadística para negocios y economía* (11.^a ed.). Cengage Learning.
- Argueta, P. (2024). *Climatología* [Accedido el 13 de febrero de 2025]. <https://insivumeh.gob.gt/?p=92918>
- Bochenek, B., y Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere* 13(2), 180. <https://doi.org/10.3390/atmos13020180>
- Bonnin, R. (2017). *Machine learning for developers*. Packt Publishing.
- Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover how they work and implement from scratch* (1.^a ed.). Machine Learning Mastery.
- Collins, S., James, R., Ray, P., Chen, K., Lassman, A., y Brownlee, J. (2013). Grids in Numerical Weather and Climate Models. *Climate Change and Regional/Local Responses* (pp. 125-145). Springer. <https://doi.org/10.5772/55922>
- CONRED. (2024a). *Memoria de Labores Cuarto Año de Gobierno 2023-2024*. https://conred.gob.gt/wp-content/uploads/DCS_20241001_MEMORIA_LABORES23_compressed-1.pdf
- CONRED. (2024b). *Protocolo específico de lluvias en la república de Guatemala 2024*. https://conred.gob.gt/wp-content/uploads/Protocolo_Operativo_Lluvias_2024234-1.pdf
- Davis, R., Hayden, B., Gay, D., Phillips, W., y Jones, G. (1997). The North Atlantic Subtropical Anticyclone. *Journal of Climate*. [https://doi.org/10.1175/1520-0442\(1997\)010<0728:TNASA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<0728:TNASA>2.0.CO;2)
- Dinku, T., Thomson, M., Cousin, R., del Corral, J., Ceccato, P., Hansen, J., y Connor, S. (2017). Enhancing National Climate Services (ENACTS) for development in Africa. *Climate and Development*, 10(7), 664–672. <https://doi.org/10.1080/17565529.2017.1405784>
- Doshi, R., Kant, K., Kumar, R., y Lakhawani, K. (2022). *Machine learning: master supervised and unsupervised learning algorithms with real examples*. BPC Publications.

- Elbasheer, M. E. E. E., Corzo, G. A., Solomatine, D., y Varouchakis, E. (2023). Machine Learning and Committee Models for Improving ECMWF Subseasonal to Seasonal (S2S) Precipitation Forecast. *Hydrol. Earth Syst. Sci.* <https://doi.org/10.5194/hess-2022-348>
- Grazzini, F., Dorrington, J., Grams, C., Craig, G., Magnusson, L., y Vitart, F. (2024). Improving forecast of precipitation extremes over northern and central Italy using machine learning. *Quarterly Journal of the Royal Meteorological Society*, 2024;1–15. <https://doi.org/10.1002/qj.4755>
- Harrington, P. (2012). *Machine learning in action*. Manning Publications Co.
- INSIVUMEH. (2023). *Boletín de Aplicaciones Climáticas. Septiembre 2023* [Accedido el 13 de febrero de 2025]. <https://insivumeh.gob.gt/wp-content/uploads/2023/09/Aplicaciones-Climaticas-BAC-202309.pdf>
- INSIVUMEH. (2025). *Pronóstico Estacional Flexible de Precipitación NextGen (3 meses)* [Accedido el 13 de febrero de 2025]. http://dl.insivumeh.gob.gt/maproom/Cuarto_Mapas_Climatologia/Pronostico/NextGen_Precip_3m.html
- Instituto Nacional de Estadística (INE). (2019). *Resultados del Censo 2018. Documentación y manuales*. <https://censo2018.ine.gob.gt/documentacion>
- IRI. (2020). *THE NEXT GENERATION OF CLIMATE FORECASTS* [Accedido el 2 de febrero de 2025]. https://iri.columbia.edu/wp-content/uploads/2020/06/Fact-Sheet_Next-Gen_small.pdf
- Ixcamparij, C. (2025). *Pronóstico WRF-BMJ GFS (Extendido A 15 Días)* [Accedido el 31 de enero de 2025]. <https://insivumeh.gob.gt/?p=5313>
- Knieval, J. (2006). Numerical Weather Prediction (NWP) and the WRF Model. *ATEC Forecasters' Conference*.
- Ko, C-M, Jeong, YY, Lee, Y-M y Kim, B-S. (2020). The Development of a Quantitative Precipitation Forecast Correction Technique Based on Machine Learning for Hydrological Applications. *Atmosphere* 2020, 11(1), 111. <https://doi.org/10.3390/atmos11010111>
- Kroese, D., Botev, Z., Taimre, T., y Vaisman, R. (2023). *Data science and machine learning: mathematical and statistical methods*.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., y Battaglia, P. (2023). GraphCast: learning skillful medium-range global weather forecasting. *Arxiv*. <https://doi.org/10.48550/arXiv.2212.12794>
- Maldonado, T. (2016). Inter-annual variability of rainfall in Central America: Connection with global and regional climate modulators. *Uppsala University*.
- Marroquín, P. (2023). *Análisis de predictibilidad en el sistema dinámico caótico del péndulo magnético mediante la aplicación de modelos de aprendizaje automático*. Universidad del Valle de Guatemala.
- Marsland, S. (2015). *Machine Learning: An Algorithm Perspective* (2.^a ed.). Taylor & Francis Group, LLC.
- Martinez, C., Goddard, L., Kushnir, Y., y Ting, M. (2019). Seasonal climatology and dynamical mechanisms of rainfall in the Caribbean. *Climate Dynamics*. <https://doi.org/10.1007/s00382-019-04616-4>
- McIveen, R. (1992). *Fundamentals of weather and climate*. Chapman; Hall.
- Mesa, O., y Peñaranda, V. (2015). Complejidad de la estructura espacio-temporal de la precipitación. *Rev. Acad. Colomb. Cienc. Ex. Fis. Nat.* <https://doi.org/10.18257/raccefy.196>
- Müller, A., y Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists* (1.^a ed.). O'Reilly Media, Inc.
- Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J., y Grover, A. (2023). ClimaX: A foundation model for weather and climate. *Arxiv*. <https://doi.org/10.48550/arXiv.2301.10343>
- Nunno, F., Granata, F., Pham, Q., y De Marinis, G. (2022). Precipitation forecasting in northern Bangladesh using a hybrid machine learning model. *Sustainability* 2022, 14(5), 2663. <https://doi.org/10.3390/su14052663>
- Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Thorsten, K., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., y Anandkumar,

- A. (2022). FourCastNet: a global data-driven high-resolution weather model using adaptive fourier neural operators. *Arxiv*. <https://doi.org/10.48550/arXiv.2202.11214>
- Portillo, L., Solórzano, I., Cardona, D., Gómez, V., Icó, L., Contreras, M., y Rodríguez, J. (2024). *Estado del Clima en Guatemala 2023*. INSIVUMEH.
- Pu, Z., y Kalnay, E. (2018). Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation. In: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H., Schaake, J. (eds) Handbook of Hydrometeorological Ensemble Forecasting. *Springer*. https://doi.org/10.1007/978-3-642-40457-3_11-1
- Raschka, S., Liu, Y., y Mirjalili, V. (2022). *Machine learning with pytorch and scikit-learn: develop machine learning and deep learning models with python*. Packt Publishing.
- Rebala, G., Ravi, A., y Churiwala, S. (2019). *An Introduction to Machine Learning* (1.^a ed.). Springer Nature.
- The COMET Program. (2015). *NWP Essentials: Structure and Dynamics. Media Gallery* [Accedido el 12 de febrero de 2025]. https://lyqx.sinaapp.com/wfiles/mstructure_dynamics/media_gallery.php.htm
- Turner, J. (2021). *Colombia, Guatemala adopt ENACTS Climate Data Initiative* [Accedido el 13 de febrero de 2025]. <https://iri.columbia.edu/news/colombia-guatemala-adopt-enacts-climate-data-initiative/>
- Vaidya, S., y Singh, S. (1999). Applying the Betts-Miller-Janjic Scheme of Convection in Prediction of the Indian Monsoon. *Weather and Forecasting* 15(3):349-356. [https://doi.org/10.1175/1520-0434\(2000\)015<0349:ATBMJS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0349:ATBMJS>2.0.CO;2)
- Wallace, J., y Hobbs, P. (2006). *Atmospheric science: an introduction survey* (2.^a ed.). Elsevier Inc.
- Wang, C., y Enfield, D. (2006). Influences of the Atlantic Warm Pool on Western Hemisphere Summer Rainfall and Atlantic Hurricanes. *Journal of Climate*. <https://doi.org/10.1175/JCLI3770.1>
- Waqas, M., Wannasingha, U., Chueasa, B., y Wangwongchai, A. (2024). Artificial intelligence and numerical weather prediction models: A technical survey. *Natural Hazards Research*. <https://doi.org/10.1016/j.nhres.2024.11.004>
- Watt, J., Borhani, R., y Katsaggelos, A. (2020). *Machine learning refined: foundations, algorithms and applications* (2.^a ed.). Cambridge University Press.
- Wolf, F., Voigt, A., y Donner, R. (2021). Aclimate network perspective on the intertropical convergence zone. *Earth Syst. Dynam.*, 12, 353–366, 2021. <https://doi.org/10.5194/esd-12-353-2021>
- WorldRiskIndex. (2024). *The WorldRiskReport 2024 - Focus: Multiple Crises*. <https://reliefweb.int/report/world/worldriskreport-2024-focus-multiple-crises>

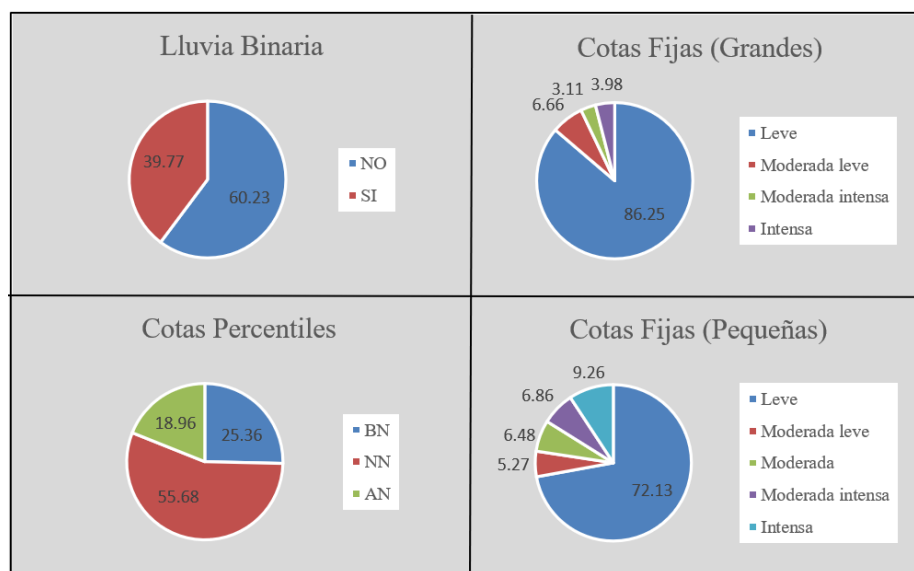
Código de programación

El código *Python* desarrollado a lo largo del proyecto se encuentra en el siguiente repositorio de *GitHub*:

https://github.com/isaac0911/ML_precipitacion.git

Debido a las políticas de propiedad de INSIVUMEH, no es posible compartir los datos meteorológicos utilizados. Por esta razón, el repositorio incluye únicamente el código fuente, excluyendo cualquier tipo de información meteorológica.

Distribución de clases

**Figura B.1:** Distribución de clases según variable categórica

Uso de gráficos interactivos

En el siguiente enlace se presenta un video con una breve explicación sobre el funcionamiento de los gráficos interactivos utilizados en este trabajo. Se describe el uso de herramientas tales como el *zoom*, la descarga de figuras y la exploración de datos mediante el cursor.

<https://youtu.be/9CGtmIofYC4>

Resultados de modelos de regresión

			MAE (mm) Evaluación general									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	5.390	5.382	5.964	5.822	5.722	5.700	5.334	5.331	5.335	5.348
		lluviosa	5.264	5.186	6.477	6.233	6.047	5.990	5.173	5.178	5.211	5.261
		seca	5.553	5.609	6.192	6.104	6.051	6.015	5.631	5.686	5.735	5.785
Intermedia	5	general	5.389	5.378	5.979	5.827	5.751	5.703	5.333	5.330	5.338	5.354
		lluviosa	5.267	5.175	6.451	6.196	6.024	5.977	5.157	5.160	5.190	5.258
		seca	5.537	5.603	6.215	6.112	6.054	6.031	5.629	5.684	5.724	5.796
Básica	3	general	5.387	5.372	5.954	5.824	5.731	5.704	5.334	5.333	5.339	5.352
		lluviosa	5.276	5.140	6.498	6.193	6.034	5.974	5.166	5.178	5.196	5.236
		seca	5.528	5.577	6.133	6.057	6.010	5.986	5.618	5.657	5.708	5.764
	5	general	5.389	5.371	5.952	5.826	5.729	5.697	5.333	5.330	5.334	5.344
		lluviosa	5.269	5.150	6.393	6.160	6.005	5.941	5.152	5.153	5.201	5.286
		seca	5.536	5.592	6.192	6.104	6.035	6.030	5.624	5.669	5.729	5.783
	7	general	5.389	5.369	5.959	5.824	5.720	5.691	5.330	5.328	5.336	5.350
		lluviosa	5.261	5.163	6.375	6.187	6.016	5.943	5.154	5.164	5.194	5.256
		seca	5.535	5.601	6.239	6.181	6.081	6.064	5.631	5.692	5.779	5.829

Cuadro D.1: MAE para modelos de regresión evaluados con datos de todo el año

			MAE (mm)									
			Evaluación época lluviosa									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	8.666	8.576	9.263	9.076	8.930	8.906	8.523	8.500	8.479	8.479
		lluviosa	8.501	8.495	9.306	9.106	8.973	8.930	8.480	8.470	8.470	8.488
		seca	8.893	8.988	9.811	9.701	9.641	9.599	9.075	9.180	9.264	9.334
Intermedia	5	general	8.668	8.578	9.304	9.061	8.965	8.896	8.512	8.491	8.482	8.491
		lluviosa	8.502	8.490	9.360	9.112	8.990	8.941	8.474	8.461	8.466	8.494
		seca	8.863	8.981	9.854	9.708	9.635	9.613	9.072	9.172	9.237	9.356
Básica	3	general	8.672	8.589	9.272	9.075	8.938	8.908	8.510	8.499	8.490	8.493
		lluviosa	8.515	8.492	9.331	9.126	8.988	8.945	8.484	8.474	8.477	8.492
		seca	8.850	8.950	9.695	9.603	9.551	9.525	9.053	9.128	9.217	9.300
	5	general	8.668	8.577	9.261	9.070	8.930	8.889	8.502	8.496	8.478	8.482
		lluviosa	8.504	8.484	9.305	9.114	8.961	8.928	8.473	8.469	8.479	8.493
		seca	8.862	8.972	9.794	9.696	9.603	9.609	9.062	9.145	9.254	9.341
	7	general	8.666	8.570	9.267	9.067	8.915	8.879	8.507	8.500	8.483	8.498
		lluviosa	8.497	8.480	9.302	9.096	8.955	8.904	8.477	8.474	8.486	8.508
		seca	8.858	8.982	9.875	9.853	9.698	9.679	9.079	9.198	9.353	9.425

Cuadro D.2: MAE para modelos de regresión evaluados con datos de la época lluviosa

			MAE (mm)									
			Evaluación época seca									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	2.102	2.176	2.653	2.556	2.503	2.482	2.133	2.150	2.181	2.206
		lluviosa	2.015	1.867	3.639	3.350	3.111	3.041	1.855	1.874	1.940	2.024
		seca	2.201	2.218	2.561	2.495	2.449	2.419	2.175	2.179	2.195	2.223
Intermedia	5	general	2.100	2.167	2.643	2.582	2.525	2.499	2.144	2.157	2.183	2.207
		lluviosa	2.020	1.849	3.532	3.271	3.047	3.002	1.828	1.847	1.903	2.011
		seca	2.199	2.212	2.563	2.504	2.459	2.437	2.173	2.183	2.198	2.224
Básica	3	general	2.091	2.144	2.624	2.562	2.513	2.488	2.148	2.157	2.177	2.200
		lluviosa	2.026	1.775	3.656	3.249	3.070	2.992	1.837	1.872	1.903	1.969
		seca	2.194	2.192	2.557	2.499	2.456	2.434	2.170	2.173	2.186	2.216
	5	general	2.098	2.153	2.633	2.570	2.515	2.493	2.153	2.153	2.179	2.196
		lluviosa	2.023	1.805	3.471	3.196	3.038	2.944	1.819	1.825	1.912	2.068
		seca	2.198	2.200	2.577	2.500	2.454	2.438	2.174	2.179	2.193	2.212
	7	general	2.100	2.156	2.640	2.570	2.515	2.491	2.143	2.145	2.179	2.190
		lluviosa	2.014	1.835	3.438	3.268	3.067	2.972	1.819	1.843	1.890	1.994
		seca	2.200	2.208	2.591	2.497	2.451	2.436	2.170	2.174	2.193	2.220

Cuadro D.3: MAE para modelos de regresión evaluados con datos de la época seca

			MdAE (mm)									
			Evaluación general									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	2.547	2.466	2.635	2.734	2.787	2.807	2.430	2.417	2.398	2.392
		lluviosa	2.340	2.157	3.501	3.500	3.426	3.386	2.146	2.142	2.162	2.199
		seca	2.773	2.755	3.082	3.120	3.159	3.139	2.789	2.800	2.793	2.773
Intermedia	5	general	2.545	2.471	2.653	2.746	2.802	2.800	2.432	2.414	2.402	2.381
		lluviosa	2.345	2.145	3.470	3.438	3.380	3.371	2.131	2.129	2.120	2.188
		seca	2.748	2.758	3.075	3.137	3.157	3.152	2.776	2.805	2.796	2.800
Básica	3	general	2.543	2.476	2.618	2.726	2.779	2.794	2.436	2.408	2.400	2.395
		lluviosa	2.359	2.062	3.447	3.405	3.366	3.358	2.143	2.149	2.146	2.155
		seca	2.740	2.745	3.033	3.086	3.121	3.121	2.771	2.786	2.785	2.776
	5	general	2.542	2.463	2.605	2.729	2.760	2.777	2.435	2.414	2.405	2.387
		lluviosa	2.345	2.091	3.406	3.372	3.342	3.308	2.108	2.099	2.093	2.192
		seca	2.745	2.755	3.053	3.106	3.124	3.136	2.785	2.799	2.800	2.777
	7	general	2.539	2.451	2.616	2.742	2.760	2.766	2.425	2.401	2.391	2.381
		lluviosa	2.329	2.115	3.397	3.432	3.377	3.309	2.111	2.094	2.100	2.151
		seca	2.745	2.765	3.101	3.131	3.146	3.160	2.768	2.789	2.800	2.788

Cuadro D.4: MdAE para modelos de regresión evaluados con datos de todo el año

			MdAE (mm)									
			Evaluación época lluviosa									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	5.924	5.779	6.243	6.267	6.223	6.223	5.728	5.698	5.663	5.631
		lluviosa	5.727	5.691	6.275	6.301	6.268	6.296	5.681	5.663	5.638	5.624
		seca	6.237	6.324	6.834	6.936	6.997	6.972	6.374	6.503	6.512	6.513
Intermedia	5	general	5.928	5.784	6.270	6.248	6.256	6.217	5.708	5.693	5.673	5.648
		lluviosa	5.730	5.698	6.344	6.345	6.279	6.265	5.676	5.667	5.644	5.634
		seca	6.185	6.309	6.877	6.960	6.946	6.961	6.375	6.488	6.494	6.570
Básica	3	general	5.942	5.792	6.236	6.240	6.206	6.235	5.725	5.704	5.673	5.639
		lluviosa	5.745	5.698	6.329	6.313	6.304	6.315	5.692	5.667	5.656	5.627
		seca	6.169	6.259	6.749	6.807	6.854	6.868	6.367	6.424	6.489	6.515
	5	general	5.928	5.781	6.240	6.254	6.181	6.195	5.704	5.709	5.672	5.646
		lluviosa	5.730	5.688	6.323	6.305	6.264	6.262	5.687	5.668	5.655	5.640
		seca	6.183	6.308	6.812	6.901	6.947	6.950	6.377	6.453	6.528	6.544
	7	general	5.928	5.776	6.221	6.212	6.188	6.199	5.715	5.698	5.668	5.635
		lluviosa	5.720	5.685	6.285	6.268	6.257	6.251	5.667	5.662	5.645	5.631
		seca	6.183	6.320	6.873	7.007	6.992	7.017	6.384	6.503	6.589	6.588

Cuadro D.5: MdAE para modelos de regresión evaluados con datos de la época lluviosa

			MdAE (mm)									
			Evaluación época seca									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	0.465	0.642	0.567	0.669	0.740	0.757	0.500	0.526	0.581	0.598
		lluviosa	0.104	0.000	1.000	1.089	1.115	1.091	0.000	0.000	0.000	0.000
		seca	0.595	0.622	0.560	0.666	0.732	0.738	0.541	0.542	0.545	0.542
Intermedia	5	general	0.463	0.626	0.565	0.695	0.756	0.767	0.525	0.539	0.578	0.598
		lluviosa	0.115	0.000	0.914	1.012	1.000	1.032	0.000	0.000	0.000	0.000
		seca	0.584	0.610	0.555	0.671	0.733	0.751	0.550	0.547	0.545	0.544
Básica	3	general	0.439	0.574	0.556	0.665	0.730	0.746	0.527	0.541	0.562	0.586
		lluviosa	0.135	0.000	0.813	0.891	0.900	0.900	0.000	0.000	0.000	0.000
		seca	0.572	0.568	0.536	0.659	0.713	0.718	0.540	0.544	0.539	0.538
	5	general	0.463	0.596	0.551	0.679	0.733	0.750	0.533	0.531	0.566	0.583
		lluviosa	0.121	0.000	0.800	0.835	0.956	0.917	0.000	0.000	0.000	0.000
		seca	0.581	0.583	0.547	0.648	0.719	0.725	0.546	0.545	0.542	0.542
	7	general	0.471	0.611	0.559	0.673	0.739	0.747	0.513	0.521	0.575	0.581
		lluviosa	0.100	0.000	0.799	0.904	0.965	0.915	0.000	0.000	0.000	0.000
		seca	0.592	0.599	0.558	0.660	0.726	0.740	0.550	0.544	0.544	0.541

Cuadro D.6: MdAE para modelos de regresión evaluados con datos de la época seca

Resultados de modelos de clasificación con cotas percentiles

			Exactitud (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	57.63	61.02	62.18	63.48	64.04	63.92	64.19	64.53	64.70
		lluviosa	27.50	38.31	39.59	41.00	41.34	37.03	38.14	40.32	41.34
		seca	56.30	57.63	57.81	58.10	58.08	57.52	57.85	57.97	58.02
Intermedia	5	general	57.48	61.76	62.42	63.30	63.95	63.92	64.23	64.51	64.65
		lluviosa	27.50	38.12	38.77	39.75	40.30	37.18	38.23	40.22	41.21
		seca	56.28	57.57	57.86	58.08	58.09	57.56	57.83	58.01	58.05
Básica	3	general	57.10	61.96	62.97	63.85	64.25	63.92	64.15	64.47	64.59
		lluviosa	26.94	38.66	39.49	40.18	40.74	36.23	38.15	39.64	40.60
		seca	56.25	57.93	58.12	58.28	58.36	57.60	57.85	57.98	58.10
	5	general	57.44	61.73	62.81	63.82	64.18	63.94	64.23	64.50	64.62
		lluviosa	27.46	38.25	39.42	40.16	40.69	36.19	37.88	39.31	40.35
		seca	56.27	57.71	57.94	58.06	58.23	57.57	57.81	58.02	58.07
	7	general	57.76	61.62	62.39	63.53	64.28	63.89	64.23	64.52	64.60
		lluviosa	27.81	38.34	39.37	40.26	40.69	36.96	37.82	39.76	40.73
		seca	56.33	57.65	57.86	58.00	58.16	57.56	57.88	58.03	58.08

Cuadro E.1: Exactitud para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año

			Exactitud (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	31.55	38.96	40.32	42.28	43.13	42.70	43.07	43.71	44.03
		lluviosa	41.46	39.92	41.08	42.97	43.85	43.26	43.64	44.02	44.05
		seca	30.22	31.39	31.18	31.33	31.24	30.03	30.37	30.45	30.56
Intermedia	5	general	31.51	39.80	40.59	42.06	43.00	42.70	43.15	43.62	43.89
		lluviosa	41.45	39.93	41.05	42.54	43.75	43.30	43.70	43.89	44.04
		seca	30.23	31.48	31.36	31.27	31.28	30.03	30.27	30.53	30.68
Básica	3	general	31.37	39.87	41.32	42.58	43.35	42.70	43.00	43.55	43.81
		lluviosa	41.23	40.27	41.24	42.58	43.41	43.19	43.52	43.66	43.80
		seca	30.22	31.71	31.63	31.50	31.57	30.13	30.37	30.50	30.68
	5	general	31.49	39.68	41.17	42.76	43.45	42.75	43.17	43.64	43.86
		lluviosa	41.42	39.94	41.24	42.80	43.54	43.11	43.45	43.65	43.65
		seca	30.21	31.47	31.42	31.26	31.42	30.05	30.28	30.56	30.68
	7	general	31.68	39.76	40.86	42.42	43.85	42.65	43.14	43.62	43.72
		lluviosa	41.56	40.26	41.16	43.06	44.06	43.30	43.62	43.87	43.91
		seca	30.24	31.69	31.40	31.21	31.33	30.13	30.37	30.58	30.76

Cuadro E.2: Exactitud para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa

			Exactitud (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	78.50	83.16	84.12	84.75	85.03	85.22	85.38	85.41	85.44
		lluviosa	11.78	36.70	38.09	39.03	38.83	30.78	32.62	36.61	38.61
		seca	82.52	83.96	84.53	84.96	85.00	85.12	85.43	85.58	85.57
Intermedia	5	general	78.92	83.79	84.33	84.61	84.97	85.22	85.38	85.47	85.49
		lluviosa	11.85	36.29	36.49	36.94	36.84	31.04	32.75	36.53	38.36
		seca	82.39	83.75	84.45	84.98	84.99	85.18	85.49	85.58	85.52
Básica	3	general	78.57	84.12	84.69	85.19	85.22	85.21	85.38	85.46	85.45
		lluviosa	11.33	37.03	37.73	37.78	38.06	29.25	32.75	35.60	37.39
		seca	82.24	84.25	84.71	85.14	85.24	85.17	85.42	85.56	85.61
	5	general	79.06	83.86	84.52	84.96	84.99	85.22	85.36	85.42	85.45
		lluviosa	11.80	36.55	37.60	37.51	37.82	29.25	32.28	34.96	37.05
		seca	82.38	84.04	84.56	84.96	85.13	85.19	85.43	85.57	85.56
	7	general	79.31	83.55	84.00	84.71	84.78	85.21	85.39	85.49	85.55
		lluviosa	12.28	36.41	37.57	37.44	37.31	30.60	32.01	35.65	37.54
		seca	82.45	83.69	84.41	84.89	85.09	85.09	85.49	85.57	85.51

Cuadro E.3: Exactitud para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca

			Precisión de la categoría <i>NN</i> (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	62.82	76.12	77.28	79.88	81.38	87.60	87.47	87.22	86.58
		lluviosa	78.12	64.33	67.33	71.97	75.32	80.01	78.52	78.06	76.59
		seca	57.00	61.16	60.64	60.51	60.29	58.68	59.50	59.97	60.06
Intermedia	5	general	62.50	77.40	78.17	79.72	81.79	87.64	87.46	87.43	86.68
		lluviosa	79.17	64.32	66.88	71.08	74.80	81.28	80.18	79.51	77.82
		seca	57.02	61.16	60.70	60.48	60.29	58.70	59.38	59.88	60.05
Básica	3	general	61.56	78.10	79.70	81.46	82.70	87.65	87.51	87.51	86.86
		lluviosa	77.37	64.73	66.99	70.22	72.90	79.50	80.30	78.58	77.17
		seca	57.01	61.71	61.43	61.25	61.30	58.46	59.43	59.85	60.06
	5	general	62.37	77.38	79.22	81.01	82.42	87.82	87.57	87.43	86.86
		lluviosa	79.26	64.58	67.49	71.50	74.86	80.08	79.29	78.02	76.92
		seca	57.04	61.33	61.03	60.77	60.71	58.61	59.43	59.92	60.09
	7	general	62.89	76.87	77.75	80.17	82.02	87.75	87.74	87.47	86.87
		lluviosa	79.83	64.30	67.29	72.10	75.57	80.64	80.19	79.26	78.24
		seca	57.08	61.58	60.89	60.64	60.51	58.69	59.53	60.10	60.24

Cuadro E.4: Precisión de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año

			Precisión de la categoría <i>NN</i> (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	30.02	33.52	34.88	37.74	39.36	47.92	46.81	47.55	46.16
		lluviosa	37.09	35.25	36.49	39.51	42.10	48.98	48.25	47.22	44.74
		seca	30.02	29.96	29.94	30.01	30.00	29.81	29.70	29.65	29.70
Intermedia	5	general	30.01	34.77	35.18	37.01	39.77	47.88	47.31	48.46	46.85
		lluviosa	36.81	34.71	36.65	38.61	41.87	49.02	49.10	46.69	44.85
		seca	30.02	29.99	29.96	29.92	29.98	29.80	29.73	29.72	29.73
Básica	3	general	29.99	34.96	36.81	38.90	41.07	48.31	48.36	48.57	47.60
		lluviosa	35.64	35.34	37.11	38.85	41.06	50.01	49.51	46.75	44.66
		seca	30.01	30.08	29.96	29.93	29.98	29.91	29.74	29.73	29.78
	5	general	30.03	34.14	36.43	38.69	41.17	48.70	48.08	48.45	47.11
		lluviosa	37.85	35.35	36.68	39.70	41.91	49.65	48.84	46.71	44.74
		seca	30.02	29.89	29.91	29.89	29.94	29.81	29.69	29.70	29.73
	7	general	30.11	34.25	35.35	38.01	41.88	47.78	48.12	48.35	46.55
		lluviosa	37.23	35.14	36.37	40.27	43.04	48.90	48.78	46.91	45.07
		seca	30.02	30.07	29.93	29.87	29.95	29.83	29.72	29.67	29.74

Cuadro E.5: Precisión de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa

			Precisión de la categoría <i>NN</i> (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	89.37	89.76	89.67	89.68	89.72	90.11	90.49	90.88	91.13
		lluviosa	90.14	85.42	87.05	88.76	89.73	89.32	87.55	87.62	87.37
		seca	85.10	90.11	89.96	90.01	89.73	89.97	90.94	91.59	91.71
Intermedia	5	general	89.79	89.97	89.68	89.36	89.56	90.14	90.49	90.95	91.18
		lluviosa	91.58	86.12	87.17	89.10	90.24	90.85	89.47	89.99	89.62
		seca	84.90	89.78	89.85	89.96	89.81	90.07	91.04	91.57	91.67
Básica	3	general	89.16	90.82	90.83	90.87	90.74	90.19	90.52	91.17	91.31
		lluviosa	89.15	86.15	86.65	88.37	89.21	88.31	89.35	88.83	88.67
		seca	84.56	90.98	90.90	90.96	91.05	90.05	91.05	91.62	91.79
	5	general	89.77	90.23	90.23	90.01	89.89	90.15	90.53	90.93	91.21
		lluviosa	91.68	86.31	87.64	89.05	90.30	89.33	88.42	88.01	88.48
		seca	84.92	90.26	90.25	90.28	90.24	89.91	90.97	91.64	91.73
	7	general	90.10	89.83	89.36	89.45	89.31	90.22	90.62	91.03	91.32
		lluviosa	92.93	86.00	87.31	89.24	90.52	89.76	89.87	89.73	90.20
		seca	85.10	90.09	89.82	89.82	89.80	89.85	90.98	91.58	91.72

Cuadro E.6: Precisión de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca

			Sensibilidad de la categoría <i>NN</i> (%)								
			Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	87.78	77.58	78.15	77.55	76.88	73.26	73.47	73.94	74.39
		lluviosa	2.05	33.24	33.53	32.79	31.40	21.43	23.80	28.42	30.97
		seca	97.95	91.34	92.64	93.35	93.72	96.34	94.77	93.90	93.74
Intermedia	5	general	88.06	77.63	77.70	77.26	76.64	73.19	73.53	73.87	74.45
		lluviosa	2.01	32.39	31.74	30.47	29.44	21.31	23.55	27.65	30.19
		seca	97.88	91.29	92.52	93.27	93.75	96.34	95.03	94.11	93.77
Básica	3	general	88.86	77.40	77.31	76.94	76.53	73.26	73.63	73.92	74.45
		lluviosa	1.28	33.38	33.39	32.06	31.51	19.96	23.62	27.05	29.37
		seca	97.91	90.79	91.59	92.11	92.23	96.95	94.93	94.19	93.92
	5	general	88.23	77.43	77.40	77.00	76.52	73.11	73.54	73.86	74.32
		lluviosa	1.96	32.77	32.73	31.39	30.53	19.73	23.23	26.45	29.07
		seca	97.85	90.95	92.05	92.73	93.11	96.56	94.89	94.02	93.77
	7	general	87.98	77.61	77.86	77.29	76.84	73.08	73.42	73.89	74.33
		lluviosa	2.64	32.70	32.66	31.16	29.90	21.00	22.72	26.88	29.19
		seca	97.84	90.70	92.23	92.89	93.42	96.37	94.85	93.75	93.53

Cuadro E.7: Sensibilidad de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año

			Sensibilidad de la categoría <i>NN</i> (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	83.11	30.81	29.69	25.70	22.91	8.89	10.13	12.65	14.93
		lluviosa	1.94	28.48	26.36	22.83	19.75	11.26	12.51	15.13	17.01
		seca	97.60	80.12	83.12	84.67	85.48	94.66	90.34	88.17	88.00
Intermedia	5	general	83.08	29.53	27.48	24.58	21.60	8.78	10.38	12.64	15.22
		lluviosa	1.96	27.58	25.98	21.96	19.56	10.93	12.34	14.62	17.05
		seca	97.41	79.71	82.66	84.24	85.57	94.69	91.37	88.99	88.15
Básica	3	general	83.96	29.32	27.36	24.74	22.88	9.11	10.79	13.10	15.46
		lluviosa	1.29	28.57	27.31	24.17	22.36	10.74	12.32	14.58	16.51
		seca	97.07	79.15	80.36	81.49	81.72	96.94	91.11	89.34	88.61
	5	general	83.16	29.10	27.08	23.99	21.80	8.48	10.48	12.53	14.81
		lluviosa	1.92	28.46	26.16	23.05	20.28	10.60	12.28	14.25	16.62
		seca	97.31	78.99	81.27	82.89	83.66	95.16	90.75	88.76	88.09
	7	general	82.78	29.98	28.28	24.58	22.43	8.60	10.16	12.65	14.74
		lluviosa	2.42	28.36	25.81	22.66	19.94	10.57	12.11	14.47	16.57
		seca	97.41	78.27	81.44	82.82	84.17	94.59	90.45	87.56	87.25

Cuadro E.8: Sensibilidad de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa

			Sensibilidad de la categoría <i>NN</i> (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	88.98	94.78	95.96	96.62	96.73	96.93	96.76	96.48	96.26
		lluviosa	1.55	34.99	36.16	36.45	35.68	25.17	27.95	33.31	36.10
		seca	97.84	95.47	96.14	96.54	96.75	96.96	96.40	96.01	95.86
Intermedia	5	general	89.25	95.32	96.17	96.63	96.87	96.88	96.75	96.39	96.23
		lluviosa	1.56	34.16	33.86	33.60	33.08	25.12	27.67	32.45	35.02
		seca	97.86	95.55	96.14	96.60	96.76	96.94	96.37	95.99	95.84
Básica	3	general	89.30	95.08	95.68	96.14	96.26	96.85	96.74	96.28	96.15
		lluviosa	0.88	35.15	35.63	34.95	34.88	23.35	27.77	31.64	34.10
		seca	97.98	95.07	95.72	96.02	96.10	96.96	96.34	95.97	95.87
	5	general	89.43	95.20	95.90	96.50	96.65	96.87	96.73	96.41	96.21
		lluviosa	1.48	34.36	35.14	34.46	34.30	23.09	27.26	30.94	33.64
		seca	97.85	95.35	96.02	96.34	96.58	97.08	96.41	95.96	95.86
	7	general	89.54	95.12	96.08	96.68	96.85	96.79	96.69	96.41	96.24
		lluviosa	2.20	34.29	35.18	34.28	33.57	24.84	26.62	31.44	33.83
		seca	97.75	95.27	96.20	96.60	96.82	97.03	96.46	96.03	95.83

Cuadro E.9: Sensibilidad de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca

			Especificidad de la categoría <i>NN</i> (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	34.31	69.22	70.94	75.30	77.76	86.88	86.69	86.30	85.41
		lluviosa	99.27	76.70	79.43	83.85	86.99	93.23	91.76	89.90	88.03
		seca	6.55	26.64	23.95	22.97	21.94	14.22	18.42	20.72	21.17
Intermedia	5	general	33.17	71.34	72.55	75.14	78.43	86.95	86.67	86.56	85.53
		lluviosa	99.33	77.27	80.12	84.33	87.46	93.80	92.64	90.99	89.12
		seca	6.67	26.69	24.26	22.94	21.90	14.28	17.80	20.27	21.11
Básica	3	general	29.84	72.56	75.10	77.86	79.76	86.94	86.71	86.66	85.76
		lluviosa	99.53	77.00	79.19	82.81	85.19	93.49	92.67	90.67	89.01
		seca	6.62	28.75	27.26	26.29	26.38	12.88	18.06	20.09	21.02
	5	general	32.69	71.38	74.33	77.17	79.36	87.18	86.80	86.57	85.78
		lluviosa	99.35	77.27	80.06	84.18	87.04	93.80	92.33	90.57	88.97
		seca	6.80	27.48	25.65	24.29	23.79	13.74	18.08	20.48	21.23
	7	general	34.35	70.48	71.83	75.82	78.70	87.10	87.02	86.62	85.79
		lluviosa	99.16	77.04	79.92	84.75	87.78	93.62	92.90	91.10	89.73
		seca	6.97	28.44	25.10	23.76	22.91	14.23	18.45	21.28	21.93

Cuadro E.10: Especificidad de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de todo el año

			Especificidad de la categoría <i>NN</i> (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	17.06	73.84	76.28	81.85	84.89	95.86	95.07	94.03	92.55
		lluviosa	98.59	77.61	80.36	85.04	88.37	94.98	94.26	92.76	91.01
		seca	2.63	19.85	16.75	15.48	14.62	4.59	8.47	10.46	10.84
Intermedia	5	general	17.07	76.29	78.33	82.10	85.99	95.91	95.05	94.24	92.61
		lluviosa	98.56	77.80	80.78	85.05	88.37	95.14	94.52	92.86	91.02
		seca	2.82	20.36	17.28	15.53	14.46	4.53	7.55	9.92	10.81
Básica	3	general	16.12	76.65	79.90	83.37	85.95	95.83	95.07	94.06	92.72
		lluviosa	99.00	77.62	80.19	83.71	86.26	95.40	94.62	92.89	91.24
		seca	3.10	21.26	19.59	18.32	18.31	2.78	7.86	9.62	10.55
	5	general	17.04	75.97	79.77	83.73	86.67	96.18	95.16	94.30	92.88
		lluviosa	98.65	77.72	80.67	85.02	87.97	95.40	94.49	93.04	91.21
		seca	2.88	20.68	18.47	16.75	16.21	4.08	8.01	10.07	10.89
	7	general	17.74	75.37	77.86	82.84	86.67	95.98	95.31	94.21	92.76
		lluviosa	98.25	77.60	80.67	85.62	88.71	95.27	94.56	92.99	91.36
		seca	2.81	22.11	18.39	16.75	15.75	4.78	8.46	11.17	11.78

Cuadro E.11: Especificidad de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de la época lluviosa

			Especificidad de la categoría <i>NN</i> (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	95.28	97.86	98.36	98.62	98.79	99.11	98.99	98.75	98.56
		lluviosa	81.12	91.06	93.72	95.31	95.80	96.31	96.71	96.89	96.64
		seca	99.29	97.93	98.39	98.65	98.81	99.05	98.68	98.34	98.21
Intermedia	5	general	95.19	97.99	98.43	98.68	98.79	99.21	99.06	98.78	98.63
		lluviosa	81.10	91.46	93.87	95.25	95.84	96.17	96.60	96.71	96.61
		seca	99.30	98.11	98.39	98.62	98.75	99.02	98.66	98.38	98.24
Básica	3	general	95.13	97.79	98.11	98.26	98.35	99.05	98.95	98.67	98.54
		lluviosa	81.45	90.99	93.20	94.50	95.05	95.75	96.12	96.43	96.26
		seca	99.37	97.76	97.98	98.22	98.31	99.01	98.71	98.36	98.26
	5	general	95.22	97.99	98.32	98.56	98.61	99.15	98.97	98.77	98.61
		lluviosa	81.39	91.93	94.17	95.07	95.66	96.27	96.68	96.88	96.48
		seca	99.31	97.97	98.35	98.51	98.65	99.02	98.67	98.41	98.30
	7	general	95.15	98.07	98.33	98.69	98.78	99.12	98.95	98.78	98.63
		lluviosa	81.30	91.82	94.15	95.48	96.01	96.58	96.90	97.14	96.87
		seca	99.28	97.99	98.37	98.71	98.92	99.12	98.71	98.38	98.21

Cuadro E.12: Especificidad de la categoría *NN* para modelos de clasificación con cotas percentiles, evaluados con datos de la época seca

ANEXO F

Resultados de modelos de clasificación de lluvia binaria

			Exactitud (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	70.33	71.25	72.85	74.13	74.45	74.34	74.67	74.90	74.90
		lluviosa	56.98	67.79	69.57	71.38	71.91	72.77	72.97	72.83	72.47
		seca	61.12	64.99	65.72	66.07	66.18	67.07	68.28	68.26	67.82
Intermedia	5	general	70.07	71.46	73.02	74.02	74.33	74.30	74.60	74.84	74.84
		lluviosa	57.09	68.05	69.87	71.36	72.03	72.70	72.93	72.56	72.13
		seca	61.12	65.04	66.01	66.41	66.68	67.07	67.78	67.88	67.80
Básica	3	general	69.40	71.67	73.12	74.08	74.39	74.11	74.46	74.70	74.74
		lluviosa	55.79	68.30	69.72	70.77	71.33	72.39	72.81	72.24	71.60
		seca	61.02	65.27	65.94	66.43	66.50	65.99	66.84	66.90	66.59
	5	general	69.98	71.70	73.15	74.15	74.50	74.39	74.55	74.79	74.76
		lluviosa	56.93	68.60	69.82	71.53	72.16	72.61	72.95	72.86	72.41
		seca	61.09	65.48	66.20	66.83	67.03	67.10	68.02	68.20	68.04
	7	general	70.39	71.62	72.98	74.15	74.53	74.37	74.61	74.73	74.76
		lluviosa	57.79	68.31	70.13	71.80	72.50	72.93	73.30	73.05	72.76
		seca	61.16	65.47	66.29	67.29	67.38	67.75	68.39	68.38	68.16

Cuadro F.1: Exactitud para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año

			Exactitud (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	53.04	61.96	64.47	66.47	66.92	66.88	67.37	67.77	67.74
		lluviosa	64.40	62.67	65.19	66.80	67.30	67.14	67.55	67.68	67.72
		seca	40.21	48.95	49.96	50.39	50.42	52.30	54.62	54.51	53.72
Intermedia	5	general	53.03	62.12	64.68	66.33	66.84	66.87	67.27	67.65	67.67
		lluviosa	64.43	62.64	65.04	66.71	67.32	67.12	67.46	67.53	67.62
		seca	40.20	49.13	50.48	51.03	51.40	52.23	53.57	53.76	53.55
Básica	3	general	52.59	62.51	64.90	66.38	66.83	66.47	67.09	67.45	67.56
		lluviosa	64.20	63.03	65.16	66.58	67.14	67.03	67.44	67.47	67.16
		seca	40.22	49.63	50.46	51.10	51.02	50.20	51.84	51.99	51.41
	5	general	52.91	62.62	65.04	66.62	67.05	66.98	67.23	67.56	67.51
		lluviosa	64.40	62.98	64.91	66.72	67.48	67.00	67.38	67.48	67.47
		seca	40.21	49.99	50.87	51.89	52.15	52.36	54.15	54.48	54.28
	7	general	53.07	62.39	64.76	66.53	67.21	66.96	67.29	67.46	67.48
		lluviosa	64.60	62.75	65.16	66.70	67.48	67.19	67.54	67.63	67.61
		seca	40.23	50.13	51.22	52.82	52.75	53.65	54.83	54.79	54.43

Cuadro F.2: Exactitud para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa

			Exactitud (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	75.23	80.56	81.25	81.82	82.00	81.82	81.98	82.05	82.08
		lluviosa	36.11	72.93	73.96	75.97	76.54	78.43	78.41	78.00	77.23
		seca	81.15	81.09	81.54	81.80	82.00	81.89	81.98	82.05	81.96
Intermedia	5	general	75.43	80.83	81.38	81.74	81.84	81.77	81.95	82.05	82.04
		lluviosa	36.03	73.48	74.72	76.03	76.75	78.30	78.42	77.60	76.65
		seca	81.16	80.99	81.58	81.85	82.03	81.96	82.04	82.05	82.10
Básica	3	general	74.77	80.85	81.37	81.80	81.97	81.79	81.86	81.98	81.95
		lluviosa	35.48	73.58	74.30	74.98	75.53	77.76	78.20	77.04	76.06
		seca	81.18	80.96	81.48	81.81	82.03	81.84	81.89	81.87	81.83
	5	general	75.49	80.81	81.29	81.71	81.97	81.83	81.91	82.05	82.03
		lluviosa	35.88	74.25	74.76	76.36	76.85	78.25	78.54	78.26	77.37
		seca	81.18	81.01	81.59	81.83	81.97	81.90	81.93	81.98	81.85
	7	general	75.92	80.88	81.23	81.80	81.87	81.81	81.94	82.02	82.07
		lluviosa	36.14	73.90	75.11	76.93	77.55	78.68	79.08	78.48	77.93
		seca	81.21	80.86	81.41	81.81	82.08	81.91	82.01	82.03	81.93

Cuadro F.3: Exactitud para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca

			Precisión (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	66.97	67.11	68.04	68.72	68.73	67.95	68.30	68.62	68.66
		lluviosa	47.92	59.35	60.67	62.24	62.61	63.55	64.07	64.02	63.61
		seca	69.72	68.16	70.71	72.50	73.94	74.55	73.98	73.66	72.90
Intermedia	5	general	66.89	67.31	68.18	68.50	68.68	67.92	68.22	68.61	68.57
		lluviosa	48.00	59.71	61.10	62.27	62.77	63.45	64.03	63.60	63.07
		seca	69.74	67.86	71.07	72.75	73.99	74.58	74.02	73.55	73.27
Básica	3	general	66.07	67.34	68.06	68.27	68.34	67.87	68.14	68.39	68.39
		lluviosa	47.20	60.00	60.95	61.49	61.92	62.97	63.88	63.13	62.35
		seca	68.74	67.94	69.77	71.64	72.29	74.11	73.67	73.25	72.45
	5	general	66.88	67.50	68.16	68.50	68.64	67.96	68.08	68.43	68.39
		lluviosa	47.89	60.52	61.13	62.53	62.94	63.34	64.10	64.05	63.56
		seca	69.47	67.94	70.27	72.40	73.37	74.43	74.04	73.69	72.84
	7	general	67.39	67.43	68.04	68.70	68.68	67.89	68.07	68.32	68.38
		lluviosa	48.44	60.06	61.51	62.87	63.43	63.82	64.58	64.34	64.07
		seca	70.08	67.61	70.29	72.98	73.75	74.29	74.00	73.36	72.78

Cuadro F.4: Precisión para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año

			Precisión (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	74.81	69.51	69.68	69.83	69.67	68.75	69.17	69.51	69.62
		lluviosa	65.96	69.28	69.43	69.28	69.13	68.37	69.17	69.52	69.71
		seca	74.37	73.25	74.92	76.27	77.48	77.77	77.40	77.29	77.19
Intermedia	5	general	74.86	69.41	69.73	69.60	69.71	68.75	69.11	69.50	69.56
		lluviosa	65.97	69.21	69.23	69.23	69.15	68.38	69.15	69.43	69.63
		seca	74.26	72.75	75.06	76.17	77.16	77.86	77.72	77.35	77.27
Básica	3	general	74.57	69.50	69.61	69.35	69.24	68.54	68.96	69.25	69.35
		lluviosa	65.57	69.50	69.49	69.26	69.25	68.34	69.15	69.32	69.18
		seca	73.93	73.38	74.22	75.62	75.86	78.11	78.05	77.94	77.46
	5	general	74.88	69.60	69.77	69.64	69.53	68.76	68.99	69.31	69.35
		lluviosa	65.91	69.50	69.22	69.22	69.25	68.32	69.06	69.34	69.46
		seca	74.16	72.62	73.98	75.72	76.44	77.65	77.85	77.65	77.06
	7	general	74.99	69.48	69.72	69.77	69.64	68.69	68.92	69.23	69.30
		lluviosa	66.18	69.20	69.33	69.11	69.18	68.48	69.13	69.45	69.62
		seca	74.79	72.35	74.17	76.07	76.46	77.17	77.37	77.00	76.75

Cuadro F.5: Precisión para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa

			Precisión (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	38.14	48.48	53.26	57.58	58.99	58.40	58.83	59.19	58.98
		lluviosa	22.05	32.15	34.74	38.62	39.80	43.48	43.46	42.45	40.67
		seca	54.72	52.27	56.18	58.81	61.07	60.41	59.25	59.15	57.85
Intermedia	5	general	38.52	50.26	54.18	57.10	57.92	57.98	58.59	59.23	58.66
		lluviosa	22.04	33.02	35.85	38.58	40.17	43.23	43.50	41.73	39.76
		seca	54.82	51.58	56.54	59.37	61.28	60.77	59.58	59.12	59.04
Básica	3	general	37.38	50.39	54.02	57.31	58.80	59.02	58.55	58.94	58.35
		lluviosa	21.91	33.18	35.16	37.07	38.18	42.21	42.98	40.57	38.66
		seca	55.36	51.23	55.03	57.80	59.78	59.82	58.58	57.90	57.24
	5	general	38.54	50.11	53.50	56.82	59.05	58.44	58.25	59.12	58.63
		lluviosa	21.99	34.06	35.66	39.08	40.34	43.14	43.71	43.09	40.89
		seca	55.10	51.67	56.32	58.80	60.39	60.37	58.72	58.42	57.18
	7	general	39.25	50.61	53.08	57.66	58.27	58.30	58.67	58.83	58.94
		lluviosa	22.06	33.53	36.16	40.14	41.56	44.05	45.01	43.51	42.13
		seca	55.33	50.52	55.03	58.99	61.53	60.35	59.38	58.74	57.74

Cuadro F.6: Precisión para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca

			Sensibilidad (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	50.53	54.73	60.19	64.51	65.92	67.48	68.09	68.29	68.21
		lluviosa	90.67	61.11	67.36	71.82	73.42	74.41	73.41	72.79	72.35
		seca	4.46	22.95	24.00	24.04	23.49	26.47	31.57	31.77	30.73
Intermedia	5	general	49.42	55.31	60.65	64.54	65.50	67.43	68.00	68.04	68.17
		lluviosa	90.70	61.21	67.34	71.53	73.37	74.44	73.32	72.94	72.67
		seca	4.45	23.44	24.91	25.25	25.39	26.45	29.60	30.41	30.33
Básica	3	general	47.88	56.23	61.44	65.41	66.66	66.65	67.56	68.01	68.19
		lluviosa	91.24	61.53	67.05	71.52	73.04	74.68	73.24	73.08	72.69
		seca	4.17	24.45	25.79	26.20	25.95	22.63	26.22	26.81	26.20
	5	general	49.01	56.00	61.32	65.16	66.39	67.73	68.15	68.30	68.26
		lluviosa	90.73	61.24	66.82	71.45	73.44	74.39	73.16	72.82	72.26
		seca	4.37	25.46	26.46	27.21	27.22	26.69	30.50	31.53	31.68
	7	general	49.93	55.78	60.84	64.64	66.42	67.81	68.42	68.29	68.31
		lluviosa	90.43	61.37	67.04	71.57	73.33	74.17	73.22	72.74	72.18
		seca	4.58	25.77	26.83	28.57	28.30	29.29	31.99	32.55	32.20

Cuadro F.7: Sensibilidad para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año

			Sensibilidad (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	33.71	66.10	73.05	78.48	80.22	82.94	83.09	83.20	82.80
		lluviosa	85.01	68.79	75.86	81.07	83.00	84.98	83.62	82.91	82.47
		seca	1.75	24.59	25.96	26.11	25.42	29.61	35.28	35.10	33.34
Intermedia	5	general	33.65	66.85	73.54	78.70	79.90	82.91	83.00	82.90	82.77
		lluviosa	85.07	68.87	75.99	80.92	83.00	84.91	83.43	82.78	82.42
		seca	1.73	25.43	27.16	27.71	27.90	29.38	32.58	33.31	32.87
Básica	3	general	32.81	67.78	74.50	79.60	81.26	82.36	82.93	83.10	83.08
		lluviosa	85.94	69.28	75.60	80.44	82.15	84.77	83.34	82.91	82.40
		seca	1.80	26.24	27.74	28.27	27.90	24.54	28.34	28.77	27.74
	5	general	33.32	67.82	74.45	79.47	81.06	83.23	83.23	83.21	82.95
		lluviosa	85.22	69.12	75.60	81.00	83.14	84.74	83.46	82.89	82.49
		seca	1.77	27.81	28.97	30.11	30.19	29.82	33.81	34.74	34.76
	7	general	33.64	67.45	73.79	78.82	81.18	83.36	83.65	83.17	82.98
		lluviosa	84.83	69.23	76.04	81.27	83.37	84.79	83.71	83.00	82.41
		seca	1.79	28.42	29.71	32.09	31.61	33.20	35.79	36.00	35.37

Cuadro F.8: Sensibilidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa

			Sensibilidad (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	46.58	18.82	19.54	20.38	20.71	18.64	20.71	21.18	22.09
		lluviosa	91.77	36.84	40.50	42.58	43.15	41.01	41.12	40.82	40.38
		seca	10.96	17.76	17.80	17.51	17.40	16.56	19.84	21.25	22.47
Intermedia	5	general	46.81	18.84	19.95	19.80	20.02	18.52	20.63	21.09	22.03
		lluviosa	91.82	37.00	40.02	41.84	42.94	41.35	41.36	41.84	41.89
		seca	10.95	17.13	17.79	17.46	17.46	17.20	20.18	21.23	22.29
Básica	3	general	46.40	19.75	20.18	20.56	20.53	17.02	19.01	20.34	21.11
		lluviosa	91.97	37.03	40.04	43.34	44.27	42.80	41.33	42.03	41.99
		seca	10.37	18.81	19.62	19.64	19.78	16.61	19.50	20.62	21.35
	5	general	46.37	18.66	19.83	19.95	20.00	18.76	20.51	21.20	21.85
		lluviosa	91.78	36.36	39.06	41.27	42.79	41.69	40.63	41.01	39.93
		seca	10.82	18.01	18.53	18.04	17.85	16.80	20.02	21.39	21.94
	7	general	46.28	18.92	19.92	19.84	19.78	18.64	20.32	21.28	21.96
		lluviosa	91.79	36.53	38.62	40.91	41.58	40.62	40.09	40.33	39.85
		seca	11.31	17.37	17.73	17.42	17.84	16.94	20.02	21.63	22.20

Cuadro F.9: Sensibilidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca

			Especificidad (%) Evaluación general								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	83.46	82.20	81.25	80.51	80.11	78.89	79.03	79.28	79.34
		lluviosa	34.63	72.23	71.03	71.09	70.91	71.69	72.68	72.85	72.54
		seca	98.71	92.89	93.41	93.95	94.51	94.01	92.63	92.46	92.42
Intermedia	5	general	83.77	82.18	81.22	80.31	80.18	78.87	78.98	79.35	79.27
		lluviosa	34.80	72.59	71.55	71.25	71.13	71.55	72.67	72.31	71.76
		seca	98.72	92.63	93.27	93.72	94.08	94.02	93.11	92.74	92.66
Básica	3	general	83.69	81.91	80.87	79.83	79.52	79.06	79.04	79.14	79.09
		lluviosa	32.28	72.79	71.49	70.28	70.19	70.86	72.52	71.69	70.88
		seca	98.74	92.35	92.59	93.12	93.40	94.76	93.78	93.50	93.39
	5	general	83.90	82.11	81.00	80.12	79.88	78.81	78.80	79.09	79.07
		lluviosa	34.50	73.49	71.82	71.59	71.30	71.43	72.81	72.89	72.51
		seca	98.72	92.03	92.57	93.12	93.44	93.92	92.91	92.53	92.16
	7	general	83.97	82.12	81.04	80.46	79.91	78.73	78.71	78.99	79.04
		lluviosa	36.14	72.92	72.17	71.96	71.96	72.10	73.35	73.25	73.15
		seca	98.70	91.81	92.48	92.98	93.32	93.27	92.54	92.16	92.01

Cuadro F.10: Especificidad para modelos de clasificación de lluvia binaria, evaluados con datos de todo el año

			Especificidad (%) Evaluación época lluviosa								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	82.63	55.62	51.34	48.09	46.55	42.30	43.32	44.15	44.69
		lluviosa	32.84	53.31	48.86	44.96	43.27	39.83	42.94	44.35	45.13
		seca	99.07	86.25	86.69	87.56	88.69	87.05	84.23	84.21	84.92
Intermedia	5	general	82.70	54.89	51.13	47.38	46.85	42.31	43.21	44.30	44.57
		lluviosa	32.82	53.11	48.29	44.94	43.30	39.90	43.02	44.20	44.97
		seca	99.08	85.42	86.19	86.73	87.36	87.21	85.70	85.07	85.20
Básica	3	general	82.87	54.46	50.20	46.14	44.74	42.13	42.85	43.51	43.80
		lluviosa	30.91	53.45	49.18	45.36	44.18	39.88	43.09	43.83	43.82
		seca	99.03	85.43	85.25	86.04	86.40	89.48	87.80	87.53	87.64
	5	general	82.88	54.66	50.62	46.96	45.61	42.10	42.73	43.60	43.88
		lluviosa	32.54	53.57	48.54	44.87	43.50	39.84	42.77	43.88	44.47
		seca	99.06	83.95	84.40	85.22	85.76	86.86	85.27	84.69	84.16
	7	general	82.82	54.64	50.94	47.71	45.84	41.84	42.26	43.41	43.74
		lluviosa	33.62	52.83	48.50	44.39	43.15	40.25	42.78	44.11	44.96
		seca	99.08	83.37	84.16	84.55	85.10	84.97	83.98	83.54	83.60

Cuadro F.11: Especificidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época lluviosa

			Especificidad (%) Evaluación época seca								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	82.04	95.24	95.92	96.43	96.58	96.84	96.55	96.53	96.35
		lluviosa	22.88	81.52	81.91	83.91	84.48	87.33	87.28	86.84	85.99
		seca	97.84	96.14	96.70	97.08	97.36	97.42	96.76	96.51	96.11
Intermedia	5	general	82.24	95.57	95.99	96.46	96.54	96.81	96.53	96.55	96.31
		lluviosa	22.76	82.16	82.97	84.16	84.79	87.09	87.23	86.11	84.91
		seca	97.85	96.18	96.75	97.16	97.38	97.36	96.74	96.51	96.32
Básica	3	general	81.52	95.38	95.92	96.36	96.58	97.19	96.80	96.63	96.42
		lluviosa	22.05	82.27	82.44	82.51	82.96	86.07	86.96	85.36	84.16
		seca	98.01	95.74	96.19	96.59	96.84	97.35	96.72	96.43	96.21
	5	general	82.42	95.58	95.90	96.39	96.70	96.83	96.50	96.51	96.33
		lluviosa	22.58	83.26	83.24	84.71	84.95	86.94	87.56	87.12	86.27
		seca	97.90	96.00	96.58	96.99	97.22	97.38	96.65	96.38	96.09
	7	general	82.97	95.61	95.81	96.54	96.63	96.83	96.60	96.46	96.36
		lluviosa	22.91	82.78	83.79	85.49	86.10	87.74	88.35	87.55	86.99
		seca	97.83	95.96	96.56	97.12	97.35	97.35	96.74	96.39	96.14

Cuadro F.12: Especificidad para modelos de clasificación de lluvia binaria, evaluados con datos de la época seca

ANEXO G

Tiempos de entrenamiento

			Tiempo de entrenamiento (s)									
			Regresión									
Comb.	No. Días	Entren.	LR	PR	RF				XGB			
					10	20	50	100	10	20	50	100
Avanzada	5	general	0.61	92.55	406.04	795.75	1938.22	4817.19	0.86	1.17	2.86	3.08
		lluviosa	0.33	28.45	162.38	316.30	766.55	1659.95	0.50	0.57	1.01	1.54
		seca	0.29	30.21	203.06	447.01	981.01	2610.80	0.48	0.57	0.87	1.53
Intermedia	5	general	0.55	45.50	438.70	603.25	1497.76	3876.21	1.54	2.03	2.69	4.96
		lluviosa	0.25	18.42	182.11	253.70	632.61	1362.68	0.78	1.09	1.44	2.49
		seca	0.25	15.87	215.68	304.09	762.95	1602.61	0.89	0.90	1.42	2.14
Básica	3	general	0.44	6.42	249.83	501.13	1931.49	2721.77	0.90	1.17	1.71	2.72
		lluviosa	0.19	3.11	118.42	297.69	742.96	1241.25	0.44	0.56	0.10	1.56
		seca	0.16	2.81	121.61	380.25	674.06	1310.12	0.40	0.50	0.90	1.37
	5	general	0.52	13.99	330.20	675.67	1146.48	2442.99	0.64	0.84	1.42	2.06
		lluviosa	0.29	6.64	143.12	289.98	457.36	1356.78	0.33	0.40	0.65	1.03
		seca	0.33	6.59	177.68	214.29	505.31	1525.28	0.32	0.37	0.63	0.96
	7	general	0.85	41.81	668.60	1312.72	4465.27	7812.54	1.35	1.04	1.53	2.57
		lluviosa	0.31	17.49	298.00	594.99	1490.79	3974.24	0.83	0.50	0.84	1.27
		seca	0.28	15.26	332.20	671.78	1658.94	3311.00	0.72	0.60	0.77	1.17

Cuadro G.1: Tiempos de entrenamiento para los modelos de regresión

			Tiempo de entrenamiento (s) Cotas percentiles								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	25.28	36.43	67.93	179.13	344.61	1.60	2.51	5.25	9.01
		lluviosa	14.45	15.80	31.75	79.98	159.55	0.84	1.32	2.71	4.51
		seca	13.48	14.05	24.21	62.64	122.38	0.78	1.30	2.54	4.73
Intermedia	5	general	12.21	27.23	53.35	133.83	261.34	1.48	2.46	4.70	8.18
		lluviosa	5.68	12.65	23.90	60.86	118.18	0.82	1.36	2.34	3.92
		seca	7.17	9.57	17.85	52.99	91.32	0.79	1.16	2.27	4.07
Básica	3	general	10.29	18.63	36.25	87.01	171.63	1.12	1.75	3.48	6.22
		lluviosa	4.34	8.29	16.54	41.88	79.18	0.50	0.79	1.62	2.78
		seca	6.22	5.62	10.78	27.33	53.80	0.47	0.94	1.64	2.83
	5	general	11.33	22.43	43.79	111.41	221.24	1.23	2.13	4.01	7.21
		lluviosa	5.82	10.21	20.22	50.89	100.54	0.67	0.98	2.12	3.48
		seca	6.40	7.51	14.90	36.47	73.72	0.63	1.12	2.06	3.74
	7	general	13.90	27.78	55.12	135.23	272.78	1.47	2.24	4.73	8.29
		lluviosa	5.97	12.40	25.32	61.73	124.75	0.87	1.19	2.29	4.16
		seca	7.92	9.52	19.18	46.69	93.51	0.84	1.31	2.28	4.07

Cuadro G.2: Tiempos de entrenamiento para los modelos de clasificación con cotas percentiles

			Tiempo de entrenamiento (s)								
			Lluvia binaria								
Comb.	No. Días	Entren.	LogReg	RF				XGB			
				10	20	50	100	10	20	50	100
Avanzada	5	general	2.53	53.05	97.84	239.64	516.98	0.97	1.36	1.93	3.15
		lluviosa	1.31	23.80	46.40	114.12	212.47	0.55	0.68	1.01	1.72
		seca	0.94	25.59	42.55	102.86	199.58	0.52	0.67	1.13	1.65
Intermedia	5	general	1.28	45.32	89.56	213.53	344.24	0.84	1.11	1.74	2.79
		lluviosa	0.81	19.51	39.73	72.84	149.64	0.44	0.63	0.96	1.55
		seca	0.46	17.89	35.92	67.21	159.93	0.48	0.62	0.87	1.34
Básica	3	general	0.90	21.59	43.77	108.23	215.91	0.49	0.71	1.19	2.01
		lluviosa	0.38	9.71	19.12	52.27	93.99	0.25	0.33	0.60	0.95
		seca	0.47	8.52	16.69	42.32	82.76	0.26	0.35	0.57	0.89
	5	general	0.92	29.82	56.55	138.21	282.42	1.31	1.56	2.52	3.91
		lluviosa	0.48	12.98	24.95	61.78	125.99	0.65	0.99	1.28	2.06
		seca	0.52	11.85	23.30	55.76	112.79	0.73	0.78	1.19	2.10
	7	general	1.52	35.01	76.51	198.59	455.03	1.62	2.30	3.04	4.95
		lluviosa	0.75	16.85	30.89	87.88	203.13	0.84	1.10	1.56	3.20
		seca	0.64	14.67	29.51	110.25	189.91	0.83	1.21	1.51	2.80

Cuadro G.3: Tiempos de entrenamiento para los modelos de clasificación de lluvia binaria