

# **PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR**

## **FACULTAD DE INGENIERÍA**



### **TEMA:**

Estimación de Radiación UV con Modelos de Machine Learning a partir de Sensores  
Hidrometeorológicos en el Distrito Metropolitano de Quito DMQ

### **AUTOR:**

Ing. Alexis Sebastián Cachaguay Alquina

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE  
INFORMACIÓN MENCIÓN DATA SCIENCE

### **TUTOR:**

Juan Francisco Chafra Altamirano, Ph.D.

**Quito, marzo 2026**

## RESUMEN

En el presente trabajo de maestría se desarrolló un modelo predictivo para la estimación del Índice Ultravioleta en Quito, basado en datos meteorológicos de 4 estaciones y técnicas de aprendizaje automático. Tras comparar algoritmos y evaluar el impacto de la heterogeneidad espacial, se seleccionó LightGBM entrenado exclusivamente con datos de la estación Belisario como el modelo final, alcanzando un  $R^2$  de 0.942, RMSE de 0.809 y un desempeño sobresaliente en rangos críticos (MAE extremo = 1.737; recall IUV >8 = 0.807).

Se implementó una aplicación web interactiva con Streamlit y Folium que permite la edición manual de sensores (RS, TMP, HUM), visualización geográfica satelital de las 33 estaciones (9 de la Secretaría de Ambiente y 24 de EPMAAPS) y actualización en tiempo real de predicciones, con indicadores de riesgo y recomendaciones sanitarias según estándares OMS.

Los resultados demuestran que el enfoque por estación individual supera al modelo global en precisión y utilidad práctica, destacando la importancia de considerar la variabilidad microclimática en regiones andinas tropicales. La solución propuesta ofrece una herramienta viable para monitoreo ambiental y prevención de riesgos asociados a la radiación ultravioleta, con potencial de escalabilidad e integración con sistemas de alerta sanitaria.

**Palabras clave:** Índice Ultravioleta, predicción horaria, LightGBM, Quito, aprendizaje automático, despliegue interactivo.

# ÍNDICE

|                                                                                |           |
|--------------------------------------------------------------------------------|-----------|
| <b>CAPÍTULO I: INTRODUCCIÓN.....</b>                                           | <b>7</b>  |
| <b>1. Introducción.....</b>                                                    | <b>7</b>  |
| 1.1    Antecedentes .....                                                      | 9         |
| 1.2    Planteamiento del problema.....                                         | 11        |
| 1.3    Justificación .....                                                     | 11        |
| 1.4    Objetivos .....                                                         | 12        |
| 1.4.1    Objetivo General .....                                                | 12        |
| 1.4.2    Objetivos Específicos.....                                            | 12        |
| 1.5    Alcance .....                                                           | 13        |
| <b>CAPITULO II. FUNDAMENTACIÓN TEORICA.....</b>                                | <b>15</b> |
| <b>2. Marco Teórico .....</b>                                                  | <b>15</b> |
| 2.1    Fundamentos Físicos de la Radiación en Altura .....                     | 15        |
| 2.1.1 Menor espesor atmosférico y Dispersión de Rayleigh.....                  | 15        |
| 2.1.2 Posición Ecuatorial y Geometría Solar.....                               | 16        |
| 2.1.3 Dinámica de la Composición Atmosférica y Dispersión de Mie.....          | 17        |
| 2.2 Justificación Física de las Variables Hidrometeorológicas.....             | 18        |
| 2.3    Modelado de Transferencia Radiativa vs. Enfoques Basados en Datos ..... | 18        |
| 2.4    Estado del Arte en Estimación de UV con ML .....                        | 19        |
| 2.5 Algoritmos de Aprendizaje Supervisado para Regresión.....                  | 20        |
| <b>CAPÍTULO III: METODOLOGÍA .....</b>                                         | <b>22</b> |
| <b>3. Metodología.....</b>                                                     | <b>22</b> |
| 3.1 Comprensión de Datos y Negocio.....                                        | 22        |
| 3.1.1 Comprensión de Negocio .....                                             | 22        |
| 3.1.1.1 Objetivos del negocio.....                                             | 22        |
| 3.1.1.2 Criterios de rendimiento .....                                         | 23        |
| 3.1.1.3 Evaluación de la situación.....                                        | 23        |
| 3.1.1.4 Traducir objetivos comerciales en objetivos de ML.....                 | 23        |
| 3.1.1.5 Plan de proyecto .....                                                 | 24        |
| 3.1.2 Comprensión de Datos .....                                               | 24        |
| 3.1.2.1 Recopilación de los datos .....                                        | 24        |
| 3.1.2.2 Descripción de los datos .....                                         | 27        |
| 3.2 Preparación de los Datos .....                                             | 28        |

|                                                                           |    |
|---------------------------------------------------------------------------|----|
| 3.2.1 Limpieza .....                                                      | 28 |
| 3.2.2 Análisis de Correlación .....                                       | 34 |
| 3.2.3 Ingeniería de datos .....                                           | 37 |
| 3.3 Modelado.....                                                         | 40 |
| 3.3.1 Preparación de los datos para el entrenamiento.....                 | 41 |
| 3.3.2 Selección y configuración de modelos.....                           | 41 |
| 3.3.3 Métricas de evaluación.....                                         | 42 |
| 3.3.4 Resultados y comparación gráfica .....                              | 43 |
| 3.3.5 Discusión preliminar y selección del modelo.....                    | 43 |
| 3.4 Evaluación.....                                                       | 44 |
| 3.4.1 Evaluación inicial del modelo multi-estación.....                   | 44 |
| 3.4.2 Análisis por estación individual y selección del modelo final ..... | 46 |
| 3.4.3 Selección del modelo final para despliegue .....                    | 52 |
| 3.5 Despliegue .....                                                      | 53 |
| 3.5.1 Objetivos y enfoque del despliegue .....                            | 53 |
| 3.5.2 Tecnologías utilizadas .....                                        | 55 |
| 3.5.3. Arquitectura y flujo de la aplicación .....                        | 56 |
| 3.5.4. Características técnicas destacadas .....                          | 57 |
| 3.5.5. Limitaciones y consideraciones .....                               | 57 |
| 3.5.6. Conclusiones del despliegue.....                                   | 57 |
| CAPITULO IV: CONCLUSIONES Y RECOMENDACIONES .....                         | 58 |
| 4.1 Conclusiones .....                                                    | 58 |
| 4.2 Recomendaciones.....                                                  | 59 |
| 5. Referencias.....                                                       | 61 |

## ÍNDICE DE FIGURAS

|                                                                                                          |    |
|----------------------------------------------------------------------------------------------------------|----|
| <b>Figura 1:</b> Disponibilidad de datos en la estación Belisario .....                                  | 25 |
| <b>Figura 2:</b> Disponibilidad de datos en la estación Centro .....                                     | 26 |
| <b>Figura 3:</b> Disponibilidad de datos en la estación Cotocollao.....                                  | 26 |
| <b>Figura 4:</b> Disponibilidad de datos en la estación Guamaní.....                                     | 27 |
| <b>Figura 5:</b> Series de tiempo de variables meteorológicas curadas de la estación Belisario .....     | 30 |
| <b>Figura 6:</b> Series de tiempo de variables meteorológicas curadas de la estación Centro.....         | 31 |
| <b>Figura 7:</b> Series de tiempo de variables meteorológicas curadas de la estación Cotocollao .....    | 32 |
| <b>Figura 8:</b> Series de tiempo de variables meteorológicas curadas de la estación Guamaní .....       | 33 |
| <b>Figura 9:</b> Análisis de correlación de Pearson entre las variables de estudio y la de objetivo..... | 34 |
| <b>Figura 10:</b> Comparativa de modelos de ML para estimación de IUV .....                              | 43 |
| <b>Figura 11:</b> IUV observado vs estimado con el modelo LightGBM multi-estación .....                  | 45 |
| <b>Figura 12:</b> Estimación IUV para la estación Belisario.....                                         | 46 |
| <b>Figura 13:</b> Estimación IUV para la estación Centro .....                                           | 46 |
| <b>Figura 14:</b> Estimación IUV para la estación Cotocollao .....                                       | 47 |
| <b>Figura 15:</b> Estimación IUV para la estación Guamaní .....                                          | 47 |
| <b>Figura 16:</b> IUV observado vs estimado con el modelo LightGBM para la estación Belisario.....       | 48 |
| <b>Figura 17:</b> IUV observado vs estimado con el modelo LightGBM para la estación Centro .....         | 49 |
| <b>Figura 18:</b> IUV observado vs estimado con el modelo LightGBM para la estación Cotocollao..         | 49 |
| <b>Figura 19:</b> IUV observado vs estimado con el modelo LightGBM para la estación Guamaní ....         | 50 |
| <b>Figura 20:</b> Despliegue del dashboard para la estimación de IUV en el DMQ.....                      | 55 |

## ÍNDICE DE TABLAS

|                                                                                                             |    |
|-------------------------------------------------------------------------------------------------------------|----|
| <b>Tabla 1:</b> Modelado de Transferencia Radiativa vs. Enfoques Basados en Datos .....                     | 19 |
| <b>Tabla 2:</b> Plan de proyecto.....                                                                       | 24 |
| <b>Tabla 3:</b> Comparación de resultados de modelos de ML .....                                            | 43 |
| <b>Tabla 4:</b> Evaluación final del modelo multi-estación .....                                            | 44 |
| <b>Tabla 5:</b> Resultados de la evaluación de modelos de ML por cada estación .....                        | 48 |
| <b>Tabla 6:</b> Resultados de la evaluación de modelos de ML por cada estación aplicando refinamiento ..... | 51 |
| <b>Tabla 7:</b> Estaciones para estimación de IUV en el DMQ .....                                           | 53 |

# CAPÍTULO I: INTRODUCCIÓN

## 1. Introducción

Los páramos andinos son ecosistemas únicos de alta montaña que funcionan como verdaderas "esponjas" del agua. En Ecuador, estos paisajes altos capturan, filtran y regulan el agua de lluvia y niebla, almacenándola en suelos orgánicos, turberas y humedales para luego liberarla gradualmente hacia ríos, lagunas y acuíferos. Los páramos del Antisana, Papallacta y Pichincha-Atacazo son vitales para el Distrito Metropolitano de Quito (DMQ): proveen más del 70 % del agua potable que consumen más de 2.8 millones de habitantes y sostienen ecosistemas downstream en la Amazonía (EPMAPS, 2023).

Sin embargo, estos ecosistemas frágiles enfrentan una amenaza creciente que es la alta radiación ultravioleta (UV) del sol. Quito, a unos 2 850 metros sobre el nivel del mar y muy cerca del ecuador, recibe niveles de UV entre los más intensos del planeta (Cordero et al., 2019). La delgada atmósfera a esta altitud filtra menos los rayos UV, y la posición ecuatorial hace que la radiación solar sea casi perpendicular durante gran parte del año. Como consecuencia, el Índice de Radiación Ultravioleta (IUV) supera frecuentemente valores extremos ( $>11$ , clasificado como “extremo” por la OMS), con picos que alcanzan 16-19 en zonas altoandinas, incluso en días nublados (Zaratti et al., 2014).

Esta radiación intensa genera estrés fotológico en las plantas y microorganismos del páramo. Especies nativas como frailejones, chuquiraguas, musgos y líquenes han desarrollado adaptaciones (vellos protectores, pigmentos antioxidantes, formas compactas) para soportar la alta UV, pero exposiciones prolongadas y extremas pueden:

- Alterar la fotosíntesis y reducir el crecimiento.
- Aumentar el daño oxidativo en células vegetales.
- Disminuir la biodiversidad endémica (muchas especies son únicas de los Andes tropicales).
- Comprometer la capacidad de los suelos para retener agua y carbono orgánico.

Estos impactos debilitan la función hidrológica principal de los páramos que es la regulación del ciclo del agua. Si la vegetación se degrada o se pierde cobertura, aumenta la erosión, disminuye la infiltración y se altera el almacenamiento en turberas y humedales, poniendo en riesgo el suministro hídrico sostenible para la ciudad y los ecosistemas dependientes.

Al mismo tiempo, estos niveles extremos de UV representan un riesgo directo para la salud de las personas que viven, trabajan o visitan estas zonas. La exposición prolongada puede causar quemaduras solares, envejecimiento prematuro de la piel, cataratas oculares y un aumento significativo en el riesgo de cáncer de piel, una de las patologías oncológicas de mayor incidencia en Ecuador (WHO, 2002). Grupos vulnerables incluyen niños, adultos mayores, trabajadores al aire libre (agricultores, recolectores de agua, guardaparques) y comunidades en páramos y zonas periurbanas. En Quito, autoridades como la Secretaría de Salud y el COE Metropolitano emiten alertas frecuentes recomendando medidas de protección como: usar protector solar SPF 30+ (reaplicar cada 2 horas), ropa de manga larga, sombreros de ala ancha, gafas con filtro UV, sombrillas y evitar la exposición directa entre las 10:00 y 16:00. Estas recomendaciones se refuerzan en campañas locales y nacionales para prevenir daños en piel y ojos, especialmente en días con IUV >11 (Secretaría de Salud del DMQ, 2024).

A pesar de su importancia, el monitoreo continuo de la radiación UV en el DMQ y sus páramos es limitado. Los sensores UV especializados son costosos (5 000 a 15 000 USD por unidad) y requieren mantenimiento frecuente, lo que impide su despliegue masivo en las 33 estaciones hidrometeorológicas de la Secretaría de Ambiente y EPMAAPS. Por ello, la información diaria y espacial sobre el IUV es insuficiente para detectar hotspots de estrés ambiental, priorizar acciones de conservación o emitir alertas oportunas para la población.

Es por ello que surge una solución práctica y de bajo costo que es aprovechar los datos que se encuentran disponibles en esas estaciones como: radiación solar global, temperatura, humedad, presión atmosférica y precipitación, con el fin de estimar el índice UV con alta precisión mediante machine learning (aprendizaje automático). Estos algoritmos pueden descubrir relaciones complejas y no lineales entre variables hidrometeorológicas accesibles y la radiación UV, logrando predicciones confiables sin sensores UV adicionales.

El objetivo principal de esta investigación es desarrollar y validar un modelo de machine learning que estime el IUV en todo el DMQ usando exclusivamente las variables de las 33 estaciones locales. El modelo se integra en un dashboard interactivo que muestre mapas del IUV estimado, facilitando la identificación de zonas de alto estrés fotológico en páramos clave y apoyando alertas preventivas para la población.

Con este enfoque se busca:

- Reducir más del 90 % los costos de monitoreo UV directo.
- Mejorar la vigilancia espacial y temporal de la radiación UV en ecosistemas estratégicos.
- Fortalecer la conservación de los páramos como fuentes de agua y sumideros de carbono.
- Proporcionar información oportuna para proteger la salud de las personas y promover prácticas de fotoprotección en la ciudadanía.
- Generar una herramienta escalable para otras regiones andinas con alta radiación UV y recursos limitados.

De esta forma, el proyecto contribuye a una gestión ambiental más inteligente y preventiva en el contexto del cambio climático, maximizando la infraestructura pública existente, protegiendo los servicios ecosistémicos vitales y salvaguardando la salud de la población quiteña.

## **1.1 Antecedentes**

La estimación de la radiación UV mediante modelos de Machine Learning (ML) ha ganado relevancia en los últimos años, especialmente en regiones de alta altitud donde los niveles de UV son elevados debido a factores como la elevación, la latitud y las condiciones atmosféricas. Diversos estudios han demostrado que los algoritmos de ML pueden predecir con precisión la radiación UV utilizando variables hidrometeorológicas y datos satelitales, reduciendo la dependencia de equipos costosos como radiómetros. Por ejemplo, en contextos de alta altitud, como el Desierto de Atacama, se han aplicado redes neuronales artificiales para estimar la radiación UV-A a partir de mediciones de irradiancia solar global horizontal (GHI), logrando errores relativos bajos del 6% en irradiancia y 3% en irradiación diaria. De manera similar, en regiones montañosas como el Tíbet, mediciones directas de radiación UV en altitudes entre 3000 y 5000 metros han resaltado la influencia de la elevación en el aumento de la intensidad UV, lo que subraya la necesidad de modelos predictivos adaptados a entornos andinos.

En entornos de alta altitud como Nepal, estudios comparativos entre mediciones satelitales (OMI) y datos terrestres han cuantificado el efecto de altitud en el índice UV, con incrementos del 6.5-8.8% por kilómetro, lo que valida la integración de variables como el ángulo zenital solar y la nubosidad en modelos predictivos. En Pakistán, en glaciares de regiones norteñas de alta elevación, algoritmos de ML como redes neuronales feed-forward (FFNN) han predicho la radiación solar superficial con alta precisión ( $R^2$  de 0.916 en KPK y 0.887 en Gilgit), incorporando variables

como temperatura, presión, precipitación y aerosoles, lo que es transferable a la predicción de UV dada su correlación con la radiación solar global. Estos enfoques destacan la robustez de modelos ensemble como Random Forest y XGBoost para capturar relaciones no lineales en datos hidrometeorológicos.

Revisiones exhaustivas sobre técnicas de ML para el pronóstico de radiación solar, que sirven de base para modelos de UV, indican que métodos como Support Vector Machines (SVM), Random Forest y redes neuronales superan a los modelos empíricos tradicionales en precisión, especialmente en escenarios con variabilidad atmosférica. Por instancia, enfoques híbridos que combinan deep learning con ML, como redes neuronales recurrentes (RNN) y Gradient Boosting, han logrado  $R^2$  superiores a 0.94 en predicciones a largo plazo de irradiancia solar, utilizando variables como humedad relativa, velocidad del viento y temperatura.

Además, en contextos urbanos y de salud pública, sistemas móviles basados en deep learning han calculado el índice UV utilizando valores de iluminancia y datos de ubicación, logrando precisiones superiores al 44% respecto a métodos tradicionales. Modelos como NNARX han estimado la radiación solar global incorporando el índice UV como variable, alcanzando coeficientes de determinación de 0.97. En China, datasets generados con Random Forest han predicho radiación UV diaria a 10 km de resolución, considerando factores como elevación y meteorología, con aplicaciones en salud pública similares a las propuestas para Quito. Enfoques híbridos explicables, como redes neuronales profundas con interpretabilidad, han predicho el índice UV a una hora de antelación con  $R^2$  de 0.96, identificando variables clave como el ángulo zenital solar y el ozono.

Estos antecedentes resaltan la viabilidad de algoritmos como Random Forest, XGBoost y redes neuronales para estimar UV en regiones andinas de alta altitud, alineándose con el enfoque este proyecto al utilizar variables hidrometeorológicas locales y técnicas de selección de características para optimizar el rendimiento ( $R^2 \geq 0.92$ ,  $RMSE \leq 0.8$ ). Sin embargo, hay una brecha en aplicaciones específicas al Distrito Metropolitano de Quito, donde la topografía única y la variabilidad climática tropical requieren

## **1.2 Planteamiento del problema**

En el Distrito Metropolitano de Quito se registran índices de radiación ultravioleta extremadamente altos y altamente variables debido a la combinación de alta altitud ( $>2.800$  m s.n.m.), cercanía a la línea ecuatorial y condiciones atmosféricas dinámicas (Cordero et al., 2019). Esta exposición representa un riesgo significativo para la salud pública y afecta procesos ecológicos esenciales en ecosistemas estratégicos de recarga hídrica.

Sin embargo, no existe una red operativa de radiómetros UV debido al elevado costo de adquisición y mantenimiento de estos instrumentos. Aunque se dispone de 33 estaciones hidrometeorológicas (9 de la Secretaría de Ambiente y 24 de EPMAPS), que registran variables fuertemente correlacionadas con la radiación UV, estas no se han utilizado sistemáticamente para estimar esta variable de forma indirecta

¿Cómo estimar con precisión la radiación ultravioleta en el Distrito Metropolitano de Quito utilizando exclusivamente datos hidrometeorológicos disponibles y técnicas de machine learning, evitando la necesidad de instrumentación especializada y generando mapas espacio-temporales continuos de alta resolución?

Este problema es relevante, viable (datos históricos disponibles), delimitado (DMQ y variables existentes), específico (estimación mediante machine learning) y de alto impacto en gestión de recursos hídricos, adaptación climática y salud pública.

## **1.3 Justificación**

El presente proyecto resulta altamente pertinente debido a su impacto directo en la salud pública y la conservación ambiental del Distrito Metropolitano de Quito. La generación del modelo de estimación de radiación ultravioleta permite identificar zonas de alto riesgo para la población (especialmente en sectores vulnerables expuestos a índices UV extremos) y detectar hotspots de estrés fotológico en ecosistemas estratégicos de recarga hídrica, como los páramos del Antisana, Papallacta y Pichincha-Atacazo. De esta forma, se facilitará la toma de decisiones informadas en campañas de prevención de cáncer de piel, protección ocular y manejo sostenible de las fuentes de agua que abastecen a más del 70 % de la ciudad.

Desde el punto de vista económico y operativo, el enfoque propuesto logra una reducción superior al 90 % en los costos asociados al monitoreo de radiación UV al eliminar la necesidad de adquirir y mantener radiómetros especializados (cuyo valor unitario oscila entre 5 000 y 15 000 USD). En su lugar, aprovecha la infraestructura hidrometeorológica ya existente de 33 estaciones operadas por la Secretaría de Ambiente y EPMAPS para generar información de alta calidad mediante técnicas de machine learning, maximizando el uso de recursos públicos y fortaleciendo la colaboración interinstitucional.

Finalmente, la metodología desarrollada posee un elevado potencial de escalabilidad y replicabilidad en otras ciudades andinas que enfrentan condiciones similares de alta radiación UV y limitaciones presupuestarias para instrumentación dedicada. Además, contribuye al avance de la ciencia de datos ambiental al validar un modelo predictivo robusto basado en variables de bajo costo, sentando un precedente técnico transferible al contexto latinoamericano y enriqueciendo el conocimiento en meteorología aplicada en regiones tropicales de montaña.

## **1.4 Objetivos**

### **1.4.1 Objetivo General**

Desarrollar y validar un modelo de Machine Learning para la estimación precisa del Índice de Radiación Ultravioleta (IUV) en el Distrito Metropolitano de Quito, utilizando variables hidrometeorológicas de la red de monitoreo local para su aplicación en gestión ambiental y salud pública.

### **1.4.2 Objetivos Específicos**

- Identificar el subconjunto óptimo de variables predictoras de la radiación UV en condiciones de alta altitud tropical andina mediante técnicas de selección de características (SHAP, RFE, importancia por permutación)
- Evaluar y seleccionar la arquitectura de algoritmo de *machine learning* más eficaz entre regresión lineal múltiple, Random Forest, XGBoost y LightGBM, utilizando validación cruzada espacio-temporal.

- Implementar un dashboard interactivo para la visualización espacial y temporal del índice de radiación UV estimado para las estaciones del DMQ.

## 1.5 Alcance

El alcance de esta investigación se define en términos geográfico, temporal, técnico-metodológico y de aplicación, con el fin de garantizar un desarrollo preciso, viable y con impacto concreto en el contexto del DMQ.

- **Alcance geográfico:** La estimación del IUV se limita al territorio del Distrito Metropolitano de Quito, abarcando aproximadamente 4 217 km<sup>2</sup>, incluyendo zonas urbanas, periurbanas y altoandinas con páramos clave como los del Antisana, Papallacta y Pichincha-Atacazo. Se utilizan exclusivamente los datos de las 33 estaciones hidrometeorológicas operativas de la Secretaría de Ambiente y EPMAAPS (9 de la Secretaría de Ambiente y 24 de EPMAAPS), distribuidas estratégicamente en el DMQ. No se incluyen áreas fuera del límite administrativo del DMQ ni estaciones externas.
- **Alcance temporal:** El modelo se desarrolla y valida con datos históricos disponibles de al menos un período completo de un año (preferentemente 2020-2024 o el rango más reciente y continuo que proporcionen las instituciones), cubriendo variabilidad estacional (seca y lluviosa) típica de Quito. La operacionalización permite generar estimaciones diarias y horarias en tiempo casi real o retrospectivo, pero el desarrollo inicial se centra en predicción/estimación retrospectiva y validación con datos observados, sin pronóstico a futuro (forecasting) más allá de la hora actual.
- **Alcance técnico-metodológico:**
  - La variable objetivo (IUV) para el entrenamiento y validación del modelo se obtendrá de los registros históricos de las cuatro estaciones de la red que, durante un determinado periodo, contaron con sensores de radiación UV espectral. Estos datos constituyen el "ground truth" o verdad de campo contra la cual se ajustarán y evaluarán los modelos de machine learning. Para la predicción operativa, el modelo final utilizará únicamente las variables hidrometeorológicas de las 33 estaciones, sin requerir sensores UV activos.
  - No se incorporan datos satelitales, mediciones de ozono ni aerosoles.

- Se comparan al menos cuatro algoritmos de machine learning.
  - Se aplican técnicas avanzadas de selección de características.
  - La evaluación utiliza validación cruzada espacio-temporal.
  - El resultado final incluye un prototipo de dashboard interactivo.
- 
- **Alcance de aplicación:** El modelo busca apoyar la gestión ambiental (identificación de hotspots de estrés fotológico en páramos y monitoreo de la función hidrológica), la conservación de ecosistemas de recarga hídrica y la emisión de alertas preventivas para fotoprotección de la población (campanas de salud pública). Contribuye indirectamente a la planificación de recursos hídricos y adaptación al cambio climático, pero no incluye implementación operativa en sistemas de alerta temprana ni evaluación de impacto post-implementación.

Este capítulo establece así las bases conceptuales y justificativas para un proyecto innovador que transforma una limitación tecnológica en una oportunidad de gestión inteligente. Al maximizar la infraestructura pública existente y alinear la ciencia de datos con prioridades ambientales y sociales, se pavimenta el camino hacia una Quito más resiliente ante los desafíos del cambio climático. Los capítulos subsiguientes profundizarán en la metodología detallada, la implementación del modelo y sus validaciones, consolidando esta propuesta como un aporte tangible a la sostenibilidad andina y a la salud pública.

## CAPITULO II. FUNDAMENTACIÓN TEORICA

### 2. Marco Teórico

El marco teórico de esta investigación proporciona los fundamentos científicos para la estimación del IUV mediante modelos de machine learning en el DMQ. Se inicia con los principios físicos que explican la intensidad extremadamente alta de la radiación UV en la atmósfera andina de alta altitud tropical, se justifica el uso de variables hidrometeorológicas como proxies físicos, se analiza la brecha existente entre los modelos tradicionales y los enfoques basados en datos, y finalmente se presenta el estado del arte y los algoritmos seleccionados, destacando sus ventajas matemáticas para el tipo de datos disponibles.

#### 2.1 Fundamentos Físicos de la Radiación en Altura

La RSG es la suma de componentes directa, difusa y reflejada que llega a la superficie, medida en  $W/m^2$  y abarcando espectros visibles (400–700 nm), infrarrojo ( $>700$  nm) y ultravioleta (UV, 100–400 nm) (OMM, 2008). La radiación UV representa menos del 10% de la RSG, pero sus impactos biológicos y ecológicos son críticos: daños en ADN celular, estrés oxidativo en vegetación y alteraciones en ciclos hidrológicos (McKinlay & Diffey, 1987).

La ubicación de Quito a aproximadamente 2.850 m s.n.m. y su posición muy cercana al ecuador ( $0^{\circ}13'S$ ) genera niveles de radiación UV entre los más intensos del planeta, con valores de IUV que superan frecuentemente la categoría de “extremo” ( $>11$ ) según la escala de la OMS, y que en ocasiones alcanzan 16–19 en zonas altoandinas (Zaratti et al., 2014; Cordero et al., 2019; Quito Informa, 2025). Esta intensidad se explica por varios fundamentos físicos característicos de la atmósfera andina que se detallan a continuación:

##### 2.1.1 Menor espesor atmosférico y Dispersión de Rayleigh

A mayor altitud, la columna de aire sobre la superficie es más delgada, lo que reduce significativamente la densidad molecular disponible para la atenuación radiativa. Este fenómeno se rige principalmente por la dispersión de Rayleigh, un proceso elástico donde los fotones interactúan con moléculas ( $N_2$ ,  $O_2$ ) de tamaño mucho menor a la longitud de onda ( $\lambda$ ).

La sección eficaz de Rayleigh presenta una dependencia crítica con la longitud de onda, definida por la relación:

$$\sigma_R = \frac{8\pi^3}{3\lambda^4 N^2} \frac{(n^2 - 1)^2}{(n^2 + 2)^2} F_k \quad (1)$$

Donde:

- **$\sigma_R$**  (Sección eficaz de Rayleigh): Representa la "probabilidad" de que un fotón sea dispersado por una molécula. A mayor  **$\sigma_R$** , más radiación se desvía.
- **$\pi$  (Pi)**: Constante matemática (3.14159...).
- **$\lambda$  (Lambda)**: Longitud de onda de la radiación incidente (en nanómetros). Está elevada a la cuarta potencia en el denominador, dictando la intensidad de la dispersión.
- **$N$** : Densidad molecular (número de moléculas por unidad de volumen). En Quito, al estar a 2.850 msnm,  **$N$**  es menor que al nivel del mar.
- **$n$** : Índice de refracción del aire ( $\approx 1.0003$ ). Define cómo se propaga la luz en el medio gaseoso.
- **$F_k$  (Factor de King)**: Un factor de corrección que ajusta la fórmula debido a que las moléculas de aire ( **$N_2$**  y  **$O_2$** ) no son esferas perfectas, sino que tienen cierta asimetría (anisotropía).

Esta dependencia de  $1/\lambda^4$  implica que las longitudes de onda más cortas, como el espectro UVB, se dispersan con mayor intensidad que el espectro visible. En Quito (presión  $\approx 750$  hPa), la reducción de la profundidad óptica de Rayleigh ( $\tau_R$ ) es de entre un 30% y 50% respecto al nivel del mar. En consecuencia, la irradiancia UVB aumenta entre un 8% y 10% por cada 1.000 m de elevación, resultando en un flujo neto en superficie notablemente superior al de regiones costeras (Blumthaler et al., 1997; Cordero et al., 2019).

### 2.1.2 Posición Ecuatorial y Geometría Solar

La ubicación geográfica en latitud cercana a cero garantiza que el ángulo cenital solar (SZA, por sus siglas en inglés) sea mínimo durante gran parte del año. Un SZA bajo reduce el camino óptico (masa de aire) que los fotones deben recorrer a través de la atmósfera, minimizando las probabilidades de absorción y dispersión, y maximizando la irradiancia directa sobre el plano horizontal.

### 2.1.3 Dinámica de la Composición Atmosférica y Dispersión de Mie

La capa de ozono absorbe la mayor parte de la UVB, pero presenta variaciones locales y estacionales que incrementan la exposición en ciertos periodos (Madronich, 1993). Asimismo, los componentes en la tropósfera baja, como aerosoles y nubes, modulan la radiación mediante la dispersión de Mie. A diferencia de Rayleigh, Mie ocurre con partículas de tamaño comparable o mayor a  $\lambda$  (0.1-10  $\mu\text{m}$ ) y se describe mediante la serie infinita de coeficientes de Riccati-Bessel:

$$\sigma_M = \frac{2\pi}{k^2} \sum_{n=1}^{\infty} (2n + 1) \text{Re}(a_n + b_n) \quad (2)$$

Esta fórmula es una sumatoria infinita porque la dispersión por partículas grandes (aerosoles/nubes) es mucho más compleja que por moléculas pequeñas.

- **$\sigma_M$  (Sección eficaz de Mie):** Al igual que en Rayleigh, mide la capacidad de las partículas de polvo, polen o gotas de agua para dispersar la luz.
- **$k$  (Número de onda):** Se define como  $\left(\frac{2\pi}{\lambda}\right)$ . Relaciona la frecuencia espacial de la onda con su longitud de onda.
- **$a_n, b_n$  (Coeficientes de Mie):** Son funciones matemáticas complejas (basadas en funciones de Riccati-Bessel)

Este mecanismo es fundamental en la región andina debido a la alta variabilidad nubosa. Aunque las nubes densas atenúan la radiación, el fenómeno de "Cloud Enhancement" (refuerzo por nubes) puede incrementar la radiación difusa hasta 1.5 veces en condiciones de nubosidad parcial, debido a la dispersión múltiple en los bordes de nubes cúmulos (Calbó et al., 2005). Esto explica por qué en Quito se registran índices de radiación extremo (IUV 11+) incluso en días con presencia nubosa intermitente.

En 2025, el INAMHI y el COE Metropolitano reportaron alertas frecuentes por índices IUV de 8 a 11 (muy alto/extremo), con picos en julio debido a olas de calor. Estos niveles se vieron influenciados por la variabilidad nubosa y una menor dispersión de Rayleigh en la capital (Quito Informa, 2025).

## 2.2 Justificación Física de las Variables Hidrometeorológicas

Debido a la ausencia de una red operativa de sensores UV especializados en el DMQ (por su elevado costo), es necesario utilizar variables hidrometeorológicas ya disponibles en las estaciones como proxies fuertemente correlacionados con la radiación UV. Estas variables tienen una justificación física directa:

- **Radiación Solar Global (RSG):** Es el predictor principal. La UV forma parte del mismo espectro solar que la RSG y en condiciones de cielo despejado existe una correlación muy fuerte (UV representa aproximadamente 3–6% de la RSG) (Kudish et al., 2005). No obstante, su relación varía bajo cielos nublados debido a la respuesta diferenciada de las longitudes de onda cortas ante la turbidez atmosférica.
- **Humedad relativa y Precipitación:** El vapor de agua y las gotas de nubes provocan dispersión selectiva (Mie) y absorción, modulando el flujo de UV. Condiciones de alta humedad suelen correlacionar negativamente con la intensidad UV debido a mayor dispersión y absorción (Calbó et al., 2005).
- **Presión atmosférica:** Actúa como proxy inverso de la altitud efectiva y del espesor óptico atmosférico. No obstante, su relación varía bajo cielos nublados debido a la respuesta diferenciada de las longitudes de onda cortas ante la turbidez atmosférica (Blumthaler et al., 1997).
- **Temperatura:** Es un indicador indirecto de estabilidad atmosférica y nubosidad. Días más cálidos y despejados tienden a presentar mayores valores de UV, mientras que inversiones térmicas (frecuentes en valles quiteños) pueden modular la dispersión (Badosa et al., 2007).
- **Ángulo cenital solar (SZA):** Calculado a partir de la geometría solar (hora, fecha y coordenadas), el SZA determina la masa de aire relativa ( $m \approx 1/\cos(SZA)$ ). Este valor dicta la longitud de la trayectoria que los fotones deben recorrer, definiendo la intensidad de la irradiancia directa según la Ley de Beer-Lambert

Estas relaciones físicas sustentan la viabilidad de estimar el IUV exclusivamente a partir de variables hidrometeorológicas locales.

## 2.3 Modelado de Transferencia Radiativa vs. Enfoques Basados en Datos

Los modelos físicos como TUV (Madronich, 1993) o libRadtran son teóricamente muy precisos, pero exigen datos de entrada que no están disponibles de forma continua y económica en el DMQ (perfiles verticales de ozono, AOD satelital, distribución vertical de nubes).

- **Brecha científica principal:** Los modelos empíricos tradicionales asumen relaciones lineales simples que no capturan las no linealidades producidas por nubosidad variable, anomalías de ozono negativo (NOA) y efectos de altitud extrema (Foyo-Moreno et al., 1999; Calbó et al., 2005).
- **Solución propuesta:** El Machine Learning permite rellenar esta brecha al aprender relaciones no lineales y complejas directamente de los datos hidrometeorológicos locales disponibles, logrando predicciones robustas y de bajo costo sin necesidad de instrumentación UV especializada ni datos satelitales complejos (Voyant et al., 2017; Liu et al., 2024).

*Tabla 1: Modelado de Transferencia Radiativa vs. Enfoques Basados en Datos*

| Característica                          | Modelos de Transferencia Radiativa (RTM)                   | Enfoques Basados en Datos (ML/DL)                        |
|-----------------------------------------|------------------------------------------------------------|----------------------------------------------------------|
| <b>Fundamento</b>                       | Leyes físicas (ecuación de transferencia radiativa)        | Algoritmos estadísticos y aprendizaje de patrones        |
| <b>Datos requeridos</b>                 | Perfiles atmosféricos detallados (ozono, aerosoles, nubes) | Datos históricos locales de sensores hidrometeorológicos |
| <b>Interpretación</b>                   | Alta (caja blanca, causalidad física)                      | Baja-media (caja negra / explicabilidad con SHAP)        |
| <b>Costo</b>                            | Alto (datos de entrada costosos + cómputo intensivo)       | Bajo (aprovecha infraestructura existente)               |
| <b>Robustez ante variabilidad local</b> | Media-baja (requiere recalibración frecuente)              | Alta (aprende de datos locales reales)                   |

## 2.4 Estado del Arte en Estimación de UV con ML

La literatura reciente muestra clara superioridad de los métodos de machine learning sobre los modelos empíricos tradicionales, especialmente en condiciones de alta variabilidad atmosférica. Esta transición se fundamenta en la capacidad de los algoritmos modernos para modelar relaciones no lineales complejas entre variables proxy y la radiación ultravioleta.

Actualmente, los modelos basados en ensambles de árboles dominan el procesamiento de datos tabulares. Algoritmos como XGBoost y CatBoost han ganado relevancia entre 2024 y 2026 debido a su robustez frente a valores atípicos y eficiencia computacional. Por ejemplo:

- En glaciares andinos de Pakistán, lograron un  $R^2$  de hasta 0.916 integrando variables térmicas y de presión (Ghazal et al., 2023)
- Datasets nacionales diarios de UV a 10 km en China, validando del uso de Random Forest con muy buen desempeño (Liu et al., 2024).

Paralelamente, los enfoques de Aprendizaje Profundo (DL) e híbridos han demostrado ser óptimos para capturar dependencias espaciotemporales. Destacan los modelos CNN-LSTM que, optimizados mediante algoritmos genéticos, permiten predicciones con horizontes temporales extendidos y errores relativos inferiores al 6% en entornos de alta radiación como el Desierto de Atacama (Mondaca et al., 2019; Prasad et al., 2023).

Asimismo, la selección de características técnicas como SHAP, Recursive Feature Elimination (RFE) e importancia por permutación son ampliamente utilizadas para identificar el subconjunto óptimo de variables, priorizando usualmente: SZA, RSG, nubosidad/temperatura, y descartando variables de bajo aporte predictivo (Lundberg & Lee, 2017).

Muy pocos estudios aplican modelos exclusivamente con variables hidrometeorológicas locales en contextos andinos tropicales de alta altitud como Quito

## 2.5 Algoritmos de Aprendizaje Supervisado para Regresión

Los algoritmos seleccionados se justifican por su capacidad para manejar las características típicas de los datos hidrometeorológicos del DMQ: valores atípicos (picos de UV en días despejados), valores faltantes (fallos de sensores) y relaciones no lineales.

- **Regresión Lineal Múltiple:** Es el modelo base y de referencia. Asume una relación lineal entre las variables predictoras y la variable objetivo. Es altamente interpretable y sirve como línea base para comparar el beneficio de los modelos no lineales. Sin embargo, es limitada cuando existen interacciones complejas o no linealidades fuertes en los datos (Hastie et al., 2009).

- **Random Forest:** Ensemble de árboles mediante bagging, muy robusto a outliers y ruido. Permite imputación interna de valores nulos y ofrece buena interpretabilidad mediante importancia de variables y SHAP (Biau & Scornet, 2016).
- **XGBoost y LightGBM:** Representan el estado del arte actual para datos tabulares (Chen & Guestrin, 2016). Ventajas matemáticas clave:
  - Boosting secuencial (corrige errores del árbol anterior), mayor precisión que RF en la mayoría de casos reales.
  - Manejo nativo y optimizado de valores faltantes (asigna dirección por defecto en las divisiones).
  - Funciones de pérdida flexibles y regularización; mejor control de sobreajuste y manejo relativo de outliers.
  - Extremadamente rápidos y escalables (LightGBM especialmente eficiente con grandes volúmenes de datos).

Esta selección de algoritmos, combinada con técnicas avanzadas de selección de características y validación espacio-temporal, permite obtener un modelo preciso, interpretable y robusto adaptado a las condiciones reales de monitoreo del DMQ.

## CAPÍTULO III: METODOLOGÍA

### 3. Metodología

En el presente estudio se aplicó la metodología CRISP-ML para el desarrollo e implementación de un modelo predictivo de IUV en el DMQ. Esta metodología, basada en un enfoque estructurado adaptado de CRISP-DM, proporciona una guía flexible y adaptable para proyectos de aprendizaje automático, permitiendo una planificación sistemática desde la comprensión del problema hasta la evaluación y despliegue del modelo. Al utilizar CRISP-ML se garantiza un proceso replicable, con énfasis en la calidad de datos, interpretabilidad de modelos y métricas específicas para valores extremos de IUV, alineado con los desafíos ambientales y de salud pública en regiones de alta altitud tropical. El enfoque integra principios de física atmosférica (e.g., proxies como ángulo zenital solar) con técnicas de ML, asegurando robustez en datos heterogéneos de estaciones hidrometeorológicas.

#### 3.1 Comprensión de Datos y Negocio

La estimación de IUV mediante ML aborda un problema crítico en Quito: la alta exposición UV debido a altitud y latitud ecuatorial, que afecta ecosistemas (estrés fotológico en páramos) y salud humana (riesgo de cáncer de piel). Utilizando variables hidrometeorológicas de bajo costo (de 33 estaciones existentes), se evita la necesidad de sensores UV especializados, reduciendo costos en >90% y maximizando infraestructura pública.

##### 3.1.1 Comprensión de Negocio

###### 3.1.1.1 Objetivos del negocio

- Maximizar la conservación de ecosistemas hídricos identificando hotspots de estrés fotológico en páramos clave (Antisana, Papallacta, Pichincha-Atacazo), optimizando la gestión de recursos hídricos que abastecen >70% del agua potable del DMQ.

- Mejorar la rentabilidad operativa al eliminar costos de adquisición/mantenimiento de radiómetros UV (~5.000-15.000 USD/unidad), reutilizando datos existentes para generar estimaciones diarias espaciales.
- Aumentar la eficiencia en alertas de salud pública, descartando predicciones inexactas en días de IUV extremo ( $>11$ ) para prevenir exposiciones riesgosas en poblaciones vulnerables.

### **3.1.1.2 Criterios de rendimiento**

Obtener un modelo con precisión predictiva alta, para alertas críticas. Facilitar interpretación de variables (e.g., vía SHAP) para entender relaciones físicas (RSG como proxy principal).

### **3.1.1.3 Evaluación de la situación**

En Quito, el monitoreo UV es limitado por costos y variabilidad atmosférica (nubosidad, altitud), llevando a decisiones basadas en experiencia subjetiva o datos puntuales, con riesgos como:

- Subestimar IUV en días nublados parciales (cloud enhancement), afectando páramos y salud.
- Mantener monitoreo ineficiente, incrementando gastos innecesarios.
- Pérdida de potencial predictivo en datos hidrometeorológicos existentes (~33 estaciones con cobertura  $>80\%$ ).

Riesgos: datos faltantes ( $>10\%$  en variables como nubosidad), desbalance en IUV extremos (pocos  $>16$ ), multicolinealidad (e.g., temperatura-humedad). No hay restricciones legales en datos (anonimizados, públicos/institucionales). Recursos disponibles: Google Colab para procesamiento, bibliotecas Python (pandas, XGBoost, SHAP).

### **3.1.1.4 Traducir objetivos comerciales en objetivos de ML**

- Desarrollar un modelo de regresión supervisada que prediga IUV a partir de proxies hidrometeorológicos, minimizando errores en la predicción
- Optimizar métricas para reducir bias en extremos (e.g., MAE estratificado  $>11$ ), evitando sobreajuste espacial/temporal vía CV espacio-temporal.

- Proporcionar explicabilidad (SHAP) para identificar variables clave (e.g., RSG, SZA), permitiendo comprensión física de predicciones.

### 3.1.1.5 Plan de proyecto

*Tabla 2: Plan de proyecto*

| Fase                     | Duración  | Recursos                                                 | Riesgos                                       |
|--------------------------|-----------|----------------------------------------------------------|-----------------------------------------------|
| Comprensión del negocio  | 1 semana  | Responsable del proyecto, tutor                          |                                               |
| Comprensión de los datos | 2 semanas | Acceso a datos históricos de estaciones (SAQ/EPMAPS)     |                                               |
| Preparación de los datos | 4 semanas | Python (pandas, scikit-learn), Google Colab              | Alta cantidad de faltantes, multicolinealidad |
| Modelado                 | 4 semanas | Plataforma para modelado (Jupyter/Colab), bibliotecas ML | Sobreajuste en series temporales              |
| Evaluación               | 2 semanas | Métricas estratificadas, SHAP                            | Bajo rendimiento en extremos IUV              |
| Despliegue               | 1 semana  | Shany, Streamlit o Power BI para dashboard               |                                               |

### 3.1.2 Comprensión de Datos

#### 3.1.2.1 Recopilación de los datos

Se utilizó registros horarios de 2017-2025 de las 4 estaciones de estudio que son: Belisario, Centro, Cotocollao y Guamaní, almacenados en archivos Excel (.xlsx) por variable. Acceso público de la página web: <https://datosambiente.quito.gob.ec/>. Los archivos clave para este modelo fueron:

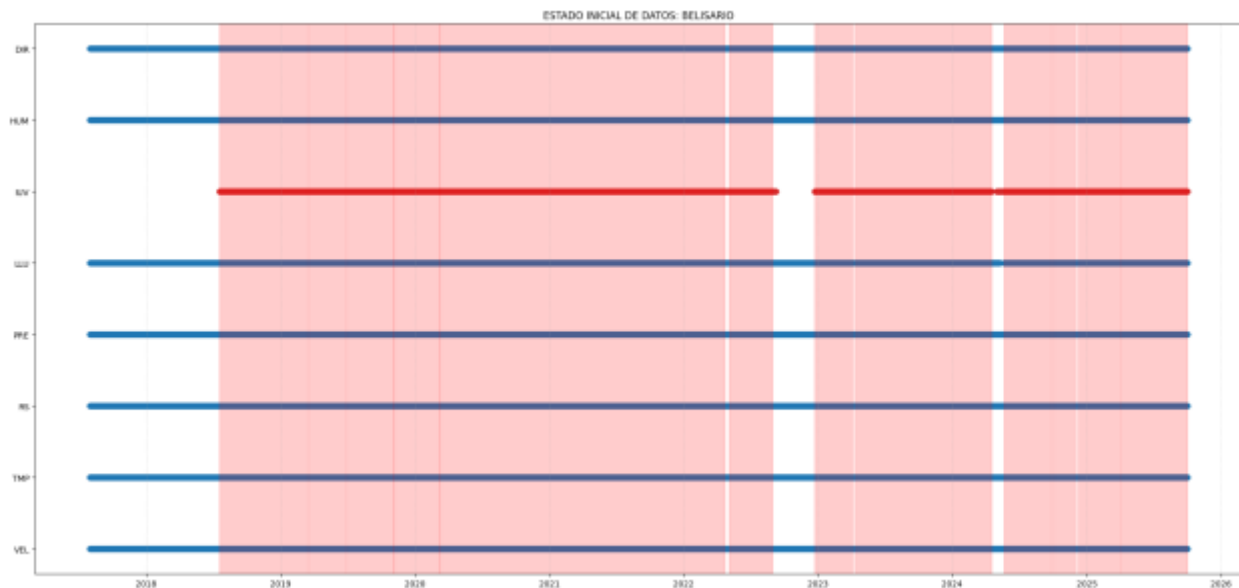
- RS.xlsx (radiación solar global).
- TMP.xlsx (temperatura)
- HUM.xlsx (humedad).
- IUV.xlsx (Índice de radiación ultra violeta)
- LLU.xlsx (precipitación)
- PRE.xlsx (presión)
- VEL.xlsx (velocidad del viento)

- DIR.xlsx (dirección del viento)

Los datos se cargaron al google drive para poder usar la herramienta Google Colab con librerías pandas, numpy, seaborn, matplotlib y hacer una revisión inicial.

En esta revisión inicial se observó la disponibilidad de datos de cada variable para las 4 estaciones de estudio y a su vez se observan los periodos de tiempo donde se puede correlacionar la variable de estimación que es IUV con las variables de estudio.

- **Estación Belisario**

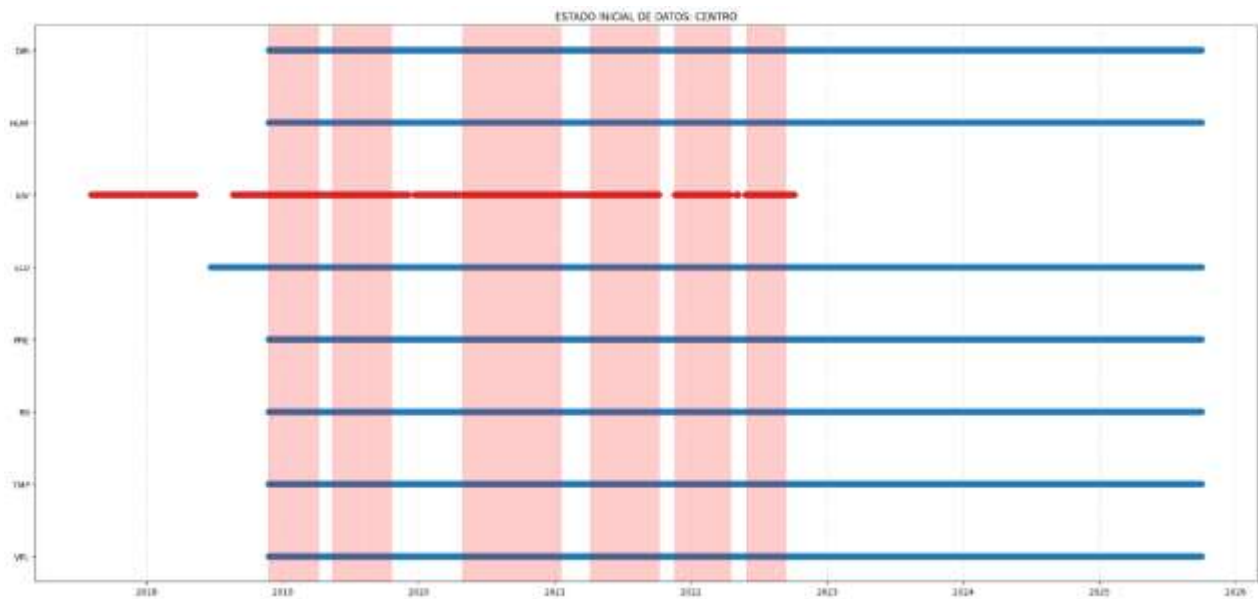


*Figura 1: Disponibilidad de datos en la estación Belisario*

En la Figura 1 se puede ver la disponibilidad de datos crudos de cada variable para la estación Belisario, además las franjas rojas sugieren que el análisis se realice en esos periodos de datos, ya que se puede comparar las variables de estudio (líneas azules) con la variable de estimación IUV (línea roja).

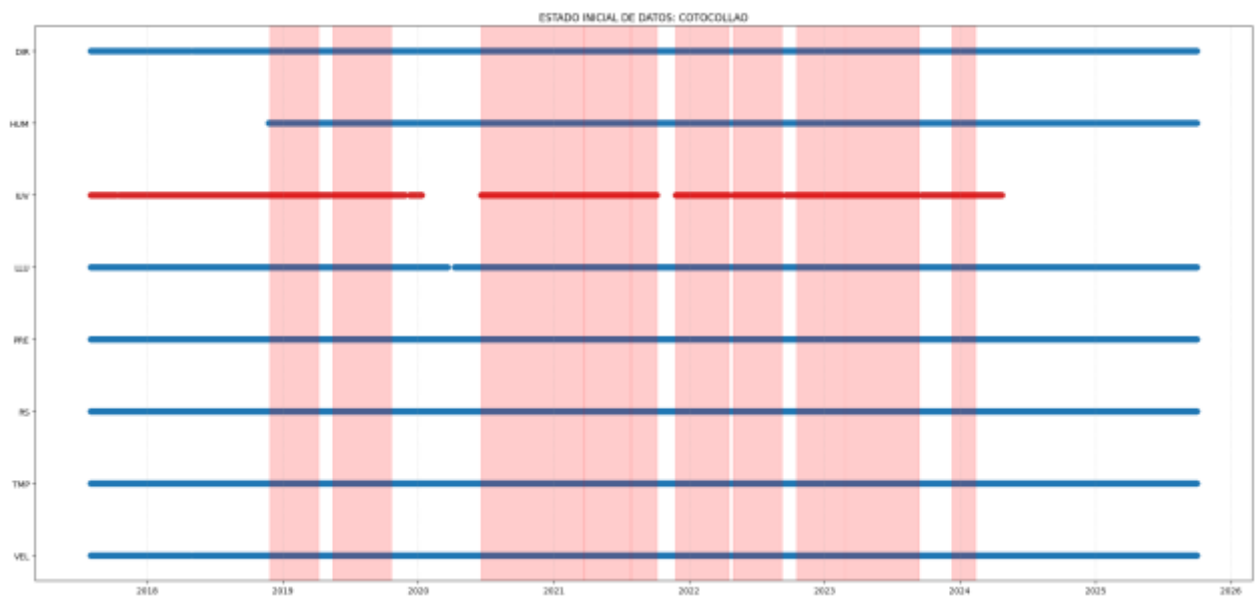
- **Estación Centro**

En la Figura 2 se puede ver la disponibilidad de datos crudos de cada variable para la estación Centro, además las franjas rojas sugieren que el análisis se realice en esos periodos de datos, ya que se puede comparar las variables de estudio con la variable de estimación IUV.



*Figura 2: Disponibilidad de datos en la estación Centro*

- **Estación Cotocollao**



*Figura 3: Disponibilidad de datos en la estación Cotocollao*

En la Figura 3 se puede ver la disponibilidad de datos crudos de cada variable para la estación Cotocollao, además las franjas rojas sugieren que el análisis se realice en esos periodos de datos, ya que se puede comparar las variables de estudio con la variable de estimación IUV.

- **Estación Guamaní**



*Figura 4: Disponibilidad de datos en la estación Guamaní*

En la Figura 4 se puede ver la disponibilidad de datos crudos de cada variable para la estación Guamaní, además las franjas rojas sugieren que el análisis se realice en esos periodos de datos, ya que se puede comparar las variables de estudio con la variable de estimación IUV

### 3.1.2.2 Descripción de los datos

La base de datos consta de 4 nuevos datasets con todas las variables en los periodos que todos los datos están completos.

- Para la estación Belisario se tiene 14 periodos óptimos / 2466 días totales
- Para la estación Centro se tiene 7 periodos óptimos / 1284 días totales.
- Para la estación Cotocollao se tiene 11 periodos óptimos / 1695 días totales.
- Para la estación Guamaní se tiene 7 periodos óptimos / 1129 días totales.

Sin embargo, se realizó un trabajo muy minucioso para poder completar las series de datos y entrenar de mejor manera a nuestro modelo.

## 3.2 Preparación de los Datos

Se preparó el dataset para ML, manejando gaps y derivando features físicas.

### 3.2.1 Limpieza

Una vez integrados los registros de las diferentes estaciones, se aplicó un proceso de limpieza secuencial diseñado específicamente para series temporales meteorológicas a resolución horaria. El objetivo fue eliminar valores físicamente imposibles, imputar huecos cortos de forma precisa y rellenar huecos prolongados preservando la estacionalidad y el ciclo diario de cada variable.

El procedimiento se implementó mediante la función `limpiar_dataset_horario` y consta de los siguientes pasos:

**1. Filtro Físico:** Se aplicaron límites basados en el conocimiento físico de cada variable:

- Para todas las variables analizadas (RS, HUM, VEL, IUV y LLU): cualquier valor negativo se consideró imposible y se marcó como NaN (dato faltante). Ejemplo: una RS de  $-5 \text{ W/m}^2$  indica un fallo electrónico del sensor.
- Límite superior específico (según especificaciones técnicas de los sensores y literatura meteorológica):
- RS: valores  $> 2000 \text{ W/m}^2 \rightarrow \text{NaN}$
- IUV: valores  $> 20 \rightarrow \text{NaN}$

Estos filtros eliminan artefactos evidentes antes de cualquier imputación.

**2. Inputación lineal:** Se utilizó interpolación lineal con un límite máximo de 2 observaciones consecutivas:

```
df[var] = df[var].interpolate(method='linear', limit=2, limit_direction='both')
```

Esta técnica es especialmente adecuada para series temporales meteorológicas porque:

- Asume una evolución gradual entre dos valores conocidos (ej. si a las 10:00 h hay 20 °C y a las 12:00 h hay 22 °C, el valor de las 11:00 h se estima en 21 °C).
- Evita la introducción de sesgos fuertes en huecos muy cortos, que son los más frecuentes por fallos puntuales de transmisión o mantenimiento.

**3. Relleno dinámico estratificado:** Cuando la interpolación lineal no pudo rellenar el hueco (porque supera las 2 horas), se utilizó el promedio histórico de la misma hora y el mismo mes:

```
perfil_mensual_horario = df.groupby([df.index.month,
df.index.hour])[var].transform('mean')
df[var] = df[var].fillna(perfil_mensual_horario)
```

Este método preserva:

- El ciclo diario (patrón de radiación solar, temperatura, etc.).
- La estacionalidad (un enero a las 14:00 h no se rellena con el promedio de julio).

De esta forma se evitó el típico “efecto escalera” o línea plana que produce el simple relleno por media global.

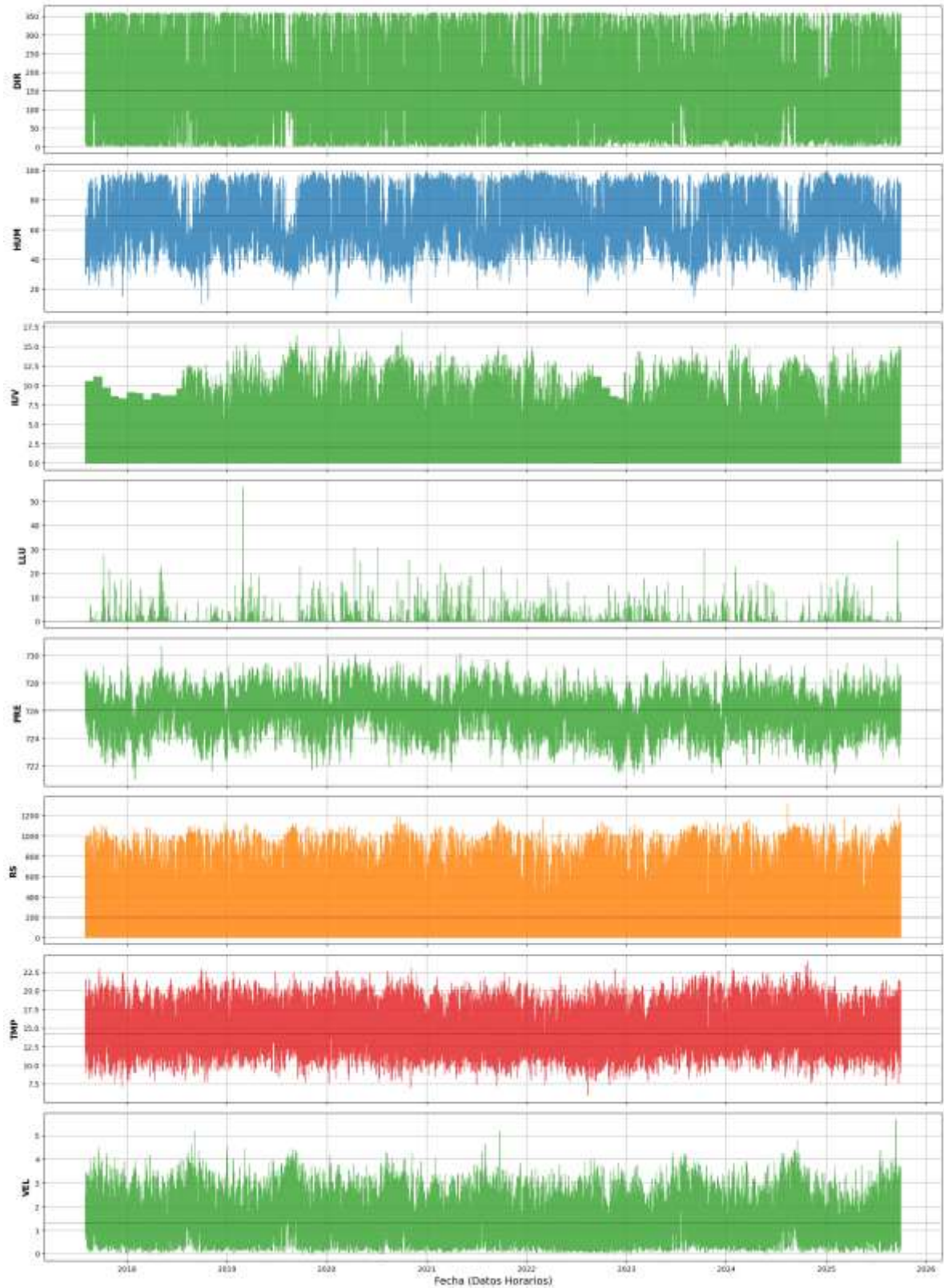
**4. Relleno de emergencia:** En casos muy raros en los que un mes completo carece de datos para cierta hora (por ejemplo, un sensor averiado todo un mes), se recurrió al promedio de esa misma hora a lo largo de todo el año

```
perfil_horario_anual = df.groupby(df.index.hour)[var].transform('mean')
df[var] = df[var].fillna(perfil_horario_anual)
```

Este último paso garantiza que no queden valores faltantes en el dataset final, aunque introduce un mayor nivel de suavizado.

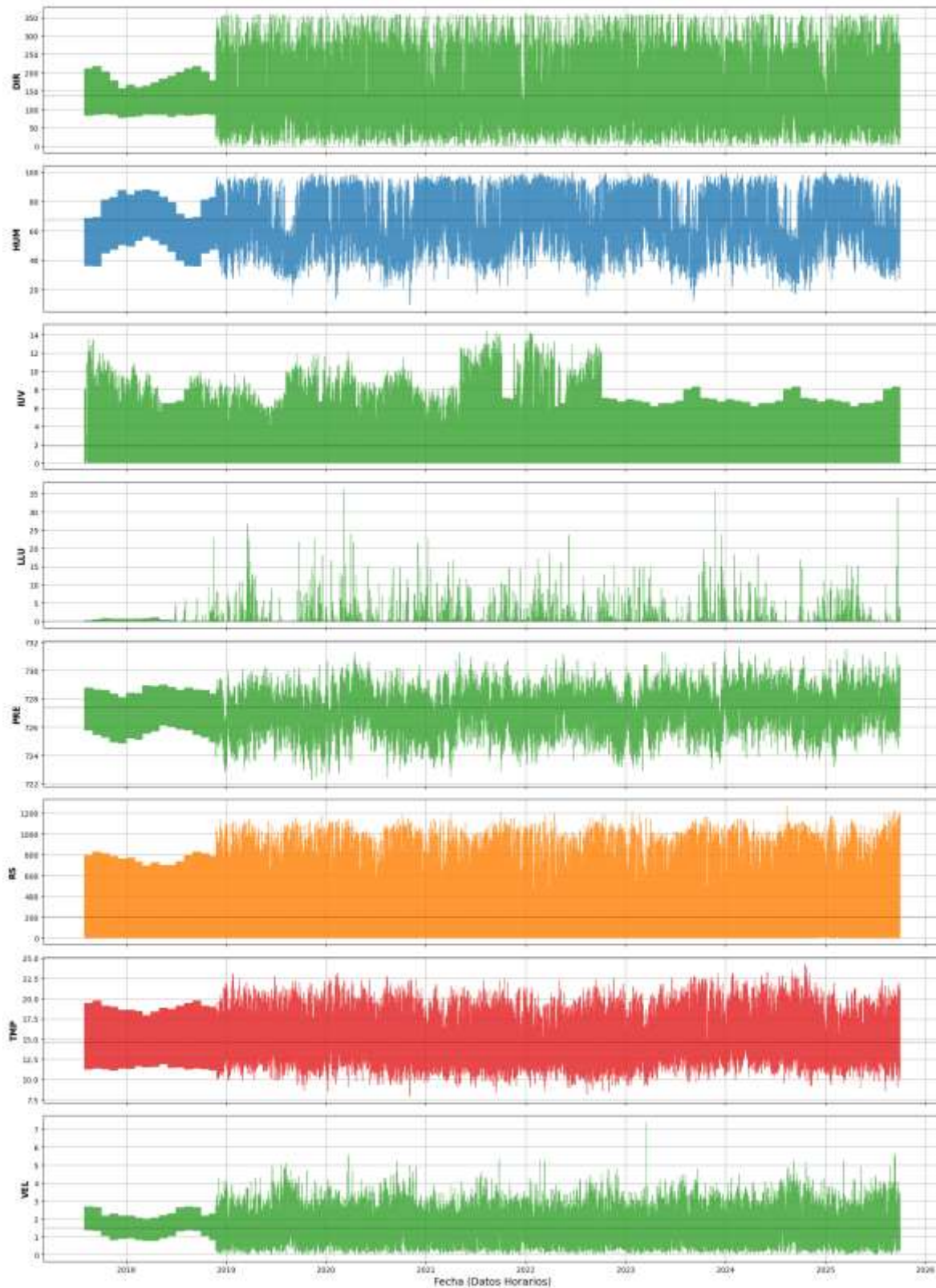
Con este procedimiento se obtuvo un dataset horario limpio con 71593 registros horarios totales, consistente y listo para los análisis posteriores, manteniendo la integridad física y temporal de las series meteorológicas.

# SERIES DE TIEMPO CURADAS: ESTACIÓN BELISARIO



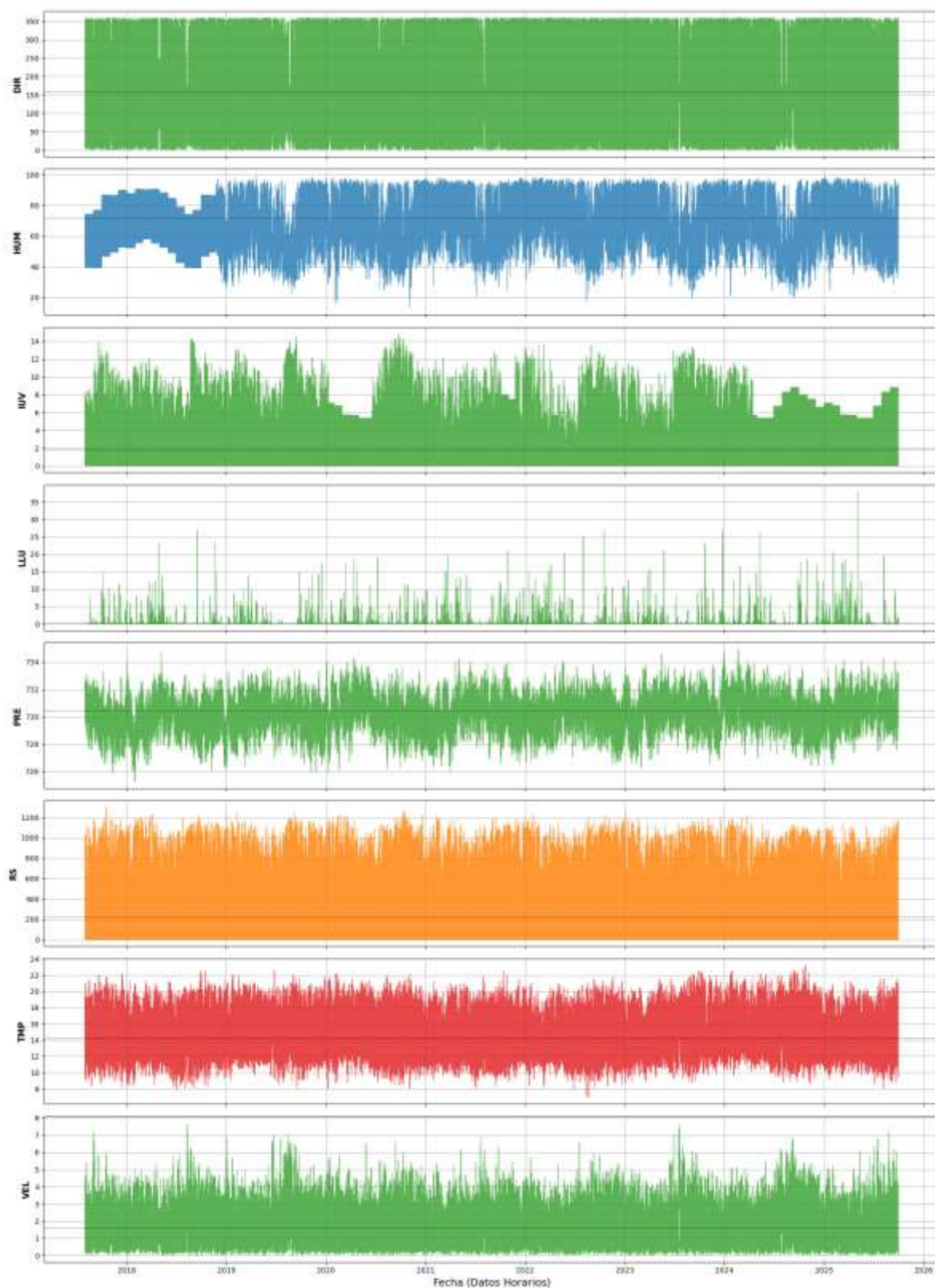
*Figura 5: Series de tiempo de variables meteorológicas curadas de la estación Belisario*

# SERIES DE TIEMPO CURADAS: ESTACIÓN CENTRO



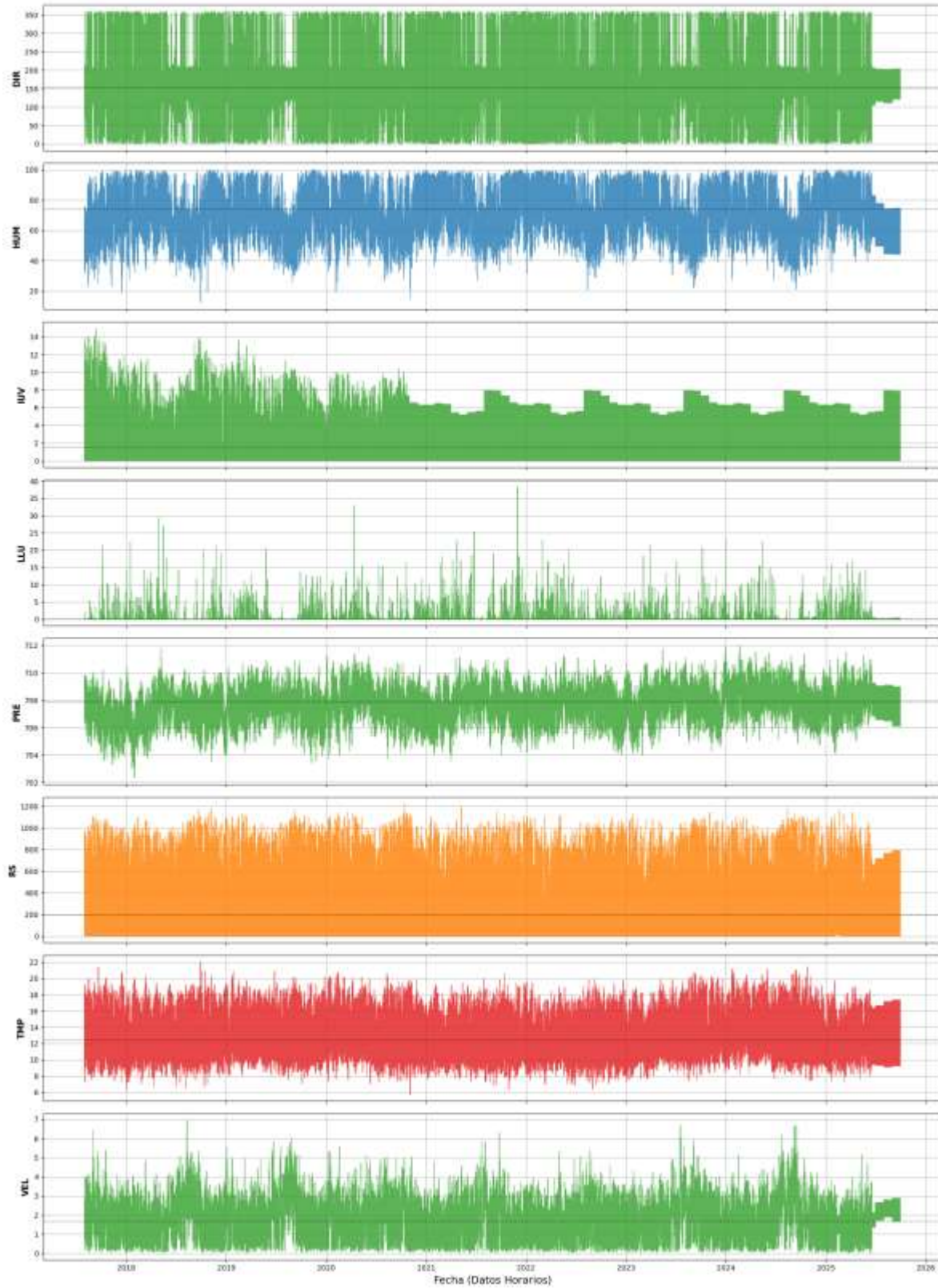
**Figura 6:** Series de tiempo de variables meteorológicas curadas de la estación Centro

# SERIES DE TIEMPO CURADAS: ESTACIÓN COTOCOLLAO



**Figura 7:** Series de tiempo de variables meteorológicas curadas de la estación Cotocollao

# SERIES DE TIEMPO CURADAS: ESTACIÓN GUAMANI

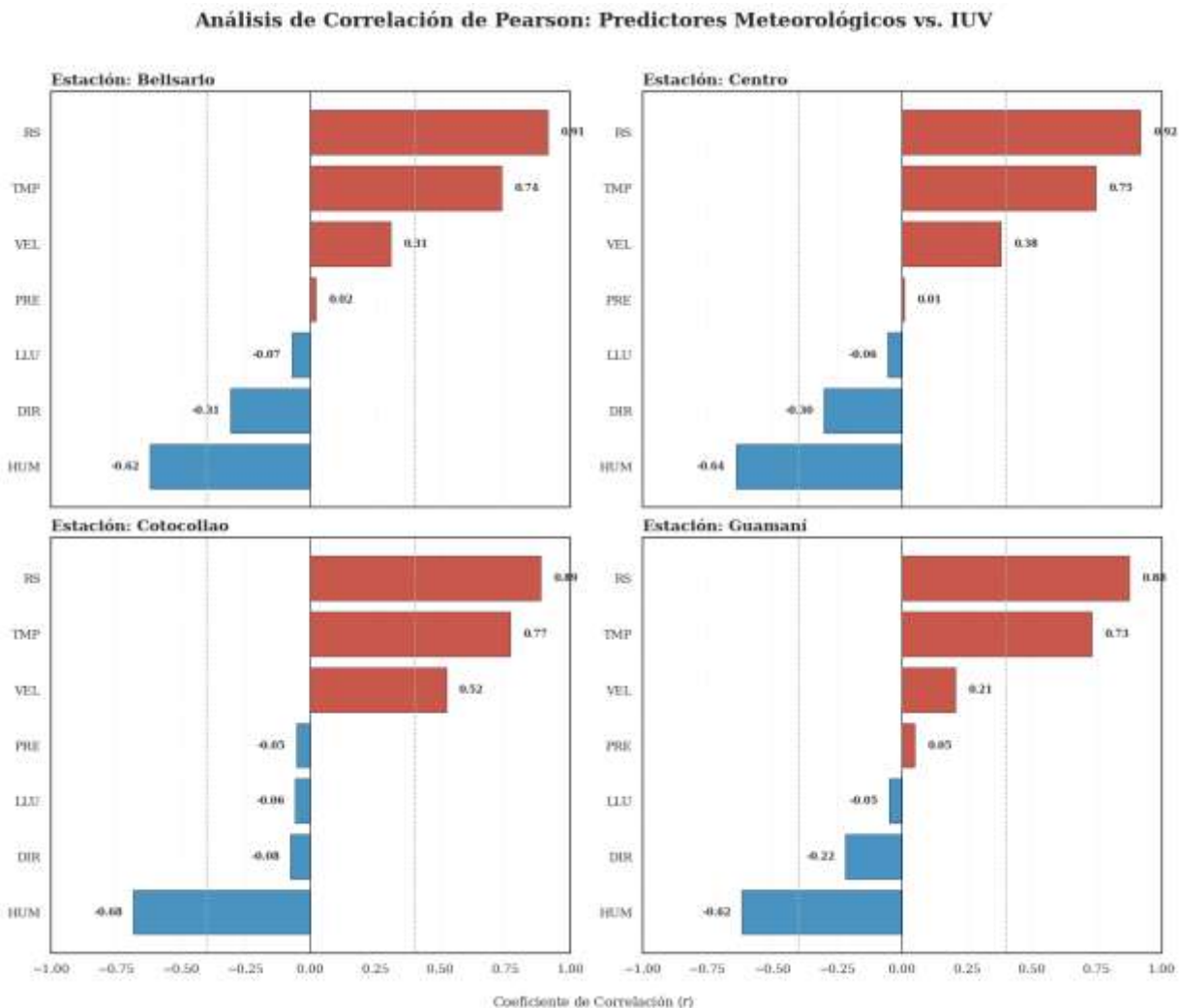


**Figura 8:** Series de tiempo de variables meteorológicas curadas de la estación Guamaní

Las Figuras 5, 6, 7 y 8 presentan las series de datos completas después del proceso de limpieza y relleno, lo que permitió mejorar su calidad y continuidad. De esta manera, se dispone de información más consistente para realizar un análisis integral y construir un modelo más preciso en la estimación del IUV.

### 3.2.2 Análisis de Correlación

Con el objetivo de identificar los predictores con mayor capacidad explicativa para la estimación del IUV, se realizó un análisis de correlación de Pearson para cada una de las estaciones de estudio. Este análisis permite cuantificar el grado de asociación lineal entre las variables meteorológicas y la variable de estudio el IUV, facilitando la selección de características (*feature selection*) para los modelos de aprendizaje automático.



**Figura 9:** Análisis de correlación de Pearson entre las variables de estudio y la de objetivo

### **a. Interpretación de Predictores Principales**

El análisis de correlación de Pearson entre el IUV y las variables meteorológicas (Figura 9) revela patrones consistentes y físicamente coherentes en las cuatro estaciones estudiadas (Belisario, Centro, Cotocollao y Guamaní).

La RS emerge como el predictora dominante, exhibiendo coeficientes de correlación muy altos y positivos en todas las estaciones ( $r = 0.88$  a  $0.92$ ). Esta fuerte asociación lineal es esperada, ya que el IUV representa fundamentalmente la fracción eritemática efectiva de la radiación ultravioleta (principalmente UVB) dentro del espectro solar total. En condiciones de cielo despejado o parcialmente nublado, la radiación global explica la mayor parte de la variabilidad del IUV, lo cual concuerda con estudios previos que reportan correlaciones superiores a  $0.85$ – $0.95$  entre irradiancia solar global y dosis UV eritemática.

La temperatura del aire muestra una correlación positiva moderada-alta ( $r = 0.73$  a  $0.77$ ), reflejando la covariación estacional y diaria: temperaturas más elevadas suelen coincidir con mayor insolación y menor nubosidad, factores que incrementan tanto la radiación incidente como el IUV.

Por el contrario, la humedad relativa presenta una correlación negativa moderada-fuerte y consistente ( $r = -0.62$  a  $-0.68$ ). Este comportamiento se explica principalmente por la relación indirecta entre alta humedad y presencia de nubosidad: las nubes actúan como un filtro efectivo, aumentando la dispersión (especialmente Mie) y la absorción de la radiación UV, lo que reduce significativamente el IUV en superficie. En menor medida, la absorción por vapor de agua también contribuye, aunque su impacto es más relevante en bandas específicas del espectro IR que en UV.

Otras variables muestran asociaciones más débiles o inconsistentes:

- Velocidad del viento: correlación positiva baja a moderada ( $r = 0.21$  a  $0.52$ ), posiblemente relacionada con mayor mezcla atmosférica y menor estancamiento de aerosoles/nubes en días ventosos.

- Presión atmosférica, Precipitación y Dirección del viento: coeficientes cercanos a cero ( $|r| \leq 0.08$  en la mayoría de casos), indicando una dependencia lineal muy baja o nula con el IUV en la escala horaria analizada.

## **b. Selección de Características para el Modelo**

Con el propósito de construir modelos de Machine Learning eficientes, robustos y con menor riesgo de sobreajuste (overfitting), se aplicó un criterio de selección basado en la magnitud de la correlación lineal. Se estableció un umbral conservador de  $|r| > 0.40$ , el cual equilibra la retención de variables con contribución explicativa relevante y la eliminación de predictores con bajo poder predictivo o potencial multicolinealidad.

### **Variables seleccionadas (principales estimadores del IUV):**

- Radiación Solar: predictora dominante por su altísima correlación.
- Temperatura del aire: captura efectos estacionales y diurnos complementarios a RS.
- Humedad Relativa: aporta información sobre nubosidad y atenuación atmosférica, mejorando la explicación en condiciones variables.

### **Variables descartadas:**

- Dirección del viento, precipitación, presión atmosférica y velocidad del viento.

La exclusión de estas variables reduce la dimensionalidad del problema, disminuye el riesgo de multicolinealidad (especialmente entre variables sinérgicas con RS) y mejora la generalización de los modelos, al enfocarse en los drivers físicos más directos del IUV.

Estos resultados son coherentes con la literatura internacional, donde la radiación solar global explica típicamente  $> 80 \%$  de la varianza del IUV en escalas horarias/diarias bajo diferentes regímenes climáticos.

### 3.2.3 Ingeniería de datos

La ingeniería de características constituyó una etapa clave para potenciar el rendimiento predictivo de los modelos de machine learning, especialmente en series temporales meteorológicas donde las relaciones no lineales, los ciclos naturales y las interacciones físicas juegan un rol determinante en la variabilidad del IUV.

Se diseñó un conjunto de transformaciones y variables derivadas que incorporan conocimiento de dominio (física solar, geometría astronómica y procesos atmosféricos) y mejoran la capacidad representativa de los datos originales. Estas transformaciones se aplicaron de forma consistente a los datasets limpios de cada estación mediante la función `aplicar_ingenieria_caracteristicas_pro`.

Las principales categorías y justificaciones de las nuevas características son las siguientes:

#### a. Codificación cíclica de variables temporales

Las variables mes y hora del día poseen una naturaleza cíclica inherente (el mes 12 es contiguo al mes 1, al igual que la hora 23 con la 0). Una codificación numérica lineal introduce discontinuidades artificiales que dificultan el aprendizaje de patrones estacionales y diurnos.

Para superar este problema se aplicó la transformación trigonométrica clásica en machine learning de series temporales:

- **Mes:**

$$\text{MES\_SIN} = \sin(2\pi \times \text{mes} / 12) \quad (3)$$

$$\text{MES\_COS} = \cos(2\pi \times \text{mes} / 12) \quad (4)$$

- **Hora:**

$$\text{HORA\_SIN} = \sin(2\pi \times \text{hora} / 24) \quad (5)$$

$$\text{HORA\_COS} = \cos(2\pi \times \text{hora} / 24) \quad (6)$$

Esta representación bidimensional permite que el modelo capture continuidad y periodicidad de forma natural, mejorando significativamente el aprendizaje de patrones estacionales (ej. mayor IUV en meses de alta insolación) y diurnos (pico alrededor del mediodía solar).

## **b. Cálculo aproximado del coseno del ángulo zenital solar.**

El ángulo zenital solar (SZA) es uno de los determinantes físicos más importantes de la intensidad de la radiación UV en superficie, ya que controla la longitud del trayecto atmosférico y, por ende, la atenuación por absorción y dispersión.

Dado que la latitud de Quito es  $\approx 0^\circ$  (ecuador), se utilizó una aproximación simplificada pero efectiva:

- **Declinación solar ( $\delta$ ):**

$$\delta = 0.409 \times \sin(2\pi \times \text{día del año} / 365 - 1.39) \quad (7)$$

(fórmula estándar de Spencer, 1971, aproximada para uso horario)

- **Ángulo horario (H):**

$$H = (\text{hora} + \text{minuto}/60 - 12) \times (\pi / 12) \quad (8)$$

(H = 0 al mediodía solar local)

- **Coseno del SZA:**

$$\cos(\text{SZA}) \approx \cos(\delta) \times \cos(H) \quad (9)$$

(válido cerca del ecuador, donde el término  $\cos(\text{lat}) \approx 1$ )

Se aplicó un recorte inferior a 0 ( $\text{clip}(\text{lower}=0)$ ) para eliminar valores negativos durante la noche, cuando no hay radiación incidente.

Esta variable actúa como un proxy directo de la geometría solar y es ampliamente utilizada en modelos empíricos y de machine learning para predecir IUV y radiación UV eritemática.

### c. Interacciones físicas y proxies atmosféricos

Se generaron características derivadas que capturan relaciones no lineales y efectos sinérgicos entre variables:

- **RS × TMP (RS\_x\_TMP):**

Representa la interacción entre la energía radiante incidente y la temperatura del aire. Días cálidos con alta radiación suelen coincidir con menor nubosidad y mayor IUV.

- **Transparencia atmosférica (ATMOSFERA\_TRANSPARENCIA = RS / (COS\_SZA + 0.05)):**

Proxy de la claridad atmosférica o factor de transmisión. Bajo cielo despejado, RS se aproxima al valor teórico de cielo claro proporcional a  $\cos(SZA)$ . Valores bajos indican alta atenuación (nubes, aerosoles), lo cual reduce fuertemente el IUV. El pequeño término  $+0.05$  evita divisiones por cero en la noche.

- **RS × HUM (RS\_x\_HUM):**

Captura el efecto moderador de la humedad relativa sobre la radiación UV. Alta humedad suele asociarse a mayor nubosidad y vapor de agua, incrementando la absorción y dispersión UV.

Estas interacciones permiten que modelos lineales y no lineales capturen efectos multiplicativos y condicionales que no son evidentes en las variables individuales.

### d. Codificación categórica de la estación

Dado que se entrenaron modelos tanto por estación como globales (multi-estación), se incluyó una codificación ordinal simple:

- Belisario  $\rightarrow 1$
- Centro  $\rightarrow 2$
- Cotacollao  $\rightarrow 3$
- Guamaní  $\rightarrow 4$

Esta variable permite al modelo aprender efectos fijos, sin introducir alta dimensionalidad.

Las transformaciones incorporaron conocimiento físico-astronómico y cíclico, reduciendo la dependencia del modelo en patrones espurios y mejorando su capacidad para generalizar en diferentes condiciones de cielo, estaciones del año y horarios. En particular:

- COS\_SZA y ATMOSFERA\_TRANSPARENCIA aportan información sobre la trayectoria óptica y la claridad atmosférica, clave para el IUV.
- Las codificaciones cíclicas eliminan discontinuidades artificiales en los patrones temporales.
- Las interacciones explícitas facilitan el aprendizaje de relaciones no lineales incluso en modelos más simples.

Con este conjunto ampliado de características se procedió al entrenamiento y evaluación de los modelos predictivos.

### **3.3 Modelado**

Una vez preparados los datos limpios y enriquecidos con las características derivadas (sección 3.1.3), se procedió al entrenamiento y evaluación comparativa de varios modelos de regresión para predecir el IUV horario en las cuatro estaciones meteorológicas de Quito. El objetivo fue identificar el algoritmo que ofreciera el mejor equilibrio entre precisión, generalización y robustez frente a la variabilidad atmosférica (nubosidad, ciclo solar, efectos estacionales).

### 3.3.1 Preparación de los datos para el entrenamiento

Se construyó un conjunto de datos global (df\_global) concatenando los registros horarios de las cuatro estaciones (Belisario, Centro, Cotocollao y Guamani) tras aplicar la limpieza y la ingeniería de características. Esto permitió entrenar modelos tanto globales (que capturan patrones comunes) como evaluar su transferibilidad entre estaciones.

Las variables predictoras (X) incluyeron las características originales y derivadas (RS, TMP, HUM, COS\_SZA, ATMOSFERA\_TRANSPARENCIA, interacciones físicas, codificaciones cíclicas MES\_SIN/MES\_COS y HORA\_SIN/HORA\_COS, y el código de estación). La variable objetivo (y) fue el IUUV.

La división de los datos se realizó mediante:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Se utilizó una partición estratificada temporal implícita (random\_state fijo) y un 20 % de los datos reservados para prueba, preservando la proporción representativa de condiciones diurnas/nocturnas y estacionales.

### 3.3.2 Selección y configuración de modelos

Se evaluaron cuatro algoritmos representativos de diferentes familias de aprendizaje supervisado:

- **Regresión Lineal Múltiple (RLM):** Baseline lineal simple, sensible a las relaciones lineales dominantes (especialmente RS-IUUV) y útil como referencia.
- **Random Forest Regressor (RF):** Ensemble de árboles de decisión con bagging, robusto a ruido, no linealidades y multicolinealidad; maneja bien interacciones sin necesidad de escalado explícito.
- **LightGBM Regressor:** Algoritmo de gradient boosting basado en histogramas, eficiente en datasets grandes y con alta cardinalidad; destaca por su velocidad y rendimiento en problemas de regresión meteorológica.
- **XGBoost Regressor:** Gradient boosting optimizado con regularización L1/L2, manejo nativo de valores faltantes y gran capacidad para capturar patrones complejos.

Las configuraciones básicas fueron:

- RF: 100 estimadores.
- LightGBM: 100 estimadores, learning\_rate=0.1, num\_leaves=31.
- XGBoost: objective='reg:squarederror', 100 estimadores por defecto.

No se realizó ajuste exhaustivo de hiperparámetros en esta etapa inicial (grid search u Optuna se reservan para el modelo final seleccionado).

### 3.3.3 Métricas de evaluación

El desempeño de cada modelo se cuantificó mediante tres métricas estándar en regresión:

- **Mean Absolute Error (MAE):** Error absoluto promedio entre predicciones y valores reales. Se interpreta directamente en la escala del IUV (0–20 aprox.); menor valor indica mayor precisión promedio. Es robusto frente a outliers.

$$MAE = (1/n) \sum |y_{real} - y_{pred}| \quad (10)$$

- **Root Mean Squared Error (RMSE):** Raíz cuadrada del error cuadrático medio. Penaliza más los errores grandes (útil en contextos donde subestimar picos altos de IUV es crítico, ej. alertas de salud pública). Mantiene las unidades del IUV.

$$RMSE = \sqrt{\left[ (1/n) \sum (y_{real} - y_{pred})^2 \right]} \quad (11)$$

- **Coefficiente de determinación (R<sup>2</sup>):** Proporción de la varianza del IUV explicada por el modelo (0 a 1; valores cercanos a 1 indican excelente ajuste). Negativo si el modelo es peor que predecir la media.

$$R^2 = 1 - \left[ \frac{\sum (y_{real} - y_{pred})^2}{\sum (y_{real} - \bar{y})^2} \right] \quad (12)$$

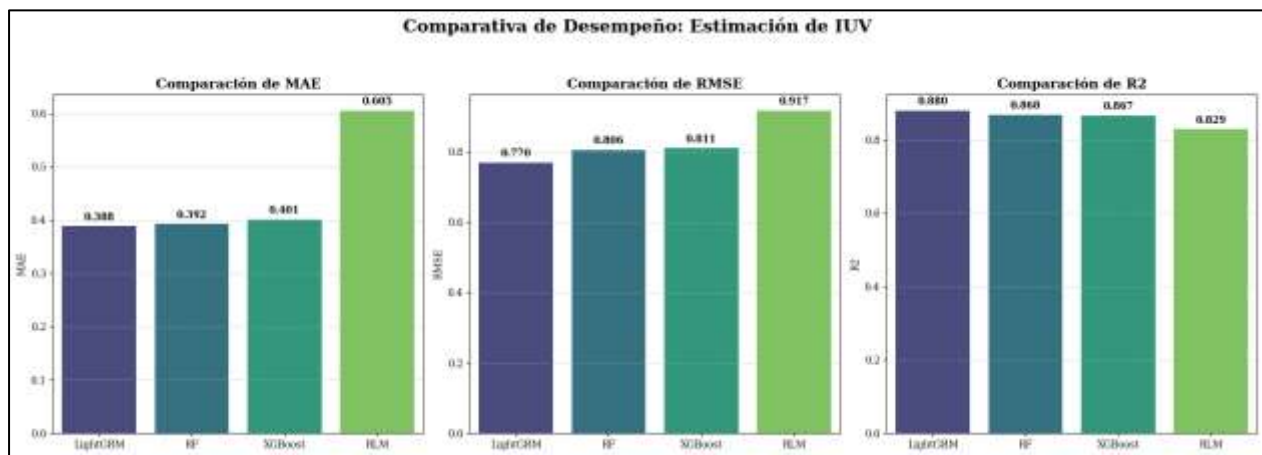
Estas métricas se calcularon sobre el conjunto de prueba para evaluar la capacidad de generalización.

### 3.3.4 Resultados y comparación gráfica

*Tabla 3: Comparación de resultados de modelos de ML*

| Nº | Modelo   | MAE   | RMSE  | $R^2$ |
|----|----------|-------|-------|-------|
| 1  | RLM      | 0.605 | 0.917 | 0.829 |
| 2  | RF       | 0.392 | 0.805 | 0.868 |
| 3  | LigthGBM | 0.388 | 0.769 | 0.879 |
| 4  | XGBoost  | 0.401 | 0.810 | 0.866 |

La Tabla 3 resume las métricas obtenidas en el conjunto de prueba para los cuatro modelos.



*Figura 10: Comparativa de modelos de ML para estimación de IUUV*

La Figura 10 presenta una comparación visual de barras para cada métrica, ordenando los modelos de mejor a peor en cada subgráfico:

- En MAE y RMSE: valores más bajos indican mejor desempeño.
- En  $R^2$ : valores más altos indican mejor ajuste.

### 3.3.5 Discusión preliminar y selección del modelo

Basado en estos resultados iniciales, se seleccionó **LightGBM** como modelo ganador para refinamiento posterior (ajuste de hiperparámetros, validación cruzada temporal, análisis de importancia de variables con SHAP/Permutation Importance). Este algoritmo equilibra alto desempeño predictivo, eficiencia computacional y robustez en datasets meteorológicos con patrones cíclicos y no lineales.

Los resultados son coherentes con estudios previos de predicción de IUV y radiación UV, donde ensembles de gradient boosting (XGBoost, LightGBM) suelen superar a modelos lineales y Random Forest en precisión horaria/multi-horaria.

### 3.4 Evaluación

#### 3.4.1 Evaluación inicial del modelo multi-estación

En la fase inicial, se desarrolló un modelo global que integró los registros horarios de las cuatro estaciones meteorológicas (Belisario, Centro, Cotocollao y Guamaní), con el objetivo de capturar patrones climáticos transversales y maximizar el volumen de datos.

Tras la evaluación comparativa (Sección 3.1.4), LightGBM se consolidó como el algoritmo superior con un  $R^2 = 0.879$ ,  $RSME = 0.769$  y  $MAE = 0.388$ . Seleccionado como el modelo base, se procedió a su refinamiento mediante las siguientes estrategias:

- Ajuste de hiperparámetros mediante RandomizedSearchCV con 20 iteraciones.
- Validación cruzada temporal (TimeSeriesSplit con 3 splits) para respetar la estructura secuencial de los datos y evitar leakage temporal.
- Entrenamiento final con early stopping basado en un conjunto de validación cronológico (15 % final del entrenamiento).

Los resultados finales en el conjunto de prueba independiente fueron:

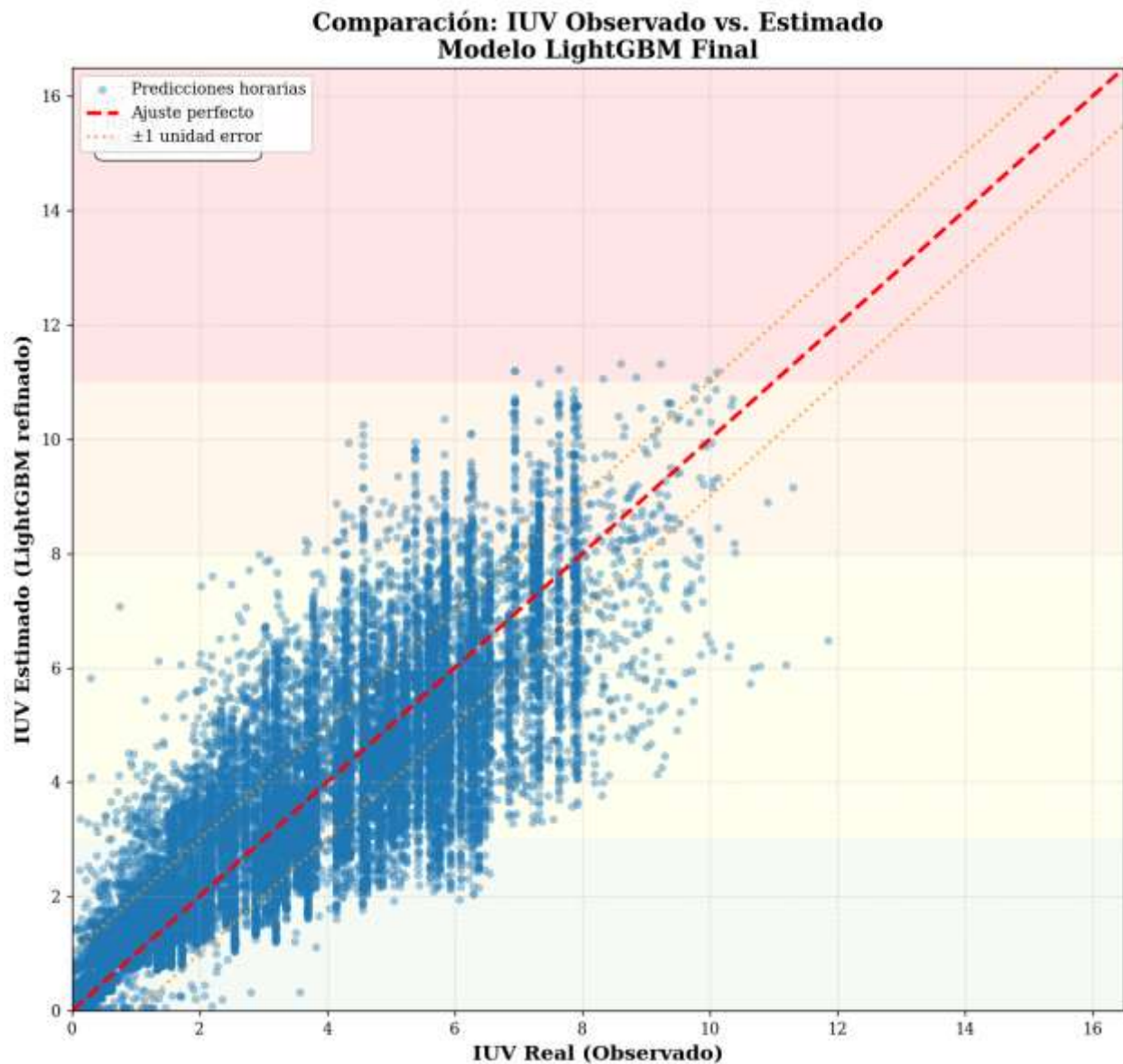
**Tabla 4:** Evaluación final del modelo multi-estación

| Indicador       | Valor |
|-----------------|-------|
| $R^2$           | 0.878 |
| RMSE            | 0.775 |
| MAE             | 0.400 |
| MAE (IUV >11)   | 4.234 |
| Recall (IUV >8) | 0.416 |

Estos indicadores reflejan un desempeño aceptable en promedio, con un coeficiente de determinación que explica aproximadamente el 88 % de la variabilidad observada y errores medios relativamente bajos ( $MAE < 0.5$ ). No obstante, el análisis estratificado reveló limitaciones importantes en los rangos de mayor relevancia práctica: el error absoluto medio en valores

extremos ( $IUV > 11$ ) superó las 4 unidades y el recall para episodios de riesgo alto fue inferior al 42 %. Estos resultados sugieren una subestimación sistemática de los picos de radiación ultravioleta y una capacidad limitada para detectar condiciones peligrosas para la salud pública.

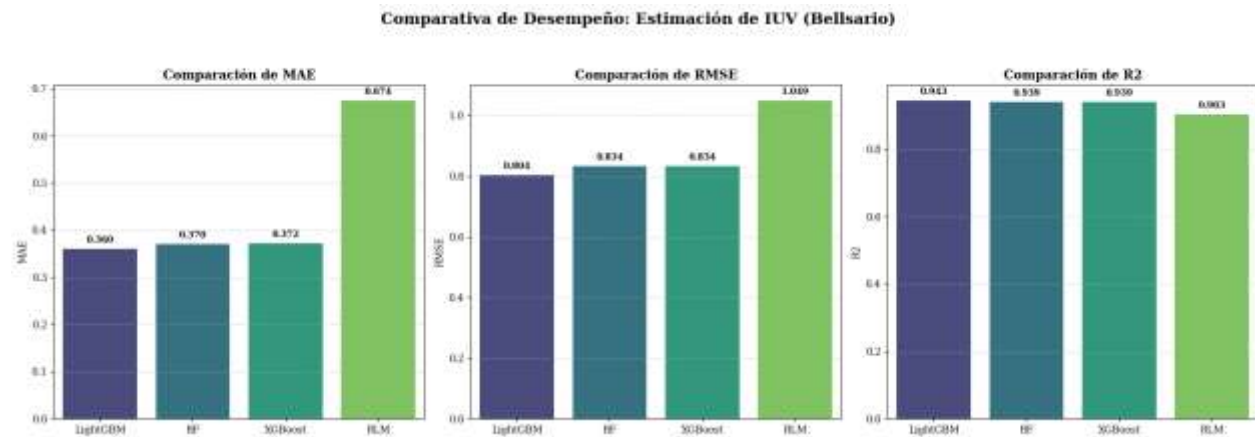
La Figura 11 presenta la comparación gráfica entre los valores observados y estimados en este modelo global, donde se aprecia una dispersión aceptable en rangos bajos y moderados, pero mayor desviación en la zona superior derecha (valores altos/extremos).



**Figura 11:** IUV observado vs estimado con el modelo LightGBM multi-estación

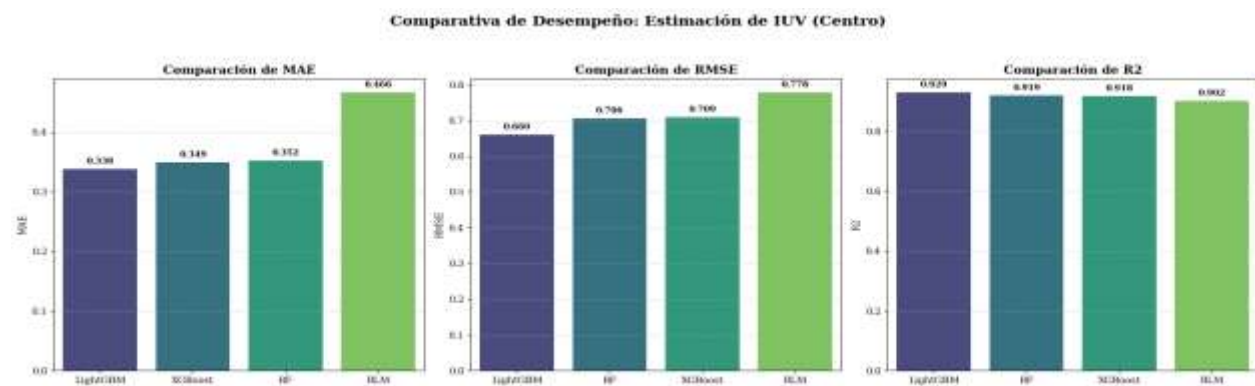
### 3.4.2 Análisis por estación individual y selección del modelo final

Ante las limitaciones observadas en el modelo global, se procedió a evaluar el desempeño predictivo de LightGBM cuando se entrenaba de manera independiente para cada estación, con el fin de determinar si la heterogeneidad espacial y microclimática entre ubicaciones influía significativamente en la calidad de las predicciones.



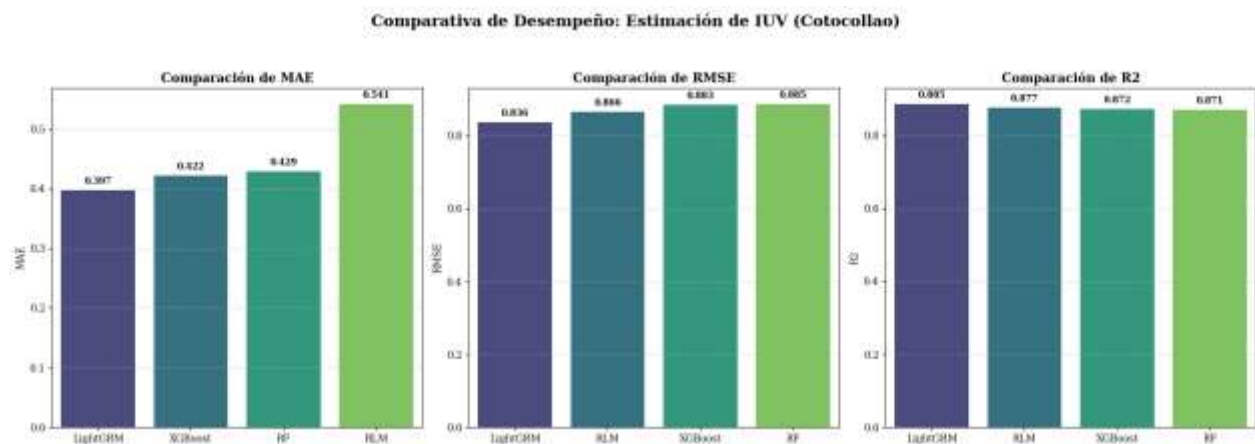
*Figura 12: Estimación IUV para la estación Belisario*

En la Figura 12 se presenta el desempeño comparativo de los distintos modelos de ML para la estimación del IUV en la estación Belisario, destacando al modelo LightGBM como el de mejor rendimiento.



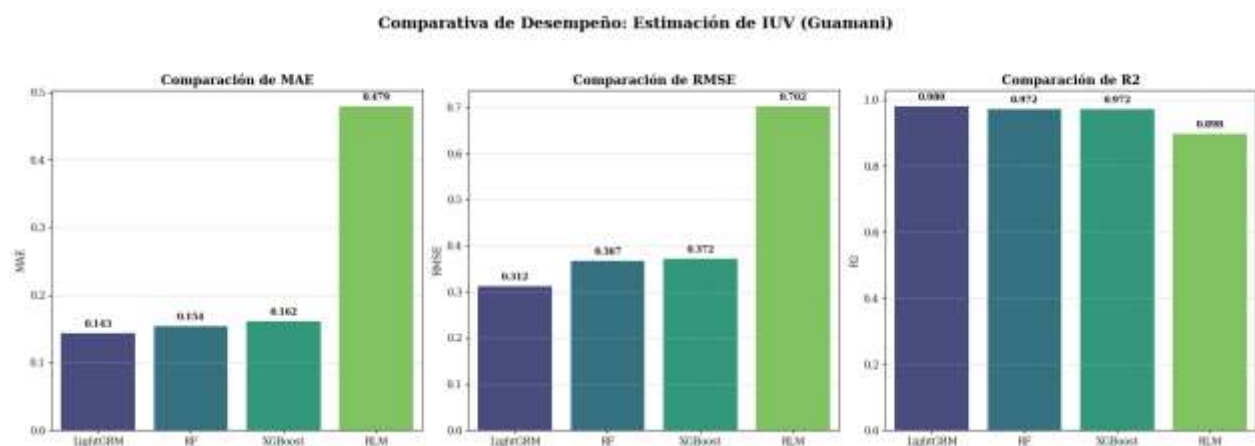
*Figura 13: Estimación IUV para la estación Centro*

En la Figura 13 se presenta el desempeño comparativo de los distintos modelos de ML para la estimación del IUV en la estación Centro, destacando al modelo LightGBM como el de mejor rendimiento.



**Figura 14:** Estimación IUV para la estación Cotocollao

En la Figura 14 se presenta el desempeño comparativo de los distintos modelos de ML para la estimación del IUV en la estación Cotocollao, destacando al modelo LightGBM como el de mejor rendimiento.



**Figura 15:** Estimación IUV para la estación Guamaní

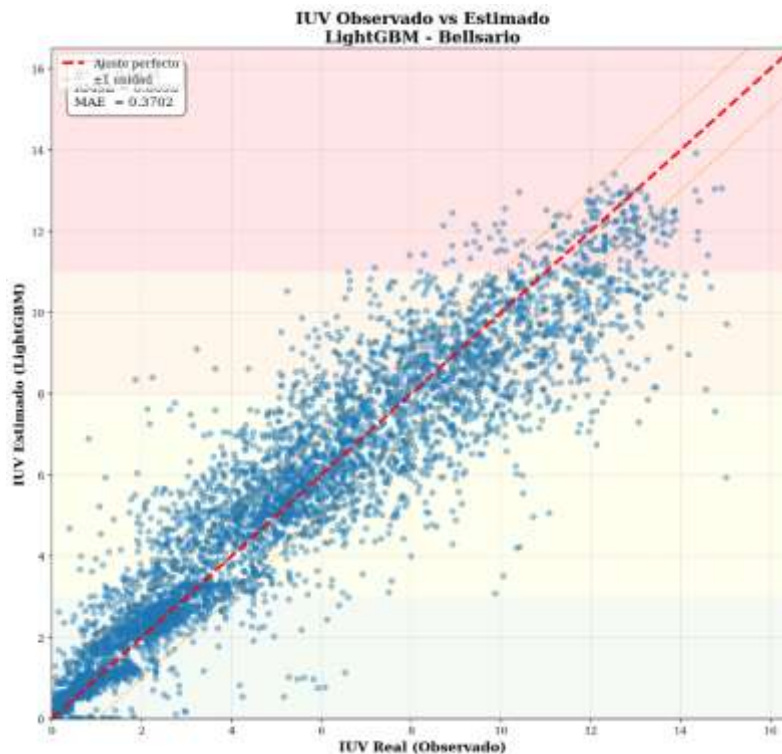
En la Figura 15 se presenta el desempeño comparativo de los distintos modelos de ML para la estimación del IUV en la estación Guamaní, destacando al modelo LightGBM como el de mejor rendimiento.

En la Tabla 5 se presentan los resultados obtenidos tras la evaluación por estación de cada modelo de ML. El modelo LightGBM se destacó como el ganador en todas las estaciones analizadas. Asimismo, la estación Guamaní registró los mejores valores de MAE, RMSE y  $R^2$  hasta esta etapa, previo al proceso de refinamiento.

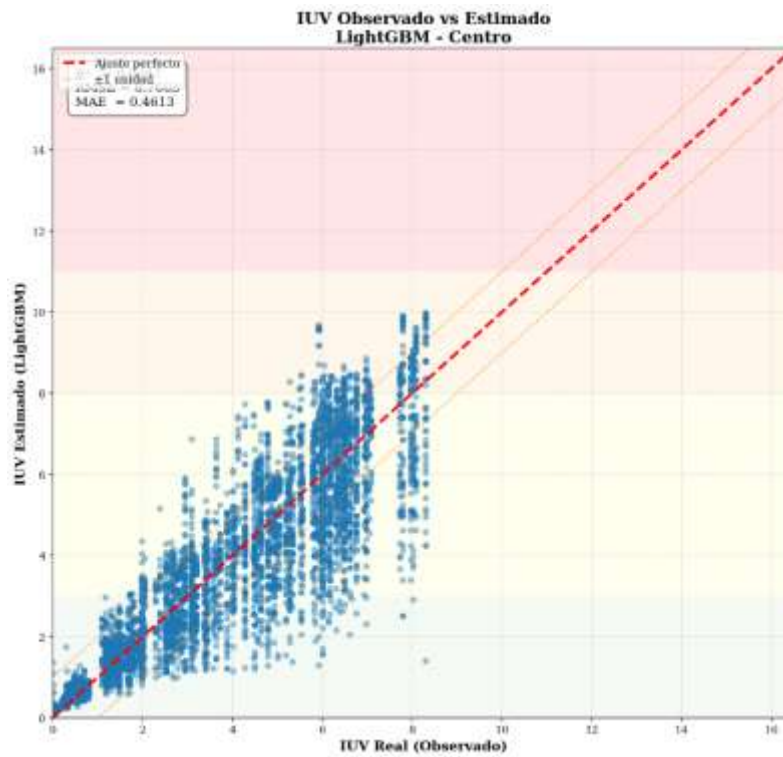
**Tabla 5:** Resultados de la evaluación de modelos de ML por cada estación

| Estación          | Modelo   | MAE   | RMSE  | R <sup>2</sup> |
|-------------------|----------|-------|-------|----------------|
| <b>Belisario</b>  | RLM      | 0.674 | 1.048 | 0.903          |
|                   | RF       | 0.370 | 0.833 | 0.938          |
|                   | LightGBM | 0.359 | 0.804 | 0.943          |
|                   | XGBoost  | 0.372 | 0.834 | 0.938          |
| <b>Centro</b>     | RLM      | 0.466 | 0.777 | 0.901          |
|                   | RF       | 0.352 | 0.706 | 0.919          |
|                   | LightGBM | 0.337 | 0.660 | 0.929          |
|                   | XGBoost  | 0.348 | 0.709 | 0.918          |
| <b>Cotocollao</b> | RLM      | 0.541 | 0.865 | 0.876          |
|                   | RF       | 0.429 | 0.885 | 0.871          |
|                   | LightGBM | 0.396 | 0.836 | 0.885          |
|                   | XGBoost  | 0.421 | 0.883 | 0.871          |
| <b>Guamaní</b>    | RLM      | 0.478 | 0.701 | 0.898          |
|                   | RF       | 0.153 | 0.367 | 0.972          |
|                   | LightGBM | 0.143 | 0.312 | 0.979          |
|                   | XGBoost  | 0.161 | 0.371 | 0.971          |

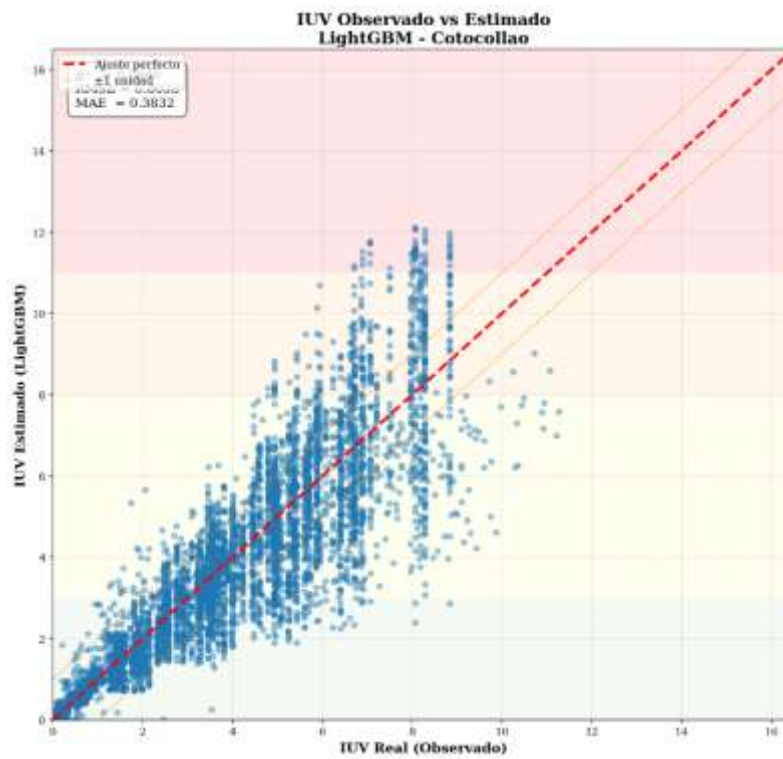
Se aplicó el mismo procedimiento de refinamiento del modelo multi-estación (validación temporal, early stopping y parámetros optimizados) a los datos filtrados de cada estación.



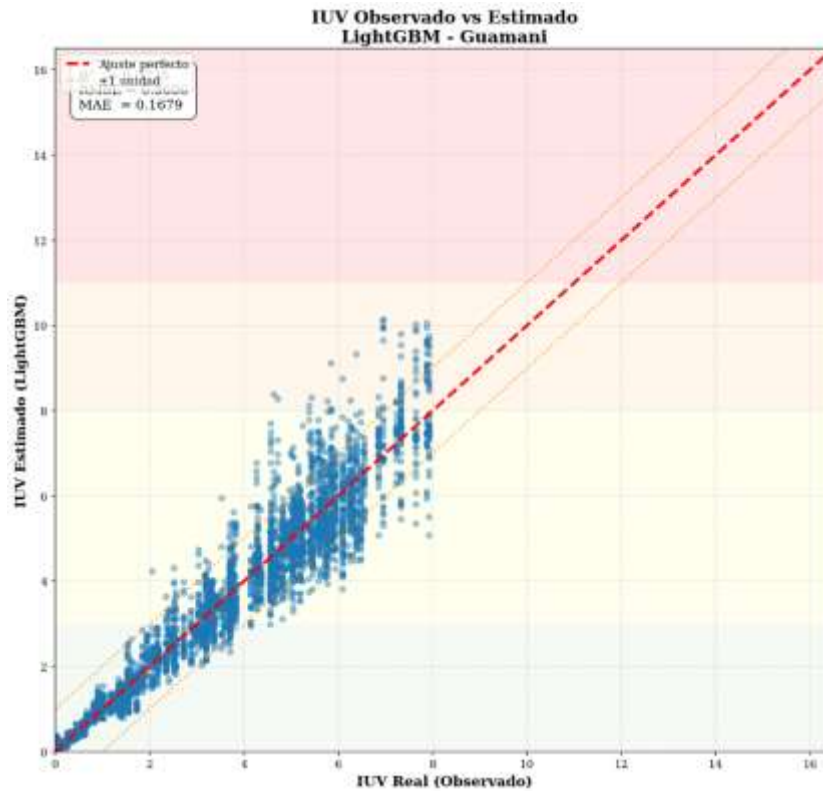
**Figura 16:** IUV observado vs estimado con el modelo LightGBM para la estación Belisario



**Figura 17:** IUV observado vs estimado con el modelo LightGBM para la estación Centro



**Figura 18:** IUV observado vs estimado con el modelo LightGBM para la estación Cotocollao



**Figura 19:** IUV observado vs estimado con el modelo LightGBM para la estación Guamaní

Las Figuras 16, 17, 18 y 19 presenta la comparación entre los valores observados y estimados del IUV mediante el modelo LightGBM en cuatro estaciones de Quito: Belisario, Centro, Cotocollao y Guamaní. Cada gráfico incluye la línea de ajuste perfecto ( $y=x$ ), la línea de ajuste del modelo y el valor del Mean Absolute Error (MAE) como métrica de desempeño.

Los resultados muestran diferencias relevantes entre estaciones. En términos de error promedio, Guamaní presenta el menor MAE (0.1679), lo que refleja un ajuste muy preciso en valores moderados de IUV. Sin embargo, el análisis detallado evidencia que la estación Belisario constituye la mejor referencia para el modelo. A pesar de tener un MAE ligeramente mayor (0.3702), Belisario se distingue por su capacidad de estimar con mayor fidelidad los valores altos de IUV, aspecto crítico en la evaluación del riesgo de exposición a radiación ultravioleta. En escenarios donde el IUV alcanza niveles extremos, contar con un modelo que reproduzca adecuadamente dichos valores resulta más relevante que minimizar el error promedio en rangos bajos o intermedios.

El modelo en Belisario mantiene parámetros de evaluación sólidos y una tendencia de ajuste cercana a la diagonal de ajuste perfecto, especialmente en los rangos superiores del IUV. Esto indica que el algoritmo logra capturar de manera consistente las condiciones críticas, evitando la subestimación de valores extremos. En contraste, el sector Centro presenta el mayor error (MAE = 0.4613), lo que sugiere una menor capacidad predictiva asociada a condiciones locales más complejas. Cotocollao (MAE = 0.3832) muestra un desempeño intermedio, mientras que Guamaní, aunque preciso en valores moderados, no refleja la misma sensibilidad en los niveles altos de IUV.

En consecuencia, se concluye que Belisario es la estación con mejor desempeño del modelo LightGBM, dado que combina un error aceptable con la capacidad de estimar adecuadamente los valores críticos de IUV. Este hallazgo tiene implicaciones prácticas y científicas, desde la perspectiva de salud pública, es preferible un modelo que no subestime los niveles elevados de radiación, ya que estos representan mayor riesgo para la población. Asimismo, metodológicamente, este resultado refuerza la importancia de evaluar el desempeño de los modelos de aprendizaje automático en contextos locales y bajo criterios de relevancia práctica, más allá de métricas globales de error.

Los resultados clave se resumen en la Tabla 6

**Tabla 6:** Resultados de la evaluación de modelos de ML por cada estación aplicando refinamiento

| Estación              | R <sup>2</sup> | RMSE  | MAE   | MAE (IUV >11) | Recall (IUV >8) |
|-----------------------|----------------|-------|-------|---------------|-----------------|
| <b>Belisario</b>      | 0.942          | 0.809 | 0.370 | 1.737         | 0.807           |
| <b>Centro</b>         | 0.904          | 0.766 | 0.461 | N/A           | 0.511           |
| <b>Cotocollao</b>     | 0.894          | 0.803 | 0.383 | 3.456         | 0.513           |
| <b>Guamani</b>        | 0.972          | 0.365 | 0.167 | N/A           | 0               |
| <b>Global (todas)</b> | 0.878          | 0.775 | 0.400 | 4.234         | 0.416           |

La estación Belisario alcanzó un R<sup>2</sup> de 0.942 (mejora de ~7.4 % respecto al modelo global), un MAE global de 0.370 y, lo más relevante, un desempeño drásticamente superior en rangos críticos: MAE en IUV >11 reducido a 1.737 (mejora de ~60 %) y recall para IUV >8 elevado a 0.807 (casi el doble del modelo global).

Esta mejora sustancial se atribuye principalmente a la menor heterogeneidad de los datos en Belisario, posiblemente relacionada con:

- Mayor consistencia en las mediciones (menos valores atípicos o missing tras limpieza).

- Menor interferencia por efectos urbanos locales (contaminación, aerosoles, sombras topográficas) en comparación con estaciones céntricas o más bajas.
- Mejor alineación entre las variables predictoras físicas (RS, COS\_SZA, ATMOSFERA\_TRANSPARENCIA) y el IUV observado, lo que permite al modelo capturar patrones más limpios y predecibles.

La Figura 16 muestra la comparación observada vs. estimada para el modelo Belisario, evidenciando un ajuste mucho más ajustado a la línea de identidad, especialmente en la zona de valores altos, y una dispersión notablemente menor que en el modelo global.

### 3.4.3 Selección del modelo final para despliegue

Dado que el objetivo principal del trabajo es obtener una herramienta de predicción precisa y confiable para la estimación del Índice Ultravioleta en Quito, con especial énfasis en la detección de condiciones de riesgo para las afectaciones de ecosistemas y la salud, se selecciona el modelo LightGBM entrenado exclusivamente con los datos de la estación Belisario como el definitivo para el despliegue.

Este modelo ofrece:

- Alto poder explicativo ( $R^2 = 0.942$ ).
- Errores bajos en promedio (MAE = 0.370, RMSE = 0.809).
- Desempeño adecuado en rangos críticos (MAE extremo  $\approx 1.74$ , recall  $>8 \approx 0.81$ ), lo cual lo hace viable para aplicaciones de alerta sanitaria y monitoreo ambiental.

Aunque el enfoque multi-estación podría ser preferible en escenarios con mayor cobertura geográfica o datos más homogéneos, los resultados demuestran que, en el contexto específico de Quito, la modelización por estación individual (particularmente Belisario) proporciona mayor precisión y utilidad práctica. Este hallazgo resalta la importancia de considerar la variabilidad espacial y microclimática en estudios de radiación ultravioleta en regiones montañosas tropicales.

El modelo seleccionado será implementado en la etapa de despliegue, permitiendo generar predicciones en tiempo real o bajo demanda para la estación Belisario como referencia principal para la ciudad.

### 3.5 Despliegue

#### 3.5.1 Objetivos y enfoque del despliegue

Una vez finalizada la etapa de modelado y evaluación, se procedió al desarrollo de una aplicación web interactiva que permite la visualización y actualización en tiempo real del Índice Ultravioleta (IUV) estimado para las 33 estaciones meteorológicas consideradas en el estudio (Tabla 7). El objetivo principal fue transformar el modelo predictivo LightGBM (seleccionado y optimizado exclusivamente con datos de la estación Belisario) en una herramienta práctica, accesible y útil para monitoreo ambiental y toma de decisiones sanitarias en Quito.

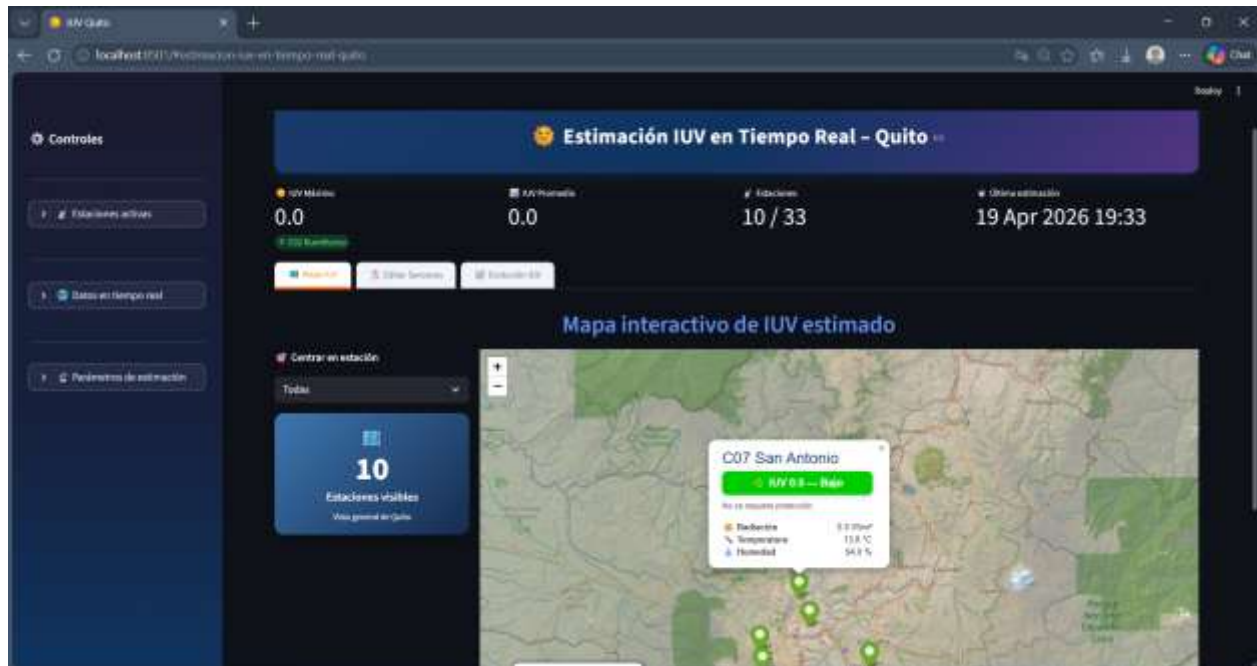
*Tabla 7: Estaciones para estimación de IUV en el DMQ*

| Nº | Nombre                | Latitud      | Longitud     |
|----|-----------------------|--------------|--------------|
| 1  | E03 Belisario         | -0.184719    | -78.495986   |
| 2  | E05 El Camal          | -0.250000    | -78.510000   |
| 3  | E06 Centro            | -0.221393    | -78.514005   |
| 4  | E08 Tumbaco           | -0.214933    | -78.403253   |
| 5  | E09 Los Chillos       | -0.297062    | -78.455248   |
| 6  | E10 Cotacollao        | -0.107777    | -78.497222   |
| 7  | E11 Carapungo         | -0.095472    | -78.449809   |
| 8  | E12 Guamaní           | -0.334253531 | -78.55316696 |
| 9  | E13 San Antonio       | -0.009222    | -78.448001   |
| 10 | C01 Maucatanbo        | -0.680536    | -78.34358    |
| 11 | C02 Rumihurco         | -0.12975     | -78.537476   |
| 12 | C04 Rumipamba         | -0.175415    | -78.531287   |
| 13 | C05 Bellavista        | -0.181639    | -78.4617     |
| 14 | C06 Yaruquí           | -0.164357    | -78.306428   |
| 15 | C07 San Antonio       | -0.018408    | -78.45541    |
| 16 | C08 IASA              | -0.391952    | -78.414397   |
| 17 | C10 La Mica Presa     | -0.546444    | -78.228218   |
| 18 | C11 Pita Campamento   | -0.496508    | -78.438100   |
| 19 | C12 Quijos Campamento | -0.472885    | -78.046635   |
| 20 | C13 Salve Faccha      | -0.230673    | -78.152494   |
| 21 | C14 Mindo Captación   | -0.149422    | -78.58812    |
| 22 | C15 Nanegalito        | 0.059827     | -78.685323   |
| 23 | C16 Guaytaloma        | -0.297609    | -78.52       |
| 24 | C18 Puengasí          | -0.236666    | -78.49309    |
| 25 | C19 El Troje          | -0.333371    | -78.523589   |
| 26 | C20 Calderón          | -0.0777329   | -78.433366   |
| 27 | C21 El Cinto          | -0.248295    | -78.569833   |

| Nº | Nombre                | Latitud    | Longitud    |
|----|-----------------------|------------|-------------|
| 28 | C22 Gualea Cruz       | 0.104583   | -78.7377839 |
| 29 | C23 La Delicia        | 0.14340772 | -78.803389  |
| 30 | C24 Itchimbia         | -0.2214536 | -78.4998945 |
| 31 | C25 San José de Minas | 0.173397   | -78.413226  |
| 32 | C26 La Granja         | -0.185533  | -78.497213  |
| 33 | C17 La Borrega        | 0.038798   | -78.695249  |

El despliegue se diseñó con los siguientes principios:

- Facilidad de uso: interfaz intuitiva para ingresar valores manuales de sensores (RS, TMP, HUM) en tiempo real.
- Visualización geográfica: mapa interactivo que muestra las 33 estaciones con su ubicación real y predicción actualizada.
- Interpretabilidad sanitaria: indicadores visuales de riesgo (colores) y recomendaciones preventivas basadas en estándares de la OMS.
- Flexibilidad: actualización inmediata de predicciones sin necesidad de recargar la aplicación.
- Simplicidad tecnológica: implementación ligera, reproducible y con bajo costo de mantenimiento.



*Figura 20: Despliegue del dashboard para la estimación de IUV en el DMQ*

### 3.5.2 Tecnologías utilizadas

La aplicación se desarrolló utilizando las siguientes herramientas y bibliotecas:

- **Streamlit:** Framework de Python para la creación rápida de aplicaciones web interactivas con interfaz de usuario amigable y sin necesidad de conocimientos avanzados de frontend.
- **Folium + streamlit-folium:** Generación de mapas interactivos basados en Leaflet.js, con soporte para marcadores personalizados, popups y capas de fondo satelital (tiles de Esri World Imagery).
- **Joblib:** Serialización y carga eficiente del modelo LightGBM entrenado.
- **Pandas y NumPy:** Manipulación de datos tabulares y cálculos numéricos para la ingeniería de características en tiempo real.
- **Python 3.10+:** Lenguaje base de la aplicación.

No se emplearon frameworks adicionales de backend (como FastAPI o Flask) ni bases de datos persistentes, ya que el caso de uso se centró en la edición manual y visualización inmediata de predicciones, sin requerir almacenamiento histórico ni autenticación de usuarios.

### 3.5.3. Arquitectura y flujo de la aplicación

La aplicación se estructura en dos pestañas principales:

#### 1. Pestaña “Mapa”:

- Muestra un mapa satelital centrado en Quito con zoom inicial de nivel 11.
- Incluye las 33 estaciones georreferenciadas (coordenadas hardcodeadas en el código).
- Cada marcador cambia de color según el rango de IUV estimado (verde, azul, naranja, rojo y púrpura).
- Al hacer clic en un marcador, se despliega un popup con:
  - Nombre y código de la estación.
  - Valor de IUV estimado.
  - Nivel de riesgo.
  - Medidas preventivas recomendadas según estándares OMS.

#### 2. Pestaña “Editar Sensores”:

- Tabla editable (usando st.data\_editor) que muestra las 33 estaciones.
- Columnas editables: RS ( $W/m^2$ ), TMP ( $^{\circ}C$ ) y HUM (%).
- Columnas fijas (deshabilitadas): código, nombre, latitud y longitud.
- Botón “Actualizar Estimaciones” que ejecuta el pipeline de predicción para todas las filas.

#### Flujo de operación:

1. El usuario abre la aplicación.
2. El mapa se carga automáticamente con las 33 estaciones (sin predicciones iniciales).
3. El usuario pasa a la pestaña de edición, modifica los valores de sensores para las estaciones deseadas (simulando lecturas en tiempo real).
4. Presiona “Actualizar Estimaciones”.
5. La función de predicción calcula las 12 features requeridas por el modelo (incluyendo componentes cíclicas, COS\_SZA y las interacciones físicas).
6. Se generan las estimaciones de IUV y el mapa se actualiza con los nuevos valores y colores.

### 3.5.4. Características técnicas destacadas

- **Ingeniería de características en tiempo real:** Se calculan automáticamente las variables derivadas necesarias para el modelo (MES\_SIN, MES\_COS, HORA\_SIN, HORA\_COS, COS\_SZA, RS\_x\_TMP, RS\_x\_HUM, ATMOSFERA\_TRANSPARENCIA).
- **Visualización geográfica:** Uso de capa satelital (Esri World Imagery) para mejor contexto urbano y topográfico de Quito. Marcadores con iconos de sol y colores OMS para rápida identificación de riesgo.
- **Persistencia de datos:** Los valores editados se mantienen entre interacciones gracias a `st.session_state`.
- **Rendimiento:** La aplicación es ligera y responde en tiempo real incluso con 33 estaciones, gracias a la eficiencia de LightGBM y la caché de Streamlit.

### 3.5.5. Limitaciones y consideraciones

- El modelo utilizado es específico de la estación Belisario, por lo que las predicciones para otras estaciones deben interpretarse como aproximaciones basadas en patrones similares.
- La entrada de datos es manual y simulada; en una implementación operativa se podría integrar con APIs de estaciones en tiempo real o sensores IoT.
- No se implementó autenticación ni persistencia de historial de predicciones, ya que el enfoque fue demostrativo y centrado en la visualización interactiva.

### 3.5.6. Conclusiones del despliegue

El despliegue desarrollado constituye una herramienta práctica y visualmente atractiva que transforma el modelo entrenado en un producto usable. Permite a cualquier usuario (investigadores, autoridades sanitarias, ciudadanos) simular condiciones reales, observar la distribución espacial del IUV y tomar decisiones informadas sobre protección solar en Quito. La aplicación puede ser desplegada de forma gratuita en plataformas como Streamlit Community Cloud, facilitando su acceso y posible escalabilidad futura.

## CAPITULO IV: CONCLUSIONES Y RECOMENDACIONES

### 4.1 Conclusiones

El presente trabajo tuvo como objetivo principal desarrollar un modelo robusto y confiable para la estimación del Índice Ultravioleta en el DMQ, utilizando datos meteorológicos de múltiples estaciones y técnicas avanzadas de aprendizaje automático.

Entre los principales hallazgos se destacan:

- La radiación solar global (RS) resultó ser el predictor dominante en todas las estaciones, con correlaciones de Pearson superiores a 0.88, seguido de la temperatura del aire (TMP) y la humedad relativa (HUM) como variables complementarias clave. Las interacciones físicas derivadas ( $RS \times TMP$ ,  $RS \times HUM$ ) y proxies atmosféricos (transparencia atmosférica, coseno del ángulo zenital) mejoraron significativamente la capacidad explicativa del modelo.
- El modelo LightGBM, tras refinamiento de hiperparámetros y validación cruzada temporal, superó consistentemente a los baselines evaluados (regresión lineal, Random Forest y XGBoost) en la estación Belisario, alcanzando un  $R^2$  de 0.942, RMSE de 0.809, MAE de 0.370, MAE en IUV extremo ( $>11$ ) de 1.737 y recall para IUV peligroso ( $>8$ ) de 0.807. Estos resultados representan un desempeño excelente y clínicamente relevante para aplicaciones de alerta sanitaria.
- El análisis comparativo por estación reveló una marcada heterogeneidad espacial y microclimática en Quito. El modelo global (multi-estación) presentó limitaciones importantes en rangos extremos (MAE  $>4$  y recall  $\sim 0.42$ ), mientras que el modelo específico de Belisario mostró una mejora sustancial ( $\sim 60\%$  en MAE extremo y duplicación del recall), lo que demuestra que la modelización por estación individual es más adecuada en contextos con alta variabilidad local.
- La aplicación desarrollada permite la visualización geográfica interactiva y la actualización en tiempo real de predicciones a partir de valores manuales de sensores, constituyendo una herramienta práctica y escalable para monitoreo ambiental y prevención de riesgos asociados a la radiación ultravioleta.

En síntesis, se logró construir un modelo predictivo preciso, interpretable y con alto potencial de aplicación real, validando la hipótesis de que el conocimiento de dominio físico (geometría solar, interacciones atmosféricas y componentes cíclicas) combinado con técnicas de gradient boosting permite superar los desafíos de la predicción de IUV en entornos tropicales de alta montaña.

## **4.2 Recomendaciones**

Basado en los resultados obtenidos y las limitaciones identificadas, se proponen las siguientes recomendaciones para futuros trabajos y aplicaciones prácticas:

### **1. Integración de datos en tiempo real**

Conectar la aplicación con APIs de estaciones meteorológicas para reemplazar la entrada manual por lecturas automáticas, permitiendo monitoreo continuo y alertas automáticas.

### **2. Mejora en la predicción de extremos**

Implementar técnicas específicas para rangos altos de IUV, como:

- Pérdida ponderada o sample weighting para dar mayor importancia a muestras con IUV >8.
- Modelado híbrido (clasificador binario + regresor condicional).
- Uso de modelos de series temporales (LSTM, Transformer) que incorporen lags y persistencia.

### **3. Incorporación de variables adicionales**

Incluir datos de ozono total, aerosoles, nubosidad (de satélite o modelos numéricos) y altitud efectiva para aumentar la explicabilidad y precisión, especialmente en episodios de cielo claro o contaminación elevada.

### **4. Escalabilidad y accesibilidad**

Desplegar la aplicación en un servidor público (Streamlit Cloud, Heroku, AWS) con autenticación básica y notificaciones push para usuarios (autoridades sanitarias, población vulnerable).

Desarrollar una versión móvil responsive o integración con WhatsApp/Telegram para alertas masivas.

## **5. Validación externa y estudios de impacto**

Realizar validación cruzada con datos de estaciones independientes (no usadas en entrenamiento) y evaluar el impacto real de las predicciones en campañas de prevención de cáncer de piel y quemaduras solares en Quito.

## 5. Referencias

An, J., Lu, Y., Huang, D. D., Ding, X., Hu, Q., Yan, R., Qiao, L., Zhou, M., Huang, C., Wang, H., Fu, Q., Yu, F., & Wang, L. (2024). Sectoral Size-Resolved Particle Number Emissions With Speciation: Emission Profile-Based Quantification and a Case Study in the Yangtze River Delta Region, China. *Journal of Geophysical Research: Atmospheres*, 129.

<https://doi.org/10.1029/2024JD041234>

Arlot, S., & Genuer, R. (2014). *Analysis of purely random forests bias*.

Badosa, J., McKenzie, R. L., Kotkamp, M., Calbó, J., González, J. A., Johnston, P. v., O'Neill, M., & Anderson, D. J. (2007). Towards closure between measured and modelled UV under clear skies at four diverse sites. *Atmospheric Chemistry and Physics*, 7, 2817–2837.

<https://doi.org/10.5194/acp-7-2817-2007>

Blumthaler, M., Ambach, W., & Ellinger, R. (1997). Increase in solar UV radiation with altitude. *Journal of Photochemistry and Photobiology B: Biology*, 39, 130–134.

[https://doi.org/10.1016/S1011-1344\(96\)00018-8](https://doi.org/10.1016/S1011-1344(96)00018-8)

Calbó, J., Pagès, D., & González, J. A. (2005). Empirical studies of cloud effects on UV radiation: A review. En *Reviews of Geophysics* (Vol. 43, pp. 1–28).

<https://doi.org/10.1029/2004RG000155>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016, 785–794. <https://doi.org/10.1145/2939672.2939785>

Foyo-Moreno, I., Vida, J., & Alados-Arboledas, L. (1999). A simple all weather model to estimate ultraviolet solar radiation (290-385 nm). *Journal of Applied Meteorology*, 38, 1020–1026. [https://doi.org/10.1175/1520-0450\(1999\)038<1020:ASAWMT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1999)038<1020:ASAWMT>2.0.CO;2)

Jiang, Y., Shi, S., Li, X., Xu, C., Kan, H., Hu, B., & Meng, X. (2024). A 10 km daily-level ultraviolet-radiation-predicting dataset based on machine learning models in China from 2005 to 2020. *Earth System Science Data*, 16, 4655–4672. <https://doi.org/10.5194/essd-16-4655-2024>

Liu, H., Hu, B., Zhang, L., Zhao, X. J., Shang, K. Z., Wang, Y. S., & Wang, J. (2017). Ultraviolet radiation over China: Spatial distribution and trends. En *Renewable and Sustainable Energy Reviews* (Vol. 76, pp. 1371–1383). Elsevier Ltd. <https://doi.org/10.1016/j.rser.2017.03.102>

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-December, 4766–4775.

Madronich, S. (1993). UV radiation in the natural and perturbed atmosphere. In *Environmental effects of UV radiation* (pp. 17–69). Lewis Publishers.

OMM. (2008). *Guía de instrumentos y métodos de observación meteorológicos* (7ma ed., Vol. 8). Organización Meteorológica Mundial.

Parra, R., Cadena, E., & Flores, C. (2019). Maximum UV index records (2010-2014) in Quito (Ecuador) and its trend inferred from remote sensing data (1979-2018). *Atmosphere*, 10. <https://doi.org/10.3390/ATMOS10120787>

Prasad, S. S., Deo, R. C., Salcedo-Sanz, S., Downs, N. J., Casillas-Pérez, D., & Parisi, A. v. (2023). Enhanced joint hybrid deep neural network explainable artificial intelligence model for 1-hr ahead solar ultraviolet index prediction. *Computer Methods and Programs in Biomedicine*, 241. <https://doi.org/10.1016/j.cmpb.2023.107737>

Quito, I. (2025, enero 21). *Recomendaciones para días con radiación extrema*. <https://www.quitoinforma.gob.ec/2025/07/26/recomendaciones-para-dias-con-radiacion-extrema/>.

Secretaría de Salud del DMQ. (2024). *Protocolo de alerta temprana y respuesta ante índices extremos de radiación ultravioleta en el DMQ*. Municipio del Distrito Metropolitano de Quito.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. En *Renewable Energy* (Vol. 105, pp. 569–582). Elsevier Ltd. <https://doi.org/10.1016/j.renene.2016.12.095>

WHO. (2002, junio 13). *Global solar UV index - a practical guide*. <https://www.who.int/publications/i/item/9241590076>.

Zaratti, F., Piacentini, R. D., Guillén, H. A., Cabrera, S. H., Liley, J. ben, & McKenzie, R. L. (2014). Proposal for a modification of the UVI risk scale. En *Photochemical and Photobiological Sciences* (Vol. 13, pp. 980–985). Royal Society of Chemistry. <https://doi.org/10.1039/c4pp00006d>

Zhao, R., & He, T. (2022). Estimation of 1-km Resolution All-Sky Instantaneous Erythral UV-B with MODIS Data Based on a Deep Learning Method. *Remote Sensing*, 14. <https://doi.org/10.3390/rs14020384>