UNIVERSIDAD DE CUENCA

Facultad de Ingeniería
Carrera de Ingeniería de Sistemas

Pronóstico de lluvia en una cuenca de alta montaña basado en técnicas de aprendizaje automático (Machine Learning) y datos del radar meteorológico de medición de lluvias de banda X CAXX

Trabajo de titulación previo a la obtención del título de Ingeniero de Sistemas

Autor:
Franklin Enrique Lara Sanmartín
C.I. 0105036024
larafranklin95@gmail.com

Director:
Ing. Ángel Oswaldo Vázquez Patiño, MSc.
C.I. 0105725634

Cuenca - Ecuador
26/02/2020

RESUMEN

La literatura relacionada al pronóstico de lluvia en zonas de alta montaña utilizando técnicas de Aprendizaje Automático no es extensa y en especial en zonas tropicales andinas como el Ecuador. Una de las posibles razones es la escasa información debido a la dificultad de implementar estaciones hidrometeorológicas en estas zonas de difícil topografía que a su vez dificulta tener resultados aceptables por la alta variabilidad espacio temporal. Para extender las observaciones de lluvia en Ecuador, se implementó una red de radares meteorológicos llamada RadarNet-Sur que da observaciones de reflectividad que pueden ser utilizadas para derivar datos de lluvia. Una de las zonas importantes que cubre este radar es la ciudad de Cuenca (> 600.000 habitantes), en donde eventos extremos (e.g., inundaciones, desbordamiento de ríos) pueden tener efectos catastróficos. Poder tener modelos de pronóstico de lluvia para prevenir los efectos de estos eventos extremos es primordial para los tomadores de decisiones. En esta tesis se utilizan las técnicas de Máquinas de Soporte Vectorial y Árboles Aleatorios para pronosticar la reflectividad a escala horaria con los datos del Radar CAAX y convertirlos a tasa de lluvia. Los resultados demuestran que es posible aprender de los datos utilizando únicamente la variable de reflectividad capturada por el radar de banda X. Por un lado, los modelos con random forest dan valores muy cercanos a los óptimos para las métricas PCC, BIAS, RMSE, permitiendo pronosticar reflectividad para transformarlo a tasa de lluvia.

Palabras Clave: Machine Learning. Support Vector Machine. Random Forest. Pronóstico. Lluvia. Alta montaña.

ABSTRACT

Literature related to the rainfall forecast in highlands using Machine Learning techniques is not extensive, especially in tropical Andean regions such as Ecuador. One of the possible reasons is the limited data due to the difficulty of implementing hydro-meteorological stations in these zones. This also because of the complex topography, which at the same time makes it difficult to obtain acceptable forecasting results since there is a high level of temporal rainfall variability. In order to obtain better rain observations in Ecuador, a network of meteorological radars called RadarNet-Sur was implemented; each radar provides observations of reflectivity which can be used to derivate rain rates. One of the most essential zones that one of the radars covers is the city of Cuenca (> 600.000 inhabitants) where extreme events such as flooding's and river overflows can have catastrophic consequences. Therefore, the development of rain forecasting models, in order to prevent the effects of the extreme events previously mentioned, is crucial for the decision-makers. In this thesis, Support Vector Machine and Random Forest techniques were used to forecast one-hour reflectivity by using the data of Radar CAAX and converting it into rain rate. The results indicate that it is possible to learn from the data only by using a reflectivity variable captured by the X-band radar. On the one hand, the random forest models provide values approximated to the optimum for the metrics PCC, BIAS and RMSE, which allow forecasting reflectivity in order to transform it into rain rate.

Key Words: Machine Learning. Support Vector Machine. Random Forest. Forecast. Rainfall. Highland.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Franklin Enrique Lara Sanmartín en calidad de autor y titular de los derechos morales y patrimoniales del trabajo de titulación "Pronóstico de lluvia en una cuenca de alta montaña basado en técnicas de aprendizaje automático (Machine Learning) y datos del radar meteorológico de medición de lluvias de banda X CAXX", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 26 de febrero de 2020

Franklin Enrique Lara Sanmartín

C.I: 0105036024

Cláusula de Propiedad Intelectual

Franklin Enrique Lara Sanmartín, autor del trabajo de titulación "Pronóstico de lluvia en una cuenca de alta montaña basado en técnicas de aprendizaje automático (Machine Learning) y datos del radar meteorológico de medición de lluvias de banda X CAXX", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor.

Cuenca, 26 de febrero de 2020

Franklin Enrique Lara Sanmartín

C.I: 0105036024

ACRÓNIMOS Y ABREVIATURAS

Acrónimo o abreviatura	Significado
ML	Machine Learning (Aprendizaje automático)
SVM	Support Vector Machine (Máquinas de vectores soporte)
KNN	k-nearest neighbors (K vecinos más cercanos)
M. S. N. M.	Metros sobre el nivel del mar
INAMHI	Instituto Nacional de Meteorología E Hidrología
PIA	Path Integrated Attenuation (Atenuación integrada del camino)
RF	Random Forest
RMSE	Raíz del Error Medio Cuadrático
NSE	Coeficiente de Eficiencia Nash-Sutcliffe
PCC	Coeficiente de Correlación de Pearson
BIAS	Sesgo Relativo
σ	Desviación Estándar
Me	Mediana

AGRADECIMIENTOS

Al Ing. Ángel Vázquez MSc, por su apoyo, guía y consejos durante la realización de este trabajo de titulación.

A la Ing. Johanna Orellana MSc, por haberme confiado realizar el estudio presentado en este trabajo, por las enseñanzas y consejos esenciales que me brindó durante este trabajo de titulación.

Al proyecto ETAPA-DIUC “Desarrollo de modelos para pronóstico hidrológico a partir de datos de radar meteorológico en cuencas de montaña” financiado por la Dirección de Investigación de la Universidad de Cuenca y Empresa Pública Municipal de Telecomunicaciones, Agua Potable, Alcantarillado y Saneamiento de Cuenca (ETAPA-EP) quienes han proporcionado los datos para la ejecución de este trabajo.

A todos mis compañeros, amigos y familiares, por haberme dado el apoyo para continuar día a día.

A la Facultad de Ingeniería y a la Universidad de Cuenca, por prepararme para la vida profesional.

DEDICATORIA

A mis padres Raúl y Victoria, mi ejemplo a seguir, nunca han dejado de apoyarme para que pueda conseguir las metas que me he propuesto.

A mi abuela Rosa, que siempre ha sido y seguirá siendo mi apoyo para conseguir mis metas.

A mi hermano Santiago, que siempre ha sido y será mi apoyo y motivación para conseguir mis metas.

A mis tutores, Ángel Vázquez y Johanna Orellana, que compartieron conmigo parte de su conocimiento y sabiduría, para facilitar el proceso de realización de mi trabajo de titulación.

A la Universidad de Cuenca, mi segundo hogar, aquí conocí personas maravillosas que aportaron para mi crecimiento académico y personal.

Franklin

1. Introducción

En zonas de alta montaña la literatura relacionada al pronóstico de lluvia es escasa y en especial en zonas andinas como Ecuador. Entre las posibles razones tenemos la dificultad presentada al momento de implementar estaciones hidrometereológicas en lugares de difícil topografía que a su vez dificulta tener resultados aceptables por la alta variabilidad espacio temporal. Además, desde el punto de vista científico, es importante caracterizar fenómenos naturales. Realizar esta caracterización es complejo; sin embargo, desde el punto de vista social, es necesaria para evitar desastres que involucran la pérdida de vidas humanas. La precipitación, y más específicamente la lluvia, es un fenómeno natural meteorológico estudiado vastamente por sus implicaciones desde las sociales hasta las económico-industriales. Así, por ejemplo, el pronóstico de lluvia permite advertir del riesgo de posibles crecientes de ríos que desencadenaría desastres en las áreas bajas.

Para extender las observaciones de lluvia en Ecuador, el proyecto regional de transferencia de Conocimiento Alemán "Operational rainfall monitoring in southern Ecuador" implementó una red de radares meteorológicos llamada RadarNet-Sur que proporciona observaciones de reflectividad que pueden ser utilizadas para derivar datos de lluvia. Una de las zonas importantes que cubre este radar es la ciudad de Cuenca (> 600.000 habitantes), en donde eventos extremos (e.g., inundaciones, desbordamiento de ríos) pueden tener efectos catastróficos. Poder tener modelos de pronóstico de lluvia para prevenir los efectos de los mencionados eventos extremos es primordial para los tomadores de decisiones.

A nivel mundial se ha realizado investigaciones para generar modelos de pronóstico de lluvia con diferentes variables predictoras y con diferentes técnicas de ML. Las técnicas más utilizadas de ML para pronóstico de lluvia son K-Nearest Neighbors (KNN) y K-Means. Por ejemplo, Dash et al. (2018) utilizaron KNN, Redes Neuronales Artificiales y Machine Learning Extreme (es una configuración del ML tradicional, donde se trabaja con una o varias capas ocultas en donde sus entradas se asignan de forma aleatoria) para generar pronósticos de lluvia, mostrando el rendimiento de cada técnica en cuanto al tiempo de ejecución y la aproximación a la observación real de los datos. Por otro lado, Chakraborty et al. (2014) presentan una metodología genérica, mediante agrupamiento incremental por K-Means estableciendo la métrica de Manhattan. Sus resultados tienen una exactitud (accuracy) aproximada de 0.83 y concluyen que este nuevo método es adecuado para zonas donde se tiene alta variabilidad temporal. Kumar et al. (2012) utilizan K-Means para realizar agrupamiento, donde los autores han utilizado los resultados numéricos generados a través del algoritmo de función de densidad, para

determinar si el valor que ingresa a ser pronosticado está dentro del clúster, utilizando los datos de las variables temperatura, viento y humedad para determinar con el modelo si llueve o no llueve. Lograron tener un error medio cuadrático de 29.69% y una precisión de 85.63%, demostrando que los datos pronosticados tienen una desviación mínima con respecto a los datos reales. Por otro lado, es importante tener en cuenta la altura en las zonas donde se realizaron estudios de pronóstico. El estudio a mayor altura ha sido realizado en la India en el estado de Kerala, con una altura promedio de 2.695 m.s.n.m. (Dash et al., 2018) y el resto de estudios se realizaron en sitios de alturas inferiores. Por lo tanto, se evidencia la relevancia del estudio del tema de predicción de lluvia. En los estudios mencionados se utilizan variables como temperatura y humedad. Desafortunadamente datos de estas variables son escasos en zonas de montaña. Adicionalmente, hasta donde conocemos no hay literatura en donde se utilice únicamente datos de reflectividad de un radar a una altura mayor de 3000 m.s.n.m. para pronosticar lluvia con técnicas de ML. Esto ha presentado un nicho interesante de aplicación de técnicas de ML para llevar a cabo la predicción de lluvia.

En Ecuador, la Dirección de Pronósticos y Alertas Hidrometeorológicas del INAMHI, realiza pronóstico de condiciones atmosféricas, las cuales se esperan para las 12 horas del día entre las 07h00 y 19h00, manteniendo como fuente de datos la red de superficie, satelital y modelos numéricos, para el Distrito Metropolitano de Quito y los cantones de las provincias de Pichincha, Guayas, Loja y Riobamba. Adicionalmente, presenta condiciones atmosféricas de diferentes zonas del país, entre ellas la provincia del Azuay por medio de un geovisor (“INAMHI – Ecuador 2018). Al revisar los resultados expuestos por el INAMHI, presentan pronósticos de lluvia, temperatura, humedad y la probabilidad de tormentas eléctricas para tres días al futuro, los cuales corresponden a todo un cantón. Sin embargo, no es posible conocer la cantidad de lluvia que va caer en un área específica de Cuenca. Cabe señalar que estos sistemas no utilizan información espacio-temporal proporcionada por radar meteorológico. Sin embargo, estos datos de radar ya fueron utilizados para realizar la predicción con Aprendizaje Profundo (Deep Learning), donde se utilizó parte de los datos de las lecturas del radar, obteniendo como resultado que es posible predecir lluvia mediante el aprendizaje de los datos (Godoy, 2019).

El objetivo de la tesis es la aplicación de dos técnicas de ML: Support Vector Machine (SVM) y Random Forest, para obtener un pronóstico de reflectividad a una hora al futuro y convertirla a tasa de lluvia con un rendimiento que permitiría su posible utilización para sistemas de alerta temprana.

2. Materiales y Métodos

2.1. Radar Meteorológico de banda X CAXX

Los radares de precipitación son instrumentos que permiten realizar observaciones atmosféricas realizando lecturas para el análisis temporal y espacial de las precipitaciones, su clasificación, entre otras. El instrumento utilizado es de polarización simple, el cual emite un haz de pulsos electromagnéticos con el que cubre un área de 31.415 km2, y al momento que una gota atraviesa por dicho haz electromagnético, la señal reflejada permite al radar captar información de reflectividad, la cual puede ser transformada a lluvia (Ellis et al., 2006; Orellana-Alvear et al., 2019).

El Radar Meteorológico CAXX, parte del proyecto regional de transferencia de Conocimiento Alemán "Operational rainfall monitoring in southern Ecuador" (Monitoreo operativo de lluvia en el sur de Ecuador), está ubicado en Cuenca, Ecuador (2° 45′ S, 79° 16′ O), en los límites del parque Nacional El Cajas a una altura de 4450 m.s.n.m. y es considerado el radar de banda X más alto del mundo (Bendix et al., 2017). Este inició sus operaciones en el año 2015. En la Figura 2.1 se muestra la ubicación del radar CAXX, con relación a la zona central del Ecuador. Este radar permite monitorear las provincias: Azuay, Cañar y parcialmente El Oro y Guayas. El radio de cobertura es de 100 km, pero mientras más lejana es la lectura, es menos precisa en el registro de la reflectividad; esto es ocasionado, entre otras cosas, por la disminución de la intensidad del haz al colisionar con las gotas más cercanas. Además, es importante conocer que los daños físicos (deterioro de componentes, humedad, etc.) que sufre el radar no permite tener lecturas del radar durante un año continuo. Y también hay que tomar en cuenta los problemas generados por factores físicos externos, tales como montañas (Orellana-Alvear et al., 2019), animales u objetos voladores como drones, helicópteros, aviones, etc.

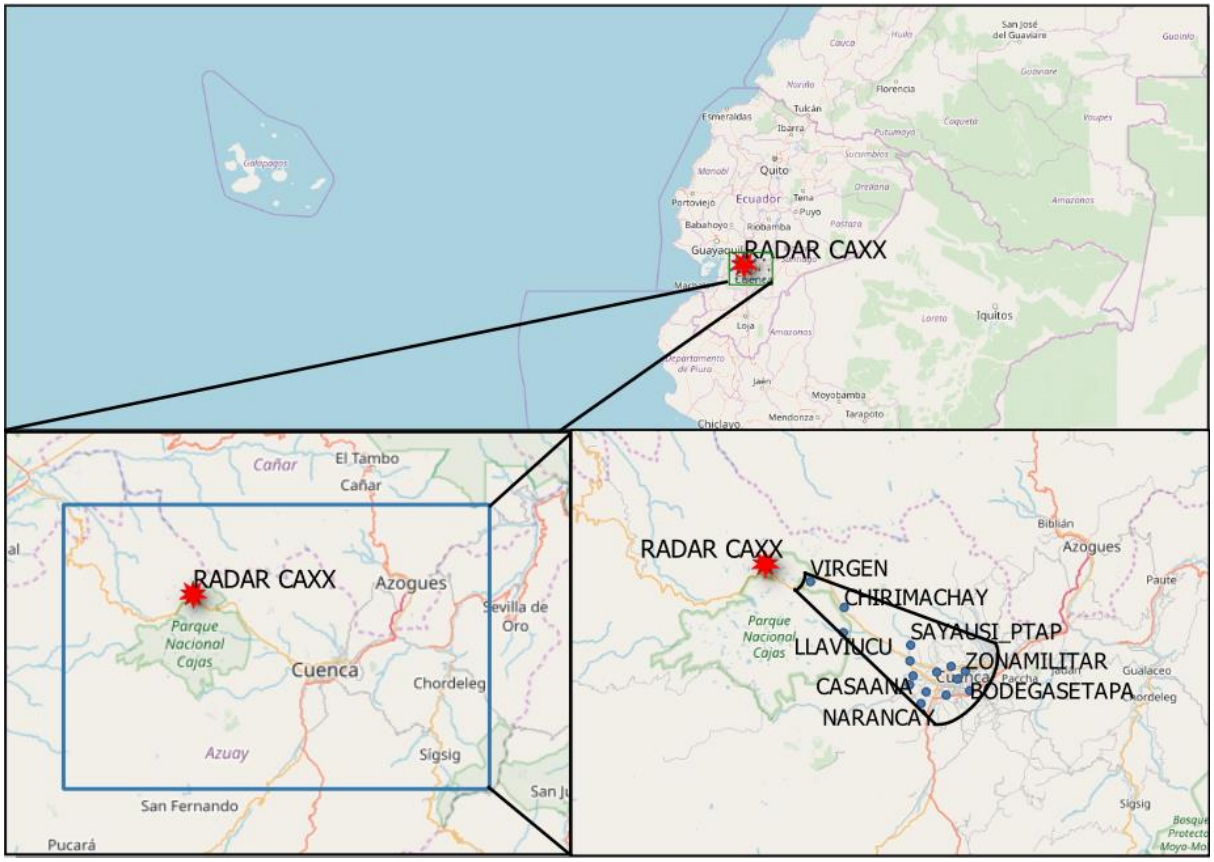

Figura 2.1: Zona de estudio y ubicación del Radar CAXX

.

2.3. Recopilación y selección de datos

Los datos de reflectividad de las precipitaciones generados por el radar meteorológico CAXX fueron proporcionados por el Departamento de Recursos Hídricos de la Universidad de Cuenca (iDRHICA) en formato NETCDF (Network Common Data Format). Dichos datos incorporan las correcciones realizadas por Orellana-Alvear et al. (2019), en cuyo trabajo se detalla la metodología aplicada. A continuación, resumimos tal preprocesamiento. Primero, se eliminan los valores extremos menores al percentil 5 y mayores al percentil 95 de la acumulación de imágenes durante el periodo de operación del magnetrón (resonador magnético del radar) por medio del eco considerando el clutter estático y dinámico. Posteriormente, se aplica el filtro de Gabella (Marra & Morin, 2015) para considerar homogeneidad de textura de las lecturas. En el segundo paso, las regiones identificadas como clutter se eliminan de la imagen y se aplica una interpolación en coordenadas polares. En el tercer paso, se realiza una corrección por atenuación a partir del método "*Path Integrated Attenuation*" (PIA) (Capozzi et al., 2014).

Los datos de las lecturas que corresponde a la reflectividad están representados en la escala -31.5 dBZ hasta 95.5 dBZ. Estos se encuentran en forma de matriz de

datos de 1.000 x 180 celdas de información, 1.000 celdas que representan 100m de cobertura de cada una y 180 celdas representando 2° cada una hasta los 360° de la rotación del radar. Teniendo lecturas continuas cada 5 minutos durante todo el día, éstas están almacenadas en un solo archivo por día y dichos datos corresponden a los períodos marzo-diciembre 2015, febrero-diciembre 2016 y enero-junio 2017, los datos en estos periodos están incompletos por problemas de lectura del radar provocados por daños de hardware del mismo. Las variables presentes en los archivos se describen en la Tabla 2.1.

Variable	Nombre estándar	Descripción	Tipo de dato
rangebin	rangebin	Distancia en kilómetros desde el radar	flotante
azimuth	Azimuth	Ángulo en grados sexagesimales de la coordenada polar	flotante
raw dBZ	Unidad de medida de la reflectividad	Decibelios (dBZ)	flotante
time	Tiempo	Cada cinco minutos	fecha

Tabla 2.1: Descripción de las variables de los archivos de datos.

2.4. Área de Estudio

El área de estudio está comprendida dentro del alcance del radar meteorológico CAXX. Se extiende entre los radios 6 km hasta los 33 km a partir de la ubicación del radar, con un ángulo de apertura de 22° (este comprende los 112° hasta los 132° azimuth). En la parte inferior derecha de la Figura 2.1 se muestra el área de estudio. Se tiene que considerar que se cubre la zona céntrica de Cuenca. Se escoge este rango debido a que en los primeros 5 km, a partir del radar, existen atenuaciones, y en distancias muy lejanas se encontró ruido generado por factores físicos, climáticos, atmosféricos y el error de atenuación (Orellana-Alvear et al., 2019).

2.5. Metodología

Los datos de ingreso al modelo se obtienen de series de tiempo, agrupamos los valores de reflectividad por horas, es decir obtenemos la mediana y desviación estándar de las 12 lecturas (cada 5 min) de una hora, para representarlas en una sola lectura por hora por cada uno de los puntos para el ángulo de apertura del radar y su distancia dentro del área de estudio.

En la Figura 2.2 mostramos la distribución de la reflectividad por cada día, para esta investigación se usan los datos correspondientes a los meses marzo-mayo

2015, febrero-mayo 2016 y febrero-marzo 2017, estos fueron seleccionados por tener presencia de reflectividades más altas y continuas con respecto al conjunto total de datos y además coinciden con la temporada de invierno en la zona de estudio. Presentamos la descripción de las variables de entrada para los modelos en la Tabla 2.2, las cuales corresponden a las lecturas del radar CAXX.

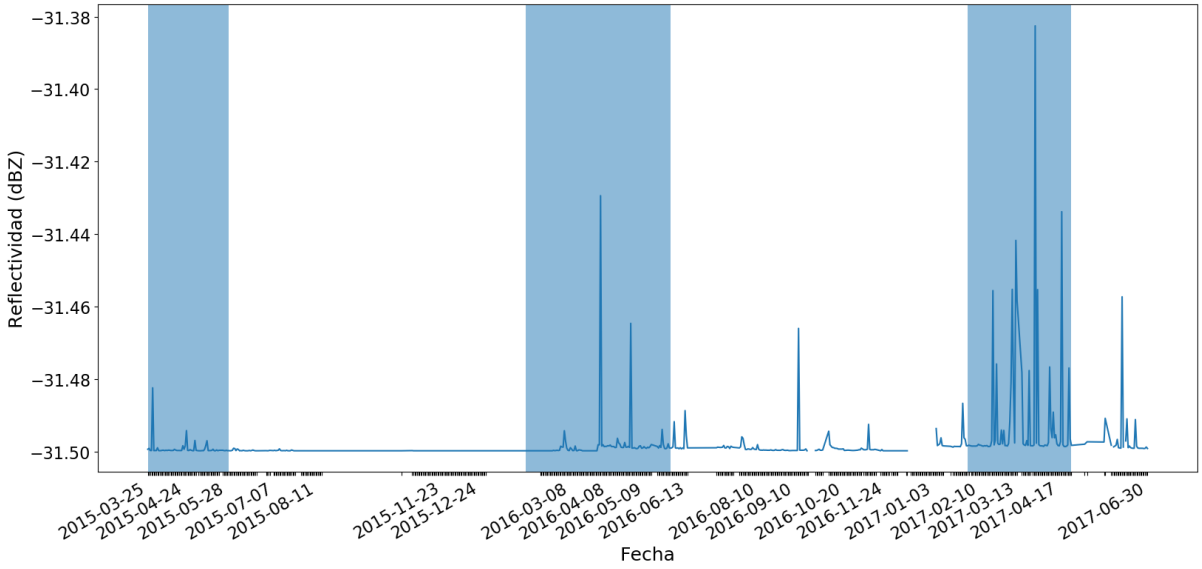

Figura 2.2: Distribución de la mediana de datos del Radar CAXX.

Variable	Descripción
Azimuth	Ángulo de apertura del radar durante la lectura.
Rangebin	Es el radio de la lectura a partir de la ubicación del radar.
dBZ	Medida de los valores de reflectividad en ese tiempo.
Mediana de dBZ (Me)	La mediana de los valores de reflectividad del espacio en el ángulo y radio.
Desviación estándar de dBZ (σ)	Es la medida usada para cuantificar la variación o dispersión de las lecturas de reflectividad.

Tabla 2.2: Descripción de las variables de entrada para los modelos de pronóstico

En la Figura 2.3, se muestra el esquema de la metodología aplicada sobre los datos para el entrenamiento y validación de los modelos de ML a partir de las lecturas del radar, dando como resultado las variables de entrada para el modelo corresponden a lecturas en el pasado, teniendo hasta 5 lecturas preliminares a la

del tiempo t, teniendo un total de 14 variables. Entre las que tenemos azimuth, rangebin, σ (t-5), Me (t-5), σ (t-4), Me (t-4), σ (t-3), Me (t-3), σ (t-2), Me (t-2), σ (t-1), Me (t-1), σ (t) y Me (t). Como variable objetivo se tiene a la mediana de la lectura en un tiempo t + 1. El conjunto de datos de entrada se divide en dos partes, 80% de los datos para el entrenamiento y el 20% para pruebas (testing), que corresponden a 11´299.182 instancias (lecturas del radar por hora) para entrenamiento y 2´824.796 para test. Estos datos fueron normalizados con la función StandardScaler que está dentro de la librería scikit-learn (scikit-learn, 2018), la cual genera una distribución general entre todos los datos. Posteriormente, se aplica la conversión mediante la relación Z–R ($Z = AR^b$) (Marshall et al, 1947) para transformar la reflectividad a tasa de lluvia pronosticada del valor de reflectividad para el periodo de datos de validación y se compara con las lecturas reales, donde: Z es el factor de reflectividad capturado por el radar, R es la tasa de precipitación y A, b son factores que deben ser constantes determinadas empíricamente.

Se utilizan los parámetros identificados para la relación Z–R en el sitio de estudio en (Orellana-Alvear et al. 2017) ya que se conoce la distribución del tamaño de las gotas dentro de un volumen de aire medido por el radar y se compara con las mediciones realizadas en tierra por los pluviómetros ubicados en la zona de estudio, además se tiene que tener presente el tipo de lluvia, ya que en nuestra zona que es geográficamente compleja cambia de un lugar a otro, donde A = 103 y b = 2.03.

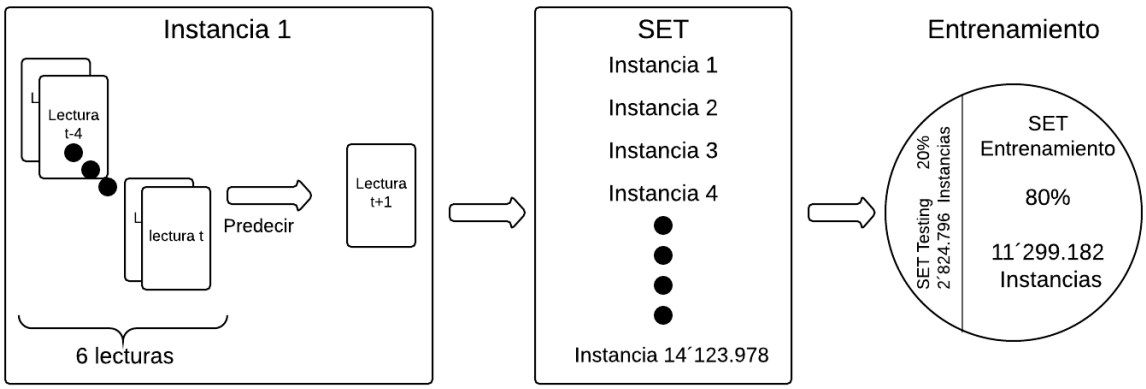

Figura 2.3: Esquema de la metodología aplicada para los datos para el entrenamiento y validación de los modelos de ML.

El rendimiento del pronóstico con el conjunto de datos de validación (testing) se evaluó mediante cinco métricas: coeficiente de correlación de pearson (PCC), raíz del error medio cuadrático (RMSE), coeficiente de eficiencia Nash-Sutcliffe (NSE), sesgo relativo (BIAS) y varianza (Su et al., 2010). PCC, RMSE, NSE y BIAS se usaron para evaluar el sesgo entre los valores pronosticados y los valores de las lecturas reales de reflectividad y la varianza para conocer la dispersión de los datos. PCC mide la fuerza de la relación de correlación entre los valores pronosticados y

los valores reales. RMSE es una media absoluta entre los valores pronosticados y los valores reales. NSE determina la magnitud relativa entre la varianza de los valores pronosticados y los valores reales. BIAS mide la tendencia de la media de la reflectividad pronosticada con la real. Varianza es una medida de dispersión entre los valores de la variable y su media. Todas las métricas de evaluación se muestran en la Tabla 2.3.

Métrica	Unidad	Fórmula	Rango	Valor óptimo
Coeficiente de Correlación de Pearson (PCC)	-	$PCC = \frac{\sum_{i=1}^{n}(Y_{oi} - y_0)(C_i - c)}{\sqrt{\sum_{i=1}^{n}(Y_{oi} - y_o)^2} * \sqrt{\sum_{i=1}^{n}(C_i - c)^2}}$	[-1,1]	1 o -1
Raíz del Error Medio Cuadrático (RMSE)	dBZ	$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(C_i - Y_{oi})^2}$	[0,+∞]	0
Coeficiente de eficiencia Nash-Sutcliffe (NSE)	-	$NSE = 1 - \frac{\sum_{i=0}^{n}(C_{i-} Y_{oi})^2}{\sum_{i=1}^{n}(C_{i-} \gamma_o)^2}$	(-∞,1]	1
Sesgo Relativo (BIAS)	-	$BIAS = \frac{\sum_{i=0}^{n}(C_{i-} Y_{oi})}{\sum_{i=1}^{n} Y_{oi}}$	(-∞,+∞)	0
Varianza	dBZ^2	$var = \frac{\sum_{i=1}^{n}(xi - \chi)^2}{N}$	(0, +∞)	0

Tabla 2.3: Fórmulas de las métricas de evaluación de modelos de aprendizaje en Random Forest.

2.6. Técnicas de Machine Learning utilizadas

2.6.1. Support Vector Machine (SVM, Máquina de Vectores de Soporte)

Las máquinas de vectores de soporte (SVM) son un conjunto de métodos de aprendizaje supervisado que se utilizan para la clasificación, regresión y detección de valores atípicos. SVM realiza un mapeo no lineal de los datos de entrenamiento pasando a un espacio de mayor dimensión, donde se pueda realizar una regresión lineal. Además, presenta las siguientes ventajas (Mohandes et al., 2004):

- Eficaz en espacios de alta dimensión, lo que nos permite trabajar con varias características de los objetos, siendo aún efectivo en casos donde el número de dimensiones es mayor que el número de muestras.
- Utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que también es eficiente en memoria.
- Versátil, se pueden especificar diferentes funciones del Kernel para la función de decisión tales como: kernel de Lineal, Polinómico, función de base Radial

(RBF) y Sigmoide. Se proporcionan núcleos comunes, pero también es posible especificar núcleos personalizados.

- En el caso de tener un número de funciones mucho mayor que el número de muestras, se tiene que evitar la adaptación excesiva en la elección de las funciones del núcleo y el término de regularización es crucial para obtener buenos resultados.
- SVM no proporciona directamente estimaciones de probabilidad, estas se calculan utilizando una costosa validación cruzada de cinco veces.

Dentro del algoritmo de SVM, existen varias funciones para realizar modelos de regresión, existe una función llamada LinearSVR, esta utiliza un kernel similar al lineal de la librería predeterminada para el entrenamiento, permitiendo tener una mayor flexibilidad en la elección de penalizaciones y funciones de pérdida, lo que permite escalar de mejor forma en el entrenamiento con la cantidad de datos, la función recibe como entrada los siguientes parámetros:

- **kernel:** Tipo de kernel que se utilizará en el algoritmo. Entre estos tenemos "lineal", "poli", "rbf".
- **max_iter:** El número máximo de iteraciones variamos en cada prueba.
- **tol:** Tolerancia para el criterio de parada es de 0.0001.
- **C:** Parámetro de penalización C del término de error es de 1.

2.6.2. Random Forest (Bosque Aleatorio)

Random Forest (RF, Bosque aleatorio) es un algoritmo de aprendizaje supervisado que se utiliza para tareas de clasificación y regresión. Genera buenos resultados la mayor parte del tiempo incluso sin ajuste de hyper-parámetros, es uno de los más actualizados por la simplicidad de su funcionamiento (Donges, 2018).

Para su funcionamiento, Random Forest crea un bosque y lo hace de forma aleatoria. El "bosque" que construye, es un conjunto de árboles de decisión, la mayoría de las veces entrenados con el método de "embolsado". La idea general del método de embolsado es que una combinación de modelos de aprendizaje aumenta el resultado general (Donges, 2018), obteniendo una predicción más precisa y estable. En la Figura 2.4, podemos ver cómo se vería un bosque al azar con dos árboles.

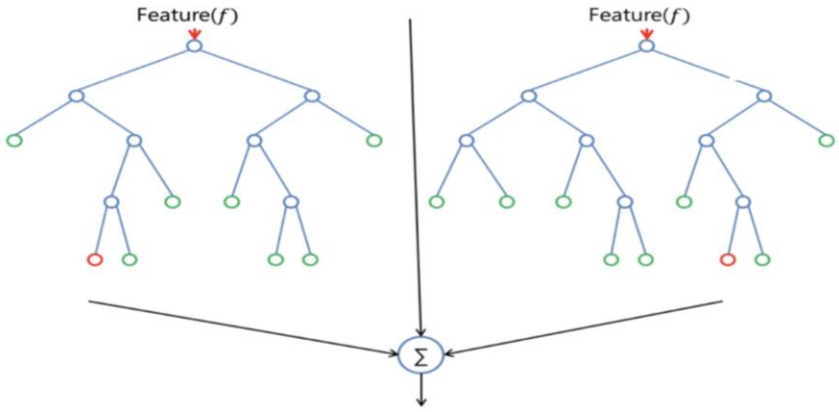

Figura 2.4: Clasificación con Random Forest. Obtenida de (Donges 2018)

Random Forest agrega aleatoriedad al modelo, mientras crecen los árboles, se busca la mejor característica de entre un conjunto aleatorio. Cada nodo que no tiene hijos es una hoja. En la implementación de este algoritmo, se pueden ingresar los siguientes parámetros:

- **n_estimadores:** El número de árboles en el bosque varían en cada prueba realizada.
- **max_depth:** Define la profundidad del árbol, en el caso de tener el valor 0 los nodos se expanden hasta que sus hojas sean puras.
- **min_samples_split:** Se establece 2 como número mínimo de muestras requeridas para dividir un nodo.
- **min_samples_leaf:** Se asigna 1 como número mínimo de muestras requeridas para estar en un nodo hoja.
- **max_features:** La cantidad de características a considerar cuando se busca la mejor división. Se deja al algoritmo utilizar de forma automática para una mejor adaptación del estudio. Donde considera a todas las características de las entradas.

Se han utilizado las mismas instancias como entradas para el entrenamiento de los modelos. Por un lado, para entrenar los modelos de SVM se utilizó una función denominada LinearSVR. Por otro lado, para entrenar Random Forest se utilizó la función de regresión que nos determina la librería Scikit-Learn. En la implementación se realizan varias pruebas modificando los parámetros de entrada de los algoritmos, con lo que podemos determinar lo valores óptimos para extraer los resultados de los presentes modelos.

3. Resultados y Discusión

En Random Forest; por un lado, la profundidad del árbol es importante, al tener un valor pequeño el árbol se tiene que ajustar dando paso a un mayor error en el aprendizaje, por lo que, se ha dejado que el valor sea automático y que el árbol se

extienda hasta tener una instancia en cada hoja, teniendo una profundidad de 63 y 152.300 nodos. Además, se utilizan todas las variables de entrada para el aprendizaje, utilizamos dos como el número requerido para dividir un nodo y determinamos una única muestra necesaria para estar en el nodo hoja. Por otro lado, el número de árboles se ha probado con los mejores valores de parámetros antes mencionados, teniendo una variación pequeña del error o la precisión (accuracy) del modelo, obteniendo como resultado los valores de la Tabla 3.1 y Figura 3.1. La varianza de los valores en el pronóstico con respecto a los reales está aproximadamente a 0.03. Finamente, se utilizan todas las variables de entrada para entrenar el modelo, la precisión (accuracy) 99.95% - 99.97% en la Figura 3.2 presentamos la importancia de las variables de entrada para el modelo.

#	Número de árboles	PCC	RMSE (dBZ)	NSE	BIAS	Varianza	Diferencia de Varianzas	Exactitud (Accuracy)
1	2	0.9997	0.173	0.9995	-1.5e-6	70.76	0.0337	99,96%
2	10	0.9998	0.164	0.9996	7.57e-8	70.65	0.0309	99.97%
3	100	0.9998	0.161	0.9996	-2.39e-6	70.56	0.0318	99.95%
4	250	0.9998	0.161	0.9996	-1.31e-6	70.47	0.0322	99.96%
5	500	0.9998	0.167	0.9996	2.62e-6	70.23	0.0321	99.96%

Tabla 3.1: Resultados de validación (testing) con los modelos de RF.

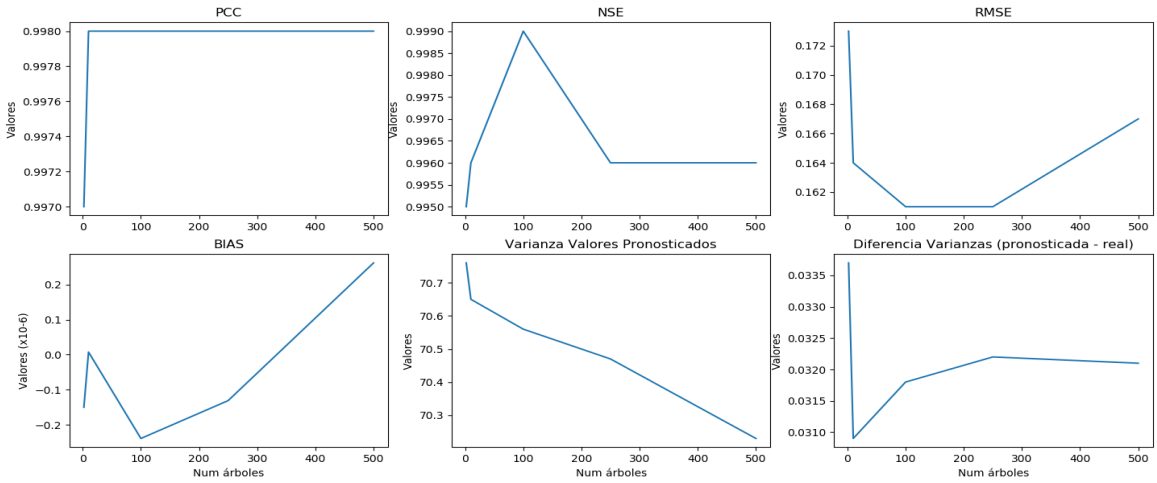

Figura 3.1: Resultados de validación (testing) con los modelos de Random Forest.

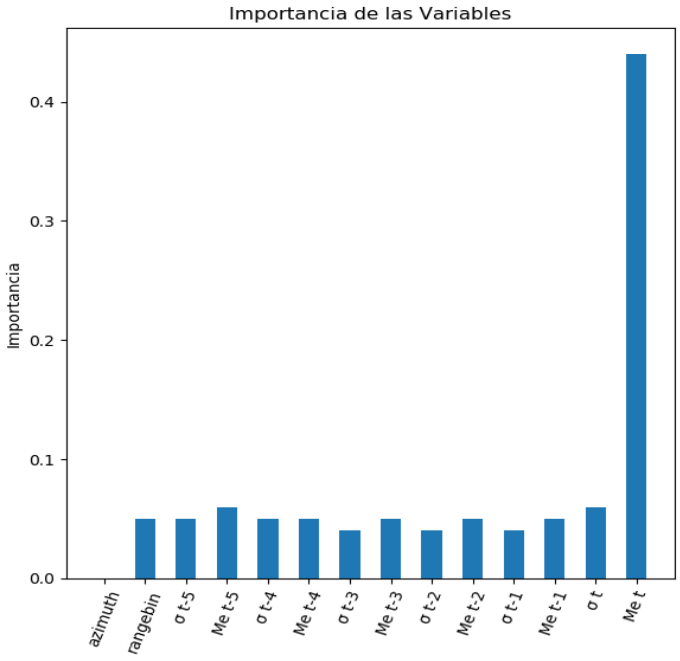

Figura 3.2: Importancia de las variables de entrada en el entrenamiento de los modelos con Random Forest.

En la Figura 3.3, presentamos la intercepción de los valores reales (azul) contra los pronosticados (rojo) de la variable objetivo de reflectividad. Vemos que esta curva no cumple con todos los valores, debido a lo variante del clima en la zona de estudio.

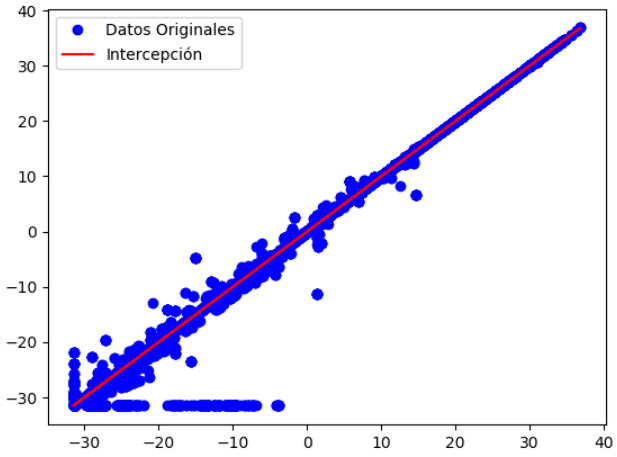

Figura 3.3: Intercepción de valores reales contra los pronosticados de la variable objetivo del modelo RF.

El resultado de Random Forest puede dar paso a un sobreajuste de los datos, sin embargo, esto genera una controversia, al tener un rango amplio de datos y capturar cientos de eventos de lluvia en la zona de estudio, se captura el complejo

comportamiento de la reflectividad durante los años en los que se realizaron las lecturas en el área del estudio. Podemos determinar que tenemos un modelo de ML que capturó el comportamiento hidrometereológico del área de estudio estableciendo que es un entrenamiento robusto o tenemos un modelo sobre ajustado con los datos de la zona. Al realizar la validación de los modelos con datos que no fueron utilizados en el entrenamiento de estos, no se puede determinar que es un sobreajuste.

Para SVM, con la función LinearSVR donde se utiliza un kernel similar al “lineal” para el entrenamiento de los modelos, se utiliza el valor de tolerancia de 0.0001, se modifica el valor del número de iteraciones que se realizan para el entrenamiento del modelo. Por un lado, mientras crece el valor de iteraciones el PCC mantiene la relación directa entre los valores pronosticados y los reales. Por otro lado, los valores del RMSE, NSE y la varianza disminuyen, sin embargo, la diferencia de las varianzas entre los pronósticos y los valores reales es minima. Finalmente, en la Tabla 3.2 y Figura 3.4 se presenta la exactitud (accuracy), la cual se incrementa cuando más iteraciones hay, sin embargo, el tiempo de aprendizaje también se incrementa, pero el valor más alto de 26.01% se presenta al tener 10.000 iteraciones, ver en la Figura 3.5.

#	iteraciones	PCC	RMSE (dBZ)	NSE	BIAS	Varianza	Diferencia de Varianzas	Accuracy
1	10	0.60	7.46	0.21	0.0038	52.38	18.19	22.47%
2	100	0.60	6.87	0.33	0.047	30.14	40.71	17.85%
3	250	0.60	6.78	0.34	0.042	29.21	41.35	17.89%
4	1.000	0.60	6.77	0.349	0.040	30.10	40.47	20.21%
5	10.000	0.60	6.76	0.35	0.039	30.09	40.46	26.01%

Tabla 3.2: Resultados de la validación(testing) de modelos de SVM.

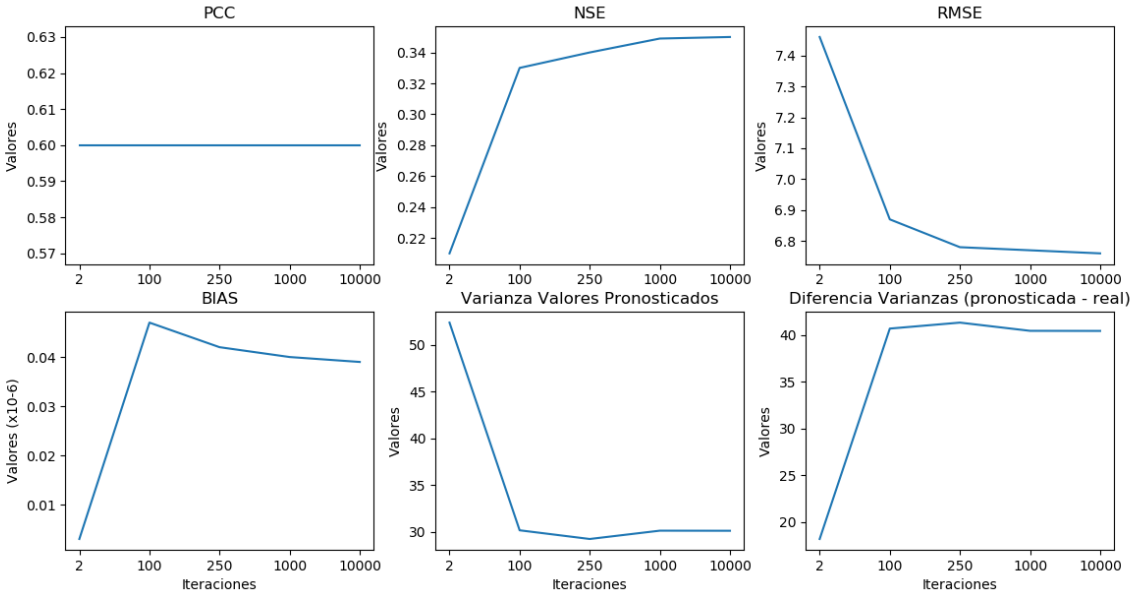

Figura 3.4: Resultados de la validación(testing) de los modelos con SVM.

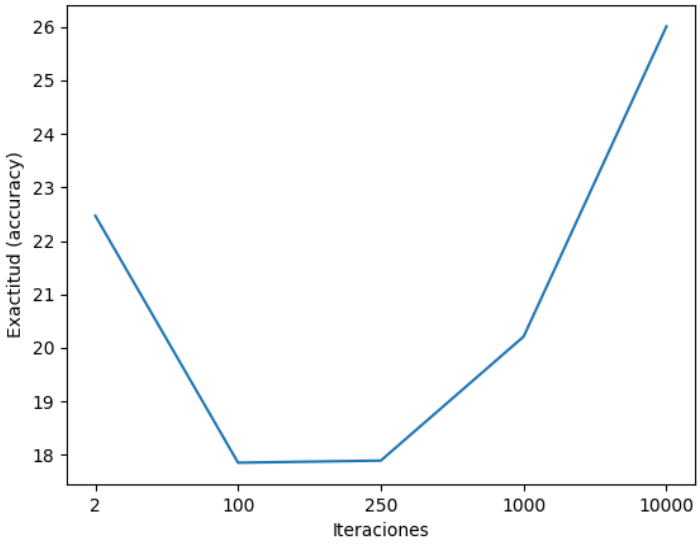

Figura 3.5: Exactitud de la validación(testing) de los modelos SVM.

En la Figura 3.6, se presentan los valores reales de la variable objetivo contra la función de ajuste del pronóstico de reflectividad. Vemos que el ajuste no cumple con la mayoría de valores, existen valores que están fuera de dicha función, esto se debe al ajuste lineal que realiza SVM por que se tienen 14 variables de entrada para el aprendizaje del modelo.

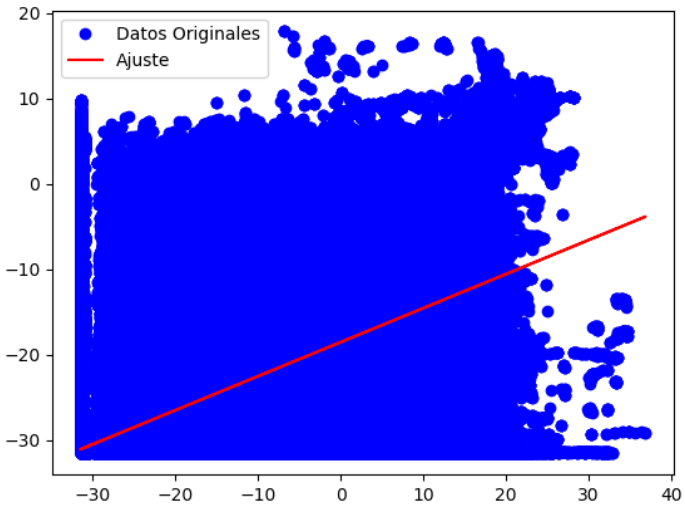

Figura 3.6: Valores reales de la variable objetivo del modelo SVM contra la función de ajuste del pronosticados de reflectividad.

Luego de obtener los valores pronosticados de reflectividad con los modelos de ML antes mencionados, los utilizamos para convertirlos a tasa de lluvia mediante la relación Z-R, e igual procesamiento con los valores de reflectividad real de las lecturas del radar. De los resultados de random forest obtenemos la media de los criterios de evaluación de la reflectividad convertida a tasa de lluvia. De SVM se toma el entrenamiento con la exactitud más alta del pronóstico de la reflectividad para luego convertirlo a tasa de lluvia y expresar los resultados en la Tabla 3.3.

Técnica	PCC	RMSE (mm/h)	NSE	BIAS
Random Forest	0.9999	6.77	0.9998	-0.0001
SVM	0.002	271.86	-9.5894	-0.9996

Tabla 3.3: Resultados de los criterios de evaluación al convertir la reflectividad a tasa de lluvia con los datos de validación(testing).

Al convertir la reflectividad a tasa de lluvia visualizamos un incremento del error entre los valores pronosticados frente a los reales, esto se debe al acarreo del error desde el pronóstico de la reflectividad, ya que crece al utilizar la relación Z-R para obtener la tasa de lluvia. Por un lado, Random Forest presenta una correlación entre la tasa de lluvia pronosticada frente a la real, conservando el patrón de los datos, ya que su PCC, BIAS y NSE están muy cercanos a los valores óptimos. Por otro lado, SVM presenta una correlación muy baja entre los valores ya que su PCC, BIAS y NSE están muy distantes de ser óptimos. Finalmente, Random Forest tiene un RMSE de 6.77 mm/h por cada pixel de 2° x 100m.

Teniendo en cuenta los umbrales de tasa de lluvia especificados por Vigilant (2014), conocemos que la diferencia entre una lluvia débil a moderada y de moderada a muy fuerte es de 15 mm/h, por ende al tener 6.77 mm/h de RMSE entre los valores pronosticados,

determinamos que es un valor de alta variabilidad de tasa de lluvia, ya que conocemos que el pronóstico se realiza por cada pixel del área de estudio, el cual puede influenciar en la toma de decisiones de alerta temprana.

4. Conclusiones y Trabajo Futuro

Durante el desarrollo de esta tesis hemos presentado la aplicación de técnicas de ML para pronosticar reflectividad y convertirla a tasa de lluvia. Primero, obtención de los datos, metodología aplicada para generar las instancias de entrada y entrenamiento de los modelos. Segundo, se realiza la validación (testing) y predicción de la reflectividad para ser convertida a tasa de lluvia mediante la relación Z-R. Las técnicas utilizadas para generar el modelo SVM (Support Vector Machine) y RF (Random Forest).

El aporte principal de este estudio es la evaluación de SVM y Random Forest para pronosticar la reflectividad y convertirla a tasa de lluvia dentro del área de estudio, está fue establecida debido al tamaño de los archivos originales. Además, el área de estudio comprende la zona urbana de Cuenca, Ecuador; donde tiene un impacto social por la prevención de posibles catástrofes naturales. Con los resultados expuestos por estos modelos, se ha demostrado que es posible 'aprender' de las lecturas del radar CAXX, considerado que los datos hidrometeorológicos tienen un comportamiento complejo en zonas de alta montaña. Al pronosticar la reflectividad y convertir estos datos mediante la relación Z-R se determina la tasa de lluvia en nuestra zona de estudio. Se utilizaron los mismos datos para el entrenamiento de los modelos tanto para SVM y Random Forest, teniendo en cuenta una división del 80% de los datos para el entrenamiento y el 20% restante para la validación (testing) del modelo.

SVM dispone de diferentes tipos de kernel para el entrenamiento, estos generan una función de ajuste para los datos de entrada. El kernel lineal genera una función lineal para ajustar el aprendizaje de los datos y rbf genera una función en un espacio n dimensional, el cual se puede ajustar a las variables de entrada de cada modelo, sin embargo, al utilizar este kernel con más de 14´000.000 instancias el modelo no converge, por ende se utilizó otra función que ofrece la librería scikit-learn denominado LinearSVR, donde usan un kernel similar al lineal para ajustar todos los datos y al conocer el comportamiento complejo de la variable estudiada genera una exactitud (accuracy) de 26.01%.

En Random Forest, al tener dos instancias para dividir un nodo, tener una instancia en cada nodo hoja y utilizar todas las variables de entrada para el modelo la profundidad es 63 y 152.300 nodos. Esto genera un equilibrio entre el BIAS y la varianza, lo que nos permite tener una buena estimación de las predicciones de reflectividad, por lo que se ha obtenido un BIAS de 0.000001 y una varianza de 70.5 como media, teniendo el pronóstico muy cercano a los valores reales pero muy dispersos por la varianza, el PCC de 0.99 nos dice que los valores están muy relacionados validando el valor del BIAS. Sin embargo, al sacar la diferencia de la varianza de los valores reales de las lecturas en contra los pronosticados existe una

diferencia inferior al 0.0335 y al tener el NSE de 0.99, nos permite comprobar que la varianza alta se produce por el complejo comportamiento de la variable estudiada dentro de la zona de estudio, pero los valores mantienen una relación directa. Dando como resultado una exactitud del 99.96% en promedio de todos los entrenamientos.

Al convertir los valores reales y los pronosticados de reflectividad mediante random forest y SVM con la relación Z-R a tasa de lluvia con los parámetros identificados en la zona de estudio por (Orellana-Alvear et al., 2019), se determinan los valores de la tasa de lluvia. Donde el modelo con random forest genera mejores resultados que SVM, teniendo resultados que están directamente relacionados entre ellos, teniendo un PCC de 0.99, BIAS de -0.0001 y el NSE de 0.99 nos expresa que la diferencia entre las varianzas de estos valores es muy baja. Además, tenemos un RMSE de 6.77, lo que nos indica que se incrementa el error al momento de realizar la conversión desde la reflectividad a tasa de lluvia. Finalmente, conociendo que la diferencia entre una lluvia moderada y fuerte es de 15 mm/h, concluimos que la diferencia entre los valores de tasa de lluvia va influir en la toma de decisiones ya que el pronostico se realiza por cada pixel del área de estudio.

Al finalizar la revisión de los resultados y conociendo que existen pluviómetros en la zona de estudio, se plantea como trabajo futuro utilizar estos datos de medición para validar los modelos de predicción o en su defecto utilizarlas como la columna objetivo en el entrenamiento. Donde ya se utilizaría la relación Z-R para transformar la reflectividad a tasa de lluvia y se podría analizar si el error se reduce evitando dicha relación no lineal.

5. Bibliografía

Bendix, J., Fries, A., Zárate, J., Trachte, K., Rollenbeck, R., Pucha-Cofrep, F., Paladines, R., Palacios, I., Orellana, J., & Oñate-Valdivieso, F. (2017). RadarNet-Sur first weather radar network in tropical high mountains. *Bulletin of the American Meteorological Society*, *98*(6), 1235-1254.

Capozzi, V., Picciotti, E., Budillon, G., & Marzano, F. (2014, septiembre 1). *X-band weather radar monitoring of precipitation fields in Naples urban areas: Data quality, comparison and analysis*.

Chakraborty, S., Nagwani, N., & Dey, L. (2014). Weather Forecasting using Incremental K-Means clustering. *arXiv preprint arXiv:1406.4756*.

Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2018). Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Computers & Electrical Engineering*, *70*, 66-73.

Donges, N. (2018, febrero 22). *The Random Forest Algorithm*. Towards Data Science. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Ellis, R. A., Sandford, A. P., Jones, G. E., Richards, J., Petzing, J., & Coupland, J. M. (2006). New laser technology to determine present weather parameters. *Measurement Science and Technology*, *17*(7), 1715. https://doi.org/10.1088/0957-0233/17/7/009

Kitchenham, B. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. 44.

Kumar, A., Sinha, R., Bhattacherjee, V., Verma, D. S., & Singh, S. (2012). *Modeling using K-means clustering algorithm*. 554-558.

Marra, F., & Morin, E. (2015). Use of radar QPE for the derivation of Intensity–Duration–Frequency curves in a range of climatic regimes. *Journal of Hydrology*, *531*, 427-440. https://doi.org/10.1016/j.jhydrol.2015.08.064

Mohandes, M. A., Halawani, T. O., Rehman, S., & Hussain, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy*, *29*(6), 939-947.

Orellana-Alvear, J., Célleri, R., Rollenbeck, R., & Bendix, J. (2019). Optimization of X-Band Radar Rainfall Retrieval in the Southern Andes of Ecuador Using a Random Forest Model. *Remote Sensing*, *11*(14), 1632. https://doi.org/10.3390/rs11141632

scikit-learn. (2018, noviembre 9). *Scikit-learn*. https://scikit-learn.org/stable/

Su, F., Gao, H., Huffman, G. J., & Lettenmaier, D. P. (2010). Potential Utility of the Real-Time TMPA-RT Precipitation Estimates in Streamflow Prediction. *Journal of Hydrometeorology*, *12*(3), 444-455. https://doi.org/10.1175/2010JHM1353.1

vigilant. (2014, junio 17). *Clasificación de la lluvia según su intensidad, para cualquier tiempo*. https://foro.tiempo.com/clasificacion-de-la-lluvia-segun-su-intensidad-para-cualquier-tiempo-t85594.0.html

Godoy Mendía, A. S. (2019-04-26). Aplicación de 'aprendizaje profundo' para el pronóstico de precipitación a partir de datos de reflectividad de radar meteorológico (Bachelor's thesis). Retrieved from http://dspace.ucuenca.edu.ec/handle/123456789/32551

6. Anexos

Anexo 1. Metodología de la revisión sistemática

Esta investigación se ha realizado siguiendo las directrices de Kitchenham (2007). La cual propone tres fases principales: A. Planificación del mapeo, llevar a cabo la revisión mediante un protocolo de revisión B. Ejecutar la revisión, realizar la revisión mediante una búsqueda planificada; y, C. Reporte de la investigación, presentación de los resultados de la investigación.

En primer lugar, la estrategia de búsqueda general establece las preguntas de búsqueda para ser respondidas por nuestro análisis de la literatura; luego, se definieron los criterios de selección de literatura para extraer las publicaciones relevantes sobre métodos de predicción de precipitaciones. Luego se dio una descripción detallada del proceso de selección de literatura. La idea clave de encontrar la literatura relevante es realizar una búsqueda inicial, filtrar los resultados, revisar las referencias de las publicaciones seleccionadas y finalizar a búsqueda agrupando y eliminando artículos duplicados.

- **Pregunta de investigación**

La siguiente pregunta de investigación ha sido respondida:

¿Cuáles son los métodos que están siendo utilizados para realizar predicción de precipitaciones? ¿La posición geográfica del radar cumple un papel fundamental para la recolección de datos, se realiza notificaciones de alertas y qué variables utilizando para realizarlo?

- **Sub preguntas de investigación**

Para responder la pregunta de investigación, se han respondido seis sub preguntas que nos guiaron para consolidar un tema tan extenso como la predicción de precipitaciones. Las Preguntas son las siguientes:

- SRQ1: ¿Qué métodos están siendo utilizados para predecir precipitaciones?
- SRQ2: ¿La predicción de precipitaciones cumple un papel fundamental para la toma de decisiones preventivas para desastres?
- SRQ3: ¿Qué efecto tienen los datos históricos en la predicción de precipitaciones?
- SRQ4: ¿Los métodos de predicción de Machine Learning generan mejores resultados que los modelos matemáticos?
- SRQ5: ¿La ubicación del radar cumple un papel fundamental para la investigación?

- SRQ6: ¿La utilización de los pluviómetros permiten afinar los modelos de predicción de Machine Learning?

La respuesta de estas preguntas de investigación nos proporciona una visión general de los enfoques de investigación actuales y nos ayudan a identificar áreas de investigación no cubiertas. Además, las respuestas a las preguntas anteriores nos permiten categorizar los métodos de predicción y definir los más utilizados por sus buenos resultados.

- **Estrategia de búsqueda**

Nuestro mapeo sistemático se basa en una búsqueda electrónica en las siguientes bibliotecas digitales:

- Google Académico.
- IEEE Xplore Digital Library.
- Springer Link.
- Science Direct.

Para cumplir con la revisión sistemática, se realizó una búsqueda avanzada en estas bases digitales con el fin de obtener citas bibliográficas de los principales autores científicos en el campo de predicción de precipitaciones. Esta estrategia puede omitir inevitablemente artículos útiles los cuales no están incluidos en las bases digitales consultadas. Además, se nota un gran incremento de investigación en el tema desde el año 2014 hasta la enero del 2019. Sin embargo, se tiene en cuenta artículos anteriores a esa fecha por su importancia del trabajo dentro del tema de investigación. Para realizar la búsqueda automáticamente en las bases digitales seleccionadas, planteamos usar la siguiente cadena de búsqueda descrita en la Tabla 1.

Concepto	Sub-String	Conector
Data Mining	Mining	Or
Machine Learning	Machine Learning	Or
Prevention	Prevention	And
Rainfall	Rainfall	And
Prediction	Prediction	And
Search string	(DATA MINING OR MACHINE LEARNING OR PREVENTION) AND (RAINFALL) AND (PREDICTION)	

Tabla 1: Criterios para la búsqueda avanzada

La búsqueda se realizó aplicando los mismos criterios de búsqueda a los mismos metadatos (nombre, título y palabras clave), para cada base digital. La cadena de búsqueda se adaptará para que puede ser aplicada en cada una.

- **Selección de estudios primarios**

Por cada búsqueda realizada en las diferentes bases digitales se escanea los artículos. Se incluyen, aquellos que cumplen con los siguientes criterios:

- Artículos indexados y publicados alrededor de los últimos 20 años en la investigación de predicción de precipitaciones.
- Artículos con métodos matemáticos para la predicción en los últimos 40 años.
- Generación de predicción con datos meteorológicos.
- Utilización de datos meteorológicos capturados por un radar.
- Utilización de modelos de predicción de precipitaciones.
- Artículos que estén citados al menos 10 veces.

Se excluirán los estudios que cumplan al menos uno de los siguientes criterios de exclusión.

- Predicción por medio de cadenas de Markov.
- Publicaciones con datos generados mediante algoritmos.
- Artículos que no estén indexados.

Según lo mencionado anteriormente, se establecieron los criterios de extracción, que sirven para la clasificación de los artículos. Los criterios se proponen teniendo en cuenta la relevancia del tema en respuesta a las preguntas establecidas en las secciones anteriores. Además, cada criterio se subdivide en varias opciones con las que mapear cada artículo. La clasificación de cada criterio se muestra a continuación en la Tabla 2.

SRQ1: ¿Qué métodos están siendo utilizados para predecir precipitaciones?		
EC1	Métodos utilizados	Estadísticos
		Matemáticos
		Inteligencia Artificial
		Machine Learning
SRQ2: ¿La predicción de precipitaciones cumple un papel fundamental para la toma de decisiones preventivas para desastres?		
EC2	Forma de exponer los resultados	Solo estudio
		Página normal
		Geovisor
EC3	Escala de la presentación de los resultados	Trimestral
		Mes
		Dia
		Horas

SRQ3: ¿Qué efecto tienen los datos históricos en la predicción de precipitaciones?		
EC4	Uso de datos históricos	Si
		No
EC5	Calidad de resultados	Buenos
		Medios
		Malos
SRQ4: ¿Los métodos de predicción de Machine Learning generan mejores resultados que los modelos matemáticos?		
EC6	Machine Learning genera mejores resultados que métodos matemáticos	Si
		No
SRQ5: ¿La ubicación del radar cumple un papel fundamental para la investigación?		
EC7	Uso del radar para la captura de los datos	Si
		No
EC8	Altitud del estudio sobre el nivel del mar	< 1.000
		1.001 < alt < 1.500
		1.501 < alt < 2.000
		>2.001
SRQ6: ¿La utilización de los pluviómetros permiten afinar los modelos de predicción de Machine Learning?		
EC9	Uso de datos capturados con pluviómetros para la predicción	Si
		No
EC10	Uso de pluviómetros para refinar modelos de predicción	Si
		No

Tabla 2: Criterios de extracción de datos

Se han encontrado alrededor de 20 artículos de los cuales se han tomado 5 como los más influyentes o importantes en nuestra investigación, con los cuales definimos el estado del arte presente en la introducción de la esta tesis.

Anexo 2. Implementación de los algoritmos de SVM y RF

La implementación de la presente tesis se encuentra en el repositorio denominado “Implementación de modelos de Machine Learning para pronóstico de reflectividad” en Bitbucket(https://cutt.ly/sr1Y2o9). Es importante solicitar los datos de reflectividad capturados por el radar de banda X CAAX al Departamento de Recursos Hídricos de la Universidad de Cuenca (iDRHICA) en formato NETCDF.

Está implementación se realizó en el lenguaje de programación Python 3.6 con librerías como: pandas, scikit-learn, matplotlib, scipy y numpy. A continuación, se presenta una descripción de los principales algoritmos (scripts) que se utilizan para el desarrollo de la tesis:

- **readData:** Lectura, comprobación de datos en formato NETCDF y retorna un archivo en formato csv con la sección de datos correspondiente a la zona de estudio.
- **showDataAsimilation:** Calculo de la mediana y desviación estándar del intervalo de los datos escogido para el estudio. Presentación de la imagen de esta sección y exporta un archivo en formato csv con los datos del intervalo por cada pixel con la mediana y la desviación estándar (azimuth, rangebin, mediana, desviación estándar). Este script elimina los valores del archivo con resultados de reflectividad nan (los cuales no se pueden remplazar o realizar un ajuste).
- **generaDataTraining:** Este archivo toma los valores generados por el script showDataInterval y realiza el proceso para obtener las instancias de ingreso a los modelos de Machine Learning, tanto para SVM y RF.
- **rf:** Se presenta la implementación de la técnica Random Forest, donde se realiza la normalización de las instancias que van ser utilizadas en el entrenamiento. Se realiza la división estratificada del conjunto de datos: 80% entrenamiento y 20% validación (testing). Se realiza el entrenamiento de la técnica y la posterior validación. En cada sección se presenta los valores de los criterios de evaluación con el conjunto de validación (testing). Además, se presenta una imagen por la importancia de las variables utilizadas, se almacena el entrenamiento del modelo. Finalmente, se almacenan en formato csv los valores de pronóstico, los reales de las lecturas tanto de la reflectividad con los datos normalizados, datos reales y la tasa de lluvia.
- **svm:** Se presenta la implementación de la técnica SVM, donde se realiza la normalización de las instancias que van ser utilizadas en el entrenamiento. Se realiza la división estratificada del conjunto de datos: 80% entrenamiento y 20% validación (testing). Se realiza el entrenamiento de la técnica y la posterior validación. En cada sección se presenta los valores de los criterios de evaluación con el conjunto de validación (testing). Además, se almacena el entrenamiento del modelo y el ajuste de los valores del pronóstico.

Finalmente, se almacenan en formato csv los valores de pronóstico, los reales de las lecturas tanto de la reflectividad con los datos normalizados, datos reales y la tasa de lluvia.

En el repositorio existen más archivos con nombres descriptivos que permiten visualizar los resultados en gráficos, entre otras actividades. Para ejecutar los scripts de Python se debe utilizar el interprete desde un IDE o desde el terminal (linux) o cmd (windows) con el comando “python nombreArchivo.py”.