ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

ESTUDIO DE LA CALIDAD DE LA CONEXIÓN EN REDES DE TELEFONÍA CELULAR EN BASE A MEDICIONES DE CAMPO Y TÉCNICAS DE MACHINE LEARNING

ESTUDIO DEL COMPORTAMIENTO DE LOS PARÁMETROS DE RF EN MODO INACTIVO Y ESTÁTICO EN REDES DE TELEFONÍA CELULAR EN BASE A MEDICIONES DE CAMPO Y TÉCNICAS DE MACHINE LEARNING.

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN TELECOMUNICACIONES

ALEX ANDRÉS PÁEZ LEMA

alex.paez@epn.edu.ec

DIRECTOR: Ph.D. PABLO ANÍBAL LUPERA MORILLO

pablo.lupera@epn.edu.ec

DMQ, agosto 2023

CERTIFICACIONES

Yo, Alex Andrés Páez Lema, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

_____

ALEX ANDRÉS PÁEZ LEMA

Certifico que el presente trabajo de integración curricular fue desarrollado por Alex Andrés Páez Lema, bajo mi supervisión.

_____

Ph.D. PABLO ANÍBAL LUPERA MORILLO

DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

ALEX PÁEZ

PABLO LUPERA

DEDICATORIA

A mi padre y a mi madre, Alex y Patricia, quienes con su amor y sacrificio me permitieron llegar hasta este punto tan importante en mi vida, todo esto es gracias a ellos.

AGRADECIMIENTO

Quiero agradecer a mi familia, en especial a mis padres, Alex y Patricia, y a mi hermano Xavier, quienes siempre estuvieron en las buenas y en las malas, mostrándome su apoyo incondicional en cada instante de esta etapa universitaria, con amor, cariño y paciencia.

A mi novia, Alisson Navas, quien siempre me ha apoyado de manera incondicional en cada uno de los objetivos que me he propuesto, por inspirarme a ser mejor, día a día.

A mi buen amigo Blue, quien, con su carisma y ternura, me llenaba de paz y alegría para empezar y terminar cada día de la mejor manera posible.

A mis amigos, Alex Ramos, Jonathan Villareal, Michelle Chiluisa, Ibeth Sotalín, Daysi Guano y Paula Altamirano, por siempre estar cuando tenían que estar, por todos esos gratos momentos compartidos en las aulas de clase.

Finalmente, expresar mi gratitud con el Dr. Pablo Lupera quien fue capaz de brindarme su apoyo a lo largo de este trabajo de integración curricular.

ÍNDICE DE CONTENIDO

RESUMEN

PALABRAS CLAVE: aprendizaje automático, calidad, factores ambientales, parámetros de RF.

Hoy en día, la calidad de la señal se vuelve un tema sumamente importante, dada la necesidad de vernos conectados en la red a través de nuestro teléfono móvil. Por tal motivo es importante conocer en qué condiciones la calidad de esta señal puede verse afectada, es por ello que en este estudio se tienen presentes parámetros de RF que permiten identificar bajo qué condiciones se tienen calidades etiquetadas como buena, media o mala; para ello se tiene al parámetro RSRQ (Reference Signal Received Quality) medido en dB que proporciona rangos bajo los cuales se tienen tales etiquetas. Además de este parámetro de RF también se incluyen otros dos, los cuales son RSSI (Received Signal Strength Indicator) y RSSNR (Reference Signal Signal to Noise Ratio), mismos que servirán como variables predictoras para el análisis de la calidad de la señal recibida por el teléfono móvil. Así como se han incluido estos parámetros de RF, se han incluido factores climáticos como variables predictoras con el objetivo de verificar si alguno de ellos repercute de manera negativa en la calidad de la señal recibida, estos factores que se consideraron son la temperatura, la presión atmosférica, la radiación ultravioleta y la velocidad del viento. Para llevar a cabo este estudio se utilizaron dos técnicas de clasificación de aprendizaje automático, estas técnicas fueron la de los árboles de decisión y los k-vecinos más cercanos. Los resultados obtenidos tienen precisiones que superan el 70%, lo cual indica un adecuado funcionamiento de estos algoritmos.

ABSTRACT

KEYWORDS: Environmental factors, machine learning, quality, RF parameters.

Nowadays, signal quality has become an extremely important issue due to the need to stay connected through our mobile phones. For this reason, it is important to understand the conditions under which signal quality can be affected. Hence, this study takes into consideration RF parameters that allow the identification of conditions categorized as good, moderate, or poor quality. To achieve this, the RSRQ parameter (Reference Signal Received Quality), measured in dB, is used to establish ranges for these quality labels. In addition to this RF parameter, two others are included: RSSI (Received Signal Strength Indicator) and RSSNR (Reference Signal Signal to Noise Ratio). These parameters will serve as predictive variables for analyzing the quality of the signal received by the mobile phone.

Alongside these RF parameters, environmental factors are also included as predictive variables to determine whether any of them have a negative impact on received signal quality. The considered factors include temperature, atmospheric pressure, ultraviolet radiation, and wind speed. To conduct this study, two machine learning classification techniques were employed: decision trees and k-nearest neighbors. The obtained results have accuracies exceeding 70%, indicating the effective performance of these algorithms.

1 INTRODUCCIÓN

El objetivo de estudio de este trabajo consistió en determinar en qué condiciones se ve comprometida la calidad de la señal, en una red celular. Para ello se hizo uso de aplicaciones que permiten capturar parámetros de RF, por ejemplo, el RSSI y el RSSNR pertinentes en este estudio. Las aplicaciones que permitieron esta recolección de datos fueron: CellMapper y Net Monitor. Además de la recolección de estos parámetros también se incluyeron las mediciones de factores ambientales tales como la temperatura, la presión atmosférica, la radiación ultravioleta y la velocidad del viento. Para identificar si la calidad de recepción de la señal es buena, media o mala se utilizó como referencia al parámetro RSRQ (Reference Signal Received Quality). Los rangos utilizados para clasificar a la señal recibida se describirán posteriormente en este documento. Los datos recolectados, antes de ser analizados, debieron preprocesarse con el objetivo de contar con valores apropiados que se ajusten a la teoría y clasifiquen adecuadamente la calidad de la señal. Una vez que los datos fueron preprocesados se procedió a realizar un análisis estadístico para determinar que la desviación de las variables con respecto a su media y saber que variables (de RF o climáticas) pueden comprometer la calidad de la señal, para ello se hizo uso de tres mecanismos muy importantes: el uso de diagramas de caja, el análisis ANOVA y el análisis Tukey; el primero de estos análisis representa una forma gráfica de visualizar la distribución de los datos. Posterior a estos análisis, se procedió a seleccionar las variables que pueden influir en la calidad de la señal, para incluirlas como variables predictoras en los respectivos modelos de aprendizaje automático. Los modelos de aprendizaje utilizados fueron modelos de clasificación, por tal motivo se hizo uso de los árboles de decisión y los k-vecinos más cercanos. Ambos modelos utilizados presentan una alta precisión, puesto que en la mayoría de los puntos de medición se supera el 70% de precisión, lo cual permite predecir de mejor manera la calidad de la señal recibida. Otro de los métodos de evaluación para uno de los modelos de aprendizaje automático (árbol de decisión) corresponde a la matriz de confusión. Los árboles de precisión tienen la gran ventaja de presentar una forma gráfica para visualizar la clasificación de los datos, por tal motivo es más sencillo saber en qué circunstancias se tiene una calidad buena, media o mala. Una vez que se han analizado los árboles de decisión en cada punto de medición, existen factores climáticos que podrían repercutir negativamente en la calidad de la señal, asimismo existen rangos en los parámetros de RF que repercuten en una inapropiada calidad de la señal recibida.

1.1 OBJETIVO GENERAL

Analizar parámetros de RF y otros externos a la red mediante técnicas de aprendizaje automático por clasificación de datos recolectados por el teléfono celular en estado estático e inactivo, con el objetivo de identificar en qué condiciones y en qué medida estos parámetros repercuten en la calidad de la señal recibida.

1.2 OBJETIVOS ESPECÍFICOS

1. Medir los parámetros de RF en el modo inactivo y estático mediante el uso de aplicaciones pertinentes para dicho propósito, entre ellas: Net Monitor y CellMapper.
2. Depurar los datos más relevantes para incluirlos en el modelo de aprendizaje propuesto.
3. Comparar los resultados de al menos dos técnicas de aprendizaje automático a fin de identificar cuál de ellas presenta una mayor precisión para el análisis de los datos recolectados.
4. Analizar los resultados obtenidos y verificar qué parámetros repercuten en mayor medida, a la calidad de la señal.

1.3 ALCANCE

Describir el alcance del componente de acuerdo con lo establecido en el Plan. Este componente corresponde al modo inactivo y estático. Las mediciones de los parámetros de RF se recolectarán en los UE mediante el uso de aplicaciones específicas instaladas y configuradas adecuadamente. El registro de las mediciones se realizará dentro de una zona específica, en cada ubicación el registro será continuo en el lapso de una semana sin el uso de ninguno de los servicios de la red celular, durante las mediciones el UE estará configurado para que se conecte a la red de cualquier tipo de tecnología de telefonía celular y se desconectará la opción de conexión a una red WiFi. Con los datos recolectados se realizará un análisis estadístico básico y se aplicarán al menos 2 técnicas de Machine Learning para observar el comportamiento de los datos y se tratará de identificar las condiciones en las cuales se presentan reducciones en los niveles de la calidad de la conexión. En los análisis se considerarán los parámetros de RF y además factores externos a la red como la ubicación del UE con respecto a la estación base, la distancia hacia la estación base, el periodo del día y el día de la semana, entre otros. El análisis de los datos se realizará con el uso de R.

1.4 MARCO TEÓRICO

1.4.1 Herramientas de recolección de datos

1.4.1.1 CellMapper

Esta es una aplicación móvil útil para la búsqueda de estaciones base que proveen servicios 2G, 3G, 4G y 5G. La aplicación se encarga de medir la intensidad de la señal así como otros parámetros de red, recolectados por usuarios finales; esto se hace con el objetivo de localizar las estaciones base y su cobertura. Esta aplicación se encuentra disponible para dispositivos móviles de Android y Windows 10 [1].

CellMapper permite la incorporación de varias herramientas de las cuales el usuario puede hacer uso, una de las más importante es la opción "*Mapa*", que permite la visualización de las estaciones base a las cuales el usuario podría ser sujeto de una posible conexión. Los datos recolectados con esta aplicación se almacenan en archivos CSV. La Figura 1 presenta la interfaz gráfica de la aplicación [1].

Figura 1. Interfaz gráfica de la aplicación (CellMapper).

1.4.1.2 Net Monitor Cell Signal Logging Lite

Esta es una aplicación móvil disponible para el sistema operativo Android, a partir de la versión 4.3 en adelante. Net Monitor permite el monitorio de tecnologías tales como GSM, WCDMA y LTE [2]. Además, permite el almacenamiento de las mediciones en archivos con formato CSV y KML para su posterior depuración y análisis de datos. Al igual que la herramienta anterior, permite la inclusión de gráficos y estadísticas de red. La Figura 2 presenta la interfaz gráfica de la aplicación [2].

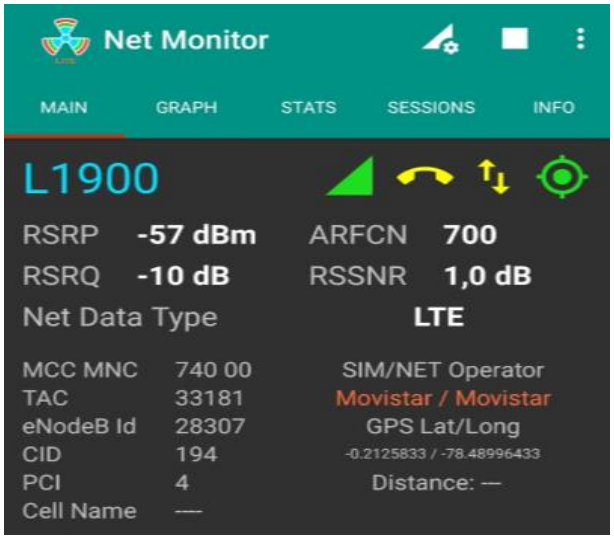

Figura 2. Interfaz gráfica de la aplicación (Net Monitor).

1.4.2 DESCRIPCIÓN ACERCA DE LOS ATRIBUTOS ADICIONALES A CONSIDERAR EN ESTE ESTUDIO.

En esta sección se consideran atributos adicionales que no pueden ser recolectados con las herramientas mencionadas anteriormente. El objetivo del registro de estos atributos es analizar su influencia en la calidad de las conexiones.

Además de los parámetros descritos anteriormente, se considerarán como parámetros adicionales ciertos factores atmosféricos como: la temperatura atmosférica, la radiación ultravioleta, la velocidad del viento y la presión atmosférica. Estos parámetros se calcularon y registraron durante el proceso de recolección de datos. El objetivo de tomar en cuenta estos parámetros es revisar si dichos factores influyen de alguna manera en la calidad de las conexiones de las redes celulares en la zona de estudio.

1.4.3 DESCRIPCIÓN DE LA ETAPA DE RECOLECCIÓN DE DATOS.

Objetivos de la recolección de datos

- Definir las áreas de interés para llevar a cabo la recolección de datos.
- Medir los parámetros de radiofrecuencia (RF) asociados a las redes celulares de tercera y cuarta generación.

1.4.4 IDENTIFICACIÓN Y DEFINICIÓN DE LOS PARÁMETROS DE RADIO FRECUENCIA QUE SE RECOLECTARON.

Los parámetros que se recolectaron son los siguientes:

- **RSSI (Received Signal Strenght Indicator):** esta medida es ampliamente utilizada en el campo de las comunicaciones inalámbricas, la cual sirve para cuantificar la intensidad de una señal recibida por un dispositivo [3]. Este parámetro se expresa en dBm y proporciona un indicador numérico de la potencia de la señal recibida. Dentro del contexto de redes celulares, este indicador está presente en redes 3G y 4G [3].

- **RSRQ (Reference Signal Received Quality):** esta medida expresada en dB es ampliamente utilizada en redes celulares, propia de redes de cuarta y quinta generación (4G y 5G), sirve para evaluar la calidad de la señal recibida en relación con el nivel de interferencia [4]. Esta medida se expresa en dB. Cuanto mayor sea el valor de RSRQ, mejor será la calidad de la señal recibida [4]. La calidad de una señal recibida se puede clasificar en tres grupos (excelente, media, mala), de acuerdo con los siguientes criterios que se pueden aplicar para redes LTE (4G) [4]:

  - **Buena:** cuando se tiene un RSRQ mayor o igual a -15dB
  - **Media:** cuando se tiene un RSRQ en un rango de -14dB a -19dB.
  - **Mala:** cuando se tiene un RSRQ menor o igual -20dB.

- **RSSNR (Reference Signal Signal to Noise Ratio):** es una medida que permite evaluar la calidad de una señal recibida [5]. Proporciona una indicación de la relación existente entre la potencia de la señal deseada y la suma de la potencia de las señales no deseadas, tales como la interferencia de otras señales, ruido térmico, ruido atmosférico, entre otros.

1.4.5 ESPECIFICACIÓN DE LAS ZONAS EN DONDE SE RECOLECTARON LOS DATOS.

La zona donde se recolectaron los datos se encuentra en los alrededores de la Escuela Politécnica Nacional, en dichas zonas se seleccionaron diez puntos geográficos

- **Punto A:** Edificio del Centro de Educación Continua (CEC).
- **Punto B:** Edificio de Química Eléctrica.
- **Punto C:** Escuela de formación de Tecnólogos (ESFOT).

- **Punto D**: Facultad de Ingeniería Eléctrica y Electrónica (Hall del edificio).
- **Punto E:** Facultad de Ingeniería Eléctrica y Electrónica (Aula S1/E001).
- **Punto F:** Estadio de la Escuela Politécnica Nacional.
- **Punto G:** Patio aledaño al Instituto de Ciencias Básicas (ICB).
- **Punto H:** Afueras de la Biblioteca Central.
- **Punto I:** Afueras de la Casa Patrimonial de la EPN.
- **Punto J:** Entrada Facultad de Ingeniería Civil.

La figura 3 muestra las zonas en donde se recolectaron los datos:

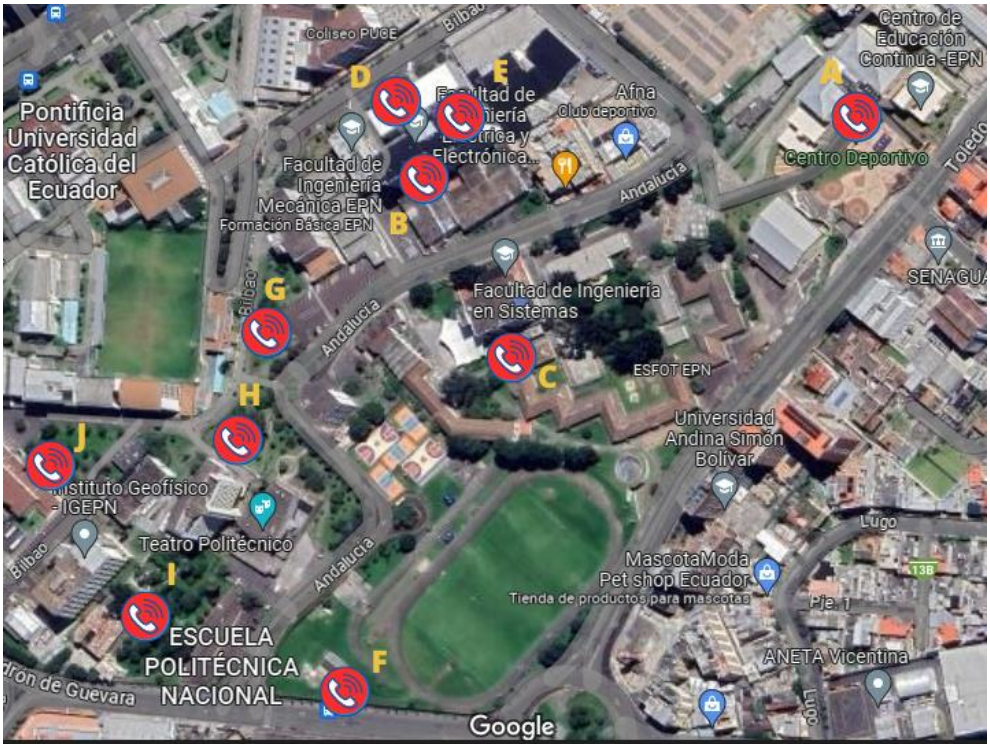

Figura 3. Puntos de recolección de datos.

La figura 4 presenta la ubicación de las estaciones base, aledañas a la zona de interés, para la recolección de datos.

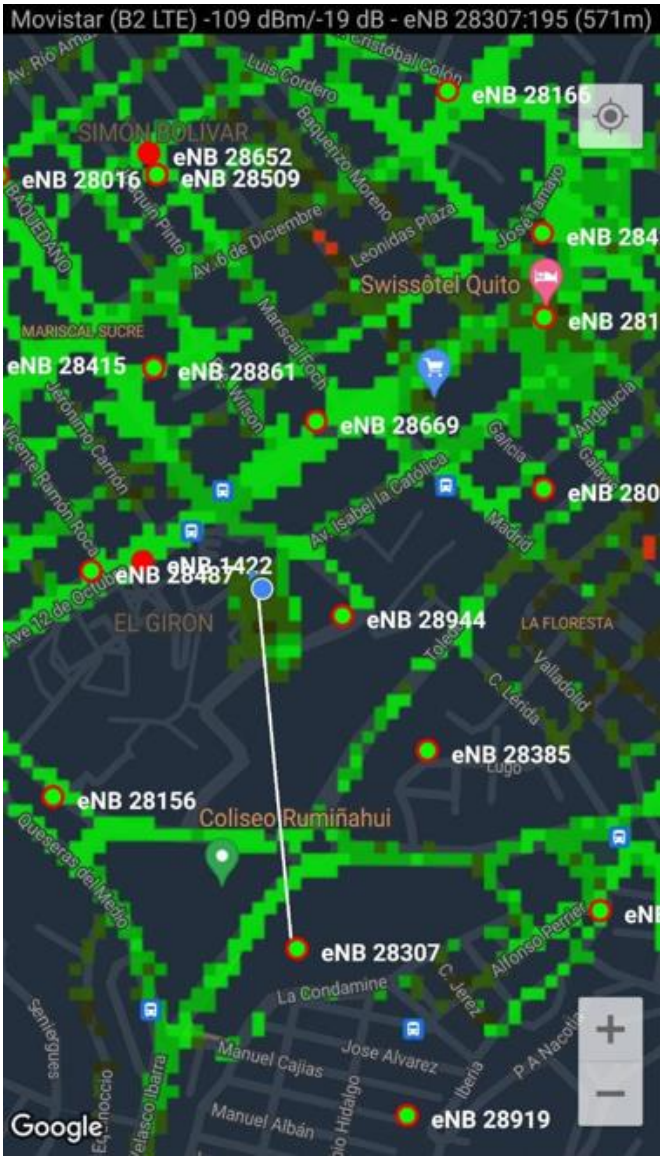

Figura 4. Estaciones base aledañas a los puntos de recolección de datos.

1.4.6 CRONOGRAMA DE LA ETAPA DE RECOLECCIÓN DE DATOS.

La etapa de recolección de datos se realizó en base al siguiente cronograma:

Tabla 1. Cronograma para la etapa de recolección de datos.

Identificación	Lugar	Fecha
A	Centro de Educación Continua (CEC)	2023-06-21
		2023-06-23
		2023-06-27
		2023-06-28
		2023-06-30
B		2023-06-21

	Edificio de Química Eléctrica	2023-06-23
		2023-06-26
		2023-06-27
		2023-06-29
		2023-06-30
C	Escuela de Formación de Tecnólogos (ESFOT)	2023-06-23
		2023-06-27
		2023-06-28
		2023-06-30
D	Hall de la Facultad de Ingeniería Eléctrica y Electrónica (FIEE).	2023-06-21
		2023-06-22
		2023-06-23
		2023-06-26
		2023-06-28
E	Aula de la Facultad de Ingeniería Eléctrica y Electrónica (FIEE).	2023-06-21
		2023-06-28
F	Estadio de la Escuela Politécnica Nacional.	2023-06-21
		2023-06-22
		2023-06-23
		2023-06-26
		2023-06-27
		2023-06-28
		2023-06-30
G	Patio del Instituto de Ciencias Básicas.	2023-06-21
		2023-06-23
		2023-06-27
		2023-06-28
		2023-06-30
H	Biblioteca Central	2023-06-22
		2023-06-23
		2023-06-26
		2023-06-27
		2023-06-28

I	Casa Patrimonial de la EPN.	2023-06-21
		2023-06-22
		2023-06-27
		2023-06-28
		2023-06-30
J	Entrada Facultad de Ingeniería Civil	2023-06-22
		2023-06-27
		2023-06-28
		2023-06-30

1.4.7 DESCRIPCIÓN DE LOS ATRIBUTOS ADICIONALES.

- **Temperatura:** a medida que la temperatura aumenta, existe una relación inversamente proporcional con el RSSI [6]. Se conoce que variaciones en la temperatura pueden causar pérdida de sincronización, degradación en la calidad del enlace, lo cual repercute en el rendimiento de la comunicación inalámbrica.

- **Velocidad del viento:** al igual que sucede con la temperatura, este parámetro puede influenciar en el rendimiento de la comunicación, no en gran medida como el anterior; sin embargo, es un parámetro pertinente para la evaluación de su influencia en la calidad de la conexión en las redes celulares [6].

- **Presión atmosférica e índice de radiación ultravioleta:** a pesar de que en otros estudios no se presentan como factores climáticos determinantes en cuanto al estudio de la calidad de conexión en redes celulares, el propósito de su inclusión en este estudio consiste en el hecho de determinar si influyen o no en la calidad de la conexión de una red celular.

- **Radiación ultravioleta:** el estudio correspondiente expondrá si este factor influye o no en la calidad de la conexión.

- **Período del día:** este parámetro será considerado con el objetivo de identificar si existen variaciones en la calidad de la señal en los siguientes dos períodos: mañana (0) o tarde (1).

1.4.8 BREVE DESCRIPCIÓN DE R Y SUS CAPACIDADES.

Para el análisis de los data sets se hará uso del software R, el cual corresponde a un entorno de software libre para el análisis estadístico [7]. R se ejecuta en una gran variedad de plataformas, estas incluyen: Windows, Unix y MacOS.

Entre la amplia variedad de técnicas estadísticas y gráficas que puede ofrecer R se encuentran los modelados lineales y no lineales, pruebas estadísticas clásicas, técnicas de clasificación, clustering, entre otras [7].

R facilita la manipulación de datos, el cálculo y la visualización de gráficos. A continuación, se presentan algunas de las facilidades que incluye este software [7]:

- Manejo y almacenamiento efectivo de los datos que se deseen manipular.
- Un conjunto de operadores para realizar cálculos de arreglos (sobre todo en matrices)
- Una colección integrada de herramientas para el análisis de datos, facilidades gráficas para el análisis y visualización de datos.
- Un lenguaje de programación bien desarrollado, simple y efectivo dentro del cual se pueden incluir diversas funciones, tales como bucles, funciones recursivas definidas por el usuario, entre otras.

1.4.9 DESCRIPCIÓN DE LOS ASPECTOS TEÓRICOS ACERCA DE LOS ANÁLISIS ESTADÍSTICOS PRESENTADOS EN ESTE ESTUDIO.

En este estudio se propone hacer uso de los siguientes análisis estadísticos, mismos que se describirán brevemente a continuación:

- **Media:** es el valor promedio de un determinado grupo de datos [8]. Es la suma de los números en la muestra, dividida entre la cantidad total de datos que existen.
- **Desviación estándar:** es una medida que permite determinar cómo varían los datos alrededor de la media, es decir, mide el grado de dispersión en una muestra [8].
- **Diagramas de caja:** un diagrama de este tipo es una gráfica que presenta la mediana, el primer y tercer cuartiles además de cualquier dato atípico que se encuentre presente en una determinada muestra [8].
- **Análisis ANOVA (Analysis of Variance):** es una herramienta de análisis usado en la estadística que permite la comparación de la media de tres o más grupos y así determinar si existen diferencias significativas entre ellas [9].
- **Análisis Tukey:** es un procedimiento que permite la comparación de medias que pueden resultar diferentes entre sí, de manera significativa [10].

1.4.10 DESCRIPCIÓN DE LOS ASPECTOS TEÓRICOS ACERCA DE LOS ANÁLISIS ESTADÍSTICOS PRESENTADOS EN ESTE ESTUDIO.

Las técnicas de Machine Learning (ML) o de aprendizaje automático se dividen principalmente en dos tipos, estos son: supervisados y no supervisados. A continuación, se presenta un breve resumen de cada uno de estos tipos de técnicas:

- **Aprendizaje supervisado:** este tipo de aprendizaje se basa en la creación de un modelo que realiza predicciones sobre nuevas instancias, para ello se proporciona al algoritmo un conjunto de datos etiquetados que consisten en ejemplos de entrada y su correspondiente salida esperada con lo cual se busca entrenar al modelo para generar predicciones razonables que generen una respuesta adecuada para los datos de entrada proporcionados. Este tipo de aprendizaje es práctico cuando se tienen datos conocidos para la salida que se tratan de estimar [11].

  Las técnicas de aprendizaje supervisado incluyen modelos de clasificación y regresión para realizar predicciones. A continuación, se describe cada una de ellas:

  - **Clasificación:** este modelo se encarga de la clasificación de los datos de entrada. Estos modelos predicen respuestas discretas [11]. Entre los algoritmos más utilizados se encuentran los árboles de decisión, los K-vecinos más cercanos, análisis discriminante, análisis bayesiano, regresión logística y redes neuronales [11].

  - **Regresión:** este modelo predice respuestas continuas en función de los datos de entrada ingresados [11].

- **Aprendizaje no supervisado:** este tipo de aprendizaje trabaja con datos no etiquetados, con lo cual no se le proporciona ninguna salida esperada. El objetivo de esta técnica de aprendizaje radica en el descubrimiento de patrones, estructuras o relaciones ocultas en los datos. El modelo de agrupamiento (clustering) es uno de los más utilizados dentro de esta técnica. Dentro de este modelo se encuentran algoritmos como k-means, k-medoids, agrupamiento jerárquico (hierarchical clustering), modelos gaussianos, modelos de Markov, entre otros [11].

1.4.11 DESCRIPCIÓN DE LAS ETAPAS DE RECOLECCIÓN, PREPROCESAMIENTO, ENTRENAMIENTO, PRUEBA, EVALUACIÓN Y AJUSTE.

Las etapas para presentar en esta sección son necesarias para la aplicación de las técnicas de aprendizaje automático y se describen de acuerdo con su uso en este estudio.

- **Etapa de recolección:** dentro de la etapa de recolección de datos, se procedió a obtener mediciones en diez puntos diferentes, dentro del campus de la Escuela Politécnica Nacional. En estos diez puntos se recolectaron datos relacionados con algunos de los parámetros de radiofrecuencia útiles para identificar el nivel y la calidad de la señal que llega a un dispositivo móvil (UE); además, como se mencionó anteriormente se consideraron parámetros adicionales, tales como la distancia y el ángulo de azimut entre la estación base (eNB) y el dispositivo móvil (UE), en conjunto con factores medioambientales como la temperatura, la presión atmosférica, el índice de radiación ultravioleta y la velocidad del viento. Todos estos parámetros, en conjunto, tienen el objetivo de procesarse a través de ciertos modelos de aprendizaje automático para determinar cómo influyen en la calidad de la señal recibida por el dispositivo móvil.

- **Etapas de preprocesamiento:** en esta etapa se procedió a eliminar todos los datos que contengan un valor numérico erróneo, de tal forma que no afecten a los resultados obtenidos tras procesarlos en un modelo de aprendizaje automático.

- **Etapa de entrenamiento:** para la etapa de entrenamiento, el modelo de aprendizaje se entrena al hacer uso de un conjunto de datos de entrenamiento. Este conjunto de datos consta de atributos de entrada y salida. Durante la etapa de entrenamiento, el modelo ajusta sus parámetros internos de manera que pueda aprender patrones y así realizar predicciones más precisas [12].

- **Etapa de prueba:** una vez que el modelo ha sido entrenado, se evalúa su rendimiento al hacer uso de un conjunto de datos de prueba. Este conjunto de datos es diferente al conjunto de datos de entrenamiento, dado que contiene patrones que el modelo no había visto anteriormente. Por lo tanto, el conjunto de datos de prueba permite medir cómo se comporta el modelo con datos desconocidos [12].

- **Etapa de evaluación:** esta etapa consiste en mediar y analizar el rendimiento del modelo en el conjunto de datos de prueba. Algunas de las métricas que permiten

evaluar el modelo son: matrices de confusión, precisión, sensibilidad, especificidad, el área bajo la curva ROC, entre otros [13].

- **Etapa de ajuste:** si el rendimiento del modelo no es satisfactorio, entonces es necesario el ajuste y la mejora del modelo, lo que implica generar modificaciones en el algoritmo de aprendizaje, agregar o eliminar atributos que permitan esta mejora [13].

1.4.12 TÉCNICAS DE APRENDIZAJE AUTOMÁTICO CONSIDERADAS.

Para este estudio se tomará a consideración al menos dos técnicas de aprendizaje automático, entre las cuales se encuentran las siguientes: árbol de decisiones, k-vecinos más cercanos. A continuación, se describirá cada una de estas técnicas aplicadas:

- **Árboles de decisión:** esta técnica consiste en un algoritmo utilizado para problemas de clasificación, la cual se basa en la construcción de un modelo en forma de árbol, donde cada nodo interno representa un atributo y cada rama del árbol representa un valor de salida [14]. El proceso de construcción en los árboles de decisión se realiza de manera recursiva al dividir el conjunto de datos en subconjuntos más pequeños, en función de los valores de los atributos, con el objetivo de maximizar la pureza en los valores de salida en cada subconjunto resultante [14]. Una vez que el árbol de decisión se haya modelado, este puede utilizarse para realizar predicciones de nuevos datos. Para ello, se sigue el camino desde la raíz del árbol hasta una de sus hojas, tomando en cuenta las decisiones en cada nodo interno basadas en los valores de los atributos.

  Un árbol de decisión tiene como punto inicial al **nodo root**, el cual hace referencia a toda la muestra de datos, los cuales luego se dividen en dos o más grupos uniformes mediante un método denominado **splitting** (división). Cuando los subnodos se dividen más se identifican como **nodos de decisión**, mientras que, los que no se dividen se denominan **nodos terminales** u **hojas** [15]. El diagrama asociado a un árbol de decisión puede apreciarse en la figura 5.

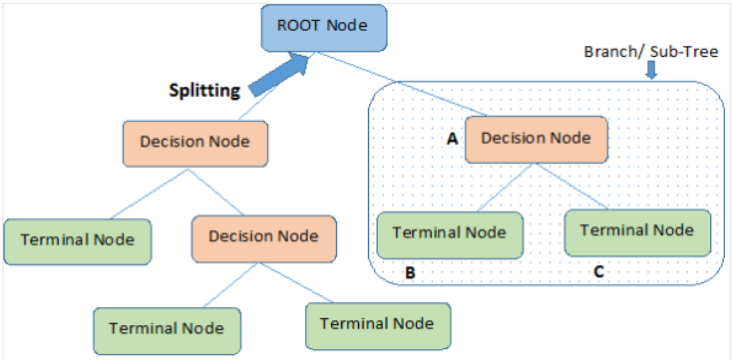

Figura 5. Esquema de un árbol de decisión.

- **K-vecinos más cercanos (k-Nearest Neighbors):** abreviado como k-NN es un algoritmo utilizado para problemas de clasificación. Esta técnica de aprendizaje se basa en la idea de que la agrupación de datos similares tiende a estar cerca unos de otros en el espacio de características [16]. El algoritmo se encarga de predecir las etiquetas del conjunto de datos de prueba al observar las etiquetas de sus vecinos más cercanos en el espacio de características del conjunto de datos de entrenamiento. El hiperparámetro **"K"** es el más importante, dado que se puede ajustar para mejorar el desempeño del modelo [16].

2 METODOLOGÍA

Este capítulo presentará el procedimiento aplicado para la obtención de los análisis realizados en este estudio. Se detallarán los análisis estadísticos y los modelos de aprendizaje automático utilizados.

2.1 ANÁLISIS ESTADÍSTICOS

2.1.1 Diagrama de caja

Un diagrama de caja refleja una representación de un conjunto de datos numéricos. Esta herramienta es sumamente útil para el análisis y la comparación de la distribución de diferentes grupos de datos [17]. Estos diagramas de caja se basan en cinco conceptos estadísticos y descriptivos de un conjunto de datos, estos son:

- **Mediana (Q2):** este es el valor que divide el conjunto de datos en dos partes iguales, con el 50% de los datos por encima y el 50% por debajo [17], [18].
- **Primer cuartil (Q1):** representa el valor en el que el 25% de los datos son menores que este y el 75% son mayores que este valor [17], [18].
- **Tercer cuartil (Q3):** representa el valor en el que el 75% de los datos son menores que este y el 25% son mayores que este [17], [18]..

- **Rango intercuartílico (IRQ):** consiste en la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), es decir, IQR=Q3-Q1. Es útil para estimar la dispersión de la distribución. Este parámetro representa el 50% de los datos centrales [17], [18].
- **Valores atípicos (outliers):** son los valores que se encuentran significativamente alejados de la mayoría de los datos y se representan como puntos individuales en el diagrama de caja. Un valor se considera atípico si es más grande que Q3+1.5IQR o si es menor que Q1-1.5IQR [17], [18]..

La figura 6 ilustra los conceptos estadísticos que forman parte de un diagrama de caja [17].

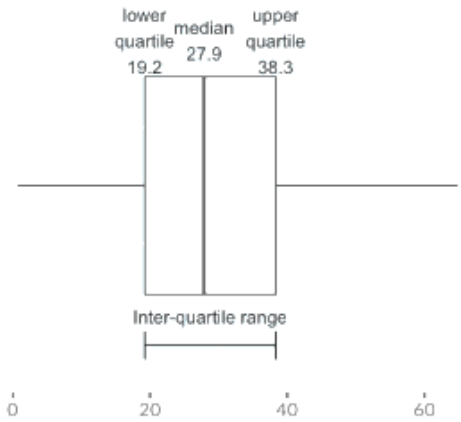

Figura 6. Representación gráfica de un diagrama de caja (boxplot).

Nota: con el objetivo de presentar la implementación de los análisis estadísticos, así como los modelos de aprendizaje automático, se mostrará como ejemplo el proceso aplicado al conjunto de datos pertenecientes al punto A, con ello se entenderá la forma en que se aplicaron los procedimientos. Para el análisis de los demás puntos, ya no se detallará paso a paso la implementación de estos modelos y análisis estadísticos dado que el procedimiento es similar para todos.

Punto B: Interiores del edificio de química eléctrica.

El comando utilizado (Código 1) para la visualización de estos diagramas de caja es el siguiente:

Código 1. Visualización del diagrama de cajas.

```
qe_uv_box<-boxplot (uv~calidad_rx, data=qe)
```

Explicación:

- La función “boxplot ()” construye un diagrama de caja con el objetivo de visualizar la distribución de la variable “uv” en función de la variable categórica “calidad_rx”, mediante el uso del conjunto de datos denominado “qe**”** [19].
  - La variable **“**uv***”*** es una variable numérica que se pretende visualizar en el diagrama de caja. En esta variable se representan los niveles de radiación ultravioleta.
  - La variable “calidad_rx” corresponde a la variable categórica, la cual actúa como factor para dividir los datos en diferentes grupos en el diagrama de caja. En esta variable se obtienen las etiquetas: excelente, media y mala.
  - La variable “qe” corresponde al conjunto de datos que contiene las variables “uv” y “calidad_rx”, utilizadas para la generación del diagrama de caja.

En el diagrama de caja resultante se tendrá un conjunto de cajas (uno por cada nivel de la variable “calidad_rx”) donde cada caja presentará la distribución de los valores de la variable **“**uv**”** para el respectivo nivel de la calidad de recepción de la señal.

El uso de los diagramas de caja para este estudio permitirá obtener una perspectiva más amplia de la distribución de los atributos adicionales con lo cual se podrá comparar y analizar la variabilidad entre uno y otro atributo en función de la calidad de recepción: excelente, media o mala.

El análisis de la distribución de los datos se realizó en base a las mediciones de campo realizadas en los diez puntos de estudio. Este análisis será de suma importancia para tener un conocimiento a priori que permita identificar qué atributos repercuten en la calidad de la señal, de tal manera, que tales atributos sean tomados en consideración para la predicción de la calidad de la señal en los modelos de aprendizaje automático.

Diagramas de caja asociadas al punto B:

Con el objetivo de visualizar los resultados preliminares obtenidos con respecto a la influencia de los atributos considerados en la calidad de la conexión, se usa la función “**boxplot ()”** [18]**,** [19]**.**

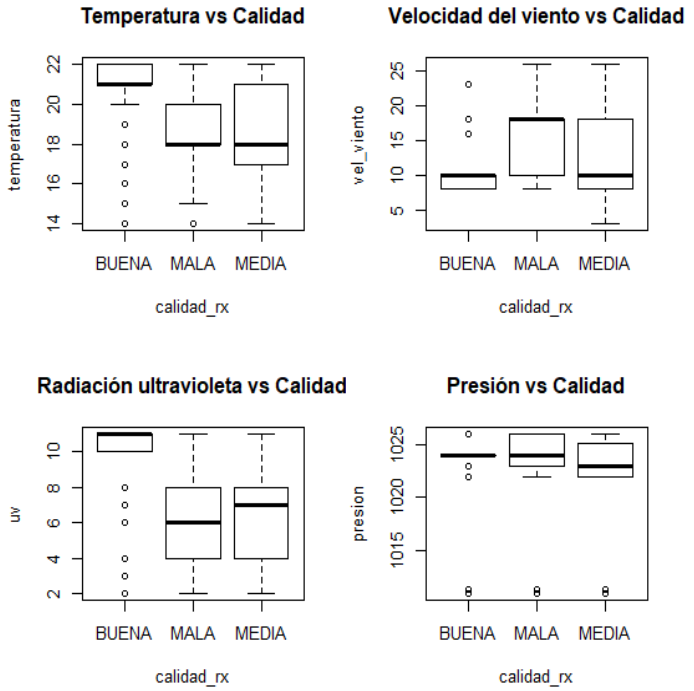

Figura 7. Diagramas de caja, asociados a los atributos del punto B.

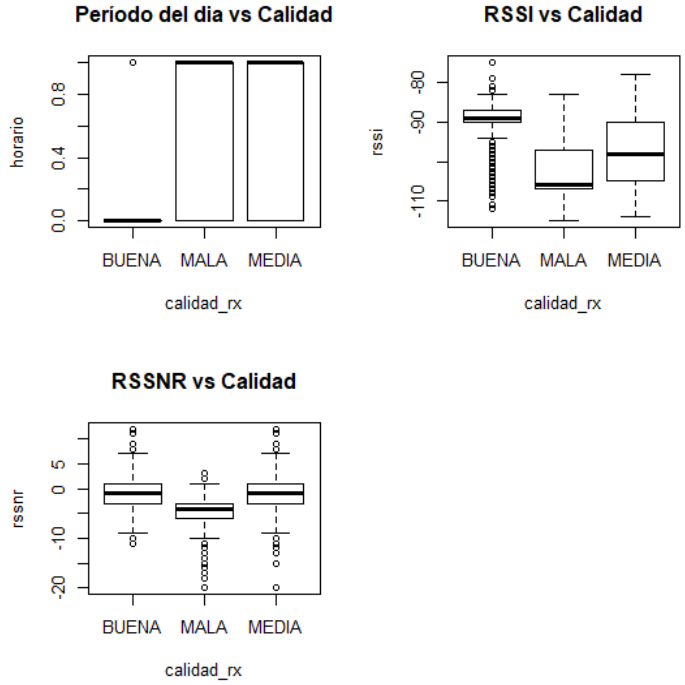

Figura 8. Diagramas de caja, asociados a los atributos del punto B.

Nota: para la variable “horario” (período del día) se ha decidido etiquetar al período “Mañana” como 0 y al período “Tarde” como 1 dado que solo se admiten variables numéricas para el manejo de los diagramas de caja.

Los diagramas de caja, presentes en las figuras 7 y 8, representan una forma gráfica para la visualización de la dispersión de los datos, sin embargo, es necesario incluir

otros análisis que permitan conocer de manera más precisa, el grado de dispersión entre una clase y otra. Para ello, es necesario el uso del análisis ANOVA y Tukey.

2.1.2 PRUEBA ESTADÍSTICA ANOVA

El análisis ANOVA resulta imprescindible para determinar si existen diferencias significativas entre las variables de estudio (variables numéricas), en función de una variable categórica.

Para presentar un análisis estadístico ANOVA en R [18], se debe tomar como referencia el siguiente comando, presentado en el código 2:

Código 2. Función ANOVA en R para el análisis estadístico.

```
>qe_anova<-summary (aov (temperatura~calidad_rx, data=qe))
>qe_anova
```

A continuación, se describirán los parámetros asociados a la función “aov ()”:

- La función “aov ()” permite identificar si existen diferencias significativas en la variable “temperatura”, en relación con la variable “calidad_rx”.

La variable “qe_anova” genera la siguiente salida (figura 9):

```
> qe_anova<-summary(aov(temperatura~calidad_rx,data=qe))
> qe_anova
              Df Sum Sq Mean Sq F value Pr(>F)
calidad_rx     2  15197    7599    1546 <2e-16 ***
Residuals   25787 126774       5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 9. resultado generado a partir de la función ANOVA.

A continuación, se incluye un breve resumen de los parámetros más importantes obtenidos en la salida:

- **Df:** hace referencia a los grados de libertad, lo que implica que existen 3 niveles en la variable “calidad_rx” (dado que 3-1 es igual a 2) [18], [19].
- **“Pr(>F)”:** el valor p hace referencia a la probabilidad de obtener un valor F igual o más extremo que el observado, al asumir que la hipótesis nula es cierta (es decir, no existen diferencias significativas). Generalmente un valor Pr(>F) <0.05 indica que existen diferencias significativas entre las medias [19].

Una vez explicada la salida generada por la función **“aov ()”,** en este caso en particular se puede mencionar que se tiene una probabilidad (Pr(>F)) < 2e-16, lo que indica que la probabilidad es extremadamente baja (tiende a cero). Este resultado, en conjunto

con los símbolos **“***”**, indican que existe una diferencia altamente significativa entre las medias de los grupos. Por lo tanto, una vez realizado el análisis ANOVA, en este ejemplo se pone de manifiesto que la variable temperatura probablemente tiene un efecto significativo en la variable calidad de recepción (“calidad_rx”), dado que existe una diferencia significativa en la media de esta variable.

2.1.3 PRUEBA DE TUKEY

La prueba de Tukey permite realizar una comparación para el análisis de las diferencias significativas entre las medias de las variables de estudio [19]. Para la implementación de esta prueba, en R, se ha ingresado el siguiente comando (código 3):

Código 3. Función TukeyHSD en R para el análisis estadístico.

```
>qe_tukey<-TukeyHSD (aov (temperatura, data=qe))
>qe_tukey
```

En este estudio la función “TukeyHSD ()” es utilizada para realizar comparaciones múltiples entre los niveles de la variable “calidad_rx” con respecto a otras variables (la variable temperatura en este caso en particular) [19].

Los resultados de las comparaciones presentan una tabla donde se encuentra cada nivel de la variable “calidad_rx” con los siguientes parámetros:

- **“diff”:** diferencia entre las medias de los niveles de la variable “calidad_rx” [19]**.**
  - Si este parámetro tiene un valor positivo, significa que la media del segundo grupo es mayor que la media del primer grupo, es decir, el segundo grupo tiene valores más altos en promedio que el primer grupo [19].
  - Si este parámetro es negativo, significa que la media del segundo grupo es menor que la media del primer grupo, es decir, el primer grupo tiene valores más bajos en promedio que el segundo grupo [19].
- **“lwr” y “upr”:** indican el límite inferior y superior, respectivamente del intervalo de confianza del 95% para la diferencia de medias [19]. Un intervalo de confianza es un rango en el cual se cree que estará el valor real de una medida estadística. Es una forma de expresar la incertidumbre en nuestras estimaciones [8].
- **"p adj”:** es el valor p ajustado y corresponde a la probabilidad de obtener una diferencia de medias tan grande como la observada, al asumir que no hay

diferencia real (hipótesis nula). Un valor “p adj” menor a 0.05 indica que la diferencia de medias es estadísticamente significativa [19].

```
> qe_tukey<-TukeyHSD(aov(temperatura~calidad_rx,data=qe))
> qe_tukey
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = temperatura ~ calidad_rx, data = qe)

$calidad_rx
                    diff        lwr        upr p adj
MALA-BUENA   -1.8893404 -1.9894912 -1.7891895     0
MEDIA-BUENA  -2.1295566 -2.2203152 -2.0387980     0
MEDIA-MALA   -0.2402162 -0.3149973 -0.1654351     0
```

Figura 10. Resultado generado por la función TukeyHSD.

El resultado proporcionado en la figura 10 presenta la siguiente información [19]:

- **MALA-BUENA:** tiene un valor **“diff”** de -1.889, lo que indica que la media del grupo “BUENA” es mayor que la media del grupo “MALA”, por aproximadamente 1.89 unidades [19].
- **MEDIA-BUENA:** tiene un valor **“diff”** de -2.12, lo que indica que la media del grupo “BUENA” es mayor que la media del grupo “MEDIA” por aproximadamente 2.12 unidades [19].
- **MEDIA-MALA:** tiene un valor **“diff”** de -0.24, lo que indica que la media del grupo “MEDIA” es mayor que la media del grupo “MALA” por aproximadamente 0.24 unidades [19].

Con base en los resultados anteriores de este ejemplo, se puede verificar que existen diferencias significativas en las medias de la variable numérica considerada para los diferentes niveles de la calidad de recepción. Por lo tanto, estas diferencias nos permiten concluir que es probable que en el punto observado la variable temperatura tenga un efecto sobre la variable “calidad_rx”.

Nota: el ejemplo presentado anteriormente, corresponde a la variable temperatura, sin embargo, el análisis se ha realizado para cada una de las variables de estudio consideradas con el objetivo de verificar qué variables probablemente tienen un mayor grado de dispersión y variabilidad en función de la variable asociada a la calidad de recepción de la señal. Una vez realizado el análisis estadístico pertinente, se procedió a implementar los algoritmos de aprendizaje automático correspondientes.

2.2 MODELOS DE APRENDIZAJE AUTOMÁTICO

Los modelos de aprendizaje automático considerados en este estudio son: árboles de decisión y k-vecinos más cercanos, dado que usan algoritmos de clasificación.

2.2.1 ÁRBOLES DE DECISIÓN

Los comandos utilizados en R para la generación del árbol de decisión son los siguientes:

Nota: para el modelamiento se utilizan datos de entrenamiento y de prueba, en este caso se tienen 19342 y 6448 datos respectivamente.

Código 4. Creación de la receta en R para el modelo de aprendizaje automático.

```
>qe_rec3<-
recipe(calidad_rx~temperatura+vel_viento+uv+rssi+horario,data=qe_train)%>%
+themis::step_downsample(calidad_rx)%>%
+recipes::prep()
>qe_rec3
```

Las funciones asociadas al código 4 son las siguientes.

- La función “recipe ()” crea una receta con la fórmula especificada. Esta receta tiene a la variable “calidad_rx” como variable objetivo y utiliza las variables temperatura, velocidad del viento, radiación ultravioleta, RSSI y período del día como variables predictoras [20].
- La instrucción “themis::step_downsample (calidad_rx)” aplica el método de submuestreo (downsampling) a la variable objetivo mediante el uso de la función “step_downsample ()”. El submuestreo es una técnica para abordar conjuntos de datos desequilibrados, donde hay una desproporción significativa en la cantidad de observaciones entre las clases de la variable objetivo [20].
- La instrucción “recipes::prep ()” tiene como objetivo preparar la receta para su posterior uso, ajustando el preprocesamiento según los datos proporcionados, esto incluye la ejecución de pasos de preprocesamiento, como centrar y escalar variables, convertir variables categóricas en variables ficticias, entre otras cosas [20].

A continuación, se hace uso del código 5 para visualizar el número de observaciones en cada nivel de la variable asociada a la calidad de la recepción de la señal.

Código 5. Funciones juice () y count () en R.

```
>juice (qe_rec3)%>%
+count (calidad_rx)
```

- **“juice ()”:** esta función se utiliza para la extracción de los resultados de la receta, previamente definida, luego se aplica la función “count ()” para contabilizar la cantidad de observaciones en cada nivel de la variable “calidad_rx”. En este caso se presentan 3237 observaciones para cada nivel asociado a la calidad de recepción de la señal, con esto se tiene un conjunto de datos equilibrado. Un equilibrio en los datos permite un mejor entrenamiento del modelo, lo que implica una mejor predicción de los datos [21].

Código 6. Invocación al modelo de aprendizaje (árbol de decisión) en R.

```
>qe_tree_spec3<-decision_tree () %>%
+set_engine(“rpart”) %>%
+set_mode("classification”)
```

- **“decision_tree ()”:** esta función es utilizada para especificar la obtención de un árbol de decisión [15].
- **“set_engine (“rpart”)”:** esta instrucción se la utiliza para la configuración del motor del modelo, en este caso, se ha elegido el algoritmo **“rpart”,** el cual es una implementación del árbol de decisión [15].
- **“set_mode ("classification”)”:** esta función se usa con el objetivo de establecer el modo del modelo a obtener; en este caso se ha elegido el modo de clasificación dado que se trata de un modelo de aprendizaje de este tipo [15].

Código 7. Ajuste del modelo para los datos de entrenamiento.

```
>qe_tree_fit3<-qe_tree_spec%>%
+fit(calidad_rx~.,data=juice(qe_rec3))
```

- **“fit ()”:** esta función, presente en el código 7, se utiliza para ajustar el modelo a los datos de entrenamiento. En este caso, se ajusta el modelo de árbol de decisión a la variable objetivo en base a todas las variables predictoras. Una vez que se ejecuta este comando, el modelo de árbol de decisión se ajusta mediante el uso de los datos de entrenamiento que posteriormente se podrá utilizar para la predicción de un conjunto de datos de prueba [15].

Código 8. Comando que ajusta el modelo de árbol de decisión.

```
> qe_fit3<-rpart (calidad_rx~
temperatura+vel_viento+uv+rssi+horario,data=qe_train,method="class",minsplit=1,minbucket=1)
```

- El comando (código 8) permite el ajuste del modelo del árbol de decisión [15]. El objetivo es predecir la calidad de recepción de la señal (“calidad_rx”).

Código 9. Comando que realiza las predicciones con los datos de prueba.

```
qe_prediccion3<-predict(qe_fit3,qe_test,type=”class”)
```

- El comando (código 9) hace uso de la función “predict ()” para realizar predicciones con modelos entrenados, en este caso, el modelo considerado es el “qe_fit3”. Este modelo se ha ajustado previamente mediante el uso del algoritmo correspondiente al árbol de decisión [22].

Código 10. Comando de R que entrega la matriz de confusión generada por el modelo.

```
> qe_fit3<-table(qe_test$calidad_rx,qe_prediccion3)
```

- El comando (código 10) se utiliza para la creación de una matriz de confusión entre las etiquetas reales y las predicciones realizadas por el modelo en un conjunto de datos de prueba [23].

```
> qe_matriz_summary3
Confusion Matrix and Statistics

          qe_prediccion3
           BUENA MALA MEDIA
  BUENA      697   26   301
  MALA         2 1083   741
  MEDIA      326  526  2746

Overall Statistics
```

Figura 11. Matriz de confusión resultante para el modelo de aprendizaje.

A continuación, se interpretará la matriz de confusión obtenida en la figura 11 :

- La primera fila de la matriz hace referencia a las instancias reales asociadas con la clase “BUENA”, en este caso existen 1024 instancias. El modelo predijo correctamente 697, pero clasificó erróneamente 26 y 301 como “MALA” y “MEDIA”, respectivamente [23].

- La segunda fila de la matriz hace referencia a las instancias reales asociadas con la clase “MALA”, en este caso existen 1826 instancias. El modelo predijo correctamente 1083, pero clasificó erróneamente 2 y 741 como “BUENA” y “MEDIA”, respectivamente [23].

- La tercera fila de la matriz hace referencia a las instancias reales de la clase “MEDIA”, en este caso existen 3598 instancias. El modelo predijo correctamente 2746, pero clasificó erróneamente 326 y 526 como “BUENA” y “MALA”, respectivamente [23].

Código 11. Comando que entrega las estadísticas generales de evaluación del modelo de aprendizaje.

```
>qe_matriz_summary3<-confusionMatrix(qe_matriz3)
```

- Este comando (código 11) presenta un resumen estadístico de evaluación del modelo de aprendizaje considerado (árbol de decisión).

```
Confusion Matrix and Statistics

          qe_prediccion3
           BUENA MALA MEDIA
  BUENA      697   26   301
  MALA         2 1083   741
  MEDIA      326  526  2746

Overall Statistics

                Accuracy : 0.7019
                  95% CI : (0.6906, 0.7131)
    No Information Rate : 0.5875
    P-Value [Acc > NIR] : < 2.2e-16
```

Figura 12. Resultado de la evaluación del modelo de aprendizaje.

- El resultado obtenido en la figura 12 indica que se tiene una precisión del 70.19%, es decir, 7 de cada 10 muestras fueron etiquetadas correctamente.

Árbol de decisión obtenido

Código 12. Comando que grafica el árbol de decisión

```
> fancyRpartPlot(qe_fit3,caption=NULL)
```

- El código 12 permite visualizar de forma gráfica (figura 13) el diagrama del árbol de decisión obtenido [24].

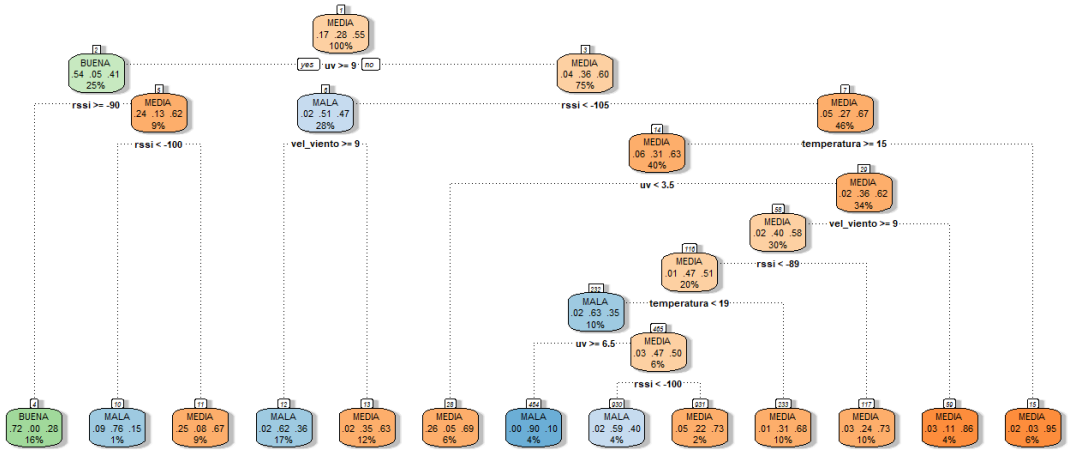

Figura 13. Diagrama del árbol de decisión generado.

2.2.2 K-VECINOS MÁS CERCANOS

A continuación, de forma breve, se describe la generación y la evaluación del modelo de los K-vecinos más cercanos en R que se utilizó en este proyecto [16].

Código 13. Especificaciones del modelo k-NN en R

```
> qe_kknn_spec<-nearest_neighbor()%>%
+ set_engine("kknn")%>%
+ set_mode("classification")
```

- En el código 13 se define el algoritmo que se va a utilizar (kknn) en conjunto con la técnica de clasificación [16].

Código 14. Ajuste del modelo de aprendizaje en R

```
>qe_kknn_fit<-qe_kknn_spec%>%
+ fit(calidad_rx~.,data=juice(qe_rec))
```

- En el código 14 se ajusta el modelo k-NN mediante el uso de los datos proporcionados en la receta (uso de variables predictoras). En este código se especifica que se realiza una clasificación con la variable objetivo "calidad_rx".

Código 15. Particiones de la validación cruzada en R.

```
>qe_validation_splits<-mc_cv(juice(qe_rec),prop=0.9,strata=calidad_rx)
```

- En el código 15 se indica que se va a utilizar el 90% de los datos para entrenamiento y el 10% para validación en cada partición, en función de la variable “calidad_rx”.

Código 16. Ajuste del modelo de k-NN mediante el uso de validación cruzada en R.

```
>qe_kknn_res<-
tune::fit_resamples(qe_kknn_spec,calidad_rx~.,qe_validation_splits,control=control_resamples(save_pred=TRUE))
```

- En el código 16 se realiza el ajuste y la evaluación del modelo en las particiones de validación cruzada, además se especifican las variables predictoras y la variable objetivo [25].

Código 17. Evaluación del rendimiento del modelo k-NN en R.

```
>qe_kknn_res%>%
+collect_metrics()
```

- Con el código 17 se recopilan las métricas de evaluación del rendimiento del modelo donde se extraen las métricas calculadas del rendimiento, durante la validación cruzada [26].

Nota: el hiperparámetro k en este y todos los puntos donde se recolectó la información es igual a 5 (k=5).

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 RESULTADOS

A continuación, se expondrán los resultados obtenidos en el presente estudio, para ello se expondrán los resultados estadísticos gráficos y teóricos. A continuación, se incluirán las técnicas de aprendizaje automático propuestas.

3.1.1 DIAGRAMAS DE CAJA OBTENIDOS

Diagramas de caja de la calidad en función de la temperatura en los diez puntos de recolección

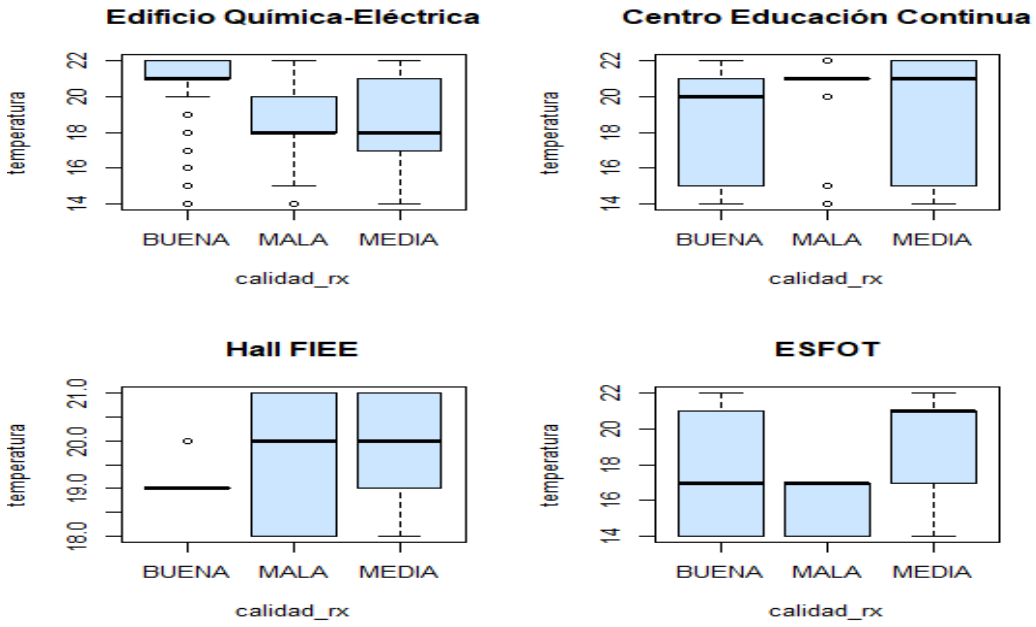

Figura 14. Diagramas de caja de la calidad en función de la temperatura.

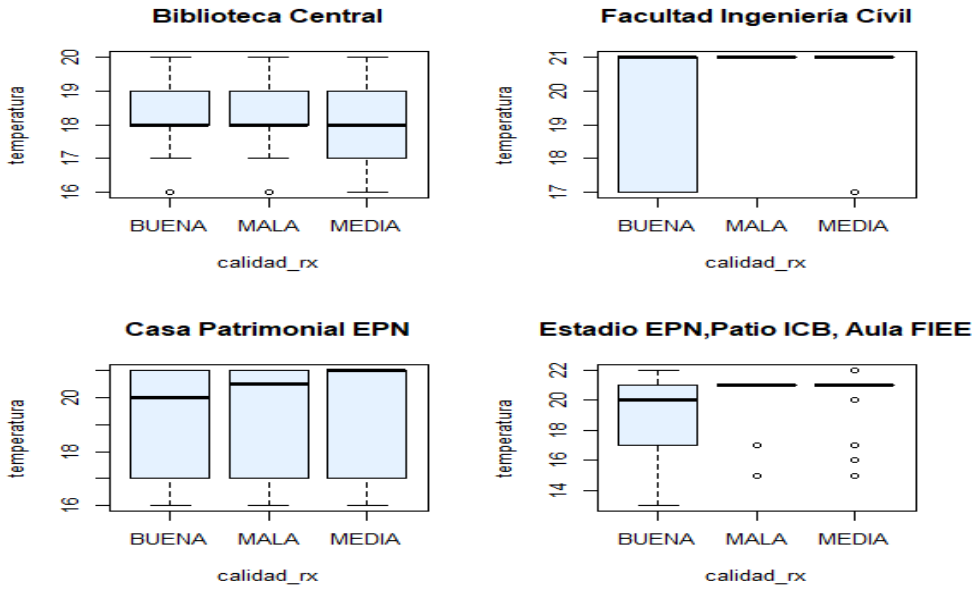

Figura 15. Diagramas de caja de la calidad en función de la temperatura.

Como se puede visualizar en las figuras 14 y 15, se presentan los diagramas de caja en cada uno de los puntos donde se recolectaron los datos, enfocados en la variable temperatura. El objetivo de incluir los diagramas de caja radica en el hecho de visualizar como varía la calidad (buena, medida y mala) en términos de la temperatura. Como se puede visualizar en la figura 14, existe una alta variación dentro de cada grupo asociado a la calidad, por otro lado, también existe una alta variación entre grupos. Asimismo, en la figura 15 se puede visualizar una alta variación dentro de cada grupo, sin embargo, en el punto I (casa patrimonial EPN) no existe una variación entre grupos, de tal manera que

esta variable no debería ser insertada dentro del modelo de aprendizaje, en este punto en particular. El propósito de esta introducción, con respecto al uso de los diagramas de caja, es para visualizar que variables serán consideradas como variables predictoras al instante de incluirlas en los modelos de aprendizaje automático y determinar si existe alguna probabilidad de que estos atributos repercutan en la calidad de recepción de la señal.

Diagramas de caja de la radiación ultravioleta en los diez puntos de recolección

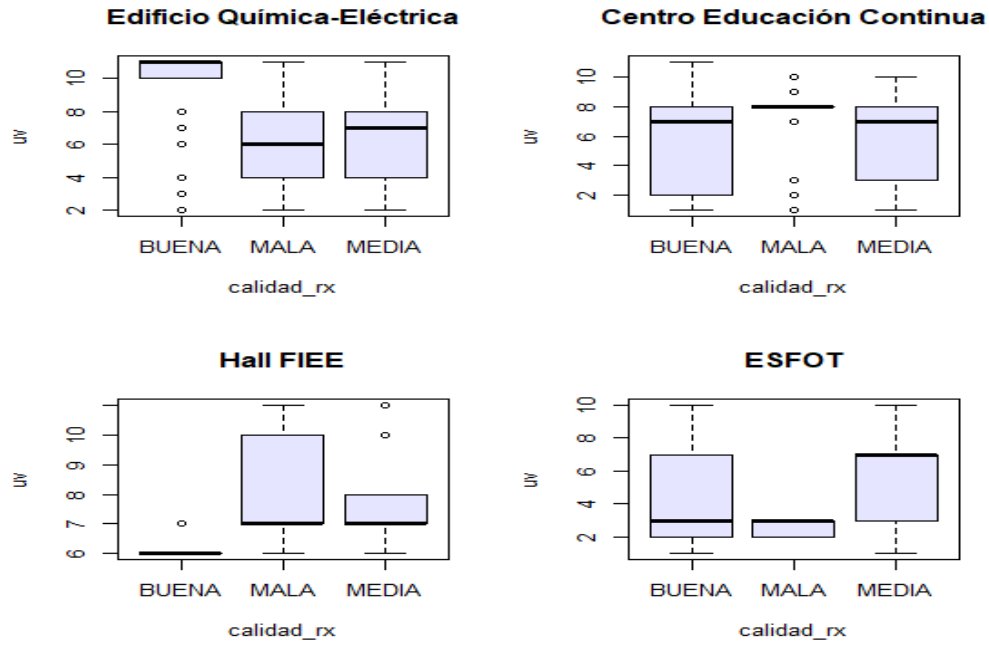

Figura 16. Diagramas de caja (radiación ultravioleta).

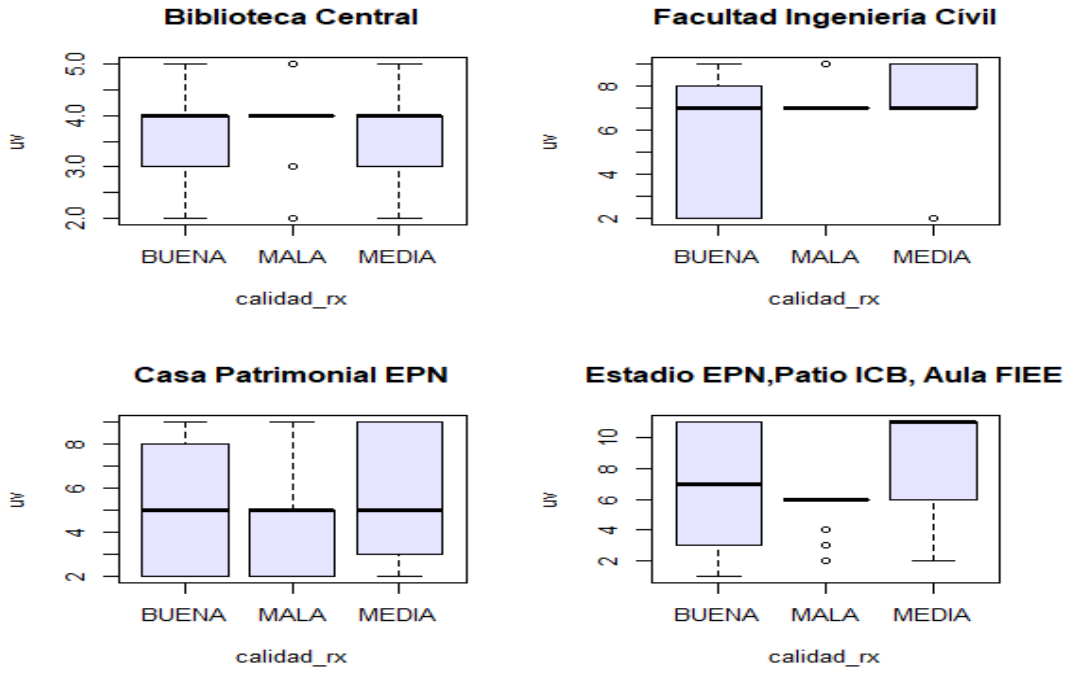

Figura 17. Diagramas de caja (radiación ultravioleta).

Las figuras 16 y 17 indican una alta variación dentro de cada grupo asociado a la calidad, sin embargo, no en todos los puntos se tiene una alta variación entre grupos, por ejemplo, en el punto C (ESFOT).

Diagramas de caja de la velocidad del viento en los diez puntos de recolección

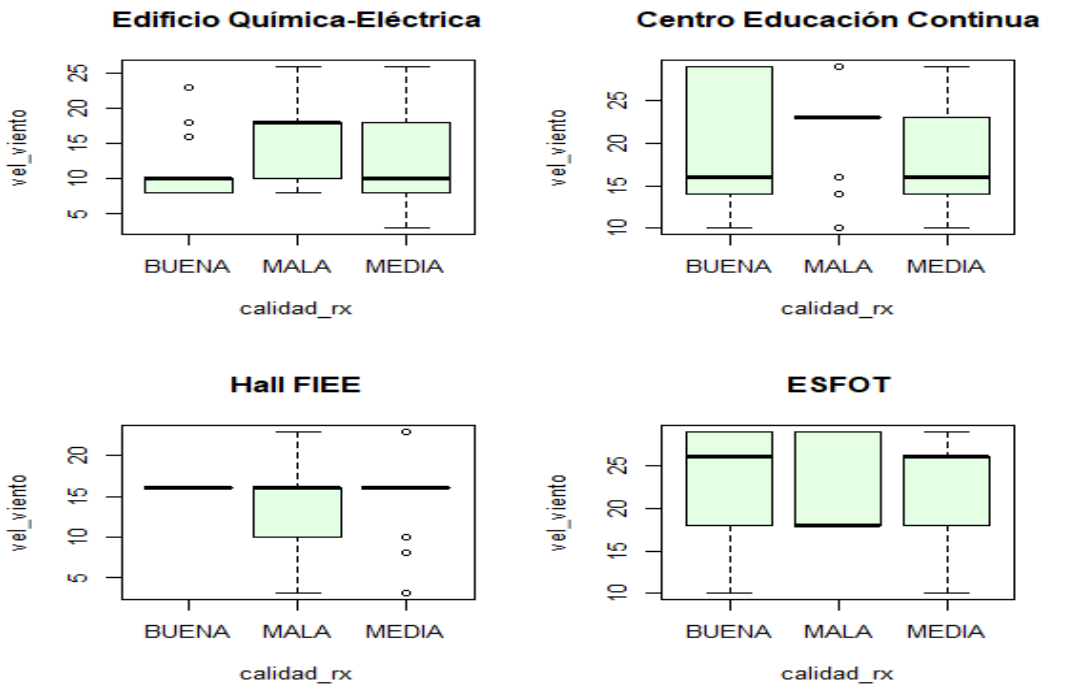

Figura 18. Diagramas de caja (velocidad del viento).

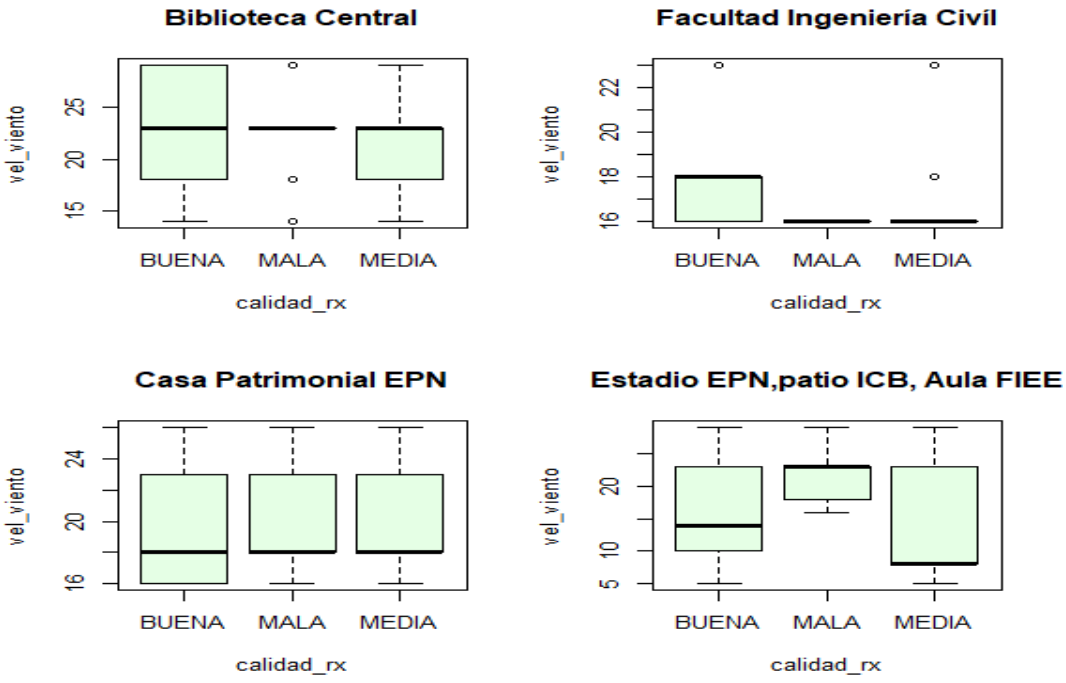

Figura 19. Diagramas de caja (velocidad del viento).

Las figuras 18 y 19 indican una alta variación dentro de cada grupo asociado a la calidad, sin embargo, no en todos los puntos se puede apreciar esta variación, por ejemplo, los puntos D y J (Hall FIEE y Facultad de Ingeniería Civil). Por otra parte, no en todos los puntos se tiene una alta variación entre grupos, por ejemplo, en los puntos () (ESFOT, Hall FIEE y casa patrimonial EPN).

Diagramas de caja de la presión atmosférica en los diez puntos de recolección

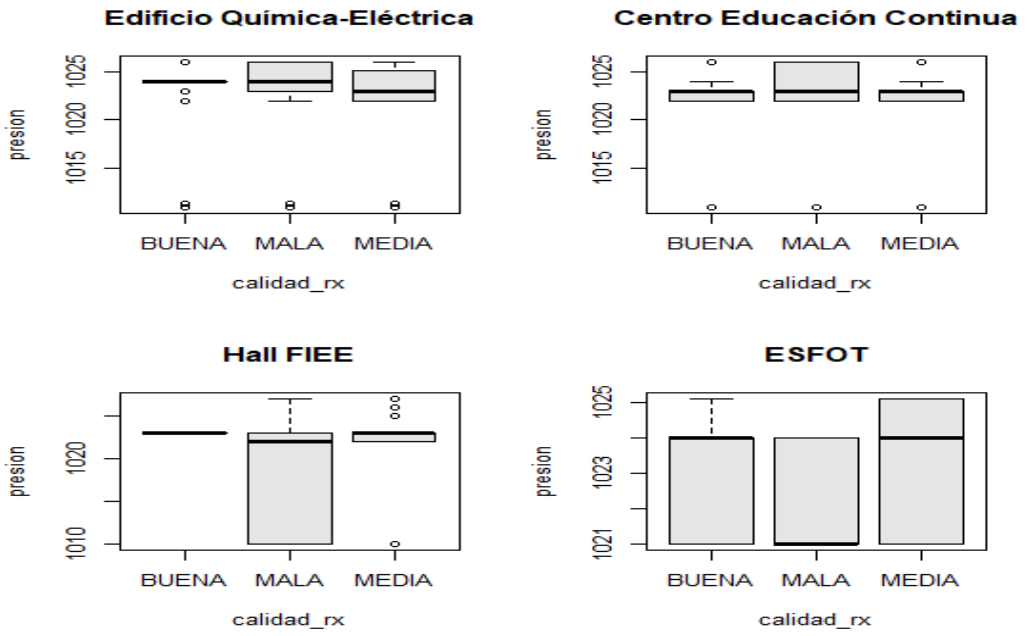

Figura 20. Diagramas de caja (presión atmosférica).

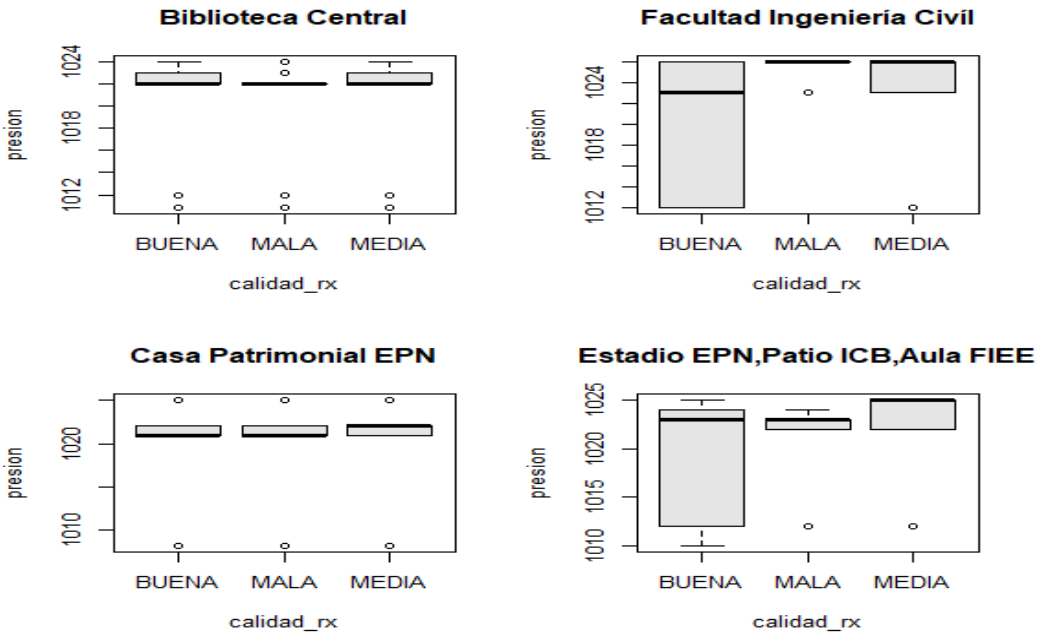

Figura 21. Diagramas de caja (presión atmosférica).

Las figuras 20 y 21 indican ambas variaciones, altas y bajas, dentro de los grupos, una variación alta se encuentra en el punto C (ESFOT), mientras que una variación baja se encuentra en el punto I (casa patrimonial EPN). Por otro lado, no existen variaciones entre grupos, claramente notorias; sin embargo, se puede notar que en la mayoría de los casos existen variaciones que podrían influir en la calidad de la señal.

Diagramas de caja del período del día en los diez puntos de recolección

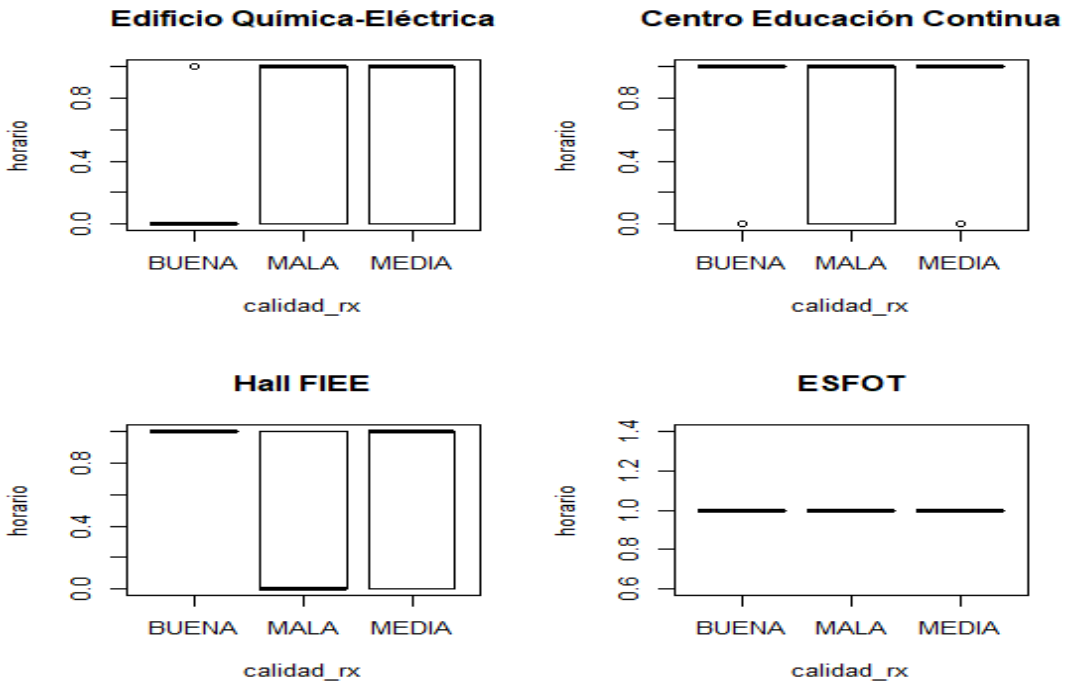

Figura 22. Diagramas de caja (período del día).

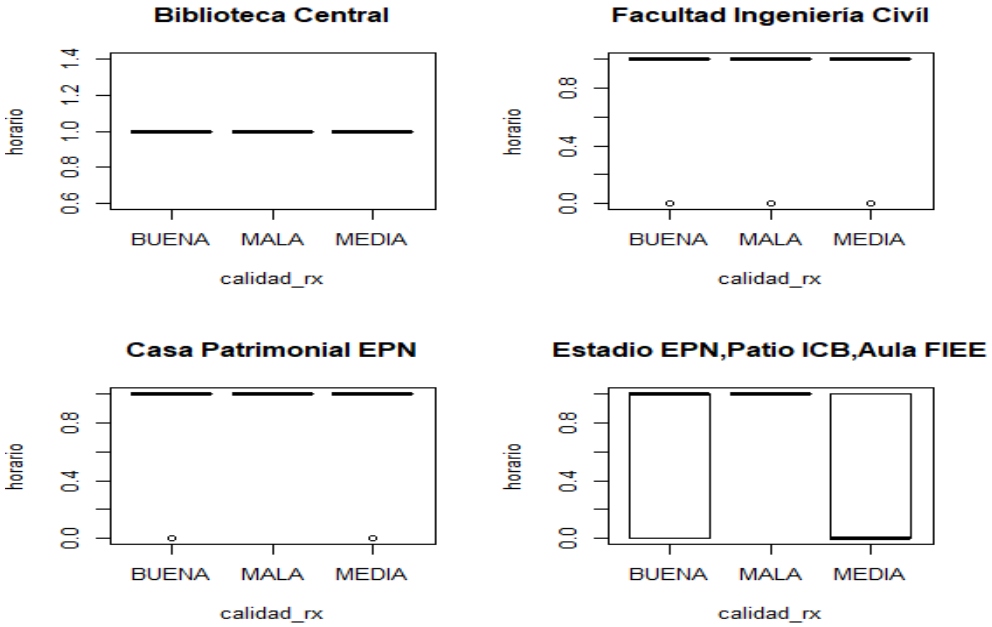

Figura 23. Diagramas de caja (período del día).

La variable “horario” corresponde al período del día mañana y tarde con las etiquetas 0 y 1, respectivamente. En las figuras 22 y 23 existen variaciones; sin embargo, estas no son ampliamente pronunciadas. A pesar de esto, se puede decir que en la mayoría de los puntos existen un impacto en la calidad de la señal, dado que ciertos puntos se presentan variaciones.

Diagramas de caja del RSSI en los diez puntos de recolección

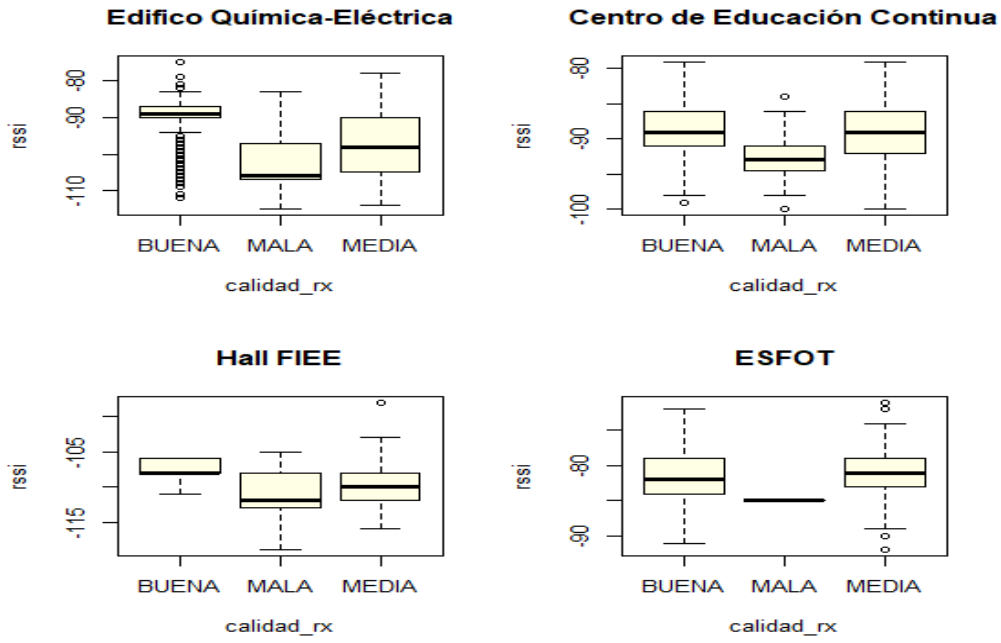

Figura 24. Diagramas de caja (RSSI).

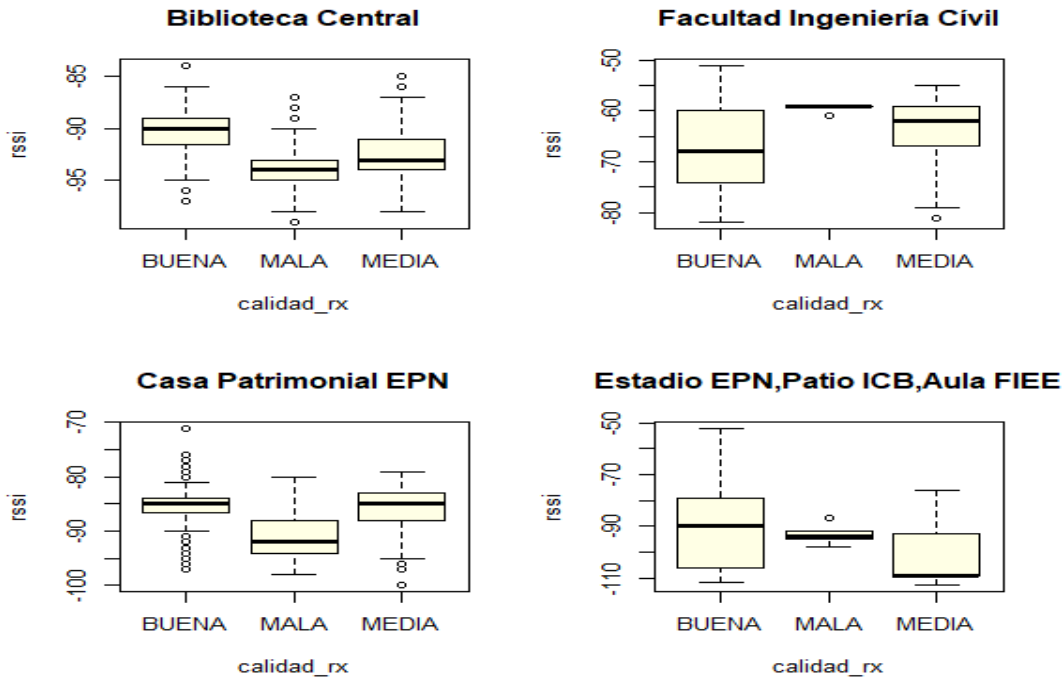

Figura 25. Diagramas de caja (RSSI).

Los diagramas presentes en las figuras 24 y 25 presentan variaciones de ambos tipos (dentro del grupo y entre grupos), se puede afirmar que claramente el RSSI es un parámetro que repercute en la calidad de la señal.

Diagramas de caja del RSSNR en los diez puntos de recolección

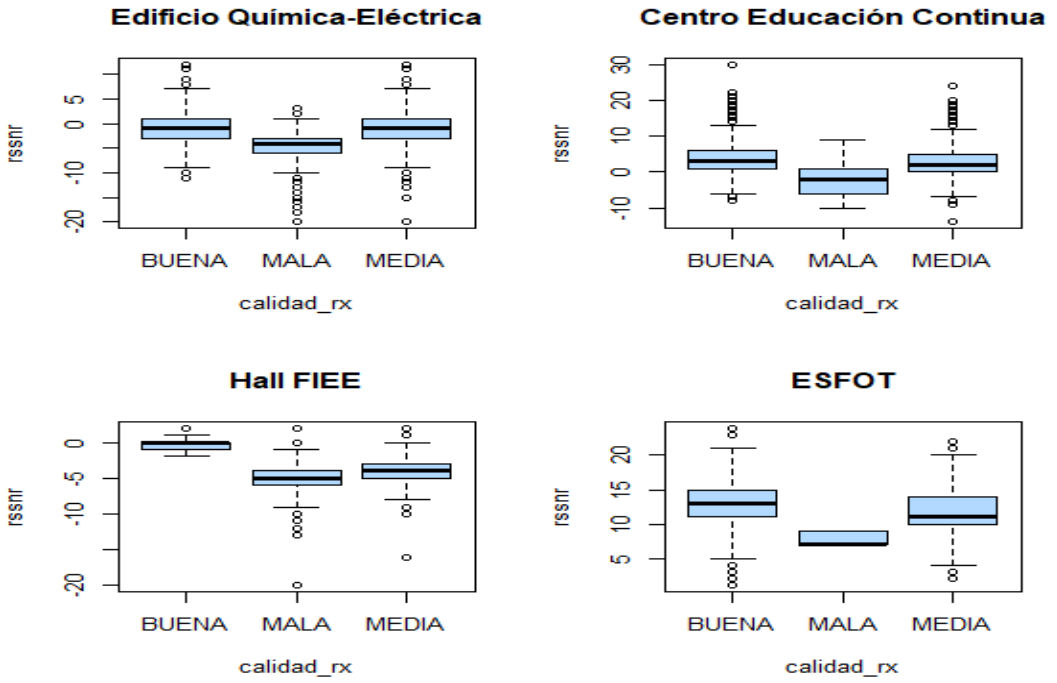

Figura 26. Diagramas de caja (RSSNR)

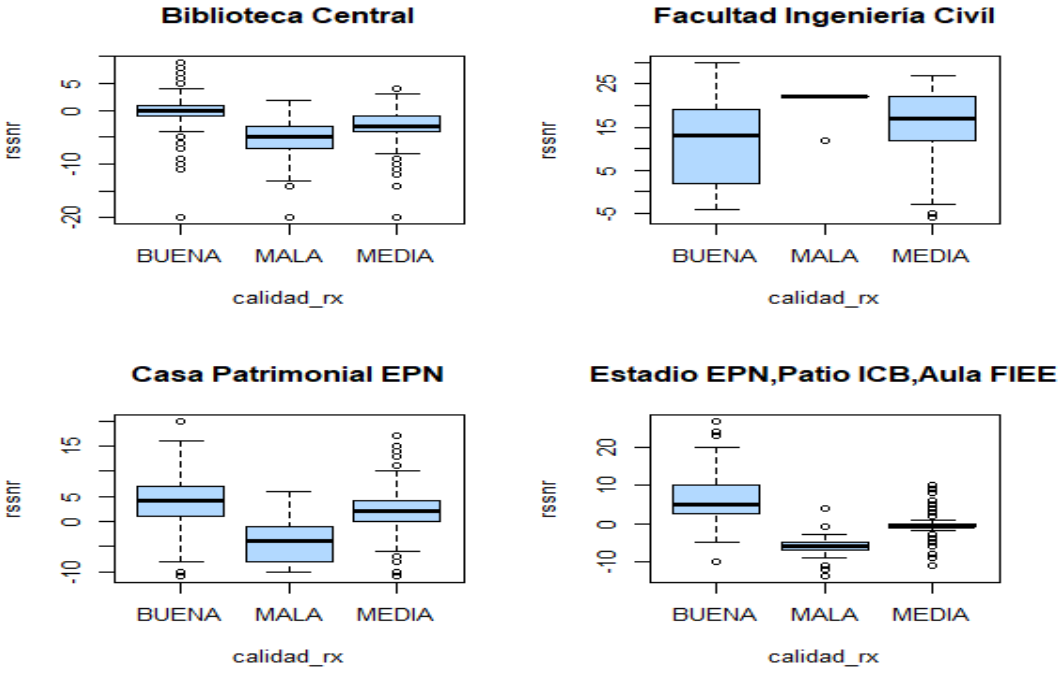

Figura 27. Diagramas de caja (RSSNR)

En base a lo que se observa de las gráficas, al igual de lo que sucede con el RSSI, el RSSNR es otro factor que repercute en la calidad de la señal recibida. Las variaciones presentes en las figuras 26 y 27 corresponden a diferencias de ambos tipos (dentro del grupo y entre grupos).

3.1.2 RESULTADOS DE LOS ANÁLISIS ANOVA, TUKEY Y MODELOS DE APRENDIZAJE AUTOMÁTICO

A continuación, se incluirán los resultados de los análisis ANOVA y Tukey, asociados a cada punto en el que se recolectaron los datos.

Como se ha podido visualizar en el diagrama de caja de cada una de las variables, resulta claro que existe una influencia de estas variables sobre la calidad de recepción de la señal, sin embargo, para saber cómo influyen realmente es necesario continuar con un análisis más preciso que me permita diferenciar una clase de otra. A continuación, se exponen los resultados obtenidos en los análisis estadísticos: ANOVA y Tukey, posterior a estos análisis se evaluarán los modelos de aprendizaje automático: árboles de decisión y k-vecinos más cercano.

Nota: cuando el valor “p adj” sea mayor a 0.05 en dos o más comparaciones (Mala-Buena, Media-Buena, Media Mala); entonces la variable climática o de RF, no se incluirá en la evaluación del modelo de aprendizaje automático [19]. Por otra parte, dentro del algoritmo de los árboles de decisión existe un mecanismo denominado “pruning” el cual se encarga de descartar variables que no permitan tomar decisiones adecuadas en los nodos del árbol de decisión. Por lo tanto, es posible que, en algunos puntos, no se encuentren todas las variables que se describen en los análisis ANOVA y Tukey [15].

Punto A: Edificio del Centro de Educación Continua (CEC).

Análisis ANOVA y Tukey

Tabla 2. Resultados obtenidos en el punto A.

Parámetro	**Resultado**	**diff**	**P adj**
Temperatura	Mala-Buena	1.69	0
	Media-Buena	0.82	0
	Media-Mala	-0.86	1.39e-5
Radiación ultravioleta	Mala-Buena	1.57	0
	Media-Buena	0.70	0
	Media-Mala	-0.87	7.3e-6

Velocidad del viento	Mala-Buena	1.96	2.8e-5
	Media-Buena	-0.73	7.32e-5
	Media-Mala	-2.69	0
Presión atmosférica	Mala-Buena	2.39	0
	Media-Buena	0.77	0
	Media-Mala	-1.61	0
Periodo del día	Mala-Buena	-0.32	0
	Media-Buena	0	0.99
	Media-Mala	0.32	0
RSSI	Mala-Buena	-4.03	0
	Media-Buena	-0.48	6e-7
	Media-Mala	3.55	0
RSSNR	Mala-Buena	-5.56	0
	Media-Buena	-1.50	0
	Media-Mala	4.07	0

De acuerdo con los resultados generados por el análisis Tukey y ANOVA, presentes en la tabla 2, un p valor menor a 0.05 en cada una de las variables, indica que las diferencias entre las medias son estadísticamente significativa en comparación con una familia de comparaciones (Mala-Buena, Media-Buena, Media-Mala). Por lo tanto, en base a este análisis se propone que todas estas variables serán consideradas para el modelo de aprendizaje automático, y servirán como variables predictoras para la variable categórica correspondiente a la calidad de recepción (“calidad_rx”).

Para la obtención de los modelos en este estudio se hará uso de dos tipos de aprendizaje automático (árboles de decisión y k-vecinos más cercanos). El objetivo de hacer uso de ambos tipos de aprendizaje será para evaluar la precisión de cada modelo, de tal forma que permita identificar qué modelo resulta más preciso al momento de hacer predicciones en este tipo de estudios.

Técnica de aprendizaje automático: árboles de decisión.

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          cec_prediccion
           BUENA MALA MEDIA
  BUENA       48    0   463
  MALA         0    0    72
  MEDIA       42    0  1588

Overall Statistics

                  Accuracy : 0.7393
                    95% CI : (0.7204, 0.7575)
    No Information Rate : 0.9593
    P-Value [Acc > NIR] : 1
```

Figura 28. Resultados obtenidos de la matriz de confusión en el punto A.

Como se puede visualizar en la figura 28, se tiene una precisión del 73.93%, lo que indica que 7 de cada 10 datos de prueba fueron etiquetados correctamente. Asimismo, se puede visualizar la matriz de confusión obtenida, la cual indica lo siguiente:

- De un total de 511 datos, 48 fueron etiquetados correctamente en la clase "BUENA".
- De un total de 72 datos, 0 fueron etiquetados correctamente en la clase "MALA".
- De un total de 1630 datos, 1588 fueron etiquetados correctamente en la clase "MEDIA"

Los resultados presentados por la matriz de confusión indican que la clase predominante en este punto en particular es la clase **"MEDIA**" en un conjunto de 2213 datos.

- **Diagrama del árbol de decisión obtenido**

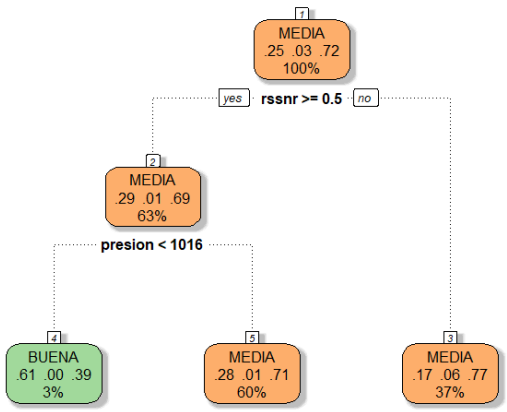

Figura 29. Diagrama de decisión obtenido en el punto A.

- El diagrama del árbol obtenido en la figura 29, presenta como nodo raíz a la variable “rssnr” (mejor variable predictora), a continuación, se desprende la variable “presión” y finalmente se tienen las etiquetas asociadas a la clase “BUENA” y “MEDIA”, con los respectivos porcentajes de cada una. Como se puede visualizar, la variable predominante en este punto en particular corresponde a la clase “MEDIA” con un 72% de los datos. Se tiene que cuando el RSSNR sea menor a 0.5dB es probable que la calidad sea “MEDIA”.

- De acuerdo con los resultados obtenidos en este punto, la variable ambiental que posiblemente incide en la calidad de la señal corresponde a la presión atmosférica. Cuando el parámetro RSSNR es mayor o igual a 0.5 y la presión atmosférica tiene un valor menor a 1016 mb, entonces los datos son etiquetados en la clase “BUENA”; por otra parte, cuando se tiene una presión atmosférica mayor a 1016 mb, entonces es probable que las conexiones sean de la clase “MEDIA”.

- Los porcentajes de cumplimiento de estas condiciones se presentan en el gráfico del árbol.

Técnica de aprendizaje automático: k-vecinos más cercanos

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.648    25 0.00921 Preprocessor1_Model1
2 roc_auc  hand_till  0.800    25 0.00670 Preprocessor1_Model1
```

Figura 30. Resultados obtenidos del modelo k-vecinos más cercanos en el punto A.

Los resultados obtenidos con el modelo de los k-vecinos más cercanos indican una precisión del 64.8%. Al igual que en el caso anterior, se tiene una precisión del modelo en donde 7 de cada 10 datos de prueba predicen correctamente la calidad de la conexión.

La precisión en este caso es ligeramente menor en comparación con el modelo del árbol de decisión.

Punto B: Edificio de Química Eléctrica.

Tabla 3. Resultados obtenidos del análisis Anova y Tukey en el punto B.

Parámetro	**Resultado**	**diff**	**P adj**
Temperatura	Mala-Buena	-1.88	0
	Media-Buena	-2.12	0

	Media-Mala	-0.24	0
Radiación ultravioleta	Mala-Buena	-3.34	0
	Media-Buena	-2.95	0
	Media-Mala	0.39	0
Velocidad del viento	Mala-Buena	4.17	0
	Media-Buena	3.14	0
	Media-Mala	-1.02	0
Presión atmosférica	Mala-Buena	-1.23	0
	Media-Buena	-2.099	0
	Media-Mala	-0.87	0
Periodo del día	Mala-Buena	0.4	0
	Media-Buena	0.4	0
	Media-Mala	-0.02	0
RSSI	Mala-Buena	-12.84	0
	Media-Buena	-8.05	0
	Media-Mala	4.79	0
RSSNR	Mala-Buena	-3.49	0
	Media-Buena	-0.26	1.8e-6
	Media-Mala	3.21	0

De acuerdo con los resultados generados por el análisis Tukey y ANOVA, presentes en la tabla 3, un p valor menor a 0.05 en cada una de las variables, indica que las diferencias entre las medias son estadísticamente significativas en comparación con una familia de comparaciones (Mala-Buena, Media-Buena, Media-Mala). Por lo tanto, en base a este análisis se propone que todas estas variables serán consideradas para el modelo de aprendizaje automático, y servirán como variables predictoras para la variable categórica correspondiente a la calidad de recepción (“calidad_rx”).

Para la obtención de los modelos en este estudio se hará uso de dos tipos de aprendizaje automático (árboles de decisión y k-vecinos más cercanos).

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          qe_prediccion
           BUENA MALA MEDIA
  BUENA      767   21   236
  MALA         0 1093   733
  MEDIA      259  324  3015

Overall Statistics

                Accuracy : 0.756
                  95% CI : (0.7454, 0.7665)
     No Information Rate : 0.6179
     P-Value [Acc > NIR] : < 2.2e-16
```

Figura 31. Resultados obtenidos de la matriz de confusión en el punto B.

Los resultados obtenidos, presentes en la figura 31 indican una precisión del 75.6%, por lo que aproximadamente 8 de cada 10 datos de prueba fueron etiquetados correctamente. Por otro lado, la matriz de confusión proporciona la siguiente información:

- De un total de 1024 datos, 767 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 1826 datos, 1093 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 3598 datos, 3015 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama del árbol de decisión obtenido**

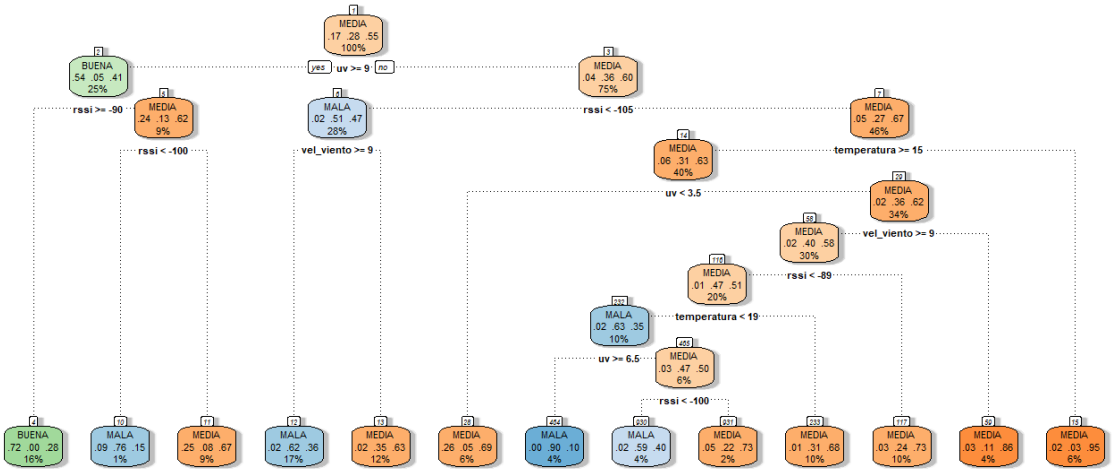

Figura 32. Árbol de decisión obtenido en el punto B.

- En este punto el nodo raíz (mejor variable predictora) corresponde a la variable radiación ultravioleta (“uv”).
- La calidad de la señal etiquetada como “BUENA” parte de la condición asociada con valores de “uv” mayores o iguales a 9 ($uv \geq 9$) y se tienen valores de RSSI mayores o iguales a -90dBm.
- Cuando no se cumple la condición anterior ($uv \geq 9$), entonces predomina la clase “MEDIA”. El predominio de la clase “MEDIA” también se presenta en ciertas condiciones climáticas en las cuales la temperatura supere los 15°C, la velocidad del viento sea menor a 9km/h o en casos en los que la radiación ultravioleta sea menor a 3.5.
- La clase etiquetada como “MALA” corresponde a valores de RSSI que se encuentran por debajo de -105dBm y con velocidades de viento que superan los 9km/h, así como con índices de radiación ultravioleta superiores a 6.5.

Técnica de aprendizaje automático: k-vecinos más cercanos

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.651    25 0.00235 Preprocessor1_Model1
2 roc_auc  hand_till  0.794    25 0.00210 Preprocessor1_Model1
```

Figura 33. Modelo de los k-vecinos más cercanos en el punto B.

Los resultados obtenidos con el modelo de los k-vecinos más cercanos indican una precisión del 65.1% en el punto B. Al igual que en el caso anterior, se tiene que aproximadamente 7 de cada 10 datos de prueba se predicen correctamente con este modelo.

La precisión en este caso es ligeramente menor en comparación con el modelo del árbol de decisión.

Punto C: Escuela de formación de Tecnólogos (ESFOT).

Tabla 4. Resultados del análisis Anova y Tuckey obtenidos del punto C.

Parámetro	**Resultado**	**Diff**	**P adj**
Temperatura	Mala-Buena	-0.86	0.78
	Media-Buena	1.82	0
	Media-Mala	2.69	0.09
Radiación ultravioleta	Mala-Buena	-1.27	0.55
	Media-Buena	1.59	0

	Media-Mala	2.86	0.05
Velocidad del viento	Mala-Buena	-1.69	0.82
	Media-Buena	-2.19	0
	Media-Mala	-0.49	0.98
Presión atmosférica	Mala-Buena	-1.04	0.29
	Media-Buena	-0.07	0.53
	Media-Mala	-0.96	0.34
Periodo del día	Mala-Buena	-6.66	0.99
	Media-Buena	-6.66	0.81
	Media-Mala	0	1
RSSI	Mala-Buena	-3.54	0.07
	Media-Buena	0.39	0.04
	Media-Mala	3.93	0.04
RSSNR	Mala-Buena	-5.25	0
	Media-Buena	-1.37	0
	Media-Mala	3.87	0.02

De acuerdo con los resultados presentados en la tabla 4, las variables predictoras que se ajustan a un “p adj” menor a 0.05 son todas menos dos (período del día y presión atmosférica), por lo tanto, se tendrán 5 variables predictoras.

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          esfot_prediccion
           BUENA MALA MEDIA
  BUENA      442    0    26
  MALA         4    0     0
  MEDIA      114    0    88

Overall Statistics

                Accuracy : 0.7864
                  95% CI : (0.7534, 0.8167)
     No Information Rate : 0.8309
     P-Value [Acc > NIR] : 0.9988
```

Figura 34. Resultados obtenidos de la matriz de confusión en el punto C.

Los resultados obtenidos presentados en la figura 34, indican una precisión del 78.64%, por lo que aproximadamente 8 de cada 10 datos de prueba fueron

etiquetados correctamente. Por otro lado, la matriz de confusión proporciona la siguiente información:

- De un total de 468 datos, 442 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 4 datos, 0 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 202 datos, 88 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama del árbol de decisión obtenido**

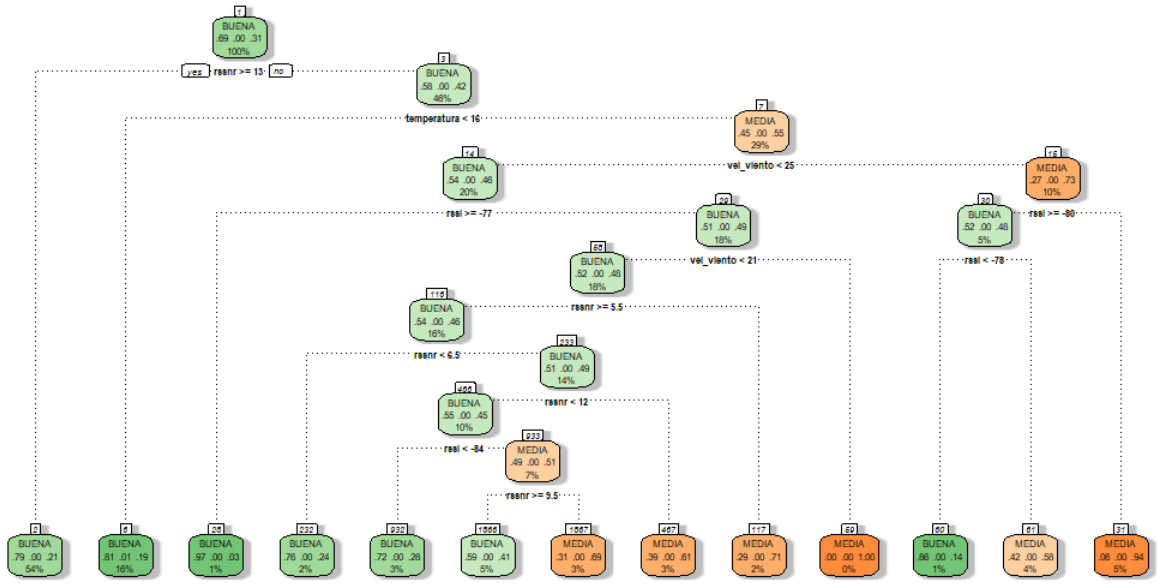

Figura 35. Árbol de decisión obtenido en el punto C.

El árbol de decisión presente en la figura 35 presenta la siguiente información:

- El nodo raíz (mejor variable predictora) corresponde a la variable RSSNR (“rssnr”), en este punto las mediciones obtenidas son mayoritariamente de calidad “BUENA” y “MEDIA”, por lo tanto, no se generan condiciones que caractericen al nivel de calidad “MALA”.
- Los datos que fueron etiquetados en la clase “BUENA” se dieron principalmente en los casos en los que el RSSNR resultó mayor o igual a 13dB, en temperaturas menores a 16°C y en velocidades del viento inferiores a 21km/h.
- Los datos que fueron etiquetados en la clase “MEDIA” se dieron principalmente en los casos en los que el RSSI resultó menor a -80dBm.

Punto D: Facultad de Ingeniería Eléctrica y Electrónica (Hall del edificio).

Tabla 5. Resultados de los análisis Anova y Tuckey obtenidos en el punto D.

Parámetro	**Resultado**	**Diff**	**P adj**
Temperatura	Mala-Buena	0.41	0.03
	Media-Buena	0.53	0.004
	Media-Mala	0.11	0.02
Radiación ultravioleta	Mala-Buena	1.98	0
	Media-Buena	1.41	0
	Media-Mala	-0.57	0
Velocidad del viento	Mala-Buena	-1.30	0.15
	Media-Buena	-0.66	0.62
	Media-Mala	0.64	0.001
Presión atmosférica	Mala-Buena	-4.78	0
	Media-Buena	-3.22	0
	Media-Mala	1.57	0
Periodo del día	Mala-Buena	-0.57	0
	Media-Buena	-0.27	5.87e-5
	Media-Mala	0.29	0
RSSI	Mala-Buena	-3.49	0
	Media-Buena	-2.56	0
	Media-Mala	0.93	0
RSSNR	Mala-Buena	-4.85	0
	Media-Buena	-3.85	0
	Media-Mala	0.99	0

Como se puede visualizar en la tabla 5, todas las variables excepto una (velocidad del viento) tienen un valor “p adj” menor a 0.05 por lo tanto, se tendrán 6 variables predictoras para cada modelo de aprendizaje automático.

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          hall_prediccion
           BUENA MALA MEDIA
  BUENA       14     0     7
  MALA         0   607    55
  MEDIA        0    72   228

Overall Statistics

                  Accuracy : 0.8637
                    95% CI : (0.8406, 0.8845)
       No Information Rate : 0.6907
       P-Value [Acc > NIR] : < 2.2e-16
```

Figura 36. Resultados obtenidos de la matriz de confusión en el punto D.

Los resultados obtenidos, presentes en la figura 36, indican una precisión del 86.37%, por lo que aproximadamente 9 de cada 10 datos de prueba fueron etiquetados correctamente. Por otro lado, la matriz de confusión proporciona la siguiente información:

- De un total de 21 datos, 14 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 662 datos, 607 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 300 datos, 228 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama de decisión obtenido**

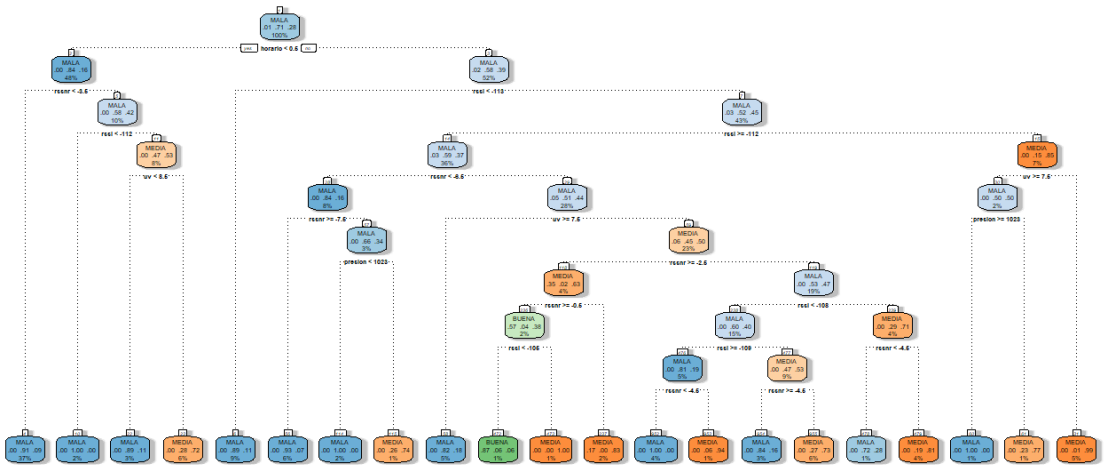

Figura 37. Diagrama del árbol de decisión obtenido en el punto D.

El árbol de decisión presente en la figura 37 presenta la siguiente información:

- La clase “BUENA” es casi nula en este punto, pues tan solo el 1% de los datos han sido etiquetados con esta clase.
- El nodo raíz (mejor variable predictora) corresponde a la variable “período del día”. Es muy importante recordar que para esta variable se tienen dos etiquetas numéricas, 1 y 0 para la tarde y mañana, respectivamente. En la totalidad de los datos la clase predominante es la clase “MALA”, tanto en la tarde como en la mañana.
- Los parámetros de RF que influyen son RSSI y RSSNR en la calidad de la conexión.
- La clase “MALA” predomina cuando se tienen valores de RSSNR menores a -3.5dB.
- De los resultados obtenidos en este punto se observa que el parámetro de radiación UV y la presión atmosférica probablemente influyen en la calidad de la conexión.
- Cuando se cumplen ciertas condiciones de los parámetros de RF y la radiación UV es mayor a 7.5 es probable que la calidad sea “MALA” [4], [5].

Técnica de aprendizaje automático: k-vecinos más cercanos.

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.81     25  0.0201 Preprocessor1_Model1
2 roc_auc  hand_till  0.882    25  0.0158 Preprocessor1_Model1
```

Figura 38. Resultados obtenidos del modelo k-vecinos más cercanos en el punto D.

Punto E: Facultad de Ingeniería Eléctrica y Electrónica (Aula S1/E001), **Punto F:** Estadio de la Escuela Politécnica Nacional y **Punto G:** Patio aledaño al Instituto de Ciencias Básicas (ICB).

Tabla 6. Resultados de los análisis Anova y Tukey de los puntos E, F y G.

Parámetro	**Resultado**	**Diff**	**P adj**
Temperatura	Mala-Buena	1.19	0
	Media-Buena	1.55	0
	Media-Mala	0.36	0.24
Radiación ultravioleta	Mala-Buena	-1.33	0
	Media-Buena	1.62	0
	Media-Mala	2.97	0
Velocidad del viento	Mala-Buena	5.31	0

	Media-Buena	-1.43	0.005
	Media-Mala	-6.74	0
Presión atmosférica	Mala-Buena	2.2	0
	Media-Buena	3.4	0
	Media-Mala	1.19	0.034
Periodo del día	Mala-Buena	0.47	0
	Media-Buena	-0.007	0.04
	Media-Mala	-0.55	0
RSSI	Mala-Buena	-3.87	0.03
	Media-Buena	-11.37	0
	Media-Mala	-7.51	0
RSSNR	Mala-Buena	-12.72	0
	Media-Buena	-7.30	0
	Media-Mala	5.41	0

Como se puede visualizar en la tabla 6, todas las variables tienen un valor “p adj” menor a 0.05 por lo tanto, se tendrán 7 variables predictoras para cada modelo de aprendizaje automático.

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          aep_prediccion
           BUENA MALA MEDIA
  BUENA    113    0     2
  MALA       0   12     8
  MEDIA     13    0    55

Overall Statistics

                Accuracy : 0.8867
                  95% CI : (0.8349, 0.9268)
    No Information Rate : 0.6207
    P-Value [Acc > NIR] : < 2.2e-16
```

Figura 39. Resultados obtenidos de la matriz de confusión de los puntos E, F y G.

Los resultados obtenidos presentes en la figura 39, indican una precisión del 88.67%. Por otro lado, la matriz confusión proporciona la siguiente información:

- De un total de 115 datos, 113 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 20 datos, 12 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 68 datos, 55 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama del árbol de decisión obtenido**

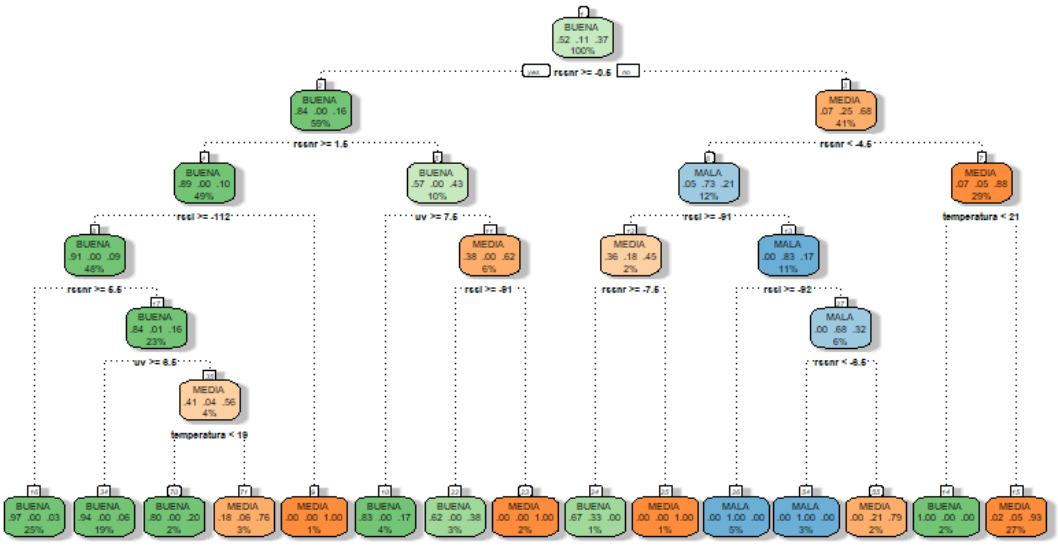

Figura 40. Diagrama del árbol de decisión obtenido en los puntos E, F y G.

- Como se puede visualizar en la figura 40, el nodo raíz (mejor variable predictora) corresponde a la variable RSSNR (“rssnr”). Las clases predominantes corresponden las clases “BUENA” y “MEDIA” con porcentajes del 52% y 37%, respectivamente.
- Dentro de la primera categoría, cuando se tiene un valor de RSSNR mayor o igual -0.5dB y niveles de RSSI mayores a -112, las etiquetas asociadas se ajustan a la clase “BUENA”.
- En este punto se observa que a temperaturas menores a 19°C es muy probable que se tengan conexiones de calidad “BUENA”. Se observa que cuando los niveles de RSSI y RSSNR se encuentran en niveles para un tipo de conexión “BUENA”, el nivel de radiación UV no afecta en dicha calidad, ya que se tienen niveles de calidad “BUENA” o “MEDIA” [4], [5].
- En este punto los niveles de calidad “MALA” se presentan principalmente en función de los valores de los parámetros RSSI y RSSNR, en donde esta calidad

se tiene principalmente cuando el RSSNR es menor a –8.6 y el RSSI es menor a –91.

Técnica de aprendizaje automático: k-vecinos más cercanos

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.910    25 0.0118  Preprocessor1_Model1
2 roc_auc  hand_till  0.970    25 0.00699 Preprocessor1_Model1
```

Figura 41. Resultados obtenidos del modelo k-vecinos más cercanos en los puntos E, F y G.

Los resultados obtenidos en la Figura 41, presentan una precisión del 91% con lo que 9 de cada 10 datos de prueba han sido etiquetados correctamente en cada una de las clases. En este punto, la precisión de este modelo es ligeramente mayor, en comparación con el modelo de los árboles de decisión.

Punto H: Afueras de la Biblioteca Central.

Tabla 7. Resultados obtenidos del análisis Anova y Tukey en el punto H.

Parámetro	**Resultado**	**diff**	**P adj**
Temperatura	Mala-Buena	-0.10	0.03
	Media-Buena	-0.46	0
	Media-Mala	-0.36	0
Radiación ultravioleta	Mala-Buena	0.15	0
	Media-Buena	-0.09	0.004
	Media-Mala	-0.24	0
Velocidad del viento	Mala-Buena	0.43	0.07
	Media-Buena	-0.93	0
	Media-Mala	-1.37	0
Presión atmosférica	Mala-Buena	1.96	0
	Media-Buena	0.52	0.001
	Media-Mala	-1.44	0
Periodo del día	Mala-Buena	-3.67e-15	0.39
	Media-Buena	0	1
	Media-Mala	3.66e-15	0.19
RSSI	Mala-Buena	-3.40	0
	Media-Buena	-2.10	0
	Media-Mala	1.29	0

RSSNR	Mala-Buena	-5.26	0
	Media-Buena	-2.75	0
	Media-Mala	2.50	0

Como se puede visualizar en la tabla 7, todas las variables, excepto una (período del día) tienen un valor “p adj” menor a 0.05, por lo tanto, se tendrán 6 variables predictoras para cada modelo de aprendizaje automático.

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          bib_prediccion
           BUENA MALA MEDIA
  BUENA       75    3   101
  MALA         0  111   152
  MEDIA       34   52   577

Overall Statistics

                Accuracy : 0.6905
                  95% CI : (0.6623, 0.7177)
     No Information Rate : 0.7511
     P-Value [Acc > NIR] : 1
```

Figura 42. Resultados obtenidos de la matriz de confusión en el punto H para el árbol de decisión.

Los resultados obtenidos presentes en la figura 42, indican una precisión del 69.05%, por lo que aproximadamente 7 de cada 10 datos de prueba fueron etiquetados correctamente. Por otro lado, la matriz de confusión proporciona la siguiente información:

- De un total de 179 datos, 75 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 263 datos, 111 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 663 datos, 577 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama del árbol de decisión obtenido**

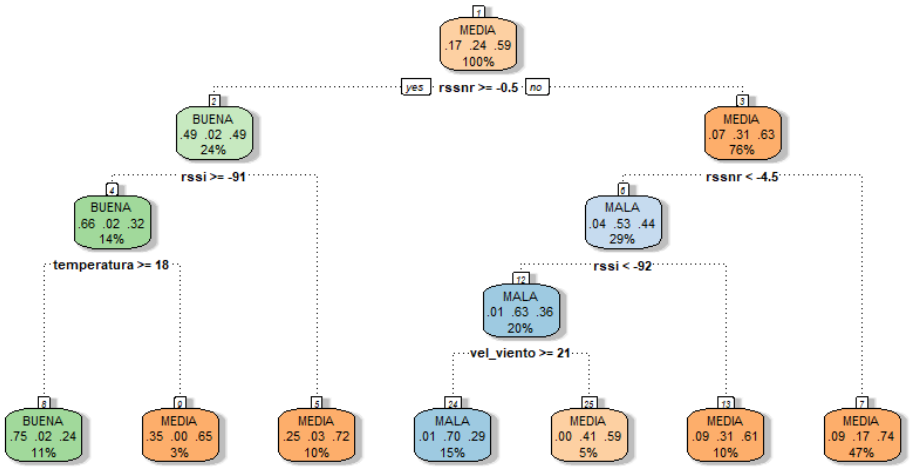

Figura 43. Diagrama del árbol de decisión obtenido en el punto H.

- Como se puede visualizar en la figura 43, el nodo raíz (mejor variable predictora) corresponde a la variable RSSNR ("rssnr"). Las clases predominantes corresponden a las clases "MALA" y "MEDIA" con porcentajes del 24% y 59%, respectivamente. Dentro de la segunda categoría con condiciones de RSSI menores a -92 y RSSNR menores a –92 y con velocidades del viento menores a 21km/h los datos están etiquetados en la clase "MEDIA", mientras que a velocidades mayores o iguales a 21km/h, los datos en una buena parte están etiquetados en la clase "MALA".

Técnica de aprendizaje automático: k-vecinos más cercanos

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.771    25 0.00535 Preprocessor1_Model1
2 roc_auc  hand_till  0.899    25 0.00330 Preprocessor1_Model1
```

Figura 44. Resultados de evaluación del modelo k-vecinos más cercanos en el punto H.

Los resultados obtenidos en la Figura 44, presentan una precisión del 77.1% con lo que aproximadamente 8 de cada 10 datos de prueba han sido etiquetados correctamente en cada una de las clases. En este punto, la precisión de este modelo es mayor, en comparación con el modelo de los árboles de decisión.

Punto I: Afueras de la Casa Patrimonial de la EPN.

Tabla 8. Resultados del análisis Anova y Tukey obtenidos en el punto I.

Parámetro	**Resultado**	**diff**	**P adj**
Temperatura	Mala-Buena	-0.02	0.99

	Media-Buena	0.32	0
	Media-Mala	0.35	0.17
Radiación ultravioleta	Mala-Buena	-0.46	0.21
	Media-Buena	0.41	0
	Media-Mala	0.87	0.004
Velocidad del viento	Mala-Buena	0.73	0.103
	Media-Buena	0.79	0
	Media-Mala	0.069	0.97
Presión atmosférica	Mala-Buena	-0.049	0.99
	Media-Buena	0.83	0
	Media-Mala	0.87	0.15
Periodo del día	Mala-Buena	0.027	0.07
	Media-Buena	0.02	0
	Media-Mala	-0.006	0.87
RSSI	Mala-Buena	-5.44	0
	Media-Buena	-0.46	0
	Media-Mala	4.98	0
RSSNR	Mala-Buena	-8.19	0
	Media-Buena	-2.37	0
	Media-Mala	5.82	0

Como se puede visualizar en la tabla 8, tan solo tres variables (RSSI, radiación ultravioleta y RSSNR) tienen un valor “p adj” menor a 0.05, por lo tanto, se tendrán 3 variables predictoras para cada modelo de aprendizaje automático.

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          patrimonial_prediccion
           BUENA MALA MEDIA
  BUENA      415    0   158
  MALA         2    0    23
  MEDIA      110    0   551

Overall Statistics

                Accuracy : 0.7673
                  95% CI : (0.7429, 0.7904)
     No Information Rate : 0.5814
     P-Value [Acc > NIR] : < 2.2e-16
```

Figura 45. Matriz de confusión del árbol de decisión en el punto I.

Se debe indicar que, en este dataset, las clases se encuentran desbalanceadas y se tiene una cantidad limitada de mediciones de la calidad “MALA”. Los resultados obtenidos, presentes en la figura 45, indican una precisión del 76.73%, por lo que aproximadamente 8 de cada 10 datos de prueba fueron etiquetados correctamente. Por otro lado, la matriz de confusión proporciona la siguiente información:

- De un total de 573 datos, 415 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 25 datos, 0 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 661 datos, 551 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama del árbol de decisión obtenido**

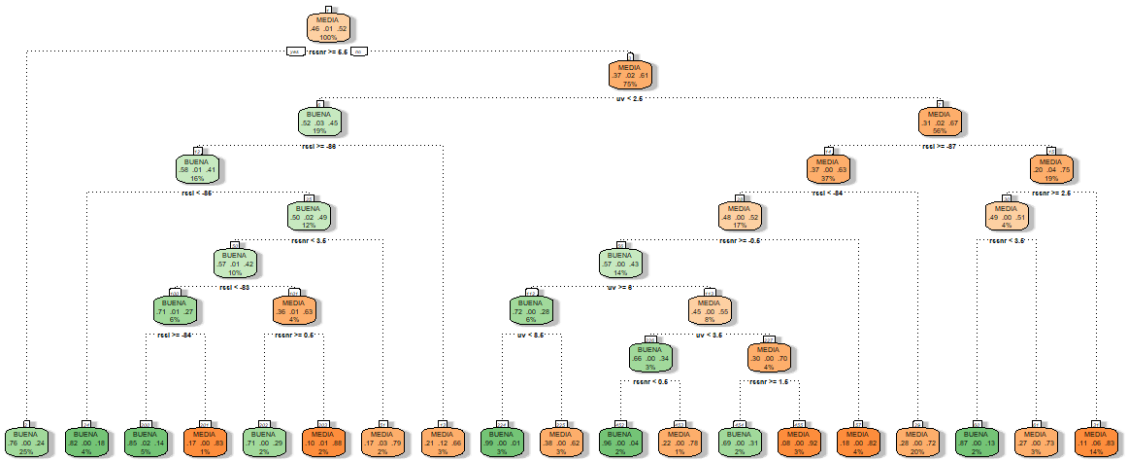

Figura 46. Diagrama del árbol de decisión obtenido en el punto I.

- Como se puede visualizar en la figura 46, el nodo raíz (mejor variable predictora) corresponde a la variable RSSNR (“rssnr”). Las clases predominantes en el dataset corresponden a las clases “BUENA” y “MEDIA” con porcentajes del 46% y 52%, respectivamente. Por el desbalanceo del dataset; las condiciones establecidas por el árbol solamente diferencia entre clases “BUENA” y “MEDIA” y no presenta ninguna condición para la clase “MALA”.
- Sin embargo, la segunda categoría $(rssnr < 6.6dB)$, es la única que presenta un factor ambiental (índice de radiación ultravioleta) que influye en la calidad. El índice de radiación ultravioleta se etiqueta en la clase “BUENA” cuando se tiene un índice menor a 8.6; por otro lado, también se etiqueta en la clase “MEDIA” cuando este índice es mayor o igual a 8.6.
- Para este punto, tan solo el índice ultravioleta influye en la calidad de recepción, sin embargo, es muy probable que no influya de manera significativa, dado los bajos porcentajes que se pueden visualizar en la gráfica.

Técnica de aprendizaje automático: k-vecinos más cercanos

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.709    25  0.0155 Preprocessor1_Model1
2 roc_auc  hand_till  0.849    25  0.0138 Preprocessor1_Model1
```

Figura 47. Evaluación del modelo k-vecinos más cercanos en el punto I.

Los resultados obtenidos en la Figura 47, presentan una precisión del 70.9% con lo que 7 de cada 10 datos de prueba han sido etiquetados correctamente en cada una de las clases. En este punto, la precisión de este modelo es menor, en comparación con el modelo de los árboles de decisión.

Punto J: Entrada Facultad de Ingeniería Civil.

Tabla 9. Resultados de los análisis Anova y Tukey obtenidos en el punto J.

Parámetro	**Resultado**	**diff**	**P adj**
Temperatura	Mala-Buena	1.94	0
	Media-Buena	1.42	0
	Media-Mala	-0.53	0.104
Radiación ultravioleta	Mala-Buena	1.99	0
	Media-Buena	1.94	0
	Media-Mala	-0.04	0.99

Velocidad del viento	Mala-Buena	-2.24	0
	Media-Buena	-1.07	0
	Media-Mala	1.16	0.006
Presión atmosférica	Mala-Buena	7.33	0
	Media-Buena	5.24	0
	Media-Mala	-2.08	0.03
Periodo del día	Mala-Buena	0.095	0.19
	Media-Buena	-0.05	0.1
	Media-Mala	-0.15	0.02
RSSI	Mala-Buena	8.42	0
	Media-Buena	3.9	0
	Media-Mala	-4.52	6.88e-5
RSSNR	Mala-Buena	9.19	0
	Media-Buena	4.04	0
	Media-Mala	-5.16	0

Como se puede visualizar en la tabla 9, todas las variables, excepto una (período del día), tienen un valor “p adj” menor a 0.05 por lo tanto, se tendrán 6 variables predictoras para cada modelo de aprendizaje automático.

Técnica de aprendizaje automático: árboles de decisión

- **Matriz de confusión y resumen estadístico**

```
Confusion Matrix and Statistics

          civil_prediccion
           BUENA MALA MEDIA
  BUENA     89    0    40
  MALA       3    9     0
  MEDIA      3    0    62

Overall Statistics

                Accuracy : 0.7767
                  95% CI : (0.7136, 0.8316)
    No Information Rate : 0.4951
    P-Value [Acc > NIR] : < 2.2e-16
```

Figura 48. Resultados de evaluación del modelo de árboles de decisión en el punto J.

Los resultados obtenidos, presentes en la figura 48, indican una precisión del 77.67%, por lo que aproximadamente 8 de cada 10 datos de prueba fueron etiquetados correctamente. Por otro lado, la matriz de

confusión proporciona la siguiente información:

- De un total de 129 datos, 89 fueron etiquetados correctamente en la clase “BUENA”.
- De un total de 12 datos, 9 fueron etiquetados correctamente en la clase “MALA”.
- De un total de 65 datos, 62 fueron etiquetados correctamente en la clase “MEDIA”.

- **Diagrama del árbol de decisión obtenido**

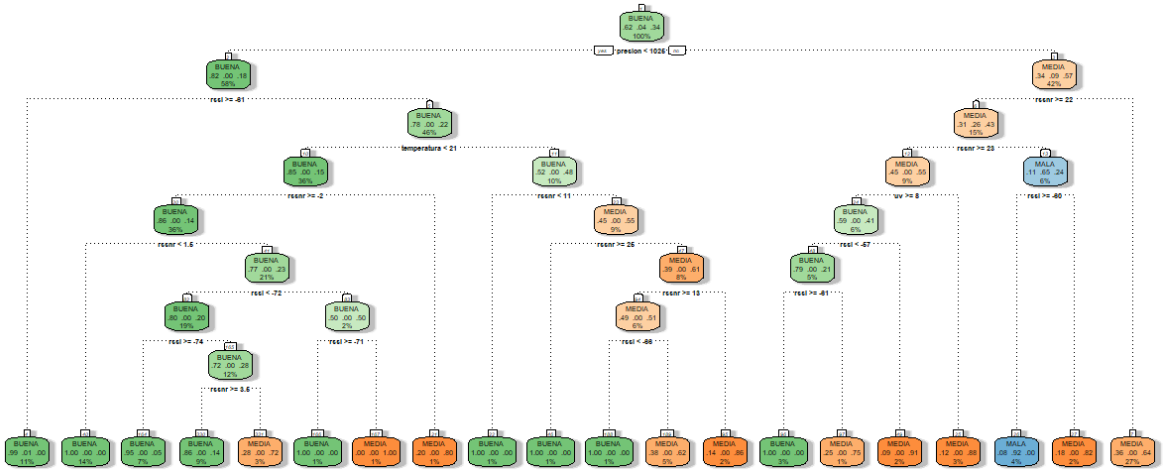

Figura 49. Diagrama del árbol de decisión obtenido en el punto J.

- Como se puede visualizar en la figura 49, el nodo raíz (mejor variable predictora) corresponde a la variable presión atmosférica (“presión”) con un valor referencial de 1026mb para diferenciar entre calidad “BUENA” y “MEDIA”.
- El nivel de calidad “MALA” se presenta solo con niveles referenciales de RSSNR y RSSI.
- Para este punto, no existe una influencia destacable de los factores ambientales sobre la calidad de recepción de la señal.

Técnica de aprendizaje automático: árboles de decisión

```
# A tibble: 2 × 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <chr>
1 accuracy multiclass 0.742    25  0.0210 Preprocessor1_Model1
2 roc_auc  hand_till  0.864    25  0.0164 Preprocessor1_Model1
```

Figura 50. Evaluación del modelo k-vecinos más cercanos en el punto J.

Los resultados obtenidos en la Figura 50 presentan una precisión del 74.2% con lo que 7 de cada 10 datos de prueba aproximadamente han sido etiquetados correctamente en cada una de las clases. En este punto, la precisión de este modelo es menor, en comparación con el modelo de los árboles de decisión.

Resumen final de los resultados obtenidos.

A continuación, en la tabla 10 se presentan los resultados obtenidos a modo de resumen de los parámetros más relevantes que se han obtenido en este estudio. En esta tabla se presentan las variables ambientales y de RF que probablemente repercuten en la calidad de la señal; es decir, aquellas que se etiquetan en mayor medida en la clase “MALA”.

Tabla 10. Resumen de los resultados obtenidos.

Punto	**Tamaño total de la muestra**			**Técnica de aprendizaje automático**		**Factor ambiental que probablemente influye**	**Parámetro de RF que probablemente influye**
				Árbol de decisión	**K-vecinos más cercanos**		
	Buena	**Mala**	**Media**	**Precisión [%]**	**Precisión [%]**		
A	511	72	1630	73.9	64.8	Ninguna	Ninguna
B	1024	1826	3598	75.6	65.1	Velocidad del viento	RSSI
C	468	4	202	78.6	79.8	Ninguna	Ninguna
D	21	662	300	86.37	81	Radiación ultravioleta	RSSNR

E, F y G	115	20	68	88.7	91	Ninguna	RSSNR
H	179	263	663	69.1	77.1	Velocidad del viento	RSSI
I	573	25	661	76.7	70.9	Ninguna	Ninguna
J	129	3	65	77.7	74.2	Ninguna	RSSI

3.2 CONCLUSIONES

- Los parámetros de RF entregados por las aplicaciones Net Monitor y CellMapper permitieron la recolección de los datos de RF pertinentes para este estudio. La aplicación Net Monitor permitió la recolección de los parámetros RSSI y RSSNR, mientras que la aplicación CellMapper permitió definir la ubicación de las estaciones base que daban cobertura en la zona de interés.

- Como parte del preprocesamiento de datos, fue muy importante depurar información poco relevante, de tal manera que los resultados no se vean comprometidos con la presencia de valores no deseados que pueden alterar los análisis de los parámetros que influyen o reflejan el nivel de la calidad de recepción de la señal.

- Como se puede visualizar en la tabla 10, se ha considerado establecer una comparación entre la precisión de dos técnicas de aprendizaje automático (árboles de decisión y k-vecinos más cercanos) para clasificar las conexiones en calidad “BUENA”, “MEDIA” y “MALA”. La tabla permite visualizar que el modelo de aprendizaje automático con la mayor precisión corresponde al modelo de los árboles de decisión, en 6 de los 8 puntos de interés; es decir, se tiene un predominio de este modelo utilizando las mismas variables predictoras en cada uno.

- Como se puede visualizar en la tabla 10, se tiene una división de los parámetros de RF y de los factores ambientales que repercuten en mayor medida con la calidad de la señal. Dentro de los parámetros de RF, valores de RSSI menores a -80dBm y valores de RSSNR menores a -4.5dB se presentan en un deterioro de la calidad de la señal, la cual se etiqueta como “MALA”. Por otra parte, dentro de los factores climáticos, velocidades del viento mayores o iguales a 21km/h e índices de

radiación ultravioleta mayores a 7.5 se presentan en un deterioro de la calidad de la señal, la cual es etiquetada como “MALA”.

- Es importante recalcar que en el punto D, el nodo raíz correspondió a la variable período del día, donde el 71% de los datos fue etiquetado en la clase “MALA”. Se tiene que en el período del día “MAÑANA” se tienen los mayores índices de radiación ultravioleta; por ende, es probable que la calidad de la señal en la mañana se vea influenciado por este factor climático.

3.3 RECOMENDACIONES

- Para tener una certeza total de cómo influyen los factores ambientales descritos en este estudio, sería indispensable contar con un software que sea capaz de registrar, segundo a segundo (como lo hace Net Monitor), los valores de temperatura, radiación ultravioleta, presión atmosférica y velocidad del viento, de tal manera que se tengan datos más precisos y en tiempo real del cambio de estos factores climáticos y así poder confirmar los resultados de este estudio.

4 REFERENCIAS BIBLIOGRÁFICAS

[1] CellMapper, «First time startup», *CellMapper*. Disponible en: https://www.cellmapper.net/First_Time_Startup, https://www.cellmapper.net/it. [Accedido: 15 de agosto de 2023]

[2] «NetMonitor Cell Signal Logging - Display network technology», *FREE-APPS-ANDROID.COM*, 8 de junio de 2018. Disponible en: https://free-apps-android.com/netmonitor-cell-signal-logging/. [Accedido: 15 de agosto de 2023]

[3] «Received signal strength indicator», *Wikipedia*. 31 de marzo de 2023. Disponible en: https://en.wikipedia.org/w/index.php?title=Received_signal_strength_indicator&oldid=1147491776. [Accedido: 15 de agosto de 2023]

[4] vadims.kalejs, «What is RSSI, SINR, RSRP, RSRQ? How does this affect signal quality?», *Mobile Signal Booster - Wide Range of Repeaters for Home and Office*, 27 de enero de 2022. Disponible en: https://www.rangeful.com/what-is-rssi-sinr-rsrp-rsrq-how-does-this-affect-signal-quality/. [Accedido: 15 de agosto de 2023]

[5] «Understand LTE Signal Strength Values». Disponible en: https://webhelp.tempered.io/webhelp/kb_lte_signal.html. [Accedido: 15 de agosto de 2023]

[6] Faleti, Idowu & Nwanya, Stephen & Njoku, Howard & Obi, Amarachukwu & Dzah, Julius. (2021). Effect of weather conditions on Network Performance of 3G and 4G Networks in Nigeria -Review.

[7] «R: What is R?» Disponible en: https://www.r-project.org/about.html. [Accedido: 15 de agosto de 2023]

[8] W. Navidi, «Estadística para ingenieros y científicos», 5th ed. México, 2022, pp.13-35

[9] «Analysis of Variance (ANOVA) Explanation, Formula, and Applications», *Investopedia*. Disponible en: https://www.investopedia.com/terms/a/anova.asp. [Accedido: 15 de agosto de 2023]

[10] «Tukey's range test», *Wikipedia*. 28 de julio de 2023. Disponible en: https://en.wikipedia.org/w/index.php?title=Tukey%27s_range_test&oldid=1167492043. [Accedido: 15 de agosto de 2023]

[11] «Machine Learning Techniques - Javatpoint», *www.javatpoint.com*. Disponible en: https://www.javatpoint.com/machine-learning-techniques. [Accedido: 15 de agosto de 2023]

[12] D.-D. Science, «7 Stages of Machine Learning — A Framework», *Medium*, 20 de julio de 2020. Disponible en: https://medium.com/@datadrivenscience/7-stages-of-machine-learning-a-framework-33d39065e2c9. [Accedido: 15 de agosto de 2023]

[13] J. Saltz, «The Machine Learning Process», *Data Science Process Alliance*, 31 de mayo de 2022. Disponible en: https://www.datascience-pm.com/machine-learning-process/. [Accedido: 21 de agosto de 2023]

[14] A. Saini, «Decision Tree Algorithm - A Complete Guide», *Analytics Vidhya*, 29 de agosto de 2021. Disponible en: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/. [Accedido: 15 de agosto de 2023]

[15] D. Bhalla, «Decision Tree in R : Step by Step Guide», *ListenData*. Disponible en: https://www.listendata.com/2015/04/decision-tree-in-r.html. [Accedido: 20 de agosto de 2023

[16] «RPubs - KNN with R». Disponible en: https://rpubs.com/pmtam/knn. [Accedido: 21 de agosto de 2023]

[17] «R Boxplot - javatpoint», *www.javatpoint.com*. Disponible en: https://www.javatpoint.com/r-boxplot. [Accedido: 20 de agosto de 2023]

[18] «BOXPLOT in R [boxplot by GROUP, MULTIPLE box plot, ...]», *R CODER*, 11 de abril de 2020. Disponible en: https://r-coder.com/boxplot-r/. [Accedido: 15 de agosto de 2023]

[19] S. Midway, *Chapter 7 Understanding ANOVA in R | Data Analysis in R*. Disponible en: https://bookdown.org/steve_midway/DAR/understanding-anova-in-r.html. [Accedido: 15 de agosto de 2023]

[20] G. Chordia, «Recipes in R», *Medium*, 8 de septiembre de 2022. Disponible en: https://gaganchordia.medium.com/recipes-all-about-data-preprocessing-in-r-d97a3466d8a5. [Accedido: 15 de agosto de 2023]

[21] «juice function - RDocumentation». Disponible en: https://www.rdocumentation.org/packages/recipes/versions/1.0.6/topics/juice. [Accedido: 15 de agosto de 2023]

[22] «How To Use the predict() Function in R Programming | DigitalOcean». Disponible en: https://www.digitalocean.com/community/tutorials/predict-function-in-r. [Accedido: 15 de agosto de 2023]

[23] «Confusion Matrix in R | A Complete Guide | DigitalOcean». Disponible en: https://www.digitalocean.com/community/tutorials/confusion-matrix-in-r. [Accedido: 15 de agosto de 2023]

[24] D. Brown, «Use Rattle to Help You Learn R», *Medium*, 3 de abril de 2021. Disponible en: https://towardsdatascience.com/use-rattle-to-help-you-learn-r-d495c0cc517f. [Accedido: 15 de agosto de 2023]

[25] «Cross-Validation Essentials in R - Articles - STHDA», 11 de marzo de 2018. Disponible en: http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/. [Accedido: 15 de agosto de 2023]

[26] Fit multiple models via resampling — fit_resamples». Disponible en: https://tune.tidymodels.org/reference/fit_resamples.html#:~:text=fit_resamples()%20computes%20a%20set,formula%20combination%20across%20many%20resamples. [Accedido: 15 de agosto de 2023]