Generación de datos sintéticos a partir de mediciones RSSI en una red SigFox utilizando Data Augmentation

Leiton Reina, Kevin Alexander y Tigse Pérez, Dennis Eduardo

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Trabajo de titulación, previo a la obtención del título de Ingeniero en Electrónica y Telecomunicaciones

Ing. Román Alcides Lara Cueva PhD.

6 de diciembre del 2023

Copyleaks

Plagiarism report

TESIS_LEITON_TIGSE_27_10_2023.pdf

Scan details

Scan time:	Total Pages:	Total Words:
October 30th, 2023 at 12:10 UTC	71	17627

Plagiarism Detection

2.9%

Types of plagiarism		Words
Identical	1%	172
Minor Changes	0.6%	112
Paraphrased	1.3%	231
Omitted Words	0%	0

AI Content Detection

N/A

Text coverage
- AI text
- Human text

Plagiarism Results: (14)

La matriz de confusión y sus métricas – Juan Barrios — 0.8%

https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-...

Skip to content Inicio Biografía COVID-19 Destac...

Evaluando los modelos de Clasificación en Aprendiz... — 0.5%

https://profesordata.com/2020/08/07/evaluando-los-modelo...

...

▷ Machine Learning: Su Definición, Propósito y Mec... — 0.5%

https://canalinnova.com/machine-learning-su-definicion-pro...

Eduardo Rodríguez

Saltar al contenido Menú Menú Software Temáticas Tec...

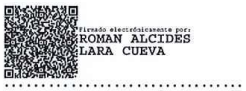
Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Certificación

Certifico que el trabajo de titulación: "**Generación de datos sintéticos a partir de mediciones RSSI en una red SigFox utilizando Data Augmentation**" fue realizado por los señores **Leiton Reina Kevin Alexander y Tigse Pérez Dennis Eduardo;** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 06 de diciembre de 2023



Ing. Román Alcides Lara Cueva PhD.

C.I.: 1713988218

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Responsabilidad de Autoría

Nosotros, **Leiton Reina Kevin Alexander con C.C 1726666132 y Tigse Pérez Dennis Eduardo con C.C 1725934978**, declaramos que el contenido, ideas y criterios del trabajo de titulación: "**Generación de datos sintéticos a partir de mediciones RSSI en una red SigFox utilizando Data Augmentation**" es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 06 de diciembre de 2023

Leiton Reina Kevin Alexander
C.C: 1726666132

Tigse Pérez Dennis Eduardo
C.C: 1725934978

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Autorización de Publicación

Nosotros **Leiton Reina Kevin Alexander con C.C 1726666132 y Tigse Pérez Dennis Eduardo con C.C 1725934978**, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Generación de datos sintéticos a partir de mediciones RSSI en una red SigFox utilizando Data Augmentation**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 06 de diciembre del 2023

Leiton Reina Kevin Alexander

C.C: 1726666132

Tigse Pérez Dennis Eduardo

C.C: 1725934978

Dedicatoria

Dedico este trabajo primero a Dios por estar presente en cada paso de mi proceso en la Universidad, y a mis padres Alba y Cástulo que me han apoyado en este camino de forma incondicional, recibiendo valiosos consejos y los más sinceros deseos para alcanzar el éxito profesional a lo largo de mi vida, les agradezco y aprecio mucho.

Leiton Reina, Kevin Alexander

El siguiente trabajo de titulación le dedico a Dios, por siempre darme la sabiduría necesaria al momento de tomar decisiones y que me ha hecho crecer como persona. De manera similar a mis padres que son el pilar fundamental en toda mi vida que siempre tuvieron el apoyo hacia mí, y a mis hermanos que me acompañaron en esta etapa de mi vida. Este éxito es por ustedes.

Tigse Pérez, Dennis Eduardo

Agradecimientos

En primer lugar, agradezco a Dios por brindarme sabiduría, responsabilidad y compromiso en cada paso que he dado, y así lograr superar cada reto que la vida ha puesto en mi camino a lo largo de mi carrera.

Agradezco a mi madre Alba, gracias a sus consejos, su apoyo, su inmenso amor de madre, y sobre todo su comprensión, gracias por que siempre me brindó buenos consejos para poder superarme y para continuar adelante en mis estudios, además de sus innumerables cuidados en mi bienestar y su total apoyo en cada decisión que he tomado.

A mi padre Cástulo, quien me brindó todo el apoyo que un padre puede dar, sus sabias palabras para alentarme a continuar y a enseñarme que en la vida nada se consigue fácil y sin esfuerzo, fue y es una gran figura a la cual admiro mucho por sus enseñanzas y la forma en la que ha salido adelante en la vida, muchas gracias.

Agradezco a mi tutor Ing. Román Alcides Lara Cueva, PhD por compartir sus conocimientos como docente y como director del presente trabajo.

De igual manera agradezco a mi hermano, a mi novia y a mis amigos por su apoyo y por acompañarme a lo largo de esta hermosa carrera, finalmente me agradezco a mí, por no rendirme, por confiar siempre en las cosas buenas y por continuar superándome a mí mismo como persona y como profesional.

Leiton Reina, Kevin Alexander

Agradezco de todo corazón a Dios por darme cada día las fuerzas necesarias para ser una mejor persona cada día, por darme la mejor familia que me acompaña en todas las etapas de mi vida y por las personas que me rodean. Agradecido por los dones que Dios me ha podido regalar para seguir con más fuerza cada día.

A mis padres, quienes con sus palabras sabias me guían en cada momento de mi vida, me han demostrado que en la vida las cosas no pueden resultar fácil, pero nunca son imposibles de conseguir. Por brindar el apoyo incondicional a cada paso que día en mi carrera estudiantil y personal, por siempre confiar en mí en los momentos más difíciles de mi vida.

A mis hermanos, quienes han sido mi motivación para ser una mejor persona y ser un ejemplo para ellos, por todo su cariño y comprensión que me dan en cada momento.

A mis amigos que he logrado conocer a lo largo de mi carrera universitaria, solo queda decir que muchas gracias por tantos momentos de aprendizaje dentro y fuera de las aulas, me han demostrado su apoyo incondicional en todos los semestres.

Tigse Pérez, Dennis Eduardo

Índice de Contenido

Índice de Tablas

Índice de Figuras

Resumen

Actualmente, el desarrollo tecnológico trata de facilitar diversas tareas que se realizan de manera cotidiana, como en la predicción de rutas de destino en el transporte, la detección de enfermedades de manera temprana dentro de la salud, mejorar el diseño de las redes inalámbricas al verificar zonas con cobertura en el aspecto tecnológico, entre otros. Estas soluciones se obtienen a partir del uso de aplicaciones de *Deep Learning* y *Machine Learning,* mismos que permiten la generación de predicciones de valores o la toma de decisiones, para ello es necesario lograr el entrenamiento de estos algoritmos a partir de una base de datos inicial. La adquisición de esta información puede convertirse en un problema debido a la escaza en la cantidad de datos que se pueda recolectar.

Una solución para la obtención de datos es *Data Augmentation*, el cual emplea un conjunto de técnicas para generar artificialmente datos, este proyecto de investigación tiene como fin la generación de datos a partir de mediciones de niveles de intensidad de señal recibida (RSSI, del inglés *Received Signal Strength Indicator*) de una base de datos en el cantón Quito. Esta base de datos inicial presenta 5174 valores, de los cuales 1608 no poseen medición RSSI para lo cual se realizó varios regresores para la generación del valor RSSI y completar los valores nulos.

Además, se desarrollaron modelos de clasificación según el indicador de calidad del enlace (LQI, del inglés *Link Quality Indicator*) para etiquetar estos datos y observar si los datos poseen una buena o mala conexión. Posteriormente, se realizó el proceso de *Data Augmentation* al generar nuevas coordenadas a partir del programa *Collect Earth* y emplear el modelo de regresión con el mejor desempeño para obtener un valor RSSI, luego se utiliza el mejor modelo de clasificación para la generación de etiquetas LQI. Finalmente, se realiza la comparación de la base de datos inicial y la base de datos aumentada a través de los mapas de calor y los mapas basados en las etiquetas LQI.

*Palabras clave***:** Machine Learning, Data Augmentation, Regresor, Clasificador

Abstract

Currently, technological development seeks to facilitate various tasks that are performed on a daily basis, such as the prediction of destination routes in transportation, early detection of diseases in health, improving the design of wireless networks by verifying areas with coverage in the technological aspect, among others. These solutions are obtained from the use of Deep Learning and Machine Learning applications, which allow the generation of value predictions or decision making, for which it is necessary to train these algorithms from an initial database. The acquisition of this information can become a problem due to the LIMITEd amount of data that can be collected.

One solution for obtaining data is Data Augmentation, which employs a set of techniques to artificially generate data. This research project aims to generate data from measurements of Received Signal Strength Indicator (RSSI) levels from a database in the Quito canton. This initial database presents 5174 values, of which 1608 have no RSSI measurement, for which several regressors were performed to generate the RSSI value and complete the null values.

In addition, Link Quality Indicator (LQI) classification models were developed to label these data and observe whether the data have a BUENO or bad connection. Subsequently, the Data Augmentation process was performed by generating new coordinates from the Collect Earth program and using the regression model with the best performance to obtain an RSSI value, then the best classification model is used to generate LQI labels. Finally, the comparison of the initial database and the augmented database is performed through heat maps and maps based on LQI labels.

Keywords: Machine Learning, Data Augmentation, Regression, Classifier.

Capítulo I

Introducción

La revolución tecnológica abarca diversas áreas de conocimiento, como el Internet de las Cosas (IoT, del inglés *Internet Of Things*), para impulsar el progreso y la eficiencia en la sociedad. La Inteligencia Artificial, con sus subcampos como el *Machine Learning* y el *Deep Learning*, se han convertido en una herramienta fundamental en numerosas aplicaciones, desde la atención médica y la automoción hasta la industria manufacturera y la gestión de datos. Estas tecnologías permiten a las máquinas aprender, adaptarse y tomar decisiones de manera autónoma, de esta manera se fortalecen las bases tecnológicas e industriales.

La Inteligencia Artificial representa uno de los avances tecnológicos más relevantes de los últimos tiempos debido a la capacidad de las máquinas para aprender y procesar información de forma independiente con grandes cantidades de datos. Esta tecnología es relevante para automatizar procedimientos que normalmente se realizan manualmente, así se mejora el rendimiento de los procesos. Se trata de una ciencia interdisciplinaria que se destaca por incluir el *Machine Learning* y el *Deep Learning.*

El *Machine Learning* es una parte esencial de la Inteligencia Artificial que se enfoca en crear nuevos algoritmos que permiten a las máquinas aprender de forma autónoma. Esto lleva a diversas soluciones para problemas de predicción de valores o la toma de decisiones basadas en información inicial. Los modelos tradicionales de *Machine Learning* buscan generar estimaciones en un conjunto de datos al utilizar variables de entrada y generar variables de salida, conocidas como predicciones. Estos modelos incluyen algoritmos supervisados y no supervisados.

El *Deep Learning* se basa en el uso de redes neuronales con múltiples capas para simular el procesamiento del cerebro humano. Estas redes neuronales buscan aprender a partir de grandes cantidades de datos, lo que mejora la automatización de tareas analíticas y físicas sin necesidad de intervención humana.

Antecedentes

Para *Machine Learning* o *Deep Learning*, es necesario disponer de un proceso previo de entrenamiento, el cual se basa en el uso de un conjunto de datos inicial para que los algoritmos generen decisiones de manera autónoma, para la ello se requiere de bases de datos que contengan la información necesaria para entrenar estos algoritmos. Sin embargo, esto puede ser problemático debido a que la recolección de información puede convertirse en un proceso limitante. Ya sea debido al costo monetario que representan los equipos de medición o a las zonas donde se lleva a cabo la recopilación de información, esto puede resultar en una base de datos con información reducida.

No obstante, también es posible obtener esta información mediante técnicas de generación de datos, como *Data Augmentation*. Esta técnica es fundamental para ampliar un conjunto de datos inicial, con el objetivo de aumentar artificialmente el tamaño de dicho grupo de datos y reducir la escasez de información. De esta manera, esta información se puede utilizar en modelos tradicionales de *Deep Learning* y en procesos de automatización.

Un ejemplo del uso de estos modelos de Inteligencia Artificial son las aplicaciones que se basan en la predicción de mediciones de la intensidad de señal recibida (RSSI, del inglés *Received Signal Strength Indicator*) en redes de área amplia de baja potencia, específicamente basadas en la tecnología SigFox.

En el trabajo (Chen et al., 2020) se describe que los sistemas de geoposicionamiento RSSI pierden eficiencia debido al entorno que los rodea, el resultado de varias trayectorias, entre otros factores, lo que da como consecuencia una precisión de ubicación deficiente, debido a esto se propone el uso de modelos de Inteligencia Artificial a través del diseño de algoritmos de localización RSSI a partir de una red neuronal para realizar la predicción de la ubicación de los nodos desconocidos. Se obtienen coordenadas predichas y se las comparan con las coordenadas reales por medio del error cuadrático medio (MSE, del

inglés *Mean Squared Error*), donde el modelo de posicionamiento posee un error menor a 1 metro y el error de posicionamiento de 12 nodos es 0,4436 m.

De igual forma se aplica *Machine Learning* y *Deep Learning* en redes de baja potencia para la predicción de mediciones RSSI, ya sea mediante el uso de modelos de regresión o clasificación a su vez que se evalúa el desempeño de estos algoritmos por medio de sus valores generados y los valores reales.

En (Maduranga & Abeysekera, 2021) las mediciones RSSI se obtuvieron por medio de un sensor móvil que ocupa 32 ubicaciones diferentes y de sensores fijos, mismos que se almacenaron en un servidor que recopila 4520 valores de RSSI, estas mediciones se emplearon para preparar códigos de *Machine Learning* supervisados con modelos de regresión como regresor lineal, regresor polinomial, regresor por árbol de decisión (DTs, del inglés *Decision Trees*) y regresor de máquina de vectores de soporte (SVM, del inglés *Support Vector Machine*) para conocer la ubicación de aplicaciones de localización que involucran IoT, se efectuaron pruebas con diversos nodos entre 3, 4 y 5 donde los mejores resultados son entregados al aumentar el número de nodos, para este caso con 5 nodos. Los resultados al emplear 5 nodos describieron los resultados de la raíz del error cuadrático medio (RMSE, del inglés *Root Mean Squared Error*), el cual es de 77.55 RMSE para el regresor lineal, 65.93 RMSE para el regresor polinomial, 71.11 RMSE en el regresor SVM y el 27.31 RMSE para el regresor por DTs. Este trabajo define que el modelo más eficiente es el regresor por DTs en comparación de los otros algoritmos.

Mientras que en (Anjum et al., 2020) se describe la necesidad de mejorar las técnicas de localización en aplicaciones de ciudades inteligentes basadas en IoT, es fundamental contar con métodos precisos para localizar y rastrear dispositivos en entornos tanto interiores como exteriores, para ello se emplean algoritmos *Machine Learning* para mejorar la precisión en la localización basada en el RSSI. Se utilizó un dispositivo final que emite información durante un lapso de tiempo, luego se detiene y envía información a otro

dispositivo diferente, de tal manera que este proceso se emplea para la recopilación de datos RSSI. Para la evaluación de la precisión en localización se usó el RMSE a través de los nuevos datos generados, donde se obtuvo el 45.75 RMSE en el regresor lineal, 46.68 RMSE en el regresor polinomial, 40.44 RMSE para el regresor SVM y el 42.55 RMSE para el regresor por DTs. Se describe que el mejor modelo de regresión es el regresor SVM frente a los demás modelos.

Mientras que en (Le et al., 2021) se explica que los modelos de propagación de ondas emplean demasiado tiempo y recursos computacionales que están asociados a factores del entorno, por ello este trabajo busca modelos para predecir el RSSI con base en las coordenadas en un área determinada. Para ello, se realizaron pruebas iniciales para generar 10000 mediciones que incluyen latitud, longitud, distancia y RSSI dentro del rango de una estación base, se desarrollaron modelos tradicionales *Machine Learning* como regresor lineal, SVM y DTs. La eficacia de los diferentes modelos se los evaluó mediante el error cuadrático medio, donde se obtiene el 6.28 RMSE para el regresor lineal, 2.81 RMSE para el regresor SVM y 1.4448 RMSE para el regresor por DTs. Este trabajo describe que el mejor modelo es el regresor por DTs.

Por otro lado, en (Sutiyo et al., 2018) expresa que los modelos de localización actuales disponen de diferentes niveles de precisión y que solo se aplican en interiores, este trabajo propone diferentes modelos de regresión para la estimación de la distancia del sensor y el punto de acceso en exteriores. Dentro de los modelos de regresión se presentan los regresores lineal, exponencial y polinomial, que emplean una base de datos inicial que dispone con la intensidad de la señal recibida RSSI y la distancia entre el sensor y el *Access Point* con un rango de 0 a 100 metros para la recolección de estas mediciones. Para realizar la evaluación de los modelos se utilizó el error cuadrático medio donde el regresor lineal posee un 0.8133 RMSE, el regresor exponencial 0.8641 RMSE y el regresor

polinomial 0.9951 RMSE. A partir de estos resultados, el regresor polinomial es el modelo más preciso en relación con a los regresores lineal y exponencial.

De forma análoga a lo anterior (Takayama et al., 2018) propone métodos de posicionamiento en interiores que utilizan la recolección de huellas dactilares de ubicación compuesta en ambientes interiores para la localización de los sensores, este trabajo utiliza modelos de regresión como SVM y bosques aleatorios (RF, del inglés *Random Forests*). Se utilizó el error cuadrático medio para determinar la eficiencia en la predicción de posicionamiento de estos modelos, donde SVM tiene un error de 2.35 metros y RF un error de 0.80 metros. Con base en estos, el modelo más eficiente es RF al poseer un menor error de predicción.

En la investigación realizada por (Pimentel & Baldovino, 2022), detallan la falta de investigación en el área de las comunicaciones inalámbricas y el uso de modelos de regresión para mejorar la predicción en la ubicación de estos dispositivos en interiores. Este trabajo usa mediciones RSSI en función de parámetros independientes como interferencia, distancias desde el transmisor y tecnologías inalámbricas de localización de IoT, mediante modelos de regresión lineal múltiples, k vecinos más cercanos, DTs y RF en base de las métricas de desempeño como el error cuadrático medio y la desviación estándar donde el mejor modelo de regresión es el modelo por DTs.

Finalmente, en el trabajo de (Cruz & Amado, 2020) en el cual se desarrolló el estudio de modelos de clasificación mediante el uso de los indicadores de intensidad de señal RSSI para la localización inalámbrica en interiores e identificar la ubicación de un usuario al usar lecturas RSSI que recibe su teléfono inteligente. En esta investigación se puede observar la verificación de la precisión de cada modelo con clasificadores como clasificador SVM con 97,83%, clasificador de DTs con 97,67%, clasificador *Naive Bayes* con 98,50% y redes neuronales con 97,33% donde se observa que el mejor clasificador es *Naive Bayes* debido a su elevada precisión.

Justificación e Importancia

A medida que las organizaciones se embarcan en el proceso de la transformación digital, este adquiere un rol esencial en la sociedad al perfeccionar diferentes sectores, como la educación, la agricultura, las empresas, la robótica, entre otros. Para todo esto, es necesario el uso de equipos y algoritmos para tratar la información. Entre estos se destacan la Inteligencia Artificial, *Machine Learning*, *Deep Learning* y el IoT.

Es importante destacar que el avance continuo en la tecnología ha llevado a una mayor integración del *Machine Learning* en una amplia gama de aplicaciones (González, 2021), resalta cómo esta evolución ha redefinido la manera en que interactuamos con la tecnología en nuestra vida diaria, con cada vez más dispositivos conectados y una cantidad masiva de datos generados, el *Machine Learning* ha demostrado ser una herramienta invaluable para extraer información significativa y patrones de estos conjuntos de datos voluminosos, esto permite la predicción de valores o la toma de decisiones.

Por otro lado, la gran necesidad del uso de redes inalámbricas y sus diversos usos en aplicaciones como el estudio de cobertura, ha logrado un mayor enfoque en la obtención de la intensidad de señal para las comunicaciones inalámbricas. La zona de cobertura es un elemento importante para el diseño de las redes inalámbricas donde, se busca conseguir una medición de la intensidad de la señal y formar una agrupación de datos para la generación de mapas de cobertura, de esta manera se conoce las zonas de cobertura donde no se obtienen mediciones o se genera una mala conexión entre los dispositivos (Joubert & Helberg, 2015).

La recolección de datos es imprescindible en el estudio de los modelos de *Machine Learning* o en análisis de mapas de cobertura para la planificación de redes inalámbricas, ya que sin una basta cantidad de datos no es posible generar un proceso de entrenamiento adecuado para los modelos de regresión y clasificación dentro de *Machine Learning* u

obtener una perspectiva confiable con la información que se posee dentro de estos mapas de cobertura.

Actualmente, la obtención de datos se convierte en un problema debido a la dificultad en la recolección de esta información. Esto puede deberse a la accesibilidad del medio donde se realiza la toma de datos o al elevado costo monetario de los equipos que se requieren para su medición (Miralbell, 2021). También se puede presentar la problemática de disponer de una base de datos inicial que no abarca todos los datos en sus variables. Por ejemplo, en el caso de estudio de este proyecto, se conocen las coordenadas, pero no se obtiene un valor RSSI. De esta forma, se reduce la utilidad de esta base de datos debido a la reducción en la información recolectada. Al momento de realizar los mapas de calor, se observa una baja resolución en las diferentes zonas o no se puede generar una predicción cercana al real debido a la escasez de información al momento de entrenar a los modelos de *Machine Learning.*

Para solventar esta problemática se puede usar *Data Augmentation*, para la generación de nueva información a partir de los datos recopilados, de esta forma se cubre la necesidad de la obtención de datos, los mismos que mejoran las predicciones que realizan los modelos de *Machine Learning* y aumentan la resolución de los mapas de cobertura.

La relevancia de este proyecto de investigación radica en su enfoque innovador para abordar una cuestión fundamental en el ámbito de la tecnología de la información y las comunicaciones. La generación de una base de datos aumentada mediante técnicas de *Data Augmentation* representa un paso crucial hacia la mejora de la calidad y precisión de los datos geoespaciales utilizados en diversos contextos. Al incorporar variables como latitud, longitud, altura, RSSI y LQI, este proyecto busca ampliar la riqueza informativa de la base de datos existente, lo que a su vez tiene el potencial de enriquecer significativamente la comprensión de la cobertura de señal y la calidad de enlace en el área del cantón Quito.

La elección de emplear modelos de regresión y clasificación de *Machine Learning* para llevar a cabo este proceso de generación de datos aumentados resalta la sofisticación y versatilidad de las herramientas tecnológicas disponibles en la actualidad. Estos modelos, al aprovechar patrones y relaciones complejas en los datos, tienen la capacidad de inferir y predecir con precisión nuevos valores, se contribuye así a la construcción de un conjunto de datos más completo y robusto. Además, la aplicación de estos nuevos datos en la mejora de la resolución de mapas de cobertura y en entornos de *Deep Learning* destaca la interconexión y complementariedad de diversas tecnologías emergentes.

Al lograr una mayor resolución en los mapas de cobertura, esta investigación abre la puerta a una toma de decisiones más informada y estratégica en lo que respecta a la expansión y optimización de las redes de comunicación en el cantón Quito. Por otro lado, la integración de los datos aumentados en aplicaciones de *Deep Learning* subraya la importancia de alimentar a estas redes neuronales con información diversa y enriquecida, lo que puede resultar en un aprendizaje más profundo y representaciones más precisas de los patrones subyacentes en los datos.

En este proyecto, se emplea un enfoque que utiliza mediciones RSSI. Estas mediciones son especialmente importantes porque se obtienen a través de los sensores utilizados para recopilar datos en el trabajo que sirve como base para este documento.

En última instancia, este proyecto de investigación no solo contribuye al avance tecnológico en términos de técnicas y aplicaciones, sino que también en su enfoque integral y su exploración de las sinergias entre diferentes áreas de la ciencia de datos y la Inteligencia Artificial, reflejan la dirección prometedora en la que se enfoca la investigación tecnológica actual.

Alcance

El presente proyecto tiene como alcance la generación de datos sintéticos a partir de una base de datos inicial que dispone de 5174 mediciones de una red SigFox en el cantón

Quito. Inicialmente, los datos, al ser analizados, presentan las siguientes variables: latitud, longitud, altura, RSSI y LQI, donde se presentan mediciones de RSSI faltantes en ciertas coordenadas; de esta manera, los valores que muestran esta escasez de medición se consideran datos nulos, con un total de 1608 datos, y los otros 3566 datos restantes poseen medición RSSI en sus respectivas coordenadas. En caso de presentarse datos nulos, se emplean modelos de regresión *Machine Learning* para generar una estimación de la medición RSSI de acuerdo a las variables de entrada: latitud, longitud, altura y LQI, con la posibilidad de presentar agrupaciones entre dichas variables para los modelos de regresión y así rellenar los datos faltantes mediante datos sintéticos. Dentro de los modelos tradicionales, se utilizan modelos de regresión lineal o modelos de regresión no lineal basados en *Machine Learning*, como regresores lineales, regresores polinomiales, regresores por SVM y regresores por DTs, para generar los valores RSSI faltantes. Estos valores se evalúan de acuerdo a su desempeño a lo largo del proyecto, mediante el error cuadrático medio (MSE, del inglés *Mean Squared Error*).

Una vez obtenidos los valores faltantes con medición RSSI para cambiar datos nulos de la base de datos inicial, se procede a la parte cualitativa. Para esto, se busca generar un algoritmo de clasificación para etiquetar los valores que dependen de las variables de entrada: LQI, latitud, longitud y altura, de esta manera se especifica la calidad de la señal para etiquetarlos como EXCELENTE, BUENO, PROMEDIO o malo de acuerdo a los modelos de clasificación basados en la base de datos original y a los rangos de los valores RSSI obtenidos en un trabajo anterior. Se realiza un balanceo de datos de acuerdo a la cantidad de etiquetas EXCELENTE, BUENO, PROMEDIO y malo, además se trata de evaluar el desempeño de cada escenario con balanceo de datos, se utiliza la matriz de confusión y sus métricas de desempeño como especificidad, sensibilidad, precisión y exactitud.

Después de conocer el desempeño de los modelos de regresión, se utiliza el modelo con el menor MSE para el proceso de *Data Augmentation* con el fin de generar datos sintéticos para completar la base de datos inicial y rellenar los valores nulos. Esto corresponde al regresor SVM. De manera similar, se elige el modelo de clasificación con el mejor desempeño para la generación de etiquetas LQI, que pertenece al clasificador por DTs.

Para la obtención de las nuevas coordenadas correspondientes al cantón Quito, se utiliza el software *Open Foris Collect* para generar 62942 datos nuevos con las variables de latitud, longitud y altura. Se genera la nueva base de datos ampliada y se procede a *usar Data Augmentation* para la generación de las mediciones RSSI y etiquetas LQI, que dependen de las variables de entrada latitud, longitud y altura para cada modelo de regresión y clasificación.

Posterior a la utilización de *Data Augmentation*, se procede a evaluar el desempeño de la nueva base de datos ampliada a través del uso de mapas de calor, se emplean las coordenadas y el valor del RSSI. Se compara la resolución de los mapas de calor con la base de datos inicial de 5174 datos y la nueva base de datos de 62942 datos. De manera similar, se realiza la comparación de las etiquetas LQI de la base de datos inicial y la base de datos ampliada, se ubican estas etiquetas sobre el cantón Quito, donde se utiliza el software *Google Earth*.

De esta forma se muestran de manera general las etiquetas de cada una de las clases, es decir el LQI, así como los mapas de calor o cobertura con los datos originales y de los datos donde se aplica *Data Augmentation,* mismos que se encuentran dentro del cantón Quito, así como en la Universidad de las Fuerzas Armadas ESPE, para cumplir con este alcance se presentan los siguientes objetivos.

Objetivos

Objetivo General

Generar datos sintéticos a partir de una base de datos de mediciones RSSI en una red SigFox utilizando *Data Augmentation.*

Objetivos Específicos

- Analizar el estado del arte relacionado a las técnicas tradicionales del aprendizaje estadístico.
- Estimar un valor adecuado del RSSI para rellenar los datos nulos con modelos de regresión.
- Desarrollar un clasificador basado en aprendizaje supervisado del LQI.
- Aplicar técnicas asociadas a *Data Augmentation* para la generación de datos sintéticos empleando modelos tradicionales de *Machine Learning*.
- Identificar rangos del LQI para etiquetar los datos en función del RSSI.
- Evaluar el desempeño del clasificador con la base de datos aumentada.

Descripción general del proyecto

El presente proyecto de investigación tiene una organización por capítulos, la cual se presenta de la siguiente manera:

El primer capítulo trata sobre el alcance del proyecto, así como sus antecedentes, su justificación e importancia, además de la descripción general del proyecto, y sus objetivos.

En el segundo capítulo se da a conocer el desarrollo del estado del arte, en donde se detallan conceptos importantes para este proyecto, conceptos como: Regresor en *Machine Learning* y sus modelos como lineal, exponencial, SVM y DTs. Además de la descripción de los modelos de clasificación, con el fin de desarrollar la metodología a seguir.

En el tercer capítulo se describe la metodología general del proyecto a realizar, tanto con los modelos de clasificación para la obtención del RSSI y regresión para la asignación de

una etiqueta LQI que depende de las variables de entrada a utilizar, además del proceso de *Data Augmentation* para la generación de una base de datos aumentada.

El capítulo cuarto trata sobre los análisis de resultados para cada modelo de regresión por medio del error cuadrático medio y clasificación de acuerdo a los parámetros de desempeños asociados a la matriz de confusión. Dentro de *Data Augmentation* se realiza una comparación entre la base de datos inicial y la base de datos aumentada a través de los mapas de cobertura y la asignación de etiquetas LQI dentro del cantón Quito.

En el último capítulo se da a conocer las conclusiones del proyecto de investigación, así como las recomendaciones, además se dan a conocer posibles propuestas para trabajos futuros relacionados con el *Machine Learning* y *Data Augmentation*.

Capítulo II

Marco Teórico

Red de baja potencia

Las redes de área amplia y de baja potencia (LPWAN, del inglés *Low Power Wide Area Networks*) son un estándar de comunicación inalámbrica que ofrece una amplia cobertura de transmisión, paquetes de datos de tamaño reducido, velocidades mínimas y una batería de larga duración. Estas tecnologías se han implementado y han demostrado un gran potencial para su uso generalizado en el IoT que requieren baja latencia y alta confiabilidad y la comunicación entre máquinas (M2M, del inglés *Machine to Machine*), especialmente en entornos con poca conectividad. Debido a esto, esta tecnología se puede aplicar en diversas áreas, como la vigilancia ambiental, ciudades inteligentes, servicios públicos, agricultura, atención médica, automatización industrial, seguimiento de activos, logística, transporte y muchos otros campos (Chaudhari & Zennaro, 2020).

En la última década han surgido nuevas tecnologías de largo alcance y bajo consumo energético como LoRa, SigFox y NB-IoT.

LoRaWAN

LoRaWAN (del inglés, *Long Range Wide Area Networks*) es el estándar LPWAN actualmente más utilizado, la capa física se basa en los fundamentos de LoRa (LoRa, del inglés *Long Range*) al ser la tecnología para transmitir la información, fue desarrollada por la compañía Cycleo y años después adquirida por la empresa Semtech. Esta tecnología de comunicación posee un alcance de cobertura de hasta 20 kilómetros, admite una conexión de millones de nodos y tiene una velocidad de datos de 50 kbps. Posee seis factores de dispersión, donde el factor de dispersión más alto admite el envío de información a una distancia más larga a su vez que reduce la velocidad de datos. Estas redes comprenden tres clases de dispositivos: la clase A que presenta un menor consumo energético porque la mayor parte de tiempo se encuentra en suspensión y se activa solo para enviar datos, la

clase B que extiende el tiempo de recepción en ventanas de tiempo adicional y la clase C que todo el tiempo está activo para la recepción de datos, excepto si se emiten datos (Patel et al., 2021).

SigFox

La tecnología de red SigFox permite establecer comunicación a largas distancias entre diferentes objetos y transceptores, lo que permite una cobertura amplia. SigFox ofrece una mejor penetración que las redes móviles tradicionales (GSM, del inglés *Global System For Mobile Comunication*), al mismo tiempo que garantiza alta confiabilidad y un consumo de energía considerablemente menor. Esta red opera en la banda de radiofrecuencia industrial, médico y científico (ISM, del inglés *Industrial Scientific and Medical*), sin requerir licencias específicas, aunque la frecuencia exacta puede variar según las regulaciones vigentes en cada país (Agha et al., 2016).

Una manera de establecer rangos RSSI es por medio de la distribución normal inversa, la media y la desviación estándar de los datos, además de asociar una etiqueta del indicador de calidad de enlace según su rango. Para estos rangos se obtiene la siguiente clasificación:

- -94 dBm ≤ RSSI ≤ -66 dBm con media -89 dBm equivale a *EXCELENTE*
- -115 dBm ≤ RSSI ≤ -95 dBm con media -107 dBm equivale a *BUENO*
- -127 dBm ≤ RSSI ≤ -116 dBm con media -120 dBm equivale a *PROMEDIO*
- -133 dBm ≤ RSSI ≤ -128 dBm con media -129 dBm equivale a *LIMITE*

Estos rangos son utilizados para realizar el etiquetado de los valores LQI en función del RSSI en este trabajo.

NB-IoT

El IoT de Banda Angosta (NB-IoT, del inglés *Narrow Band Internet of Things*) tiene como objetivo mantener un bajo consumo de energía, proporcionar un bajo costo, ser escalable,

admitir miles de conexiones en una zona, y tener un extenso rango de comunicación con baja velocidad de datos. Además, permite una carga útil de 1600 bytes por mensaje y es recomendado para aplicaciones con una latencia de hasta 10 segundos y baja velocidad. NB-IoT emplea el acceso múltiple por división de frecuencia de portadora única (FDMA, del inglés *Frequency Division Multiple Access*) y la modulación por desplazamiento de fase en cuadratura (QPSK, del inglés *Quadrature Phase-Shift Keying*) (Patel et al., 2021).

Intensidad de señal recibida o RSSI

El RSSI es un valor PROMEDIO que representa la potencia de la señal recibida por un equipo receptor o antena, se mide usualmente en milivatios de decibelios, una unidad del nivel que se emplea para expresar que una relación de potencia se expresa en decibelios con referencia a un milivatio (Rábanos & Riera, 2015).

Indicador de calidad del enlace o LQI

El LQI es un valor que representa la calidad de la señal en un enlace de comunicación entre un emisor y un receptor. Este indicador se determina en función de la intensidad de la señal y la relación señal-ruido (SNR, del inglés *Signal to Noise Ratio*), y está directamente relacionado con la probabilidad de que los datos se transmitan correctamente. Al considerar la relación entre la calidad del enlace y la intensidad de la señal recibida, es posible evaluar la calidad del enlace en función de la distancia (Dong et al., 2016).

En algunos estudios, los datos de LQI se utilizan para crear un mecanismo de detección de intrusiones externas que podrían originarse desde nodos, computadoras portátiles u otros dispositivos de alto riesgo. Además, los datos de RSSI y LQI se emplean para determinar el número de nodos activos en una red.

Recopilar datos de RSSI y LQI en topologías inalámbricas desconocidas implica registrar la intensidad de la señal recibida y la calidad del enlace en diferentes puntos de la red. Esto permite obtener información sobre la fortaleza de la señal y la calidad de los enlaces de

comunicación, lo que a su vez puede ayudar en el análisis y la optimización de la red inalámbrica (Wu et al., 2019).

Relación en dBm y el número de estaciones base

SigFox asigna el indicar de calidad de enlace en función de:

- RSSI.
- El número de estaciones base.
- Zona de redundancia del receptor.

De esta forma se obtienen diversos modos de trabajo basado en las configuraciones de radio, los rangos mostrados a continuación son solo de referencia y no son utilizados en este trabajo ya que están basados en el número de estaciones base que nos entrega SigFox.

Al emplear configuraciones de radio RC1, RC3, RC5, RC6 y RC7 se obtiene:

- -122 dBm < RSSI equivalente a EXCELENTE para 3 estaciones base
- -135 dBm < RSSI ≤ -122 dBm equivalente a BUENO para 3 estaciones base
- -122 dBm < RSSI equivalente a BUENO para 1 o 2 estaciones base
- -135 dBm < RSSI ≤ -122 dBm equivalente a PROMEDIO para 1 o 2 estaciones base
- RSSI ≤ -135 dBm equivalente a LIMITE para cualquier estación base

Por otro lado, al usar configuraciones de radio RC2 y RC4 se obtiene:

- -114 dBm < RSSI equivalente a EXCELENTE para 3 estaciones base
- -127 dBm < RSSI ≤ -114 dBm equivalente a BUENO para 3 estaciones base
- -114 dBm < RSSI equivalente a BUENO para 1 o 2 estaciones base
- -127 dBm < RSSI ≤ -114 dBm equivalente a PROMEDIO para 1 o 2 estaciones base

- RSSI ≤ -127 dBm equivalente a LIMITE para cualquier estación base

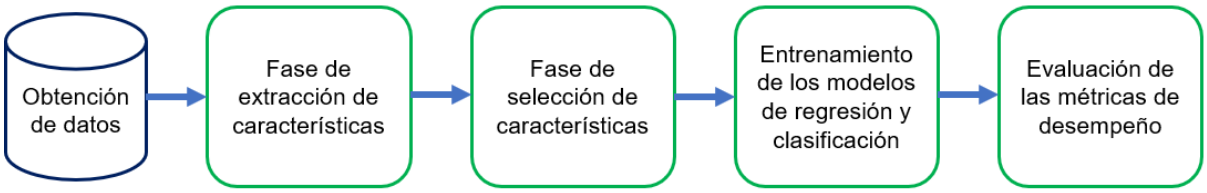
Además, la cantidad de estaciones base se encuentra asociado por la cifra de estaciones base a quienes les llegaron el mensaje y cumplen con los rangos RSSI (Sigfox support, n.d.).

Machine Learning

El *Machine Learning* es un campo derivado de la Inteligencia Artificial que permite que un sistema aprenda de datos de entrada en lugar de depender de una programación explícita por parte de los usuarios. Utiliza una variedad de algoritmos que aprenden iterativamente de los datos de entrada, lo que le permite realizar predicciones sobre resultados futuros de manera efectiva (Silva & Moya, 2019).

Figura 1

Diagrama de bloques Machine Learning



Nota. La Figura 1 ilustra el diagrama de bloques del proceso *Machine Learning* para los algoritmos de clasificación y regresión.

La Figura 1 describe el proceso general para la generación de modelos *Machine Learning*, mismo que dispone con la primera fase de la obtención de datos, es necesario adquirir una basta cantidad de datos para el entrenamiento de los modelos, la fase de extracción de características donde se identifican posibles elementos a ser descartados, la fase de selección de características se identifican las posibles variables de entrada de los modelos, la fase de entrenamiento donde se emplea una porción de los datos originales para entrenar los modelos y la fase de evaluación de los modelos donde se emplean los

datos restantes para la obtención de las métricas de desempeño, el regresor utiliza el MSE y el clasificador la matriz de confusión.

Entrenar algoritmos simples de *Machine Learning*

Los algoritmos de *Machine Learning* son secciones de código diseñadas para ayudar a los usuarios a explorar y analizar conjuntos de datos, que a menudo son complejos, con el objetivo de encontrar significado en ellos. Cada algoritmo consiste en un conjunto de instrucciones paso a paso que guían a una máquina para lograr un objetivo específico. Estos algoritmos se utilizan principalmente en modelos de *Machine Learning* para identificar y detectar patrones definidos por los usuarios, los cuales pueden ser utilizados para realizar predicciones o clasificar información (Microsoft, 2022).

Modelos de entrenamiento: preprocesamiento de datos

La preparación de datos es un proceso fundamental en la construcción, entrenamiento y prueba en modelos de *Machine Learning*. Consiste en realizar ajustes y transformaciones en los datos antes de utilizarlos en el entrenamiento del modelo. Es una tarea crucial para asegurar que el conjunto de datos sea adecuado y pueda ser utilizado de manera efectiva. Un problema común en este proceso es la mezcla de variables continuas y discretas, lo cual requiere una atención especial para manejar de forma adecuada esta combinación de tipos de datos (Tabladillo, 2023).

El método *Splitting* se basa en dividir el conjunto de datos original en dos subgrupos disjuntos: uno se utiliza para el entrenamiento del modelo y el otro se reserva para las pruebas y evaluación de diferentes modelos. La proporción entre estos grupos puede variar, pero es común utilizar alrededor del 70% de los datos para el entrenamiento y el 30% restante para las pruebas. También es aceptable utilizar una proporción de 80:20. Es importante seleccionar las muestras (los datos) de manera aleatoria para asegurar que ambos conjuntos sean representativos y evitar sesgos en la evaluación del modelo (Sanz & Rodríguez, n.d.).

Árbol de decisión o DTs

Los clasificadores de árboles de decisión son ampliamente reconocidos como uno de los métodos más populares para representar y clasificar datos. Investigadores de diversos campos, como el *Machine Learning*, el reconocimiento de patrones y la estadística, han abordado el problema de extender los árboles de decisión al utilizar los datos disponibles.

En áreas como el análisis de enfermedades médicas, la clasificación de textos, la clasificación de usuarios de teléfonos inteligentes, el procesamiento de imágenes, entre otros, se han propuesto diversas formas de emplear clasificadores de árboles de decisión. Estos clasificadores son técnicas comúnmente utilizadas en la minería de datos, ya que pueden manejar grandes volúmenes de información. Son capaces de realizar suposiciones sobre las clases categóricas, clasificar el conocimiento en función de conjuntos de entrenamiento y etiquetas de clase, y clasificar datos nuevos.

Los algoritmos de clasificación en el *Machine Learning* abarcan una variedad de técnicas, y en este trabajo se enfoca en el algoritmo del DTs en general como una de las opciones más utilizadas (Taha Jijo & Mohsin Abdulazeez, 2021).

La entropía mide la impureza o la aleatoriedad de un grupo de datos, posee un valor que varía entre 0 y 1 (Taha Jijo & Mohsin Abdulazeez, 2021). La ecuación 1 describe la entropía.

$$\text{Entropy(S)} = \sum_{i=1}^{c} P_i \log 2^{P_i}, \qquad (1)$$

donde P_i es la relación entre el número de muestra del subconjunto y el valor del atributo i-ésimo, y c es el número de estados.

La ganancia de información es una métrica utilizada para la segmentación y a menudo se denomina información mutua. Esto informa intuitivamente cuánto conocimiento se tiene del valor de una variable aleatoria. Es lo opuesto a la entropía, cuanto mayor sea su valor,

mejor (Taha Jijo & Mohsin Abdulazeez, 2021). La ecuación 2 describe la ganancia de información.

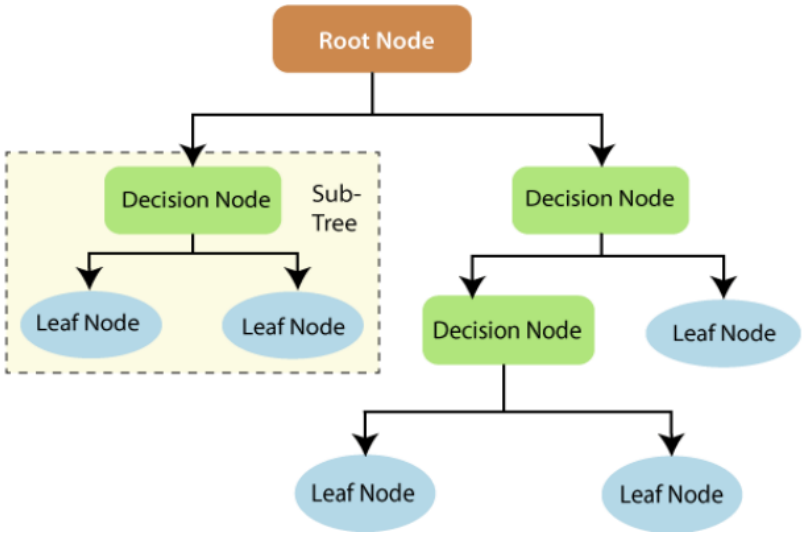
$$\text{Gain}(\text{S}, \text{A}) = \sum v \in v(A) \frac{|S_v|}{|S|} \text{Entropy}(S_v), \quad (2)$$

donde el rango del atributo A es $v(A)$, y S_v es un subconjunto del conjunto S igual al valor del atributo v.

La Figura 2 muestra la estructura de un DTs con sus diferentes nodos (Taha Jijo & Mohsin Abdulazeez, 2021).

Figura 2

Representación gráfica del DTs



Nota. La Figura 2 ilustra una estructura de un *Decision Tree* (DT). Tomado de (Taha Jijo & Mohsin Abdulazeez, 2021).

Máquinas de vectores soporte o SVM

SVM es un algoritmo de aprendizaje correspondiente a la categoría general de métodos de kernel, el cual puede ser tanto para clasificación (clasificación binaria y multi clasificación), como también para regresión; este método fue propuesto en 1992, en el

trabajo realizado por Boser, Guyon y Vapnik denominado: "*A Training Algorithm for Optimal Margin Classifiers*" (Boser Bernhard et al., 1992).

El uso de técnicas como SVM en el análisis de bases de datos permite abordar problemas de clasificación y regresión en conjuntos de datos complejos.

En el análisis de bases de datos, SVM puede ser utilizado para realizar tareas de clasificación, donde se asigna una etiqueta o categoría a cada registro de la base de datos. Por ejemplo, en un conjunto de datos de clientes de una empresa, SVM podría utilizarse para predecir si un cliente es propenso a abandonar la compañía o no.

Además, SVM puede aplicarse en problemas de regresión, donde se busca predecir un valor numérico en función de las características presentes en los datos. Por ejemplo, en el análisis de datos de precios de viviendas, SVM podría utilizarse para predecir el precio de una propiedad en función de sus características, como el número de habitaciones, la ubicación, etc.

La definición matemática del SVM no lineales, está asociada al tipo de núcleo (*kernel*) que se va a utilizar. La función de decisión se expresa como una suma ponderada de funciones *kernel* aplicadas a los vectores de soporte:

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i \times K(x, x_i), \quad (3)$$

donde:

- $f(x)$ es la función de decisión
- β_0 es el término de sesgo.
- α_i son los multiplicadores de Lagrange asociados a los vectores de soporte.
- $K(x, x_i)$ es la función *kernel* que mide la semejanza entre el vector de entrada x y un vector de soporte x_i.

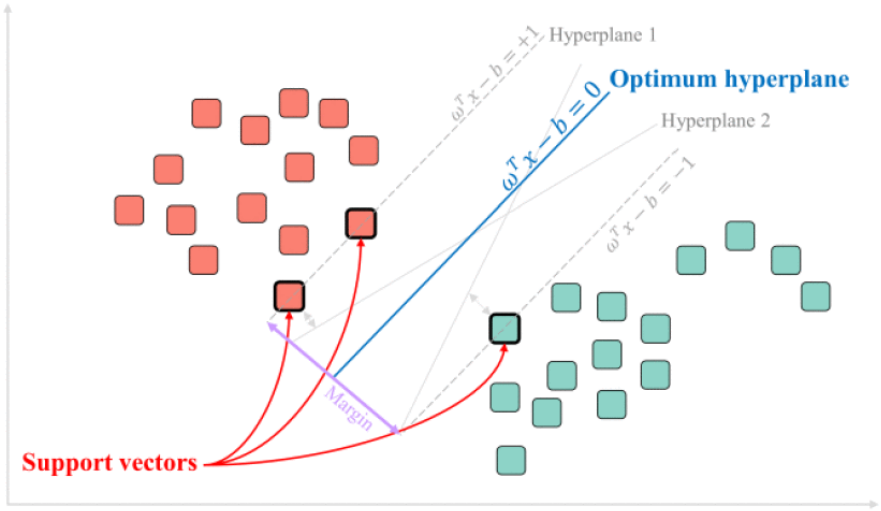
La elección del *kernel* ya sea lineal, polinomial o gaussiana asigna la estructura de la superficie de decisión. Es esencial encontrar los parámetros óptimos de β_0 y α_i para una adecuada separación de clases.

Además, los algoritmos SVM han sido implementados en numerosos campos de investigación, como la categorización de texto e hipertexto, la detección de pliegues de proteínas y homología remota, la clasificación de imágenes, la bioinformática (clasificación de proteínas y cáncer), el reconocimiento de caracteres escritos a mano, la detección de rostros, el control predictivo generalizado, entre otros (Cervantes et al., 2020).

La Figura 3 muestra el procedimiento que lleva a cabo el modelo SVM (Cervantes et al., 2020).

Figura 3

Ejemplo de SVM



Nota. En la Figura 3 se representa un ejemplo de SVM de datos linealmente separables. Tomado de (Sheykhmousa et al., 2020)

Capítulo III

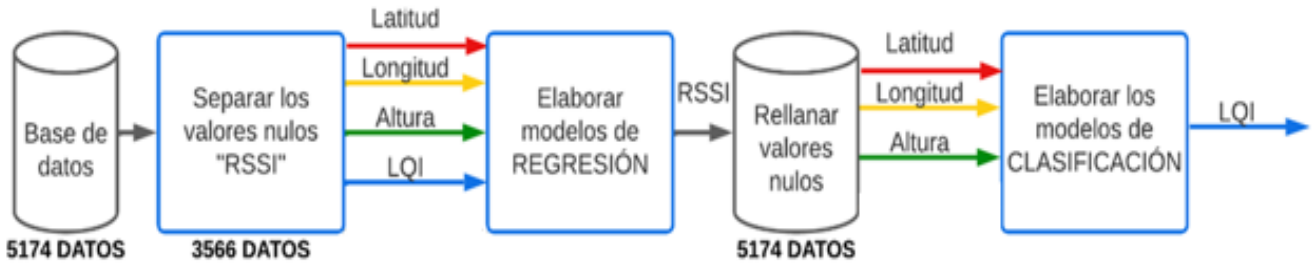
Metodología

En el presente capítulo se propone la metodología a ser utilizada en el presente trabajo de titulación, la misma que se aprecia en la Figura 4 y describe el diagrama de bloques para el proceso de regresión y clasificación de los diferentes modelos tradicionales de *Machine Learning,* se especifican las variables de entrada y las variables de salida para cada modelo, además de la cantidad de datos que se van a emplear en los modelos de regresión y clasificación.

Para ello es necesario la obtención de la base de datos inicial del trabajo previo denominado “Desarrollo de un Site Survey de la Red LPWAN Sigfox mediante un prototipo de Geoposicionamiento en el cantón Quito” donde se tienen las variables como latitud, longitud, altura, RSSI y LQI. Esta base de datos dispone de un total de 5174 datos, pero con la deficiencia de 1608 registros que no poseen medición RSSI, de esta forma se utilizan los 3566 registros restantes para el entrenamiento de los modelos *Machine Learning*.

Figura 4

Diagrama de bloques de los modelos de regresión y clasificación



Nota. La Figura 4 representa el diagrama de bloques de los modelos de regresión y clasificación.

Al conseguir la base de datos con mediciones RSSI, se empieza por el desarrollo de los algoritmos de regresión como el regresor lineal, polinomial, SVM y DTs por ser los modelos más populares dentro de *Machine Learning*, mismos que se emplean para la generación de

mediciones RSSI y dependen de las variables de entrada (latitud, longitud, altura y LQI) para el proceso de entramiento. Para la evaluación de estos modelos de regresión se utiliza el MSE, de esta manera se selecciona el modelo más eficiente para la generación de mediciones RSSI y rellenar los 1608 registros que no poseen esta variable.

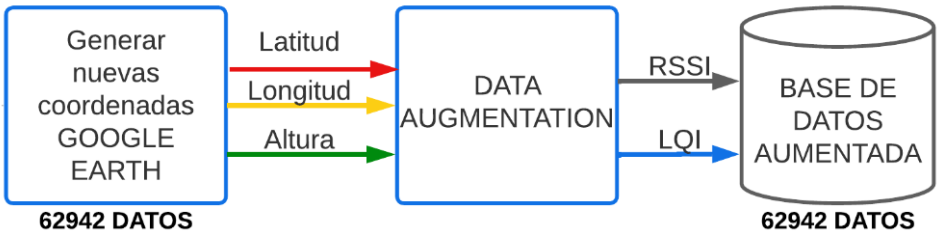
Después se elaboran los modelos de clasificación como el clasificador por SVM y DTs para la generación de una etiqueta LQI que varía entre EXCELENTE, BUENO, PROMEDIO y límite. Estos clasificadores emplean como variables de entrada la latitud, longitud y altura para el entrenamiento, posteriormente se emplean las métricas de desempeño de la matriz de confusión como precisión, exactitud, especificidad y sensibilidad para la selección del mejor modelo de clasificación para la asignación de las etiquetas LQI.

Además, se realiza la elaboración de otro clasificador diferente al de los modelos de clasificación *Machine Learning* para la generación de las mismas etiquetas LQI y realizar una comparación de predicción entre este clasificador y los clasificadores *Machine Learning.* Este clasificador se encuentra en función de los diferentes rangos RSSI donde cada rango tiene una etiqueta definida en EXCELENTE, BUENO, PROMEDIO y límite.

Posterior se generan datos nuevos por medio de *Data Augmentation*, donde es necesario la obtención de nuevas coordenadas que se encuentren dentro del cantón Quito. La Figura 5 describe este proceso al emplear el software *Open Foris* el mismo que permitió conseguir 62942 nuevas coordenadas (latitud, longitud), y además basándose en los mapas de Quito la altura. Al obtener estas coordenadas se emplean el mejor modelo de *Machine Learning* dentro de regresión para la obtención del RSSI y clasificación para la generación de etiquetas LQI, de esta forma se genera una base de datos aumentada con las variables latitud, longitud, altura, RSSI y LQI.

Figura 5

Diagrama de bloques del proceso de Data Augmentation



Nota. La Figura 5 representa el diagrama de bloques del proceso de *Data Augmentation.*

Finalmente se importa la base de datos inicial y la base de datos aumentada en MongoDB para visualizar los mapas de cobertura y realizar la comparación con los datos originales para verificar si existe una mejora en la resolución de estos mapas, además de realizar la visualización de los valores de LQI separados por clases (EXCELENTE, BUENO, PROMEDIO y malo) mismos que se encuentran distribuidos en todo el cantón Quito.

Desarrollo de los modelos de regresión

Al usar la base de datos inicial, se debe realizar la separación de los datos nulos correspondientes a un valor de 1608 datos, de la base original con 5174 datos, de los cuales se emplean 3566 datos para el entrenamiento de estos modelos de regresión. De esta manera se logra obtener un modelo capaz de generar mediciones RSSI de acuerdo a las variables de entrada, mismo que es utilizado en el proceso de *Data Augmentation*.

En primera instancia se realiza los modelos de regresión para este proyecto, de los cuales se proponen variar el número de variables de entrada para verificar el comportamiento y eficiencia de dichos modelos. Debido a esto se presentan los siguientes escenarios:

1. Emplear 2 variables (latitud y longitud)
2. Emplear 3 variables (latitud, longitud y altura),
3. Finalmente emplear 4 variables (latitud, longitud, altura y el valor LQI)

Tabla 1

Base de datos original con medición RSSI y sin medición RSSI

Variables de entrada	Número de variables
latitud y longitud	2
latitud, longitud y altura	3
latitud, longitud, altura y LQI	4

Nota. Esta tabla muestra la cantidad de variables de entrada que se pueden emplear en los diferentes modelos de regresión

El LQI al ser un tipo de valor *categorical* no permite una utilización matemática clara similar a las demás variables, la solución a esto es asignar un valor numérico a cada etiqueta al asignar “4” a EXCELENTE, “3” a BUENO, “2” a PROMEDIO y “1” a límite conforme a la Tabla 2. De esta manera se puede trabajar con la variable LQI semejante a las variables de tipo numérico.

Tabla 2

Asignación numérica para cada etiqueta LQI

Etiqueta LQI	Valor numérico
EXCELENTE	4
BUENO	3
PROMEDIO	2
Límite	1

Nota. Esta tabla indica la disposición numérica que ocupan cada una de las etiquetas del LQI.

Dentro del uso de los modelos de *Machine Learning* es necesario emplear un porcentaje del grupo para el entrenamiento y el otro porcentaje restante para el proceso de validación

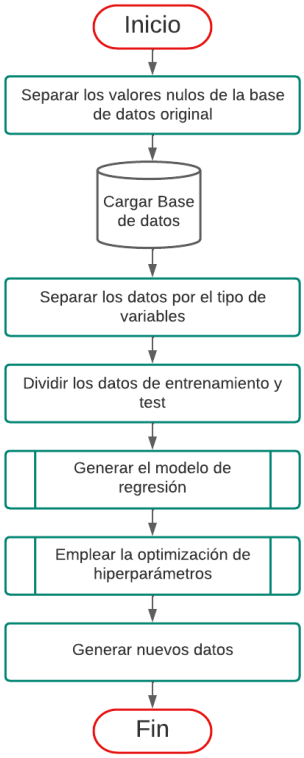
de los 3566 datos que cuentan con medición RSSI, para ello se realizan pruebas mediante la variación de estos porcentajes de la siguiente forma:

- 70% para entrenamiento (2496 datos) y 30% para validación (1070 datos).
- 60% para entrenamiento (2140 datos) y 40% para validación (1426 datos).
- 50% para entrenamiento (1783 datos) y 50% para validación (1783 datos).

El proceso para la generación de las mediciones RSSI en función de las variables de entrada (latitud, longitud, altura y LQI) puede conllevar varios pasos a seguir para la construcción del algoritmo requerido, debido a esto los diagramas de flujos son una manera de entender de manera gráfica y sencilla el progreso del mismo. La Figura 6 describe el diagrama con los diferentes pasos que se realizaron en los modelos de regresión.

Figura 6

Diagrama de flujo del modelo de regresión



Nota. La Figura 6 representa el diagrama de flujo requerido para los algoritmos de los modelos de regresión utilizados.

Al comenzar con el algoritmo se necesita cargar esta base de datos para luego separarla en las variables de entrada (latitud, longitud, altura y LQI) al ser los posibles casos expuestos en la Tabla 1, y la variable de salida el valor RSSI. Posteriormente se designan los grupos con las respectivas variables predictoras para el proceso de entrenamiento y validación.

Después de obtener el grupo de datos para el entrenamiento se debe utilizar esta información para el desarrollo de los diferentes modelos de regresión, este proyecto propone los siguientes modelos de regresión: regresor lineal, regresor polinomial, regresor SVM y regresor DTs, de esta manera se obtiene un modelo matemático para cada regresor en función de la matriz X de entrada (variables de entrada) y la variable de salida (mediciones RSSI).

Para obtener algoritmos eficientes se debe realizar la optimización de los hiperparámetros en los modelos de *Machine Learning* para no emplear los valores de los parámetros que son designados por defectos en las funciones de regresión y clasificación por el software Matlab.

Los hiperparámetros varían según los modelos a utilizar, para el caso de regresión se tienen: La función del regresor lineal *fitlm* que no posee parámetros por defecto, por lo que solo se designa las variables de entrada y las variables de salida. Para la función del regresor polinomial se puede realizar la variación del parámetro *polyijk* donde *ijk* representa el grado del polinomio, mismo que se les asigna a las variables de entrada y se realiza la variación del grado de estas variables hasta encontrar un modelo óptimo. La Tabla 3 muestra el grado que ocupa cada variable de entrada para obtener el mejor modelo posible con el menor error cuadrático medio.

Tabla 3

Grado del polinomio para cada variable de entrada

Variables de entrada	Grado del polinomio
latitud y longitud	latitud=7; longitud=2
latitud, longitud y altura	latitud=7; longitud=2; altura=1
latitud, longitud, altura y LQI	latitud=7; longitud=2; altura=1; LQI=2

Nota. Esta tabla muestra las variables de entrada del regresor polinomial y el grado de cada variable.

La función del regresor SVM *fitrsvm* se puede modificar los parámetros *BoxConstraint* que define la restricción de cuadro para los coeficientes alfa, *Standardize* que centra y escala cada columna de las variables predictoras según la media ponderada de la columna y la desviación estándar, *Epsilon* describe la mitad del ancho de la banda insensible a épsilon y *KernelScale* que divide todos los elementos de la matriz predictora por este valor, después se aplica la norma del núcleo adecuada para calcular la matriz de Gram. Para generar un modelo con el mejor desempeño posible, para ello se debe utilizar la optimización de hiperparámetros correspondiente a la siguiente función:

fitrsvm(X,datosTrainRSSI,'OptimizeHyperparameters','auto',
'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName', 'expected-improvement-plus')).

Finalmente, la función del regresor DTs *fitrtree* tiene como parámetro la variable *MinLeafSize* que especifica el número mínimo de observaciones de nodos de rama, la función a emplear para la optimización de los hiperparámetros es:

fitrtree(X,datosTrainRSSI,'OptimizeHyperparameters','auto',
'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName', 'expected-improvement-plus')).

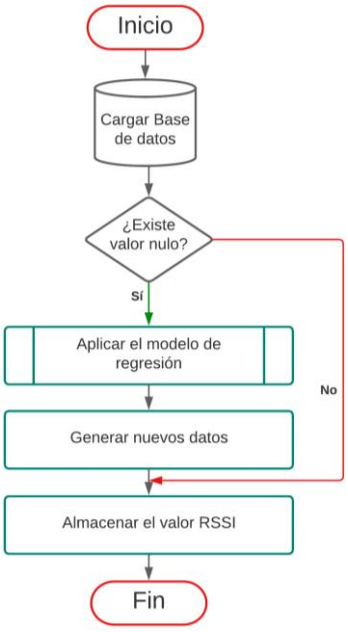
Desarrollo del algoritmo para el relleno de los datos nulos

Una vez obtenido el regresor que presente el menor error posible, es decir el regresor con el modelo SVM con tres variables y que se adapte de mejor manera a la base de datos, se cargan los datos que no contienen el valor de RSSI de los 5174 datos y se aplica el modelo.

En la Figura 7 indica el diagrama de flujo para el proceso de rellenar los datos nulos donde es necesario cargar la base de datos original y se procede a verificar en cada una de las filas con la variable RSSI si existe un valor nulo, de existir este caso aplica el modelo regresión para obtener un valor RSSI y posteriormente este valor se lo almacena en la misma fila.

Figura 7

Diagrama de flujo para rellenar los datos nulos



Nota. En la Figura 7 se describe el procedimiento a realizar para generar las mediciones RSSI en las coordenadas que no poseen esta medición.

En el caso de tener una fila con valor RSSI se debe mantener el mismo valor y continuar con la siguiente fila. Finalmente, al completar todas las filas con una medición RSSI se lo

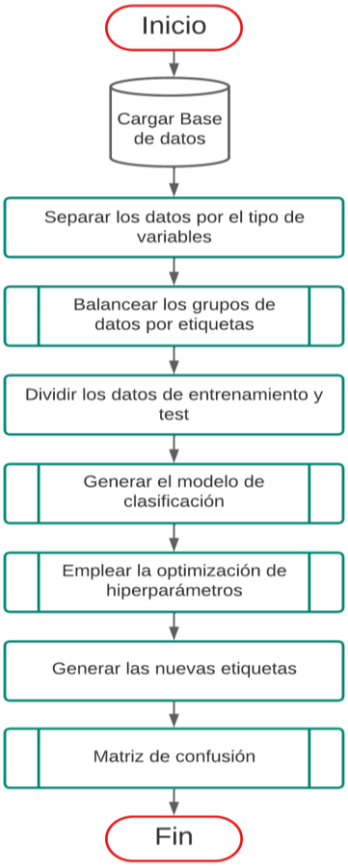
guarda en un archivo en formato Excel para acoplarlo a la base de datos inicial y obtener una base de datos completa al obtener un total de 5174 datos.

Desarrollo de los modelos clasificación

Para el desarrollo del clasificador se lo realizó al tomar como variables de entrada la latitud, longitud y altura, además que mientras más información se les entregue a los modelos de clasificación mejor será el resultado y el rendimiento de los modelos realizados.

Figura 8

Diagrama de flujo para el modelo de clasificación



Nota. En la Figura 8 se describe el diagrama de flujo para los modelos de clasificación.

La Figura 8 representa el procedimiento del algoritmo del modelo de clasificación empleado para generas etiquetas LQI y generar el mejor modelo de clasificación para cada caso.

Se comienza por cargar la base de datos con 5174 datos y se separa las variables a utilizar entre las variables de entrada (latitud, longitud y altura) y la variable de salida (LQI), de esta manera se obtienen las variables requeridas para el modelo de clasificación.

Antes de aplicar los modelos se realizó un procedimiento de balanceado de datos el cual consiste en separar los datos de cada clase, en este caso las clases LQI son 4: EXCELENTE, BUENO, PROMEDIO, LIMITE.

Al realizar un conteo de los valores de LQI dentro de los 5174 datos se tiene que existen:

- 1218 datos PROMEDIO
- 383 datos EXCELENTE
- 3330 datos BUENO
- 243 datos LIMITE

El procedimiento de balancear los datos consiste en separar primero los valores de la clase que tenga el número de datos más bajo (clase LIMITE) e igualar este número a las demás clases, es decir hasta que cada clase tenga los 243 primeros valores, una vez separados los datos se prueban en el modelo y se evalúan los resultados.

Luego de evaluar los resultados con el número de datos más bajo (clase LIMITE), se realiza el mismo procedimiento con la siguiente clase (clase EXCELENTE), es decir hasta que cada clase tengan los primeros 383 valores a excepción de la clase inferior a esta que solo cuenta con 243 datos, y se prueba con el modelo. Este procedimiento se repite hasta utilizar todas las clases del LQI y se obtienen diferentes resultados, el resultado con menor tasa de error es el utilizado para la realizar la clasificación.

Finalmente, se procede a realizar la generación de etiquetas LQI en función de las variables de entradas del modelo y se genera la matriz de confusión para obtener los parámetros de desempeño para cada clase y para cada modelo, se deben encontrar los

indicadores o las métricas de la matriz de confusión, estas se las encuentra de la siguiente manera:

La Exactitud

La Exactitud o *Accuracy* se refiere a lo cerca que está el resultado de una medición del valor verdadero.

$$A(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (4)$$

donde:

TP o Verdadero positivo: valor real positivo y la prueba predijo también que era positivo.

TN o Verdadero negativo: valor real negativo y la prueba predijo también que el resultado era negativo.

FN o Falso negativo: valor real positivo, y la prueba predijo que el resultado es negativo.

FP o Falso positivo: valor real negativo, y la prueba predijo que el resultado es positivo.

La Precisión

Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión.

$$P(\%) = \frac{TP}{TP + FP} \times 100 \quad (5)$$

La Sensibilidad o *Recall*

También se conoce como Tasa de Verdaderos Positivos (*True Positive Rate*) o TP. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

$$R(\%) = \frac{TP}{TP + FN} \times 100 \quad (6)$$

La Especificidad o Specificity

También conocida como la Tasa de Verdaderos Negativos (*True Negative Rate*) o TN. Se trata de los casos negativos que el algoritmo ha clasificado correctamente, además también expresa cuan bien puede el modelo detectar esa clase.

$$S(\%) = \frac{TN}{TN + FP} \times 100, \quad (7)$$

BER

La tasa de error balanceada (BER, del inglés *Balanced Error Rate*), se enfoca en cuántos eventos se clasifican incorrectamente de manera equitativa, se debe tener en cuenta tanto los errores de clasificación positiva como los errores de clasificación negativa. Un BER bajo indica un buen equilibrio entre ambas tasas y, por lo tanto, un buen rendimiento del clasificador.

$$BER(\%) = 1 - \left(\frac{R + S}{200}\right), \quad (8)$$

donde:

R representa la sensibilidad o *Recall*

S representa la especificidad o *Specificity*

Balanceado de Datos

Dado que se mantienen cantidades diferentes en el número de etiquetas LQI para cada clase de EXCELENTE, BUENO, PROMEDIO y límite, se puede presentar el inconveniente en la predicción de las clases minoritarias debido a la existencia de las clases predominantes. Por ello es necesario equilibrar la cantidad de etiquetas de cada clase mediante el balanceo de datos, de esta manera se puede comprobar la eficiencia de clasificación al variar la cantidad de etiquetas según la clase a tratar. Este procedimiento se lo realiza para los modelos de clasificación seleccionados, mismos que son el clasificador SVM y DTs.

Datos Balanceados a clase LIMITE

En este caso el valor más bajo en el número de datos de cada grupo es el de la clase LIMITE con 243 datos, por lo tanto, se igualan los primeros 243 datos de las demás clases (PROMEDIO, EXCELENTE y BUENO), y se prueba el modelo con estos valores. La Tabla 4 describe la cantidad total de datos que se utilizan para la generación del modelo de clasificación.

Tabla 4

Cantidad de datos balanceado en función de LIMITE

Etiquetas	Datos utilizados	Datos originales
LIMITE	243	243
EXCELENTE	243	383
PROMEDIO	243	1218
BUENO	243	3330

Nota. Esta tabla muestra la cantidad de datos utilizados en cada clase para el caso LIMITE.

Datos Balanceados a clase EXCELENTE

La siguiente clase cuenta con un número total de 383 datos, por lo tanto, se igualan las clases restantes a este valor a excepción de la clase anterior (clase LIMITE) en donde se utilizan todos sus datos y se prueba el modelo, de esa manera se obtiene la matriz de confusión para este caso.

La Tabla 5 describe la cantidad total de datos que se utilizan para la generación del modelo de clasificación para el caso *EXCELENTE* donde se emplean las 243 etiquetas de la clase LIMITE y 383 etiquetas para todas las demás clases.

Tabla 5

Cantidad de datos balanceado en función de EXCELENTE

Variables de entrada	Datos usados	Datos originales
LIMITE	243	243
EXCELENTE	383	383
PROMEDIO	383	1218
BUENO	383	3330

Nota. Esta tabla muestra la cantidad de datos utilizados en cada clase para el caso EXCELENTE

Datos Balanceados a clase PROMEDIO

Esta clase cuenta con un total de 1218 datos, semejante a los anteriores procesos se realiza el balanceado para esta clase al igualar los valores de las clases a este valor a excepción de las clases anteriores (clase LIMITE, y clase EXCELENTE), y de esta manera se obtiene la matriz de confusión para este caso.

La Tabla 6 indica la cantidad total de datos que se utilizan en cada clase para el desarrollo del modelo de clasificación en el caso PROMEDIO.

Tabla 6

Cantidad de datos balanceado en función de PROMEDIO

Variables de entrada	Datos usados	Datos originales
LIMITE	243	243
EXCELENTE	383	383
PROMEDIO	1218	1218
BUENO	1218	3330

Nota. Esta tabla muestra la cantidad de datos utilizados en cada clase para el caso PROMEDIO.

Datos Balanceados a clase BUENO

Para el último grupo de datos de la clase BUENO se basa en la utilización de todos los datos, es decir que no existe el proceso de balanceado en si para este caso, por lo tanto, se aplica el modelo para este caso.

La Tabla 7 muestra los siguientes valores para cada caso: 3330 datos BUENO, 1218 datos PROMEDIO, 383 datos EXCELENTE, y 243 datos LIMITE, finalmente se obtiene la matriz de confusión.

Tabla 7

Cantidad de datos balanceado en función de BUENO

Variables de entrada	Datos usados	Datos originales
LIMITE	243	243
EXCELENTE	383	383
PROMEDIO	1218	1218
BUENO	3330	3330

Nota. Esta tabla muestra la cantidad de datos utilizados en cada clase para el caso *BUENO.*

Modelo SVM para clasificación de datos

SVM es otro algoritmo de aprendizaje supervisado utilizado para la realización de la clasificación de datos utilizado en los 5174 datos para la obtención del valor LQI. De manera similar al caso de regresión se utiliza los mismos argumentos para realizar la optimización de hiperparámetros al modificar el valor de las variables de *Box Constraint, Kernel Scale, Standardize* y *Kernel Function*.

Modelo Decision tree DTs en clasificación de datos

Se debe seleccionar el argumento *'PruneCriterion','impurity'* para señalar el criterio de poda al utilizar la medida de impureza descrita por el criterio de división y el argumento

Prune para la poda del DTs, de esta manera de obtiene un modelo de clasificación con el mejor rendimiento posible.

***Data Augmentation* (DA)**

La aplicación de DA en la base de datos se la realizó con los algoritmos de aprendizaje supervisado de regresión y clasificación realizados previamente. Se considera que las variables de entrada para dichos algoritmos son tres (latitud, longitud, y altura) y además que se deben generar los valores de RSSI y los valores de LQI, al considerar estas coordenadas, se realizaron los siguientes procesos que se muestran a continuación:

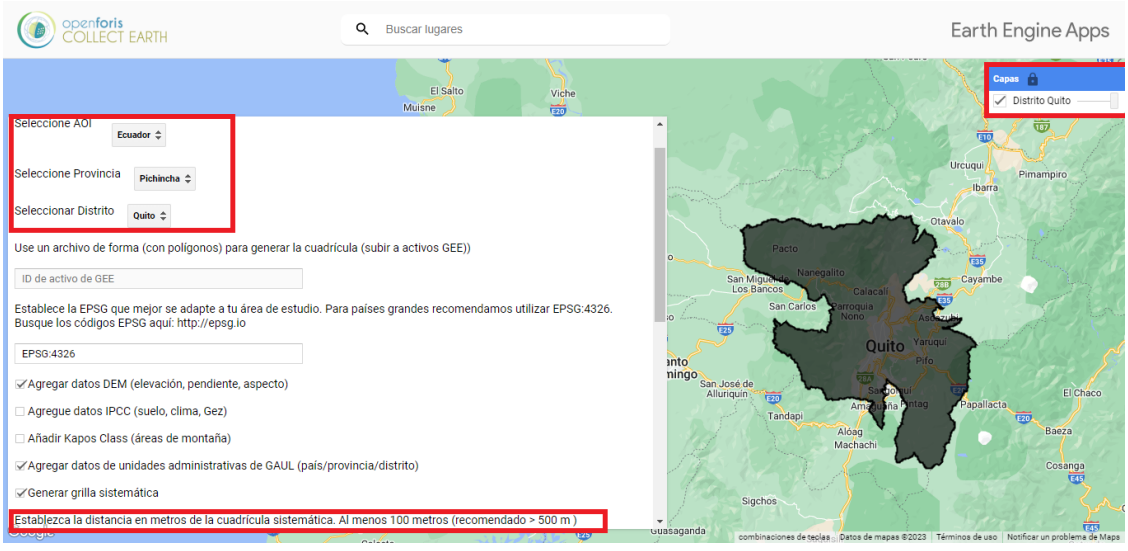
Creación de nuevas coordenadas

Para la creación de las nuevas coordenadas se utilizó la herramienta *Collect Earth*, que es parte de *Open Foris* la cual permite la recopilación de datos de *Google Earth* donde se puede obtener las variables latitud, altitud y altura.

En la Figura 9 se indica la división del Cantón Quito lo cual se puede observar y delimitar detalladamente los sectores a los cuales se deben enfocar.

Figura 9

Sectorización del cantón Quito o Distrito Quito



Nota. En la Figura 9 se observa el distrito Quito obtenido de *Collect Earth*.

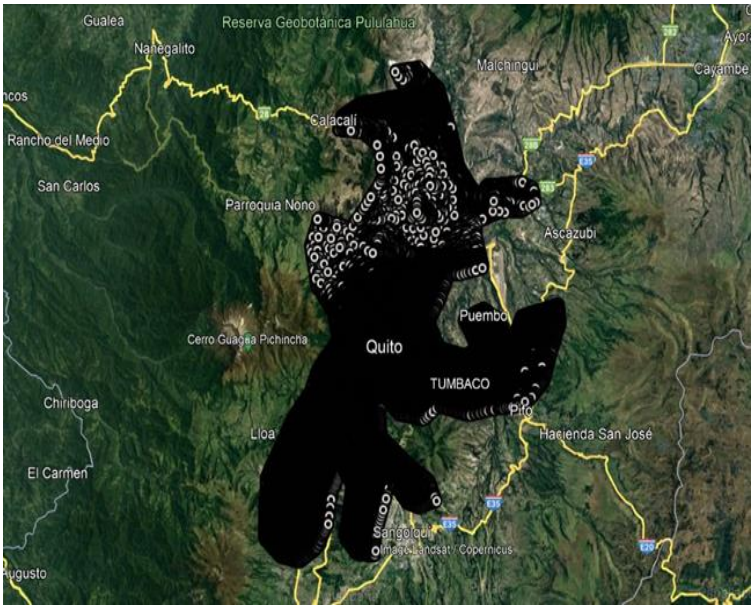
Luego de observar geográficamente como se encuentra dividido todo el distrito de Quito se realiza el trazado del contorno alrededor de las zonas donde se encuentra la mayor concentración de la población y evitar zonas verdes o áreas verdes donde no se puede evidenciar correctamente las zonas de cobertura para el caso de estudio.

Dentro de la herramienta *Collect Earth* se da la posibilidad de seleccionar la opción de “dibujar forma”, esto permite generar el contorno de las zonas urbanas y facilita la reducción del área del cantón Quito.

La Figura 10 indica el contorno de las zonas urbanas de Quito para la obtención de las coordenadas geográficas.

Figura 10

Contorno de las zonas urbanas del cantón Quito



Nota. En la Figura 10 se observa el contorno seleccionado para la obtención de las coordenadas realizado por la herramienta *Collect Earth* e importadas para su mejor visualización al software *Google Earth*.

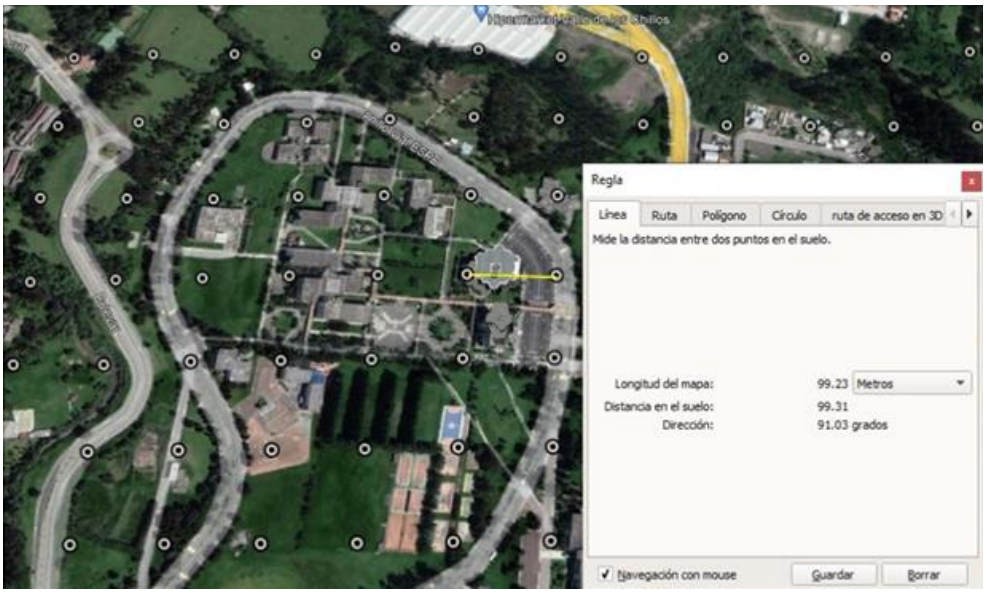
Una vez realizado el contorno de las zonas urbanas se procede a realizar una grilla con 100 metros de separación entre cada punto, de esta manera se obtiene un total de 62492

nuevos puntos que cubren todo el cantón Quito y parte de Sangolquí cantón Rumiñahui que es donde se encuentra la Universidad de las Fuerzas Armadas ESPE.

La Figura 11 indica la separación de 100 metros entre cada punto dentro de la zona seleccionada previamente en la herramienta *Collect Earth*, esta es la menor distancia de separación que puede proporcionar esta herramienta.

Figura 11

Grilla con 100 metros de separación entre cada punto



Nota. En la Figura 11 se observa la grilla realizada para la obtención de nuevos puntos donde se encuentran las nuevas coordenadas, obtenido de *Google Earth*.

El siguiente paso consiste en extraer las coordenadas de cada uno de los puntos mostrados, esto se lo realiza por parte de la herramienta *Collect Earth*, la misma que posee la opción de descargar la información en formato *.csv* los detalles donde la información requerida con la ayuda de esta herramienta es la latitud, la longitud, la elevación o altura y el distrito al que pertenece.

Es decir que al conocer las coordenadas de latitud y longitud se obtuvo también la altura para cada uno de estos puntos, esto es muy importante ya que la altura es una variable

adicional permite reducir el porcentaje de error al utilizar los modelos tradicionales que poseen los mejores resultados.

Utilización del algoritmo de regresión

Anteriormente este algoritmo permite obtener los valores de potencia RSSI faltantes para completar la base de 5174 datos, por lo tanto, se puede reutilizar este código para encontrar en base a las nuevas coordenadas los valores RSSI en función de la nueva base de 62942 datos.

El mejor modelo de aprendizaje supervisado encontrado para regresión y que se adapta de forma adecuada a los datos es el modelo de SVM, esto se evidencia en el capítulo de resultados, en donde se encuentra el error cuadrático medio MSE para cada uno de los regresores realizados.

Utilización del algoritmo de clasificación

El algoritmo de clasificación anteriormente sirvió para realizar un modelo de aprendizaje supervisado en base a los valores de LQI de la base de datos original que cuenta con 5174 datos.

El mejor modelo de aprendizaje supervisado es el modelo SVM para clasificación de acuerdo a los parámetros de desempeño otorgados por la matriz de confusión (precisión, exactitud, sensibilidad y especificidad) además de la obtención del BER, de igual manera los resultados a detalle de todos los modelos para clasificación se muestran en el capítulo de resultados. Este algoritmo tiene como fin reconocer en base a las coordenadas de un punto cual es el valor de LQI para ese caso, en otras palabras, si este punto se encuentra en la categoría EXCELENTE, BUENO, PROMEDIO, o LIMITE.

Se encontró cada valor de LQI para la nueva base de datos con 62942 valores nuevos, de esta manera se obtiene una base de datos aumentada o *Data Augmentation* misma que

se encuentra completa y en donde se aplicaron varias técnicas tradicionales de *Machine Learning* para su elaboración.

Utilización del valor de RSSI para la generación del LQI

También se realizó la obtención de las etiquetas LQI para cada una de las clases (LIMITE, PROMEDIO, BUENO y EXCELENTE) en función de la variable RSSI a diferencia de los modelos de clasificación que emplean como entrada las variables latitud, longitud y altura, de esta manera se puede obtener una comparación entre las etiquetas predichas con el modelo de clasificación tradicional *Machine Learning* y el modelo de clasificación en función de la variable RSSI.

Figura 12

Diagrama de flujo para el clasificador en función del RSSI

Nota. En la Figura 12 se observa el proceso y la asignación de toma el algoritmo para la asignación de las etiquetas LIMITE, PROMEDIO, BUENO y EXCELENTE en función del RSSI.

La Figura 12 indica el nombre de las etiquetas que se les asigna en función del RSSI, mismas fueron especificadas en el capítulo 2.

La Tabla 8 muestra los rangos pertinentes para la correcta asignación de cada etiqueta LQI, para ello es importante la obtención del valor RSSI para poder utilizar este modelo ya que es la variable de entrada y las etiquetas varían en función de la misma.

Tabla 8

Clasificación del LQI en función del RSSI

LQI	Rangos [dBm]
LIMITE	$[-94 \rightarrow -66]$
PROMEDIO	$[-115 \rightarrow -95]$
BUENO	$[-127 \rightarrow -116]$
EXCELENTE	$[-133 \rightarrow -128]$

Nota. Esta tabla muestra los rangos para cada etiqueta LQI en función de la variable RSSI.

Capítulo IV

Resultados

Regresor

Los resultados realizados para el regresor se presentan de acuerdo a valores de MSE y RSME que es básicamente comparar un valor predicho y un valor observado o conocido el mismo que se utiliza para el aprendizaje de máquina en este caso para encontrar los valores de potencia RSSI.

Estos valores permiten conocer que tan efectivo es el regresor que se ha realizado en función de los datos empleados, además podemos verificar si al aumentar el número de variables este valor aumenta o disminuye. Es importante aclarar que mientras el valor del MSE sea menor se obtendrá un modelo de predicción eficiente.

También es importante realizar la selección del porcentaje a utilizar para el entrenamiento y validación. Para ello, se varían los porcentajes descritos anteriormente en la metodología. De acuerdo a los resultados dentro de estas variaciones, se establece que el mejor porcentaje para el proceso de entrenamiento es el 70%, y el 30% para el proceso de validación de los modelos de *Machine Learning*.

Regresor dos variables

Se realizaron cuatro regresores Lineal, Polinomial, donde el regresor lineal no posee parámetros libres, el regresor polinomial posee parámetros libres de acuerdo al grado del polinomio (latitud: 7, longitud: 2).

Además, se aplicaron modelos tradicionales de *Machine Learning* como SVM, y DTs, para estos dos últimos modelos se debe utilizar la optimización de hiperparámetros y generar los valores para los parámetros libres, además de realizar un proceso de pruebas y errores se establecen los siguientes parámetros.

Parámetros libres en SVM:

- *Standardize: true*
- *KernelFunction: gaussian*
- *KernelScale:* 0.22068
- *Epsilon:* 0.012794
- *BoxConstraint:* 23.468

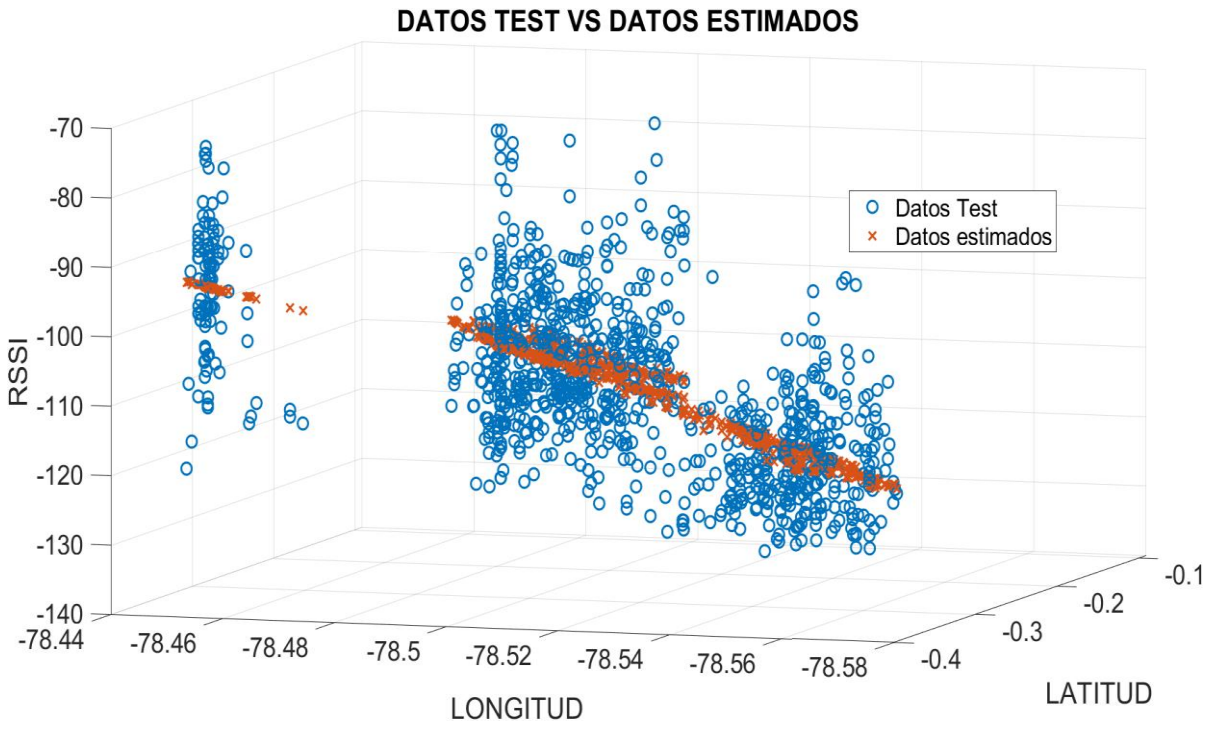
Parámetros libres en DTs:

- *MinLeafSize:* 11

Con estos parámetros libres se obtienen los siguientes resultados para cada modelo de regresión:

Figura 13

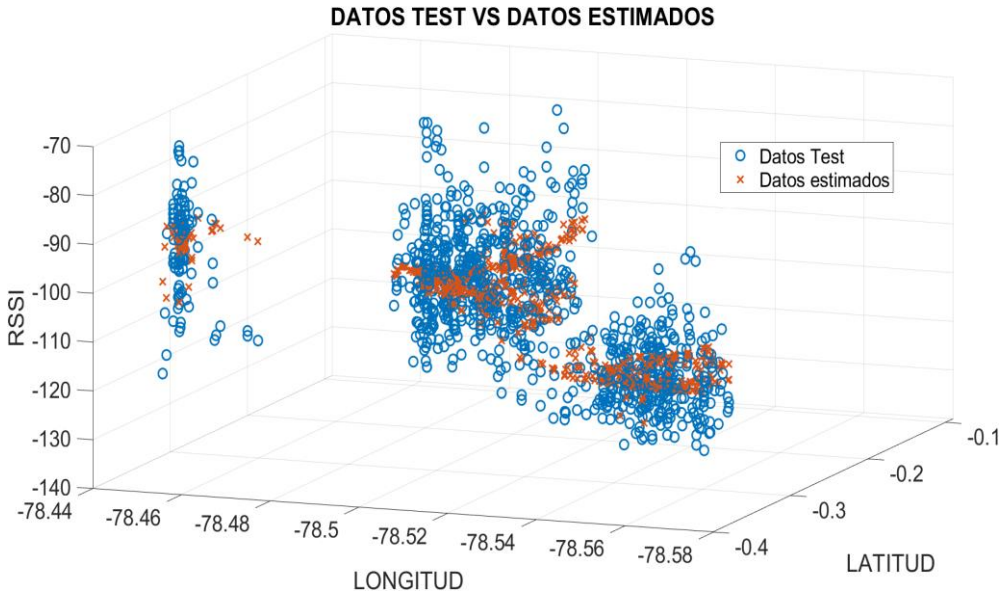
Regresor Lineal 2 variables (latitud y longitud)



Nota. En la Figura 13 se observa el regresor lineal realizado con 2 variables (latitud y longitud).

Figura 14

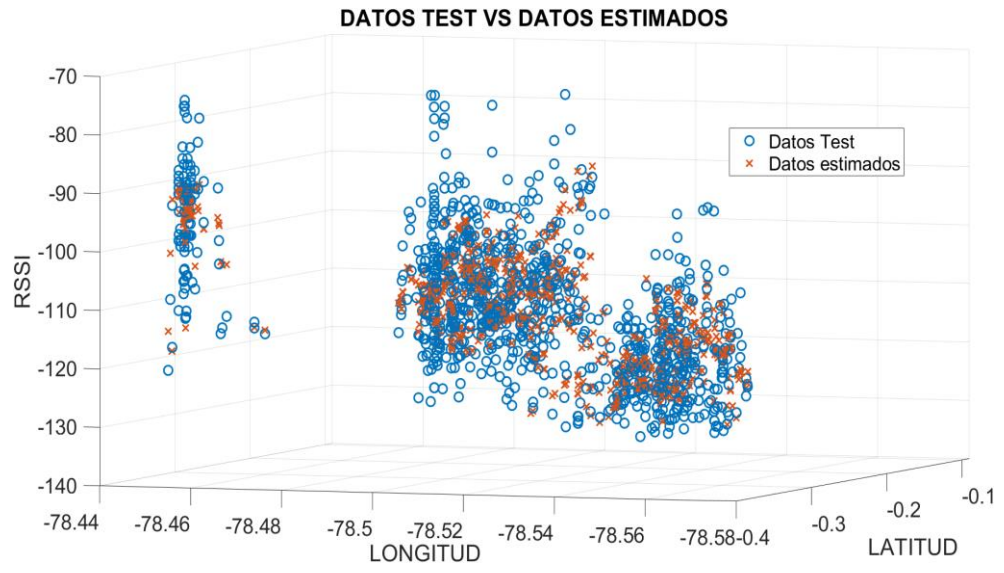
Regresor polinomial 2 variables



Nota. En la Figura 14 se observa el regresor polinomial realizado con 2 variables (latitud y longitud).

Figura 15

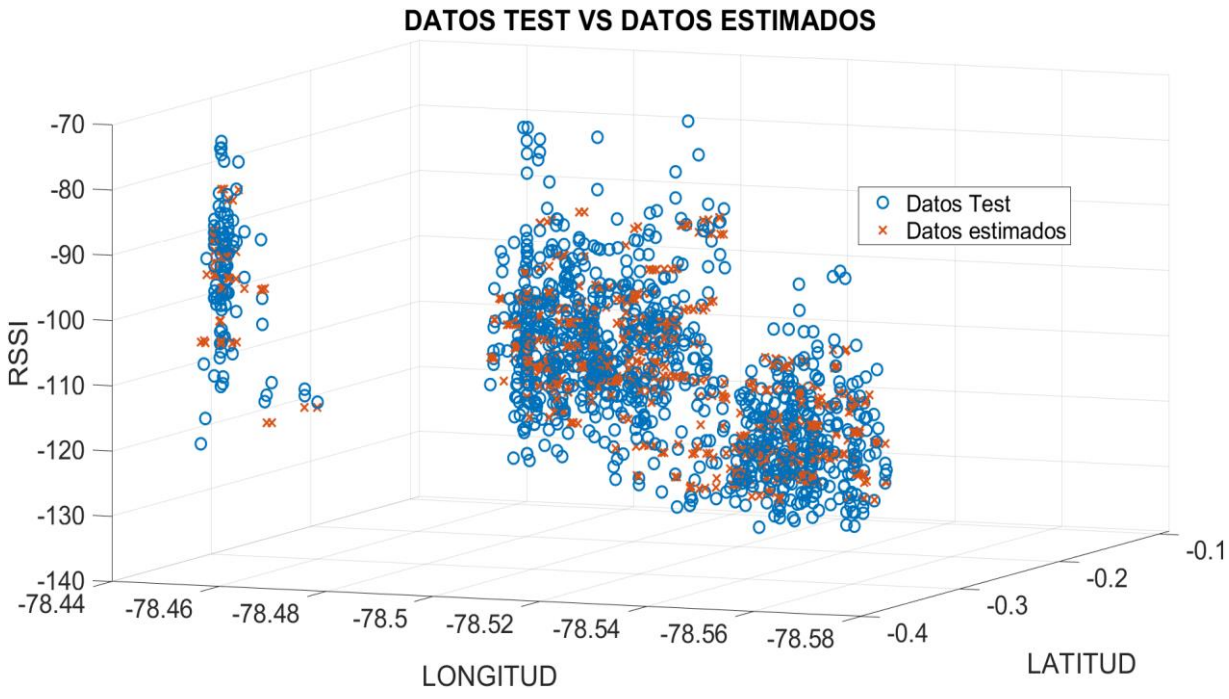
SVM de 2 variables



Nota. En la Figura 15 se observa el regresor modelo SVM realizado con 2 variables (latitud y longitud).

Figura 16

DTs 2 variables



Nota. En la Figura 16 se observa el modelo de DTs del regresor realizado con 2 variables (latitud y longitud).

Es importante mencionar que para cada regresor sin importar el número de variables se utilizó el mejor caso, por ejemplo, para el regresor polinomial se utilizó el grado con el que menor error se consiguió, de igual manera para los regresores de *Machine Learning* se consideró utilizar los mejores parámetros para conseguir el mejor rendimiento en cada modelo de aprendizaje.

La Tabla 9 muestra los resultados obtenidos por cada modelo de regresión en base a las dos variables de entradas utilizados (latitud y longitud), se puede evidenciar que el modelo lineal posee el error cuadrático más elevado debido a que gráficamente se pudo observar que los datos no poseen una tendencial lineal.

El modelo SVM tiene un valor menor de MSE de 53.8831 de esta manera SVM es el modelo con el mejor desempeño de los 4 algoritmos desarrollados.

Tabla 9

Resultados MSE y RMSE al usar 2 variables (latitud y longitud)

Regresor	MSE	RMSE
Lineal	85.7351	9.2593
Polinomial	69.4176	8.3317
SVM	53.8831	7.3008
DTs	57.2417	7.5658

Nota. Se observan los resultados de MSE y RSE de los regresores al usar 2 variables (latitud y longitud), de esta manera el modelo de SVM es el mejor con un MSE de 53.2417.

Regresor tres variables

Se añade una variable adicional para observar si al momento de aumentar más información a los distintos regresores y en especial a los modelos tradiciones de *Machine Learning* el error disminuiría.

Los parámetros libres para el regresor polinomial de acuerdo al grado de sus variables son: latitud=7, longitud=2 y altura=1. Para los otros modelos de regresión sus parámetros libres son:

Parámetros libres en SVM:

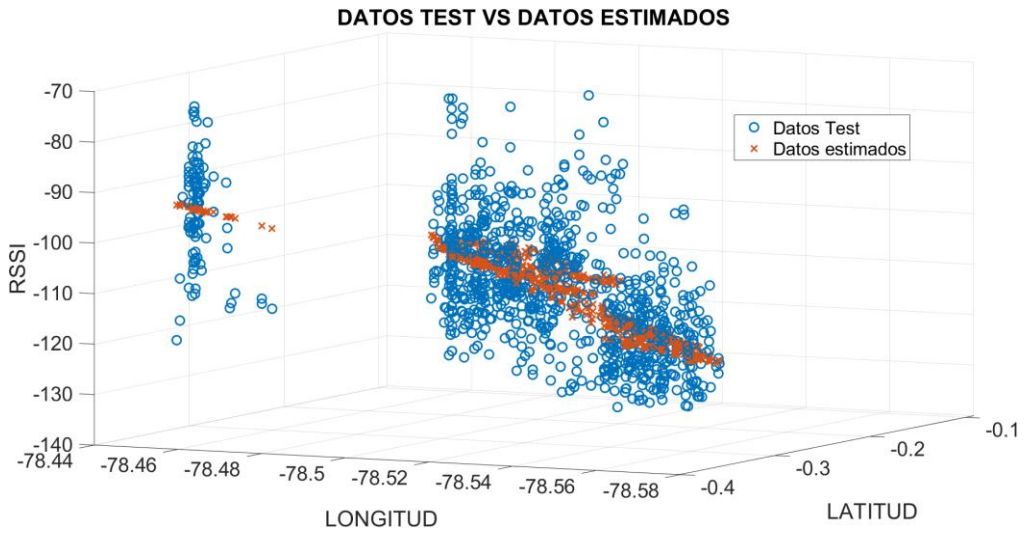
- *Standardize: true*
- *KernelFunction: gaussian*
- *KernelScale:* 0.21068
- *Epsilon:* 0.012794
- *BoxConstraint:* 23.468

Parámetros libres en DTs:

- *MinLeafSize:* 6

Figura 17

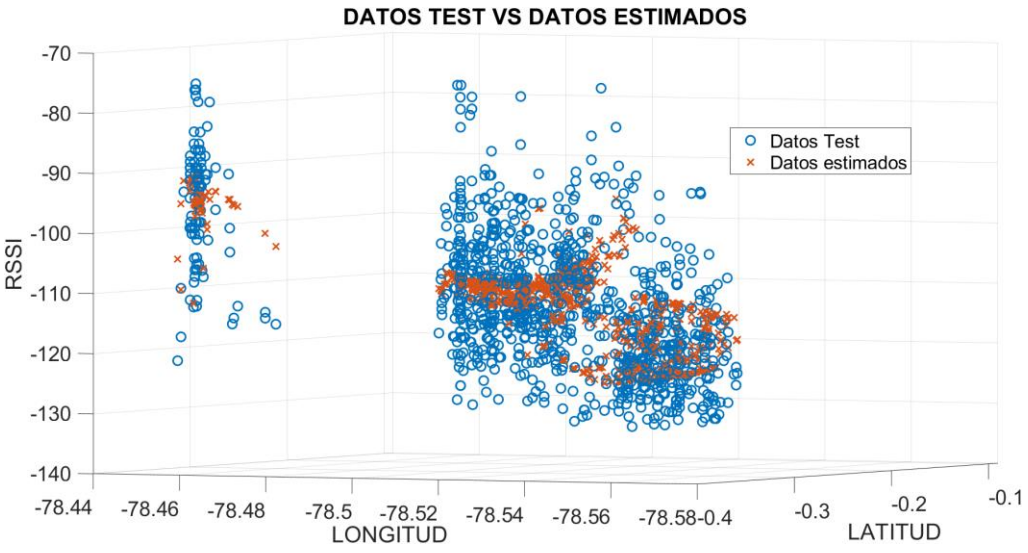
Regresor Lineal 3 variables



Nota. En la Figura 17 se observa el regresor lineal realizado con 3 variables (latitud, longitud y altura).

Figura 18

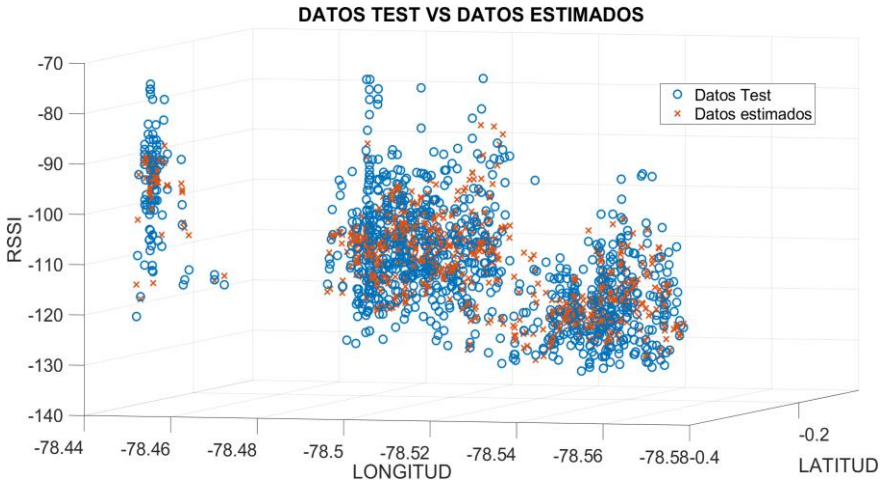
Regresor polinomial 3 variables



Nota. En la Figura 18 se observa el regresor polinomial realizado con 3 variables (latitud, longitud y altura).

Figura 19

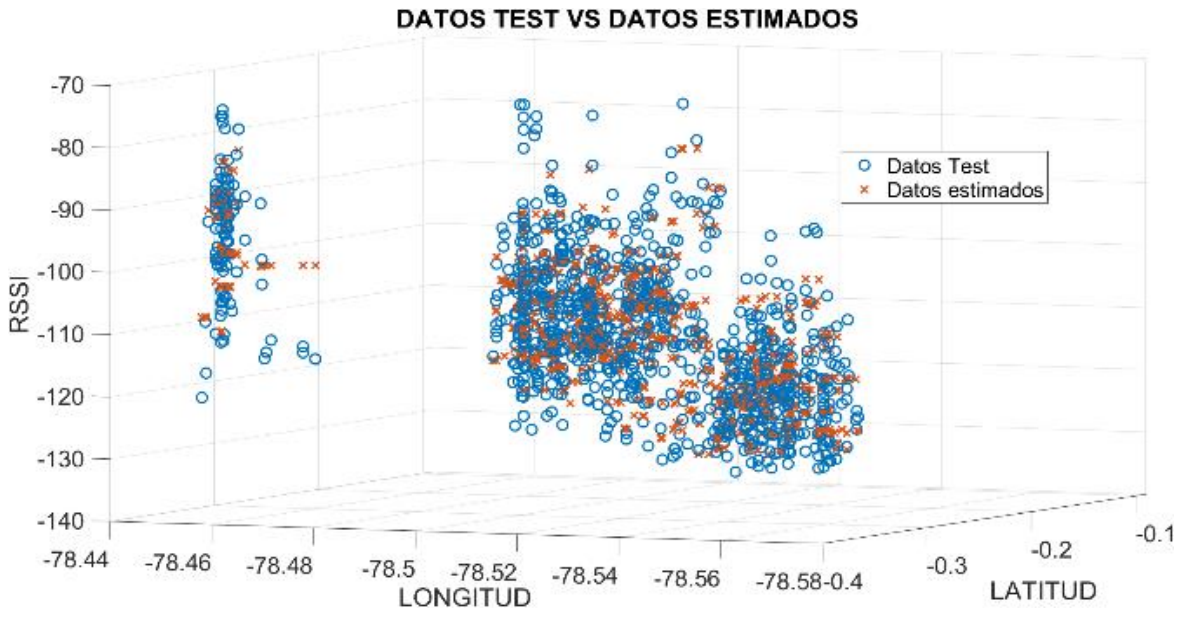
SVM de 3 variables



Nota. En la Figura 19 se observa el regresor modelo SVM realizado con 3 variables (latitud, longitud y altura).

Figura 20

DTs 3 variables



Nota. En la Figura 20 se observa el modelo de DTs del regresor realizado con 3 variables (latitud, longitud y altura).

La Tabla 10 describe los resultados generados por cada modelo de regresión en base a las tres variables de entradas (latitud, longitud y altura) donde el peor modelo de clasificación es el regresor lineal y el mejor es el SVM. Además, se puede decir que al aumentar otra variable a los modelos de regresión el error cuadrático medio decreció.

Tabla 10

Resultados MSE y RMSE al usar 3 variables (latitud, longitud y altura)

Regresor	MSE	RMSE
Lineal	85.6060	9.2524
Polinomial	66.2791	8.1412
SVM	53.7159	7.3291
DTs	56.7289	7.5319

Nota. Se observa que los resultados de MSE y RSE de los regresores al usar 3 variables (latitud, longitud y altura), disminuyeron los errores en todos los regresores, sin embargo, el modelo de SVM resulto el mejor con un 53.7159 de MSE.

Regresor cuatro variables

Finalmente se aumentó una variable adicional para observar cual es el rendimiento de los regresores al considerar el valor del LQI como un valor adicional, para esto se tuvo que modificar el valor *categorical* de las etiquetas LQI a un valor numérico y trabajar de acuerdo a esto.

Los parámetros libres para cada modelo de regresión son los siguientes:

Parámetros libres en el regresor polinomial:

- latitud=7; longitud=2; altura=1; LQI=2

Parámetros libres en SVM:

- *Standardize: true*

- *KernelFunction: gaussian*
- *KernelScale:* 0.21068
- *Epsilon:* 0.012794
- *BoxConstraint:* 23.468

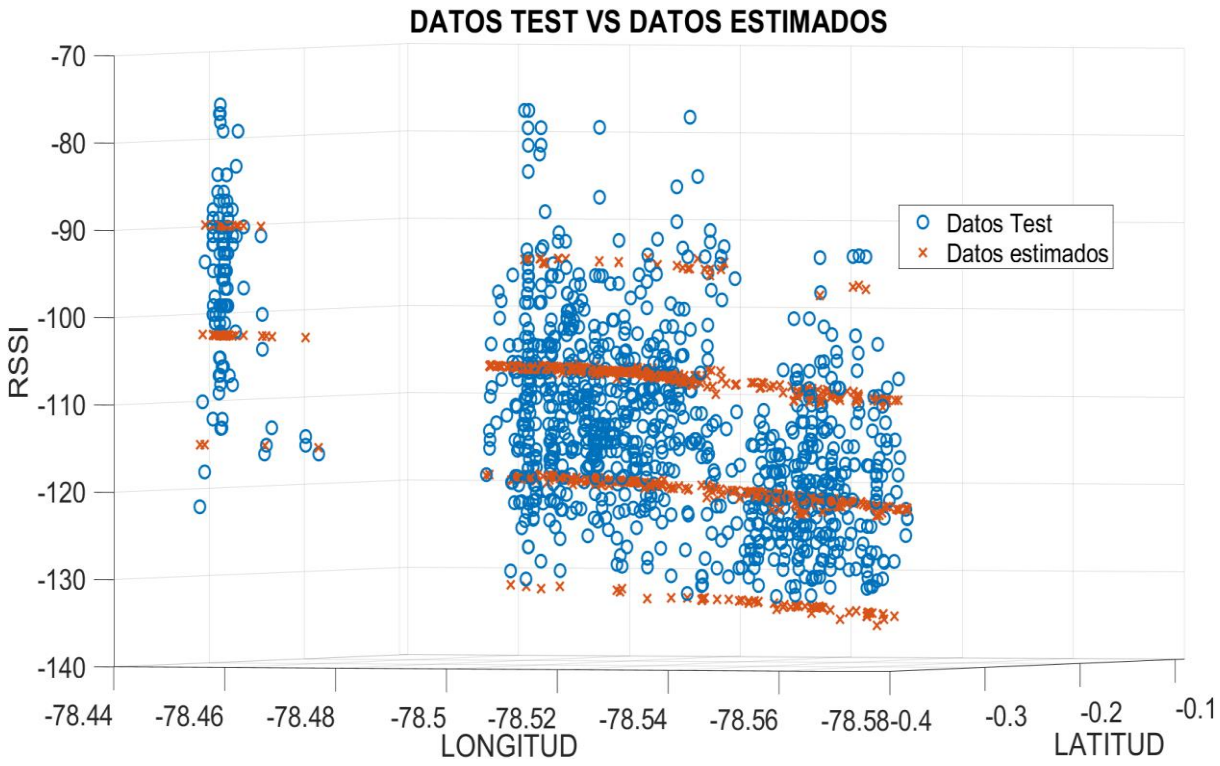
Parámetros libres en DTs:

- *MinLeafSize:* 7

A partir de estos valores se obtiene los siguientes resultados:

Figura 21

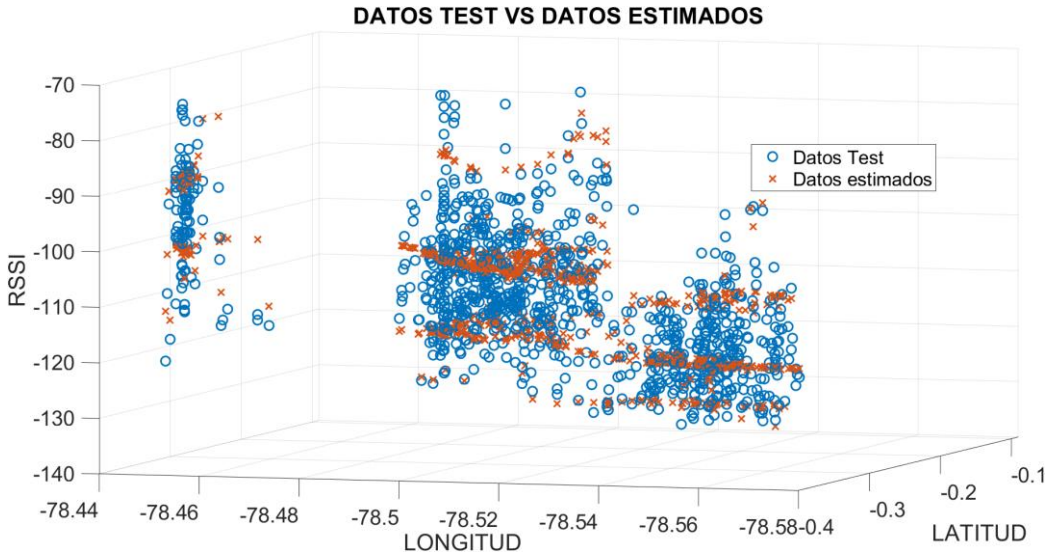
Regresor Lineal 4 variables



Nota. En la Figura 21 se observa el regresor lineal realizado con 4 variables (latitud, longitud, altura y LQI).

Figura 22

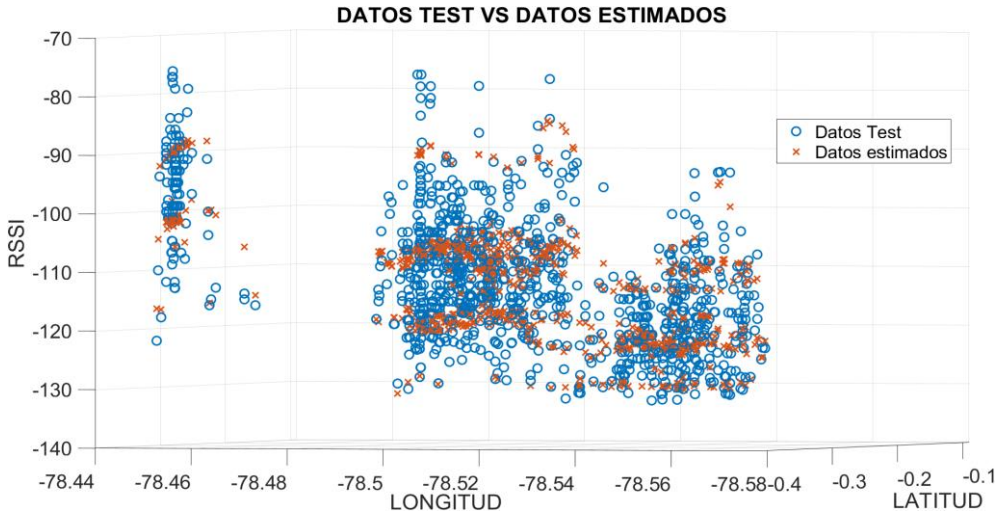
Regresor polinomial 4 variables



Nota. En la Figura 22 se observa el regresor polinomial realizado con 4 variables (latitud, longitud, altura y LQI).

Figura 23

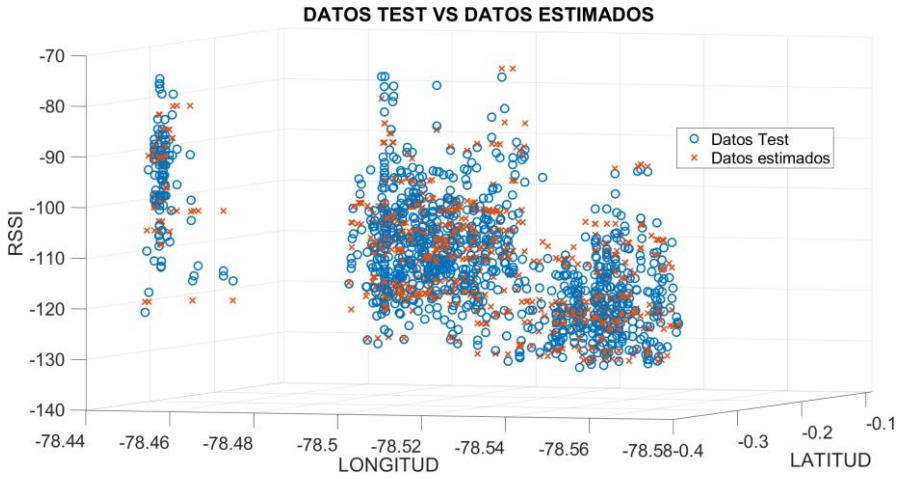
SVM de 4 variables



Nota. En la Figura 23 se observa el regresor modelo SVM realizado con 4 variables (latitud, longitud, altura y LQI).

Figura 24

DTs 4 variables



Nota. En la Figura 24 se observa el modelo de DTs del regresor realizado con 4 variables (latitud, longitud, altura y LQI).

Tabla 11

Resultados MSE y RMSE al usar 4 variables (latitud, longitud, altura y LQI)

Regresor	MSE	RMSE
Lineal	24.1198	4.9112
Polinomial	20.8077	4.5615
SVM	20.2045	4.4949
DTs	17.9020	4.2311

Nota. Se observa que los resultados de MSE y RSE de los regresores al usar 4 variables (latitud, longitud, altura y LQI), disminuyeron considerablemente en comparación con los regresores anteriores, en este caso el mejor modelo de regresor es el DTs con un valor MSE de 17.9020.

La Tabla 11 muestra los resultados obtenidos al utilizar un regresor de 4 variables (latitud, longitud, altura y LQI) donde el mejor modelo de regresión es el DTs al superar a los modelos SVM de los anteriores casos.

La Tabla 12 muestra todos los resultados para el caso de los regresores de 2 variables, 3 variables y 4 variables, se observa que al incrementar el número de variables en estos modelos el error cuadrático medio disminuye.

Tabla 12

Resultados MSE y RMSE para todas las variables de entrada

	2 variables		**3 variables**		**4 variables**	
Regresor	**MSE**	**RMSE**	**MSE**	**RMSE**	**MSE**	**RMSE**
Lineal	85.7351	9.2593	85.6060	9.2524	24.1198	4.9112
Polinomial	69.4176	8.3317	66.2791	8.1412	20.8077	4.5615
SVM	53.8831	7.3008	53.7159	7.3291	20.2045	4.4949
DTs	57.2417	7.5658	56.7289	7.5319	17.9020	4.2311

Nota. Se observa que los resultados de MSE y RSE de los regresores de 2 variables, 3 variables y 4 variables donde se obtienen resultados eficientes en el regresor con más variables predictoras.

Sin embargo, pese a que los resultados fueron satisfactorios en los modelos de regresión de 4 variables no se puede utilizar, esto es debido a que el valor LQI depende exclusivamente de los valores de potencia RSSI, por lo que el regresor quedaría limitado a contar con este valor de RSSI y generar las respectivas etiquetas LQI para poder ser aplicado en estos modelos, al momento de generar nuevas coordenadas solo se puede obtener las variables latitud, longitud y altura.

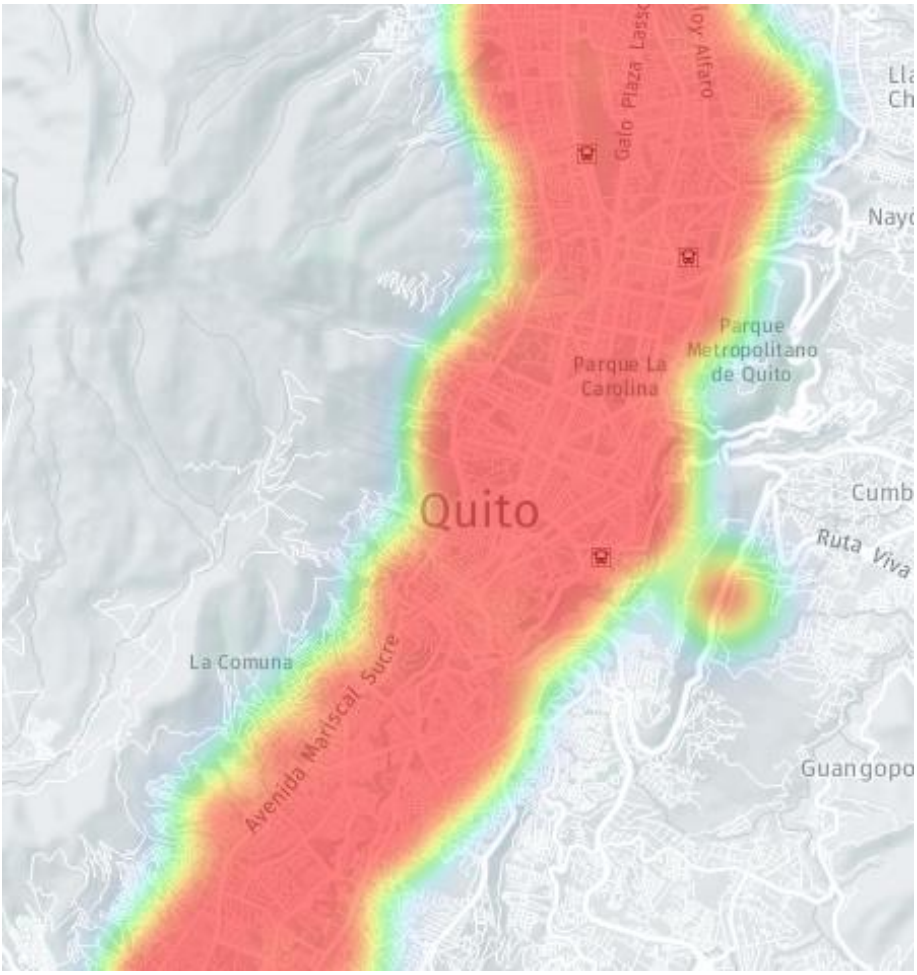
Debido a esta razón para el desarrollo de este proyecto de graduación se utilizó el regresor de tres variables al tomar en cuenta que el modelo SVM de *Machine Learning* mostró el mejor desempeño.

Mapas de calor con la base de datos original

Con el regresor antes mencionado (SVM de 3 variables) se obtuvieron los 1608 valores de RSSI faltantes de la base de datos original, de esta manera se completó esta base y se procedió a encontrar los mapas de calor resultantes que nos indican el área de cobertura de las mediciones de potencia dentro del cantón Quito.

Figura 25

Mapa de calor del cantón Quito



Nota. En la Figura 25 se muestra el mapa de calor del cantón Quito en donde se representa la intensidad de señal RSSI, realizado con 5174 datos. Obtenido de MongoDB.

Figura 26

Mapa de calor zona norte y zona sur del cantón Quito

a) b)

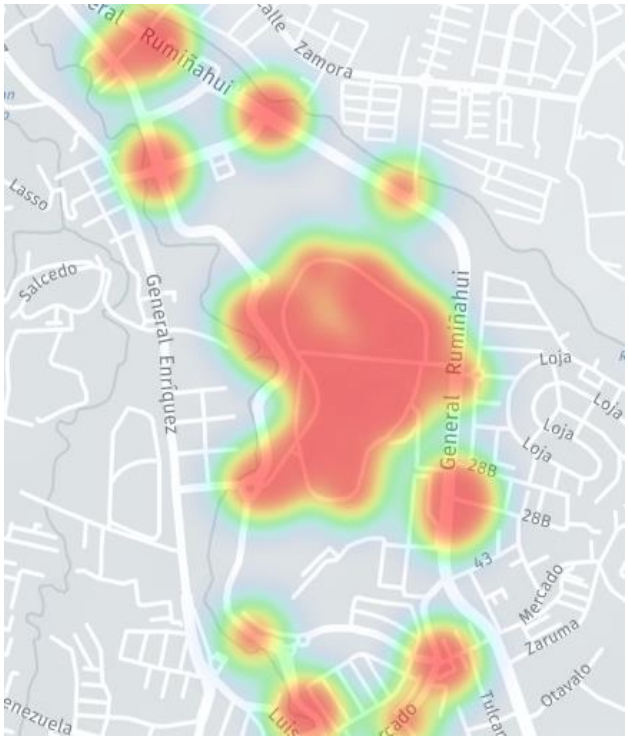
Nota. En la Figura 26 se muestra el mapa de calor de: a) la zona norte del cantón Quito y b) la zona sur del cantón Quito, realizado con 5174 datos.

Es importante mencionar que la base de datos cuenta también con coordenadas del cantón Rumiñahui que pertenecen a Sangolquí lugar donde se encuentra la Universidad de las Fuerzas Armadas ESPE.

La Figura 27 indica el mapa de cobertura con las mediciones realizadas en la Universidad de las Fuerzas Armadas ESPE donde se puede observar una baja resolución en los alrededores de la universidad y una mayor concentración en los niveles de señal en la parte central de la universidad.

Figura 27

Mapa de calor de la Universidad de las Fuerzas Armadas ESPE



Nota. En la Figura 27 se muestra el mapa de calor correspondiente a los valores de RSSI de las coordenadas tomadas en la Universidad de las Fuerzas Armadas ESPE.

Clasificador

Los resultados de los modelos de clasificación van a depender del modelo al que pertenece, al número de variables y a la clase que corresponda, esto se debe al procedimiento realizado llamado balanceado de datos.

Además, se debe realizar el mismo proceso de optimización de hiperparámetros para encontrar los mejores valores en los parámetros libres tanto en el clasificador SVM y DTs.

Parámetros libres en SVM:

- *Standardize: true*
- *KernelFunction: gaussian*
- *KernelScale:* 0.037684

- *Epsilon:* 0.012794
- *BoxConstraint:* 757.33

Parámetros libres en DTs:

- *MinLeafSize:* 7

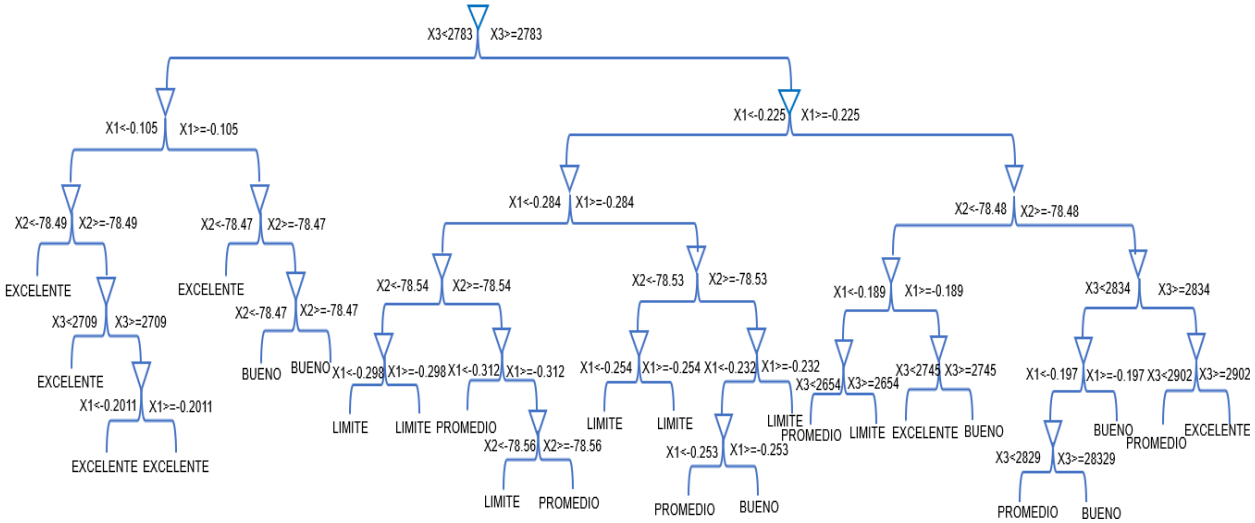
A continuación, se presentan los resultados correspondientes a los dos modelos de aprendizaje supervisado realizados.

Modelo de aprendizaje DTs

La Figura 28 describe un ejemplo de la estructura que toma el modelo *Decision Tree* al depender de las variables de entrada latitud (X1), longitud (X2) y altura (X3) donde se le asigna una etiqueta.

Figura 28

Modelo Decision tree DTs



Nota. En la Figura 28 se visualiza el modelo de DTs en clasificación de datos de manera general obtenido mediante MATLAB para cuatro clases EXCELENTE, BUENO, PROMEDIO, y LIMITE.

Clasificador al usar tres variables (latitud, longitud y altura)

Proceso: Balanceado al usar la clase LIMITE

Figura 29

Matriz de confusión clase LIMITE

Nota. En la Figura 29 se visualiza la matriz de confusión en donde se observa que 41 datos PROMEDIOS, 71 datos EXCELENTE, 46 datos BUENO, y 51 datos LIMITE son verdaderos positivos.

Tabla 13

Obtención de los parámetros de desempeño clase LIMITE modelo DTs

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	55.41	90.32	81.44	66.13	0.27
EXCELENTE	84.52	93.24	90.72	83.53	0.11
BUENO	67.65	86.99	82.47	61.33	0.23
LIMITE	78.46	92.04	89	73.91	0.15

BER TOTAL	=	0.19

Nota. Como resultado se observan las cuatro clases de la matriz de confusión del clasificador de datos, los resultados de los indicadores, el BER para cada caso y el BER general de toda la matriz de confusión.

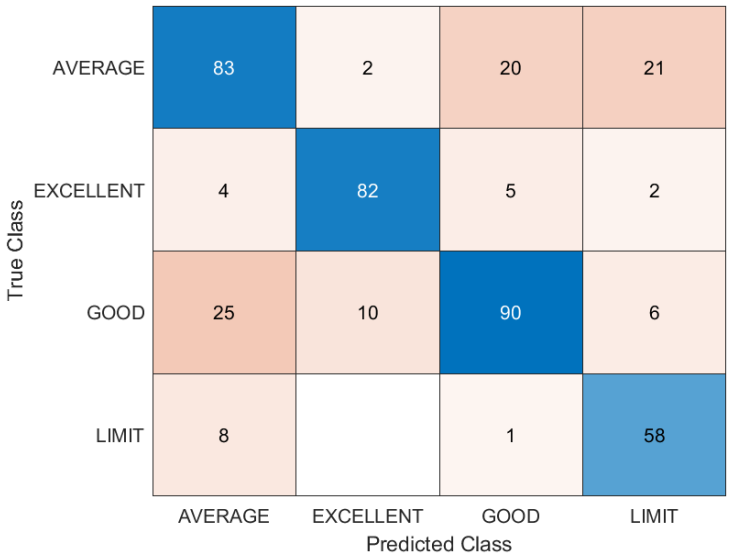
La Tabla 13 describe los resultados obtenidos en cada parámetro de desempeño al balancear la clase LIMITE.

En este caso el BER es de 0.19, en donde los resultados con mejor desempeño son los de la clase EXCELENTE con una sensibilidad de 84.52, una especificidad de 93.24, una exactitud de 90.72 y una precisión de 83.53, de igual forma el BER en esta clase es el más bajo de las demás con 0.11.

Proceso: Balanceado al usar la clase EXCELENTE

Figura 30

Matriz de confusión clase EXCELENTE



Nota. En la Figura 30 se visualiza la matriz de confusión en donde se observa que 83 datos PROMEDIOS, 82 datos EXCELENTE, 90 datos BUENO, y 58 datos LIMITE son verdaderos positivos.

En la Tabla 14 se observa que para este caso el BER es de 0.16, en donde los resultados con mejor desempeño son los de la clase EXCELENTE con una sensibilidad de 88.17, una especificidad de 96.3, una exactitud de 94.48 y una precisión de 87.23, de igual forma el BER en esta clase es el más bajo de las demás con 0.08.

Tabla 14

Obtención de los parámetros de desempeño clase EXCELENTE modelo DTs

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	65.87	87.29	80.82	69.17	0.23
EXCELENTE	88.17	96.3	94.48	87.23	0.08
BUENO	68.7	90.91	83.93	77.59	0.20
LIMITE	86.57	91.71	90.89	66.67	0.11

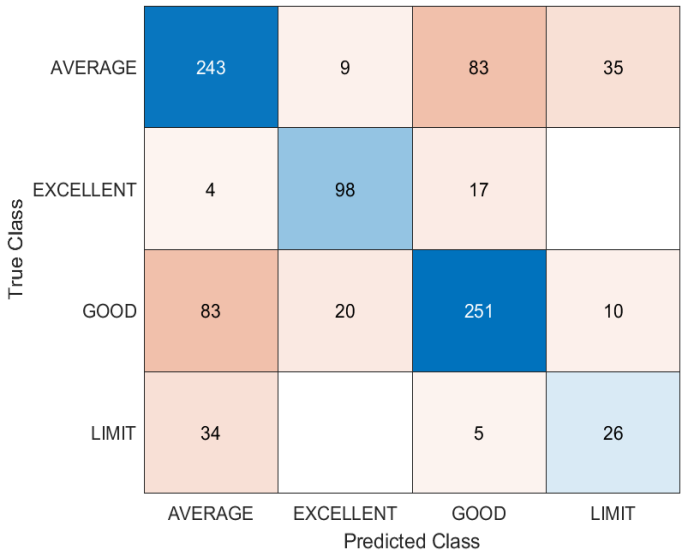
BER TOTAL	=	0.16

Nota. Como resultado se observan las cuatro clases de la matriz de confusión del clasificador de datos, los resultados de los indicadores, el BER para cada caso y el BER general de toda la matriz de confusión.

Proceso: Balanceado al usar la clase PROMEDIO

Figura 31

Matriz de confusión clase PROMEDIO



Nota. En la Figura 31 se visualiza la matriz de confusión en donde se observa que 243 datos PROMEDIO, 98 datos EXCELENTE, 251 datos BUENO, y 26 datos LIMITE son verdaderos positivos.

Tabla 15

Obtención de los parámetros de desempeño clase PROMEDIO modelo DTs

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	65.68	77.92	72.99	66.76	0.28
EXCELENTE	82.35	96.37	94.55	77.17	0.11
BUENO	69.96	81.05	76.25	70.51	0.25
LIMITE	40	94.73	90.85	36.62	0.33

BER TOTAL	=	0.24

Nota. Como resultado se observan las cuatro clases de la matriz de confusión del clasificador de datos, los resultados de los indicadores, el BER para cada caso y el BER general de toda la matriz de confusión.

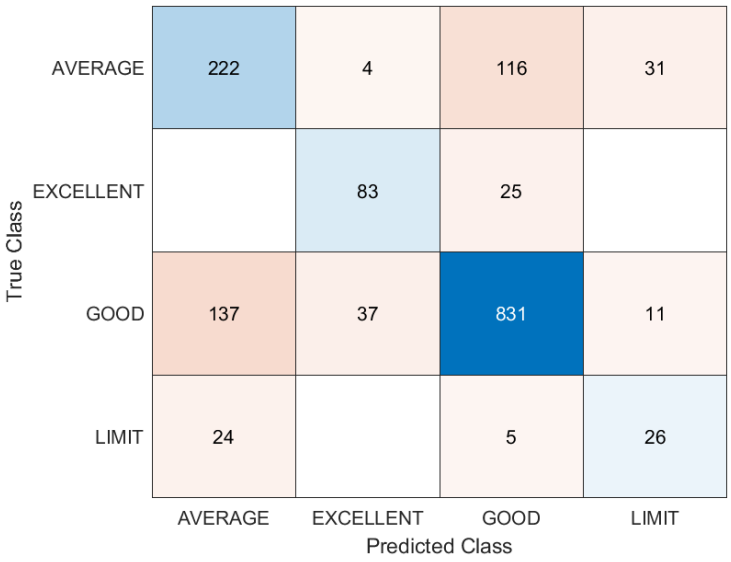
Para este caso el BER es de 0.24, este valor es mayor al de las anteriores clases, de igual forma la clase EXCELENTE con una sensibilidad de 82.35, una especificidad de 96.37, una exactitud de 94.55 y una precisión más baja de 77.17, es la más baja en comparación a las otras clases, el BER en esta clase es el más bajo de las demás con 0.11.

Proceso: Sin realizar balanceado de datos

Para el proceso sin balanceo se emplean todos los datos disponibles de las clases EXCELENTE, BUENO, PROMEDIO y LIMITE, de esta manera se procede a utilizar el modelo de clasificación para obtener los siguientes resultados tanto para la matriz de confusión y VER. Para este caso la Figura 32 describe los resultados de acuerdo a la clase *BUENO.*

Figura 32

Matriz de confusión clase BUENO



Nota. En la Figura 32 se visualiza la matriz de confusión en donde se observa que 222 datos PROMEDIO, 83 datos EXCELENTE, 831 datos BUENO, y 26 datos LIMITE son verdaderos positivos.

Tabla 16

Obtención de los parámetros de desempeño clase BUENO modelo DTs

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	59.52	86.34	79.9	57.96	0.27
EXCELENTE	76.85	97.16	95.75	66.94	0.13
BUENO	81.79	72.76	78.67	85.06	0.23
LIMITE	47.27	97.19	95.43	38.24	0.28

BER TOTAL	=	0.23

Nota. En este proceso no se realizó el balanceado, y de igual manera se observan las cuatro clases de la matriz de confusión del clasificador de datos, los resultados de los indicadores, el BER para cada caso y el BER general de toda la matriz de confusión.

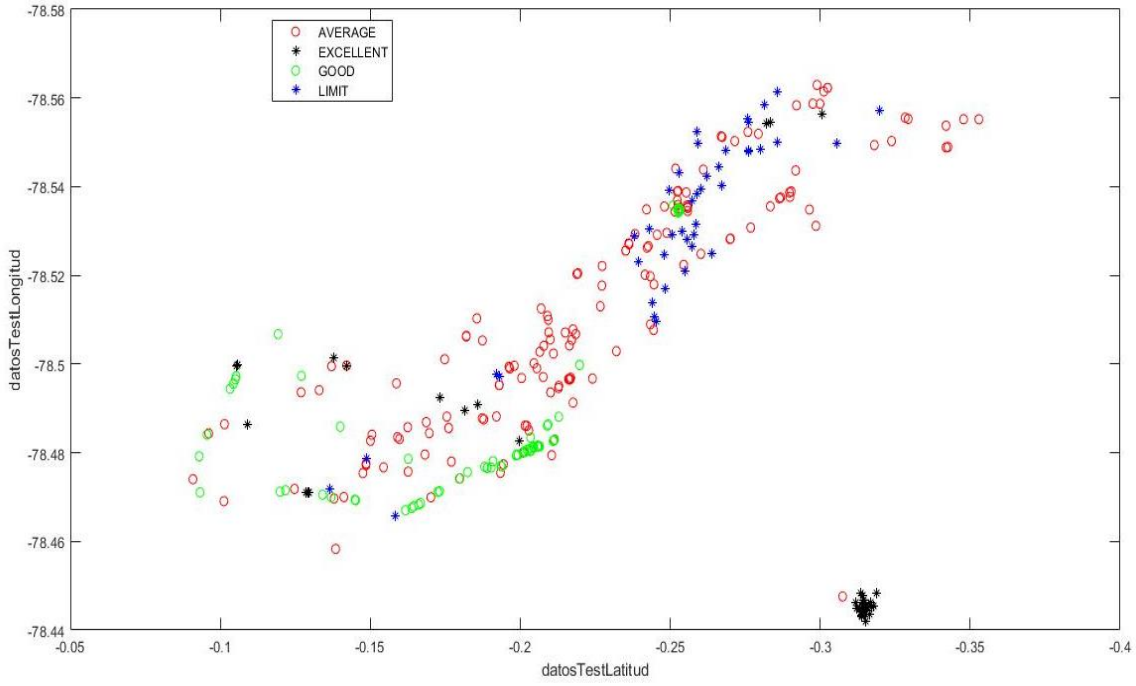
Para este caso el BER general es de 0.23 sin realizar el balanceo, lo que nos indica que es mejor que el proceso con balanceo de la clase PROMEDIO, pero no superior al proceso de balanceo con la clase EXCELENTE.

Modelo de aprendizaje SVM

La Figura 33 muestra el diagrama de dispersión por grupo generado el por el modelo de clasificación SVM.

Figura 33

Modelo SVM



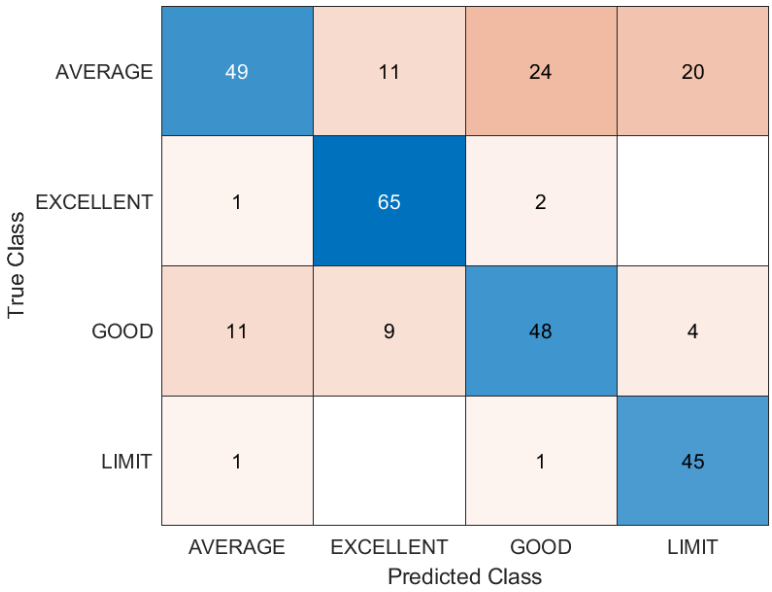
Nota. En la Figura 33 se visualiza el modelo SVM en clasificación de datos de manera general obtenido mediante MATLAB para cuatro clases EXCELENTE, BUENO, PROMEDIO, y LIMITE.

Clasificador al usar tres variables (latitud, longitud y altura)

Proceso: Balanceado al usar la clase LIMITE

Figura 34

Matriz de confusión clase LIMITE



Nota. En la Figura 34 se visualiza la matriz de confusión en donde se observa que 49 datos PROMEDIOS, 65 datos EXCELENTE, 48 datos BUENO, y 45 datos LIMITE son verdaderos positivos.

Tabla 17

Obtención de los parámetros de desempeño clase LIMITE modelo SVM

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	47.12	93.05	76.63	79.03	0.30
EXCELENTE	95.59	91.03	92.1	76.47	0.07
BUENO	66.67	87.67	82.47	64	0.23
LIMITE	95.75	90.16	91.07	65.22	0.07

BER TOTAL	=	0.17

Nota. Se observa las métricas de la matriz de confusión para cada clase, y se observa que la clase EXCELENTE tiene el menor BER en este caso.

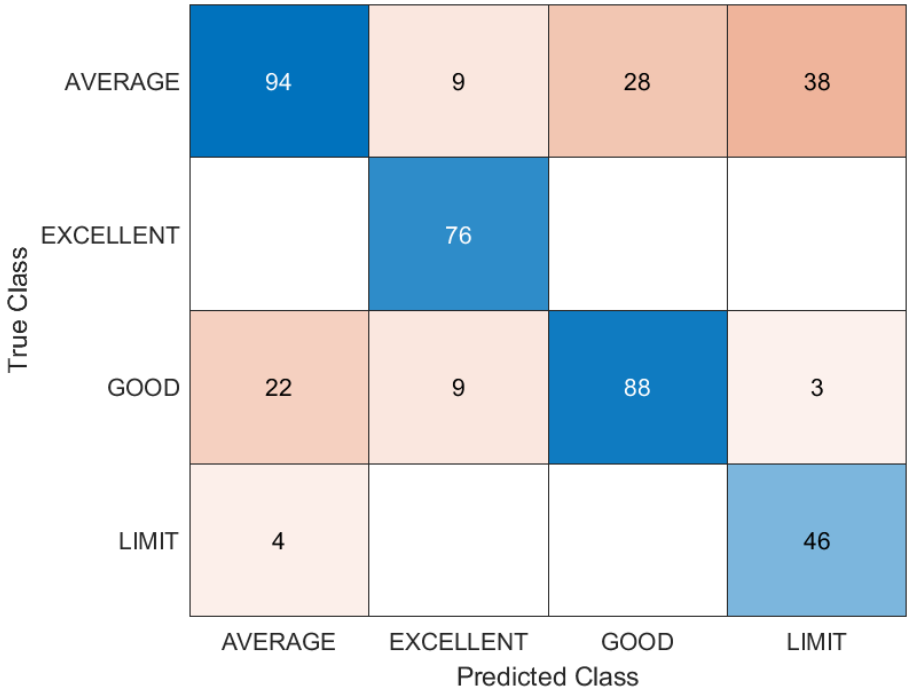
En la Tabla 17 se observa el BER general de toda la matriz de confusión de 0.17 resultado de aplicar el modelo de clasificación SVM al considerar que se encuentra balanceado a la clase LIMITE con 243 datos.

Proceso: Balanceado al usar la clase EXCELENTE

La Figura 35 muestra la matriz de confusión para la clase *EXCELENTE* donde los falsos positivos se reducen a diferencia del modelo de clasificación al emplear el DTs, se obtiene así una mejor predicción.

Figura 35

Matriz de confusión clase EXCELENTE



Nota. En la Figura 35 se visualiza la matriz de confusión en donde se observa que 94 datos PROMEDIOS, 76 datos EXCELENTE, 88 datos BUENO, y 46 datos LIMITE son verdaderos positivos.

Tabla 18

Obtención parámetros de desempeño clase EXCELENTE modelo SVM

Clases	Sensibilidad (%)	Especificidad (%)	Exactitud (%)	Precisión (%)	BER
PROMEDIO	55.62	89.52	75.78	78.33	0.27
EXCELENTE	100	94.72	95.68	80.85	0.03
BUENO	72.13	90.51	85.13	75.86	0.19
LIMITE	92	88.83	89.21	52.87	0.1

BER TOTAL	=	0.15

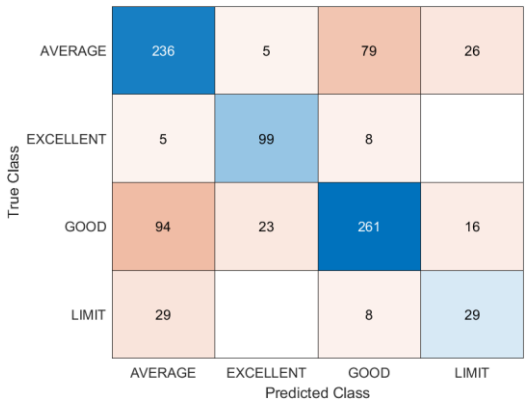
Nota. Se observa las métricas de la matriz de confusión para cada clase, en este caso la clase EXCELENTE es la de menor BER con un valor de 0.026393.

En la Tabla 18 se observa el BER general de toda la matriz de confusión de 0.15 resultado de aplicar el modelo de clasificación SVM al considerar que se encuentra balanceado a la clase EXCELENTE con 383 datos.

Proceso: Balanceado al usar la clase PROMEDIO

Figura 36

Matriz de confusión clase PROMEDIO



Nota. En la Figura 36 se visualiza la matriz de confusión en donde se observa que 236 datos PROMEDIO, 99 datos EXCELENTE, 261 datos BUENO, y 29 datos LIMITE son verdaderos positivos.

La Figura 36 muestra la matriz de confusión para la clase PROMEDIO donde se obtienen resultados similares a la matriz de confusión generado por el modelo de clasificación DTs.

Tabla 19

Obtención parámetros de desempeño clase PROMEDIO modelo SVM

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	68.21	77.62	74.07	64.84	0.27
EXCELENTE	88.39	96.53	95.33	77.95	0.08
BUENO	66.24	81.87	75.16	73.32	0.26
LIMITE	43.94	95.07	91.39	40.85	0.3

BER TOTAL	=	0.23

Nota. Se observa las métricas de la matriz de confusión para cada clase en donde permanece la clase EXCELENTE como la de menor BER de 0.075405 en este caso.

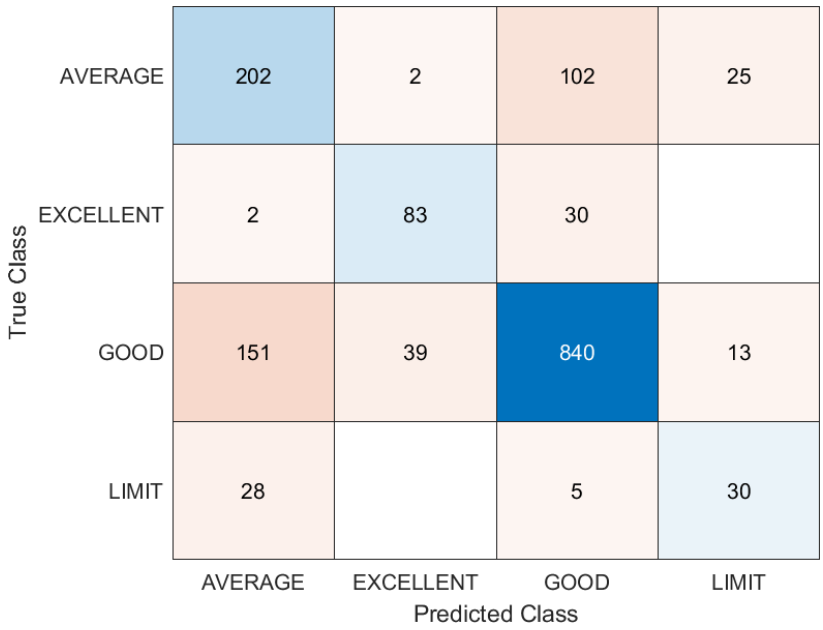
En la Tabla 19 se observa el BER general de toda la matriz de confusión de 0.27 resultado de aplicar el modelo de clasificación SVM al considerar que se encuentra balanceado a la clase PROMEDIO con 1218 datos, en este caso el error es más grande que la en la anterior clase por lo ya que no es considerado factible el proceso de balancear datos con esta clase.

Proceso: Sin realizar balanceado de datos

La Figura 37 muestra la matriz de confusión para la clase BUENO donde se consigue un aumento de los falsos positivos al discrepar de la matriz de confusión generado por el modelo de clasificación DTs.

Figura 37

Matriz de confusión clase BUENO



Nota. En la Figura 37 se visualiza la matriz de confusión en donde se observa que 202 datos PROMEDIO, 83 datos EXCELENTE, 841 datos BUENO, y 30 datos LIMITE son verdaderos positivos.

Tabla 20

Obtención de los parámetros de desempeño clase BUENO modelo SVM

Clases	**Sensibilidad (%)**	**Especificidad (%)**	**Exactitud (%)**	**Precisión (%)**	**BER**
PROMEDIO	61.03	85.18	80.03	52.74	0.27
EXCELENTE	72.17	97.15	95.3	66.94	0.15
BUENO	80.54	73.08	78.09	85.98	0.23
LIMITE	47.62	97.45	95.43	44.12	0.27

BER TOTAL	=	0.23

Nota. Se observa las métricas de la matriz de confusión sin el proceso del balanceado de datos, al igual que al aplicar balanceado de datos el BER por clase con menor error permanece la clase EXCELENTE con un valor de 0.15.

En la Tabla 20 se observa el BER general de toda la matriz de confusión de 0.23 resultado de aplicar el modelo de clasificación SVM al considerar que los datos no se encuentran balanceados.

Valores de LQI base original

Los valores de LQI indican a que categoría pertenece cada coordenada tomada en relación con los niveles de intensidad de la señal, esta va a depender si la medición fue tomada cerca de una estación base que generalmente es donde la calidad de la medición es EXCELENTE o a su vez si tiene barreras naturales o no tiene línea de vista y por ende la calidad de la señal resultará menor.

La Tabla 21 muestra la cantidad de etiquetas por cada clase; LIMITE, PROMEDIO, BUENO y EXCELENTE, mismas que son empleados en los modelos de regresión.

Tabla 21

Número de etiquetas por cada clase LQI base original

Etiqueta	Número de etiquetas	Color de etiqueta
LIMITE	243	Rojo
PROMEDIO	1218	Verde
BUENO	3330	Amarillo
EXCELENTE	383	Azul
TOTAL	5174	

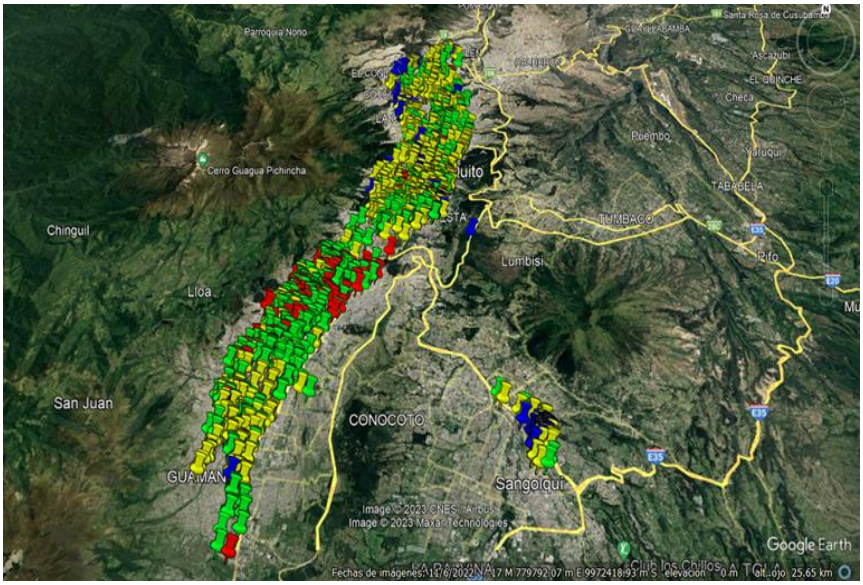
Nota. Se observa la cantidad de etiquetas dispuestas por cada clase de LQI entre LIMITE, PROMEDIO, BUENO y EXCELENTE y se respectivo color identificativo.

Categorización de forma geográfica base original

Para visualizar los resultados de manera gráfica se emplea los 5174 datos originales completos con el valor RSSI y etiquetas LQI donde se obtiene el siguiente resultado.

Figura 38

Categorización de valores LQI correspondiente al canto Quito

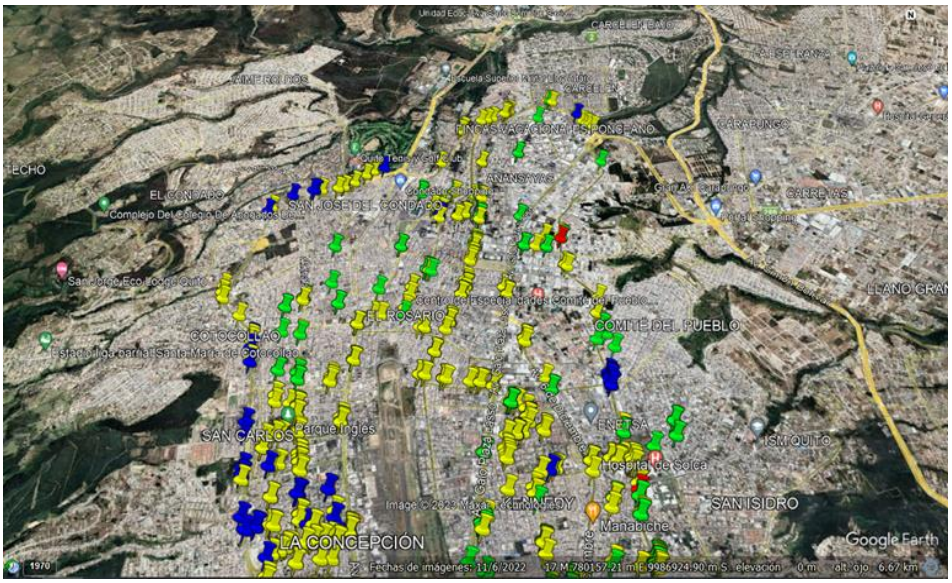


Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

Nota. En la Figura 38 se muestran las cuatro categorías correspondientes al valor LQI en base a las mediciones RSSI de 5174 datos distribuidas en todo el cantón Quito y en Sangolquí correspondiente a la Universidad de las Fuerzas Armadas ESPE.

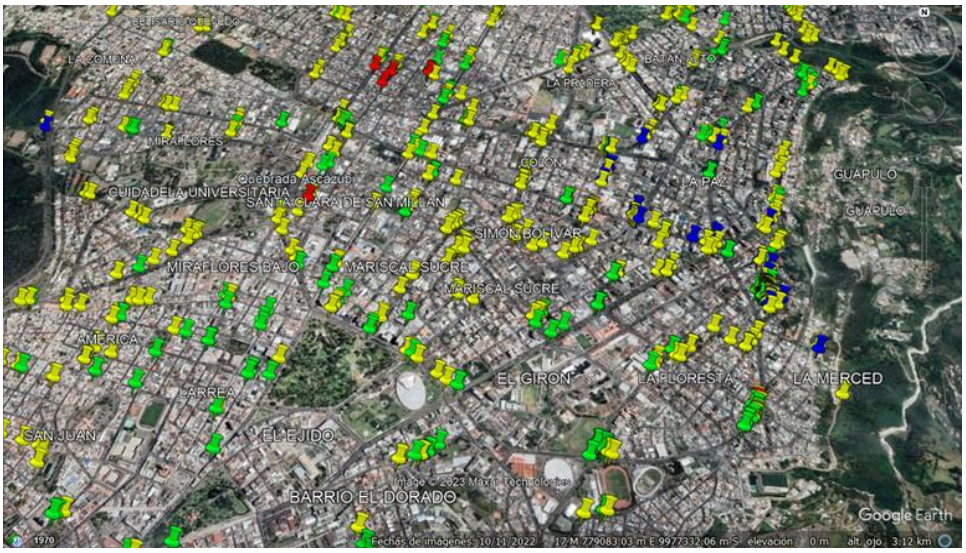
Figura 39

Categorización de valores LQI en el canto Quito por zonas



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

a)



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

b)

Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

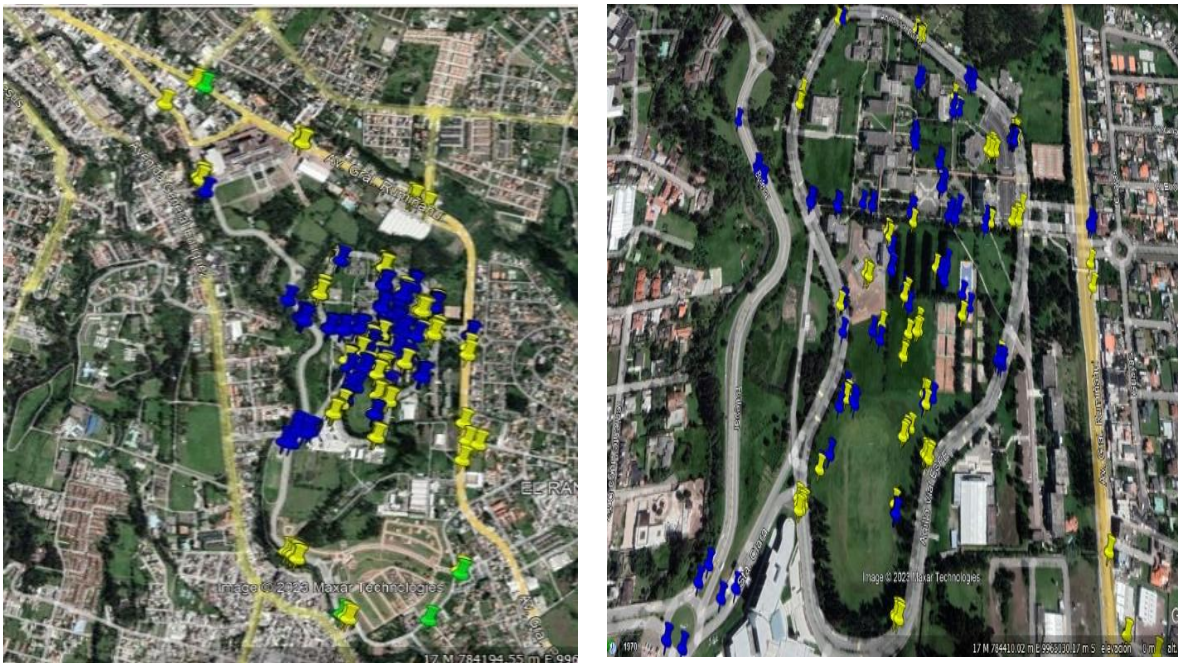
c)

Nota. En la Figura 39 se observa el valor LQI separado por zonas del cantón Quito: a) zona norte, b) zona centro y c) zona sur.

Como se muestra en la Figura 39 en las zonas de norte y centro no se observan valores LIMITE, y solo valores de EXCELENTE, BUENO y PROMEDIO, en su lugar en donde se observan más valores LIMITE es en la zona sur rodeada de valores PROMEDIO, esto demuestra que en ciertas zonas del sur de Quito no existen estaciones base o a su vez existen interferencias.

Figura 40

Categorización LQI en la Universidad de las Fuerzas Armadas ESPE



a) b)

Nota. En la Figura 40 se observa el valor LQI en el cantón Rumiñahui en Sangolquí lugar donde se encuentra la Universidad de las Fuerzas Armadas ESPE, a) representa la zona donde se realizaron las mediciones en Sangolquí y b) representa las mediciones dentro de la ESPE.

Como se observa en la figura anterior no se presentan valores que se encuentren en la categoría LIMITE, solo se observan valores BUENO, PROMEDIO, y EXCELENTE sobre todo en las inmediaciones de la ESPE.

Base de datos al aplicar *Data Augmentation* (DA)

Los resultados de DA al utilizar el mejor modelo encontrado de acuerdo a los datos de la base original entrega como resultado un total de 62942 nuevos datos como se describe en el capítulo de metodología, con estos nuevos datos se obtuvieron los mapas de calor en base a los valores de la intensidad de la señal RSSI mismos que sirven para observar de

manera gráfica el rendimiento del modelo de acuerdo a lo realizado en el cantón Quito y en la Universidad de las Fuerzas Armadas ESPE.

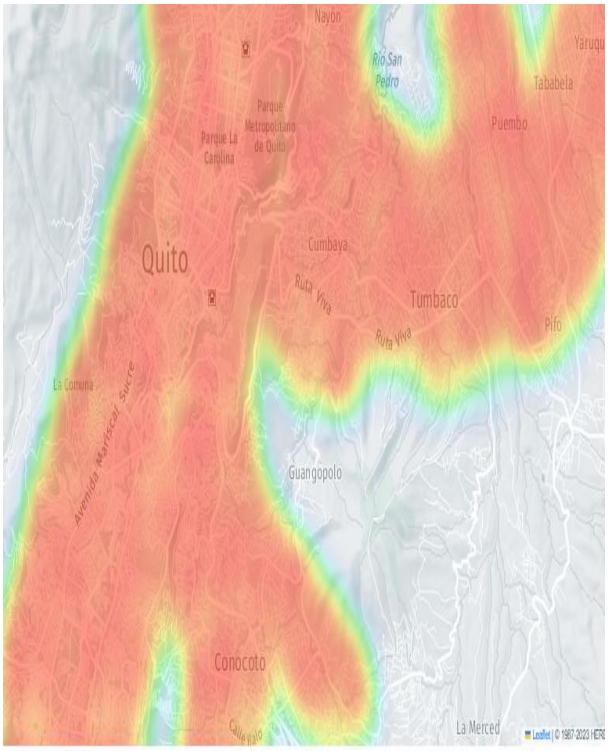
El modelo que se aplicó para el relleno de estos valores es el modelo de SVM con 3 variables (latitud, longitud y altura), pese a que el modelo con 4 variables era mejor este mismo no se puede aplicar debido a que la cuarta variable LQI no se genera en el software *Open Foris Collect*, solo se obtiene la coordenada con latitud, longitud y altura.

Mapas de calor con *Data Augmentation*

A continuación, se muestra el mapa de calor que indica los valores de la intensidad de la señal RSSI de cantón Quito con 62942 nuevos datos.

Figura 41

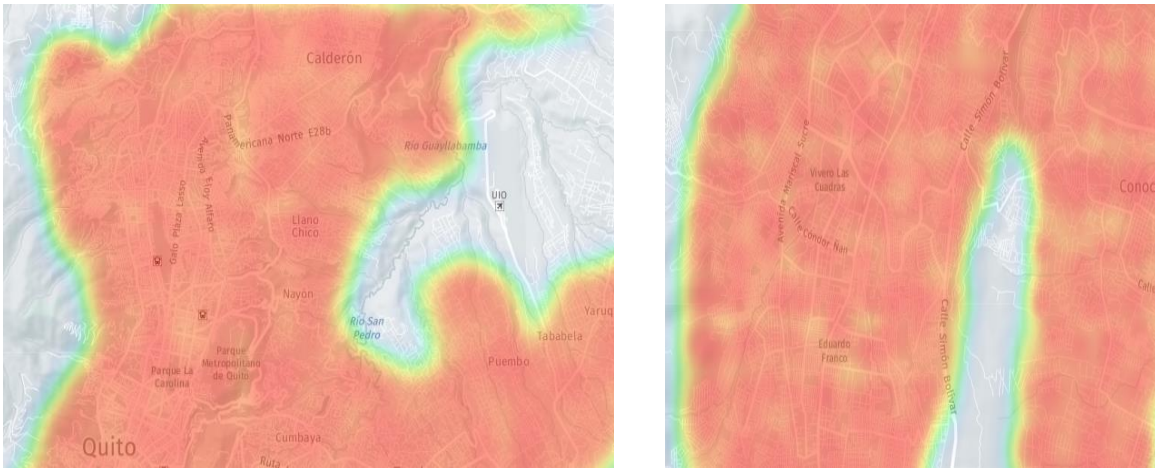
Categorización de valores LQI en el canto Quito separado por zonas



Nota. En la Figura 41 se muestra el mapa de calor del cantón Quito en donde se representa la intensidad de señal RSSI, realizado con 62942 nuevos datos.

Figura 42

Mapa de calor zona norte y zona sur del cantón Quito



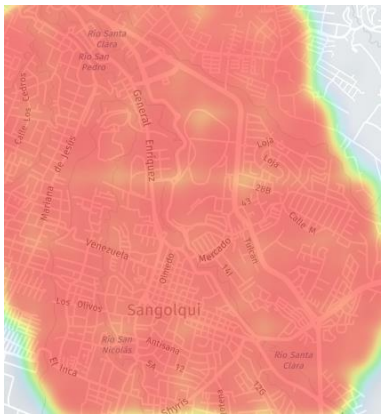
a) b)

Nota. En la Figura 42 se muestra el mapa de calor de: a) la zona norte del cantón Quito y b) la zona sur del cantón Quito, realizado con 62942 nuevos datos.

Al igual que los mapas de calor de la base original 5174 datos aquí también se cuenta con coordenadas del cantón Rumiñahui que pertenecen a Sangolquí lugar donde se encuentra la Universidad de las Fuerzas Armadas ESPE.

Figura 43

Mapa de calor en la Universidad de las Fuerzas Armadas ESPE



Nota. En la Figura 43 se muestra el mapa de calor correspondiente a los valores de RSSI en base a las coordenadas tomadas en la Universidad de las Fuerzas Armadas ESPE.

Categorización de forma geográfica Data Augmentation

Modelo: SVM

Dentro de los nuevos valores generados de RSSI y de LQI se pueden obtener varios elementos en cada etiqueta.

La Tabla 22 muestra la cantidad de etiquetas por cada clase; LIMITE, PROMEDIO, BUENO y EXCELENTE generados por el modelo de clasificación SVM.

Tabla 22

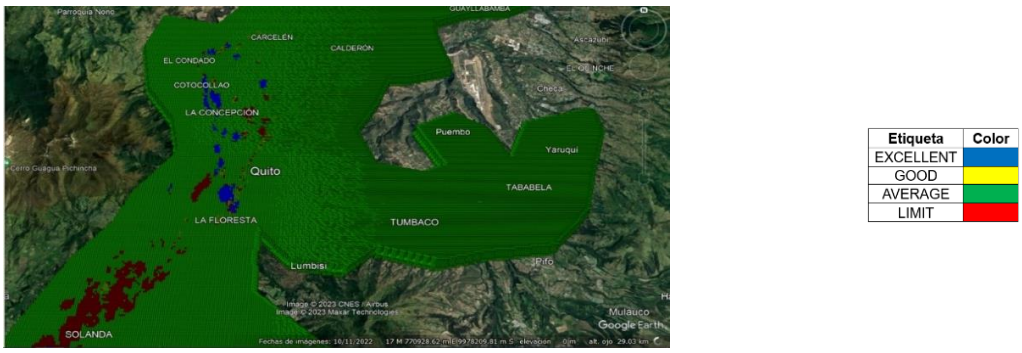
Número de etiquetas por cada clase LQI modelo SVM

Etiqueta	Número de etiquetas	Color de etiqueta
LIMITE	352	Rojo
PROMEDIO	62014	Verde
BUENO	275	Amarillo
EXCELENTE	301	Azul
TOTAL	62942	

Nota. Se observa la cantidad de etiquetas generadas por el modelo SVM para cada clase de LQI entre LIMITE, PROMEDIO, BUENO y EXCELENTE y se respectivo color identificativo.

Figura 44

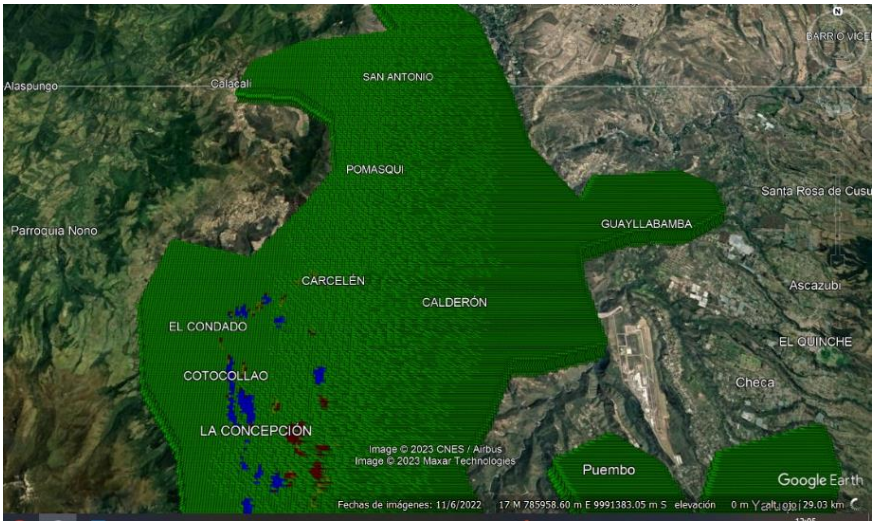
Categorización de valores LQI correspondiente al canto Quito



Nota. En la Figura 44 se muestran las cuatro categorías correspondientes al valor LQI en base a las mediciones RSSI de los 62942 nuevos datos distribuidos en todo el cantón Quito.

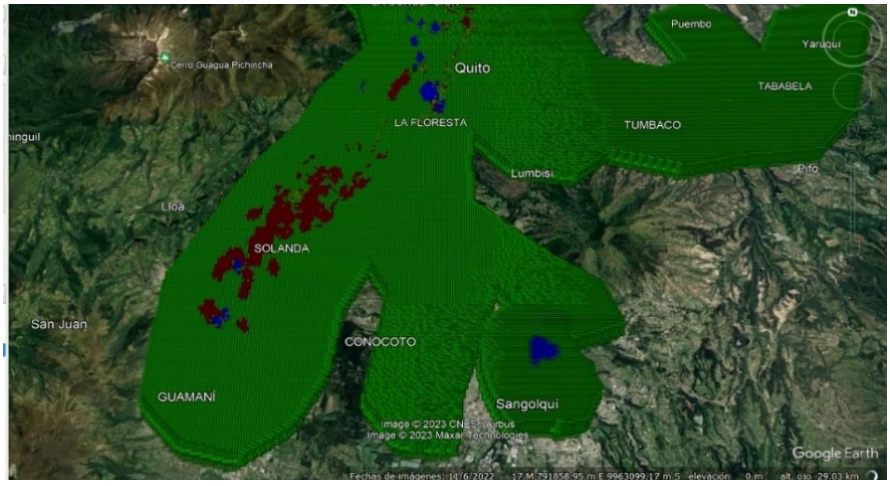
Figura 45

Categorización de valores LQI del canto Quito separado por zonas



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

a)



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

b)

Nota. En la Figura 45 se observa el valor LQI separado por zonas del cantón Quito: a) zona norte y b) zona sur.

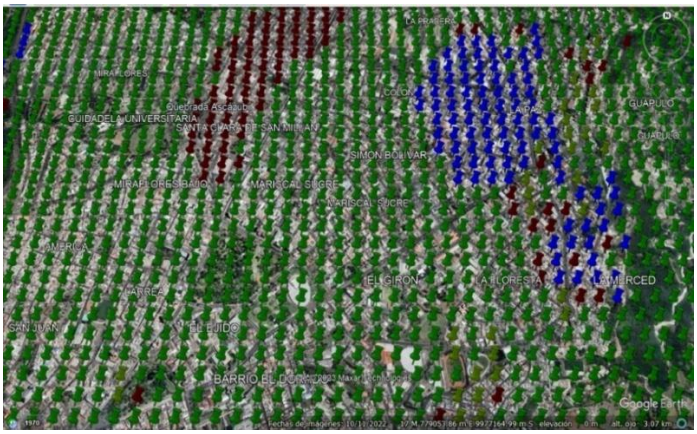
Es importante resaltar que las coordenadas entre si tiene una separación de 100 metros en forma de cuadrilla, es por eso que las coordenadas totales que cubren todo el cantón Quito son 62942. A continuación, se realizó un acercamiento de los mapas y se efectuaron las capturas de los sectores norte, centro y sur donde se observa dicha separación entre coordenadas.

Figura 46

Categorización zonas de Quito, separación de coordenadas.

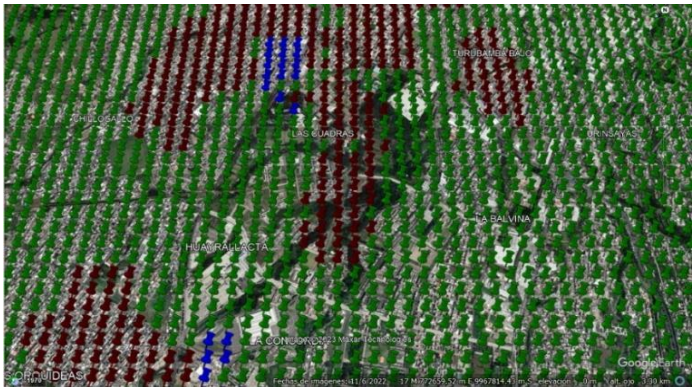
Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

a)



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

b)



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

c)

Nota. En la Figura 46 se observa las categorías del valor LQI que corresponden a las zonas: a) Norte de Quito donde se observa el DA en el sector del parque Bicentenario, b) Centro de

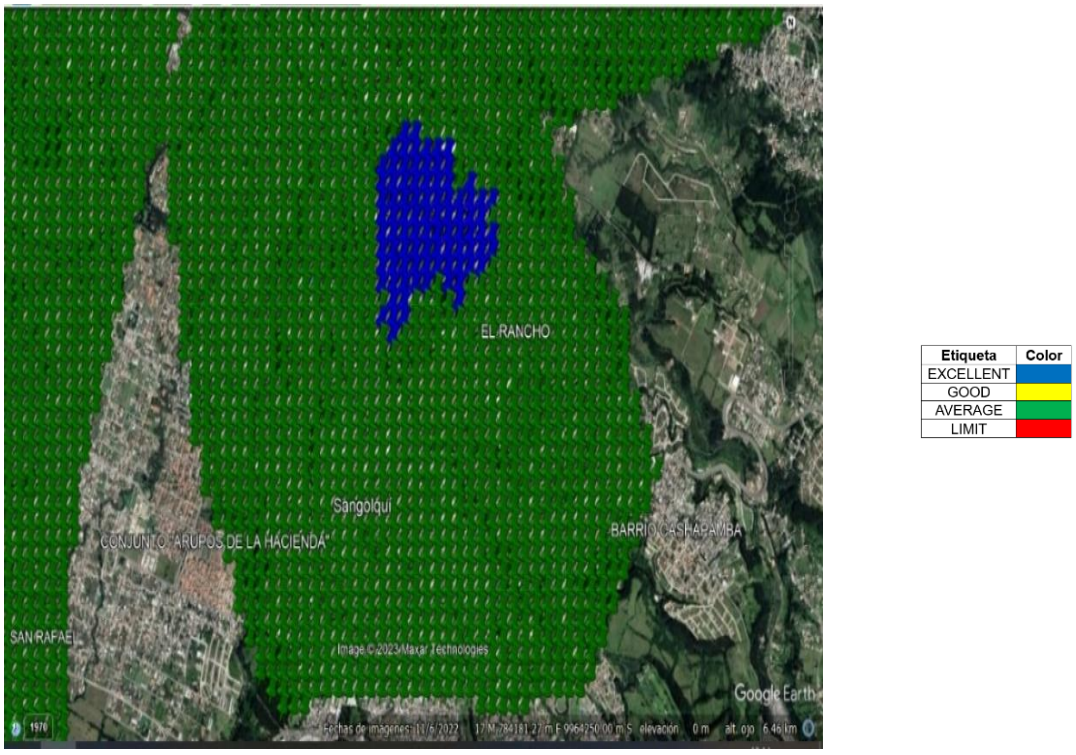
Quito donde se observan las coordenadas en el sector de la Floresta, el Girón y la ciudadela Universitaria, y c) Sur de Quito donde se observa el sector de Chillogallo, y las Cuadras.

En las zonas norte de Quito se encuentran las categorías de color verde que representan PROMEDIO correspondientes al valor de RSSI de cada coordenada, a partir de las zonas centro de Quito se observan categorías PROMEDIO, EXCELENTE y en pocas zonas valores LIMITE, en las zonas sur las categorías LIMITE y PROMEDIO son las que más presencia tienen.

Esto se debe a las mediciones realizadas de RSSI mismas que se encuentran en la base original de 5174 datos por lo tanto al utilizar DA existe una relación con esos valores y DA lo rellenara de acuerdo a la relación que existe entre: las coordenadas latitud y longitud, y la altura.

Figura 47

Categorización cantón Rumiñahui, Sangolquí lugar ESPE



a)

b)

Nota. En la Figura 47 se observa: a) sector seleccionado Sangolquí correspondiente al cantón Rumiñahui, b) zona que cubre a la universidad de las Fuerzas Armadas misma que presenta las categorías de PROMEDIO representada por el color verde y EXCELENTE que representa el color azul.

Al hacer una comparación entre la figura 40 y la figura 47 se observa que las etiquetas LQI tienen relación con la base original y con la base de datos aumentada por *Machine Learning,* lo que afirma que el modelo de clasificación SVM está correcto, ya que se ve un incremento de las etiquetas del mismo tipo, EXCELENTE para el caso de la ESPE, de manera similar al realizar una comparación entre la figura 39 y 46 se observan que las etiquetas siguen el mismo patrón, LIMITE para el caso del SUR, BUENO y PROMEDIO para el centro y norte de Quito respectivamente.

Categorización geográfica Data Augmentation en función del valor de RSSI

De forma similar al caso de la generación de las etiquetas por medio del modelo de clasificación SVM y DTs, se pueden obtener estas etiquetas al utilizar los rangos RSSI previamente generados por el modelo de regresión SVM.

En este caso el valor del LQI para asignar las categorías correspondientes dependen del rango en el que este se encuentre el valor del RSSI.

La Tabla 23 describe los posibles rangos según el valor de RSSI que pueden obtener cada de una de las clases de LIMITE, PROMEDIO, BUENO y EXCELENTE.

Tabla 23

Clasificación de datos por rangos de LQI con Data Augmentation

LQI	Rangos [dBm]
LIMITE	$[-94\,,\,-66]$
PROMEDIO	$[-115\,,\,-95]$
BUENO	$[-127\,,\,-116]$
EXCELENTE	$[-133\,,\,-128]$

Nota. En esta tabla se observan los rangos en dBm para las cuatro clases del valor LQI el mismo que depende de la intensidad de la señal RSSI.

Esto permite conocer a que clase pertenece el valor de RSSI que se encuentre dentro de alguno de estos rangos. Una vez definido esto, se puede realizar un código que nos permita clasificar cada coordenada para conocer su clase, graficarla y conocer su geolocalización.

Tabla 24

Número de etiquetas por cada clase LQI con Data Augmentation

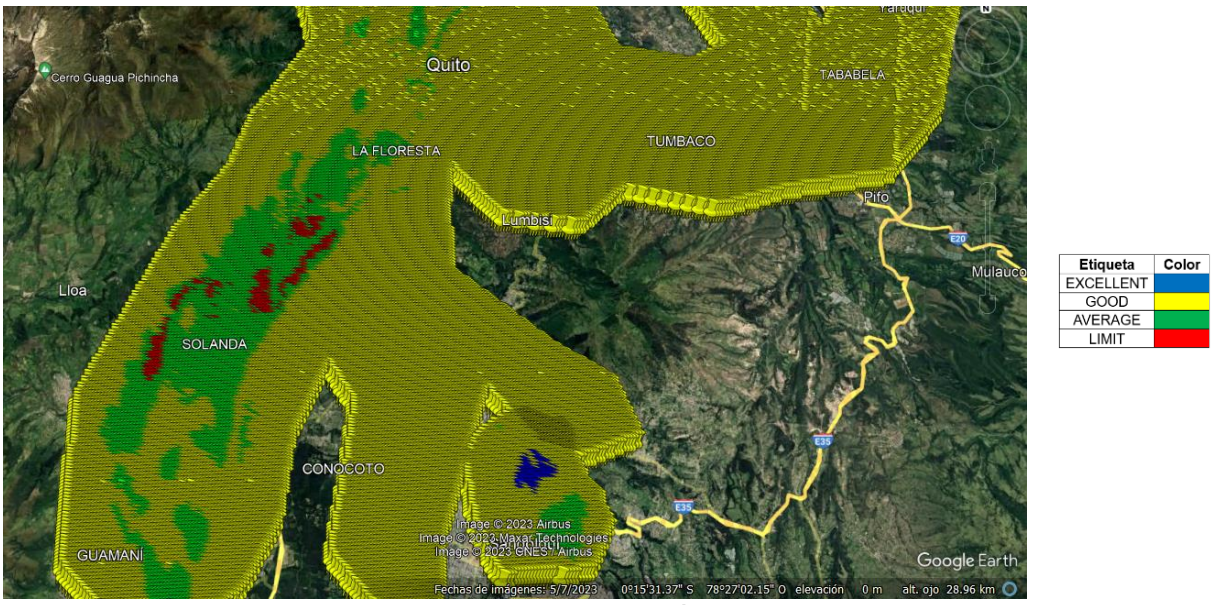
Etiqueta	Número de etiquetas	Color de etiqueta
LIMITE	401	Rojo
PROMEDIO	5162	Verde
BUENO	57062	Amarillo
EXCELENTE	317	Azul
TOTAL	62942	

Nota. Se observa la cantidad de etiquetas generadas por el modelo SVM para cada clase de LQI entre LIMITE, PROMEDIO, BUENO y EXCELENTE y se respectivo color identificativo.

La Tabla 24 muestra la cantidad de etiquetas por cada clase; LIMITE, PROMEDIO, BUENO y EXCELENTE generados por el modelo de clasificación SVM.

Figura 48

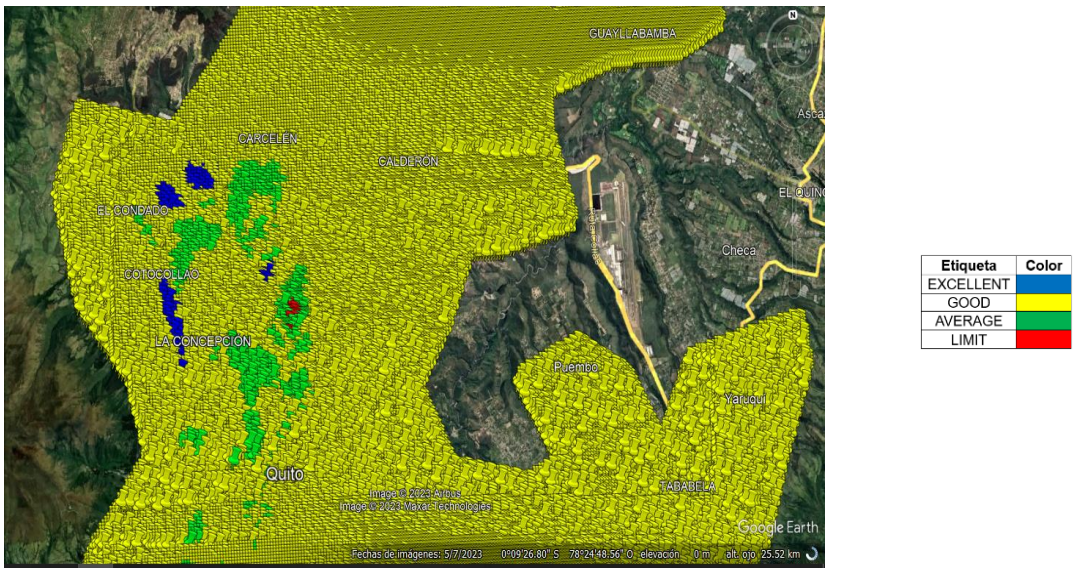
Categorización de valores LQI del canto Quito



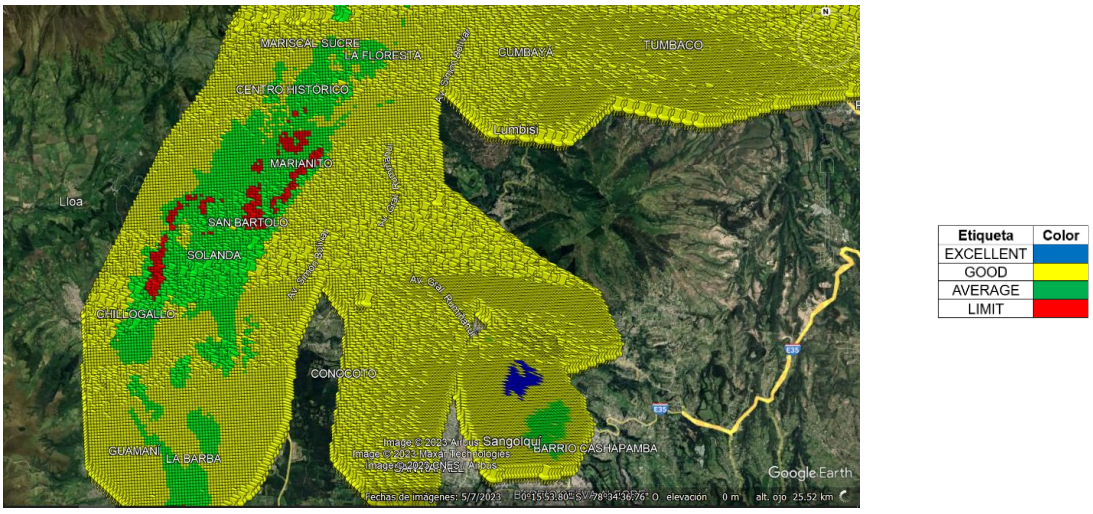
Nota. En la Figura 48 se muestran las cuatro categorías correspondientes al valor LQI mismas que se obtiene en función del valor de RSSI y dependen del rango al que pertenezca.

Figura 49

Categorización LQI por rangos del canto Quito separado por zonas



a)



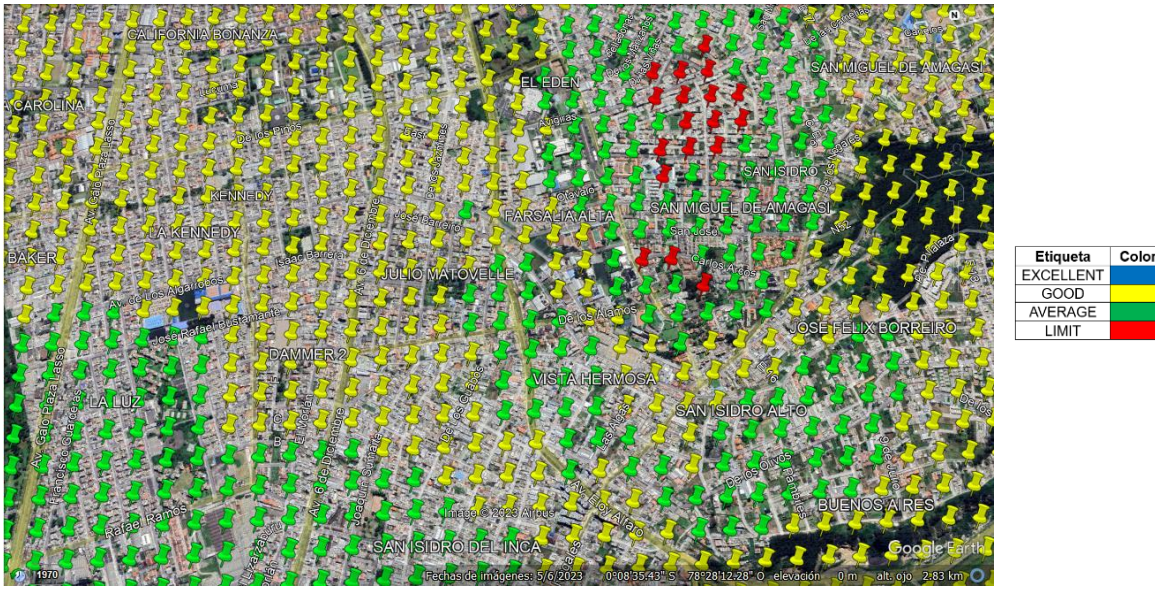
b)

Nota. En la Figura 49 se observa el valor LQI separado por zonas del cantón Quito: a) zona norte y b) zona sur.

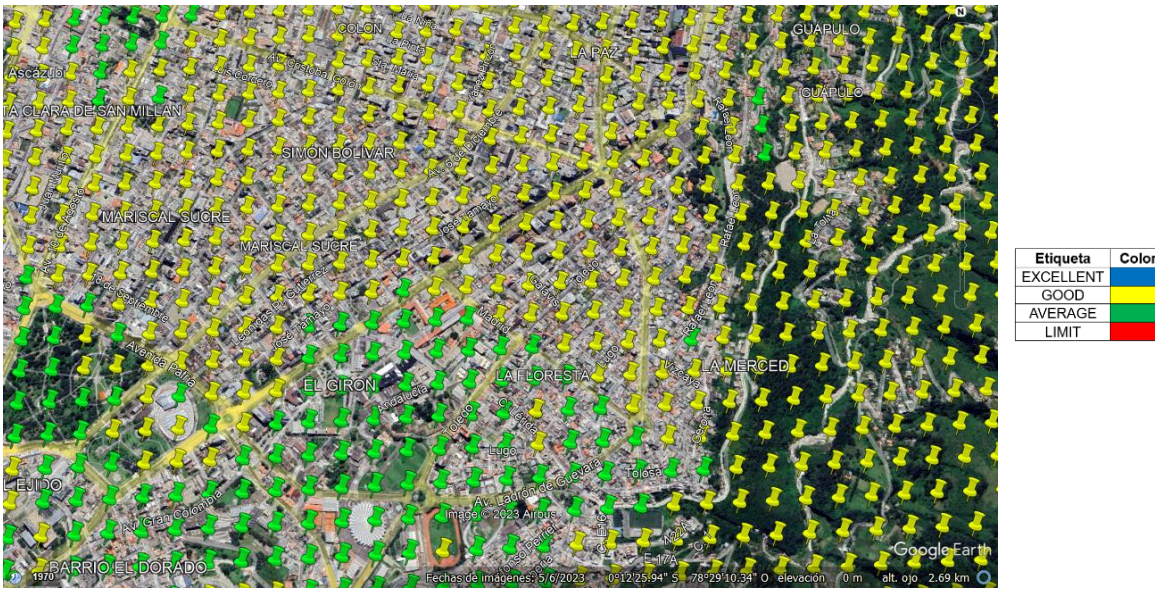
A continuación, se realizó un acercamiento de los mapas y se efectuaron las capturas de los sectores norte, centro y sur donde se observa dicha separación entre coordenadas, así como el aumento de coordenadas por *Data Augmentation*.

Figura 50

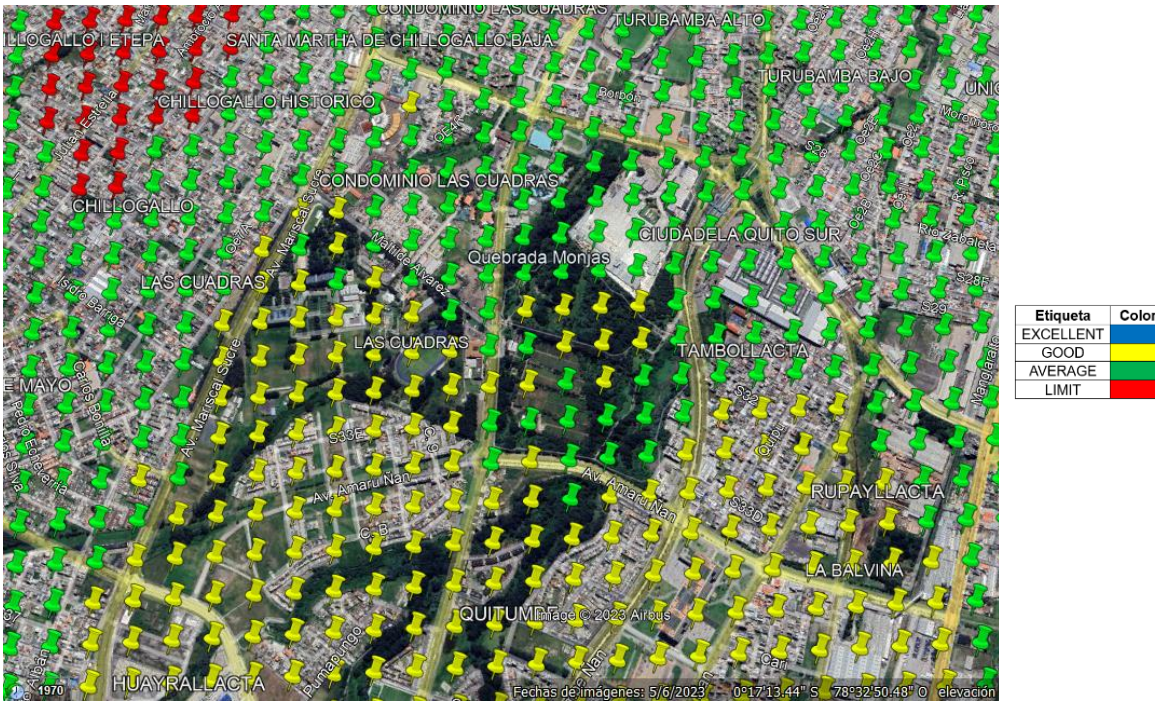
Categorización zonas de Quito, separación de coordenadas



a)



b)



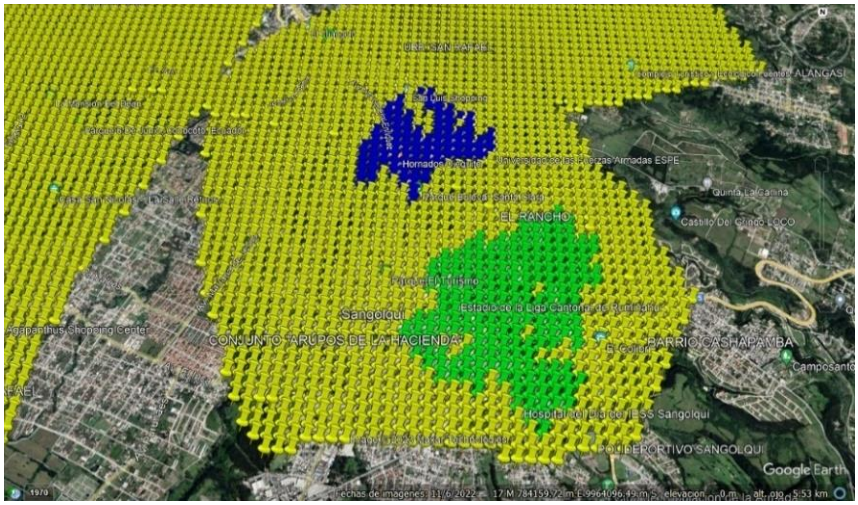
c)

Nota. En la Figura 50 se observan las categorías del valor LQI obtenido por rangos mismas que corresponden a las zonas: a) Norte de Quito donde se observa presencia de valores BUENO en el sector del parque del Bicentenario, la Kennedy, b) Centro de Quito donde se observan las coordenadas en el sector de la Floresta, el Girón y la ciudadela Universitaria y c) Sur de Quito donde se observa el sector de la Ciudadela Quito Sur, y las Cuadras.

En las zonas del norte de Quito se observa presencia de valores BUENO en el sector del parque del Bicentenario, la Kennedy, mientras que en el sector de La Luz y Vista hermosa presentan valores PROMEDIO, además se pueden ver pequeñas áreas donde la categoría LIMITE aparece a diferencia del anterior caso que en su mayoría solo presentaba valores PROMEDIO, en la zona centro de Quito se observan las coordenadas en el sector de la Floresta, el Girón y la ciudadela Universitaria con presencia de categorías PROMEDIO y BUENO, y al sur de Quito se observa el sector de la Ciudadela Quito Sur, y las Cuadras que presentan valores PROMEDIO, BUENO y LIMITE.

Figura 51

Categorización cantón Rumiñahui, Sangolquí lugar ESPE



Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

a)

Etiqueta	Color
EXCELLENT	
GOOD	
AVERAGE	
LIMIT	

b)

Nota. En la Figura 51 se observa: a) Sector seleccionado Sangolquí correspondiente al cantón Rumiñahui, b) Zona que cubre a la universidad de las Fuerzas Armadas ESPE.

En el sector seleccionado de Sangolquí se observa gran cantidad de valores BUENO, alrededor de la ESPE se observa claramente valores EXCELENTE y en sentido sur lugar donde se encuentra el estadio de la liga cantonal de Rumiñahui presentan valores PROMEDIO.

Capítulo V

Conclusiones, Recomendaciones y Trabajos Futuros

Conclusiones

En el proceso de investigación y análisis, se ha realizado una revisión exhaustiva del estado del arte en donde se identificaron las principales técnicas tradicionales del aprendizaje automático. Este análisis ha proporcionado una comprensión sólida de las prácticas y enfoques convencionales que se han utilizado en el campo de *Machine Learning*. Como resultado, se definió una base de conocimientos sólida en torno a estas técnicas, lo que contribuyó en el desarrollo de soluciones en el presente trabajo de titulación. Esta revisión ha permitido identificar las fortalezas y limitaciones de las técnicas tradicionales, lo que sirvió como punto de partida para futuros avances y enfoques más innovadores en el campo del aprendizaje automático.

Se identificó que para el regresor los mejores resultados fueron obtenidos en base a los valores de MSE y RMSE, donde el regresor de menor error corresponde al modelo de SVM con un MSE de 53.73 dBm y un RMSE de 7.33 dBm.

Con este resultado se concluye el mejor modelo para el regresor y se completó los 2145 valores restantes o valores nulos correspondientes al RSSI, de esta manera se rellenó la base de datos inicial de 5137 datos.

Se identificaron los mejores resultados de los clasificadores en base a Machine Learning tradicional con los valores correspondientes al 70% para entrenamiento y 30% para validación, además mediante el proceso de balanceamiento de datos se estimó que el mejor resultado alcanzado por ambos modelos de clasificación (DTs y SVM) fue el balanceamiento de la clase EXCELENTE con 383 valores de los 5137.

De esta manera se obtuvo para el modelo de SVM balanceado a la clase EXCELENTE una exactitud del 95.68% de la clase y un BER total de 0.15, por otro lado, para el modelo de DTs balanceado a la clase EXCELENTE se obtuvo una exactitud del 94.48% de la clase

y un BER total de 0.16. Por lo tanto, se concluye que el mejor modelo que se adapta mejor a la base de datos para clasificación es el SVM con un BER del 0.15.

Para la obtención de las nuevas coordenadas correspondientes al cantón Quito, se concluyó que el software *Open Foris Collect* fue el ideal para generar datos con las variables necesarias de acuerdo a nuestra nueva base, se generan 62942 datos nuevos.

Se llegó a definir que al utilizar los mejores modelos de regresión y clasificación en las nuevas variables se obtiene una base de datos completa con todas las variables el cual está mucho mayor a la base de datos inicial.

Finalmente, se identificó luego de una comparación visual entre la base de datos aumentada y la base de datos original, que la zona de cobertura de la base de datos aumentada es considerablemente mayor en comparación con la base de datos original ya que ahora está abarca áreas donde antes no existía cobertura o, en su defecto era muy débil.

De manera similar, se obtuvo gráficos de los valores de LQI para cada coordenada, y se identificó que los modelos de aprendizaje siguieron el estándar de la base de datos original, es decir que se respetaron las áreas de mejor cobertura como las de peor cobertura. Se concluye que es posible desarrollar diversos modelos de aprendizaje donde se usan técnicas tradicionales de *Machine Learning* para clasificación como para regresión en una base de datos inicial pequeña y llevarla a una base de datos mucho mayor.

Recomendaciones y Trabajos Futuros

Se recomienda el uso de la optimización de los hiperparámetros en los modelos SVM y DTs para evitar el uso de los parámetros por defecto que se emplean en el software Matlab, de esta forma se pueden obtener modelos de aprendizaje con mejores resultados.

Se recomienda al momento de realizar la generación de datos se debe seleccionar un software que puede generar las variables latitud, longitud y altura porque son variables necesarias en los modelos de entrenamiento, además de efectuar un trazado de área

acorde al caso de estudio dentro del cantón Quito al reducir las áreas con nivel forestal alto y seleccionar las zonas urbanas.

De acuerdo al trabajo realizado se puede realizar la publicación de un artículo científico donde se describan otros modelos tradicionales de Machine Learning de regresión y clasificación desarrollados y los diferentes resultados obtenidos para la generación de datos sintéticos en una red Sigfox o en una base de datos similar.

Se pueden desarrollar otros modelos de regresión y clasificación al utilizar Deep Learning para verificar el comportamiento de dichos modelos respecto a los datos de latitud, longitud, altura y LQI que se tienen en la base de datos original, de acuerdo a esto se podrán tener modelos con diferente desempeño.

Se recomienda el uso de software que pueden representar gráficamente todos los puntos o etiquetas requeridas con la finalidad de no perder información y se obtenga una visualización gráfica completa para realizar comparaciones entre modelos y tomar decisiones de manera correcta.

Referencias

Agha, K., Pujolle, G., & Yahiya, T. (2016). *Mobile and Wireless Networks* (2nd ed., Vol. 2). WILEY. https://books.google.com.ec/books?id=vhYlCgAAQBAJ&pg=PA241&dq=sigfox+definition&hl=es&sa=X&ved=2ahUKEwjwu6vM_db7AhUSSzABHaTzDMMQ6AF6BAgCEAI#v=onepage&q=sigfox%20definition&f=false

Anjum, M., Khan, M. A., Hassan, S. A., Mahmood, A., Qureshi, H. K., & Gidlund, M. (2020). *RSSI Fingerprinting-based Localization Using Machine Learning in LoRa Networks*. https://doi.org/10.1109/IOTM.0001.2000019

Boser Bernhard, Guyon Isabelle, & Vapnik VN. (1992). *A Training Algorithm for Optimal Margin Classifiers.* (ACM Workshop, Vol. 5th). http://doi.org/10.1.1.21.3818.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/J.NEUCOM.2019.10.118

Chaudhari, B., & Zennaro, M. (2020). *LPWAN Technologies for IoT and M2M Applications* (1st ed.).

Chen, M., Zhao, J., & Wang, X. (2020). An Optimal Algorithm Design of RSSI Indoor Location based on Neural Network. *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, 84–88. https://doi.org/10.1109/ICAICE51518.2020.00022

Cruz, J. C. Dela, & Amado, T. M. (2020). Development of Machine Learning-based Predictive Models for Wireless Indoor Localization Application with Feature Ranking via Recursive Feature Elimination Algorithm. *2020 IEEE International Conference on*

Signal Processing, Communications and Computing (ICSPCC), 1–6. https://doi.org/10.1109/ICSPCC50002.2020.9259526

Dong, C., Ding, J., & Lin, J. (2016). Segmented polynomial RSSI-LQI ranging modelling for ZigBee-based positioning systems. *Chinese Control Conference, CCC*, *2016-August*, 8387–8390. https://doi.org/10.1109/CHICC.2016.7554693

González, S. (2021, December 10). *Por qué el Machine Learning es un gran aliado para la ciberseguridad*. Welivesecurity.

Joubert, P. J., & Helberg, A. S. J. (2015). An investigation into the use of the kriging for Wi-Fi RSSI estimation in complex indoor environments. *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, 1326–1331. https://doi.org/10.1109/WCNC.2015.7127661

Le, T. G., Quach, H. T., Le, T. T. D., & Tran, M. H. (2021). *RSSI prediction using Machine Learning models*.

Maduranga, P., & Abeysekera, R. (2021). Supervised Machine Learning for RSSI based Indoor Localization in IoT Applications. *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2021921305

Microsoft. (2022). *Algoritmos de aprendizaje automático | Microsoft Azure*. https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-are-machine-learning-algorithms

Miralbell, X. (2021). Grado en Ingeniería en Tecnologías Industriales Time series data augmentation. *Escola Tècnica Superior d'Enginyeria Industrial de Barcelona*. https://upcommons.upc.edu/bitstream/handle/2117/361029/time-series-data-augmentation.pdf?sequence=1&isAllowed=y

Patel, A., Rajarajan, M., Kesswani, N., Misra. Rajiv, & Bharadwaj, V. (2021). *Internet of Things and Connected Technologies*. Springer International Publishing.

Pimentel, A. A., & Baldovino, R. G. (2022). IoT indoor localization using design of experiment analysis and multi-output regression models. *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*, 1–5. https://doi.org/10.1109/IPRECON55716.2022.10059563

Rábanos, J., & Riera, J. (2015). *Comunicaciones móviles* (3rd ed.). Editorial Universitaria Ramón Areces. https://books.google.com.ec/books?id=lnqnDAAAQBAJ&pg=PA627&dq=rssi+definicion&hl=es&sa=X&ved=2ahUKEwiXvde1i9f7AhUESDABHW7HDwIQ6AF6BAgDEAI#v=onepage&q=rssi%20definicion&f=false

Sanz, J., & Rodríguez, B. (n.d.). *Manual Práctico de Inteligencia Artificial En Entornos Sanitarios - Google Books*. Retrieved April 16, 2023, from https://www.google.com.ec/books/edition/Manual_Pr%C3%A1ctico_de_Inteligencia_Artific/aQWtEAAAQBAJ?hl=es&gbpv=1&dq=porcentaje+de+entrenamiento+y+test+machine+learning&pg=PA50&printsec=frontcover

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

Sigfox support. (n.d.). *Link Quality: general knowledge*. Retrieved July 4, 2023, from https://support.sigfox.com/docs/link-quality:-general-knowledge

Silva, H., & Moya, S. (2019). *DESARROLLO Y ANÁLISIS DE SISTEMAS DE ESTIMACIÓN Y DETECCIÓN DE OBJETIVOS DE RADAR MEDIANTE ALGORITMOS DE MACHINE Y DEEP LEARNING* [Universidad de las Fuerzas Armadas ESPE]. http://repositorio.espe.edu.ec/bitstream/21000/20499/1/T-ESPE-039346.pdf

Sutiyo, Hidayat, R., Sunarno, & Mustika, I. W. (2018). Regression Analysis for Estimated Distance in Fingerprinting-Based WLAN Outdoor Localization System. *2018 4th International Conference on Science and Technology (ICST)*, 1–4. https://doi.org/10.1109/ICSTC.2018.8528593

Tabladillo, M. (2023). Preparación de datos para el aprendizaje automático mejorado. *Microsoft*. https://learn.microsoft.com/es-es/azure/architecture/data-science-process/prepare-data

Taha Jijo, B., & Mohsin Abdulazeez, A. (2021). *Classification Based on Decision Tree Algorithm for Machine Learning*. *02*(01), 20–28. https://doi.org/10.38094/jastt20165

Takayama, T., Umezawa, T., Komuro, N., & Osawa, N. (2018). An Indoor Positioning Method Based on Regression Models with Compound Location Fingerprints. *2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)*, 1–7. https://doi.org/10.1109/UPINLBS.2018.8559728

Wu, X., Huang, J., Ling, J., & Shu, L. (2019). BLTM: Beta and LQI Based Trust Model for Wireless Sensor Networks. *IEEE Access*, *7*, 43679–43690. https://doi.org/10.1109/ACCESS.2019.2905550

Apéndices