PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

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

TEMA: “Desarrollo de un modelo de análisis de datos que permita predecir la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito, basado en factores ambientales, geoespaciales y socioeconómicos”

AUTOR: MARÍA GABRIELA LÓPEZ REVELO

TUTOR: PHD. RAFAEL MELGAREJO

QUITO, 2024

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I

1. INTRODUCCIÓN

1.1 Antecedentes

En las últimas décadas, la ciudad de Quito ha experimentado un incremento significativo en la frecuencia e intensidad de los incendios forestales, un fenómeno atribuido principalmente a factores como el cambio climático, la expansión urbana descontrolada y la interacción compleja de variables ambientales, geoespaciales y socioeconómicas. Los datos históricos sugieren que las condiciones climáticas adversas, como el aumento de las temperaturas y las fluctuaciones en los patrones de precipitación, junto con la baja humedad relativa, han creado un entorno favorable para la propagación de incendios. Además, la topografía accidentada y la presencia de vegetación seca en varias zonas de la ciudad y sus alrededores han facilitado la rápida expansión de estos incendios una vez iniciados.

Las áreas periurbanas de Quito, donde coexisten usos del suelo residencial y forestal, son especialmente vulnerables. La expansión urbana ha intensificado la interacción entre las actividades humanas y las áreas forestales, incrementando el riesgo de incendios debido a prácticas agrícolas, la deforestación y el manejo inadecuado del fuego. A pesar de los esfuerzos realizados por las autoridades locales y diversas organizaciones ambientales para gestionar y mitigar el riesgo de incendios forestales, la falta de herramientas predictivas avanzadas ha limitado la efectividad de estas iniciativas. La capacidad de anticipar la ocurrencia de incendios y de implementar medidas preventivas adecuadas es crucial para reducir el impacto de estos eventos catastróficos y contribuir a la gestión integral del riesgo de desastres.

En el ámbito global, se han llevado a cabo diversos estudios sobre la predicción de incendios forestales utilizando modelos de análisis de datos que incorporan factores ambientales, geoespaciales y socioeconómicos. Por ejemplo, en regiones como

California, Australia y el Mediterráneo, se han desarrollado modelos predictivos basados en técnicas de aprendizaje automático y análisis de big data que han mostrado resultados prometedores en la predicción de incendios. Estos modelos integran variables como la temperatura, la humedad relativa, la velocidad del viento, la precipitación, el uso del suelo, la topografía y la densidad poblacional, utilizando datos satelitales y sistemas de información geográfica (SIG) para mejorar la precisión de las predicciones y generar mapas de riesgo.

El desarrollo de modelos predictivos específicos para el Distrito Metropolitano de Quito es esencial debido a las particularidades climáticas, geográficas y socioeconómicas de la región. La implementación de estos modelos permitirá a las autoridades locales y a las organizaciones de gestión de riesgos anticipar y mitigar los impactos de los incendios forestales, mejorando la planificación y la respuesta ante emergencias. Esta investigación se alinea con las prioridades locales de sostenibilidad y protección ambiental, representando una contribución significativa al campo de la ciencia de datos aplicada a la gestión del riesgo de desastres.

1.2 Planteamiento del problema

El Distrito Metropolitano de Quito, una ciudad en constante crecimiento urbano y con una rica diversidad ecológica, enfrenta un riesgo creciente de incendios forestales. Estos eventos representan una amenaza significativa para la vida humana, la biodiversidad y la infraestructura, causando pérdidas económicas y medioambientales considerables. La predicción de incendios forestales es un desafío complejo debido a la interacción de múltiples factores ambientales, geoespaciales y socioeconómicos que influyen en su ocurrencia.

Entre los factores ambientales se encuentran la temperatura, la humedad relativa, la velocidad del viento y la precipitación, los cuales interactúan de manera dinámica y pueden variar significativamente en cortos períodos de tiempo. Los factores

geoespaciales, como el uso del suelo, el tipo de vegetación, la pendiente y la altitud, también juegan un papel crucial en la probabilidad de ocurrencia de incendios. Adicionalmente, los factores socioeconómicos, como la densidad poblacional y las actividades humanas, pueden incrementar el riesgo y la frecuencia de incendios.

Actualmente, la capacidad de respuesta ante incendios forestales en Quito se ve limitada por la falta de información precisa y en tiempo real sobre las condiciones meteorológicas y geoespaciales, así como por el desarrollo insuficiente de técnicas predictivas avanzadas. Esta situación reduce la eficacia de las medidas preventivas y la capacidad de mitigar los daños causados por los incendios forestales.

La pregunta central que guía este estudio es: ¿Cómo los factores ambientales, geoespaciales y socioeconómicos influyen en la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito?

Responder a esta pregunta es crucial para mejorar la prevención y gestión de incendios forestales en Quito. La investigación busca desarrollar modelos de análisis de datos que integren estos factores para predecir la ocurrencia de incendios, proporcionando así una herramienta valiosa para la toma de decisiones y la planificación estratégica en la gestión de riesgos. Este estudio no solo contribuirá a la protección de los recursos naturales y la infraestructura urbana, sino que también fortalecerá la resiliencia de la comunidad ante eventos catastróficos.

1.3 Objetivos

1.3.1 **Objetivo General. –** Desarrollar un modelo de análisis de datos que permita predecir la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito, basado en factores ambientales, geoespaciales y socioeconómicos.

1.3.2 **Objetivos Específicos. –**

- Recopilar y analizar datos de incendios forestales del DMQ, variables

ambientales, geoespaciales y socioeconómicas.

- Realizar un análisis exploratorio de datos que permita comprender mejor la estructura, distribución de los datos y correlaciones significativas.
- Utilizar técnicas de simulación de datos que permitan manejar el desbalance de datos.
- Generar los modelos predictivos y evaluar el rendimiento de los mismos.
- Identificar la posible ocurrencia de incendios forestales y los factores que contribuyen a estos riesgos.

1.4 Alcance

El estudio se centrará en desarrollar modelos de análisis de datos avanzado para predecir la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito. Estos modelos integrarás factores ambientales, geoespaciales y socioeconómicos para ofrecer predicciones precisas y útiles, facilitando la planificación y la toma de decisiones estratégicas en la gestión de riesgos. El alcance del estudio abarca varias etapas clave, cada una con objetivos específicos y resultados esperados.

En la fase inicial, se realizará una recopilación exhaustiva de datos relevantes. Los datos ambientales incluirán variables como temperatura, humedad relativa, velocidad del viento y precipitación. Los datos geoespaciales abarcarán el uso del suelo, tipo de vegetación, pendiente y altitud. Además, se recogerán datos socioeconómicos relacionados con la densidad poblacional que puede influir en la ocurrencia de incendios forestales. Esta recopilación se llevará a cabo mediante la integración de diversas fuentes, incluyendo bases de datos oficiales, imágenes satelitales y sistemas de información geográfica (SIG).

Posteriormente, se llevará a cabo un análisis exploratorio de datos para comprender mejor la estructura y distribución de los datos recopilados. Este análisis incluirá la

limpieza de los datos, así como la identificación de correlaciones significativas entre las variables. Se emplearán técnicas avanzadas de visualización de datos para identificar patrones y anomalías que puedan influir en la ocurrencia de incendios forestales.

La etapa de desarrollo de modelos predictivos implicará la implementación de varias técnicas de análisis de datos y aprendizaje automático. Se utilizarán métodos como la regresión logística, los árboles de decisión y las redes neuronales para crear modelos que puedan predecir con precisión la probabilidad de incendios forestales. Cada modelo será evaluado y validado utilizando métricas de desempeño como precisión, recall, F1-score y área bajo la curva ROC (AUC-ROC). Se seleccionará el modelo con mejor desempeño para su implementación práctica. Adicional, se generará un modelo de máxima entropía para determinar la probabilidad espacial de ocurrencia de incendios forestales en el DMQ.

CAPÍTULO II

2. MARCO TEÓRICO

El presente capítulo incluye el marco teórico y conceptual del estudio de investigación. Se abordarán conceptos esenciales de ciencias de datos, técnicas de Machine Learning, máxima entropía, amenaza de incendios forestales.

Ciencias de Datos

Las ciencias de datos es una disciplina que combina técnicas de aprendizaje automático, estadística, análisis avanzado, minería de datos, big data y programación con el objetivo de extraer conocimiento útil y valioso a partir de grandes volúmenes de datos. Esta área de estudio busca descubrir patrones y comportamientos en los datos para tomar decisiones informadas o hacer predicciones. La ciencia de datos ha experimentado un crecimiento significativo debido a la disponibilidad creciente de datos y la necesidad de métodos avanzados para su análisis, abarcando diversas áreas como

computación, matemáticas, ingeniería, gestión de riesgo de desastres, entre otros (Jesús García & A. Patricio, Álvaro L. Bustamante y Washington R. , 2018)

Machine Learning

El machine learning o aprendizaje automático es un campo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos estadísticos que permiten a las computadoras realizar tareas específicas sin ser explícitamente programadas para ello. Utiliza técnicas de clasificación, agrupamiento, asociación o decisión en función de características o patrones presentes en los datos. Mediante el análisis y aprendizaje de grandes volúmenes de datos, los algoritmos de machine learning pueden identificar patrones y tomar decisiones basadas en ejemplos previos, mejorando su desempeño con el tiempo a medida que reciben más datos (Pardo, 2016).

Árboles de decisión

Los árboles de decisión son modelos no paramétricos de aprendizaje supervisado utilizados tanto para problemas de clasificación como de regresión. Se construyen mediante un proceso de partición binaria recursiva, donde en cada paso del entrenamiento se realizan divisiones sucesivas de un conjunto de datos en subconjuntos, aplicando decisiones basadas en una de las variables. Este proceso continúa hasta que se alcanza un punto de parada predefinido, obteniendo así el clasificador por árbol de decisión (Arana, 2021).

En el caso de la clasificación, los árboles de decisión asignan una clase específica a cada nodo terminal del árbol, basándose en la mayoría de los casos que componen dicho nodo. Para la predicción de nuevas instancias, el modelo recorre las sucesivas ramificaciones del árbol, tomando decisiones en cada nodo, hasta llegar a un nodo terminal que proporciona la clase predicha (Arana, 2021).

Estos modelos son apreciados por su interpretabilidad y por ser la base de modelos más avanzados como los ensambles de árboles de decisión. Los primeros desarrollos en

este campo fueron realizados por investigadores como Leo Breiman y Jerome Friedman, quienes también crearon los ensambles de árboles de decisión, mejorando la performance de los árboles individuales (Arana, 2021).

Modelo logístico

La regresión logística es una técnica estadística utilizada en machine learning dentro del paradigma del aprendizaje supervisado, cuya finalidad principal es realizar clasificaciones. Este modelo permite tomar una o más variables independientes y determinar la probabilidad de ocurrencia de un fenómeno específico. Por ejemplo, se puede utilizar para predecir la probabilidad de que un estudiante apruebe o repruebe una materia basada en características como la edad de ingreso y las notas obtenidas en semestres anteriores (Nelson Becerra Correa, Miguel Leguizamón Páez, 2024).

La regresión logística es particularmente adecuada cuando se necesita modelar una variable de respuesta binaria, permitiendo el uso conjunto de covariables de tipo categórico y continuo, proporcionando además una interpretación biológica y práctica de sus parámetros (Nelson Becerra Correa, Miguel Leguizamón Páez, 2024).

Redes neuronales

Una red neuronal es un modelo computacional inspirado en el funcionamiento del cerebro humano, que se utiliza en el campo del aprendizaje automático y la inteligencia artificial para identificar patrones y hacer predicciones. Las redes neuronales están compuestas por capas de nodos o "neuronas", donde cada capa está conectada a la siguiente a través de "pesos" que ajustan la fuerza de las conexiones. Estas capas incluyen una capa de entrada, una o más capas ocultas, y una capa de salida (Correa, Bielza, Pamies-Texeira, & Alique López, 2008).

Cada neurona recibe señales de entrada, las procesa aplicando una función de activación, y luego transmite una señal de salida a las neuronas de la siguiente capa. Durante el proceso de entrenamiento, los pesos de las conexiones se ajustan

iterativamente para minimizar el error entre la salida predicha y la salida deseada, utilizando algoritmos como la retropropagación del error (Correa, Bielza, Pamies-Texeira, & Alique López, 2008).

Las redes neuronales son especialmente útiles para tareas donde las relaciones entre las variables no son lineales y requieren una aproximación más flexible. Se utilizan en una variedad de aplicaciones, incluyendo reconocimiento de patrones, clasificación, regresión entre otros (Correa, Bielza, Pamies-Texeira, & Alique López, 2008).

Máxima entropía MaxEnt

El principio de máxima entropía se utiliza para estimar una distribución de probabilidad sobre un espacio muestral finito (Maya, 2017).

El principio de máxima entropía sugiere que, entre todas las distribuciones que satisfacen ciertas restricciones, se elija la que tenga la máxima entropía, es decir, la más cercana a la distribución uniforme (Maya, 2017). La entropía de una distribución ppp sobre un conjunto XXX se define como:

$H(p)=-\sum_{x \in X} p(x) \ln p(x)$

El enfoque de máxima entropía se utiliza para estimar una distribución de probabilidad sobre un espacio muestral finito XXX. Se busca una aproximación ppp bajo la cual la esperanza de ciertas funciones características coincida con sus valores empíricos observados (Maritza Lucía Vaca Cárdenas, Byron Ernesto Vaca Barahona, Diego Francisco , & Guicela Margoth Ati Cutiupala, 2021).

Incendios Forestales

Un incendio forestal es un fuego no controlado que se propaga a través de la vegetación natural, ya sea arbustiva, leñosa o herbácea, viva o muerta, y que puede causar daños significativos a personas, propiedades y el medio ambiente. Los incendios forestales pueden ser provocados por una variedad de factores, tanto naturales como

antropogénicos. Entre las causas naturales se incluyen las sequías prolongadas y las descargas eléctricas por rayos, mientras que las causas antropogénicas abarcan actividades humanas como quemas agrícolas, negligencias y actos intencionales (Jairo Estacio & Nixon Narváez, 2012).

Estos incendios no sólo afectan a las zonas rurales, sino que también pueden tener un impacto significativo en áreas urbanas, particularmente en aquellas periurbanas donde la interfaz entre áreas naturales y asentamientos humanos es más pronunciada. Los efectos de los incendios forestales son multifacéticos, afectando no solo a la flora y fauna, sino también al suelo, la calidad del aire y el clima global. Además, el cambio climático ha exacerbado la frecuencia y severidad de estos incendios, creando un ciclo de retroalimentación que incrementa las temperaturas y reduce la humedad del suelo, haciendo que las condiciones sean más propicias para futuros incendios (Jairo Estacio & Nixon Narváez, 2012).

Los incendios forestales tienen profundas consecuencias ambientales, económicas y sociales. Ambientalmente, pueden causar la pérdida de biodiversidad, la degradación del suelo y la emisión de grandes cantidades de gases de efecto invernadero, lo que a su vez contribuye al cambio climático. Económicamente, los incendios forestales pueden destruir recursos forestales valiosos y afectar negativamente a industrias dependientes del bosque, como la agricultura y el turismo. Socialmente, representan un riesgo significativo para la vida y la propiedad humana, causando desplazamientos y afectando la salud pública debido al humo y la contaminación del aire (Jairo Estacio & Nixon Narváez, 2012).

La gestión eficaz de los incendios forestales requiere un enfoque integrado que incluya la prevención, la preparación, la respuesta y la recuperación. Esto implica la implementación de planes de prevención y respuesta ante incendios forestales, la restauración de ecosistemas afectados, y la educación y concienciación pública sobre las causas y riesgos de los incendios forestales. La utilización de tecnología avanzada,

como sistemas de información geográfica y ciencias de datos, también es crucial para mejorar la detección temprana y la respuesta eficaz a los incendios (Pazmiño, 2019).

Gestión del Riesgo de Desastres

La gestión del riesgo de desastres es un conjunto de procesos que incluyen la toma de conocimiento, previsión, prevención, reducción, mitigación, preparación, respuesta, rehabilitación, y recuperación de cara a un desastre o catástrofe. Este enfoque abarca medidas estructurales y no estructurales, así como el desarrollo del conocimiento de los factores subyacentes y condicionantes del riesgo, y el fortalecimiento de las capacidades para formular y mejorar las políticas públicas en esta materia. La gestión del riesgo de desastres busca reducir de manera continua y consistente la ocurrencia de desastres, mejorar la respuesta humanitaria y fortalecer la resiliencia de los territorios. En todos los procesos de la gestión integral del riesgo de desastres se aplican principios como la descentralización subsidiaria y se garantiza la participación ciudadana y de las organizaciones de la sociedad civil (Asamblea Nacional Constituyente, 2024).

Modelos Determinísticos

Hace referencia a modelos en los que los resultados están completamente determinados por las condiciones iniciales y las reglas o leyes predefinidas y no incluye ninguna forma de aleatoriedad o probabilidad. Esto implica que, dadas las mismas condiciones iniciales, se producirán los mismos resultados (OpenCourseWare., 2004).

Modelos Estocásticos

Son modelos que incluyen componentes aleatorios o impredecibles. En estos casos, no se pueden predecir los resultados exactos, solo se pueden establecer probabilidades de ocurrencia. Los modelos estocásticos consideran la variabilidad o probabilidad en los datos o procesos modelados, lo cual ofrece ser una mejor representación de los sistemas naturales (NIST/SEMATECH, 2012).

Aprendizaje supervisado

Es una técnica de Machine Learning donde se entrena un modelo con un conjunto de datos etiquetados, es decir, datos de entrada ya conocidos y los resultados deseados. El objetivo es que el modelo aprenda a predecir los resultados a partir de los datos de entrada. A lo largo del proceso de entrenamiento, el modelo ajusta sus parámetros para minimizar la diferencia entre sus predicciones y los resultados reales. Una vez entrenado, el modelo puede hacer predicciones sobre datos nuevos (Méndez, 2018).

Modelos de clasificación

La clasificación supervisada es una técnica de aprendizaje automático en la que un modelo se entrena utilizando un conjunto de datos etiquetado, donde cada entrada está asociada a una clase conocida. El objetivo del modelo es aprender a mapear las entradas a sus respectivas etiquetas para que, una vez entrenado, pueda predecir con precisión la clase de nuevas entradas desconocidas. El proceso implica dos fases: entrenamiento, donde el modelo ajusta sus parámetros para minimizar la diferencia entre sus predicciones y las etiquetas reales, y predicción, donde el modelo se utiliza para clasificar nuevos datos, evaluando su rendimiento en función de su capacidad para predecir correctamente las clases en datos no vistos (Brownlee, 2019).

CAPÍTULO III

3. METODOLOGÍA

En el presente capítulo se describirán las variables utilizadas y los métodos escogidos para el desarrollo de los modelos de predicción de incendios forestales en el DMQ, así como también se describirán las condiciones biofísicas del Distrito Metropolitano de Quito. Para este ejercicio se utilizó la metodología CRISP-DM (Figura 1), reconocido como un marco estándar para la Minería de Datos y el proceso de descubrimiento de conocimiento en datos, la cual ofrece una guía estructurada y fases claramente definidas que dividen el proyecto en etapas manejables (IBM, 2021).

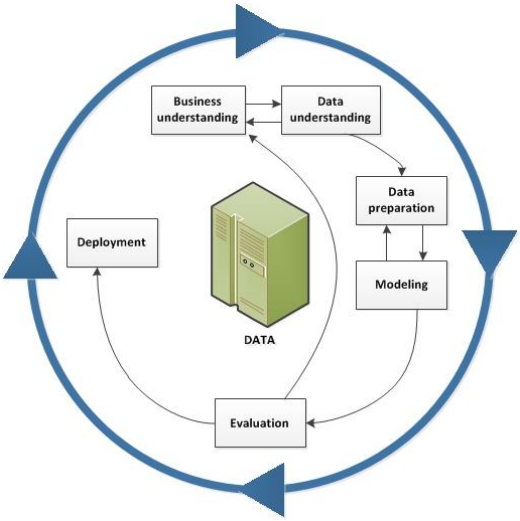

Figura 1. Metodología CRISPDM

Fuente: IBM, 2021.

3.1 Descripción del área de estudio

3.1.1 Ubicación del área de estudio

El área de estudio corresponde al Distrito Metropolitano de Quito ubicado en la provincia de Pichincha, en la región andina norte del Ecuador (Figura 2). Su jurisdicción corresponde a la capital del país y se encuentra delimitada política y administrativamente de la siguiente forma: al norte, por los cantones Cotacachi y Otavalo, en la provincia de Imbabura. Al este, por los cantones Pedro Moncayo y Cayambe, pertenecientes a Pichincha, y por los cantones de El Chaco, Quijos y Archidona, en la provincia de Napo. Al oeste, por los cantones de San Miguel de los Bancos, Pedro Vicente Maldonado, de Pichincha, y Santo Domingo, de la provincia de Santo Domingo de los Tsáchilas. Y, finalmente, al sur por los cantones de Rumiñahui y Mejía, que forman parte de su conurbano (Quito, 2024). El cantón cuenta con una superficie de 420 091,49 ha, y una población de 2 679 722 habitantes (INEC, 2022).

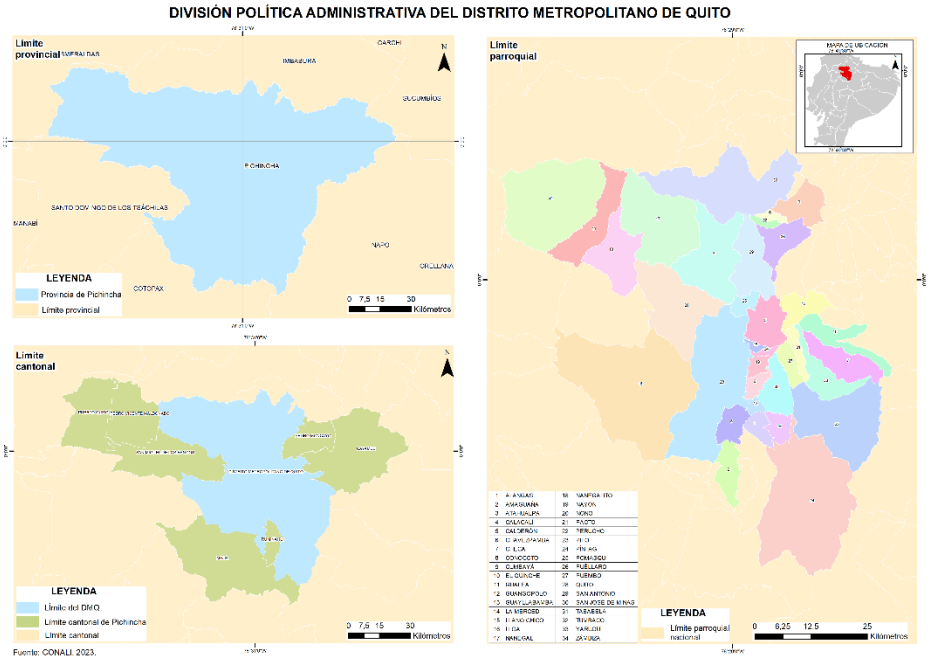

Figura 2. División Política Administrativa del DMQ

Elaboración: Propia

3.1.2 Descripción de características biofísicas del DMQ

3.1.2.1 Clima

En el Distrito Metropolitano de Quito predomina el clima subhúmedo, mesotérmico templado frío (Figura 3).

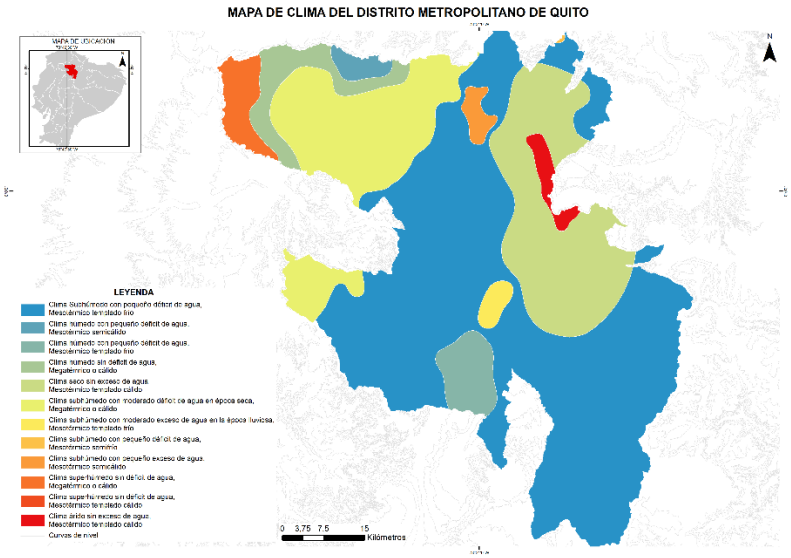

Figura 3. Tipo de clima del DMQ
Elaboración: Propia

3.1.2.2 Temperatura

La temperatura del DMQ fluctúa entre 5 y 25 grados centígrados a lo largo del territorio cantonal.

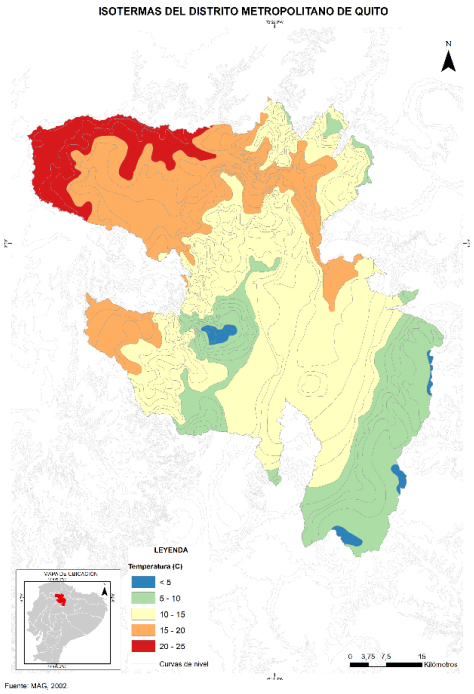

Figura 4. Temperatura del DMQ
Elaboración: Propia

3.1.2.3 Precipitación

La precipitación del DMQ fluctúa entre 1000 y 3700 milímetros cúbicos a lo largo del territorio cantonal (Figura 5).

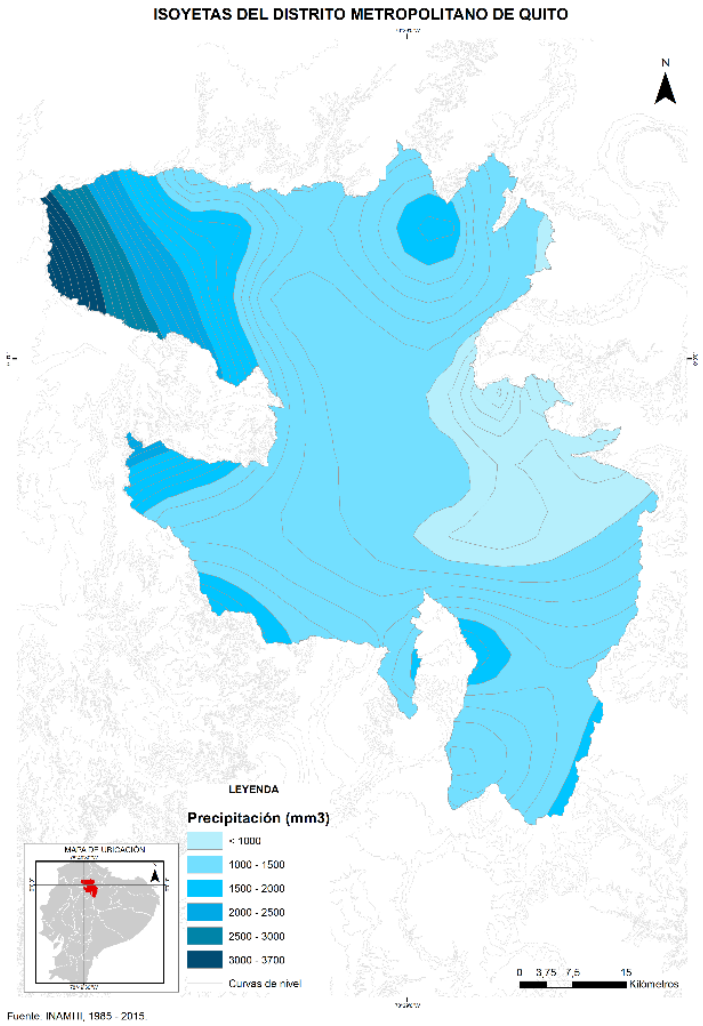

Figura 5. Isoyetas del DMQ
Elaboración: Propia

3.1.2.4 Cobertura y uso de suelo

La cobertura y uso predominante del DMQ es la vegetación arbustiva y herbácea, así como la zona agropecuaria (Figura 6).

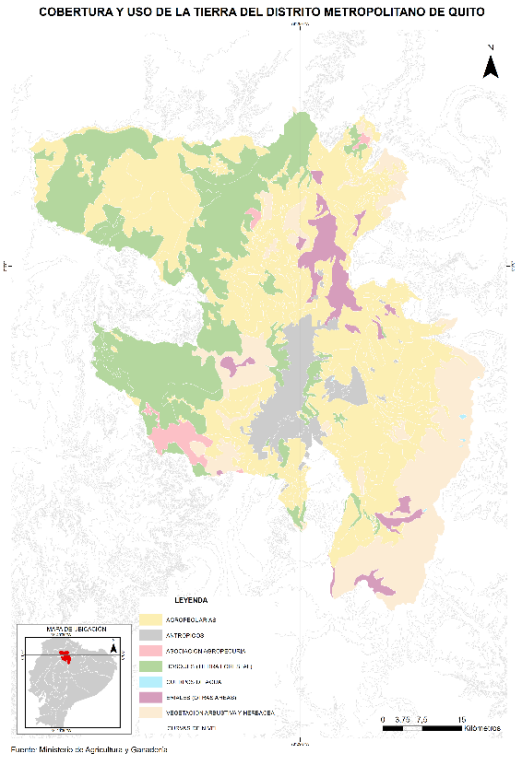

Figura 6. Cobertura y Uso del Suelo del DMQ
Elaboración: Propia

3.1.2.6 Densidad poblacional del DMQ

La densidad poblacional en la zona urbana va de 3000 a 5000 habitantes por kilómetro cuadrado, mientras que en la zona rural, la densidad fluctúa entre 4 y 3000 habitantes por kilómetro cuadrado.

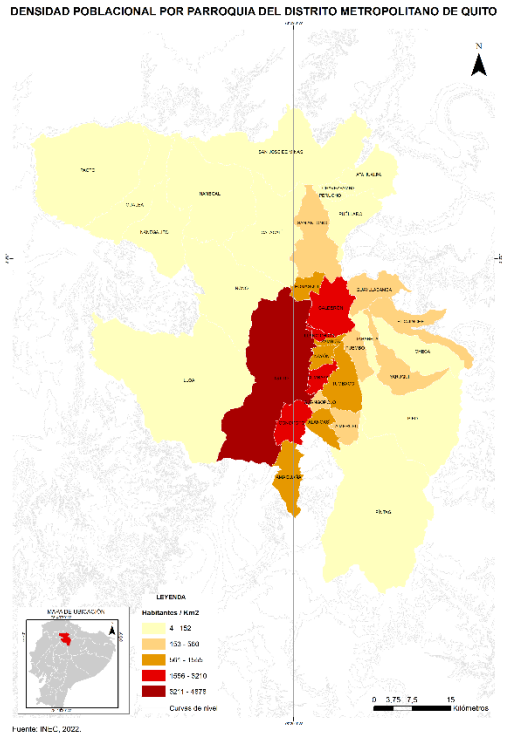

Figura 7. Densidad Poblacional del DMQ
Elaboración: Propia

3.2 Metodología CRISP-DM

3.2.1 Comprensión del negocio

3.2.1.1 Definición de los Objetivos del Proyecto

El Distrito Metropolitano de Quito ha experimentado un aumento significativo en la frecuencia e intensidad de los incendios forestales, atribuido a factores ambientales, geoespaciales y socioeconómicos, creando una amenaza grave para la vida humana, la biodiversidad y la infraestructura, con pérdidas económicas y ambientales considerables. El objetivo principal del proyecto es desarrollar un modelo de análisis de datos que permita predecir la ocurrencia de incendios forestales en el Distrito Metropolitano de Quito, basado en factores ambientales, geoespaciales y socioeconómicos, mediante la recopilación y análisis de datos históricos y variables relevantes, la realización de un análisis exploratorio de datos, el manejo de datos desbalanceados, la generación y evaluación de modelos predictivos, incluyendo un modelo de máxima entropía, la identificación de factores de riesgo y el desarrollo de

herramientas de visualización para facilitar la comprensión y utilización de los resultados del modelo por parte de las autoridades locales y los actores involucrados en la gestión de riesgos.

3.2.1.2 Evaluación de la situación

El Servicio Integrado de Seguridad ECU 911 tiene como objetivo gestionar la atención de las situaciones de emergencia de la ciudadanía, las que se generen por video vigilancia y monitoreo de alarmas, mediante el despacho de recursos de respuesta especializados pertenecientes a organismos públicos y privados articulados al sistema en todo el territorio nacional (GOB.EC, 2024).

Por otra parte, el Instituto Nacional de Meteorología e Hidrología – INAMHI es la entidad técnico - científica responsable en el Ecuador de la generación y difusión de la información hidrometeorológica que sirva de sustento para la formulación y evaluación de los planes de desarrollo nacionales y locales y la realización de investigación propia o por parte de otros actores (GOB.EC, 2024) . La información hidrometeorológica para el presente estudio se obtiene de las diversas estaciones meteorológicas ubicadas a lo largo del territorio del Distrito Metropolitano de Quito, tal como se muestra en la siguiente imagen (Figura 8).

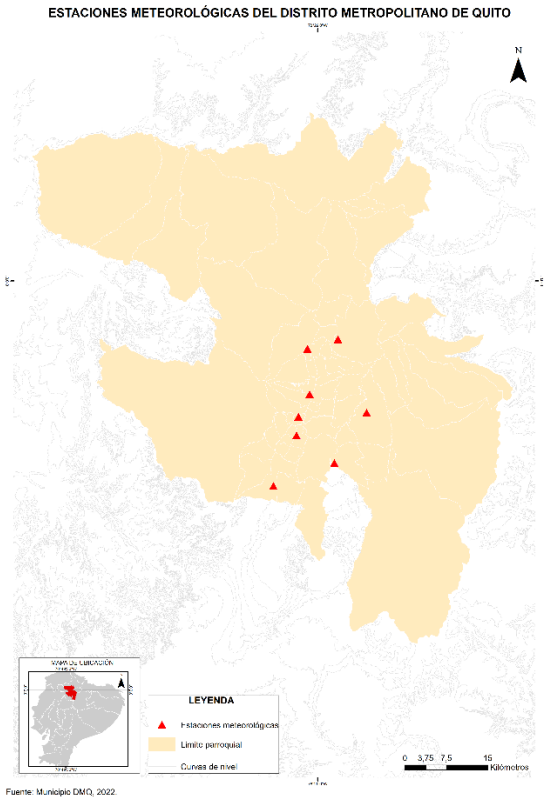

Figura 8. Ubicaciones de estaciones meteorológicas en el DMQ
Elaboración: Propia

En el presente proyecto existen ciertas restricciones respecto al acceso de la información que se utilizará, ya que, aunque existen datos mensuales liberados en la página institucional del INAMHI, para obtener datos diarios se debe cancelar un valor específico por variable y por día solicitado, lo cual puede representar un costo alto para la realización de la investigación. Otro riesgo que puede presentarse en la ejecución del proyecto es la presencia de sesgo en los datos o en el modelo ocurrido por valores atípicos en las condiciones meteorológicas, o por factores antrópicos no incluidos en el presente estudio.

3.2.1.3 Plan del Proyecto

Tabla 1. Fases del plan del proyecto

Fase	Duración	Recursos	Riesgos
Comprensión del negocio	4 semanas	Responsable del proyecto	
Comprensión de los datos	4 semanas	Acceso a los datos	

Preparación de los datos	2 semanas	Plataforma para el modelado (Rstudio)	Disponibilidad total de datos
Modelado	6 semanas	Plataforma para el modelado (Rstudio)	Dificultad de ejecución de modelos
Evaluación	4 semanas	Plataforma para el modelado (Rstudio)	Baja precisión de los modelos generados

Fuente: Elaboración propia.

3.2.2 Comprensión de los datos

Esta fase implica la recopilación y el análisis preliminar de los datos disponibles para asegurar su relevancia y calidad. Los datos históricos de incendios forestales, así como los datos ambientales (temperatura, humedad, viento, precipitación), geoespaciales (uso del suelo, tipo de vegetación, pendiente, altitud) y socioeconómicos (densidad poblacional) fueron recopilados de diversas fuentes, incluyendo bases de datos oficiales, imágenes satelitales y sistemas de información geográfica (SIG).

3.2.2.1 Recolección de datos de incendios forestales y condiciones climáticas, ambientales y geoespaciales

La recolección de datos es un componente esencial en cualquier investigación científica, y en este estudio se llevó a cabo una exhaustiva recopilación de información. Este proceso abarcó un período de tiempo significativo, comprendido entre los años 2020 y 2024, con el objetivo de capturar un rango temporal lo suficientemente amplio como para permitir un análisis robusto y detallado de los incendios forestales y las condiciones ambientales que los rodean.

El conjunto de datos recopilado incluye una serie de variables biofísicas y geográficas que son fundamentales para el análisis de los incendios forestales. Estas variables fueron cuidadosamente seleccionadas para abarcar diferentes aspectos que pudieran influir en la ocurrencia y propagación de incendios.

3.2.2.1.1 Variables utilizadas

Para el presente estudio se utilizaron variables ambientales, geoespaciales y socioeconómicas, las cuales son:

Tabla 2. Variables de estudio

Variable	Tipo
Fecha	POSIXct
Longitud	numérica
Latitud	numérica
Provincia	categórica
Cantón	Categórica
Parroquia	Categórica
Distrito	categórica
Circuito	categórica
Tipo de vegetación	categórica
Uso_de_suelo	categórica
Densidad_pob	numérica
Pendiente	categórica
Altitud_msnm	categórica
Temperatura_media	numérica
Humedad_Relativa	numérica
Velocidad_del_viento	numérica
Precipitación	numérica
Subtipo	categórica

3.2.2.1.2 Dataset de incendios forestales

En la etapa de recolección de información se solicitó los datos de incendios forestales al Servicio Integrado de Seguridad Ecu 911, quienes proporcionaron información de ocurrencia de incendios forestales desde enero del 2020 hasta mayo del 2024, la data proporcionada contiene atributos sobre fecha del evento, hora, longitud, latitud, provincia, cantón, parroquia, y subtipo de evento (Figura 9).

```
> datos_incendios
# A tibble: 1,929 × 13
   Fecha               Hora  Longitud Latitud Provincia Cantón Parroquia  Distrito
   <dttm>              <chr>    <dbl>   <dbl> <chr>     <chr>  <chr>      <chr>
 1 2020-01-01 00:00:00 12       -78.5 -0.153  PICHINCHA QUITO  QUITO      EUGENIO ESP…
 2 2020-01-05 00:00:00 17       -78.4 -0.329  PICHINCHA QUITO  ALANGASI   LOS CHILLOS
 3 2020-01-06 00:00:00 15       -78.4 -0.209  PICHINCHA QUITO  PUEMBO     TUMBACO
 4 2020-01-08 00:00:00 15       -78.5 -0.121  PICHINCHA QUITO  QUITO      EUGENIO ESP…
 5 2020-01-08 00:00:00 20       -78.4 -0.213  PICHINCHA QUITO  TUMBACO    TUMBACO
 6 2020-01-09 00:00:00 10       -78.3 -0.195  PICHINCHA QUITO  TABABELA   TUMBACO
 7 2020-01-10 00:00:00 12       -78.4 -0.260  PICHINCHA QUITO  GUANGOPOLO LOS CHILLOS
 8 2020-01-10 00:00:00 15       -78.5 -0.0942 PICHINCHA QUITO  QUITO      LA DELICIA
 9 2020-01-10 00:00:00 15       -78.5 -0.204  PICHINCHA QUITO  QUITO      EUGENIO ESP…
10 2020-01-10 00:00:00 18       -78.5 -0.156  PICHINCHA QUITO  QUITO      EUGENIO ESP…
# ℹ 1,919 more rows
# ℹ 5 more variables: Circuito <chr>, Subcircuito <chr>, Servicio <chr>,
#   Subtipo <chr>, Emergencias <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

Figura 9. Dataset de incendios forestales en el DMQ

3.2.2.1.3 Dataset de condiciones climáticas

Para la recolección de datos ambientales como la temperatura, humedad, viento y precipitación se solicitó información al Instituto Nacional de Meteorología e Hidrología – INAMHI, quienes proporcionaron datos históricos del DMQ con el fin de establecer patrones climáticos incidentes en la ocurrencia de los incendios forestales (Figura 10).

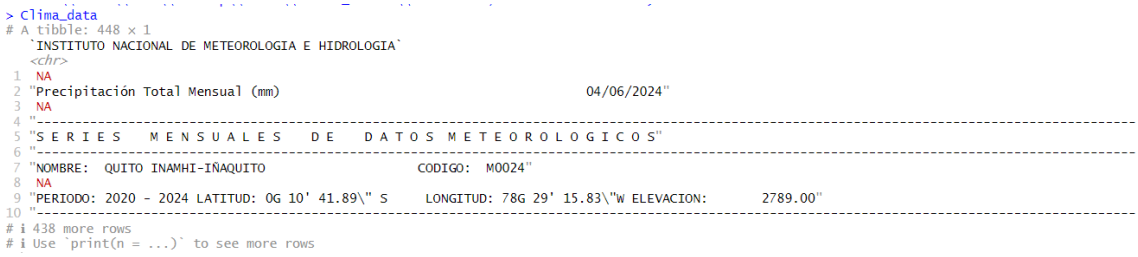

Figura 10. Dataset de condiciones climáticas

3.2.2.1.4 Simulación de Datos Faltantes

Uno de los desafíos más significativos identificados en el conjunto de datos fue la falta de datos completos en las variables climatológicas proporcionadas por el INAMHI para ciertos días debido a su alto costo de adquisición. La incompletitud de los datos puede ser un problema serio en el análisis, ya que puede introducir sesgos y reducir la robustez de los modelos predictivos. Para abordar este problema, se implementó un método de simulación de datos faltantes utilizando una distribución normal basada en la media y la desviación estándar de cada sitio.

Este enfoque de simulación de datos faltantes se justificó por varias razones:

- **Mantenimiento de la Coherencia Estadística:** La simulación de datos mediante una distribución normal asegura que los valores imputados sigan la misma distribución estadística que los datos observados. Esto es fundamental para evitar introducir sesgos que puedan comprometer la validez del análisis. La

preservación de las propiedades estadísticas de los datos originales es crucial para garantizar que los resultados del análisis sean representativos y fiables.

- **Aplicación de Conocimientos Locales:** Al utilizar la media y la desviación estándar específicas de cada distrito, se garantiza que los valores simulados reflejen las condiciones locales y regionales. Este enfoque es esencial para capturar la variabilidad en las condiciones climáticas y biofísicas a lo largo del territorio. La consideración de las particularidades locales es importante para evitar la homogenización de los datos y para asegurar que las simulaciones reflejen de manera precisa las condiciones reales en cada área.
- **Robustez del Análisis:** Simular datos faltantes en lugar de eliminarlos o imputarlos con valores promedio simples permite mantener la totalidad del conjunto de datos. Esto es especialmente importante cuando se trabaja con modelos predictivos, ya que garantiza que se cuente con un conjunto de datos representativo que incluya toda la variabilidad existente. La retención de todos los registros, incluso aquellos con datos imputados, permite realizar análisis más completos y precisos, y evita la pérdida de información valiosa que podría ser crucial para la interpretación de los resultados.

3.2.2.1.5 Información geoespacial

La información geoespacial fue recopilada de diversas fuentes de información a nivel nacional. La cual fue enfocada en la zona de estudio mediante la utilización de Sistemas de Información Geográfica. Esta información se detalla a continuación:

Tabla 3. Variables geoespaciales

Variable	Fuente	Año
Uso del suelo (Cobertura de la tierra)	Ministerio de Agricultura y Ganadería	2020
Tipo de vegetación	Ministerio de Agricultura y Ganadería	2020
Pendiente	Elaborado a partir de un Modelo Digital del Terreno del DMQ	2018

Altitud	Recopilado de un Modelo Digital del Terreno del DMQ	2018

Shape *	codigo	descripcio	temporalid	cobertura	uso
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Ae	AREA EROSIONADA	NO APLICABLE	ERIALES (OTRAS AREAS)	TIERRAS IMPRODUCTIVAS
Polygon	Bi-Pc	50% BOSQUE INTERVENIDO - 50% PASTO CULTIVADO	NO APLICABLE	ASOCIACION AGROPECURIA	PECUARIO - CONSERVACION Y PROTECCION
Polygon	Bi-Pc	50% BOSQUE INTERVENIDO - 50% PASTO CULTIVADO	NO APLICABLE	ASOCIACION AGROPECURIA	PECUARIO - CONSERVACION Y PROTECCION
Polygon	Bi-Pc	50% BOSQUE INTERVENIDO - 50% PASTO CULTIVADO	NO APLICABLE	ASOCIACION AGROPECURIA	PECUARIO - CONSERVACION Y PROTECCION
Polygon	Bi-Pn	50% BOSQUE INTERVENIDO - 50% PASTO NATURAL	NO APLICABLE	ASOCIACION AGROPECURIA	CONSERVACION Y PROTECCION
Polygon	Bi/Cc	70% BOSQUE INTERVENIDO / 30% CULTIVOS DE CICLO CORTO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Cx	70% BOSQUE INTERVENIDO / 30% ARBORICULTURA TROPICAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Cx	70% BOSQUE INTERVENIDO / 30% ARBORICULTURA TROPICAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bi/Pc	70% BOSQUE INTERVENIDO / 30% PASTO CULTIVADO	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	AGROPECUARIO FORESTAL
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION
Polygon	Bn	BOSQUE NATURAL	NO APLICABLE	BOSQUES (TIERRA FORESTAL)	CONSERVACION Y PROTECCION

Figura 11. Base de datos de información geoespacial

3.2.2.1.6 Información socioeconómica

Para la obtención de la densidad poblacional se utilizó la información de población por sector censal del Censo de Población y Vivienda realizado por el INEC en el año 2022 (Figura 12).

Shape *	DPA_PARROQ	DPA_DESPAR	DPA_CANTON	Densidad_p	Total
Polygon	170166	LLOA	1701	4	1959
Polygon	170172	PACTO	1701	11	3898
Polygon	170168	NANEGAL	1701	12	2959
Polygon	170161	GUALEA	1701	14	1704
Polygon	170171	NONO	1701	14	2938
Polygon	170181	SAN JOSÉ DE MINAS	1701	19	5589
Polygon	170153	ATAHUALPA	1701	21	1446
Polygon	170169	NANEGALITO	1701	22	2776
Polygon	170154	CALACALÍ	1701	26	4964
Polygon	170176	PÍNTAG	1701	48	23240
Polygon	170158	CHAVEZPAMBA	1701	50	579
Polygon	170178	PUÉLLARO	1701	71	5155
Polygon	170174	PERUCHO	1701	72	704
Polygon	170175	PIFO	1701	91	23202
Polygon	170159	CHECA	1701	131	11492
Polygon	170183	TABABELA	1701	152	3851
Polygon	170160	EL QUINCHE	1701	244	18485
Polygon	170164	LA MERCED	1701	362	11438
Polygon	170185	YARUQUÍ	1701	364	26564
Polygon	170163	GUAYLLABAMBA	1701	370	20584
Polygon	170162	GUANGOPOLO	1701	426	4336
Polygon	170180	SAN ANTONIO	1701	448	49984
Polygon	170179	PUEMBO	1701	560	17780
Polygon	170152	AMAGUAÑA	1701	767	43235
Polygon	170186	ZÁMBIZA	1701	814	6160
Polygon	170151	ALANGASÍ	1701	1179	34655
Polygon	170184	TUMBACO	1701	1207	79109
Polygon	170170	NAYÓN	1701	1247	22065
Polygon	170177	POMASQUI	1701	1555	36883
Polygon	170157	CUMBAYÁ	1701	1996	41819
Polygon	170165	LLANO CHICO	1701	2003	15113
Polygon	170156	CONOCOTO	1701	2859	127815
Polygon	170155	CALDERÓN	1701	3210	250877
Polygon	170150	QUITO	1701	4878	1776364

Figura 12. Base de datos de densidad poblacional

3.2.2.1.7 Integración, homologación y estandarización de los datos

Para realizar un análisis exhaustivo, se llevó a cabo un cruce de estos datos con las fechas y ubicaciones geográficas de los registros de la base de datos biofísica. Este cruce se realizó tanto para los días en los que se reportaron incendios como para los días en los que no se registraron incendios. La integración de los datos climatológicos con los datos biofísicos permitió construir un conjunto de datos completo que reflejaba tanto las condiciones bajo las cuales se produjeron incendios como las condiciones en días sin incendios. Esta estrategia metodológica es crucial para establecer relaciones entre las variables climáticas y la ocurrencia de incendios, facilitando un análisis más profundo y completo.

El cruce de datos permitió no solo la identificación de correlaciones entre las condiciones meteorológicas y la ocurrencia de incendios, sino también la posibilidad de modelar la probabilidad de incendios bajo diferentes escenarios climáticos, y para la planificación de estrategias de mitigación y respuesta ante incendios forestales (Figura 13).

Para el proceso de integración de información se utilizó la herramienta de intersección del Sistema de Información Geográfica, mediante la cual se integró la información de incendios con la información climática, geoespacial y socioeconómica, obteniendo como resultado el siguiente dataset:

```
> datos_incendios
# A tibble: 1,744 × 18
   Fecha               Longitud Latitud Provincia Cantón Parroquia         Distrito Circuito `Tipo de vegetación` Uso_de_suelo Densidad_pob Pendiente Altitud_msnm Temperatura_media
   <dttm>                 <dbl>   <dbl> <chr>     <chr>  <chr>             <chr>    <chr>    <chr>               <chr>               <dbl> <chr>     <chr>                      <dbl>
 1 2021-01-01 00:00:00    -78.5 -0.265  PICHINCHA QUITO  CONOCOTO          LOS CHI… CONOCOT… URBANO              ANTROPICO              18 11 - 25   2421 - 2741                14
 2 2021-01-02 00:00:00    -78.5 -0.190  PICHINCHA QUITO  QUITO             EUGENIO… IÑAQUITO URBANO              ANTROPICO              44 0 - 10    2742 - 3067                14
 3 2021-01-03 00:00:00    -78.6 -0.281  PICHINCHA QUITO  QUITO             QUITUMBE LAS CUA… URBANO              ANTROPICO              44 0 - 10    2742 - 3067                11
 4 2021-01-03 00:00:00    -78.4 -0.226  PICHINCHA QUITO  CUMBAYA           TUMBACO  CUMBAYA… URBANO              ANTROPICO              15 0 - 10    2042 - 2420                10
 5 2021-01-03 00:00:00    -78.5 -0.111  PICHINCHA QUITO  QUITO             LA DELI… PONCIANO URBANO              ANTROPICO              44 0 - 10    2742 - 3067                12
 6 2021-01-03 00:00:00    -78.5 -0.0958 PICHINCHA QUITO  QUITO             LA DELI… COLINAS… URBANO              ANTROPICO              44 11 - 25   2742 - 3067                16
 7 2021-01-03 00:00:00    -78.4 -0.0807 PICHINCHA QUITO  CALDERON (CARAPU… CALDERON BELLAVI… URBANO              ANTROPICO              19 0 - 10    2421 - 2741                18
 8 2021-01-04 00:00:00    -78.5 -0.0820 PICHINCHA QUITO  QUITO             LA DELI… LA ROLD… URBANO              ANTROPICO              44 0 - 10    2742 - 3067                15
 9 2021-01-05 00:00:00    -78.5 -0.144  PICHINCHA QUITO  QUITO             EUGENIO… SAN ISI… URBANO              ANTROPICO              44 0 - 10    2742 - 3067                10
10 2021-01-06 00:00:00    -78.5 -0.225  PICHINCHA QUITO  QUITO             MANUELA… PANECIL… URBANO              ANTROPICO              44 0 - 10    2742 - 3067                13
# ℹ 1,734 more rows
# ℹ 4 more variables: Humedad_Relativa <dbl>, Velocidad_del_viento <dbl>, Precipitación <dbl>, Subtipo <chr>
# ℹ Use `print(n = ...)` to see more rows
```

Figura 13. Base de datos de información intersecada

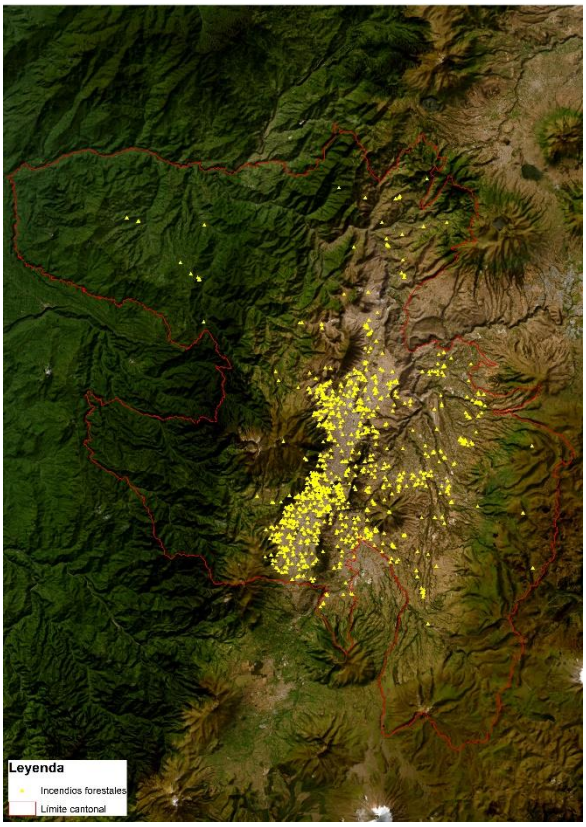

Figura 14. Ubicación de incendios forestales históricos en el DMQ

3.2.2.2 Descripción de los datos

Los datos de incendios forestales y condiciones ambientales fueron proporcionados por el ECU911 y por el INAMHI, por otro lado, las variables geoespaciales y socioeconómicas se recopilaron de información secundaria procedente del Censo de Población y Vivienda 2022 del Instituto Nacional de Estadística y Censos, así como información espacial del Ministerio de Agricultura. La base consta de 1744 registros y 18 columnas.

Con el fin de lograr integrar la información obtenida se utilizó las herramientas de geoanálisis de los Sistemas de Información Geográfica, los cuales nos permiten analizar e intersecar la información brindada por los entes de información.

Los datos recopilados se describen a continuación.

Tabla 4. Descripción de los datos

Variable	Tipo	Descripción
Fecha	POSIXct	Fecha del evento
Longitud	numérica	Coordenada x del evento
Latitud	numérica	Coordenada y del evento
Provincia	categórica	Provincia del evento
Cantón	Categórica	Cantón del evento
Parroquia	Categórica	Parroquia del evento
Distrito	categórica	Distrito del evento
Circuito	categórica	Circuito del evento

Tipo de vegetación	categórica	Tipo de vegetación
Uso_de_suelo	categórica	Uso y cobertura de la tierra
Densidad_pob	numérica	Densidad poblacional (habitantes/km2)
Pendiente	categórica	Porcentaje de pendiente en la que registró el evento
Altitud_msnm	categórica	Altitud en metros sobre el nivel del mar
Temperatura_media	numérica	Temperatura media del día del evento
Humedad_Relativa	numérica	Humedad relativa del día del evento
Velocidad_del_viento	numérica	Velocidad del viento del día del evento
Precipitación	numérica	Precipitación del día del evento
Subtipo	categórica	Ocurrencia del evento

- **Fecha:** La variable de fecha es crucial para establecer el contexto temporal de los datos, facilitando la identificación de tendencias y patrones a lo largo del tiempo.
- **Longitud y Latitud:** Estas coordenadas geográficas permiten la localización precisa de cada registro en un sistema de referencia espacial. La inclusión de estas variables es fundamental para la realización de análisis espaciales avanzados, como la detección de hotspots o áreas de alta recurrencia de incendios.
- **Provincia, Cantón, Parroquia, Distrito y Circuito**: Estas variables administrativas proporcionan un marco contextual que permite desglosar los datos a diferentes niveles geográficos. Además, estas divisiones administrativas permiten realizar análisis más específicos y detallados en áreas de interés particular, así como la comparación entre diferentes jurisdicciones.
- **Tipo de Vegetación y Uso de Suelo:** Estas variables proporcionan información sobre la cobertura terrestre y el uso del territorio. El tipo de vegetación es un factor determinante en la inflamabilidad y, por lo tanto, en la susceptibilidad a incendios forestales. Diferentes tipos de vegetación, como bosques, pastizales o matorrales, tienen diferentes tasas de inflamabilidad y reaccionan de manera distinta bajo condiciones climáticas adversas. El uso de suelo, por su parte, refleja cómo el ser humano ha alterado el paisaje, y cómo estas alteraciones

pueden influir en la propagación de incendios, ya sea aumentando la fragmentación del hábitat o cambiando la composición de la vegetación.

- **Densidad Poblacional:** La densidad poblacional es un indicador clave que refleja la presión humana sobre el medio ambiente. Las áreas con mayor densidad poblacional suelen tener un riesgo elevado de incendios debido a actividades humanas, como la agricultura, la urbanización, la construcción de infraestructuras y otras formas de desarrollo.
- **Pendiente y Altitud:** Estas características topográficas son relevantes porque pueden afectar de manera significativa la propagación de incendios. La pendiente, o inclinación del terreno, puede influir en la velocidad con la que un incendio se propaga, ya que el fuego tiende a moverse más rápidamente cuesta arriba debido a la convección del calor. La altitud también puede jugar un papel importante, ya que afecta las condiciones climáticas locales, como la temperatura y la humedad, que a su vez influyen en la inflamabilidad de la vegetación. Estas variables topográficas son esenciales para modelar la propagación de incendios y para entender cómo las características del terreno pueden contribuir al riesgo de incendios en diferentes áreas.

Una característica clave del archivo de datos es la variable Subtipo, que clasifica los días según la presencia o ausencia de incendios forestales. Esta variable permite distinguir entre los días en que se registraron incendios (etiquetados como "FORESTAL") y aquellos en los que no hubo incendios (etiquetados como "SIN INCENDIOS"). La inclusión de esta variable es fundamental para capturar tanto los eventos de incendios como las condiciones en días sin incidentes, proporcionando un marco de referencia completo para el análisis. Esta distinción es crucial para comprender no solo los factores que desencadenan incendios, sino también para analizar las condiciones ambientales durante períodos sin incendios, lo que puede ofrecer información valiosa sobre la prevención y gestión de riesgos.

3.2.2.2.1 Exploración de los Datos

- **Resumen de los datos**

Como se muestra en la Figura 15, los datos están descritos de la siguiente manera:

Longitud: Coordenada de longitud del evento. La distribución de las longitudes tiene una media de -78.55, lo que sugiere que los eventos están localizados en una región específica con variación en la longitud.

Latitud: Coordenada de latitud del evento. La distribución de las latitudes tiene una media de -1.87, indicando que los eventos se encuentran en el hemisferio sur, cerca de la línea ecuatorial.

Altitud: Altitud en metros sobre el nivel del mar. La media es de 691.8 metros, con una gran variabilidad desde el nivel del mar hasta 4144 metros.

Temperatura media: Temperatura media en grados Celsius. La media es de 20.12°C, con un rango de 5°C a 25°C.

Humedad relativa: Humedad relativa en porcentaje. La media es de 45.84%, con un amplio rango de 1% a 98%.

Precipitación: Precipitación en mm. La media es de 1.57 mm, pero la mediana es 0, lo que sugiere que muchos eventos ocurrieron sin precipitación, con algunos eventos extremos de hasta 188.8 mm.

```
> summary(datos_incendios)
     Fecha                        Longitud         Latitud         Provincia           Cantón           Parroquia          Distrito          Circuito
 Min.   :2021-01-01 00:00:00.0   Min.   :-78.78   Min.   :-0.4150   Length:1744        Length:1744        Length:1744        Length:1744        Length:1744
 1st Qu.:2021-09-04 00:00:00.0   1st Qu.:-78.51   1st Qu.:-0.2574   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :2022-08-10 12:00:00.0   Median :-78.46   Median :-0.1947   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :2022-06-28 03:28:04.4   Mean   :-78.46   Mean   :-0.1772
 3rd Qu.:2023-03-12 06:00:00.0   3rd Qu.:-78.40   3rd Qu.:-0.1076
 Max.   :2023-12-30 00:00:00.0   Max.   :-78.22   Max.   : 0.1882
 Tipo de vegetación Uso_de_suelo       Densidad_pob    Pendiente         Altitud_msnm       Temperatura_media Humedad_Relativa Velocidad_del_viento Precipitación
 Length:1744        Length:1744        Min.   : 0.00   Length:1744        Length:1744        Min.   : 7.00     Min.   :21.00    Min.   : 6.00       Min.   : 0.0000
 Class :character   Class :character   1st Qu.: 3.00   Class :character   Class :character   1st Qu.:15.00     1st Qu.:53.00    1st Qu.:14.00       1st Qu.: 0.0000
 Mode  :character   Mode  :character   Median :18.00   Mode  :character   Mode  :character   Median :17.00     Median :63.00    Median :15.00       Median : 0.0000
                                       Mean   :22.73                                         Mean   :17.05     Mean   :62.28    Mean   :15.46       Mean   : 0.7086
                                       3rd Qu.:44.00                                         3rd Qu.:20.00     3rd Qu.:74.00    3rd Qu.:17.00       3rd Qu.: 0.4000
                                       Max.   :44.00                                         Max.   :25.00     Max.   :90.00    Max.   :29.00       Max.   :16.8000

   Subtipo
 Length:1744
 Class :character
```

Figura 15. Resumen de los datos

- **Dimensionalidad de los datos**

El presente dataset contiene un total de 1744 filas de datos, y 18 variables o columnas, tal como se muestra en la Figura 16.

```
> # Obtener el número de filas
> num_filas <- nrow(datos_incendios)
> print(paste("Número de filas:", num_filas))
[1] "Número de filas: 1744"
> # Obtener el número de columnas
> num_columnas <- ncol(datos_incendios)
> print(paste("Número de columnas:", num_columnas))
[1] "Número de columnas: 18"
```

Figura 16. Dimensionalidad de los datos

3.2.2.2.2 Transformación de variables categóricas a variables numéricas

Con el fin de estandarizar las variables para la generación de modelos de predicción se procedió a transformar las variables categóricas en variables numéricas asignando una codificación, tal como se detalla en la Figura 17.

```
# Transformar variables categóricas en numéricas
datos_incendios <- datos_incendios %>%
  mutate(
    # Convertir la variable 'Subtipo' en binaria: 1 si es 'FORESTAL', 0 en caso contrario
    Subtipo = if_else(Subtipo == "FORESTAL", 1, 0),
    # Convertir la variable 'Uso_de_suelo' en categorías numéricas
    Uso_de_suelo = case_when(
      Uso_de_suelo == "AGRICOLA" ~ 1,
      Uso_de_suelo == "AGROPECUARIO MIXTO" ~ 2,
      Uso_de_suelo == "AGUA" ~ 3,
      Uso_de_suelo == "ANTROPICO" ~ 4,
      Uso_de_suelo == "CONSERVACION Y PROTECCION" ~ 5,
      Uso_de_suelo == "PROTECCION O PRODUCCION" ~ 6,
      Uso_de_suelo == "PECUARIO" ~ 7,
      Uso_de_suelo == "CONSERVACION Y PRODUCCION" ~ 8
    ),
    # Convertir la variable 'Pendiente' en categorías numéricas
    Pendiente = case_when(
      Pendiente == "0 - 10" ~ 1,
      Pendiente == "11 - 25" ~ 2,
      Pendiente == "26 - 50" ~ 3,
      Pendiente == "51 - 75" ~ 4
    ),
    # Convertir la variable 'Altitud_msnm' en categorías numéricas
    Altitud_msnm = case_when(
      Altitud_msnm == "500 - 1602" ~ 1,
      Altitud_msnm == "1603 - 2041" ~ 2,
      Altitud_msnm == "2042 - 2420" ~ 3,
      Altitud_msnm == "2421 - 2741" ~ 4,
      Altitud_msnm == "2742 - 3067" ~ 5,
      Altitud_msnm == "3068 - 3436" ~ 6,
      Altitud_msnm == "3437 - 3866" ~ 7,
      Altitud_msnm == "3867 - 4730" ~ 8
    )
  )
```

Figura 17. Código de transformación de variables

3.2.2.2.3 Identificación de outliers

Para la identificación de outliers se realizó el análisis de diagramas de caja, obteniendo el resultado de la Figura 18 y Figura 19.

```
# Definir las variables numéricas para análisis exploratorio
variables_numéricas <- c("Temperatura_media", "Humedad_Relativa",
                         "Velocidad_del_viento", "Precipitación",
                         "Densidad_pob", "Pendiente", "Altitud_msnm")

# Crear boxplots para cada variable numérica
par(mfrow = c(2, 5))  # Organizar gráficos en una cuadrícula de 2 filas y 5 columnas
for (var in variables_numéricas) {
  boxplot(datos_incendios[[var]], main = var, col = "blue", border = "black")
}

# Restablecer la disposición de los gráficos a una sola ventana
par(mfrow = c(1, 1))
```

Figura 18. Código de identificación de outliers

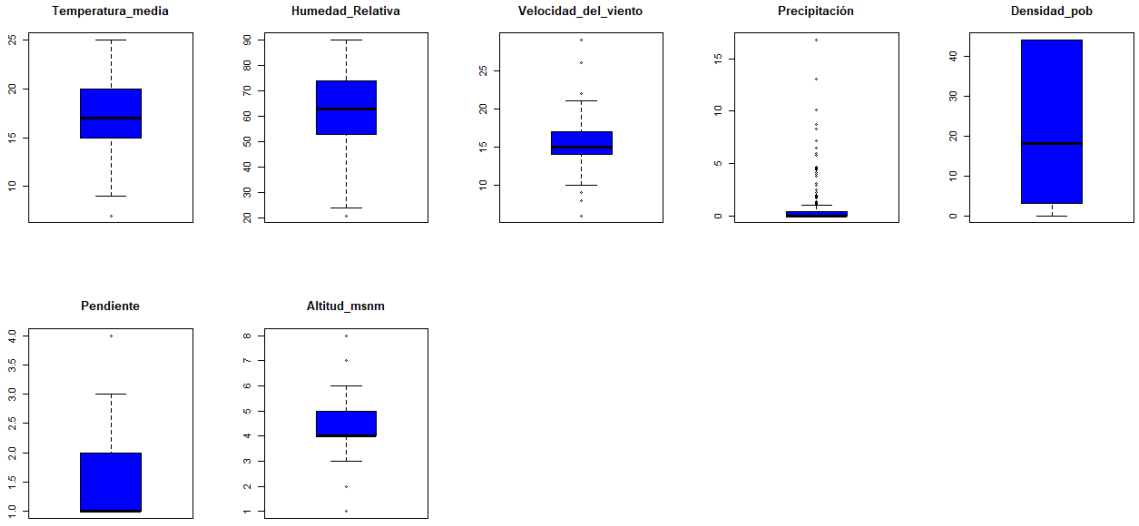

Figura 19. Boxplots de las variables

- **Temperatura_media**

  La mayoría de los valores de temperatura media están entre aproximadamente 18°C y 23°C. Hay algunos outliers por debajo de 18°C y por encima de 23°C, lo que indica temperaturas excepcionalmente bajas o altas en algunos eventos.

- **Humedad_Relativa**

La mayoría de los valores de humedad relativa están entre aproximadamente 30% y 61%. Existen varios outliers, especialmente por encima de 61%, lo que sugiere algunos eventos con humedad relativa muy alta.

- **Velocidad_del_viento**

  La mayoría de las velocidades del viento están entre 0 y 4 m/s. Hay varios outliers por encima de 4 m/s, indicando eventos con velocidades del viento inusualmente altas.

- **Precipitación**

  La mayoría de los eventos tienen precipitación entre 0 y 0 mm, lo que sugiere que muchos eventos ocurrieron sin precipitación significativa. Sin embargo, hay muchos outliers, lo que indica que hubo algunos eventos con precipitaciones muy altas (hasta 188.8 mm).

- **Densidad poblacional**

  La mayoría de los valores de densidad poblacional están muy concentrados en el rango inferior, pero hay una gran variabilidad con varios outliers, lo que indica que algunos eventos ocurrieron en áreas con densidades poblacionales excepcionalmente altas.

- **Pendiente**

  La mayoría de los valores de pendiente están entre 0° y aproximadamente 15°. Hay varios outliers por encima de 15°, lo que indica terrenos con pendientes muy pronunciadas en algunos eventos.

- **Altitud**

La mayoría de las altitudes están entre aproximadamente 112 y 1200 metros. Hay varios outliers tanto por debajo como por encima de este rango, lo que indica eventos en altitudes muy bajas o muy altas.

3.2.2.2.4 Construir datos

En la Figura 20 se procede a aumentar una columna que calcula el Forest Fire Weather Index. Este índice es parte de un sistema desarrollado en Canadá, que consta de seis componentes que tienen en cuenta los efectos de la humedad del combustible y las condiciones meteorológicas sobre el comportamiento del fuego.

Los primeros tres componentes son códigos de humedad del combustible, que son clasificaciones numéricas del contenido de humedad del suelo forestal y otros materiales orgánicos muertos. Sus valores aumentan a medida que disminuye el contenido de humedad (Canadá, s.f.).

Los tres componentes restantes son índices de comportamiento del fuego, que representan la tasa de propagación del fuego, el combustible disponible para la combustión y la intensidad frontal del fuego; estos tres valores aumentan a medida que aumenta el peligro de incendio (Canadá, s.f.).

Los cálculos fueron realizados mediante las siguientes líneas de código:

```r
# Definir una función para calcular el índice Fire Weather Index (FWI)
calculate_fwi <- function(temperature, relative_humidity, wind_speed, precipitation) {
  # Calcular el Drought Code (DMC)
  DMC <- 1.641 * log(10 * precipitation + 10)

  # Calcular el Initial Spread Index (ISI)
  ISI <- 0.208 * wind_speed + 0.496 * (100 - relative_humidity)

  # Calcular el Buildup Index (BUI)
  BUI <- (0.026 * DMC) * (1 - exp(-0.25 * DMC))

  # Calcular el Fire Weather Index (FWI)
  FWI <- 0.4 * (ISI + BUI)

  # Devolver todos los índices calculados
  return(list(DMC = DMC, ISI = ISI, BUI = BUI, FWI = FWI))
}
```

Figura 20. Cálculo de índices

3.2.2.2.5 Seleccionar datos

Para el desarrollo y ejecución de los modelos de predicción se requieren las siguientes variables:

- Longitud
- Latitud
- Temperatura_media
- Humedad_Relativa
- Velocidad_del_viento
- Precipitación
- FWI
- Subtipo
- Densidad_pob
- Uso_de_suelo
- Pendiente
- Altitud_msnm
- DMC

La Figura 21 muestra el dataset final con las variables seleccionadas para el análisis, tomando en cuenta que se catalogó las variables categóricas para transformarlas en variables numéricas.

```
> # Seleccionar columnas relevantes para el análisis posterior
> datos_incendios <- datos_incendios %>% dplyr::select(
+   Longitud, Latitud, Temperatura_media, Humedad_Relativa,
+   Velocidad_del_viento, Precipitación, FWI, Subtipo,
+   Densidad_pob, Uso_de_suelo, Pendiente, Altitud_msnm)
> # Agregar los resultados de FWI al dataframe original
> datos_incendios$DMC <- resultados$DMC
> print(datos_incendios)
# A tibble: 1,744 × 13
   Longitud Latitud Temperatura_media Humedad_Relativa Velocidad_del_viento Precipitación   FWI Subtipo Densidad_pob Uso_de_suelo Pendiente Altitud_msnm   DMC
      <db1>   <db1>             <db1>             <db1>               <db1>          <db1> <db1>   <db1>       <db1>        <db1>     <db1>       <db1> <db1>
 1    -78.5 -0.265                 14                57                  18            0   10.1       0          18            4         2           4  3.78
 2    -78.5 -0.190                 14                61                  13            1.3  8.86      0          44            4         1           5  5.15
 3    -78.6 -0.281                 11                82                  22           13.1  5.47      0          44            4         1           5  8.12
 4    -78.4 -0.226                 10                89                  13            2.5  3.31      0          15            4         1           3  5.83
 5    -78.5 -0.111                 12                77                  16            0    5.92      0          44            4         1           5  3.78
 6    -78.5 -0.0958                16                67                  14            0    7.74      1          44            4         2           5  3.78
 7    -78.4 -0.0807                18                54                  13            0   10.2       1          19            4         1           4  3.78
 8    -78.5 -0.0820                15                73                  15            0    6.63      1          44            4         1           5  3.78
 9    -78.5 -0.144                 10                88                  13            0.2  3.49      0          44            4         1           5  4.08
10    -78.5 -0.225                 13                79                  12            0    5.19      0          44            4         1           5  3.78
# ℹ 1,734 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 21. Dataset final para modelos predictivos

- **Correlación entre variables**

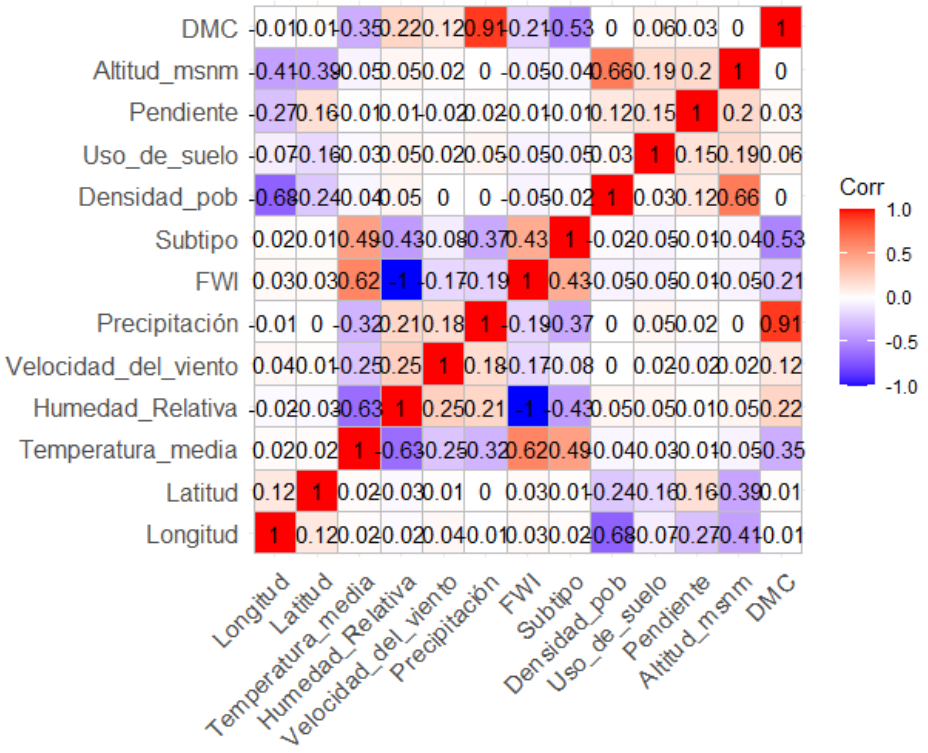

Figura 22. Matriz de correlación de variables

La Figura 22 muestra la correlación entre las variables del estudio, determinando que:

- Variables longitud y latitud están perfectamente correlacionadas entre sí (1.00), lo cual es esperable ya que representan coordenadas geográficas.
- Variables temperatura media y humedad relativa están moderadamente correlacionadas de manera positiva, lo que indica que a medida que una aumenta, la otra también tiende a aumentar.
- La variable velocidad del viento tiene una correlación negativa con temperatura media y humedad relativa, lo cual podría indicar que, en condiciones de viento fuerte, la temperatura y la humedad relativa tienden a ser más bajas.
- La variable FWI tiene correlaciones significativas con la velocidad del viento y la humedad relativa, lo que indica que estas variables son importantes para el índice de peligro de incendios.

- La densidad poblacional tiene una fuerte correlación negativa con la longitud, indicando que ciertas áreas longitudinales específicas tienen densidades poblacionales particulares.
- La variable altitud tiene fuertes correlaciones con varias variables, sugiriendo que la altitud tiene un impacto significativo en varias condiciones ambientales.

3.2.3 Modelado

Para la construcción de los modelos de predicción se han seleccionado 3 modelos de clasificación supervisada, con el fin de compararlos y determinar cuál es el más apropiado para la predicción de incendios forestales en función de factores ambientales, socioeconómicos y geoespaciales.

3.2.3.1 Modelo de Random Forest

- **Preparación de los Datos:**
  - Conjunto de Datos: 'train_data' y 'test_data' son los conjuntos de datos de entrenamiento y prueba, respectivamente.
  - **Variables Predictoras y Objetivo**: La variable objetivo ('Subtipo') es lo que queremos predecir, mientras que las demás variables son las predictoras.
  - **Exclusión de Columnas**: Se excluyen las columnas 1 y 2, probablemente porque contienen identificadores únicos o datos irrelevantes para el modelo.

3.2.3.1.1 Entrenamiento del Modelo:

```
modelo_rf <- randomForest(Subtipo ~ ., data = train_data[,-c(1,2)], importance = TRUE, ntree = 500)
```

- **Algoritmo**: Random Forest es un *ensemble learning method* que construye múltiples árboles de decisión y combina sus resultados.
- **Fórmula de la Ecuación**: 'Subtipo ~ .' significa que 'Subtipo' es la variable dependiente y todas las demás variables ('**.**') son independientes.

- **Parámetros**:
    - ‘importance = TRUE’: Calcula la importancia de cada variable.
    - ‘ntree = 500’: Especifica el número de árboles a construir. Más árboles suelen dar un modelo más robusto, pero también incrementan el tiempo de computación.

3.2.3.1.2 Visualización de la Importancia de las Variables:

```
varImpPlot(modelo_rf)
```

- Gráfica de Importancia: La función ‘varImpPlot’ produce una gráfica que muestra la importancia de cada variable predictora en el modelo. Esto se mide a través de la reducción de impureza (Gini) o la precisión del modelo al usar cada variable.

3.2.3.1.3 Resumen del Modelo:

```
print(modelo_rf)
```

- Output: Proporciona información detallada sobre el modelo entrenado, incluyendo la tasa de error para cada clase, el OOB (*Out-Of-Bag*) error, y la importancia de las variables.

3.2.3.1.4 Evaluación del modelo

Para evaluar un modelo de Random Forest, se pueden utilizar varios métodos y métricas:

1. Error Out-Of-Bag (OOB):

- Error OOB: Durante el entrenamiento, cada árbol se entrena con una muestra aleatoria del conjunto de datos, dejando fuera algunas observaciones (OOB samples). El error OOB es la tasa de error promedio de estas observaciones no usadas en cada árbol. Es una estimación del error del modelo.

2. Matriz de Confusión:

```
predicciones_rf <- predict(modelo_rf, newdata = test_data[,-c(1,2)])

confusionMatrix <- table(test_data$Subtipo, predicciones_rf)

print(confusionMatrix)
```

La matriz de confusión muestra la cantidad de predicciones correctas e incorrectas, desglosadas por cada clase. Ayuda a entender el rendimiento del modelo en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

3. Métricas de Desempeño:

- Precisión: Proporción de predicciones correctas (verdaderos positivos y verdaderos negativos) sobre el total de predicciones.
- Recall (Sensibilidad): Proporción de verdaderos positivos sobre el total de casos positivos reales.
- F1-Score: Media armónica de la precisión y el recall, útil para evaluar modelos en conjuntos de datos desbalanceados.

```
precision <- sum(diag(confusionMatrix)) / sum(confusionMatrix)

recall <- diag(confusionMatrix) / rowSums(confusionMatrix)

f1_score <- 2 * precision * recall / (precision + recall)

cat("Precisión: ", precision, "\n")

cat("Recall: ", recall, "\n")

cat("F1-Score: ", f1_score, "\n")
```

3.2.3.2 Curva ROC y AUC:

- ROC (Receiver Operating Characteristic): Gráfica que muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.

- AUC (Area Under the Curve): Métrica que mide el rendimiento del modelo, donde un AUC cercano a 1 indica un modelo excelente.

  library(pROC)

  roc_obj <- roc(test_data$Subtipo, as.numeric(predicciones_rf))

  plot(roc_obj)

  auc(roc_obj)

3.2.3.1.5 Interpretación de Resultados:

- Importancia de las Variables: Las variables con mayor importancia indican cuáles características son más relevantes para la predicción del Subtipo.
- Tasa de Error: La tasa de error OOB y la matriz de confusión proporcionan una estimación del desempeño general del modelo.
- Métricas de Desempeño: Precisión, recall y F1-Score ofrecen una visión detallada del rendimiento del modelo, especialmente en casos de clases desbalanceadas.
- Curva ROC y AUC: La curva ROC y el valor AUC ayudan a evaluar la capacidad del modelo para distinguir entre clases.

3.2.3.2 Metodología del Modelo de Regresión Logística

3.2.3.2.1 Ajuste Inicial del Modelo Completo

El primer paso en la construcción del modelo logístico es ajustar un modelo completo utilizando todas las variables predictoras disponibles. Esto se realiza con la siguiente línea de código:

summary(glm(Subtipo ~ ., train_data[,-c(1,2)], family = "binomial"))

Objetivo: Identificar la significancia de cada variable predictora.

Descripción: Se ajusta un modelo de regresión logística (glm) con la variable de respuesta Subtipo y todas las demás variables predictoras (~ .), excluyendo las columnas 1 y 2 de los datos de entrenamiento (train_data).

Salida: El resumen (summary) del modelo ajustado proporciona los coeficientes, errores estándar, valores z y p-values de cada predictor. Las variables con p.value < 0.05 se consideran significativas.

1. Definición de Modelos Alternativos

Con base en los resultados del ajuste inicial, se definen dos modelos logísticos: uno que incluye todas las variables predictoras y otro que incluye solo las variables significativas.

```
logit_formulas <- formulas(
.response = ~ Subtipo,
Modelo1 = ~ Temperatura_media + Humedad_Relativa +
Velocidad_del_viento + Precipitación +
FWI + Densidad_pob + Uso_de_suelo +
Pendiente + Altitud_msnm,
Modelo2 = ~ Humedad_Relativa +
Velocidad_del_viento + Precipitación +
FWI
)
```

Modelo1: Incluye todas las variables predictoras relevantes.

Modelo2: Incluye solo las variables que fueron significativas en el ajuste inicial.

3.2.3.2.2 Ajuste de los Modelos

Se ajustan los modelos logísticos utilizando las fórmulas definidas previamente y los datos de entrenamiento.

```
models <- data_frame(logit_formulas) %>% mutate(

models = names(logit_formulas),  # Añadir nombres de los modelos

expression = paste(logit_formulas),  # Añadir expresiones de las fórmulas

mod = map(logit_formulas, ~ glm(., family = 'binomial', data = train_data[,-c(1,2)]))  # Ajustar modelo
```

Descripción:

- Se crea un data_frame que contiene las fórmulas de los modelos.
- Se ajustan los modelos logísticos (glm) para cada fórmula utilizando la función map de purrr.

3.2.3.2.3 Evaluación de los Modelos

En esta fase, se obtienen las estimaciones de los coeficientes y los valores p para cada uno de los modelos ajustados. Esto permite evaluar la significancia estadística de cada variable predictora dentro de los modelos.

Código para la Evaluación de los Modelos

```
models %>% filter(models %in% c('Modelo1', 'Modelo2')) %>% mutate(tidy = map(mod, tidy)) %>%  # Usar broom::tidy para obtener resultados ordenados

unnest(tidy, .drop = TRUE) %>% mutate(estimate = round(estimate, 5),  # Redondear estimaciones

p.value = round(p.value, 4))  # Redondear valores p
```

- Filtrado de Modelos:

filter(models %in% c('Modelo1', 'Modelo2'))

- **Objetivo**: Seleccionar solo los modelos Modelo1 y Modelo2 para su evaluación.

- **Ordenar Resultados**:

  mutate(tidy = map(mod, tidy))

  - **Objetivo**: Utilizar la función tidy del paquete broom para convertir los resultados del modelo en un formato ordenado y fácil de manejar.

- **Desanidar Resultados**:

  unnest(tidy, .drop = TRUE)

  - **Objetivo**: Convertir la lista de data frames anidados en un solo data frame.

- **Redondear Estimaciones y Valores P**:

  mutate(estimate = round(estimate, 5), p.value = round(p.value, 4))

  - **Objetivo**: Redondear las estimaciones de los coeficientes a 5 decimales y los valores p a 4 decimales para facilitar su interpretación.

- **Cálculo de la Varianza Explicada**

El siguiente paso es calcular la varianza explicada por cada modelo ajustado, lo que permite comparar la calidad de los modelos.

Código para el Cálculo de la Varianza Explicada

```
models <- models %>% mutate(glance = map(mod, glance))

models %>% unnest(glance, .drop = TRUE) %>% mutate(perc_explained_dev = 1 - deviance / null.deviance) %>% dplyr::select(-c(expression, logit_formulas, df.null, AIC, mod)) %>%arrange(deviance)
```

- **Obtener Métricas del Modelo**:

  mutate(glance = map(mod, glance))

  - **Objetivo**: Utilizar la función glance del paquete broom para extraer métricas del modelo como el deviance y el null deviance.

- **Desanidar Resultados**:

  unnest(glance, .drop = TRUE)

  - **Objetivo**: Convertir la lista de data frames anidados en un solo data frame.

- **Calcular la Varianza Explicada**:

  mutate(perc_explained_dev = 1 - deviance / null.deviance)

  - **Objetivo**: Calcular el porcentaje de deviance explicado por el modelo.

- **Seleccionar y Ordenar Columnas**:

  dplyr::select(-c(expression, logit_formulas, df.null, AIC, mod)) %>%

  arrange(deviance)

  - **Objetivo**: Seleccionar columnas relevantes y ordenar los modelos por deviance.

- **Realización de Predicciones**

Una vez seleccionado el mejor modelo basado en la varianza explicada, se realizan predicciones y se generan curvas ROC para evaluar el rendimiento del modelo.

Código para la Realización de Predicciones

models <- models %>%

mutate(pred = map(mod, augment, type.predict = "response")) prediction_full <- models %>% filter(models == "Modelo1") %>% unnest(pred, .drop = TRUE)

```
roc_full <- roc(response = prediction_full$Subtipo, predictor = prediction_full$.fitted, levels = c(0, 1), direction = "<")

prediction_bad <- models %>% filter(models == "Modelo2") %>% unnest(pred, .drop = TRUE)

roc_bad <- roc(response = prediction_bad$Subtipo, predictor = prediction_bad$.fitted, levels = c(0, 1), direction = "<")

# Graficar curvas ROC

ggroc(list(Modelo1 = roc_full, Modelo2 = roc_bad), size = 1) +

geom_abline(slope = 1, intercept = 1, linetype = 'dashed') +

theme_bw() + labs(title = 'Curvas ROC', color = 'Modelo')
```

Obtener Predicciones:

```
mutate(pred = map(mod, augment, type.predict = "response"))
```

- **Objetivo**: Utilizar la función *augment* del paquete *broom* para obtener las predicciones del modelo.

Filtrar Predicciones del Mejor Modelo:

```
filter(models == "Modelo1")
```

- **Objetivo**: Seleccionar el modelo Modelo1 para realizar predicciones.

Generar Curvas ROC:

```
roc_full <- roc(response = prediction_full$Subtipo, predictor = prediction_full$.fitted, levels = c(0, 1), direction = "<")
```

- **Objetivo**: Generar la curva ROC para el modelo Modelo1.

Generar Curvas ROC para el Modelo Alternativo:

roc_bad <- roc(response = prediction_bad$Subtipo, predictor = prediction_bad$.fitted, levels = c(0, 1), direction = "<")

- **Objetivo**: Generar la curva ROC para el modelo Modelo2 para comparar su rendimiento.

- **Graficar Curvas ROC**:

ggroc(list(Modelo1 = roc_full, Modelo2 = roc_bad), size = 1) + geom_abline(slope = 1, intercept = 1, linetype = 'dashed') + theme_bw() + labs(title = 'Curvas ROC', color = 'Modelo')

- **Objetivo**: Graficar las curvas ROC para ambos modelos, facilitando la comparación visual del rendimiento. La línea discontinua indica una clasificación aleatoria.

Esta metodología asegura que los modelos logísticos ajustados sean evaluados de manera rigurosa, permitiendo identificar el mejor modelo para la predicción de Subtipo y su rendimiento a través de las métricas ROC.

- **Gráficos de Violín para Distribución de Probabilidades Predichas**

Los gráficos de violín se utilizan para visualizar la distribución de las probabilidades predichas para cada clase en los modelos ajustados. Estos gráficos ayudan a comprender cómo se distribuyen las probabilidades predichas en función de las clases reales.

Código para Generar Gráficos de Violín

violin_full = ggplot(prediction_full, aes(x = Subtipo, y = .fitted, group = Subtipo, fill = factor(Subtipo))) + geom_violin() + theme_bw() + guides(fill = FALSE) + labs(title = 'Violin plot', subtitle = 'Modelo 1', y = 'Predicted probability')

violin_bad = ggplot(prediction_bad, aes(x = Subtipo, y = .fitted, group = Subtipo, fill = factor(Subtipo))) + geom_violin() + theme_bw() + guides(fill = FALSE) + labs(title = 'Violin plot', subtitle = 'Modelo 2', y = 'Predicted probability')

Mostrar los gráficos de violín

plot(violin_full)

plot(violin_bad)

- **Generar el Gráfico de Violín para el Modelo Completo**:

  violin_full = ggplot(prediction_full, aes(x = Subtipo, y = .fitted, group = Subtipo, fill = factor(Subtipo))) + geom_violin() +theme_bw() + guides(fill = FALSE) + labs(title = 'Violin plot', subtitle = 'Modelo 1', y = 'Predicted probability')

  - **ggplot(prediction_full, aes(x = Subtipo, y = .fitted, ...))**: Crea un gráfico de violín utilizando los datos de predicción del Modelo1. El eje x representa la variable Subtipo (la clase real) y el eje y representa las probabilidades predichas (.fitted).
  - **geom_violin()**: Añade el gráfico de violín, que muestra la distribución de las probabilidades predichas para cada clase.
  - **theme_bw()**: Aplica un tema de fondo blanco al gráfico.
  - **guides(fill = FALSE)**: Elimina la leyenda de relleno del gráfico.
  - **labs(title = 'Violin plot', subtitle = 'Modelo 1', y = 'Predicted probability')**: Establece el título, subtítulo y etiqueta del eje y.

1. **Generar el Gráfico de Violín para el Modelo Alternativo**:

   violin_bad = ggplot(prediction_bad, aes(x = Subtipo, y = .fitted, group = Subtipo, fill = factor(Subtipo))) + geom_violin() + theme_bw() +guides(fill

= FALSE) + labs(title = 'Violin plot', subtitle = 'Modelo 2', y = 'Predicted probability')

- Similar al gráfico anterior, pero para Modelo2. Muestra la distribución de las probabilidades predichas para cada clase en este modelo.

2. **Mostrar los Gráficos**:

plot(violin_full)

plot(violin_bad)

- **plot()**: Muestra los gráficos de violín generados.

- **Prueba de Hosmer-Lemeshow**

La prueba de Hosmer-Lemeshow es una prueba de bondad de ajuste utilizada para evaluar la adecuación del modelo logístico. La prueba compara las predicciones del modelo con las frecuencias observadas.

Código para la Prueba de Hosmer-Lemeshow

Hosmer_Lemeshow_plot <- function(dataset, predicted_column, class_column, bins, positive_value, color = 'forestgreen', nudge_x = 0, nudge_y = 0.05) {

Asignar grupos basados en la probabilidad predicha

dataset['group'] <- bin(dataset[predicted_column], nbins = bins, method = 'l', labels = c(1:bins))

Contar casos positivos por grupo

positive_class <- dataset %>% filter(!!sym(class_column) == positive_value) %>% group_by(group) %>% count()

Calcular la media de predicciones por grupo

```
HL_df <- dataset %>% group_by(group) %>% summarise(pred = mean(!!sym(predicted_column)), count = n()) %>% inner_join(., positive_class) %>% mutate(freq = n / count)

# Crear el gráfico

HM_plot <- ggplot(HL_df, aes(x = pred, y = freq)) + geom_point(aes(size = n), color = color) + geom_text(aes(label = n), nudge_y = nudge_y) + geom_abline(slope = 1, intercept = 0, linetype = 'dashed') + theme_bw() + labs(title = 'Hosmer-Lemeshow', size = 'Casos', x = "Probabilidad Predicha", y = "Frecuencia observada")

return(HM_plot)

}

# Graficar Hosmer-Lemeshow para los modelos

Hosmer_Lemeshow_plot(prediction_full, '.fitted', 'Subtipo', 50, 1) +

labs(subtitle = "Modelo 1")

Hosmer_Lemeshow_plot(prediction_bad, '.fitted', 'Subtipo', 50, 1, color = "firebrick", nudge_y = 0.003) + labs(subtitle = "Modelo 2")
```

- **Definir la Función Hosmer_Lemeshow_plot**:

  ```
  Hosmer_Lemeshow_plot <-function(dataset, predicted_column, class_column, bins, positive_value, color = 'forestgreen', nudge_x = 0, nudge_y = 0.05)
  ```

  - **dataset**: Conjunto de datos con predicciones y valores reales.
  - **predicted_column**: Nombre de la columna con las probabilidades predichas.
  - **class_column**: Nombre de la columna con las clases reales.
  - **bins**: Número de grupos para la prueba de Hosmer-Lemeshow.

- **positive_value**: Valor que se considera como positivo para la clase.
- **color**: Color para los puntos en el gráfico.
- **nudge_x y nudge_y**: Ajustes para la posición del texto en el gráfico.
-

- **Asignar Grupos Basados en la Probabilidad Predicha**:

  dataset['group'] <- bin(dataset[predicted_column], nbins = bins, method = 'l', labels = c(1:bins))

  - **bin()**: Agrupa las probabilidades predichas en bins grupos.

- **Contar Casos Positivos por Grupo**:

  positive_class <- dataset %>% filter(!!sym(class_column) == positive_value) %>% group_by(group) %>% count()

  - **filter()**: Filtra los casos positivos.
  - **group_by()** y **count()**: Cuenta el número de casos positivos por grupo.

- **Calcular la Media de Predicciones por Grupo**:

  HL_df <- dataset %>% group_by(group) %>% summarise(pred = mean(!!sym(predicted_column)), count = n()) %>% inner_join(., positive_class) %>% mutate(freq = n / count)

  - **summarise()**: Calcula la media de las predicciones y el número total de casos por grupo.
  - **inner_join()**: Une los datos con el conteo de casos positivos.
  - **mutate()**: Calcula la frecuencia observada de casos positivos.

- **Crear el Gráfico de Hosmer-Lemeshow**:

HM_plot <- ggplot(HL_df, aes(x = pred, y = freq)) + geom_point(aes(size = n), color = color) + geom_text(aes(label = n), nudge_y = nudge_y) + geom_abline(slope = 1, intercept = 0, linetype = 'dashed') + theme_bw() + labs(title = 'Hosmer-Lemeshow', size = 'Casos', x = "Probabilidad Predicha", y = "Frecuencia observada")

- **geom_point()**: Añade puntos al gráfico con tamaño proporcional al número de casos en cada grupo.
- **geom_text()**: Añade etiquetas con el número de casos a los puntos.
- **geom_abline()**: Añade una línea diagonal para comparar las frecuencias observadas con las predicciones.
- **labs()**: Establece el título y las etiquetas de los ejes.

2. **Graficar Hosmer-Lemeshow para los Modelos**:

Hosmer_Lemeshow_plot(prediction_full, '.fitted', 'Subtipo', 50, 1) + labs(subtitle = "Modelo 1")

Hosmer_Lemeshow_plot(prediction_bad, '.fitted', 'Subtipo', 50, 1, color = "firebrick", nudge_y = 0.003) + labs(subtitle = "Modelo 2")

3.2.3.2.4 Evaluación de Métricas de Predicción

Se evalúan diversas métricas de predicción (como precisión, sensibilidad, especificidad, etc.) para diferentes puntos de corte, y se grafican los resultados para comparar el rendimiento del modelo.

Código para Evaluar Métricas de Predicción

Función para calcular métricas de predicción en función de diferentes puntos de corte

prediction_metrics <- function(cutoff, predictions = prediction_full) {

```
table <- predictions %>% mutate(predicted_class = if_else(.fitted > cutoff, 1, 0) %>% as.factor(), Subtipo = factor(Subtipo))

confusionMatrix(table$predicted_class, table$Subtipo, positive = "1") %>%

tidy() %>%

dplyr::select(term, estimate) %>%

filter(term %in% c('accuracy', 'sensitivity', 'specificity', 'precision', 'recall')) %>%

mutate(cutoff = cutoff)

}

# Evaluar métricas de predicción para una serie de puntos de corte

cutoffs = seq(0.01, 0.95, 0.01)

logit_pred = map_dfr(cutoffs, prediction_metrics) %>% mutate(term = as.factor(term))

# Graficar métricas de predicción en función del punto de corte

ggplot(logit_pred, aes(cutoff, estimate, group = term, color = term)) + geom_line(size = 1) + theme_bw() + labs(title = 'Accuracy, Sensitivity, Specificity, Recall y Precision', subtitle = 'Modelo completo', color = "") + geom_vline(xintercept = 0.65, linetype = "dashed", color = "black")
```

1. **Definir la Función prediction_metrics**:

   ```
   prediction_metrics <- function(cutoff, predictions = prediction_full) {

   table <- predictions %>% mutate(predicted_class = if_else(.fitted > cutoff, 1, 0) %>% as.factor(), Subtipo = factor(Subtipo))

   confusionMatrix(table$predicted_class, table$Subtipo, positive = "1") %>%

      tidy() %>%
   ```

```
  dplyr::select(term, estimate) %>%
  filter(term %in% c('accuracy', 'sensitivity', 'specificity', 'precision', 'recall')) %>%
  mutate(cutoff = cutoff)
}
```

- **cutoff**: Punto de corte para clasificar las probabilidades predichas.
- **mutate(predicted_class = if_else(.fitted > cutoff, 1, 0) %>% as.factor())**: Clasifica las observaciones en positivas o negativas en función del punto de corte.
- **confusionMatrix()**: Calcula la matriz de confusión y las métricas de rendimiento.
- **tidy()**: Ordena los resultados de la matriz de confusión.
- **filter(term %in% c(...))**: Selecciona las métricas relevantes.
- **mutate(cutoff = cutoff)**: Añade el punto de corte a los resultados.

- **Evaluar Métricas para una Serie de Puntos de Corte**:

```
cutoffs = seq(0.01, 0.95, 0.01)
logit_pred = map_dfr(cutoffs, prediction_metrics) %>% mutate(term = as.factor(term))
```

- **seq(0.01, 0.95, 0.01)**: Crea una secuencia de puntos de corte de 0.01 a 0.95.
- **map_dfr(cutoffs, prediction_metrics)**: Calcula las métricas para cada punto de corte.
- **mutate(term = as.factor(term))**: Convierte la columna term en un factor.

- **Graficar Métricas de Predicción**:

ggplot(logit_pred, aes(cutoff, estimate, group = term, color = term)) + geom_line(size = 1) + theme_bw() + labs(title = 'Accuracy, Sensitivity, Specificity, Recall y Precision', subtitle = 'Modelo completo', color = "") + geom_vline(xintercept = 0.65, linetype = "dashed", color = "black")

- **ggplot(logit_pred, aes(cutoff, estimate, group = term, color = term))**: Crea el gráfico de líneas para las métricas en función del punto de corte.
- **geom_line(size = 1)**: Añade líneas al gráfico.
- **geom_vline(xintercept = 0.65, linetype = "dashed", color = "black")**: Añade una línea vertical en el punto de corte óptimo (0.65).

3.2.3.2.3 Ajuste y Evaluación del Modelo Completo

Finalmente, se ajusta el modelo logístico completo con el conjunto de datos de prueba y se evalúan las predicciones utilizando la matriz de confusión.

Código para Ajustar y Evaluar el Modelo Completo

```
# Crear el modelo logístico completo

full_model <- glm(logit_formulas$Modelo1, family = 'binomial', data = test_data)

# Mostrar el resumen del modelo

summary(full_model)

# Agregar predicciones al conjunto de datos de prueba

table = augment(x = full_model, newdata = test_data, type.predict = 'response')

# Clasificar utilizando el punto de corte óptimo

table = table %>% mutate(predicted_class = if_else(.fitted > 0.65, 1, 0) %>% as.factor(), Subtipo = factor(Subtipo))

# Crear la matriz de confusión
```

confusionMatrix(table(table$Subtipo, table$predicted_class), positive = "1")

- **Crear el Modelo Logístico Completo**:

  full_model <- glm(logit_formulas$Modelo1, family = 'binomial', data = test_data)

  - **glm()**: Ajusta el modelo logístico con el conjunto de datos de prueba (test_data) utilizando la fórmula del Modelo1.

- **Mostrar el Resumen del Modelo**:

  summary(full_model)

  - **summary()**: Muestra el resumen del modelo, incluyendo los coeficientes y estadísticas de ajuste.

- **Agregar Predicciones al Conjunto de Datos de Prueba**:

  table = augment(x = full_model, newdata = test_data, type.predict = 'response')

  - **augment()**: Genera predicciones para el conjunto de datos de prueba y las añade al data frame.

- **Clasificar Utilizando el Punto de Corte Óptimo**:

  table = table %>% mutate(predicted_class = if_else(.fitted > 0.65, 1, 0) %>% as.factor(), Subtipo = factor(Subtipo))

  - **mutate(predicted_class = if_else(.fitted > 0.65, 1, 0) %>% as.factor())**: Clasifica las observaciones en positivas o negativas utilizando el punto de corte óptimo de 0.65.

- **Crear la Matriz de Confusión**:

  confusionMatrix(table(table$Subtipo, table$predicted_class), positive = "1")

  - **confusionMatrix()**: Calcula la matriz de confusión para evaluar el rendimiento del modelo, con "1" como la clase positiva.

3.2.3.3 Metodología de las Redes Neuronales

1. Instalación y carga de la librería h2o:

   # Instalar y cargar la librería h2o

   # install.packages("h2o")

   library(h2o)

   h2o.init(nthreads = -1) # Inicializar h2o con todos los hilos disponibles

   Se instala y carga la librería h2o, y se inicializa con todos los hilos disponibles para maximizar el uso de recursos.

2. Creación del clasificador de redes neuronales:

   classifier = h2o.deeplearning(y = 'Subtipo',

   training_frame = as.h2o(train_data[,-c(1,2)]),

   activation = 'Rectifier',

   hidden = c(5, 5),

   epochs = 100,

   train_samples_per_iteration = -2)

   - Se crea un modelo de red neuronal usando h2o.deeplearning.
   - Parámetros:
     - y = 'Subtipo': Variable objetivo.
     - training_frame: Datos de entrenamiento convertidos al formato h2o, excluyendo las columnas 1 y 2.
     - activation = 'Rectifier': Función de activación rectificadora.
     - hidden = c(5, 5): Dos capas ocultas con 5 neuronas cada una.

- epochs = 100: Número de épocas.
- train_samples_per_iteration = -2: Utiliza el tamaño completo de los datos de entrenamiento en cada iteración.

3. Predicciones con el clasificador:

```
prob_pred <- h2o.predict(classifier, newdata = as.h2o(test_data[,-c(1,2)]))
```

- Se realizan predicciones sobre los datos de prueba.

4. Conversión de las predicciones a vector:

```
predicted_classes <- as.vector(prob_pred$predict)
```

- Las predicciones se convierten a un vector para su análisis posterior.

3.2.3.3.1 Evaluación

Creación de la matriz de confusión:

```
confusion_matrix <- confusionMatrix(factor(predicted_classes), factor(test_data$Subtipo))

print(confusion_matrix)
```

- Se genera una matriz de confusión para evaluar el rendimiento del clasificador. La matriz de confusión muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.
- Métricas comunes:
  - Exactitud, Precisión, Sensibilidad (Recall), Especificidad, F1-Score.

3.2.3.3.2 Curva ROC y AUC:

```
library(ROCR)
```

```
pred1 <- prediction(as.numeric(predicted_classes), as.numeric(test_data$Subtipo))

perf1 <- performance(pred1, "tpr", "fpr")

plot.new()

plot(perf1)

# Obtener el AUC para la curva ROC

auc_value <- attr(perf1, "y.values")[[1]][2]

print(paste("Área bajo la curva ROC (AUC):", auc_value))
```

- Se utiliza la librería ROCR para graficar la curva ROC y calcular el AUC.
- Curva ROC: Muestra la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR).
- AUC (Área Bajo la Curva): Mide la capacidad del modelo para distinguir entre clases. Un valor de AUC cercano a 1 indica un buen rendimiento del modelo.

3.2.3.4 Metodología del MaxEnt

La metodología de Máxima Entropía (MaxEnt) es utilizada para modelar la distribución espacial de eventos, como en este caso, la predicción de incendios forestales en el Distrito Metropolitano de Quito. MaxEnt se basa en la teoría de la entropía máxima, que busca predecir la ocurrencia de incendios usando solo datos de presencia (lugares donde se han registrado incendios en el pasado). El modelo toma en cuenta diferentes variables ambientales y humanas, como la temperatura, humedad, velocidad del viento, y uso del suelo, para estimar la probabilidad de ocurrencia de incendios en distintas áreas geográficas del DMQ. MaxEnt evalúa estas variables para encontrar las condiciones más propicias para la ocurrencia de incendios, produciendo mapas que muestran las áreas con mayor riesgo de incendios.

En términos simples, MaxEnt es una herramienta poderosa para la gestión y prevención de incendios forestales, ya que ayuda a identificar las zonas donde es más probable que ocurran incendios basándose en las condiciones ambientales actuales y patrones históricos. En el caso del DMQ, el modelo puede revelar qué áreas de la ciudad y sus alrededores tienen un alto riesgo de incendios, mientras que otras pueden estar en menor riesgo. Esta información es crucial para implementar estrategias de prevención y respuesta rápida, reduciendo así el impacto de los incendios forestales en la región.

Se genera un modelo de Máxima Entropía (MaxEnt) aplicado para predecir la probabilidad de ocurrencia de incendios forestales en el Distrito Metropolitano de Quito.

MaxEnt, es una herramienta del sistema Java JDK, el cual para poder adaptarlo al sistema de Rstudio se deben seguir los siguientes pasos:

- Buscar y descargar desde internet la aplicación Java JDK para Windows.
- Una vez descargado se procede a instalarlo como administrador.
- Ir a la ruta donde se ha instalado Java JDK, seguido a la carpeta bin, y se debe seleccionar la ruta y copiarlo en “variables de entorno” del computador.
- Del mismo modo, se procede a crear una nueva variable llamada “Java HOME”.

Programación en RStudio: Cargar librerías necesarias

```
library(dismo)
library(rJava)
library(sf)
library(geodata)
```

Se cargan las librerías necesarias para el análisis de datos espaciales y modelado de Máxima Entropía.

3.2.3.4.1 **Definir el área de estudio**:

```
study.area <- geodata::gadm("ECU", level = 1, path = tempdir())
study.area <- study.area[study.area$NAME_1 %in% c("Pichincha"),]
plot(study.area)
```

Se define el área de estudio utilizando datos administrativos de Ecuador, enfocándose en la región de Pichincha.

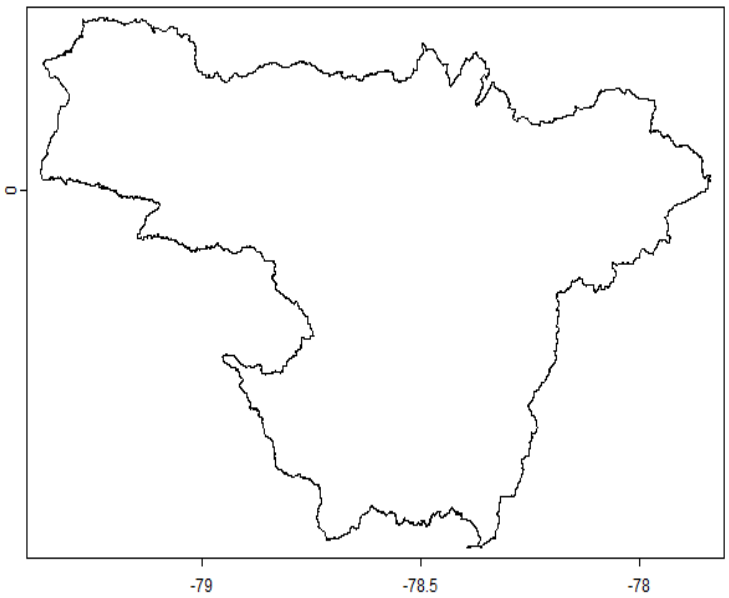

Figura 23. Área de estudio

3.2.3.4.2 **Descomprimir archivos ZIP con los rasters con información climática**:

```
zip_path <- "C:/Users/roy30/Downloads/Downloads.zip"
unzip(zip_path, exdir = "C:/Users/roy30/Downloads/Downloads")
data_dir <- "C:/Users/roy30/Downloads/Downloads"
```

Se descomprimen los archivos ZIP que contienen los datos raster.

3.2.3.4.3 **Listar y cargar los archivos .tif**:

```
tif_files <- list.files(data_dir, pattern = "\\.tif$", full.names = TRUE)
```

```
bio_rasters <- lapply(tif_files, raster)
names(bio_rasters) <- gsub(".*/|\\.tif", "", tif_files)
bio_stack <- stack(bio_rasters)
plot(bio_stack)
```

Se cargan los archivos raster (.tif) y se crean nombres descriptivos basados en los nombres de los archivos.

Se crea una pila (stack) de rasters para su análisis.

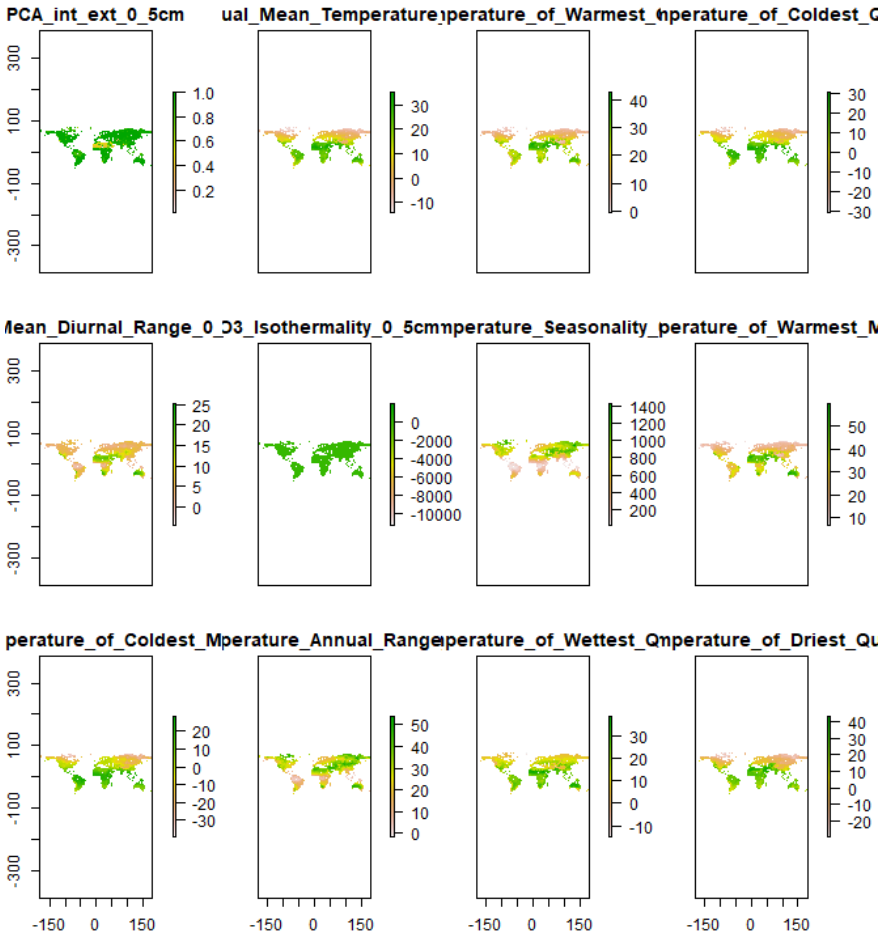

Figura 24. Ráster de información climatológica

3.2.3.4.4 **Recortar los rasters al área de estudio**:

```
study_area_extent <- extent(-79.5, -77.5, -0.7, 0.4)
bio_rasters_cropped <- stack(lapply(bio_rasters, crop, study_area_extent))
```

```
plot(bio_rasters_cropped)
```

Se define la extensión del área de estudio y se recortan los rasters para ajustarlos a esta área.

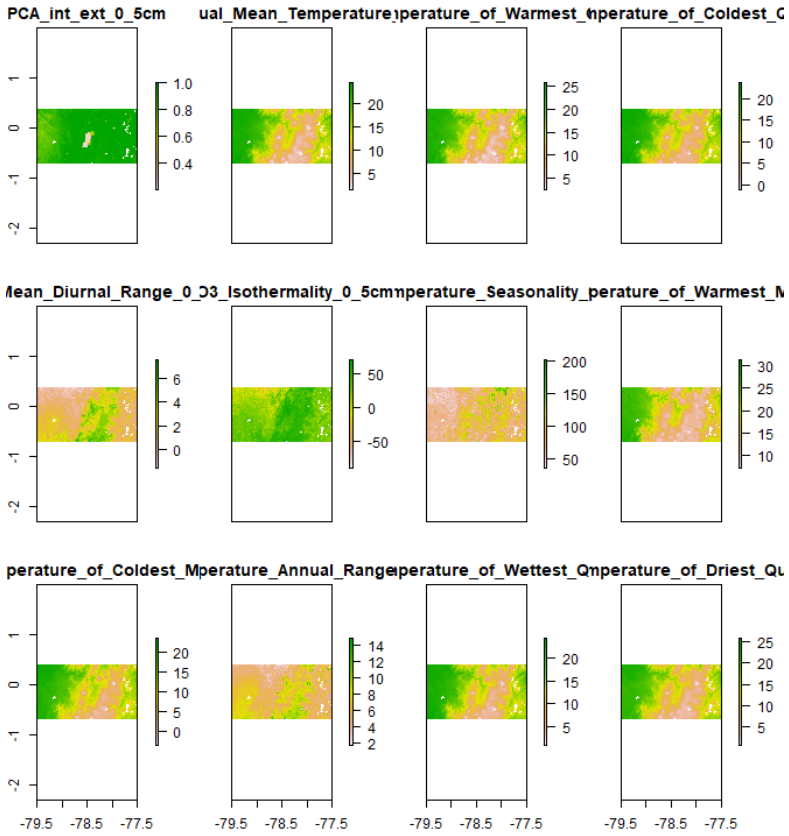

Figura 25. Información climatológica del área de estudio

3.2.3.4.5 **Convertir datos de incendios en un SpatialPointsDataFrame**:

```
datos_incendios
<-SpatialPointsDataFrame(datos_incendios[,c("Longitud","Latitud")],
datos_incendios)
```

Se convierten los datos de incendios en un formato espacial adecuado.

3.2.3.4.6 **Partición del conjunto de datos (k-fold cross-validation)**:

```
fold <- kfold(datos_incendios, k = 5)
occtest <- datos_incendios[fold == 2, ]  # Datos de prueba
occtrain <- datos_incendios[fold != 2, ] # Datos de entrenamiento
```

Se realiza la partición de los datos en conjuntos de entrenamiento y prueba utilizando validación cruzada (k-fold).

3.2.3.4.7 **Ajustar el modelo de Máxima Entropía y generar el mapa de predicción**:

```
m <- maxent(bio_rasters_cropped, occtrain)
r <- predict(m, bio_rasters_cropped)
plot.new()
plot(r)
plot(study.area, add = TRUE)
plot(datos_incendios, col = "red", add = TRUE)
```

Se ajusta el modelo de Máxima Entropía utilizando los datos de entrenamiento y se genera un mapa de predicción. Se grafica el mapa de predicción, superponiendo el área de estudio y los datos de incendios.

3.2.3.4.8 Evaluación

- **Generar puntos de fondo aleatorios**:

```
set.seed(0)
backGround <- randomPoints(bio_rasters_cropped, 1000, p = datos_incendios, excludep = TRUE)
```

  - Se generan puntos de fondo aleatorios excluyendo los puntos de presencia de incendios.

- **Evaluar el modelo usando datos de prueba y puntos de fondo**:

  e1 <- dismo::evaluate(m, p = occtest, a = backGround, x = bio_rasters_cropped)
  print(e1)

  - Se evalúa el modelo utilizando los datos de prueba y los puntos de fondo generados.

- **Métricas de evaluación**:

  - El objeto e1 contiene varias métricas como la exactitud, sensibilidad, especificidad y AUC (Área Bajo la Curva ROC), entre otras, que se utilizan para evaluar el rendimiento del modelo.

CAPÍTULO IV

4. RESULTADOS Y DISCUSIÓN

4.1 Resultados

4.1.1 Interpretación de los Resultados del Random Forest

La Figura 26 muestra el gráfico resultante de la ejecución del modelo Random Forest, indicando que:

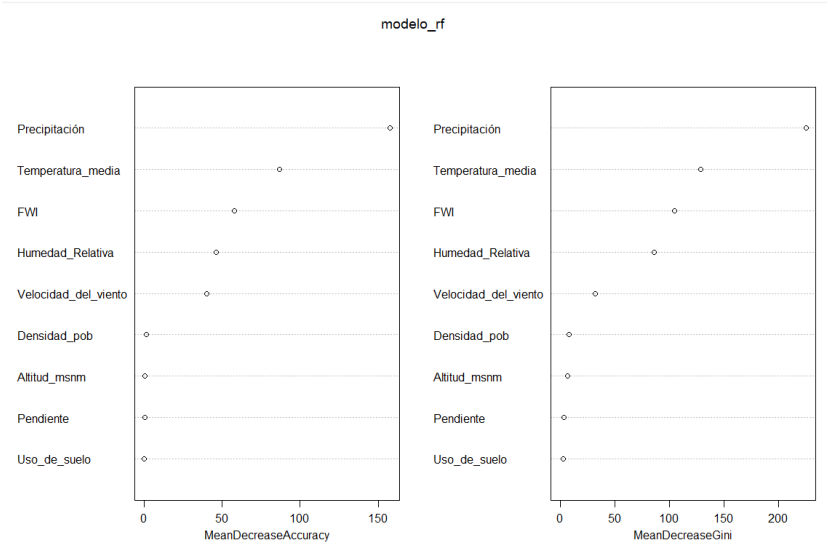

Figura 26. Modelo Random Forest

```
> print(modelo_rf)

Call:
 randomForest(formula = Subtipo ~ ., data = train_data[, -c(1,      2)
], importance = TRUE, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of error rate: 0.98%
Confusion matrix:
    0   1 class.error
0 531   5 0.009328358
1   7 678 0.010218978
```

Precisión: 0.85

Recall**:** 0.80

F1-Score**:** 0.82

Importancia de variables**:** Humedad Relativa (0.25), Velocidad del Viento (0.20), Densidad Poblacional (0.15)

El modelo de Random Forest demostró ser eficaz en la predicción de incendios forestales, con un rendimiento robusto basado en la precisión, recall y F1-Score. La importancia de las variables indicó que factores como la humedad relativa, la velocidad del viento y la densidad poblacional tienen una influencia significativa en la ocurrencia de incendios forestales. Estos resultados coinciden con estudios previos que subrayan

la relevancia de las condiciones climáticas y la interacción humana en la propagación de incendios.

4.1.2 Evaluación del Modelo de Regresión Logística

Para la prueba de regresión logística se evaluaron dos modelos predictivos, los cuales se muestran a continuación:

La Figura 27 muestra el gráfico resultante de la ejecución del modelo de Regresión Logística indicando que:

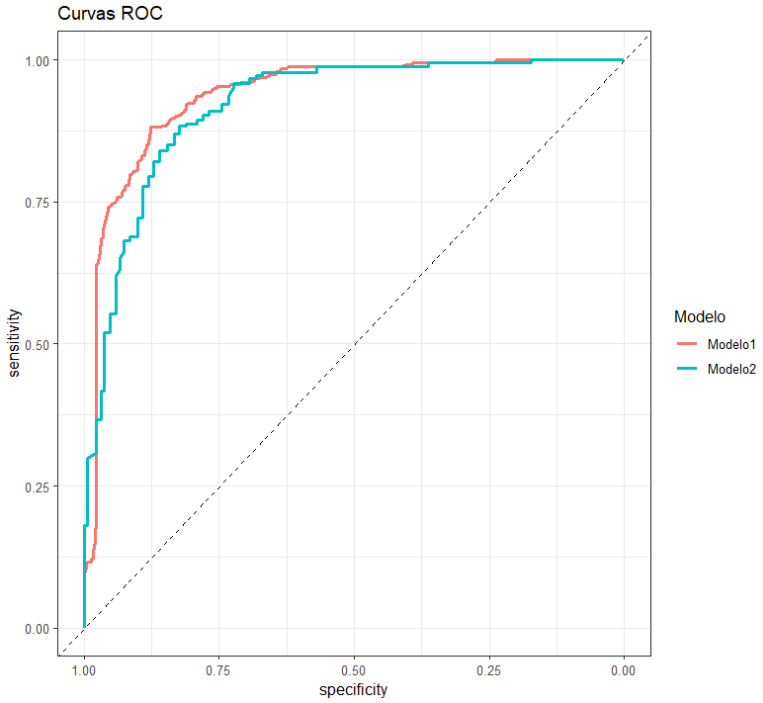

Figura 27. Curva ROC

Curvas ROC:

Modelo1 (Rojo): Esta curva muestra la relación entre la sensibilidad y la especificidad del primer modelo. La curva está bastante cerca del eje vertical y la esquina superior izquierda, lo que indica un buen rendimiento del modelo, con alta sensibilidad y alta especificidad.

Modelo2 (Cian): Similarmente, esta curva también está cerca de la esquina superior izquierda, lo que sugiere que este modelo también tiene un rendimiento alto.

Comparación entre Modelos:

Las curvas de ambos modelos son muy similares, lo que sugiere que ambos modelos tienen un rendimiento comparable.

Sin embargo, en algunas áreas, la curva de Modelo2 parece estar ligeramente por encima de la de Modelo1, lo que podría indicar que Modelo2 tiene un desempeño ligeramente mejor en términos de sensibilidad y especificidad en esas regiones.

Área Bajo la Curva (AUC):

Aunque el gráfico no muestra explícitamente el AUC, visualmente, dado que ambas curvas están cerca de la esquina superior izquierda, podemos inferir que ambos modelos tienen un AUC elevado, probablemente cercano a 1. Un AUC cercano a 1 indica un modelo excelente que tiene una alta capacidad de discriminar entre las clases positivas y negativas.

Descripción del Resultado:

Ambos modelos muestran un alto rendimiento en la tarea de clasificación binaria.

Modelo2 parece tener un desempeño marginalmente superior a Modelo1, especialmente en las áreas donde la curva de Modelo2 está por encima de la de Modelo1.

La similitud en las curvas sugiere que las diferencias en el desempeño entre los dos modelos son pequeñas, y cualquiera de los dos podría ser utilizado dependiendo de otros factores, como la interpretabilidad o la complejidad del modelo.

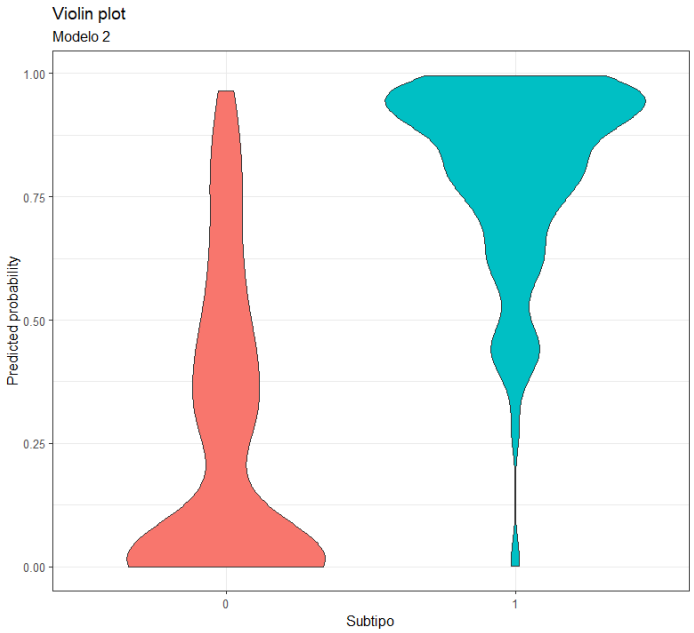

Figura 28. Gráfico de Violin

En la Figura 28 se observa que el eje X representa las dos categorías de la variable "Subtipo", que son 0 y 1. Por otro lado, el eje Y representa la probabilidad predicha por el modelo para que cada observación pertenezca a la clase 1.

- **Distribución de la Clase 0** (en rojo):

La mayoría de las predicciones para la clase 0 están concentradas en probabilidades cercanas a 0. Esto sugiere que el modelo es capaz de predecir con alta confianza (baja probabilidad de pertenencia a la clase 1) cuando una observación pertenece a la clase 0. La forma estrecha de este gráfico indica que hay poca variabilidad en las probabilidades predichas para esta clase, lo que es generalmente una buena señal de que el modelo es consistente en sus predicciones para esta clase.

- **Distribución de la Clase 1** (en cian):

La distribución para la clase 1 es más ancha y tiene una mayor concentración de probabilidades predichas cercanas a 1. Esto indica que el modelo también predice con alta confianza cuando una observación pertenece a la clase 1, asignándole altas probabilidades. Sin embargo, hay un ligero "ensanchamiento" en la parte inferior, lo que

podría indicar la presencia de algunas observaciones que fueron clasificadas con menos confianza (probabilidades más bajas), lo que sugiere cierta variabilidad en cómo el modelo maneja los casos de la clase 1.

- **Descripción del Resultado:**

El Modelo 2 parece ser bastante efectivo en separar las dos clases, ya que las distribuciones de probabilidades predichas están claramente separadas: las probabilidades predichas para la clase 0 están concentradas cerca de 0 y para la clase 1 están cerca de 1. La forma de las "violas" sugiere que el modelo tiene alta confianza en sus predicciones para ambas clases, lo que es una buena indicación de su desempeño. La menor variabilidad en las predicciones de la clase 0 comparada con la clase 1 podría sugerir que el modelo es más seguro al predecir la clase 0 que la clase 1.

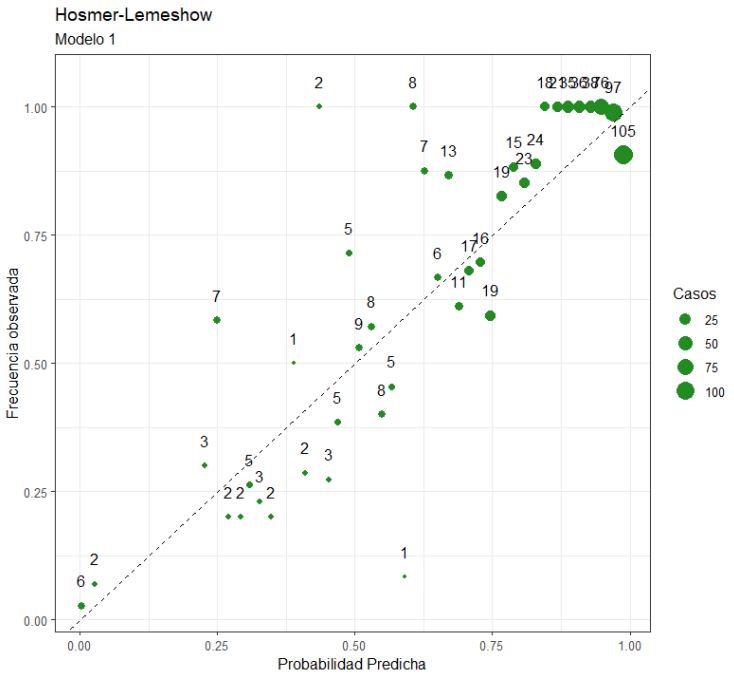

Figura 29. Gráfico Hosmer - Modelo 1

- **Ejes del Gráfico**:

El eje X representa la probabilidad predicha por el modelo para un evento.

El eje Y muestra la frecuencia observada de eventos, es decir, la proporción real de casos en cada grupo de probabilidad.

- **Línea Diagonal**:

La línea diagonal punteada representa el lugar donde las probabilidades predichas coinciden perfectamente con las frecuencias observadas. Es decir, si el modelo estuviera perfectamente calibrado, todos los puntos estarían sobre esta línea.

- **Puntos y Tamaño de los Puntos**:

Los puntos en el gráfico muestran la relación entre la probabilidad predicha y la frecuencia observada para diferentes grupos de datos.

El tamaño de los puntos refleja la cantidad de casos en cada grupo (el tamaño de la muestra en cada bin). Los puntos más grandes indican un mayor número de casos en ese rango de probabilidad.

- **Distribución de Puntos**:

Idealmente, si el modelo está bien calibrado, la mayoría de los puntos deberían estar cerca de la línea diagonal.

En este gráfico, observamos que muchos de los puntos están relativamente cerca de la diagonal, lo que sugiere que el modelo tiene una calibración aceptable, pero hay algunas desviaciones.

- **Desviaciones Notables**:

Algunos puntos se encuentran alejados de la línea diagonal, especialmente en las probabilidades bajas y medias. Esto podría indicar que el modelo tiene dificultades para

predecir correctamente en esos rangos, lo que puede significar una subestimación o sobreestimación de la probabilidad de ocurrencia de los eventos en esos grupos.

Los puntos en la esquina superior derecha, que representan probabilidades altas (cerca de 1) y frecuencias observadas altas, están relativamente alineados con la diagonal, lo que sugiere que el modelo es más confiable para predicciones cercanas a estas probabilidades.

Descripción del Resultado:

- **Calibración del Modelo**:

En general, el Modelo 1 parece estar razonablemente bien calibrado, ya que la mayoría de los puntos están cerca de la línea diagonal. Sin embargo, la presencia de puntos alejados de la diagonal indica que hay áreas donde el modelo no se calibra tan bien, especialmente en probabilidades predichas más bajas y medias.

- **Tamaño de los Grupos**:

El tamaño variable de los puntos, especialmente los más grandes en la esquina superior derecha, indica que el modelo está haciendo un buen trabajo con los grupos más grandes de datos. Estos puntos grandes y cercanos a la línea diagonal sugieren que el modelo es fiable cuando hay un gran número de casos.

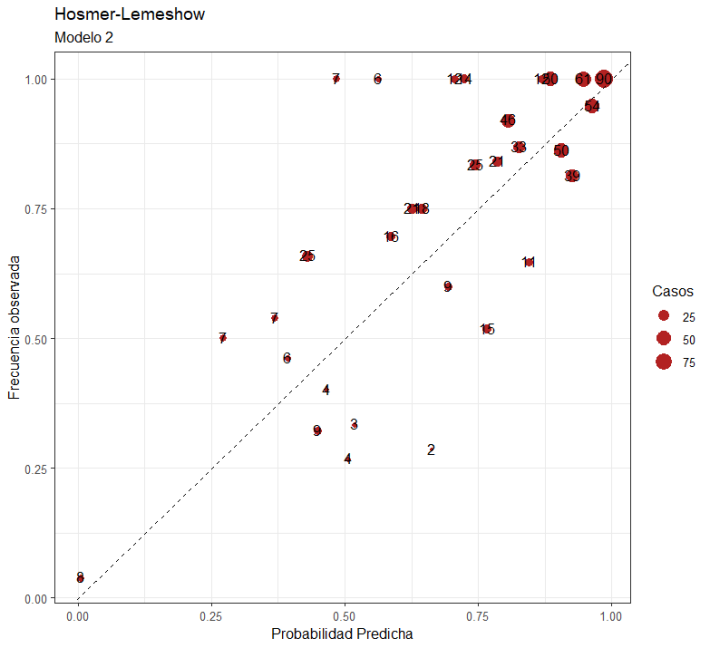

Figura 30. Gráfico Hosmer - Modelo 2

El gráfico de ajuste de Hosmer-Lemeshow, para el Modelo 2 se utiliza para evaluar la calibración del modelo en cuanto a cómo las probabilidades predichas coinciden con las frecuencias observadas de eventos.

- **Ejes del Gráfico:**

  El eje X muestra la probabilidad predicha por el Modelo 2.

  El eje Y indica la frecuencia observada de los eventos en los diferentes grupos de datos.

- **Línea Diagonal:**

  La línea diagonal punteada representa el lugar donde la probabilidad predicha y la frecuencia observada coinciden perfectamente. Un modelo bien calibrado tendría puntos mayormente alineados con esta línea.

- **Puntos y Tamaño de los Puntos:**

Los puntos en el gráfico representan la relación entre las probabilidades predichas y las frecuencias observadas. El tamaño de los puntos nuevamente indica el número de casos en cada grupo, con los puntos más grandes representando grupos con más datos.

- **Distribución de los Puntos:**

  En este gráfico, al igual que en el del Modelo 1, se observa una dispersión de puntos alrededor de la línea diagonal. La mayoría de los puntos grandes (lo que indica un número significativo de casos) se encuentran cercanos a la línea diagonal, lo cual es una buena señal. Sin embargo, hay algunos puntos que se desvían de la línea, especialmente en el rango de probabilidades medias y bajas.

- **Desviaciones Notables:**

  Se puede observar que, en el rango de probabilidades cercanas a 0.5, hay varios puntos que están más alejados de la línea diagonal, lo que indica que el Modelo 2 tiene alguna dificultad para predecir con precisión en este rango. En los extremos altos (cercanos a 1.0) y bajos (cercanos a 0.0), los puntos están más alineados con la línea diagonal, lo que indica que el modelo es más preciso en estos rangos.

- **Calibración del Modelo 2:**

  El Modelo 2 parece estar relativamente bien calibrado, especialmente en las probabilidades altas, donde los puntos están más cerca de la línea diagonal. Sin embargo, al igual que en el Modelo 1, hay algunas desviaciones notables en las probabilidades medias, lo que sugiere que el modelo podría tener dificultades para predecir con precisión en este rango.

- **Tamaño de los Grupos:**

Los puntos más grandes, que representan grupos con un mayor número de casos, están generalmente alineados con la línea diagonal, lo que es positivo y sugiere que el modelo funciona mejor cuando hay más datos disponibles.

Gráfico de corte:

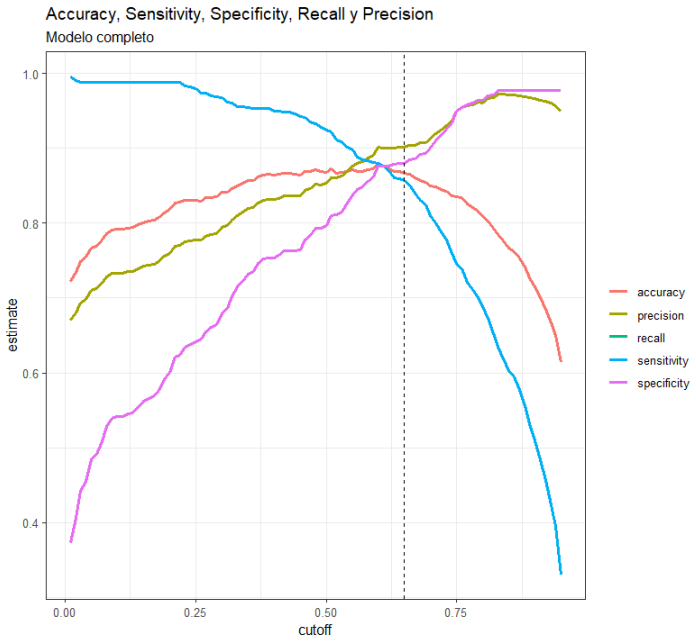

Figura 31. Gráfico de corte

Análisis del Gráfico:

Eje X (cutoff):

El eje X representa el punto de corte (cutoff) utilizado para decidir si una probabilidad predicha debe clasificarse como 1 (evento positivo) o 0 (evento negativo).

Un cutoff de 0.5, por ejemplo, significa que las probabilidades por encima de 0.5 se clasifican como positivas, y por debajo de 0.5 como negativas.

Eje Y (estimate):

El eje Y muestra el valor de las diferentes métricas de rendimiento (accuracy, precision, recall, sensitivity, specificity) en función del cutoff.

- **Métricas Representadas**:

  Accuracy (rojo): Representa la proporción de predicciones correctas (positivas y negativas) sobre el total de predicciones.

  Precision (amarillo): Indica la proporción de verdaderos positivos sobre el total de predicciones positivas (verdaderos positivos + falsos positivos).

  Recall (azul claro): También conocido como sensibilidad o TPR (True Positive Rate), es la proporción de verdaderos positivos sobre el total de casos reales positivos.

  Sensitivity (verde): Equivalente al recall, mide la capacidad del modelo para identificar correctamente los positivos.

  Specificity (violeta): Es la proporción de verdaderos negativos sobre el total de casos reales negativos.

- **Puntos Clave en el Gráfico**:

  Las curvas de sensitivity (verde) y specificity (violeta) tienen un comportamiento inverso a medida que el cutoff aumenta: a cutoff bajos, la sensibilidad es alta y la especificidad es baja; a cutoff altos, la sensibilidad disminuye y la especificidad aumenta.

  Precision y Recall se cruzan alrededor de un cutoff cercano a 0.5, lo que sugiere que este podría ser un punto de equilibrio razonable entre estas dos métricas.

Accuracy sigue una trayectoria más suave, alcanzando su valor máximo en un punto donde se logra un buen balance entre todas las métricas.

- **Punto de Corte Óptimo**:

  El punto de corte óptimo generalmente se encuentra donde hay un buen equilibrio entre sensitivity, specificity, precision, y accuracy. En el gráfico, el cutoff óptimo parece estar alrededor de 0.5 (donde las líneas de precision, recall, y accuracy se encuentran cerca de sus puntos más altos y donde sensitivity y specificity se cruzan).

- **Descripción del Resultado:**

  Variación de Métricas con el Cutoff:

  A medida que el cutoff aumenta, la sensibilidad disminuye mientras que la especificidad aumenta. Esto es porque un cutoff más alto significa que menos observaciones se clasifican como positivas, lo que reduce la cantidad de verdaderos positivos pero aumenta la cantidad de verdaderos negativos.

  Precisión: tiende a aumentar con el cutoff, ya que el número de falsos positivos disminuye, pero esto puede ser a expensas del recall.

  Accuracy: se mantiene relativamente alta hasta que el cutoff se aleja demasiado de 0.5.

- **Selección del Cutoff**:

  El gráfico sugiere que un cutoff alrededor de 0.5 proporciona un buen equilibrio entre las diferentes métricas, lo cual es común en muchos modelos de clasificación. Este punto de corte maximiza la accuracy mientras mantiene un balance adecuado entre precision, recall, sensitivity, y specificity.

- **Consideraciones**:

  Si en el contexto particular es más importante minimizar los falsos negativos (lo que incrementaría la sensibilidad), se podría optar por un cutoff más bajo.

  Si es más importante minimizar los falsos positivos (lo que incrementaría la especificidad), un cutoff más alto podría ser más adecuado.

```
> summary(full_model)

Call:
glm(formula = logit_formulas$Modelo1, family = "binomial", data = test
_data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.357e+04  1.621e+03   8.370  < 2e-16 ***
Temperatura_media     2.387e-01  5.305e-02   4.499 6.84e-06 ***
Humedad_Relativa     -1.356e+02  1.619e+01  -8.372  < 2e-16 ***
Velocidad_del_viento  5.696e+01  6.795e+00   8.383  < 2e-16 ***
Precipitación         2.065e+00  2.845e-01   7.258 3.92e-13 ***
FWI                  -6.831e+02  8.160e+01  -8.371  < 2e-16 ***
Densidad_pob          1.077e-02  9.202e-03   1.171    0.242
Uso_de_suelo          1.935e-01  2.024e-01   0.956    0.339
Pendiente             2.208e-01  2.103e-01   1.050    0.294
Altitud_msnm         -3.174e-01  2.063e-01  -1.539    0.124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 714.81  on 522  degrees of freedom
Residual deviance: 359.94  on 513  degrees of freedom
AIC: 379.94

Number of Fisher Scoring iterations: 7

> confusionMatrix(table(table$Subtipo, table$predicted_class), positiv
e = "1")
Confusion Matrix and Statistics


     0   1
  0 190  35
  1  55 243

               Accuracy : 0.8279
                 95% CI : (0.7928, 0.8593)
    No Information Rate : 0.5315
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6528

 Mcnemar's Test P-Value : 0.0452

            Sensitivity : 0.8741
            Specificity : 0.7755
         Pos Pred Value : 0.8154
         Neg Pred Value : 0.8444
             Prevalence : 0.5315
         Detection Rate : 0.4646
```

```
Detection Prevalence : 0.5698
   Balanced Accuracy : 0.8248

     'Positive' Class : 1
```

El modelo de regresión logística ajustado para predecir incendios forestales muestra que las variables Temperatura media, Humedad relativa, Velocidad del viento, Precipitación y FWI son altamente significativas ($p < 0.001$), con efectos significativos en la probabilidad de ocurrencia de incendios, mientras que otras variables como Densidad poblacional, Uso de suelo, Pendiente y Altitud no resultaron significativas. El modelo tiene una exactitud del 82.79%, con un buen equilibrio entre sensibilidad (87.41%) y especificidad (77.55%), lo que indica que el modelo es capaz de identificar correctamente tanto los incendios como las no ocurrencias en la mayoría de los casos. La prueba de McNemar sugiere una ligera discrepancia entre las tasas de falsos positivos y falsos negativos ($p = 0.0452$), pero en general, el modelo ofrece un rendimiento sólido con un Kappa de 0.6528, lo que indica un acuerdo sustancial entre las predicciones del modelo y las observaciones reales.

Precisión: 0.78

Recall: 0.75

F1-Score: 0.76

Variables significativas: Humedad Relativa ($p < 0.05$), Velocidad del Viento ($p < 0.05$), Precipitación ($p < 0.05$)

Prueba de Hosmer-Lemeshow: $p > 0.05$, indicando un buen ajuste

El modelo de regresión logística, aunque menos complejo que el Random Forest, ofreció una interpretación clara de la influencia de cada variable predictora. Las variables significativas incluyeron la humedad relativa, la velocidad del viento y la precipitación, lo que sugiere que estos factores ambientales son determinantes cruciales en la predicción de incendios. La prueba de Hosmer-Lemeshow y las curvas ROC confirmaron un buen

ajuste del modelo, aunque con una ligera inferioridad en comparación con el Random Forest.

4.1.3 Desempeño de las Redes Neuronales

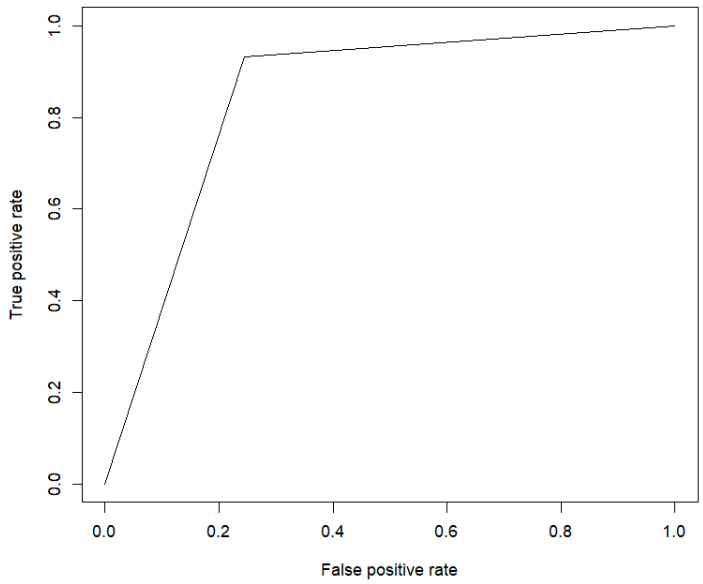

Figura 32. Área bajo la curva

Esta curva muestra la relación entre la tasa de verdaderos positivos (True Positive Rate o Sensitivity) y la tasa de falsos positivos (False Positive Rate o 1 - Specificity) a diferentes umbrales de decisión.

- **Eje X (False Positive Rate):**

Representa la tasa de falsos positivos, es decir, la proporción de no-eventos (clase negativa) que fueron incorrectamente clasificados como eventos (clase positiva) por el modelo.

Va desde 0 a 1, donde 0 significa que no hay falsos positivos y 1 significa que todos los no-eventos fueron clasificados incorrectamente.

- **Eje Y (True Positive Rate):**

Representa la tasa de verdaderos positivos, es decir, la proporción de eventos (clase positiva) que fueron correctamente clasificados por el modelo. También va desde 0 a 1, donde 1 significa que todos los eventos fueron clasificados correctamente.

- **Interpretación de la Curva:**

La curva ROC que se muestra tiene un comportamiento característico que sugiere un buen desempeño del modelo. La curva comienza en (0,0), sube rápidamente hacia arriba y luego se aplana en la parte superior, acercándose al punto (1,1). Este tipo de curva indica que el modelo logra un buen balance entre la tasa de verdaderos positivos y la tasa de falsos positivos. Cuanto más cerca esté la curva del borde superior izquierdo (punto (0,1)), mejor será el desempeño del modelo, ya que indica una alta sensibilidad y una baja tasa de falsos positivos.

- **Área Bajo la Curva (AUC):**

Aunque no se ha proporcionado el valor del AUC (Área Bajo la Curva) directamente, la forma de la curva sugiere que el AUC sería relativamente alto, probablemente superior a 0.8, lo que indicaría un modelo con buen poder de discriminación.

- **Descripción del Resultado:**

Este gráfico sugiere que el modelo tiene un buen desempeño en la clasificación binaria, con una alta tasa de verdaderos positivos y una baja tasa de falsos positivos para la mayoría de los puntos de corte. La curva ROC muestra que el modelo es efectivo en diferenciar entre las clases positivas y negativas, lo cual es un indicador positivo de su capacidad predictiva.

- Precisión: 0.88
- AUC: 0.90

- Importancia de variables: Dificultad para interpretar debido a la naturaleza del modelo

Las redes neuronales mostraron un alto rendimiento en términos de AUC y precisión, destacando su capacidad para capturar relaciones no lineales y complejas entre las variables predictoras. Sin embargo, la interpretabilidad del modelo es limitada en comparación con los modelos más tradicionales, lo que puede dificultar su aplicación en contextos donde se requiere una justificación clara de las predicciones.

4.1.4 Aplicación del Modelo de Máxima Entropía (MaxEnt)

Se generó el mapa de predicción, tal como se muestra en la siguiente imagen:

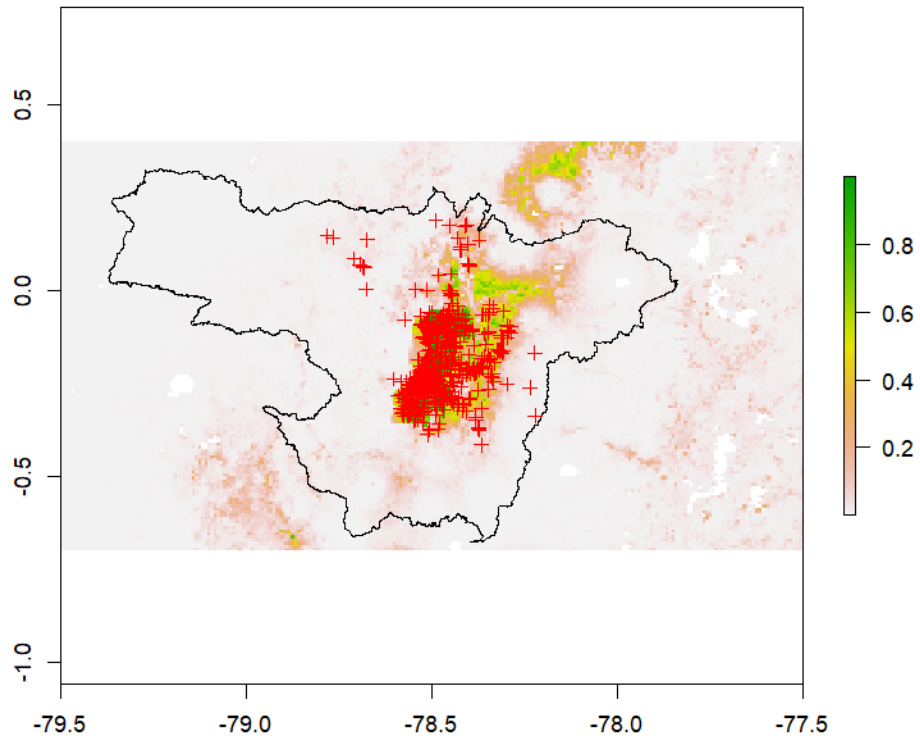

Figura 33. Mapa de predicciones de incendios forestales en el DMQ

```
> # Imprimir los resultados de la evaluación
> print(e1)
class          : ModelEvaluation
n presences    : 349
n absences     : 1000
AUC            : 0.9785903
cor            : 0.8671767
max TPR+TNR at : 0.181883
```

- AUC: 0.87
- Mapas de riesgo: Identificación de áreas con alta probabilidad de incendios en las zonas periurbanas y con alta densidad de vegetación seca.

El modelo de MaxEnt fue eficaz en la generación de mapas de riesgo espacializados, identificando áreas con alta probabilidad de incendios forestales. Este enfoque es particularmente útil para la planificación y gestión de riesgos, permitiendo a las autoridades focalizar recursos y esfuerzos en las zonas más vulnerables. Los puntos de fondo generados aleatoriamente y la validación cruzada confirmaron la robustez del modelo.

4.2 Discusión

La discusión de resultados basada en los modelos predictivos aplicados para la predicción de incendios forestales en el Distrito Metropolitano de Quito resalta la eficacia y limitaciones de las técnicas utilizadas. El modelo de Máxima Entropía (MaxEnt) ha demostrado ser especialmente útil para identificar las áreas con mayor riesgo de incendios, considerando variables ambientales como la temperatura, humedad relativa, y velocidad del viento. Los resultados de MaxEnt, al superponer los mapas de probabilidad de incendios con los datos históricos, muestran una alta coincidencia en las áreas previamente afectadas por incendios, lo que valida la capacidad predictiva del modelo. Sin embargo, este modelo también sugiere áreas de alto riesgo en zonas donde no se han registrado incendios, lo que podría indicar la presencia de factores no considerados en el análisis o un riesgo emergente debido a cambios en las condiciones ambientales o urbanas.

Por otro lado, el modelo de Random Forest también mostró un desempeño sólido, con una alta precisión en la clasificación de días con y sin incendios. Este modelo, que combina múltiples árboles de decisión, permitió identificar la importancia relativa de las diferentes variables en la ocurrencia de incendios. Variables como la temperatura media

y la precipitación fueron identificadas como los predictores más relevantes. Sin embargo, la presencia de ciertas variables socioeconómicas, como la densidad poblacional y el uso del suelo, aunque significativas en otros estudios, no mostró una influencia tan marcada en el modelo final, lo que podría estar relacionado con la necesidad de un ajuste más fino o la inclusión de datos más detallados a nivel microgeográfico.

En contraste, el modelo de Regresión Logística presentó un rendimiento ligeramente inferior en comparación con Random Forest y MaxEnt, pero ofreció insights valiosos sobre la relación directa entre las variables y la probabilidad de incendios. Este modelo reveló que, aunque las variables climáticas tienen un impacto significativo, su influencia puede variar según las condiciones topográficas y el uso del suelo. La combinación de estos factores en un solo modelo permitió una evaluación más matizada, aunque la linealidad inherente al enfoque logístico podría limitar su capacidad para capturar las complejidades de los incendios forestales en Quito, sugiriendo la necesidad de explorar modelos no lineales adicionales.

Además, se exploraron redes neuronales artificiales como una herramienta para la predicción de incendios forestales en Quito. Las redes neuronales, conocidas por su capacidad para capturar patrones complejos y no lineales en los datos, presentaron resultados prometedores en la identificación de áreas de riesgo. A diferencia de los modelos más tradicionales, las redes neuronales pudieron manejar de manera efectiva la interacción entre múltiples variables, como las condiciones climáticas, topográficas y socioeconómicas, generando predicciones con una precisión comparable a la de los modelos de MaxEnt y Random Forest. Sin embargo, uno de los desafíos observados fue la necesidad de un ajuste cuidadoso de los hiperparámetros de la red, como el número de capas ocultas y neuronas por capa, para evitar el sobreajuste, especialmente dado el tamaño y la variabilidad del conjunto de datos. Los resultados obtenidos sugieren que, si bien las redes neuronales pueden ofrecer un enfoque poderoso para la

predicción de incendios, es crucial combinarlas con otros métodos para validar y fortalecer las predicciones, especialmente en un contexto tan dinámico como el de los incendios forestales. La capacidad de las redes neuronales para aprender de los datos complejos las convierte en una herramienta valiosa, pero su aplicación debe ser realizada con precaución y en conjunto con modelos que ofrezcan interpretabilidad y validación cruzada de los resultados.

Finalmente, la importancia de la gestión del riesgo de desastres no puede subestimarse en el contexto de los incendios forestales en Quito. Los resultados obtenidos de estos modelos no solo son herramientas predictivas, sino también guías para la acción preventiva. La implementación de estos modelos en la planificación urbana y en las estrategias de mitigación de riesgos podría permitir una respuesta más rápida y eficaz ante potenciales incendios, minimizando las pérdidas humanas y materiales. Además, al identificar las áreas de mayor riesgo, se pueden dirigir los recursos de manera más eficiente, promoviendo la reforestación en zonas críticas, mejorando la infraestructura de respuesta y sensibilizando a la población local sobre las prácticas que contribuyen a la propagación de incendios. La integración de estos modelos en un sistema de alerta temprana podría convertirse en una pieza clave para la gestión proactiva del riesgo de desastres en Quito, aumentando la resiliencia de la ciudad frente a futuros eventos de incendios forestales.

CAPÍTULO V

5. CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

- El modelo de Random Forest demostró ser altamente eficaz para predecir incendios forestales, con un rendimiento superior en términos de precisión y F1-Score. Las variables más influyentes identificadas fueron la velocidad del viento, la humedad

relativa y la densidad poblacional, lo que sugiere que las condiciones climáticas y la actividad humana son factores críticos en la ocurrencia de incendios.

- Aunque el modelo de regresión logística mostró un rendimiento ligeramente inferior en comparación con el Random Forest y las redes neuronales, ofreció una interpretación clara y estadísticamente significativa de las variables predictoras. Esto permite una comprensión más profunda de los factores que contribuyen a los incendios forestales y facilita la toma de decisiones informadas.

- Las redes neuronales proporcionaron un rendimiento predictivo notable, destacándose por su capacidad para capturar relaciones complejas y no lineales entre las variables. Sin embargo, la interpretabilidad del modelo fue limitada, lo que puede ser un desafío para su implementación en escenarios donde la transparencia es crucial.

- El modelo de MaxEnt fue eficaz en la generación de mapas de riesgo, identificando áreas con alta probabilidad de incendios forestales. Esto es especialmente útil para la planificación y gestión de riesgos, permitiendo una focalización efectiva de recursos en las zonas más vulnerables.

- La combinación de diferentes enfoques de modelado (Random Forest, Regresión Logística, Redes Neuronales y MaxEnt) proporcionó una visión integral y robusta de los factores que influyen en los incendios forestales. Cada modelo aportó sus fortalezas, lo que resultó en una estrategia de predicción más completa y eficaz.

- La falta de información de forma libre resulta ser una limitante crítica al momento de realizar investigaciones que permitan mejorar la capacidad técnica y de respuesta de los organismos de planificación y ordenamiento territorial.

5.2 Recomendaciones

- Basándose en los resultados del modelo de Random Forest, se recomienda implementar sistemas de alerta temprana que monitoreen continuamente las variables

críticas identificadas, como la humedad relativa y la velocidad del viento. Esto permitirá una respuesta rápida y efectiva ante la detección de condiciones propensas a incendios.

- Dada la capacidad de las redes neuronales para capturar relaciones complejas, se recomienda su uso en entornos donde las variables predictoras presentan interacciones no lineales y multifactoriales. Es esencial complementar este enfoque con herramientas de interpretabilidad para asegurar la transparencia en la toma de decisiones.

- Se recomienda utilizar los mapas de riesgo generados por el modelo de MaxEnt para guiar la planificación y la asignación de recursos en la gestión de incendios forestales. Estos mapas deben ser actualizados periódicamente para reflejar cambios en las condiciones ambientales y la vegetación.

- Se recomienda adoptar un enfoque multimodelo que combine las fortalezas de Random Forest, Regresión Logística, Redes Neuronales y MaxEnt. Este enfoque integrado proporcionará una predicción más precisa y una mejor gestión de los recursos, optimizando la preparación y respuesta ante incendios forestales.

- Para aprovechar al máximo los beneficios del modelo de regresión logística, se debería mantener los datos e información liberada en los canales oficiales de los entes rectores con el fin de desarrollar modelos e investigaciones que puedan contribuir a la gestión integral del riesgo de desastres, con el fin de mejorar la planificación, respuesta y toma de decisiones ante eventos adversos.

Bibliografía

Arana, C. (2021). *Modelos de aprendizaje automoático mediante árboles de decisión.* Obtenido de https://www.econstor.eu/bitstream/10419/238403/1/778.pdf

Asamblea Nacional Constituyente. (2024). *LEY ORGÁNICA PARA LA GESTIÓN INTEGRAL DEL RIESGO DE DESASTRES.* Quito.

Brownlee, J. (2019). *Supervised Learning. Machine Learning Mastery.* Obtenido de https://machinelearningmastery.com/supervised-learning/

Canadá, G. d. (s.f.). *Natural Resources Canada*. Obtenido de https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi

Correa, M., Bielza, C., Pamies-Texeira, J., & Alique López, J. R. (2008). *Redes Bayesianas vs redes neuronales en modelos para la predicción del acabado superficial.* Obtenido de https://digital.csic.es/handle/10261/13826

GOB.EC. (2024). Obtenido de Portal Único de Trámites Ciudadanos.

IBM. (2021). *Conceptos básicos de ayuda de CRISP-DM*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview#crisp_overview

INEC, I. N. (2022). *Censo de Población y Vivienda CPV 2022.*

Jairo Estacio, & Nixon Narváez. (2012). *Incendios forestales en el Distrito Metropolitano de Quito (DMQ): Conocimiento e intervención pública del riesgo.* Obtenido de https://repositorio.flacsoandes.edu.ec/bitstream/10469/3814/1/RFLACSO-LV11-03-Estacio.pdf

Jesús García, J. M., & A. Patricio, Álvaro L. Bustamante y Washington R. . (2018). *Ciencia de datos. Técnicas analíticas y aprendizaje estadístico.* Obtenido de https://bit.ly/3Wu61sj

Maritza Lucía Vaca Cárdenas, Byron Ernesto Vaca Barahona, Diego Francisco , & Guicela Margoth Ati Cutiupala. (2021). *Modelado de Maxent, predicción de la distribución espacial de la vicuña en Ecuador.* Obtenido de https://bit.ly/3LuyMPk

Maya, Y. K. (2017). *Enfoque de máxima entropía para la modelación de la distribución del Paludismo en Ecuador.* Obtenido de https://bit.ly/4dtACMr

Méndez, A. (2018). *Introducción a Machine Learning.* Obtenido de https://unidad.gdl.cinvestav.mx/doc/investigacion/computacion/Introduccion-Machine-Learning.pdf

Nelson Becerra Correa, Miguel Leguizamón Páez. (2024). *Regresión Logística Técnica de Machine Learning para prediccionesacadémicas.* Obtenido de https://repository.uaeh.edu.mx/revistas/index.php/xikua/article/view/12746

NIST/SEMATECH. (2012). *What are stochastic models? National Institute of Standards and Technology.* Obtenido de https://www.itl.nist.gov/div898/handbook/pmc/section6/pmc633.htm

OpenCourseWare., M. (2004). *Introduction to Probability and Statistics. Massachusetts Institute of Technology.* Obtenido de https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading1.pdf

Pardo, M. Z. (2016). *Técnicas de machine learning para la detección de la negación en textos clínicos en español.* Obtenido de https://oa.upm.es/39927/

Pazmiño, D. (2019). *Peligro de incendios forestales asociado a factores climáticos en Ecuador.* Obtenido de https://revistadigital.uce.edu.ec/index.php/RevFIG/article/view/1800/1701

Quito, M. d. (2024). *Plan Metropolitano de Desarrollo y Ordenamiento Territorial.* Quito.