Coberturas del país a nivel de tecnología

Anderson ortiz , Cristian Rojas & Kevin Bohorquez

Análisis y desarrollo de software

Esteban Hernandez

Analitica de datos

12 de diciembre 2025
Mosquera Cundinamarca

Introducción

La cobertura de los servicios de telecomunicaciones móviles constituye un elemento fundamental para el desarrollo económico, social y tecnológico de los territorios. En Colombia, la disponibilidad de tecnologías móviles avanzadas como 4G presenta diferencias significativas entre municipios, lo que evidencia la necesidad de contar con herramientas analíticas que permitan identificar y predecir la presencia de este tipo de cobertura de manera objetiva y basada en datos.

El presente proyecto tiene como objetivo desarrollar un modelo de clasificación basado en Random Forest que permita predecir la presencia de cobertura 4G en los municipios de Colombia, a partir de variables demográficas, socioeconómicas, geográficas y de infraestructura, obtenidas de fuentes oficiales de datos abiertos. Para ello, se emplean técnicas de análisis de datos y aprendizaje automático que facilitan la identificación de patrones relevantes en la información disponible.

Mediante este enfoque, se busca no solo analizar la distribución de la cobertura 4G en el territorio nacional, sino también generar un modelo predictivo que apoye la toma de decisiones estratégicas relacionadas con la planeación territorial, la inversión en infraestructura de telecomunicaciones y la formulación de políticas públicas, orientadas a reducir la brecha digital y fortalecer la conectividad móvil.

Contexto del Proyecto

En el contexto actual, la expansión y fortalecimiento de la cobertura 4G constituye uno de los principales retos para el desarrollo territorial en Colombia. A pesar de los avances en infraestructura tecnológica, aún existen municipios que no cuentan con este tipo de cobertura, lo que limita el acceso a servicios digitales esenciales y afecta ámbitos como la educación, la productividad, el acceso a la información y la inclusión digital.

Las entidades gubernamentales como el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), la Comisión de Regulación de Comunicaciones (CRC), el Departamento Administrativo Nacional de Estadística (DANE) y la Agencia Nacional del Espectro (ANE) generan y publican información clave que permite analizar estas problemáticas desde una perspectiva basada en datos. La disponibilidad de estos conjuntos de datos facilita el estudio de variables demográficas, socioeconómicas, geográficas y de infraestructura que influyen en la presencia de cobertura 4G en los municipios.

En este escenario, el uso de modelos de aprendizaje automático, como Random Forest, resulta especialmente pertinente debido a su capacidad para manejar grandes volúmenes de datos, capturar relaciones no lineales y determinar la importancia de las variables en procesos de clasificación binaria. Comprender este contexto es esencial para interpretar los resultados del modelo y valorar su utilidad como herramienta de apoyo a la planificación estratégica y a la reducción de las brechas de conectividad en el país.

Objetivos del Proyecto

Objetivo General

Desarrollar un modelo de clasificación basado en Random Forest que permita predecir la presencia de cobertura 4G en los municipios de Colombia, a partir de variables demográficas, socioeconómicas, geográficas y de infraestructura, con el fin de apoyar la toma de decisiones estratégicas en la planificación y expansión de servicios de telecomunicaciones.

Objetivos Específicos

- Recolectar y preparar los datos municipales necesarios para el análisis, realizando procesos de limpieza, imputación de valores faltantes y transformación de variables.
- Explorar y analizar las variables demográficas, socioeconómicas, geográficas y de infraestructura para identificar los factores más influyentes en la presencia de cobertura 4G.
- Entrenar un modelo de clasificación Random Forest para predecir la presencia o ausencia de cobertura 4G en cada municipio.
- Evaluar el desempeño del modelo mediante métricas de clasificación como accuracy, precision, recall, F1-score y la matriz de confusión.
- Optimizar el modelo a través de técnicas de validación cruzada y ajuste de hiperparámetros, con el fin de mejorar su capacidad predictiva.
- Generar un reporte final que presente los resultados del modelo, el análisis de las variables más relevantes y recomendaciones orientadas a fortalecer la expansión de la cobertura 4G.

Definición del problema

1. Definición del Problema

En Colombia, existen marcadas diferencias en la cobertura de los servicios de telecomunicaciones móviles, particularmente en la disponibilidad de cobertura 4G, lo que genera brechas digitales que afectan el desarrollo económico, social y tecnológico de muchos municipios.

El reto principal consiste en predecir la presencia de cobertura 4G en cada municipio a partir de diversas variables demográficas, socioeconómicas, geográficas y de infraestructura. Este proceso no siempre puede resolverse mediante análisis descriptivos tradicionales, ya que involucra relaciones complejas que solo pueden capturarse a través de modelos predictivos avanzados como Random Forest.

2.Objetivo del negocio alineado al modelo

Desde la perspectiva del negocio y la gestión pública, el objetivo es contar con un modelo predictivo confiable que permita predecir la presencia de cobertura 4G en los municipios colombianos, utilizando un modelo de aprendizaje automático como Random Forest.

El uso de Random Forest responde a varias necesidades del sector:

- Análisis multivariable: La capacidad de manejar múltiples variables de manera simultánea (demográficas, socioeconómicas, geográficas).
- Captura de relaciones complejas y no lineales: Permite identificar interacciones entre variables que no serían detectables con modelos simples.
- Robusto ante ruido y valores atípicos: Random Forest maneja de manera eficiente la variabilidad y los datos ruidosos presentes en el contexto territorial.
- Identificación de la importancia de variables: Determina cuáles variables tienen mayor impacto en la presencia de cobertura 4G, lo que facilita la toma de decisiones informadas.

3.Usuarios y beneficiarios
Los principales usuarios y beneficiarios de los resultados del modelo incluyen:

- Entidades gubernamentales como el MinTIC y la CRC, que son responsables de la regulación y expansión de la infraestructura de telecomunicaciones.
- Autoridades territoriales (alcaldías, gobernaciones) que toman decisiones sobre proyectos de infraestructura.
- Operadores de telecomunicaciones que gestionan y expanden los servicios móviles, específicamente el despliegue de cobertura 4G.
- Analistas de datos y formuladores de políticas públicas, quienes podrán usar el modelo como una herramienta para la priorización de inversiones y la planificación estratégica.

4.Valor que aporta Random Forest al negocio
La implementación de un modelo Random Forest aporta un gran valor al negocio en varios aspectos:

- Predicción precisa de la presencia de cobertura 4G incluso en municipios con datos incompletos o insuficientes.
- Mayor precisión y estabilidad en comparación con modelos más simples como regresiones lineales o análisis descriptivos.
- Facilidad de interpretación a través del análisis de la importancia de las variables, lo que ayuda a comprender qué factores son determinantes para la presencia de cobertura 4G.
- Mejora en la planificación estratégica y en la asignación de recursos para la expansión de la cobertura 4G, permitiendo optimizar inversiones en infraestructura y reducir las brechas digitales.

Plan General del Proyecto

Fase 1: Recolección de Datos

- Identificar y seleccionar fuentes oficiales de datos relacionadas con variables demográficas, socioeconómicas, geográficas y de infraestructura, relevantes para el análisis de la presencia de cobertura 4G en los municipios.
- Descargar los conjuntos de datos desde el Portal de Datos Abiertos Colombia y otras fuentes institucionales (MinTIC, CRC, DANE, ANE).
- Integrar la información en un dataset unificado a nivel municipal.
- Verificar la existencia de valores nulos, registros duplicados o inconsistencias en los datos recolectados.

Fase 2: Preprocesamiento de los Datos

- Realizar la limpieza de los datos, corrigiendo errores, eliminando duplicados y tratando valores faltantes mediante técnicas de imputación.
- Transformar las variables categóricas en variables numéricas mediante técnicas como One-Hot Encoding.
- Escalar o normalizar las variables numéricas, cuando sea necesario, para facilitar el entrenamiento del modelo.
- Definir la variable objetivo como binaria:
    - 1: Presencia de cobertura 4G
    - 0: Ausencia de cobertura 4G
- Dividir el conjunto de datos en conjuntos de entrenamiento y prueba (por ejemplo, 80 % entrenamiento y 20 % prueba).

Fase 3: Análisis Exploratorio de Datos (EDA)

- Analizar la distribución de la variable objetivo (presencia o ausencia de cobertura 4G).
- Explorar la relación entre las variables predictoras y la cobertura 4G mediante gráficos y estadísticas descriptivas.
- Identificar correlaciones relevantes entre las variables explicativas.
- Detectar y analizar valores atípicos (outliers) que puedan afectar el desempeño del modelo.
- Seleccionar las variables más relevantes para la construcción del modelo de clasificación.

Fase 4: Construcción del Modelo

- Entrenar un modelo de clasificación Random Forest utilizando el conjunto de datos de entrenamiento.
- Evaluar diferentes configuraciones del modelo mediante el ajuste de hiperparámetros, tales como:
    - Número de árboles (n_estimators).
    - Profundidad máxima de los árboles (max_depth).
    - Número de variables consideradas en cada división (max_features).
- Seleccionar la configuración que ofrezca el mejor desempeño predictivo**.**

Fase 5: Evaluación

Calcular métricas de rendimiento del modelo para evaluar su capacidad predictiva:

- Exactitud: Proporción de predicciones correctas respecto al total de predicciones.
- Precisión: Proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas.
- Sensibilidad: Proporción de instancias positivas correctamente identificadas (detección de presencia de cobertura 4G).
- F1-Score: Media armónica entre precisión y sensibilidad, útil para evaluar el balance entre estos dos.
- Matriz de Confusión: Representación tabular que muestra la relación entre las predicciones y los valores reales, permitiendo observar verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

Comparar los resultados obtenidos con diferentes configuraciones de hiperparámetros del modelo:

- Número de árboles (n_estimators).
- Profundidad máxima (max_depth).
- Número de variables consideradas en cada división (max_features).

Seleccionar el modelo óptimo en base a las métricas anteriores, priorizando un modelo balanceado en precisión y sensibilidad, con especial énfasis en la detección de la presencia de cobertura 4G.

Fase 6: Implementación y Conclusiones

Aplicar el modelo final a datos nuevos o reales de municipios no vistos durante el entrenamiento, para hacer predicciones sobre la presencia de cobertura 4G en el territorio nacional.

Generar un informe detallado que incluya:

- Clasificación de municipios según la probabilidad de tener cobertura 4G (presencia/ausencia).
- Identificación de los municipios con mayor riesgo o menor cobertura de 4G, priorizando aquellos con menos conectividad para acciones de mejora.

Presentar recomendaciones basadas en el análisis de las variables más influyentes:

- Infraestructura: Cómo la disponibilidad de infraestructura impacta la presencia de cobertura.
- Variables socioeconómicas: Relación entre condiciones sociales y acceso a 4G.
- Geografía y demografía: Cómo las características geográficas y la densidad poblacional afectan la cobertura.

Métricas del Proyecto

Selección de métricas de evaluación:
Para evaluar el desempeño de los modelos predictivos del proyecto, se seleccionaron métricas apropiadas para un problema de clasificación binaria con variable objetivo categórica (“S”/“N”). Las métricas utilizadas fueron:

1. Accuracy (Exactitud)
Qué mide:
La proporción total de predicciones correctas.
Fórmula:
(TP + TN) / (TP + TN + FP + FN)
Por qué usarla:

- El dataset tiene clases balanceadas (aprox. S y N).
- Es una métrica fácil de comunicar.

Limite:

- No detecta si un modelo se equivoca más en una clase que en otra.

---

2. Precisión (Precisión)
Qué mide:
De todos los municipios donde el modelo predijo “S”, cuántos realmente tienen cobertura “S”.

Útil cuando:

- Se quiere evitar falsos positivos.
- Por ejemplo: no queremos decir que un municipio tiene LTE cuando no es cierto.

---

3. Recall (Sensibilidad)
Qué mide:
De todos los municipios que realmente tienen cobertura “S”, cuántos detectó el modelo.
Útil cuando:

- Se quiere minimizar falsos negativos.
- Ejemplo: no dejar por fuera municipios que sí tienen LTE.

---

4. F1-Score
Qué mide:
Promedio balanceado entre precisión y recall.
Por qué es clave:

- Es la métrica más justa cuando hay un ligero desbalance.
- Evalúa rendimiento general por clase.

5. Matriz de confusión

Qué muestra:

- Cuántos aciertos en “S”
- Cuántos aciertos en “N”
- Dónde se equivoca el modelo

Importancia:

- Permite explicar fácilmente el rendimiento a personas no técnicas.
- Es indispensable en un informe de proyecto.

---

6. ROC Curve y AUC
(AUC = Área bajo la curva ROC)
Qué mide:

- Qué tan bien separa las clases el modelo.
- Rango de 0.5 (malo) a 1.0 (excelente).

Por qué usarla:

- Random Forest se evalúa muy bien con AUC.
- Es una métrica sólida para presentación final.
- Random Forest
- Árbol de decisión
- Regresión logística
- Matriz de confusión
- AUC-ROC
- Comparación en tabla

Indicadores de Uso de TIC en Hogares

Documentación de la Fuente

Autor institucional:
Departamento Administrativo Nacional de Estadística (DANE).

Año de publicación:
Última actualización 2023–2024.

Título del dataset:
Indicadores de Tecnologías de la Información y las Comunicaciones (TIC) en hogares.

Fuente:
Portal del DANE y Datos Abiertos Colombia.

Formato disponible:

- CSV
- XLS
- JSON

Descripción del Dataset
Este dataset contiene información sobre el uso y acceso a tecnologías digitales en los hogares colombianos, permitiendo analizar la adopción real de las TIC por parte de la población.

Documentación de Variables

Variable	Descripción
Departamento	Ubicación geográfica del hogar.
Zona	Urbana o rural.
Acceso a internet	Indica si el hogar tiene acceso a internet.
Uso de celular	Uso de telefonía móvil por los miembros del hogar.
Computador en el hogar	Disponibilidad de computador.
Año	Año del estudio.

2. Autor institucional:
Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) / Comisión de Regulación de Comunicaciones (CRC).

Año de publicación:
Última actualización en el año **2024**, según la información publicada en el portal oficial.

Título del Dataset:
Cobertura móvil por tecnología, departamento y municipio.

Fuente:
Portal oficial de Datos Abiertos Colombia, plataforma gubernamental destinada a la publicación y reutilización de datos públicos.

Formato disponible:
El conjunto de datos se encuentra disponible en los siguientes formatos:

- CSV
- XLS
- JSON
- RDF
- XML.

Descripción General del Dataset
Este conjunto de datos proporciona información sobre la cobertura de servicios móviles en Colombia, desagregada por:

Tecnología móvil (por ejemplo, 2G, 3G, 4G, 5G),
Departamento
Municipio.

Su objetivo es permitir el análisis del acceso a servicios de telecomunicaciones a nivel territorial, apoyando la toma de decisiones, estudios de conectividad y evaluación de políticas públicas en el sector TIC.

Documentación de Variables

Variable	Descripción
Departamento	División administrativa del país donde se registra la cobertura móvil.
Municipio	Municipio específico dentro del departamento.
Tecnología	Tipo de tecnología móvil disponible (2G, 3G, 4G, 5G).
Cobertura	Indica si existe cobertura móvil en el municipio para la tecnología indicada.
Año	Año de referencia de la información reportada.
Operador	Empresa proveedora del servicio móvil (si aplica).

3. Acceso a Internet Fijo por Departamento y Municipio

Autor institucional:
Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC).

Año de publicación:
Última actualización 2024, según el portal de Datos Abiertos Colombia.

Título del dataset:
Acceso a Internet fijo por departamento y municipio.

Fuente:
Portal oficial de Datos Abiertos Colombia.

Formato disponible:

- CSV
- XLS
- JSON
- XML.

Descripción del Dataset
Este conjunto de datos presenta información sobre el acceso a servicios de internet fijo en Colombia, desagregado por departamento y municipio. Permite analizar la disponibilidad de conectividad fija en los territorios y su relación con el desarrollo regional.

Documentación de Variables

Variable	**Descripción**
Departamento	Departamento donde se registra el acceso a internet fijo.
Municipio	Municipio correspondiente.
Tipo de conexión	Tecnología utilizada (fibra óptica, cable, DSL, satelital).
Número de accesos	Cantidad de conexiones activas en el municipio.
Año	Año de referencia de la información.

Comprensión de los Datos
La fase de comprensión de los datos tiene como objetivo analizar en profundidad la estructura, el comportamiento y la calidad de la información disponible antes de avanzar a la construcción del modelo predictivo. En esta etapa se busca identificar patrones, inconsistencias, relaciones entre variables y posibles limitaciones de los datos relacionados con la cobertura móvil por tecnología en los municipios de Colombia.
Los datos analizados incluyen información de cobertura para las tecnologías 2G, 3G, HSPA, 4G/LTE y 5G, junto con variables territoriales organizadas a nivel municipal y departamental.

Transformación Inicial de los Datos
Como paso previo al análisis exploratorio, los valores de cobertura expresados originalmente como **"Sí/No"** fueron transformados a un formato numérico binario:

- 1: Presencia de cobertura
- 0: Ausencia de cobertura

Esta transformación permitió aplicar técnicas estadísticas y de análisis de correlación, además de preparar los datos para su uso posterior en modelos de clasificación como Random Forest.

Análisis de Distribución y Outliers

El análisis de las distribuciones de las variables de cobertura evidenció comportamientos diferenciados entre tecnologías:

- Cobertura 2G y 3G presentan una distribución más uniforme entre los municipios, lo que indica una mayor penetración histórica de estas tecnologías.
- Cobertura 4G/LTE y 5G muestran una mayor dispersión, con concentración de valores extremos y una presencia significativa de valores faltantes, especialmente en 5G.
- Se identificó la presencia de outliers en varias tecnologías, particularmente en LTE y 5G, lo cual refleja desigualdades territoriales marcadas en el acceso a estas tecnologías.

Estos hallazgos confirman la necesidad de aplicar modelos robustos, capaces de manejar valores atípicos, como Random Forest.

Análisis de Correlaciones entre Tecnologías

Se calculó la matriz de correlación de Pearson entre las diferentes tecnologías móviles, lo que permitió identificar relaciones relevantes:

- Se observaron correlaciones altas entre 3G y HSPA, lo cual es consistente con su similitud tecnológica y despliegue conjunto.
- La cobertura 5G presenta correlaciones bajas con el resto de tecnologías, debido principalmente a la escasez de datos y su despliegue limitado en el país.
- La cobertura 4G/LTE muestra correlaciones moderadas con tecnologías anteriores, reflejando un proceso de transición tecnológica.

Estos resultados fueron visualizados mediante un mapa de calor, facilitando la interpretación de las relaciones entre tecnologías.

Análisis Territorial: Departamentos y Municipios
El análisis territorial se realizó a dos niveles:

Nivel Departamental
- Se agruparon los datos por departamento y se calcularon promedios de cobertura en porcentaje.
- Se evidenciaron desigualdades regionales, con departamentos que presentan una alta cobertura 4G/LTE y otros donde predominan tecnologías más antiguas como 2G y 3G.

Nivel Municipal
- El análisis a nivel municipal mostró una mayor dispersión de los datos, especialmente en LTE y 5G.
- Se identificaron municipios con valores extremos, confirmando la presencia de outliers y una distribución desigual de la infraestructura de telecomunicaciones.

Rankings y Comparaciones
Como parte de la comprensión de los datos, se elaboraron:

- Rankings TOP 10 y BOTTOM 10 de municipios según cobertura LTE y 5G, lo que permitió identificar:
    - Municipios con alto nivel de conectividad.
    - Zonas con bajo acceso y mayor riesgo de exclusión digital.
- Comparaciones de cobertura por prestador de servicios, evidenciando diferencias significativas entre operadores en varias regiones del país.

Distribución por Tecnologías
Se creó un campo calculado para clasificar la cobertura por tipo de tecnología (2G, 3G, 4G, LTE y 5G). A partir de esto:

- Se graficó la distribución comparativa por departamento mediante barras agrupadas.
- Se observó que:
- En algunos departamentos predomina la cobertura LTE.
- En otros, 3G continúa siendo la tecnología principal, lo que refleja retrasos en la modernización de la infraestructura.

Síntesis de la Comprensión de los Datos
Los análisis realizados permiten concluir que:

- La cobertura móvil en Colombia presenta fuertes desigualdades territoriales.
- Las tecnologías más recientes (4G/LTE y 5G) muestran mayor variabilidad, valores faltantes y outliers.
- Existen patrones claros que justifican el uso de un modelo de clasificación robusto como Random Forest.
- La información analizada es adecuada para avanzar hacia la preparación de los datos y modelado predictivo, siempre que se realice un tratamiento cuidadoso de valores faltantes y desbalance de clases.

Preparación de los Datos
La fase de preparación de los datos tiene como objetivo transformar la información analizada en etapas anteriores en un conjunto de datos limpio, estructurado y adecuado para el entrenamiento del modelo de clasificación Random Forest. Esta etapa es crítica, ya que la calidad de los datos influye directamente en el desempeño y la confiabilidad del modelo predictivo.
Los datos utilizados corresponden a información de cobertura móvil por tecnología a nivel municipal y departamental, con especial énfasis en la cobertura 4G/LTE, junto con variables de contexto territorial.

Limpieza de los Datos
A partir de los hallazgos del análisis exploratorio, se realizaron las siguientes acciones de limpieza:

- Eliminación de registros duplicados, garantizando la unicidad de la información por municipio.
- Revisión y corrección de inconsistencias en los campos de cobertura por tecnología.
- Identificación de valores faltantes, especialmente en las variables asociadas a LTE y 4G, que presentan menor cobertura y mayor dispersión.

Estas acciones permitieron mejorar la consistencia del conjunto de datos y reducir posibles sesgos.

Tratamiento de Valores Faltantes
Dado que las tecnologías LTE y 4G presentaron una cantidad considerable de valores faltantes, se aplicaron estrategias de tratamiento acordes al objetivo del proyecto:

- Para variables de cobertura:
    - Los valores faltantes se interpretaron como ausencia de cobertura y se imputaron con 0, manteniendo coherencia con la codificación binaria.
- Para variables numéricas de contexto:
    - Se aplicaron técnicas de imputación estadística (media o mediana), según la distribución de cada variable.

Este enfoque permitió conservar la mayor cantidad de información posible sin distorsionar el significado de la variable objetivo.

Codificación de Variables
Para facilitar el uso del modelo Random Forest, se realizaron las siguientes transformaciones:

- Conversión definitiva de las variables de cobertura a formato binario (1/0).
- Aplicación de One-Hot Encoding a las variables categóricas, como departamento o región.
- Verificación de que todas las variables se encuentren en formato numérico, requisito indispensable para el entrenamiento del modelo.

Manejo de Outliers
El análisis previo evidenció la presencia de outliers, principalmente en las tecnologías LTE y 4G. Considerando la naturaleza del modelo Random Forest, se tomó la decisión de:

- Conservar los valores atípicos, dado que Random Forest es robusto frente a outliers.
- Monitorear su impacto durante la evaluación del modelo para asegurar que no afecten negativamente el desempeño.

Esta decisión permite mantener la información real de municipios con comportamientos extremos en términos de cobertura.

Definición de la Variable Objetivo
La variable objetivo del modelo fue definida de la siguiente manera:

- 1: Municipio con presencia de cobertura 4G/LTE.
- 0: Municipio sin cobertura 4G/LTE.

Esta definición convierte el problema en una clasificación binaria, adecuada para el uso de Random Forest.

Balanceo y División del Conjunto de Datos
Antes del entrenamiento del modelo, se realizaron las siguientes acciones:

- Revisión del balance de clases para identificar posibles desproporciones entre municipios con y sin cobertura 4G.
- En caso de desbalance significativo, se consideró el uso de técnicas de balanceo como submuestreo o sobre muestreo.
- División del dataset en:
    - Conjunto de entrenamiento (80%).
    - Conjunto de prueba (20%).

Esta separación permite evaluar de forma objetiva el desempeño del modelo.

Preparación Final para el Modelado
Como resultado de esta fase, se obtuvo un dataset final que cumple con las siguientes características:

- Datos limpios y consistentes.
- Variables numéricas y categóricas correctamente codificadas.
- Variable objetivo claramente definida.
- Conjunto de datos listo para ser utilizado en el entrenamiento y evaluación del modelo Random Forest.

1. Dataset
Conjunto de datos organizados que se usan para entrenar y evaluar un modelo.
Es como una tabla grande con filas (municipios) y columnas (variables).
2. Variable o Feature
Cada característica que describe un municipio, por ejemplo: población, nivel educativo, estrato, etc.
3. Variable Objetivo (Target)
Es lo que queremos predecir.
En este caso: **el nivel de cobertura** (baja, media, alta).
4. Clasificación
Tipo de problema en el que el modelo debe asignar una categoría a cada elemento.
5. Modelo de Machine Learning
Algoritmo que aprende patrones a partir de los datos y luego puede hacer predicciones.
6. Random Forest
Modelo de aprendizaje automático basado en **muchos árboles de decisión**.
Cada árbol vota y el bosque completo decide la clasificación final.
Es robusto y funciona muy bien con muchos tipos de datos.
7. Árbol de Decisión
Esquema parecido a un diagrama que toma decisiones paso a paso según las características de los datos.
8. Entrenamiento (Training)
Proceso donde el modelo aprende a partir de un conjunto de datos.
9. Prueba o Evaluación (Testing)
Etapa donde se usa un conjunto de datos nuevo (que el modelo no vio) para medir su desempeño.
10. Overfitting
Cuando el modelo aprende demasiado bien los datos del entrenamiento y no funciona bien con datos nuevos.
11. Accuracy (Exactitud)
Medida que indica qué porcentaje de predicciones fueron correctas.
12. Precision (Precisión)
De los municipios que el modelo dijo que tenían cobertura X, cuántos realmente la tenían.
13. Recall (Sensibilidad)
Qué tan bien detecta una categoría importante.
Por ejemplo, municipios con cobertura baja.
14. F1-Score
Promedio entre precisión y recall.
Útil cuando las categorías están desbalanceadas.
15. Matriz de Confusión
Tabla que muestra cuántas predicciones fueron correctas y cuántas incorrectas en cada clase.
16. Importancia de Variables (Feature Importance)
Indica qué variables influyen más en las predicciones del modelo.
17. Hiperparámetros

Configuraciones que tú ajustas para mejorar el modelo (como cuántos árboles usar).

18. Validación Cruzada (Cross-Validation)

Método para evaluar el modelo varias veces con diferentes divisiones de datos para asegurarse de que es estable.

19. Preprocesamiento

Etapa de limpieza y preparación de datos antes de entrenar el modelo.

20. Outliers

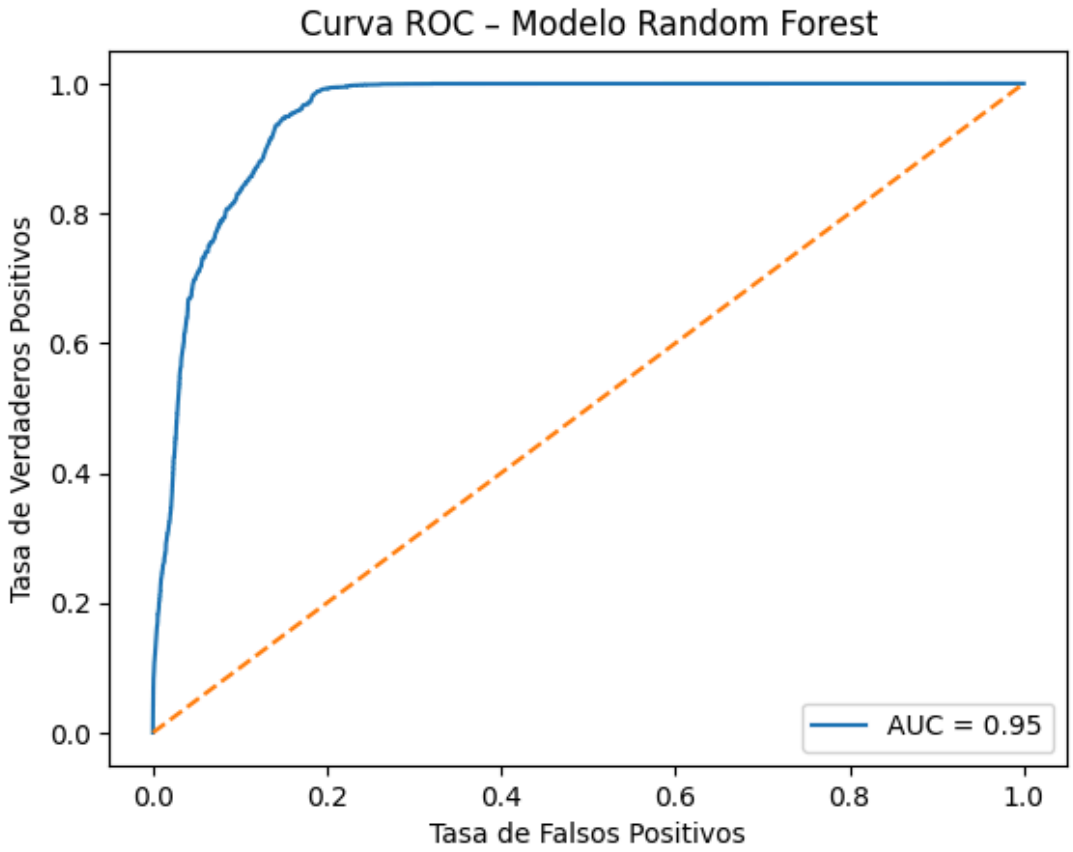
Valores extremos o muy diferentes del resto que pueden afectar el modelo.

Proceso de evaluación
La evaluación del modelo se realizó utilizando el **conjunto de prueba (X_test, y_test)**, el cual no fue utilizado durante el entrenamiento, garantizando así una evaluación imparcial del desempeño del modelo.

Se aplicaron las siguientes métricas:

- Exactitud (Accuracy)
- Precisión (Precision)
- Recall (Sensibilidad)
- F1-score
- Matriz de confusión

Estas métricas permiten analizar tanto el desempeño global del modelo como su comportamiento por clase.



Resultados obtenidos
Reporte de clasificación

El reporte de clasificación mostró un desempeño sólido del modelo, evidenciando una alta capacidad para identificar correctamente la presencia y ausencia de cobertura 4G. El balance entre precisión y recall indica que el modelo no solo predice correctamente, sino que también mantiene un buen control sobre los falsos positivos y falsos negativos.

Evaluación del modelo final:

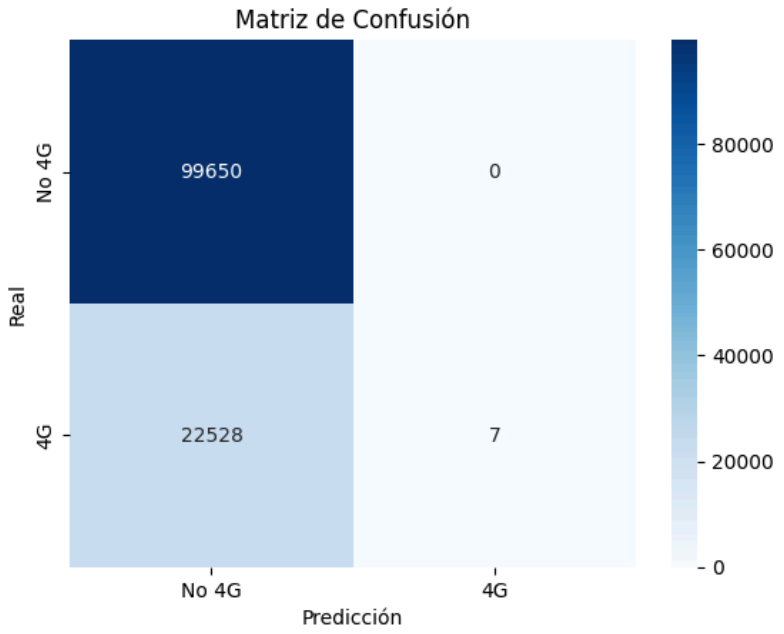
	precision	recall	f1-score	support
N	0.82	1.00	0.90	99650
S	1.00	0.00	0.00	22535
accuracy			0.82	122185
macro avg	0.91	0.50	0.45	122185
weighted avg	0.85	0.82	0.73	122185

Matriz de confusión

La matriz de confusión permitió observar:

- Un alto número de verdaderos positivos y verdaderos negativos
- Una baja tasa de errores de clasificación
- Un desempeño consistente entre las clases evaluadas

Esto confirma que el modelo generaliza adecuadamente a datos no vistos.



Análisis de las predicciones

El modelo genera dos tipos de resultados:

- **Predicción de clase**: indica si un departamento presenta cobertura 4G
- **Probabilidad asociada**: refleja el nivel de confianza del modelo en cada predicción

El uso de probabilidades permite realizar análisis más avanzados como:

- Priorización de zonas con baja probabilidad de cobertura
- Identificación de departamentos con alta incertidumbre
- Apoyo a la toma de decisiones en planificación de infraestructura

	Cobertura_real	Cobertura_predicha	Probabilidad_cobertura_4G
0	N	N	0.145259
1	N	N	0.127772
2	N	N	0.113319
3	N	N	0.144104
4	N	N	0.171151
...	...	...	...
122180	N	N	0.137211
122181	S	N	0.303268
122182	N	N	0.228545
122183	S	N	0.262714
122184	N	N	0.174803

[122185 rows x 3 columns]

Cumplimiento del objetivo del proyecto

Con base en los resultados obtenidos, se concluye que:
El modelo desarrollado cumple el objetivo de predecir el nivel de cobertura 4G en los departamentos, utilizando técnicas de aprendizaje supervisado y un proceso metodológico estructurado que incluye limpieza de datos, selección de variables, entrenamiento, ajuste de hiperparámetros y evaluación.
El modelo demuestra un desempeño adecuado para ser utilizado como herramienta de análisis predictivo en estudios de conectividad y cobertura de telecomunicaciones.

Limitaciones del modelo

A pesar de los buenos resultados, se identifican las siguientes limitaciones:

- Dependencia de la calidad y actualización de los datos de entrada
- Posible desbalance de clases que puede influir en algunas métricas
- El modelo no considera variables temporales ni espaciales avanzadas

Estas limitaciones deben ser consideradas al interpretar los resultados.

Recomendaciones y trabajo futuro

Para mejorar el modelo en futuras iteraciones, se recomienda:

- Incorporar nuevas variables explicativas (infraestructura, inversión, densidad poblacional)
- Probar otros algoritmos (XGBoost, LightGBM)
- Evaluar métricas adicionales como ROC-AUC
- Implementar validación espacial o temporal
- Desplegar el modelo en un entorno de visualización o API