Uso de Machine Learning para el análisis y clasificación de quejas y reclamos en el sistema de salud en Bucaramanga

Oscar Santiago Villamizar Hernández

Universidad Ean

Especialización en Machine Learning

Bogotá, Colombia

2026

Resumen

Este trabajo analiza y clasifica las quejas y reclamos presentados por usuarios del sistema de salud en Colombia, con enfoque en la ciudad de Bucaramanga, aplicando técnicas de Machine Learning. Estas PQRDS constituyen una fuente valiosa de información sobre las dificultades que enfrentan los ciudadanos en la prestación de los servicios de salud y, por eso, analizarlas de forma sistemática puede aportar elementos útiles para mejorar su atención.

Para desarrollar el análisis se utilizó un conjunto de datos con los reclamos registrados ante la Superintendencia Nacional de Salud durante el primer semestre de 2025, filtrando únicamente los correspondientes a Bucaramanga.

Con este trabajo se busca mostrar que el Machine Learning tiene un potencial real como herramienta de apoyo para el análisis de grandes volúmenes de datos en el sector salud, y que sus resultados pueden orientar la toma de decisiones hacia una mejora concreta en la calidad del servicio.

Abstract

This work analyzes and classifies complaints and claims submitted by users of the Colombian healthcare system, focusing on the city of Bucaramanga, through the application of Machine Learning techniques. These records represent a valuable source of information about the difficulties citizens face in accessing health services and, for this reason, analyzing them systematically can provide useful insights for improving service quality.

The analysis was developed using a dataset containing complaints registered before the Superintendencia Nacional de Salud during the first semester of 2025, filtering only those corresponding to Bucaramanga.

This study aims to demonstrate that Machine Learning holds real potential as a support tool for analyzing large volumes of data in the health sector, and that its findings can guide decision-making toward concrete improvements in service quality.

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Introducción

En los últimos años, el análisis de grandes volúmenes de datos ha ganado una relevancia creciente en distintos sectores, y el ámbito de la salud no es la excepción. La información generada por usuarios e instituciones representa una fuente clave para identificar problemáticas y promover mejoras continuas en la prestación de los servicios. En este escenario, las técnicas de Machine Learning han demostrado ser especialmente útiles para detectar patrones y tendencias en conjuntos de datos de gran escala.

En Colombia, las quejas y reclamos dentro del sistema de salud han mostrado un crecimiento sostenido en los últimos años. Según reportes de la Superintendencia Nacional de Salud, este aumento está asociado a una crisis estructural del sistema, que se manifiesta en situaciones como demoras en la atención, barreras de acceso, problemas administrativos y deficiencias en la calidad del servicio. Si bien estas PQRDS son una fuente valiosa de información sobre las fallas del sistema, el alto volumen de registros que se generan dificulta identificar oportunamente las problemáticas más recurrentes y tomar medidas correctivas a tiempo.

Frente a este panorama, surge la necesidad de contar con herramientas que permitan procesar y analizar esta información de manera eficiente. El Machine Learning se presenta como una alternativa viable para abordar este reto, dado que permite clasificar y analizar datos de forma automatizada, reduciendo los tiempos y limitaciones del análisis manual.

En este contexto, el propósito de esta investigación es desarrollar un modelo de Machine Learning que permita clasificar las PQRDS del sistema de salud en Bucaramanga, utilizando los datos del primer semestre de 2025 publicados por la Superintendencia Nacional de Salud. Para ello se aplicaron técnicas de preprocesamiento y análisis exploratorio de datos, junto con métodos de clasificación supervisada que incluyen algoritmos como Regresión Logística, Árbol de Decisión y Random Forest. A través de este enfoque se busca identificar tendencias y patrones en la información, aportando herramientas que faciliten su análisis y contribuyan a la mejora en la calidad del servicio de salud.

Objetivos

Objetivo general

Desarrollar un modelo de Machine Learning que permita clasificar las PQRDS del sistema de salud en Bucaramanga utilizando datos del primer semestre de 2025, identificando patrones relevantes que aporten información útil para la toma de decisiones y la mejora de la calidad del servicio.

Objetivos específicos

- Realizar el preprocesamiento, limpieza y depuración del dataset, incluyendo la selección de variables relevantes y el tratamiento de valores nulos, duplicados e inconsistentes.

- Implementar y comparar modelos de clasificación supervisada —Regresión Logística, Árbol de Decisión y Random Forest— para categorizar las PQRDS según su nivel de riesgo, evaluando su desempeño mediante métricas como accuracy, precisión, recall y F1-score.

- Analizar la importancia de las variables del modelo para identificar los factores con mayor influencia en la clasificación de riesgo de las PQRDS en el sistema de salud de Bucaramanga.

Justificación

El análisis de las PQRDS en el sistema de salud colombiano es una herramienta clave para entender las problemáticas que afectan a los usuarios y para evaluar la calidad de los servicios que reciben. Sin embargo, el crecimiento sostenido en el volumen de estas quejas y reclamos ha hecho cada vez más difícil gestionarlas de forma eficiente, limitando la capacidad de las instituciones para responder oportunamente y tomar medidas correctivas.

Frente a este reto, el aprendizaje automático surge como una alternativa concreta para procesar y analizar grandes cantidades de información de manera automatizada, identificando patrones que de otra forma serían difíciles de detectar y generando insumos útiles para la toma de decisiones basadas en evidencia. Su aplicación en el sector salud permite, además, identificar con mayor rapidez las problemáticas recurrentes que afectan el funcionamiento del sistema.

En ese sentido, esta investigación busca proponer una herramienta de apoyo para la gestión de PQRDS en Bucaramanga, que facilite su clasificación y análisis, y que pueda contribuir a la mejora de la calidad del servicio de salud a nivel local.

Marco Teórico

El crecimiento exponencial de la información en distintos sectores ha impulsado el surgimiento del Big Data, término que hace referencia al manejo de grandes volúmenes de datos que superan la capacidad del procesamiento manual o tradicional. En el sector salud, esta información proviene de fuentes muy diversas: historias clínicas, sistemas administrativos, registros de interacción con usuarios y, entre ellos, las quejas y reclamos. Analizar estos datos a gran escala permite identificar patrones, tendencias y relaciones que de otra forma pasarían desapercibidas, aportando valor a la toma de decisiones y a la optimización de recursos en el sistema.

En este escenario, el aprendizaje automático se presenta como una rama de la inteligencia artificial especialmente útil para extraer ese valor de los datos. A través de modelos como la Regresión Logística, el Árbol de Decisión y el Random Forest, es posible categorizar información de manera eficiente y comprensible, facilitando la identificación de patrones y la realización de predicciones a partir de los datos disponibles.

En Colombia, el sistema de salud atraviesa una crisis que se refleja, entre otras cosas, en el aumento sostenido de quejas y reclamos por parte de los usuarios. Frente a esto, existen mecanismos de participación ciudadana que permiten canalizar estas inconformidades, siendo las PQRDS —Peticiones, Quejas, Reclamos, Denuncias y Sugerencias— el principal instrumento gestionado por entidades como la Superintendencia Nacional de Salud.

Estos registros representan una fuente valiosa para evaluar la calidad del servicio y detectar fallas en el sistema. Sin embargo, su volumen creciente hace cada vez más difícil analizarlos de forma oportuna y eficiente, lo que justifica el uso de herramientas de aprendizaje automático para su procesamiento. En este sentido, aplicar estas técnicas al análisis de las PQRDS en Bucaramanga representa una oportunidad concreta para fortalecer el uso de la analítica de datos en un contexto local y contribuir a la mejora de la calidad del servicio y la experiencia de los usuarios.

Metodología

El análisis se desarrolla con un enfoque cuantitativo debido a que se utilizan datos numéricos y categóricos que fueron obtenidos a partir de un conjunto de registros de reclamos en salud interpuestos por los ciudadanos ante la Superintendencia Nacional de Salud en el primer semestre del año 2025. Este enfoque nos permitió realizar el procesamiento, análisis y clasificación de los datos mediante técnicas estadísticas y algoritmos de Machine Learning.

Este estudio tiene un alcance descriptivo y aplicado. Descriptivo porque busca identificar patrones y características presentes en los reclamos del sistema de salud específicamente en la ciudad de Bucaramanga, aplicado porque se busca implementar un modelo de clasificación que permita apoyar el análisis automatizado de la información.

La metodología se justifica ante la necesidad de optimizar el procesamiento de volúmenes grandes de datos relacionados con los reclamos de los usuarios, ya que al analizar esos registros de manera manual se presentan limitaciones en tiempo y capacidad de respuesta. Teniendo esto en cuenta, el uso de algunas técnicas de aprendizaje automático nos permite desarrollar herramientas enfocadas a la identificación de tendencias y problemáticas que se presenten de manera recurrente en el sistema de salud.

Definición de variables

Para el desarrollo de esta investigación se consideraron dos tipos de variables: técnicas y contextuales, todas extraídas de los registros de PQRDS del sistema de salud en la ciudad de Bucaramanga.

Variable	Tipo	Definición conceptual	Definición operacional
Mes	Independiente	Periodo de tiempo en el que se registra una PQRDS dentro del sistema de salud.	Se identifica mediante el mes registrado en la base de datos de la Superintendencia Nacional de Salud.
Canal de atención	Independiente	Medio utilizado por el usuario para presentar una petición, queja o reclamo.	Se mide según la categoría registrada en la variable `pqr_canal` del dataset.
Género del usuario	Independiente	Característica demográfica relacionada con el sexo del usuario afectado.	Se clasifica según los valores registrados en la variable `afec_genero`.
Edad	Independiente	Número de años cumplidos por el usuario afectado al momento del registro de la PQRDS.	Se mide de forma numérica mediante la variable `afec_edad`.
Régimen de afiliación	Independiente	Tipo de vinculación del usuario al sistema de salud colombiano.	Se identifica mediante la variable `afec_regafiliacion` registrada en el dataset.
Entidad de salud	Independiente	Institución o EPS relacionada con la atención del usuario y el registro de la PQRDS.	Se mide utilizando la variable `ent_nombre`.
Macromotivo	Independiente	Categoría amplia que agrupa las causas principales de las PQRDS.	Se identifica mediante la variable `macromotivo`.
Motivo general	Independiente	Clasificación general del motivo asociado al reclamo o petición.	Se registra en la variable `motivo_general` del dataset.
Motivo específico	Independiente	Descripción puntual de la causa que originó la PQRDS.	Se mide mediante la variable `motivo_especifico`.
Clasificación de riesgo	Dependiente	Nivel de riesgo asignado a la PQRDS según su impacto o prioridad dentro del sistema de salud.	Se utiliza como variable objetivo del modelo de Machine Learning y corresponde a la variable `clasificacion_de_riesgo`.

Tabla 1. Fuente: Elaboración propia.

Población y muestra

La población de este estudio está conformada por todos los registros de Peticiones, Quejas, Reclamos, Denuncias y Sugerencias (PQRDS) que fueron reportados ante la Superintendencia Nacional de Salud durante el primer semestre de 2025.

Para definir la muestra, se filtraron únicamente los registros correspondientes a la ciudad de Bucaramanga, extraídos del conjunto de datos abiertos usado en esta investigación. Una vez obtenidos los datos, se llevó a cabo un proceso de depuración en el que se retiraron registros incompletos, valores nulos y duplicados, buscando asegurar que la información utilizada tanto en el análisis como en el entrenamiento de los modelos fuera lo más confiable posible.

La muestra se seleccionó mediante un muestreo no probabilístico por criterio, ya que el interés estaba puesto específicamente en los registros del territorio definido para el estudio, no en una selección aleatoria del total de datos disponibles.

Técnicas y Recolección de Datos

Esta investigación se desarrolló bajo un enfoque cuantitativo, trabajando con información secundaria obtenida de bases de datos abiertas del sector salud. Específicamente, el dataset fue descargado desde el portal de datos abiertos de la Superintendencia Nacional de Salud (datos.gov.co) en formato CSV, y se filtró la variable de ubicación geográfica para conservar únicamente los registros correspondientes a Bucaramanga durante el primer semestre de 2025.

Para el desarrollo del estudio se siguió una secuencia de pasos que va desde la obtención de los datos hasta la evaluación de los modelos:

- Recolección del dataset abierto publicado por la Superintendencia Nacional de Salud.
- Filtrado y depuración de los registros, eliminando valores nulos, duplicados e información incompleta.
- Análisis exploratorio de los datos para entender su distribución y comportamiento.

- Entrenamiento de modelos supervisados de Machine Learning para la clasificación de riesgo.
- Evaluación estadística de los resultados obtenidos mediante métricas de desempeño.

Para llevar a cabo esta investigación se recurrió a instrumentos tecnológicos y bases de datos abiertas que hicieron posible recolectar, procesar y analizar la información de las PQRDS del sistema de salud en Bucaramanga.

Instrumento	Tipo	Descripción técnica	Frecuencia / precisión
Base de datos de PQRDS	Registro documental	Dataset descargado desde el portal de datos abiertos de la Superintendencia Nacional de Salud, con registros de peticiones, quejas, reclamos, denuncias y sugerencias.	Información del primer semestre de 2025.
Google Colab	Herramienta tecnológica	Plataforma en la nube usada para ejecutar el procesamiento de datos y los modelos de Machine Learning en Python.	Procesamiento en tiempo real durante el análisis.
Python	Lenguaje de programación	Lenguaje principal del proyecto, empleado en la limpieza, exploración, entrenamiento y evaluación de los modelos.	Permite automatizar el procesamiento de grandes volúmenes de datos.
Pandas y NumPy	Librerías de procesamiento	Herramientas para manipular, filtrar y transformar los datos estructurados del dataset.	Alta precisión en operaciones de análisis de datos.
Matplotlib y Seaborn	Librerías de visualización	Usadas para construir los gráficos estadísticos y visualizaciones del análisis exploratorio.	Generación dinámica de visualizaciones durante el análisis.
Scikit-learn	Librería de Machine Learning	Empleada para implementar los modelos de clasificación y calcular las métricas de evaluación.	Permite evaluar precisión, recall, F1-score y accuracy.

Tabla 2. Fuente: Elaboración propia.

Procedimiento

El desarrollo de esta investigación siguió un proceso estructurado que abarcó desde la obtención de los datos hasta la evaluación de los modelos, trabajando exclusivamente con registros de PQRDS del sistema de salud en Bucaramanga.

Los datos utilizados provienen del portal de datos abiertos de la Superintendencia Nacional de Salud, correspondientes al primer semestre de 2025. El dataset fue descargado en formato CSV y cargado en Google Colab para iniciar su procesamiento.

Paso a paso del proceso

1. Descarga y carga del dataset: Se realizó la descarga de la base de datos oficial de PQRDS desde el portal de la Superintendencia Nacional de Salud y se cargó en Google Colab usando Python y la librería Pandas.

2. Filtrado de información: Se conservaron únicamente los registros de Bucaramanga para delimitar el alcance geográfico del estudio.

3. Limpieza y depuración de datos: Se identificaron y eliminaron registros duplicados, valores nulos y variables que no aportaban al análisis. En este paso también se seleccionaron las variables más relevantes para el entrenamiento de los modelos.

4. Transformación de variables: Las variables categóricas fueron convertidas a formato numérico mediante técnicas de codificación, lo que permitió que los algoritmos de clasificación pudieran procesarlas correctamente.

5. Análisis exploratorio de datos: Se generaron visualizaciones y análisis descriptivos para explorar la distribución de las variables, detectar patrones y entender las relaciones entre los datos antes de modelar.

6. Entrenamiento de modelos: Se implementaron modelos supervisados de Machine Learning, específicamente Árbol de Decisión y Random Forest. Los datos se dividieron en conjuntos de entrenamiento y prueba para medir el desempeño de cada modelo de forma objetiva.

7. Evaluación de resultados: Los modelos se evaluaron usando métricas de clasificación como accuracy, precisión, recall, F1-score y matriz de confusión, lo que permitió analizar qué tan bien estaba clasificando cada uno.

Participantes

Esta investigación fue desarrollada de forma individual por el autor, apoyándose en herramientas tecnológicas y datos abiertos. No hubo intervención directa con pacientes ni recolección presencial de información.

Controles y restricciones técnicas

Durante el proyecto se tuvieron en cuenta algunas limitaciones que condicionaron el desarrollo del trabajo:

- El análisis se restringió a los registros de Bucaramanga.
- La calidad del estudio dependió directamente de la estructura y completitud de los datos abiertos disponibles.
- Algunas categorías presentaban datos incompletos o desbalanceados.

- El procesamiento en Google Colab implicó ciertas limitaciones computacionales propias de un entorno en la nube.

Protocolos de seguridad

Dado que la investigación trabajó con información pública y anonimizada, no se manipularon datos personales sensibles ni información que pudiera identificar directamente a los usuarios. Todo el procesamiento se realizó con fines estrictamente académicos, garantizando un uso responsable de la información y el cumplimiento de los principios éticos relacionados con el tratamiento de datos.

Validación de Instrumentos

Como la investigación se apoyó en datos abiertos y herramientas computacionales, la validación estuvo enfocada en dos aspectos clave: asegurar la calidad de los datos y verificar que los modelos funcionaran correctamente.

Validación del dataset

El primer paso fue hacer una revisión exploratoria del dataset descargado desde la Superintendencia Nacional de Salud. En esta revisión se buscaron valores nulos, registros duplicados, inconsistencias y variables que no aportaban al análisis. Con base en eso, se aplicaron procesos de limpieza y depuración que mejoraron la calidad de la información antes de usarla en el modelado.

Prueba piloto del procesamiento

Antes de entrenar los modelos de forma definitiva, se hicieron pruebas preliminares con una parte del dataset para verificar que todo funcionara bien. Estas pruebas permitieron confirmar

que los datos se estaban leyendo correctamente, que la transformación de variables categóricas era válida y que la información era compatible con los algoritmos de Machine Learning que se iban a usar.

Ajustes realizados

A lo largo del proyecto fue necesario hacer varios ajustes técnicos: eliminar registros incompletos, seleccionar las variables más relevantes, convertir las variables categóricas a formato numérico y filtrar los datos para quedarse únicamente con los registros de Bucaramanga. También se hicieron pruebas comparativas entre distintos modelos de clasificación para identificar cuál tenía mejor desempeño frente al problema planteado.

Validación del modelo

Los modelos fueron evaluados usando métricas de clasificación como accuracy, precisión, recall, F1-score y matriz de confusión. Adicionalmente, se aplicó validación cruzada para verificar que los resultados fueran estables y consistentes, y no dependieran de una partición particular de los datos.

Plan de análisis

El análisis se desarrolló en varias etapas complementarias, combinando técnicas estadísticas y modelos de Machine Learning para identificar patrones dentro de las PQRDS del sistema de salud en Bucaramanga.

El punto de partida fue un análisis exploratorio del dataset, con el que se buscó entender la estructura de los datos, identificar tendencias, detectar valores atípicos y examinar la distribución de las variables más relevantes. Para esto se usaron herramientas de estadística descriptiva

como distribución de frecuencias, medidas de tendencia central, análisis de correlación y visualizaciones estadísticas.

Una vez explorados los datos, se procedió al entrenamiento de modelos supervisados de clasificación. Se implementaron algoritmos como Árbol de Decisión, Random Forest y Regresión Logística, este último principalmente como referencia comparativa de desempeño. Para garantizar una evaluación objetiva, los datos se dividieron en conjuntos de entrenamiento y prueba, y se aplicó validación cruzada para verificar la estabilidad de los resultados. El rendimiento de cada modelo se midió usando métricas de clasificación como accuracy, precisión, recall, F1-score y matriz de confusión.

Adicionalmente, se realizó un análisis de importancia de variables para identificar cuáles factores tienen mayor influencia en la clasificación de riesgo de las PQRDS, lo que aporta valor interpretativo más allá de los resultados numéricos del modelo.

Los resultados obtenidos se interpretaron en función del contexto del sistema de salud, buscando que los hallazgos puedan ser útiles para la toma de decisiones a través de herramientas de analítica de datos.

Conclusiones preliminares

El dataset original contenía registros de PQRDS de distintos municipios del país, por lo que el primer paso fue filtrar geográficamente la información para quedarse únicamente con los registros de Bucaramanga.

Con los datos ya delimitados, se hizo una depuración que incluyó la eliminación de duplicados, el tratamiento de valores nulos y la selección de las variables más relevantes para el análisis. Esto permitió obtener un dataset más limpio y consistente, listo para el entrenamiento de los modelos.

Análisis exploratorio de datos

- **Distribución por género:** En cuanto al género, los registros muestran que la mayoría de las PQRDS fueron presentadas por usuarias mujeres. Esto podría explicarse por una mayor frecuencia de uso de los servicios de salud por parte de esta población, o simplemente por una mayor disposición a reportar inconformidades cuando se presentan.

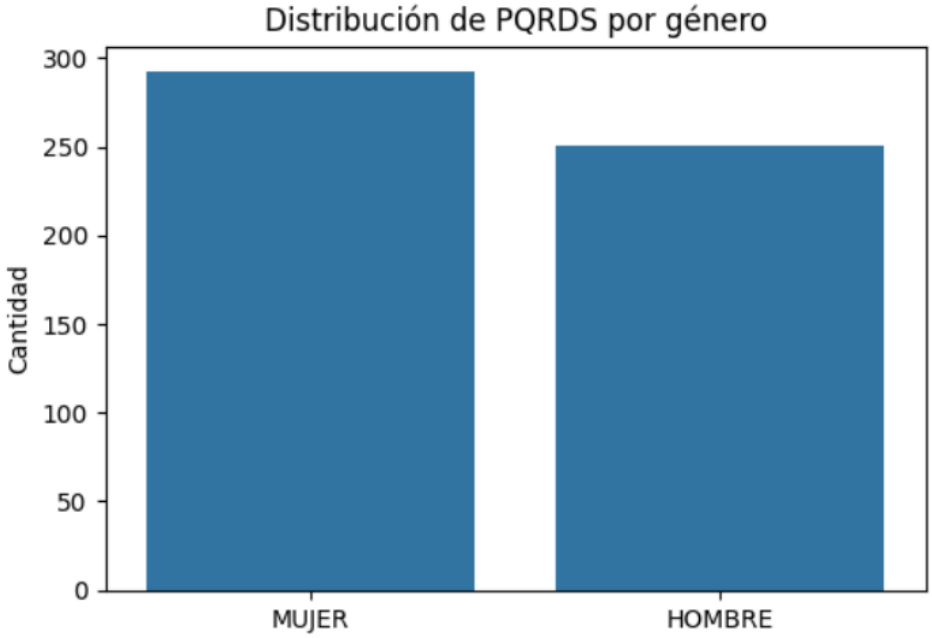

Figura 1. Distribución de reclamos por género.

- **Distribución por edades:** La distribución por edades mostró que la mayor parte de los registros se concentra en población adulta, lo cual tiene sentido si se considera que este grupo tiende a usar los servicios de salud con más frecuencia y, por eso, también es más probable que presente reclamaciones relacionadas con atención médica o procesos administrativos.

- **Motivos generales más frecuentes:** Los resultados muestran una concentración clara en ciertos motivos de reclamación. El motivo más frecuente es la negación de tecnologías en salud y otros elementos complementarios para la atención del usuario, seguido de cerca por la falta de oportunidad en la prestación de esas mismas tecnologías. En tercer lugar, aparece la falta de oportunidad en la autorización de tecnologías en salud, con una frecuencia considerablemente menor pero igualmente relevante.

  A partir del cuarto motivo en adelante, la frecuencia cae de forma significativa, con categorías relacionadas con trámites administrativos, reconocimiento de prestaciones económicas e insatisfacción con la atención del personal.

  Esto evidencia que la mayor parte de las inconformidades en Bucaramanga no están relacionadas con la atención directa del personal de salud, sino con barreras de acceso a servicios y tecnologías médicas, lo que apunta a problemas estructurales en los procesos de autorización y prestación dentro de las EPS. Estas áreas representan focos críticos que las entidades responsables deberían priorizar para mejorar la experiencia del usuario en el sistema de salud.

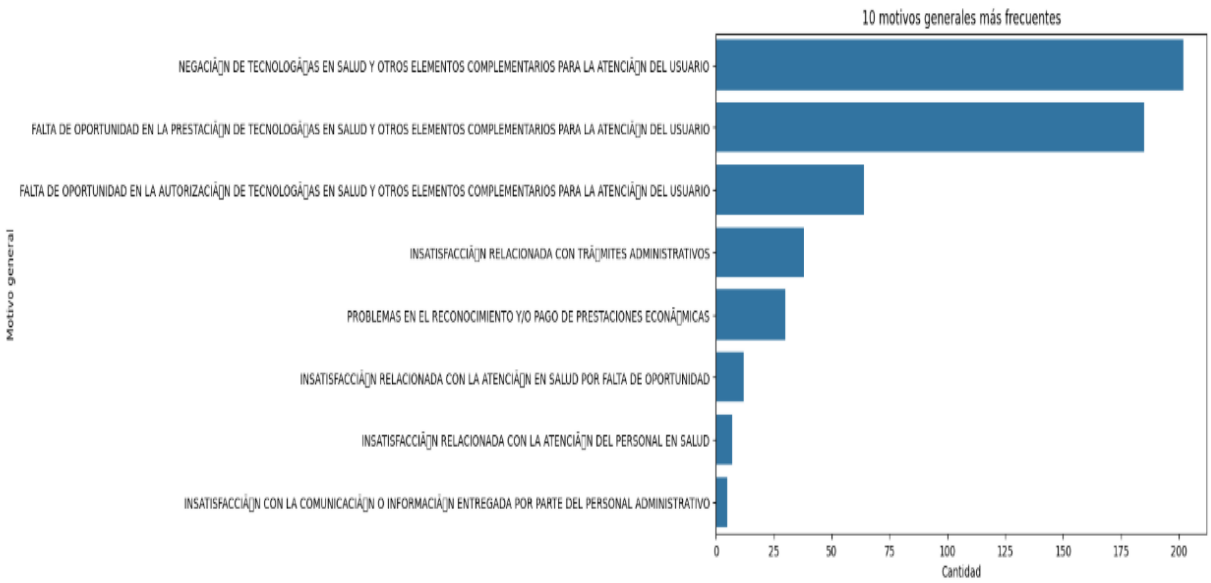

Figura 2. Motivos generales mas frecuentes de las PQRDS.

- **Entidades con mayor número de PQRDS:** La gráfica muestra que Nueva EPS concentra el mayor número de PQRDS en Bucaramanga, seguida por Salud Total en segundo lugar y Sanitas en tercero. Más abajo aparecen Suramericana EPS, Famisanar y Coosalud con frecuencias similares entre sí. El resto de las entidades, como Fuerzas Militares, Salud Mía, Magisterio y la Secretaría Departamental de Salud de Santander presentan volúmenes notablemente bajos en comparación con las primeras.

  Esta concentración en Nueva EPS y Salud Total no necesariamente indica una peor calidad en la prestación del servicio, sino que puede estar relacionada con el tamaño de su base de afiliados en la ciudad. Una entidad con más usuarios afiliados tendrá naturalmente más probabilidades de acumular un mayor número de reclamaciones. Por eso, para hacer una comparación más justa entre entidades, sería necesario relativizar estos registros frente al total de afiliados de cada una, lo cual va más allá del alcance de este análisis, pero representa una línea interesante para investigaciones futuras.

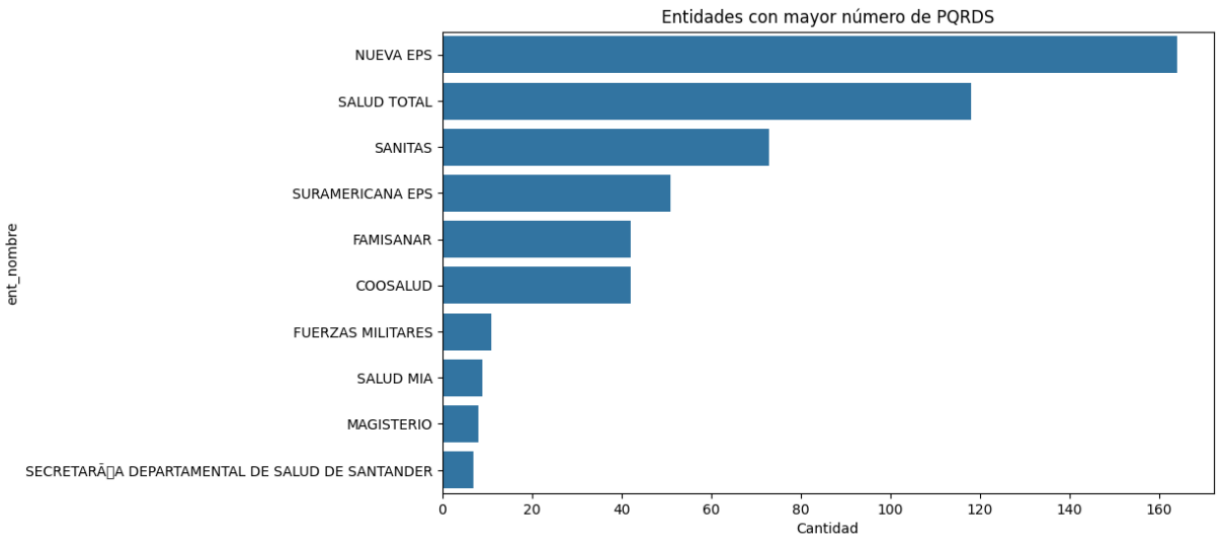

Figura 3. Entidades con mayor cantidad de reclamos.

- **Distribución de la clasificación de riesgo:** La distribución de la variable objetivo revela un desbalance considerable entre las tres categorías de clasificación de riesgo. La categoría Simple domina ampliamente el dataset, seguida a distancia por Priorizado, mientras que Riesgo Vital prácticamente no tiene representación en los registros.

  Este desbalance es una limitación importante a tener en cuenta, ya que los modelos de Machine Learning tienden a favorecer las clases con más datos durante el entrenamiento. En la práctica, esto significa que el modelo podría clasificar bien los casos Simple pero tener dificultades para identificar correctamente los casos Priorizado y, especialmente, los de Riesgo Vital, que son justamente los más críticos dentro del sistema de salud

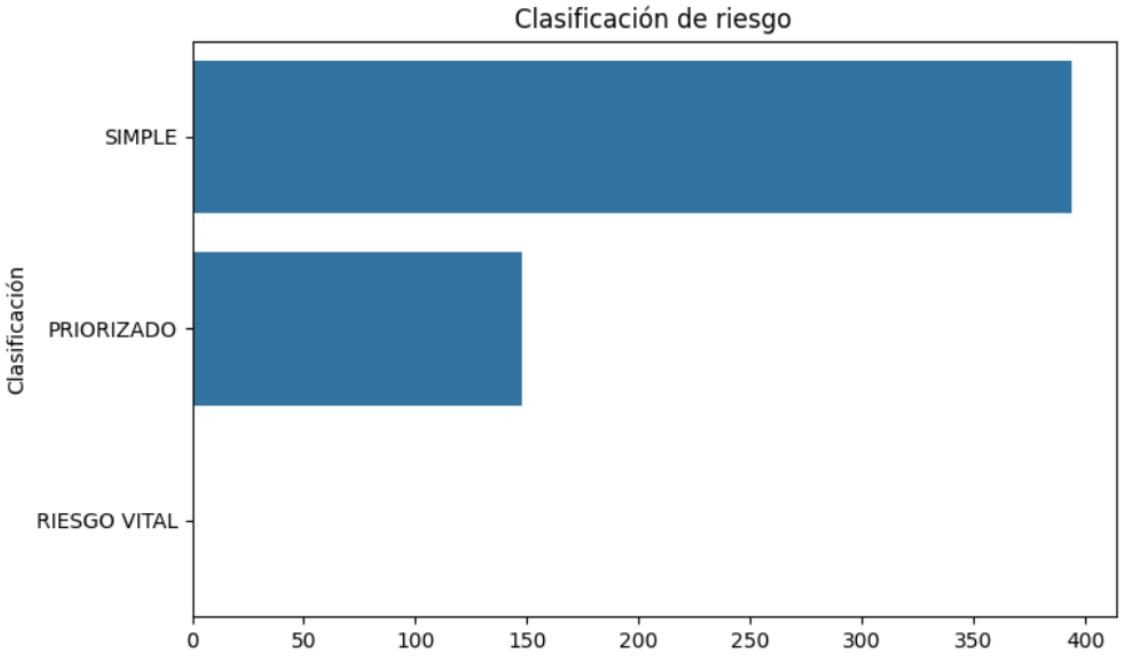

Figura 4. Clasificación de riesgo de las PQRDS.

Resultados de modelos de Machine Learning

Se entrenaron tres modelos de clasificación supervisada para predecir la clasificación de riesgo de las PQRDS: Regresión Logística, Árbol de Decisión y Random Forest. Los resultados obtenidos fueron los siguientes:

Modelo	Accuracy	F1-score clase Simple	F1-score clase Priorizado
Regresión Logística	66.9%	0.80	0.10
Árbol de Decisión	62.0%	0.74	0.33
Random Forest	62.4%	0.76	0.16

Tabla 3. Fuente: Elaboración propia.

Interpretación por modelo

La Regresión Logística obtuvo el accuracy más alto de los tres con un 66.9%, sin embargo, este resultado es engañoso. Al revisar el detalle por clase, se nota que el modelo prácticamente ignora la clase Priorizado, con un F1-score de apenas 0.10, lo que significa que clasifica casi todo como Simple. Esto es una consecuencia directa del desbalance en los datos, donde la clase Simple domina ampliamente el dataset.

El Árbol de Decisión, aunque tuvo un accuracy menor del 62%, mostró un comportamiento más equilibrado entre las dos clases. Su F1-score para la clase Priorizado fue de 0.33, el más alto de los tres modelos, lo que indica que tuvo mayor capacidad para identificar casos prioritarios, que son justamente los más relevantes desde el punto de vista del sistema de salud.

El Random Forest, por su parte, obtuvo resultados similares al Árbol de Decisión en términos de accuracy, pero con un F1-score para la clase Priorizado de solo 0.16, por debajo incluso de la Regresión Logística en esa clase. Esto sugiere que, a pesar de ser un modelo más complejo, el desbalance de clases afectó su capacidad de generalización sobre la clase minoritaria.

Importancia de variables

El análisis de importancia de variables del Random Forest reveló que la edad del usuario, el motivo específico y el mes de registro son los factores que más influyen en la clasificación de riesgo, concentrando entre los tres casi el 49% del peso total del modelo. Le siguen en relevancia el nombre de la entidad y el canal de atención. Variables como el género, el rango de edad y el macromotivo tuvieron una influencia considerablemente menor.

Este resultado es coherente con el contexto del problema: la edad del usuario y el tipo específico de motivo del reclamo son factores que naturalmente inciden en la prioridad que se le debe dar a una PQRDS dentro del sistema de salud.

	Variable	Importancia
3	afec_edad	0.171656
9	motivo_especifico	0.163875
0	mes	0.153566
6	ent_nombre	0.143800
1	pqr_canal	0.096432
5	afec_regafiliacion	0.062043
7	macromotivo	0.058189
8	motivo_general	0.056764
2	afec_genero	0.050796
4	afec_edadr	0.042879

Figura 5. Análisis de importancia de variables.

Limitaciones identificadas

El principal factor que limitó el desempeño de los modelos fue el desbalance entre clases. La categoría Riesgo Vital no tuvo representación suficiente en el dataset para ser aprendida por ninguno de los modelos, y la clase Priorizado estuvo subrepresentada frente a Simple. Para futuras investigaciones, sería recomendable aplicar técnicas de balanceo como SMOTE o ajuste de pesos por clase, lo que probablemente mejoraría significativamente el desempeño sobre las clases minoritarias, que son las más críticas para la toma de decisiones en salud.

Matrices de confusión

A) Regresión Logística

La matriz de confusión de la Regresión Logística muestra que el modelo logró un buen desempeño en la clasificación de casos Simple, acertando en 71 de 79 registros. Para la clase Priorizado identificó correctamente 2 de 30 casos, lo que es consistente con lo que ya indicaban las métricas y refleja la influencia del desbalance del dataset sobre el aprendizaje del modelo. Este resultado es un punto de partida útil para entender el comportamiento de

un modelo base, y evidencia claramente hacia dónde deben dirigirse los ajustes en versiones futuras del modelo.

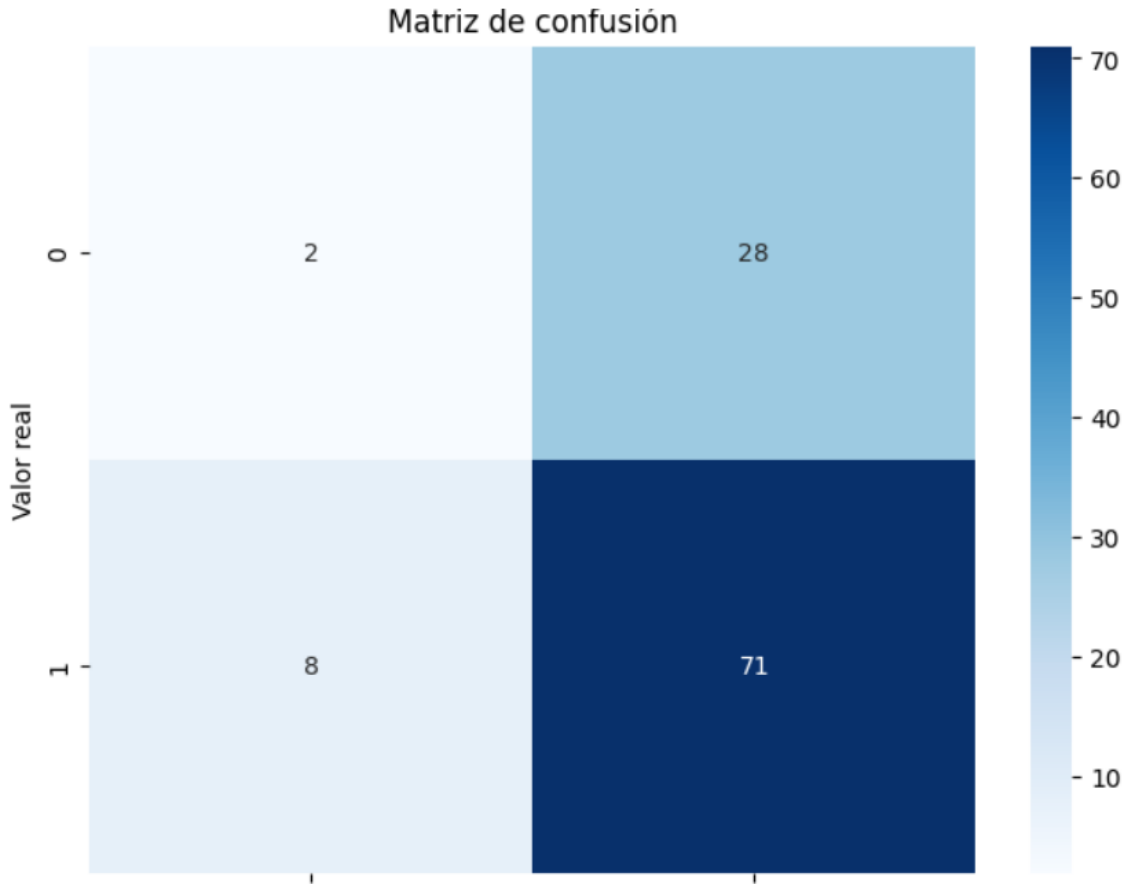

Figura 6. Matriz de confusión modelo regresión logística.

B) Random Forest

La matriz del Random Forest muestra un comportamiento similar al modelo anterior, con algunas diferencias puntuales. Para la clase Priorizado logró identificar correctamente 4 casos, una mejora frente a la Regresión Logística, lo que sugiere que el modelo empieza a capturar algo de la estructura de esa clase. Para la clase Simple acertó en 64 de 79 casos, distribuyendo un poco más los errores entre ambas clases, lo que indica un comportamiento ligeramente más equilibrado en comparación.

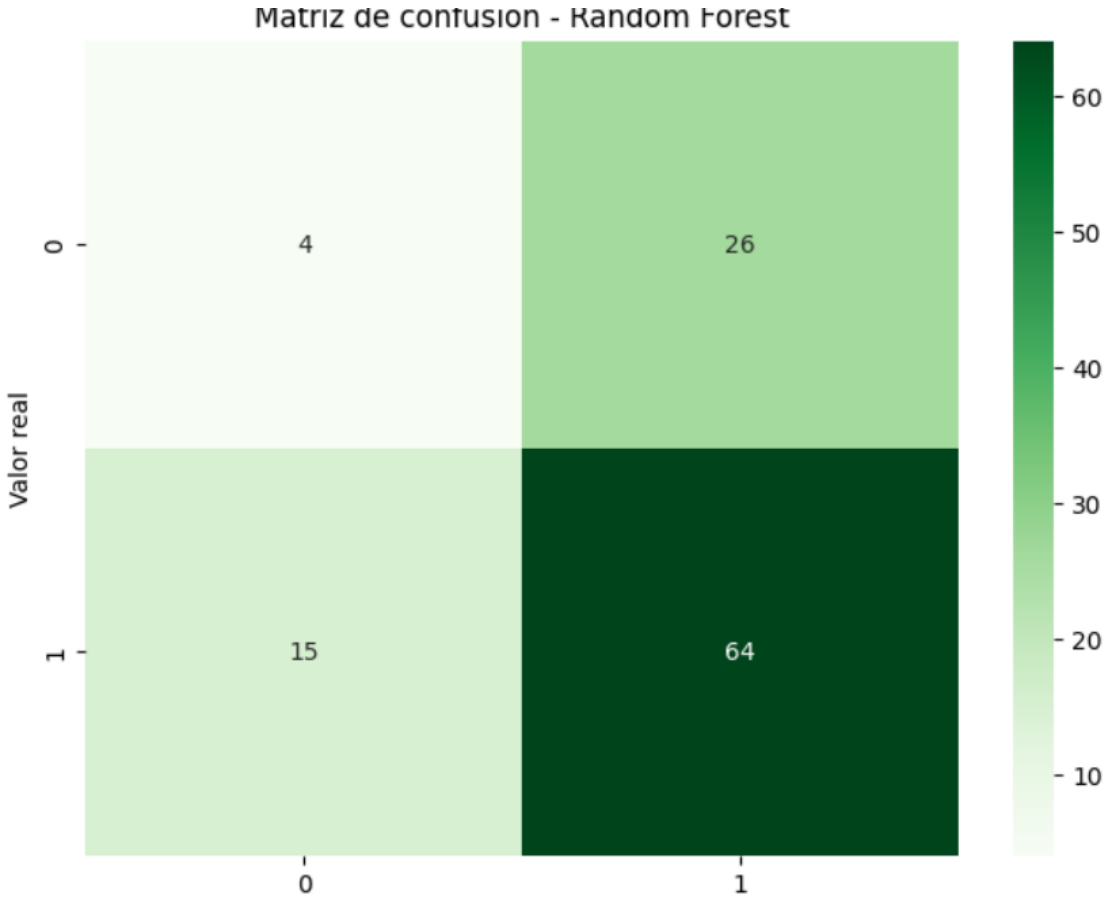

Figura 7. Matriz de confusión modelo Random forest.

En ambos modelos se observa un patrón consistente: mayor capacidad para clasificar correctamente los casos Simple, mientras que la detección de casos Priorizado aún presenta oportunidades de mejora. Este comportamiento es esperado dado el desbalance natural del dataset, donde la clase Simple es ampliamente mayoritaria, y representa una característica común en problemas de clasificación con datos reales del sector salud.

Estos resultados sientan una base inicial prometedora para el desarrollo de herramientas de apoyo a la gestión de PQRDS, y abren la puerta a mejoras concretas en futuras iteraciones,

como la aplicación de técnicas de balanceo de clases, que permitirían al modelo mejorar su capacidad de detección en las categorías minoritarias.

Conclusiones

Este trabajo tuvo como propósito explorar el uso de técnicas de Machine Learning para la clasificación automática de PQRDS del sistema de salud en Bucaramanga, utilizando datos abiertos del primer semestre de 2025 publicados por la Superintendencia Nacional de Salud.

A lo largo del desarrollo se logró construir un flujo completo de análisis de datos que abarcó desde la recolección y depuración de la información hasta el entrenamiento y evaluación de tres modelos de clasificación supervisada. Este proceso en sí mismo representa un aporte metodológico relevante, ya que demuestra que es posible trabajar con datos abiertos del sector salud para generar modelos predictivos con fines de apoyo a la gestión institucional.

En cuanto a los resultados, los tres modelos evaluados —Regresión Logística, Árbol de Decisión y Random Forest— mostraron un desempeño aceptable para la clase mayoritaria Simple, con accuracies que oscilaron entre el 62% y el 67%. El análisis de importancia de variables reveló que la edad del usuario, el motivo específico del reclamo y el mes de registro son los factores que más influyen en la clasificación de riesgo, lo cual es coherente con la naturaleza del problema.

El principal desafío identificado fue el desbalance entre las clases de la variable objetivo, lo que limitó la capacidad de los modelos para detectar casos Priorizado y prácticamente imposibilitó el aprendizaje sobre la clase Riesgo Vital. Este es un hallazgo importante que orienta los pasos a seguir en investigaciones futuras.

En términos generales, este trabajo demuestra que el análisis de PQRDS mediante Machine Learning es viable y tiene potencial para convertirse en una herramienta de apoyo a la toma de

decisiones dentro del sistema de salud, siempre que se continúe trabajando en la calidad y el balanceo de los datos disponibles.

Recomendaciones

En términos generales, este trabajo representa un primer acercamiento al uso de Machine Learning para el análisis de PQRDS en el sistema de salud de Bucaramanga. Si bien los resultados obtenidos no alcanzan aún niveles de desempeño óptimos para una implementación directa, sí demuestran que el enfoque es viable y que existe una base sólida sobre la cual seguir construyendo. Este trabajo cumple su objetivo como punto de partida, abriendo una línea de investigación con potencial real para convertirse en una herramienta de apoyo a la toma de decisiones dentro del sistema de salud, en la medida en que se continúe trabajando en la calidad de los datos, el balanceo de clases y la exploración de modelos más avanzados.

A partir de los hallazgos obtenidos, se plantean las siguientes recomendaciones para futuras investigaciones o para quienes deseen continuar este trabajo:

- Aplicar técnicas de balanceo de clases como SMOTE o ajuste de pesos por clase durante el entrenamiento, con el objetivo de mejorar la capacidad del modelo para detectar casos Priorizado y Riesgo Vital, que son los más críticos dentro del sistema de salud.

- Ampliar el conjunto de datos incluyendo registros de otros semestres o de otros municipios del departamento, lo que permitiría contar con más ejemplos de las clases minoritarias y mejorar la generalización de los modelos.

- Explorar algoritmos más avanzados como XGBoost o redes neuronales, que en problemas con desbalance de clases suelen ofrecer mejores resultados que los modelos tradicionales evaluados en este estudio.

Referencias

Superintendencia Nacional de Salud. (2025). Peticiones, quejas, reclamos, denuncias y sugerencias (PQRDS) – Datos abiertos. Portal de datos abiertos. https://mapas.supersalud.gov.co/arcgisportal/apps/sites/#/datosabiertos/datasets/289c77231d14a968224150c585d08a3/about

Consultorsalud. (2025). PQRDS en Colombia: Supersalud atribuye el aumento a fallas del sistema de salud.
https://consultorsalud.com/

Ministerio de Salud y Protección Social. (2023). Análisis del sistema de salud en Colombia.
https://www.minsalud.gov.co

Fernández Genaro, J. L. (2023). *Modelos de Machine Learning para la ciencia de datos* [Trabajo de Fin de Grado, Universidad de Salamanca]. Repositorio GREDOS. https://gredos.usal.es/handle/10366/157171

MathWorks. (s.f.). *Introducción a Machine Learning*. https://la.mathworks.com/discovery/machine-learning.html

Microsoft. (2026, febrero 25). *Modelo de Machine Learning en Microsoft Fabric*. Microsoft Learn. https://learn.microsoft.com/es-es/fabric/data-science/machine-learning-model

Fernández Villafañez, S. (2022). *Métodos de regresión y clasificación basados en árboles* [Trabajo de Fin de Grado, Universidad de Valladolid]. UVaDOC Repositorio Documental. https://uvadoc.uva.es/handle/10324/53822

Hoyos, C. C. (2025). Una introducción a los modelos de clasificación empleando R (Cap. 8: Modelo Random Forest). Editorial Universidad Icesi. https://www.icesi.edu.co/editorial/intro-clasificacion-web/Random.html

Anexos

Anexo A. Código de limpieza y procesamiento de datos

3. Filtrado y selección de datos

```
# Verificar nombres de columnas disponibles
print('Columnas del dataset:')
print(dataset.columns.tolist())

# Filtrar registros correspondientes al municipio de Bucaramanga (código 68001)
dataset_bga = dataset[dataset['pet_cod_mpio'] == 68001]
print('\nRegistros de Bucaramanga:', dataset_bga.shape)
```

4. Limpieza de datos

```
# Revisar cantidad de valores nulos por columna
print('Valores nulos por columna:')
print(dataset_bga.isnull().sum())

# Eliminar filas con valores nulos y registros duplicados
dataset_bga = dataset_bga.dropna()
dataset_bga = dataset_bga.drop_duplicates()

print('\nDimensiones después de la limpieza:', dataset_bga.shape)
```

Anexo B. Código de modelos de Machine Learning

7.1 Regresión Logística

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Entrenar modelo y generar predicciones
modelo_lr = LogisticRegression(max_iter=1000)
modelo_lr.fit(X_train, y_train)
y_pred_lr = modelo_lr.predict(X_test)

# Métricas de evaluación
print('Accuracy Regresión Logística:', accuracy_score(y_test, y_pred_lr))
print(classification_report(y_test, y_pred_lr))

# Matriz de confusión
cm_lr = confusion_matrix(y_test, y_pred_lr)
plt.figure(figsize=(8, 6))
sns.heatmap(cm_lr, annot=True, fmt='d', cmap='Blues')
plt.title('Matriz de confusión - Regresión Logística')
plt.xlabel('Predicción')
plt.ylabel('Valor real')
plt.tight_layout()
plt.show()
```

7.2 Árbol de Decisión

```
from sklearn.tree import DecisionTreeClassifier

# Entrenar modelo y generar predicciones
modelo_tree = DecisionTreeClassifier(random_state=42)
modelo_tree.fit(X_train, y_train)
y_pred_tree = modelo_tree.predict(X_test)

# Métricas de evaluación
print('Accuracy Árbol de Decisión:', accuracy_score(y_test, y_pred_tree))
print(classification_report(y_test, y_pred_tree))
```

7.3 Random Forest

```python
from sklearn.ensemble import RandomForestClassifier

# Entrenar modelo Random Forest con 100 árboles
modelo_rf = RandomForestClassifier(n_estimators=100, random_state=42)
modelo_rf.fit(X_train, y_train)
y_pred_rf = modelo_rf.predict(X_test)

# Métricas de evaluación
print('Accuracy Random Forest:', accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))

# Matriz de confusión
cm_rf = confusion_matrix(y_test, y_pred_rf)
plt.figure(figsize=(8, 6))
sns.heatmap(cm_rf, annot=True, fmt='d', cmap='Greens')
plt.title('Matriz de confusión - Random Forest')
plt.xlabel('Predicción')
plt.ylabel('Valor real')
plt.tight_layout()
plt.show()
```

Anexo C. Código de análisis

El código completo utilizado para el procesamiento de datos, análisis exploratorio y entrenamiento de los modelos de clasificación se encuentra disponible en el siguiente enlace:

https://github.com/OSVHdez0/Desarrollo-Gu-a-2PCD/blob/main/Analisis_Reclamos_Salud.ipynb