**Comportamiento de Criminalidad en Colombia 2003-2023**

Proyecto-G24

1. **Contexto del problema.**

La criminalidad en Colombia ha sido una preocupación persistente, a lo largo de las últimas décadas, el país ha enfrentado desafíos significativos en términos de seguridad ciudadana debido a diversos factores socioeconómicos, políticos y culturales.

Durante años, Colombia ha luchado contra altas tasas de criminalidad, vinculadas en gran medida a actividades delictivas relacionadas con el narcotráfico, la violencia asociada a grupos armados, la pobreza, la desigualdad social y la falta de oportunidades. Si bien ha habido avances notables en la reducción de algunos tipos de delitos, el problema de la seguridad urbana persiste y varía significativamente entre diferentes ciudades y regiones.

Este contexto define un desafío crucial en Colombia: entender la evolución de la violencia e inseguridad en las ciudades principales, ofreciendo la oportunidad de informar políticas públicas y estrategias de prevención del delito que sean más efectivas y estén respaldadas por evidencia empírica.

1. **Pregunta de negocio y alcance del proyecto.**

2.1 Pregunta de negocio:

¿Cómo ha evolucionado la tasa de criminalidad en las ciudades de Colombia durante los últimos 20 años y cuáles son los factores que influyen en esta evolución?

2.2 Alcance del proyecto:

El análisis de la evolución de las tasas de criminalidad entre 2003 y 2023 busca comprender estas dinámicas complejas. La intención es identificar tendencias, patrones estacionales y diferencias entre ciudades para comprender mejor los factores subyacentes que influyen en la violencia e inseguridad. Es clave considerar las limitaciones de los datos, como el subregistro de delitos no denunciados y los sesgos en la recolección de información, para realizar un análisis preciso y completo.

Se desarrollará una herramienta de visualización interactiva para resumir gráficamente estas tendencias delictivas, apoyando la toma de decisiones para reducir la actividad delictiva en estas ciudades mediante una presentación clara y accesible de la información recopilada.

1. **Breve descripción de conjuntos de datos a emplear.** Resalte posibles cambios con respecto a la segunda entrega.

Una vez consolidados todos los data sets, las relaciones clave podrían ser identificadas a través de la distribución temporal de incidentes ('FECHA HECHO', 'Fecha Hecho', 'FECHA HECHOS'), el análisis geográfico de delitos por departamento y municipio ('COD\_DEPTO', 'DEPARTAMENTO', 'COD\_MUNI', 'MUNICIPIO'), y la comprensión de la tipología del delito ('DESCRIPCION CONDUCTA') en relación con la cantidad de incidentes ('CANTIDAD'). Además, las variables adicionales como 'SEXO', 'tipo', 'DEPTO\_GIE', 'MUNICIPIO\_GIE', 'Zona' podrían ser exploradas para comprender cómo influyen en la ocurrencia delictiva. La columna 'CASOS' puede ser utilizada para evaluar la gravedad o magnitud de los incidentes y su relación con otros factores. Estas relaciones permitirían una comprensión más profunda de la dinámica delictiva en términos temporales, geográficos y tipológicos.

1. **Modelos**

Es importante poder predecir si el índice de homicidios aumentará o bajará en cada departamento de Colombia por varias razones:

* Permite focalizar recursos y programas de seguridad de manera preventiva en las zonas con mayor probabilidad de aumento en la tasa de homicidios. Esto optimiza la asignación de presupuesto.
* Las autoridades pueden comunicar y preparar a la ciudadanía y fuerza pública ante los posibles cambios esperados en los índices delictivos. Esto mejora reacción ante incrementos de violencia.
* El análisis predictivo facilita evaluar el impacto real de políticas públicas actuales y pasadas, y reorientar las estrategias de reducción de violencia de manera oportuna y evidenciada.
* Los resultados de los modelos predictivos permiten estimar necesidades futuras de talento humano, programas sociales y recursos en el sector seguridad por región.
* La predicción de homicidios brinda insumos planificación en otros campos como el sector salud, para proyectar necesidades de talento médico o infraestructura hospitalaria derivadas del posible aumento en tasas lesividad.
* El análisis predictivo de homicidios salva vidas al permitir mayor preparación, eficiencia del gasto público y evaluación evidenciada de política criminal a nivel regional.

4.1 Modelos desarrollados y su evaluación.

El objetivo del modelo será predecir si la cantidad de homicidios va a aumentar o disminuir en función de diversas características como el departamento, municipio, sexo y zona. El modelo de clasificación se entrena con datos históricos para identificar patrones que puedan relacionarse con cambios en la cantidad de homicidios, con la finalidad de anticipar si esa cantidad se incrementará (1) o disminuirá (0) en futuros periodos. Esto puede ser útil para anticipar tendencias en la incidencia de homicidios y, en consecuencia, para la planificación de medidas preventivas o de respuesta por parte de autoridades o entidades pertinentes.

* + 1. RandomForestClassifier

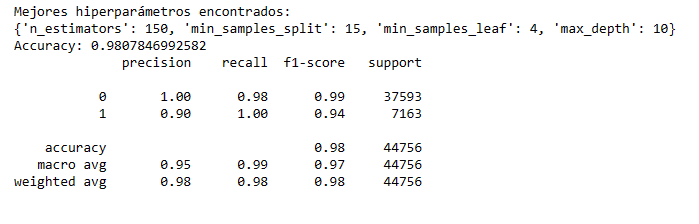
El uso de RandomForestClassifier para predecir si la cantidad de homicidios subirá o bajará de acuerdo con la naturaleza del problema se basa en las siguientes características:

* Flexibilidad del modelo: RandomForest es un método versátil que puede manejar datos no lineales y complejos. Puede aprender patrones en los datos sin supuestos estrictos sobre la relación entre las características y la variable objetivo.
* Capacidad para identificar patrones: Este modelo es capaz de detectar relaciones no lineales entre las características y la variable objetivo. En el caso de la cantidad de homicidios, pueden existir patrones complejos y no lineales que podrían ser capturados por este algoritmo.
* Manejo de múltiples características: RandomForest puede manejar múltiples características de manera efectiva. En este caso, estás utilizando características como el departamento, municipio, sexo y zona para predecir cambios en la cantidad de homicidios.
* Resistencia al sobreajuste (overfitting): Random Forest tiende a ser menos propenso al sobreajuste que otros modelos más complejos. Esto puede ser beneficioso cuando se tienen datos limitados o ruidosos.
* Robustez y desempeño: En general, RandomForest es robusto y puede funcionar bien en una amplia gama de problemas, proporcionando buenos resultados incluso con un ajuste mínimo.

RESULTADOS:

El alto valor de precisión y recall para ambas clases, así como la alta precisión general del modelo (98%), sugieren que es capaz de predecir con precisión si la cantidad de homicidios subirá o bajará. Esto es respaldado por los valores de f1-score cercanos a 1, indicando un buen equilibrio entre precisión y recall. Estos resultados se derivan de las métricas comunes utilizadas en problemas de clasificación, como la precisión, recall y f1-score.

* Precisión: Mide la proporción de predicciones positivas correctas sobre todas las predicciones positivas. En este caso, la precisión general es del 98%, lo que significa que el 98% de las veces que el modelo predice que la cantidad de homicidios subirá o bajará, acierta.
* Recall (Sensibilidad): Mide la proporción de verdaderos positivos predichos correctamente sobre todos los verdaderos positivos. En la clase 0 (cantidad no subirá), el recall es del 98%, lo que indica que el modelo identifica correctamente el 98% de las veces cuando la cantidad de homicidios no subirá. En la clase 1 (cantidad subirá), el recall es del 100%, lo que significa que el modelo predice correctamente cada vez que la cantidad de homicidios efectivamente subirá.
* F1-score: Es una medida combinada de precisión y recall. Es útil cuando las clases están desbalanceadas. Un valor de 1 indica una precisión y recall perfectos, y en este caso, tienes valores muy altos, lo que sugiere un modelo muy preciso y sensible.
* Accuracy (Exactitud): Es la proporción de predicciones correctas sobre el total de predicciones. En este caso, el modelo tiene una exactitud del 98%, lo que significa que clasifica correctamente el 98% de todas las muestras.



Predicción:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* + 1. XGBoost

El modelo utiliza datos como la fecha en la que ocurrieron los homicidios, el departamento y municipio donde ocurrieron, posiblemente información demográfica (sexo, zona), y la cantidad de homicidios en períodos anteriores. Estos tipos de modelos de clasificación tienen varias características distintivas:

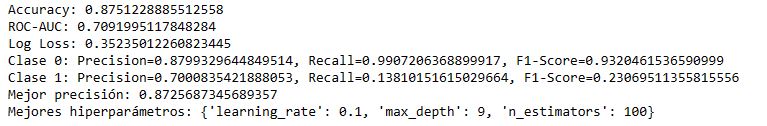
* Clasificación binaria: Este tipo de modelo clasifica los datos en dos categorías distintas: en este caso, predecir si la cantidad de homicidios va a subir o bajar. Esto se traduce en la necesidad de tener un conjunto de datos etiquetado con estas dos clases.
* Utiliza datos históricos: Estos modelos se basan en datos previos de la cantidad de homicidios, junto con otras características relevantes, como la fecha, ubicación geográfica, posiblemente información demográfica (sexo, zona), entre otros, para hacer predicciones futuras.
* Aprendizaje supervisado: Estos modelos se entrenan usando un conjunto de datos previamente etiquetado, lo que significa que requieren datos históricos con las respuestas conocidas (si la cantidad de homicidios subió o bajó en períodos anteriores).
* Algoritmos de Gradient Boosting: Estos algoritmos funcionan construyendo varios modelos débiles (generalmente árboles de decisión) secuencialmente, donde cada uno intenta corregir los errores del anterior.

RESULTADOS:

El modelo tiene un buen desempeño general con un alto accuracy del 87.51%. Sin embargo, hay algunas observaciones importantes:

* Accuracy (Exactitud): Esta métrica indica la proporción de predicciones correctas sobre el total de predicciones. El modelo tiene una tasa de predicciones correctas del 87.51%, lo cual es bastante bueno.
* ROC-AUC: El área bajo la curva ROC (Receiver Operating Characteristic) es una medida de la capacidad del modelo para distinguir entre clases. Aquí, tenemos un valor de 0.7092, lo que sugiere una capacidad moderada del modelo para discernir entre las clases.
* Log Loss (Pérdida Logarítmica): Es una medida de la incertidumbre del modelo. Cuanto menor sea este valor, mejor. tenemos un valor de 0.3523, lo cual es bastante bajo y señala que el modelo tiene un buen nivel de certeza en sus predicciones.
* Precision, Recall, F1-Score para Clase 0: Estas métricas evalúan la calidad de las predicciones para la clase mayoritaria, posiblemente la categoría de "baja" en tu contexto. Aquí, el modelo muestra excelentes resultados, con altos valores en precision, recall y F1-score, lo que indica que predice muy bien esta clase.
* Precision, Recall, F1-Score para Clase 1: Estas métricas miden la capacidad del modelo para predecir la clase minoritaria, posiblemente la categoría de "Sube". Aunque la precisión es baja, el recall y F1-score sugieren que el modelo puede identificar esta clase en cierta medida.

Los mejores hiperparámetros encontrados ('{'learning\_rate': 0.1, 'max\_depth': 9, 'n\_estimators': 100}') indican que un learning rate más bajo, árboles más profundos y un número mayor de estimadores contribuyen a un mejor rendimiento del modelo, como se observa a continuación.



4.1.3 Red Neuronal

Las redes neuronales, especialmente las redes neuronales profundas como las redes neuronales convolucionales (CNN) o las redes neuronales recurrentes (RNN), como el LSTM que usamos, pueden aprender patrones complejos en los datos. Las razones por las cuales una red neuronal podría ser apropiada para este proyecto de predicción de homicidios:

* Capacidad para capturar patrones complejos: Las redes neuronales son excelentes para capturar patrones complejos y no lineales en los datos. En un conjunto de datos de homicidios, donde las relaciones pueden ser intrincadas y no fácilmente discernibles, las redes neuronales podrían encontrar patrones que otros modelos podrían pasar por alto.
* Flexibilidad en la estructura: Las redes neuronales pueden ser configuradas con múltiples capas y diferentes arquitecturas, lo que permite adaptar la red al problema específico y experimentar con diferentes configuraciones para mejorar el rendimiento.
* Capacidad para manejar grandes cantidades de datos: Si tienes una gran cantidad de datos, las redes neuronales pueden ser efectivas, ya que a menudo funcionan mejor con conjuntos de datos más grandes.
* Predicciones secuenciales o temporales: Si tu proyecto incluye datos secuenciales en el tiempo (como patrones de homicidios a lo largo de los años o meses), las redes neuronales recurrentes (RNN) o las redes neuronales convolucionales (CNN) pueden ser útiles para capturar estas relaciones secuenciales.
* Adaptabilidad a distintos tipos de datos: Las redes neuronales pueden manejar una variedad de datos, ya sean datos numéricos, categóricos, de imágenes o de texto, lo que las hace versátiles para problemas multidimensionales.

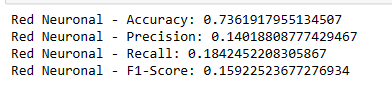
Sin embargo, las redes neuronales no son siempre la mejor opción. Pueden requerir más datos para entrenar de manera efectiva, más tiempo de cómputo y ajuste de hiperparámetros, y suelen ser más complejas de interpretar que algunos otros modelos como los árboles de decisión.

RESULTADOS:

Parece que los resultados de la red neuronal en cuanto a precisión, recall y F1-score no son tan altos como los obtenidos con otros modelos. Esto podría significar que, para este conjunto de datos específico, la red neuronal no está logrando un rendimiento tan bueno como los modelos de XGBoost o Random Forest.

* La precisión indica cuántas de las predicciones positivas son realmente correctas. En este caso, aproximadamente el 14% de las predicciones positivas son verdaderamente correctas, lo que podría ser bajo dependiendo del contexto.
* El recall, que mide la proporción de verdaderos positivos que se identificaron correctamente, también es relativamente bajo, alrededor del 18%
* El F1-score, que combina precisión y recall en una sola métrica, también refleja un rendimiento moderado, alrededor del 15.9%.

En general, estos resultados sugieren que la red neuronal puede no ser la mejor opción para este conjunto de datos específico o que quizás necesita una configuración diferente o más ajuste de hiperparámetros para mejorar su rendimiento. Es posible que los modelos de XGBoost o Random Forest sean más adecuados para este problema en particular, dado su rendimiento más alto en las métricas evaluadas.



4.1.3 Elección del modelo

La elección del modelo depende de varios factores, no solo de la precisión. La precisión es importante, pero no siempre cuenta toda la historia. Aquí hay algunas consideraciones adicionales al elegir un modelo:

* Precisión: Si solo te guías por la precisión, el modelo de Random Forest parece tener la precisión más alta.
* Interpretación: Algunos modelos como Random Forest o XGBoost pueden ofrecer más interpretabilidad que las redes neuronales. Si es crucial comprender cómo se toman las decisiones, estos modelos podrían ser preferibles.
* Tiempo de entrenamiento y predicción: Las redes neuronales suelen llevar más tiempo de entrenamiento que otros modelos. Si el tiempo de respuesta es importante, es mejor considerar modelos como Random Forest o XGBoost, que tienden a ser más rápidos.
* Capacidad de generalización: A veces, un modelo que tiene una precisión ligeramente inferior en el conjunto de datos de entrenamiento puede generalizar mejor a datos nuevos. Esto se conoce como "overfitting" (sobreajuste) y es importante evaluar cómo los modelos se desempeñan en conjuntos de datos que no han visto antes.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Por lo anterior se elige el modelo de Random Forest, por términos de tiempo de procesamiento y precisión del modelo.

• Descripción del tablero desarrollado y la funcionalidad que éste ofrece.

• Principales resultados y conclusiones.