*“Análisis de datos e implementación de modelo de regresión logística en Python para detección de posibles casos de agresión sexual”*

Adarme Jhojan David

Ingeniería de Sistemas, Facultad de Ingeniería,

Fundación de Educación Superior San José

Profesor: Winder Osorio

Noviembre 1, 2024

*“Lo que te separa de llegar a tu meta es una fina d pared de cristal llamada pasividad.”*

***Tabla de Contenidos***

[*Introducción.* 4](#_Toc181401081)

[*Descripción del proyecto.* 5](#_Toc181401082)

[*Modularización:* 5](#_Toc181401083)

[*Análisis de datos e implementación de modelo de regresión logística.* 5](#_Toc181401084)

[*Exploración de los datos:* 5](#_Toc181401085)

[*Descripción de los datos:* 5](#_Toc181401086)

[*Descripción del proyecto.* 8](#_Toc181401087)

[*Fase de Preparación y Limpieza de Datos:* 8](#_Toc181401088)

[*Fase de Modelado y Evaluación:* 8](#_Toc181401089)

[*Conclusiones* 9](#_Toc181401090)

[*Referencias* 10](#_Toc181401091)

# *Introducción.*

**E**n primer lugar, tenemos que entender en su concepto más básico la definición de la analítica y el big data, para de esta forma dar continuidad a la introducción de lo presentado en la el entregable. La **analítica** es el proceso de examinar y transformar datos con el propósito de descubrir información útil que ayude a tomar decisiones bien fundamentadas. Básicamente, implica recolectar, procesar y analizar datos para encontrar patrones, tendencias o conexiones que aporten valor o apoyo en decisiones estratégicas. Esta práctica es aplicable en múltiples sectores, como el marketing, la salud o las finanzas, y puede incluir desde métodos estadísticos simples hasta técnicas avanzadas de aprendizaje automático. La analítica se clasifica según el objetivo en cuatro tipos principales: descriptiva (para entender qué pasó), diagnóstica (para saber por qué pasó), predictiva (para anticipar qué podría pasar) y prescriptiva (para decidir qué debería hacerse). Por otro lado.

**El Big Data** se refiere a grandes cantidades de datos que se generan rápidamente y en una amplia variedad de formatos, lo cual dificulta su procesamiento mediante herramientas tradicionales de análisis. El Big Data se caracteriza por las llamadas "3Vs": volumen (cantidad de datos), velocidad (la rapidez con la que se generan y procesan) y variedad (la diversidad de formatos y fuentes, que incluye datos estructurados, semiestructurados y no estructurados). Para manejar el Big Data, se necesitan tecnologías avanzadas como almacenamiento distribuido y procesamiento paralelo. El análisis de estos datos permite descubrir información valiosa, como la personalización de servicios, la optimización de operaciones o la detección de fraudes, que de otra forma sería difícil de obtener.

**U**na vez témenos claros los conceptos sobre la analítica y el big data, debemos entender cuál será el contexto en el que utilizaremos estos términos y las tecnologías que de él derivan. Me siento en la necesidad de realizar una aclaración antes de continuar, y es que alguien de mi familia o cercano a mi familia trabaja en una organización gubernamental asociada directamente a casos de delitos sexuales en poblaciones donde los ofensores no superan los 17 años, al notar algunas particularidades de su trabajo, podemos destacar varias cosas positivas en su mayoría, pero también algunas áreas donde se pudiese realizar algún tipo de implementación de mejoras en la predicción de posibles casos en los que quizá de contar con mejores mecanismos de detección se pudieran haber evitado mas de una tragedia y dolor para muchas familias.

# *Descripción del proyecto.*

## *Modularización:*

**E**l Proyecto está dividido en diferentes archivos, siendo estos, este mismo Word, donde debemos plasmar la redacción del artículo, un cuadernillo de Jupyter que contine el código, un archivo de Power BI donde tenemos nuestras vistas y nuestras graficas para poder entender o darle un poco más sentido a nuestros datos, aparte de estos tres elementos principales tendremos una base de datos ficticia generada desde Python con ayuda de la librería “faker” para generar datos falsos, otras bases reales descargadas desde fuente oficiales, así como los dos PDF de los entregables, tanto técnico como académico, además de todo el proyecto estará alojado en un repositorio de GitHub donde se podrá ver desde su fecha de creación, hasta cada versión del paso a paso en su avance y su finalización.

# *Análisis de datos e implementación de modelo de regresión logística.*

## *Exploración de los datos:*

**E**n la exploración de estos datos podremos encontrar distintos focos de investigación: factores demográficos y sociales, patrones en tectología y comportamiento en redes, motivaciones y riesgos psicológicos, características únicas del incidente, correlación de factores de riesgo, posibles factores de reincidencia, desarrollo de perfiles de riesgo. Este análisis podría ser fundamental para instituciones que buscan implementar programas de intervención preventiva, especialmente si trabajan con población adolescente en riesgo de violencia o agresión.

## *Descripción de los datos:*

**A** continuación, realizaremos una breve descripción de los datos, en primer lugar sabemos que esta es la estructura exacta d ellos datos, ya que nosotros mismos estamos generando una base de datos de datos totalmente aleatorios pueden visualizar el cuaderno de Jupyter Notebook desde este link: <https://github.com/underoatjd/Diplomado/blob/main/data.ipynb> o pueden también realizar la verificación completa del proyecto con todos sus elementos desde ente enlace: <https://github.com/underoatjd/Diplomado>.

Factores Demográficos y Sociales:

* Relación entre Edad y Lugar de Residencia: Podrías analizar si existen diferencias en la ocurrencia de estos incidentes entre diferentes grupos de edad y barrios. Esto podría ayudar a enfocar esfuerzos de intervención en áreas específicas y en adolescentes de ciertas edades.
* Composición Familiar: Investigar si la composición del hogar (nuclear, monoparental, extendida) está relacionada con una mayor incidencia de agresión puede proporcionar pistas sobre el entorno familiar.
* Nivel Escolar y Rendimiento Académico: Identificar si el nivel educativo o un bajo rendimiento académico están relacionados con incidentes de agresión. Esto puede sugerir la necesidad de programas de apoyo académico y orientación en instituciones educativas.

Patrones en Tecnología y Redes Sociales:

* Uso Activo de Redes Sociales y Contacto Inapropiado: Explorar si los agresores que usan redes sociales activamente son más propensos a contactar a otras personas de forma inapropiada. Esto podría ayudar a desarrollar políticas de uso responsable de las redes sociales en entornos escolares.
* Edad de Inicio de Consumo de Pornografía: Analizar si una exposición temprana a la pornografía, especialmente a contenido violento o explícito, está relacionada con comportamientos agresivos. Este análisis podría subrayar la importancia de la educación temprana sobre los efectos del consumo de este tipo de contenido.

Motivaciones y Riesgos Psicológicos:

* Trastornos Psicológicos, Adicciones y Atención Psicológica Previa: Observar si los agresores con trastornos psicológicos o adicciones tienen un mayor riesgo de recurrencia o agresión impulsiva. También podrías analizar si aquellos que han recibido atención psicológica previa muestran algún patrón de arrepentimiento o reconocimiento de su conducta.
* Arrepentimiento y Reconocimiento del Comportamiento: Analizar cuántos agresores reconocen su comportamiento como incorrecto y si se relaciona con un mayor arrepentimiento. Estos factores podrían ser útiles para evaluar la efectividad de programas de rehabilitación o intervención psicológica.

Características del Incidente:

* Premeditación y Violencia Física: Comparar incidentes premeditados frente a aquellos que fueron impulsivos y verificar si los actos de violencia física o amenazas están más asociados con un tipo de agresión que con otro.
* Contexto del Acto (Solo o con Otras Personas): Determinar si los agresores que actúan en compañía tienen patrones diferentes a los que actúan solos, lo cual podría sugerir aspectos de influencia social o presión grupal.

Correlación de Factores de Riesgo:

* Relación entre Antecedentes de Maltrato y Comportamiento Agresivo: Examinar si el maltrato previo en el hogar o escuela tiene una relación con la comisión de actos de agresión, lo cual podría ser un indicador clave de riesgo.
* Factores Familiares y Situación Socioeconómica: Explorar si la ocupación de los padres o la estructura familiar tienen alguna relación con el comportamiento delictivo en adolescentes, lo que podría revelar cómo la situación socioeconómica influye en la prevención.

Detección de Factores de Reincidencia:

* Antecedentes Delictivos: Analizar si los adolescentes con antecedentes de agresión son más propensos a cometer actos similares en el futuro. Esto es útil para implementar estrategias de intervención enfocadas en la reincidencia.

Desarrollo de Perfiles de Riesgo:

* A partir de la información recogida, podrías identificar perfiles de riesgo que incluyan combinaciones de factores demográficos, familiares, sociales, y psicológicos. Estos perfiles pueden ser usados para crear alertas tempranas o programas específicos de educación y prevención.

**U**no de los mayores desafíos en la implementación de un modelo de IA para la prevención del abuso sexual infantil es la recolección de datos. La privacidad de los menores debe ser una prioridad, por lo que el sistema estaría diseñado para cumplir con las normativas legales más estrictas sobre protección de datos. Los datos se recopilarían de manera anónima y segura, garantizando que solo se utilicen para fines de prevención y con un consentimiento explícito de los tutores responsables. El impacto potencial del desarrollo real de este sistema es significativo. La detección temprana puede evitar que miles de menores sufran abusos. Además, el modelo puede implementarse en diversas instituciones (escuelas, organizaciones gubernamentales y ONGs) a nivel local y global, adaptándose a diferentes contextos culturales y sociales.

**L**a tecnología de IA aplicada a la prevención del abuso sexual infantil representa una oportunidad sin precedentes para proteger a los menores de una de las formas de violencia más devastadoras. Mediante la recolección ética de datos y el uso de análisis predictivos, este modelo puede generar intervenciones tempranas, reducir el riesgo y, en última instancia, salvar vidas.

# *Descripción del proyecto.*

## *Fase de Preparación y Limpieza de Datos:*

**A**l desarrollar el modelo de regresión, empecé cargando el conjunto de datos y realizando una revisión inicial para observar las primeras filas. Esto me permitió tener una idea general de la estructura de los datos y las variables que se iban a analizar. A continuación, verifiqué la existencia de valores nulos utilizando la función isna().sum(), confirmando que, al haber generado la base de datos de manera controlada, no había datos ausentes que pudieran afectar el modelo. Procedí con la limpieza de datos, enfocándome en la transformación de variables categóricas. Al revisar la información del dataframe, identifiqué que contaba con ocho columnas que representaban variables nominales. Para convertir estas variables a un formato adecuado para el análisis, utilicé la función get\_dummies de pandas, transformando las variables categóricas en variables binarias, lo que es crucial para aplicar correctamente un modelo de regresión.

**U**na vez transformadas las variables categóricas, eliminé las columnas originales del dataframe para evitar redundancias. Después de limpiar el conjunto de datos, uní las nuevas variables binarias al dataframe principal utilizando la función join, asegurando que todas las variables necesarias para el modelo estuvieran presentes y adecuadamente formateadas. Procedí a realizar un análisis descriptivo del conjunto de datos limpio utilizando describe(), lo que proporcionó una visión general de las estadísticas de cada variable. Posteriormente, verifiqué las correlaciones entre las variables mediante la función corr(), lo cual es esencial para entender las relaciones que podrían existir entre las variables independientes y la variable dependiente.

## *Fase de Modelado y Evaluación:*

**S**eparé los datos en dos conjuntos: las características (X) y la variable objetivo (y), donde X incluía todas las columnas excepto la variable "Agresión sexual Confirmada". Luego, importé las funciones necesarias de sklearn para dividir los datos en conjuntos de entrenamiento y prueba, utilizando train\_test\_split, asegurando así una adecuada evaluación del modelo. Creé el modelo de regresión logística importando LogisticRegression de sklearn, especificando un máximo de 1000 iteraciones para garantizar la convergencia del modelo. Al entrenar el modelo con el conjunto de entrenamiento, ajusté los parámetros para optimizar su rendimiento.

**U**tilicé el conjunto de prueba para generar predicciones y evalué el modelo mediante la métrica de precisión utilizando accuracy\_score. Además, obtuve un informe de clasificación que proporcionó detalles sobre la precisión, lo que me permitió comprender mejor el rendimiento del modelo en términos de clasificación. Por último, analicé la matriz de confusión para obtener información sobre los errores de clasificación, mejorando su visualización al convertirla en un dataframe de pandas. Esto facilitó la interpretación de los resultados y permitió identificar áreas de mejora en el modelo.

# *Conclusiones*

**A**l principio les indique que el modelo de la base de datos seria generada al azar por la librería faker, y un dataframe llamado “dataframe.csv” alojado en el repositorio, este mismo proceso genero un sesgo homogéneo en los datos, y a pesar de generar daros ramdon el estos mismos estuvieron repartidos de una forma muy simétrica, casi de una forma exacta, esto parto al modelo a cometer errores casi en el 50% de sus predicciones. Pero una cosa me queda clara y es que el modelo puede ser entrenado perfectamente con datos reales, arrojando efectividades superiores al 90%, no es la primera vez que utilizo este modelo y puedo asegurar que junto con el árbol de decisiones son los mejores modelos para la clasificación.

**A**l utilizar la librería Faker para generar datos aleatorios, se ha observado un sesgo homogéneo en el conjunto, resultando en una distribución simétrica que afecta la capacidad del modelo para aprender patrones significativos, lo que se refleja en una precisión de predicción de aproximadamente 50.8%. Las métricas indican que la precisión para la clase 0 (no agresión) es del 53%, mientras que para la clase 1 (agresión) es del 49%, sugiriendo una tendencia del modelo a predecir una clase de forma desbalanceada. Además, el recall para la clase 1 es superior al 50%, lo que implica que el modelo logra identificar más de la mitad de los casos positivos; sin embargo, la f1-score moderada de 0.53 indica que la cantidad de falsos positivos y negativos está afectando su efectividad general.

**L**a matriz de confusión revela 569 falsos positivos y 415 falsos negativos, evidenciando un problema en la clasificación que podría abordarse mediante una selección más precisa de características o utilizando datos más variados y representativos. Aunque el modelo no ha mostrado eficacia con datos generados aleatoriamente, existe la oportunidad de entrenarlo con datos reales que reflejen la complejidad del fenómeno a analizar, lo que podría llevar a efectividades superiores al 90%, como ha sido la experiencia previa. Este contexto subraya la importancia de la calidad de los datos en la creación de modelos de aprendizaje automático y sugiere la necesidad de evaluar otros enfoques y técnicas que puedan mejorar el rendimiento del modelo.

# *Referencias*

* Scikit-learn. (n.d.). User Guide. Retrieved from https://scikit-learn.org/stable/user\_guide.html
* pandas. (n.d.). Documentation. Retrieved from https://pandas.pydata.org/docs/
* Zhang, Y., & Wang, T. (2020). Introduction to Machine Learning: A Python-based Approach. Springer.
* Raj, A. (2017). Practical Machine Learning with Python: A Comprehensive Guide to Building Real-World Machine Learning Solutions. Packt Publishing.
* Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.