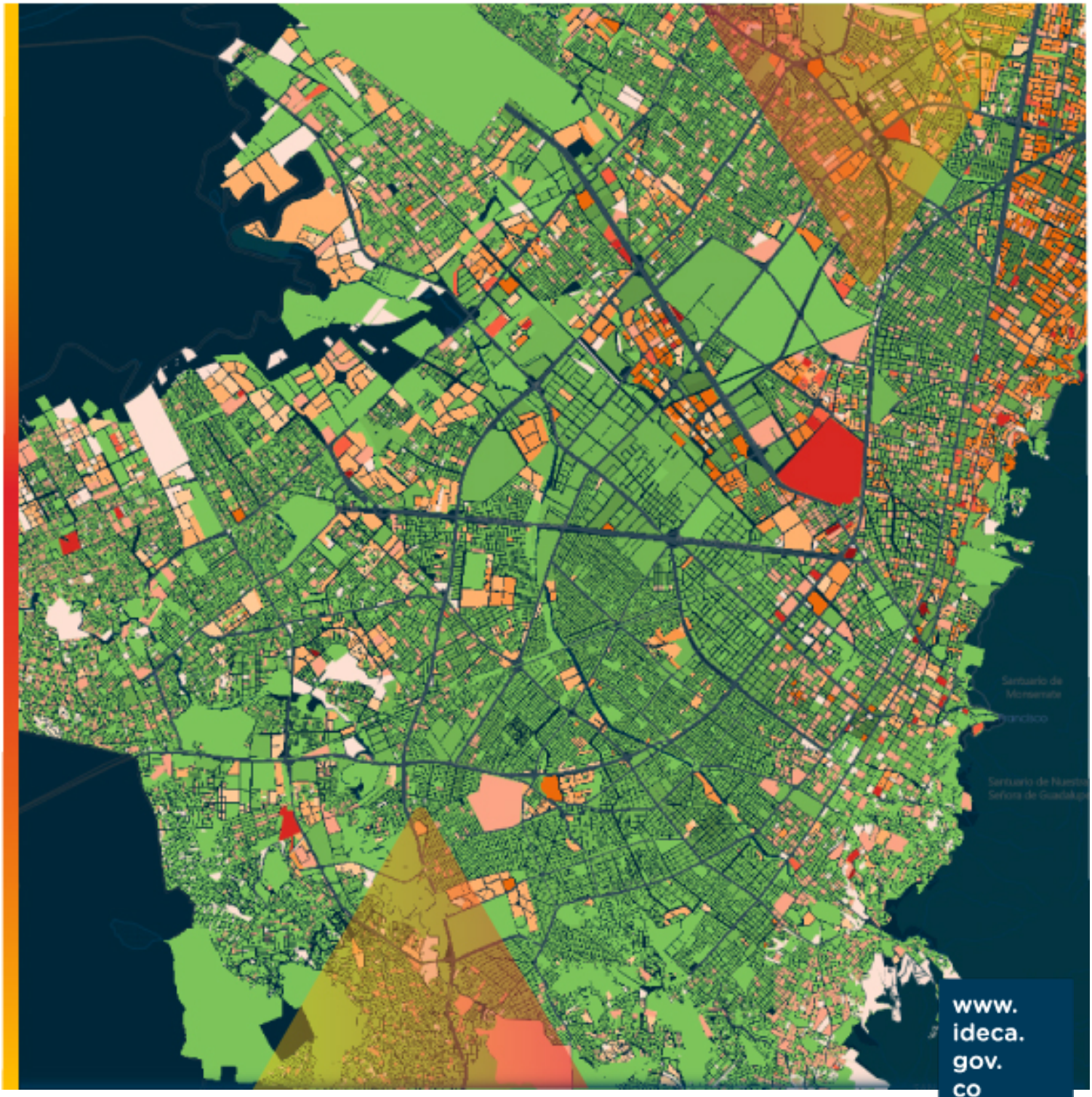
ideca



Caso de Uso Abuso y Violencia

—

Bogotá, D.C., 30/12/2025

ALCALDÍA MAYOR DE BOGOTÁ D.C.

ideca

ideca

Caso de Uso Abuso y Violencia



Gerente IDECA
Salomé Naranjo Luján

Subgerente de Analítica de Datos
Diego Ricardo Ibarra Rodríguez

Profesional Especializado
Subgerencia de Analítica de Datos
Martha Nubia Rojas Guerrero

Equipo de Trabajo

Contratista
Subgerencia de Analítica de Datos
Lilian Daniela Beltrán Barrera

Fecha de creación o actualización: 30/12/2025

Página web: www.ideca.gov.co

Correo electrónico: ideca@catastrobogota.gov.co

Licencia: Attribution 4.0 International (CC BY 4.0)

Control de Versiones

CAMBIOS

Fecha	Autor	Versión	Cambio efectuado
30/10/2025	Martha Rojas G.	1.0	Primera versión del documento. No hay cambios para registrar.
30/11/2025	Martha Rojas G. Lilian Daniela Beltrán B.	2.0	Se incluye todo lo relacionado los ensayos realizados mediante la aplicación de los cuatro modelos de ML.
30/12/2025	Martha Rojas G. Lilian Daniela Beltrán B.	3.0	Se incluye todo lo relacionado con el aplicativo de predicción.

REVISORES

Nombre	Dependencia
Diego Ricardo Ibarra Rodríguez	Subgerencia de Analítica de Datos

Tabla de Contenido

Tabla de Contenido ____ ***4***

Resumen ejecutivo ____ ***7***

Introducción ____ ***7***

Antecedentes ____ ***8***

Alcance ____ ***9***

Objetivos ____ ***9***

1. Inicio ____ **10**
1.2. Requerimientos ____ 10
1.3. Estudio de Mercado ____ 11

2. Idear ____ **13**
2.1. Proceso de Ideación ____ 13
2.2. Entendimiento de los Datos ____ 13
2.3. Ingeniería de Características ____ 18
2.4. Limpieza y Transformación ____ 19
2.5. Entorno de Desarrollo y Visualización. ____ 20
2.6. Requisitos ____ 20
2.7. Objetivos ____ 21
2.8. Propuesta ____ 21
2.9. Revisión con el cliente – Análisis descriptivo univariado de los datos ____ 22
2.10. Revisión con el cliente – Análisis descriptivo bivariado de los datos ____ 27
2.1. Diseño y Elección del Modelo ____ 32
2.2. Actividades ____ 40

3. Prototipar ____ **41**
3.1. Prototipo / Modelado: Aplicativo Abuso y Violencia ____ 41
3.2. Testeo 0.1: Escenario Base ____ 44
3.3. Testeo 0.2: Escenario Hiperparámetros ____ 44
3.4. Testeo 0.3: Escenario Balanceo de Clases ____ 44
3.5. Revisión Cliente ____ 45

4. Operacionalizar – Industrializar ____ **45**
4.1. Transferencia ____ 55
4.2. Retrospectiva ____ 56
4.3. Finalización ____ 56

5. Resultados ____ **57**

6. Conclusiones 58
7. Bibliografía 59

Lista de tablas

Tabla 1. Variables Módulo Abuso y Violencia – Histórico- 15
Tabla 2. Variables Módulo Abuso y Violencia – 2025- 17
Tabla 3. Variables Explicativas y Variables Respuesta empleadas en el modelo. 21
Tabla 4. Variables Explicativas y Variables Respuesta 27
Tabla 5. Cronograma de Actividades 40

Lista de gráficas

Gráfica 1. Estado del arte consultado. 12
Gráfica 2. Portal de Datos del Obce 14
Gráfica 3. Datos del Obce objeto de análisis 14
Gráfica 4. Datos de Abuso y Violencia 2014 - 2024 14
Gráfica 5. Datos de Abuso y Violencia 2025 y Posterior 15
Gráfica 6. Análisis inicial de las variables 22
Gráfica 7. Variables explicativas 23
Gráfica 8. Análisis inicial de las variables (por jornada, localidad, tipo colegio, grado) . 24
Gráfica 9. Análisis inicial – Variable explicativa Edad 25
Gráfica 10. Análisis Variables Respuesta 26
Gráfica 11. Análisis del tipo de violencia respecto a la jornada 27
Gráfica 12. Análisis Tipo de Violencia respecto a los Rangos de Edad 28
Gráfica 13. Análisis Tipo de Violencia respecto al Nivel Académico 29
Gráfica 14. Análisis Tipo de Violencia respecto a la Localidad 30
Gráfica 15. Análisis Tipo de Violencia respecto al Género 31
Gráfica 16. Experimentos realizados 33
Gráfica 17. Datos de las violencias desbalanceados 33
Gráfica 18. Mejor modelo "XGBOOST" 34
Gráfica 19. Ajuste de hiperparámetros al mejor modelo XGBOOST 35
Gráfica 20. Sobremustreo "Técnica Smote" 36
Gráfica 21. Submuestreo 37
Gráfica 22. XGBOOST con el Modelo Base 38
Gráfica 23. XGBOOST con el Modelo Over-Under 38
Gráfica 24. XGBOOST con mejor Accuracy y f1-score 39
Gráfica 25. Diseño de Prototipo Entregable 42
Gráfica 26. Diseño de Prototipo Entregable 43
Gráfica 27. Tablero de predicción para Abuso y Violencia 46

Gráfica 28. Tablero de predicción para Abuso y Violencia – Módulo carga de datos ... 47
Gráfica 29. Módulo carga de datos – Ejemplo datos cargados ... 48
Gráfica 30. Tablero de predicción para Abuso y Violencia – Módulo análisis descriptivo ... 49
Gráfica 31. Tablero de predicción para Abuso y Violencia – Distribución de variables categóricas ... 50
Gráfica 32. Tablero de predicción para Abuso y Violencia – Distribución por Localidad 51
Gráfica 33. Tablero de predicción para Abuso y Violencia – Tabla de reportes ... 52
Gráfica 34. Tablero de predicción para Abuso y Violencia – Módulo resultados de predicción ... 52
Gráfica 35. Distribución de la predicción tipo violencia ... 53
Gráfica 36. Distribución de la predicción tipo agresor ... 53
Gráfica 37. Tablero de predicción para Abuso y Violencia – Despliegue en Azure - ... 54
Gráfica 38. Módulos Tablero de predicción para Abuso y Violencia – Despliegue en Azure - ... 55
Gráfica 39. Módulos Tablero de predicción para Abuso y Violencia – Transferencia del conocimiento - ... 56

Resumen ejecutivo

El caso de uso "Abuso y Violencia", desarrollado de manera conjunta con la Secretaría de Educación, tuvo como objetivo ajustar un modelo predictivo orientado a determinar el tipo de violencia y el tipo de agresor más probables a los que un estudiante podría ser susceptible, con base en variables como Localidad, Género, Nivel Académico, Edad y Jornada.

Se realizaron diversas pruebas de ajuste utilizando metodologías como XGBoost, Random Forest, MLP y Regresión Logística, siendo XGBoost la técnica que presentó los mejores resultados en términos de desempeño predictivo.

Adicionalmente, se diseñó un aplicativo en Streamlit que permite a los usuarios finales generar predicciones a partir de nuevos conjuntos de datos de manera sencilla e interactiva. Este aplicativo facilita la interpretación de los resultados, agiliza el proceso de consulta y hace más accesible la herramienta para el personal de la Secretaría, contribuyendo así a la toma de decisiones informadas y oportunas.

Introducción

El proyecto surge de la necesidad de avanzar más allá de lo que la Secretaría de Educación Distrital ha realizado en los últimos años. Esta entidad ha recopilado información detallada sobre los casos de abuso y violencia ocurridos en los colegios de Bogotá, identificando variables relevantes como el tipo de agresor, el tipo de violencia, la edad y el género del agresor, así como diversas características sociodemográficas de la víctima, entre ellas localidad, género, edad, nivel académico, jornada, entre otras.

A partir de esta información, la SED desarrolló análisis descriptivos robustos que hoy permiten a los usuarios identificar las problemáticas que afectan con mayor frecuencia a las víctimas de abuso y violencia. Sin embargo, con el propósito de aprovechar aún más el valor de los datos, surgió la iniciativa de ir más allá del enfoque estrictamente descriptivo. El objetivo fue desarrollar una solución que no solo describiera, sino que también **predijera** qué niños podrían ser más propensos a experimentar eventos de abuso o violencia. Esto permitiría tomar decisiones preventivas y generar intervenciones oportunas por parte de los profesionales encargados.

Antecedentes

Durante los últimos años, la Secretaría de Educación Distrital (SED) ha consolidado un esfuerzo sistemático para recopilar, organizar y centralizar la información relacionada con los casos de abuso y violencia reportados en las instituciones educativas de Bogotá. Esta labor ha permitido construir una base de datos amplia y detallada, que incluye variables relacionadas con el tipo de agresión, las características del agresor, las condiciones sociodemográficas de las víctimas y otros factores relevantes para la comprensión de estas problemáticas.
A partir de esta información, la SED por medio del Observatorio de Convivencia Escolar (Obce) ha desarrollado análisis descriptivos que facilitan la identificación de patrones, perfiles y tendencias en los eventos de violencia escolar. Estos análisis han sido útiles para visibilizar las dinámicas del fenómeno y orientar algunas estrategias de intervención. Sin embargo, su alcance se ha centrado principalmente en describir lo ocurrido, lo cual limita la capacidad de anticiparse a situaciones de riesgo y de actuar de manera preventiva.

Ante esta necesidad, surge el interés de avanzar hacia un enfoque analítico predictivo que permita aprovechar integralmente el potencial de los datos. La posibilidad de identificar, a través de modelos estadísticos y de aprendizaje automático, qué estudiantes podrían estar en mayor riesgo de experimentar un evento de abuso o violencia, abre la puerta a intervenciones más oportunas y efectivas.
En este contexto, se plantea el desarrollo de una herramienta tecnológica que integre análisis descriptivos, modelos predictivos y visualizaciones interactivas, facilitando la toma de decisiones basada en datos por parte de profesionales y equipos técnicos. Esta iniciativa responde a la búsqueda de fortalecer las capacidades institucionales para la prevención y atención temprana de la violencia escolar, utilizando métodos modernos de ciencia de datos y una plataforma accesible para distintos niveles de usuario.

Alcance

Con la aplicación de la Metodología de Analítica de Datos al Caso de Uso "Abuso y Violencia" se desarrolló el piloto de un aplicativo en Streatmlit con la capacidad de generar predicciones a partir de nuevos conjuntos de datos de manera sencilla e interactiva. Este aplicativo facilita la interpretación de los resultados, agiliza el proceso de consulta y hace más accesible la herramienta para el personal de la Secretaría, contribuyendo así a la toma de decisiones informadas y oportunas.

Objetivos

Objetivo General

Ajustar un modelo predictivo orientado a determinar el tipo de violencia y el tipo de agresor más probables a los que un estudiante podría ser susceptible, con base en variables como Localidad, Género, Nivel Académico, Edad y Jornada.

Objetivos Específicos

Entender las fuentes de datos suministradas por el Observatorio de Convivencia Escolar - Obce (adscrito a la Secretaría de Educación – SED) a través del sitio público del Obce.

Realizar diversas pruebas (experimentos) con la aplicación de cuatro modelos de Machine Learning: XGBoost, Random Forest, MLP y Regresión Logística.

Diseñar un aplicativo que permita a los usuarios finales (Docentes) generar predicciones de posibles violencias que puede llegar a tener un estudiante.

1. Inicio

El Observatorio de Convivencia Escolar (Obce) adscrito a la Secretaria de Educación (SED) manifestó la necesidad de apoyo en la Estrategia de Artificación para el uso reflexivo de los datos.

1.2. Requerimientos

La Secretaría de Educación tiene un interés en la compresión de los entornos educativos lo cual implica una aproximación a las particularidades territoriales y su incidencia en la prevalencia de violencias hacia NNJ. A partir de ello, el OBCE en el material Bogodatos para la Convivencia escolar, ha avanzado en un análisis territorial de las situaciones que afectan la convivencia escolar (módulos del sistema de alertas), que incluye el cálculo de la tasa por cada 1.000 estudiantes en cada una de las localidades de Bogotá, y la elaboración de unos mapas de calor que visualizan los hallazgos. No obstante, este análisis se ha desarrollado exclusivamente con los reportes del sistema de alertas y no ha incluido variables geográficas diversas, ni tampoco se ha desarrollado análisis estadísticos predictivos y/o correlacionales con características espaciales de la ciudad.

El conocimiento de las particularidades territoriales y su incidencia en la generación de violencias podría tener un efecto positivo en la aproximación de las comunidades educativas a las problemáticas de convivencia escolar. Esto, debido a que proporciona datos situados que pueden impactar no solo en su visibilización, sino en la creación de espacios de reflexión pedagógica y construcción de conocimiento para la transformación desde dichas comunidades.

A esto se suma, los necesidades y retos identificados en los ejercicios de difusión y apropiación social del material con las comunidades educativas:

(i) uso de estrategias pedagógicas diferenciadas por cursos de vida y para diversos actores de la comunidad educativa, en el abordaje de los datos;

(ii) fomento de la construcción de conocimiento que contribuya a la prevención de violencias y situaciones que afectan la salud mental y el bienestar de las y los estudiantes.

(iii) Lo anterior plantea desafíos concernientes a quién, cómo y para qué generar conocimiento sobre datos de convivencia escolar, los cuales se pretenden resolver en la investigación de Artificación para el uso reflexivo de los datos.

1.3. Estudio de Mercado

Una vez identificada la necesidad presentada por el Obce, se procedió a realizar consultas referentes a estudios que contemplaran el análisis de temas referentes a abuso y violencia en población escolar, y que a la vez hubieran empleado técnicas de analítica de datos a manera de predicción como se describe a continuación.

En primera instancia se consultó el portal del Observatorio de Convivencia Escolar (Obce)[1] donde se encontaron boletines de información periódica mensual relacionada con los reportes realizados al sistema de alertas **(Gráfica 1)**; así como, entender los seis módulos a través de los cuales se encuentran clasificadas las situaciones de vulneración en la población escolar, a saber:

1. Abuso y violencia.
2. Accidentalidad escolar.
3. Conducta suicida.
4. Consumo de sustancias psicoactivas.
5. Maternidad y paternidad tempranas.
6. Trastornos de aprendizaje y del comportamiento.

En segunda instancia, y en el marco de estudio del arte se consultó el documento "Modelos de probabilidad de vulneración para la prevención de violencia contra niñas, niños y adolescentes en Colombia" (ICBF, 2022)

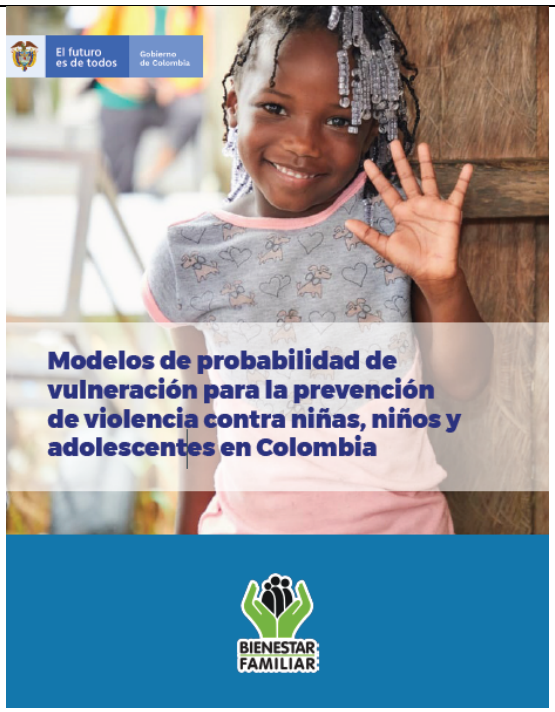
[1] Obce consultado en línea en mayo 2025: https://www.educacionbogota.edu.co/portal_institucional/gestion-educativa/observatorio-de-convivencia-escolar

(Gráfica 1), que consigna: i) Definiciones sobre los tipos o formas de violencia; ii) Indicadores de violencia contra niñas, niños y adolescentes; iii) Características de los datos empleados: población infantil, hogar, vivienda, municipio; iv) Descripción de los modelos probabilísticos empleados: Random Forest y Gradient Boosted Tees.

Por lo anterior, el objetivo del boletín del ICBF fue brindar insumos que ayuden a prevenir las violencias, pues los modelos desarrollados permiten estimar la probabilidad de que una niña, niño o adolescente registrado en la base de datos del Sisbén IV pueda ser identificado como una posible víctima de violencia física, psicológica y negligencia (maltrato)[2].

G1 *Gráfica 1. Estado del arte consultado.*
Fuente: Enlaces a los portales Obce e ICBF.





[2] Modelos de probabilidad de vulneración para la prevención de violencia contra niñas, niños y adolescentes en Colombia. ICBF. 2022. Consultado en línea en julio 2025: https://www.icbf.gov.co/system/files/modelos_de_probabilidad_de_vulneracion_para_la_prevencion_de_violencia_contra_ninas_ninos_y_adolescentes_en_colombia.pdf

2. Idear

En este capítulo se recopilan las ideas y o necesidades expresadas por el Equipo de Profesionales del Observatorio de Convivencia Escolar (Obce), en el marco del Módulo de Abuso y Violencia, así como posibles estrategias de solución.

2.1. Proceso de Ideación

El módulo de abuso y violencias comprende el reporte de seis tipos de violencias:

-Física

-Sexual

-Psicológica

-Negligencia

-Abandono

-Económica.

Idea No.1: Abordar la conducta suicida como problemática prioritaria para la Secretaría de Educación Distrital (SED) en relación con la garantía de derechos de los niñas, niños y jóvenes (NNJ). No obstante, tras revisar las investigaciones que actualmente se adelantan en colaboración con otras entidades y universidades, se identifica una alta concentración de esfuerzos en esta área.

No obstante, se puede explorar un análisis de las particularidades territoriales asociadas a la prevalencia de violencias hacia NNJ.

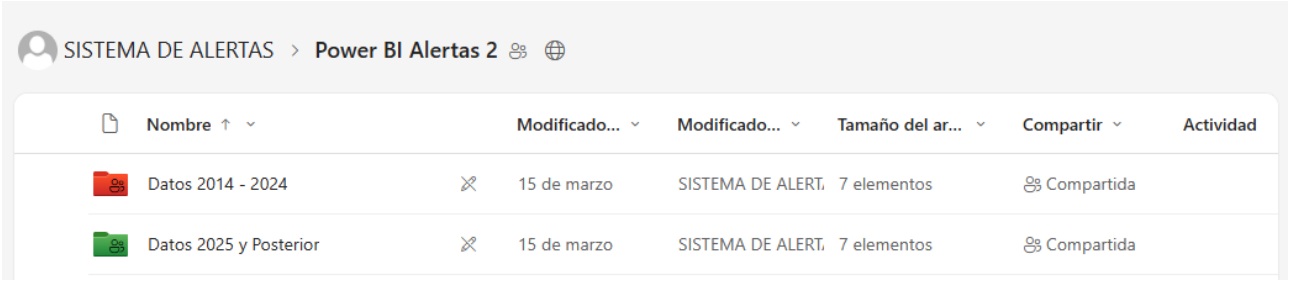
2.2. Entendimiento de los Datos

De acuerdo con lo manifestado por el Obce, se accedió al sitio: https://oape.educacionbogota.edu.co/ **(Gráfica 2)**, donde se encuentra disponible una base de datos abierta que puede ser consultada a través de Power BI y a la vez descargada.

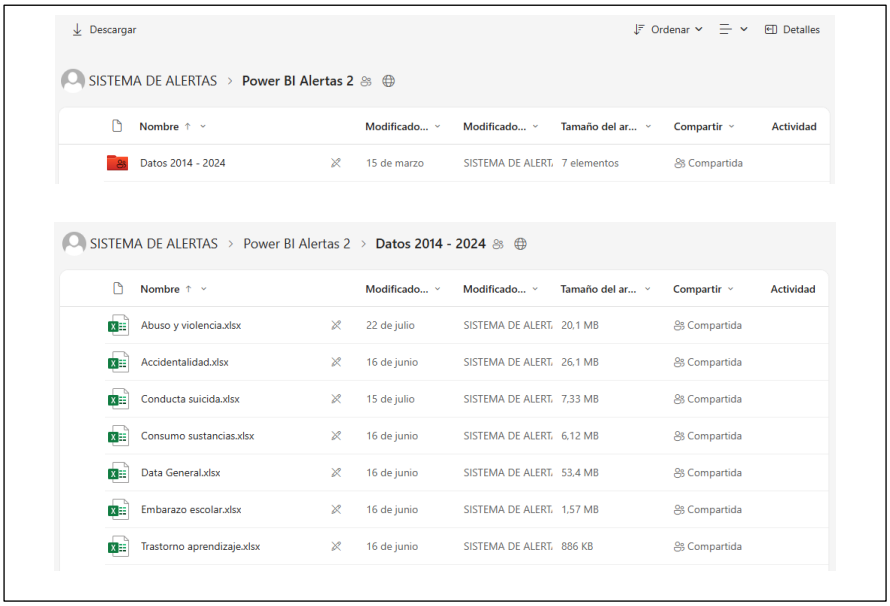
G2 *Gráfica 2. Portal de Datos del Obce*
Fuente: https://oape.educacionbogota.edu.co/



G3 *Gráfica 3. Datos del Obce objeto de análisis*
Fuente: https://oape.educacionbogota.edu.co/



G4 *Gráfica 4. Datos de Abuso y Violencia 2014 - 2024*
Fuente: https://oape.educacionbogota.edu.co/



G5 *Gráfica 5. Datos de Abuso y Violencia 2025 y Posterior*
Fuente: https://oape.educacionbogota.edu.co/



T1 Tabla 1. Variables Módulo Abuso y Violencia – Histórico-
Fuente: Elaboración propia adaptada de la Plataforma Obce

Variable	Descripción de la Variable	Necesario en el procesamiento SI/NO
Lugar de ocurrencia	Lugar dónde ocurrió el evento de abuso y/o violencia.	NO
Tipo de violencia	Tipo de violencia asociada al evento: Abandono, Económica, Física, Negligencia, Psicológica o Sexual.	SI
Tipo de agresor	Tipo de agresor asociado al evento: Administrativo, Escolar, Externos o Familiar	SI
Relación con el agresor	Relación con el individuo asociado al evento de abuso y/o violencia.	NO
Edad del agresor	Edad del agresor asociado al evento	NO
Género del agresor	Género del agresor asociado al evento	NO
Jornada	Jornada escolar de la víctima de abuso y/o violencia.	SI
Género	Género de la víctima de abuso y/o violencia	SI

Fecha_Nac	Fecha de nacimiento de la víctima de abuso y/o violencia.	NO
Edad	Edad de la víctima de abuso y/o violencia	NO
NivelAcad	Nivel Académico de la víctima de abuso y/o violencia	SI
Clasificación edad	Clasificación edad víctima evento el evento de abuso y/o violencia. Adolescencia, Infancia, Primera infancia o Adultez	SI
Localidad	Localidad en la que se presentó el evento de abuso y/o violencia.	SI
Trimestre	Trimestre en el que ocurrió el evento de abuso y/o violencia	NO
Semestre	Semestre en el que ocurrió el evento de abuso y/o violencia	NO
año	Año en el que ocurrió el evento de abuso y/o violencia	NO
País	País de residencia de la víctima	NO
Víctima del conflicto armado	Determina si el individuo fue alguna vez víctima del conflicto armado: SI o NO	NO
Discapacidad	Determina si la víctima presenta alguna discapacidad : SI o NO	NO
Etnia	Etnia asociada a la víctima del envento de abuso y/o violencia	NO
Tipo de colegio	Tipo de colegio al que pertenecía la víctima: Privado , Público, entre otros	NO
Grado	Grado de escolaridad de la víctima	NO
UPZ	Unidad de Planeamiento Zonal de la víctima	NO
Hechos violentos	Determina si el individuo fue víctima de algún hecho violento previamente	NO
Violencia por discrimación / hostigamiento	Determina si el individuo fue víctima de discriminación u hostigamiento: SI o NO	NO
Tipo de hostigamiento	Determina el tipo de hostigamiento del cual fue víctima	NO
Fecha Caso	Fecha de ocurrencia del evento de abuso y/o violencia	NO
Lugar Real	Lugar de ocurrencia del evento de abuso y/o violencia	NO
Nom_Loca	Localidad en la que se presentó el evento de abuso y/o violencia.	NO

T2 Tabla 2. Variables Módulo Abuso y Violencia – 2025-
Fuente: Elaboración propia adaptada de la Plataforma Obce

Variable	Descripción de la Variable	Necesario en el procesamiento SI/NO
Lugar de ocurrencia	Lugar dónde ocurrió el evento de abuso y/o violencia.	NO
Tipo de violencia	Tipo de violencia asociada al evento: Abandono, Económica, Física, Negligencia, Psicológica o Sexual.	SI
Tipo de agresor	Tipo de agresor asociado al evento: Administrativo, Escolar, Externos o Familiar	SI
Relación con el agresor	Relación con el individuo asociado al evento de abuso y/o violencia.	NO
Edad del agresor	Edad del agresor asociado al evento	NO
Género del agresor	Género del agresor asociado al evento	NO
Jornada	Jornada escolar de la víctima de abuso y/o violencia.	SI
Género	Género de la víctima de abuso y/o violencia	SI
Fecha_Nac	Fecha de nacimiento de la víctima de abuso y/o violencia.	NO
Edad	Edad de la víctima de abuso y/o violencia	NO
NivelAcad	Nivel Académico de la víctima de abuso y/o violencia	SI
Clasificación edad	Clasificación edad víctima evento el evento de abuso y/o violencia. Adolescencia,Infancia, Primera infancia o Adultez	SI
Localidad	Localidad en la que se presentó el evento de abuso y/o violencia.	SI
Trimestre	Trimestre en el que ocurrió el evento de abuso y/o violencia	NO
Semestre	Semestre en el que ocurrió el evento de abuso y/o violencia	NO
año	Año en el que ocurrió el evento de abuso y/o violencia	NO
País	País de residencia de la víctima	NO
Víctima del conflicto armado	Determina si el individuo fue alguna vez víctima del conflicto armado: SI o NO	NO
Discapacidad	Determina si la víctima presenta alguna discapacidad : SI o NO	NO
Etnia	Etnia asociada a la víctima del envento de abuso y/o violencia	NO

Tipo de colegio	Tipo de colegio al que pertenecía la víctima: Privado , Público, entre otros	NO
Grado	Grado de escolaridad de la víctima	NO
UPZ	Unidad de Planeamiento Zonal de la víctima	NO
Hechos violentos	Determina si el individuo fue víctima de algún hecho violento previamente	NO
Violencia por discrimación / hostigamiento	Determina si el individuo fue víctima de discriminación u hostigamiento: SI o NO	NO
Tipo de hostigamiento	Determina el tipo de hostigamiento del cual fue víctima	NO
Fecha Caso	Fecha de ocurrencia del evento de abuso y/o violencia	NO
Lugar Real	Lugar de ocurrencia del evento de abuso y/o violencia	NO
Nom_Loca	Localidad en la que se presentó el evento de abuso y/o violencia.	NO
Víctima - Agresor	Determina si el individuo es víctima o es la persona agresora	NO

2.3.Ingeniería de Características

Durante el proceso de revisión de la data se identificaron los siguientes hallazgos:

Hallazgo No.1: Edades.

Durante la revisión de la información se encontraron registros de individuos con edades superiores a los 18 años, en algunos casos hasta de 80 años. Por otro lado, hay registros con edad asociada de 0 años. En estos casos, sería importante saber si deben mantenerse así o si podrían corresponder a inconsistencias en los datos.

A lo anterior, el Obce manifestó aclaración sobre los rangos de edad que la SED trabaja, así:

-Primera infancia 0 a 5 años
-Infancia 6 a 11 años
-Adolescencia 12 a 17 años
-Adultez + 18 años

Hallazgo No.2: Variables "Lugar de ocurrencia" y "Relación con el agresor".

Para estas variables se identificó una cantidad considerable de categorías asociadas al suceso. Sería útil definir categorías más claras o agrupaciones que faciliten su análisis y manejo. Para ello, se adjunta un archivo en Excel con el análisis detallado de cada variable y el porcentaje de participación de cada categoría respecto al total. Ahí podrán consultar las categorías presentes en cada caso.

Hallazgo No.3: Población de análisis.

Se observó que la base correspondiente al año 2025 incluye una variable llamada "¿Víctima – Agresor?" con tres categorías: Víctima, Agresor y Sin información. ¿Se debería considerar únicamente la categoría Víctima para el análisis?

Además, la base histórica no contiene este campo, lo que genera la duda de si en ese conjunto de datos solo se incluyen registros correspondientes a víctimas.

Hallazgo No.4: Nuevas variables:

Se identifica la necesidad de obtener variables adicionales a la información ya suministrada, específicamente relacionadas con la composición familiar del individuo. Idealmente, buscamos conocer más sobre su entorno familiar, incluyendo variables que indiquen con quién vive, si ambos padres trabajan, entre otros aspectos similares que serían útiles para el análisis.

2.4.Limpieza y Transformación

La Limpieza y estructuración de los datos consistió en la depuración de los registros y normalización de las variables de los datos descargados de: https://oape.educacionbogota.edu.co/, así:

En lo referente al Análisis de Completitud, se identificaron y seleccionaron las variables más relevantes.

Para la información de 2025 existía una variable llamada Víctima – Agresor, la cual presentaba 3 categorías: Víctima, Agresor y Sin Información – se tomaron solamente registros asociados a la categoría de víctimas.

Se corrigieron algunos temas de atipicidad en los datos correspondientes al atributo "edad", tomando registros entre los 3 y los 20 años.

Finalmente, una vez revisada toda la data descargada del sitio público, se resalta que la base contiene información correspondiente a los años 2014 hasta 2025; y luego de realizados los ajustes mencionados en la limpieza de la data, se tomó como ventana de trabajo los datos comprendidos desde 2019 hasta 2025, lo cual permitió tener una base de datos preliminar de 113375 registros.

Adicionalmente, se dispone de información sociodemográfica como edad, género, localidad, entre otras variables, así como de datos que caracterizan el suceso de abuso y violencia, tales como: **tipo de violencia, tipo de agresor, edad y género del agresor.**

2.5. Entorno de Desarrollo y Visualización.

El proyecto fue desarrollado utilizando un entorno tecnológico basado principalmente en Python para el procesamiento, análisis y modelado de los datos. Python permitió integrar herramientas robustas para la limpieza, transformación y preparación de la información, así como para la construcción y validación de modelos predictivos, empleando librerías especializadas como pandas, NumPy, scikit-learn y matplotlib.

Para facilitar la interacción con los resultados, se implementó una interfaz dinámica mediante Streamlit, la cual permitió transformar los análisis y modelos en una aplicación web intuitiva y accesible para los usuarios. A través de esta plataforma, los profesionales pueden visualizar indicadores descriptivos, explorar patrones de abuso y violencia, y consultar las predicciones generadas por el modelo, todo en tiempo real y sin necesidad de conocimientos técnicos avanzados.

2.6. Requisitos

-Bases de datos normalizadas

-Diccionarios de datos

-Lenguaje de programación en Phyton

-Librerías: pandas, NumPy, scikit-learn y matplotlib.

-Interfaz dinámica mediante Streamlit

2.7. Objetivos

- Ajustar un modelo predictivo orientado a determinar el tipo de violencia y el tipo de agresor más probables a los que un estudiante podría ser susceptible, con base en variables como Localidad, Género, Nivel Académico, Edad y Jornada.
- Poner a disposición la herramienta final, con el propósito de que cualquier profesional pueda analizar la información y obtener conclusiones relevantes que permitan tomar decisiones de manera oportuna.

2.8. Propuesta

Se propone desarrollar un modelo predictivo multietiqueta que, a partir de variables sociodemográficas (como edad, género, localidad y UPZ), permita predecir las características del posible suceso de abuso o violencia que podría experimentar un estudiante. Este modelo utilizaría como variables respuesta aspectos como el lugar de ocurrencia, el tipo de violencia, la relación con el agresor y los hechos violentos, con el fin de establecer asociaciones que permitan identificar perfiles de riesgo y apoyar la toma de decisiones en prevención y atención.

La propuesta de solución comienza con el análisis de los datos suministrados por el Obce a través del enlace público: https://oape.educacionbogota.edu.co/ con lo cual se seleccionaron las siguientes variables:

T3 Tabla 3. Variables Explicativas y Variables Respuesta empleadas en el modelo.
Fuente: Elaboración propia adaptada de los datos publicados en la Plataforma Obce

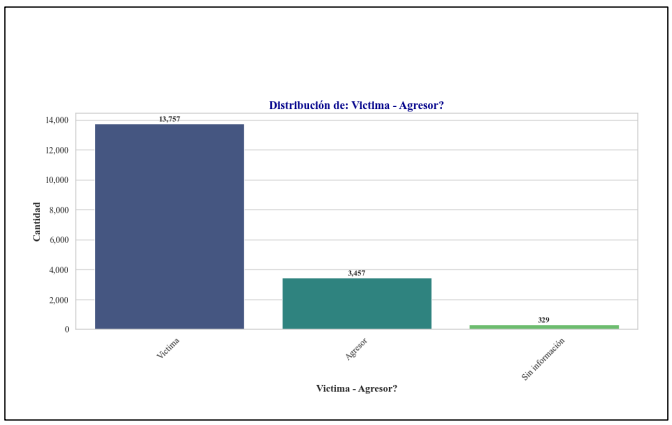
Variables Explicativas	Variables Respuesta
Localidad Género Nivel Académico Clasificación Edad Jornada	Tipo de violencia Tipo de agresor

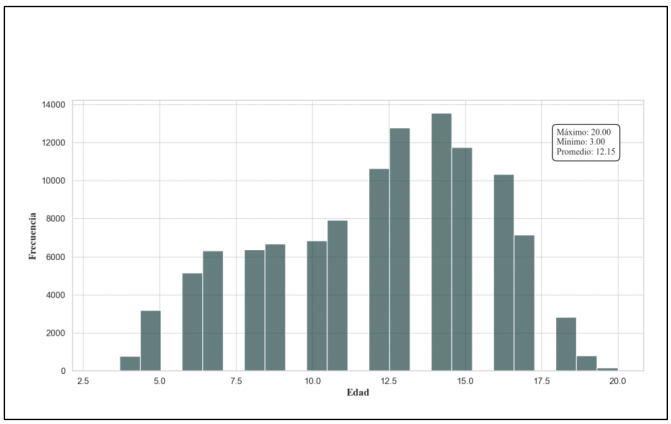
2.9. Revisión con el cliente – Análisis descriptivo univariado de los datos

La propuesta presentada se sustentó con base en los análisis descriptivos realizados a la data establecida para la ventana de tiempo 2019 hasta 2025.

G6 *Gráfica 6. Análisis inicial de las variables*

Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



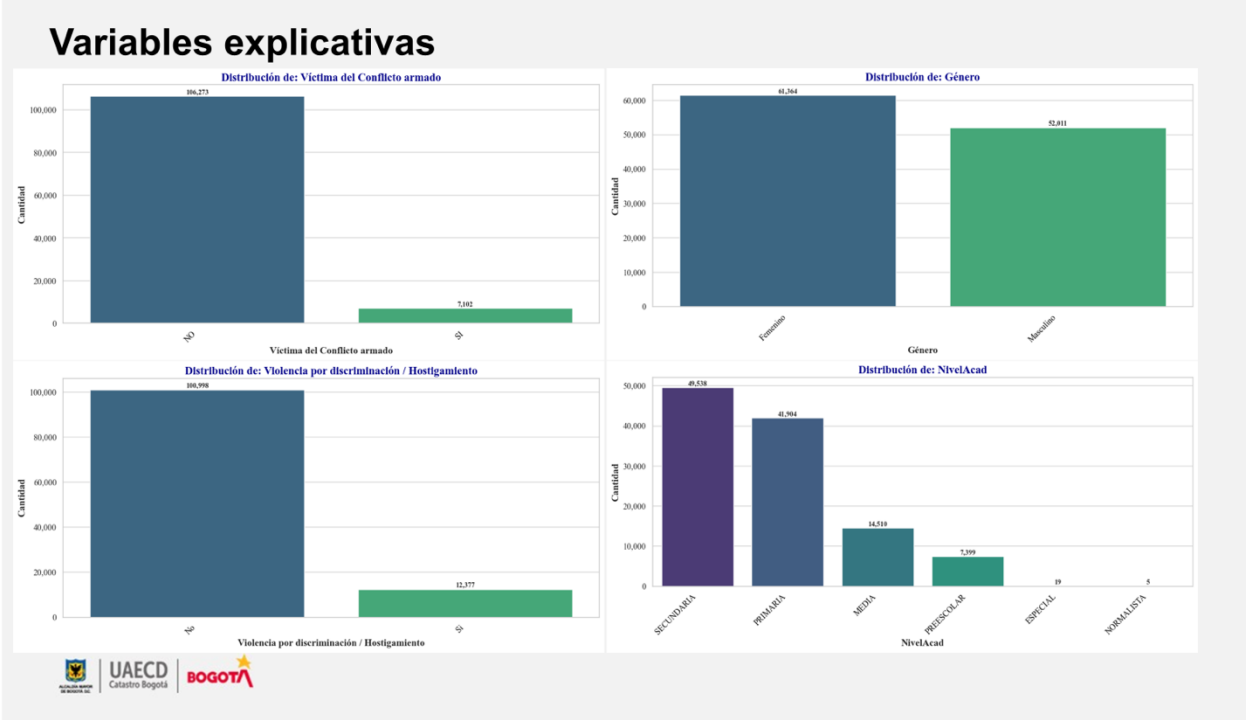


Adicionalmente, se tomaron edades entre los 3 y 20 años, omitiendo registros con edades atípicas. Una vez realizados estos ajustes se obtiene una base preliminar con un total de 113.375 registros.

Finalmente, se dispone de información sociodemográfica como edad, género, localidad, entre otras variables, así como de datos que caracterizan el suceso de abuso y violencia, tales como: tipo de violencia, tipo de agresor, edad y género del agresor.

G7 *Gráfica 7. Variables explicativas*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

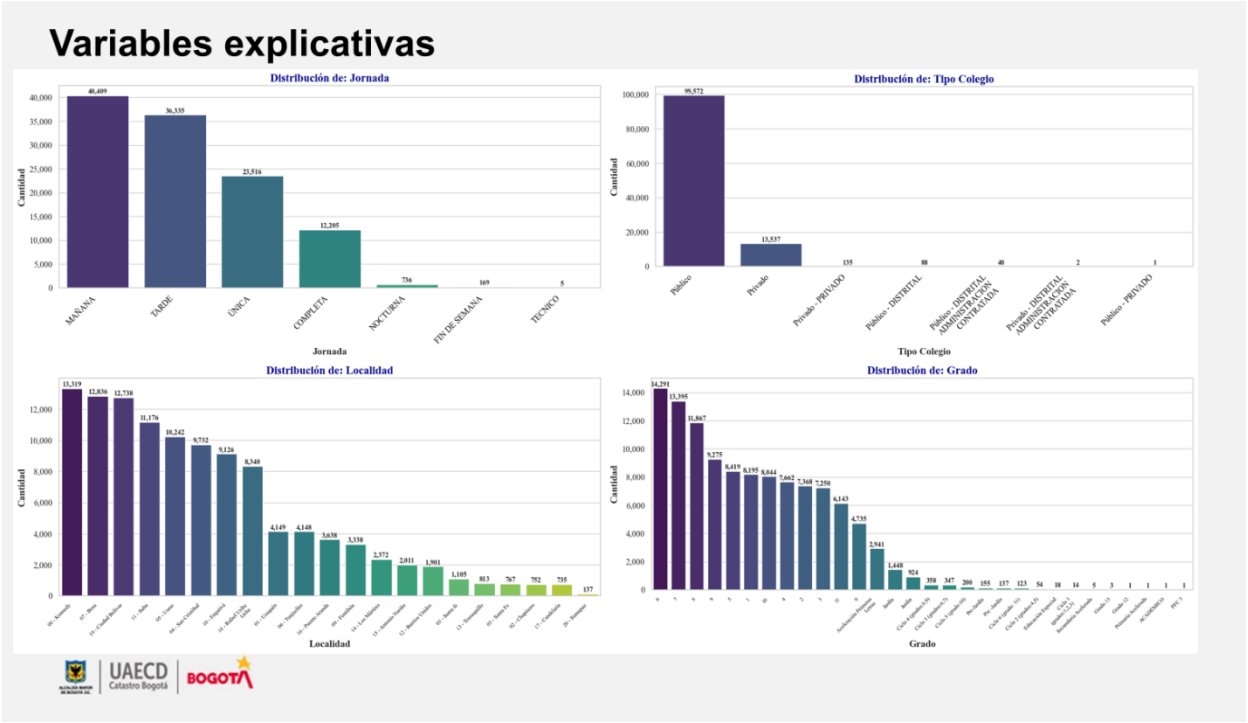
De acuerdo con las variables explicativas consideradas para el modelo, se realizaron los siguientes análisis:



En cuanto a la distribución de las Víctimas del Conflicto armado, se observa **(Gráfica 7)** que la mayoría no fueron víctimas del conflicto.

Respecto al tipo de violencia por discriminación/hostigamiento, se observa que la mayoría de los estudiantes no han sufrido este tipo de violencia.
Finalmente, en cuanto a la distribución por nivel académico, se observa que la mayor cantidad de eventos se presentan en Secundaria, seguido de Primaria.

G8 *Gráfica 8. Análisis inicial de las variables (por jornada, localidad, tipo colegio, grado)*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



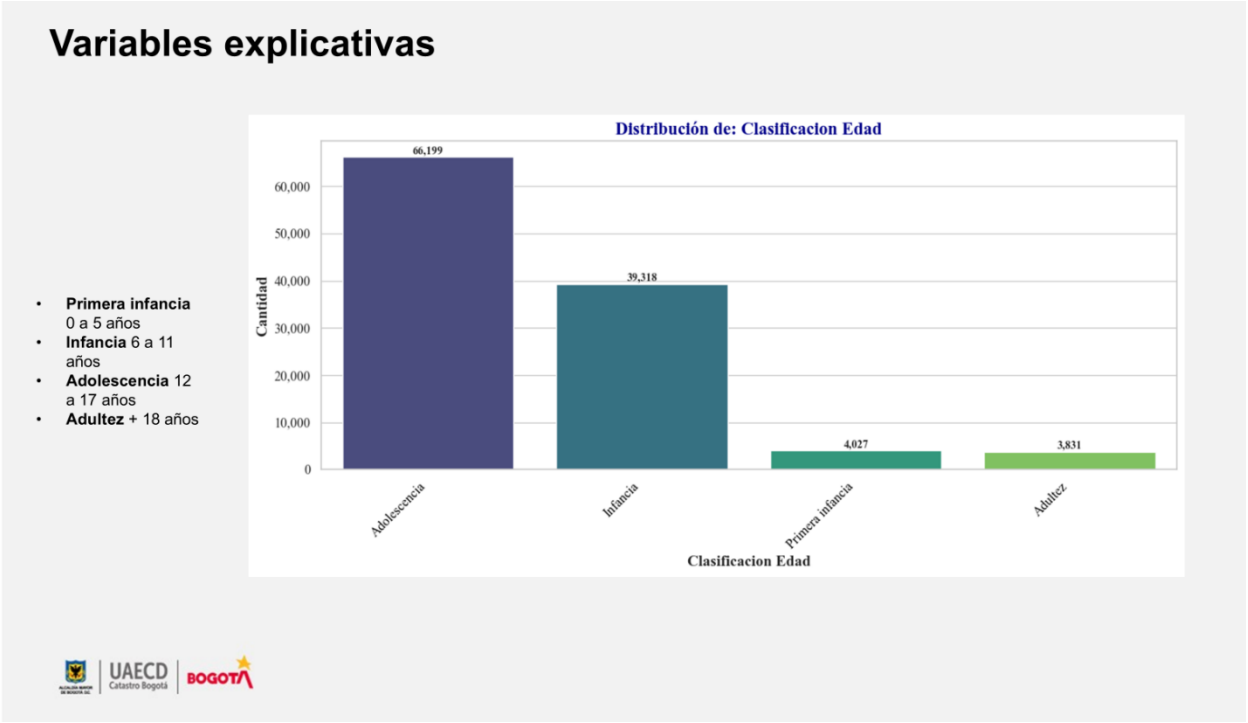
En cuanto a la variable explicativa "Jornada," la mayor cantidad de eventos se presentan en la jornada de la mañana, seguido de la tarde.

Para la variable "Tipo de colegio" se observa una diferencia considerable entre el Colegio Público y el Privado.

Respecto a la variable "Localidad", las localidades de Kennedy, Bosa, Ciudad Bolívar y Suba, retienen la mayor cantidad de casos.

Esta información de grado se concilia con la distribución "Nivel académico" **(Gráfica 8)**, porque hay correspondencia entre ambas.

G9 *Gráfica 9. Análisis inicial – Variable explicativa Edad*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

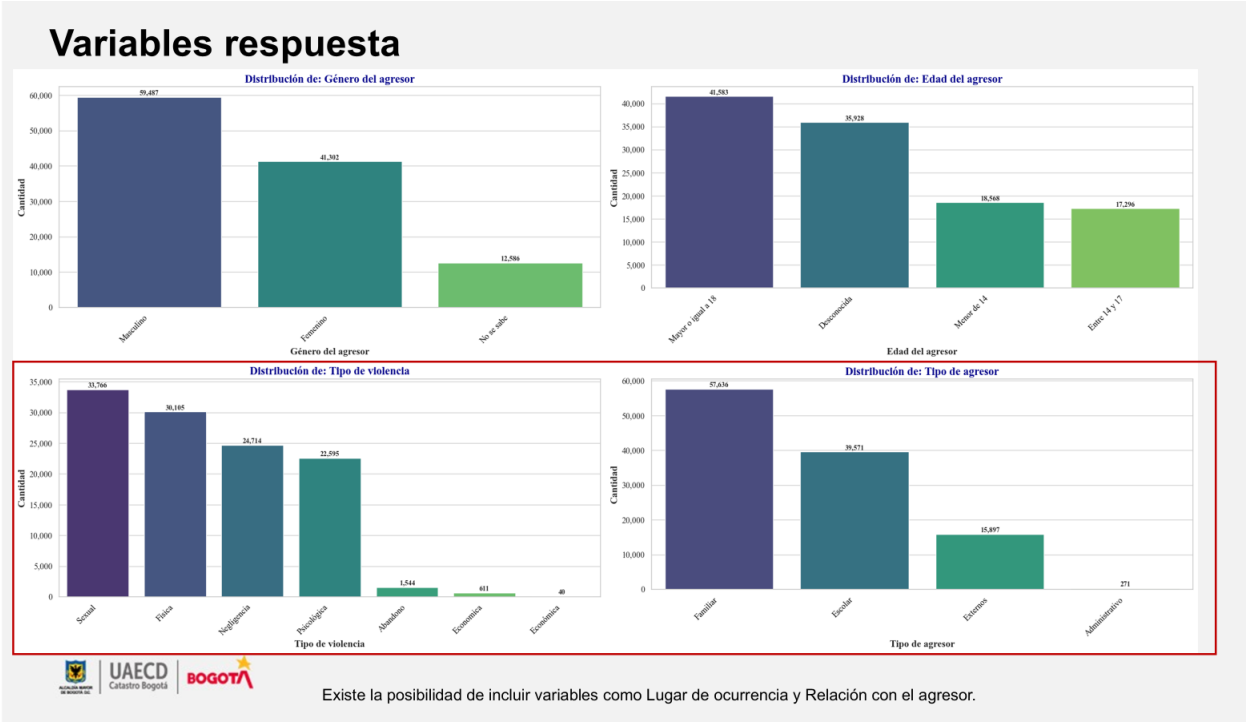


La variable edad se categorizó de acuerdo con las tipologías enviadas por la SED (Anexo 1), así:

-Primera infancia: 0 -5 años
-Infancia: 6 a 11 años
-Adolescencia: 12 a 17 años
-Adultez: Mayores a 18 años

Por lo anterior, se observa que, en las etapas de la vida, la mayor cantidad de eventos se presentan en la Adolescencia, seguido por la Infancia.

G10 *Gráfica 10. Análisis Variables Respuesta*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



Recordemos que las variables que se determinaron como "Variables respuesta" fueron:
-Tipo de violencia
-Tipo de agresor
-Relación con el agresor
-Lugar de ocurrencia

Las anteriores variables permitieron analizar temas como como el género del agresor y el tipo de agresor; sin embargo, se considera de mayor relevancia enfocarse en Tipo de violencia y en Tipo de agresor.
En cuanto a Tipo de violencia, se observa que la violencia sexual y la física son las que más se presentan, seguida por la "Violencia por Negligencia". Y en tipo de agresor, la mayoría de los casos se genera por un "Familiar".
Existen dos variables más que son el lugar de ocurrencia del evento y la relación con el agresor, estas variables tienen múltiples categorías y eso se comunicó en

reunión pasada a la Secretaría (Obce), y se llegó al acuerdo de colaborar en reducir las categorías para poderlas incluir dentro del análisis condensando más las categorías.

2.10. Revisión con el cliente – Análisis descriptivo bivariado de los datos

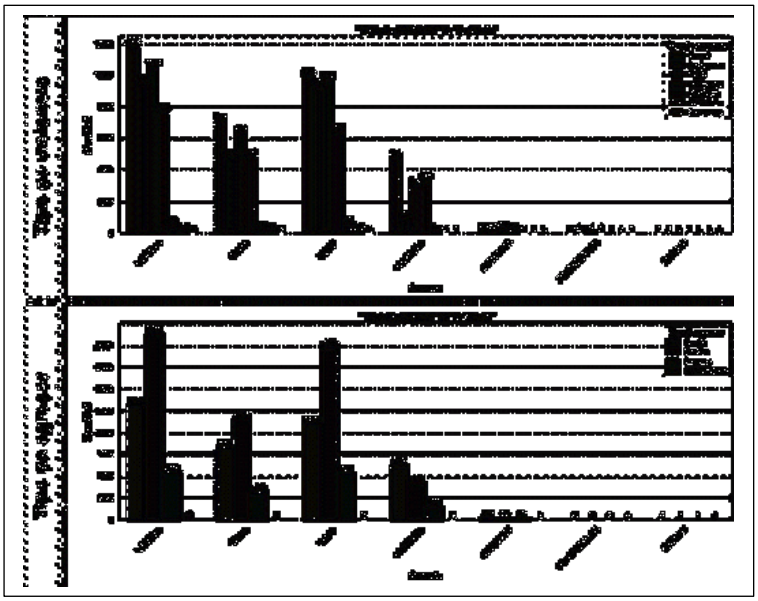
Adicionalmente, se realizó un análisis descriptivo Bivariado a la información, de acuerdo con las variables seleccionadas como respuesta y las sugeridas como explicativas (Tabla 4).

T4 Tabla 4. Variables Explicativas y Variables Respuesta
Fuente: Elaboración propia.

Variables Explicativas	Variables Respuesta
Localidad Género Nivel Académico Clasificación Edad Jornada	Tipo de violencia Tipo de agresor

Obteniendo los siguientes resultados de análisis:

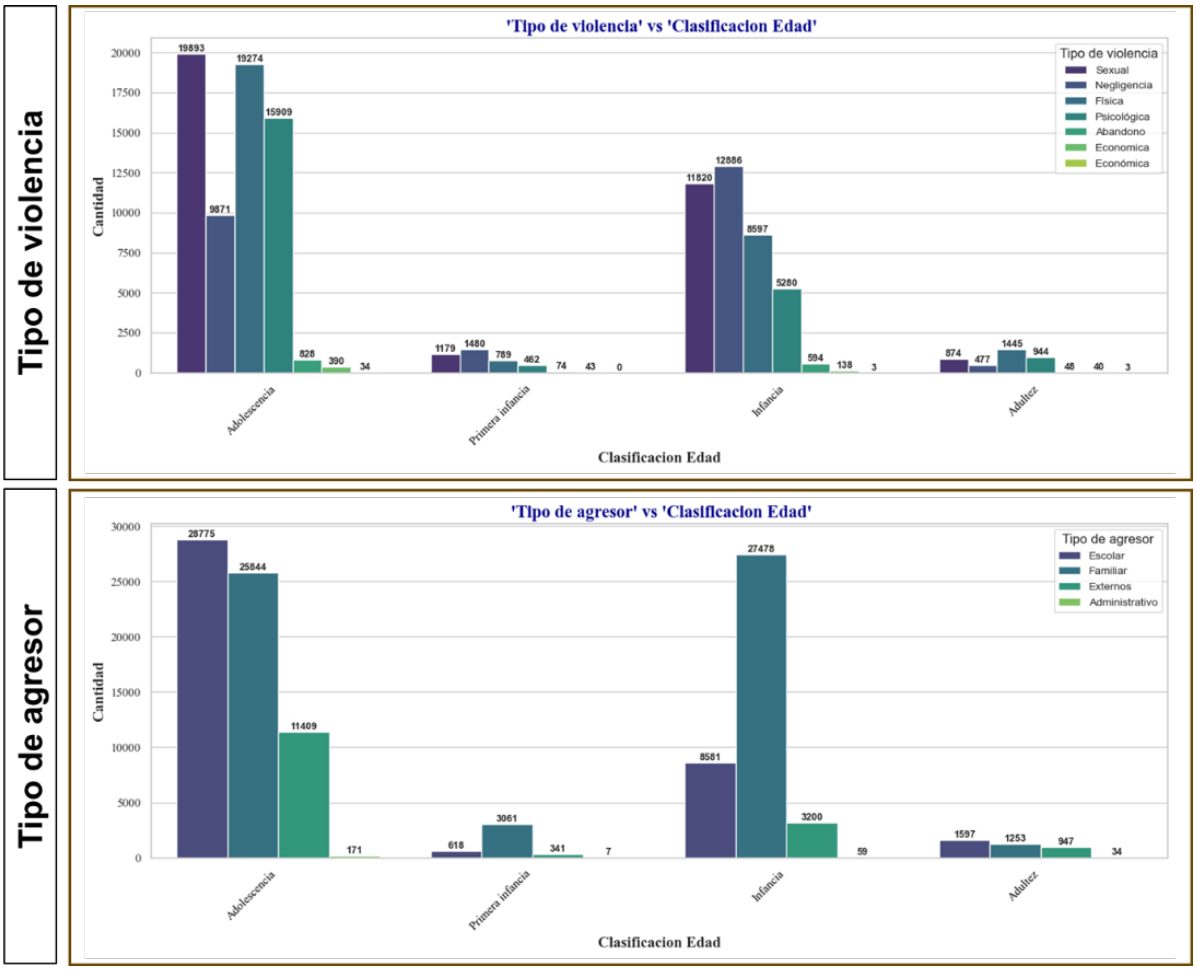
G11 *Gráfica 11. Análisis del tipo de violencia respecto a la jornada*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



Al analizar el tipo de violencia respecto de la jornada se observa que en todas las jornadas predomina el tipo de violencia sexual, seguida por la violencia física.

Y comparando el tipo de agresor con la jornada se observa que en todas las jornadas predomina el tipo de agresor "Familiar", a excepción de la jornada completa que el tipo de agresor es "Escolar".

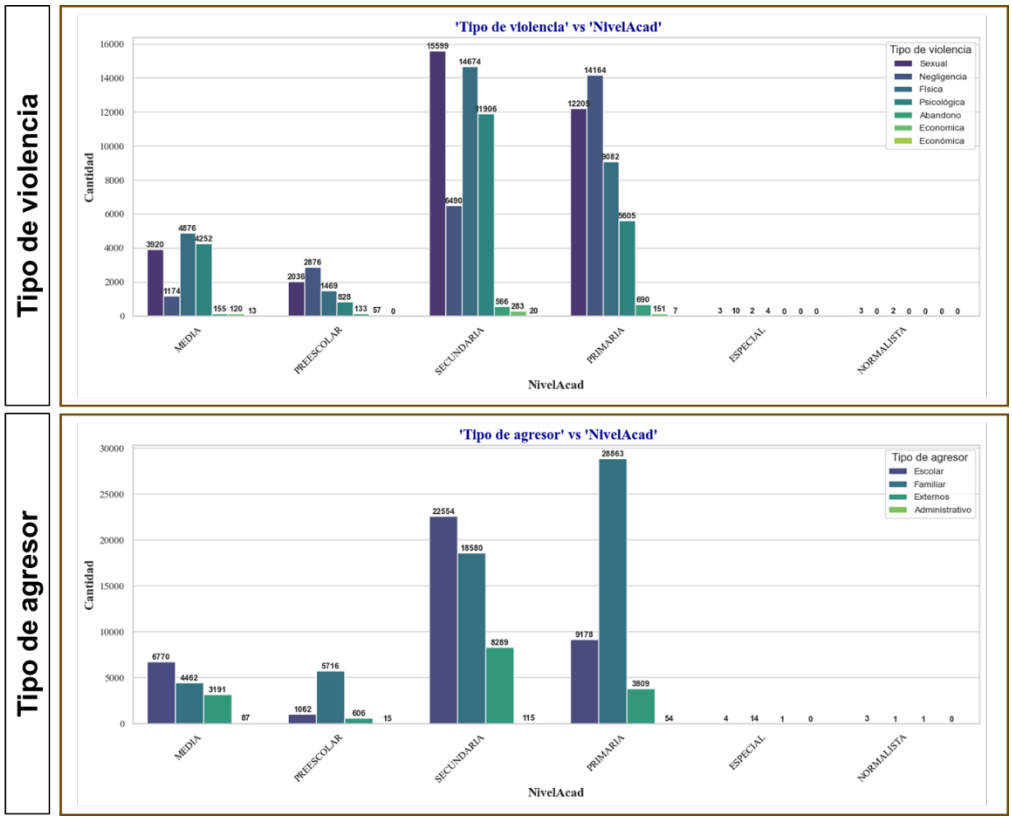
G12 *Gráfica 12. Análisis Tipo de Violencia respecto a los Rangos de Edad*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



En la **Gráfica 12**, se analiza el tipo de violencia respecto a la clasificación de la edad de los estudiantes, y se observa que en la adolescencia predomina la violencia sexual; en la infancia y primera infancia predomina la violencia por **"negligencia"**; y en la adultez predomina la violencia **"física"**.

Si se revisa por el tipo de agresor, esto es un poco correspondiente, es decir, en la adolescencia el tipo de agresor es mayoritariamente **"escolar"**; mientras que en primera infancia e infancia el agresor es **"familiar"**.

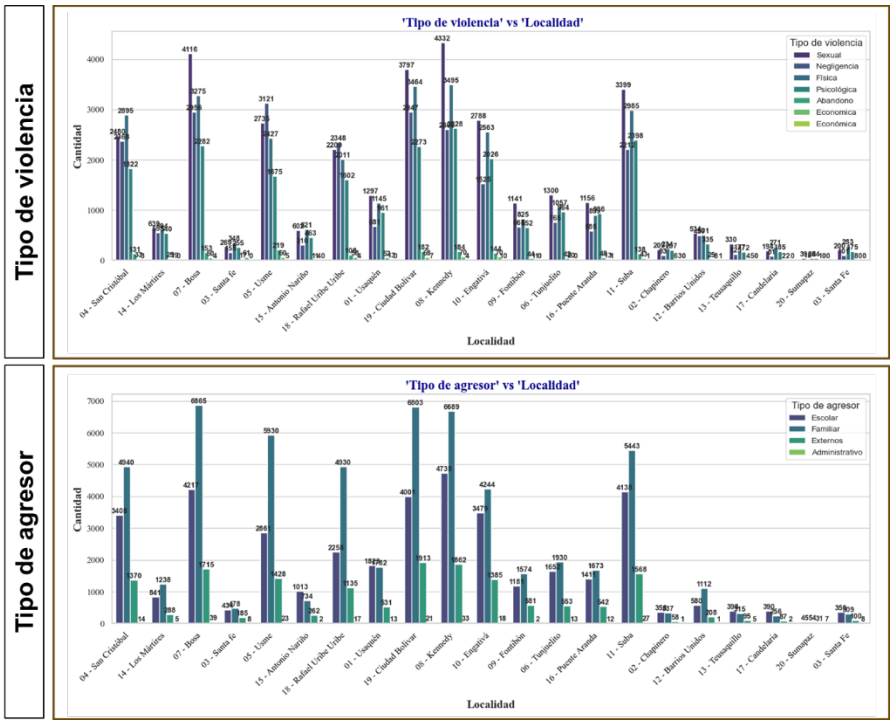
G13 *Gráfica 13. Análisis Tipo de Violencia respecto al Nivel Académico*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



En el análisis del **Tipo de violencia respecto al Nivel Académico**, se observa que en secundaria es donde se da una mayor cantidad de violencia "sexual", seguido por la violencia física; en cambio, en Primaria la violencia predominante es la "negligencia".

En el análisis de **Tipo de agresor Vs Nivel Académico**, se observa que en primaria el tipo de agresor tiende a ser un familiar; mientras que en secundaria el tipo de agresor es escolar.

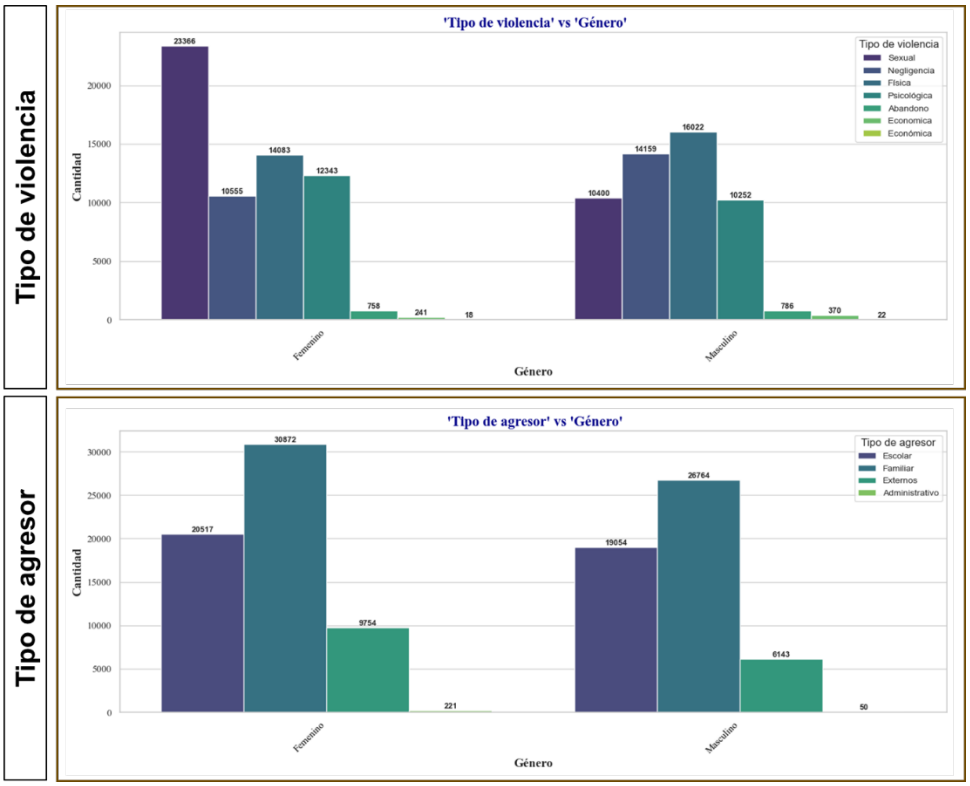
G14 *Gráfica 14. Análisis Tipo de Violencia respecto a la Localidad*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



En el análisis del Tipo de Violencia respecto a la Localidad, se observa que las localidades que agrupan la mayor cantidad de casos son: Bosa, Ciudad Bolívar, Kennedy y Suba; y el tipo de violencia predominante es "sexual", seguida de la violencia "física"; sin embargo, en San Cristóbal la violencia física es la predominante.

Al revisar por Tipo de Agresor, se observa que mayoritariamente corresponde a un "Familiar" en la mayor parte de las localidades.

G15 *Gráfica 15. Análisis Tipo de Violencia respecto al Género*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



Si se discrimina por género se observa **las mujeres** la "violencia sexual" es la predominante; mientras en **los hombres** la "violencia física" es mayoritaria.
Y en cuanto al Tipo de Agresor respecto al género, se observa que para los dos (femenino y masculino) el tipo de agresor predominante es el "familiar".

2.1. Diseño y Elección del Modelo

El procesamiento de la información y diseño del modelo se realizó a través del planteamiento de tres escenarios por medio de los cuales se probaron 9 experimentos aplicando cuatro modelos de Machine Learning (ML), como se describe a continuación:

Escenario Base:. En este escenario se hicieron cuatro experimentos, aplicando los siguientes cuatro modelos (un experimento por modelo), manteniendo los hiperparámetros por defecto y sin modificar los conjuntos de datos originales :

-Modelo Logístico: Un modelo estadístico sencillo que calcula la probabilidad de que algo ocurra (por ejemplo, "sí" o "no") usando relaciones lineales entre las variables.

-Modelo Random Forest: Un conjunto de muchos árboles de decisión que votan entre sí. Cada árbol aporta su opinión y el modelo elige la respuesta más consensuada.

-Modelo XGBOOST: Un método avanzado que combina muchos "pequeños árboles de decisión" para hacer predicciones muy precisas, aprendiendo de los errores que va cometiendo.

-Modelo MLP: Un tipo de red neuronal que imita cómo procesa información el cerebro: toma datos de entrada, los transforma a través de varias capas y produce una predicción.

Escenario Hiperparámetros: Basado en los resultados del escenario previo, se seleccionó la metodología con mejor rendimiento, una vez seleccionada se ajusta una grilla de hiperparámetros que permitan determinar si existe o no una mejora en el rendimiento del modelo. Se decidió solamente seleccionar una de las metodologías debido al costo computacional.

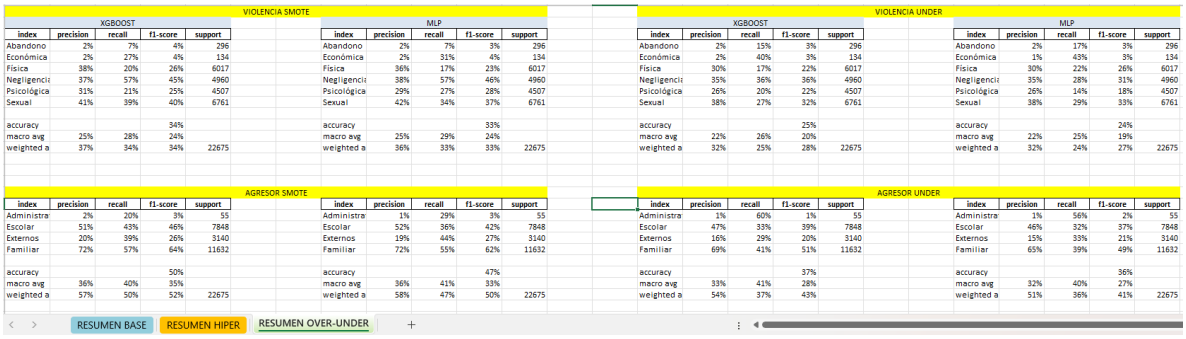
Escenario Balanceo de Clases (Over-Under): En este escenario se decidió tomar los dos modelos con mejor rendimiento en el escario base, siendo MLP y XGBoost; una vez seleccionados estos modelos se les aplicó Sobre Muestreo usando la técnica de Smote para un equivalente a dos experimentos más; y por otro lado, se les aplicó Submuestreo, con lo cual se obtuvieron dos experimentos adicionales.

Detalle de los experimentos realizados

En total se desarrollaron 9 experimentos aplicando cuatro modelos de Machine Learning (ML).

Los 9 experimentos desarrollados se materializan en este archivo de Excel **(Gráfica 16)** soportada con las siguientes tres hojas de cálculo:

G16 *Gráfica 16. Experimentos realizados*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



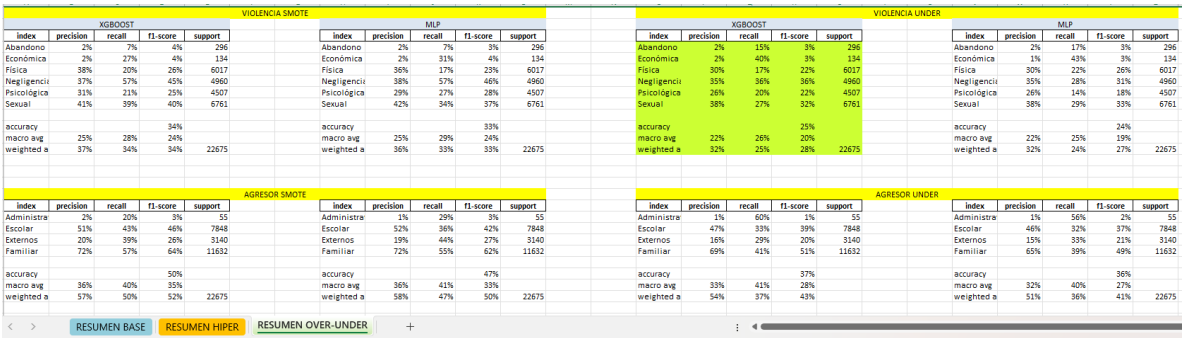
Hoja resumen base: En esta hoja se puede encontrar cada uno de los resultados obtenidos en cada una de las metodologías probadas: Regresión Logistica, XGBoost, Random Forest y MLP. Para cada una de las metodologías se ajustaron 2 modelos correspondiente a cada una de las variables respuesta analizadas.

G17 *Gráfica 17. Datos de las violencias desbalanceados*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

LOGISTIC				
index	**precision**	**recall**	**f1-score**	**support**
Abandono	0%	0%	0%	296
Económica	0%	0%	0%	134
Física	37%	33%	35%	6017
Negligencia	40%	47%	43%	4960
Psicológica	34%	5%	9%	4507
Sexual	39%	63%	48%	6761
accuracy			39%	
macro avg	25%	25%	22%	
weighted avg	37%	39%	35%	22675

Según la **Gráfica 17,** en la columna **support** se almacena la cantidad de individuos que hay en cada clase en el conjunto de prueba, las violencias de **Abandono** y **Económica** tiene menor cantidad de individuos a comparación de las violencias de negligencia, psicológica y sexual, por lo tanto, al entrenar un modelo de Machine Learning con este desbalance, el modelo presentará sesgo y muy probablemente no sea capaz de predecir las clases minoritarias, en la **Gráfica 17** se pueden observar las métricas para cada clase, en el caso concreto de Abandono y económica sus métricas son muy bajas o casi 0, precisamente por el problema de desbalanceo mencionado previamente.

G18 *Gráfica 18. Mejor modelo "XGBOOST"*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



VIOLENCIA SMOTE

XGBOOST

index	precision	recall	f1-score	support
Abandono	2%	7%	4%	296
Económica	2%	27%	4%	134
Física	38%	20%	26%	6017
Negligencia	37%	57%	45%	4960
Psicológica	31%	21%	25%	4507
Sexual	41%	39%	40%	6761
accuracy			34%	
macro avg	25%	28%	24%	
weighted a	37%	34%	34%	22675

MLP

index	precision	recall	f1-score	support
Abandono	2%	7%	3%	296
Económica	2%	31%	4%	134
Física	36%	17%	23%	6017
Negligencia	38%	57%	46%	4960
Psicológica	29%	27%	28%	4507
Sexual	42%	34%	37%	6761
accuracy			33%	
macro avg	25%	29%	24%	
weighted a	36%	33%	33%	22675

VIOLENCIA UNDER

XGBOOST

index	precision	recall	f1-score	support
Abandono	2%	15%	3%	296
Económica	2%	40%	3%	134
Física	30%	17%	22%	6017
Negligencia	35%	36%	36%	4960
Psicológica	26%	20%	22%	4507
Sexual	38%	27%	32%	6761
accuracy			25%	
macro avg	22%	26%	20%	
weighted a	32%	25%	28%	22675

MLP

index	precision	recall	f1-score	support
Abandono	2%	17%	3%	296
Económica	1%	43%	3%	134
Física	30%	22%	26%	6017
Negligencia	35%	28%	31%	4960
Psicológica	26%	14%	18%	4507
Sexual	38%	29%	33%	6761
accuracy			24%	
macro avg	22%	25%	19%	
weighted a	32%	24%	27%	22675

AGRESOR SMOTE

XGBOOST

index	precision	recall	f1-score	support
Administra	2%	20%	3%	55
Escolar	51%	43%	46%	7848
Externos	20%	39%	26%	3140
Familiar	72%	57%	64%	11632
accuracy			50%	
macro avg	36%	40%	35%	
weighted a	57%	50%	52%	22675

MLP

index	precision	recall	f1-score	support
Administra	1%	29%	3%	55
Escolar	52%	36%	42%	7848
Externos	19%	44%	27%	3140
Familiar	72%	55%	62%	11632
accuracy			47%	
macro avg	36%	41%	33%	
weighted a	58%	47%	50%	22675

AGRESOR UNDER

XGBOOST

index	precision	recall	f1-score	support
Administra	1%	60%	1%	55
Escolar	47%	33%	39%	7848
Externos	16%	29%	20%	3140
Familiar	69%	41%	51%	11632
accuracy			37%	
macro avg	33%	41%	28%	
weighted a	54%	37%	43%	

MLP

index	precision	recall	f1-score	support
Administra	1%	56%	2%	55
Escolar	46%	32%	37%	7848
Externos	15%	33%	21%	3140
Familiar	65%	39%	49%	11632
accuracy			36%	
macro avg	32%	40%	27%	
weighted a	51%	36%	41%	22675

RESUMEN BASE | RESUMEN HIPER | RESUMEN OVER-UNDER

En conclusión **"el base"** significa que el modelo se entrenó con los datos tal cual vienen, sin realizar ningún ajuste al desbalanceo de clases, siendo el XGBOOST quien presenta mejor rendimiento **(Gráfica 18).**

Hoja Resumen hiper: Una vez identificado, a partir del escenario base, que el modelo con mejor desempeño era XGBoost, se procedió a realizar un ajuste de sus hiperparámetros con el propósito de evaluar si era posible mejorar su capacidad predictiva **(Gráfica 19)**. Este ejercicio consistió en probar, mediante una grilla de valores, diferentes combinaciones de hiperparámetros para determinar cuál podía aumentar su rendimiento.
Es importante señalar que este proceso es computacionalmente costoso; por esta razón, se decidió aplicarlo únicamente a XGBoost y no a las demás metodologías. Tras completar el ejercicio, se concluyó que el ajuste de hiperparámetros no generó mejoras significativas respecto al escenario base inicial.

G19 *Gráfica 19. Ajuste de hiperparámetros al mejor modelo XGBOOST*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

VIOLENCIA				
XGBOOST				
index	**precision**	**recall**	**f1-score**	**support**
Abandono	0%	0%	0%	296
Económica	57%	6%	11%	134
Física	37%	31%	34%	6017
Negligencia	41%	47%	44%	4960
Psicológica	34%	6%	10%	4507
Sexual	39%	64%	49%	6761
accuracy			39%	
macro avg	35%	26%	25%	
weighted avg	38%	39%	35%	22675

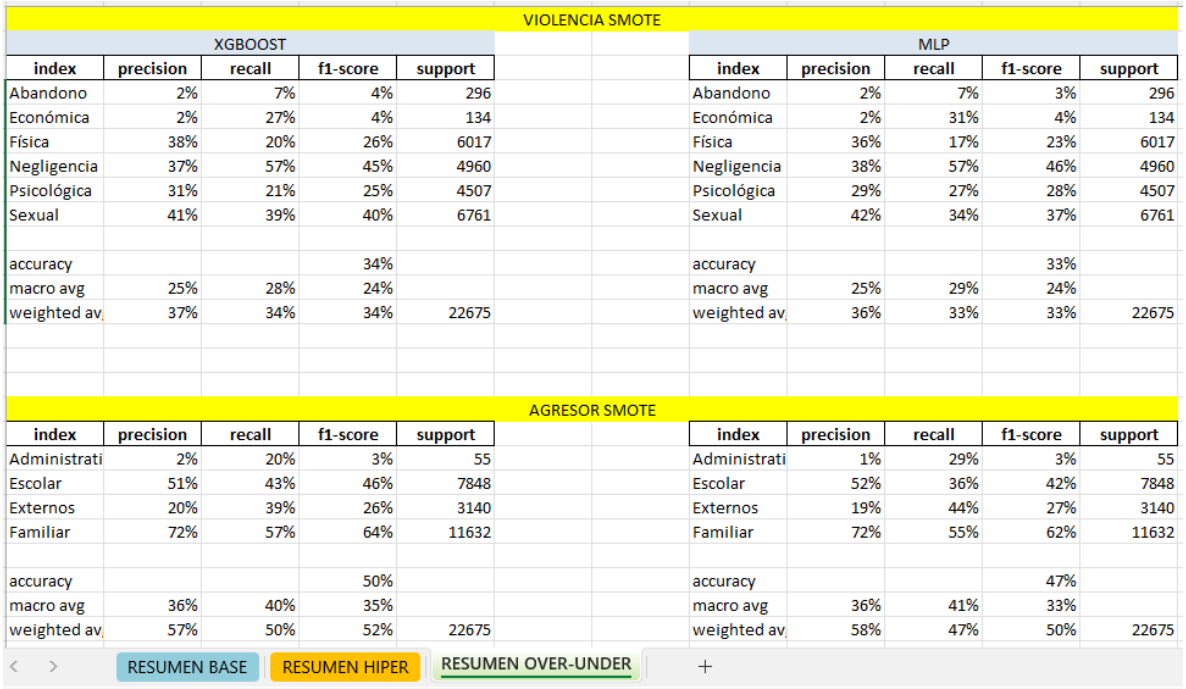
AGRESOR				
index	**precision**	**recall**	**f1-score**	**support**
Administrativo	0%	0%	0%	55
Escolar	47%	64%	54%	7848
Externos	50%	0%	0%	3140
Familiar	66%	68%	67%	11632
accuracy			57%	
macro avg	41%	33%	30%	
weighted avg	57%	57%	53%	22675

Hoja resumen OVER-UNDER: (Gráfica 20) Como se mencionó previamente, existía un problema de desbalance en las clases de las variables *Tipo de agresor* y *Tipo de violencia*, lo cual probablemente influyó en el bajo rendimiento predictivo de los modelos ajustados inicialmente. Con el fin de mejorar dicho rendimiento, se decidió aplicar técnicas de balanceo mediante **sobre muestreo** y **submuestreo**. Estas técnicas no se aplicaron a todas las metodologías evaluadas en el escenario base, sino únicamente a las dos con mejor desempeño: **XGBoost** y **MLP**.

En el caso del sobre muestreo, se empleó la técnica **SMOTE**, la cual genera observaciones sintéticas de las clases minoritarias con el objetivo de equilibrar la cantidad de registros por clase sin perder información. No obstante, un aspecto negativo de esta estrategia es que las observaciones generadas son

"artificiales", lo que puede introducir ciertas distorsiones durante el entrenamiento del modelo.

G20 *Gráfica 20. Sobremustreo "Técnica Smote"*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



VIOLENCIA SMOTE

XGBOOST

index	precision	recall	f1-score	support
Abandono	2%	7%	4%	296
Económica	2%	27%	4%	134
Física	38%	20%	26%	6017
Negligencia	37%	57%	45%	4960
Psicológica	31%	21%	25%	4507
Sexual	41%	39%	40%	6761
accuracy			34%	
macro avg	25%	28%	24%	
weighted av	37%	34%	34%	22675

MLP

index	precision	recall	f1-score	support
Abandono	2%	7%	3%	296
Económica	2%	31%	4%	134
Física	36%	17%	23%	6017
Negligencia	38%	57%	46%	4960
Psicológica	29%	27%	28%	4507
Sexual	42%	34%	37%	6761
accuracy			33%	
macro avg	25%	29%	24%	
weighted av	36%	33%	33%	22675

AGRESOR SMOTE

index	precision	recall	f1-score	support
Administrati	2%	20%	3%	55
Escolar	51%	43%	46%	7848
Externos	20%	39%	26%	3140
Familiar	72%	57%	64%	11632
accuracy			50%	
macro avg	36%	40%	35%	
weighted av	57%	50%	52%	22675

index	precision	recall	f1-score	support
Administrati	1%	29%	3%	55
Escolar	52%	36%	42%	7848
Externos	19%	44%	27%	3140
Familiar	72%	55%	62%	11632
accuracy			47%	
macro avg	36%	41%	33%	
weighted av	58%	47%	50%	22675

RESUMEN BASE | RESUMEN HIPER | RESUMEN OVER-UNDER | +

Con la metodología de submuestreo **(Gráfica 21)**, lo que se busca es seleccionar una muestra reducida de las clases mayoritarias de manera que igualen, o al menos se aproximen mucho más, al tamaño de las clases minoritarias. Esta técnica puede resultar útil en conjuntos de datos grandes; sin embargo, implica una pérdida de información, pues se descartan observaciones para equilibrar la distribución de clases.

G21 *Gráfica 21. Submuestreo*

Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

VIOLENCIA UNDER

XGBOOST

index	precision	recall	f1-score	support
Abandono	2%	15%	3%	296
Económica	2%	40%	3%	134
Física	30%	17%	22%	6017
Negligencia	35%	36%	36%	4960
Psicológica	26%	20%	22%	4507
Sexual	38%	27%	32%	6761
accuracy			25%	
macro avg	22%	26%	20%	
weighted av	32%	25%	28%	22675

MLP

index	precision	recall	f1-score	support
Abandono	2%	17%	3%	296
Económica	1%	43%	3%	134
Física	30%	22%	26%	6017
Negligencia	35%	28%	31%	4960
Psicológica	26%	14%	18%	4507
Sexual	38%	29%	33%	6761
accuracy			24%	
macro avg	22%	25%	19%	
weighted av	32%	24%	27%	22675

AGRESOR UNDER

index	precision	recall	f1-score	support
Administrati	1%	60%	1%	55
Escolar	47%	33%	39%	7848
Externos	16%	29%	20%	3140
Familiar	69%	41%	51%	11632
accuracy			37%	
macro avg	33%	41%	28%	
weighted av	54%	37%	43%	

index	precision	recall	f1-score	support
Administrati	1%	56%	2%	55
Escolar	46%	32%	37%	7848
Externos	15%	33%	21%	3140
Familiar	65%	39%	49%	11632
accuracy			36%	
macro avg	32%	40%	27%	
weighted av	51%	36%	41%	22675

En conclusión, de los experimentos realizados, el modelo que presenta los mejores resultados es XGBoost con sobremuestreo. Si bien las métricas no alcanzan los niveles ideales esperados en un modelo de clasificación, esta configuración permite generar predicciones en todas las clases a pesar del desbalanceo original. Aunque su accuracy no es el más alto, algunas clases muestran valores de recall y precision aceptables, lo que lo convierte en la alternativa más adecuada dentro de las condiciones del proyecto.

Algunas precisiones:

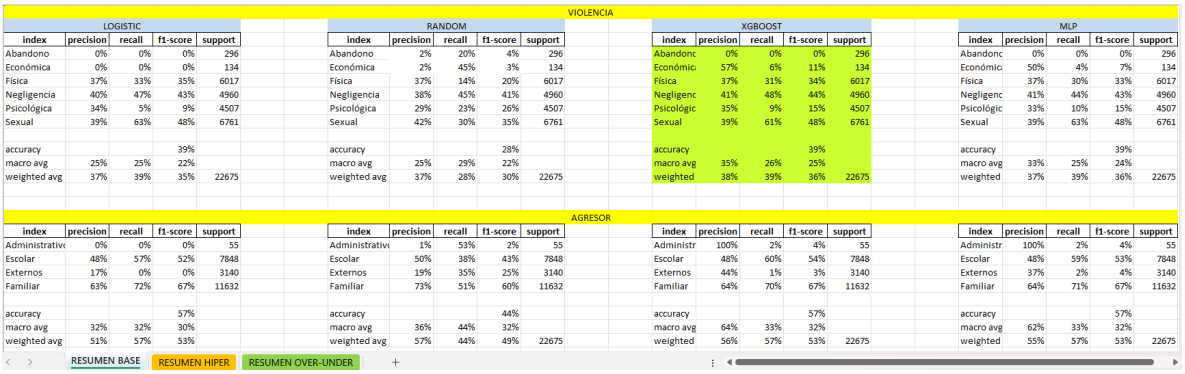
Precision: mide cuando clasificas algo como positivo, si realmente es positivo. Ejp: si se clasifica a un individuo en clase violencia económica, que realmente haya tenido violencia económica, esto mide la precisión.

Recall: de todos los que estaban en clase abandono económica cuántas detectó.

F1-score: es la media armónica entre ambos. Es como el balance entre los dos, siendo el que resume el resultado de los dos (Precisión y Recall).

G22 *Gráfica 22. XGBOOST con el Modelo Base*

Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



VIOLENCIA

LOGISTIC

index	precision	recall	f1-score	support
Abandono	0%	0%	0%	296
Económica	0%	0%	0%	134
Física	37%	33%	35%	6017
Negligencia	40%	47%	43%	4960
Psicológica	34%	5%	9%	4507
Sexual	39%	63%	48%	6761
accuracy			39%	
macro avg	25%	25%	22%	
weighted avg	37%	39%	35%	22675

RANDOM

index	precision	recall	f1-score	support
Abandono	2%	20%	4%	296
Económica	2%	45%	3%	134
Física	37%	14%	20%	6017
Negligencia	38%	45%	41%	4960
Psicológica	29%	23%	26%	4507
Sexual	42%	30%	35%	6761
accuracy			28%	
macro avg	25%	29%	22%	
weighted avg	37%	28%	30%	22675

XGBOOST

index	precision	recall	f1-score	support
Abandonc	0%	0%	0%	296
Económic	57%	6%	11%	134
Física	37%	31%	34%	6017
Negligenc	41%	48%	44%	4960
Psicológic	35%	9%	15%	4507
Sexual	39%	61%	48%	6761
accuracy			39%	
macro avg	35%	26%	25%	
weighted	38%	39%	36%	22675

MLP

index	precision	recall	f1-score	support
Abandonc	0%	0%	0%	296
Económic	50%	4%	7%	134
Física	37%	30%	33%	6017
Negligenc	41%	44%	43%	4960
Psicológic	33%	10%	15%	4507
Sexual	39%	63%	48%	6761
accuracy			39%	
macro avg	33%	25%	24%	
weighted	37%	39%	36%	22675

AGRESOR

index	precision	recall	f1-score	support
Administrativ	0%	0%	0%	55
Escolar	48%	57%	52%	7848
Externos	17%	0%	0%	3140
Familiar	63%	72%	67%	11632
accuracy			57%	
macro avg	32%	32%	30%	
weighted avg	51%	57%	53%	

index	precision	recall	f1-score	support
Administrativ	1%	53%	2%	55
Escolar	50%	38%	43%	7848
Externos	19%	35%	25%	3140
Familiar	73%	51%	60%	11632
accuracy			44%	
macro avg	36%	44%	32%	
weighted avg	57%	44%	49%	22675

index	precision	recall	f1-score	support
Administr	100%	2%	4%	55
Escolar	48%	60%	54%	7848
Externos	44%	1%	3%	3140
Familiar	64%	70%	67%	11632
accuracy			57%	
macro avg	64%	33%	32%	
weighted	56%	57%	53%	22675

index	precision	recall	f1-score	support
Administr	100%	2%	4%	55
Escolar	48%	59%	53%	7848
Externos	37%	2%	4%	3140
Familiar	64%	71%	67%	11632
accuracy			57%	
macro avg	62%	33%	32%	
weighted	55%	57%	53%	22675

RESUMEN BASE | RESUMEN HIPER | RESUMEN OVER-UNDER

G23 *Gráfica 23. XGBOOST con el Modelo Over-Under*

Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

VIOLENCIA SMOTE

XGBOOST

index	precision	recall	f1-score	support
Abandono	2%	7%	4%	296
Económica	2%	27%	4%	134
Física	38%	20%	26%	6017
Negligencia	37%	57%	45%	4960
Psicológica	31%	21%	25%	4507
Sexual	41%	39%	40%	6761
accuracy			34%	
macro avg	25%	28%	24%	
weighted av	37%	34%	34%	22675

MLP

index	precision	recall	f1-score	support
Abandono	2%	7%	3%	296
Económica	2%	31%	4%	134
Física	36%	17%	23%	6017
Negligencia	38%	57%	46%	4960
Psicológica	29%	27%	28%	4507
Sexual	42%	34%	37%	6761
accuracy			33%	
macro avg	25%	29%	24%	
weighted av	36%	33%	33%	22675

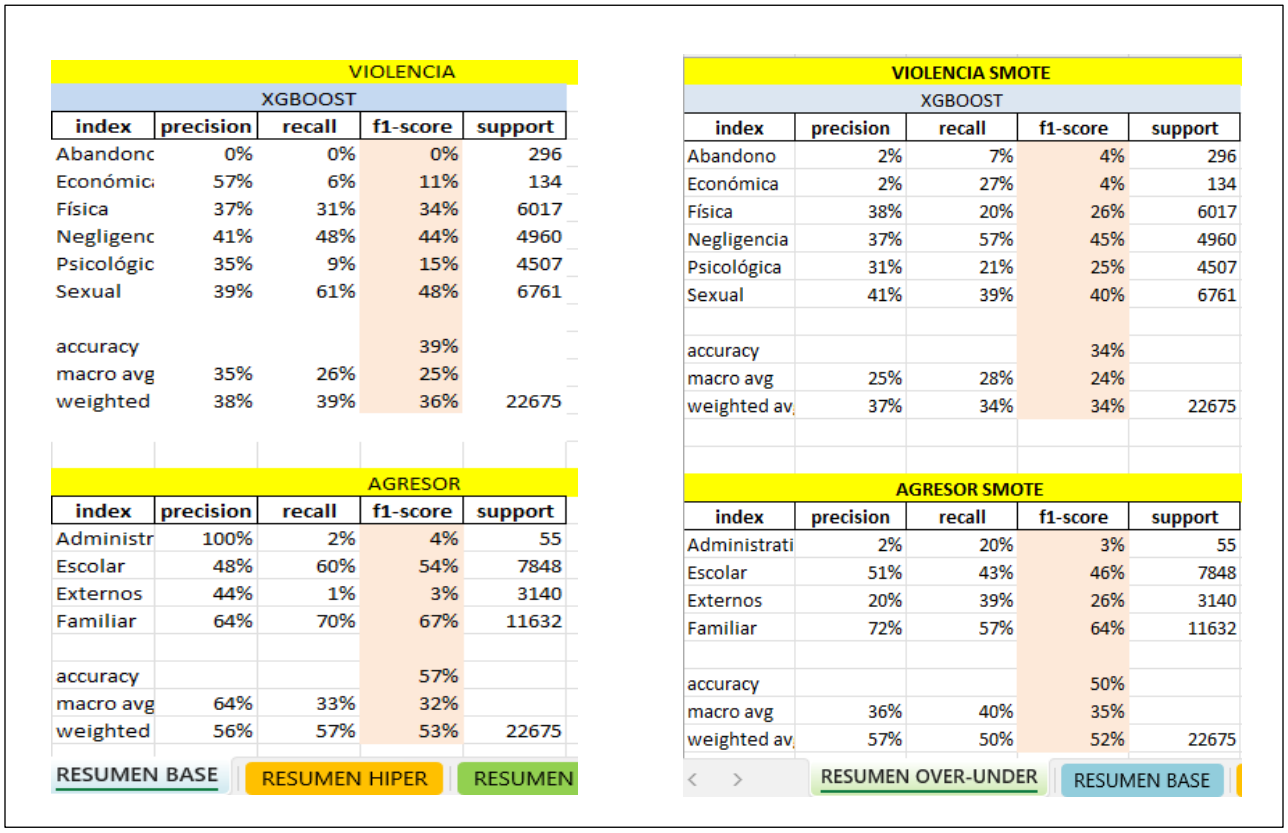
AGRESOR SMOTE

index	precision	recall	f1-score	support
Administrati	2%	20%	3%	55
Escolar	51%	43%	46%	7848
Externos	20%	39%	26%	3140
Familiar	72%	57%	64%	11632
accuracy			50%	
macro avg	36%	40%	35%	

index	precision	recall	f1-score	support
Administrati	1%	29%	3%	55
Escolar	52%	36%	42%	7848
Externos	19%	44%	27%	3140
Familiar	72%	55%	62%	11632
accuracy			47%	
macro avg	36%	41%	33%	

RESUMEN BASE | RESUMEN HIPER | RESUMEN OVER-UNDER

G24 *Gráfica 24. XGBOOST con mejor Accuracy y f1-score*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



VIOLENCIA
XGBOOST

index	precision	recall	f1-score	support
Abandonc	0%	0%	0%	296
Económic	57%	6%	11%	134
Física	37%	31%	34%	6017
Negligenc	41%	48%	44%	4960
Psicológic	35%	9%	15%	4507
Sexual	39%	61%	48%	6761
accuracy			39%	
macro avg	35%	26%	25%	
weighted	38%	39%	36%	22675

AGRESOR

index	precision	recall	f1-score	support
Administr	100%	2%	4%	55
Escolar	48%	60%	54%	7848
Externos	44%	1%	3%	3140
Familiar	64%	70%	67%	11632
accuracy			57%	
macro avg	64%	33%	32%	
weighted	56%	57%	53%	22675

RESUMEN BASE | RESUMEN HIPER | RESUMEN

VIOLENCIA SMOTE
XGBOOST

index	precision	recall	f1-score	support
Abandono	2%	7%	4%	296
Económica	2%	27%	4%	134
Física	38%	20%	26%	6017
Negligencia	37%	57%	45%	4960
Psicológica	31%	21%	25%	4507
Sexual	41%	39%	40%	6761
accuracy			34%	
macro avg	25%	28%	24%	
weighted av	37%	34%	34%	22675

AGRESOR SMOTE

index	precision	recall	f1-score	support
Administrati	2%	20%	3%	55
Escolar	51%	43%	46%	7848
Externos	20%	39%	26%	3140
Familiar	72%	57%	64%	11632
accuracy			50%	
macro avg	36%	40%	35%	
weighted av	57%	50%	52%	22675

RESUMEN OVER-UNDER | RESUMEN BASE

2.2.Actividades

En esta sección se describe de forma general las actividades que se realizaron durante el desarrollo de la solución de este caso de uso (Tabla 5).

T5 Tabla 5. Cronograma de Actividades
Fuente: Elaboración propia

Fases	Actividades	Mayo	Junio	Julio	Agt.	Sep.	Oct.	Nov.	Dic.
INICIO	1. Reuniones con el cliente interno para entender necesidades y expectativas								
	2. Análisis inicial de datos existentes								
	3. Depuración y normalización de datos								
	4. Reunión de seguimiento con cliente interno								
IDEAR	5. Modelos ML propuestos								
	6. Ensayos								
	7. Resultados								
PROTOTIPAR	8. Propuesta de Prototipo								
	9. Pruebas del prototipo								

Fases	Actividades	Mayo	Junio	Julio	Agt.	Sep.	Oct.	Nov.	Dic.
	10. Presentación Interna del aplicativo								
	9. Presentación Externa (cliente) del aplicativo								
OPERACIONALIZAR – INDUSTRIALIZAR	10. Despliegue de la solución en el entorno de producción								
	11. Pruebas con cliente final								
	12. Reunión interna de seguimiento								

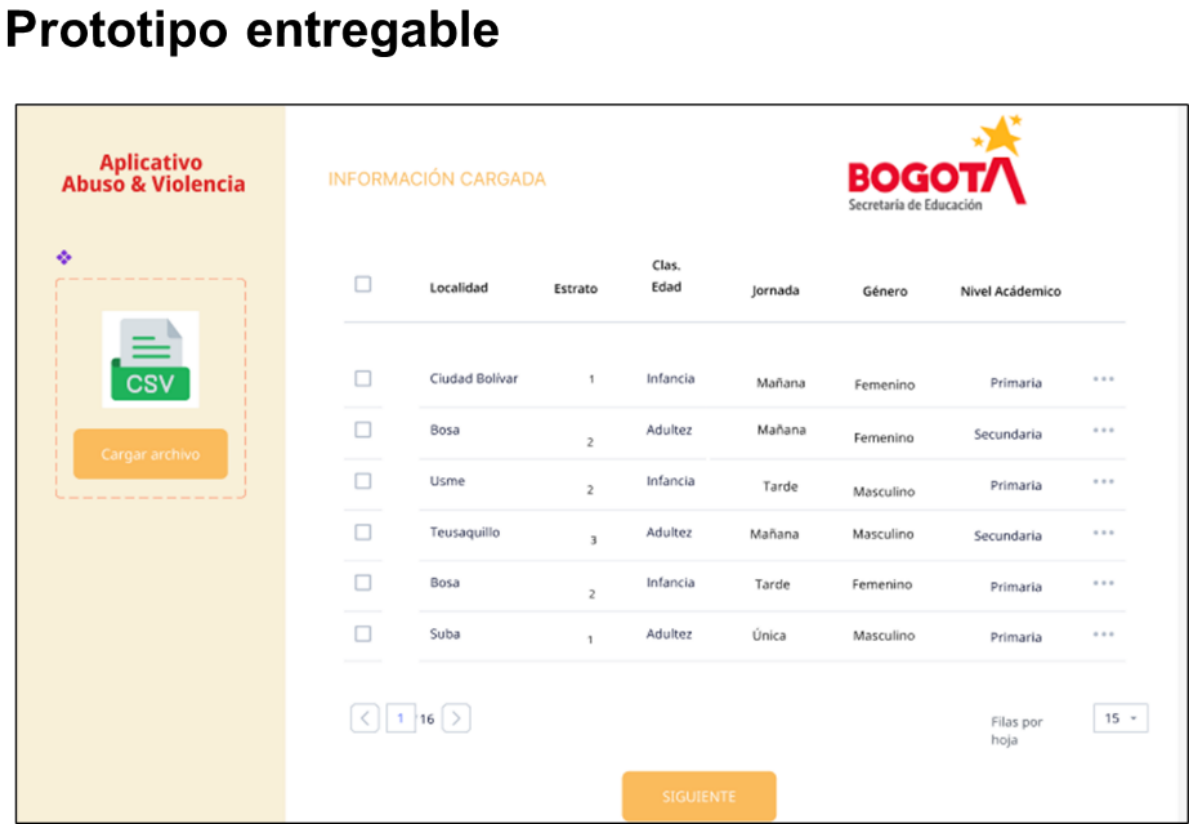
3. Prototipar

Una vez analizada la data suministrada por el Obce y realizados los análisis descriptivos univariados y bivariados, se diseñó un MockUp (Prototipo), como se describe a continuación.

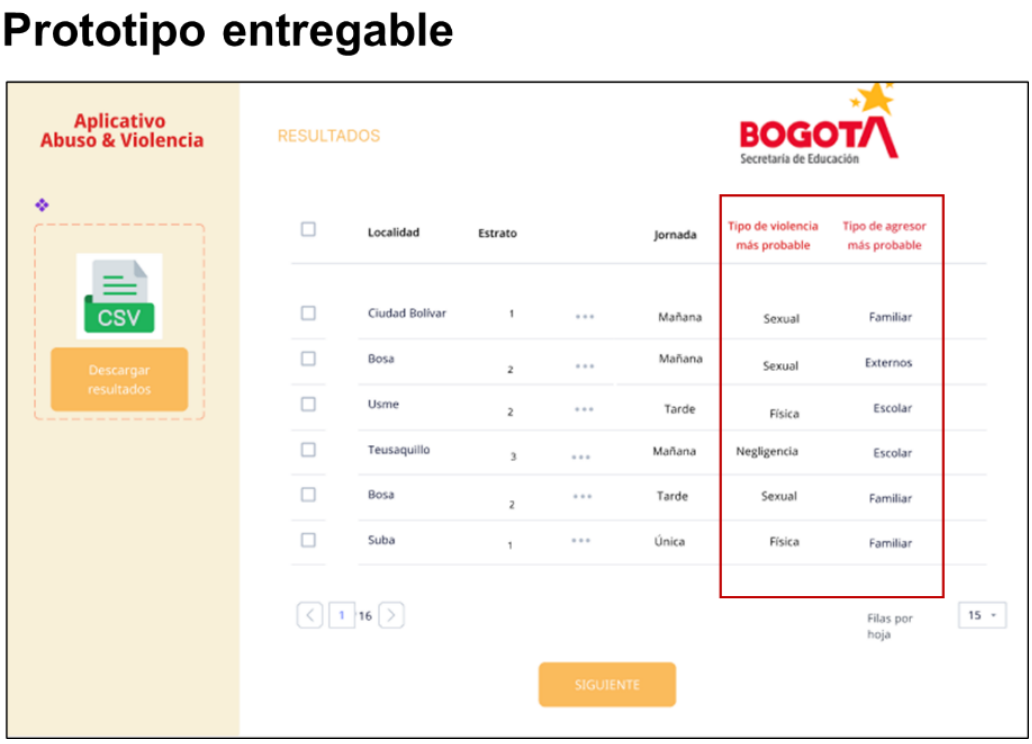
3.1. Prototipo / Modelado: Aplicativo Abuso y Violencia

El prototipo (MockUp) representa el aplicativo de lo que se entregaría, y corresponde a una interfaz donde se pueda cargar un archivo con el listado de los estudiantes a los que se desee realizarles el estudio sobre la posibilidad de tener algún tipo de abuso y/o violencia– **Gráfica 25.**

G25 *Gráfica 25. Diseño de Prototipo Entregable*
Fuente: Perfilamiento de Datos y Estadística – Desarrollo inicial en Phyton



G26 *Gráfica 26. Diseño de Prototipo Entregable*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



Luego de cargado el archivo con los datos de los estudiantes, el aplicativo arrojaría para cada estudiante el tipo de violencia y el tipo de agresor más probable **(Gráfica 17)**.

El modelo asignaría una categoría de violencia y agresor a cada estudiante.

NOTA: Si se obtuviera información de estudiantes sin eventos se podrían determinar los que no presenten ningún riesgo.

Con el fin de elegir el mejor modelo que prediga la posible violencia y abuso que pueda llegar a tener un estudiante, se realizó el planteamiento de tres escenarios por medio de los cuales se probaron 9 experimentos aplicando cuatro modelos de Machine Learning (ML), como se describe en los siguientes testeos.

3.2. Testeo 0.1: Escenario Base

Escenario Base: En este escenario se hicieron cuatro experimentos, aplicando los siguientes cuatro modelos (un experimento por modelo), manteniendo los hiperparámetros por defecto y sin modificar los conjuntos de datos originales :

-Modelo Logístico: Un modelo estadístico sencillo que calcula la probabilidad de que algo ocurra (por ejemplo, "sí" o "no") usando relaciones lineales entre las variables.

-Modelo Random Forest: Un conjunto de muchos árboles de decisión que votan entre sí. Cada árbol aporta su opinión y el modelo elige la respuesta más consensuada.

-Modelo XGBOOST: Un método avanzado que combina muchos "pequeños árboles de decisión" para hacer predicciones muy precisas, aprendiendo de los errores que va cometiendo.

-Modelo MLP: Un tipo de red neuronal que imita cómo procesa información el cerebro: toma datos de entrada, los transforma a través de varias capas y produce una predicción.

3.3. Testeo 0.2: Escenario Hiperparámetros

Escenario Hiperparámetros: Basado en los resultados del escenario previo, se seleccionó la metodología con mejor rendimiento, una vez seleccionada se ajusta una grilla de hiperparámetros que permitan determinar si existe o no una mejora en el rendimiento del modelo. Se decidió solamente seleccionar una de las metodologías debido al costo computacional.

3.4. Testeo 0.3: Escenario Balanceo de Clases

Escenario Balanceo de Clases (Over-Under): En este escenario se decidió tomar los dos modelos con mejor rendimiento en el escario base, siendo MLP y XGBoost; una vez seleccionados estos modelos se les aplicó Sobre Muestreo usando la técnica de Smote para un equivalente a dos

experimentos más; y por otro lado, se les aplicó Submuestreo, con lo cual se obtuvieron dos experimentos adicionales.

En total se desarrollaron 9 experimentos aplicando cuatro modelos de Machine Learning (ML).

3.5. Revisión Cliente

La revisión con el cliente se hizo mediante presentación ante el Jefe del Obce y su equipo técnico; a la vez, se compartieron las grabaciones de entrega del aplicativo que permite realizar la predicción sobre el posible caso de abuso que pueda experimentar un estudiante; el aplicativo como se describe en los capítulos siguientes permite visualizar resultados descriptivos y las posibles predicciones que pueda llegar a presentar el estudiante.

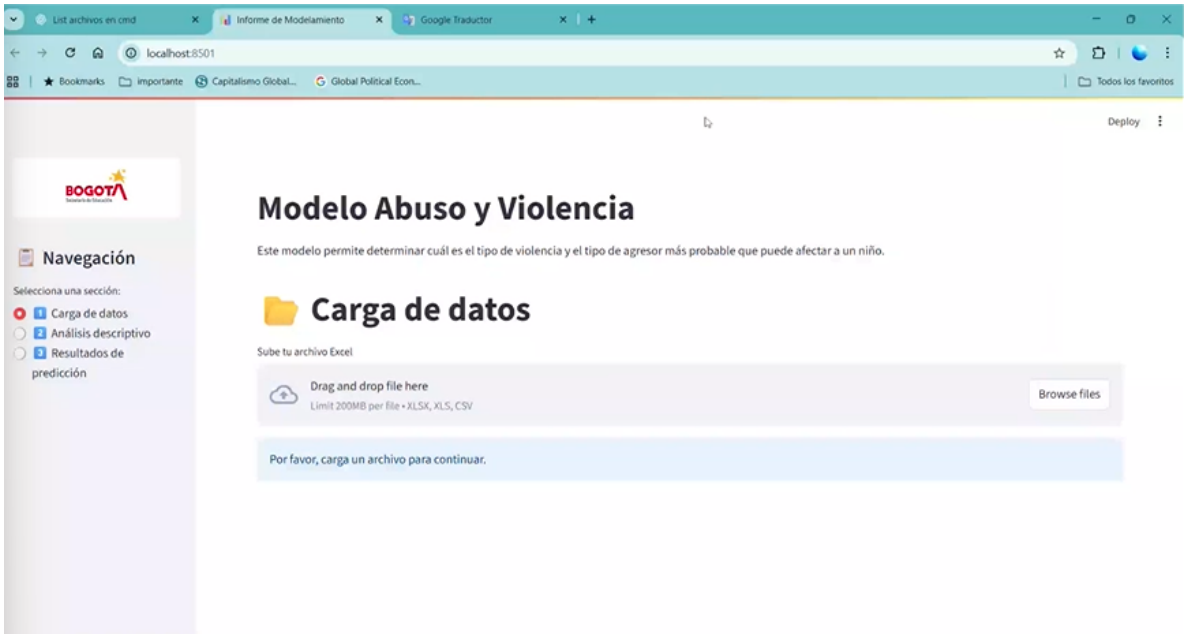
4. Operacionalizar – Industrializar

A continuación, se presenta el aplicativo como materialización del MockUp, sujeto a modificaciones.

Está compuesto por tres módulos:

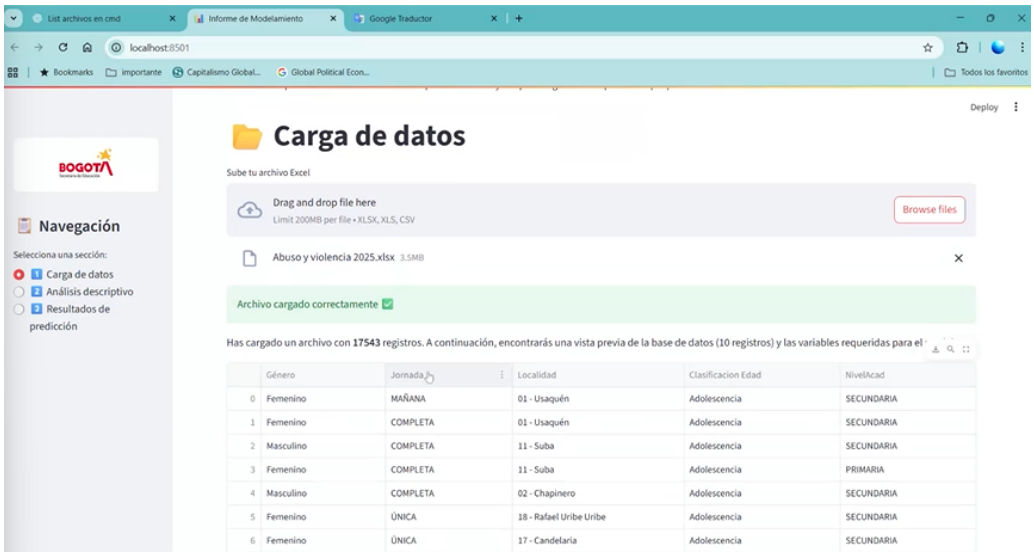
1. Carga de datos.
2. Análisis descriptivo.
3. Resultados de la predicción.

G27 *Gráfica 27. Tablero de predicción para Abuso y Violencia*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



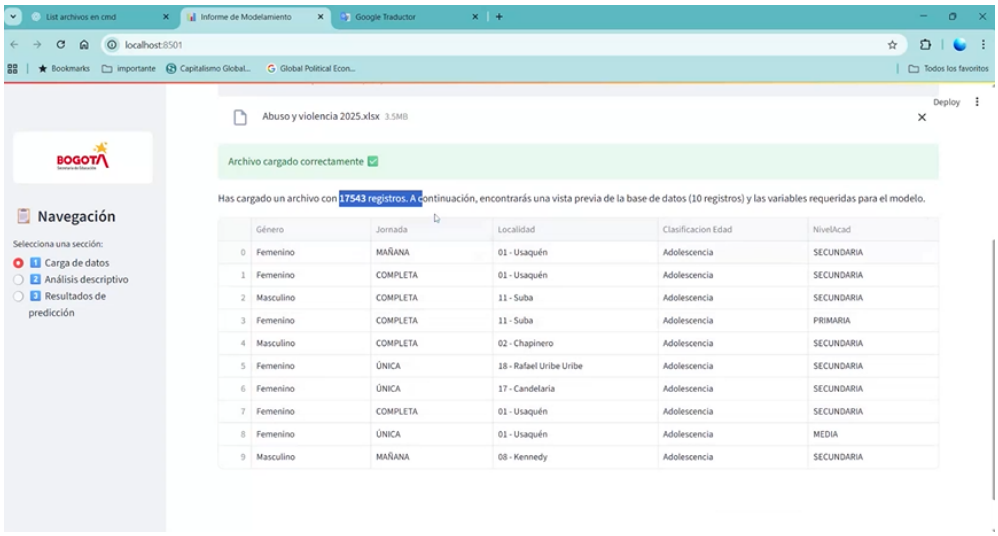
Se cargan los datos con los individuos a los cuales se les desea realizar una predicción; se pueden cargar datos en formato .xlsx o .csv, para que el aplicativo funcione satisfactoriamente es indispensable que el archivo contenga las 5 variables de interés: Género, Jornada, Nivel Académico, Clasificación Edad y Localidad.

G28 *Gráfica 28. Tablero de predicción para Abuso y Violencia – Módulo carga de datos*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



Una vez cargada la información, el aplicativo indicará mediante un mensaje que los datos han sido cargados satisfactoriamente. Además, mostrará una muestra del conjunto de datos y la cantidad total de registros procesados.

G29 *Gráfica 29. Módulo carga de datos – Ejemplo datos cargados*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

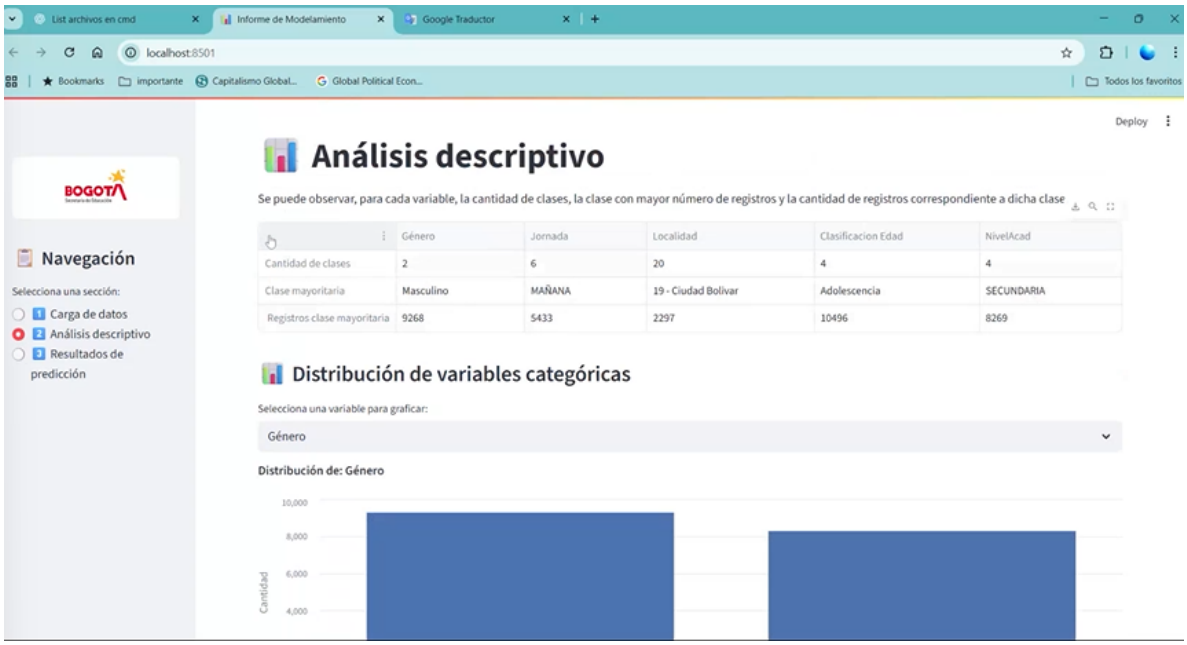


	Género	Jornada	Localidad	Clasificacion Edad	NivelAcad
0	Femenino	MAÑANA	01 - Usaquén	Adolescencia	SECUNDARIA
1	Femenino	COMPLETA	01 - Usaquén	Adolescencia	SECUNDARIA
2	Masculino	COMPLETA	11 - Suba	Adolescencia	SECUNDARIA
3	Femenino	COMPLETA	11 - Suba	Adolescencia	PRIMARIA
4	Masculino	COMPLETA	02 - Chapinero	Adolescencia	SECUNDARIA
5	Femenino	ÚNICA	18 - Rafael Uribe Uribe	Adolescencia	SECUNDARIA
6	Femenino	ÚNICA	17 - Candelaria	Adolescencia	SECUNDARIA
7	Femenino	COMPLETA	01 - Usaquén	Adolescencia	SECUNDARIA
8	Femenino	ÚNICA	01 - Usaquén	Adolescencia	MEDIA
9	Masculino	MAÑANA	08 - Kennedy	Adolescencia	SECUNDARIA

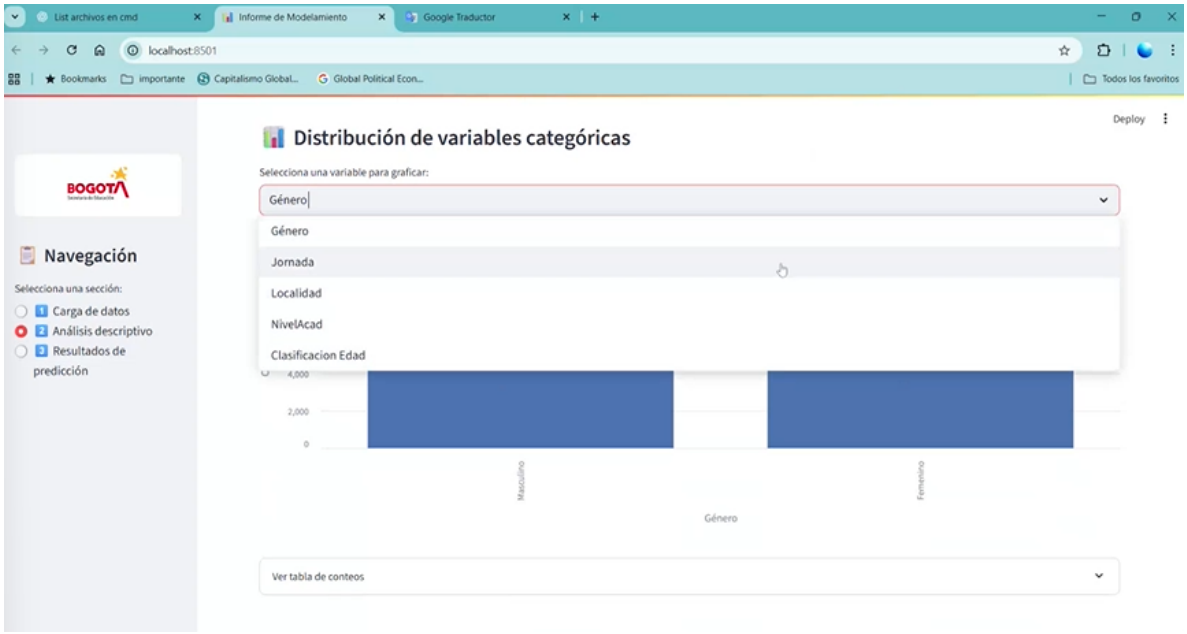
En el **módulo 2: Análisis descriptivo**, se presenta un resumen conciso de las variables cargadas, donde se visualiza información como la cantidad de clases por variable, la clase con mayor número de registros y el total de observaciones correspondientes a dicha clase mayoritaria.

Adicionalmente, en la misma página se muestran diagramas de barras que ilustran la distribución de cada variable, así como tablas con los conteos por clase. Estas tablas incluyen la opción de descarga para facilitar su consulta y análisis posterior.

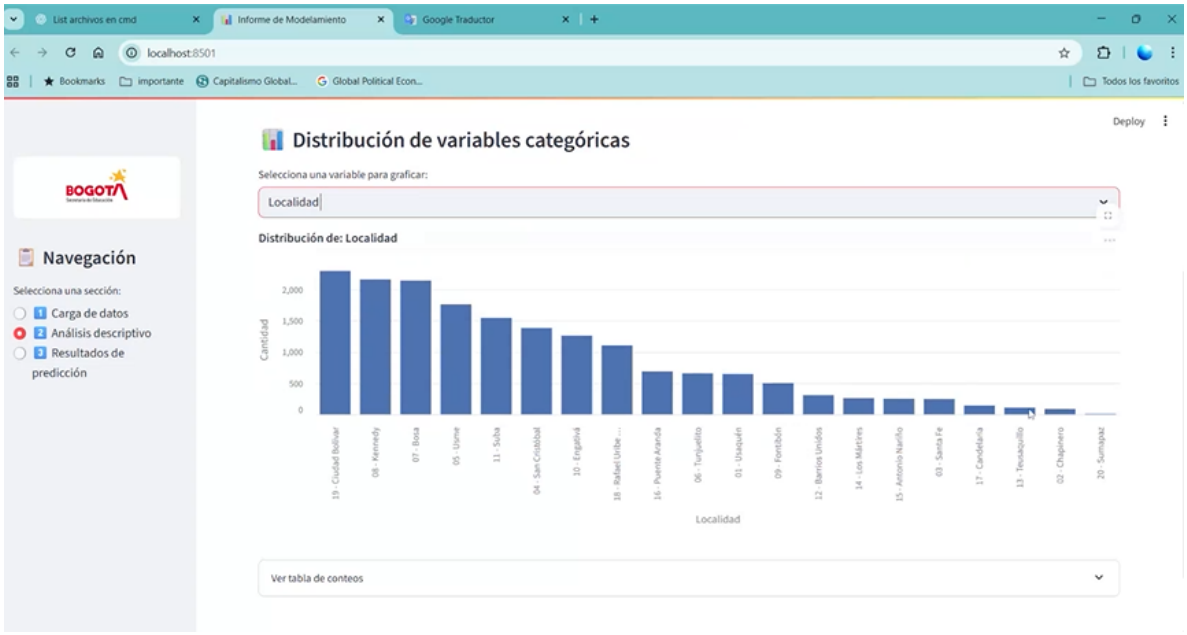
G30 *Gráfica 30. Tablero de predicción para Abuso y Violencia – Módulo análisis descriptivo*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



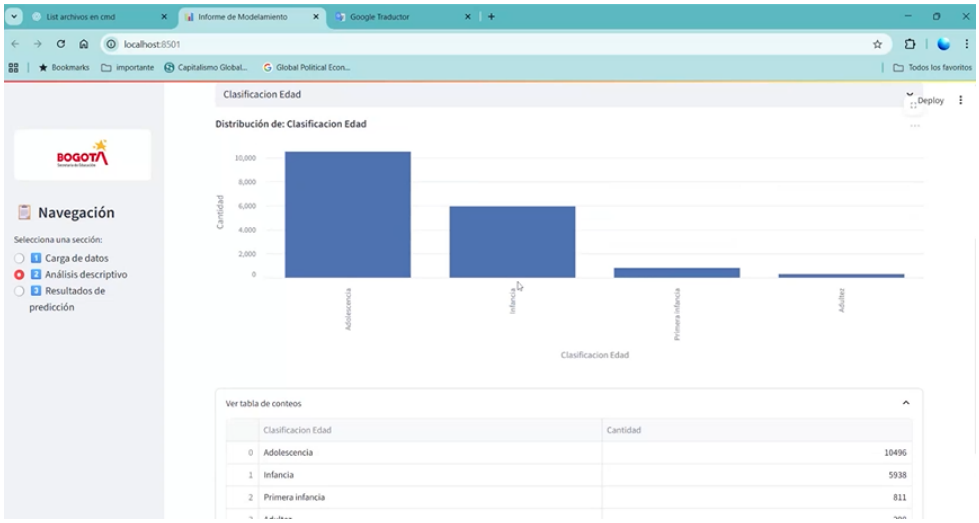
G31 *Gráfica 31. Tablero de predicción para Abuso y Violencia – Distribución de variables categóricas*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



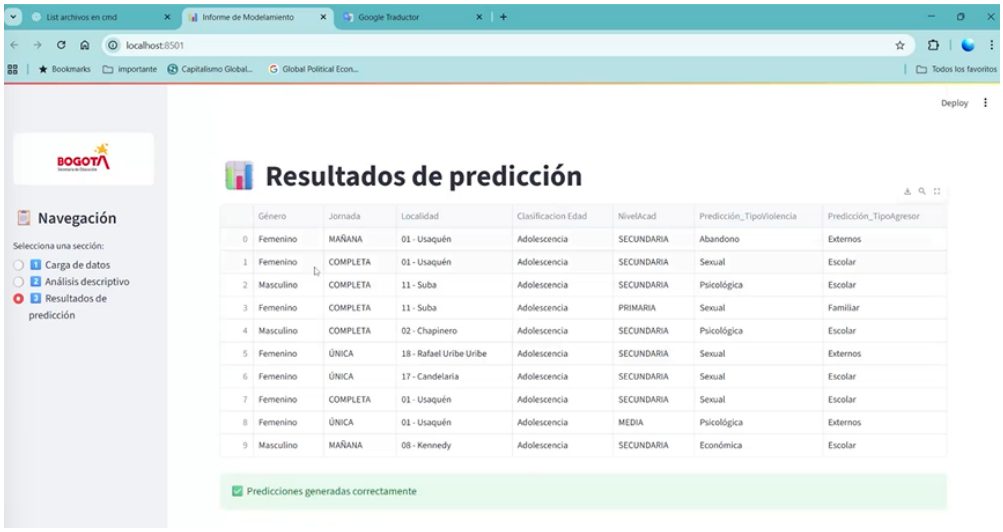
G32 *Gráfica 32. Tablero de predicción para Abuso y Violencia – Distribución por Localidad*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



G33 *Gráfica 33. Tablero de predicción para Abuso y Violencia – Tabla de reportes*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



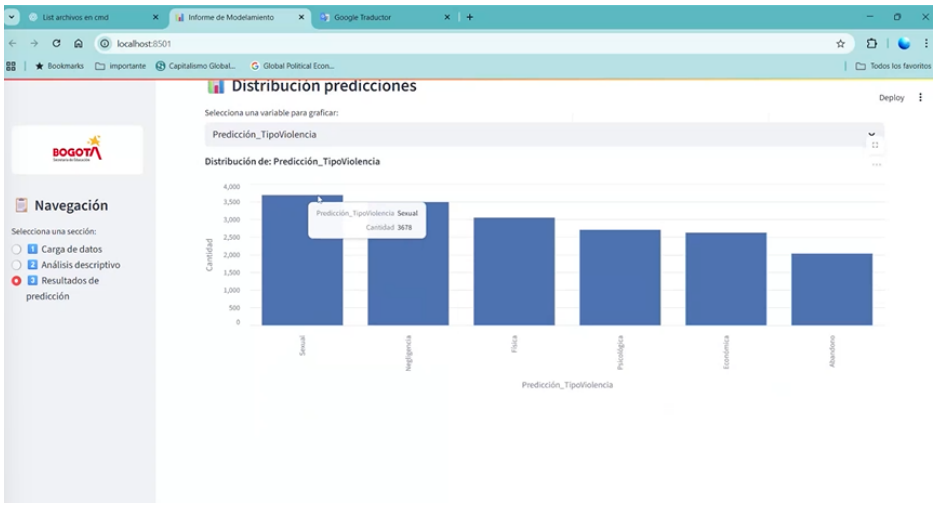
G34 *Gráfica 34. Tablero de predicción para Abuso y Violencia – Módulo resultados de predicción*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



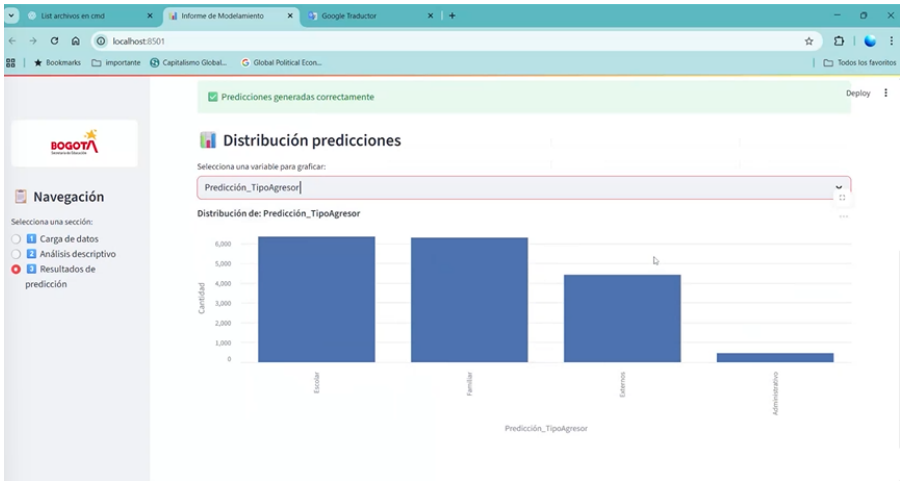
En el módulo 3: Resultado de predicción, se visualiza la información original cargada junto con dos nuevas variables que corresponden a las predicciones generadas por el modelo para cada individuo. Adicionalmente, para cada

caso se incluyen diagramas de barras que permiten observar cómo se distribuyen las clases predichas en cada una de las nuevas variables.
Información que se exportaría con las predicciones para cada niño (en un escenario real).

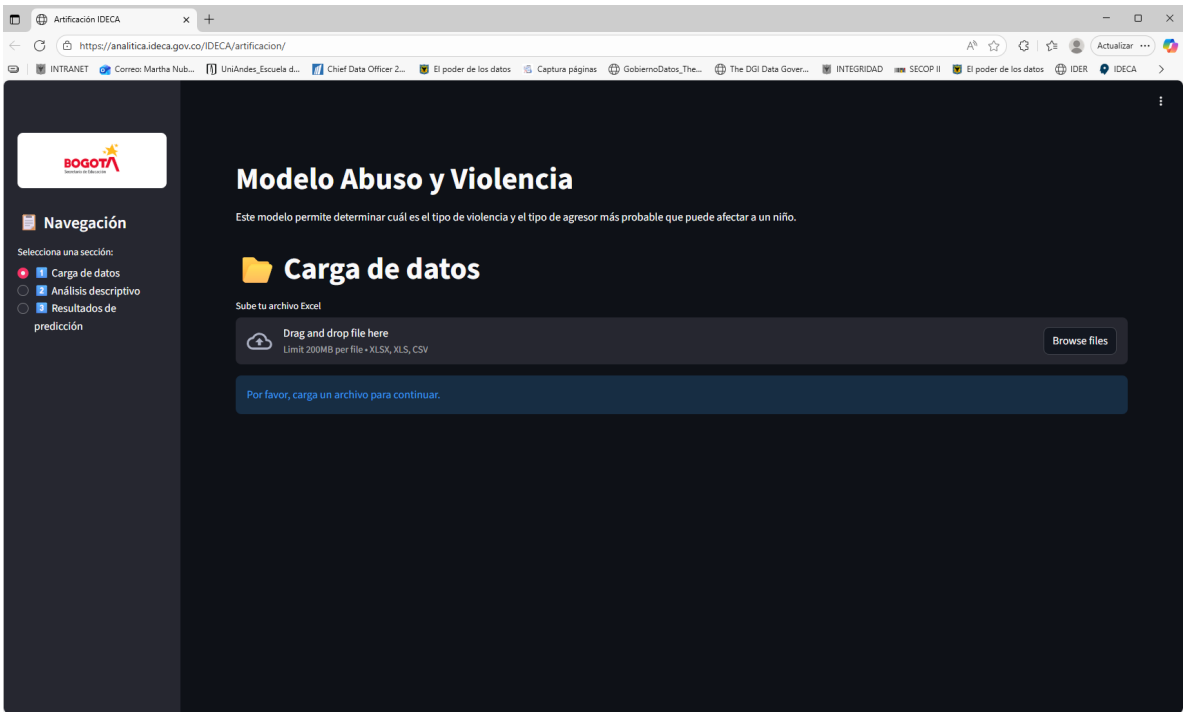
G35 *Gráfica 35. Distribución de la predicción tipo violencia*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



G36 *Gráfica 36. Distribución de la predicción tipo agresor*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

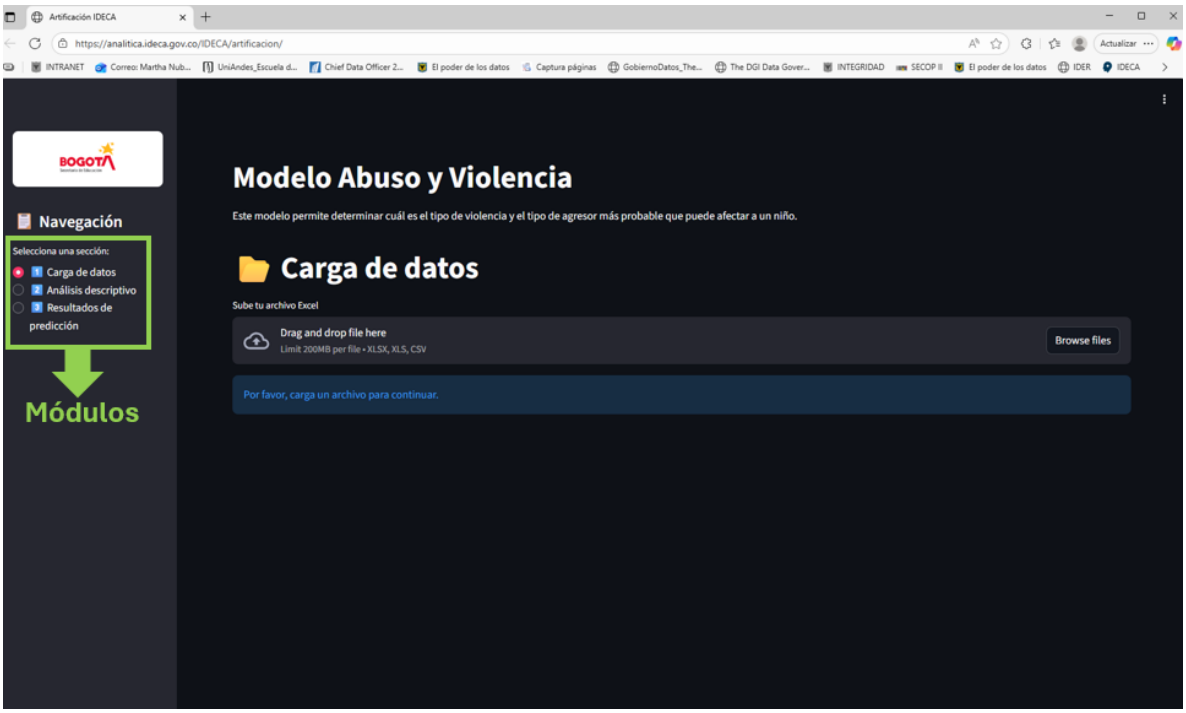


G37 *Gráfica 37. Tablero de predicción para Abuso y Violencia – Despliegue en Azure -*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



Se adjunta el código que soporta el tablero del Proyecto de Artificación con la SED para su despliegue en Azure.
Se realizó la creación de un contenedor sobre nuestro servidor linux catavms001, el cual se configuró con el ambiente python requerido para la ejecución de la librería streamlit con la que se desarrolló la aplicación. Se realizó el despliegue y se solucionaron algunos problemas de librerías. Se realizaron pruebas y se evidencia correcta ejecución.
Se configuró la URL que aloja la aplicación, quedando de la siguiente forma:
https://analitica.ideca.gov.co/IDECA/artificacion/

G38 *Gráfica 38. Módulos Tablero de predicción para Abuso y Violencia – Despliegue en Azure -*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton

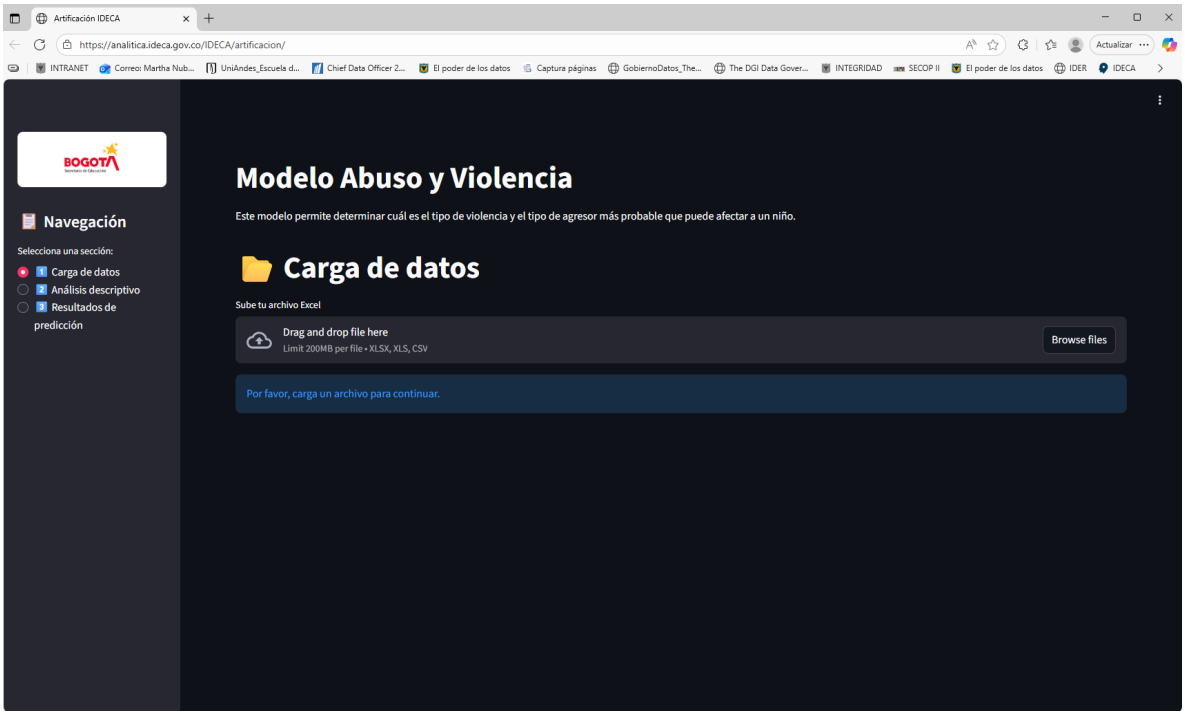


El aplicativo clasifica a todos los niños dentro de alguna de las dos categorías de violencia debido a la ausencia de registros de niños que no hayan presentado ningún tipo de violencia; es decir, no se contó con datos de casos negativos. Esto limita la capacidad del modelo para diferenciar entre presencia y ausencia de violencia, y explica en parte el comportamiento observado.

4.1. Transferencia

En esta fase de transferencia, se explicó al Equipo del Observatorio de Convivencia Escolar en qué consiste el aplicativo, los módulos diseñados, la forma de cargar los datos, así como los resultados que arroja.

G39 *Gráfica 39. Módulos Tablero de predicción para Abuso y Violencia – Transferencia del conocimiento -*
Fuente: Perfilamiento de Datos y Estadística - Desarrollo Inicial en Phyton



4.2. Retrospectiva

El Equipo del Observatorio de Convivencia Escolar explorará el aplicativo de predicción en cuanto a probar los tres módulos: cargue de datos, presentación de estadísticas descriptivas y predicción del posible caso de abuso y violencia, para lo cual se facilitó el enlace de accedo al aplicativo por medio de la siguiente URL: https://analitica.ideca.gov.co/IDECA/artificacion/

4.3. Finalización

Esta primera etapa del Caso de Uso finaliza en la Fase de Operacionalización (según la Metodología de Analítica de Datos), pues se diseñó y entregó un aplicativo de predicción en Streamlit que permite a los usuarios finales generar

predicciones a partir de nuevos conjuntos de datos de manera sencilla e interactiva. Este aplicativo facilita la interpretación de los resultados, agiliza las consultas y hace más accesible la herramienta para el personal de la Secretaría, contribuyendo a una toma de decisiones más informada y oportuna.

5. Resultados

- ✓ Se desarrolló el código que soporta el tablero del Proyecto de Artificación con la SED para su despliegue en Azure.
- ✓ Se realizó la creación de un contenedor sobre nuestro servidor linux catavms001, el cual se configuró con el ambiente python requerido para la ejecución de la librería streamlit con la que se desarrolló la aplicación. Se realizó el despliegue y se solucionaron algunos problemas de librerías. Se realizaron pruebas y se evidencia correcta ejecución.
- ✓ Se configuró la URL que aloja la aplicación, quedando de la siguiente forma: https://analitica.ideca.gov.co/IDECA/artificacion/

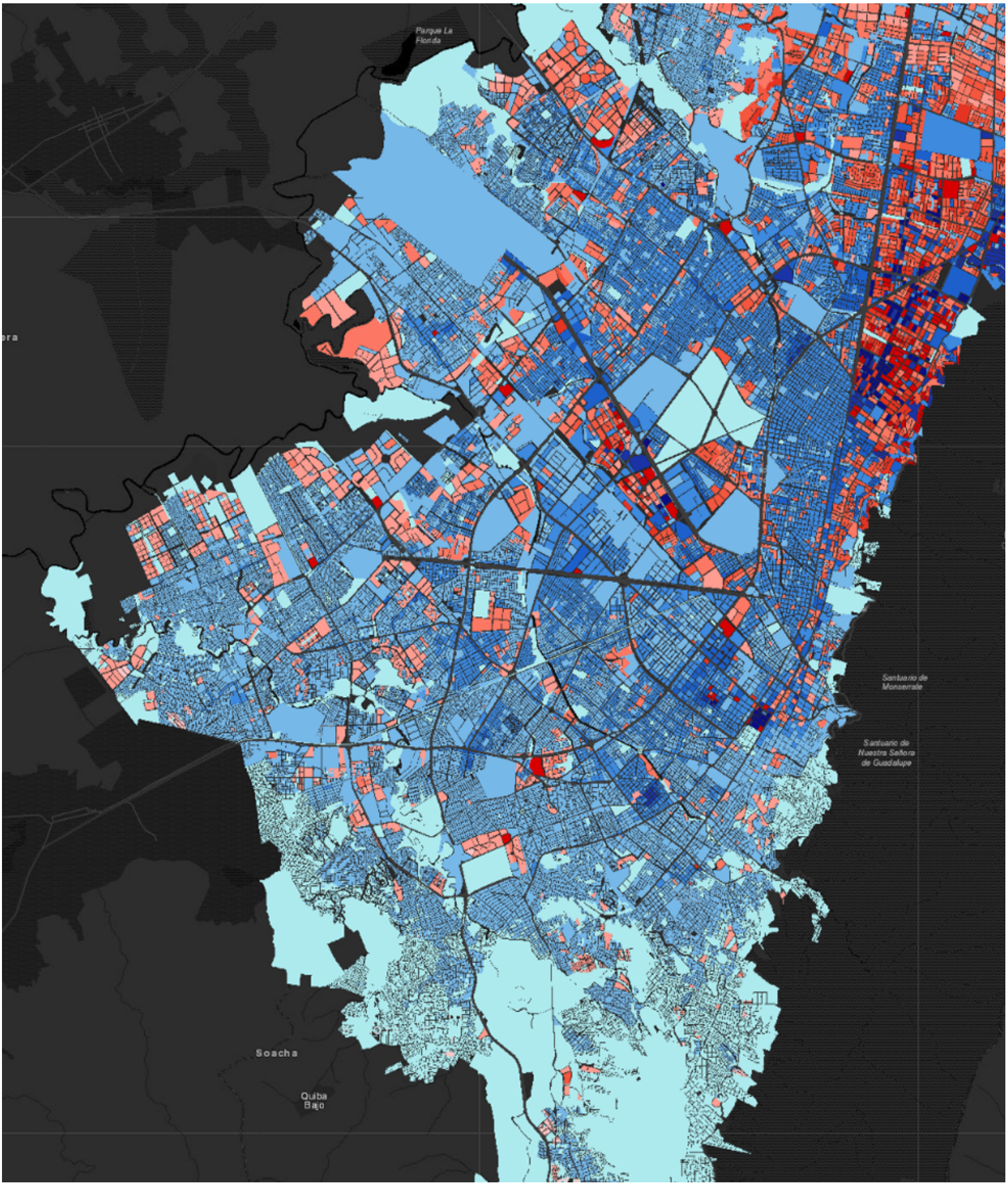
6. Conclusiones

- Los resultados obtenidos son modestos, principalmente por la falta de variables adicionales que permitan una caracterización más completa del entorno del niño tales como Información sobre condiciones socioeconómicas, situación del hogar, composición familiar o factores contextuales habría permitido construir un modelo con mayor capacidad predictiva y menor generalización. Sin embargo, dadas las limitaciones de los datos disponibles, los resultados alcanzados representan el máximo nivel de detalle posible en esta etapa.

7. Bibliografía

- ***Instituto Colombiano de Bienestar Familiar (ICBF). Modelos de probabilidad de vulneración para la prevención de violencia contra niñas, niños y adolescentes en Colombia. Consultado en Línea (Julio, 2025):*** modelos_de_probabilidad_de_vulneracion_para_la_prevencion_de_violencia_contra_ninas_ninos_y_adolescentes_en_colombia.pdf

- ***Observatorio de Convivencia Escolar. Boletines y Memorias Técnicas. Consultado en Línea (Mayo, 2025): https://www.educacionbogota.edu.co/portal_institucional/gestion-educativa/observatorio-de-convivencia-escolar***



ideca

ALCALDÍA MAYOR DE BOGOTÁ D.C.

ideca