



Informe Técnico: Resultados modelo de machine learning

Proyecto: Fortalecimiento de capacidades CTel

Fundación CINDE

Abril 21 de 2023



Tabla de contenido

Introducción	1
Marco teórico.....	4
Metodología.....	12
Datos	31
Resultados	42
Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima	47
Situaciones de victimización por intimidación.....	51
Normalización de las violencias	54
Prototipo funcional de solución web	59
Conclusiones y recomendaciones	61
Referencias	63
Anexos.....	66



Introducción

El objetivo de este proyecto es contribuir a fortalecer capacidades en CTel del sector educativo rural y urbano para garantizar ambientes educativos inclusivos y diferenciales en el marco de la emergencia derivada por el COVID-19. En especial, aportar al objetivo específico de “Establecer ambientes educativos para la escucha activa de niños, niñas y jóvenes en situaciones de vulneración derivadas del COVID-19” mediante el uso de modelos de *machine learning* para predecir si un joven se encuentra en una alta o baja cantidad de situaciones de vulnerabilidad y mediante la analítica de datos para explorar cuáles variables se relacionan con aquellos jóvenes que encuentran con una alta probabilidad de estar en esta situación.

En este caso la vulnerabilidad se analiza desde tres tipos de situaciones: i) situaciones de ansiedad, depresión o estrés postraumático, ii) situaciones de victimización por intimidación y iii) situaciones de normalización de la violencia. Para lograr esto, se analizan las interacciones de más de 250 jóvenes, entre las edades de 12 y 18 años, que asisten a 11 colegios públicos en Medellín, Colombia.

Los jóvenes en esta franja de edad a menudo se enfrentan a una serie de desafíos y riesgos psicosociales en su entorno escolar y comunitario. Estos desafíos pueden incluir la presión académica, la intimidación, la violencia en la comunidad, la discriminación y otros factores que pueden tener un impacto negativo en su bienestar emocional y mental. De ahí que la detección temprana de estos problemas sea crucial para proporcionar la intervención adecuada y el apoyo necesario a los jóvenes en situación de ansiedad.

Los datos que se emplean en este proyecto provienen de las interacciones que los jóvenes realizan en una plataforma virtual donde puedan establecer diálogos confiables sobre la vida, el amor y el sexo para prevenir casos de abuso sexual y violencia.



A partir de estos datos se construyen modelos de *Machine Learning*, o aprendizaje automático. Estos modelos son útiles para este caso por tres motivos. En primer lugar han demostrado que pueden ser una herramienta poderosa e innovadora en la detección temprana de estos fenómenos (Reynolds, Kontostathis y Edwards (2011), Dadvar, Trieschnigg y De Jong (2014), Pestian et al. (2016), Hani et al. (2019), Azeez et al.(2021)). Esto a raíz de que utilizando algoritmos de machine learning se pueden analizar datos del tipo texto, como lo son comentarios, para identificar patrones y tendencias ocultas que permitan priorizar la atención sobre los jóvenes que más lo necesitan e identificar posibles pistas que hagan alusión a qué siente o piensa el joven.

Asimismo, los modelos de machine learning son útiles puesto que no requieren de un trabajo operativo y costoso cada vez que se quisiera hacer una predicción respecto a la situación baja o alta de vulnerabilidad. Lo contrario sucede cuando el análisis de estas interacciones se realiza de manera manual, lo que representa un arduo trabajo operativo y costoso.

La tercera ventaja consiste en que los modelos se pueden emplear para identificar nuevos casos potenciales de vulnerabilidad para estas tres situaciones para otros jóvenes que compartan sus mismas características y que interactúen en esta plataforma virtual sin tener que tomar el instrumento.

En este informe, se discutirán los resultados de emplear modelos de *machine learning* para predecir la probabilidad de encontrarse en una alta o baja cantidad de situaciones de i) ansiedad, depresión, estrés post traumático y efectos sobre autoestima (de ahora en adelante, ansiedad), ii) victimización por intimidación (de ahora en adelante, victimización) y/o iii) normalización de las violencias (de ahora en adelante, violencia), sobre jóvenes que asisten a colegios públicos entre las edades de 12 y 18 años en la ciudad de Medellín y corregimientos aledaños. Se encuentra que los modelos de *machine learning* pueden ser una herramienta valiosa para identificar y abordar estos problemas de manera temprana, lo que puede tener un impacto significativo en los programas de atención que se diseñan al interior de los



colegios, en la salud mental y emocional de los jóvenes, y en la prevención de problemas más graves a largo plazo.

Adicional a esta introducción, este documento cuenta con cinco secciones más. La siguiente describe el marco teórico con artículos relevantes sobre el tema. Posterior a eso, se encuentra la metodología empleada para predecir la probabilidad de encontrarse en una situación de i) ansiedad, depresión, estrés post traumático y efectos sobre autoestima, ii) victimización por intimidación y/o iii) violencia. En la cuarta sección se describen los datos empleados. La quinta sección describe los resultados de los modelos y el análisis de estos en relación con las variables que se relacionan o caracterizan a los jóvenes que se encuentran con una alta probabilidad de encontrarse en alguna de estas tres situaciones de vulnerabilidad. En la última sección se presentan comentarios y recomendaciones.



Marco teórico

El uso de modelos de aprendizaje automático (*machine learning*) para predecir si un joven se encuentra en una situación de vulnerabilidad es una tendencia reciente. La mayoría de los estudios (ver Tabla 1) se concentran en clasificar comentarios o usuarios como bullying y no bullying dentro de un contexto virtual (cyberbullying). Por ende, gran parte de los estudios emplean datos provenientes de redes sociales como YouTube y Twitter.

La magnitud de los datos que utilizan varía dependiendo si reportan el número de comentarios o la cantidad de personas. Por ejemplo, en términos de comentarios el rango está entre 1608 conversaciones (Hani et al, 2019) y 350.000 tweets (Azeez et al., 2021). Si se tiene en cuenta el número de personas hay una gran diversidad, por ejemplo, (Pestian et al., 2016) emplean datos de 92 pacientes con información de 84000 encuentros entre pacientes y doctores, Dadvar, Trieschnigg y De Jong (2014) emplean 54050 comentarios realizados por 3825 personas en la plataforma de Youtube; mientras que, Reynolds, Kontostathis y Edwards (2011) cuentan con datos para 18,554 personas.

Los principales modelos empleados son de clasificación puesto que se busca predecir si un comentario, por ejemplo, se puede considerar como bullying o no bullying. En otras palabras, los modelos de clasificación del tipo binarios o dicotómicas. Además, la mayoría de los modelos corresponden a árboles de decisión, bosques aleatorios, máquina de soporte vectorial, *Naive Bayes*, redes neuronales y regresión logística.

Generalmente, se prueban varios modelos y se selecciona el mejor modelo empleando métricas como exactitud, precisión o sensibilidad.



Imagen 1. Matriz de confusión para variables dicotómicas

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente: *RPubs - Matriz De Confusión - Evaluación De Modelos De Predicción*, n.d.

La Tabla 1 describe las fórmulas utilizadas para el cálculo de cada métrica de acuerdo con (Arce, 2022):

Tabla 1. Métricas obtenidas de la matriz de clasificación

Métrica	Fórmula	Descripción
Exactitud ó Accuracy	$\frac{(VP + VN)}{(VP + FP + FN + VN)}$	Proporción de predicciones correctas que el modelo hizo en la muestra de prueba.
Precisión	$\frac{(VP)}{(VP + FP)}$	Porcentaje de casos positivos detectados.
Sensibilidad ó Recall	$\frac{(VP)}{(VP + FN)}$	Proporción de casos positivos que fueron correctamente clasificadas.
Valor-F ó F1-Score	$\frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$	Resume la precisión y sensibilidad en una sola métrica.

Fuente: Arce (2022)

Se pueden encontrar diversos resultados en las investigaciones donde cada una puede escoger una métrica diferente para seleccionar el mejor modelo, y esto dificulta realizar comparaciones entre los estudios. Por ejemplo, Azeez et al. (2021) encontrando que el mejor



modelo era un bosque aleatorio con una exactitud de 77%, una precisión del 73% y una sensibilidad del 94%. Mientras que, Hani et al. (2019) encuentra para redes neuronales una precisión promedio de 91,76%. Muneer y Fati (2020) encontraron un nivel de precisión similar empleando un modelo de regresión logística, con una precisión media de alrededor del 90,57% y teniendo como base 37,373 twitters.

En síntesis, no se encontró en la literatura estudios que empleen comentarios de jóvenes bajo una dinámica pedagógica para predecir si un joven se encuentra expuesto a una alta o baja cantidad de situaciones de vulnerabilidad del tipo ansiedad, violencia y/o por una situación de victimización. De igual forma, son pocos los estudios en donde se analizan los resultados más allá de la evaluación de las métricas de los modelos utilizados. La cantidad de datos con las que se cuenta en el proyecto (ver sección Datos) dista de manera significativa frente a otros estudios, siendo similar a estudios como Pestian et al., (2016), Alanazi y Alves-Foss (2020) y Chavanne et al., (2023). Finalmente, siguiendo la línea de modelos utilizados se utilizan bosques aleatorios, Naive Bayes, SGBosting y Elastic Net.



Tabla 2. Estudios

Año	Título	Algoritmos empleados	Datos de entrada	Datos de salida
2011	Using Machine Learning to Detect Cyberbullying	Máquina de soporte vectorial (SVM). Árboles de decisión. K - vecinos más cercanos (KNN).	Texto de la página Formspring.me.	Clasifica los casos positivos y negativos de cyberbullying.
2014	Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies	Sistema de evaluación Multi - Criterio (MCES). de Naive Bayes. Arbol de decisión. Máquina de soporte vectorial (SVM). Híbrido entre resultados de MCES y los algoritmos de clasificación.	Datos recolectados sobre cuentas de usuarios de Youtube, como su nick, su actividad, comentarios realizados, entre otros.	Clasifica a los usuarios como "Acosadores" y "No acosadores".
2015	Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network	Frecuencia de término - Frecuencia inversa de documento (TF - IDF). N-grama. Test Chi cuadrado. Máquina de soporte vectorial (SVM). Regresión logística.	Texto proveniente de comentarios de distintas fuentes.	Clasifica los textos en "Ofensivo" y "No Ofensivo".



2016	A Controlled Trial Using Natural Language Processing to Examine the Language of Suicidal Adolescents in the Emergency Department	Máquina de soporte vectorial (SVM).	Datos recolectados de pacientes con un Índice de Confiabilidad (IR) destacado.	Clasifica posibles usuarios suicidas.
2016	Automatic Detection of Cyberbullying on Social Networks based on Bullying Features	Máquina de soporte vectorial (SVM).	Publicaciones de Twitter.	Clasifica los textos en "Bullying" y "No Bullying".
2017	A Bullying-Severity Identifier Framework Based on Machine Learning and Fuzzy Logic	Máquina de soporte vectorial (SVM). Lógica difusa.	Comentarios de redes sociales de estudiantes.	Clasifica la existencia de Bullying, así como sus participantes y sus formas.
2017	Deep Learning for Hate Speech Detection in Tweets	Frecuencia de término - Frecuencia inversa de documento (TF - IDF). Clasificador mediante redes neuronales (CNN). Redes neuronales LSTM.	Publicaciones de Twitter.	Clasifica las palabras que son usadas en los discursos de odio.



2019	Automatic Cyber Bullying Detection in Arabic Social Media	Léxico de sentimientos mediante aproximación PMI, Chi cuadrado y Entrópica.	Texto de diferentes fuentes y redes sociales, como Youtube, Twitter, Microsoft Flow, etc.	Clasifica los textos en "Bullying" y "No Bullying".
2019	Social Media Cyberbullying Detection using Machine Learning	Máquina de soporte vectorial (SVM) y redes neuronales	Texto de Formspring	Clasifica los textos en byllying y no bulling
2020	A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter	Regresión Logística (LR), Máquina de Refuerzo de Gradiente Ligero (LGBM), Descendente Estocástico de Gradiente (SGD), Bosque Aleatorio (RF), AdaBoost (ADB), Naive Bayes (NB) y Máquina de Vectores de Soporte (SVM).	Publicaciones de Twitter	Clasifica los textos en ofensivos y no ofensivos
2021	Cyberbullying Detection in Social Networks: Artificial Intelligence Approach	Naive Bayes, KNN, Regresión logística, Arboles de decisión, Random forest, LSVC, Stochastic Gradient,	Comentarios de Twitter	Clasifica los textos en "Bullying" y "No Bullying".



20 22	AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naïve Bayes and Optimized Linear Support Vector Machine.	Máquina de soporte vectorial (SVM). Naive Bayes Multinomial (MNB).	Texto de diferentes fuentes y redes sociales, como Youtube, Twitter, Wikipedia, etc.	Clasifica los textos en "Bullying" y "No Bullying". En caso que algún mensaje sea clasificado como "Bullying", cancela su envío.
20 22	Anxiety onset in adolescents: a machine-learning prediction	Bosques aleatorios. Máquina de soporte vectorial (SVM). Regresión logística.	Neuroimágenes	Clasifica la ansiedad clínica en los adolescentes.
20 22	Leveraging Deep Learning Models to Understand the Daily Experience of Anxiety in Teenagers over the Course of a Year	Redes neuronales LSTM.	Datos de estudiantes, en tres tiempos distintos del año escolar. Los datos se recolectaron de acuerdo a la escala de ansiedad generalizada (GAD).	Predice el puntaje de ansiedad.



20 22	Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis	Redes neuronales artificiales. Bosques aleatorios. Máquina de soporte vectorial (SVM). Árboles de decisión. Naive Bayes.	Datos sobre hábitos de salud de estudiantes de diferentes grados y edades.	Clasifica la depresión y la ansiedad en "Normal", "Moderado" y "Severo".
----------	--	---	--	---

Fuente: Elaboración propia

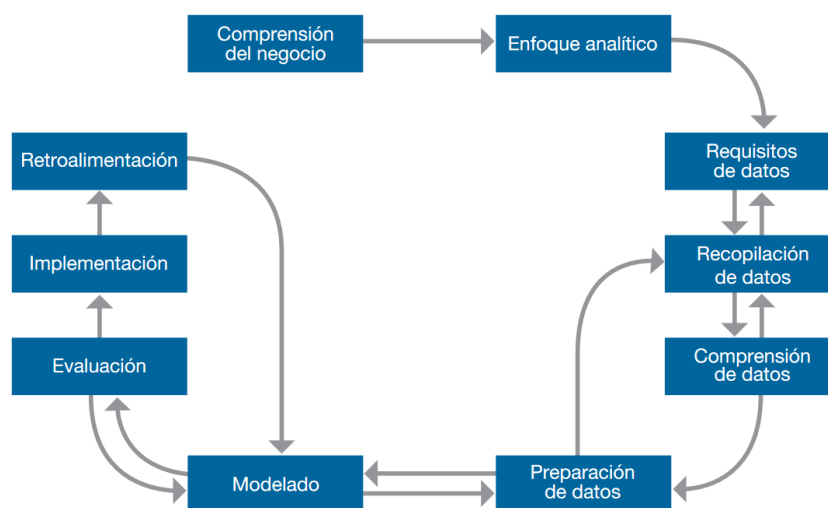
Nota: Ver referencias bibliográficas.



Metodología

En este proyecto se sigue la metodología de *Data Science* de IBM (ver imagen 2). A continuación, se describe lo realizado en cada una de las etapas.

Imagen 2. Metodología de *Data Science*



Fuente: IBM

1.1 Comprensión del negocio y enfoque analítico

El objetivo de este proyecto es poder determinar si un joven se encuentra expuesto a una alta o baja cantidad de situaciones de vulnerabilidad. Para lograr este objetivo desde el enfoque analítico se eligen tres tipos de vulnerabilidad: A. Estar en una situación de victimización por intimidación, B. Estar en una situación de normalización de violencia, y C. Estar en una situación de depresión, ansiedad, estrés post traumático y efectos sobre autoestima.

Los datos que permiten predecir si un joven está expuesto a una alta o baja cantidad de situaciones de vulnerabilidad provienen de un instrumento que los jóvenes toman al inicio del



curso “Entre pares nos cuidamos (niñas, niños y jóvenes abordando estrategias pedagógicas para el cuidado de sí y de los otros)” provisto en la plataforma de Mangus, y las variables independientes se construyen a partir de la caracterización sociodemográfica y los comentarios que escriben en las actividades de las lecciones del curso.

1.2 Recopilación de datos

Se utilizan 3 fuentes de datos: caracterización de los estudiantes, resultados del instrumento de medición de vulnerabilidad y comentarios de los jóvenes en las diferentes lecciones. Estas tres se encuentran en la plataforma provista por Mangus y corresponden a datos recolectados entre enero y abril de 2023. Un mayor detalle descriptivo de los datos se encuentra en la sección Datos.

1.3 Compresión y preparación de datos

1.3.1 Procesamiento y limpieza

La etapa de procesamiento y limpieza consiste en términos generales en la preparación y transformación de los datos para su posterior análisis. Esto implica, por ejemplo, identificación y manejo de datos faltantes, la eliminación de duplicados, la normalización de variables, y la codificación de variables categóricas. A continuación, se describe las actividades realizadas para cada una de las fuentes:

1.3.1.1 Caracterización

- Retirar a los jóvenes que sin id de ingenia.
- Calcular la edad de los jóvenes empleando la fecha de nacimiento. Si no tenía fecha de nacimiento, pero había diligenciado el campo de edad, se dejó este valor.
- Retirar jóvenes por fuera del rango de 12 a 18
- Retirar jóvenes sin sexo



- Retirar jóvenes sin estrato

Este proceso de limpieza llevó a pasar de tener 538 jóvenes a 484 jóvenes.

1.3.1.2 Instrumento

- Eliminar individuos que no fuera estudiantes
- Eliminar individuos que no hubieran respondido todas las preguntas del cuestionario
- Eliminar individuos que no tuvieran información completa en la caracterización

Este proceso de limpieza llevó a pasar de tener datos de instrumento para 447 jóvenes a 350 jóvenes.

1.3.1.3 Comentarios

- Eliminar comentarios duplicados
- Unificar comentarios por usuario y lección
- Eliminar números, puntuación y símbolos
- Pasar todo a minúsculas
- Eliminar tildes
- Reemplazar la ñ por n
- Eliminar stopwords, palabras que no aportan ningún significado (e.j. “yo”, “eso”, “la”, “cual”, “más”, “que”)
- Eliminar patrones de letras (e.j. “aaaaa...”, “jajaja...”, “lalala...”)
- Eliminar palabras relacionadas con las preguntas (e.j. nombres de canciones)
- Lematización (e.j. “cuidado”, “cuidada”, “cuidadoso”, “cuidadosa” = “cuidar”). El diccionario de lematización se construyó manualmente.
- Eliminar nombres de los estudiantes
- Filtrar por las lecciones con id: 132,135,142,143,144 y 145



Este tipo de limpieza se ha empleado en estudios como Chavan y SS (2015), Kwartler (2017), Zhao et al. (2016) y AlHarbi et al. (2020).

A continuación, un ejemplo de cómo un comentario se transformó con la limpieza:

Tabla 3. Ejemplo de limpieza de comentario

Antes de la limpieza	Después de la limpieza
1.yo me cuido siempre estando acompañado por un adulto o persona que nos pueda cuidar,si es el caso de estar solos siempre tener un instrumento de proteccion como:gas pimienta,tener en el celular previamente marcado 123.	siempre ir acompañado adulto poder cuidar casos soledad siempre tener instrumento proteccion gas pimienta tener celular previamente marcado tener camara seguridad siempre contacto
2.recomiendo a mi familia y a mi escuela que para cuidarnos deberian tener camaras de seguridad o siempre tener un docente en los lugares mas remotos de la institucion,para la familia es estar con todo momento en contacto con ellos.	recomendar familia colegio cuidar deber tener camara seguridad siempre tener docente lugares remoto colegio familia momentos contacto

Fuente: Elaboración propia

1.3.2 Unión de las tres fuentes de datos

Posterior a la limpieza, se realiza el cruce de datos entre caracterización e instrumentos empleando la columna de “identificación”, de esto resulta una tabla para 306 jóvenes. Posterior a esto, se cruza con los datos de comentarios empleando la columna de “id_usuario” y se obtienen datos para **274** jóvenes.



La distribución de los jóvenes por lección se observa en la Tabla 4.

Tabla 4. Número de jóvenes por lección seleccionada

Lección	Pregunta	No. Obs.
Caperucita roja	¿Qué estrategias se nos ocurren para cuidarnos y cuidar a otras niñas, niños y jóvenes, ante posibles violencias? ¿Qué mensaje le darías a la escuela y a la familia para fortalecer los entornos protectores?	228
Me llegó carta	Luego de leer la carta de Charles Chaplin, resalta las palabras que más te gustaron relacionadas con el cuidado y el amor propio...	221
Seis claves	Te invitamos a responder las claves para la reflexión, de acuerdo con el infográfico. 1.¿Cómo podemos identificar riesgos y violencias para tomar medidas de protección?...	206
Introducción	¿Qué tan cómodo me siento en espacios como la familia y la escuela conversando sobre mi identidad de género?	194
Reflexionemos nuestro entorno	Escucha las canciones que te hemos propuesto y sugiere, a partir de tus reflexiones, el material que conoces para la prevención de las violencias y la movilización ante la misma en los diferentes entornos.	174
Reconociendo otras vivencias	Selecciona una imagen de la galería y a partir de ella cuéntanos una anécdota sobre situaciones similares que hayas vivido o visto en tu contexto escolar...	172

Fuente: Elaboración propia



1.3.3 Creación de variables

1.3.3.1 Variable dependiente

Para la construcción de la variable dependiente se utilizan los datos del instrumento, en la Tabla 5 se pueden observar las secciones utilizadas para cada uno de los modelos y en el Anexo I se encuentran todas las situaciones que cubrían cada una de las secciones seleccionadas.

Tabla 5. Sección utilizada del instrumento para la construcción de la variable dependiente

Modelo	Sección utilizada del instrumento
Ansiedad	Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima
Victimización	Situaciones de victimización por intimidación
Violencia	Normalización de las violencias

Fuente: Elaboración propia

Para definir como alto o bajo en el número de situaciones en los que se encuentra o se ve expuesto el estudiante para cada uno de los tipos de vulnerabilidad seleccionados se utilizó una aproximación frecuentista. En este caso se definió un umbral de qué porcentaje de las respuestas deberían estar en Nunca y/o Casi Nunca para que el estudiante estuviera calificado con una baja exposición o si se encontraba en una baja cantidades de situaciones para cada uno de los tipos de vulnerabilidad (ver Tabla 6).

Para construir la variable dependiente del modelo relacionado a situaciones de normalización de la violencia no se tuvieron en cuenta las preguntas 38 (Intervengo cuando hay una



situación de agresión, acoso e intimidación), 39 (Denuncio ante adultos una situación de agresión, acoso e intimidación), y 41 (Veo atención o acciones contundentes ante una situación de agresión, acoso e intimidación) porque estas son preguntas que en caso de responder “Casi Siempre” o “Siempre” se considera algo positivo para el estudiante y segaría la metodología de cálculo de la variable dependiente. La pregunta 43 (Por ser hombre debo ser tolerante ante situaciones de agresión, acoso e intimidación) solo se tiene en cuenta para los hombres.

Tabla 6. Definición del nivel bajo y alto por modelos

Nivel	Ansiedad	Victimización	Violencia
BAJO	Si el 90% o más de las preguntas marca Nunca o si la suma de Nunca y Casi nunca suma 90% o más	Si el 90% o más de las preguntas marca Nunca, o si la suma de Nunca y Casi nunca suma 83% o más, y Siempre no tiene ninguna marca	Si el 90% o más de las preguntas marca Nunca, o si la suma de Nunca y Casi nunca suma 87% o más
ALTO	Si no cumple con la condición de ser bajo, entonces es Alto.		

Fuente: Elaboración propia

1.3.3.2 Variables independientes

Las variables independientes son más de 400. Estas se agrupan en distintos conjuntos de acuerdo con sus características. A continuación se describen cada uno:

- Variables sociodemográficas

En este caso se utilizaron algunas de las variables provenientes del archivo de caracterización como estrato, etnia y sexo (ver Tabla 7).



Tabla 7. Variables sociodemográficas

Variable	Valores	Observaciones
estrato	Categórica: estrato 1, estrato 2, estrato 3, estrato 4, estrato 5 y otro	
etnia	“N” o “S”,	“S” si pertenece a un grupo étnico (Afrocolombiano(a), Indígena, Negro(a), Palenquero(a), Raizal u Otro) y “N” si no pertenece.
edad	numérico entre 12 y 18	
sexo	“m”, “h”, “otro”	“m” pertenece a femenino, “h” pertenece a masculino y “otro” pertenece a cualquier otro valor.

Fuente: Elaboración propia

- Variables relacionadas con las características de los comentarios

En este caso se calcularon nuevas variables a partir de la estructura de los comentarios en términos de número de caracteres y palabras.

Tabla 8. Características de los comentarios

Variable	Descripción
n_caracteres	Número de caracteres por comentario por lección por joven. Estas 6 variables se calculan antes de la limpieza.



n_palabras	Número de palabras por comentario por lección por joven. Estas 6 variables se calculan antes de la limpieza.
dif_nchar_median	Diferencia entre el número de caracteres por comentario por lección respecto a la mediana del número de caracteres en esa lección. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_npal_median	Diferencia entre el número de caracteres por comentario por lección respecto a la mediana del número de caracteres en esa lección. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_nchar_median_ans	Diferencia entre el número de caracteres por comentario por lección respecto a la mediana del número de caracteres en esa lección para los bajos en ansiedad. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_npal_median_ans	Diferencia entre el número de palabras por comentario por lección respecto a la mediana del número de palabras en esa lección para los bajos en ansiedad. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_nchar_median_vict	Diferencia entre el número de caracteres por comentario por lección respecto a la mediana del número de caracteres en esa lección para los bajos en victimización. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_npal_median_vict	Diferencia entre el número de palabras por comentario por lección respecto a la mediana del número de palabras en esa lección para los bajos en victimización. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_nchar_median_vio	Diferencia entre el número de caracteres por comentario por lección respecto a la mediana del número de caracteres en esa lección para



	los bajos en violencia. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.
dif_npai_median_vio	Diferencia entre el número de palabras por comentario por lección respecto a la mediana del número de palabras en esa lección para los bajos en violencia. La mediana se calcula antes de la limpieza. Y estas también son 6 variables.

Fuente: Elaboración propia

A continuación se muestra la mediana en el número de caracteres y palabras por lección se puede observar en la Tabla 9.

Tabla 9. Cálculo de la mediana del número de caracteres y palabras por lección

Lección	Mediana	Mediana
	Número de caracteres	Número de palabras
132	299	57
135	133	25
142	228	39
143	255.5	45
144	145	26
145	159.5	31

Fuente: Elaboración propia

- Variables de presencia de palabras con mayor frecuencia por lección

Para construir estas variables se utilizan modelos de n-gram¹, con n igual a 1 y n igual a 2 . Posterior a eso, se calcula la frecuencia de los unigramas y bigram por lección.

¹ Los N-gramas son un grupo de n-elementos de un texto organizados en una secuencia continua que se utiliza para dividir el texto en n fragmentos. Por ejemplo, la oración "Usted es gracioso" se podría dividir en su unigrama como "usted", "es", "gracioso", y el bigram sería "usted es", "es gracioso".



Para el caso de los unigramas se obtuvo el top 20 de palabras más frecuentes por lección (ver Tabla 10 para un ejemplo). Para el caso de bigramas, se obtuvo el top 15 (ver Tabla 11 para un ejemplo) por lección. Finalmente, se crean las variables correspondientes que pueden tomar un valor de 1 o de 0. El 1 indica que el comentario del joven para esa lección contiene el unigrama o bigrama, y el 0 indica que no se encontró.

Tabla 10. Top 20 de palabras más frecuentes de la lección 132

Lección	Palabra	Total
132	sentir	222
132	feliz	190
132	hacer	174
132	vida	119
132	poder	117
132	ser	116
132	bien	113
132	cosas	104
132	querer	104
132	mismo	86
132	tristeza	77
132	amistad	74
132	gustar	74
132	familia	71
132	malo	66
132	siempre	62
132	saber	57
132	bueno	56



132	mejor	56
132	decir	55

Fuente: Elaboración propia

Tabla 11. Top 15 de los bigramas más frecuentes de la lección 132

Lección	Bigrama	Total
132	sentir bien	43
132	momentos feliz	39
132	ser feliz	32
132	momentos tristeza	29
132	sentir feliz	25
132	hacer sentir	22
132	no poder	22
132	no querer	20
132	mucho cosas	19
132	feliz no	16
132	no dejar	16
132	no sentir	16
132	sin importancia	13
132	querer decir	12
132	sentir orgulloso	12

Fuente: Elaboración propia

- Variables extraídas del análisis de sentimiento



El análisis de sentimiento es una técnica de procesamiento del lenguaje natural para determinar la polaridad emocional de un texto, en este caso los comentarios de los jóvenes. Se emplea el diccionario de sentimientos de la NRC para calcular la presencia de ira, miedo, tristeza y negatividad. Posterior a esto, se calcula la mediana de la frecuencia de los sentimientos seleccionados para el nivel bajo para cada lección (ver Tabla 12). Finalmente, se calcula la diferencia entre el valor que cada comentario tiene para cada emoción y la mediana respectiva. De igual manera, se incluye el número de palabras asociadas a las emociones de ira, miedo, tristeza y negatividad por lección.

Tabla 12. Cálculo de la mediana de los sentimientos del nivel bajo por cada lección

Lección	Miedo	Tristeza	Negatividad	Miedo	Negatividad
	Violencia	Violencia	Violencia	Ansiedad	Ansiedad
132	1	1	2	1	2
135	0	0	0	0	0
142	2	1	2	1	2
143	3	2	3.5	2	3
144	0	1	1	0	1
145	0	1	1	0	1

Fuente: Elaboración propia

- Variables valor de TF-IDF para palabras comunes entre todos los jóvenes.

TF-IDF significa "Término de frecuencia, Frecuencia Inversa del Documento". Es una forma de evaluar la importancia de las palabras en un documento basado en la frecuencia con que aparecen a través de varios documentos. Cuanto mayor sea la puntuación, más relevante será esa palabra para ese individuo en particular. Esta técnica se aplica a las palabras más frecuentes utilizadas por los estudiantes.



- Variables de frecuencia de palabras que hacen alusión a elementos positivos y negativos

Las palabras seleccionadas fueron bueno, amistad, malo, feliz, soledad, comodidad, miedo, respeto, acoso, angustia y tristeza. Se cuenta el número de veces que aparecen estas palabras para cada joven independiente de la lección.

- Variables de frecuencia de palabras alerta por lección y joven

Se obtiene una lista de palabras que podrían ser indicadores de alarma o alerta relacionadas con temáticas de ansiedad y violencia. En este caso se contaron las veces que aparecieron las palabras de del tipo violencia en la lección 143 (Seis Claves), y para las palabras de alerta de ansiedad se contaron para las lecciones 132 (Me llegó la carta), 135 (Introducción) y 145 (Reflexionemos nuestro entorno). Estas lecciones se escogieron después de hacer un análisis estadístico de diferencias en medias entre los altos y bajos para cada uno de los tipos de vulnerabilidad.

Tabla 13. Palabras alerta por tipo de vulnerabilidad

Tipo	Palabras
Ansiedad	"tristeza", "psicologo", "triste", "solo", "soledad", "desesperanza", "desmotivacion", "agotamiento", "irritabilidad", "insomnio", "pesimismo", "culpa", "estres", "mortificar", "mortificado", "desinteres", "desesperacion", "preocupacion", "negativo", "inutil", "no resaltar", "fatiga", "desinteres", "aislamiento", "suicidio", "humor", "verguenza", "no agradecer", "agujero", "dificultad", "suicida", "dificultad", "sin esperanza", "inseguridad", "desaparecer", "no escapar", "estresado", "incomodo", "miedo", "molestien", "criticas", "bullying", "temor", "angustia", "bulin", "morir", "trauma", "bullyg", "bullyn", "inseguro", "denigracion", "angustia",



	"dolio", "solo", "morir", "traumas", "mundo caer", "apodos", "sin autoestima", "baja autoestima", "ansiedad"
Violencia	"abuso", "bullying", "maltrato", "acoso", "violacion", "agresion", "amexas", "coaccion", "control", "dominacion", "marginacion", "humillacion", "exclusion", "rechazo", "discriminacion", "racismo", "xenofobia", "homofobia", "intimidacion", "chantaje", "manipulacion", "aislamiento", "ignorada", "burlas", "gritos", "ofensas", "insultos", "agresiones", "violencia", "temor", "acosan", "maltrato", "angustia", "dolio", "doler", "miedo", "vulnerable", "traumas"

Fuente: Elaboración propia

1.3.4 Selección de variables

Teniendo en cuenta que el número de variables independientes es mayor al número de observaciones (estudiantes) se emplea una técnica para hacer una preselección de variables independientes que tienen una mayor correlación con las variables dependientes. Para esto se usa un método estadístico de hipótesis conocido como "Test de Chi Cuadrado", el cual es una prueba no paramétrica, y donde se usa el estadístico de prueba:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

Este estadístico sigue una distribución χ^2 con 1 grado de libertad cuando se realiza una única comparación. O_{ij} es el valor observado de individuos en un nivel de vulnerabilidad y E_{ij} es el valor esperado de individuos en un nivel de vulnerabilidad para esa misma característica. El nivel de confianza usado para seleccionar las variables es del 85%. Este método permite contrastar la hipótesis nula de que la variable independiente no está relacionada con la



variable dependiente, frente a la hipótesis alterna de que no se cumple lo supuesto en la hipótesis nula. Por lo tanto, cuando se rechaza la hipótesis nula, se decide seleccionar a la variable independiente.

Adicionalmente, se retiran las variables que resultan ser una constante para todos los jóvenes.

Tabla 14. Número de variables seleccionadas después de aplicar el método de Test de Chi Cuadrado

Modelo	Número de variables independientes iniciales	Número de variables seleccionadas
Ansiedad	434	43
Victimización	434	44
Violencia	434	41

Fuente: Elaboración propia

Las pruebas estadísticas, al igual que los modelos, son realizadas con el software R (R Core Team, 2016).

1.3.5 Estandarización de variables numéricas

Este paso se realiza para que todas las variables numéricas tengan un rango común, de esa forma son comparables y son tratadas con la misma escala. Para esto se realizan los siguientes pasos:

- **Centrado:** Se resta la media de la variable a todos los valores, lo que resulta en una distribución con media cero. Esto se realiza para eliminar la diferencia de escala y asegurarse de que todos los valores numéricos están centrados alrededor del cero.



- **Escalamiento:** Se divide cada valor de la variable centrada por la desviación estándar de la variable. Esto hace que la desviación estándar sea igual a uno. Esto se hace para asegurarse de que los datos estén en la misma escala y tengan una variabilidad similar.

1.3.6 Separación de los datos en entrenamiento y prueba

Los datos se separan en el conjunto de entrenamiento² (80% de los datos) y prueba³ (20%) para cada uno de los tres modelos. Esta es una práctica común que consiste en emplear los datos de entrenamiento para enseñarle al modelo cómo hacer predicciones correctas, y luego se utilizan los datos de prueba para evaluar si el modelo puede hacer predicciones precisas para jóvenes cuyos datos no ha visto con anterioridad. Este paso es importante para poder evaluar el rendimiento del modelo de manera objetiva y evitar el sobreajuste.

1.4 Modelos

Para este proyecto se emplean 4 modelos:

- **Random Forest o Bosques aleatorios**

Este es un modelo de ensamble que combina varios árboles de decisión para realizar la clasificación. Cada árbol se entrena con un subconjunto aleatorio de datos y variables independientes, y luego sus predicciones se combinan para obtener la predicción final.

² El conjunto de entrenamiento se utiliza para entrenar el modelo, es decir, para ajustar los parámetros de los modelos. El objetivo es que el modelo aprenda patrones y relaciones en los datos de entrenamiento para poder hacer predicciones precisas en vulnerabilidad cuando se tengan datos para nuevos jóvenes.

³ El conjunto de prueba se utiliza para evaluar el rendimiento del modelo una vez ha sido entrenado, dado que son datos que el modelo no ha visto antes. Si el modelo tiene un buen rendimiento en el conjunto de prueba, es un indicio de que el modelo es capaz de hacer predicciones precisas en datos nuevos de jóvenes sobre los cuales no se conoce el nivel de vulnerabilidad (no se les ha aplicado el instrumento).



- **Naive Bayes**

Es un modelo probabilístico basado en el teorema de Bayes. Este se entrena calculando las probabilidades de las variables independientes en las diferentes clases y luego utiliza estas probabilidades para predecir la clase de un nuevo dato basándose en la probabilidad condicional.

- **Stochastic Gradient Boosting (SGBoosting)**

Es un tipo de algoritmo de machine learning que pertenece a la familia de métodos de ensamble. Este modelo combina múltiples modelos de aprendizaje débil, como árboles de decisión, en un modelo fuerte y más preciso. Su principal característica es que utiliza una técnica de muestreo aleatorio para entrenar los árboles de decisión, lo que lo hace más eficiente en términos de tiempo y memoria en comparación con otros métodos de boosting. El SGBoosting mejora iterativamente la predicción del modelo al corregir los errores del modelo anterior en el proceso de entrenamiento.

- **Elastic - Net Regularized Generalized Linear Model**

Este modelo combina tres técnicas: la regresión lineal generalizada, la regularización L1 (Lasso) y la regularización L2 (Ridge). La regresión lineal generalizada es una técnica que permite predecir una variable dependiente a partir de variables independientes. La regularización L1 y L2 son técnicas que ayudan a evitar el sobreajuste y mejoran la capacidad de generalización del modelo.

Adicionalmente, se emplea un método de validación cruzada repetida. En este caso los datos de entrenamiento se dividen en 10 partes, donde una parte se emplea para validación y con las otras 9 se entrena el modelo. Esto se hace para 10 iteraciones y se repite 5 veces. Esta validación se realiza para disminuir el sobreajuste de los modelos. Adicionalmente, el método de selección de hiper parámetros de los modelos se hace en forma de grilla con el objetivo de mejorar el nivel de exactitud. Este procedimiento ayuda a obtener una evaluación más robusta



del desempeño de los modelos y a encontrar los mejores hiper parámetros para cada uno de estos.

Finalmente, se calcula la matriz de confusión para cada uno de los modelos y se selecciona aquel que tiene la mejor sensibilidad (ver Tabla 1). Se selecciona la métrica de sensibilidad pues se quiere escoger aquel modelo que logre clasificar en alto al mayor porcentaje de jóvenes que reconocer encontrarse en lo que se ha denomina alta cantidad de situaciones de vulnerabilidad, de acuerdo con los resultados del instrumento empleado.



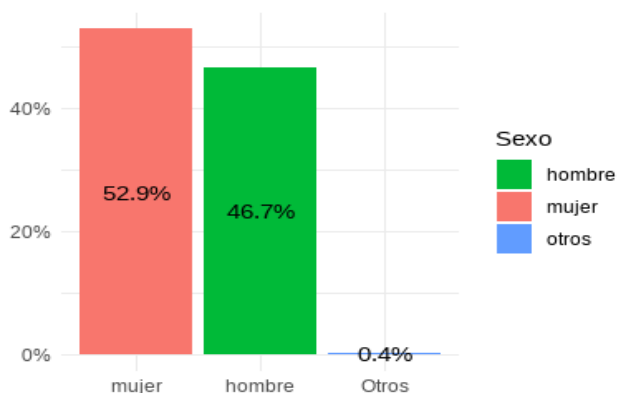
Datos

A continuación se describen los datos de los 274 usuarios con los que se realizó el entrenamiento de los modelos. Primero, se describen las variables sociodemográficas y después las secciones del instrumento, el cual se empleó para construir la variable dependiente de cada modelo.

1. Descripción de variables sociodemográficas

Del total de usuarios, el 52.92 % son mujeres, en su mayoría entre las edades de 15 (17.52 %) y 16 (13.14 %) años. Por otro lado, el 46.72 % son hombres, en su mayoría entre las edades de 15 (14.96 %) y 16 (9.12 %) años (ver Gráfica 1).

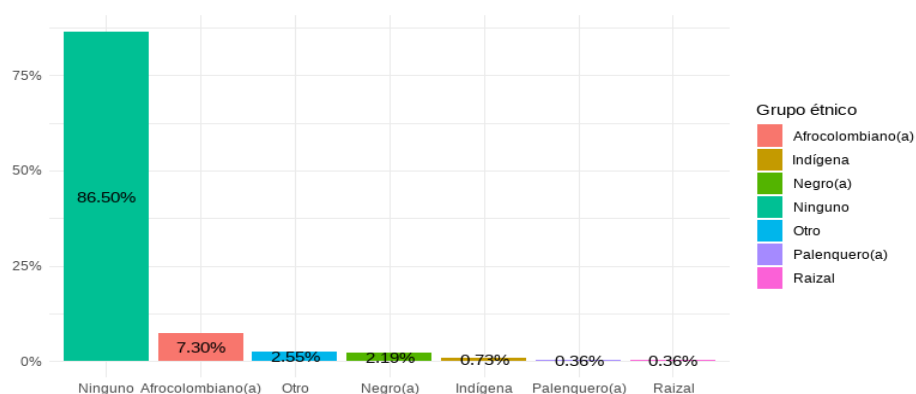
Gráfico 1. Distribución de la población estudiada por sexo.



Fuente: Elaboración propia

El 86.5 % de los usuarios no se reconocen como parte de una etnia. Seguidamente, el (7.3 %) de los usuarios se reconoce como Afrocolombiano(a) (ver Gráfica 2).

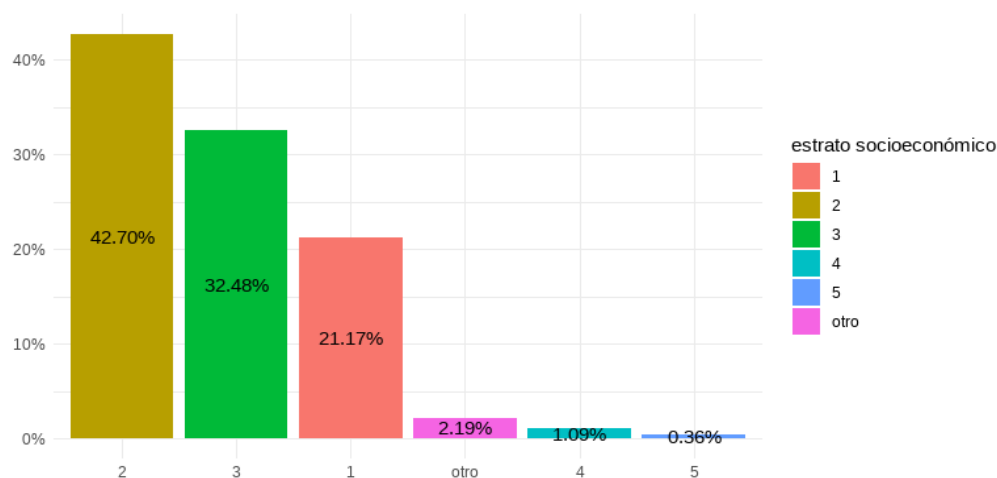
Gráfico 2. Distribución de la población estudiada por etnia.



Fuente: Elaboración propia

El mayor porcentaje de los usuarios 42.7 % pertenece al estrato 2 , seguido del estrato 3 al cual pertenece el (32.48 %) de los usuarios (ver Gráfica 3).

Gráfico 3. Distribución de la población estudiada por estrato.

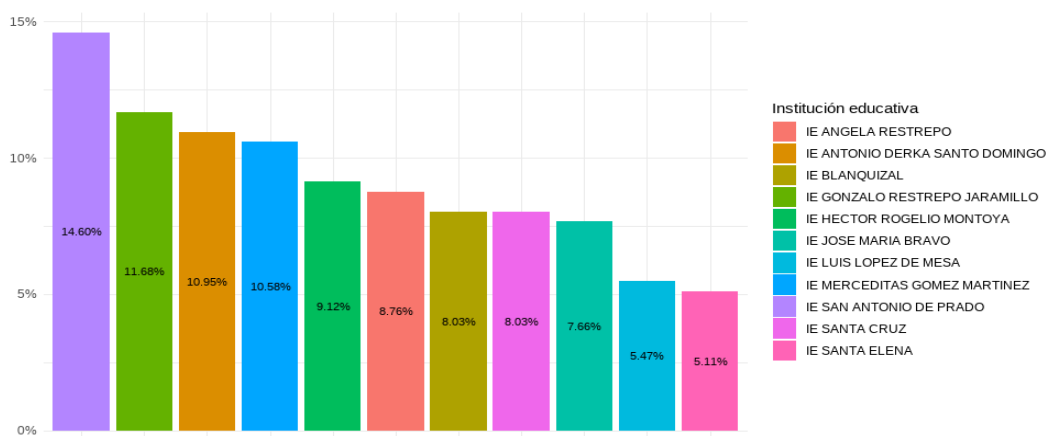


Fuente: Elaboración propia

En cuanto a la institución educativa a la que pertenecen los usuarios, el estudio se realiza con usuarios de 11 instituciones educativas, de los cuales aproximadamente el 14.6 % pertenece a IE San Antonio de Prado, seguido de la IE Gonzalo Restrepo Jaramillo a la que pertenece el 11.7 % de los usuarios (ver Gráfica 4).



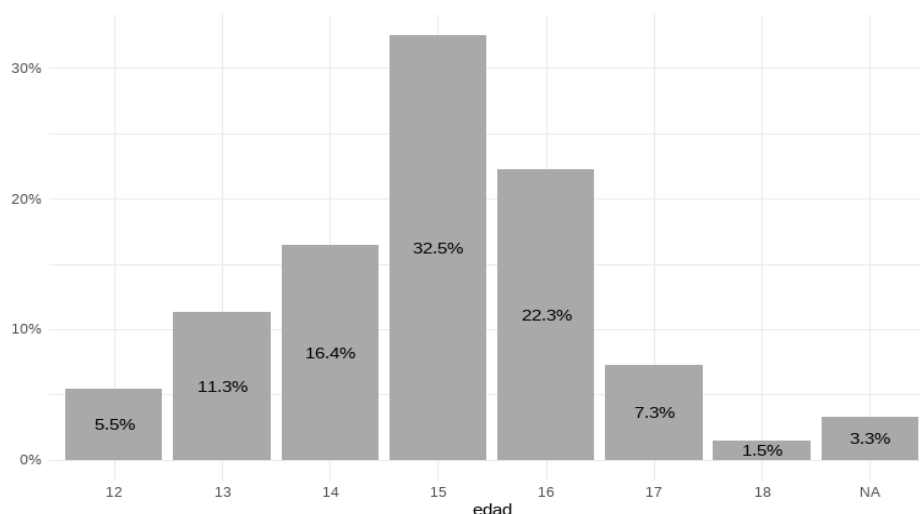
Gráfico 4. Distribución de la población estudiada por institución educativa.



Fuente: Elaboración propia

En cuanto a la edad de los participantes, estos se encuentran en un rango que va desde los 12 hasta los 18 años. La mayoría de los usuarios, aproximadamente el 32.48 % de ellos, tiene 15 años. Seguidamente, un 22.26 % de los usuarios tiene 16 años (ver Gráfica 5).

Gráfico 5. Distribución de la población estudiada por edad.



Fuente: Elaboración propia

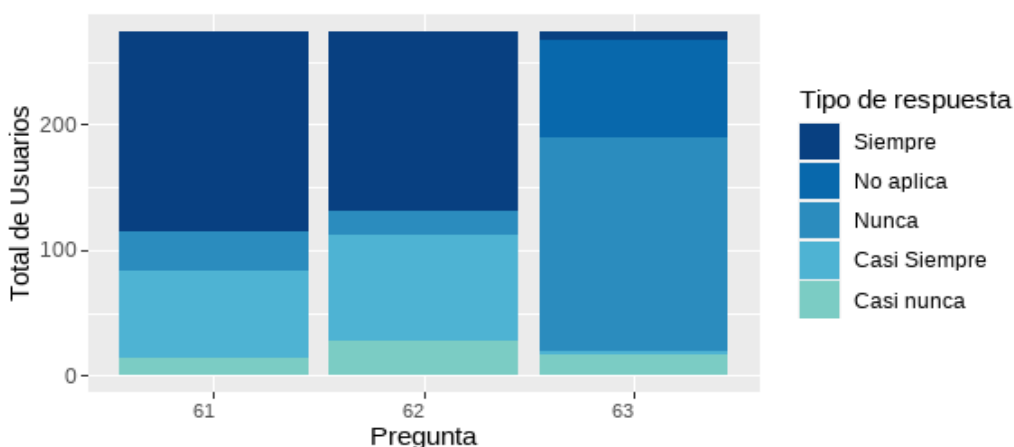
2. Descripción de las secciones del instrumento



a. Sección: Condiciones de vida digna

Para la sección alusiva a Condiciones de vida digna se observa que en la pregunta 61. Mi familia cuenta con condiciones materiales para nuestro bienestar y la pregunta 62. Mi familia es un entorno protector y promotor de mi tranquilidad y bienestar físico y emocional la respuesta que predomina es Siempre o Casi Siempre lo cual es positivo (ver Gráfica 6).

Gráfico 6. Distribución de respuestas para la categoría Condiciones de vida digna.



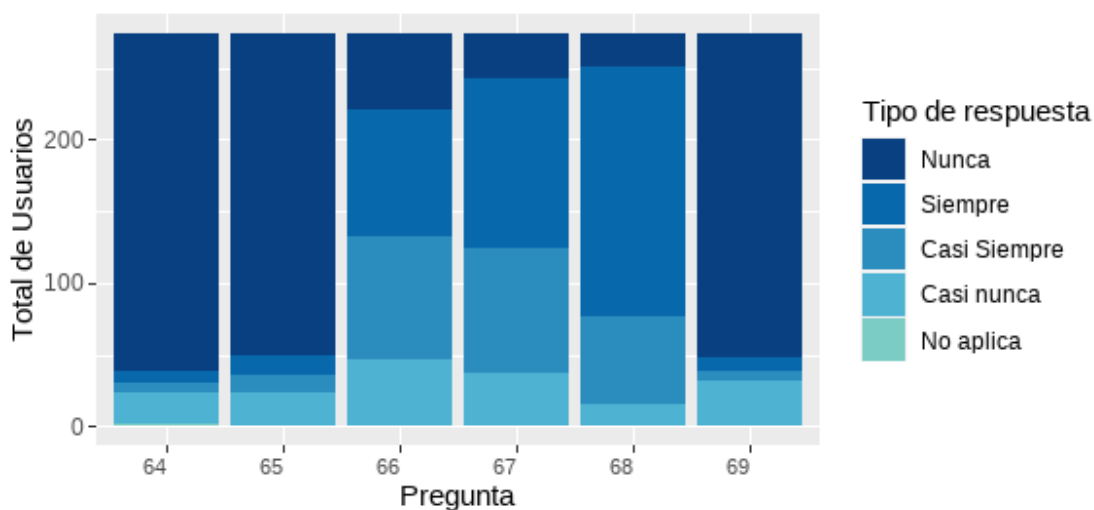
Fuente: Elaboración Propia

b. Sección: Dinámicas Familiares

En esta sección se observa que en las preguntas 66. Soy incluido en la toma de decisiones y/o en propuestas que surgen en mi familia , 67. Identifico espacios para el diálogo respetuoso con los integrantes de mi familia y 68. Los adultos de mi familia promueven el cuidado y el buen trato hacia los demás, la respuesta que predomina es Siempre lo cual es positivo (ver Gráfica 7).



Gráfico 7. Distribución de respuestas para la categoría Dinámicas familiares.



Fuente: Elaboración Propia

c. Sección: Dinámicas territoriales y familia

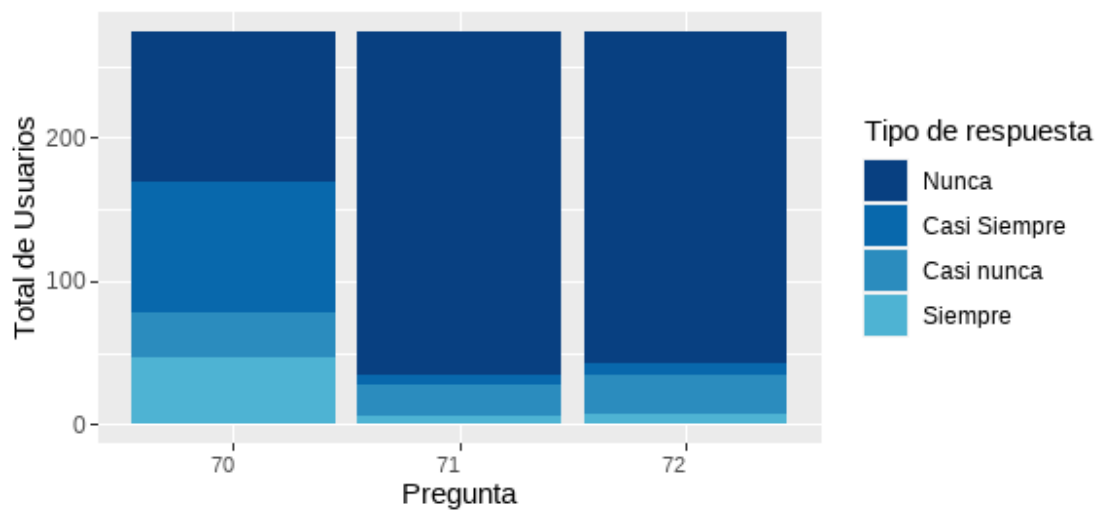
Se observa que en las preguntas 71. Creo que es normal que niñas, niños y jóvenes obtengan ingresos por la comercialización de sus cuerpos , 72. Considero que el consumo de sustancias psicoactivas debe ser permitido para niñas, niños y jóvenes, la respuesta que predomina es Nunca lo cual es positivo (ver Gráfica 8).

d. Sección: Intimidaciones por parte de respondientes

Se observa que en general, la respuesta que predomina es Nunca, lo cual es un buen indicio. Por otra parte, resalta que en las preguntas 32. Me burlo de otros y 33. Riego chismes de otros, de manera mal intencionada, la respuesta Casi nunca tiene una gran proporción, seguido de Casi Siempre lo cual es un aspecto a tener en cuenta (ver Gráfica 9).

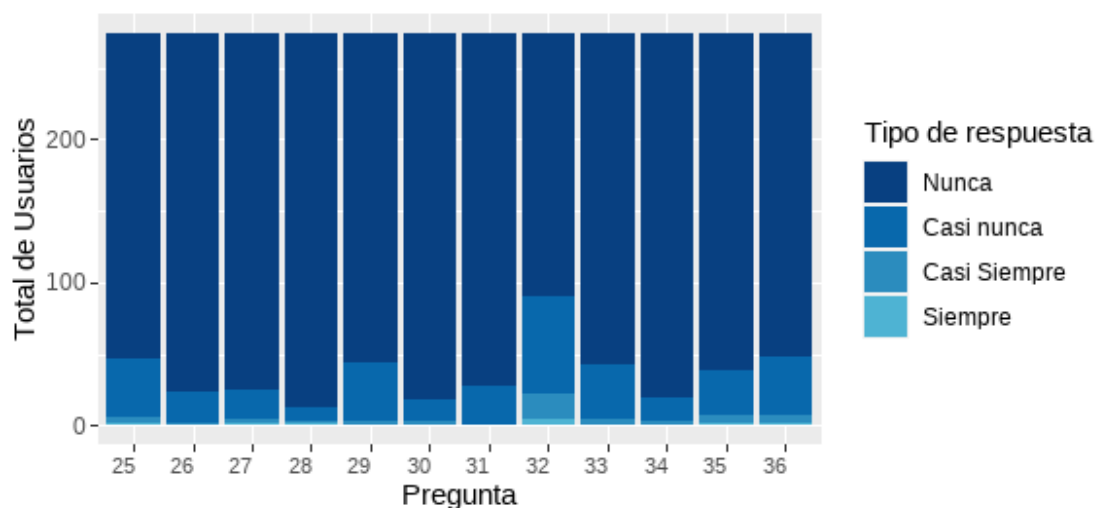


Gráfico 8. Distribución de respuestas para la categoría Dinámicas territoriales y familia.



Fuente: Elaboración Propia

Gráfico 9. Distribución de respuestas para la categoría Intimidaciones por parte de respondientes.



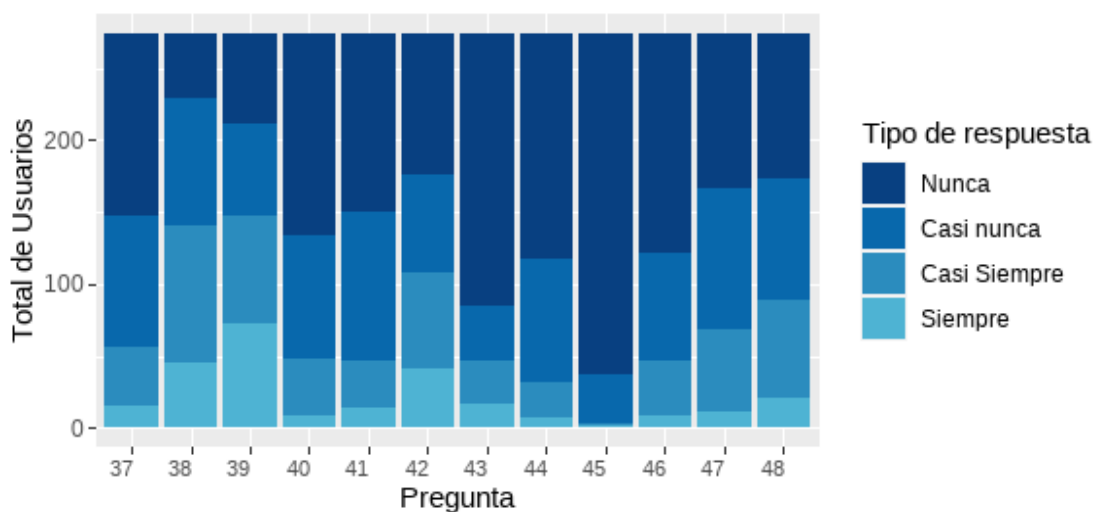
Fuente: Elaboración Propia

e. Sección: Normalización de las violencias



En esta sección se observa que en las preguntas 38. Intervengo cuando hay una situación de agresión, acoso e intimidación y 39. Denuncio ante adultos una situación de agresión, acoso e intimidación, las respuestas Casi Siempre y Siempre tienen una gran proporción, lo cual es positivo. Por otro lado, se observa que aunque en la mayoría de las preguntas la respuesta que predomina es Nunca lo cual es algo positivo, es necesario prestar atención a la proporción de Casi Siempre y Siempre presentes en todas las preguntas de esta sección del instrumento (ver Gráfica 10) y que se encuentran con una mayor proporción que para otras secciones.

Gráfico 10. Distribución de respuestas para la categoría Normalización de las violencias.



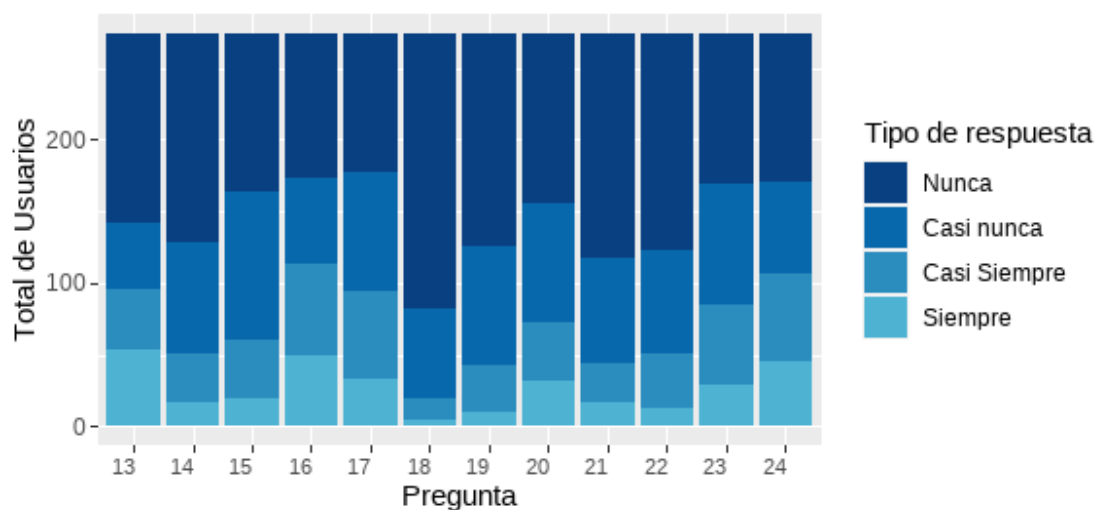
Fuente: Elaboración Propia

- f. Sección: Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima

En esta parte del instrumento se observa que en aunque en general, la respuesta que predomina es Nunca, en las preguntas 16. Me suelen sudar las manos sin saber por qué, 17. Tengo sensación de peligro o miedo sin saber por qué, 23. Me vienen recuerdos horribles mientras estoy despierto y 24. Me vienen nervios, ansiedad o angustia sin saber por qué, las respuestas Casi Siempre y Siempre tienen una proporción significativa, a lo cual es necesario prestar atención. (ver Gráfica 11).



Gráfico 11. Distribución de respuestas para la categoría Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima.



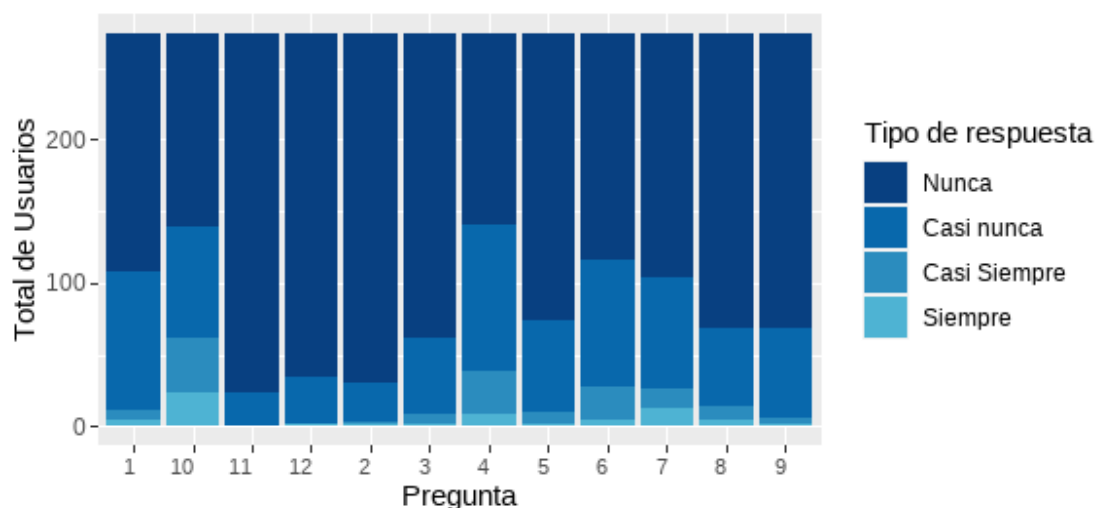
Fuente: Elaboración Propia

g. Sección: Situaciones de victimización por intimidación

Para la categoría se observa que en general, las respuestas que predominan son Nunca y Casi nunca, lo cual es algo positivo. Por otro lado, se observa que en las preguntas 06. Me chiflan, me gritan y 10. Me llaman por apodos que me incomodan, la respuesta Casi Siempre tiene una proporción significativa, a lo cual es necesario prestar atención. (ver Gráfica 12).



Gráfico 12. Distribución de respuestas para la categoría Situaciones de victimización por intimidación.



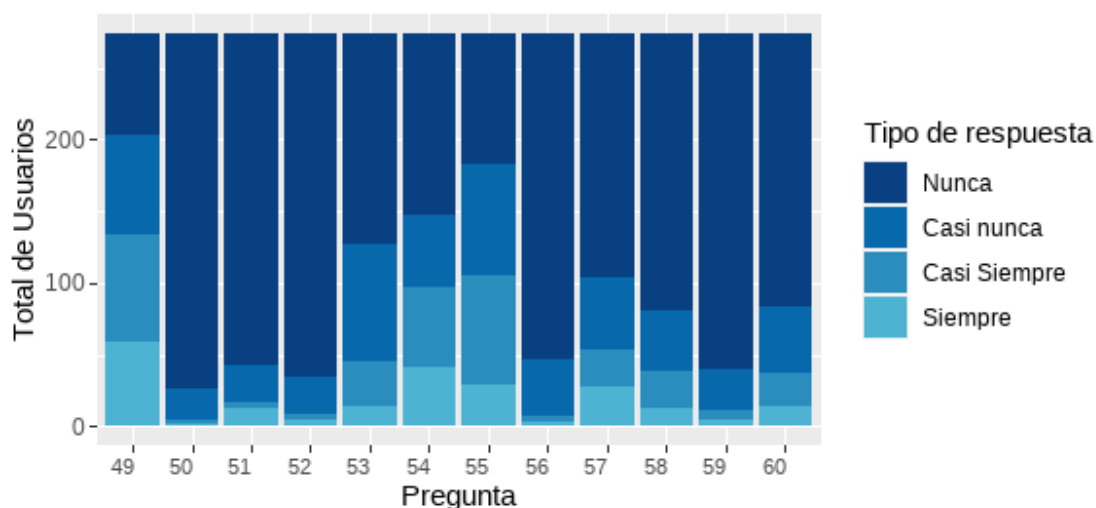
Fuente: Elaboración Propia

h. Sección: Violencia de Género contra la mujer

Se encuentra que la respuesta que predomina es Nunca, lo cual es algo positivo. Por otro lado, se observa que en la pregunta 49. En mi familia se conversa sobre derechos sexuales y reproductivos la respuesta principal es Casi Siempre siendo esto un resultado positivo. Sin embargo, si observamos las preguntas 53. Me siento insegura o inseguro al compartir con algunos compañeros, 54. Creo que no puedo expresar debilidad ante situaciones de acoso, porque me hace ver menos hombre o mujer y 55. Siento que algunos comentarios de mis compañeros son ofensivos contra las mujeres, podemos ver que las respuestas Casi Siempre y Siempre tienen una proporción significativa, a lo cual es necesario prestar atención. (ver Gráfica 12).



Gráfico 13. Distribución de respuestas para la categoría Violencia de Género contra la mujer.



Fuente:

Elaboración Propia

Siguiendo lo expuesto en la sección de metodología, se tomaron tres secciones del instrumento para construir los modelos. A continuación se muestra el número de usuarios en el estado bajo y alto para cada tipo de vulnerabilidad. Se observa que para el caso de la variable dependiente de ansiedad y violencia los datos están balanceados, es decir, hay un porcentaje similar en estado bajo y alto. Sin embargo, para el caso de victimización, la mayoría se encuentra en un estado bajo.

Tabla 14. Distribución de los estados por tipo de situaciones de vulnerabilidad

ESTADO	Ansiedad		Victimización		Violencia	
	No. Usuarios	Frec. relativa	No. Usuarios	Frec. relativa	No. Usuarios	Frec. relativa
BAJO	118	56.9%	225	82.1%	110	59.6%



ALTO	156	43.1%	49	17.9%	162	40.4%
------	-----	-------	----	-------	-----	-------

Fuente: Elaboración Propia



Resultados

A continuación se muestran los resultados del mejor modelo para cada uno de los tipos de vulnerabilidad analizados y una tabla resumen con las métricas de evaluación en comparación con los otros modelos. Las matrices de confusión de los otros modelos se encuentran en el Anexo II. A partir de estos resultados se encuentra que el mejor modelo para predecir un estado alto de situaciones de Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima⁴ y de situaciones de victimización por intimidación es el modelo de Elastic Net y para el caso de situaciones de normalización de la violencia es el modelo SGBosting.

En la tabla 15 se observa la matriz de confusión del modelo Elastic Net para situaciones de ansiedad y en la tabla 16 el resumen de métricas de los 4 modelos. Se encontró que los 4 modelos analizados presentan resultados relativamente similares para exactitud y precisión.

El modelo de Elastic Net presenta los valores máximos para las 4 métricas analizadas. La exactitud indica que de cada 10 predicciones, acierta en 7 (entre altos y bajos). La precisión indica que de cada 10 predicciones que indica alto, acierta en 7; la especificidad indica que de cada 10 jóvenes expuestos o que viven una baja cantidad de situaciones de ansiedad, el modelo acierta a 5. Finalmente, la sensibilidad indica que de cada 10 jóvenes expuestos o que viven una alta cantidad de situaciones de ansiedad, el modelo acierta a aproximadamente 9.

⁴ Para el caso de ansiedad, aunque el bosque aleatorio tiene un mayor nivel de sensibilidad que el elastic net, selecciona el modelo de elastic net puesto que presenta un mejor desempeño en todas las otras métricas.



Tabla 15. Matriz de confusión del modelo Elastic Net para situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	27	12
	Bajo	4	11

Fuente: Elaboración propia

Tabla 16. Métricas de evaluación de los modelos relacionados a Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima

Métrica	Bosque Aleatorio	NB	SGBosting	ElasticNet
Exactitud	66.67%	62.96%	66.67%	70.37%
Precisión	65.12%	62.79%	69.70%	69.23%
Especificidad	34.78%	30.43%	56.52%	47.83%
Sensibilidad	90.32%	87.10%	74.19%	87.10%

Fuente: Elaboración propia

En la tabla 17 se observa la matriz de confusión del modelo Elastic Net para situaciones de victimización por intimidación y en la tabla 18 el resumen de métricas de los 4 modelos. Se



encontró que los 4 modelos analizados presentan resultados diferentes en todas las métricas, con excepción de la exactitud.

Tabla 17. Matriz de confusión del modelo ElasticNet para situaciones de victimización por intimidación

		Observaciones reales	
		Alto	Bajo
Predicción n	Alto	7	7
	Bajo	2	36

Fuente: Elaboración propia

El modelo de Elastic Net presenta el valor máximo únicamente para la métrica de sensibilidad. La exactitud indica que de cada 10 predicciones, acierta en 8 (entre altos y bajos). La precisión indica que de cada 10 predicciones que indica alto, acierta en 5; la especificidad indica que de cada 10 jóvenes expuestos o que viven una baja cantidad de situaciones de victimización por intimidación, el modelo acierta a 8. Finalmente, la sensibilidad indica que de cada 10 jóvenes expuestos o que viven una alta cantidad de situaciones de victimización por intimidación, el modelo acierta a aproximadamente 8.

En la tabla 17 se observa la matriz de confusión del modelo Elastic Net para situaciones de normalización de la violencia y en la tabla 18 el resumen de métricas de los 4 modelos. Se encontró que los 4 modelos analizados presentan resultados diferentes en todas las métricas.



Tabla 18. Métricas de evaluación para los modelos de situaciones de victimización por intimidación

Métrica	Bosque Aleatorio	NB	SGBosting	ElasticNet
Exactitud	80.77%	84.62%	82.69%	82.69%
Precisión	0%	57.14%	NA	50%
Especificidad	97.67%	93.02%	100%	83.72%
Sensibilidad	0%	44.44%	0%	77.78%

Fuente: Elaboración propia

El modelo de SGBosting presenta el valor máximo para las cuatro métricas analizadas. La exactitud indica que de cada 10 predicciones, acierta aproximadamente 7 (entre altos y bajos). La precisión indica que de cada 10 predicciones que indica alto, acierta en 7; la especificidad indica que de cada 10 jóvenes expuestos o que viven una baja cantidad de situaciones de normalización de la violencia, el modelo acierta a 5. Finalmente, la sensibilidad indica que de cada 10 jóvenes expuestos o que viven una alta cantidad de situaciones de normalización de la violencia, el modelo acierta a aproximadamente 9.



Tabla 19. Matriz de confusión del modelo SGBoosting para violencia

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	29	7
	Bajo	3	15

Fuente: Elaboración propia

Tabla 20. Métricas de evaluación para los modelos de violencia

Métrica	Bosque Aleatorio	NB	SGBoosting	ElasticNet
Exactitud	66.67%	59.26%	81.48%	74.07%
Precisión	71.88%	60.87%	80.56%	73.68%
Especificidad	59.09%	18.18%	68.18%	54.55%
Sensibilidad	71.88%	87.5%	90.62%	87.5%

Fuente: Elaboración propia

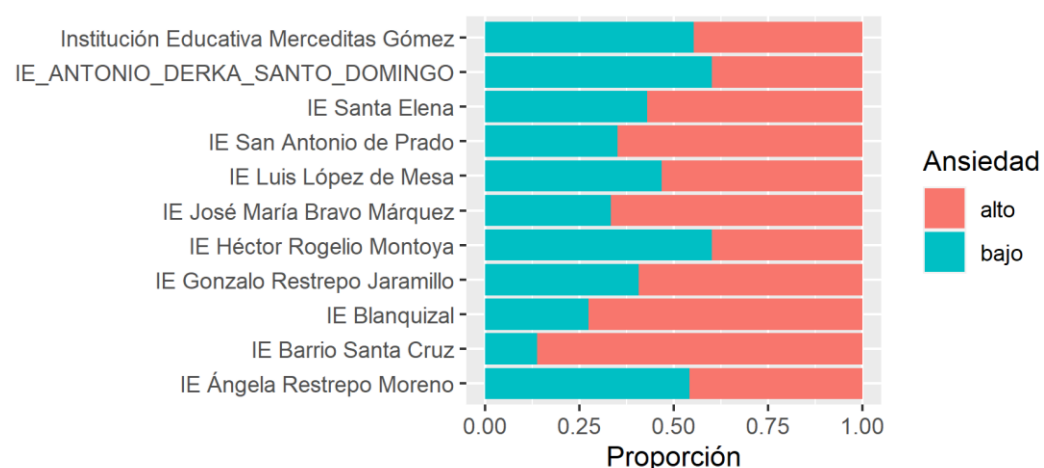
A partir de estos resultados se procede a analizar las variables relacionadas con los estados altos o bajos para cada una de las vulnerabilidades seleccionadas.



Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima

Cuando se analiza el estado real de situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima por institución educativa se encuentra que 5 de cada 10 jóvenes se encuentran en un estado alto. Esto se intensifica en algunas instituciones como IE Barrio Santa Cruz, donde puede ser 8 de cada 10 y en IE Blanquizal con de 7 de cada 10 (ver Gráfico 14).

Gráfico 14. Proporción de estado de situaciones de ansiedad por institución educativa

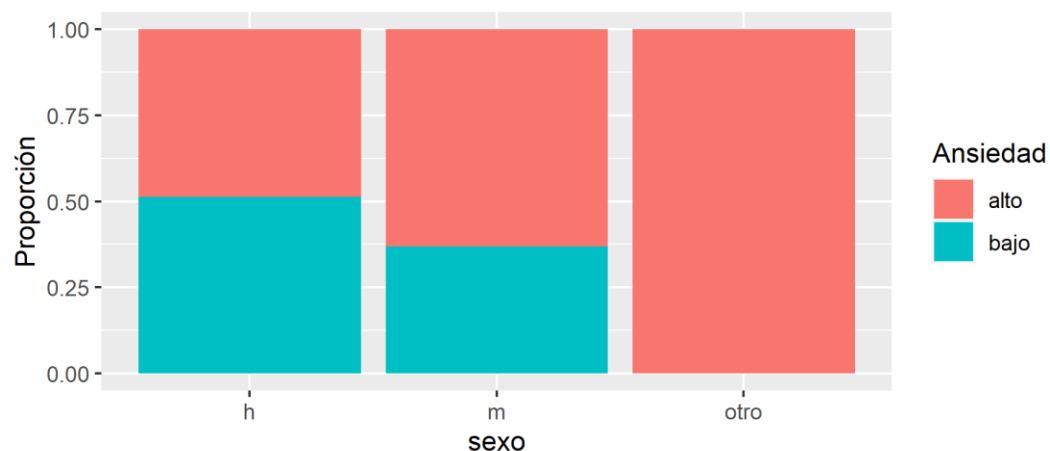


Fuente: Elaboración propia

En términos del sexo que reconoce el estudiante, se observa una mayor proporción de mujeres con un estado alto que en el caso de los hombres. Sin embargo, el porcentaje de hombres no deja de ser significativo, puesto que 1 de cada 2 se encuentra en un estado alto (ver Gráfico 15).



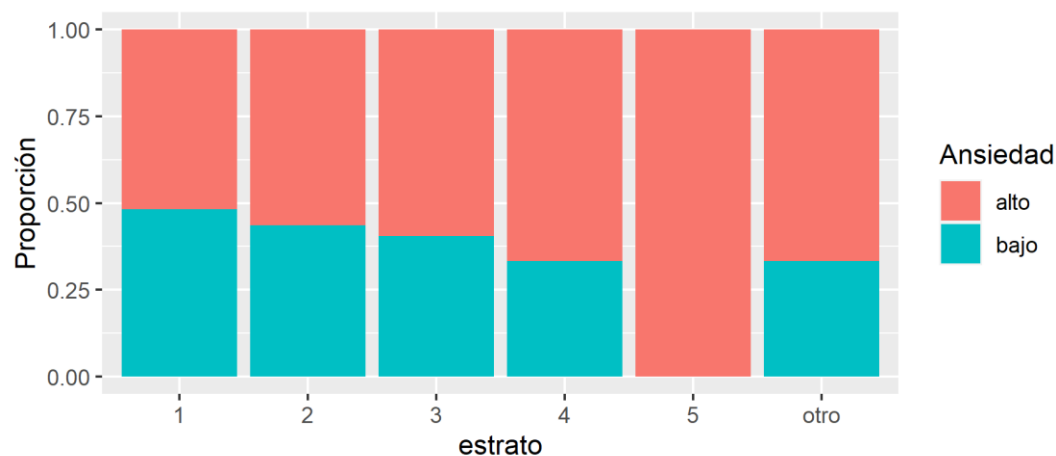
Gráfico 15. Proporción de estado real de situaciones de ansiedad por sexo



Fuente: Elaboración propia

En cuanto al estrato, se observa que a mayor estrato hay más proporción de jóvenes con un estado alto; sin embargo, hay que tener en cuenta que son muy pocos los jóvenes de este estudio que pertenecen a estrato 4 (3 jóvenes) y 5 (1 joven) (ver Gráfico 16).

Gráfico 16. Proporción de estado de situaciones de ansiedad por estrato



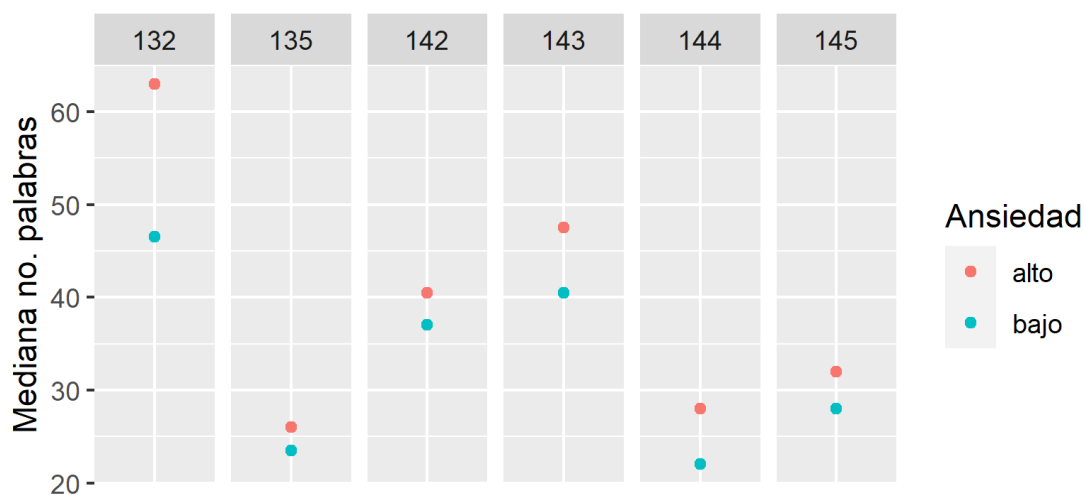
Fuente: Elaboración propia

En el caso del número de palabras, se observa que el 50% de los jóvenes con un estado alto de situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima



tienden a escribir más palabras en sus comentarios que en el caso de los que se encuentran en un estado bajo. Esta diferencia no es tan relevante para la lección 135, 142 y 145 (ver Gráfico 17).

Gráfico 17. Mediana en el número de palabras por lección por estado de situaciones de ansiedad



Fuente: Elaboración propia

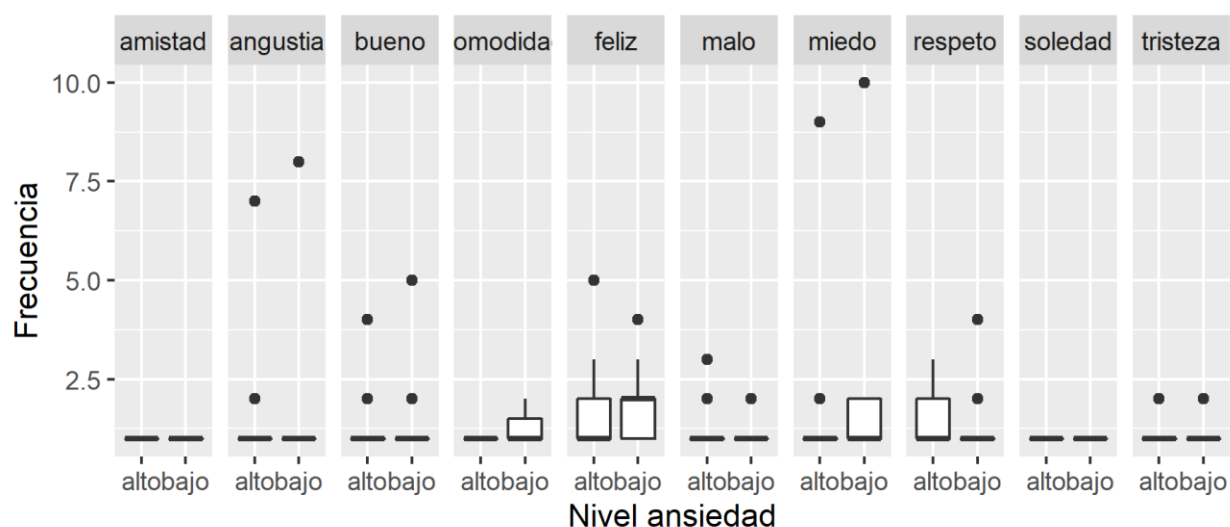
En cuanto a la frecuencia de palabras claves se observan diferencias en la frecuencia de uso de palabras relacionadas con feliz, miedo y respeto (ver Gráfico 18). Los jóvenes con estados bajos de situaciones de ansiedad expresan en sus comentarios más veces la palabra feliz, que los que están en un estado alto. Para el caso del miedo, los que se encuentran en un estado bajo de situaciones de ansiedad expresan más está que los bajos. Finalmente, para el caso de la palabra respeto, hay más estudiantes con estados altos de estas situaciones de ansiedad que emplean esta palabra en sus comentarios que los que se encuentran en un estado bajo..

En cuanto a la frecuencia de palabras relacionadas con una emoción de tristeza se observa que hay un mayor uso por parte de los jóvenes que se encuentran en un estado alto de situaciones ansiedad (ver Gráfico 19) que los que están en un estado bajo. Esto ocurre en la lección donde deben hablarse a sí mismos como la lección 132 (Me llegó la carta). De igual



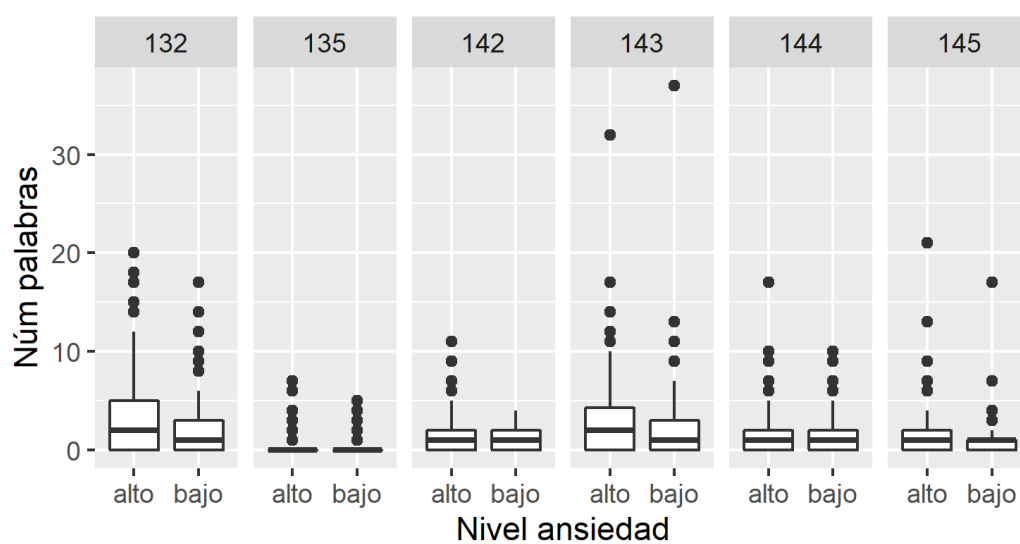
forma, en la lección de prácticas de cuidado como la 143 (Seis claves) o 145 (reflexiones sobre nuestro entorno) emplean más palabras asociadas a esta emoción, tal vez haciendo referencia a más situaciones de tristeza donde se necesita más cuidado.

Gráfico 18. Frecuencia de palabras claves por estado de situaciones de ansiedad



Fuente: Elaboración propia

Gráfico 19. Frecuencia de palabras relacionadas con una emoción de tristeza



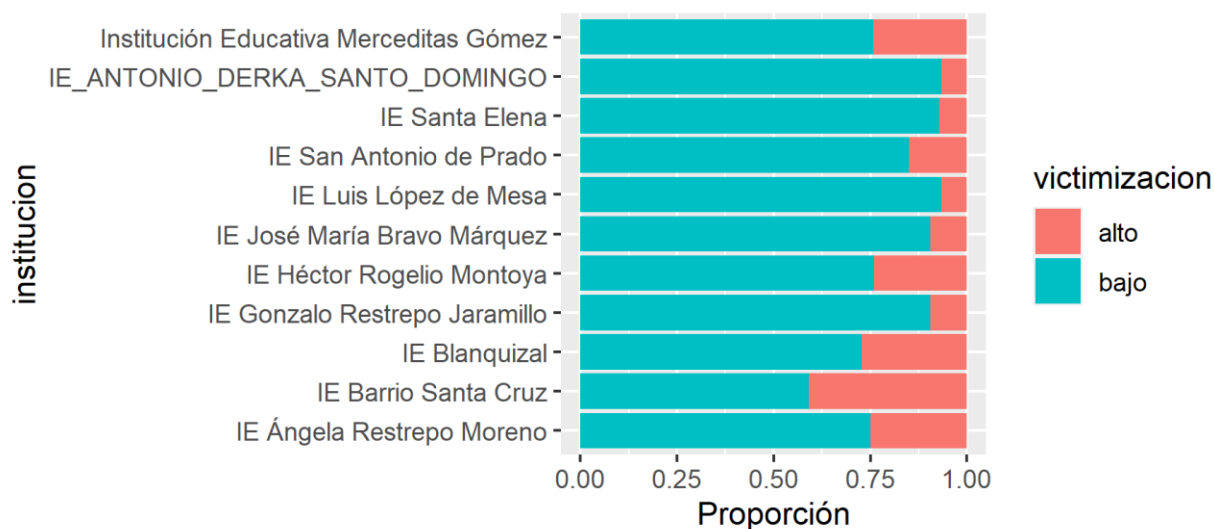


Fuente: Elaboración propia

Situaciones de victimización por intimidación

En la mayoría de las instituciones se observa entre uno o dos jóvenes se encuentran en un estado alto de situaciones de victimización por intimidación por cada 10 jóvenes del estudio. Resaltan nuevamente las instituciones IE Blanquizal y IE Barrio Santa Cruz, seguidas de IE Ángela Restrepo Moreno, IE Merceditas Gómez, y IE Héctor Rogelio Montoya (ver Gráfico 20).

Gráfico 20. Proporción de estados reales de situaciones de victimización por intimidación por institución educativa



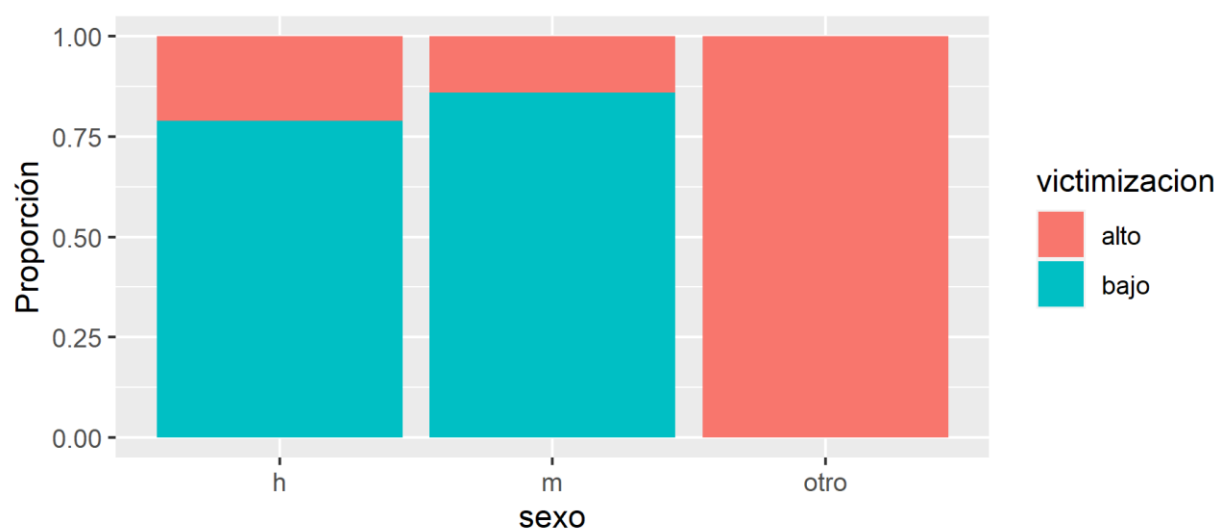
Fuente: Elaboración propia

En términos del sexo que reconoce la persona, se observa una mayor proporción de hombres con estado alto de situaciones de victimización por intimidación. La diferencia entre sexos no es tan marcada como el caso de situaciones de ansiedad y situaciones de normalización de



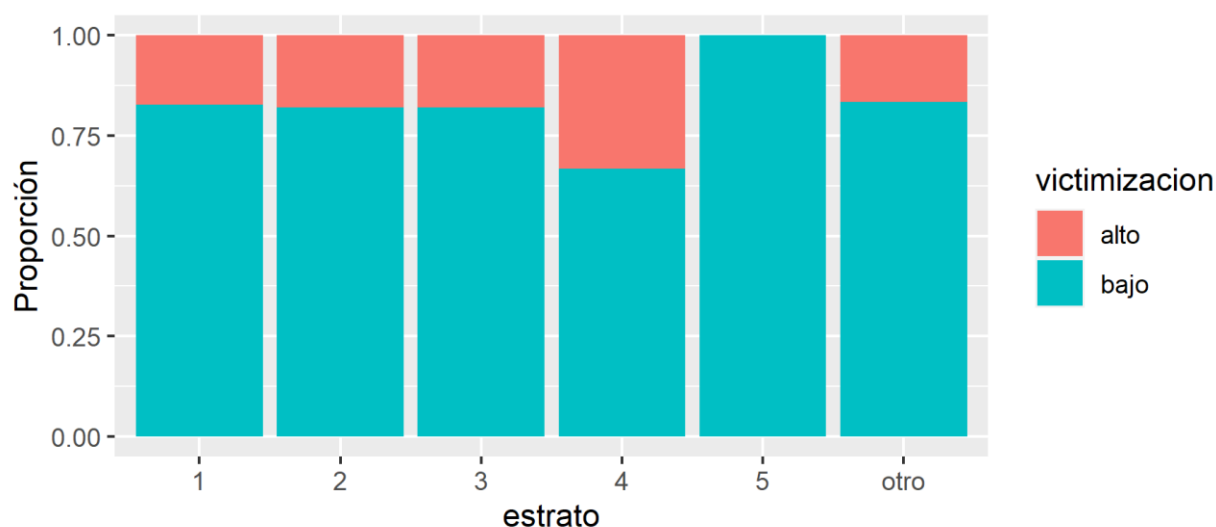
violencia (ver Gráfico 21). Frente al estrato, se observan valores relativamente similares para estrato 1, 2 y 3, y el mayor porcentaje de jóvenes en una situación de victimización se encuentran en el estrato 4 (ver Gráfico 22).

Gráfico 21. Proporción de estados reales de situaciones de victimización por intimidación por sexo



Fuente: Elaboración propia

Gráfico 22. Proporción de estados reales de situaciones de victimización por intimidación por estrato



Fuente: Elaboración propia

El joven que se encuentra en un estado alto de situaciones de victimización por intimidación se caracteriza en términos generales por utilizar menos palabras, que sus pares en un estado bajo (ver Gráfico 23). Por ejemplo el 50% de los jóvenes en un estado alto de situaciones de victimización por intimidación empleaban aproximadamente 45 palabras o menos en la lección 132 (Me llegó la carta), versus las más de 60 palabras que utiliza el 50% de los jóvenes con estado bajo de situaciones de victimización por intimidación.

Gráfico 23. Mediana en el número de palabras por lección para situaciones de victimización por intimidación

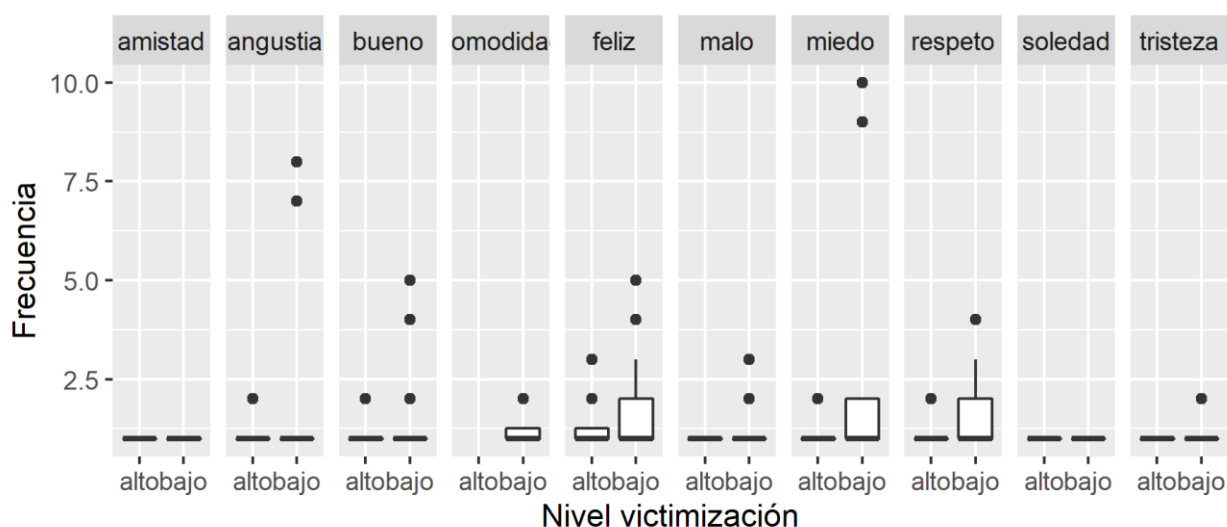




Fuente: Elaboración propia

En cuanto a la frecuencia de palabras claves se observa para el caso de situaciones de victimización por intimidación que donde resaltan diferencias es con el uso de palabras relacionadas con feliz, miedo y respeto nuevamente (Gráfico 24). Los jóvenes con un estado bajo de situaciones de victimización por intimidación expresan en sus comentarios más veces estas tres palabras que los jóvenes con estados altos.

Gráfico 24. Frecuencia de palabras claves por estado en situaciones de victimización por intimidación



Fuente: Elaboración propia

Normalización de las violencias

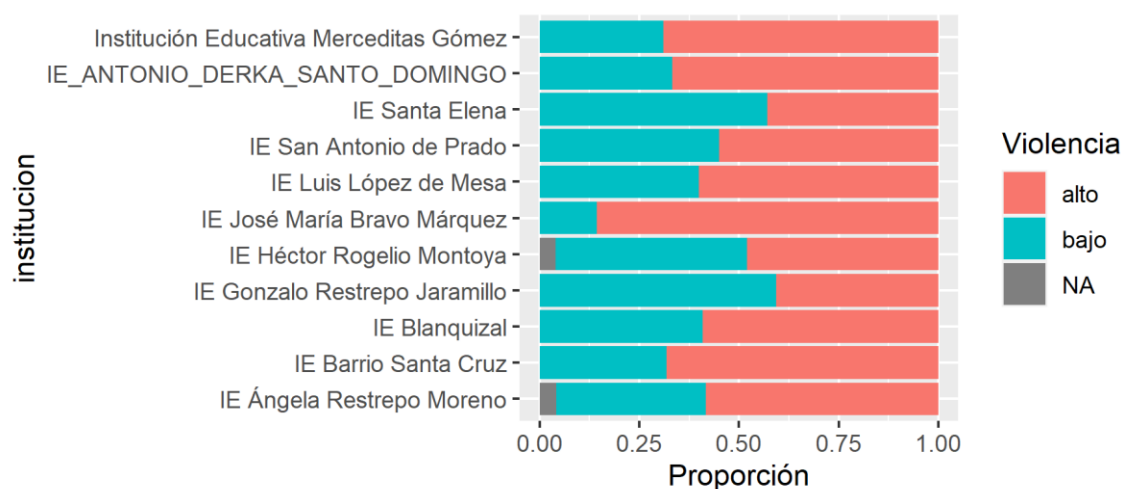
De manera similar al tipo de vulnerabilidad de ansiedad, los estados de normalización de la violencia son altos en casi todas las instituciones de este proyecto pues el 50% de los jóvenes o más se encuentran en este estado. Dentro de estas instituciones las mayores proporciones



de estados altos se encuentran en IE José María Bravo Márquez y nuevamente la institución IE Barrio Santa Cruz,(ver Gráfico 25).

En términos del sexo que reconoce la persona, se observa una mayor proporción de hombres con un estado alto de situaciones de normalización de las violencias que en el caso las mujeres. Sin embargo, el porcentaje de mujeres no deja de ser relevante, puesto que aproximadamente 5 de cada 10 estaría en un estado alto (ver Gráfico 26). Para el caso del estrato, se encuentran una mayor proporción de estudiantes con estados altos en el estrato 3 que en los estratos 1 y 2 (ver Gráfico 27). A diferencia de las situaciones de ansiedad y las situaciones de victimización por intimidación, no se observan diferencias significativas en la mediana del número de palabras por lección entre altos y bajos para el caso de situaciones de normalización de la violencia (ver Gráfico 28).

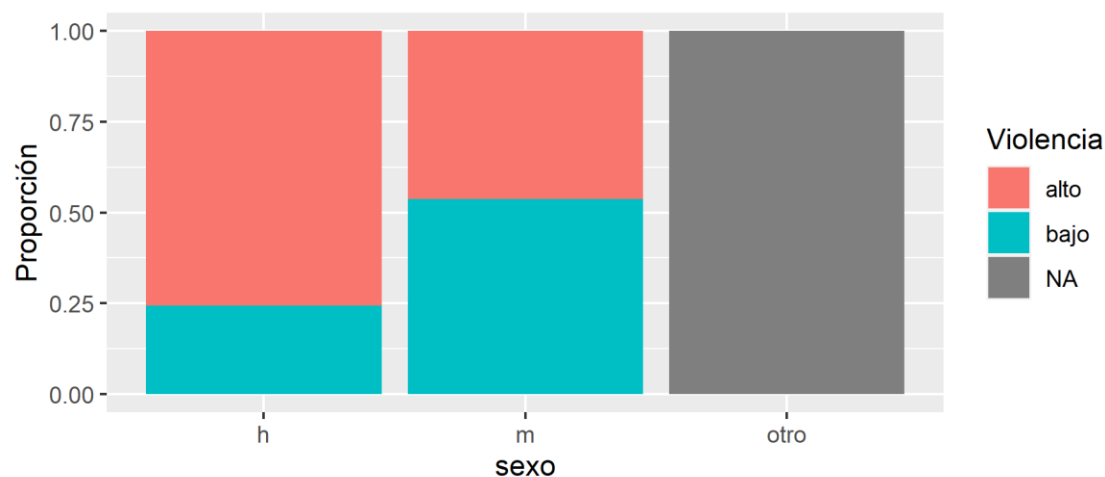
Gráfico 25. Proporción de estados reales (alto o bajo) de situaciones de normalización de la violencia por institución educativa



Fuente: Elaboración propia

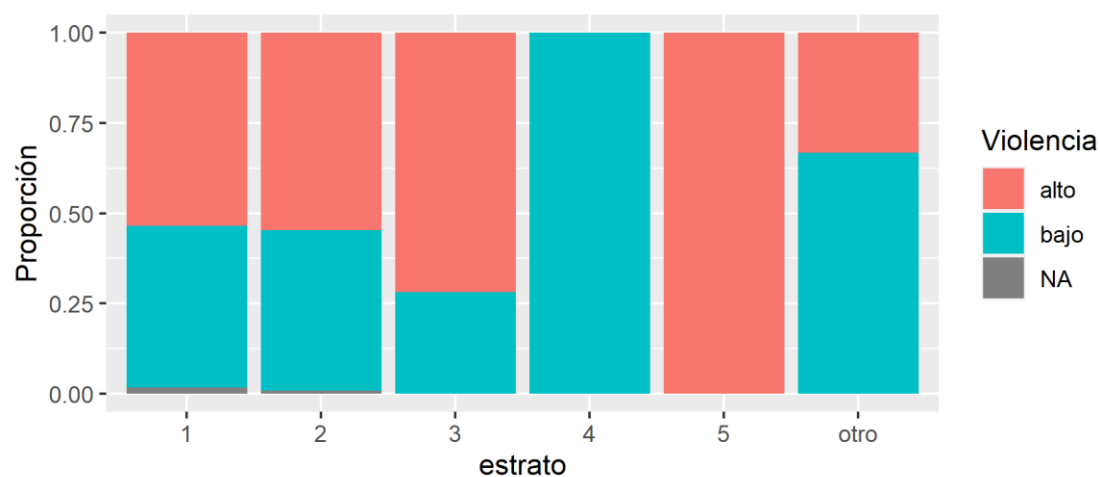


Gráfico 27. Proporción de estado real en situaciones de normalización de la violencia por sexo



Fuente: Elaboración propia

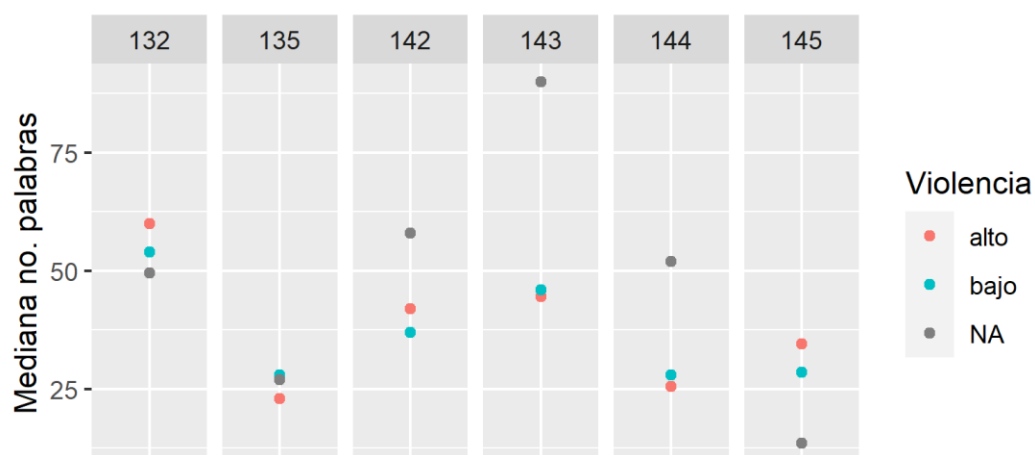
Gráfico 28. Proporción de estado real en situaciones de normalización de la violencia por estrato



Fuente: Elaboración propia



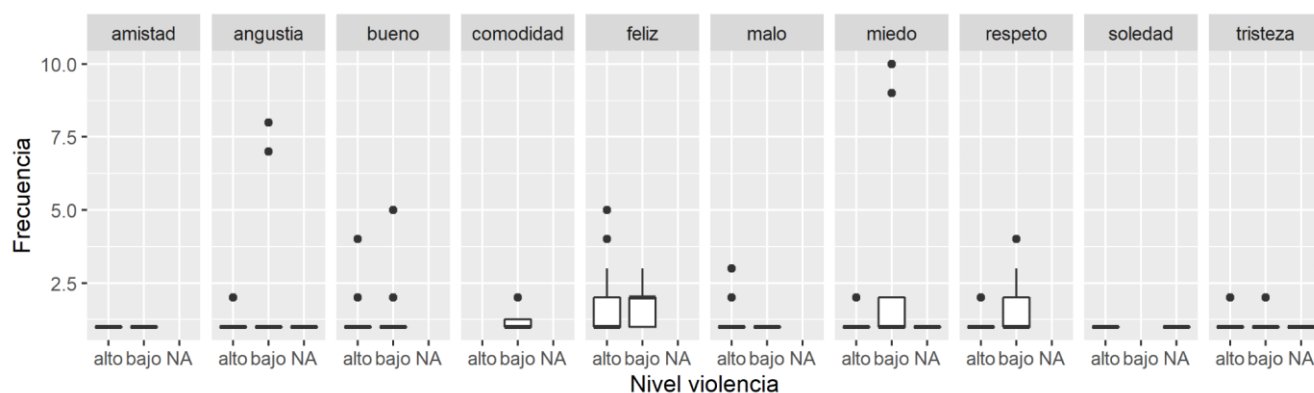
Gráfico 27. Mediana en el número de palabras por lección por estado de situaciones de normalización de la violencia



Fuente: Elaboración propia

En cuanto a la frecuencia de palabras claves se observa para el caso de violencia que donde resaltan diferencias es con el uso de palabras relacionadas con feliz, comodidad, miedo y respeto nuevamente. Los jóvenes con bajos niveles de violencia expresan en sus comentarios más veces estas cuatro palabras que los jóvenes con altos niveles.

Gráfico 28. Frecuencia de palabras claves por estado de situaciones de normalización de la violencia

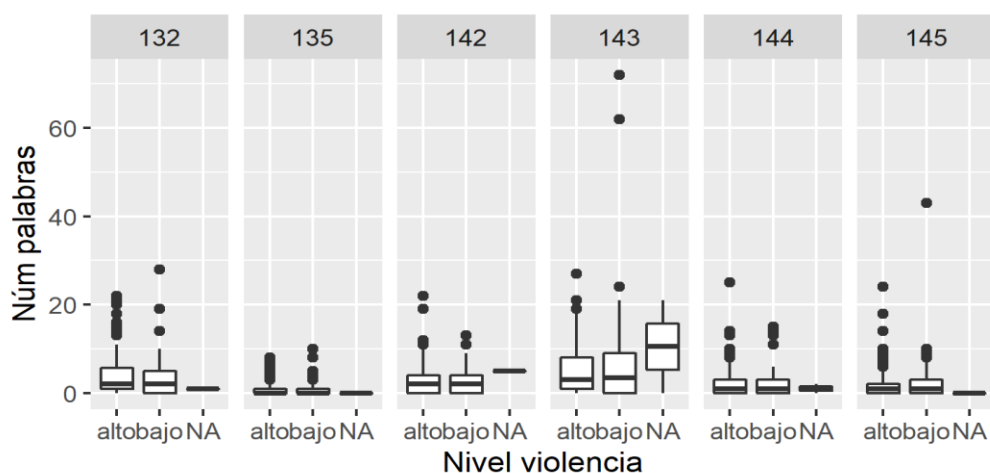


Fuente: Elaboración propia



En cuanto a la frecuencia de palabras relacionadas con una emoción negativa se observa que para el caso de situaciones por normalización de las violencias no hay mucha diferencia entre los jóvenes de alto y bajo. Solo se observa unas diferencias que pueden ser significativas en la lección 132 (Me llegó la carta) o en la lección 143 (Seis claves).

Gráfico 29. Frecuencia de palabras relacionadas con una emoción negativa por estado de situaciones de normalización de la violencia



Fuente: Elaboración propia

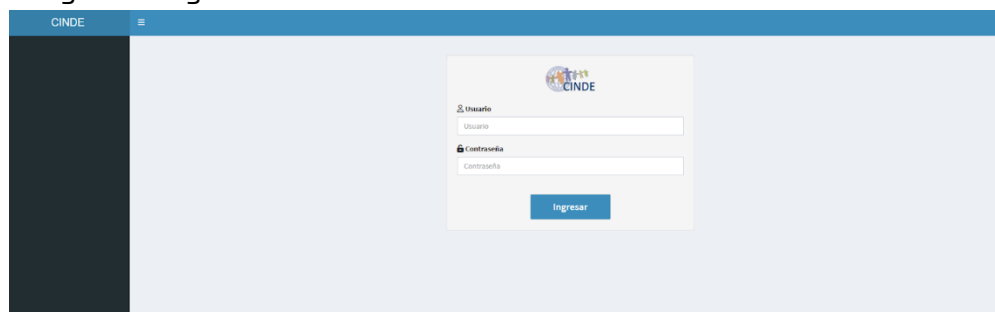


Prototipo funcional de solución web

Este tipo de modelos de *machine learning* son muy útiles cuando se pueden seguir usando. Por ejemplo, para predecir los estados altos o bajos de las diferentes situaciones analizadas para estudiantes de otros colegios con características similares, o cada año para tener insumos para las políticas internas de las instituciones educativas que hicieron parte de este proyecto.

A continuación se evidencia el uso del prototipo funcional. En la Imagen 3 se observa la pantalla de inicio de sesión. El usuario ingresa el nombre del usuario y contraseña y puede ingresar a la solución.

Imagen 3. Página de inicio de sesión



Posterior a eso, observa la interfaz donde debe cargar los archivos⁵ para el procesamiento de los datos y predicción (Imagen 4) los cuales son:

1. Archivo de caracterización
2. Archivo de comentarios
3. Archivo de texto entregable

⁵ Se debe conservar siempre la misma estructura en los archivos, de lo contrario no será posible generar el resultado.



Una vez los archivos estén cargados da clic en el botón “Procesar archivos”

Imagen 4. Página de parámetros

Una vez el procesamiento de los archivos concluya, el usuario de la solución podrá previsualizar el archivo resultante en pantalla. Este contendrá los datos de caracterización del usuario, predicción para cada tipo de vulnerabilidad, y las variables empleadas en los modelos (ver Imagen 5). Finalmente, dar clic en el botón “Descargar archivo de resultado” lo cual descargará un archivo de Excel llamado “archivo_procesado.xlsx”.

Imagen 5. Previsualización del archivo resultante



Conclusiones y recomendaciones

En resumen, con este proyecto se pudo evidenciar la utilidad de emplear modelos de aprendizaje automático para predecir el estado alto o bajo de situaciones de situaciones de ansiedad, depresión, estrés post traumático y efectos sobre autoestima, de situaciones de victimización por intimidación y de situaciones de normalización de la violencia para los estudiantes entre 12 y 18 años de las 11 instituciones educativas de este estudio.

Para el caso de situaciones de ansiedad, se encontró que el mejor modelo fue un Elastic Net con una sensibilidad de 87.10%; para el segundo caso, el mejor modelo también fue un Elastic Net con una sensibilidad de 77.78%, y para situaciones de normalización de la violencia el mejor modelo fue SGBooting con una sensibilidad del 90.62%.

Los algoritmos de aprendizaje automático como SGBooting y Elastic Net, pueden analizar diferentes tipos de datos, como los comentarios de los jóvenes, para identificar patrones y relaciones que pueden no ser evidentes a través de métodos estadísticos tradicionales. Logrando predicciones precisas y confiables que se pueden utilizar para la identificación temprana e intervención de este tipo de situaciones en los adolescentes. Esto puede ayudar a las escuelas, secretaría de educación y otros interesados a implementar sistemas de apoyo e intervenciones apropiadas para abordar las preocupaciones relacionadas con los tres tipos de situaciones analizadas de manera oportuna. Lo anterior porque los resultados del modelo permiten priorizar la población que tienen una mayor probabilidad de encontrarse en una alta cantidad de situaciones de ansiedad, victimización por intimidación y/o de normalización de las violencias.

Adicionalmente, es importante contar con las siguientes recomendaciones. En primer lugar, al compartir los resultados del estudio con escuelas u otros interesados, es importante que los datos estén anonimizados y desprovistos de cualquier información de identificación personal (por ejemplo, la identificación real de los jóvenes) para proteger la privacidad de los adolescentes.



En segundo lugar, se necesita capturar más datos para seguir mejorando las métricas de los modelos ya que es importante tener en cuenta que los modelos de aprendizaje automático no son perfectos. Estos están sujetos a sesgos, limitaciones y posibles errores. De ahí que sea crucial validar y evaluar minuciosamente la precisión y robustez del modelo con estudiantes de otros colegios antes de utilizarlo en entornos reales y a gran escala. También puede ser necesario monitorear y actualizar regularmente el modelo para asegurar su eficacia y relevancia a lo largo del tiempo.

En tercer lugar, seguir priorizando las lecciones de la plataforma virtual donde los jóvenes hablen de sus vivencias o sus sentimientos como lo fueron la lección 132, 135 y 144. Por ahora nuevos jóvenes que hagan parte de la plataforma virtual de mangus como mínimo deberán llenar los datos de caracterización (edad, sexo y estrato) y las lecciones (132, 135, 136, 142, 143, 144 y 145).

Finalmente, para el futuro uso de la herramienta se recomienda que se obtenga el consentimiento informado de los adolescentes y sus padres, o tutores legales, antes de recopilar los datos de la plataforma virtual. El consentimiento informado debe explicar claramente el propósito que le puede dar la escuela o ente gubernamental al uso de los resultados del modelo y de los datos, los posibles riesgos y beneficios, y los derechos de los participantes. Los adolescentes podrían tener la opción de optar por no participar o retirar sus datos en cualquier momento sin consecuencias negativas por parte de la institución educativa, en caso de que el modelo se use en alianza con las instituciones educativas.



Referencias

Alanazi, I., & Alves-Foss, J. (2020). Cyber bullying and machine learning: a survey. *Int Journal of Computer Science and Information Security (IJCSIS)*, 18(10).

AlHarbi, B. Y., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M., Alshobaili, J. F., & Ibrahim, D. M. (2019). Automatic cyber bullying detection in Arabic social media. *Int. J. Eng. Res. Technol*, 12(12), 2330-2335.

Arce, J. I. B. (2022). La matriz de confusión y sus métricas. *Juan Barrios*. <https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/>

Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2354-2358). IEEE.

Chavanne, A. V., Paillère Martinot, M. L., Penttilä, J., Grimmer, Y., Conrod, P., Stringaris, A., ... & Artiges, E. (2023). Anxiety onset in adolescents: a machine-learning prediction. *Molecular Psychiatry*, 28(2), 639-646.

Dadvar, M., Trieschnigg, D., & De Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings 27* (pp. 275-281). Springer International Publishing.



Hill, A. J., & Zhou, W. (2021). Peer discrimination in the classroom and academic achievement. *Journal of Human Resources*, 0919-10460R3.

Ige, T., & Adewale, S. (2022). AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. *arXiv preprint arXiv:2207.11897*.

Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

Lee, H., & Song, E. (2022). Peer discrimination toward rural migrant students and academic performance in urban China: A machine learning approach. *Cities*, 131, 104027.

Nikiforos, S., Tzanavaris, S., & Kermanidis, K. L. (2020). Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. *Journal of Computers in Education*, 7, 531-551.

Pestian, J. P., Grupp-Phelan, J., Bretonnel Cohen, K., Meyers, G., Richey, L. A., Matykiewicz, P., & Sorter, M. T. (2016). A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide and Life-Threatening Behavior*, 46(2), 154-159.

Qasrawi, R., Polo, S. P. V., Al-Halawa, D. A., Hallaq, S., & Abdeen, Z. (2022). Assessment and prediction of depression and anxiety risk factors in schoolchildren: machine learning techniques performance analysis. *JMIR Formative Research*, 6(8), e32736.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops* (Vol. 2, pp. 241-244). IEEE.



RPubs - Matriz de Confusión - Evaluación de modelos de predicción. (n.d.).
<https://rpubs.com/chzelada/275494>

Sedano, C. R., Ursini, E. L., & Martins, P. S. (2017). A bullying-severity identifier framework based on machine learning and fuzzy logic. In *Artificial Intelligence and Soft Computing: 16th International Conference, ICAISC 2017, Zakopane, Poland, June 11-15, 2017, Proceedings, Part I 16* (pp. 315-324). Springer International Publishing.

Wang, B., Nemesure, M. D., Park, C., Price, G., Heinz, M. V., & Jacobson, N. C. (2022, July 18). Leveraging Deep Learning Models to Understand the Daily Experience of Anxiety in Teenagers over the Course of a Year. <https://doi.org/10.31234/osf.io/y7m69>

Zhao, R., Zhou, A., & Mao, K. (2016, January). Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking* (pp. 1-6).



Anexos

Anexo I Apartados del instrumento

Tabla 21. Situaciones de victimización por intimidación

Situaciones incluidas en el instrumento
No me dejan participar, me excluyen
Me obligan a hacer cosas peligrosas para mí
Rompen mis cosas a propósito
Me insultan
Me pegan coscorrones, puñetazos, patadas
Me chiflan, me gritan
Dicen a otros que no estén conmigo o que no me hablen
Me desprecian
Me amenazan para que haga cosas que no quiero
Me obligan a hacer cosas que están mal

Tabla 22. Situaciones de ansiedad, depresión, estrés post traumático y efectos sobre la autoestima

Situaciones incluidas en el instrumento
Si volviera a nacer pediría ser diferente a como soy
Creo que nadie me aprecia
Tengo sueños y pesadillas horribles



Me suelen sudar las manos sin saber por qué
A veces tengo una sensación de peligro o miedo sin saber por qué
Al venir al colegio siento miedo o angustia
A veces me encuentro sin esperanza
A veces creo que no tengo remedio
Algunas veces tengo ganas de morirme
Algunas veces me odio a mi mismo
A veces me viene recuerdos horribles mientras estoy despierto
Me vienen nervios, ansiedad o angustia sin saber por qué

Tabla 23. Situaciones de normalización de la violencia

Situaciones incluidas en el instrumento
Permanezco callado ante una situación de agresión, acoso e intimidación
Intervengo cuando hay una situación de agresión, acoso e intimidación
Denuncio ante adultos una situación de agresión, acoso e intimidación
Permito la realización de juegos agresivos
Veo atención o acciones contundentes ante una situación de agresión, acoso e intimidación
Los adultos tienen derecho a corregirme física y verbalmente por mi mal comportamiento
Por ser hombre debo ser tolerante ante situaciones de agresión, acoso e intimidación
Mi relación con los otros está mediada por palabras groseras y agresivas
Los otros tienen razones válidas para agredirme



Recibir golpes, insultos, comentarios agresivos o exclusiones es algo común en la escuela

Los adultos pasan por alto las situaciones de agresión, acoso e intimidación

Los adultos minimizan las denuncias por agresión, acoso e intimidación

Anexo II Matrices de confusión

Tabla 24. Matriz de confusión del modelo bosque aleatorio para estado de situaciones de ansiedad

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	28	15
	Bajo	3	8

Fuente: Elaboración propia

Tabla 25. Matriz de confusión del modelo naive bayes para estado de situaciones de ansiedad

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	27	16
	Bajo	4	7

Fuente: Elaboración propia

Tabla 26. Matriz de confusión del modelo SGBooting para estado de situaciones de ansiedad

Observaciones reales	
----------------------	--



		Observaciones reales	
		Alto	Bajo
Predicción	Alto	28	10
	Bajo	8	13

Fuente: Elaboración propia

Tabla 27. Matriz de confusión del modelo bosque aleatorio para estado de situaciones de victimización por intimidación

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	0	1
	Bajo	9	42

Fuente: Elaboración propia

Tabla 28. Matriz de confusión del modelo naive bayes para estado de situaciones de victimización por intimidación

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	4	3
	Bajo	5	40

Fuente: Elaboración propia

Tabla 29. Matriz de confusión del modelo SGBooting para estado de situaciones de victimización por intimidación

		Observaciones reales	
		Alto	Bajo



		Observaciones reales	
		Alto	Bajo
Predicción	Alto	0	0
	Bajo	9	43

Fuente: Elaboración propia

Tabla 30. Matriz de confusión del modelo bosque aleatorio para estado de situaciones de normalización de la violencia

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	23	9
	Bajo	9	13

Fuente: Elaboración propia

Tabla 31. Matriz de confusión del modelo naive bayes para estado de situaciones de normalización de la violencia

		Observaciones reales	
		Alto	Bajo
Predicción	Alto	28	18
	Bajo	4	4

Fuente: Elaboración propia

Tabla 32. Matriz de confusión del modelo ElasticNet para estado de situaciones de normalización de la violencia



		Observaciones reales	
		Alto	Bajo
Predicción	Alto	28	10
	Bajo	4	12

Fuente: Elaboración propia