



CINDE: Resultados 1era versión

— Miércoles, 5 de Abril —

Contenido

- Objetivo
- Datos
- Procesamiento y limpieza
- Variables dependientes
- Variables independientes
- Modelo
- Resultados de modelos



Objetivo

Objetivo

Apoyar en el proceso de identificación de jóvenes en situación de vulnerabilidad mediante la analítica de datos y el desarrollo de un modelo de machine learning .

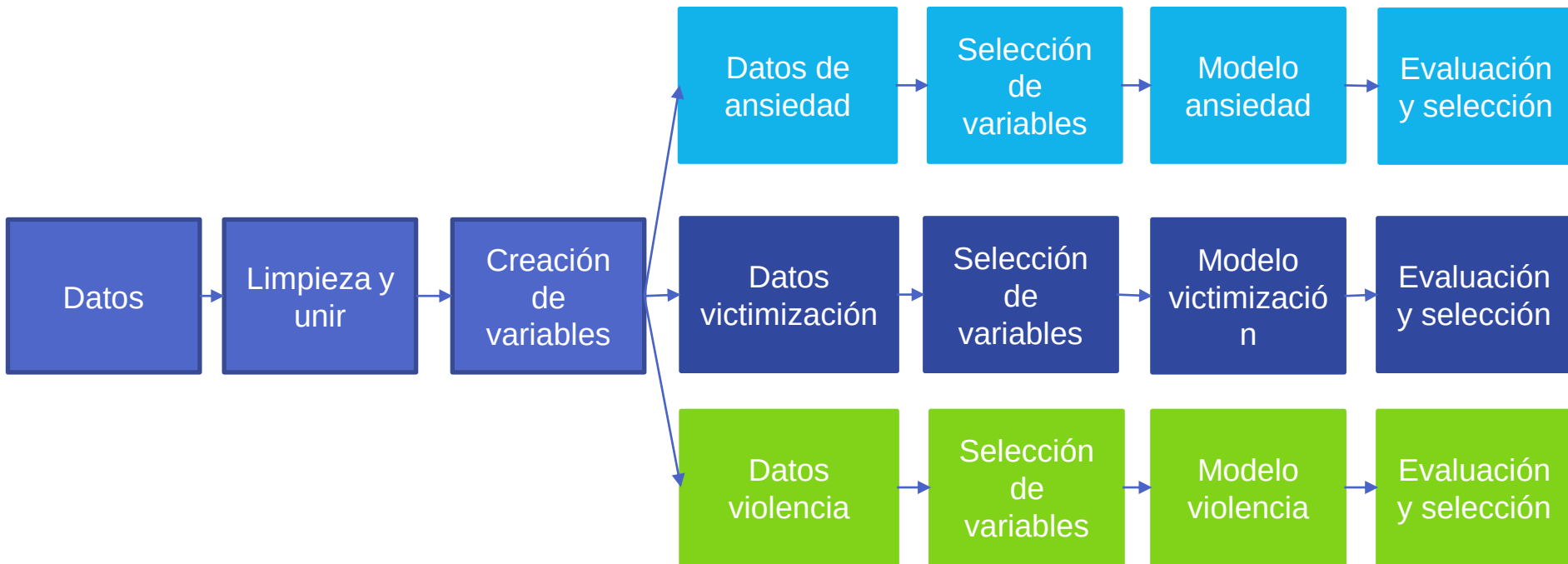
Específicamente en situación de vulnerabilidad por:

- Estar en una situación de **victimización por intimidación**
- Estar en una situación de **normalización de violencia**
- Estar en una situación de **depresión, ansiedad, estrés post traumático y efectos sobre autoestima**



Proceso

Proceso



Datos

Observaciones iniciales

- Caracterización de jóvenes:
 - 436 jóvenes
- Instrumento:
 - 312 jóvenes
- Comentarios
 - Lección con más jóvenes : 228 jóvenes (Lección 04 Caperucita roja)
 - Lección con menos jóvenes: 83 jóvenes (Lección 01 Bienvenida y presentación)

Nota: Fecha de corte 26 de marzo de 2023



Procesamiento y limpieza

Procesamiento y limpieza : Datos de caracterización

Como parte del proceso de procesamiento y limpieza se realizaron los siguientes pasos:

- Retirar jóvenes por fuera del rango de 12 a 18 (se dejaron los que no tenían edad)
- Retirar jóvenes sin sexo
- Retirar jóvenes sin estrato



Procesamiento y limpieza: Datos instrumento

Como parte del proceso de procesamiento y limpieza se realizaron los siguientes pasos:

- 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



Procesamiento y limpieza: Datos comentarios

Como parte del proceso de procesamiento y limpieza de los comentarios, se realizaron los siguientes pasos:

- Dejar lecciones con más de 170 jóvenes
- Eliminar comentarios duplicados
- Unificar comentarios por usuario y lección
- Eliminar números, puntuación y símbolos
- Pasar todo a minúsculas
- Eliminar tildes



Procesamiento y limpieza: Datos comentarios

Como parte del proceso de procesamiento y limpieza de los comentarios, se realizaron los siguientes pasos:

- 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”)



Procesamiento y limpieza: Datos comentarios

Ejemplo

Antes	Después
<p>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.</p> <p>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.</p>	<p>siempre ir acompañado adulto poder cuidar casos soledad siempre tener instrumento proteccion gas pimienta tener celular previamente marcado tener camara seguridad siempre contacto</p> <p>recomendar familia colegio cuidar deber tener camara seguridad siempre tener docente lugares remoto colegio familia momentos contacto</p>



Observaciones finales

- Caracterización de jóvenes:
 - Antes de limpieza: 436 jóvenes
 - Después de limpieza: 385 jóvenes con información completa (88%)
- Instrumento:
 - Antes de limpieza: 312 jóvenes
 - Después de limpieza inicial: 296 jóvenes (95%)
- Unión de caracterización e instrumentos
 - 278 jóvenes
- Unión de caracterización, instrumentos y comentarios:
 - **240 jóvenes**



Lecciones seleccionadas

Lección	Pregunta	# Observaciones
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 infografico. 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
Introducción	¿Cómo soy cuidado o cuidada en espacios como la escuela, la familia y la comunidad?	170



Variables del modelo

Modelo y cálculo de variables dependientes

La clasificación depende del porcentaje de preguntas que respondieron en las diferentes etiquetas:

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
MEDIO	Son los estudiantes que no tienen el nivel bajo ni alto		
ALTO	Si Siempre es 33% o más, Casi siempre es 50% o más, suma de Casi siempre y siempre son más del 33%		



Variable dependiente - nivel_ansiedad

Posterior al cálculo de la variable dependiente nivel_ansiedad, se obtuvo la siguiente distribución por categoría:

Nivel	# Observaciones	Frec. relativa
BAJO	97	40.4%
MEDIO	55	22.9%
ALTO	88	36.6%



Variable dependiente - nivel_victimizacion

Posterior al cálculo de la variable dependiente nivel_victimizacion, se obtuvo la siguiente distribución por categoría:

Nivel	# Observaciones	Frec. relativa
BAJO	193	80.4%
MEDIO	38	15.8%
ALTO	9	3.8%



Variable dependiente - nivel_violencia

Posterior al cálculo de la variable dependiente nivel_violencia, se obtuvo la siguiente distribución por categoría:

Nivel	# Observaciones	Frec. relativa
BAJO	95	39.6%
MEDIO	78	32.5%
ALTO	65	27.1%



Variables independientes

Se trabaja con un total de **384** variables independientes.

Las variables sociodemográficas utilizadas en el modelo son:

Variable	Valores
estrato	numérico entre 1 y 5
etnia	"N" o "S", dependiendo si pertenece o no a un grupo étnico
edad	numérico entre 12 y 18
sexo	"m", "h", "otro"



Variables independientes

Características de los comentarios

Variable	Descripción
n_caracteres	número de caracteres en el comentario
n_palabras	número de palabras en el comentario
dif_nchar_median	desviación en el número de caracteres respecto a la mediana del número de caracteres en el comentario de los jóvenes con un nivel bajo por lección
dif_npal_median	desviación en el número de palabras respecto a la mediana del número de palabras en el comentario de los jóvenes con un nivel bajo



Variables independientes

Presencia del top 10 de palabras con mayor presencia en lecciones: numérico entre 0 y 1, 1 indica que el joven utilizó la palabra en su comentario y 0 indica que no

Palabras lección 132		
mismo	bien	feliz
tristeza	cosas	ser
gustar	querer	vida
amistad		



Variables independientes

Frecuencia de términos – Frecuencia inversa del document (TF-IDF) para palabras comunes entre todos los jóvenes.

Cuanto mayor sea la puntuación, más relevante será esa palabra para ese individuo en particular.



Variables independientes

Presencia de parejas de palabras más comunes en las lecciones:

Ejemplo	Valores
sentir comodidad	numérico entre 0 y 1, 1 indica que el joven utilizó la pareja en su comentario y 0 indica que no



Variables independientes

Presencia de palabras alerta

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", "molesten", "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"



Modelos

Modelos utilizados

Para llevar a cabo este análisis, se utilizaron tres modelos de aprendizaje automático diferentes:

- Random Forest (90 combinaciones diferentes)
- Naive Bayes
- General Linear Model

Se dividió el conjunto de datos en dos partes: el conjunto de entrenamiento y el conjunto de prueba. Se utilizó un enfoque de partición 80/20, donde el 80% de los datos se usaron para entrenar los modelos y el 20% restante se usó para evaluar su desempeño.



Modelo: Ansiedad

Matriz de confusión - Nivel de ansiedad

Modelo: Random Forest

		Predicción		
		Alto	Medio	Bajo
Observación	Alto	10	0	7
	Medio	2	1	7
	Bajo	5	5	14

		Predicción		
		Alto	Medio	Bajo
Observación	Alto	Verdaderos Positivos	Falsos Negativos	Falsos Negativos
	Medio	Falsos Positivos	Verdaderos Positivos	Falsos Negativos
	Bajo	Falsos Positivos	Falsos Positivos	Verdaderos Positivos



Matriz de confusión - Nivel de ansiedad

Modelo: Random Forest

		Predicción			Total	Tasa de acierto	Tasa de error*
		Alto	Medio	Bajo			
Observación	Alto	10	0	7	17	58.8%	41.2%
	Medio	2	1	7	10	10%	70%
	Bajo	5	5	14	24	58.3%	41.7%



Matriz de confusión - Nivel de ansiedad

Modelo: NB

		Predicción			Total	Tasa de acierto	Tasa de error*
		Alto	Medio	Bajo			
Observación	Alto	7	3	7	17	41.2%	41.2%
	Medio	3	0	7	10	0%	70%
	Bajo	8	2	14	24	58.3%	41.7%



Matriz de confusión - Nivel de ansiedad

Modelo: GLM

		Predicción			Total	Tasa de acierto	Tasa de error*
		Alto	Medio	Bajo			
Observación	Alto	14	0	3	17	82.4%	17.6%
	Medio	5	1	4	10	10%	40%
	Bajo	4	5	15	24	62.5%	37.5%



Selección del modelo de ansiedad

		Modelo RF Tasa de error*	Modelo NB Tasa de error*	Modelo GLM Tasa de error*
Observación	Alto	41.2%	41.2%	17.6%
	Medio	70%	70%	40%
	Bajo	41.7%	41.7%	37.5%



Modelo seleccionado para ansiedad: GLM

Total: 49 variables

- Sociodemográficas: sexo
- Presencia de palabras: confianza, tranquilidad, riesgo, familia
- Frecuencia de palabras asociadas a sentimientos: tristeza, rabia y negativas
- Frecuencia de palabra: angustia
- Diferencia en el número de caracteres respecto a la mediana de jóvenes en bajo



Ejemplo

Usuario 530: Real: Alto / Predicción : Alto

Lección	Comentario
03 Me llegó carta	1. Cuento con los valores de: empatía, sororidad, amabilidad, respeto, entusiasmo. 2. Los momentos de felicidad más significativos en mi vida han sido cuando me siento bien con mi familia, cuando estamos todos juntos, y <i>cuando me he sentido estable mentalmente</i> .
02 Reconociendo otras vivencias	Me identifico con la <i>imagen de la salud mental</i> , ya que siempre procuro poner como prioridad mi bienestar mental y físico, creo que ambos son proporcionales, para así sentirme bien conmigo misma y transmitir buenas energías a los demás y convivir de una manera sana igualmente.



Ejemplo

Usuario 291: Real: Bajo / Predicción : Bajo

Lección	Comentario
03 Me llegó carta	1.R:honestidad,honradez,tolerancia,comprensión,respeto,solidaridad. 2.R:cuando me regalaron un computador sin importar todas las dificultades. 3.R:me siento estupendo ya que comparto una ambiente agradable con las personas que me rodean(colegio),me siento integrado en todo momento y no me desprecian. soy la mejor persona ya que mi compañerismo es enorme y rompe todas las barreras. 4.R:si 5.R:si
02 Reconociendo otras vivencias	No cuenta con datos



Ejemplo

Usuario 291: Real: Alto / Predicción : Bajo

Lección	Comentario
03 Me llegó carta	<p>yo sebas Sanchez tengo unos valores que me identifican, los cuales son: respeto, honestidad, lealtad, honradez, tolerancia y responsabilidad, mis capacidades mentales y físicas son: matemático, pensamiento rápido y socialmente, las físicas son: la altura, deportista y fuerza. una vez pase por un momento triste que fue cuando mi bisabuela se murió, fue un momento muy duro para mi ya que era el ser mas querido y bondadoso que existía en este mundo. mi mejor momento fue cuando yo pude ir al mar de niño ya que era unos de mis sueños de pequeño y también cuando operaron a mi mama y le fue super bien y todo le salió bien gracias a Dios.</p> <p>me siento feliz, por que tengo una vida tranquila de felicidad y sin problemas, también estable. yo soy feliz siendo una persona tranquila y responsable, pensando positivo y haciendo las cosas a conciencia y tomando buenas decisiones.</p>
02 Reconociendo otras vivencias	no eh vivido ninguna de las imagines o casos que hay



Modelo: Victimización

Matriz de confusión - Nivel de victimización

Modelo seleccionado GLM

		Predicción			Total	Tasa de acierto	Tasa de error*
		Alto	Medio	Bajo			
Observación	Alto	0	1	0	1	0%	0%
	Medio	0	2	10	12	16.7%	83.3%
	Bajo	0	5	33	38	86.8%	13.2%



Modelo: Violencia

Matriz de confusión - Nivel de violencia

Modelo seleccionado : Random forest

		Predicción			Total	Tasa de acierto	Tasa de error*
		Alto	Medio	Bajo			
Observación	Alto	7	4	6	17	41.2%	35.3%
	Medio	1	9	5	15	60%	33.3%
	Bajo	2	7	4	19	86.8%	52.6%



Pasos a seguir

Siguientes pasos

- Correr nuevamente los modelos cuando se tengan más datos con la expectativa de calibrarlos mejor
- Construcción de la solución web

