Informe Final – Detección de Riesgo Suicida en Jóvenes mediante Aprendizaje Automático

---

Introducción

En la actualidad, el suicidio juvenil constituye una de las principales causas de muerte prevenibles en América Latina, con un incremento sostenido en las últimas décadas. La provincia de Tierra del Fuego, por sus particularidades geográficas, demográficas y sociales, presenta indicadores especialmente sensibles en esta materia. Frente a esta realidad, el presente trabajo se propuso explorar la posibilidad de aplicar técnicas de aprendizaje automático como apoyo a los dispositivos clínicos existentes, con el fin de anticipar situaciones de riesgo y colaborar en la toma de decisiones tempranas dentro del primer nivel de atención.

Este proyecto se enmarca en la Tecnicatura en Ciencia de Datos e Inteligencia Artificial y combina el procesamiento, análisis y modelado de datos con un enfoque netamente aplicado a problemáticas reales del sistema público de salud.

Objetivo

El objetivo principal fue construir un modelo supervisado de clasificación que permita anticipar casos de riesgo suicida en población joven (12 a 29 años), a partir de registros clínicos disponibles en el sistema de salud provincial. Se definió como variable objetivo "riesgo_suicida" a aquellos casos donde se refiriera explícitamente ideación, intento o riesgo suicida, conforme al contenido de los campos subjetivos registrados.

En todo momento, se priorizó la sensibilidad del modelo, entendiendo que resulta más relevante identificar correctamente a quienes se encuentran en riesgo, incluso si ello implica incurrir en falsos positivos.

Dataset: origen, extracción y variables

Los datos fueron extraídos mediante consultas SQL ejecutadas sobre el sistema clínico de gestión del Ministerio de Salud de Tierra del Fuego. Se obtuvieron registros anonimizados de pacientes que entre 2018 y 2024 tuvieron al menos un contacto con servicios de salud mental. Las tablas relevadas fueron:

- pacientes_salud_mental.csv: datos sociodemográficos, sexo, edad, cobertura, centro de atención, localidad, nivel de estudio.
- diagnosticos.csv: diagnósticos estructurados (CIE-10), subjetivos, fecha.
- internaciones.csv: frecuencia, duración promedio, especialidad, tipo de egreso.

Se consolidó todo a nivel paciente, aplicando filtros, validaciones y transformaciones específicas.

Variables construidas

A partir de los datos originales, se generaron variables derivadas para representar patrones clínicos y sociales relevantes:

- **n_diagnosticos, n_internaciones, dias_estada_avg:** cantidades y promedios a nivel paciente.
- **tipo_egreso_mas_frecuente:** categorización del tipo de alta más habitual.
- **internacion_salud_mental:** booleano construido a partir de la especialidad.
- **refiere_ansiedad, refiere_estres, refiere_insomnio:** presencia textual extraída de los campos subjetivos mediante expresiones regulares que respetan variantes léxicas.
- **cobertura_cat:** recategorización de obra social (sin cobertura / estatal / privada).
- **nivel_estudio, centro_atencion_mas_frecuente:** según predominancia.

- **riesgo_suicida:** variable objetivo, extraída exclusivamente de textos subjetivos, con control de codificación cruzada con CIE-10 (sólo se validó si el código comenzaba con “F”).

ETL: limpieza, consolidación y transformación

Durante la etapa de procesamiento se realizaron las siguientes tareas:

- Validación de claves (paciente_nro), control de duplicados.
- Limpieza de fechas, cálculo de edad exacta.
- Conversión de texto libre a expresiones binarias.
- Agregación de registros a nivel paciente.
- Imputación de faltantes y one-hot encoding para variables categóricas.
- Unificación de las tres tablas en un dataset final de 10.255 registros únicos.

Análisis exploratorio

- La clase positiva (riesgo_suicida) representó un 8.5% del total.
- Más del 60% de los pacientes no contaban con obra social.
- La institución con mayor concentración fue el HRU, seguida por el HRRG.
- La coocurrencia de síntomas subjetivos mostró: ansiedad (16.8%), estrés (6.1%) e insomnio (4.2%).
- Se observaron casos extremos con más de 700 diagnósticos y más de 200 internaciones.
- Se descartaron variables con más del 60% de nulos, y en otras se imputó “SIN DATOS”.

Modelado: regresión logística y árbol de decisión

Se seleccionaron variables en función de su valor predictivo, consistencia clínica y distribución. Las variables categóricas se codificaron con one-hot y se aplicó remuestreo para abordar el desbalance.

Regresión logística sin balanceo

- **Accuracy: 0.92**
- **Recall clase 1: 0.1264**
- **F1-score clase 1: 0.2115**

El modelo logró detectar apenas 22 verdaderos positivos sobre 174, incurriendo en una alta tasa de falsos negativos. A pesar de su precisión general, su rendimiento sobre la clase minoritaria fue insuficiente.

Regresión logística balanceada

- **Accuracy: 0.8108**
- **Recall clase 1: 0.65**
- **F1-score clase 1: 0.37**

El remuestreo permitió mejorar significativamente la sensibilidad: el modelo logró captar 113 verdaderos positivos, lo que representó una mejora sustancial en términos del objetivo principal, aunque a costa de perder algo de precisión global.

Árbol de decisión balanceado

- **Accuracy: 0.87**
- **Recall clase 1: 0.32**
- **F1-score clase 1: 0.29**

El árbol de decisión se entrenó sobre el mismo conjunto balanceado. Si bien tuvo menor recall que la regresión balanceada, permitió visualizar de forma jerárquica la importancia relativa de las variables.

Las más relevantes fueron:

- **n_diagnosticos**
- **refiere_ansiedad**
- **internacion_salud_mental**
- **cobertura_cat**
- **n_internaciones**
- **centro_atencion_mas_frecuente**

Comparativa final de modelos

Modelo	**Accuracy**	**Recall (Clase 1)**	**F1-score (Clase 1)**
Regresión logística	**0.92**	**0.13**	**0.21**
Regresión balanceada	**0.81**	**0.65**	**0.37**
Árbol de decisión balanceado	**0.87**	**0.32**	**0.29**

Conclusiones finales

El proyecto logró cumplir con el objetivo principal: desarrollar un modelo supervisado que permita detectar con prioridad los casos de riesgo suicida entre jóvenes, aprovechando datos del sistema de salud público provincial.

La regresión logística con clases balanceadas resultó ser la estrategia más eficaz en términos de sensibilidad, que era el criterio central del trabajo. El árbol, en tanto, aportó claridad interpretativa y permitió observar con mayor detalle la lógica de decisiones subyacente al fenómeno clínico.

Además, se detectaron patrones administrativos y asistenciales que merecen ser tenidos en cuenta por las políticas públicas: la falta de cobertura sanitaria, la alta concentración institucional, la frecuencia de internaciones y la coocurrencia de determinados síntomas subjetivos.

Proyecciones y líneas futuras

- Incorporar técnicas más sofisticadas de sobremuestreo como SMOTE o ADASYN.
- Explorar modelos más complejos como Random Forest o Gradient Boosting.
- Realizar validación cruzada en cohortes nuevas.
- Incorporar análisis longitudinales para observar evolución temporal del riesgo.
- Fortalecer el trabajo colaborativo con equipos clínicos que puedan retroalimentar los hallazgos.

Este trabajo no busca reemplazar el criterio clínico, sino complementar la capacidad institucional de respuesta temprana con herramientas que permitan anticiparse al agravamiento de las situaciones de riesgo.