CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título X Tesis de Postgrado

Título del trabajo: ___ Diseño de un modelo de predicción de accidentabilidad en actividades y deportes al aire libre como base para el desarrollo de una aplicación móvil_____

Nombre del candidato(a): __**Oliver Franciso Esteban Bravo Martinez**_____

Carrera / Grado: _____**Magíster en Tecnologías de la Información**_____

Campus: ____San Joaquin_____ **Departamento:** _____Informática_____

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, **José Luis Martí Lara**, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

X El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 06 - 04 - 2026 **Firma:** _____

Estudiante o Candidato(a):

Fecha: ___03-04-2026_____ **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Diseño de un modelo de predicción de accidentabilidad en actividades y deportes al aire libre como base para el desarrollo de una aplicación móvil

Oliver Bravo Martínez

4/17 Myoora Rd, Toorak VIC 3142, Melbourne, Australia

Oliver.bravo@alumnos.usm.cl

Resumen: El incremento en la práctica de actividades al aire libre, como el montañismo, escalada, senderismo y ciclismo de montaña, ha generado un aumento de la tasa de accidentabilidad, algunos con consecuencias graves o fatales, lo que demuestra una necesidad de desarrollo de herramientas que permitan gestionar riesgos asociados a estas actividades. Se propone el desarrollo de un modelo predictivo basado en técnicas de aprendizaje automático, con el objetivo de determinar la probabilidad de accidentabilidad basada en ciertas variables, considerando datos históricos, condiciones ambientales e información en tiempo real o asíncrona. Para validar la propuesta, se recopilaron, analizaron y procesaron datos relevantes utilizando metodologías de aprendizaje automático, evaluando distintos algoritmos y ajustando los parámetros para optimizar la precisión del modelo. Se aplicaron métricas de desempeño para seleccionar la mejor configuración y se realizaron pruebas con datos reales y simulados. Los resultados de este trabajo muestran la capacidad del modelo para determinar el nivel de riesgo con un alto grado de exactitud. Además, se presenta una propuesta de diseño de una aplicación móvil que registre actividades e integre el modelo de predicción desarrollado, lo que permitirá alimentar constantemente el modelo. Se espera que esta herramienta ayude a la prevención de accidentes y a una gestionar mejor las actividades en lugares remotos.

Palabras Clave: Prevención de accidentes, Gestión de riesgos, Aprendizaje automático, Aplicaciones Móviles.

1 Introducción

1.1 Contexto, motivación y problemática

En los últimos años, se ha observado un crecimiento sostenido en la práctica de actividades deportivas y recreativas al aire libre, especialmente en entornos naturales y remotos como montañas, parques nacionales y rutas de senderismo, así como deportes más clásicos como Ciclismo, SKI o Snowboard. Esto ha sido potenciado por diversos factores como el creciente interés por el bienestar físico y mental, una mayor conciencia sobre la salud y la necesidad de reconexión con la naturaleza, además esto se intensificó posterior a los periodos de confinamiento experimentados durante pandemia (Wunsch et al., 2022; Ahsan & Abualait, 2024). El contacto con la naturaleza ha mostrado tener efectos positivos en la salud mental y física, como reducción del estrés, mejora del estado de ánimo (Passmore & Howell, 2014) y el fortalecimiento del sistema inmunológico. El aumento en las actividades al aire libre, también se refleja en el crecimiento de la industria de deporte y actividades al aire libre que conlleva a su vez una mayor cantidad de marcas en el mercado, y un aumento en las ventas y servicios relacionados. Igualmente, el aumento en la participación en estas actividades ha incrementado la ocurrencia de incidentes y accidentes de acuerdo con diversas entidades respetadas en el ámbito (Comité de Seguridad FEDME[1], 2020), lo que confirma la necesidad de desarrollar estrategias preventivas más eficaces, basadas en datos y orientadas a la gestión del riesgo.

La motivación para desarrollar esta investigación nace tanto de la experiencia personal del autor en actividades al aire libre como del haber presenciado o participado en situaciones de riesgo, incluyendo incidentes y accidentes que, en retrospectiva, podrían haberse evitado mediante una adecuada gestión del riesgo o el uso de herramientas tecnológicas preventivas. Estas situaciones han permitido identificar vacíos importantes en la preparación, planificación y toma de decisiones que enfrentan quienes practican deportes en entornos naturales y remotos.

La principal problemática es que gran parte de los participantes en actividades al aire libre carece de conocimientos adecuados sobre los riesgos en entornos naturales. Esta limitación se manifiesta en el desconocimiento de factores críticos como condiciones climáticas, exposición, características del terreno, nivel de dificultad de la actividad y la

[1] Federación española de deportes de Montaña y Escalada

propia experiencia del individuo. La ausencia de esta información dificulta una gestión del riesgo efectiva y oportuna, incrementando la probabilidad de exposición a situaciones peligrosas.

La falta de conocimiento puede generar una percepción alterada del riesgo. Por un lado, puede llevar a una subestimación de los peligros reales, lo que puede causar toma de decisiones inadecuadas; por otro lado, puede provocar una aversión al riesgo, que reduzca la participación o genere comportamientos inseguros por desconocimiento. Por ejemplo, una exposición prolongada en un terreno con condiciones adversas puede ser tan riesgosa como atravesar rápidamente una zona inestable sin la preparación adecuada. En ambos casos, la falta de herramientas de soporte que integren y comuniquen información relevante en tiempo real o eventualmente asíncrona, representa una brecha significativa para la prevención de accidentes y la correcta gestión de los riesgos.

1.2 Definición del problema

Actualmente, existen diversas aplicaciones utilizadas para el registro de actividades deportivas al aire libre, cuyo enfoque principal se centra en el monitoreo del rendimiento físico o la documentación de rutas. Ejemplos de estas aplicaciones son Suda (Suda Outdoors, 2024), Strava (Strava Inc., 2024), Wikiloc (Wikiloc Outdoor S.L., 2024), AllTrails (AllTrails, LLC, 2024) y Komoot (Komoot GmbH, 2024) y repositorios colaborativos de rutas como Wikiexplora (Wikiexplora, 2024) y Andeshandbook (Sociedad Geográfica Andeshandbook, 2024). Sin embargo, estas herramientas no consideran ni integran variables críticas para la prevención de accidentes, como las condiciones climáticas, el tipo de terreno, la concurrencia de personas (a excepción de algunas funcionalidades limitadas en zonas urbanas) o el historial de incidentes en zonas específicas. Tampoco incorporan mecanismos de análisis predictivo que integren datos históricos en tiempo real o asíncrono para generar alertas, estimaciones o evaluaciones de riesgo.

A partir de la revisión de antecedentes, se observa una falta de modelos de predicción de accidentabilidad que utilicen tanto variables internas (relacionadas con el usuario) como externas (provenientes de servicios en línea, como datos meteorológicos, geográficos o variables inferidas) siendo que los datos están disponibles tanto en las aplicaciones como en servicios de libre acceso como clima y datos geográficos. Asimismo, no existen índices actualizados de accidentabilidad ni sistemas de documentación de accidentes e incidentes accesibles y ampliamente difundidos que permitan a los usuarios conocer los riesgos asociados a una zona o actividad específica, una problemática reconocida por las principales entidades del sector (Comité de Seguridad FEDME, 2018; Comité de Seguridad FEDME, 2020). Esta falta de información estructurada y de herramientas predictivas representa una limitación importante para la gestión preventiva de la seguridad en actividades al aire libre o lugares remotos, y evidencia la necesidad de desarrollar soluciones tecnológicas que integren múltiples fuentes de datos para apoyar la toma de decisiones informadas y la correcta gestión del riesgo.

1.3 Propuesta de solución

A partir del análisis de los antecedentes, la presente propuesta tiene como objetivo el desarrollo de un modelo predictivo capaz de procesar datos de diversas fuentes, como aplicaciones o plataformas digitales, registros de información de usuarios basado en bases de datos de actividad y servicios accesibles en línea como **OpenWeatherMap**, **National Oceanic and Atmospheric Administration**, **OpenStreetMap** y **Servicio de Datos de Elevación (SRTM/DEM)** que se puedan consumir libremente. A través del análisis de estas variables, se busca identificar posibles correlaciones significativas que permitan estimar el nivel de riesgo asociado, tanto a una actividad específica como en el entorno geográfico en el que se pretende llevar a cabo; este enfoque quiere apoyar la toma de decisiones mediante el uso de herramientas basadas en datos. Como parte del desarrollo del modelo propuesto, el primer paso consiste en la identificación y selección de las fuentes de datos de acuerdo con la información disponible. Una vez definidas las fuentes y de acuerdo con la disponibilidad de los datos, se continua con la recolección de la información como data cruda. Posteriormente, se realiza una depuración que incluye la limpieza de datos y la eliminación de ruido, para asegurar su calidad e idoneidad para el análisis. Esta etapa permite distinguir las variables más relevantes para el modelo. A continuación, los datos se transforman y estructuran mediante técnicas de preprocesamiento, con el objetivo de facilitar su clasificación, agrupamiento, asociación o detección de patrones de comportamiento. Este proceso permite identificar las variables que aportan valor al diseño del modelo predictivo. Al completar esta fase, se da inicio al desarrollo del modelo de aprendizaje automático, utilizando las variables

seleccionadas como base. La **Figura 1** presenta una visión general del proceso, ilustrando la relación entre las variables consideradas, las etapas de procesamiento de datos y los modelos resultantes.

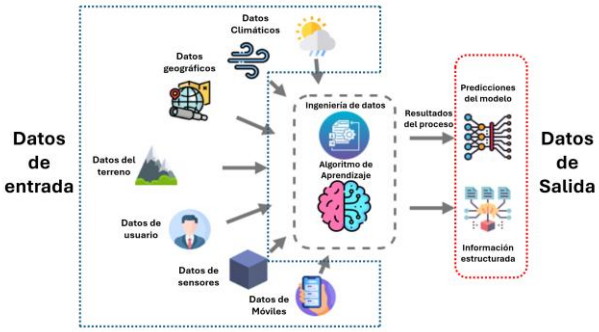

Figura 1. Flujo *Metodológico de Procesamiento de Datos para la Estimación de Riesgo*
Nota: Elaboración propia.

El desarrollo del modelo predictivo requiere rigurosidad al momento de la recolección y gestión de datos. Inicialmente, se desarrolla una base de datos de usuarios para organizar las variables disponibles y analizar su potencial impacto, en paralelo, se alimenta con información de accidentalidad. Esta información es adaptada, procesada e incorporada al modelo mediante las fases de entrenamiento para optimizar su capacidad predictiva. Tanto los datos procesados como los resultados del entrenamiento se almacenan de forma segura y se visualiza para su análisis e interpretación. La **Figura 2** presenta una arquitectura conceptual de referencia (utilizando servicios de AWS), aunque su configuración final puede variar y está pensada para el procesamiento masivo de datos para el desarrollo de la aplicación, se presenta para visualizar conceptualmente el procesamiento requerido.

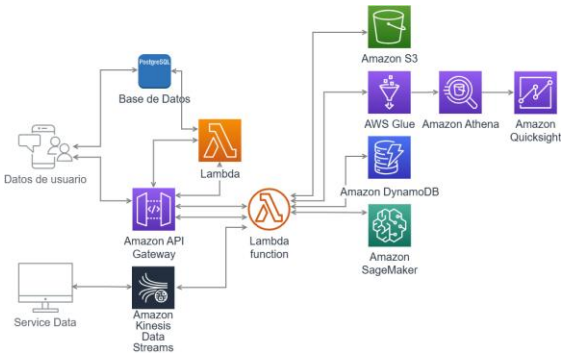

Figura 2. *Diagrama de Arquitectura Conceptual para la Captura y Procesamiento de Datos en AWS*
Nota: Elaboración propia.

1.4 Organización del informe

El informe está estructurado para facilitar la comprensión del diseño y validación del modelo predictivo, comenzando con una **Introducción** que da contexto de la problemática, posteriormente los **Objetivos** del estudio para delimitar el alcance. A continuación, el **Marco Teórico** y **Estado del Arte** donde se entrega el contexto y las bases de estudio además de analizar modelos y teóricas existentes relevantes para el modelo.

El desarrollo de la tesina se compone de tres secciones: la **Metodología** (definición de recolección, preprocesamiento de datos y técnicas de aprendizaje automático), el **Diseño y Arquitectura del Modelo** (justificación de variables, configuración y refinamiento), y la **Implementación y Resultados** (documentación de pruebas, validación y métricas de desempeño). Finalmente, se presenta el **Diseño de la Aplicación Móvil** (con *mock-ups* y manual de marca de la aplicación, muy importante para la captura de datos en el futuro), finalizando con las **Conclusiones y Trabajo Futuro**, donde se sintetizan los resultados, validación de la hipótesis y se sugieren líneas de investigación futuras.

2 Objetivos e hipótesis de trabajo

2.1 Objetivo General

"Desarrollar y validar un modelo predictivo basado en técnicas de aprendizaje automático que integre variables internas y externas con el fin de estimar el nivel de riesgo asociado a la práctica de actividades al aire libre en entornos naturales o remotos, alcanzando, al menos un 75% de exactitud en la predicción de la probabilidad de accidentalidad."

2.2 Objetivos específicos

- Determinar los factores críticos de accidentabilidad identificando al menos 10 variables significativas que incidan en el riesgo asociado a actividades al aire libre, asegurando la calidad y exactitud de los datos mediante una tasa de errores y valores nulos inferior al 10%.

- Alcanzar al menos un 75% de exactitud en la estimación del nivel de riesgo mediante el diseño, entrenamiento y optimización de modelos de aprendizaje automático, con el fin de desarrollar un algoritmo confiable para la toma de decisiones preventivas en actividades al aire libre.

- Definir la arquitectura y el diseño de la aplicación, integrando el modelo predictivo para asegurar su operatividad, escalabilidad y disponibilidad en un entorno productivo.

2.3 Definición de hipótesis

"Un modelo predictivo basado en técnicas de aprendizaje automático puede estimar con al menos un 75% de exactitud la probabilidad de accidentabilidad de un usuario en una zona geográfica determinada, considerando variables internas (como perfil, experiencia y comportamiento del usuario) y externas (como condiciones climáticas, geográficas y concurrencia), en el contexto de actividades deportivas o recreativas al aire libre en entornos urbanos o remotos."

3 Marco teórico y estado del arte

La accidentabilidad en actividades al aire libre está ligada a una amplia variedad de factores, por lo que resulta fundamental identificar y analizar tanto los factores internos de los participantes (experiencia y preparación) como los factores externos (clima, geografía o concurrencia). Diversos estudios en Chile y España (Villota, 2017; Fica, 2019) han documentado estos riesgos y destacan la importancia de incorporar variables relevantes en modelos predictivos (Dalipi et al., 2015; Basso et al., 2018). Por lo tanto, la documentación de estos factores resulta valioso para el diseño de modelos eficaces de gestión del riesgo (Schubert, 2001, 2007, 2009). En consecuencia, este capítulo se estructura sobre dos pilares fundamentales: la base técnica de herramientas de aprendizaje automático para la predicción de riesgos y el análisis del estado del arte de la industria tecnológica relacionado a este tipo de actividades.

3.1 Marco teórico

3.1.1 Gestión del riesgo en entornos naturales

La gestión del riesgo en entornos naturales constituye un proceso en sí mismo, al analizar, evaluar y monitorear los peligros asociados a una actividad (Schubert, 2001). Este proceso no es estático, sino que requiere una planificación y evaluación continua además de una cultura de seguridad que permita al usuario tomar decisiones informadas antes y durante la actividad (Taibo Vázquez, 2022).

Históricamente, la evaluación del riesgo se basa en el conocimiento técnico y la experiencia personal del individuo (Sánchez, 2019). Sin embargo, la complejidad de las variables externas y la alta proporción de participantes inexpertos o principiantes requiere un acercamiento más objetivo y automático (Villota, 2017). En este contexto, el modelo

predictivo propuesto se introduce en la fase de análisis y evaluación del riesgo previo o durante la actividad, ya que ofrece una herramienta capaz de cuantificar la probabilidad de accidentalidad a partir de variables dinámicas, complementando la capacidad de gestión del riesgo del usuario, especialmente de aquellos con menor formación o experiencia (Fica, 2019).

El riesgo en actividades al aire libre o lugares remoto se define como "la combinación de la probabilidad de que ocurra un evento adverso y la severidad de sus consecuencias". La gestión de este riesgo requiere diferenciar entre peligro (una condición intrínseca al medio, como por ejemplo la caída de rocas) y riesgo (la probabilidad de que el usuario se exponga a ese peligro).

Organizaciones clave en la prevención, como la FEDME, promueven activamente una cultura de seguridad basada en la planificación y la prevención (Comité de Seguridad FEDME, 2017; Guía Ilustrada, 2020). A pesar de esto, la literatura revisada por la propia FEDME destaca la ausencia de índices actualizados y datos que permitan analizar la accidentalidad de manera correcta, esto por su parte limita la capacidad de generar herramientas predictivas (Comité de Seguridad FEDME, 2020).

3.1.2 Enfoque predictivo basado en aprendizaje automático

Para el desarrollo de la investigación y la implementación del modelo, se busca un enfoque basado en técnicas de aprendizaje automático, por su capacidad para adaptarse a la incorporación de nuevas variables. Dado que la relación entre los factores de accidentabilidad y las condiciones del entorno es de naturaleza no lineal, se compararon los algoritmos que se detallan a continuación.

3.1.2.1 Regresión logística

Corresponde a un clasificador lineal probabilístico que utiliza la función logística (sigmoide) para identificar combinaciones lineales de características en un intervalo [0,1]. Se implementa como base para establecer un umbral mínimo de desempeño. Su principal limitación es la linealidad en los límites de decisión, lo que restringe su capacidad para modelar interacciones complejas. Se define la siguiente fórmula matemática para la regresión logística (Hosmer & Lemeshow, 2013).

$$\sigma(z) = \frac{1}{(1 + e^{-z})}$$

donde:

$\sigma(z)$: Es el resultado de la función sigmoide (probabilidad entre 0 y 1).
e: Es la base del logaritmo natural (número de Euler, aproximadamente 2.71828).
z: Es la entrada de la función, comúnmente representada como la combinación lineal de las características de entrada ($z = \beta$).

3.1.2.2 Arboles de decisión

Representan modelos no paramétricos que segmentan el espacio de características mediante reglas de decisión binarias recursivas. Se utilizan para identificar patrones locales en los datos. Su principal debilidad es la alta varianza, lo que suele derivar en un ajuste excesivo a los datos de entrenamiento (Breiman et al., 1984).

$$G = 1 - \sum_{i=1}^{C} (p_i)^2$$

donde:

G: Es el índice de impureza de Gini.
C: Es el número total de clases (ej. Accidente / No Accidente).
p_i: Es la probabilidad de que un elemento sea clasificado en una categoría específica dentro de ese nodo.

3.1.2.3 Bosques aleatorios (*Random Forest*)

Es una técnica de ensamble basada en *Bagging* (*Bootstrap Aggregating*). Su fundamento es la construcción de múltiples árboles de decisión independientes, donde la predicción final es el resultado del promedio (regresión) o votación mayoritaria (clasificación) de todos ellos. Esto reduce la varianza global del sistema sin aumentar el sesgo (Breiman, 2001).

La predicción final para clasificación se expresa como:

$$\hat{Y} = moda\{T_1(x), T_2(x), \dots, , T_n(x)\}$$

donde:

$\hat{Y}$: Es la clase predicha por el bosque.
$T_n(x)$: Es la predicción individual del n-ésimo árbol de decisión.
n: Es el número total de árboles en el ensamble.

3.1.2.4 Máquinas de Soporte Vectorial (SVM)

Este algoritmo busca el hiperplano óptimo que maximiza el margen de separación entre clases en un espacio de n-dimensiones. Ante datos no lineales, utiliza funciones kernel para proyectar los datos a un espacio de mayor dimensionalidad donde la separación sea posible (Cortes & Vapnik, 1995).

El objetivo es minimizar la siguiente función de costo:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i$$

donde:

w: Es el vector normal al hiperplano de separación.
C: Es el parámetro de regularización que controla el compromiso entre el margen y el error de clasificación.
ε_i: Son las variables de holgura que permiten errores controlados en los márgenes.

3.1.2.5 K-vecinos más próximos

Es un algoritmo de aprendizaje basado en instancias que clasifica un punto de datos según la etiqueta de los "k" puntos más cercanos en el espacio vectorial. No asume una distribución previa de los datos, lo que lo hace útil para identificar focos de riesgo geográfico (Cover & Hart, 1967).

La proximidad se calcula comúnmente mediante la Distancia Euclídea:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

donde:

$d(p,q)$: Es la distancia entre el nuevo punto p y un punto histórico q.
n: Es el número de características (variables) consideradas.
p_i, q_i: Son los valores de la característica i para cada punto.

3.1.2.6 *XGBoost* (Aumento de Gradiente Extremo)

Es un algoritmo de ensamble secuencial donde cada árbol es entrenado para corregir los errores residuales del árbol anterior mediante el descenso de gradiente. Incluye términos de regularización para controlar la complejidad y mejorar la generalización (Chen & Guestrin, 2016).

La función objetivo a optimizar en cada iteración t es:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

donde:

l: Es la función de pérdida que mide la diferencia entre la realidad y la predicción.
$f_t(x_i)$: Es el nuevo árbol que se añade en el paso t .
$\Omega(f_t)$: Es el término de regularización (L1/L2) que penaliza la complejidad del modelo para evitar el sobreajuste.

3.1.3 Metodologías de Trabajo en Proyectos de Ciencia de Datos

El ciclo de vida de un proyecto basado en aprendizaje automático requiere un marco metodológico que asegure la calidad y la trazabilidad del procesamiento de datos. En la industria y la academia, existen al menos tres marcos de trabajo predominantes:

3.1.3.1 KDD (*Knowledge Discovery in Databases*)

Es una de las metodologías pioneras, centrada en el proceso iterativo de extraer conocimiento útil a partir de grandes volúmenes de datos. Se compone de cinco etapas: selección, preprocesamiento, transformación, minería de datos e interpretación (Fayyad et al., 1996). Si bien es rigurosa en el tratamiento técnico del dato, carece de una fase explícita orientada a los objetivos de negocio o seguridad del usuario.

3.1.3.2 SEMMA (*Sample, Explore, Modify, Model, Assess*)

Desarrollada por el Instituto SAS, esta metodología se enfoca en el aspecto técnico y estadístico del modelado. Su estructura es lineal: muestreo de datos, exploración de tendencias, modificación (limpieza), modelado y evaluación. Se diferencia de KDD al ser más práctica, aunque sigue omitiendo la integración con las necesidades específicas del entorno o el despliegue final de la solución.

3.1.3.3 CRISP-DM (*Cross-Industry Standard Process for Data Mining*)

Es el estándar más utilizado actualmente debido a su enfoque cíclico e integral. A diferencia de las anteriores, introduce como fase inicial el "Entendimiento del Negocio", lo cual permite alinear los algoritmos con el objetivo real de la tesis: la prevención de riesgos. Sus seis fases (Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Despliegue) permiten retroceder a etapas previas para ajustar el modelo según se descubran nuevos patrones en la accidentalidad (Chapman et al., 2000).

3.1.4 Métricas de Evaluación de Calidad

La evaluación del modelo representa una fase muy importante y se mide de acuerdo con métricas estadísticas estandarizadas, requeridas para validar el cumplimiento de la hipótesis definida. Debido a que el modelo debe estimar un nivel de riesgo y determinar la probabilidad de accidentalidad, se emplean métricas tanto de regresión como de clasificación:

3.1.4.1 Métricas de Estimación (Regresión)

Se utilizan para cuantificar la magnitud del error en la predicción del índice de riesgo numérico.

- **Error Absoluto Medio (MAE):** Representa el promedio de las diferencias absolutas entre los valores reales y las predicciones.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Permite conocer la magnitud promedio del error en las mismas unidades que la variable medida, ofreciendo una interpretación directa del error del modelo (Chai & Draxler, 2014). Se utiliza cuando se requiere mostrar el error en una escala lineal y fácil de interpretar.

- **Raíz del Error Cuadrático Medio (RMSE)**: Es la raíz cuadrada del promedio de los errores cuadrados.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

A diferencia del error absoluto medio, esta métrica penaliza los errores grandes. En contextos de seguridad, es vital para identificar y reducir desviaciones críticas (Chai & Draxler, 2014). Se utiliza cuando se requiere identificar y penalizar errores grandes.

- **Coeficiente de Determinación (R^2):** Indica la proporción de la varianza de la variable dependiente que es explicada por el modelo.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Determina qué tan bien se ajustan las predicciones a la distribución real de los datos, siendo un indicador de la calidad global del modelo de regresión (James et al., 2013). Se utiliza en las fases de selección de variables y ajuste global para determinar qué tan bien el modelo captura la complejidad y variabilidad de los datos analizados.

3.1.4.2 Métricas de Clasificación

Evalúan la capacidad del modelo para categorizar correctamente los eventos.

- **Matriz de Confusión**: Es una herramienta que desglosa los aciertos y errores en cuatro categorías: Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN).

$$C = \begin{bmatrix} VP & FP \\ FN & VN \end{bmatrix}$$

Proporciona un mapa detallado del desempeño del clasificador, permitiendo identificar si el modelo tiene sesgos hacia una clase específica.

- **Precisión**: Mide la proporción de las predicciones positivas que fueron correctas.

$$Precision = \frac{VP}{VP + FP}$$

Asegura la confiabilidad de las alertas emitidas por el sistema, minimizando las "falsas alarmas" que podrían generar desconfianza en el usuario.

- **Sensibilidad (*Recall*)**: Mide la capacidad del modelo para detectar todos los casos positivos reales.

$$Sensibilidad = \frac{VP}{VP + FN}$$

Una sensibilidad alta garantiza que la mayor cantidad de casos reales sean detectados, minimizando los Falsos Negativos (Powers, 2011).

- **Exactitud (*Accuracy*)**: Representa la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de casos evaluados.

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

Una exactitud alta garantiza que el modelo clasifique correctamente la mayoría de los eventos de riesgo, siendo la métrica fundamental para validar el desempeño global del clasificador (Sokolova & Lapalme, 2009).

- **Puntaje F1 (*F1-Score*)**: Es la media armónica entre la Precisión y la sensibilidad.

$$F1 = 2 \cdot \frac{Precision \cdot Sensibilidad}{Precision + Sensibilidad}$$

Proporciona una medida balanceada de desempeño, especialmente útil cuando existe un desequilibrio en los datos (Sokolova & Lapalme, 2009).

- **Curva ROC y AUC**: La curva ROC representa la relación entre la tasa de verdaderos positivos (TVP) y la tasa de falsos positivos (TFP). El AUC (*Area Under the Curve*) resume este gráfico en un solo valor.

$$AUC = \int_0^1 TVP(TFP)d(TFP)$$

Permite evaluar la capacidad de discriminación del modelo independientemente del umbral de decisión, siendo fundamental para comparar el desempeño entre distintos algoritmos (Fawcett, 2006).

3.2 Estado del Arte

La literatura especializada muestra una gran cantidad de modelos de predicción basados en aprendizaje automático, los cuales han sido aplicados exitosamente en distintos contextos para estimar resultados a partir de variables específicas. En el caso particular de esta investigación, el registro de accidentalidad se encuentra separado de un modelo de predicción como, por ejemplo, los registros de accidentalidad en montaña de Chile o de España que se tomaron como referencia (Villota, 2017; Fica, 2019). Por otro lado, si bien existen modelos de gestión del riesgo orientados a usuarios más expertos o con formación técnica que facilitan la gestión del riesgo, estas herramientas suelen excluir a personas aficionadas o con menor conocimiento técnico. Este último grupo representa una proporción significativa de los participantes en actividades al aire libre y está expuesto a los mismos riesgos que los grupos con mayor experiencia (Schubert, 2001, 2007, 2009).

La revisión bibliográfica evidencia que, a nivel local, la cantidad de literatura asociada al tema es limitada y el tema no se encuentra muy desarrollado, las aproximaciones que se han dado en el ámbito están relacionados principalmente a la documentación de accidentes o literatura relacionada a cómo enfrentar situaciones de riesgo e incluso a la gestión del riesgo que finalmente requiere de un conocimiento más técnico. Por otra parte, la literatura asociada al desarrollo de la tesina está orientada a modelos de deportes en particular a la predicción de accidentabilidad en otros ámbitos, como accidentes de tráfico o en la construcción, utilizando modelos para predecir la probabilidad de accidentabilidad, lo que permite aplicar los conocimientos desarrollados en estos ámbitos, pero no pueden ser asociados a deportes o actividades al aire libre.

Un ejemplo relevante en el ámbito deportivo es el desarrollo de un modelo predictivo aplicado al SKI (Dalipi et al., 2015), en el cual se implementó un modelo de red neuronal artificial con el objetivo de predecir la ocurrencia de accidentes y lesiones asociadas a la práctica de este deporte. Este modelo fue diseñado para asistir a los equipos médicos y patrullas de esquí en la prevención y gestión de emergencias en terreno. Para su construcción, se desarrolló una base de datos que recopiló información sobre accidentes, lo que permitió alimentar el modelo y estimar la frecuencia y la probabilidad de ocurrencia de los eventos. Las variables consideradas en el modelo incluyeron variables

como la cantidad de días de práctica, la afluencia de personas, la acumulación de nieve, la temperatura y la precipitación. Los resultados obtenidos demostraron una capacidad predictiva alta; específicamente, el modelo alcanzó un coeficiente de correlación R de 0.994 y un error cuadrático medio MSE de 0.003. No obstante, los autores señalan que la incorporación de un mayor volumen de datos y de variables adicionales como indicadores de riesgo real y coordenadas geográficas, podría mejorar aún más la precisión del modelo. Esta ampliación permitiría, por ejemplo, la generación de mapas de riesgo que identifiquen zonas con alta probabilidad de avalanchas o accidentabilidad, como indican los autores.

Otro caso relevante corresponde al desarrollo de un modelo de predicción de accidentabilidad en el contexto del tránsito vehicular, particularmente en autopistas urbanas (Basso et al., 2018), cuyas consecuencias presentan similitudes con las que se buscan abordar en esta investigación. En este estudio, se aplicó una metodología de aprendizaje supervisado para predecir accidentes en tiempo real. Tras preseleccionar las variables mediante bosques aleatorios, los autores compararon el desempeño de Máquinas de Soporte Vectorial (SVM) y Regresión Logística. El modelo considera datos capturados directamente desde los usuarios, mediante pórticos de autopistas y variables externas, tales como temperatura, presión atmosférica y precipitaciones. Esta combinación de fuentes de información permitió identificar las variables con mayor correlación con la ocurrencia de accidentes, y posteriormente, calibrar el modelo con datos históricos. Si bien los resultados obtenidos alcanzaron una precisión cercana al 70% en la estimación de la probabilidad de accidentabilidad, este valor representa un desempeño aceptable para una primera aproximación. Además, valida la aplicabilidad de este tipo de metodologías en contextos similares. Por lo tanto, se considera que los enfoques utilizados en este estudio pueden ser adaptados y aplicados al desarrollo de un modelo predictivo para actividades deportivas y recreativas al aire libre o en lugares remotos.

4 Desarrollo de la solución

Esta sección se enfoca en describir la metodología de la construcción, la implementación y la validación del modelo predictivo y de la investigación, siguiendo las definiciones y criterios establecidos en el capítulo anterior. Para asegurar el óptimo cumplimiento, el proceso se articula de acuerdo con una secuencia de etapas que considera la recopilación de información, el preprocesamiento de datos y la selección y optimización del modelo.

4.1 Metodología de trabajo

La metodología de trabajo se basa en el estándar CRISP-DM, el cual proporciona un marco estructurado para el desarrollo de proyectos de ciencia de datos y aprendizaje automático. Esta elección asegura que el proceso de trabajo sea un ciclo iterativo alineado con los objetivos estratégicos de la investigación. La elección de esta metodología sobre otros marcos de trabajo, como KDD o SEMMA, se justifica debido a su estructura cíclica y la fuerte orientación hacia los objetivos de la investigación. Comparado con KDD, que es un proceso más lineal y centrado en la extracción técnica de conocimiento, o SEMMA, que está muy enfocado en la implementación estadística de herramientas específicas, CRISP-DM prioriza la fase de entendimiento del negocio.

Esta investigación se desarrolla utilizando un enfoque cuantitativo, buscando no solo comprender el fenómeno, sino también desarrollar una solución tecnológica y validarla en etapas posteriores. La naturaleza de la investigación está enfocada en la predicción de la accidentabilidad, lo que implica recolección y análisis de datos numéricos para poder establecer relaciones entre sí y construir un modelo que pueda cumplir con el objetivo del estudio.

La investigación combina componentes empíricos y cuantitativos para la recopilación de datos, preprocesamiento y análisis, así como para el entrenamiento y la evaluación del modelo predictivo. En el diseño de la aplicación se considera toda la información recopilada, la información proveniente de los modelos evaluados, el resultado de la investigación y las mejores prácticas en cuanto al diseño de aplicaciones para potenciar la recolección de datos en etapas futuras y, por ende, poder incrementar la exactitud del modelo a medida que la aplicación va procesando una mayor cantidad de datos reales.

4.1.1 Herramientas y Tecnologías Utilizadas

El proyecto fue desarrollado íntegramente en el entorno de **Python 3.13**, aprovechando las bibliotecas líderes en ciencia de datos, indicadas en la **Tabla 1**.

Tabla 1: Herramientas y tecnologías

Herramienta/Librería	Propósito Específico	Versión
Python	Lenguaje de programación base.	3.13
Pandas	Estructuración y limpieza de datos	2.0.3
NumPy	Soporte para cálculos.	1.24.3
Scikit-learn (sklearn)	Framework principal de aprendizaje automático.	1.3.0
Imbalanced-learn	Balanceo avanzado de datasets.	0.11.0
Matplotlib / Seaborn	Generación de la capa visualización	3.7.2 / 0.12.2
Visual Studio Code	Entorno de desarrollo integrado para la edición y ejecución del código.	1.90.0

4.2 Comprensión del negocio

Esta fase se basa en los fundamentos establecidos en la introducción, los objetivos y el estado del arte de esta investigación, donde se identificó la creciente accidentabilidad en deportes de montaña y la falta de herramientas predictivas integradas. El propósito aquí es transformar los objetivos en requerimientos técnicos para el modelo de aprendizaje automático.

4.2.1 Alineación con los objetivos del proyecto

Como se indicó en la sección de Objetivos, el propósito principal es el diseño de un modelo capaz de determinar la probabilidad de accidentes. Bajo el marco CRISP-DM, esto se traduce en un objetivo de minería de datos de clasificación, donde el éxito será determinado por la capacidad del modelo para identificar correctamente situaciones de riesgo, utilizando las métricas de desempeño detalladas en el marco teórico.

4.2.2 Criterios de éxito

Para que el modelo pueda usarse dentro de la aplicación móvil propuesta, debe cumplir con los siguientes criterios:

- **Relevancia Predictiva**: El modelo debe identificar datos ambientales y de usuario que permitan una alerta temprana, en línea con los hallazgos en la literatura (Basso et al., 2018).
- **Utilidad de la Solución**: La predicción debe generar una entrada para la aplicación móvil y entregar una recomendación clara al usuario, cumpliendo así con el objetivo de desarrollo tecnológico planteado.

4.2.3 Requerimientos y Restricciones de Datos

El modelo depende de la integración de diversas fuentes como registros históricos de accidentes, ubicación geográfica de la actividad y variables meteorológicas. La principal restricción identificada es la falta de conexión y de información de accidentabilidad en zonas remotas, lo que exige un modelo robusto que funcione incluso con información limitada, parcial o asíncrona.

4.2.4 Definición del Problema de Minería de Datos

El problema se define como una tarea de aprendizaje supervisado. La finalidad en este caso es la seguridad de los individuos donde se requiere que el sistema procese un vector de entrada (clima, ubicación, experiencia) y entregue una probabilidad de riesgo que la aplicación móvil pueda comunicar de forma sencilla al usuario.

4.3 Compresión de los datos

En esta fase, el trabajo se centra en entender la estructura y el valor de la información recolectada para lograr un modelo sólido. La precisión de un sistema predictivo depende de la calidad de los datos; por ello, se realizó un análisis para identificar las variables que podrían tener el mayor impacto en la accidentabilidad al aire libre.

A pesar de la escasa disponibilidad de registros históricos bien documentados, se logró definir un conjunto de atributos que abarcan dimensiones climatológicas, geográficas, personales y temporales. Este trabajo permitió establecer los lineamientos para la adquisición de información, sirviendo como base técnica para el diseño de la base de datos y la posterior estrategia de procesamiento.

4.3.1 Recopilación de Datos

La viabilidad y la precisión de cualquier modelo predictivo dependen directamente de la calidad y la exactitud de los datos que lo alimentan. En esta etapa, se llevó a cabo un proceso detallado para identificar, acceder y recopilar las variables consideradas como relevantes para la predicción de la accidentabilidad en actividades al aire libre. Este proceso se enfrentó a grandes desafíos debido a la disponibilidad de datos y en particular de bases de datos de accidentes bien documentados.

Se estableció el conjunto de características y atributos necesarias para el modelo predictivo. En esta fase del desarrollo, el foco consistió en la especificación de las variables definidas (Climatológicas, Geográficas, de Actividad, Personales, Salud, Temporales y de Concurrencia).

Para la correcta recopilación de estas variables se incluyó la delimitación del tipo de variable, formatos, rangos y protocolos de adquisición, elementos que constituyeron el requisito base para el diseño de la arquitectura de la base de datos y la estrategia de minería de datos.

4.3.2 Fuentes de Datos

La obtención de datos para cada categoría de variable se realizó a partir de diversas fuentes, buscando maximizar el detalle de la información:

- **Datos de Accidentabilidad**: La recopilación de incidentes y accidentes específicos en actividades al aire libre representó un desafío complejo debido a la dispersión y la falta de registros oficiales centralizados públicamente accesible. Únicamente se encontraron resumen de datos de accidentabilidad que, para esta investigación y modelo, no eran suficientes, frente a la escasez de bases de datos públicas y detalladas de accidentes, la información sobre incidentes y accidentes se basó en las estadísticas y recopilación de accidentes de la revista escalando y el libro publicado por Rodrigo Fica (Fica, 2019). que abarca una gran cantidad de incidentes y accidentes de hace al menos 20 años. Este enfoque permitió construir un conjunto de datos representativo y reales de accidentabilidad para la validación del modelo.

- **Datos Climáticos**: Se emplearon servicios de API de OpenWeatherMap, Servicio Meteorológico Nacional de Chile (DMC), entre otras bases de datos para obtener datos históricos y en tiempo real de temperatura, precipitación, humedad, velocidad y dirección del viento, y presión atmosférica para las ubicaciones de interés donde ocurrieron los incidentes y accidentes de la base de datos de accidentabilidad.

- **Datos Geográficos**: La información sobre datos geográficas donde ocurrieron los incidentes y accidentes como altitud se obtuvo a través de Google Maps y datos provenientes de GPS. Las características del terreno y la vegetación se infirieron de bases de datos GIS abiertas como OpenStreetMap, Andeshandbook wikiloc y otras plataformas donde se describen rutas de montaña, escalada, trekking, entre otros.

- **Datos de Actividad, Personales, Temporales, de Concurrencia y Salud**: Para el entrenamiento inicial, se emplearon datos provenientes del registro de accidentabilidad (Fica, 2013, 2019), agregados y en algunos casos inducidos en base a los antecedentes de accidentabilidad. No obstante, el diseño del modelo prevé

que la futura aplicación móvil será la principal fuente de recopilación continua de estas variables directamente de los usuarios.

4.3.3 Proceso y Desafíos de la Recopilación

La recopilación de los datos se desarrolló en dos diferentes etapas, según el tipo de dato requerido. La primera fue la solicitud registro de accidentabilidad a entidades o personas que hubiesen documentado este tipo de información. No fue posible obtener bases de datos de accidentabilidad por parte de las autoridades competentes en el ámbito, como el FEDME, debido a la política de protección de datos, ni de FEACH, debido a que no cuentan con un registro de accidentabilidad. Posteriormente se hizo el requerimiento de los registros al autor de los libros de accidentabilidad Rodrigo Fica, que finalmente declinó debido a que el trabajo de recopilación fue realizado con un objetivo diferente. Ante estas situaciones, se determinó que la recopilación de datos de accidentabilidad se desarrollara de forma manual, utilizando como fuente principal los libros del autor previamente mencionados (Fica, 2013, 2019).

La segunda etapa de recolección de datos se desarrolló en base al registro de accidentabilidad, en la cual se consideraron principalmente las variables geográficas, temporales y climáticas. Para esta recopilación se utilizó una amplia cantidad de metodologías como *scripts* automatizados desarrollados en Python, para obtener los datos climáticos, se utilizó Google Earth y Google Maps, para georreferenciar el lugar del accidente y un procesamiento por base de datos para obtener más información sobre las variables temporales.

Uno de los principales desafíos fue la ausencia de un conjunto de datos unificado y completo que contuviera todas las variables identificadas incluyendo además las variables temporales, climáticas y geográficas requeridas. Esto implicó un esfuerzo significativo en la integración de múltiples fuentes y en el modelado de datos faltantes. La privacidad y la ética en la recopilación de datos personales fueron consideradas primordiales, asegurando que cualquier información sensible o personal no fuese incluida en el modelo.

4.3.4 Estrategia de Recolección

La estrategia de recolección se desarrolló con un enfoque híbrido que combina una fuente primaria interna (el registro histórico de eventos de accidentabilidad) con la integración de fuentes secundarias externas (datos temporales, geográficos y climáticos). Este enfoque es crucial para construir un conjunto de características lo suficientemente robusto para el entrenamiento del modelo predictivo.

El proceso de recopilación de información se llevó a cabo en las siguientes etapas:

a) **Recopilación de Datos Primarios**

- **Fuente Principal**: A partir de la información disponible y de la documentación previamente señalada, se seleccionaron todas las variables útiles y se estructuró una base de datos, para generar un archivo centralizado de Antecedentes de Accidentabilidad. Dentro de este conjunto de datos se definió la variable objetivo que se denominó como “Consecuencia” en base al resultado del evento identificado y se categorizo en 4 niveles, 1: leve, 2: media, 3: Alta y 4: Fatal. Además, se incluyeron las antecedentes personales y de actividad (Edad, experiencia, tipo de actividad, experiencia, modalidad de participación, localidad, etc.).

  El registro de datos comprende información obtenida desde diciembre de 2022 hasta el diciembre de 2023, donde se identificaron 255 registros, de los cuales cada registro identifica un participante; sin embargo, el evento puede incluir más de un individuo. La base se estructuró con un correlativo de evento y un correlativo de participante, la cual a su vez se gestionó con un correlativo general que fue la combinación del ID de la actividad agregado al ID del participante. Finalmente, es importante destacar que se identificó un marcado desbalance entre las categorías, siendo la clase "Fatal (4)" la que concentra la mayor cantidad de registros.

b) **Enriquecimiento con Fuentes Secundarias:**

- Las variables climatológicas, geográficas y temporales se obtienen de servicios en línea de acceso libre y APIs.
- Se emplearon *scripts* para consultar, a gran escala, servicios meteorológicos y servicios geográficos.
- Cada registro de accidente en la base de datos principal se enriquece mediante una coincidencia de clave compuesta, utilizando la fecha, la hora y las coordenadas geográficas del evento, asegurando que las condiciones ambientales añadidas sean precisas para el momento y lugar exacto de la ocurrencia del evento.

c) **Consolidación:**

- La estrategia de recolección finaliza con la integración de todos los atributos en una única base de datos estructurada, con las características listas para la fase de limpieza y preprocesamiento. Esta base consolidada es el insumo directo para el desarrollo de los algoritmos de Aprendizaje Automático posterior a la depuración.

Esta metodología garantiza que la base de datos final contenga tanto la consecuencia del evento como el conjunto completo de factores endógenos y exógenos que influyeron en su ocurrencia.

Para visualizar la distribución espacial de los eventos recopilados, se elaboró un Mapa de accidentes georreferenciado en la **Figura 3**. En esta representación, cada punto geográfico ha sido codificado mediante una escala de colores que corresponde al nivel de consecuencia, esto permite identificar visualmente patrones de riesgos o zonas críticas donde la severidad de los accidentes es mayor, facilitando el análisis de la variable geográfica.

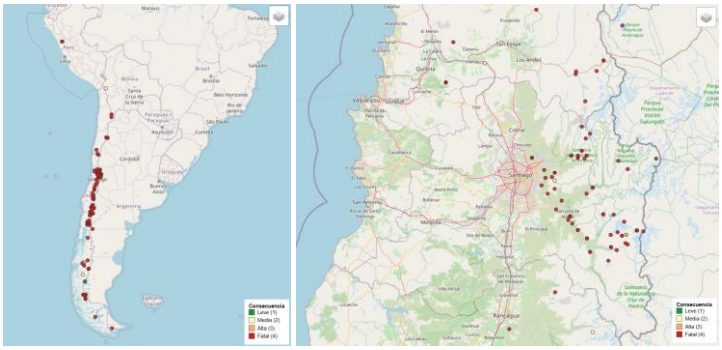

Figura 3. *Mapa de accidentes georreferenciado con color por consecuencia*
Nota: Elaboración propia.

4.4 Preprocesamiento de los de Datos

4.4.1 Limpieza de Datos

La etapa de limpieza es fundamental para asegurar la calidad y fiabilidad de los datos y que el modelo sea entrenado de forma adecuada. El proceso de limpieza se enfoca en tres áreas principales para ajustar o corregir las deficiencias generadas durante la recopilación e integración de fuentes, a continuación, se describen la limpieza de datos realizada.

a) **Tratamiento de Valores Faltantes:** Se realiza una revisión detallada de los valores nulos o incompletos. Para las características continuas o valores numéricos continuos (como datos climatológicos o altitud), se aplica un método de deducción basado en los registros históricos o promedios. En el caso de variables categóricas que no fue posible completar, son ajustadas con la moda o se le asigna una categoría de faltante o no definido. Esta metodología es muy importante para cumplir con el objetivo de mantener la tasa de errores y valores nulos inferior al 10% establecida en la metodología.

b) **Detección y Gestión de Atípicos:** Se identifican los valores atípicos que podrían distorsionar el proceso de entrenamiento del modelo. Para las variables numéricas (ej. Edad, Duración de la Actividad), se utilizan métodos estadísticos para detectar observaciones que se encuentran fuera de los límites aceptables por ejemplo percentiles. Estos valores se evalúan individualmente: si fueron generados por errores de registro, se elimina o se ajusta; si representan condiciones extremas reales, se procede a ajustar su valor para que no influyan exageradamente en el algoritmo, pero manteniendo la información del resto de las variables del registro.

c) **Estandarización y Corrección de Inconsistencias:** Se realiza la estandarización de las variables, asegurando que categorías idénticas tengan una representación única y no múltiples valores similares. Además, se validan los formatos de dato, verificando que las fechas, horas y coordenadas geográficas se ajusten a un estándar uniforme, corrigiendo cualquier error de codificación como puntuación, inconsistencia de mayúsculas/minúsculas o cualquier otro error conceptual o de tipografía que pueda impedir la correcta lectura o interpretación única de la categoría en la siguiente fase de preprocesamiento.

4.4.2 Transformación de Datos

La transformación de datos es el proceso metodológico que convierte el conjunto de datos limpio en un formato adecuado para el entrenamiento del modelo. Esta etapa se centra en la aplicación de técnicas de ingeniería de características y escalado.

4.4.2.1 Ingeniería de Características

Esta técnica se utiliza para crear nuevas variables a partir de las características existentes para poder potencial el modelo, buscando maximizar la capacidad de discriminación y exactitud del modelo, así como probar diferentes variables y ver su impacto en la predicción de accidentabilidad. A continuación, se describen todas las transformaciones y procesamiento que se realizaron a las categorías y variables de la base de datos consolidada y limpia de accidentabilidad:

- **Transformación de la Variable Temporal:** Las variables de fecha y hora se descomponen o transforman en características más significativas para el modelo (ej: Se incluye día de la semana, Estación del Año, mes). Adicionalmente, se generan una característica binaria para identificar si el evento ocurrió durante un fin de semana largo, la cual está directamente asociada al factor concurrencia.

- **Conversión de Variables Categóricas:** Las variables nominales y ordinales se convierten a un formato numérico. Para variables nominales (ej. país, día de la semana, estación), se aplica la técnica de codificación "*One-Hot Encoding*", creando una columna binaria (0 y 1) para cada categoría única, asignado el valor 1 cuando la variable nominal esta asignada y 0 cuando no está asignada al registro. Para variables ordinales (ej. Experiencia, Complejidad, Calidad del equipamiento), se utiliza la Codificación Ordinal, asignando un valor numérico secuencial (ej. Baja=1, Media=2, Alta=3) para preservar la relación de orden en cada uno de los valores, esto finalmente representa una relación equitativa de acuerdo con el valor asignado no existiendo asimetrías entre los grados definido para cada una de las variables numéricas asignadas.

- **Generación de Variables Geográficas:** Esta fase se centra en extraer el valor predictivo de las variables de accidentabilidad, ubicación, temporales y clima, identificando las limitaciones en la disponibilidad de datos de fuentes secundarias. En particular en esta etapa, de la ubicación solo se pudieron determinar e inferir nuevas variable como la Altura, el hemisferio, el país, la localidad y la región, pero en una iteración posterior, es posible se pueden determinar valores muchos más específico, como inclinación, tipo de terreno, vegetación, disponibilidad de agua u otras variables que de acuerdo a las primeras aproximaciones con eventos de accidentabilidad pueden aportar una mayor certeza a una predicción más precisa y aportar para otros análisis adicionales que pueden apoyar a los usuarios en la toma de decisiones.

- **Sensación Térmica**: Se incluye una variable adicional definida como Sensación Térmica la cual es una variable inferida, calculada para reflejar un impacto más realista del estrés ambiental sobre el cuerpo, lo

cual es un predictor más robusto de riesgos como la hipotermia o agotamiento por exposición a factores climáticos más que solo considerar la temperatura absoluta. Para el cálculo, se crea un factor de ajuste de acuerdo con las variables ambientales, donde la Temperatura promedio se ajusta basado en la intensidad del viento y las precipitaciones, utilizando la lógica explicada en la **Tabla 2**.

Tabla 2: *Matriz de Factores de Ajuste para el Cálculo de la Sensación Térmica Inferida*

Factor de Ajuste	Ajuste aplicado
Precipitaciones (Baja/Media/Alta)	1∘C/2∘C/3∘C
Velocidad del Viento (Baja/Media/Alta)	1∘C/3∘C/5∘C

Para el caso de velocidad del viento existen tablas específicas para determinar la sensación térmica como se observa en la **Tabla 3**, desarrollada por la agencia estatal de meteorología española, que no se considera, ya que no se cuenta con la velocidad del viento exacta para el modelo, pero si se puede utilizar para el cálculo de la sensación térmica para iteraciones posteriores.

Tabla 3. *Tabla de sensación térmica desarrollada por agencia estatal de meteorología española*

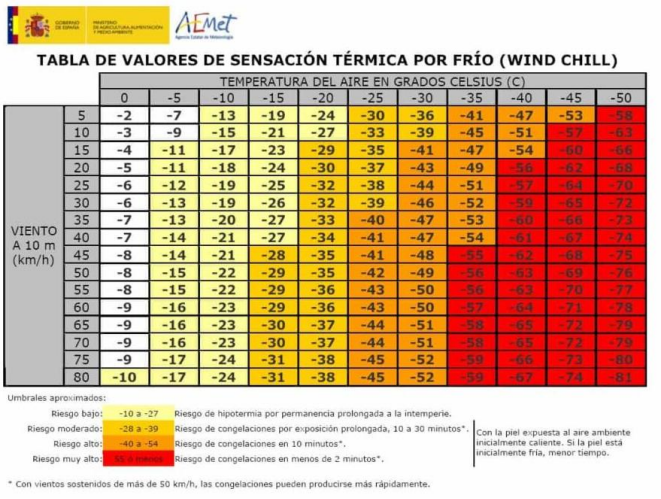

TABLA DE VALORES DE SENSACIÓN TÉRMICA POR FRÍO (WIND CHILL)

VIENTO A 10 m (km/h)	TEMPERATURA DEL AIRE EN GRADOS CELSIUS (C)										
	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50
5	-2	-7	-13	-19	-24	-30	-36	-41	-47	-53	-58
10	-3	-9	-15	-21	-27	-33	-39	-45	-51	-57	-63
15	-4	-11	-17	-23	-29	-35	-41	-47	-54	-60	-66
20	-5	-11	-18	-24	-30	-37	-43	-49	-56	-62	-68
25	-6	-12	-19	-25	-32	-38	-44	-51	-57	-64	-70
30	-6	-13	-19	-26	-32	-39	-46	-52	-59	-65	-72
35	-7	-13	-20	-27	-33	-40	-47	-53	-60	-66	-73
40	-7	-14	-21	-27	-34	-41	-47	-54	-61	-67	-74
45	-8	-14	-21	-28	-35	-41	-48	-55	-62	-68	-75
50	-8	-15	-22	-29	-35	-42	-49	-56	-63	-69	-76
55	-8	-15	-22	-29	-36	-43	-50	-56	-63	-70	-77
60	-9	-16	-23	-29	-36	-43	-50	-57	-64	-71	-78
65	-9	-16	-23	-30	-37	-44	-51	-58	-65	-72	-79
70	-9	-16	-23	-30	-37	-44	-51	-58	-65	-72	-79
75	-9	-17	-24	-31	-38	-45	-52	-59	-66	-73	-80
80	-10	-17	-24	-31	-38	-45	-52	-59	-67	-74	-81

Umbrales aproximados:

Riesgo bajo:	-10 a -27	Riesgo de hipotermia por permanencia prolongada a la intemperie.
Riesgo moderado:	-28 a -39	Riesgo de congelaciones por exposición prolongada, 10 a 30 minutos*.
Riesgo alto:	-40 a -54	Riesgo de congelaciones en 10 minutos*.
Riesgo muy alto:	55 o menos	Riesgo de congelaciones en menos de 2 minutos*.

Con la piel expuesta al aire ambiente inicialmente caliente. Si la piel está inicialmente fría, menor tiempo.

* Con vientos sostenidos de más de 50 km/h, las congelaciones pueden producirse más rápidamente.

- **Hemisferio y País**: Las variables geográficas y de demarcación se incorporan al modelo con métodos de codificación específicos de acuerdo con el tipo de dato. La variable hemisferio se incluye como una variable binaria (Codificación Binaria) para identificar la diferencia estacional por hemisferio. La variable País se transforma utilizando "*One-Hot Encoding*" para capturar el efecto de características geográficas, regulatorias u otras inherentes del entorno donde ocurrió el evento y particularmente en el país en este caso. Esta transformación es muy importante, debido a que la mayor parte del conjunto de datos proviene de Chile, lo que permitirá al modelo aprender las particularidades del riesgo específicas en comparación con otras.

4.4.2.2 Desafíos y Delimitación en la Ingeniería de Características

Se identificaron desafíos en la disponibilidad de datos que delimitan el alcance de esta fase en la iteración actual:

- **Variables de Ubicación:** La variable Localidad se descarta para la predicción debido a que existen demasiados registros con nombres únicos, lo que finalmente de acuerdo con la investigación, no aporta valor al modelo debido a las dispersiones de valores, y si se realizara la transformación de esta variable generaría un número excesivo de columnas binarias con bajo poder predictivo individual.

- **Coordenadas geográficas:** La ubicación geográfica en base a Latitud y Longitud se utilizaron solo para el proceso de enriquecimiento del modelo como la obtención de altura, hemisferio, localidad, regios y clima histórico, pero no se incluirán directamente como características predictivas en el modelo final, debido a su complejidad para poder incluirlo directamente en un modelo predictivo en una etapa inicial.

- **Zonas remotas:** Por definición una zona remota se considera un lugar que en tiempo o distancia se encuentra alejado de ciertos servicios, en base a esto, la cercanía a centros asistenciales o lugares de rescate, o eventualmente a centros urbanos si bien es una variable crucial para predecir la gravedad o el aislamiento, se considera fuera del alcance de esta tesis debido a la falta de una base de datos pública y estructurada de estas ubicaciones que puedan incluirse en el modelo.

- **Factores de Terreno y Tiempo:** El cálculo de la pendiente promedio del terreno, tipo de terreno, vegetación y las horas restantes de luz al momento del evento se identifican, como potenciales variables de alto valor predictivo para el riesgo de accidentes. No obstante, su desarrollo requiere la integración de modelos de elevación, mapas, variables geográficas, entre otras para obtener el relieve exacto y el tipo de terreno, y base de datos temporales para determinar la hora precisa del ocaso en función de la fecha y la ubicación. Debido a que la implementación rigurosa de estas fuentes de datos sobrepasa el alcance de la iteración actual de la tesis, su generación se considera una extensión crítica y prioritaria para la próxima iteración del modelo predictivo.

4.4.3 Integración de Datos

El objetivo de la integración de datos es consolidar las diversas fuentes de información transformadas y enriquecidas en un único conjunto de datos monolítico. Este conjunto final, también denominado matriz de características, es la entrada directa requerida por los algoritmos de aprendizaje automático.

El proceso de integración se realizó mediante una operación de unión relacional, utilizando el campo ID Evento como la clave primaria para cada registro. Esto asegura que todas las variables temporales, geográficas y climáticas inferidas y transformadas correspondan con el evento de accidentalidad original.

Las cuatro fuentes de información transformadas que se integraron son:

- **Maestro de Accidentabilidad**: Proporciona las variables objetivo (Consecuencia) y las características del participante y actividad (Edad, Experiencia, Género, Tipo de actividad, etc.).

- **Información de Fechas**: Aporta las características temporales codificadas con *One-Hot Encoding* (Día de la semana, Estación del Año, finde semana largo) así como otras variables numéricas como (día, mes, año).

- **Información Geográfica**: Incluye la Altura del evento y las variables geográficas codificadas (País, Tipo de terreno, localidad, región).

- **Información Meteorológica**: Contribuye con las variables climáticas (Temperatura, Viento, Precipitación) y la Sensación Térmica inferida.

El resultado de esta integración es una tabla final y homogénea, donde cada fila representa una observación (un participante en un evento) y cada columna representa una característica predictiva lista para la fase de escalado y el posterior entrenamiento del modelo.

4.5 Modelado

En esta fase, el objetivo principal es la construcción y el ajuste de los modelos predictivos capaces de identificar patrones de riesgo en las actividades al aire libre. El proceso se centra en transformar los datos preprocesados, evaluando distintos algoritmos para determinar cuál ofrece la mejor precisión.

4.5.1 Selección del Algoritmo de Aprendizaje Automático

Para abordar el problema definido en la hipótesis, la selección del algoritmo de clasificación se presenta como una parte fundamental para lograr el resultado deseado. A partir de la base de datos obtenida tras las etapas de recopilación y procesamiento, se evaluaron diversas alternativas de forma conceptual y empírica para determinar el modelo ideal.

En este proceso, se descartó el uso de Redes Neuronales y XGBoost, debido a que el tamaño de la base de datos es muy reducido y la naturaleza tabular de los datos elevaban el riesgo de sobreajuste en ambos casos. Igualmente, se desestimó el algoritmo de K-Vecinos Más Próximos, dado que su rendimiento disminuye significativamente en espacios de alta dimensionalidad y ante la presencia de variables categóricas complejas como las de este estudio.

En la fase de evaluación, se compararon tres modelos para evaluar su respuesta con los datos recopilados. Los resultados obtenidos, que se presentan en la **Tabla 4**, permitieron comparar la Regresión Logística, Máquinas de Soporte Vectorial y Bosques Aleatorios; se observa que la Regresión Logística y el SVM presentaron limitaciones para reflejar la complejidad de los accidentes de montaña, lo que se reflejó en niveles inferiores de Exactitud. Por el contrario, el modelo de Bosques Aleatorios presentó los mejores resultados con un 90,9% de Exactitud y un AUC de 0,78, demostrando que es la mejor opción para el desarrollo de este modelo.

Tabla 4. *Tabla comparativa de modelos de clasificación*

Modelo	Exactitud	Sensibilidad	Área bajo la curva
Regresión Logística	0,883	0,246	0,695
Máquina de Soporte Vectorial	0,896	0,250	0,306
Bosques Aleatorios	0,909	0,496	0,780

4.5.2 División y transformación de los datos

La base de datos inicial resultante contó con 255 registros y 57 columnas. Para asegurar la integridad del modelado, se inició con el proceso de limpieza y escalado numérico, estandarizando las variables continuas como Altura, Edad, Duración de la actividad y Sensación Térmica, que se presentan en diferentes formatos, por esto, se implementaron funciones de ajuste dentro del modelado. Posteriormente, se aplicó *StandardScaler* para normalizar los datos a una media de cero ($\mu = 0$) y desviación estándar de uno ($\sigma = 1$) evitando que variables con valores muy altos (como la Altura) sesguen el cálculo de pesos en el modelo.

En cuanto a la codificación de variables, las dimensiones nominales como Tipo de actividad y Modalidad de participación, fueron transformadas en un formato binario mediante *One-Hot Encoding*. Esta conversión resultó en una matriz de diseño de 16 características predictivas adicionales, consolidando una matriz de diseño final de 60 columnas.

Debido a la distribución de la muestra, se realizó un remapeo de la variable objetivo para optimizar la estratificación. Ante la baja frecuencia de la categoría "Leve (1)", esta se agrupó con la clase "Media (2)", resultando en un modelo de tres niveles de severidad: Media, Alta y Fatal. Finalmente, el conjunto de datos se dividió de forma estratificada en proporciones de 75% para entrenamiento (191 registros) y 25% para prueba (64 registros), garantizando que la representatividad de cada clase se mantuviera proporcional en ambos subconjuntos.

4.5.3 Entrenamiento original y Balanceo del Modelo

El entrenamiento del modelo final incorporó una técnica avanzada de balanceo para mitigar el sesgo debido a la alta cantidad de casos con la etiqueta "Fatal (4)" de la base de datos. Se aplicó la técnica SMOTE **(***Synthetic Minority Over-sampling Technique***)** solo al conjunto de entrenamiento (X_train, y_train). Se utilizó una técnica de balanceo de los estratos "Media(2)" y "Alta(3)" debido a que la cantidad de datos de estas clases de la variable objetivo tiene muy pocos registros, y con SMOTE se pueden crear muestras sintéticas de las clases minoritarias extrapolando entre vecinos cercanos. Esto obligó al modelo a enfocarse más en los patrones de estas clases poco representadas en la base

de datos, debido a que generalmente la información de accidentabilidad con consecuencia “Baja(1)” o “Media(2)”, no se registra.

Debido al tamaño extremadamente pequeño de las clases minoritarias, el parámetro por defecto de SMOTE (n_neighbors=6) fue ajustado a k_neighbors=2 para evitar errores y permitir el sobre muestreo.

De acuerdo con uno de los objetivos de este estudio, se busca desarrollar al menos dos modelos para poder predecir la accidentabilidad, el primer modelo considera los datos sin el balanceo y el segundo modelo considera los datos balanceados por lo que, el clasificador de bosques aleatorios es configurado con los datos originales y los datos balanceados para ponderar las clases minoritarias. En primera instancia fue entrenado utilizando los datos originales (X_train y Y_train) y posteriormente con los datos sobremuestreados (X_train_res y Y_train_res).

4.6 Evaluación del Modelo

En esta sección se analiza el desempeño de los modelos entrenados. El objetivo es validar la capacidad predictiva del sistema, comparando los resultados del Bosque Aleatorio original frente a la variante procesada con técnicas de balanceo.

4.6.1 Métricas de Desempeño

Para cuantificar los resultados, se seleccionaron indicadores estándar de clasificación que permiten una visión completa del rendimiento. Se analizan la Exactitud, Sensibilidad, Precisión, y el área bajo la curva, para determinar el rendimiento de los modelos en cada categoría de riesgo.

4.6.1.1 Modelo original

Para iniciar el proceso de evaluación, en la **Tabla 5** se muestra las métricas globales de desempeño, donde se observa una exactitud de 0.86. A su vez, en la **Tabla 6** el valor general de las métricas del modelo oculta un fenómeno de sesgo por mayoría. Con el análisis detallado por clase, el modelo muestra dificultades para determinar los incidentes de menor gravedad a consecuencia al bajo volumen de muestras en cada categoría, en comparación con los valores obtenidos en la categoría “Fatal (4)”.

Tabla 5: *Métricas del modelo original*

Métrica Global	Valor
Accuracy (Exactitud)	0.86
F1-Score Ponderado General	0.87
Promedio Macro (F1)	0.72

Tabla 6: *Análisis Detallado por Clase del modelo original*

Clase	Precisión	Recall	F1-Score	Nº Casos
Media (2)	0.25	0.25	0.25	4
Alta (3)	0.30	1.00	0.50	2
Fatal (4)	0.96	0.90	0.93	58

Como se analizó previamente en la fase de compresión de los datos la **Figura 4** presenta la distribución de la variable de entrenamiento por consecuencia. Esta distribución explica los errores de clasificación reflejados en la **Figura 5**, correspondiente a la matriz de confusión y en la capacidad de discriminación por nivel de riesgo expuesta en la **Figura 6** mediante la curva ROC.

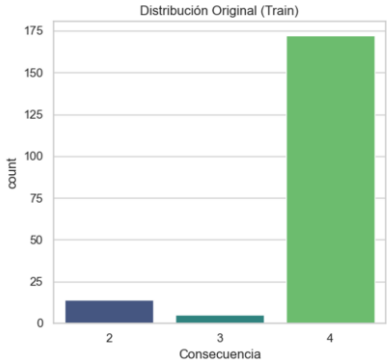

Figura 4. *Distribución de variable de entrenamiento consecuencia por tipo*

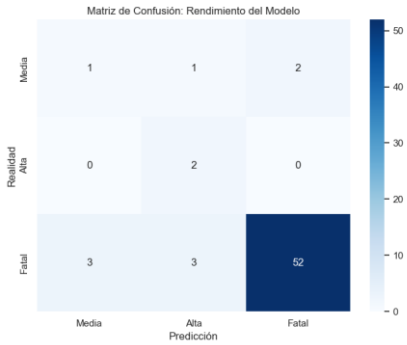

Figura 5. *Matriz de confusión del modelo original*

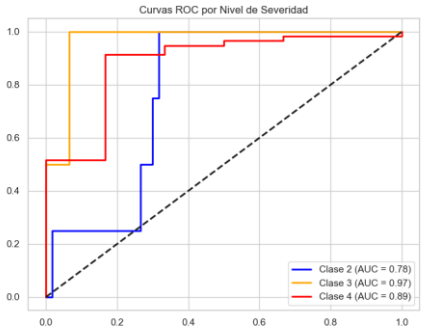

Figura 6. *Curva ROC por categoría*

Resultados:

De acuerdo con los resultados obtenidos del procesamiento con el modelo original se deduce lo siguiente:

1. **Predicción de fatalidades:** Tomando como referencia la **Tabla 6**, el modelo muestra su mayor robustez en la clase Fatal (4), con un F1-Score de 93%. La precisión del 96% es un indicador crítico de confiabilidad: cuando el modelo indica una fatalidad, existe una certeza muy alta de que las condiciones conducen a ese desenlace.
2. **Sensibilidad de consecuencia alta:** Enfocándose en la misma **Tabla 6**, se observa que la clase "Alta (3)" se obtuvo una sensibilidad de 100%. Este hallazgo es fundamental desde una perspectiva preventiva, ya que el modelo identificó la totalidad de los accidentes graves en el conjunto de prueba. No obstante, dado el bajo soporte ($n = 2$), este resultado debe interpretarse con cautela, ya que la precisión del 33% indica una tendencia a sobreestimar riesgos moderados como graves.
3. **Precisión y Sensibilidad de consecuencia media:** Para la categoría consolidada (Leve/Media), se observa un rendimiento limitado con una precisión y sensibilidad del 25%. Esto confirma que el modelo, en su estado

original, tiende a desplazar las predicciones de riesgo menor hacia categorías de mayor gravedad, actuando bajo un principio de precaución o sesgo conservador.

4.6.1.2 Modelo Balanceado

Tras identificar un sesgo debido a la clase "Fatal (4)" en el modelo base, se implementó una estrategia de balanceo mediante la técnica SMOTE. Los resultados obtenidos se presentan en la **Tabla 7**, mientras que el rendimiento por clase se muestra en la **Tabla 8**.

Tabla 7: *Métricas del modelo balanceado*

Métrica Global	Valor
Accuracy (Exactitud)	0.84
F1-Score Ponderado General	0.86
Promedio Macro (F1)	0.55

Tabla 8: *Análisis Detallado por Clase del modelo balanceado*

Clase	Precisión	Recall	F1-Score	Nº Casos
Media (2)	0.25	0.25	0.25	4
Alta (3)	0.20	0.50	0.29	2
Fatal (4)	0.95	0.90	0.92	58

El balanceo de los datos se observa en la **Figura 7**, que compara la distribución de la variable consecuencia entre el modelo original y el balanceado. Los efectos del cambio en la capacidad predictiva se analizan utilizando la matriz de confusión presentada en la **Figura 8** y la curva ROC por categoría mostrada en la **Figura 9.**

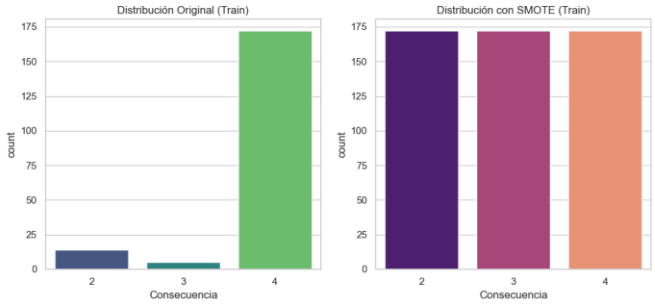

Figura 7. *Distribución de variable de entrenamiento consecuencia por tipo en los modelos original y balanceado mediante SMOTE*

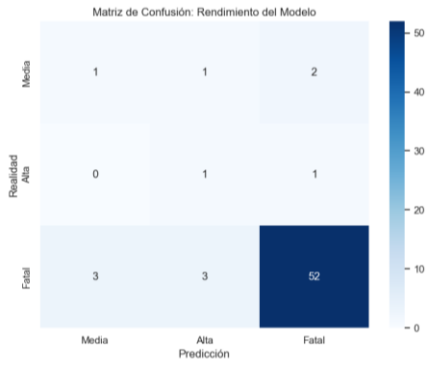

Figura 8. *Matriz de confusión del modelo balanceado*

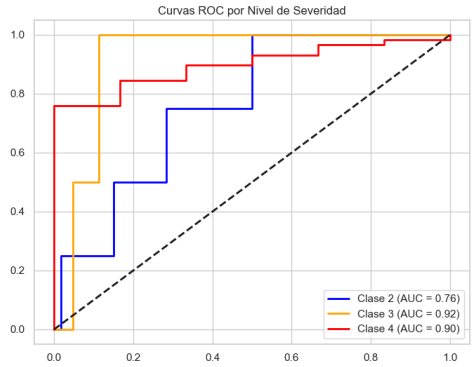

Figura 9. *Curva ROC por categoría*

Resultados:

De acuerdo con los resultados obtenidos del procesamiento con el modelo balanceado, se deduce lo siguiente:

1. **Estabilidad en la predicción crítica:** A pesar de forzar el aprendizaje en clases minoritarias, el modelo mantuvo una alta solidez para la categoría "Fatal (4)", con un F1-Score de 0.92. La precisión se mantuvo en un 95%, lo que implica que el balanceo no redujo la capacidad del modelo para identificar riesgos de muerte.
2. **Disminución en la detección de consecuencias altas:** Se observa un cambio significativo en la sensibilidad del modelo. La precisión bajó a un 20% y la sensibilidad de la clase "Alta (3)" se situó en 50%. Esto significa que el modelo ahora menos capaz de clasificar correctamente patrones de severidad alta.
3. **Análisis comparado con el modelo original:** La disminución de la exactitud global (de 0.86 a 0.84) es la consecuencia de lograr un modelo más equilibrado. En el contexto de la aplicación propuesta, no existen grandes cambios entre ambos modelos que sea relevante para privilegiar uno sobre otro.

4.6.2 Nivel de cumplimiento de la hipótesis

Los resultados obtenidos confirman el cumplimiento de la hipótesis, alcanzando un modelo predictivo con una exactitud del global del 86% mediante el algoritmo de bosques aleatorios, superando el 75% propuesto inicialmente. Asimismo, se identificaron más de 10 variables con incidencia directa en la accidentabilidad en actividades al aire libre. Basándose en estos hallazgos, las siguientes secciones describen el despliegue y la arquitectura de la aplicación móvil que integra el modelo desarrollado.

4.7 Despliegue

La fase final de la investigación contempla la conceptualización de **Wise Trek**, una aplicación móvil diseñada para alimentar el modelo de aprendizaje automático. Su objetivo es ofrecer evaluaciones de riesgo de accidentabilidad en tiempo real a usuarios que planifican y ejecutan actividades al aire libre.

4.7.1 Arquitectura de la Solución

El diseño propuesto sigue un modelo de **Arquitectura Cliente-Servidor** para asegurar la escalabilidad y la baja latencia en la predicción del riesgo:

1. **Cliente (Aplicación Móvil - Frontend):** Es responsable de la recolección de datos del usuario, la geolocalización y la visualización de la alerta de riesgo.
2. **Servidor (Backend):** Servicio web que gestiona la lógica de la aplicación y el preprocesamiento de datos (escalado y *One-Hot Encoding*).

3. **Motor de Predicción (Modelo ML):** El modelo de bosques aleatorios entrenado es cargado por el *backend* para generar predicciones de severidad: Baja (1), Media (2), Alta (3) y Fatal (4) utilizando los datos de los usuarios.

4.7.2 Flujo de Usuario y Captura de Datos

El diseño de la experiencia de usuario en Wise Trek (Aplicación móvil) se ha estructurado para recolectar de manera eficiente las variables críticas que alimentan el modelo predictivo. La captura de información se divide en tres capas para asegurar que el modelo cuente con las variables necesarias para una estimación:

1. **Perfil del Usuario (Variables Estáticas):** Datos ingresados al configurar la aplicación, incluyendo información demográfica (edad, género), nivel de experiencia, estado de salud y equipamiento disponible.
2. **Planificación de Actividad (Variables Manuales):** Información específica de cada salida que el usuario define antes de iniciar, como el tipo de deporte, duración estimada, modalidad (individual o grupal) y la ruta seleccionada.
3. **Entorno y Sensores (Variables Automatizadas):** Datos obtenidos en tiempo real mediante APIs y sensores del dispositivo para garantizar precisión, tales como geolocalización (altura, terreno) y condiciones meteorológicas (temperatura, viento y precipitaciones).

4.7.3 Procesamiento de Predicción

Una vez consolidadas las variables, el flujo sigue los siguientes pasos:

1. **Envío del Payload:** Al presionar el botón de genera actividad, el frontend envía un vector de datos al backend.
2. **Transformación en Tiempo Real:** El servidor aplica el *StandardScaler* y la codificación *One-Hot* para que los datos coincidan con el formato de entrenamiento del modelo.
3. **Ejecución del Motor ML:** El modelo procesa la entrada y devuelve una probabilidad de severidad, basado en las categorías definidas.

4.7.4 Comunicación del Nivel de Riesgo

El flujo culmina con la visualización del resultado, diseñada bajo el principio de comunicación instantánea:

1. **Interpretación Visual**: El resultado se entrega como una categoría de riesgo (Baja, Media, Alta, Fatal) asociada a la paleta de colores del manual de marca.
2. **Acción Preventiva**: Si el modelo predice un riesgo Fatal (4) o Alto (3), el flujo de la aplicación se bloquea para mostrar una pantalla de alerta roja con recomendaciones de seguridad basado en la actividad que se va a ejecutar.
3. **Retroalimentación Educativa**: Se muestra al usuario cuáles fueron los factores más importantes para predecir el nivel de riesgo, permitiendo al usuario tener herramientas para poder gestionar el riesgo de la actividad.

5 Conclusiones y Trabajo Futuro

La implementación de un Clasificador de Bosque Aleatorio como modelo de aprendizaje automático permitió confirmar parcialmente la hipótesis de que las variables identificadas tienen un alto poder predictivo en la severidad de los accidentes, aunque con limitaciones importantes:

1. **Alta Precisión en la Categoría Crítica:** El modelo demostró ser altamente efectivo para la predicción de fatalidades, con una Precisión del 96% y una Puntuación F1 de 0.98 para la clase "Fatal (4)". Este hallazgo es el más valioso del estudio, confirmando que variables como la Altitud, la temperatura, la duración de la actividad, la edad y la Sensación Térmica son determinantes para la predicción. La Puntuación F1 Ponderada de 0.95 valida la calidad general del ajuste.

2. **Dificultad en la Clasificación de Categorías Minoritarias:** La limitación más significativa es la incapacidad del modelo para categorizar la clase "Media (2)" que a su vez considera la clase “Baja (1)” (Puntuación F1 de 0.25). A pesar de emplear técnicas de balanceo y la ponderación de clases, el desequilibrio provocó un sesgo, donde el modelo tendió a asignar esos pocos casos a la clase mayoritaria.

3. **Validación del Procesamiento de Datos:** El flujo de preprocesamiento que incluyó el cálculo de la sensación térmica, el escalado de variables y la codificación para gestionar las categorías resultó ser un efectivo para el modelamiento. Este proceso es fundamental para alimentar el indicador de "Riesgo Calculado" que el usuario visualizará en la interfaz de la aplicación.

En resumen, El aprendizaje automático permite predecir la categoría más alta (Fatal), proporcionando un buen antecedente para los usuarios, Sin embargo, aún no logra la granularidad necesaria para diferenciar con consistencia entre las categorías media y alta e incluso baja debido a la falta de datos.

Para superar las limitaciones encontradas y llevar el trabajo a un nivel de implementación completo, se proponen las siguientes líneas de trabajo futuro:

1. **Ampliación de la base de datos:** La prioridad más importante es obtener más registros de excursiones o salidas sin incidentes, incidentes de severidad “Baja(1)” “Media(2)” y “Alta(3)” utilizando la misma estructura ya diseñada. Una base de datos más amplia y equilibrada es la única vía para que cualquier algoritmo pueda aprender los patrones distintivos de estas clases minoritarias.

2. **Evaluación de Algoritmos Alternativos:** Explorar el uso de modelos de *boosting* como *XGBoost* o *LightGBM*, los cuales suelen superar a bosques aleatorios en el manejo de desequilibrios y estructuras complejas de datos, y podrían ofrecer una mejor diferenciación de clases utilizando la ampliación de la base de datos.

3. **Implementación de la Aplicación Móvil (MVP):** Desarrollar un Producto Mínimo Viable (MVP) de la aplicación móvil que conecte el *frontend* con la base de datos y el modelo predictivo actual. Esto permitiría realizar pruebas piloto con usuarios reales, recopilar más datos e ir mejorando el modelo y la aplicación basado en la retroalimentación de los usuarios.

Agradecimientos

Deseo expresar mi más profundo agradecimiento a mi familia, cuyo amor y paciencia inquebrantable en mis capacidades constituyeron el pilar emocional que me sostuvo durante las etapas más exigentes de este proceso. Su apoyo ha sido fundamental para alcanzar este logro.

De manera especial, extiendo mi gratitud a José Miguel Jorquera, compañero invaluable y apoyador clave en el desarrollo de la tesis y, lo que es más importante, en la visión y proyección futura de su implementación práctica. Su colaboración y entusiasmo fueron esenciales para mantener el impulso del proyecto.

Finalmente, agradezco al Club Andino Universitario (CAU). La pasión por la montaña que compartimos en el club no solo inspiró la temática central de esta investigación, sino que también proporcionó el contexto esencial y la comprensión profunda de los desafíos del riesgo en actividades al aire libre, enriqueciendo significativamente el modelo predictivo propuesto.

6 Referencias Bibliográficas

[1] AllTrails, LLC. (2024). AllTrails: Trail Guides & Maps for Hiking, Camping, and Running.
[2] Ahsan, M., & Abualait, T. (2024). Machine learning applications in mountain safety: A systematic review. Journal of Outdoor Safety and Risk Management.
[3] Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. Transportation Research Part C: Emerging Technologies, 86, 202-219.
[4] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
[5] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Wadsworth & Brooks.
[6] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247-1250.
[7] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS.
[8] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
[9] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
[10] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
[11] Dalipi, F., Mendoza, D. M. A., Imran, A. S., & Yayilgan, S. Y. (2015). An Intelligent Model for Predicting the Occurrence of Skiing Injuries. IEEE.
[12] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
[13] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37.
[14] Fica, R. (2013). Crónicas del Anticristo.
[15] Fica, R. (2019). No me olviden
[16] Hosmer, D. W., & Lemeshow, S. (2013). Applied Logistic Regression. Wiley.
[17] Hun, J., Lee, D., & Kim, C. (2021). Machine Learning-Based Models for Accident Prediction at a Korean Container Port. Journal of Coastal Research, 114(SI), 241-245.
[18] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
[19] Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.
[20] Koc, K., Kunt, M., & Kisa, F. (2021). Accident Prediction In Construction Using Hybrid Wavelet-Machine Learning, 147.
[21] Komoot GmbH. (2024). Komoot: Route Planner & Navigation for Cycling and Hiking.
[22] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134.
[23] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77.
[24] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.
[25] Sakí, M. J., & Meira, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. CRC Press.
[26] Sánchez, C. (2019). Evaluación y planificación de riesgos en entornos naturales.
[27] Schubert, P. (2001). Seguridad y riesgo - Análisis y prevención de accidentes de escalada. Ediciones Desnivel.
[28] Schubert, P. (2007). Seguridad y riesgo en roca y hielo Vol. II. Ediciones Desnivel.
[29] Schubert, P. (2009). Seguridad y riesgo en roca y hielo Vol. III. Ediciones Desnivel.
[30] Sociedad Geográfica Andeshandbook. (2024). Andeshandbook: Guía de cerros y rutas de los Andes.
[31] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.
[32] Strava Inc. (2024). Strava: The Subscription Service for Athletes.
[33] Suda Outdoors. (2024). SUDA: Explora y comparte tus rutas al aire libre.
[34] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285.
[35] Taibo Vázquez, J. M. (2022). La gestión del riesgo y la seguridad en actividades de montaña.
[36] Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5(2), 207–232.
[37] Villota, S. (2017). Accidentabilidad en montaña - Estadística de rescates en España y campañas de prevención.
[38] W3C (2022). HTML5 (W3C Recommendation).
[39] Weinstein, N. D. (1980). Unrealistic optimism about future life events. Journal of Personality and Social Psychology, 39(5), 806–820.
[40] Wikiexplora. (2024). Wikiexplora: La guía del aire libre y de expediciones.
[41] Wikiloc Outdoor S.L. (2024). Wikiloc: Rutas del Mundo.
[42] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

7 Anexos

7.1 Manual de marca de la aplicación

Para poder definir una identidad visual única, simple y que permita al usuario navegar de una manera más fácil e intuitiva, se desarrolló un manual de marca donde se especifican los lineamientos gráficos y de usabilidad que va a tener la aplicación. En este manual se describe desde la motivación del desarrollo de aplicación hasta los detalles de aplicación de cada una de las secciones que se proyecta en la aplicación incluyendo la identidad visual, paleta de colores, tipografía, Imágenes e iconografía.

7.1.1 Identidad Visual

La identidad visual de **Wisetrek** refleja nuestros valores fundamentales: seguridad, innovación y conexión con la naturaleza. Cada elemento gráfico ha sido diseñado para transmitir confianza y tecnología, manteniendo una estética clara y funcional que facilite la experiencia del usuario.

Nuestra identidad se compone de los siguientes pilares:

- **Logotipo**: El símbolo y la tipografía que representan la marca de forma única.
- **Paleta cromática**: Colores que evocan profesionalismo, aventura y tranquilidad.
- **Tipografía**: Fuentes seleccionadas para garantizar legibilidad y coherencia en todos los soportes.
- **Estilo fotográfico**: Imágenes que muestran escenarios al aire libre y personas en movimiento
- **Iconografía**: Íconos simples y consistentes que refuerzan la comunicación visual.

7.1.2 Paleta de colores principales

La identidad visual de Wisetrek refleja los valores fundamentales de la aplicación: seguridad, innovación y conexión con la naturaleza. Cada elemento gráfico ha sido diseñado para transmitir confianza y tecnología, manteniendo una estética clara y funcional que facilite la experiencia del usuario. En la **Figura A.1** se detalla la paleta de colores de la aplicación, reflejando los valores fundamentales y en la **Figura A.2** los botones de estado de la aplicación.

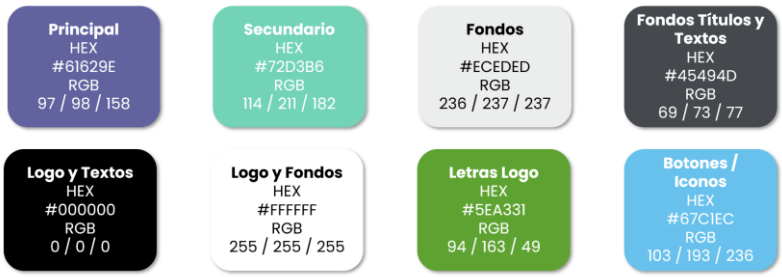

Figura A.1. Paleta de colores de la aplicación

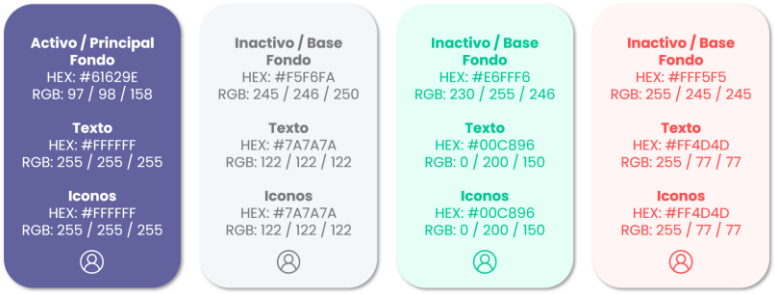

Figura A.2. Uso de Colores en Estados

7.1.3 Tipografía Concepto y Uso

La tipografía oficial de Wisetrek es Poppins, seleccionada por su legibilidad en pantallas digitales y su estilo moderno y geométrico. Su uso consistente refuerza la identidad visual y garantiza una experiencia clara y profesional.

- **Elementos clave**

Fuente única: **Poppins**
Características: Moderna, geométrica, alta legibilidad
Aplicación: Toda la interfaz de la aplicación, diseños relacionados a Wisetrek y comunicaciones.

7.1.4 Especificaciones Tipográficas

La tipografía oficial de Wisetrek es Poppins, seleccionada por su legibilidad en pantallas digitales y su estilo moderno y geométrico. Su uso consistente refuerza la identidad visual y garantiza una experiencia clara y profesional. Como se detalla en la **Figura A.3**, se define una los tipos de elementos que permite diferenciar niveles de información mediante variaciones de peso y tamaño, optimizando la lectura de datos críticos en la interfaz.

Elemento	Fuente	Peso	Tamaño	Color
Títulos / Nombres	Poppins	Semibold (600)	16–18 px	#2C2C2C
Texto Secundario	Poppins	Regular (400)	13–14 px	#7A7A7A
Botones / Íconos Inferiores	Poppins	Medium (500)	12 px	#2C2C2C / #6C63FF

Figura A.3. Especificaciones tipográficas por elemento

7.1.5 Encabezado

Destacan la información clave y guían la navegación. Utilizan la tipografía Poppins en peso semibold y color #FFFFFF o #2C2C2C en contraste.

Figura A.4. Diseño de encabezado de la aplicación

7.1.6 Espacio y Composición

Enfoque en la Seguridad: Eliminando la saturación visual, el usuario puede enfocarse en la información importante, como la estimación del nivel de riesgo o los datos de las rutas, facilitando la toma de decisiones seguras como se observa en la **Figura A.5**.

- Margen exterior: 16 px
- Espaciado entre tarjetas: 12 px
- Espaciado interno en tarjetas: 16 px
- Radio general de esquinas: 8 px
- Sombra estándar: rgba(0, 0, 0, 0.05) 0px 2px 4px

Figura A.5. Diagrama de espacios y composición de la aplicación

7.1.7 iconografía

Características principales

- Estilo: Líneas simples y formas geométricas, alineadas con la estética minimalista de la marca.
- Tamaño recomendado: 50 px para íconos principales.
- Colores: Uso del color principal (#61629E) para íconos activos y tonos neutros (#7A7A7A) para estados inactivos. En contraste con Blanco (#FFFFFF) con color secundario de fondo (#72D3B6)
- Consistencia: Mantener proporciones y evitar efectos como sombras, degradados o contornos no autorizados.

Figura A.6. Visualización y aplicación de iconografía en la aplicación

7.1.8 Barra de acciones

Íconos simples y legibles, acompañados de texto en Poppins Medium. Colores principales: #61629E y #7A7A7A para la opción Inactiva.

Figura A.7. Barra de acciones o barra inferior de navegación

7.1.9 Mock up desarrollados de la aplicación

Los *mockups* visualizan la interacción y el *layout* de la aplicación, siguiendo las directrices del manual de diseño, centrándose en el ingreso de datos y la visualización del riesgo.

7.1.10 Pantalla de Inicio (TU / Actividad)

- **Estructura:** Presenta un encabezado superior con el color Primario (#61629E) y un listado de Tarjetas blancas.
- **Propósito:** Mostrar las actividades planificadas o recientes del usuario. Cada tarjeta de actividad podría incluir un pequeño indicador de riesgo calculado previamente.
- **Barra de acciones:** Permite al usuario navegar por la aplicación.

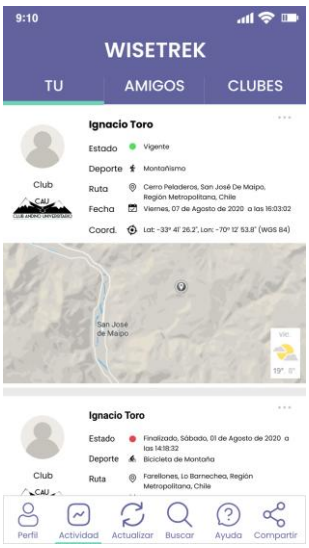

Figura A.8. Pantalla de inicio Tu/Actividad

7.1.10.1 Pantalla de Seguimiento de Actividad

- **Encabezado Superior**: Presenta un menú de pestañas (Tu, Amigos, Clubes) sobre el color Primario (#61629E).
- **Selector de Vista**: Un control segmentado para alternar entre "Track" (seguimiento) y "Configuración".
- **Cuerpo Principal**: Tarjeta blanca que contiene la información del perfil del usuario, el deporte y un mapa con la ruta trazada.

- **Panel de Datos**: Ubicación específica (Cerro Peladeros, San José de Maipo), coordenadas GPS exactas y un indicador visual de Riesgo Calculado.

Figura A.9. Pantalla de actividad en curso