



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Sistemas e Informática

Unidad de Posgrado

**Modelo de predicción basado en aprendizaje
automático para la explicabilidad del riesgo de
accidente vehicular por conducta de manejo**

TESIS

Para optar el Grado Académico de Doctor en Ingeniería de
Sistemas e Informática

AUTOR

Javier Enrique LACHERRE VARGAS

ASESOR

Dr. Ciro RODRIGUEZ RODRIGUEZ

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

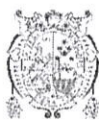
Referencia bibliográfica

Lacherre, J. (2025). *Modelo de predicción basado en aprendizaje automático para la explicabilidad del riesgo de accidente vehicular por conducta de manejo*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Javier Enrique Lacherre Vargas
Tipo de documento de identidad	DNI
Número de documento de identidad	09956084
URL de ORCID	https://orcid.org/0000-0002-4109-466X
Datos de asesor	
Nombres y apellidos	Ciro Rodriguez Rodriguez
Tipo de documento de identidad	DNI
Número de documento de identidad	06445707
URL de ORCID	https://orcid.org/0000-0003-2112-1349
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	David Santos Mauricio Sánchez
Tipo de documento	DNI
Número de documento de identidad	06445495
Miembro del jurado 1	
Nombres y apellidos	José Alfredo Herrera Quispe
Tipo de documento	DNI
Número de documento de identidad	40362859
Miembro del jurado 2	
Nombres y apellidos	Luis Soto Soto
Tipo de documento	DNI
Número de documento de identidad	22409532
Datos de investigación	

Línea de investigación	C.0.5.8. Inteligencia Artificial
Grupo de investigación	Tecnologías de la Información aplicados a la ciencia de los datos - ITDATA
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Facultad de Ingeniería de Sistemas e Informática País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Calle: Av. Amezaga Latitud: -12.05642315 Longitud: -77.08433269
Año o rango de años en que se realizó la investigación	2023 - 2024
URL de disciplinas OCDE	Ingeniería de Sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE DOCTOR EN INGENIERÍA DE SISTEMAS E INFORMÁTICA

A los veinticuatro (24) días del mes de marzo de 2025, siendo las 3:00 pm., se reunieron en el Auditorio, Profesor: Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Dr. David Santos Mauricio Sánchez (Presidente)
Dr. José Alfredo Herrera Quispe (Miembro)
Dr. Luis Soto Soto (Miembro)
Dr. Ciro Rodríguez Rodríguez (Miembro asesor)

Se inició la Sustentación invitando al candidato a Doctor **JAVIER ENRIQUE LACHERRE VARGAS**, para que realice la exposición oral de la tesis para optar el Grado Académico de Doctor en Ingeniería de Sistemas e Informática, siendo la Tesis intitulada:

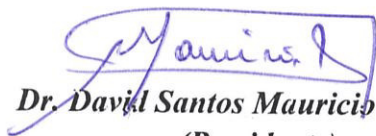
“MODELO DE PREDICCIÓN BASADO EN APRENDIZAJE AUTOMÁTICO PARA LA EXPLICABILIDAD DEL RIESGO DE ACCIDENTE VEHICULAR POR CONDUCTA DE MANEJO”

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

Diecisiete (17) Muy Bueno

Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Magíster **JAVIER ENRIQUE LACHERRE VARGAS** el Grado Académico de Doctor en Ingeniería de Sistemas e Informática.

Siendo las 16:37 horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.


Dr. David Santos Mauricio Sánchez
(Presidente)



Dr. José Alfredo Herrera Quispe
(Miembro)


Dr. Luis Soto Soto
(Miembro)


Dr. Ciro Rodríguez Rodríguez
(Miembro Asesor)



UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, **Ciro Rodriguez Rodriguez** en mi condición de asesor acreditado con **Dictamen N° 000229-2023-UPG-VDIP-FISI/UNMSM** de la tesis, cuyo título es “Modelo de predicción basado en aprendizaje automático para la explicabilidad del riesgo de accidente vehicular por conducta de manejo”, presentado por el magíster **Javier Enrique Lacherre Vargas** para optar el grado de **Doctor en Ingeniería de Sistemas e Informática**, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el **5 %** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado correspondiente.

Lima, 5 de diciembre de 2024

Firma del Asesor

Nombres y Apellidos del asesor: **Ciro Rodriguez Rodriguez**

DNI: **06020241**

DEDICATORIA

A mi esposa, **Juani**, por su amor y apoyo constante y, sobre todo, su gran paciencia durante todo este recorrido que empezó en pandemia. Gracias por creer en mí.

A mi hijo **Leandro**, quien es mi gran tesis de vida. Este esfuerzo ha requerido robarte horas que debieron ser dedicadas a ti, pero lo hice con la esperanza que sea un ejemplo de que, con dedicación y perseverancia, se pueden lograr grandes metas. Espero que siempre te sientas orgulloso y motivado a seguir tus propios sueños con determinación y coraje.

A mis padres **Abilia e Isaías**, quienes han sido mi mayor ejemplo de dedicación, esfuerzo y ética; su apoyo constante y sus valores me han inspirado a siempre dar lo mejor de mí.

AGRADECIMIENTOS

A mi asesor de tesis, el doctor **Ciro Rodríguez** le agradezco por el gran apoyo, su motivación y sus enseñanzas. Gracias por creer en mi potencial y brindarme siempre el aliento para cerrar este hito en mi crecimiento.

Quiero agradecer a todos mis profesores, ya que en cada materia que tuve la oportunidad de cursar con ellos, no solo adquirí conocimientos, sino que cada lección impartida por su experiencia representó un paso significativo hacia la consecución de esta etapa.

Quiero expresar mi agradecimiento a mis colegas de trabajo e investigación por su respaldo y colaboración en este camino.

ÍNDICE GENERAL

DEDICATORIA	II
AGRADECIMIENTOS	III
ÍNDICE GENERAL	IV
LISTA DE TABLAS	VII
LISTA DE FIGURAS	VIII
RESUMEN	IX
ABSTRACT	X
CAPÍTULO I. INTRODUCCIÓN	1
1.1. Situación problemática	1
1.2. Formulación del problema	4
1.2.1. <i>Problema general</i>	4
1.2.2. <i>Problemas específicos</i>	4
1.3. Justificación de la investigación	4
1.3.1. <i>Justificación teórica</i>	4
1.3.2. <i>Justificación práctica</i>	5
1.4. Objetivos de la investigación	6
1.4.1. <i>Objetivo general</i>	6
1.4.2. <i>Objetivos específicos</i>	6
CAPÍTULO II. MARCO TEÓRICO	7
2.1. Antecedentes de investigación	7
2.1.1. <i>Factores</i>	7
2.1.2. <i>Predicción</i>	9
2.1.3. <i>Explicabilidad</i>	10
2.2. Bases teóricas	11
2.2.1. <i>Conducta de manejo</i>	11
2.2.2. <i>Riesgo de accidente vehicular</i>	12
2.2.3. <i>Factores</i>	13
2.2.4. <i>Modelo de predicción</i>	14
2.2.5. <i>Modelo ensamblado</i>	16
2.2.6. <i>Explicabilidad</i>	18
2.3. Marco conceptual	18
CAPÍTULO III. METODOLOGÍA	20
3.1. Hipótesis general	20
3.2. Hipótesis específicas	20

3.3.	Identificación de variables	21
3.4.	Operacionalización de variables	21
3.5.	Matriz de consistencia	23
3.6.	Tipo y diseño de investigación	26
3.7.	Unidad de análisis	26
3.8.	Población de estudio	27
3.9.	Tamaño de muestra	27
3.10.	Selección de muestra	27
3.11.	Técnicas de recolección de datos	29
3.12.	Análisis e interpretación de información	30
CAPÍTULO IV. DESARROLLO DE LA PROPUESTA.....		32
4.1.	Definir e implementar un estudio naturalista de recolección de datos de telemetría	32
4.1.1.	<i>Definición del caso de estudio: Corredor minero sur.</i>	33
4.1.2.	<i>Definición del área de estudio</i>	34
4.1.3.	<i>Selección de vehículos, conductores y equipamiento</i>	37
4.1.4.	<i>Identificar factores de predicción de RAV por conducta de manejo</i>	39
4.1.5.	<i>Construcción del conjunto de datos</i>	43
4.1.6.	<i>Análisis de calidad del conjunto de datos</i>	51
4.2.	Diseño del modelo ensamblado de predicción de riesgo de accidente vehicular	57
4.2.1.	<i>Indicadores de desempeño</i>	57
4.2.2.	<i>División del conjunto de datos</i>	58
4.2.3.	<i>Aprendizaje individual</i>	59
4.2.4.	<i>Aprendizaje conjunto</i>	64
4.2.5.	<i>Herramientas de desarrollo</i>	67
4.2.6.	<i>Análisis de efectividad de modelo</i>	67
4.3.	Validación del modelo ensamblado de predicción de riesgo de accidente vehicular	70
4.3.1.	<i>Buscar hiper parámetros</i>	71
4.3.2.	<i>Ensamble optimizado</i>	72
4.3.3.	<i>Explicabilidad del modelo</i>	72
CAPÍTULO V. RESULTADOS Y DISCUSIÓN.....		73
5.1.	Análisis, interpretación y discusión de resultados	73
5.1.1.	<i>Estudio naturalista de recolección de datos de telemetría</i>	73
5.1.2.	<i>Modelo ensamblado de predicción de riesgo de accidente vehicular</i>	77
5.1.3.	<i>Modelo ensamblado optimizado</i>	78
5.2.	Prueba de hipótesis	84
5.2.1.	<i>Hipótesis específica 1</i>	84
5.2.2.	<i>Hipótesis específica 2</i>	85
5.2.3.	<i>Hipótesis específica 3</i>	87
5.2.4.	<i>Hipótesis general</i>	88
CONCLUSIONES		90
RECOMENDACIONES		91
REFERENCIAS BIBLIOGRÁFICAS		92
ANEXO A		101
ANEXO B		102

ANEXO C	103
----------------------	------------

LISTA DE TABLAS

Tabla 1.....	15
Tabla 2.....	15
Tabla 3.....	17
Tabla 4.....	21
Tabla 5.....	23
Tabla 6.....	28
Tabla 7.....	29
Tabla 8.....	30
Tabla 9.....	39
Tabla 10.....	50
Tabla 11.....	52
Tabla 12.....	53
Tabla 13.....	56
Tabla 14.....	59
Tabla 15.....	61
Tabla 16.....	61
Tabla 17.....	64
Tabla 18.....	65
Tabla 19.....	66
Tabla 20.....	67
Tabla 21.....	68
Tabla 22.....	68
Tabla 23.....	69
Tabla 24.....	69
Tabla 25.....	70
Tabla 26.....	71
Tabla 27.....	74
Tabla 28.....	75
Tabla 29.....	77
Tabla 30.....	78
Tabla 31.....	79
Tabla 32.....	79
Tabla 33.....	81
Tabla 34.....	85

LISTA DE FIGURAS

Figura 1	11
Figura 2	13
Figura 3	18
Figura 4	32
Figura 5	33
Figura 6	34
Figura 7	35
Figura 8	35
Figura 9	36
Figura 10	37
Figura 11	38
Figura 12	44
Figura 13	44
Figura 14	47
Figura 15	47
Figura 16	48
Figura 17	54
Figura 18	55
Figura 19	57
Figura 20	71
Figura 21	80
Figura 22	82
Figura 23	83
Figura 24	83

RESUMEN

Este estudio tiene como objetivo desarrollar un modelo de predicción basado en aprendizaje automático para explicar el riesgo de accidentes vehiculares en vehículos pesados del sector minero, específicamente en carreteras rurales. Con datos de telemetría recolectados en tiempo real, se busca identificar patrones de comportamiento del conductor y condiciones del vehículo que logren predecir accidentes en tiempo real. Se usa un modelo ensamblado, que combina varios algoritmos de aprendizaje automático, con el fin de mejorar la exactitud y sensibilidad en la predicción de los riesgos.

El estudio se estructura en tres fases principales. Primero, se lleva a cabo un estudio naturalista que recolecta datos de telemetría de vehículos pesados en operaciones mineras en carreteras rurales. En segundo lugar, se desarrolla y valida un modelo de predicción que combina modelos individuales para optimizar los resultados y lograr un equilibrio entre exactitud, sensibilidad y capacidad explicativa. Por último, se evalúa el modelo con un conjunto de datos independiente, y se demuestra que mantiene altos niveles de exactitud y fiabilidad.

Los resultados obtenidos muestran que el modelo propuesto puede identificar de manera efectiva los riesgos de accidentes en función de 15 características del comportamiento del conductor. Además, el estudio confirma que un modelo ensamblado mejora significativamente los resultados en comparación con modelos individuales.

Este enfoque está desarrollado para el sector minero, pero podría extenderse a otros sectores del transporte pesado. Por lo tanto, el estudio proporciona un modelo innovador para mejorar la seguridad vial y reducir el riesgo de accidentes en carreteras rurales.

ABSTRACT

This study aims to develop a machine learning-based prediction model to explain the risk of vehicle accidents in heavy vehicles in the mining sector, specifically on rural roads. Using telemetry data collected in real time, we seek to identify patterns of driver behavior and vehicle conditions that can predict accidents in real time. An ensemble model, which combines several machine learning algorithms, is used to improve the accuracy and sensitivity in risk prediction.

The study is structured in three main phases. First, a naturalistic study is carried out that collects telemetry data from heavy vehicles in mining operations on rural roads. Second, a prediction model is developed and validated that combines individual models to optimize the results and achieve a balance between accuracy, sensitivity and explanatory capacity. Finally, the model is evaluated with an independent data set, and it is shown that it maintains high levels of accuracy and reliability.

The results obtained show that the proposed model can effectively identify accident risks based on 15 characteristics of driver behavior. Furthermore, the study confirms that an assembled model significantly improves results compared to individual models.

This approach is developed for the mining sector, but could be extended to other heavy transport sectors. The study therefore provides an innovative model for improving road safety and reducing the risk of accidents on rural roads.

CAPÍTULO I. INTRODUCCIÓN

1.1. Situación problemática

En el mundo se producen alrededor de 1,19 millones de muertes al año a causa de los accidentes de tránsito (Organización Mundial de la Salud, 2023). En ese sentido, la Organización de las Naciones Unidas (ONU) plantea 17 objetivos de desarrollo sostenible (ODS) para el año 2030 donde el ODS-3 de “Buena salud y bienestar” tiene como una de sus metas “reducir a la mitad el número de muertes y lesiones causadas por accidentes de tránsito en el mundo” (Naciones Unidas, 2018).

En Perú, las víctimas de accidentes de tránsito fatales en el año 2023 llegaron a 3,317 y se pueden resaltar que las causas principales que originan los siniestros de tránsito son imprudencia del conductor con 28%, exceso de velocidad con 26%, ebriedad del conductor con 7%, invasión de carril con 3% y desacato de señal de tránsito con 1%. La suma de todas estas causas que están relacionadas a la conducta de manejo (CM) suma 65% (Observatorio Nacional de Seguridad Vial, 2024).

En la actualidad, los diferentes reportajes periodísticos o artículos en revistas que hablan constantemente sobre la conducción automatizada o vehículos autónomos nos pueden hacer pensar que estudiar los accidentes de tránsito y predicción de riesgo de accidente vehicular (RAV) es un tema obsoleto y resuelto, sin embargo, de acuerdo a Kashevnik et al. (2021) esto aún se encuentra muy lejos.

La predicción del RAV presenta una alta complejidad debido a la diversidad de escenarios involucrados. Esto exige la consideración de múltiples variables que influyen en la seguridad vial, por lo que resulta crucial garantizar la fiabilidad en la recolección de datos. Además, es muy importante la precisión debido a que está en juego vidas humanas. En ese sentido Orsini et al. (2021) señalan que tener datos desbalanceados influyen en la precisión. Esto se debe a que de toda la data que se recolecta, aquellos registros que presentan riesgo de accidente tienden a ser muy pocos (X. Shi et al., 2019) y esto puede generar sesgos en la predicción del RAV (Z. Ma et al., 2021).

Uno de los métodos usados mundialmente para la predicción de accidentes es el manual de seguridad vial de 2010 (Highway Safety Manual [HSM]). Sin embargo, este método está enfocado en predecir la cantidad de accidentes en un tramo y de esta manera ver que tan riesgoso es ese tramo y esto lo realiza usando factores como la frecuencia de vehículos en el tramo y algunos datos de geografía de la zona (Das et al., 2021). La desventaja es que usa datos históricos (J. Hu et al., 2020) y no se realiza una predicción en tiempo real del RAV (Tola et al., 2022). En Perú, también se adoptó este manual, pero, en la práctica genera problemas dado que se necesita datos históricos que usualmente no se tiene para todas las zonas del país.

Hoy en día existen diversos tipos de sensores en los vehículos que generan alertas cuando detectan comportamientos anómalos. Sin embargo estas alertas muchas veces se activan por falsos positivos y esto hace que los conductores desconfíen de estos sensores (Basso et al., 2021) a tal punto que los conductores rechazan su uso (H. Wang et al., 2020) e incluso intentan sabotearlos debido a que emiten alarmas sonoras con mucha frecuencia.

Otro factor a considerar es el incremento del tráfico por el alto volumen de vehículos que se compra anualmente. Por ejemplo, en Perú se ha incrementado bastante el número de motos y esto está relacionado a su vez con el incremento de la migración. No considerar estas características complejas (Alkinani et al., 2020) ni los cambios en el tráfico puede afectar negativamente la precisión en la predicción del RAV (Jiang et al., 2020).

Shangguan et al. (2022) destacan en su estudio la importancia de identificar la intención del conductor, especialmente al realizar un cambio de carril. En las carreteras rurales de Perú, muchas de las cuales cuentan con solo dos carriles y presentan numerosas curvas pronunciadas, esta maniobra se convierte en una acción de alto riesgo. Por ello, no anticipar la intención de conducción antes de ejecutar la maniobra puede complicar la predicción del riesgo de accidentes (Xu et al., 2022).

En cuanto al factor de fatiga y somnolencia, existen diversos desafíos para identificar con exactitud el estado del conductor, principalmente porque muchos sensores no son estandarizados y requieren configuraciones individuales. Esto significa que la detección depende, por ejemplo, de la altura del conductor en relación con el ángulo de la cámara que captura las imágenes para monitorear la fatiga. Esta alta variabilidad interindividual genera dudas sobre la exactitud en la predicción del riesgo de accidentes (Jacobé de Naurois et al., 2018). Además, aunque existen sensores que se adhieren a la piel, estos suelen ser intrusivos o semi-intrusivos, lo que puede afectar la experiencia de conducción y ser rechazados por los conductores. Esto, a su vez, incrementa la tasa de falsos positivos, afectando la exactitud en la evaluación del riesgo (Liu et al., 2021).

Así también, en el estudio de Ma et al. (J. Ma et al., 2019) concluyó que conducir bajo los efectos del alcohol incrementa significativamente el RAV y se posiciona como uno de los principales factores de mortalidad en accidentes de tráfico.

Finalmente, cada uno de los factores estudiados influye en la conducta de manejo, y según Guido et al. (2022), el factor humano es la principal causa del RAV. Sin embargo, al igual que Brühwiler et al. (2022) recomiendan considerar otros aspectos, como la geometría de las carreteras, para mejorar el rendimiento de los modelos de predicción.

1.2. Formulación del problema

1.2.1. Problema general

¿De qué manera el desarrollo de un modelo de predicción basado en aprendizaje automático permite explicar el riesgo de accidentes vehiculares por conducta de manejo en vehículos pesados del sector minero en carretera rural?

1.2.2. Problemas específicos

- PE1: ¿Cómo la ejecución de un estudio naturalista de recolección de datos de telemetría en vehículos pesados del sector minero en carretera rural, permite generar un conjunto de datos adecuado para el desarrollo y validación de un modelo de predicción basado en aprendizaje automático del riesgo de accidentes?
- PE2: ¿Cómo diseñar y validar un modelo ensamblado de predicción que combine de manera efectiva múltiples modelos de aprendizaje automático para optimizar la exactitud y sensibilidad del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural en comparación con modelos individuales?
- PE3: ¿Cómo la evaluación de la eficiencia predictiva del modelo ajustado en la identificación del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carreteras rurales, con un conjunto de datos independiente, asegura que el modelo tenga altos niveles de exactitud y explicabilidad?

1.3. Justificación de la investigación

1.3.1. Justificación teórica

En el mundo se producen alrededor de 1,19 millones de muertes al año a causa de los accidentes de tránsito y esto tiene un costo de 3% de su PBI para la mayoría de los países. Uno de los objetivos mundiales para el

Desarrollo Sostenible al año 2030 es reducir a la mitad el número de víctimas fatales y lesiones por accidentes de tránsito en todo el mundo (Organización Mundial de la Salud, 2023). En el caso de Perú, la meta es reducir las muertes por accidentes de 10 a 4.31 por cada 100,000 habitantes para 2025 (Plan Estratégico Nacional de Seguridad Vial PENsv 2017-2021, 2017) y en 2023, el 65% de los accidentes de tránsito estuvieron relacionados con el comportamiento de conducción (Observatorio Nacional de Seguridad Vial, 2024).

La justificación teórica se concentra en los siguientes puntos:

- Aportar un modelo de predicción en tiempo real basado en aprendizaje automático para explicar el riesgo de accidentes por conducta de manejo en vehículos del sector minero en carretera rural.
- Alinear al Objetivo de Desarrollo Sostenible al año 2030 para promover el bienestar y reducir a la mitad el número lesiones y víctimas fatales por accidentes de tránsito en todo el mundo.
- Incrementar el conocimiento con la literatura revisada acerca del tema de investigación considerando que existen pocos trabajos sobre la predicción de riesgo de accidentes de vehículos pesados en carreteras rurales de dos carriles y en dos sentidos.
- Sensibilizar y alertar acerca de la importancia de abordar el tema de los accidentes de tránsito relacionados al comportamiento de conducción (CC), que sigue siendo la principal causa de accidentes vehiculares.

1.3.2. Justificación práctica

La investigación contribuye con un modelo predictivo de RAV por CC en una carretera rural de dos carriles y dos sentidos el cual está basado en la aplicación de técnicas de aprendizaje automático.

Diversos estudios en el área han demostrado que la aplicación de técnicas de aprendizaje automático para la predicción y prevención de accidentes puede mejorar significativamente la seguridad vial en entornos urbanos (Zhao et al., 2022) y logísticos (Harb et al., 2008). Así mismo, los resultados del estudio ayudarán a crear más sistemas que sean usados por

personas jurídicas, principalmente del sector transporte logístico que luego pueda extenderse a personas naturales para que tomen conciencia de reducir el riesgo de accidente en carreteras rurales mediante el envío de mensajes o alertas anticipadas en cabina y de esta manera mejorar la conducta de manejo y prevenir los accidentes fatales.

1.4. Objetivos de la investigación

1.4.1. *Objetivo general*

Desarrollar un modelo de predicción basado en aprendizaje automático para explicar el Riesgo de Accidentes Vehiculares por conducta de manejo en vehículos pesados del sector minero en carretera rural.

1.4.2. *Objetivos específicos*

- OE1: Ejecutar un estudio naturalista de recolección de datos de telemetría en vehículos pesados del sector minero en carretera rural, que permita generar un conjunto de datos adecuado para el desarrollo y validación de un modelo de predicción basado en aprendizaje automático del riesgo de accidentes.
- OE2: Diseñar y validar un modelo ensamblado de predicción que combine de manera efectiva múltiples modelos de aprendizaje automático para optimizar la exactitud y sensibilidad del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural en comparación con modelos individuales.
- OE3: Evaluar la eficiencia en la predicción del modelo ajustado en la identificación del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carreteras rurales, con un conjunto de datos independiente, asegura que el modelo tenga niveles altos de exactitud y explicabilidad.

CAPÍTULO II. MARCO TEÓRICO

2.1. Antecedentes de investigación

Desde 1975, han destacado estudios como el de Chipman y Morgan (1975) sobre la predicción del RAV, quienes exploraron diversos factores que influyen en el riesgo, como los puntos de demérito, la edad, el género, la clase de licencia y el historial de accidentes. Sus hallazgos identificaron que los puntos de demérito son el factor que más influye en el riesgo de accidentes, lo que ofrece una oportunidad para prevenirlos cuando se relacionan con el CC. En los siguientes años, la investigación ha llevado a la implementación de cambios y regulaciones en los ámbitos ambiental, vehicular, de tránsito, del conductor y de la gestión. Estas medidas lo que buscan es reducir los riesgos usando elementos como dispositivos de desaceleración y barreras de protección central en las vías para la mitigación del riesgo (Geng et al., 2022). Asimismo, se han implementado algunos métodos para prevenir colisiones, identificar peatones, alertar sobre cambio de carril, identificar distracciones y somnolencia, entre otras (Celaya-Padilla et al., 2019). Esto ha generado que los gobiernos implementen manuales de seguridad, como el Manual de Seguridad Vial (AASHTO, 2010), que es uno de los más utilizados en el mundo (Das et al., 2021), sin embargo, no considera el CC en su modelo estadístico (Tola et al., 2022).

A continuación, se presentan los antecedentes sobre factores, predicción y explicabilidad del riesgo de accidentes vehiculares.

2.1.1. Factores

Se han realizado investigaciones sobre factores que ayudan a predecir accidentes de tráfico utilizando métodos de aprendizaje automático (ML, por sus siglas en inglés). Xu et al. (2022) encontraron que existe una fuerte

correlación entre la conducción agresiva y aspectos relacionados con el conductor, el vehículo y el entorno. De manera similar, Li et al. (2023) consideran el entorno, el vehículo, el conductor y el tráfico. Asimismo, Niu et al. (2021) y Yang et al. (2021) también incluyen la gestión. Es fundamental estudiar cada los factores, de tal manera que ayude a mitigar el riesgo de accidentes y mejorar la seguridad vial (Zhong et al., 2022).

Galatioto et al. (2018) encuentran que es posible predecir el número de víctimas y nivel de gravedad usando datos históricos y simulando situaciones para anticipar los riesgos en carretera. Este estudio considera factores como la demografía, el uso del suelo, la geometría de la vía, el tráfico y el contexto del accidente. Un estudio similar en Arabia Saudita, que abarca datos desde 1996 hasta 2013, evalúa factores demográficos, datos del conductor, datos geográficos y detalles de los accidentes. Esto ayuda para predecir accidentes hasta el año 2030 (Alrajhi & Kamel, 2019).

J. Ma et al. (2019) indican que conducir bajo los efectos del alcohol fue uno de los principales factores de mayor influencia en el accidente por lo que analiza los datos del sistema integrado de registros de tráfico del estado de California en USA desde el 2010 al 2012 para analizar las causas principales de muertes por accidente de tránsito y predecir la conducta de manejo en tiempo real. En este estudio se consideran los factores relacionados al accidente y factores ambientales.

Papadimitriou et al. (2019) mencionan que el 70% de los casos se ven influenciados por el uso del celular para lo cual recopila datos del mismo dispositivo para evaluar el uso del celular al conducir y asociarlos al CC evaluando factores relacionados al conductor, vehículo, tránsito e infracciones de tránsito. Además del celular, se usas otros sensores como unidad de medición inercial, GPS y el bus CAN que generan datos de conducción enviando mensajes de alta calidad en tiempo real y de esta manera la tecnología puede ayudar a reducir los accidentes mejorando la predicción de riesgo de accidente (Fridman et al., 2019).

Nour et al. (2020) realizan estudios para identificar el riesgo de accidente y nivel de gravedad en base a data histórica del reino unido desde 2005 al 2019. Para realizar el estudio se evaluaron factores relacionados al accidente, vehículo y entorno y uno de sus resultados resaltantes es que, el límite de velocidad no tiene mucha influencia sobre el nivel de gravedad de lesión por el accidente.

2.1.2. Predicción

En cuanto a la predicción, se han utilizado diversos algoritmos de inteligencia artificial para predecir el riesgo de accidentes vehiculares. Geng et al. (2022) presentan un marco de modelado extenso para evaluar la seguridad de camiones en carreteras rurales de dos carriles utilizando Extreme Gradient Boosting (XGBoost), logrando una exactitud del 96.67%. En Peng et al. (2021), se señala que Long Short-Term Memory (LSTM) es adecuado para extraer información significativa y continua del vehículo, como aceleraciones y desaceleraciones, por lo que lo aplican para la predicción del Riesgo de Accidente Vehicular por Conducta de Manejo (RAVCC), obteniendo una exactitud del 93.5%.

Jiang et al. (2020) realizan el análisis de datos de tráfico en tiempo real para identificar el comportamiento agresivo y mejorar la predicción por lo que usan 2 ventanas de tiempos distintos y obtienen una buena precisión de la predicción. Así también, se realiza un estudio que analiza una ventana de 30 segundos para predecir los próximos 10 segundos y de esta manera detectar un comportamiento agresivo y una ventana de 60 segundos para detectar la conducción distraída usando variables fisiológicas para lograr una predicción de 87% (Panagopoulos & Pavlidis, 2020).

Hu et al. (2020) aprovechan los datos de conducción de los autos conectados durante un mes para predecir el riesgo de accidente en intersecciones de choques frecuentes. Los resultados muestran que el modelo propuesto tiene una predicción notable al usar este tipo de medios de comunicación de los autos conectados.

Orsini et al. (2021) tienen como objetivo predecir accidentes relacionados a colisiones traseras usando sensores de radar para medir la distancia con el vehículo circundante. Este modelo predice la ocurrencia de situaciones inseguras en 5 minutos con una exactitud robusta.

2.1.3. Explicabilidad

Por otro lado, también se han utilizado diversos algoritmos para explicar el RAVCC. En Masello et al. (2023), se aplica SHapley Additive exPlanations (SHAP) y se encuentra que el límite de velocidad es un factor muy relevante en los eventos de mayor riesgo. En ese mismo sentido, el estudio de Alfai et al. (2022), basado en el método de importancia de características de Random Forest (RF), descubrió que los predictores más significativos para el RAVCC son la velocidad media del vehículo, la velocidad instantánea y su aceleración longitudinal.

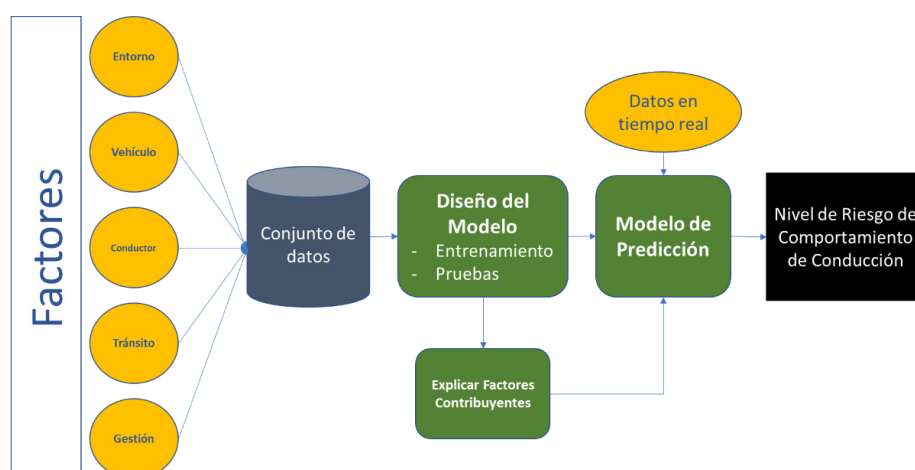
Los estudios revelan que se observa una tendencia a utilizar algoritmos que evalúan el riesgo en función de la frecuencia de accidentes y del RAVCC, pero no discriminan si se trata de vehículos ligeros o pesados, ni asocian factores relacionados con la gestión del viaje del vehículo, como el tiempo estimado de retraso hasta el punto final o la influencia de si un vehículo pesado viaja cargado o vacío. Por otro lado, se enfocan en factores que contribuyen a explicar la frecuencia o la gravedad del accidente, pero no en identificar los factores que contribuyen específicamente al RAVCC. Esto es importante, ya que cada vez existen más regulaciones para incorporar mecanismos que permitan la lectura de datos de trayectoria y seguridad para las organizaciones. De este modo, la lectura de estos datos, la predicción en tiempo real y la explicación de las causas pueden ayudar a mitigar la cantidad de accidentes.

2.2. Bases teóricas

Para realizar el estudio de la predicción del RAVCC se manejará en base a un modelo conceptual (ver Figura 1) que permite entender las entradas y salidas del proceso a realizar. Para ello se identifican los factores a usar para la clasificación del riesgo. Una vez identificados se realizará la elaboración de un conjunto de datos que será probado en el modelo de aprendizaje automático para luego explicar los factores más influyentes. Con estos factores se finaliza el modelo para ser usado con datos en tiempo real y finalmente, determinar el nivel de riesgo en línea.

Figura 1

Modelo conceptual del diseño de riesgo de accidente vehicular



Nota: Este gráfico incluye las categorías de factores, el conjunto de datos, el diseño del modelo, el modelo resultante, el nivel de riesgo resultante y la explicación de factores.

2.2.1. Conducta de manejo

Nos referimos a identificar cómo se desarrollan las acciones del conductor al realizar la actividad de conducción en diversas situaciones (Halim et al., 2023) desde un punto de inicio hacia un punto final considerando factores como por ejemplo el tiempo de retraso al punto destino. La conducta

de manejo puede ser clasificada en agrupaciones que presenten patrones similares y permitan estimar los niveles de riesgo de conducción (X. Shi et al., 2019). Esta clasificación se puede realizar considerando (1) recopilación de datos, (2) características de entrada, (3) un modelo clasificador y (4) salida, por ejemplo: normal o agresivo (Silva & Eugenio Naranjo, 2020).

Conducción normal: Se refiere al conductor que presenta una conducta segura en un corto tiempo y el valor de los factores relacionados al comportamiento del conductor en un viaje no sobrepasan el umbral permitido.

Conducción agresiva: Se refiere al conductor que tiende a presentar una variación del valor de alguno de los factores en un corto tiempo de un viaje y están relacionados al comportamiento del conductor sobrepasando un umbral permitido (Silva & Eugenio Naranjo, 2020) por ejemplo, al tener un tiempo para la colisión menor a 1 segundo y somnoliento (Alkinani et al., 2020).

Conducción peligrosa: Se refiere a una conducta aún más riesgosa que la agresiva, donde el conductor no solo sobrepasa los umbrales permitidos, sino que también exhibe patrones de comportamiento extremadamente arriesgados, como la aceleración brusca, la desaceleración inesperada o giros bruscos continuos, lo que incrementa significativamente el riesgo de accidentes. Este tipo de comportamiento es más extremo que la conducción agresiva y puede ser fatal en situaciones críticas.

2.2.2. Riesgo de accidente vehicular

Es la ponderación de la probabilidad que suceda algo peligroso y puede producir un daño grave a consecuencia de un factor de peligro (Real Academia Española, 2024). Estos factores son características o elementos que se presentan o producen al realizar la actividad de conducción vehicular y, de acuerdo a Santos et al. (2022) y Silva et al. (2020) se encuentran en las dimensiones de Entorno, Tránsito, Vehículo, Conductor y Accidente donde algunos factores tendrán mayor influencia que otros y dependiendo de eso la ponderación del riesgo será mayor. A esta clasificación se agregará en el presente estudio la dimensión de gestión que incluirá algunos factores

relacionados directamente a la gestión en vehículos pesados al realizar el traslado de mercadería.

2.2.3. Factores

Los factores asociados al comportamiento del conductor son variables que influyen en sus decisiones y acciones al volante, incluyendo características demográficas, estado físico y mental, conocimientos y habilidades de conducción, actitudes hacia la seguridad vial, uso de dispositivos electrónicos, condiciones del entorno y normas sociales. Estos factores determinan en gran medida el riesgo de incidentes viales y la efectividad de las medidas de prevención. En el estudio Lacherre et al. (2024) se identificaron 115 factores agrupadas en 5 dimensiones (Ver Figura 2):

- Entorno: entorno y distribución geográfica.
- Tránsito: relacionado con los vehículos que rodean al objeto de estudio.
- Vehículo: características de su modo estático o en movimiento.
- Conductor: relacionado con el ser humano que conduce el vehículo.
- Gestión: flota eficiente de vehículos y control y coordinación de conductores.

Figura 2

Dimensiones de factores de RAVCC



Nota: Incluye la cantidad de factores asociados a cada uno de los 5 dominios.

La dimensión más estudiada es el vehículo y los factores más utilizados debido a su influencia directa en la capacidad del conductor para controlar el vehículo en diversas situaciones de riesgo son la velocidad, la aceleración y la distancia entre dos vehículos. Además, estos factores también determinan el nivel de gravedad de un accidente. Otros factores cruciales incluyen la ubicación geográfica, determinada a través del GPS, ya que permite comprender otros elementos relacionados, como el entorno geográfico.

2.2.4. Modelo de predicción

Es una herramienta basada en técnica estadística, lógica difusa o de aprendizaje automático que permite pronosticar el comportamiento ante algún suceso con el fin de pronosticar resultados probables a futuro como la conducta de manejo (Sarker et al., 2021). Estos modelos usan los factores como entrada para realizar la predicción, pero al obtener el resultado de predicción, no se puede conocer cuál de los factores ha tenido una mayor contribución en la generación del efecto y debido a esto es que se le conoce como una técnica de “caja negra” (Orsini, Gecchele, Rossi, & Gastaldia, 2021). Por lo que, para entender esta caja negra e identificar la contribución de los factores se usan técnicas de explicabilidad.

En Lacherre et al. (2024), se analizan 80 estudios de un total de 1674 estudios potenciales, y se identifican para predecir RAVCC, 13 algoritmos de ML y 7 algoritmos híbridos de ML. Además, en dicho estudio se identifica que la mayoría de los estudios son para vehículos ligeros, es decir, hay una escasez de investigaciones para vehículos pesados. La Tabla 1 muestra los únicos 3 estudios identificados para predecir RAVCC en vehículos pesados y la Tabla 2, muestra los 7 algoritmos híbridos para predecir RAVCC en vehículos ligeros.

Tabla 1

Estudios de RAVCC en vehículos pesados basados en ML

Algoritmo	Estudio	Descripción	País	Dataset	Métrica
Sequential Covering	Geng et al. (2022)	Clasifica las maniobras de conducción basándose en datos de series temporales extraídos	China	Propio	Exactitud: 96.25%
AIO-HMM	Seo et al. (2022)	Predice el comportamiento del conductor en intersecciones con semáforos, utilizando datos de mirada y reconocimiento de semáforos	Corea del Sur	Propio	DR: 72.9%
XGBoost	Niu et al. (2021)	Predice conductas peligrosas como conducción distraída y agresiva, en un corto plazo	China	Propio	Exactitud: 80.24%

Fuente. Datos tomados de Lacherre et al. (2024)

Tabla 2

Estudios híbridos de RAVCC en vehículos ligeros basados en ML

Algoritmo	Estudio	Descripción	Países	Dataset	Métrica
MCS: BL+KNN+SVM+MLP	Elassad et al. (2020)	Aprendizaje conjunto basado en el uso de varios clasificadores base que operan sobre	Marruecos	Propio	F1 Score: 93.56%

		características fusionadas y un meta clasificador.			
GMM: HC+FA	J. Wang et al. (2022)	Identifica el comportamiento de conducción agresivo basado en viajes de manera más completa y efectiva.	China	SH-NDS	Exactitud: 87%
LSTM: Ensemble Classifier	H. Wang et al. (2022)	Realiza un método de reconocimiento de comportamiento de conducción agresivo basado en el aprendizaje conjunto.	China	Propio	Exactitud: 90.5%
LSTM + HMM	Xu et al. (2022)	Se realiza un método de predicción del comportamiento de conducción agresiva temporal multivariante.	China	SH-NDS	Exactitud: 84%
BCDU-Net+ ConvLSTM	Z. Hu et al. (2023)	Predice el riesgo de accidentes de tráfico basado en características espacio-temporales profundas contenidas en la secuencia del accidente de tráfico.	China	Propio	Exactitud: 98.48%
ANN+Backpropagation Levenberg-Marquardt	Lattanzi & Freschi, (2021)	Predecir comportamientos de conducción inseguros aprovechando la variedad de sensores ya presentes en los automóviles modernos.	Turquía	DBD	Exactitud: 83%
CNN+GRU+XGBoost	L. Shi et al. (2022)	Predice el riesgo de accidentes basado en datos de conducción continua de alta frecuencia y alta resolución.	EEUU	SHRP 2	Exactitud: 97.5%

Fuente. Datos tomados de (Lacherre et al., 2024)

2.2.5. Modelo ensamblado

El Modelo ensamblado es una técnica del aprendizaje automático que combina las predicciones de múltiples modelos individuales (también llamados aprendices base) para mejorar la exactitud y robustez del modelo

final. La idea central del aprendizaje conjunto es que la agregación de varios modelos puede compensar los errores individuales, lo que conduce a un rendimiento superior en comparación con cualquier modelo individual (Sagi & Rokach, 2018).

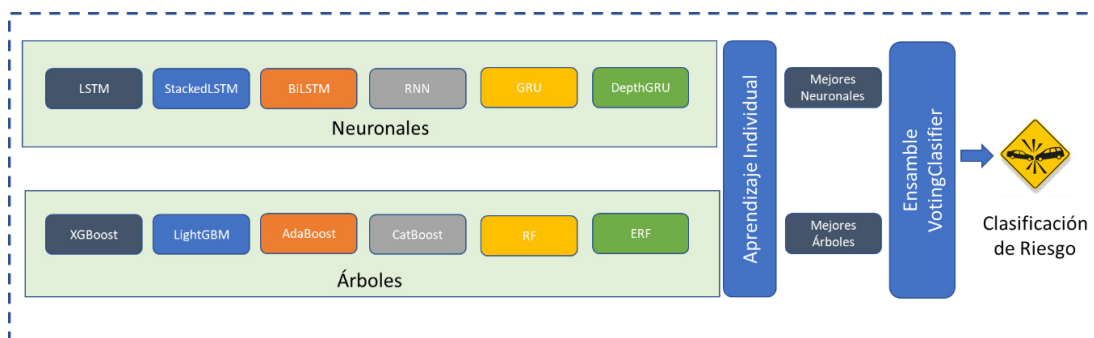
A continuación, se presenta una tabla comparativa entre modelos individuales y modelos ensamblados (ver Tabla 3):

Tabla 3

Comparación entre modelos individuales y modelos ensamblados

Criterio	Modelo Individual	Modelo Ensamblado	Cita
Desempeño (Precisión /Exactitud)	Se basa en un único algoritmo, lo que puede hacerlo sensible a datos atípicos y aumentar el riesgo de sobreajuste.	Combina múltiples algoritmos, reduciendo el error y la varianza, lo que generalmente mejora la precisión global del sistema.	(Dietterich, 2000) (Breiman, 2001)
Sensibilidad	Puede tener limitaciones para capturar patrones complejos en datos heterogéneos.	La combinación de diferentes modelos permite detectar patrones variados, mejorando la capacidad de respuesta ante variaciones en los datos.	(Kuncheva, 2004)
Interpretabilidad	Generalmente más sencillo de interpretar debido a su estructura única y menos complejidad.	Puede resultar menos intuitivo; sin embargo, el uso de técnicas explicativas (por ejemplo, LIME o SHAP) facilita la interpretación de sus predicciones.	(Ribeiro et al., 2016)

En este estudio, se propone implementar un método de ensamble que aprovecha los mejores algoritmos con el fin de clasificar el RAVCC para camiones logrando una óptima capacidad predictiva en función de las métricas de sensibilidad y exactitud (ver Figura 3).

Figura 3**Modelo Ensamblado**

Nota: Incluye los modelos que incluye la categoría de neuronales y categoría de árboles de forma individual. Luego los mejores modelos neuronales y mejores modelos de árboles se incluyen en un modelo de ensamble.

2.2.6. Explicabilidad

La Inteligencia Artificial Explicable (XAI, por sus siglas en inglés) permite realizar una interpretación adecuada del proceso de predicción (Yang et al., 2021), utilizando modelos para analizar la importancia y la dependencia de los factores que contribuyen a explicar el resultado (Nikolaou et al., 2023). De esta manera, se garantiza la confianza y la transparencia en las predicciones, lo que facilita su aplicación en el ámbito de la seguridad vial (Xu et al., 2022).

En el estudio de Lacherre et al. (2024), se identificaron 18 estudios que utilizan 6 métodos para explicar los factores con mayor contribución al RAVCC. De estos, el método de Importancia de Factores de RF y GB fue el más utilizado, con un 50%, seguido de SHAP con un 33%. Estos estudios se enfocan principalmente en explicar el comportamiento del conductor en relación con el riesgo de accidentes, siendo China y Estados Unidos los principales países en los que se aplicaron estos estudios.

2.3. Marco conceptual

GPS: Sistema de Posicionamiento Global (por sus siglas en inglés, Global Positioning System).

EDATA: Accesorio usado en conjunto con el GPS para recibir los datos de telemetría de la computadora del Vehículo.

TELEMETRIA: es la tecnología que permite la recolección y transmisión de datos en tiempo real sobre el estado y desempeño de un vehículo.

CAN BUS: es un protocolo de comunicación que permite evaluar los sistemas electrónicos en un vehículo. Puede evaluar estado de frenos, motor y seguridad y de esta manera tener los datos en tiempo real.

KEYPAD: es un accesorio que se integra con el GPS y tiene un teclado que identifica al conductor mediante el ingreso de su clave.

4G LTE CAT-4: es una tecnología de red para celulares que permite velocidades de descarga hasta 150 Mbps y de subida hasta 50 Mbps. Es ideal para ser usado por aplicaciones que demandan muchos datos.

API: permite que los desarrolladores integren servicios externos sin necesidad de conocer cómo funciona el servicio internamente.

RPM: Se refiere a las revoluciones por minutos y se utiliza para evaluar la velocidad de rotación del motor.

CAPÍTULO III. METODOLOGÍA

3.1. Hipótesis general

El desarrollo de un modelo de predicción basado en aprendizaje automático explica significativamente el riesgo de accidentes vehiculares por conducta de manejo en vehículos pesados del sector minero en carretera rural.

3.2. Hipótesis específicas

- HE1: La ejecución de un estudio naturalista de recolección de datos de telemetría en vehículos pesados del sector minero en carretera rural genera un conjunto de datos adecuado para el desarrollo y validación de un modelo de predicción basado en aprendizaje automático del riesgo de accidentes.
- HE2: El diseño y validación de un modelo ensamblado de predicción que combina de manera efectiva múltiples modelos de aprendizaje automático optimiza la exactitud y sensibilidad del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural en comparación con modelos individuales.
- HE3: La evaluación de la eficiencia predictiva del modelo ajustado en la identificación del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carreteras rurales, con un conjunto de datos independiente, asegura que el modelo tenga altos niveles de exactitud y explicativabilidad.

3.3. Identificación de variables

Las variables son características, atributos o propiedades de una persona, un objeto o una situación que pueden cambiar (adquirir diversos valores), y cuya variación es susceptible de medirse.

Las variables independiente y dependiente que se identificaron en el proyecto son las siguientes:

- Variable Independiente (VI): El desarrollo de un modelo de predicción basado en aprendizaje automático.
- Variable Dependiente (VD): riesgo de accidentes vehiculares por conducta de manejo.

3.4. Operacionalización de variables

A continuación (ver Tabla 4), se presenta la matriz de operacionalización de variables que incluye la definición de la variable y forma de medición. Esta matriz facilita la identificación de los indicadores específicos que permitirán evaluar el impacto de las variables en la investigación, proporcionando un marco claro y estructurado para la validación del modelo predictivo propuesto.

Tabla 4

Matriz de operacionalización de variables

Variable	Definición conceptual	Definición Operacional	Indicadores	Escala de medición
Variable Independiente: Desarrollo de un modelo de predicción	Es el proceso de construcción de un sistema	Desarrollo, entrenamiento y comparación entre diferentes	Modelos de predicción	Escala Numérica (Exactitud, Sensibilidad)

basado en aprendizaje automático	que combina múltiples modelos de aprendizaje automático para predecir la conducta de manejo	modelos de predicción con datos de telemetría y comportamiento del conductor		
	Sarker et al. (2021).			
Variable Dependiente:	Clasificación de riesgo que ocurra un accidente vehicular en base a condiciones de manejo y telemetría	Clasificación del riesgo de accidente en tres niveles (normal, peligroso, riesgoso), con valores numéricos asignados (normal=0, peligroso=1, riesgoso=2).	Influencia de factores para clasificar el riesgo de accidente aplicando SHAP.	Escala numérica (Influencia de factores en el RAVCC)
Riesgo de accidentes vehiculares por conducta de manejo	Silva & Eugenio Naranjo (2020).			

3.5. Matriz de consistencia

En la Tabla 5 se observa la matriz de consistencia con los problemas, objetivos e hipótesis identificadas.

Tabla 5

Matriz de consistencia

Aspecto	General	Específico 1	Específico 2	Específico 3
Problema	¿De qué manera el desarrollo de un modelo de predicción basado en aprendizaje automático permite explicar el riesgo de accidentes vehiculares por conducta de manejo en vehículos pesados del sector minero en carretera rural?	¿Cómo la ejecución de un estudio naturalista de recolección de datos de telemetría en vehículos pesados del sector minero en carretera rural permite generar un conjunto de datos adecuado para el desarrollo y validación de un modelo de predicción basado en aprendizaje automático del riesgo de accidentes?	¿Cómo diseñar y validar un modelo ensamblado de predicción que combine de manera efectiva múltiples modelos de aprendizaje automático para optimizar la exactitud y sensibilidad del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural en comparación con modelos individuales?	¿Cómo la evaluación de la eficiencia predictiva del modelo ajustado en la identificación del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carreteras rurales, con un conjunto de datos independiente, asegura que el modelo tenga altos niveles de exactitud y explicabilidad?
Objetivo	Desarrollar un modelo de predicción basado en aprendizaje automático para explicar el riesgo de	Ejecutar un estudio naturalista de recolección de datos de telemetría en vehículos pesados del	Diseñar y validar un modelo ensamblado de predicción que combine de manera	Evaluar la eficiencia en la predicción del modelo ajustado en la identificación del riesgo de

	accidentes vehiculares por conducta de manejo en vehículos pesados del sector minero en carretera rural.	sector minero en carretera rural, que permita generar un conjunto de datos adecuado para el desarrollo y validación de un modelo de predicción basado en aprendizaje automático del riesgo de accidentes.	efectiva múltiples modelos de aprendizaje automático para optimizar la exactitud y sensibilidad del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural en comparación con modelos individuales.	accidentes vehiculares en vehículos pesados del sector minero en carreteras rurales, con un conjunto de datos independiente, asegura que el modelo tenga niveles altos de exactitud y explicabilidad.
Hipótesis	El desarrollo de un modelo de predicción basado en aprendizaje automático explica significativamente el riesgo de accidentes vehiculares por conducta de manejo en vehículos pesados del sector minero en carretera rural.	La ejecución de un estudio naturalista de recolección de datos de telemetría en vehículos pesados del sector minero en carretera rural genera un conjunto de datos adecuado para el desarrollo y validación de un modelo de predicción basado en aprendizaje automático del riesgo de accidentes.	El diseño y validación de un modelo ensamblado de predicción que combina de manera efectiva múltiples modelos de aprendizaje automático optimiza la exactitud y sensibilidad del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural en comparación con modelos individuales.	La evaluación de la eficiencia predictiva del modelo ajustado en la identificación del riesgo de accidentes vehiculares en vehículos pesados del sector minero en carreteras rurales, con un conjunto de datos independiente, asegura que el modelo tenga altos niveles de exactitud y explicabilidad.
Variables	Variable independiente: Desarrollo de un modelo de	Variable independiente: Estudio naturalista de recolección de datos.	Variable independiente:	Variable independiente: Evaluación rigurosa del modelo ajustado.

	predicción basado en aprendizaje automático. Variable dependiente: Riesgo de accidentes vehiculares por conducta de manejo.	Variable dependiente: Conjunto de datos generado para el modelo predictivo.	Modelo ensamblado de predicción. Variable dependiente: Exactitud y sensibilidad del modelo predictivo.	Variable dependiente: Exactitud, sensibilidad y capacidad explicativa del modelo.
Técnicas de recolección de datos	Recolección de datos de telemetría y datos históricos de eventos de riesgo de accidentes para alimentar el modelo de aprendizaje automático.	Instalación de dispositivos de telemetría en vehículos pesados del sector minero. Recolección de datos del comportamiento de conducción en carretera rural.	Recopilar datos sobre el rendimiento de modelos individuales (árboles y neuronales.) que se utilizarán para comparar con el modelo ensamblado. Medición de exactitud y sensibilidad en la predicción del riesgo de accidentes.	División de Datos: Utilizar un conjunto de datos que no haya sido utilizado para el entrenamiento del modelo, asegurando que la evaluación sea objetiva. Desempeño de Modelos Previos: Obtener datos sobre el rendimiento de otros modelos que se están utilizando para comparación, incluyendo sus métricas de exactitud y sensibilidad.

3.6. Tipo y diseño de investigación

El tipo de investigación que se realiza en el estudio es una investigación aplicada con un enfoque cuantitativo y de tipo no experimental.

Es *investigación Aplicada* debido a que el objetivo de la investigación es desarrollar un modelo de predicción basado en aprendizaje automático para explicar y mejorar la capacidad de predicción del riesgo de accidentes vehiculares. Se trata de una investigación aplicada porque busca una solución práctica para un problema específico en un entorno real (vehículos pesados en carreteras rurales del sector minero).

Tiene un *enfoque cuantitativo* porque se utilizan datos numéricos recolectados a través de telemetría y otras mediciones para desarrollar y evaluar modelos predictivos. Las técnicas de recolección de datos incluyen pruebas experimentales con modelos de predicción y análisis estadístico (como el uso de ANOVA) para evaluar la exactitud y sensibilidad del modelo. Esto implica un análisis cuantitativo riguroso para validar los resultados.

El tipo de investigación es *No Experimental* porqué, aunque se está desarrollando y evaluando un modelo predictivo, las variables no se manipulan directamente. En lugar de intervenir activamente en el comportamiento de los conductores o modificar las condiciones de la carretera, se recolectan datos a través de un estudio naturalista (observación en condiciones reales sin intervención). Esto hace que el diseño sea no experimental, ya que no se realiza manipulación directa de las variables independientes.

3.7. Unidad de análisis

Se define como el objeto sobre el cual se recopilarán los datos para abordar el problema y evaluar las hipótesis planteadas (Hernández et al., 2014). Por lo que se define la unidad de análisis como los *conductores de vehículos pesados del sector minero en carreteras rurales*.

Estos conductores son observados y evaluados en cuanto a su comportamiento de manejo (por ejemplo, velocidad, frenado, aceleraciones bruscas) a través de datos de telemetría. Estos datos serán utilizados para generar un conjunto de información que permitirá el desarrollo y validación de un modelo predictivo del riesgo de accidentes.

Por tanto, los conductores y los datos asociados a sus vehículos representan las unidades de análisis principales, ya que son los "sujetos" sobre los cuales se recolecta información para analizar el riesgo de accidentes.

3.8. Población de estudio

Se identifica como la población de estudio los *conductores de vehículos pesados del sector minero en carreteras rurales de la zona sur del Perú*.

Esta población incluye a todos los conductores que operan vehículos pesados en rutas mineras rurales en esta región geográfica específica. Las conclusiones del estudio serán válidas para este conjunto de conductores y el entorno operacional del sector minero en dicha área.

3.9. Tamaño de muestra

Para realizar el estudio se usará una estrategia no probabilística con 5 conductores que se encargan de generar datos de conducta de manejo antes de aplicar el modelo de aprendizaje automático propuesto.

3.10. Selección de muestra

Se realizará un muestreo intencional donde el investigador, en conjunto con un representante de la organización participante, seleccionará basados en el juicio experto a cada conductor.

La muestra se ha conformado por 5 conductores profesionales que trabajan de forma diaria para la organización dueña de los vehículos. La selección se fundamenta en los siguientes aspectos:

- **Experiencia y Formalidad:** Los conductores son formales, poseen licencia activa acorde al tipo de vehículo que manejan y cuentan con la experiencia necesaria para operar en las rutas del corredor minero sur. Esta experiencia garantiza que el comportamiento registrado es representativo de las condiciones reales de manejo en la zona.
- **Representatividad del Contexto de Estudio:** Aunque la muestra incluye 5 conductores, cada uno realiza múltiples viajes durante el período de recolección de datos (7 días). Esto permite capturar variaciones en el comportamiento tanto en trayectos con carga (Ruta 1) como sin carga (Ruta 2), lo que enriquece el conjunto de datos para la validación del modelo.
- **Amplitud de Datos:** La realización de dos trayectos diarios (ida y regreso) por cada conductor genera, un número sustancial de viajes monitoreados (aproximadamente 30 viajes en total), lo que favorece la robustez y validez estadística del modelo de predicción.

A continuación, se incluye una tabla resumen de la muestra (ver Tabla 6):

Tabla 6

Perfil de la Muestra y Detalles de la Recolección de Datos

Característica	Detalle
Número de conductores	5
Perfil de conductores	Conductores formales, con licencia activa y experiencia en el manejo de vehículos pesados del sector minero.
Área de Estudio	Corredor minero sur, tramo específico entre Matarani (Provincia de Ilay) y Uchumayo (Provincia de Arequipa)
Rutas Monitoreadas	- Ruta 1: Desde Uchumayo a Matarani (107 km, vehículo cargado).

	- Ruta 2: Desde Matarani a Uchumayo (120 km, vehículo vacío).
Periodo de Recolección de Datos	Del 30 de agosto de 2023 al 5 de septiembre de 2023 (7 días).
Número Estimado de Viajes	Aproximadamente 30 viajes.

3.11. Técnicas de recolección de datos

Para este proyecto, las técnicas de recolección de datos serán las presentadas en la Tabla 7:

Tabla 7

Técnicas de recolección de datos del RAVCC

Técnica	Descripción	Lugar/Condiciones
a. Observación directa mediante telemetría	Se instalarán dispositivos de telemetría en los vehículos pesados de los conductores participantes. Estos dispositivos recolectarán datos en tiempo real sobre variables de conducta de manejo, como la velocidad, frenado, y aceleraciones bruscas.	Carreteras rurales del sector minero en la zona sur del Perú, bajo condiciones operativas reales.
b. Recolección de datos históricos	Se recolectarán datos históricos de eventos de riesgos de accidentes vehiculares para complementar los datos de telemetría y proporcionar contexto adicional para el modelo predictivo.	Análisis de registros de eventos de riesgos de accidentes del sector minero.
c. Pruebas con modelos de aprendizaje automático	Los datos recolectados serán usados para evaluar varios modelos de aprendizaje automático en un entorno controlado, con el fin de determinar cuál tiene mejor rendimiento en términos de exactitud y sensibilidad para predecir el riesgo de accidentes.	Entorno controlado para la ejecución de pruebas con los datos recolectados.

d. Evaluación de la efectividad del modelo	Se recolectarán datos después de aplicar los modelos predictivos para comparar su eficacia en la identificación del riesgo de accidentes. Se medirán métricas como exactitud y sensibilidad.	Evaluación posterior a la implementación de los modelos predictivos utilizando datos reales y simulados.
---	--	--

3.12. Análisis e interpretación de información

Dentro de las técnicas analíticas para comprobar las hipótesis se realizarán las que se presentan en la Tabla 8:

Tabla 8

Técnicas a usar para validación de hipótesis y realizar las conclusiones

Hipótesis	Proceso de Clasificación, Registro y Codificación	Técnicas Analíticas para Comprobar Hipótesis
Hipótesis general	a) Clasificación: Los datos de telemetría serán clasificados en tres categorías de riesgo: normal, riesgoso y peligroso. b) Registro: Los datos serán recolectados mediante sensores de telemetría instalados en los vehículos pesados. El conjunto de datos incluirá variables como velocidad, frenado brusco, aceleración, entre otras. Estos datos serán registrados automáticamente en una base de datos estructurada. c) Codificación: Las variables dependientes (riesgo de accidente) se codificarán como: 0 (normal), 1 (peligroso), y 2 (riesgoso), mientras que las variables predictoras serán codificadas en función de los parámetros de conducción.	SHAP (Shapley Additive Explanations): Esta técnica interpreta el impacto de las variables predictoras y muestra cómo los factores contribuyen en la clasificación del riesgo normal, peligroso o riesgoso.
Hipótesis Específica 1	a) Clasificación: Los datos de telemetría se clasificarán en diversas categorías	a) Regresión Logística: Permitirá modelar la relación entre los

relacionadas con la conducta de conducción (frenado, aceleración, etc.).

b) Registro: Los datos son recolectados en tiempo real usando sensores instalados en el vehículo. Los datos se registran y almacenan en una base de datos que luego serán usados por el modelo.

c) Codificación: Cada comportamiento de conducción tendrá un código.

comportamientos de conducción y el riesgo de accidentes.

b) Análisis de

Coeficientes: Evalúa el impacto de la variable y la calidad en el ajuste del modelo.

c) Importancia de Características con XGBoost y Random

Forest: Identifica las variables más importantes que contribuyen a predecir el riesgo.

Hipótesis Específica 2

a) Clasificación: Los datos se clasifican en subconjuntos para entrenar, validar y probar los modelos ensamblados.

b) Registro: Los resultados de cada modelo se registran y comparan para identificar qué modelo o ensamble de modelos tiene el mejor rendimiento.

c) Codificación: Los resultados de cada modelo se codifican de acuerdo a sus métricas de exactitud y sensibilidad.

a) Prueba t de Student:

Compara las medias de exactitud entre el modelo ensamblado y los modelos individuales para verificar si existen diferencias significativas.

b) Prueba de Wilcoxon:

Compara el desempeño del modelo ensamblado con los modelos individuales.

c) Análisis de Varianza (ANOVA):

Compara el desempeño del modelo ensamblado con otros modelos individuales.

a) Evaluación de

Métricas (Exactitud, Sensibilidad): Evalúa exactitud y sensibilidad.

b) Aplicación de SHAP:

Se entiende la contribución de cada variable a la predicción del riesgo.

Hipótesis Específica 3

a) Clasificación: Se dividen los datos en conjuntos de entrenamiento y prueba. Luego, el conjunto independiente se usa para la evaluación del modelo ajustado.

b) Registro: Se registran las predicciones del modelo ajustado y evalúan las métricas.

c) Codificación: Los resultados son codificados usando exactitud, sensibilidad y explicabilidad.

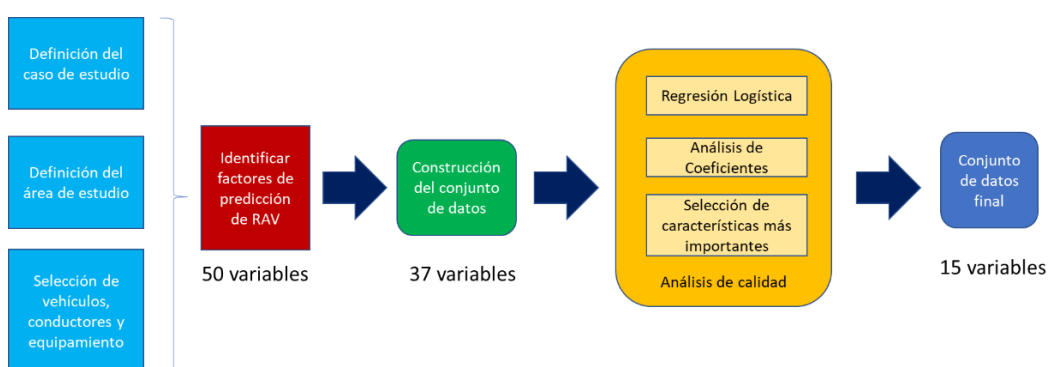
CAPÍTULO IV. DESARROLLO DE LA PROPUESTA

4.1. Definir e implementar un estudio naturalista de recolección de datos de telemetría

En esta sección se establece el contexto de cómo se llevará a cabo la investigación (ver Figura 4), y se pone en práctica el plan para observar y analizar el comportamiento de los conductores en la vida real. Primero se define el caso de estudio y el área de estudio. Luego, se selecciona los conductores, vehículos y equipamiento, asegurando que se generen datos fiables y representativos. Seguidamente, se lleva a cabo la identificación de factores de predicción del RAVCC que sirven para la construcción del conjunto de datos a partir de la recopilación de datos que se genera con los dispositivos de telemetría, para luego realizar el análisis de calidad. Finalmente, se elabora el conjunto de datos final, que se usa en el modelado y análisis posterior.

Figura 4

Contexto de la propuesta



Nota: Incluye la definición del estudio, 50 factores iniciales, 37 factores después de validar datos, y 15 factores más importantes para el conjunto de datos final.

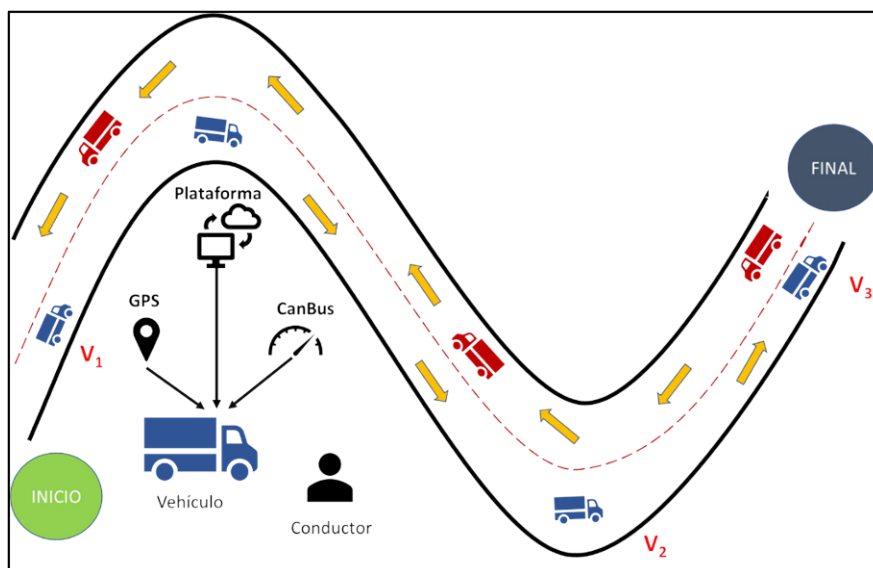
4.1.1. Definición del caso de estudio: Corredor minero sur.

En los últimos años se viene investigando sobre la conducción automatizada, sin embargo, aún es necesario medir algunos factores que influyen en los accidentes de tránsito.

En la Figura 5, se representa el escenario de investigación, donde se lleva a cabo el estudio en vehículos pesados. Estos vehículos son operados por conductores profesionales que trabajan para una entidad encargada de los viajes en este corredor. Cada vehículo está equipado con un equipo GPS y un dispositivo lector EDATA para recopilar datos de telemetría de acuerdo al protocolo Can Bus relacionados con el comportamiento del conductor. Estos datos se recolectan a lo largo de todo el trayecto de ida y regreso; y se transmiten a una plataforma de seguimiento que integra información sobre la gestión del viaje.

Figura 5

Escenario de investigación



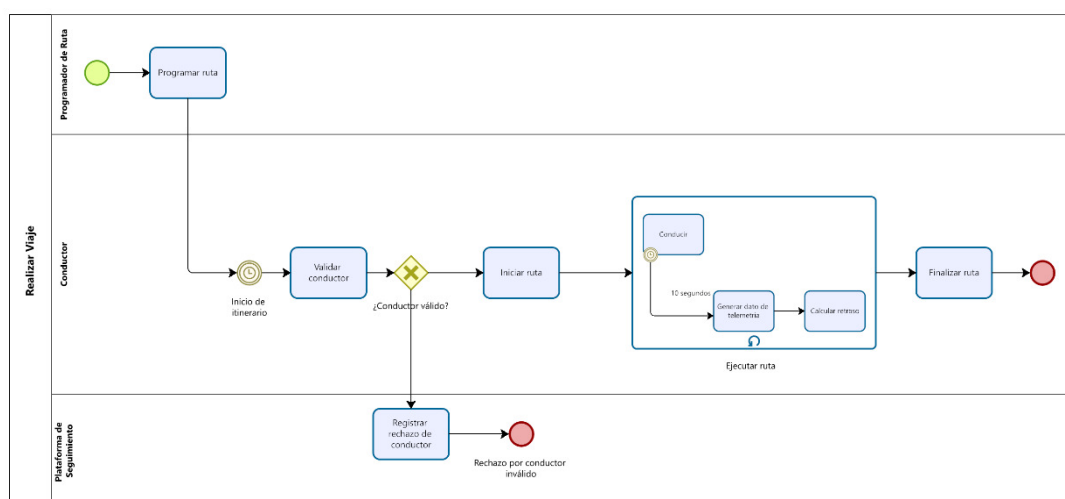
Nota: Incluye el vehículo, conductor, dispositivos, ruta de viaje y plataforma

Basado en el escenario descrito, se lleva a cabo una serie de actividades detalladas en la Figura 6. En primer lugar, se programa la ruta del vehículo, definiendo el camino y estableciendo la fecha de inicio y llegada estimada. Al momento de iniciar el viaje, el conductor ingresa su clave en el teclado de validación para su identificación. Si el conductor no es quien fue

asignado, se cancela el viaje; en caso contrario, se procede con la ejecución del mismo. Durante el trayecto, se recopilan datos de telemetría a una frecuencia de una posición por cada 10 segundos de forma regular y una posición adicional por cada evento generado relacionado al comportamiento del conductor, para luego registrarlas con otros datos de la gestión del viaje. Este proceso se realiza de manera continua hasta que el vehículo alcance su destino y finalice el viaje.

Figura 6

Proceso de viaje



4.1.2. Definición del área de estudio

El estudio se lleva a cabo en las carreteras del corredor minero sur, con énfasis en un tramo específico ubicado en el departamento de Arequipa. Este tramo abarca desde Matarani, perteneciente a la provincia de Islay, hasta Uchumayo, en la provincia de Arequipa.

La recolección de datos se efectúa en dos trayectos distintos: uno de ida, con el vehículo cargado, y otro de regreso, cuando el vehículo está vacío. Las rutas analizadas son las siguientes: La Ruta 1 cubre desde la provincia de Uchumayo hasta Matarani en el departamento de Arequipa, abarcando una distancia de 107 km, y se completa en aproximadamente 5 horas sin paradas (ver Figura 7). Este recorrido se realiza principalmente con el vehículo

cargado. La Ruta 2, por su parte, va desde la provincia de Matarani hasta Uchumayo en el departamento de Arequipa, con una distancia de 120 km y un tiempo de recorrido promedio de 6 horas sin paradas (ver Figura 8). Este trayecto se lleva a cabo principalmente cuando el vehículo está vacío.

Figura 7

Ruta 1: Uchumayo – Matarani

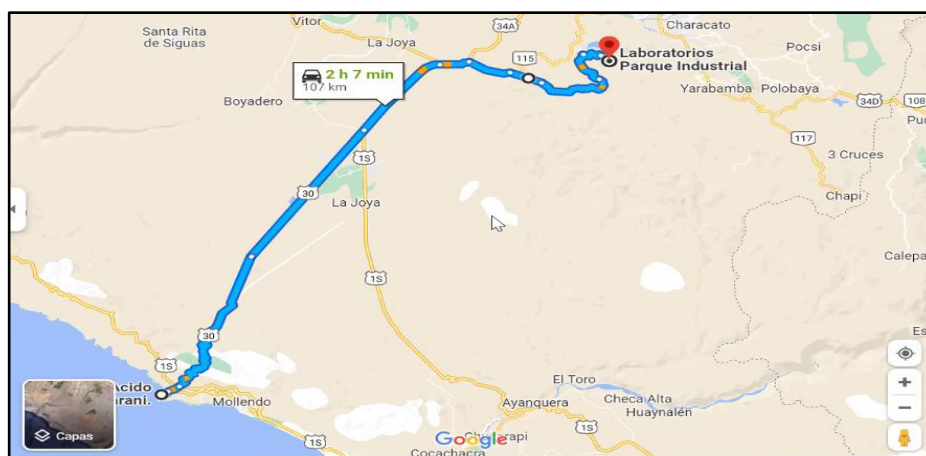
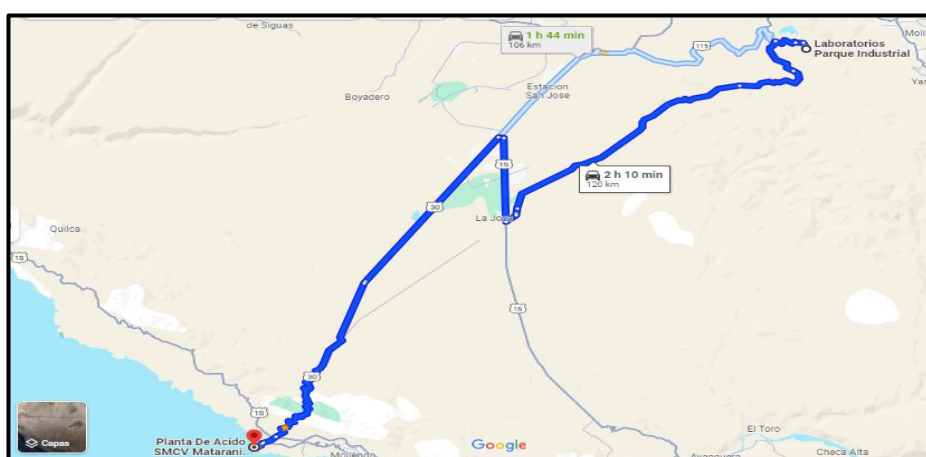


Figura 8

Ruta 2: Matarani – Uchumayo



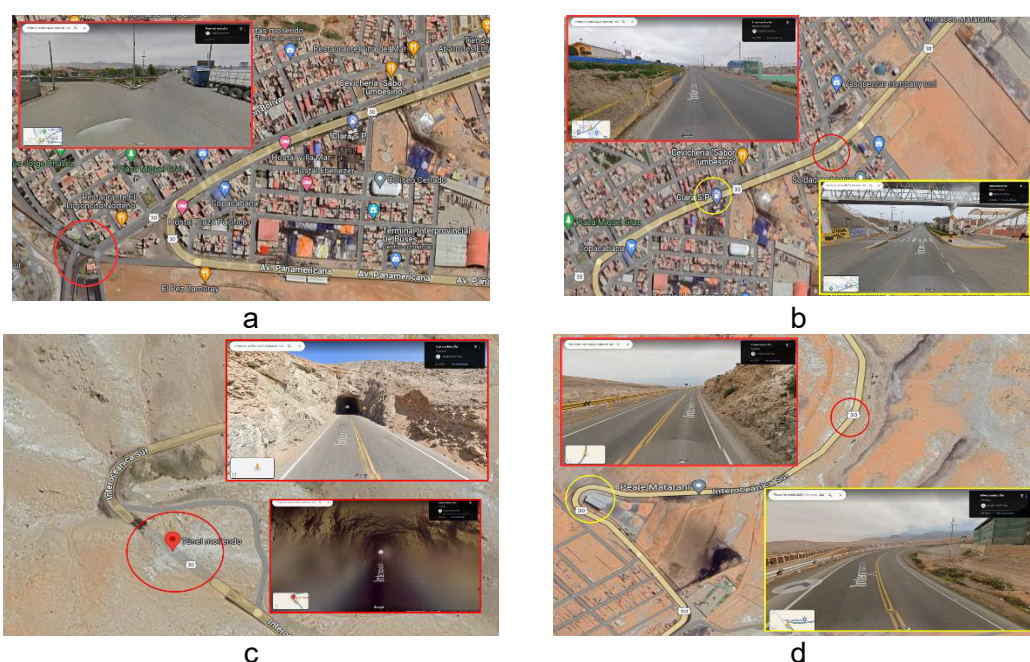
La elección de esta área para llevar a cabo el estudio cobra relevancia significativa debido a que la ciudad de Arequipa representa el segundo punto

con mayor incidencia de accidentes viales en Perú. En el año 2022, se registraron 5,017 siniestros de tránsito en esta localidad, con un saldo lamentable de 2,676 personas lesionadas y 191 vidas perdidas. Es destacable que un 63.2% de estos accidentes ocurrieron en carreteras, y un 55.24% tuvieron lugar en zonas rurales. Dentro de estos incidentes, 53 vehículos pesados se vieron involucrados, distribuidos en 40 casos de transporte de carga y 13 de transporte de personas (Observatorio Nacional de Seguridad Vial, 2024).

En la Figura 9 se puede observar, a manera de ejemplo, los diferentes tipos de vías y calidad del camino que se presentan en el recorrido que hacen los vehículos en ambas rutas.

Figura 9

Ejemplos de caminos en la ruta.



Nota: (a) vía sin línea de división de carril, (b) vía con línea de división de carril, (c) vía con túnel y (d) vías con curvas.

4.1.3. Selección de vehículos, conductores y equipamiento

Los vehículos considerados en el presente estudio son de tipo pesado de marca Internacional y modelo 9200 (ver Figura 10). Son remolcadores capaces de transportar carga de todo tipo, ya sea carga cerrada (contenedores) o plataformas, seca o húmeda (cisternas). Para el escenario de estudio se consideran 4 vehículos que realizan el recorrido seleccionado de forma constante.

Figura 10

Vehículo pesado Internacional 9200



Los conductores formales considerados en el presente estudio son 5 que realizan trabajos de conducción diariamente en la ruta en estudio y trabajan para la organización dueña de los vehículos. Los conductores poseen licencia de conducir activa y acorde al tipo de vehículo de manejo.

Hoy en día, los vehículos tienen la capacidad de transmitir diversos datos que permiten identificar el comportamiento del conductor. Estos datos de telemetría son generados por varios dispositivos y, mediante comunicación celular, se pueden enviar a plataformas para su utilización en distintos propósitos. En el estudio realizado, se emplean dispositivos como GPS, EDATA y KEYPAD para recopilar los datos de telemetría. La instalación de

estos dispositivos se realiza estratégicamente para fomentar una conducción segura y apropiada por parte del conductor.

El GPS es un dispositivo de rastreo y telemática diseñado para satisfacer las necesidades de seguimiento de vehículos, tanto privados como comerciales. Incorpora una antena celular y GPS, y utiliza tecnología celular 4G LTE CAT-4 para la transmisión de datos. Este dispositivo incluye internamente un módulo denominado ESAFE, encargado de supervisar la conducción. Este módulo analiza datos de acelerómetros tridimensionales para detectar eventos como aceleración brusca, frenado súbito, cambio de carril, giros bruscos y detección de accidentes.

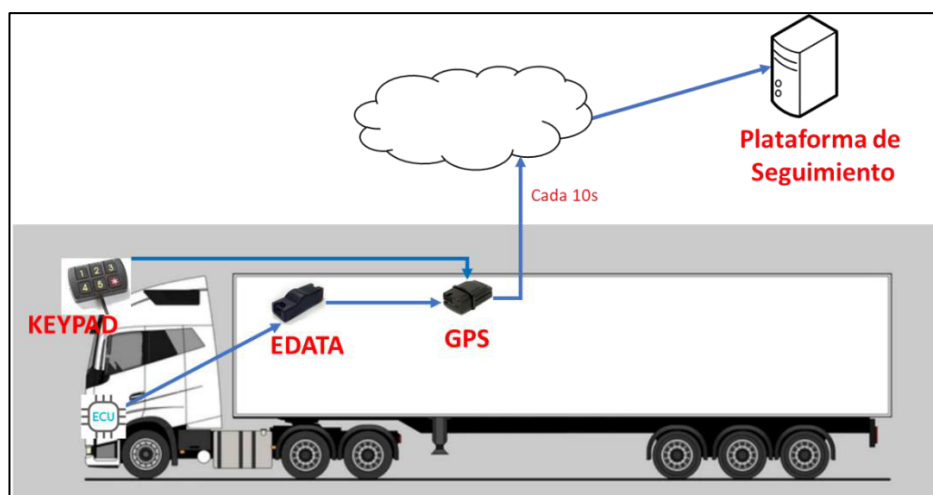
EDATA es un dispositivo que transmite el análisis en tiempo real sobre el estado del vehículo obtenido de la computadora del vehículo usando el protocolo Can Bus.

El KEYPAD es un accesorio que consta de 6 botones y permite que uno o más conductores autorizados se identifiquen cuando encienden el vehículo. Para poner en marcha el vehículo, el conductor debe ingresar su código correspondiente a través del KEYPAD. En caso de que el código no sea válido, se registra un evento de conductor inválido.

En la Figura 11 se representa el flujo de transmisión de datos, comenzando con la inserción del código del conductor. Si este código es válido, se envía un evento de acceso permitido a la plataforma de seguimiento mediante el equipo GPS, el cual dispone de un módulo para transmitir datos vía señal celular. Una vez que el vehículo está en marcha, se transmiten datos provenientes del EDATA y GPS a la plataforma de seguimiento, la cual integra estos datos de gestión a medida que el conductor avanza en su recorrido.

Figura 11

Arquitectura Tecnológica propuesta



4.1.4. Identificar factores de predicción de RAV por conducta de manejo

De acuerdo al estudio de Lacherre et al. (2024), se identificaron 115 factores agrupados en 5 dominios: Entorno, Tránsito, Vehículo, Conductor y Gestión. En este estudio se usan 50 factores extraídos de diferentes fuentes de datos, como la plataforma de gestión y las APIs de mapas de clima (Open-Meteo.com, 2024). En la Tabla 9 se describen los parámetros seleccionados a ser usados en este estudio:

Tabla 9

Factores de predicción de RAV por conducta de manejo

N°	Parámetro	Descripción	Domini nio	Tipo Variable	Diccionario
1	Latitud	Valores de coordenadas	Vehículo	NC	Posición de la ruta
2	Longitud	Valores de coordenadas	Vehículo	NC	Posición de la ruta
3	Altitud	Metros sobre el nivel del mar	Vehículo	NC	1 a 4800 (metros)
4	Velocidad	Velocidad del equipo GPS Km/h	Vehículo	NC	0 a 100 (km/h)
5	FechaInicio	Fecha de la última buena posición	Vehículo	NC	2023-08-30 al 2023-09-05
6	Radio	Orientación, rumbo actual	Vehículo	NC	0° a 359°

		relativo al norte verdadero			
7	EstadoIgnición	Encendido o apagado del vehículo	Vehículo	CB	0: Vehículo IGN apagado 1: Vehículo IGN encendido
8	EstadoValidación	Calificación de la plataforma para definir la posición	Gestión	CB	0: Posición no válida 1: Posición válida
9	Distancia	Distancia expresada en metros con respecto a la posición anterior	Gestión	NC	0 a 11.54 (Km)
10	TipoLocalización	Clasificación de la plataforma para las posiciones	Gestión	CO	0: Normal 1: Parada 2: Evento 3: Alarma 4: Exceso Velocidad
11	Horometro	Intervalo de tiempo que estuvo encendido el vehículo	Vehículo	NC	37,107,500 a 109,528,100 (seg)
12	EstadoManejo	Vehículo en movimiento o detenido	Vehículo	CB	0: No conduciendo 1: Conduciendo
13	TiempoDetenido	Acumulado de tiempo en minutos del vehículo detenido	Vehículo	ND	0 a 8 (minutos)
14	Total de Horas de Operación del Motor	Indica las horas que el motor estuvo operando	Vehículo	ND	8036 a 80388 (horas)
15	Activación del freno	Indica si el conductor pisó el pedal de freno	Conductor	CB	0: Freno sin pisar 1: Frenando
16	Temperatura del Refrigerante del Motor	Indica la temperatura del refrigerante del vehículo	Vehículo	ND	0° a 105° (Celcius)
17	Presión del Aceite del Motor	Indica la presión del aceite del motor	Vehículo	ND	0 a 364 (Kilopascal)

18	Velocidad del Motor	Indica las revoluciones del motor por minuto	Vehículo	ND	0 a 2252 (RPM)
19	Nivel de Combustible	Porcentaje de nivel del tanque de combustible	Vehículo	ND	0 a 100 (%)
20	Distancia Total del Vehículo en Alta Resolución	Acumulado de los kilómetros recorridos	Vehículo	ND	1 a 5000
21	Velocidad del Vehículo Basada en las Ruedas	Indica la velocidad basada en la rueda	Vehículo	ND	0 a 71 (Km)
22	Boleo de Motor	Cuando el vehículo está en movimiento sin marcha puesta (neutro)	Vehículo	CD	0: Evento no generado 1: Evento Generado
23	Sobrerrevoluciones	Indica si el motor tiene un RPM no adecuado o anormal	Vehículo	CD	0: Evento no generado 1: Evento Generado
24	Accidente	Si el vehículo tiene un choque o daño	Tránsito	CD	0: Evento no generado 1: Evento Generado
25	Aceleración	Conductor realiza aceleración brusca	Conductor	CD	0: Evento no generado 1: Evento Generado
26	Frenada	Conductor realiza un frenado brusco	Conductor	CD	0: Evento no generado 1: Evento Generado
27	Ingreso a óvalo	El vehículo entra a una vía con curvas prolongadas	Conductor	CD	0: Evento no generado 1: Evento Generado
28	Cambio de carril	El vehículo se desplaza de derecha a izquierda o viceversa	Vehículo	CD	0: Evento no generado 1: Evento Generado
29	Adelantamiento	Vehículo acelera y cambia de carril para realizar un adelantamiento	Vehículo	CD	0: Evento no generado 1: Evento Generado

30	Rompe Muelles	Cuando el vehículo pasa por encima de una superficie elevada	Vehículo	CD	0: Evento no generado 1: Evento Generado
31	Inicio Trocha	Cuando el vehículo inicia su recorrido en pista no asfaltada	Tránsito	CD	0: Evento no generado 1: Evento Generado
32	Fin Trocha	Cuando el vehículo finaliza su recorrida en pista no asfaltada	Tránsito	CD	0: Evento no generado 1: Evento Generado
33	Giros bruscos a la izquierda	Conductor realiza un giro brusco a la izquierda	Conductor	CD	0: Evento no generado 1: Evento Generado
34	Giros bruscos frenando a la izquierda	Conductor realiza un giro brusco frenando a la izquierda	Conductor	CD	0: Evento no generado 1: Evento Generado
35	Giros bruscos acelerando a la izquierda	Conductor realiza un giro brusco acelerando a la izquierda	Conductor	CD	0: Evento no generado 1: Evento Generado
36	Giros bruscos a la derecha	Conductor realiza un giro brusco a la derecha	Conductor	CD	0: Evento no generado 1: Evento Generado
37	Giros bruscos frenando a la derecha	Conductor realiza un giro brusco frenando a la derecha	Conductor	CD	0: Evento no generado 1: Evento Generado
38	Giros bruscos acelerando a la derecha	Conductor realiza un giro brusco acelerando a la derecha	Conductor	CD	0: Evento no generado 1: Evento Generado
39	Rango Horario	Indica si se está conduciendo de día o noche	Entorno	CB	0: Día (6:00 a 18:00) 1: Noche (18:00 a 6:00)
40	Clima	Indica el clima en tiempo real para una posición dada	Entorno	CO	0..99 Código de tipo de clima
41	Antigüedad	Antigüedad del vehículo	Vehículo	ND	5 años
42	Cargado	Vehículo con carga	Gestión	CB	0: Vehículo sin carga

		(materiales) y sin carga			
43	Tipo Via	Indica el tipo de vía por la que se circula	Entorno	CN	1: Vehículo con carga 0: Urbano 1: Carretera 2: Trocha
44	Límites de Velocidad	Límites de velocidad existentes en una ruta determinada	Entorno	ND	30km-40km-80km
45	Horas de Manejo	Horas conducidas de manera continua sin descanso	Conductor	ND	N° horas
46	Licencia	Verificar si la licencia esta vencida	Conductor	CB	0: Licencia vencida 1: Licencia no vencida
47	Multas	Verificar la cantidad de multas por conductor o vehículo	Conductor	ND	N° multas
48	Desvíos de rutas	Si el chofer toma una ruta alterna	Entorno	CB	0: Conductor fuera de ruta 1: Conductor en la ruta
49	Tiempo Atraso (EDT)	Tiempo excedente por cada segmento	Gestión	ND	N° minutos
50	Posición en curva	Si el vehículo se encuentra en una curva	Entorno	CB	0: Vehículo recto 1: Vehículo en curva

Nota: Tipo Variable (NC: Numérico-Continuo, CB: Categórico-Binario, ND:

numérico-Discreto, CO: Categórico-Ordinal)

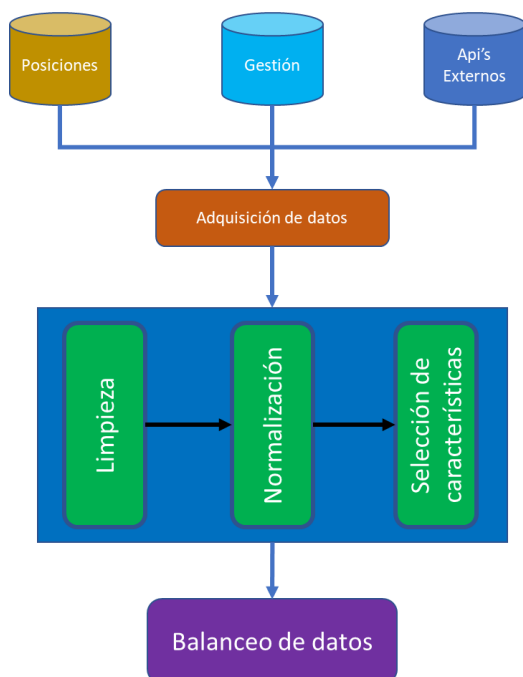
4.1.5. Construcción del conjunto de datos

En esta fase se llevan a cabo una serie de pasos para obtener los datos del vehículo y almacenarlos en un conjunto de datos. A continuación, se efectúa un análisis exploratorio y una depuración de los datos. Posteriormente, se procede a la estandarización para asegurar que todos los datos se encuentren en una misma escala, seguido de la selección de las

características más relevantes. Las actividades correspondientes al preprocesamiento de datos se detallan en la Figura 12.

Figura 12

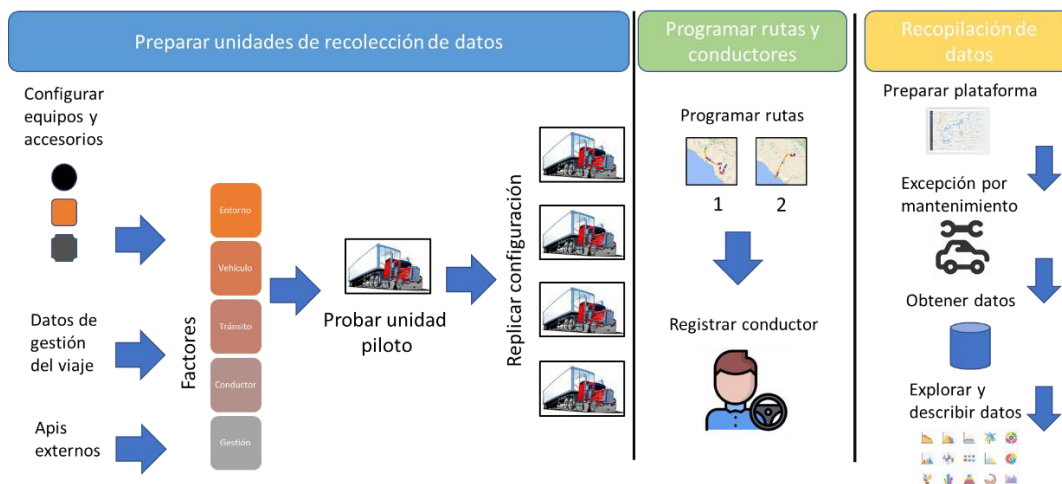
Actividades para el preprocesamiento de datos



Adquisición de datos. Para elaborar el conjunto de datos se sigue una secuencia de pasos que consiste en configurar los equipos a usar en las unidades, programar rutas de viaje y conductores, y finalmente la recolección de los datos (ver Figura 13).

Figura 13

Secuencia de pasos para elaborar el conjunto de datos



Nota: Preparación de las unidades con los equipos y accesorios de telemetría, programación de rutas y conductores y recopilación de los datos.

A continuación, se describen los pasos para la elaboración del conjunto de datos:

1. Preparación de las unidades de recolección de datos: Identificación de los equipos y accesorios a configurar e instalar en los vehículos y realizar la integración con servicios de gestión.
 - Configuración de equipos y accesorios: Se requiere configurarlos para almacenar y/o reportar datos en intervalos temporales mediante una señal celular 4G/3G/2G. Esto incluye datos de GPS (fecha, ubicación y velocidad), eSafe (giros, aceleraciones y frenadas), eDATA (velocidad y odómetro), y KeyPad (identificación del conductor).
 - Datos de gestión de viaje: Para recabar información del viaje, las unidades equipadas con GPS y sus accesorios deben ser registradas en la plataforma de seguimiento. Una vez registradas, cada vez que se reciba la ubicación de un vehículo, se debe almacenar la información de la ruta, que incluya detalles como el tipo de vehículo, área de circulación, límite de velocidad, tiempo de retraso al destino y si el vehículo viaja cargado.
 - Pruebas de unidades piloto: Una vez configurados los equipos y accesorios, se llevan a cabo pruebas en un vehículo para garantizar el correcto funcionamiento de las configuraciones. Se verifica la recepción

y posición de la señal, la autenticación del conductor, la detección de giros y la captura de datos de gestión.

- Estandarización de configuración en unidades: Luego de las pruebas exitosas con la unidad piloto se debe replicar la configuración a realizar en todas las unidades del grupo de estudio.

2. Programación de rutas y conductores

- Se realiza la programación de la ruta 1 y ruta 2 para los distintos viajes a realizar diariamente por cada una de las unidades que se encuentran consideradas dentro del estudio.
- Registro de conductores: Se registra los conductores y su identificación para que puedan autenticarse al encender la unidad y se identifique los viajes que realiza cada conductor.

3. Recopilación de datos por rutas

- Preparar ambiente para recolección de datos: Se tiene controles en la plataforma, base de datos, vehículos y plan de datos para asegurar que los datos se están recolectando sin pérdida de datos.
- Excepción de recolección por mantenimiento de unidades: En caso se presente algún inconveniente con algún vehículo por posible mantenimiento o accidente. Este debe registrar la fecha de inicio y fecha de término de mantenimiento para excluir los datos al momento de generar el análisis estadístico.
- Obtención de datos por cada ruta por un periodo de tiempo: En la fecha de inicio de la prueba se procede a activar los vehículos para recolectar la información. Los datos del estudio naturalista se recolectan en formato de serie temporal y se registran de acuerdo a una frecuencia por el conjunto de dispositivos donde cada vehículo tiene una identificación única y, por lo tanto, se pueden extraer su trayectoria y los eventos asociados al comportamiento del conductor.
- Descripción de datos recopilados: Se describe los datos para entender el comportamiento de los conductores y vehículos en las rutas. Esto sirve como punto de partida para encontrar los valores media, desviación estándar, mínimo, máximo y otros (ver Figura 14). Adicionalmente, se hace uso de gráficos de cajas e histograma para

entender los datos medios y límites de valores de las variables secuenciales (ver Figura 15) y de las variables categóricas (ver Figura 16).

Figura 14

Descripción de variables

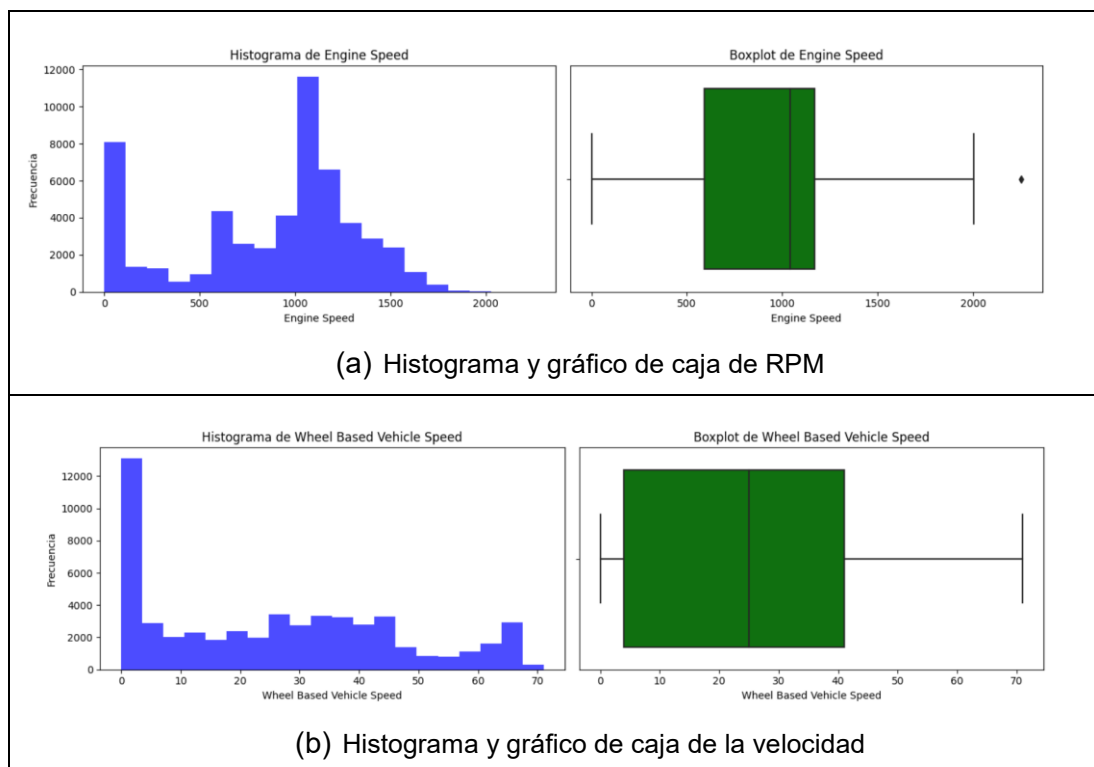
	he.Altitud	h.Radio	KM Max	Engine Coolant Temperature	Engine Oil Pressure	Engine Speed
count	57830.000000	57830.000000	56853.000000	18039.000000	18103.000000	18184.000000
mean	1509.120370	169.614612	49.547253	80.048894	215.046567	995.476848
std	1023.006159	109.328783	21.159465	11.865087	73.029290	386.378597
min	23.000000	0.000000	30.000000	20.000000	0.000000	0.000000
25%	461.000000	71.000000	30.000000	82.000000	204.000000	686.000000
50%	1365.000000	183.000000	40.000000	83.000000	248.000000	1078.000000
75%	2638.000000	266.000000	80.000000	85.000000	252.000000	1236.000000
max	2849.000000	359.000000	80.000000	105.000000	364.000000	2252.000000

Nota: Se incluyen algunas variables como altitud, radio, velocidad máxima, temperatura del refrigerante, presión del aceite del motor y RPM.

En la Figura 14 se muestra el análisis exploratorio de algunas variables como la cantidad de registros, la media, la desviación estándar y los valores por cada cuartil. Por ejemplo, existe un máximo de 57,830 registros donde la altitud media es de 1,509 msnm, una velocidad máxima de 80 km/h, una temperatura promedio del motor de 80 y un RPM promedio de 995.

Figura 15

Descripción de variables secuenciales

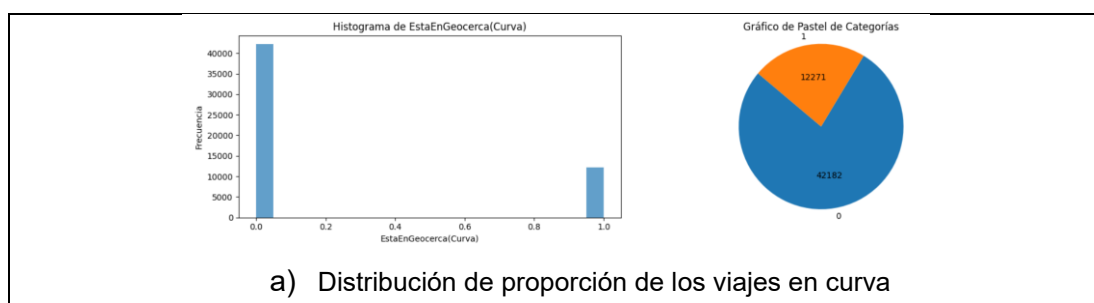


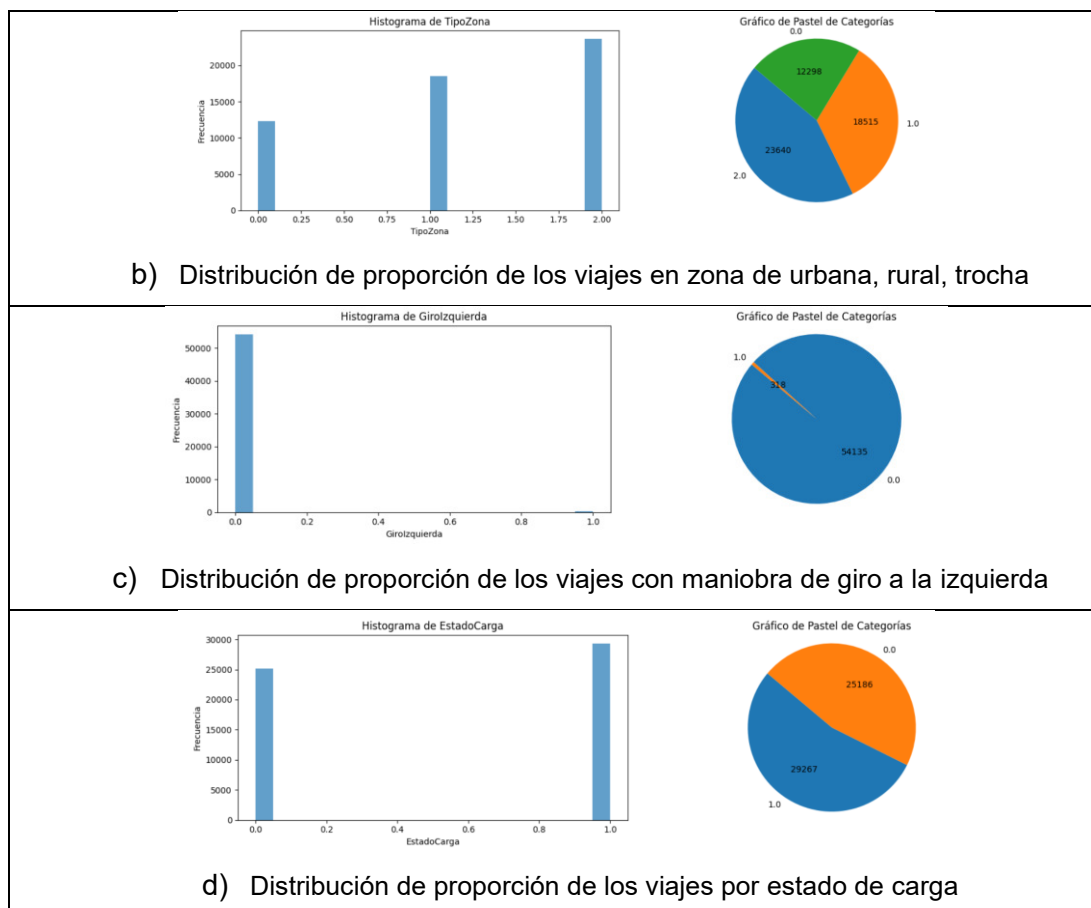
Nota: histograma ejemplo de RPM y Velocidad

En la Figura 15 se puede observar (a) las RPM y (b) la velocidad, destacando los datos que se encuentran fuera de rango en cada uno de los gráficos. Por ejemplo, en el gráfico de RPM se puede ver que algunas posiciones fuera de rango superan las 2000 RPM.

Figura 16

Descripción de variables categóricas





En la Figura 16 se puede observar en el gráfico (a) que existe una proporción considerable del viaje que se realiza en zonas de curvas. En el gráfico (b) se muestra la distribución de los viajes de acuerdo al tipo de zona. En el gráfico (c) se identifica que los eventos de giros a la izquierda son más frecuentes en los viajes. En el gráfico (d) se aprecia que la distribución de viajes es similar cuando el vehículo está cargado, así como, como cuando está vacío.

Preprocesamiento.

1. Limpieza de datos: Se lleva a cabo la verificación de datos que implica controles de limpieza y coherencia de los datos. Inicialmente, se evalúa la calidad de los datos, buscando información vacía, incorrecta, o sesgada. Luego, se depura los datos sin procesar, eliminando valores inexactos y actualizando con valores calculados aquellos que resulten incoherentes o falten en el conjunto de datos.

El resultado de esta actividad obtuvo un conjunto de datos de 37 variables (ver Tabla 10).

Tabla 10

Factores considerados en conjunto de datos después de limpieza

N°	Parámetro	Descripción	Dominio
1	Altitud	Metros sobre el nivel del mar	Vehículo
2	Radio	Orientación, rumbo actual relativo al norte verdadero	Vehículo
3	EstadoIgnition	Encendido o apagado del vehículo	Vehículo
4	Tiempo de Manejo	Es el tiempo de manejo del conductor	Gestión
5	EstadoManejo	Vehículo en movimiento o detenido	Vehículo
6	Activación del freno	Indica si el conductor pisó el pedal de freno	Conductor
7	Temperatura del Refrigerante del Motor	Indica la temperatura del refrigerante del vehículo	Vehículo
8	Presión del Aceite del Motor	Indica la presión del aceite del motor	Vehículo
9	Velocidad del Motor	Indica las revoluciones del motor por minuto	Vehículo
10	Nivel de Combustible	Porcentaje de nivel del tanque de combustible	Vehículo
11	Velocidad del Vehículo Basada en las Ruedas	Indica la velocidad basada en la rueda	Vehículo
12	Boleo de Motor	Cuando el vehículo está en movimiento sin marcha puesta (neutro)	Vehículo
13	Sobre revoluciones	Indica si el motor tiene un RPM no adecuado o anormal	Vehículo
14	Accidente	Si el vehículo tiene un choque o daño	Tránsito
15	Aceleración	Conductor realiza aceleración brusca	Conductor
16	Frenada	Conductor realiza un frenado brusco	Conductor
17	Ingreso a óvalo	El vehículo entra a una vía con curvas prolongadas	Conductor
18	Cambio de carril	El vehículo se desplaza de derecha a izquierda o viceversa	Vehículo
19	Adelantamiento	Vehículo acelera y cambia de carril para realizar un adelantamiento	Vehículo
20	Rompe Muelles	Cuando el vehículo pasa por encima de una superficie elevada	Vehículo
21	Inicio Trocha	Cuando el vehículo inicia su recorrido en pista no asfaltada	Tránsito
22	Fin Trocha	Cuando el vehículo finaliza su recorrida en pista no asfaltada	Tránsito

23	Giros bruscos a la izquierda	Conductor realiza un giro brusco a la izquierda	Conductor
24	Giros bruscos frenando a la izquierda	Conductor realiza un giro brusco frenando a la izquierda	Conductor
25	Giros bruscos acelerando a la izquierda	Conductor realiza un giro brusco acelerando a la izquierda	Conductor
26	Giros bruscos a la derecha	Conductor realiza un giro brusco a la derecha	Conductor
27	Giros bruscos frenando a la derecha	Conductor realiza un giro brusco frenando a la derecha	Conductor
28	Giros bruscos acelerando a la derecha	Conductor realiza un giro brusco acelerando a la derecha	Conductor
29	Rango Horario	Indica si se está conduciendo de día o noche	Entorno
30	Clima	Indica el clima en tiempo real para una posición dada	Entorno
31	Antigüedad	Antigüedad del vehículo	Vehículo
32	Cargado	Vehículo con carga (materiales) y sin carga	Gestión
33	Tipo Vía	Indica el tipo de vía por la que se circula	Entorno
34	Límites de Velocidad	Límites de velocidad existentes en una ruta determinada	Entorno
35	Posición	Posición del vehículo	Vehículo
36	ID del Conductor	ID del conductor	Gestión
37	Posición en curva	Si el vehículo se encuentra en una curva	Entorno

2. Normalización de datos: El propósito es estandarizar los datos de tal manera que se tengan rangos comparables.

4.1.6. Análisis de calidad del conjunto de datos

Regresión logística

Para el estudio se usa regresión logística multiclase con el objetivo de clasificar el riesgo de accidente. Además de la clasificación, se analiza el impacto de algunas variables independientes, como 'Velocidad del motor', 'Nivel de combustible' y 'Tipo de vía', en el riesgo de accidentes. Como resultado se obtuvo las 20 características más relevantes (ver Tabla 11).

Tabla 11*Selección de características más importantes con Regresión logística*

Nº	Característica	Nº	Característica
1	Estado Ignición	11	Clima
2	Estado de Manejo	12	Estado de Carga
3	Posición en Curva	13	Altitud
4	Tipo Vía	14	Radio
5	Posición	15	Límite de Velocidad
6	Activación del freno	16	Temperatura del Refrigerante del Motor
7	Giro Izquierda	17	Presión del Aceite del Motor
8	Giro Izquierda Frenando	18	Velocidad del Motor
9	Giro Izquierda Acelerando	19	Nivel de Combustible
10	Rango Horario	20	Velocidad del Vehículo Basada en las Ruedas

Análisis de coeficientes. Con las 20 características más importantes se identifican los factores que influyen de manera más significativa en la predicción del riesgo de accidentes. Este análisis se apoya en la prueba de Wald y en el p-value. La prueba de Wald evalúa si los coeficientes de las variables independientes son significativamente diferentes de cero, lo que indica si estas variables tienen un impacto relevante en la variable dependiente. Un p-value bajo sugiere que la variable tiene una influencia significativa en el riesgo de accidentes, mientras que un p-value alto indica que su efecto no es estadísticamente relevante. De esta manera, la prueba de Wald analiza cada variable independiente del modelo para determinar si tiene un efecto significativo en la variable dependiente.

Dado que el modelo es multiclase y se enfoca en diferentes categorías del riesgo de accidente, se evalúan las clases correspondientes, como se muestra en la Tabla 12.

Tabla 12

Evaluación de la Significancia de las Características mediante la Prueba de Wald

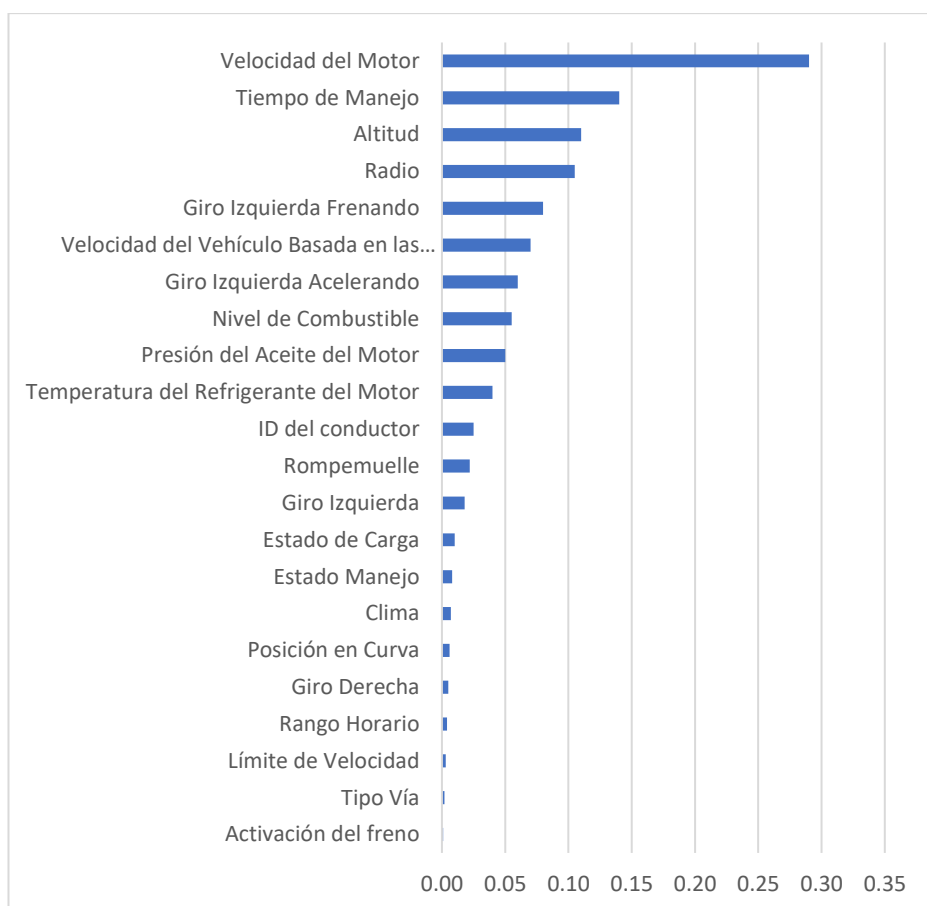
Característica	Prueba Wald Clase 1	p-value Clase 1	Prueba Wald Clase 2	p-value Clase 2	Interpretación
Estado de Ignición	5.5727	0.0182	0.6688	0.4135	Significativo.
Estado de Manejo	9.5004	0.0021	1.8733	0.1711	Significativo.
Posición	0.2760	0.5993	0.6492	0.4204	No Significativo
Tipo de vía	3.2233	0.0691	0.0114	0.9151	Marginalmente significativo.
Posición en Curva	27.7676	0.0000	7.7337	0.0054	Muy significativo.
Activación del freno	4.0957	0.0430	0.0000	0.9990	Significativo.
Giro Izquierda	0.0000	0.9976	0.0000	1.0000	No Significativo
Giro Izquierda Frenando	0.0000	0.9967	207.4804	0.0000	Muy significativo.
Giro Izquierda Acelerando	197.9134	0.0000	0.0000	0.9996	Muy significativo.
Rango Horario	0.0070	0.9334	0.3133	0.5757	No Significativo
Clima	3.1290	0.0769	9.0278	0.0027	Significativo.
Estado de carga	0.5546	0.4564	6.0410	0.0140	Significativo.
Altitud	4.1000	0.0429	2.5509	0.1102	Significativo.
Radio	3.6130	0.0619	0.0326	0.8568	Marginalmente significativo.
Límite de velocidad	0.3570	0.5502	0.5567	0.4556	No Significativo
Temperatura del Refrigerante del Motor	10.1337	0.0015	3.4251	0.0648	Significativo.
Presión del Aceite del Motor	3.4116	0.0647	0.4440	0.5052	Marginalmente significativo.
Velocidad del Vehículo Basada en las Ruedas	8.8945	0.0029	3.6404	0.0564	Significativo.
Nivel de Combustible	34.9645	0.0000	4.0957	0.0430	Muy significativo.
Velocidad del motor	1.1213	0.2896	3.8890	0.0967	Marginalmente significativo.

Seleccionar características más importantes. Según Lacherre et al. (2024), el modelo más efectivo y ampliamente utilizado para identificar las características más importantes es XGBoost, seguido de Random Forest. Por lo que el objetivo principal de este análisis es identificar las características más consistentes, no redundantes y relevantes para su uso en la construcción del modelo predictivo. En este contexto, se utiliza la "Importancia de características", una métrica que evalúa la contribución de cada característica a la capacidad predictiva del modelo. Una vez completada la eliminación de las características menos relevantes y obtenido el conjunto final de datos, estos se dividen en tres partes: entrenamiento, validación y pruebas.

Primero, Se realiza una selección de características utilizando el modelo Random Forest, que determina las más relevantes en función de las importancias calculadas por el modelo, evaluando cada característica según su capacidad para reducir la impureza en los nodos de los árboles de decisión que lo conforman, utilizando para ello el Índice de *Gini* como medida de dicha impureza (ver Figura 17).

Figura 17

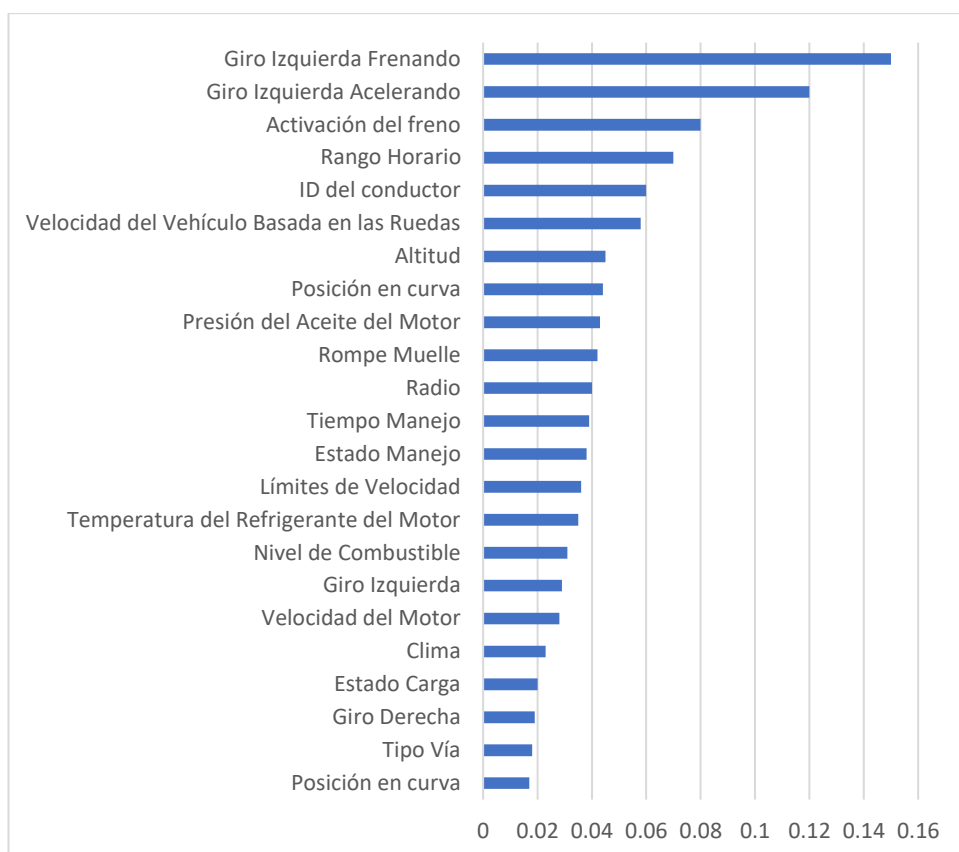
Importancia de Características usando Random Forest de acuerdo a Gini



Luego, se lleva a cabo una selección de características utilizando el modelo XGBoost, que determina las más relevantes en función de las importancias calculadas por el modelo. Este algoritmo evalúa cada característica según su contribución a la mejora del rendimiento del modelo, utilizando la *frecuencia*. Esto significa que se basa en el número de veces que cada característica se utilizó para dividir los árboles en el modelo (ver Figura 18).

Figura 18

Importancia de Características usando XGBoost



Así también se hace un análisis con el modelo XGboost usando el índice F1, Exactitud y Sensibilidad siendo las principales 15 características las siguientes (ver Tabla 13).

Tabla 13

Principales características con XGBoost usando F1, Exactitud y Sensibilidad

Característica	F1	Exactitud	Sensibilidad
Posición en Curva	X	X	X
Tipo Vía	X	X	
Activación del Freno	X	X	X
Giros bruscos frenando a la izquierda	X	X	X
Giros bruscos acelerando a la izquierda	X	X	X
Rompe Muelle	X	X	X
Cargado	X	X	X
Altitud	X	X	X
Radio	X	X	X
Temperatura del Refrigerante del Motor	X	X	X
Presión del Aceite del Motor	X	X	X

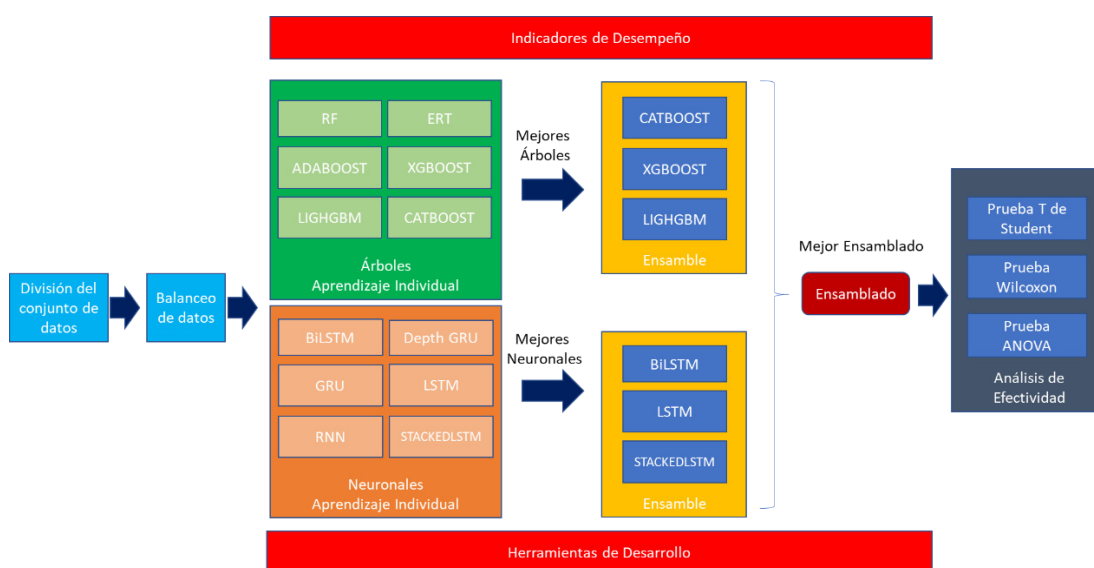
Nivel de Combustible	X	X	X
Velocidad del Vehículo Basada en las Ruedas	X	X	X
Tiempo de Manejo	X	X	X
Posición en Curva	X	X	X

4.2. Diseño del modelo ensamblado de predicción de riesgo de accidente vehicular

En esta fase, se llevan a cabo las etapas para diseñar un modelo de predicción fundamentado en un conjunto de datos equilibrado. Se dividen en conjuntos de entrenamiento, validación y pruebas, y finalmente se procede al entrenamiento del método ensamblado (ver Figura 19).

Figura 19

Modelo ensamblado de predicción de riesgo de accidente vehicular



4.2.1. Indicadores de desempeño

- Para realizar la evaluación del desempeño del modelo híbrido de riesgo de accidente vehicular se usan las siguientes métricas:
 - Exactitud: Evalúa la fracción de predicciones precisas hechas por un modelo respecto al total de predicciones realizadas. En términos

simples, la exactitud determina la proporción de predicciones correctas efectuadas por el modelo.

La fórmula para calcular la exactitud es la siguiente:

$$\text{Exactitud} = (\text{Número de predicciones correctas}) / (\text{Número total de predicciones})$$

- b. Sensibilidad: Evalúa la habilidad de un modelo para identificar correctamente todos los casos positivos en un conjunto de datos. En términos simples, la sensibilidad examina la proporción de casos positivos reales que el modelo ha identificado de manera precisa.

La fórmula para calcular la sensibilidad es la siguiente:

$$\text{Sensibilidad} = (\text{Verdaderos Positivos}) / (\text{Verdaderos Positivos} + \text{Falsos Negativos})$$

La sensibilidad es vital en escenarios donde detectar todos los casos positivos es esencial y los falsos negativos son costosos o cruciales. En contextos como la prevención de accidentes vehiculares, maximizar la sensibilidad resulta fundamental para identificar la mayor cantidad posible de riesgos, aun si esto implica un aumento en los falsos positivos.

4.2.2. División del conjunto de datos

En el proceso, los datos se dividen en tres conjuntos distintos: entrenamiento (60%), validación (20%) y prueba (20%). El conjunto de entrenamiento, se emplea para entrenar el modelo, permitiéndole aprender los patrones y estructuras presentes en los datos. El conjunto de validación, se utiliza para ajustar los hiper parámetros del modelo. Por último, el conjunto de prueba se usa para evaluar el rendimiento final del modelo para predecir sobre datos no observados.

Después de la división de los datos, se aplica la técnica de oversampling SMOTE que balancea las clases. Este método aumenta el número de instancias de la clase minoritaria para equilibrar la proporción entre las clases. Esto mejora el modelo para identificar y aprender patrones relevantes de la clase minoritaria y de esta manera se reduce el sesgo hacia la clase mayoritaria.

4.2.3. Aprendizaje individual

Para llevar a cabo esta actividad, primero se identifican los algoritmos necesarios para entrenar de forma individual cada modelo. Se realizó una exhaustiva investigación en la literatura disponible para seleccionar los algoritmos clave (Lacherre et al., 2024). Posteriormente, se procedió a ajustar manualmente algunos parámetros y evaluar la mejora de su rendimiento. En la búsqueda de las mejores métricas se usa algoritmos de ML y DL de tal manera que al ensamblarlos mejore sus resultados. Por cada modelo se realiza su matriz de confusión para evaluar las métricas de sensibilidad y exactitud con el fin de encontrar los mejores modelos.

Modelos de Árboles. Por cada modelo se realizan 5 experimentos y se cambian parámetros para ajustar el modelo y de esta manera tener las métricas de Exactitud y Sensibilidad (ver Tabla 14).

Tabla 14

Experimentos para los modelos de árboles

Exp.	Modelo	Parámetro 1	Parámetro 2	Exactitud (%)	Sensibilidad (%)
1	Random Forest	max_depth: 10	min_samples_split = 12	83	92
1	Extremely Randomized Trees	max_depth: 10	min_samples_split = 12	86	87
1	AdaBoost	n_estimators: 90	Learning Rate = 0.1	64	72
1	XGBoost	max_depth: 10	Learning Rate = 0.1	98	74
1	LightGBM	max_depth = 10	learning_rate = 0.1	100	48
1	CatBoost	depth = 10	learning_rate = 0.1	99	38
2	Random Forest	max_depth: 12	min_samples_split = 14	90	83
2	Extremely Randomized	max_depth: 12	min_samples_split = 12	88	88

	zed Trees				
2	AdaBoos t	n_estimator s: 100	Learning Rate = 0.01	54	54
2	XGBoost	max_depth: 12	Learning Rate = 0.01	99	66
2	LightGB M	max_depth = 12	learning_rate = 0.01	98	69
2	CatBoost	depth = 12	learning_rate = 0.01	94	82
3	Random Forest	max_depth: 14	min_samples_s plit = 16	94	84
3	Extremel y Randomi zed Trees	max_depth: 14	min_samples_s plit = 16	91	83
3	AdaBoos t	n_estimator s: 105	Learning Rate = 0.05	31	69
3	XGBoost	max_depth: 14	Learning Rate = 0.05	99	66
3	LightGB M	max_depth = 14	learning_rate = 0.05	100	52
3	CatBoost	depth = 14	learning_rate = 0.05	99	43
4	Random Forest	max_depth: 8	min_samples_s plit = 10	68	86
4	Extremel y Randomi zed Trees	max_depth: 8	min_samples_s plit = 10	82	83
4	AdaBoos t	n_estimator s: 95	Learning Rate = 0.02	20	62
4	XGBoost	max_depth: 8	Learning Rate = 0.02	98	69
4	LightGB M	max_depth = 8	learning_rate = 0.02	99	47
4	CatBoost	depth = 8	learning_rate = 0.02	94	77
5	Random Forest	max_depth: 12	min_samples_s plit = 12	90	83
5	Extremel y Randomi zed Trees	max_depth: 12	min_samples_s plit = 12	88	88
5	AdaBoos t	n_estimator s: 100	Learning Rate = 0.001	20	62

5	XGBoost	max_depth: 12	Learning Rate = 0.001	94	81
5	LightGBM	max_depth = 12	learning_rate = 0.001	89	91
5	CatBoost	depth = 12	learning_rate = 0.001	89	91

El siguiente paso es obtener la media y desviación estándar de cada modelo (ver Tabla 15).

Tabla 15

Métricas agrupadas por modelo de árboles

Modelo	Media de Exactitud (%)	Desviación estándar de Exactitud (%)	Media de Sensibilidad (%)	Desviación estándar de Sensibilidad (%)
Random Forest	85.0	9.21	85.6	3.69
Extremely Randomized Tree	87.0	3.37	86.6	2.42
AdaBoost	39.0	45.18	65.0	4.28
XGBoost	97.6	2.07	71.2	6.55
LightGBM	97.2	4.07	61.4	16.83
CatBoost	95.0	3.52	72.2	18.71

XGB, CB y LGBM son los modelos con mejores resultados en cuanto a exactitud. XGB presenta la menor desviación estándar, lo que indica consistencia en su rendimiento.

Modelos Neuronales. Por cada modelo neuronal se realizan 5 experimentos y se cambian los parámetros para ajustar el modelo (ver Tabla 16).

Tabla 16

Experimentos para los modelos neuronales

Experimento	Modelo	Parámetro 1	Parámetro 2	Exactitud (%)	Sensibilidad (%)
-------------	--------	-------------	-------------	---------------	------------------

1	LSTM	Unidades : 8	Tasa de Aprendizaje: 0.01	3	63
1	BiLSTM	Unidades : 8	Tasa de Aprendizaje: 0.01	85	79
1	StackedLSTM	Unidades : 8	Tasa de Aprendizaje: 0.01	57	80
1	GRU	Unidades : 8	Tasa de Aprendizaje: 0.01	0	53
1	Depth GRU	Unidades : 8	Tasa de Aprendizaje: 0.01	0	50
1	RNN	Unidades : 8	Tasa de Aprendizaje: 0.01	0	53
2	LSTM	Unidades : 12	Tasa de Aprendizaje: 0.008	1	53
2	BiLSTM	Unidades : 12	Tasa de Aprendizaje: 0.008	91	80
2	StackedLSTM	Unidades : 12	Tasa de Aprendizaje: 0.008	89	83
2	GRU	Unidades : 12	Tasa de Aprendizaje: 0.008	0	67
2	Depth GRU	Unidades : 12	Tasa de Aprendizaje: 0.008	17	65
2	RNN	Unidades : 12	Tasa de Aprendizaje: 0.008	0	39
3	LSTM	Unidades : 16	Tasa de Aprendizaje: 0.003	69	87
3	BiLSTM	Unidades : 16	Tasa de Aprendizaje: 0.003	96	75
3	StackedLSTM	Unidades : 16	Tasa de Aprendizaje: 0.003	91	83
3	GRU	Unidades : 16	Tasa de Aprendizaje: 0.003	0	67

3	Depth GRU	Unidades : 16	Tasa de Aprendizaje: 0.003	29	69
3	RNN	Unidades : 16	Tasa de Aprendizaje: 0.003	26	65
4	LSTM	Unidades : 30	Tasa de Aprendizaje: 0.001	95	77
4	BiLSTM	Unidades : 30	Tasa de Aprendizaje: 0.001	95	72
4	StackedLSTM	Unidades : 30	Tasa de Aprendizaje: 0.001	94	81
4	GRU	Unidades : 30	Tasa de Aprendizaje: 0.001	88	80
4	Depth GRU	Unidades : 30	Tasa de Aprendizaje: 0.001	92	81
4	RNN	Unidades : 30	Tasa de Aprendizaje: 0.001	90	78
5	LSTM	Unidades : 50	Tasa de Aprendizaje: 0.001	96	74
5	BiLSTM	Unidades : 50	Tasa de Aprendizaje: 0.001	97	55
5	StackedLSTM	Unidades : 50	Tasa de Aprendizaje: 0.001	94	81
5	GRU	Unidades : 50	Tasa de Aprendizaje: 0.001	92	73
5	Depth GRU	Unidades : 50	Tasa de Aprendizaje: 0.001	93	76
5	RNN	Unidades : 50	Tasa de Aprendizaje: 0.001	92	84

El siguiente paso es obtener la media y desviación estándar de cada modelo (ver Tabla 17).

Tabla 17

Métricas agrupadas por modelo neuronal

Modelo	Media de Exactitud (%)	Desviación estándar de Exactitud (%)	Media de Sensibilidad (%)	Desviación estándar de Sensibilidad (%)
BiLSTM	92.80	4.40	72.20	9.06
Depth GRU	46.20	38.91	68.20	10.65
GRU	36.00	44.11	68.00	8.90
LSTM	52.80	42.60	70.80	11.74
RNN	41.60	41.44	63.80	16.39
StackedLSTM	85.00	14.13	81.60	1.20

BiLSTM, StackedLSTM y LSTM son los modelos con mejores resultados en cuanto a exactitud. Sin embargo, se observa que el mejor resultado en base a Sensibilidad es StackedLSTM.

4.2.4. Aprendizaje conjunto

El aprendizaje conjunto es una técnica del aprendizaje automático que combina las predicciones de múltiples modelos individuales (también llamados aprendices base) para mejorar la exactitud y robustez del modelo final. La idea central del aprendizaje conjunto es que la agregación de varios modelos puede compensar los errores individuales, lo que conduce a un rendimiento superior en comparación con cualquier modelo individual (Sagi & Rokach, 2018).

Dentro de los algoritmos empleados destaca el VotingClassifier, un clasificador de votación en conjunto que permite combinar múltiples clasificadores individuales para predecir la clase de una muestra mediante votación. Este clasificador puede hacer uso de distintos algoritmos aprendices base, permitiendo tanto votación dura como blanda para la combinación de las predicciones de los clasificadores. En este estudio, se propone implementar un método de ensamble que aprovecha los mejores algoritmos con el fin de clasificar el RAVCM para camiones logrando una óptima capacidad predictiva en función de las métricas de sensibilidad y exactitud. Para realizar esto, se selecciona los principales métodos base tomando los

parámetros del experimento que obtuvo el mejor resultado individual y con estos modelos se realizará un ensamblado con los mejores 3 modelos para luego seleccionar al que tenga el mejor resultado de estos experimentos.

Modelo de Árboles. Según los modelos individuales, XGBoost, LGBM y CatBoost son los tres modelos más efectivos en términos de exactitud. El siguiente paso consiste en desarrollar un modelo de aprendizaje conjunto que combine estos tres modelos basados en árboles. A partir de esta combinación, se realizarán cinco experimentos variando dos parámetros con el fin de evaluar los resultados en las métricas de exactitud y sensibilidad (ver Tabla 18).

Tabla 18

Experimentos para el modelo ensamblado de árboles.

Experimento	Modelo Ensamblado	Parámetro 1	Parámetro 2	Exactitud (%)	Sensibilidad (%)
1	XGB	max_depth=10	learning_rate=0.01	97	77
	CatBoost	depth=10	learning_rate=0.01		
	LightGBM	max_depth=10	learning_rate=0.01		
2	XGB	max_depth=12	learning_rate=0.08	97	74
	CatBoost	depth=12	learning_rate=0.08		
	LightGBM	max_depth=12	learning_rate=0.08		
3	XGB	max_depth=14	learning_rate=0.05	96	82
	CatBoost	depth=14	learning_rate=0.05		
	LightGBM	max_depth=14	learning_rate=0.05		
4	XGB	max_depth=16	learning_rate=0.03	96	85
	CatBoost	depth=16	learning_rate=0.03		
	LightGBM	max_depth=16	learning_rate=0.03		
5	XGB	max_depth=10	learning_rate=0.001	91	91

CatBoost	depth=14	learning_rate=0.001
LightGBM	max_depth=12	learning_rate=0.001

Modelo de Neuronales. De acuerdo a los modelos individuales, los modelos BiLSTM, StackedLSTM y LSTM son los 3 mejores modelos considerando la exactitud. El siguiente paso es realizar un modelo de aprendizaje conjunto combinando los 3 modelos neuronales y con esta combinación realizar 5 experimentos cambiando 2 parámetros para evaluar los resultados de las métricas de Exactitud y sensibilidad (ver Tabla 19).

Tabla 19

Experimentos para el modelo ensamblado de neuronales

Experimento	Modelo Ensamblado	Parámetro 1	Parámetro 2	Exactitud (%)	Sensibilidad (%)
1	LSTM BiLSTM StackedLSTM	Unidades =12	learning_rate=0.01	79	85
2	LSTM BiLSTM StackedLSTM	Unidades =16	learning_rate=0.008	80	86
3	LSTM BiLSTM StackedLSTM	Unidades =30	learning_rate=0.005	94	79
4	LSTM BiLSTM StackedLSTM	Unidades =50	learning_rate=0.003	95	82
5	LSTM BiLSTM StackedLSTM	Unidades =80	learning_rate=0.001	97	77

4.2.5. Herramientas de desarrollo

Para la construcción del modelo, se emplea Windows 11 Pro como software principal, utilizando Python como lenguaje de programación. Se destacan las librerías TensorFlow, Keras y Scikit-learn, que desempeñan un papel fundamental en el desarrollo de los modelos.

El hardware empleado para llevar a cabo los procesos de entrenamiento, validación y pruebas posee las siguientes características: un procesador 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz, 1.69 GHz de velocidad, y una memoria RAM de 16 GB.

4.2.6. Análisis de efectividad de modelo

Prueba T de Student. Para realizar este análisis se aplica la T de Student entre el grupo de modelos individuales y el grupo de modelos de ensamblado. En primer lugar, se aplica para la métrica de Exactitud (ver Tabla 20) y luego para la métrica de Sensibilidad (ver Tabla 21).

Tabla 20

Prueba T de Student para la métrica Exactitud

Grupo	Media Exactitud (%)	Desviación Estándar (%)	N	Valor t	p-value
Modelos Individuales	71.27	24.78	12	-2.18	0.049
Modelos Ensamblados	92.70	3.82	2		

La diferencia en la media de exactitud entre los modelos individuales y los modelos ensamblados es estadísticamente significativa al nivel de significancia del 5%, dado que el valor p es menor que 0.05.

Los resultados de la prueba T de Student para comparar la exactitud de los modelos individuales y los modelos ensamblados indican que hay una diferencia estadísticamente significativa entre ambos grupos.

Tabla 21

Prueba T de Student para la métrica Sensibilidad.

Grupo	Media Sensibilidad (%)	Desviación Estándar (%)	N	Valor t	p-value
Modelos Individuales	72.37	8.21	12	-3.67	0.003
Modelos Ensamblados	81.80	0	2		

La diferencia en la media de sensibilidad entre los modelos individuales y los modelos ensamblados es estadísticamente significativa al nivel de significancia del 5%, dado que el valor p es menor que 0.05.

Esto sugiere que los modelos ensamblados no solo mejoran la exactitud, sino también la sensibilidad en la predicción del riesgo de accidentes vehiculares.

Prueba Wilcoxon. La prueba de Wilcoxon es una prueba no paramétrica que puede ser útil cuando no se puede asumir que la diferencia entre dos grupos sigue una distribución normal, como en el caso de la prueba T de Student. En este caso, compararíamos las diferencias en sensibilidad y exactitud entre los modelos individuales y los modelos ensamblados. Por lo tanto, se aplica la prueba de Wilcoxon a los datos de exactitud (ver Tabla 22) y sensibilidad (ver Tabla 23) de los modelos individuales y ensamblados.

Tabla 22

Prueba Wilcoxon para la métrica Exactitud

Grupo	Media Exactitud (%)	Rangos absolutos	Suma Rangos
Modelos Individuales	71.27	1, 2, 3, ..., 12	78
Modelos Ensamblados	92.70	1, 2	3
Estadístico W:		p-value: 0.046	
3			

Tabla 23*Prueba Wilcoxon para la métrica Sensibilidad*

Grupo	Media Sensibilidad (%)	Rangos absolutos	Suma Rangos
Modelos Individuales	72.37	1, 2, 3, ..., 12	78
Modelos Ensamblados	81.80	1, 2	3
Estadístico W: 3	p-value: 0.004		

La prueba de Wilcoxon revela que hay diferencias estadísticamente significativas en la exactitud y sensibilidad entre los modelos individuales y ensamblados.

Exactitud: El valor p de 0.046 indica que los modelos ensamblados tienen una exactitud significativamente mayor que los modelos individuales.

Sensibilidad: El valor p de 0.004 sugiere que los modelos ensamblados también tienen una sensibilidad significativamente mayor que los modelos individuales.

Prueba Análisis de Varianza (ANOVA). El análisis de varianza (ANOVA) permite comparar las medias de múltiples grupos y determinar si hay diferencias significativas entre ellos. En este caso, se aplica ANOVA para comparar las medias de exactitud (ver Tabla 24) y sensibilidad (ver Tabla 25) de los modelos individuales y ensamblados.

Tabla 24*Prueba ANOVA para la métrica Exactitud*

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Media de los Cuadrados	F-statistic	p-value
Entre Grupos	700.95	1	700.95	10.93	0.005

Dentro de los Grupos	751.35	12	62.61
Total	1452.30	13	

Tabla 25

Prueba ANOVA para la métrica Sensibilidad

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Media de los Cuadrados	F-statistic	p-value
Entre Grupos	224.18	1	224.18	8.34	0.013
Dentro de los Grupos	322.42	12	26.87		
Total	546.60	13			

Los resultados de ANOVA muestran que hay diferencias estadísticamente significativas tanto en la exactitud como en la sensibilidad entre los modelos individuales y ensamblados.

Exactitud: Con un F-statistic de 10.93 y un p-value de 0.005, podemos concluir que la diferencia en la exactitud entre los modelos individuales y los modelos ensamblados es significativa.

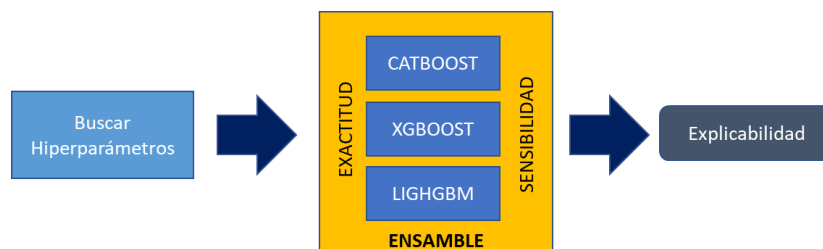
Sensibilidad: Con un F-statistic de 8.34 y un p-value de 0.013, también podemos concluir que la diferencia en la sensibilidad entre los dos grupos es significativa.

4.3. Validación del modelo ensamblado de predicción de riesgo de accidente vehicular

En esta etapa se realiza la calibración de los hiper parámetros para luego realizar la validación del modelo. A continuación, se describe las actividades realizadas (ver Figura 20).

Figura 20

Actividades para la validación del modelo ensamblado ajustado



4.3.1. Buscar hiper parámetros

Realizar la optimización de hiper parámetros es importante en el desarrollo de modelos de aprendizaje automático. El objetivo de esta búsqueda es encontrar una combinación que optimice el modelo.

Entre el estudio se evaluarán hiper parámetros como la tasa de aprendizaje, la profundidad del modelo y otros factores que influyen en la optimización. Se realizará una optimización de hiper parámetros considerando: (1) modelos usados en ensamble de mayor resultado (2) considerar la optimización de hiper parámetros de los modelos individuales, buscando la mejor exactitud y sensibilidad y (3) Uso del método GridSearchCV para optimizar los hiper parámetros (ver Tabla 26).

Tabla 26

Modelos y parámetros usados para la búsqueda de hiper parámetros por métrica

Métrica	Modelo	Parámetros
Exactitud	XGBOOST	n_estimators: [400]
		max_depth: [8, 10, 16, 20, 24, 30]
	LIGHTGBM	learning_rate: [0.01, 0.008, 0.005, 0.001]
		num_leaves: [25]
		n_estimators: [500]
		max_depth: [10, 12, 16, 20, 24, 30]
		learning_rate: [0.01, 0.008, 0.005, 0.001]

Sensibilidad	CATBOOST	iterations: [150] depth': [12, 16, 20] learning_rate: [0.01, 0.001]
	XGBOOST	n_estimators: [400] max_depth': [8, 10, 16, 20, 24, 30] learning_rate: [0.01, 0.008, 0.005, 0.001]
	LIGHTGBM	num_leaves: [25] n_estimators': [500] max_depth': [10, 12, 16, 20, 24, 30] learning_rate: [0.01, 0.008, 0.005, 0.001]
	CATBOOST	iterations: [250] depth: [12, 16, 20, 24, 30] learning_rate: [0.01, 0.008, 0.005, 0.001]

4.3.2. *Ensamble optimizado*

El siguiente paso es realizar el ensamblado de los modelos individuales considerando los hiper parámetros de mejor resultado (XGBOOST, LIGHTGBM y CATBOOST). Para el ensamblado de modelos se usa VotingClassifier con su método SOFT.

4.3.3. *Explicabilidad del modelo*

En este trabajo se aplica explicabilidad con SHAP para el modelo ensamblado. En primer lugar, se resume el conjunto de datos de entrenamiento mediante K-means en 50 clusters para hacer más simple los cálculos. Luego, se crea un explicador con KernelExplainer que usa muestras resumidas y el clasificador para estimar cómo cada característica afecta en la predicción. Finalmente, se calcula los valores de SHAP, utilizando 100 muestras.

Después, se genera un gráfico de barras resumen utilizando los valores de SHAP para mostrar la importancia promedio de las características en las predicciones del modelo. Se definen las etiquetas de las clases que representan diferentes conductas de manejo y los nombres de las características del conjunto de datos, lo que permite identificar de manera clara cuáles variables tienen el mayor impacto en la clasificación de la conducta del conductor.

CAPÍTULO V. RESULTADOS Y DISCUSIÓN

5.1. Análisis, interpretación y discusión de resultados

5.1.1. Estudio naturalista de recolección de datos de telemetría

En el estudio naturalista se utilizaron 4 unidades vehiculares que realizaron el mismo recorrido de ida y vuelta durante 7 días. Participaron 5 conductores, y en total se generaron 57,830 registros de datos con 50 características. Tras la limpieza de los datos y la eliminación de columnas irrelevantes, se conservaron 57,830 registros y 37 columnas.

Para el análisis de predicción del riesgo de accidentes, se empleó un modelo de regresión logística multiclase con 20 características. El impacto de cada variable en el modelo se evaluó con el análisis de coeficientes, usando la prueba de Wald y p-value. Como resultado se encontró que el Estado de Manejo, la Altitud y la Velocidad del Vehículo Basada en las Ruedas fueron estadísticamente significativas. Asimismo, se encontró que la Posición en Curva y los Giros a la Izquierda al Acelerar y Frenar son altamente significativas, lo que significa que los cambios bruscos de dirección durante la aceleración o frenado son clave para predecir el riesgo de accidentes.

Adicionalmente, se evaluó la importancia de las características usando Random Forest y XGBoost. Estos algoritmos identificaron las características más relevantes en función del Índice de Gini y la frecuencia de uso en las divisiones de los árboles de decisión.

Finalmente, se integraron todos los resultados para determinar 14 características que tuvieron mayor consistencia y fiabilidad para el conjunto de datos (ver Tabla 27).

Tabla 27*Comparación de la Importancia de Características en Diferentes Modelos**(RLM, Random Forest y XGBoost)*

Característica	RLM (COEF)	RF (GINI)	XGB (F1)	XGB (ACC)	XGB (RECALL)	XGB WEIGHT
Posición en Curva	X	X	X	X	X	X
Tipo Vía	X		X	X		X
Activación del Freno	X	X	X	X	X	X
Giros bruscos frenando a la izquierda	X	X	X	X	X	X
Giros bruscos acelerando a la izquierda	X	X	X	X	X	X
Rompe Muelle		X	X	X	X	X
Cargado	X	X	X	X	X	X
Altitud	X	X	X	X	X	X
Radio	X	X	X	X	X	X
Temperatura del Refrigerante del Motor	X	X	X	X	X	X
Presión del Aceite del Motor	X	X	X	X	X	X
Nivel de Combustible	X	X	X	X	X	X
Velocidad del Vehículo	X	X	X	X	X	X
Basada en las Ruedas						
Tiempo de Manejo			X	X	X	X

Con estos resultados se puede confirmar que el estudio naturalista de recolección de datos de telemetría en vehículos pesados del sector minero ha permitido generar un conjunto de datos robusto y adecuado para la predicción del riesgo de accidentes. Las variables más significativas, como la velocidad del vehículo basada en las ruedas, la altitud y la posición en curva, son factores clave. Adicionalmente, los giros bruscos, demuestran la relevancia, ya que las maniobras abruptas tienen un impacto directo en el riesgo de accidentes. Estos resultados aportan nuevas evidencias específicas para el

sector minero en carreteras rurales sobre todo porque son vehículos pesados que no transitan a velocidades altas.

Adicionalmente, el estudio se desarrolla en un entorno controlado, donde los datos de telemetría provienen de conductores de una misma empresa que operan en rutas fijas con vehículos homogéneos, lo que garantiza una alta consistencia interna y fortalece la robustez del modelo predictivo. Sin embargo, esta configuración podría introducir un sesgo, ya que, al aplicar el modelo en un contexto diferente, variables como las características de los vehículos, las habilidades y comportamientos de los conductores y la variabilidad de las rutas podrían cambiar. Aunque el modelo predictivo podría aplicarse en escenarios similares, su implementación en contextos distintos requeriría ajustes con datos específicos del nuevo entorno. Para mejorar su generalización y validez, futuras investigaciones deberían incluir muestras más diversas en cuanto a tipos de conductores, rutas y condiciones operativas.

La Tabla 28 presenta un ejemplo de datos de telemetría en distintos contextos, evidenciando diferencias significativas en la variabilidad de las métricas registradas. En el entorno minero, los conductores son empleados entrenados que operan en rutas fijas, lo que genera datos más homogéneos y menor variabilidad en eventos como frenadas bruscas o aceleraciones agresivas. En contraste, en el contexto urbano, los conductores poseen distintos niveles de experiencia y estilos de manejo, lo que introduce una mayor dispersión en estas variables. Además, mientras que en la minería las condiciones del terreno son constantes y predecibles, lo que estabiliza la velocidad y el consumo de combustible, en las rutas urbanas, factores como el tráfico, los semáforos y las intersecciones generan fluctuaciones más pronunciadas en estas métricas, así como en los tiempos de ralentí.

Tabla 28

Ejemplo de Datos de Telemetría en Contextos Distintos.

Campo	Descripción	Contexto 1: Minería	Contexto 2: Urbano
-------	-------------	------------------------	-----------------------

ID Conductor	Identificación única del conductor	ID_001, ID_002 (empleados fijos, alta experiencia)	ID_101, ID_102, ID_103 (conductores con distintos niveles de experiencia y estilos de manejo)
ID Vehículo	Identificación única del vehículo	VH_001, VH_002 (vehículos de la misma marca y modelo)	VH_201, VH_202, VH_203 (vehículos de diferentes marcas, años y condiciones mecánicas)
Tipo de Carretera	Superficie y características de la vía	Terreno irregular, sin pavimentar, con pendientes	Pavimentado, con semáforos, intersecciones, tráfico variable
Velocidad Promedio (km/h)	Velocidad media en el trayecto	30-40 km/h (limitada por seguridad)	20-60 km/h (según tráfico y semáforos)
Aceleraciones Bruscas (por viaje)	Eventos de aceleración agresiva	0-2 (conductores entrenados, menor variabilidad)	2-8 (dependiendo del tráfico y estilo de manejo del conductor)
Frenadas Bruscas (por viaje)	Eventos de frenado repentino	1-3 (terreno predecible, baja variabilidad)	5-15 (tráfico, peatones, cruces impredecibles)
RPM Promedio	Revoluciones por minuto del motor	1,500 - 2,200 RPM (operación regulada)	1,000 - 3,500 RPM (dependiendo del tipo de conducción y tráfico)
Consumo de Combustible (L/100 km)	Eficiencia de consumo	35-45 L/100 km (alta carga, terreno exigente)	8-15 L/100 km (variable según el tráfico y tipo de conducción)
Tiempo en Ralentí (%)	Tiempo en que el motor está encendido sin moverse	5-10% (pocas paradas, flujo constante)	15-30% (atascos, semáforos, espera en intersecciones)
Desviación de Ruta (%)	Variación respecto al trayecto planeado	0-2% (rutas fijas, predefinidas)	10-30% (ajustes constantes por tráfico, desvíos)
Clima Registrado	Condiciones climáticas al momento del viaje	Polvo, lluvias ligeras, niebla frecuente	Lluvia, sol intenso, niebla, cambios abruptos

5.1.2. Modelo ensamblado de predicción de riesgo de accidente vehicular

En el diseño del modelo de ensamblado de predicción de RAV se tienen los siguientes resultados basados en las métricas de Exactitud y Sensibilidad (ver Tabla 29).

Tabla 29

Comparación de resultados de modelos individuales y ensamblados en base a exactitud y sensibilidad

Modelo	Media de Exactitud (%)	Desviación estándar de Exactitud (%)	Media de Sensibilidad (%)	Desviación estándar de Sensibilidad (%)
Random Forest	85.00	9.21	85.60	3.69
Extremely Randomized Tree	87.00	3.37	86.60	2.42
AdaBoost	39.00	45.18	65.00	4.28
XGBoost	97.60	2.07	71.20	6.55
LightGBM	97.20	4.07	61.40	16.83
CatBoost	95.00	3.52	72.20	18.71
BiLSTM	92.80	4.40	72.20	9.06
Depth GRU	46.20	38.91	68.20	10.65
GRU	36.00	44.11	68.00	8.90
LSTM	52.80	42.60	70.80	11.74
RNN	41.60	41.44	63.80	16.39
StackedLSTM	85.00	14.13	81.60	1.20
Ensamblado árboles	95.40	2.24	81.80	5.98
Ensamblado Neuronal	90.00	7.82	81.80	3.43

Se puede observar que, basados en exactitud, el mejor modelo de árboles es XGBoost con 97,60% y de redes es BiLSTM. Sin embargo, basado en sensibilidad el mejor modelo en arboles es Extremely Randomized Tree y en redes es StackedLSTM. Sin embargo, al realizar el ensamblado y tener en cuenta ambos parámetros el mejor modelo es el ensamblado de árboles con una exactitud de 95.40% y una sensibilidad de 81.80%.

5.1.3. Modelo ensamblado optimizado

Resultados de búsqueda de hiper parámetros. Al aplicar los parámetros encontrados en la búsqueda de hiper parámetros a los modelos que son parte del ensamblado de árboles y neuronales, se tienen los siguientes resultados (ver Tabla 30):

Tabla 30

Parámetros y métricas optimizadas para los modelos usados en el ensamblado

Métrica	Modelo	Parámetros optimizados	Exactitud	Sensibilidad
Exactitud	XGBOOST	alpha: 0.01 learning_rate: 0.01 max_depth: 20 n_estimators: 400	99%	72%
	LIGHTGBM	learning_rate: 0.01 max_depth: 16 n_estimators: 500 num_leaves: 25	99%	72%
	CATBOOST	depth: 16 iterations:150 l2_leaf_reg: 0.1	97%	67%
Sensibilidad	XGBOOST	learning_rate: 0.01 alpha: 0.01 max_depth: 20 n_estimators: 400	99%	72%
	LIGHTGBM	learning_rate: 0.01 max_depth: 16 n_estimators: 500 num_leaves: 25	99%	72%
	CATBOOST	depth: 16 iterations:250 l2_leaf_reg: 0.1 learning_rate: 0.01	98%	72%

Resultados de ensamble optimizado. Debido a que los mejores resultados de forma individual se obtuvieron el ensamblado de árboles, entonces se realiza el entrenamiento con estos modelos optimizados y los resultados son los siguientes (ver Tabla 31):

Tabla 31

Resultados de métricas del modelo de ensamble optimizado

Modelo	Exactitud (Pruebas)	Sensibilidad (Pruebas)
Ensamblado de árboles	99%	81%

Las métricas del modelo ensamblado revelan que, en el conjunto de pruebas, el modelo presenta una alta exactitud del 99%, lo que indica que clasifica correctamente casi todas las muestras. La sensibilidad, situada en el 81%, sugiere que el modelo es bastante efectivo para identificar los casos positivos, aunque aún existe un pequeño porcentaje de falsos negativos. Esto implica que, si bien el modelo tiene un buen rendimiento general, hay margen para mejorar la detección de casos positivos, lo que podría ser crucial dependiendo del contexto del problema.

Al comparar el modelo propuesto con el mejor modelo de vehículo pesado y el mejor modelo ensamblado presentados en el estudio de Lacherre et al. (2024), se evidencia que el enfoque propuesto supera a los modelos desarrollados en investigaciones previas (ver Tabla 32).

Tabla 32

Comparativo con otros modelos de estudios previos

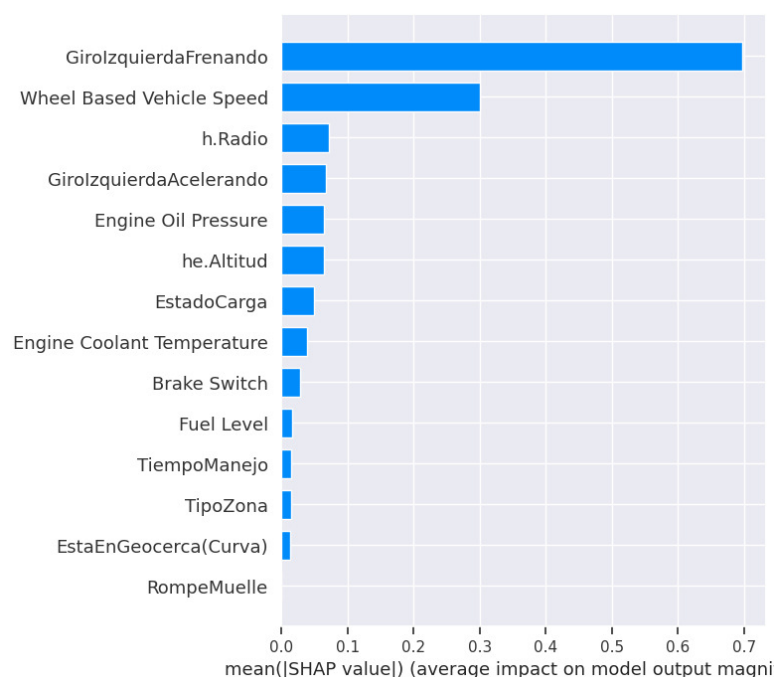
Estudio	Modelo	Exactitud	Descripción
Geng et al. (2022)	Sequential Covering	96.25%	Clasifica las maniobras de conducción en vehículos pesados basándose en datos de series temporales
Z. Hu et al. (2023)	BCDU-Net+ ConvLSTM	98.48%	Predice el riesgo de accidentes de tráfico de vehículos

			ligeros basado en características espacio-temporales
Estudio	Ensamblado de árboles	99%	

Resultados de Explicabilidad. Como resultado general de las tres clases, se presenta el siguiente gráfico de barras, el cual ilustra las características que más contribuyen a la explicación del modelo. Según el gráfico, se observa que las dos variables más significativas en la determinación del riesgo de accidente vehicular son el giro a la izquierda al frenar y la velocidad del vehículo. Esto puede explicarse por el hecho de que, en carreteras rurales, los vehículos tienden a realizar maniobras de adelantamiento con frecuencia en vías de un solo carril, lo que puede dar lugar a estas dos conductas: girar a la izquierda e incrementar la velocidad para sobrepasar al vehículo que se desea adelantar (ver Figura 21).

Figura 21

Características que mejor explican la Predicción del RAV



En la Tabla 33 se describe la influencia de cada característica en la predicción de las diferentes clases de riesgo (0: Normal, 1: Riesgosa, 2: Peligrosa). La influencia está indicada como Positiva (empuja la predicción hacia esa clase) o Negativa (empuja la predicción hacia otra clase). Para un mayor detalle sobre la explicación de la predicción del riesgo, se puede consultar el contenido detallado en los Anexos A, B y C.

Tabla 33

Influencia de factores en la predicción del RAVCC

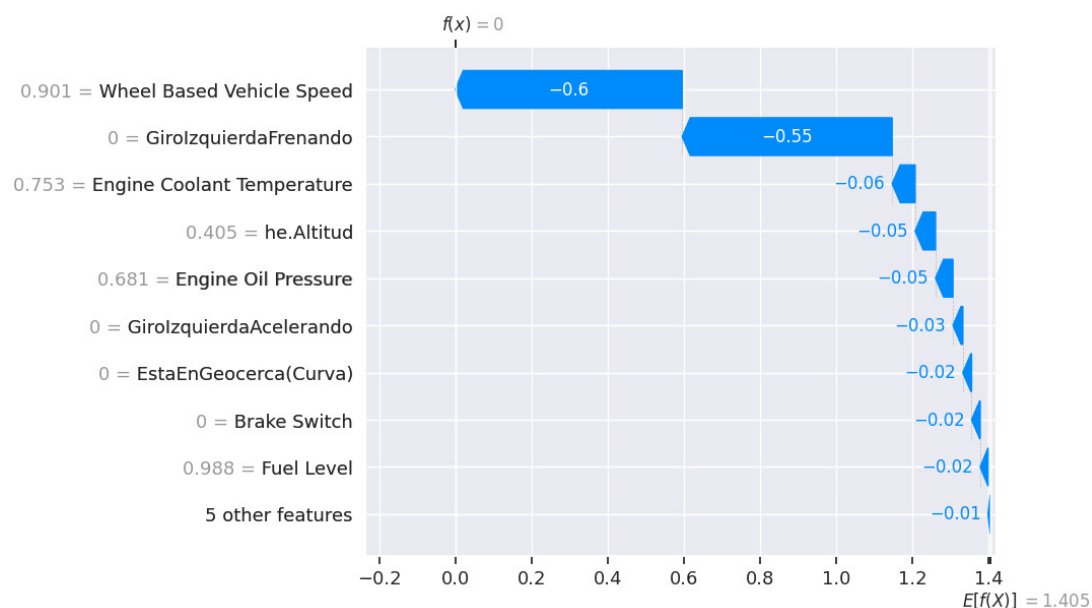
Característica	Conducción Normal	Conducción Agresiva	Conducción Peligrosa
Velocidad basada en ruedas	Negativa	Positiva	Positiva
Giro Izquierda Frenando	Negativa	Negativa	Positiva
Radio	Negativa	Negativa	Positiva
Nivel de Combustible	Negativa	Negativa	Positiva
Tipo de Zona	Negativa	Positiva	Positiva
Giro Izquierda Acelerando	Negativa	Negativa	Positiva
En Curva	Negativa	Positiva	Positiva
Pedal de Freno	Negativa	Negativa	Positiva
Altitud	Negativa	Positiva	Positiva

Presión de aceite	Negativa	Positiva	Positiva
Estado de Carga	Negativa	Negativa	Positiva
Temperatura Refrigerante	Negativa	Negativa	Positiva
Tiempo de Manejo	Negativa	Negativa	Positiva
Rompe Muelle	Negativa	Negativa	Positiva

En la Figura 22 se realiza un análisis de la clase “*Conducta Normal*” para lo cual se ha tomado un registro para realizar el análisis de explicabilidad de como aportan las variables. Se puede observar que las variables Velocidad basada en ruedas y Giro Izquierda Frenando aportan negativamente a la predicción del modelo.

Figura 22

Análisis de impacto de variables respecto a la clase “Conducta Normal”

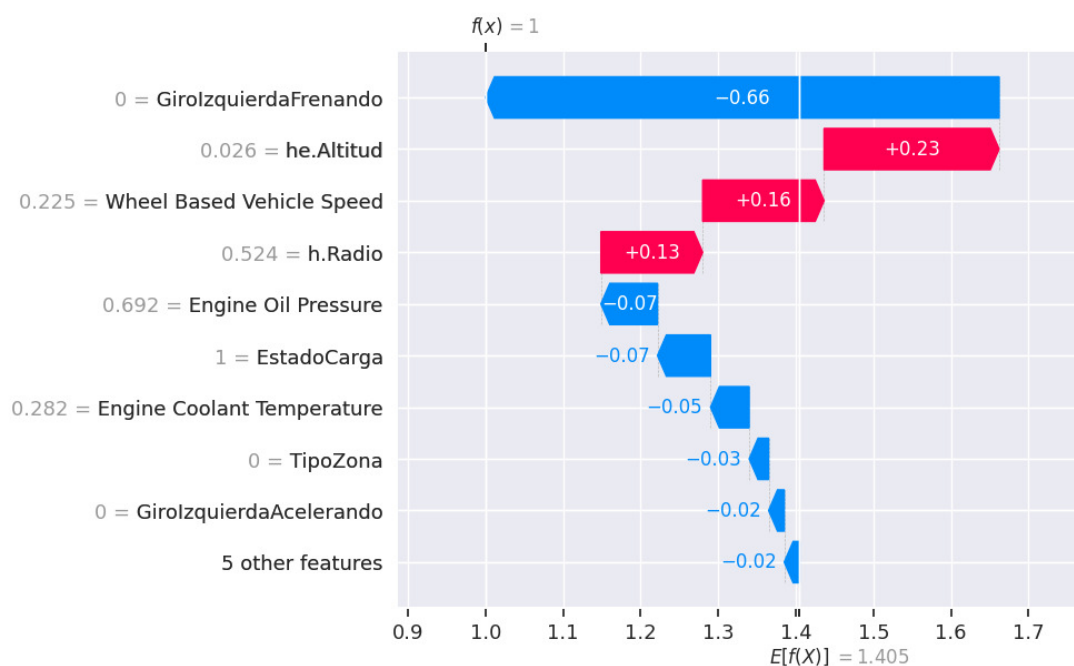


En la Figura 23, se puede ver que las características 'Altitud' y 'Velocidad basada en la rueda' ayudan a que el modelo prediga la clase 'Conducta agresiva', lo que significa que cuando estos valores son altos, hay una mayor probabilidad de que se presente una conducción agresiva. Por otro lado, la

característica 'Giro Izquierda Frenando' tiene un efecto negativo, lo que sugiere que cuando esta situación se presenta, disminuye la probabilidad de que el modelo considere que hay una conducta agresiva.

Figura 23

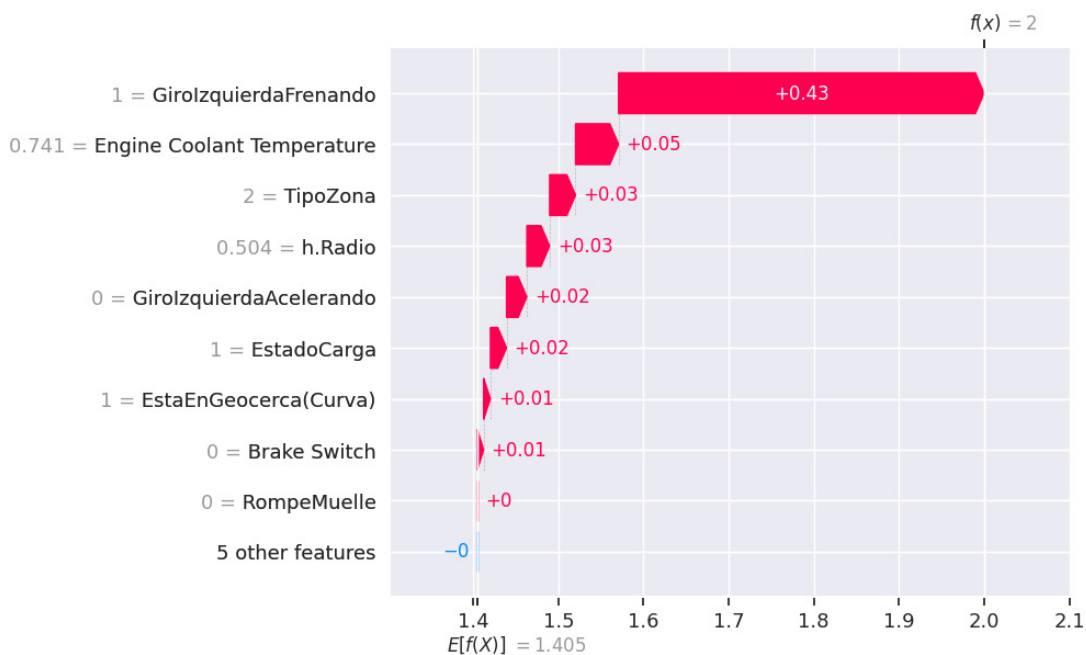
Análisis de impacto de variables respecto a la clase “Conducta Agresiva”



En la Figura 24 se realiza un análisis de la clase “Conducta Peligrosa” para lo cual se ha tomado un registro para realizar el análisis de explicabilidad de como aportan las variables. Se puede observar que las variables Giro Izquierda Frenando es la que aporta significativamente a la predicción del modelo. Esto significa que cuando estas características están presentes, el modelo tiene más probabilidades de predecir un resultado de riesgo de accidente.

Figura 24

Análisis de impacto de variables respecto a la clase “Conducta Peligrosa”



5.2. Prueba de hipótesis

5.2.1. Hipótesis específica 1

Para validar la hipótesis planteada, se llevó a cabo un estudio naturalista que permitió la recolección de datos de telemetría en vehículos pesados del sector minero, en carreteras rurales. A través de un riguroso proceso de análisis y modelado, se pudo confirmar que los datos obtenidos fueron adecuados para desarrollar un modelo de predicción del riesgo de accidentes basado en aprendizaje automático.

En primer lugar, se recolectaron más de 57,000 registros con 50 características iniciales, que luego de un proceso de limpieza y eliminación de columnas irrelevantes, se redujeron a 14 variables. Estas características incluyeron parámetros clave como el comportamiento del conductor (frenadas bruscas, giros bruscos, aceleraciones), las condiciones del vehículo (presión del aceite, velocidad basada en las ruedas, temperatura del motor) y las condiciones del entorno (altitud, tipo de vía).

Para evaluar la calidad en la predicción, se entrenó un modelo de regresión logística multiclase, Random Forest y XGBoost. Estos modelos demostraron

una buena capacidad predictiva, con métricas satisfactorias, como una alta exactitud y sensibilidad; lo cual indica que el conjunto de datos era representativo y robusto.

Adicionalmente, se realizó un análisis de coeficientes utilizando la prueba de Wald y el p-value para evaluar la significancia de las variables. Los resultados revelaron que variables como la altitud, la posición en curva y la velocidad del vehículo basada en las ruedas eran altamente significativas, lo que refuerza la relevancia de los datos recolectados para predecir eventos de riesgo.

Por lo tanto, la hipótesis queda validada al demostrar que el conjunto de datos generado es representativo, de calidad y adecuado para el desarrollo de modelos predictivos basados en aprendizaje automático que explican de manera significativa el riesgo de accidentes en vehículos pesados del sector minero en carretera rural.

5.2.2. Hipótesis específica 2

Para realizar la prueba de la hipótesis se realizaron 3 pruebas para los modelos individuales y ensamblados. Las pruebas son T de Student, Wilcoxon y ANOVA (ver Tabla 34):

Tabla 34

Resultado de evaluación comparativa de modelos

Métrica	Prueba	Grupo	Media (%)	Desviación Estándar (%)	t-statistic/ W/F-statistic	p-value	Resultado
Exactitud	T de Student	Modelos Individuales	71.27	24.78	-2.18	0.049	Diferencia significativa
		Modelos Ensamblados	92.70	3.82			
	Prueba de Wilcoxon	Modelos Individuales	71.27	24.78	W = 3	0.046	Diferencia significativa
		Modelos Ensamblados	92.70	3.82			

		Modelo s Ensam blados	92.70	3.82			
	ANOVA	Modelo s Individu ales	71.27	24.78	F = 10.93	0.0 05	Diferen cia significa tiva
		Modelo s Ensam blados	92.70	3.82			
Sensib ilidad	Prueba T de Student	Modelo s Individu ales	72.37	8.21	-3.67	0.0 03	Diferen cia significa tiva
		Modelo s Ensam blados	81.80	0			
	Prueba de Wilcoxo n	Modelo s Individu ales	72.37	8.21	W = 3	0.0 04	Diferen cia significa tiva
		Modelo s Ensam blados	81.80	0			
	ANOVA	Modelo s Individu ales	72.37	8.21	F = 8.34	0.0 13	Diferen cia significa tiva
		Modelo s Ensam blados	81.80	0			

Todas las pruebas realizadas (T de Student, Wilcoxon y ANOVA) muestran diferencias significativas entre los modelos individuales y ensamblados para ambas métricas, exactitud y sensibilidad.

La consistencia en los resultados a través de diferentes pruebas estadísticas refuerza la conclusión de que los modelos ensamblados no solo mejoran la exactitud y consistencia de las predicciones, sino que también

proporcionan un rendimiento significativamente superior en comparación con los modelos individuales.

5.2.3. Hipótesis específica 3

Se enfoca en la importancia de la evaluación de la eficiencia predictiva del modelo ajustado para la identificación precisa del riesgo de accidentes vehiculares en vehículos pesados del sector minero, usando un conjunto de datos independiente. Esta evaluación es muy importante de tal manera que confirme que modelo tiene altos niveles de exactitud y explicabilidad.

Evaluación de Modelos Individuales. Al observar los resultados de los modelos individuales, se destacan los siguientes desempeños:

- XGBoost: Con parámetros optimizados como alpha: 0.01, learning_rate: 0.01, max_depth: 20 y n_estimators: 400, logra exactitud del 99% y sensibilidad del 72%.
- LightGBM: Con learning_rate: 0.01, max_depth: 16, n_estimators: 500 y num_leaves: 25, también alcanzó una exactitud del 99% y una sensibilidad del 72%.
- CatBoost: Con depth: 16, iterations: 150, l2_leaf_reg: 0.1 y learning_rate: 0.01, presentó una exactitud del 97% y una sensibilidad del 67%. Sin embargo, cuando se ajusta los parámetros a iterations: 250, la sensibilidad llega al 72%, con una exactitud del 98%.

A pesar de la alta exactitud de estos modelos, la sensibilidad, que mide la capacidad del modelo para identificar correctamente las instancias de riesgo de accidentes, fue 72%.

Desempeño del Modelo Ensamblado. El modelo ensamblado mejora los resultados al combinar los modelos individuales. Se llega a una exactitud de 99% y una sensibilidad mejorada al 81%. Esta mejora en sensibilidad es muy importante, ya que significa una mayor capacidad para identificar situaciones de alto riesgo, lo que ayuda para prevenir los accidentes.

Explicabilidad del Modelo. El análisis de explicabilidad muestra que algunas características influyen de manera significativa en las predicciones del modelo. Por ejemplo:

En la clase "Conducta Agresiva" las variables 'Altitud' y 'Velocidad basada en la rueda' contribuyen positivamente, esto significa que altitudes y velocidades altas aumentan la probabilidad de conducta agresiva. Sin embargo, 'Giro Izquierda Frenando' tiene un efecto negativo, lo que reduce la probabilidad de la clase.

En el caso de la clase "Conducta Peligrosa", 'Giro Izquierda Frenando' es un factor significativo que incrementa la probabilidad de riesgo, lo cual va de la mano con el comportamiento común de adelantamiento que se da en carreteras rurales.

En la clase "Conducta Normal", las variables 'Velocidad basada en ruedas' y 'Giro Izquierda Frenando' contribuyen negativamente, reduciendo la probabilidad de un riesgo de accidente. Esto ayuda para diferenciar entre conductas normales y riesgosas.

Por lo tanto, la hipótesis se valida por la mejora en sensibilidad alcanzada por el modelo ensamblado (del 72% al 81%), lo cual permite identificar correctamente situaciones de alto riesgo en comparación con los modelos individuales. Además, se valida por que se tiene una alta exactitud del modelo ensamblado (99%), garantizando un rendimiento bastante sólido. El análisis de explicabilidad también refuerza esto, al mostrar cómo las características como 'Altitud', 'Velocidad basada en la rueda' y 'Giro Izquierda Frenando' influyen en las predicciones. Esto asegura que el modelo es preciso y eficiente en sus predicciones, y además comprensible para la toma de decisiones de seguridad en el sector minero.

5.2.4. Hipótesis general

La hipótesis general plantea que el desarrollo de un modelo de predicción basado en aprendizaje automático puede explicar significativamente el riesgo de accidentes vehiculares debido a la conducta de manejo en vehículos pesados en carretera rural. Esta afirmación se

demuestra por la capacidad de los modelos de aprendizaje automático para analizar grandes volúmenes de datos de telemetría que son obtenidos en el mundo real con variables que están asociadas al comportamiento del conductor, como giro a la izquierda frenando, velocidad, tipo de zona, tiempo de manejo continuo, y factores del vehículo, como la temperatura del refrigerante y presión del aceite. Estos datos permiten construir un modelo que, al identificar estos patrones de conducción, puedan clasificar el riesgo en niveles como normal, riesgoso y peligroso. Una vez clasificado se puede proveer explicaciones sobre el nivel de influencia de cada característica. Las explicaciones que se realizan con los factores, incrementan la forma de interpretar el resultado del modelo lo que ayuda para que se implementen medidas preventivas en carreteras.

CONCLUSIONES

El estudio ha demostrado que el desarrollo de un modelo de predicción basado en aprendizaje automático puede explicar de manera significativa el riesgo de accidentes vehiculares en vehículos pesados del sector minero en carretera rural.

En relación con OE1/HE1, se confirma que la recolección de más de 57,000 registros (reduciéndose a 14 variables clave) mediante un estudio naturalista proporcionó un conjunto de datos robusto y representativo para el desarrollo del modelo de predicción del riesgo de accidentes en vehículos pesados del sector minero en carretera rural. Respecto a OE2/HE2, el diseño del modelo ensamblado, que combina algoritmos como XGBoost, LightGBM y CatBoost, permitió aumentar la sensibilidad del 72% de los modelos individuales al 81% en el modelo ensamblado, alcanzando una exactitud del 99%, lo cual demuestra la optimización del rendimiento predictivo. Finalmente, en cuanto a OE3/HE3, la evaluación con un conjunto de datos independiente, respaldada por pruebas estadísticas (T de Student, Wilcoxon y ANOVA) y un análisis de explicabilidad que resalta variables significativas como altitud y velocidad basada en la rueda, ratifica que el modelo es preciso, explicable y aplicable para mejorar la seguridad vial en el sector minero.

Este estudio aporta al sector minero proporcionando un modelo que puede usarse en las estrategias de gestión de seguridad vial en carreteras rurales. La integración de modelos de predicción del riesgo con los datos de telemetría, permite que las empresas mineras puedan mejorar la seguridad de sus operaciones de traslado y reducir el número de accidentes, lo que a su vez puede ayudar en reducción de costos asociados a accidentes y daños.

RECOMENDACIONES

Se recomienda que empresas del sector minero y otras áreas del transporte pesado implementen sistemas de telemetría en sus flotas de vehículos e implementen alertas en tiempo real basadas en los resultados del modelo. Esta tecnología permite obtener datos en tiempo real sobre el comportamiento de los conductores y las condiciones del vehículo y esto ayuda para prevenir riesgos y tomar de decisiones de manera oportuna. Esta integración contribuiría en mejorar la seguridad vial.

También, se recomienda que, para ampliar la validez externa del modelo, futuras investigaciones incorporen variables de gestión adicionales que puedan influir en el riesgo de accidentes y consideren una mayor diversidad en tipos de vías y contextos del transporte pesado.

Es importante que las empresas capaciten a sus conductores en la importancia de la recolección de datos de telemetría y de esta manera evitar los boicots que a veces realizan los propios conductores a los diferentes dispositivos que se usan. El uso de estas tecnologías puede aprovecharse solo si los conductores comprenden su relevancia y los beneficios asociados a la mejora continua del modelo.

El modelo desarrollado en este estudio puede ser aplicable a una variedad de entornos. Por este motivo, se debe fomentar la colaboración entre empresas del sector privado, instituciones académicas y autoridades gubernamentales para desarrollar políticas y regulaciones que promuevan el uso de la telemetría y los modelos predictivos de accidentes.

REFERENCIAS BIBLIOGRÁFICAS

AASHTO. (2010). *Highway Safety Manual*. American Association of State Highway and Transportation Officials.

Alkinani, M. H., Khan, W. Z., & Arshad, Q. (2020). Detecting Human Driver Inattentive and Aggressive Driving Behavior Using Deep Learning: Recent Advances, Requirements and Open Challenges. *IEEE Access*, 8, 105008-105030.
<https://doi.org/10.1109/ACCESS.2020.2999829>

Alrajhi, M., & Kamel, M. (2019). A Deep-Learning Model for Predicting and Visualizing the Risk of Road Traffic Accidents in Saudi Arabia: A Tutorial Approach. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(11), Article 11.
<https://doi.org/10.14569/IJACSA.2019.0101166>

Al-refai, G., Elmoaqet, H., & Ryalat, M. (2022). In-Vehicle Data for Predicting Road Conditions and Driving Style Using Machine Learning. *Applied Sciences (Switzerland)*, 12(18). <https://doi.org/10.3390/app12188928>

Basso, F., Pezo, R., Varas, M., & Villalobos, M. (2021). A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. *Accident Analysis & Prevention*, 162, 106409. <https://doi.org/10.1016/j.aap.2021.106409>

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
<https://doi.org/10.1023/A:1010933404324>

Brühwiler, L., Fu, C., Huang, H., Longhi, L., & Weibel, R. (2022). Predicting individuals' car accident risk by trajectory, driving events, and geographical context. *Computers, Environment and Urban Systems*, 93, 101760.
<https://doi.org/10.1016/j.compenvurbsys.2022.101760>

Celaya-Padilla, J. M., Galván-Tejada, C. E., Lozano-Aguilar, J. S. A., Zanella-Calzada, L. A., Luna-García, H., Galván-Tejada, J. I., Gamboa-Rosales, N. K., Rodríguez, A. V., & Gamboa-Rosales, H. (2019). «Texting & driving» detection using deep convolutional

neural networks. *Applied Sciences (Switzerland)*, 9(15).

<https://doi.org/10.3390/app9152962>

Chipman, M. L., & Morgan, P. (1975). The role of driver demerit points and age in the prediction of motor vehicle collisions. *Journal of Epidemiology & Community Health*, 29(3), 190-195. <https://doi.org/10.1136/jech.29.3.190>

Das, S., Tsapakis, I., & Khodadadi, A. (2021). Safety performance functions for low-volume rural minor collector two-lane roadways. *IATSS Research*, 45(3), 347-356. Scopus. <https://doi.org/10.1016/j.iatssr.2021.02.004>

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Proceedings of the First International Workshop on Multiple Classifier Systems*, 1-15.

Elassad, Z. E. A., Mousannif, H., & Moatassime, H. A. (2020). A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems. *Transportation Research Part C: Emerging Technologies*, 118. <https://doi.org/10.1016/j.trc.2020.102708>

Fridman, L., Brown, D. E., Glazer, M., Angell, W., Dodd, S., Jenik, B., Terwilliger, J., Patsekin, A., Kindelsberger, J., Ding, L., Seaman, S., Mehler, A., Sipperley, A., Pettinato, A., Seppelt, B. D., Angell, L., Mehler, B., & Reimer, B. (2019). MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation. *IEEE Access*, 7, 102021-102038. IEEE Access. <https://doi.org/10.1109/ACCESS.2019.2926040>

Galatioto, F., Catalano, M., Shaikh, N., McCormick, E., & Johnston, R. (2018). Advanced accident prediction models and impacts assessment. *IET Intelligent Transport Systems*, 12(9), 1131-1141. <https://doi.org/10.1049/iet-its.2018.5218>

Geng, Z., Ji, X., Cao, R., Lu, M., & Qin, W. (2022). A Conflict Measures-Based Extreme Value Theory Approach to Predicting Truck Collisions and Identifying High-Risk Scenes on

Two-Lane Rural Highways. *Sustainability (Switzerland)*, 14(18).

<https://doi.org/10.3390/su141811212>

Guido, G., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S., Vitale, A., Astarita, V., Park, Y., & Geem, Z. W. (2022). Evaluation of Contributing Factors Affecting Number of Vehicles Involved in Crashes Using Machine Learning Techniques in Rural Roads of Cosenza, Italy. *Safety*, 8(2), Article 2. <https://doi.org/10.3390/safety8020028>

Halim, Z., Sulaiman, M., Waqas, M., & Aydın, D. (2023). Deep neural network-based identification of driving risk utilizing driver dependent vehicle driving features: A scheme for critical infrastructure protection. *Journal of Ambient Intelligence and Humanized Computing*, 14(9), 11747-11765. <https://doi.org/10.1007/s12652-022-03734-y>

Harb, R., Radwan, E., Yan, X., Pande, A., & Abdel-Aty, M. (2008). Freeway Work-Zone Crash Analysis and Risk Identification Using Multiple and Conditional Logistic Regression. *Journal of Transportation Engineering*, 134(5), 203-214. [https://doi.org/10.1061/\(ASCE\)0733-947X\(2008\)134:5\(203\)](https://doi.org/10.1061/(ASCE)0733-947X(2008)134:5(203))

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la Investigación* (6ta México, D.F.). McGraw-Hill.

Hu, J., Huang, M.-C., & Yu, X. (2020). Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. *Accident Analysis & Prevention*, 144, 105665. <https://doi.org/10.1016/j.aap.2020.105665>

Hu, Z., Zhou, J., & Zhang, E. (2023). Improving Traffic Safety through Traffic Accident Risk Assessment. *Sustainability (Switzerland)*, 15(4). <https://doi.org/10.3390/su15043748>

Jacobé de Naurois, C., Bourdin, C., Bougard, C., & Vercher, J.-L. (2018). Adapting artificial neural networks to a specific driver enhances detection and prediction of

drowsiness. *Accident Analysis & Prevention*, 121, 118-128.

<https://doi.org/10.1016/j.aap.2018.08.017>

Jiang, F., Yuen, K. K. R., & MingLee, E. W. (2020). A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. *Accident Analysis & Prevention*, 141, 105520.

<https://doi.org/10.1016/j.aap.2020.105520>

Kashevnik, A., Shchedrin, R., Kaiser, C., & Stocker, A. (2021). Driver Distraction Detection Methods: A Literature Review and Framework. *IEEE Access*, 9, 60063-60076.

<https://doi.org/10.1109/ACCESS.2021.3073599>

Kuncheva, L. L. (2004). Classifier Selection. En *Combining Pattern Classifiers* (pp. 189-202).

John Wiley & Sons, Ltd. <https://doi.org/10.1002/0471660264.ch6>

Lacherre, J., Castillo-Sequera, J. L., & Mauricio, D. (2024). Factors, Prediction, and Explainability of Vehicle Accident Risk Due to Driving Behavior through Machine Learning: A Systematic Literature Review, 2013–2023. *Computation*, 12(7), Article 7. <https://doi.org/10.3390/computation12070131>

Lattanzi, E., & Freschi, V. (2021). Machine Learning Techniques to Identify Unsafe Driving Behavior by Means of In-Vehicle Sensor Data. *Expert Systems with Applications*, 176. <https://doi.org/10.1016/j.eswa.2021.114818>

Li, D., Wang, Y., & Xu, W. (2023). A Deep Multichannel Network Model for Driving Behavior Risk Classification. *IEEE Transactions on Intelligent Transportation Systems*, 24(1), 1204-1219. <https://doi.org/10.1109/TITS.2022.3201378>

Liu, Z., Ren, S., & Peng, M. (2021). Identification of Driver Distraction Based on SHRP2 Naturalistic Driving Study. *Mathematical Problems in Engineering*, 2021. <https://doi.org/10.1155/2021/6699327>

Ma, J., Ding, Y., Cheng, J. C. P., Tan, Y., Gan, V. J. L., & Zhang, J. (2019). Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City

Management Perspective. *IEEE Access*, 7, 148059-148072. IEEE Access.

<https://doi.org/10.1109/ACCESS.2019.2946401>

Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis & Prevention*, 160, 106322. <https://doi.org/10.1016/j.aap.2021.106322>

Masello, L., Castignani, G., Sheehan, B., Guillen, M., & Murphy, F. (2023). Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis and Prevention*, 184.

<https://doi.org/10.1016/j.aap.2023.106997>

Naciones Unidas. (2018). *La Agenda 2030 y los Objetivos de Desarrollo Sostenible: Una oportunidad para América Latina y el Caribe*. Comisión Económica para América Latina y el Caribe (CEPAL).

<https://bibliotecadigital.aecid.es/bibliodig/es/consulta/registro.do?control=ES-MAAEC20190011211>

Nikolaou, D., Ziakopoulos, A., Dragomanovits, A., Roussou, J., & Yannis, G. (2023). Comparing Machine Learning Techniques for Predictions of Motorway Segment Crash Risk Level. *Safety*, 9(2). <https://doi.org/10.3390/safety9020032>

Niu, Y., Li, Z. M., & Fan, Y. X. (2021). Analysis of truck drivers' unsafe driving behaviors using four machine learning methods. *International Journal of Industrial Ergonomics*, 86. <https://doi.org/10.1016/j.ergon.2021.103192>

Nour, M. K., Naseer, A., Alkazemi, B., & Jamil, M. A. (2020). Road Traffic Accidents Injury Data Analytics. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(12), Article 12.

<https://doi.org/10.14569/IJACSA.2020.0111287>

Observatorio Nacional de Seguridad Vial. (2024). *Información de datos abiertos*.

Información de datos abiertos. <https://www.onsv.gob.pe/datosabiertos>

Open-Meteo.com. (2024, mayo 29).  API meteorológica histórica | Open-Meteo.com.

<https://open-meteo.com/en/docs/historical-weather-api>

Organización Mundial de la Salud. (2023). *Global status report on road safety 2023*.

<https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023>

Orsini, F., Gecchele, G., Rossi, R., & Gastaldi, M. (2021). A conflict-based approach for real-time road safety analysis: Comparative evaluation with crash-based models.

Accident Analysis & Prevention, 161, 106382.

<https://doi.org/10.1016/j.aap.2021.106382>

Orsini, F., Gecchele, G., Rossi, R., & Gastaldia, M. (2021). A conflict-based approach for real-time road safety analysis: Comparative evaluation with crash-based models.

Accident Analysis & Prevention, 161, 106382.

<https://doi.org/10.1016/j.aap.2021.106382>

Panagopoulos, G., & Pavlidis, I. (2020). Forecasting Markers of Habitual Driving Behaviors Associated With Crash Risk. *IEEE Transactions on Intelligent Transportation Systems*,

21(2), 841-851. IEEE Transactions on Intelligent Transportation Systems.

<https://doi.org/10.1109/TITS.2019.2910157>

Papadimitriou, E., Argyropoulou, A., Tselentis, D. I., & Yannis, G. (2019). Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving.

Safety Science, 119, 91-97. <https://doi.org/10.1016/j.ssci.2019.05.059>

Peng, L., Wang, Y., Zhang, F., Zhang, J., & Li, Z. (2021). Evaluation of emergency driving behaviour and vehicle collision risk in connected vehicle environment: A deep learning approach. *IET Intelligent Transport Systems*, 15(4), 584-594.

<https://doi.org/10.1049/itr2.12053>

Plan Estratégico Nacional de Seguridad Vial PENsv 2017-2021, DECRETO SUPREMO N° 019-2017-MTC (2017). https://www.mtc.gob.pe/cnsv/documentos/PENsv_2017-2021.pdf

Real Academia Española. (2024). *Diccionario de la lengua española*. <https://dle.rae.es/>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
<https://doi.org/10.1145/2939672.2939778>

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4), e1249. <https://doi.org/10.1002/widm.1249>

Santos, K., P.Dias, J., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, 80, 254-269.
<https://doi.org/10.1016/j.jsr.2021.12.007>

Sarker, S., Haque, Md. M., & Dewan, M. A. A. (2021). *Driving Maneuver Classification Using Domain Specific Knowledge and Transfer Learning*.
<https://ieeexplore.upc.elogim.com/document/9455147/>

Seo, H., Shin, J., Kim, K.-H., Lim, C., & Bae, J. (2022). Driving Risk Assessment Using Non-Negative Matrix Factorization with Driving Behavior Records. *IEEE Transactions on Intelligent Transportation Systems*, 23(11), 20398-20412.
<https://doi.org/10.1109/TITS.2022.3193125>

Shangguan, Q., Fu, T., Wang, J., Fang, S., & Fu, L. (2022). A proactive lane-changing risk prediction framework considering driving intention recognition and different lane-changing patterns. *Accident Analysis and Prevention*, 164.
<https://doi.org/10.1016/j.aap.2021.106500>

- Shi, L., Qian, C., & Guo, F. (2022). Real-time driving risk assessment using deep learning with XGBoost. *Accident Analysis and Prevention*, 178.
<https://doi.org/10.1016/j.aap.2022.106836>
- Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129, 170-179. <https://doi.org/10.1016/j.aap.2019.05.005>
- Silva, I., & Eugenio Naranjo, J. (2020). A Systematic Methodology to Evaluate Prediction Models for Driving Style Classification. *Sensors*, 20(6), Article 6.
<https://doi.org/10.3390/s20061692>
- Tola, A. M., Demissie, T. A., Saathoff, F., & Gebissa, A. (2022). Crash Distribution Dataset: Development and Validation for the Undivided Rural Roads in Oromia, Ethiopia. *Transport and Telecommunication Journal*, 23(1), 11-24.
<https://doi.org/10.2478/ttj-2022-0002>
- Wang, H., Gu, M., Wu, S., & Wang, C. (2020). A driver's car-following behavior prediction model based on multi-sensors data. *EURASIP Journal on Wireless Communications and Networking*, 2020(1), 10. <https://doi.org/10.1186/s13638-020-1639-2>
- Wang, H., Wang, X., Han, J., Xiang, H., Li, H., Zhang, Y., & Li, S. (2022). A Recognition Method of Aggressive Driving Behavior Based on Ensemble Learning. *Sensors*, 22(2).
<https://doi.org/10.3390/s22020644>
- Wang, J., Xu, W., Fu, T., & Jiang, R. (2022). Recognition of Trip-Based Aggressive Driving: A System Integrated With Gaussian Mixture Model Structured of Factor-Analysis, and Hierarchical Clustering. *IEEE Transactions on Intelligent Transportation Systems*, 23(11), 20442-20451. <https://doi.org/10.1109/TITS.2022.3175528>
- Xu, W., Wang, J., Fu, T., Gong, H., & Sobhani, A. (2022). Aggressive driving behavior prediction considering driver's intention based on multivariate-temporal feature

data. *Accident Analysis and Prevention*, 164.

<https://doi.org/10.1016/j.aap.2021.106477>

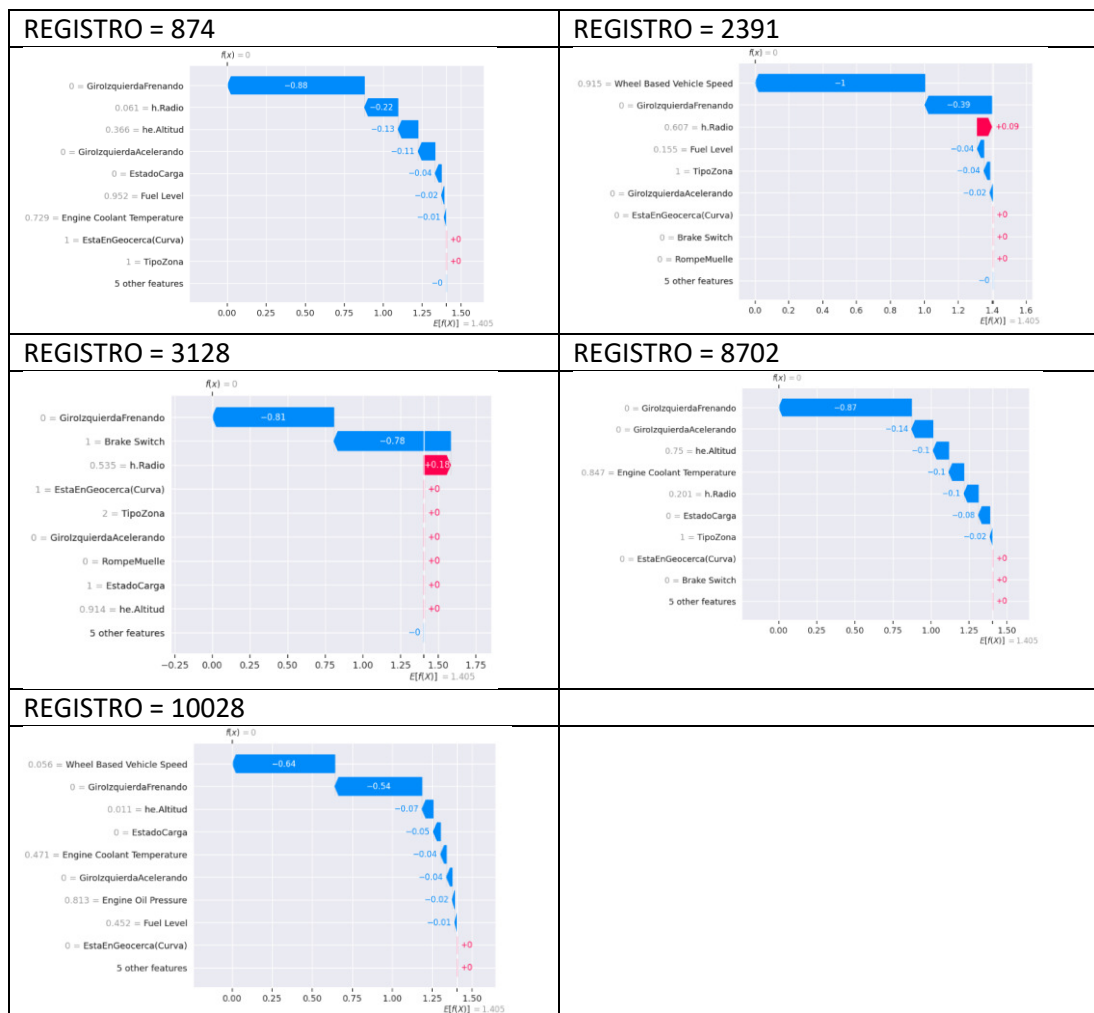
Yang, C., Chen, M. Y., & Yuan, Q. (2021). The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. *Accident Analysis & Prevention*, 158. <https://doi.org/10.1016/j.aap.2021.106153>

Zhao, L., Xu, T., Zhang, Z., & Hao, Y. (2022). Lane-Changing Recognition of Urban Expressway Exit Using Natural Driving Data. *Applied Sciences (Switzerland)*, 12(19). <https://doi.org/10.3390/app12199762>

Zhong, S., Fu, X., Lu, W., Tang, F., & Lu, Y. (2022). An Expressway Driving Stress Prediction Model Based on Vehicle, Road and Environment Features. *IEEE Access*, 10, 57212-57226. <https://doi.org/10.1109/ACCESS.2022.3165570>

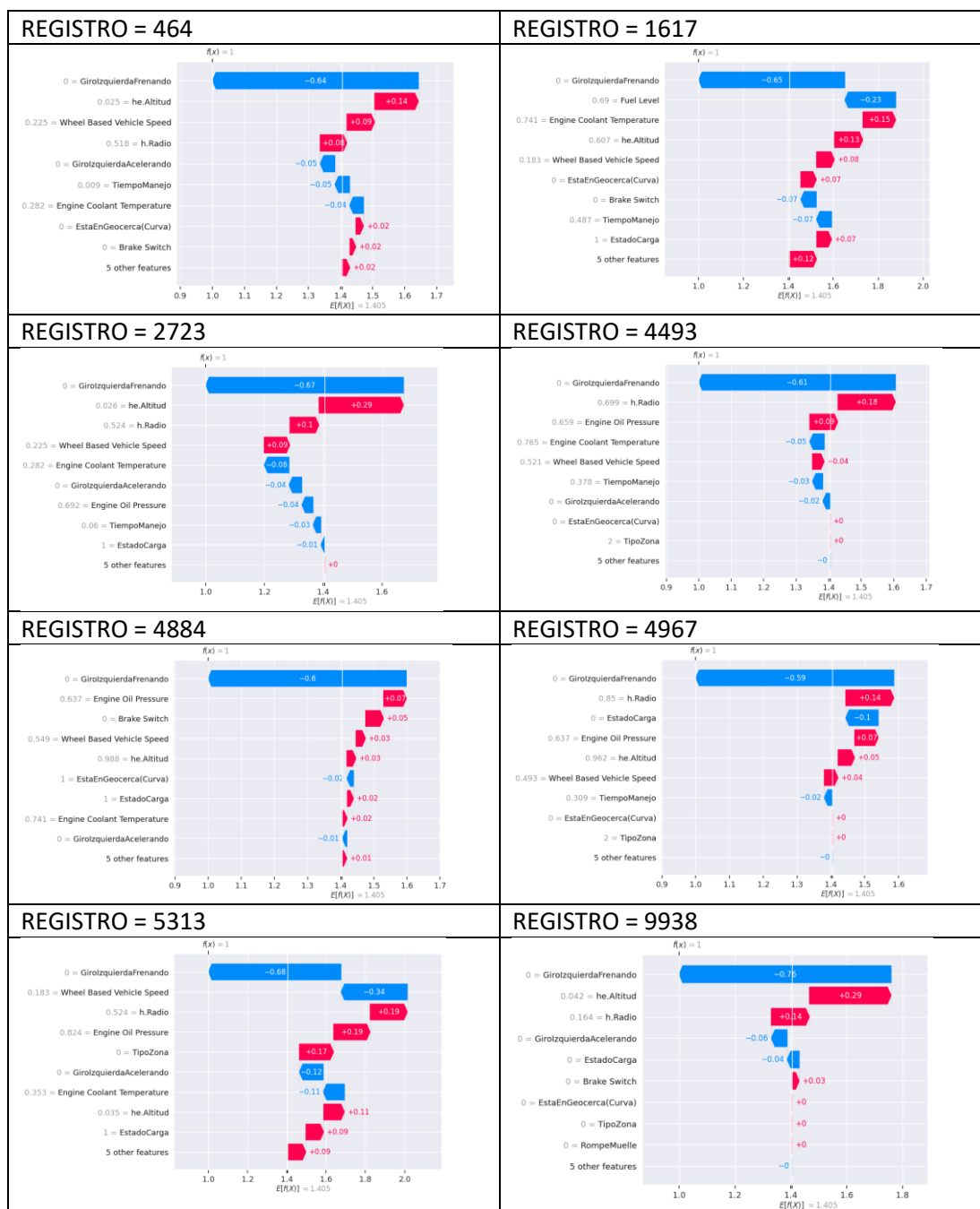
ANEXO A

EXPLICACIÓN DE CLASE “0” - NORMAL



ANEXO B

EXPLICACIÓN DE CLASE “1” – AGRESIVO



ANEXO C

EXPLICACIÓN DE CLASE “2” – PELIGROSO

