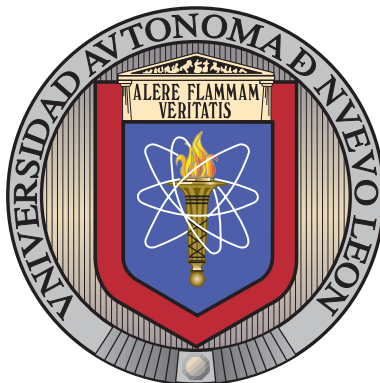


UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA

SUBDIRECCIÓN DE ESTUDIOS DE POSGRADO



AUMENTO DE LA SEGURIDAD DE DISTRIBUCIÓN
A TRAVÉS DE LA DEFINICIÓN DE VARIABLES DE
RIESGO PRESENTES EN EL TRANSPORTE DE
CARGA PESADA

POR

ARELY ISMERAI HERNÁNDEZ GONZÁLEZ

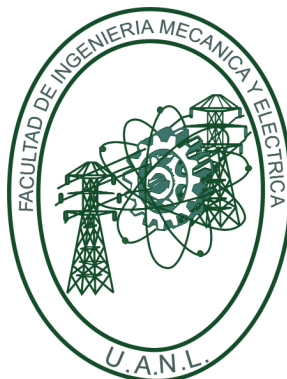
COMO REQUISITO PARA OBTENER EL GRADO DE
MAESTRÍA EN LOGÍSTICA Y CADENA DE SUMINISTRO

FEBRERO 2023

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA

SUBDIRECCIÓN DE ESTUDIOS DE POSGRADO



AUMENTO DE LA SEGURIDAD DE DISTRIBUCIÓN
A TRAVÉS DE LA DEFINICIÓN DE VARIABLES DE
RIESGO PRESENTES EN EL TRANSPORTE DE
CARGA PESADA

POR

ARELY ISMERAI HERNÁNDEZ GONZÁLEZ

COMO REQUISITO PARA OBTENER EL GRADO DE
MAESTRÍA EN LOGÍSTICA Y CADENA DE SUMINISTRO

FEBRERO 2023

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
Facultad de Ingeniería Mecánica y Eléctrica
Posgrado

Los miembros del Comité de Evaluación de Tesis recomendamos que la Tesis "Aumento de la seguridad de distribución a través de la definición de variables de riesgo presentes en el transporte de carga pesada", realizada por el/la estudiante Arely Ismerai Hernández González, con número de matrícula 2037580, sea aceptada para su defensa como requisito para obtener el grado de Maestría en Logística y Cadena de Suministro.

El Comité de Evaluación de Tesis

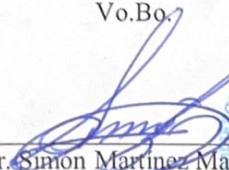
Dr. Tomás Eloy Salais Fierro
Director

Dra. Jania Astrid Saucedo Martínez
Revisor

MLyCS Blanca Idalia Pérez Pérez
Revisor

Dr. José Daniel Mosquera Artamanov
Revisor

Vo.Bo.


Dr. Simon Martinez Martinez
Subdirector de Estudios de Posgrado



Institución 190001

Programa 642597

Acta Núm. 4130

Ciudad Universitaria, a 15 de febrero 2023.

*Aquí puedes poner tu dedicatoria
si es que tienes una.*

*Si no tienes una, puedes borrar
la línea `\include{Dedicatoria}` en el
archivo **`MiTesis.tex`** pues no es obligatoria.*

ÍNDICE GENERAL

Agradecimientos	xiv
Resumen	xv
1. Introducción	1
1.1. Descripción del problema	2
1.2. Objetivo General	3
1.2.1. Objetivos Específicos	3
1.3. Hipótesis	4
1.4. Justificación de la investigación	4
1.5. Metodología	6
1.6. Estructura de tesis	7
2. Antecedentes	9
2.1. La función del transporte de carga en la cadena de suministro y la logística	9
2.2. Efectos sobre el transporte de carga pesada	13

2.3. Gestión de riesgos para el área de transporte	16
2.4. Acciones para mejorar la seguridad de las rutas (Estado del arte) . .	19
2.4.1. Propuestas de organismos internacionales para mejorar la se- guridad del transporte	19
2.4.2. El análisis del desempeño del conductor profesional	21
2.4.3. Estudios cualitativos	22
2.4.4. Estudios cuantitativos	25
2.4.5. Análisis de las herramientas	31
2.4.6. Herramientas para el análisis de datos	35
2.4.7. Minería de datos	39
2.4.8. Evaluación de desempeño de los modelos	41
2.5. Conclusiones del capítulo	41
3. Metodología	45
3.1. Descripción de la metodología	46
3.2. Fase 1: Definición de la problemática	48
3.3. Fase 2: Recopilación y pre-procesamiento de datos	49
3.3.1. Recopilación de datos	49
3.3.2. Pre-procesamiento de datos	51
3.4. Fase 3: Selección de la o las herramientas	54
3.4.1. Algoritmos de agrupación	56

3.4.2. Algoritmos de clasificación	56
3.5. Fase 4: Modelación y validación	58
3.5.1. Modelación	59
3.5.2. Validación de los modelos	60
3.5.3. Análisis de resultados	60
3.6. Conclusiones del capítulo	61
4. Experimentación y Análisis	63
4.1. Fase 1: Definición de problemática	64
4.2. Fase 2: Recopilación y pre-procesamiento de datos	64
4.2.1. Recopilación	64
4.2.2. Pre-procesamiento de datos	68
4.3. Fase 3: Selección de la o las herramientas	73
4.3.1. Normalización	74
4.3.2. Técnicas de agrupación	75
4.3.3. Problemáticas de desbalance de clases y semi-etiquetado . . .	78
4.4. Fase 4: Modelación y Validación	80
4.4.1. Modelación	80
4.4.2. Validación y Análisis	84
4.4.3. Principales factores de riesgo	85
4.4.4. Clasificación de conductores	87

4.5. Conclusiones del capítulo	89
5. Conclusiones	91
5.1. Conclusiones generales	91
5.2. Recomendaciones	94
5.3. Contribuciones y trabajo futuro	95
5.3.1. Trabajo futuro	96

ÍNDICE DE FIGURAS

1.1. Metodología de tesis	7
2.1. Total de carga movida entre los distintos medios de transporte en México. Fuente: (INEGI, 2018)	11
2.2. Criterios para la búsqueda de literatura	32
2.3. Relación entre estudios cualitativos y cuantitativos	33
2.4. Términos para el estudio del conductor. Fuente: Martinez <i>et al.</i> (2018); Ellassad <i>et al.</i> (2020)	34
2.5. Algoritmos de clasificación	36
2.6. Descripción de algoritmos. Fuente: (Constantinescu <i>et al.</i> , 2010; Junior <i>et al.</i> , 2017; Aggarwal, 2015)	37
2.7. Tipos de aprendizaje	39
2.8. Proceso de minería de datos. Fuente: (?Fayyad y Stolorz, 1997; Corso, 2007)	40
3.1. Metodología propuesta para análisis de conductores. Fuente: Adaptación de propuestas por (?Fayyad y Stolorz, 1997; Wu y Chaipiyaphan, 2020)	47

3.2. Proceso de recopilación de datos	50
3.3. Proceso de estudio de datos	52
3.4. Ejemplo de tratamiento de datos (Elaboración propia)	53
3.5. Proceso de etiquetas y clases (Elaboración propia)	55
3.6. Ventajas y desventajas de los clasificadores. Fuente: Pedamkar (2021); Garg (2018)	57
3.7. Tipos de modelos de clasificación	58
3.8. Proceso de experimentación y validación	59
3.9. Proceso de desarrollo de modelo de aprendizaje semi-supervisado . . .	60
3.10. Criterios para elegir la métrica de evaluación de modelos dependiendo del tipo de datos. (Borja-Robalino <i>et al.</i> , 2020; Corso, 2007; Badra, 2020)	61
4.1. Registro de datos para el análisis de conducción	66
4.2. Elección de variables para análisis	67
4.3. Cantidad de incidentes registrados por mes del año 2020	68
4.4. Registro por tipo de accidente (Año 2020)	69
4.5. Contabilización de frecuencias en reporte Driver Behaviour	69
4.6. Gráfica del desbalanceo de clases	71
4.7. Dataset con etiquetado de variable objetivo con valor = 1	71
4.8. Separación entre conjunto de datos de entrenamiento y prueba	73
4.9. Arriba: Datos originales. Abajo: Datos después de normalización. .	74

4.10. Gráfica para determinar el número de componentes que expliquen la variabilidad de los datos en un 80 %	76
4.11. Gráfica de la distorsión e inercia para el uso del Método del Codo . .	77
4.12. Incidentes dentro de los datos de entrenamiento	78
4.13. Ejemplo de ajuste y aprendizaje de los datos por el algoritmo de <i>LabelPropagation</i>	81
4.14. Ejemplo de ajuste y aprendizaje de los datos por el algoritmo de <i>RandomForestClassifier</i>	83
4.15. Gráfica de la Curva ROC	85
4.16. Diagrama de nivel de importancia por factor	86
4.17. Diagrama de clasificación de conductores	87
4.18. Lista de conductores clasificados en alto riesgo	88
4.19. Lista de incidentes reales registrados	88

ÍNDICE DE TABLAS

2.1. Resumen de revisión de literatura. (Cualitativo: 1 — Cuantitativo: 2)	32
2.2. Métricas de evaluación de algoritmos. Fuente: (Borja-Robalino <i>et al.</i> , 2020; Corso, 2007; Schoonjans, 2021; Badra, 2020)	42
3.1. Variables del entorno de conducción	50
3.2. Conceptos de minería de datos	54
4.1. Filtración y selección de datos	66
4.2. Resultado de matriz de correlaciones	72
4.3. Total de datos por tipo de conjunto	73
4.4. Resultado de clústers determinado por $k=7$	78
4.5. Resultado de transformación de datos con técnica de SMOTE y RandomUnderSampler	79
4.6. Resultado de eliminación de etiquetas en el 50 % de los datos	80
4.7. Resultados de Matriz de Confusión	81
4.8. Resultados de Matriz de Confusión en conjunto de datos de prueba	82

4.9. Resultados de Matriz de Confusión en conjunto de datos de prueba con modelo de Random Forest	83
4.10. Factores principales de riesgo por nivel importancia	85

AGRADECIMIENTOS

Aquí puedes poner tus agradecimientos. (No olvides agradecer a tu comité de tesis, a tus profesores, a la facultad y a CONACyT en caso de que hallas sido beneficiado con una beca).

RESUMEN

Arely Ismerai Hernández González.

Candidato para obtener el grado de Maestría en Logística y Cadena de Suministro.

Universidad Autónoma de Nuevo León.

Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio: AUMENTO DE LA SEGURIDAD DE DISTRIBUCIÓN A TRAVÉS DE
LA DEFINICIÓN DE VARIABLES DE RIESGO PRESENTES EN EL TRANSPORTE
DE CARGA PESADA.

Número de páginas: 103.

OBJETIVOS Y MÉTODO DE ESTUDIO: Aquí debes poner tus objetivos y métodos
de estudio. (Este es el formato).

CONTRIBUCIONES Y CONCLUSIONES: Y aquí tus contribuciones y conclusiones. (Tam-
bién es parte del formato).

Firma del asesor: _____
Dr. Tomás Eloy Salais Fierro

CAPÍTULO 1

INTRODUCCIÓN

El transporte de carga pesada es de los principales modos de transporte que se utilizan a nivel global. Parte de sus ventajas principales es el tener acceso a casi cualquier destino y por tal motivo se ha vuelto un elemento fundamental para la planeación y desarrollo de los sistemas logísticos. Para México la situación no es muy distinta porque a través de este modo de transporte es que se mueve gran parte de la carga nacional e internacional.

A raíz del aumento de la exigencia de dicha función, esto ha provocado que haya un creciente número de incidentes de tránsito que interrumpen el flujo de la cadena de suministro y al mismo tiempo pone en riesgo la vida de los conductores profesionales. Por lo tanto, hay un mayor interés por identificar cuáles son los factores que provocan las vulnerabilidades dentro de los sistemas logísticos, cómo obtener y analizar la información y al final cómo poder crear una mejor gestión de riesgos.

En este primer capítulo se pretende introducir la relevancia del tema, presentando con datos respecto al nivel de accidentabilidad, la importancia de la identificación de los factores que aumentan el riesgo de accidente y la necesidad de atacar esta problemática por el bien del sistema.

1.1 DESCRIPCIÓN DEL PROBLEMA

Para tener una ventaja competitiva en la función de distribución, se debe evaluar la percepción de calidad por parte de los clientes en el servicio de transporte que se encuentra en función de las exigencias del mercado y del que se involucran varios aspectos como por ejemplo, puntualidad de entrega, seguridad de transporte, control, etcétera.

El transporte es una parte esencial de los sistemas logísticos porque en este último eslabón de la cadena, recae todo el esfuerzo de una planeación que concluye en un producto que parte hacia su usuario final. El sistema de transporte hace que las mercancías y productos sean móviles y además proporciona una eficacia oportuna y regional para que sea promovido el valor añadido bajo el principio del menor costo. (yu Tseng *et al.*, 2005) Por lo tanto, un accidente, ya sea por situaciones externas, internas o humanas, generará una disrupción en el flujo que no permitirá lograr los objetivos establecidos desde la planeación.

Para lograr un proceso logístico o de transporte de producto más seguro, se debe desarrollar una gestión de riesgos logísticos para estudiar cuáles son las posibles vulnerabilidades dentro del proceso. (Vilko y Hallikas, 2012) Los accidentes logísticos son eventos que ocurren constantemente en las operaciones por lo que se le ha denominado “vulnerabilidad de entrega”. (Wu y Chaipiyaphan, 2020)

Debido al aumento de las exigencias del mercado existe una mayor demanda de personal calificado para las operaciones logísticas y así mismo un trabajo especializado dado que los conductores profesionales son sometidos a ritmos de trabajo intensos para cumplir con las entregas en tiempo y forma. Estas altas exigencias han tenido un efecto negativo para la salud de los conductores que esto ocasiona un aumento de la posibilidad de sufrir accidentes por lo que se verán afectados, no sólo las compañías, sino también todos los miembros de la sociedad. (Sanz *et al.*, 2017) Se afirma que la mayoría de los accidentes de tráfico, se atribuyen directamente a

factores humanos exclusivamente o como un factor contribuyente. (Öz *et al.*, 2013)

La compañía, una importante multinacional de producción de gases industriales, dónde es aplicado este estudio cuenta con una cultura organizacional que tiene como principal objetivo contar con un ambiente laboral seguro. Se enfatiza constantemente, desde todos los niveles, la necesidad de desarrollar mecanismos dónde se proteja la integridad del trabajador. Por lo tanto, aún con la implementación de distintas herramientas y políticas para la prevención y disminución de incidentes de tránsito en el área de distribución, se siguen registrando cifras de incidentes no deseados para la organización que representa un riesgo para los conductores profesionales, el equipo de transporte y el producto.

De modo que la motivación de este estudio estará enfocada en identificar el método que resulte mejor para la identificación de los factores que aumentan la probabilidad de riesgo para la función del transporte de carga pesada y por consiguiente para los conductores profesionales del área de distribución, y determinar cuál es el nivel de riesgo presente para poder brindar una fuente de información que ayude a la administración a desarrollar una gestión de riesgos más asertiva.

1.2 OBJETIVO GENERAL

Desarrollar una herramienta que identifique los factores que aumentan la probabilidad de incidente y que de manera predictiva determine el grado de riesgo para un conductor profesional de transporte de carga pesada que permita la implementación de mejores medidas de seguridad

1.2.1 OBJETIVOS ESPECÍFICOS

- Definir variables de riesgo

- Clasificar a los conductores entre alto riesgo y no riesgo
- Crear una herramienta basada en algoritmos para realizar mejores análisis de seguridad y tomar decisiones más asertivas para disminuir el impacto del factor humano en la accidentabilidad del transporte

1.3 HIPÓTESIS

Al identificar cuáles son las variables que tienen mayor prevalencia para la generación de un incidente con la ayuda de un modelo que se basa en algoritmos y definiendo el nivel de riesgo presente, se puede mejorar el método de captación de datos y su gestión para el desarrollo de medidas de seguridad que disminuyan la vulnerabilidad de la función de conducir.

1.4 JUSTIFICACIÓN DE LA INVESTIGACIÓN

Un alto número de las lesiones graves y muertes registradas mundialmente son a causa de los accidentes de tráfico; según cifras de la OMS (Organization, 2018) en el 2018 se registraron 1.35 millones de muertes, más los millones de personas con lesiones y efectos adversos en su salud a largo plazo. Por esta razón la OMS insta a tomar medidas drásticas y cumplirlas para poder salvar más vidas.

En la última década los accidentes de tráfico a nivel global se han convertido en un tema de creciente estudio, ya no sólo por sus efectos inmediatos como lo son los disturbios en el tráfico, cambios de rutas, embotellamiento, retrasos, etcétera, sino además como una materia de salud pública. Después de años de investigación, ha sido ampliamente aceptado que una significativa reducción del impacto de los accidentes puede ser alcanzado a través de efectivos métodos de detección y estrategias de respuesta correspondientes, (Zhang *et al.*, 2017) por lo que se han efectuado

estudios e investigaciones desde varios enfoques (médicos, fisiológicos, psicológicos, del vehículo e incluso del impacto de las condiciones climatológicas) para tratar de mitigar la gravedad de los incidentes.

Las cifras en México registran que en el 37 % de los percances en tramos federales se encuentra involucrado el transporte de carga, quienes a su vez tienden a ser más propensos a estar implicados en fatalidades y en significantes costos que en el caso de los accidentes de vehículos (CAINTRA, 2017). En nuestro país el 80 % de las causas de accidente se atribuyen a factores humanos, un 9 % a factores mecánicos, 7 % a motivos climatológicos y 4 % debido a la infraestructura carretera (SCT, 2016). Entre el 20 y 30 % de los accidentes están relacionados con vehículos comerciales, pero a pesar de este conocimiento los estudios enfocados hacia conductores profesionales del transporte de carga son escasos en México. Además, las investigaciones sobre el impacto que tiene esta función en la logística y cadena de suministro también están limitadas y no logran reflejar las transformaciones que han sufrido en su trabajo dadas las exigencias y la presión del mercado (Sanz, 2017).

Existe un aumento del interés social por la reducción de la severidad de los incidentes, por lo que la gestión de riesgo ha evolucionado para ser una tarea básica o central tanto en la industria como en el gobierno. En el caso del sector del transporte, el objetivo general de la gestión de riesgos es reducir la probabilidad y la gravedad de los accidentes. El autor Abkowitz (2002) recalca que es importante reconocer que la seguridad del transporte y la gestión de riesgos tradicional comparten un objetivo en común, que es el siguiente:

“Reducir la probabilidad y las consecuencias de las catástrofes para proteger la salud humana, la calidad de vida y el medio ambiente”.

Dentro de las organizaciones el sistema de transporte se ha vuelto la actividad económica más importante entre todos los componentes de los sistemas logísticos porque alrededor de un tercio del gasto dentro de las compañías se están destinando al área de transporte. Por ese motivo, se vuelve imprescindible realizar mejoras en

el área que involucra más gasto operativo lo cual tendrá mejores efectos en el costo. Además, también influye sobre los resultados de producción y ventas. (yu Tseng *et al.*, 2005)

La gestión de riesgo en la cadena de suministro se puede considerar dentro de las compañías como una función estratégica porque influyen sobre los resultados operativos, de mercado y financieros, por consiguiente, la eficacia y el rendimiento de una organización pueden aumentar cuando la estrategia, toma en cuenta el contexto y la realidad ambiental para reducir la incertidumbre. (Narasimhan y Talluri, 2009) En este caso, lograr que disminuya la vulnerabilidad del sistema de transporte ofreciendo mejores condiciones de seguridad para el conductor y el producto.

Se pueden determinar cuáles son los riesgos presentes para el área de transporte a través de la adecuada recopilación de información y su posterior análisis para determinar cuáles son los elementos que en su conjunto propician los incidentes. El correcto tratamiento de los datos se podrá transformar en una herramienta que ayuda a las personas a desarrollar medidas de gestión de riesgos que disminuyan la incertidumbre en las rutas de distribución.

1.5 METODOLOGÍA

El estudio será llevado a cabo bajo un esquema que ayuda a la sistematización del proceso de investigación y de la obtención de la información para lograr el objetivo del análisis. Se inicia con una descripción del problema para distinguir las palabras clave y para definir el marco de estudio que ayuden a la búsqueda de literatura que aporten los conceptos y herramientas que se han utilizado para resolver este tipo de problemáticas. Luego, se hace una recopilación de la información disponible dentro de la compañía, referente al tema de seguridad y transporte para, posteriormente, con base en los hallazgos en la literatura y en suma con los datos de la empresa, se seleccione la o las herramientas para realizar la experimentación. Finalmente, se

procede a la aplicación del modelo, análisis, visualización y validación de resultados. En la Figura 1.1 se muestra en resumen la secuencia de la metodología para llevar a cabo el estudio.

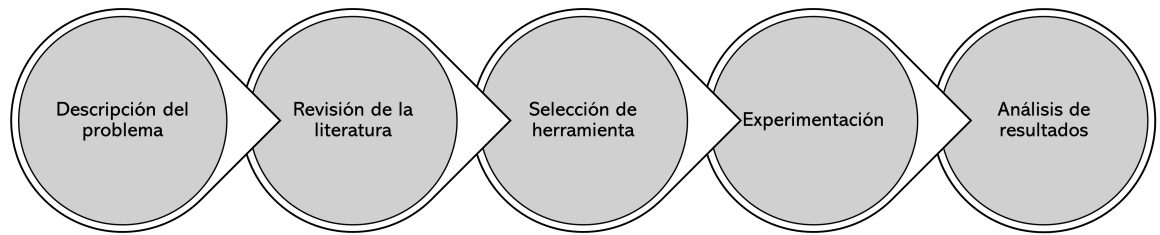


FIGURA 1.1: Metodología de tesis

1.6 ESTRUCTURA DE TESIS

Este trabajo de estudio se desarrolla en capítulos en dónde la información se ha estructurado de la siguiente manera:

- **Capítulo 1:** Introducción al tema y su pertinencia, así como datos generales sobre la problemática
- **Capítulo 2:** Se da a conocer cuál ha sido la investigación y aportación previa sobre la problemática, el estado actual y cuáles han sido los enfoques, métodos y/o propuestas de solución. La revisión de dicha literatura consiste en la com-

paración entre los métodos más utilizados entre los estudios más recientes para la selección de la o las herramientas más adecuadas para el caso de estudio

- **Capítulo 3:** Consiste en la explicación de la implementación del método basado en minería de datos para la fase de experimentación
- **Capítulo 4:** Se lleva a cabo la fase de experimentación y análisis de los resultados
- **Capítulo 5:** Finalmente, en este capítulo se presentan las conclusiones de la evolución del estudio de análisis de riesgo, hallazgos, conclusiones personales, además de recomendaciones para futuras investigaciones

CAPÍTULO 2

ANTECEDENTES

En este capítulo se pretende exponer el estado actual de los estudios que abordan el tema de la seguridad para las rutas de distribución. Se presentan conceptos enfocados a la importancia de la seguridad del conductor y cómo esto aporta a la logística y a la relación con el buen funcionamiento de la cadena; también, los métodos y/o herramientas que se han utilizado en los años recientes y qué resultados se han obtenido en las experimentaciones.

2.1 LA FUNCIÓN DEL TRANSPORTE DE CARGA EN LA CADENA DE SUMINISTRO Y LA LOGÍSTICA

Antes de incursionar en los conceptos de seguridad y de los métodos actuales, se desea plantear la repercusión que tiene el área del transporte para la logística y la cadena de suministro. Para definir transporte de una manera sencilla, nos referimos a la función que se encarga de todas las actividades relacionadas, directa o indirectamente, con la necesidad de situar los productos en los puntos de destino correspondientes, tomando en cuenta las condicionantes de seguridad, servicio y costo (García, 2014). Por ende, el transporte es una actividad fundamental, el elemento clave en el desarrollo y planeación de un sistema logístico (yu Tseng *et al.*, 2005).

La definición del **Council of Logistics Management** (1991), nos dice que: *“la logística es parte del proceso de la cadena de suministro que planifica, implementa y controla el flujo eficiente y eficaz hacia adelante y hacia atrás y el almacenamiento de bienes, servicios e información relacionada entre el punto de origen y el punto de consumo con el fin de satisfacer las necesidades de los clientes”*. Bajo estas definiciones, en conclusión, se trata de la distribución física de los bienes materiales desde el principio hasta el final, que es posible con la utilización de diferentes mecanismos de transporte que si se gestionan de manera eficiente pueden ser una ventaja competitiva para las compañías.

El transporte de carga terrestre tiene una fuerte ventaja frente a los otros modos de transporte ya que cuenta con el acceso a casi cualquier destino siendo así un factor relevante para el desarrollo económico y la competitividad de las cadenas de suministro, no sólo nacional, sino también internacional. Por lo tanto, se ha convertido en el modo y medio por el que se mueve la mayor cantidad de carga en México; en la Figura 2.1 se muestran los porcentajes de la carga de acuerdo con datos de la Secretaría de Comunicaciones y Transportes. Además, no menos interesante, este sector genera cerca del 49 % del empleo en el sector transporte del país (Sanz, 2017).

Cómo se ha previsto, para poder tener una ventaja competitiva en la función de distribución, se vuelve fundamental evaluar cuál es la percepción de calidad por parte de los clientes en el servicio de transporte que se encuentra en función de las exigencias del mercado y del que se abarcan los siguientes aspectos:

- Rapidez y puntualidad de entrega
- Fiabilidad en las metas prometidas
- Seguridad e higiene en el transporte
- Cumplimiento de las condiciones impuestos por el cliente (ejemplo: horarios de entrega)
- Información y control de transporte

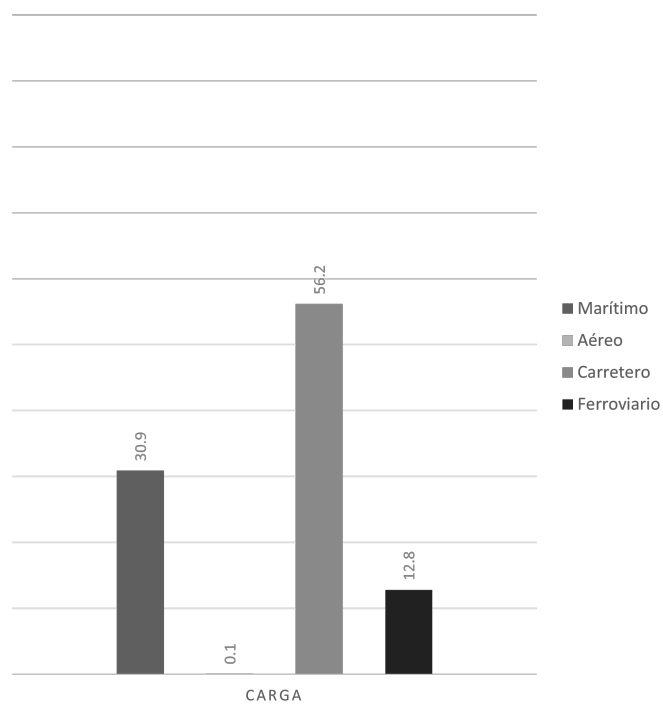


FIGURA 2.1: Total de carga movida entre los distintos medios de transporte en México. Fuente: (INEGI, 2018)

Entonces debido al aumento de las exigencias del mercado es que existe una mayor demanda de personal calificado para las operaciones logísticas y así mismo un trabajo que sea especializado dado que los conductores profesionales son sometidos a ritmos de trabajo intensos para cumplir con las entregas en tiempo y forma. Estas altas exigencias a su vez se han transformado en un efecto negativo para la salud de los conductores como, por ejemplo, trabajar bajo cierto nivel de estrés y fatiga, que a resultado en un aumento de la probabilidad de sufrir algún incidente por lo que se verán afectados, no sólo las compañías, sino también todos los miembros de la sociedad (Sanz *et al.*, 2017).

El artículo publicado en el año 2003 por (Abkowitz, 2002) nos dice que, debido al aumento del interés social por la reducción de la severidad de los incidentes, la gestión de riesgo ha evolucionado para ser una tarea básica o central tanto en la industria como en el gobierno. En el caso del sector del transporte, el objetivo general de la gestión de riesgos es reducir la probabilidad y la gravedad de los accidentes. El autor recalca que es importante reconocer que la seguridad del transporte y la gestión de riesgos tradicional comparten un objetivo en común, que es el siguiente:

Reducir la probabilidad y las consecuencias de las catástrofes para proteger la salud humana, la calidad de vida y el medio ambiente”

Por lo tanto, la motivación de este estudio estará enfocada en identificar cuáles son los elementos característicos que aumentan la probabilidad de riesgo para los conductores profesionales del área de distribución y cómo esto influye sobre una cadena de suministro para proponer una gestión de riesgos logísticos más asertiva.

2.2 EFECTOS SOBRE EL TRANSPORTE DE CARGA PESADA

Es importante reconocer cuáles son las cifras o el nivel de incidencia en el área de la transportación. Al menos en la última década los accidentes de tráfico a nivel global se han convertido en un tema de creciente estudio, ya no sólo por sus efectos inmediatos como lo son los disturbios en el tráfico, cambios de rutas, embotellamiento, retrasos, etcétera, sino además como una materia de salud pública. Después de años de investigación, ha sido ampliamente aceptado que una significativa reducción del impacto de los accidentes puede ser alcanzado a través de efectivos métodos de detección y estrategias de respuesta correspondientes, (Zhang *et al.*, 2017), por lo que, se han efectuado estudios e investigaciones desde varios enfoques (médicos, fisiológicos, psicológicos, del vehículo e incluso del impacto de las condiciones climatológicas) para tratar de mitigar la gravedad de los incidentes.

Un alto número de las lesiones graves y muertes registradas mundialmente son a causa de los accidentes de tráfico; según cifras de la OMS (Organization, 2018) en el 2018 hubo un registro de 1.35 millones de muertes, más los millones de personas con lesiones y efectos adversos en su salud a largo plazo. Actualmente se estima que es la octava causa de muerte entre todos los grupos de edades globalmente y se pronostica que será la séptima causa para el 2030 si no se efectúan contramedidas efectivas. La OMS resalta que es necesario actuar de una manera drástica implementando medidas y que se hagan cumplir con el objetivo global de salvar vidas.

Para México el panorama no es muy distinto ante esta creciente problemática. Según cifras del INEGI, en el año 2019 se registraron 691,046 accidentes de tránsito terrestre en zonas urbanas y suburbanas y el número de víctimas mortales fue de un total de 4,125 y 91,713 heridos en total. Dentro de las cifras se reportó que 62,438 fueron de camionetas y camiones de carga. Respecto a los percances en los tramos federales, en el año 2016 hubo 12,567 accidentes de los cuales el 37 % estuvo

involucrado un transporte de carga y en el 2015, se registraron 1,443 catalogados como fatales (CAINTRA, 2017). En consecuencia de este preocupante asunto, en el año 2016, se celebró un Foro en el Senado de la República con la finalidad de generar acciones para mejorar la seguridad de carretera, en el que se identificaron los principales factores de accidentabilidad en las carreteras federales para el transporte de carga. Se le atribuyeron a los factores humanos con un 80 %, mecánicos con un 9 %, climatológicos con un 7 % y la infraestructura carretera un 4 % (SCT, 2016).

De acuerdo con otro análisis, se afirma que la mayoría de los accidentes de tráfico se atribuyen directamente a factores humanos exclusivamente o como un factor contribuyente (Öz *et al.*, 2013). En un estudio realizado para presentar el panorama actual de los conductores profesionales en el país, nos indica que los accidentes de vehículos de carga permanecen siendo la causa más común de lesiones y muerte prematura. Entre el 20 y 30 % de los accidentes están involucrados con vehículos comerciales, pero a pesar de este conocimiento los estudios enfocados hacia conductores profesionales del transporte de carga son escasos en México. Según esta misma investigación, hasta la fecha del estudio mismo, sólo existían 60 publicaciones respecto a esta profesión y sólo la mitad fueron desarrolladas en el último lustro. Además, las investigaciones sobre el impacto que tiene esta función en la logística y cadena de suministro también están limitadas y no logran reflejar las transformaciones que han sufrido en su trabajo dadas las exigencias y la presión del mercado (Sanz, 2017). También es importante señalar que dentro de los errores más peligrosos que los conductores cometen son, ir a exceso de velocidad, manejar bajo el efecto del alcohol u otras sustancias tóxicas y la distracción, que son acciones tomadas recurrentemente como causa de las exigencias de la función misma (Aksjonov *et al.*, 2018).

Esta problemática se ha tornado una preocupación global, por lo que en los Estados Unidos la agencia responsable de atender estos temas, la FMCSA (Federal Motor Carrier Safety Administration), reporta que la fatiga está relacionada con el 13 % de los choques de transportes de carga, el 8 % con la distracción externa, 8 % con manejo agresivo. Además, un análisis realizado por HOS (Hours of Service

- FMCSA) halló que hay tres factores principales que contribuyen a la fatiga de conducción: ritmo circadiano, privación del sueño, fatiga acumulada y por fatiga industrial o *“time-on-task”*. Estos factores han complicado los efectos en la fatiga del conductor y en consecuencia la seguridad (Chen y Xie, 2019). Estos mismos factores han propiciado en Australia que entre el 20 y el 30 % de los accidentes de vehículos comerciales estén relacionados con el cansancio (Howard *et al.*, 2004). Además de la fatiga, un estudio realizado por la NHTSA (National Highway Traffic Safety Administration) mostró que más del 23 % de las muertes están relacionadas con el factor del estilo de conducción de los estadounidenses. Es por esto por lo que es necesario que se desarrollen métodos eficientes para modelar, clasificar y entender los varios estilos de manejo (Wang *et al.*, 2017). Adicionalmente, con la misma intención de atender esta problemática, existen fundaciones de seguridad de tráfico que, en conjunto con fabricantes de vehículos, realizan un esfuerzo constante para desarrollar nuevas ideas al mejoramiento de la seguridad en las rutas (Aksjonov *et al.*, 2018).

Bajo el conocimiento de los índices y padeciendo asimismo las consecuencias para el personal de transporte, es por ello, por lo que, los estudios también han tomado una mayor relevancia en el sector industrial por la necesidad de brindar mejor seguridad a los conductores profesionales y disminuir los costos incurridos tras los incidentes. A pesar de que se han realizado numerosas investigaciones respecto a las actitudes y comportamientos que influyen sobre el riesgo de accidente para los conductores no profesionales, hasta la fecha de esta publicación, son pocos los estudios que examinan los comportamientos de los conductores profesionales que conducen vehículos de compañías y/o que están largas jornadas frente al volante. Debido a la naturaleza de su función de conducción, el conflicto entre la movilidad (por carreteras y ciudades) y seguridad puede tener una composición diferente aunado a ciertos factores adicionales. El papel de la cultura y clima organizacional para la conducción profesional y su seguridad, sabiendo que tiene un vínculo con otros factores relacionados con la conducción, también sigue sin estudiarse (Öz *et al.*, 2013).

Desde el punto de vista de Junior *et al.* (2017), “en la gestión de los transportes de carga, la clasificación de perfiles de comportamiento de los conductores, que pueda ser automatizada, continua y si es posible en tiempo real, permitiría a los directivos instaurar campañas con el objetivo de mejorar la puntuación de los conductores y en su defecto disminuir los accidentes, eficiencia en los recursos y alargar la vida útil del vehículo.

Estando frente a un tema tan amplio y a su vez complejo, se ha abordado desde distintos enfoques por lo crítico que se vuelve el poder identificar cuáles son aquellos elementos o factores de riesgo que se encuentran presentes en las rutas de distribución para poder desarrollar mejores mecanismos que puedan brindar un ambiente más seguro para los conductores profesionales. Al mismo tiempo, encontrar cuáles son las herramientas que soporten a la función de los administradores para llevar a cabo un mejor análisis de la información disponible y de manera práctica.

Existe un área de oportunidad para México en relación con este tema debido a que, hasta el momento, la mayoría de las investigaciones en nuestro país se han enfocado principalmente hacia la evaluación de la salud de los conductores o en su caso, los daños infligidos por esta función, estilo de vida, jornadas laborales e información sociodemográfica (Sanz *et al.*, 2017). En la opinión de Abkowitz (2002), aplazar la inversión por más tiempo en estrategias nuevas de control para el ámbito de la gestión de riesgos puede dejar a la sociedad en general más vulnerable.

2.3 GESTIÓN DE RIESGOS PARA EL ÁREA DE TRANSPORTE

Repasando la definición de logística, el **Council of Supply Chain Management Professionals** nos dice que es “*el proceso de planificación, implementación y control de los procedimientos para el transporte y almacenamiento eficiente y eficaz de los bienes, incluidos los servicios y la información relacionada, desde el punto de*

origen hasta el punto de consumo con el fin de cumplir con los requisitos del cliente”, por lo tanto, con base en este concepto se puede interpretar que el transporte es una parte esencial de los sistemas logísticos porque en este último eslabón de la cadena, recae todo el esfuerzo de una planeación que concluye en un producto que parte hacia su usuario final. El sistema de transporte hace que las mercancías y productos sean móviles y además proporciona una eficacia oportuna y regional para que sea promovido el valor añadido bajo el principio del menor costo (yu Tseng *et al.*, 2005). Por lo tanto, un accidente, ya sea por situaciones externas, internas o humanas, generará una disrupción en el flujo que no permitirá lograr los objetivos establecidos desde la planeación. Las funciones que son primordiales en la logística incluyen tanto, la logística de “salida” como la de “entrada”, por lo que el control de las áreas de inventario y la transportación se vuelven funciones críticas (Choi *et al.*, 2016).

Para tener un proceso más seguro en la logística se debe contar con una gestión de riesgos logísticos que estudie de cerca cuáles son las vulnerabilidades en los procesos. (Vilko y Hallikas, 2012) Los accidentes logísticos son eventos que ocurren constantemente en las operaciones, como se demuestra en los datos en la Sección 2.2, por lo que se le han denominado “vulnerabilidad de entrega” (Wu y Chaipiyaphan, 2020). Desde el punto de vista de Datta (2017), una forma de disminuir las vulnerabilidades en los procesos logísticos se logra a través de la resiliencia en la cadena de suministro, que esto significa para una compañía tener la capacidad de estudiar el ambiente, prevenir disrupciones no deseadas y finalmente, estratégicamente asignar los recursos que gestionen los riesgos.

En la última década, la gestión de riesgos ha ido evolucionando como parte de una actividad principal tanto en el gobierno como en la industria. Esta se centra en la capacidad para medir cuál es la probabilidad de un evento potencial y sus consecuencias asociadas (Abkowitz, 2002). Entonces, ¿por qué es importante invertir en la seguridad de transporte? Porque si no se cuenta con un sistema de transporte bien desarrollado, la logística no traerá los beneficios completos, que son: mejor eficiencia

en la logística, reducción de los costos de operación y la promoción de un servicio de calidad (yu Tseng *et al.*, 2005). Así que, como señala Duncan (1972), la gestión de riesgo en la cadena de suministro se puede considerar dentro de las compañías como una función estratégica porque influyen sobre los resultados operativos, de mercado y financieros. Por lo tanto, la eficacia y el rendimiento de una organización pueden aumentar cuando la estrategia, toma en cuenta el contexto y la realidad ambiental para reducir la incertidumbre (Narasimhan y Talluri, 2009).

Dentro de las organizaciones el sistema de transporte es la actividad económica más importante entre todos los componentes de los sistemas logísticos porque alrededor de un tercio del gasto dentro de las compañías se destinan al área de transporte. Por lo tanto, una mejora en el área que involucra más gasto operativo tendrá mejores efectos en el costo. Además, también influye sobre los resultados de producción y ventas (yu Tseng *et al.*, 2005). Algo que es importante señalar es que, Abkowitz (2002) sostiene que, la eficacia en el proceso de gestión de riesgos en el transporte está muy conectada con la calidad de la recopilación de información. Para poder determinar cuál es la amenaza y/o vulnerabilidad es necesario que haya acceso a la información por parte del responsable de la gestión de riesgos para que, con base a los datos, sea capaz de definir los escenarios, consecuencias y las probabilidades de incidente.

Existe un área emergente bajo el dominio de la gestión de riesgos en la cadena de suministro que nos introduce lo relevante que es el tema de la seguridad del producto. Es decir, también concientizar que los productos son expuestos a varios riesgos y vulnerabilidades mientras son movidos a través de la cadena de suministro hasta la entrega con el cliente final. Esto quiere decir que, además de la importancia de estudiar los efectos sobre el conductor y el transporte, tener en mente que el producto también es expuesto a riesgos, dónde en su conjunto, las vulnerabilidades pueden provocar las disrupciones en la cadena logística o desenlaces negativos desencadenados por eventos inesperados. Por definición, la seguridad de producto se refiere a la disminución en la probabilidad de que el manejo de un producto provo-

que enfermedades, lesiones, muertes o consecuencias negativas para las personas, los bienes y los equipos (Maruchek *et al.*, 2011).

Por último, con el fin de resaltar la relevancia para el campo logístico y de cadena de suministro, la definición de riesgo, según la NOM 31000 dice que se trata de “el efecto de incertidumbre sobre los objetivos” (Organization, 2018). Así que, como concluyen Narasimhan y Talluri (2009), el tema de gestión de riesgos seguirá siendo una materia importante para los investigadores como asimismo para los profesionales de la gestión de la cadena de suministro.

2.4 ACCIONES PARA MEJORAR LA SEGURIDAD DE LAS RUTAS (ESTADO DEL ARTE)

A través del proceso de revisión de literatura se han identificado una amplia gama de metodologías aplicadas con el propósito de identificar los factores influyentes de riesgo y también herramientas desarrolladas para el análisis del desempeño de transporte. Abarcan desde evaluaciones diseñadas por expertos, encuestas que se han aplicado a los conductores en diferentes países, revisión de políticas y manuales para enfrentar riesgos y, en años más recientes, con el rápido avance en la era digital, el análisis con el apoyo de cámaras, sensores y de sistemas como el OBC (*On-Board Computer*).

2.4.1 PROPUESTAS DE ORGANISMOS INTERNACIONALES PARA MEJORAR LA SEGURIDAD DEL TRANSPORTE

Las organizaciones internacionales hacen esfuerzos por exponer y proponer planes de acción para que tanto los gobiernos como las organizaciones privadas implementen mejores medidas y mecanismos de seguridad.

En el año 2021 la ONU ha convocado a las naciones a formar parte de un plan global que ha llamado, *Global Plan: Decade of action for road safety*. El objetivo principal es lograr en el periodo que abarca del 2021 - 2030, el poder disminuir en un 50 % las muertes y lesiones “*preventibles*”(añaden al reto) que son a causa de los accidentes de tránsito. La intención, adicionalmente, es que haya una integración de labores entre gobiernos locales, organizaciones y compañías para alcanzar la meta. El manual expone los 5 pilares, al que denominaron el **Safe System Approach**, que son acciones para trabajar sobre:

1. Transporte multimodal y planeación del territorio
2. Infraestructura vial segura
3. Seguridad en el vehículo
4. Uso seguro de carretera
5. Respuesta tras el accidente

La Norma Internacional elaboró una guía respecto a la Administración y/o Gestión de Riesgos, la ISO 31000. En este manual se estipula que, debido a que las organizaciones se enfrentan a factores internos y externos que influyen para generar incertidumbre respecto a si podrán lograr los objetivos, esta norma es de apoyo debido a su naturaleza iterativa dónde se ayuda para establecer una estrategia, lograr objetivos y tomar las mejores decisiones informadas. El documento proporciona un plan general que es útil para cualquier tipo de organización para afrontar riesgos de diferentes naturalezas y replicable para distintos trabajos (ISO31000, 2018).

2.4.2 EL ANÁLISIS DEL DESEMPEÑO DEL CONDUCTOR PROFESIONAL

La percepción de seguridad vial no debería ser desvinculada del análisis del comportamiento de los conductores debido a que se tiene el conocimiento de que la mayoría de los accidentes de tránsito son provocados por factores humanos. El análisis del conductor favorece el mejoramiento de los sistemas de transporte para ser más inteligentes y resilientes (Elassad *et al.*, 2020). Pero Martinez *et al.* (2018) esbozan que el comportamiento de conducción es un concepto complejo porque en términos generales es acerca de intentar trazar cómo es que se manipula el vehículo en la escena de conducir y de su entorno.

En referencia al explorar los factores humanos de conducción, Öz *et al.* (2013) nos proponen que pueden ser analizados bajo dos componentes: estilo/comportamiento de conducción y habilidades/desempeño. El comportamiento se refiere a la manera en la que el conductor decide conducir como, por ejemplo, la velocidad, nivel de atención, etcétera. Es decir, abarca lo que los conductores usualmente “hacen”. En cambio, el desempeño se trata del procesamiento de información y de las habilidades de seguridad, en otras palabras, explica lo que los conductores “pueden” hacer.

En una propuesta con la intención de crear un marco de estudio del conductor, se resumieron y estructuraron los parámetros más influyentes, de acuerdo con una revisión de literatura realizada por Elassad *et al.* (2020), en un sistema que denominaron “*Driver-Vehicle-Environment*” (DVE), en el que se presentan, de manera armonizada e integrada, las interacciones entre el conductor, el vehículo y el entorno, para la evaluación del comportamiento de conductor.

En otro artículo, Siordia *et al.* (2010) también coinciden en que para tener un sistema completo de análisis se deben considerar e integrar valores de los tres elementos básicos de seguridad vial: conductor, carretera y vehículo.

En la siguiente sección, entendiendo que el principal factor de accidentabilidad

es el humano y que además el entorno de conducir es un sistema complejo dónde se involucran una multitud de variables, se recurre a una exploración de estudios que se hayan realizado en los últimos 15 años que se encuentren enfocados al análisis del desempeño y de los estilos de conducción de los conductores profesionales, así como también la búsqueda de cuáles son las mejores herramientas para identificar qué factores son los que cuentan con una mayor participación e influencia sobre el aumento del riesgos de padecer un incidente.

2.4.3 ESTUDIOS CUALITATIVOS

El estrés, el cansancio o somnolencia y la distracción se han convertido en componentes que han influido significativamente en los accidentes. Datos estadísticos dados por reportes británicos y americanos demuestran que, evitando que haya conducción cuando se detecta cansancio o distracción sería un importante enfoque para prevenir accidentes serios (Kaplan *et al.*, 2014).

Aproximadamente el 30 % de los artículos que se leyeron para realización de este estudio, utilizaron la vía de las encuestas para obtener los datos cualitativos y los factores comunes entre los conductores. Un tipo de encuesta era aplicado directamente a los conductores y en otros se trataba de evaluar a la administración u organizaciones. Según Zambrano *et al.* (2020), un punto fundamental para la elaboración y estructuración de los análisis con este tipo de enfoque es poder plantear la problemática definiendo correctamente que es “riesgo”. Nos define el factor de riesgo como, “*un elemento presente dentro de las condiciones de trabajo asociadas a un problema de seguridad o en todo caso de salud laboral*”.

Por ejemplo en México, Sanz *et al.* (2017) desarrollaron una encuesta para determinar y jerarquizar agentes que causan un aumento de la probabilidad de accidente y su repercusión en la cadena de suministro. Dicha encuesta fue enviada a expertos en el área de capacitación, contratación y de condiciones laborales de

los choferes de transporte. El estudio se basó, primero en estructurar las preguntas con el Método Delphi y posteriormente utilizaron un análisis de escalonamiento multidimensional, obteniendo los siguientes resultados:

Principales consecuencias en la cadena:

1. Accidentes de tráfico
2. Entregas incorrectas

Principales condiciones de riesgo:

1. Derivadas del tiempo, ritmo de trabajo y descanso
2. Falta de capacitación

En el país de Irán, se desarrolló una investigación para identificar los riesgos ocupacionales; se aplicó un cuestionario en el cual se descubrió que, de los 400 conductores entrevistados, el 25 % ya había sido parte de un accidente una o más veces. Se incluyeron preguntas respecto a las horas de conducción, tiempo de descanso, si habían experimentado somnolencia durante el manejo y si alguna vez se quedaron dormidos al volante, además de información de antecedentes médicos. Al analizar las repuestas se dedujo que existe una estrecha relación entre las horas de conducción sin pausa y los accidentes. Las herramientas estadísticas utilizadas en este estudio fueron la prueba exacta de Fisher y la prueba de chi-cuadrada de Pearson (Jahangiri *et al.*, 2013).

Zambrano *et al.* (2020) implementaron el Método L.E.S.T. (*Laboratoire d'Economie et Sociologie du Travail*), el cuál es conocido para evaluar las condiciones laborales de la forma más objetiva y global posible, dónde al final se indique si el trabajador se desempeña en un ambiente satisfactorio, molesto o nocivo. El diagnóstico considera 16 variables agrupadas en 5 dimensiones que son: el entorno físico, carga física, carga

mental y aspectos psicosociales y tiempo de trabajo. En este estudio enfocado a los conductores, los resultados indicaron que los factores que aumentan el riesgo son:

- Carga dinámica
- Ruido
- Iluminación
- Complejidad de la función
- Relación con los superiores
- Jornada laboral

En Australia se aplicó un cuestionario en dónde se tomaron medidas antropométricas a 3,268 conductores comerciales; asimismo se aplicó un laboratorio de polisomnografía con lo que se pretendió demostrar los riesgos ocupacionales que se enfrentan los conductores a raíz de las largas jornadas que derivan en exceso de somnolencia, respiración alterada por el sueño y una prevalencia de obesidad, que se traducen en un factor de riesgo importante para tener un accidente. Para encontrar la relación entre el exceso de somnolencia y los factores relacionados al trabajo, con un modelo de regresión logística se encontró que aumenta en un 50 % el nivel de somnolencia si se duerme menos de 7 horas en los días libres (Howard *et al.*, 2004).

En un estudio hecho en Colombia por Useche *et al.* (2017), se pretendía describir la asociación entre los indicadores de fatiga, estrés laboral, indicadores de salud y los accidentes de tráfico entre los conductores. Si se examina la asociación se contribuye al incremento del conocimiento relacionado con los factores psicosociales de riesgo a los que son sujetos. Como método se aplicaron cuestionarios divididos en las siguientes secciones:

1. Variables demográficas (edad, años de experiencia): percepción de riesgo de conducción, datos de accidentes sufridos en los últimos años

2. Escalas para medir fatiga: *Checklist for Individual Strengh* (CIS)
3. Escala para recuperación: *Need for Recovery after Work Scale* (NFR)
4. Modelo de demanda-control: *Job Content Questionnaire* (JCQ)
5. *Effort-Reward Imbalance Model: ERI Questionnaire*

Las herramientas de análisis fueron el coeficiente de correlación de Pearson y la prueba de chi-cuadrada. Se concluyó que la fatiga constituye un factor relevante a considerar para el diseño de intervenciones enfocadas para mejorar las condiciones laborales y promover el bienestar de los operadores, con respecto a su relación a factores adversos como estrés laboral, indicadores de riesgos de conducción y resultados de salud negativos.

Estos estudios presentados demuestran cuáles son parte de los efectos en la salud a causa de la labor de conducción profesional. Se deja entrever que no son riesgos laborales de ciertos países sino que se trata de una problemática global dónde existe una similitud de factores. La motivación principal de estos autores, además de identificar los daños adversos a la salud, con un enfoque hacia los factores fisiológicos y psicológicos que han aumentado la probabilidad de accidentes, es presentar el panorama actual de esta función y cómo es preciso que las organizaciones tomen acción para disminuir dichos efectos. En conclusión, autores llaman a definir dichos elementos o variables para analizarlos e instaurar condiciones de trabajo que sean idóneas para mantener la salud integral de los trabajadores y así poder tener una disminución en la prevalencia de la accidentabilidad en esta difícil labor.

2.4.4 ESTUDIOS CUANTITATIVOS

De acuerdo con Moosavi *et al.* (2019), en investigaciones realizadas anteriormente se ha utilizado una menor escala de datos, lo que ha limitado el impacto y la replicabilidad porque se han hecho con información, por ejemplo, de una ciudad y/o

de algunas cuantas carreteras, un solo tipo de unidad, o de una compañía; o grandes datos que no están actualizados o que no consideran variables como, clima, aceleración, frenado, excesos de velocidad durante la ejecución de la ruta, tráfico, etcétera. Por ejemplo, se tiene el conocimiento de que existe una relación entre los niveles de somnolencia de un conductor, sus características de comportamiento, respuestas fisiológicas y el desempeño de conducción, pero a pesar de las varias propuestas de métodos de detección considerando estos datos, los resultados de los estudios han sido insuficientes porque no han sido tan precisos (Gwak *et al.*, 2018). Estos datos pueden ser variables que representen en menor o mayor escala una influencia directa o indirecta como detonantes de incidentes. Con la intención de ampliar el conocimiento de cómo dichas variables influyen, se han desarrollado métodos capaces de captar datos que antes no eran considerados y encontrar la relación entre ellas. Cuando el nivel de manipulación de datos, la exploración e inferencia están más allá de la capacidad humana, es cuando se recurre a las tecnologías informáticas para poder automatizar el proceso (Riquelme *et al.*, 2006).

El avance de la tecnología ha brindado la oportunidad para poder analizar la información que se recolecta desde varias fuentes o sistemas que ayudan a la reducción de la incertidumbre con el desarrollo e implementación de una mejor gestión de riesgos para la prevención, con mayores niveles de seguridad a través de mecanismos y herramientas digitales e incluso con predicciones de accidentabilidad. Adicionalmente, se ha logrado captar e integrar una gran cantidad de datos y hasta en ciertos ámbitos, en tiempo real. Los siguientes ejemplos son algunos de los estudios que se han realizado en los últimos años.

Un estudio realizado por Öz *et al.* (2013) pretendía encontrar el nivel de influencia de cómo es que el ambiente/clima o la cultura organizacional tienen efectos sobre el manejo seguro de los conductores profesionales. Para ello, se hicieron distinciones claras entre el comportamiento de conducción y el desempeño de conducción, ya que el autor sostiene que la gran mayoría de los incidentes son a causa de factores humanos. El objetivo de la investigación consistió en desarrollar una herramienta

cuantitativa llamada, “*Transportation Companies’ Climate Scale*” (TCCS) en la que se miden 3 dimensiones del clima de seguridad organizacional que se compone de 33 elementos. Las dimensiones son: gestión general de seguridad, prácticas y prevenciones específicas, trabajo y presión de tiempo. Se utilizó un modelo de análisis de regresión jerárquico y los resultados demostraron que, si los conductores perciben que la organización se preocupa por ordenar la carga de trabajo y se ocupa de la presión del tiempo dando prioridad a la seguridad, se registran menos infracciones y/o errores. En resumen, indica que la cultura organizativa tiene impacto sobre el comportamiento de los empleados.

Uno de los enfoques que más se ha abordado en los últimos años es la identificación o clasificación del estilo de conducción porque juega un rol vital para la seguridad de carretera o rutas. Esto se debe al aumento de la utilización de algoritmos de aprendizaje que pueden identificar patrones de comportamiento. En un artículo escrito por Higgs y Abbas (2015) se definen los conceptos de “comportamiento del conductor”, como la función que hace mapas de los estados del tráfico con las acciones del conductor y “patrón de conducción” como el número del mapeo de funciones y la correspondiente identificación y tamaño del espacio del estado cubierto.

Para mejorar el rendimiento de la clasificación del estilo de conducción, en un artículo de análisis comparativo, se utilizó un método de aprendizaje semi-supervisado de support vector machine (S3VM) con una cantidad pequeña de datos etiquetados y posteriormente hacer una comparación contra el SVM de aprendizaje supervisado. El objetivo era que el modelo diera los datos de salida como: conductor agresivo o conductor normal. Se concluyó además que entre los dos modelos, S3VM obtuvo un mejor desempeño que SVM para la clasificación donde existe una cantidad limitada de datos etiquetados y una gran cantidad de datos sin etiqueta (Wang *et al.*, 2017).

El comportamiento de conducción puede verse influido por diferentes facto-

res, por lo que, partiendo de esta premisa, Higgs y Abbas (2015) realizaron una investigación de las características de una amplia gama de conductas con la cual se vinculan los estados de manejo en acciones. Se utilizó el modelo GHR (*Gazis-Herman-Rothery*) para expresar la función de transferencia entre estados y acciones y para demostrar que existe un mejor desempeño si se complementa con la aplicación de modelos de segmentación y agrupación (*k-means*). Los resultados mostraron que la segmentación ayuda a entender que existen factores no estudiados o considerados que causan cambios en la manera de conducir y además se identificaron 30 estilos de conducción que pueden ser modelados independientemente para un mejor análisis.

Con el previo conocimiento de que la somnolencia o cansancio son de las principales causas de accidentes de tráfico, se hizo uso de algoritmos con base en *Machine Learning* aplicando *Logistic Regression* (LR), *Support Vector Machine* (SVM), *k-Nearest Neighbor Classifier* (kNN) y *Random Forest* (RF) para la clasificación de los conductores por los niveles de somnolencia propuesto por Zilberg que inician desde cero hasta 4.

- 0: Alerta
- 1: Somnolencia ligera
- 2: Somnolencia moderada
- 3: Somnolencia significativa
- 4: Somnolencia extrema

También se buscaba encontrar la relación entre los niveles de somnolencia y el desempeño de conducción, características de comportamiento y las respuestas fisiológicas. RF (*Random Forest*) consiguió la precisión más alta y se concluye que la somnolencia va en aumento con el transcurso del tiempo y que hubo cambios significativos en el desempeño de conducción (ya no se mantenían en medio del carril) a medida que aumentaba el grado de somnolencia (Gwak *et al.*, 2018).

Un análisis realizado por Aksjonov *et al.* (2018), que con la intención de predecir y modelar el comportamiento de conducción de nuevos conductores, optaron por implementar dos modelos de regresión no lineal debido a que el desempeño de conducción es una actividad altamente no lineal que no puede ser limitada en varias clases. Los modelos fueron:

- ANN - *Artificial Neural Network*
- ANFIS - *Adaptative Network Based Fuzzy Inference System*

Los métodos no lineales son precisos para predecir de forma eficiente desde la información de muestra. En este estudio se demostró que ambas herramientas, ANN y ANFIS, dieron predicciones muy satisfactorias del comportamiento

En un caso de estudio realizado por Chen y Xie (2019) dónde se tomaron datos de dos compañías transportistas en Estados Unidos, se procedió a realizar un análisis para reconocer los patrones de conducción. Con base en los antecedentes de ambas compañías, se ratificó la prevalencia de fatiga debido a las condiciones mismas de la profesión que privan el sueño y a las largas jornadas laborales frente al volante, por lo tanto, se decidió utilizar dos herramientas de *machine learning*, para identificar los patrones únicos de conducción que se asocian con una mayor probabilidad de choque para los transportes comerciales. Se utilizaron dos métodos: análisis de clúster y regresión logística de tiempo discreto. Se obtuvo como resultado que, con el método de clúster, se clasificaron en 10 grupos los conductores que representan los estilos de conducción. Se volvió a comprobar que, en general, largas jornadas de conducción aumentan las probabilidades de incidente a través de la mayoría de los grupos (patrones), lo cual sigue siendo consistente con análisis previos. Además, se concluyen dos puntos importantes: los horarios dónde existe un menor riesgo es durante la mañana y con descansos por la tarde, por el contrario, el riesgo aumenta si se conduce por la tarde. Por último, los autores instan a incluir para futuras investigaciones más factores como, por ejemplo, edad del conductor, experiencia y género.

En un último ejemplo, en un artículo escrito por Siordia *et al.* (2010) presentan un sistema multidisciplinario para la detección automática de riesgo de conducción en donde se considera la integración de los datos del conductor, carretera y vehículo. Se utilizaron estos datos para entrenar 5 algoritmos de minería de datos: (CART - *Classification and Regression Tree*, RIP - *Regulatory Interaction Predictor*, KNN - *K-Nearest Neighbors*, ANN - *Artificial Neural Networks*, SVM - *Support Vector Machine*) para predecir y clasificar los niveles de riesgo. El algoritmo que obtuvo mejores resultados fue CART, pero se debe considerar que estaba fuertemente relacionado con el entorno de la simulación del estudio. Se concluye también que la evaluación del riesgo de conducción es un problema complejo y por lo tanto deben considerarse la participación de expertos en seguridad de tráfico para crear sistemas útiles. Los datos de salida del sistema son: seguro, aceptable y riesgoso (bajo, medio y alto).

En la compañía del caso de estudio existe un antecedente de aplicación de tecnología de análisis con *machine learning* para desarrollar una herramienta que analice el comportamiento de conducción y cómo es que el estilo de conducir afecta en el consumo de combustible. El objetivo principal es, identificar y promover un estilo de conducción que resulte en un consumo de combustible más eficiente para el vehículo. Los datos para dicho análisis fueron obtenidos desde dos fuentes principales: Telematics y el de planeación. Tras el análisis de resultados se concluyó que las buenas prácticas de conducción se traducen en un menor consumo y que, debido al tipo de datos que se procesan la herramienta resultó ser replicable para otras unidades en diferentes regiones del mundo.

En estos artículos lo que marca la distinción contra los estudios hechos con base cualitativa, es el método de obtención de la información. Marcan una diferencia donde el proceso de adquisición de datos ha sido a través de medios digitales o electrónicos y que concretamente sea información de registro numérico o en categorías. A diferencia de las herramientas estadísticas utilizadas en los estudios cualitativos, en estos métodos se posibilita el análisis lo más cercano al tiempo real y

con alta probabilidad de ser un desarrollo replicable para otras organizaciones. Un ejemplo de este tipo de análisis es pensar en que un conductor puede cambiar su estilo de conducir cada cierto tiempo dependiendo de los diferentes factores que se han identificado; por lo tanto, es precisamente una ventaja que con el análisis de datos o valores numéricos se tiene la capacidad de visualizar las variaciones mucho antes posible. Por último, otro punto a destacar es que, los estudios con enfoque un cualitativo han tendido a diagnosticar los efectos inmediatos que ha provocado la función de conducción sobre las personas; esto quiere decir, aportar una evaluación y conocimiento de los daños ocasionados y no cómo como tal un método para la prevención, por tal motivo, el análisis a desarrollar en este estudio será de manera cuantitativa con el que se pretende hacer un diagnóstico solamente con base en los datos.

2.4.5 ANÁLISIS DE LAS HERRAMIENTAS

Después de desempeñar la lectura de una variedad de estudios efectuados en los años recientes, se procede a construir un resumen de los hallazgos en cuánto a las herramientas y técnicas utilizadas para, posteriormente, determinar aquella herramienta que se ajuste mejor a nuestro caso de estudio, esto quiere decir, con base en la información disponible y que cuente con la calidad para su futuro tratamiento. De acuerdo al método de revisión de literatura se estipularon los siguientes criterios que se muestran en la Figura 2.2.

En la Tabla 2.1 se presenta el resumen de la revisión de literatura en donde se contabiliza las frecuencias del uso de las herramientas y de seccionadas dependiendo del tipo de enfoque de estudio.

Tal cómo se informa en la tabla, los estudios con un enfoque cuantitativo se han vuelto la preferencia de los análisis en los últimos 15 años (Figura 2.3). Sobre todo, el rasgo común entre los objetivos es la intención de encontrar la relación entre

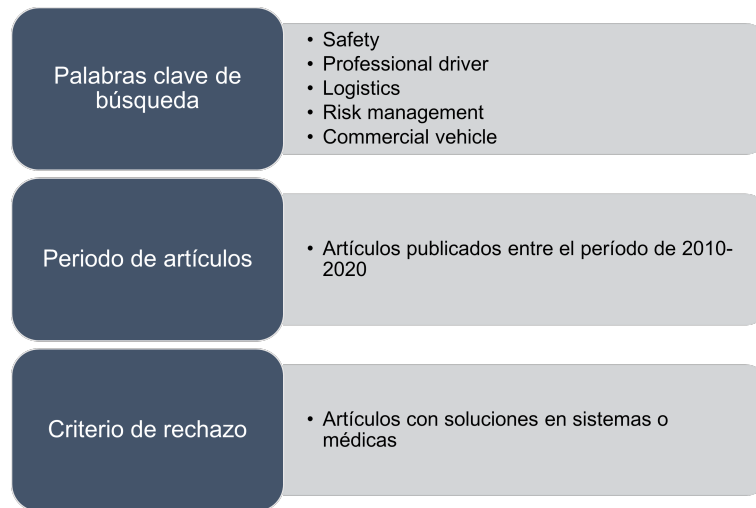


FIGURA 2.2: Criterios para la búsqueda de literatura

N	Autor	Enf	Supervisados						No Supervisados		Ambos		Estadísticos									
			RF	DT	SVM	NNC	LR	BN	Km	CA	ANN	ANFIS	KDD	DELPHI	AEM	STAT	TA	DEA	RPBL	SSIS		
1	Chen Wang (2019)	2	1	1																		
2	Chen Chen (2015)	2					1		1													
3	Berrones Sanz (2020)	1											1	1								
4	Iranitalab (2017)	2	1		1	1	1		1	1												
5	Howard (2004)	1					1								1							
6	Jahangiri (2013)	1													1							
7	Aksjonov (2018)	2									1	1										
8	Wang (2017)	2			1				1													
9	Gwak (2018)	2	1		1	1	1															
10	Higgs (2015)	2							1													
11	Wu (2020)	2		1									1									
12	Warmerdan (2016)	1															1					
13	Pereira (2020)	2																1				
14	Öz (2012)	1					1															
15	Newnam (2015)	1															1					
16	Hosseinzadeh (2021)	2			1														1			
17	Wang (2018)	2			1				1													
18	Useche (2018)	1					1															
19	Peters (2021)	1															1					
20	Hwang (2018)	2		1																		
21	Pozueco (2020)	2																				
22	Ferreira (2017)	2	1		1			1			1									1		
23	Siordia (2010)	2		1	1	1					1											
24	Quintero (2012)	2									1											
25	Constantinescu (2010)	2							1	1												
			4	4	7	3	6	1	6	2	4	1	1	1	1	2	3	1	1	1		

TABLA 2.1: Resumen de revisión de literatura. (Cualitativo: 1 — Cuantitativo: 2)

el conductor, la manipulación del vehículo y los factores externos que provoquen un estilo de conducción de peligro.

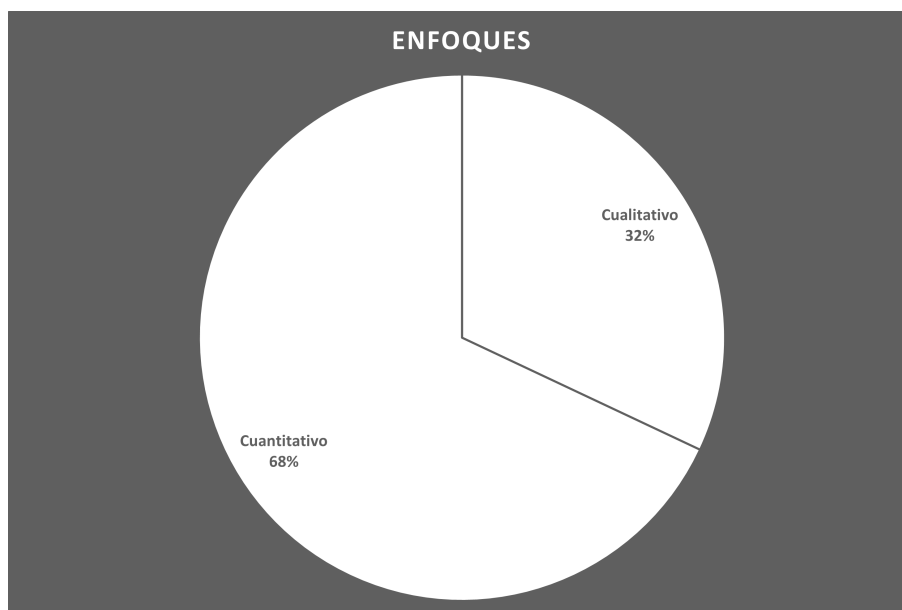


FIGURA 2.3: Relación entre estudios cualitativos y cuantitativos

A razón de que el comportamiento del conductor es un término amplio que conlleva diferentes conceptos y por lo tanto múltiples variables es por ello que la hipótesis es que, si se define y se traduce el entendimiento del comportamiento de conducción a estilos de conducción, se podrán categorizar de acuerdo a similitudes, hacer mediciones y obtener predicciones del comportamiento del conductor (Meiring y Myburgh, 2015). Para comprender mejor la relación entre los términos definidos por los autores en el entorno de los estilos de conducción, en la Figura 2.4 se muestra el contexto de los análisis. En el centro se encuentra el **estilo de conducción** que es el resultado de toda una combinación de factores que influyen directa o indirectamente sobre el. El entorno completo de la función de conducir está condicionado por: los factores internos que serían el entorno del conductor y del vehículo, que en cada uno existen diversas funciones; y también por aquellos factores externos que influyen aunque se encuentren fuera del control tanto del operador como de la unidad, como por ejemplo, el clima.

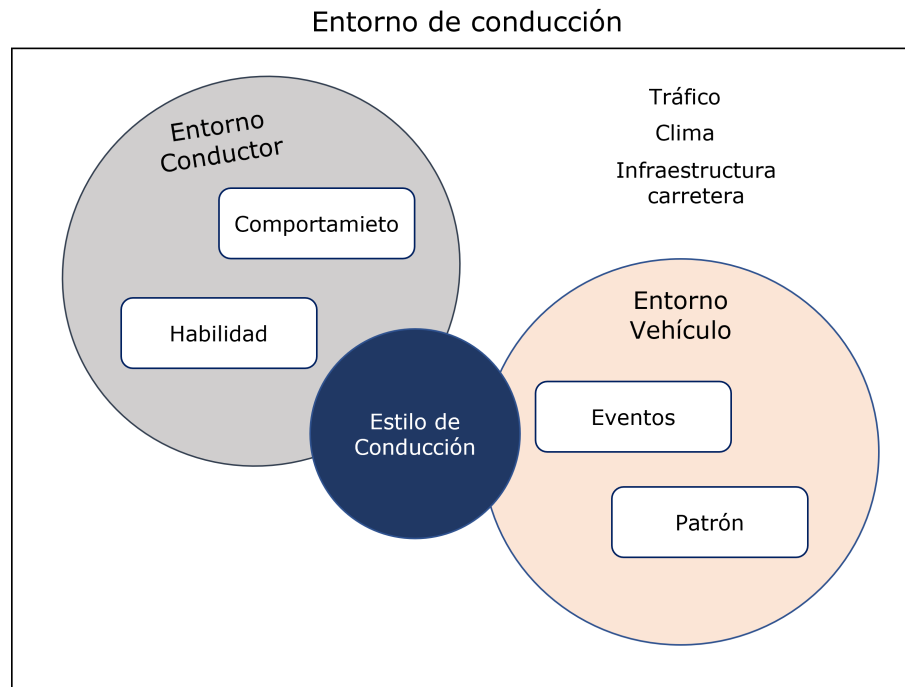


FIGURA 2.4: Términos para el estudio del conductor. Fuente: Martinez *et al.* (2018); Ellassad *et al.* (2020)

Una consideración relevante para este tipo de análisis es que últimamente han arrojado resultados más precisos y a su vez, la obtención de información proviene de fuentes más confiables debido al aumento de la utilización del registro de datos a través de los sistemas digitales. Cabe mencionar que incluso gran parte de las compañías y organizaciones están optando por soluciones con enfoque al análisis de datos. Stern *et al.* (2018) en dónde se realiza una revisión de literatura sobre las nuevas tecnologías, por ejemplo, los dispositivos de registro electrónico, los sistemas de seguridad a bordo y los sistemas telematics para localizar y dar seguimiento al transporte, tienen el potencial para ser de ayuda a los investigadores para recabar datos sobre horas de conducción, kilómetros recorridos, reducción en el control del vehículo y cualquier otra información relevante para el análisis de la exposición y el rendimiento del conductor. Los sistemas de transporte están evolucionando de ser un sistema independiente conducido por la tecnología para ser un sistema integrador de sistemas que sea impulsado por el análisis de datos; por ejemplo, mejorar las

aplicaciones de Sistemas Inteligentes de Transporte (ITS) para que ahora se basen en la calidad y en el tamaño de los datos (Bhavsar *et al.*, 2017).

En la compañía de venta y distribución de gases industriales cuenta actualmente con sistemas de Telematics en los transportes de carga dónde se registra la información del vehículo, quién conduce y de eventos durante las rutas tales como: frenados y aceleraciones bruscas, exceso de velocidad, distancias, etcétera. Esto ha permitido el conocimiento de ciertas características de riesgo durante el transcurso de la ruta pero aún no se ha llevado a cabo un análisis para identificar los estilos de conducción de su propio personal. De acuerdo con Constantinescu *et al.* (2010) este tipo de información y análisis se vuelve un medio valioso específicamente para las personas responsables de la administración de las flotas y puede ser útil para mejorar y para hacer más seguro el estilo de conducción de los varios individuos encargados de la distribución. Al mismo tiempo, tales datos también son importantes para la retroalimentación del conductor y para conocer el riesgo que representa su propio estilo de conducción. Aunque este tipo de soluciones pueden ser discutibles en si son o no eficaces, o que en su lugar sean empleados métodos activos dónde se interfiera directamente sobre el vehículo ante la inminente presencia o detección de un riesgo, la realidad es que se ha observado que el simple hecho de ser monitoreado genera en los conductores la necesidad de corregir su estilo de conducir (Martinez *et al.*, 2018).

2.4.6 HERRAMIENTAS PARA EL ANÁLISIS DE DATOS

Las herramientas que de acuerdo a la revisión han tenido una mayor presencia en la literatura son los algoritmos de *Machine Learning* (aprendizaje automático); sobre todo, los más comúnmente implementados han sido aquellos que funcionan bajo un esquema de aprendizaje supervisado pero, dependiendo del caso, también hay una aplicación de modelos para agrupamiento, de aprendizaje no supervisado, que en conjunto buscan obtener una solución para los problemas de clasificación. Según la Tabla 2.1 los algoritmos de aprendizaje automático más utilizados son

cinco y se muestran en la Figura 2.5.

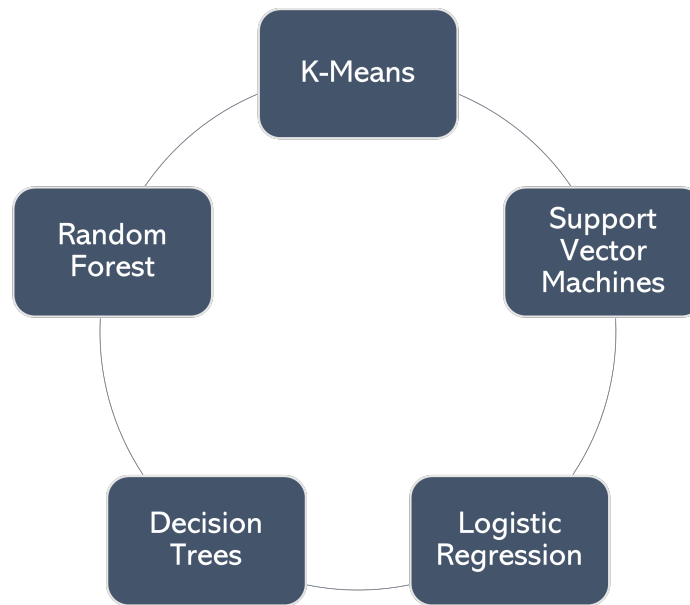


FIGURA 2.5: Algoritmos de clasificación

Novakovic *et al.* (2017) nos dicen que, existe una amplia gama de algoritmos de clasificación disponibles, en dónde cada uno cuenta con sus fortalezas y debilidades, pero no existe tal cosa dónde se coloque a uno como el “mejor” para resolver todos los problemas de clasificación o que se identifique a uno para ser utilizado para cierto tipo de problemáticas.

En la Tabla 2.6 se presentan las características generales de los algoritmos para clasificación.

Una observación importante es que en la mayoría de las experimentaciones fue necesaria la combinación de diferentes modelos y/o la comparación entre los mismos para obtener los mejores resultados, es decir, los que aporten una mayor precisión. Los modelos o algoritmos más utilizados funcionan de acuerdo al tipo de aprendizaje y aportan soluciones como de clasificación, agrupación y otros para encontrar patrones. Por ejemplo, en artículos como en el de Iranitalab y Khattak (2017) primero iniciaban aplicando un algoritmo de *K-Means* (agrupación), de aprendizaje no supervisado, dónde no hay parámetros para las agrupaciones por el desconocimiento

Support Vector Machine	Método de aprendizaje supervisado que es utilizado para la regresión y la clasificación. Se definen naturalmente para la clasificación binaria de datos numéricos
Logistic Regression	Clasificador probabilístico que construye un modelo que cuantifica la relación entre las variables de características y la variable objetivo (clase) como una probabilidad. Típicamente denominado “clasificador discriminativo”
Decision Tree	Metodología de clasificación en la que el proceso de clasificación se modela con la aplicación de un conjunto de decisiones jerárquicas sobre las variables que son características ordenados en una estructura en forma de árbol
Random Forest	Conjunto de árboles de decisión que votan juntos en una clasificación. Cada árbol se construye al azar y selecciona un subconjunto de puntos de datos. Se sabe que los RF son eficaces para evitar el sobreajuste
K-means	Método de aprendizaje no supervisado y dónde se utiliza la metodología estadística para categorizar objetos individuales en grupos con características similares (homogéneo). Utilizado cuando no se conoce el número de grupos y se desea establecer agrupaciones para analizar la pertenencia a un conjunto

FIGURA 2.6: Descripción de algoritmos. Fuente: (Constantinescu *et al.*, 2010; Junior *et al.*, 2017; Aggarwal, 2015)

de la similitud entre los estilos (Constantinescu *et al.*, 2010) y después utilizar un modelo de clasificación, en el cuál el problema de clasificación consistirá en aprender la estructura, desde el conjunto de datos ya divididos en grupos que son denominados clases (Aggarwal, 2015).

Se propone utilizar los métodos de machine learning para realizar el análisis de los datos y el desarrollo del modelo de clasificación. Un modelo de clasificación son aquellos que usan como variable dependiente a una variable categórica y variables independientes de cualquier tipo para hacer una separación de las clases objetivo a través de un modelo estadístico o de *machine learning* (Garro, 2019).

La actividad de publicaciones utilizando *machine learning* para este campo de estudio ha tenido un crecimiento explosivo, teniendo un pico entre los años 2017 y 2018. En una revisión de literatura hecha por Ellassad *et al.* (2020) identificó que los modelos han sido utilizados principalmente para dos aplicaciones:

1. Comparación de las técnicas de *machine learning*

2. Comparación entre las técnicas estadísticas y *machine learning*.

Para el segundo punto, los resultados de comparación han demostrado que en la mayoría de las experimentaciones los modelos de *machine learning* han superado a los modelos que no lo son obteniendo mejores niveles de precisión.

Este tipo de algoritmos que están basados en datos son apropiados para procesar una gran cantidad de información e incluso mejorar algoritmos previos. Además, cuentan con una capacidad de adaptación que pueden lograr especificarse en la individualidad del conductor (Martinez *et al.*, 2018).

En la actualidad en diferentes países los estudios están recurriendo a la implementación de los algoritmos de *Machine Learning*, que es una rama de la inteligencia artificial, para el estudio de los datos en diferentes áreas para crear modelos que aprendan de los datos y mejoren su precisión con el tiempo. Dichos algoritmos encuentran patrones y características en cantidades masivas para tomar decisiones y predicciones basados en nuevos datos. Otra definición nos dice que, "*se trata de una colección de herramientas o técnicas que permiten a las computadoras automatizar la creación y programación de modelos basados en datos a través de descubrir los patrones sistemáticamente que sean estadísticamente significativos en la información disponible*", (Bhavsar *et al.*, 2017).

El ideal de este sistema de aprendizaje es emular la manera en que los humanos aprendemos a procesar señales sensoriales para lograr una meta; en otras palabras, se puede decir que son capaces de mejorar su rendimiento utilizando información desde la experiencia (Novakovic *et al.*, 2017). *Machine Learning* funciona bajo cuatro tipos de métodos de aprendizaje que son: supervisado, sin supervisión, semi-supervisado y de refuerzo (Figura 2.7) (Bhavsar *et al.*, 2017).

Las técnicas del aprendizaje automático operan construyendo algoritmos que son guiados por datos en lugar de apoyarse en programadores humanos para proveer instrucciones codificadas y que inicialmente sean clasificadas dependiendo del tipo

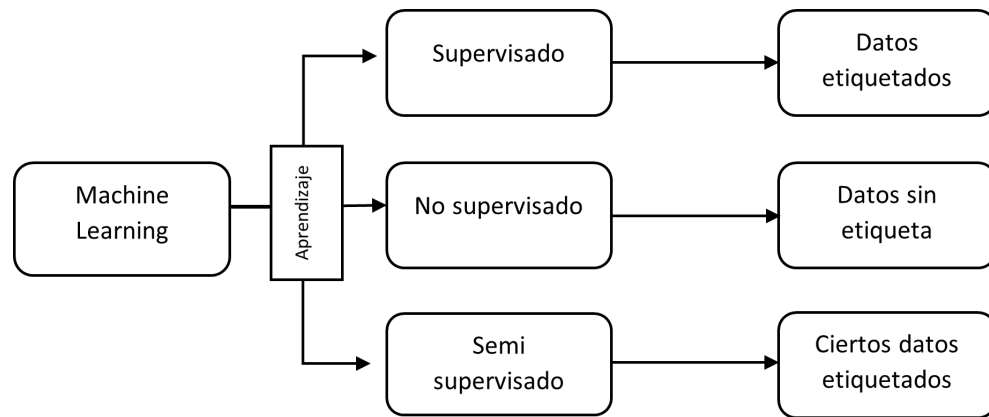


FIGURA 2.7: Tipos de aprendizaje

de aprendizaje.

Es importante señalar que para la compañía del caso de estudio, la problemática surge a raíz de la necesidad de contar con una herramienta que sea más precisa y confiable para la identificación de riesgos o factores que pueden aumentar la accidentabilidad en las rutas de distribución de gas. El principal propósito es el desarrollar un análisis y crear un sistema que ayude a mejorar la gestión de riesgos para poder disminuir la incertidumbre en el sistema logístico y las posibles vulnerabilidades de entrega y a su vez, disminuir los efectos de los riesgos ocupacionales para los conductores profesionales.

2.4.7 MINERÍA DE DATOS

Un concepto esencial, que se trata de un proceso típico para llevar a cabo un análisis estadístico es la *minería de datos*. Consiste en un conjunto de métodos o técnicas que están dirigidas para extraer conocimiento inherente en las bases de datos (Corso, 2007). Facilitan la exploración de los datos que sería complejo para los humanos visualizar y comprender grandes cantidades de datos. Las habilidades de análisis humano no pueden escalar para alta dimensionalidad ni volúmenes masivos de información (Fayyad y Stolorz, 1997).

La masificación de los datos ha sido el resultado de los avances tecnológicos y de la digitalización en casi todas las áreas de la vida moderna. Por eso se ha vuelto común dónde se enfatice en la extracción de ideas concisas y sobre todo que sean procesables en la información disponible para objetivos específicos. Es aquí dónde entra en función las técnicas de minería de datos que ayuda a aliviar la recopilación de datos que se encuentra sin estructura, arbitraria o que se encuentra en un formato que no es adecuado para su tratamiento. (Aggarwal, 2015) Las técnicas se basan en la secuencia de pasos de la Figura 2.8, lo cual puede ser un proceso reiterativo; esto quiere decir que se puede volver desde el paso 1 si así lo requiere. Las bases de este proceso se encuentra en la inteligencia artificial por lo tanto ambos van de la mano (Corso, 2007).

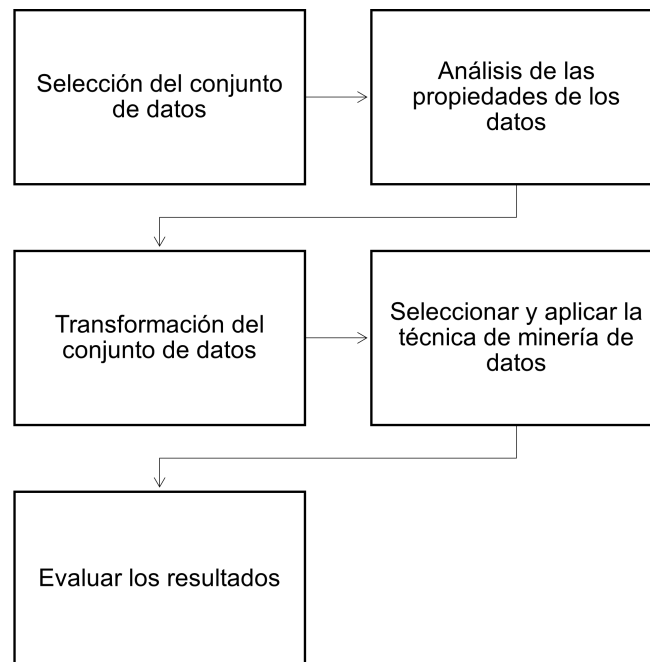


FIGURA 2.8: Proceso de minería de datos. Fuente: (?Fayyad y Stolorz, 1997; Corso, 2007)

2.4.8 EVALUACIÓN DE DESEMPEÑO DE LOS MODELOS

Un paso importante y necesario para la aplicación de cualquier algoritmo de clasificación es hacer una evaluación del desempeño del modelo. Para que el proceso de evaluación sea considerado válido, lo mejor es que sea probado con un conjunto de datos que no haya sido utilizado para el entrenamiento del modelo. Es por ello que previo a la aplicación del algoritmo, el conjunto de datos se debe separar en dos grupos: de entrenamiento (*training*) y de prueba (*testing*). El algoritmo se entrena con el grupo de entrenamiento y para evaluar la eficacia del modelo, se utiliza el conjunto de datos de prueba. (Bhavsar *et al.*, 2017; Iranitalab y Khattak, 2017)

En el caso de los algoritmos de aprendizaje supervisado, existe un tipo de problema de instrucción en donde el modelo no generaliza bien entre los datos observados y los que no son vistos, a esto se le llama sobreajuste o mejor conocido como *overfitting*. Por la existencia del *overfitting*, el modelo trabaja bien en el conjunto de datos de entrenamiento (*training*), pero no se ajusta en el conjunto de prueba (*testing*) (Ying, 2019).

Entonces, ¿cómo es que se puede definir que el modelo generaliza bien con los datos no vistos? Dentro de los métodos de evaluación de desempeño para la aplicación en los algoritmos de clasificación, los más utilizados son: *accuracy*, *recall* (sensibilidad), *F1-Score*, matriz de confusión, curva ROC y AUC. En la Tabla 2.2 se muestran las características generales de cada una de las técnicas de evaluación.

2.5 CONCLUSIONES DEL CAPÍTULO

El tema de seguridad en el transporte de carga ha acogido una preocupación y en consecuencia un aumento de estudios y análisis de posibles soluciones tanto por parte de las organizaciones internacionales así como en las compañías. Los mecanismos de seguridad han trascendido más allá de sólo hacer instalación de sistemas

Métricas	Características
<i>Accuracy</i>	Proporción de resultados verdaderos contra el número total de casos examinados. Facilidad de cálculo y de comprensión para evaluar la efectividad general del algoritmo.
<i>Recall</i>	Medida para conocer cuántos valores positivos son correctamente clasificados. En un modelo perfecto, sería igual a 1 para cada clase.
<i>F1-Score</i>	Métrica utilizada para problemas dónde el conjunto de datos está desbalanceado. Combinación de precisión y recall para obtener valor más objetivo.
<i>Matriz de confusión</i>	Permite visualizar en una tabla de contingencia la distribución de los errores cometidos por el clasificador.
<i>ROC</i>	Representación gráfica de 2 dimensiones para clasificadores dicotómicos, la sensibilidad (eje Y) contra especificidad (eje X) por clase.
<i>AUC</i>	Medida de comparación para la gráfica ROC que sirve para conocer qué tan bien el parámetro puede distinguir entre dos grupos/clases

TABLA 2.2: Métricas de evaluación de algoritmos. Fuente: (Borja-Robalino *et al.*, 2020; Corso, 2007; Schoonjans, 2021; Badra, 2020)

como GPS, Telematics e ITS o de estudios de los padecimientos médicos a causa de las prolongadas jornadas de conducción. Las más recientes investigaciones han recurrido, aprovechando el acceso a la recopilación y registro de datos a través de los sistemas electrónicos, a los análisis enfocados en datos para entender más a fondo cuáles son los factores influyentes que en conjunto provocan el aumento de riesgo de accidente. Los puntos relevantes a destacar de este capítulo son:

- El transporte de carga se ha vuelto parte esencial de los sistemas logísticos. Incluso, ciertos autores lo atribuyen como una función estratégica para las organizaciones.
- Las exigencias sobre el área de distribución ha provocado efectos negativos

sobre el conductor y esto ha aumentado la cantidad de accidentes de tráfico en todo el mundo.

- La mayoría de los accidentes de tránsito están atribuidos al factor humano. La fatiga es un componente principal, pero según los estudios más recientes, el estilo de conducción está siendo uno de los mayores actores en esta problemática.
- Para paliar esta problemática la gestión de riesgos ha ido evolucionando para reducir la incertidumbre de los procesos. Por lo tanto, es primordial contar con una gestión de riesgos para la seguridad del transporte.
- El no invertir en la seguridad del transporte. representa un alto costo para las organizaciones a razón de que una disrupción en la cadena será perjudicial para cumplir con las metas.
- Organismos internacionales, como la ONU, están promoviendo planes y metas para disminuir las lesiones y muertes, a las que han catalogado ahora como “prevenibles” a causa de los accidentes de tránsito. El llamado es para que se sumen tanto gobierno como empresas.
- El análisis del estilo de conducción ha presentado mejores resultados con el uso de algoritmos de la rama de aprendizaje automático (*machine learning*)
- Se identifica que los modelos que han logrado una mejor clasificación de los estilos son: *Support Vector Machine*, *Logistic Regression*, *Decision Tree*, *Random Forest* y *K-Means*.
- La diversidad en la implementación de distintos algoritmos en el análisis de datos se debe a qué no existe tal cosa como el “mejor” que sea aplicable ante cualquier tipo de problema de clasificación; por lo que se debe totalmente al tipo y calidad de la información y al caso de problema que se quiere resolver.

Esta exploración en la literatura nos muestra que existe un interés creciente en el tema de la seguridad y particularmente en el área de la movilidad. Si ha

existido una barrera anteriormente, fue la razón de la poca capacidad de registro de información concreta y confiable para estudiar las causas que provocan el aumento del riesgo. Ahora al contar con las fuentes y la masificación del registro de datos, es posible, como lo plantea la ONU, hacer estudios para la prevención de daños, lesiones y muertes. Para las compañías esto significa garantizar una administración de riesgos efectiva en el eslabón final que parte hacia el cliente y no menos importante, el cuidado de los conductores y el producto en transportación. Estos argumentos son importantes que deben ser tomados en cuenta para la planeación logística que soportaran los objetivos y la disminución de costos por enfrentar las posibles vulnerabilidades de entrega.

Por último, en este capítulo se propone utilizar las técnicas y procesos del aprendizaje automático (*machine learning*) para desarrollar una herramienta que beneficie a las áreas de seguridad para analizar con datos y extraer el conocimiento que existe implícitamente. Además que esto ayudará a la administración en la creación de una mejor gestión para la prevención de riesgos y la retroalimentación en la logística de distribución y sus profesionales de la función.

CAPÍTULO 3

METODOLOGÍA

La función de distribución ha ido adquiriendo su nivel de significancia en las cadenas de suministro debido a la globalización y por consiguiente, el crecimiento del mercado. Este crecimiento también ha conseguido que las tareas de logística se hagan más complejas por el aumento de las rutas, distancias y unidades que deben administrarse para realizar las entregas en tiempo y forma de acuerdo a las especificaciones del cliente. En dicha complejidad, los principales actores que han sufrido las transformaciones en el área son los conductores que en ellos recae el cumplimiento de la planeación y metas de toda una organización. Por tal motivo, se vuelve fundamental desarrollar mejores estrategias dándole el mismo nivel de importancia en la seguridad así cómo en los otros procesos de una cadena para disminuir o eliminar, de ser posible, las vulnerabilidades de la función para garantizar un buen nivel de servicio. Es por esta razón que se propone utilizar los algoritmos de clasificación de aprendizaje automático que de acuerdo con la revisión de literatura se han empleado para identificar, segmentar y clasificar los estilos de conducir que pueden resultar en un riesgo para el conductor.

La finalidad de este capítulo es proponer un método de análisis que sea de apoyo para el área de seguridad o gestión de riesgos y que sea una opción viable para atender la problemática expuesta en la sección 1.1 de este trabajo. La metodología consta de las siguientes etapas: definición de la problemática, recopilación y pre-

procesamiento de datos, selección de la o las herramientas y, por último, modelación y validación. En cada fase existen sub-procesos que hacen función de soporte para los procesos de análisis y aprendizaje, dónde de acuerdo con la minería de datos, son herramientas que no se pueden desvincular uno del otro.

3.1 DESCRIPCIÓN DE LA METODOLOGÍA

La metodología que se recomienda llevar a cabo el estudio tiene base, de acuerdo a la literatura, en el proceso y subprocesos de la minería de datos. Este proceso es un flujo de trabajo que contiene fases como: la limpieza de datos, extracción de características y diseño de algoritmo (Aggarwal, 2015). El sistema es ajustable para trabajar con diferentes tipos de problemas y esencialmente para encontrar el método que se adapte mejor de acuerdo a las condiciones de la información disponible.

La visión es que este proceso sea incorporado como parte de las actividades de análisis y gestión de riesgos en las áreas de la administración del transporte para extraer conocimiento desde la propia información de la compañía y así brindar una mejor retroalimentación sobre el comportamiento del conductor en las rutas. Incluir esta etapa facilitará la toma de decisiones más asertivas para la prevención de incidentes.

Entonces, para la adaptación de este sistema de análisis será crucial el lograr unificar la información que actualmente se encuentra dividida entre diferentes funciones. Para una mejor visualización del proceso, en la Figura 3.1 se muestran las etapas de la técnica de minería de datos.

En términos generales, cada fase debe cumplir ciertos objetivos que se describen a continuación:

- **Fase 1: Definición de la problemática.** Definir qué es lo que se quiere lograr y con base a ello iniciar la búsqueda de la información necesaria.

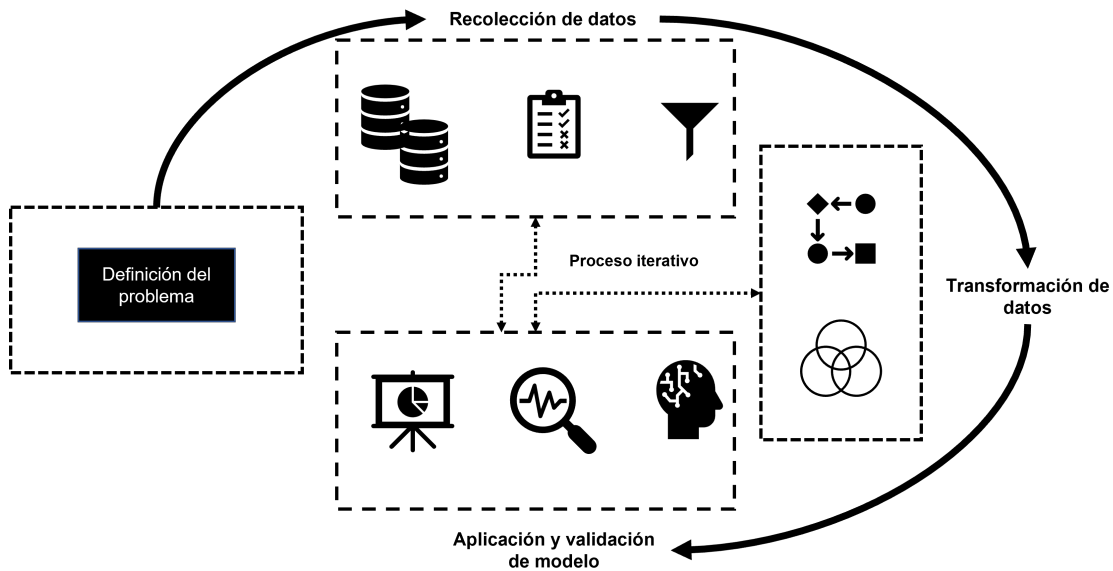


FIGURA 3.1: Metodología propuesta para análisis de conductores. Fuente: Adaptación de propuestas por (?Fayyad y Stolorz, 1997; Wu y Chaipiyaphan, 2020)

- **Fase 2: Recolección de datos y pre-procesamiento de datos.** Proceso de selección, reducción y pre-tratamiento de datos. Básicamente, estructurar la información para que tenga la calidad necesaria para su transformación. En este caso de estudio, recopilar los datos referentes a los conductores y encontrar los registros más confiables sobre los vehículos y cómo se relacionan con el comportamiento de manejo del conductor.
- **Fase 3: Selección de la o las herramientas.** Después del análisis del estado de la información, se eligen la o las herramientas que transformen los datos para mejorar su calidad y facilitar el proceso de aprendizaje para los algoritmos
- **Fase 4: Modelación y validación.** Por último, con el dataset de entrenamiento preparado, se ajustan los datos en los algoritmos para el aprendizaje del modelo y finalmente se hacen las aplicaciones sobre el conjunto de datos de prueba para validar el desempeño del modelo con las métricas utilizadas en *Machine Learning* para medir la eficiencia.

Es importante resaltar que entre las fases 2 y 3, puede presentarse la necesidad

de hacer un proceso reiterativo dónde pueden agregarse o eliminarse datos para probar si existen cambios significativos en la eficiencia de la información o si el modelo mejora o disminuye su precisión.

3.2 FASE 1: DEFINICIÓN DE LA PROBLEMÁTICA

La definición del problema es un importante primer paso para cualquier aplicación de las técnicas de aprendizaje automático que tendrá un efecto consecutivo a todos los siguientes pasos hasta que sea desarrollado el modelo (Bhavsar *et al.*, 2017).

La problemática surge a raíz de la necesidad de tener un mecanismo más preciso y asertivo para la identificación de riesgos o factores que pueden aumentar la accidentabilidad en las rutas de distribución del gas. El fin es de proveer una un ambiente más seguro para los conductores y disminuir los efectos de los riesgos ocupacionales.

Como se ha mencionado en el capítulo 2, la compañía cuenta con una política de priorizar y enfatizar el desempeño de las funciones en un ambiente seguro. A pesar de los esfuerzos constantes hechos por la organización con la implementación de estrategias para entender mejor el entorno de la transportación, sigue siendo incierto cuáles son los atributos o patrones característicos que incrementan la accidentabilidad.

3.3 FASE 2: RECOPIACIÓN Y PRE-PROCESAMIENTO DE DATOS

3.3.1 RECOPIACIÓN DE DATOS

El entorno de trabajo de un conductor está influenciado por factores internos, externos e híbridos (Figura 3.1), sin embargo aunque las compañías utilizan sistemas logísticos para satisfacer las necesidades de los clientes y mantener una supervisión de las rutas, pocas han intentado o logrado utilizar los datos derivados de dichos sistemas para realizar evaluaciones del desempeño de las entregas. Desempeñar diagnósticos para identificar las vulnerabilidades del transporte en un sistema logístico es fundamental para poder gestionar de una manera eficaz el riesgo logístico (Wu y Chaipiyaphan, 2020).

La toma de decisiones informada de los posibles riesgos presentes hace posible una gestión de riesgo que es capaz de actualizar la probabilidad de un incidente y con ello mejorar el sistema de prevención, lo que hace que este tipo de análisis sea un proceso de desarrollo dinámico a diferencias de los métodos convencionales de gestión de riesgos (Parhizkar *et al.*, 2020).

Tal como se explica en el capítulo 2, existe un entorno de conducción, resumido por diferentes autores, el cuál sirve como una guía para conocer cuál es el tipo de información que se requiere para desarrollar un análisis dependiendo del enfoque que se pretende estudiar. La Tabla 3.1 muestra cuáles son algunos tipos de datos que se pueden obtener de la relación conductor-vehículo-entorno.

En la Figura 3.2 se representa la secuencia de pasos general para la recopilación de datos.

1. Con el conocimiento de cuáles son el tipo de variables influyentes del entorno

Conductor	Vehículo	Entorno
Histórico informativo (ejemplo): <i>experiencia</i> <i>accidentes</i>	Eventos de conducción (ejemplo): <i>aceleraciones</i> <i>frenos</i> <i>excesos de velocidad</i> <i>giros de volante</i>	Condición de carretera
		Clima
		Tráfico
		Políticas
		Regulaciones de gobierno
Comportamiento	Histórico de rutas	Clima organizacional

TABLA 3.1: Variables del entorno de conducción

de conducción, se procede a la búsqueda y recopilación de las observaciones. Buscar en dónde se almacena la información con el fin de obtener la mayor cantidad de antecedentes o históricos de la información. Por ejemplo, desde cuándo se registran las observaciones, cuál es el método y formato de recopilación, y qué tan confiable o precisa es la técnica de registro de datos.

2. Seleccionar qué tipo de información cumple de acuerdo a las variables del entorno de conducción
3. Limpiar o filtrar datos, por ejemplo, que no estén completos, con espacios en blanco o no estandarizados



FIGURA 3.2: Proceso de recopilación de datos

3.3.2 PRE-PROCESAMIENTO DE DATOS

En la aplicación de la minería de datos, hay un reto importante a remarcar; cada aplicación es única por lo que, es complicado crear técnicas generales y/o que logren ser reutilizables para distintas aplicaciones (Aggarwal, 2015).

Entonces, partiendo de la recopilación de datos, considerando que los registros son de propiedades diferentes entre sí, esto quiere decir que ya sean en forma cualitativa o cuantitativa, requieren pasar por un proceso de pre-procesamiento. En muchos casos, la información se encuentra al alcance y disponible pero no en la estructura ni en la categorización deseable para realizar un análisis. Esto se vuelve una actividad importante dónde se estudia qué datos están siendo utilizados efectivamente y cuáles se desestiman por falta de conocimiento de la influencia que dichos factores pueden tener sobre el desempeño, en este caso, de la seguridad en los transportes de carga pesada.

El proceso para el manejo y adquisición de conocimiento de los datos es conocido bajo el término de KDD (*Knowledge Discovery in Databases*) y que ha tenido su expansión de uso en los últimos años y que se trata precisamente de la preparación de los datos, la selección y limpieza de los mismo, la incorporación de conocimiento previo, y la propia interpretación de los resultados de minería de datos (Riquelme *et al.*, 2006). Los campos de minería de datos y aprendizaje automático se encuentran entrelazados. La minería de datos utiliza los algoritmos de *machine learning* para interrogar grandes bases de datos y descubrir conocimiento oculto en la información, mientras que varios algoritmos de machine learning emplean métodos de minería de datos para pre-procesar la información antes de aprender las tareas deseadas (Naqa y Murphy, 2015).

Para la estructuración de la información y de los datos, Wang *et al.* (2019) nos dice que primero es necesario establecer unos lineamientos. Ellos proponen:

- ¿Cómo definir el riesgo de conducción?

- Determinar qué factores deben ser utilizados para medir o predecir el riesgo
- Encontrar la manera precisa de capturar los patrones de riesgo de conducción

Acerca de la relación modelo-objetivo, existe en la literatura algunas recomendaciones para los diferentes objetivos, por ejemplo: si hay un problema de clasificación se propone utilizar máquinas de vectores de soporte o árboles de decisión, si hablamos de un caso de regresión se puede optar por los árboles de regresión o redes neuronales o si se requiere de una agrupación, los modelos jerárquicos pueden ser la opción, etcétera (Riquelme *et al.*, 2006).

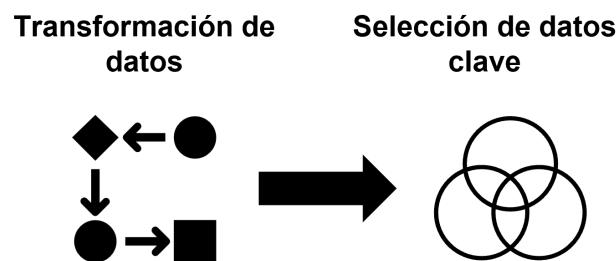


FIGURA 3.3: Proceso de estudio de datos

El proceso de pre-procesamiento de datos se resume en las dos actividades presentadas en la Figura 3.3.

1. Transformación de datos procurando que sigan un mismo formato o normalización, observar la dimensionalidad, datos etiquetados o no y si las clases son balanceadas o no
2. Análisis de los factores que influyen más o menos sobre el comportamiento del modelo

Existen métodos para el pre-procesamiento de los datos. La meta de la reducción de datos es poder representarla de una manera más compacta ya que a medida

que la cantidad de información es menor será más sencilla la implementación de algoritmos más sofisticados y/o que puedan ser “pesados” en términos computacionales (Aggarwal, 2015).

Un ejemplo de una secuencia de transformación de datos se muestra en la Figura 3.4. Es una forma práctica de representar un proceso típico de preparación de la información.

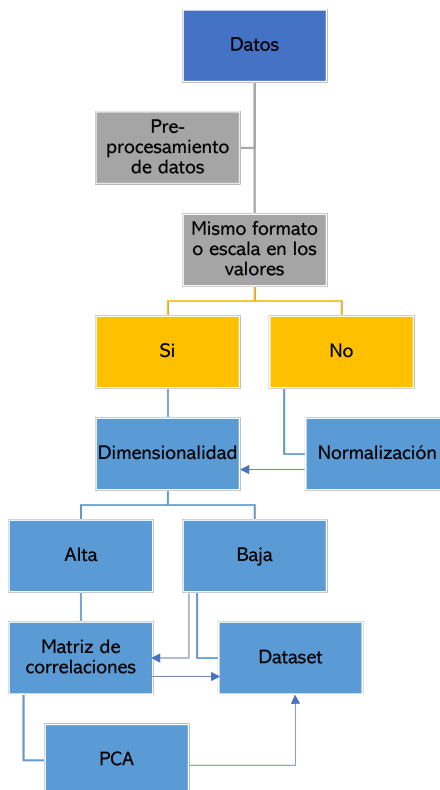


FIGURA 3.4: Ejemplo de tratamiento de datos (Elaboración propia)

El objetivo final de esta fase se resume como la integración de los datos que provienen desde las distintas fuentes o sistemas de almacenamiento, para la construcción de un objeto denominado, *dataset*. A manera de aportar una sencilla explicación, se trata de un trabajo en dónde los factores (variables) se encuentran en las columnas y los casos u observaciones en las filas.

La importancia de conocer realmente a qué tipo de datos nos enfrentamos es la llave que será condicionante para definir qué técnicas o métodos serán los apropiados

para el procesamiento de información y su correcto análisis. Bhavsar *et al.* (2017) nos dicen que, es de suma importancia recalcar que, sin importar el tipo de aprendizaje que se elija para hacer el estudio, la calidad, el tipo y el tamaño de los datos son factores significantes que afectan la eficiencia, la precisión y lo robusto del algoritmo de clasificación.

3.4 FASE 3: SELECCIÓN DE LA O LAS HERRAMIENTAS

Al llegar a esta fase del proceso, según Riquelme *et al.* (2006) la elección del modelo se determina bajo dos condicionantes: qué tipo de datos y cuál es el objetivo que se pretende lograr. En este paso se encuentra la parte fundamental de la minería de datos en dónde se pretende extraer el conocimiento de la información. Es la parte esencial del proceso en dónde se utilizan algoritmos específicos para obtener conocimiento y encontrar patrones. Las definiciones de estos conceptos esenciales se encuentran en la Tabla 3.2.

Conocimiento (<i>Knowledge</i>)	Patrón
En bases de datos, se trata del proceso para identificar patrones válidos, potencialmente útiles, nuevos y comprensibles	La obtención de patrón se entiende como la designación del ajuste de un modelo a los datos, buscar una estructura a partir de los datos y/o la descripción de alto nivel de un conjunto de datos

TABLA 3.2: Conceptos de minería de datos

Para este punto los datos deben encontrarse en una estructura de formato *dataset*; con la información preparada o transformada, existen dos características relevantes que deben valorarse previo a la implementación de cualquier algoritmo: si los datos se encuentran etiquetados (identificados con cierto grado de categorización) y si están balanceados o no. Es de suma importancia prestar atención a este paso (Figura 3.5), debido a que, en las situaciones dónde se presente un caso de clases desbalanceadas, se verá afectado el desempeño del clasificador, por lo que podemos utilizar diferentes técnicas o métodos para balancear la información. Recordar que el

buen funcionamiento del algoritmo depende enormemente de la calidad del conjunto de datos de entrenamiento desde el inicio (Li *et al.*, 2014). También la condición de las etiquetas es fundamental porque, tal cómo se menciona en la sección de antecedentes, la selección de la herramienta queda condicionado al tipo de aprendizaje.

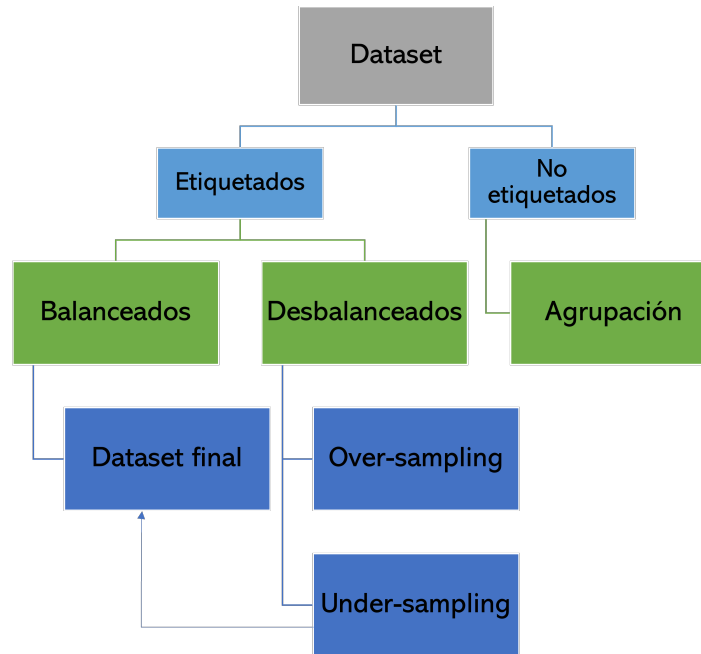


FIGURA 3.5: Proceso de etiquetas y clases (Elaboración propia)

Luego del proceso de transformación de los datos es cuando es posible, ahora si, proceder a la fase de selección de la técnica o algoritmo ideal para estudiar nuestra problemática. De acuerdo con los antecedentes, recordemos que nos enfrentamos ante una problemática de clasificación de estilos de conducción. Los algoritmos más utilizados para lograr dicho objetivo son aquellos de agrupación y clasificación.

Por lo tanto Aggarwal (2015) nos dice que, mientras el problema de la agrupación es el determinar grupos similares de puntos de datos, el de clasificación es el de aprender la estructura de un conjunto de datos, divididos en grupos, que se conocen como categorías. El aprendizaje de las categorías se logra típicamente con un modelo.

3.4.1 ALGORITMOS DE AGRUPACIÓN

Uno de los algoritmos más utilizados para la agrupación de observaciones o registros en clases de objetos similares es el algoritmo de *k-means*.

La técnica de *k-means* se trata de un algoritmo de aprendizaje no supervisado. Esto quiere decir que no se introducen datos “etiquetados” para hacer la función de agrupación, a diferencia de aquellos algoritmos que si son supervisados. Para simplificar el concepto, *k-means* desempeña una división de los objetos y los agrupa de acuerdo con las similitudes que comparten y la separación es por lo que no tienen en común. El término K se trata de un número; por lo que es imprescindible previamente establecer cuántas agrupaciones debe crear el sistema (Kumar, 2021).

Para dicho propósito, uno de los métodos más utilizados para determinar la cantidad de agrupaciones óptimas (K) es el *Método Elbow*. Se trata de un método que examina el porcentaje de la varianza que se explica en función del número de agrupaciones (Bholowalia y Kumar, 2014).

3.4.2 ALGORITMOS DE CLASIFICACIÓN

Antes de describir la particularidad de los algoritmos de clasificación, es pertinente insistir que, no existe tal cosa como catalogar que algún o ciertos algoritmos en particular sean los que funcionen bien ante un específico grupo de datos (*data set*). De hecho, el contar con una diversidad de algoritmos de clasificación sea ha vuelto un desafío para los analistas para elegir el clasificador adecuado para su problema (Zhongguo *et al.*, 2017).

Los algoritmos de clasificación supervisada son una técnica que se emplea para identificar y categorizar nuevas observaciones con base en los datos de entrenamiento. El sistema aprende de la información obtenida y luego clasifica los nuevos datos en clases. La diferencia entre una regresión y una clasificación se encuentra en que la

variable de salida será una categoría y no un valor (Jaiswal, 2019).

De acuerdo con los antecedentes, son principalmente cuatro algoritmos clasificadores que se han empleado para resolver este tipo de problemática y son los siguientes:

1. Regresión Logística - *Logistic Regression* (LR)
2. Máquina de Vectores de Soporte - *Support Vector Machines* (SVM)
3. Árboles de Decisión - *Decision Trees* (DT)
4. Bosques Aleatorios - *Random Forest* (RF)

Para poder elegir la herramienta que se ajuste mejor a los datos es importante considerar las características de cada uno que se presentan en la Tabla 3.6. La información sirve como una guía para conocer las ventajas y desventajas de cada uno de los cuatro tipos de clasificadores más utilizados.

Clasificador	Ventajas	Desventajas
LR	Utilizado para comprender la influencia de muchas variables independientes en una única variable de resultado	Útil sólo para problemas binarios. Asume que las variables predictoras son independientes entre sí y asume que no hay faltantes en los datos
SVM	Efectivo en datos con alta dimensión. Buena eficiencia de memoria. Preferidos por menor potencia de cálculo y precisión	Limitaciones de velocidad, tamaño de datos y kernel. No proporciona estimaciones de probabilidad directamente.
DT	Sencillo de comprender y visualizar. Funciona bien con poca preparación de datos. Trabaja tanto con datos numéricos y categóricos. Utilizados para análisis de decisiones	Complejos árboles que no generalicen bien. Pueden ser inestables al tener pequeñas variaciones que generen un árbol completamente diferente. Debe contar con un criterio de "parada" para el tamaño del árbol.
RF	Poca preparación de datos y sencillo para crear modelos. Tiene alta precisión y resuelve problemas predictivos. Detecta automáticamente valores atípicos. Reducción en el sobreajuste.	Requiere más tiempo, alto costo computacional y memoria.

FIGURA 3.6: Ventajas y desventajas de los clasificadores. Fuente: Pedamkar (2021); Garg (2018)

Por último, otra de las características que sirve para diferenciar entre el funcionamiento de los algoritmos clasificadores es que se dividen en dos categorías identificados como: lineales y no lineales (Figura 3.7).



FIGURA 3.7: Tipos de modelos de clasificación

3.5 FASE 4: MODELACIÓN Y VALIDACIÓN

Dentro de esta última fase se materializa el trabajo del procesamiento de los datos llevado a cabo en las fases anteriores para dar inicio a la implementación del modelo de análisis. De acuerdo con el conocimiento exploratorio de la compañía, el tipo de desarrollo de modelo será de acuerdo a un algoritmo semi-supervisado. Esto quiere decir que se requiere utilizar dos métodos diferentes, tal como lo especifica el nombre, primero trabajar con un agrupamiento no supervisado y posteriormente la aplicación de técnica una supervisada.

Las tres etapas que se llevan a cabo en esta fase son las que se muestran en la Figura 3.8.

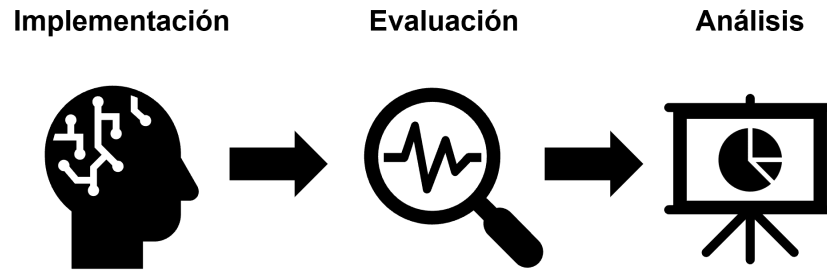


FIGURA 3.8: Proceso de experimentación y validación

3.5.1 MODELACIÓN

La fase de experimentación y el desarrollo del modelo de análisis será manejado con el lenguaje de programación conocido como *Python*. Este sistema de lenguaje de programación cuenta con las siguientes ventajas:

1. Programas de acceso libre y gratuito
2. Sencillo lenguaje de programación
3. Cada técnica utilizada en este estudio se encuentra disponible y accesible en librerías gratuitas
4. Librerías con todos los modelos de aprendizaje automático y que cuentan con actualizaciones cada cierto periodo
5. Actualmente en la compañía dónde se realiza este caso de estudio, cuenta con un proyecto dónde se imparte capacitación para enseñar este lenguaje de programación a los empleados de cualquier departamento con el propósito de que puedan realizar análisis estadísticos

Para realizar el entrenamiento de los datos debe hacerse una separación de la información en un porcentaje de 80 % como datos de entrenamiento y 20 % como datos de prueba. El desarrollo del modelo será como se muestra en la Figura 3.9.

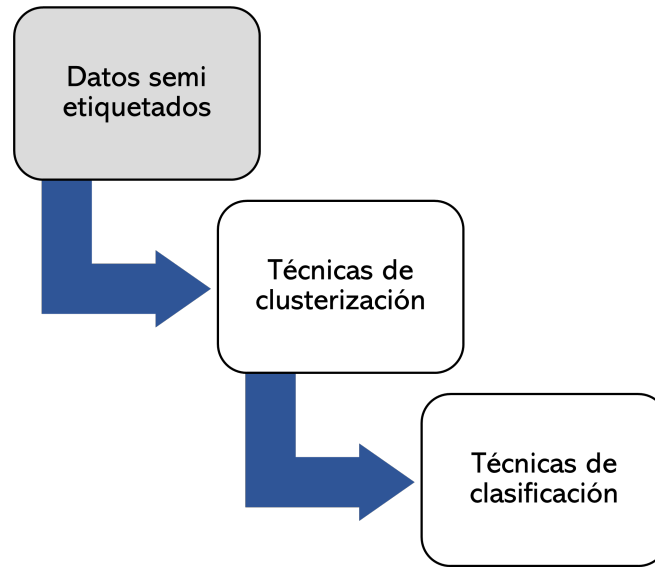


FIGURA 3.9: Proceso de desarrollo de modelo de aprendizaje semi-supervisado

3.5.2 VALIDACIÓN DE LOS MODELOS

Según Borja-Robalino *et al.* (2020) nos dice que dependiendo del tipo de datos, esto quiere decir, la estructura del *dataset*, se deben cumplir ciertos criterios para elegir el método conveniente de evaluación para medir el desempeño de los modelos. En la Figura 3.10 podemos ver los métodos de evaluación dependiendo de las características mismas de los datos. Para comprender el diagrama, cuando hablamos de *datos desbalanceados* nos referimos a un problema de clasificación en dónde las clases no se encontrarán representadas igualitariamente (Brownlee, 2020).

3.5.3 ANÁLISIS DE RESULTADOS

Los hallazgos posteriormente al desarrollo del modelo y su implementación serán para analizar cuáles son los factores más influyentes sobre la variable objetivo que es el “Incidente”. Con el soporte de conocer aquellos conductores que sean clasificados como parte del nivel de riesgo podrán obtener una retroalimentación más confiable. Por otra parte, el estudio aporta a la compañía la oportunidad de



FIGURA 3.10: Criterios para elegir la métrica de evaluación de modelos dependiendo del tipo de datos. (Borja-Robalino *et al.*, 2020; Corso, 2007; Badra, 2020)

enfaticar esfuerzos sobre las variables que se relacionan con el riesgo y plantear soluciones para disminuir ese factor.

3.6 CONCLUSIONES DEL CAPÍTULO

La metodología que se propone llevar a cabo es con base en la minería de datos y bajo un esquema del tipo aprendizaje semi-supervisado. Este tipo de método es una guía útil que lleva paso a paso, desde la recolección de la información, transformación de los datos, pre-procesamiento y finalmente, el análisis de los datos con la ayuda de los modelos. Otra ventaja que surge es la oportunidad de poder identificar el tipo de información que será útil y necesaria para el desarrollo de proyectos de mejora o para poder llevar a cabo análisis más profundos y abrir la oportunidad para generar nuevas ideas de análisis para cualquier otro tipo de proceso para satisfacer las necesidades de la compañía. Es sumamente importante hoy en día que se fomente optimizar los mecanismos de registro de información para contar con mayor cantidad de datos

confiables y precisos para soportar todo lo anteriormente mencionado.

La conclusión de utilizar un método de aprendizaje semi-supervisado surge debido a las condiciones actuales en la que se encuentra la información de la compañía. Esto sucede a consecuencia de no contar con un proyecto similar de este tipo de estudio, localmente hablando, por lo que no hay antecedentes de sugerencia para tener una calidad satisfactoria para este tipo de modelos; por tal motivo, los datos requieren ser procesados y tratados con diferentes técnicas para el entrenamiento del modelo, una clasificación de los conductores bajo riesgo o no riesgo.

CAPÍTULO 4

EXPERIMENTACIÓN Y ANÁLISIS

Para el desarrollo de este capítulo se trabaja bajo la estructura metodológica expuesta en el capítulo 3. Se explora la información histórica de la compañía para la selección de los datos y/o reportes que cumplan con los criterios y variables del marco de análisis de conducción que previamente se han identificado en el proceso de revisión de literatura. Posteriormente, dicha información pasará por el proceso de transformación y análisis de acuerdo con la minería de datos. Se hará uso de los métodos de aprendizaje no supervisado y supervisado, con las técnicas de agrupación y clasificación, así como la aplicación de herramientas complementarias para el tratamiento de los datos que refuerzan el proceso de entrenamiento. Finalmente, se concluye con la evaluación del modelo para corroborar el buen funcionamiento midiendo con los métricas de *Precision*, *F1-Score*, *Recall* y con el uso de matrices de confusión y la validación con la curva ROC y AUC.

El principal objetivo del capítulo es presentar cómo se lleva a cabo la implementación del método para el análisis con la propuesta de minería de datos y para el desarrollo de un modelo con aprendizaje semi-supervisado. También exponer la importancia de la calidad de los datos y su proceso de tratamiento para el manejo del análisis de la información, aprendizaje y la eficiencia del algoritmo. Por último, se pretende demostrar concretamente la contribución hacia la logística y cadena de suministro que la utilización de los análisis de datos con *machine learning* puede

aportar y ser una herramienta predilecta para la mejora de procesos.

4.1 FASE 1: DEFINICIÓN DE PROBLEMÁTICA

Los objetivos específicos que trazan el camino para el desarrollo del análisis son los siguientes:

1. Definir principales variables de riesgo
2. Clasificar a los conductores entre alto riesgo y bajo riesgo
3. Crear herramienta basada en algoritmos

4.2 FASE 2: RECOPIACIÓN Y PRE-PROCESAMIENTO DE DATOS

4.2.1 RECOPIACIÓN

El primer paso es la recopilación de la información histórica que cumpla con los antecedentes de la literatura. Recordemos que en la Tabla 3.1 se presenta un resumen de las variables que más se han utilizado para el análisis de este tipo de problemáticas y estudios.

En la compañía se identifican principalmente dos sistemas de registro de información que tienen relación con los conductores y las unidades de transporte. El primero es el *Reporte Mensual de Seguridad* en el que se registran manualmente todos los incidentes que suceden mes tras mes. En el segundo, se trata del sistema de telematics para las unidades en dónde se recopilan electrónicamente mediciones

del manejo del transporte y los nombres de los conductores que utilizaron la unidad por viaje. Algunos ejemplos de las mediciones registrados son la distancia recorrida, conteos por eventos como, aceleraciones, frenados bruscos, exceso de velocidad, etcétera. A través del servicio del proveedor del sistema telematics, se genera un reporte llamado *Driver Behaviour Report* en el que se registran estos datos y está disponible para ser descargado con diferentes opciones de rangos de tiempo.

En la Figura 4.1 se visualiza que el trabajo de selección de información de ambos reportes dará como resultado en un conjunto de datos para operar un análisis de conducir. La Tabla 4.2.1 contiene algunos ejemplos como parte de la exploración y filtración de las observaciones. En el caso del Reporte Mensual de Seguridad, hubo una reducción de los 75 tipos de criterios que son llenados con observaciones a tan sólo utilizar 10 debido a que, primero se hizo una filtración por el país, que en este caso sólo se elige a México; en segundo, se prioriza la información que incluyen los incidentes relacionados con el área de distribución; y por último, se descartan los criterios y observaciones que sean descriptivos, sin un estándar o categorizaciones, así como todos aquellos que contienen grandes cantidades de espacios en blanco o sin información. Recordemos que este registro se realiza manualmente por los usuarios del sistema de recopilación de datos y esto ocasiona una falta de precisión en la información. Las variables seleccionadas son las presentadas en la Figura 4.2. Estos datos cumplen o se acercan a los factores utilizados en la literatura del diagrama del marco de conducción (Capítulo 2) para poder hacer la relación entre el usuario y la unidad con el objetivo de obtener el patrón del estilo de conducir de riesgo. En resumen, contamos de manera básica con la información del conductor, el incidente, la fecha y las mediciones obtenidas de la unidad en relación con la manipulación del transporte.

El año que se ha elegido para el desarrollo del estudio es el año 2020. Esto se debe a que el Reporte Mensual de Seguridad tuvo una renovación a partir del año 2019 con el propósito de cambiar el método de registro de los incidentes laborales. Se realiza una exploración general de la información para conocer cifras y se determina



FIGURA 4.1: Registro de datos para el análisis de conducción

	Cantidad de criterios	Tipo de mediciones o datos	Cantidad de criterios seleccionados
Reporte mensual de seguridad	75 criterios	Ejemplo: * Datos del conductor * Fecha de incidente * Tipo de incidente * Descripción * Lugar * País	10 criterios
Driver behaviour report	34 criterios	Ejemplo: * Conteo de frenados bruscos * Distancia recorrida * Duración total del viaje * Exceso de velocidad	16 criterios

TABLA 4.1: Filtración y selección de datos



FIGURA 4.2: Elección de variables para análisis

que existe un registro de un total de 77 accidentes distribuidos a través del año 2020. La cantidad de accidentes por mes se grafican en un diagrama de barras para su visualización y se muestran en la Figura 4.3. (Observación: No hay accidentes registrados en el mes de junio) Así mismo, el reporte registra cuál fue el tipo de impacto del accidente y se grafican las cifras para conocer las frecuencias de los tipos resultando en el siguiente diagrama de la Figura 4.4. El tipo de impacto con mayor frecuencia fue el de accidente por alcance. Por último, como dato general, de los 77 incidentes registrados se observa que en 54 estuvieron involucrados empleados de la compañía y en 23 se trató de personal por contrato. La relación en porcentaje es de un 70 % personal de la compañía y 30 % por *outsourcing*.

En el caso del reporte Driver Behaviour se contabiliza la cantidad de viajes y la cantidad de conductores registrados en el sistema dentro de un periodo de 6 meses (Enero - Junio 2020). La elección de trabajar con un periodo de 6 meses se debe precisamente por la cantidad de datos y la capacidad de memoria de la computadora con la que se realiza el estudio. En la Figura 4.5 se muestra el resultado de las

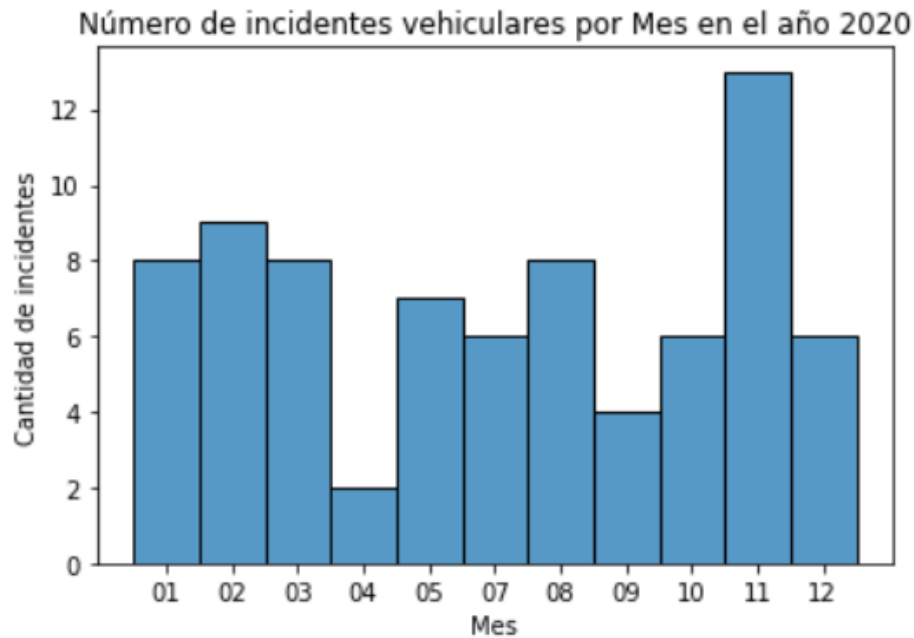


FIGURA 4.3: Cantidad de incidentes registrados por mes del año 2020

frecuencias.

4.2.2 PRE-PROCESAMIENTO DE DATOS

El segundo paso de esta fase es el pre-procesamiento de la información. Tiene como objetivo crear un dataset con la calidad en dónde cada uno de las observaciones se encuentran en un mismo formato, esto quiere decir estandarizadas, para que en el proceso de entrenamiento el algoritmo pueda tener un mejor rendimiento y precisión de aprendizaje. Esta parte del estudio es fundamental para que en la siguiente fase sea posible aplicar las técnicas y herramientas del análisis de datos.

La información disponible del Reporte Mensual no se encontraba en un formato estandarizado lo cual dificulta la filtración y búsqueda de datos. Al momento de realizar la concatenación de la información entre el Reporte Mensual y el Driver Behaviour fue necesario realizar un trabajo manual, esto significa, buscar entre los reportes los nombres de los conductores que coincidieran o en ciertas ocasiones hacer

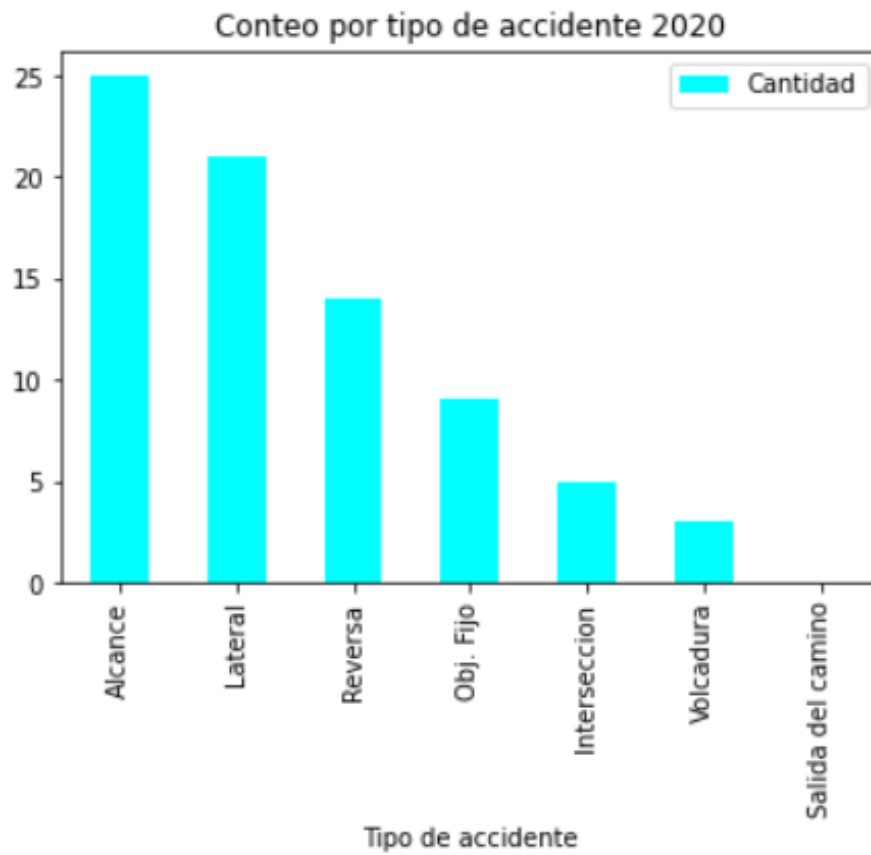


FIGURA 4.4: Registro por tipo de accidente (Año 2020)

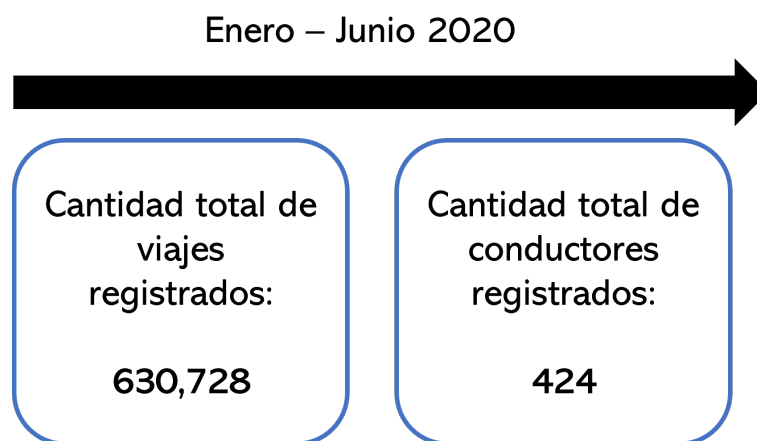


FIGURA 4.5: Contabilización de frecuencias en reporte Driver Behaviour

una suposición de tratarse de la misma persona. Además, para poder agregar la variable objetivo que es una clasificación binaria entre incidente y no incidente, también se tuvo que realizar la búsqueda manual entre las fechas y horarios de los accidentes registrados.

Este proceso manual de búsqueda disminuyó las observaciones de incidente que produce un problema de desbalanceo de clases. De contar con 77 registros de accidente, desafortunadamente la reducción de datos concluyó siendo de tan sólo 7 incidentes. Esto es por causa de dos razones:

1. La elección de ventana de tiempo de 6 meses
2. Conductores registrados en el reporte de seguridad pero no en el Driver Behaviour

La problemática de desbalanceo de clases (visualizado en la Figura 4.6) trae como consecuencia que los parámetros de aquellos viajes identificados como “incidente” sean muy pocos para que el algoritmo los pueda aprender y sea capaz de hacer una correcta clasificación. Posteriormente, se hace el etiquetado parcial de los datos añadiendo una columna al dataset llamada **Incidente**, siendo aquellos accidentes clasificados con un valor de **1**. El resto de los viajes quedarán por el momento con un valor de **0**. La Figura 4.7 es el ejemplo de cómo quedaron etiquetados los identificados y registrados accidentes.

4.2.2.1 CORRELACIÓN CONTRA VARIABLE INCIDENTE

La concatenación entre los reportes nos permite hacer un primer análisis de correlación contra la variable objetivo que es el incidente (1). Con ello se pretende identificar, previo al tratamiento de datos, cuáles son los principales factores de riesgo. Por cuestión del desbalanceo de clases los resultados muestran valores muy pequeños. En la Tabla 4.2 se ordenan las correlaciones del valor más alto hasta el

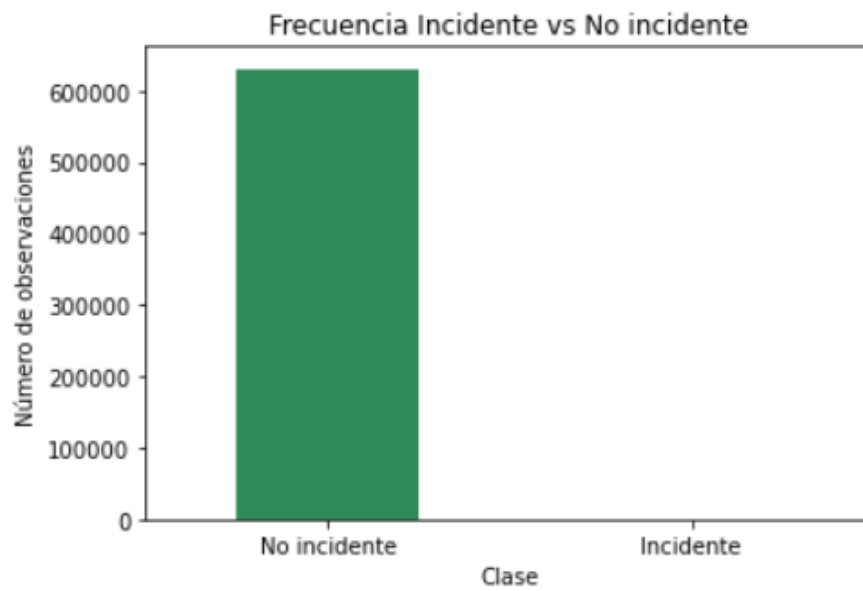


FIGURA 4.6: Gráfica del desbalanceo de clases

	Persona Involucrada	Fecha	Horario	Incidente
0		05-01-20	4:30:00 PM	1
23		10-01-20	1:30:00 PM	1
37		11-01-20	1:30:00 PM	1
42		15-01-20	7:40:00 AM	1
78		27-01-20	7:30:00 AM	1

FIGURA 4.7: Dataset con etiquetado de variable objetivo con valor = 1

Variable	Correlación
Incidente	1.000000
Exceso de velocidad (máximo)	0.003398
Distancia	0.003292
Duración total	0.003167
Frenado brusco (máximo)	0.002030
Aceleración brusca (máximo)	0.000524
Exceso de revoluciones (máximo)	0.000114
Duración aceleración brusca	-0.000045
Duración exceso de velocidad	-0.000070
Ocurrencia exceso de velocidad	-0.000126
Ocurrencia aceleración brusca	-0.000165
Duración exceso de revoluciones	-0.000169
Duración frenado brusco	-0.000218
Duración exceso de inactividad	-0.000257
Ocurrencia exceso revoluciones	-0.000278
Ocurrencia frenado brusco	-0.000394
Ocurrencia exceso de inactividad	-0.000759

TABLA 4.2: Resultado de matriz de correlaciones

de menor relación. Según los resultados, los factores que inciden prioritariamente sobre la variable incidente son la distancia total recorrida en el viaje, la duración total, manejar a exceso de velocidad y los frenados bruscos. Recordemos que este no es el resultado final de factores de riesgo sino se trata de resultados previo a un procesamiento de los datos.

4.2.2.2 SEPARACIÓN ENTRE DATOS DE ENTRENAMIENTO Y PRUEBA

Esta parte del proceso es importante porque los datos de entrenamiento serán utilizados para cumplir con toda la fase de preparación y aprendizaje del modelo. La partición de los datos en este análisis se hacen de acuerdo con la Figura 4.8 en dónde también se muestra el proceso general que llevarán ambos conjuntos de datos.

Con la separación, los conjuntos de datos quedan definidos como se muestran en la Tabla 4.3. Los 7 incidentes quedan divididos entre los datos de entrenamiento con 3 y en los de prueba con 4.

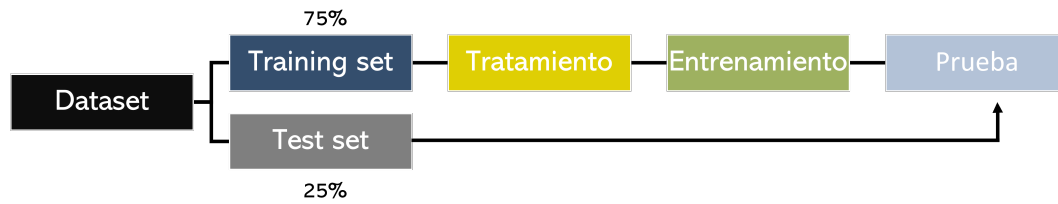


FIGURA 4.8: Separación entre conjunto de datos de entrenamiento y prueba

	Datos de entrenamiento	Datos de prueba
Total de datos	473,043	157,678
Total de incidentes	3	4

TABLA 4.3: Total de datos por tipo de conjunto

Esta fase concluye para introducirnos en la fase 3 en dónde se procede a dar tratamiento a el conjunto de datos de entrenamiento de acuerdo con las características identificadas en esta fase del proceso de desarrollo del modelo.

4.3 FASE 3: SELECCIÓN DE LA O LAS HERRAMIENTAS

El proceso para la selección de la o las herramientas para trabajar los datos y para hacer el tratamiento de la información se muestran en el diagrama de la Figura 3.4. Dicho diagrama sintetiza las técnicas y/o métodos del análisis de datos, además de aportar una ayuda visual para identificar las características particulares del dataset y con ello hacer una implementación para la transformación **únicamente** en los datos de entrenamiento.

Desde una perspectiva general, nos enfrentamos ante un dataset en el cual tenemos las siguientes características que se enlistan a continuación:

- Variedad en los tipos de mediciones

- Datos semi-etiquetados
- Problemática de desbalanceo de clases
- Baja correlación

Para poder afrontar estas particularidades, en esta fase se hace uso de las técnicas propuestas en el diagrama de la Figura 3.5. En conclusión, si no se aplican estas transformaciones en los datos, el modelo no podrá generalizar bien la información y afectará en la precisión del algoritmo.

4.3.1 NORMALIZACIÓN

El dataset cuenta con registros de distintos tipos de mediciones como distancias, tiempos y ocurrencias. Se requiere de una normalización o estandarización de los datos que facilitarán el procesamiento de la información por los algoritmos en el proceso de entrenamiento, tanto de agrupación como de clasificación. La normalización se hace con la aplicación de la función *StandardScaler()* de la librería en Python *sklearn.preprocessing*. Para ver el resultado, en la imagen se muestra el antes con los datos originales y el después de la transformación que sufren.

	Distance	Duration_TotalDurationSecs	OverSpeedingOccurences	OverSpeedingMax	OverSpeedingDuration_TotalDurationSecs
278245	138.6	8357.0	36	89	772.0
564113	3.8	502.0	0	0	0.0
425205	2.7	639.0	0	41	0.0
201304	111.7	6785.0	0	82	0.0
224002	2.2	463.0	0	50	0.0

	Distance	Duration_TotalDurationSecs	OverSpeedingOccurences	OverSpeedingMax	OverSpeedingDuration_TotalDurationSecs
278245	4.510575	2.619218	29.163197	1.753855	12.767404
564113	-0.284599	-0.256163	-0.049919	-1.526892	-0.027464
425205	-0.323729	-0.206013	-0.049919	-0.015536	-0.027464
201304	3.553675	2.043776	-0.049919	1.495819	-0.027464
224002	-0.341515	-0.270439	-0.049919	0.316225	-0.027464

FIGURA 4.9: **Arriba:** Datos originales. **Abajo:** Datos después de normalización.

4.3.2 TÉCNICAS DE AGRUPACIÓN

Las primeras herramientas que se utilizan para el proceso de preparación son aquellas de agrupación de aprendizaje no supervisado. Debido a que no existe un antecedente de un análisis de conducción ni una previa categorización de los conductores, se procede a implementar un algoritmo de clusterización para encontrar las similitudes entre los datos sin especificar ningún tipo de parámetro. Los resultados que se obtienen son parte de un análisis exploratorio inicial dónde queremos que el algoritmo encuentre patrones entre los datos.

Antes de la aplicación del algoritmo, debido a la baja correlación entre los parámetros, ayudaremos al modelo utilizando una herramienta llamada **PCA** (*Principal Component Analysis*). La técnica se encuentra disponible en la librería de Python *sklearn.decomposition*. Cuando se enfrenta ante una grande cantidad de variables correlacionadas, calculando los componentes principales permitirá resumir el conjunto con un número menor de variables que serán representativas que explican de manera colectiva la mayor parte de la variabilidad del conjunto original (James *et al.*, 2021). Conservamos el 80 % de la varianza y con ello determinamos a cuántos componentes o variables disminuir nuestro *dataset*. Según la Figura 4.10, podemos resumir la variabilidad de nuestros datos en un total de 8 componentes. Hacemos la transformación de datos y se procede a la aplicación de la primera técnica de clusterización o agrupamiento.

Una de las características del algoritmo de *K-Means* es que debemos indicarle en cuántos grupos queremos que agrupe la información. En nuestro caso no contamos con un análisis previo para definir un número de clústers por lo que se recurre al *Método del codo* para determinar el número óptimo de agrupaciones que se definen como k . Para el uso del método se requiere de graficar la distorsión y la inercia que hacen referencia a las siguientes descripciones:

- **Distorsión:** es el cálculo de la media de las distancias al cuadrado de los

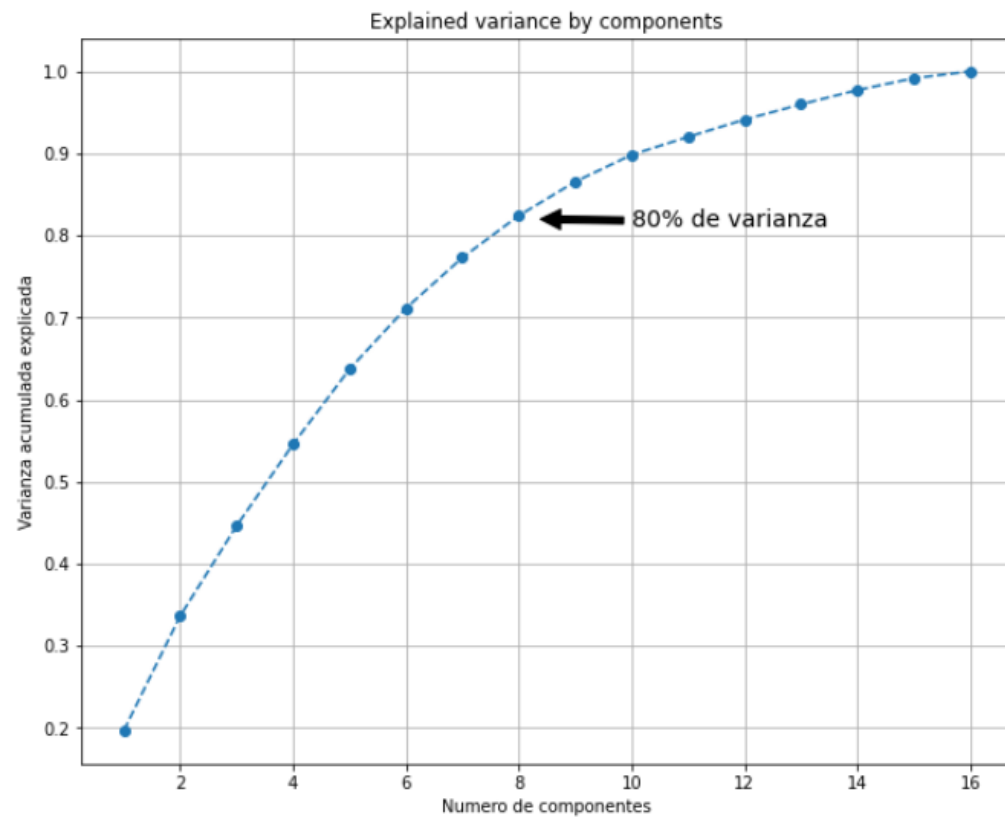


FIGURA 4.10: Gráfica para determinar el número de componentes que expliquen la variabilidad de los datos en un 80 %

centros de los clústeres. La métrica más utilizada es la distancia euclidiana.

- **Inercia:** es la suma de las distancias al cuadrado de las muestras a su centro del clúster más cercano.

Según el método, se determina el número k óptimo de clústers cuando el “codo”, el punto que marca dónde la distorsión e inercia comienzan a decrecer de manera lineal. Al realizar la gráfica (Figura 4.11) fue complicado determinar el número exacto para el valor k debido a la falta de definición del codo; pero finalmente se estipuló que el valor k será igual a 7.

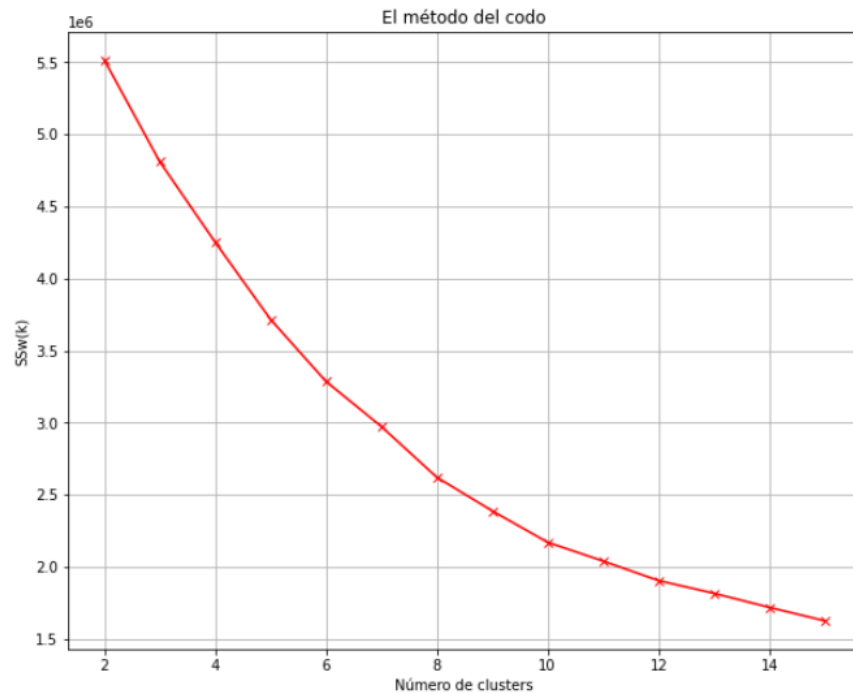


FIGURA 4.11: Gráfica de la distorsión e inercia para el uso del Método del Codo

Le indicamos al algoritmo que queremos $k=7$ clústers y las observaciones se agrupan como se muestra en la Tabla 4.4. El algoritmo de *k-means* se encuentra disponible en la librería de *sklearn.cluster*. Las observaciones etiquetadas como “Incidente” se agruparon en los clúster 1 y 3, con lo que se puede hacer una suposición dónde los viajes que se encuentren cercanos a estos clúster pueden ser de riesgo. Dos incidentes se agrupan en el grupo 3 y uno en el grupo 1 (Figura 4.12).

Clúster	Observaciones
0	162,900
1	32,869
2	4
3	237,270
4	38,041
5	1,644
6	318

TABLA 4.4: Resultado de clústers determinado por k=7

HarshBrakingDuration_TotalDurationSecs	ExcessiveldleOccurences	ExcessiveldleDuration_TotalDurationSecs	Clusters KMeans PCA	Incidente
-0.086613	-0.301496	-0.102215	3	1
-0.086613	-0.301496	-0.102215	1	1
-0.086613	-0.301496	-0.102215	3	1

FIGURA 4.12: Incidentes dentro de los datos de entrenamiento

4.3.3 PROBLEMÁTICAS DE DESBALANCE DE CLASES Y SEMI-ETIQUETADO

Después del primer análisis exploratorio de los datos trataremos las problemáticas del desbalanceo de clases y de la parcialidad de etiquetas en la información. Es necesario corregir porque se cuenta con muy pocos parámetros para que los modelos aprendan de las características de cada incidente lo que limita su predictibilidad y precisión.

4.3.3.1 SMOTE Y RANDOMUNDERSAMPLER

Las técnicas de SMOTE y RandomUnderSampler ayudarán con la función de balancear las clases.

- **SMOTE:** algoritmo que genera un aumento de datos creando puntos de datos

	0	1
Datos originales	473,043	3
Datos transformados	59,130	47,304

TABLA 4.5: Resultado de transformación de datos con técnica de SMOTE y RandomUnderSampler

sintéticos basados en los puntos originales. Con esta técnica no se generan duplicados sino puntos que son ligeramente diferentes a los originales.

- **RandomUnderSampler:** involucra elegir aleatoriamente ejemplos de la clase mayoritaria y eliminarlos de los datos de entrenamiento.

Como tenemos una diferencia muy significativa entre la clase mayoritaria (identificados con 0) contra la clase minoritaria (identificados con 1) es que se decide aumentar los datos minoritarios en un 10 % en relación con el total de datos y reducir la clase mayoritaria en un 80 %. La transformación del dataset concluye tal como en la Tabla 4.5. Con este cambio se espera que el modelo pueda generalizar mejor.

Preparación para técnica de LabelPropagation

La técnica de propagación de etiqueta (*Label Propagation*) es para crear un modelo de aprendizaje semi-supervisado. Siendo parte de las técnicas de clusterización, el algoritmo funciona con un aprendizaje transductivo que pretende etiquetar a los que no lo están explotando la información derivada de los datos etiquetados. Básicamente, aquellos puntos que se encuentren más cercanos y conectados al nodo de un punto etiquetado tendrán alta probabilidad de pertenecer a la misma clase.

Para poder utilizar esta técnica primero es necesario eliminar la etiqueta aleatoriamente de la cantidad que se elija (libre elección) en los datos de entrenamiento. Se opta por eliminar el 50 % de las etiquetas y se deben identificar con un -1. Los datos quedan distribuidos como en la Tabla 4.6.

	-1	0	1
Cantidad de datos	53,217	29,619	23,598

TABLA 4.6: Resultado de eliminación de etiquetas en el 50 % de los datos

Hasta este punto el conjunto de datos de entrenamiento ha quedado preparado para la fase de modelación para el aprendizaje semi-supervisado. Han sido implementadas diferentes herramientas para transformar y mejorar la calidad de la información con el fin de que en la siguiente fase los algoritmos puedan aprender los parámetros de cada clase, identificarlos y clasificar las observaciones con mayor precisión.

4.4 FASE 4: MODELACIÓN Y VALIDACIÓN

En esta última fase se hará uso del conjunto de datos de entrenamiento para desarrollar el modelo. Finalmente, se aplicará el modelo a los datos de prueba y se aplicarán las técnicas de validación para analizar los resultados.

4.4.1 MODELACIÓN

Label Propagation

Utilizando los datos de entrenamiento, teniendo el 50 % con etiquetas y el resto identificado con -1, permitimos al algoritmo de *LabelPropagation*, que trabaja bajo un esquema de clusterización para la clasificación, que aprenda de nuestros datos. El algoritmo se encuentra disponible en la librería de *sklearn.semi_supervised*. Le indicamos que funcione bajo el kernel de “knn” para el aprendizaje (Ver Figura 4.13). Se realiza el ajuste de los datos y le indicamos que haga la predicción sobre los datos no etiquetados.

```
Lp = LabelPropagation(kernel = "knn")
Lp.fit( X, nolabeldata)
Predicted_labels = Lp.predict( X[nolabels])
```

FIGURA 4.13: Ejemplo de ajuste y aprendizaje de los datos por el algoritmo de *LabelPropagation*

Actual / Predicción	TP	FP
Ha dado positiva	29,497	14
	FN	TN
Ha dado negativa	0	23,706

TABLA 4.7: Resultados de Matriz de Confusión

Para ver los resultados de la clasificación se hace uso de una **Matriz de Confusión**. Se clasifican las observaciones como se muestra en la Tabla 4.7. El modelo clasificó 14 observaciones como incidente cuando realmente no fueron incidentes.

Estos valores y resultado de clasificación nos indica que tenemos un modelo con muy alta precisión. Como siguiente paso de comprobación del funcionamiento, se utiliza el conjunto de datos de prueba con el modelo para hacer la predicción. Recordemos que se trata del 25 % del total de datos, los cuales no han sido modificados. Esto quiere decir que conservaron sus valores originales. Para observar el resultado de la clasificación, nuevamente utilizamos la Matriz de Confusión y se obtiene la siguiente clasificación en la Tabla 4.8. El modelo no pudo clasificar correctamente los únicos 4 incidentes, siendo clasificados como “no incidente”. Debido a la casi nula cantidad de registros categorizados como incidente es que el modelo tiene una alta precisión pero es preferible no contar con *Falsos Negativos* como en el caso de nuestro modelo.

Actual / Predicción	TP	FP
Ha dado positiva	157,678	0
	FN	TN
Ha dado negativa	4	0

TABLA 4.8: Resultados de Matriz de Confusión en conjunto de datos de prueba

Para intentar mejorar la precisión del modelo, se toman las etiquetas estimadas en el conjunto de datos de entrenamiento y utilizarlas para ajustar el modelo de aprendizaje supervisado.

Random Forest Classifier

Con base en la literatura, conociendo las ventajas y desventajas de cada clasificador (Figura 3.6), este algoritmo es capaz de procesar la información y obtener mejores resultados en el caso de la situación actual de nuestros datos. Otro punto relevante a consideración es que la modelación de un estilo de conducir tenderá a ser un problema no lineal.

El algoritmo se encuentra en la librería *sklearn.ensemble* importando *RandomForestClassifier*. Se realiza el ajuste del modelo con los datos de entrenamiento estimados de la variable objetivo (clases) en el algoritmo de *LabelPropagation* y se aplica para hacer las predicciones directamente sobre el conjunto de datos de prueba (Ver Figura 4.14). Revisamos el comportamiento del modelo para comprobar si hay una mejora con el uso del aprendizaje supervisado y se obtienen las siguientes clasificaciones de la Tabla 4.9. Podemos observar que nuevamente a nuestro modelo no le fue posible identificar y clasificar los 4 incidentes registrados en los datos de prueba, a pesar del intento de mejorar la precisión bajo un aprendizaje supervisado.

```

forest = RandomForestClassifier(n_estimators = "250")
tran_labels = Lp.transduction_
forest.fit(X, tran_labels)
y_classes = forest.predict(X_test)

```

FIGURA 4.14: Ejemplo de ajuste y aprendizaje de los datos por el algoritmo de *RandomForestClassifier*

Actual / Predicción	TP	FP
Ha dado positiva	157,678	0
	FN	TN
Ha dado negativa	4	0

TABLA 4.9: Resultados de Matriz de Confusión en conjunto de datos de prueba con

modelo de Random Forest

4.4.2 VALIDACIÓN Y ANÁLISIS

El último paso de nuestra metodología es validar el funcionamiento del modelo y también presentar finalmente cuáles son los principales factores de riesgo así como la clasificación de los conductores que se encuentran en alto o bajo riesgo. Lo primero que calculamos son los valores de *precision*, *recall* y *f1-score*. Se obtienen los siguientes resultados:

- **Precision:** 0.9993
- **Recall:** 0.9997
- **F1-Score:** 1.000

Estas métricas utilizadas comúnmente en *Machine Learning* tratan de obtener una idea más detallada de lo bien que funciona un clasificador en lugar de sólo limitarse al resultado general de precisión. Analizando el cálculo de las métricas, nos indican que estamos ante un modelo que se puede decir que funciona muy bien, aunque sabemos que no pudo identificar el objetivo de nuestro análisis.

Curva ROC y AUC

La siguiente herramienta para evaluar el funcionamiento de nuestro modelo es la curva ROC y AUC. Primero hacemos una validación cruzada para obtener las probabilidades de clase de cada observación en el conjunto de datos de prueba y con esos valores crear un diagrama dónde grafiquemos la sensibilidad y la especificidad para formar la curva ROC (Figura 4.15). Al analizar el diagrama se observa que la “curva” se encuentra completamente junto a la línea diagonal. Calculamos el área bajo la curva (AUC) y obtenemos un valor de **0.4991**. Este resultado nos indica que desafortunadamente nuestro modelo clasificador no es capaz de identificar correctamente la clase “1” de entre la clase “0”.

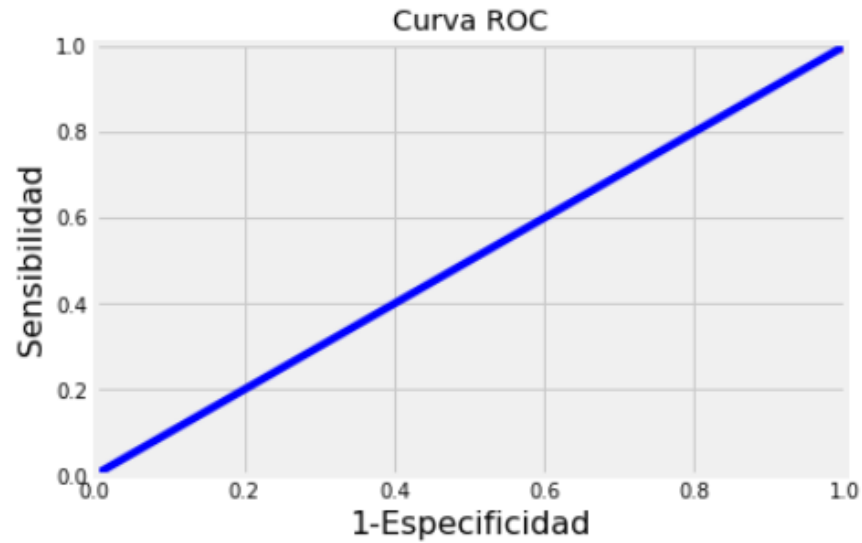


FIGURA 4.15: Gráfica de la Curva ROC

Features	Nivel de importancia
Duración total	0.31
Frenado brusco (máximo)	0.17
Distancia	0.16
Exceso de velocidad	0.14
Exceso de revoluciones	0.14

TABLA 4.10: Factores principales de riesgo por nivel importancia

4.4.3 PRINCIPALES FACTORES DE RIESGO

Para conocer los principales factores de riesgo, extraemos del modelo por nivel de importancia las variables de mayor a menos relevancia para la clasificación. En el diagrama (Figura 4.16) se grafican conforme a su nivel de importancia. Los cinco factores de mayor riesgo para un conductor son enumerados en la Tabla 4.10.

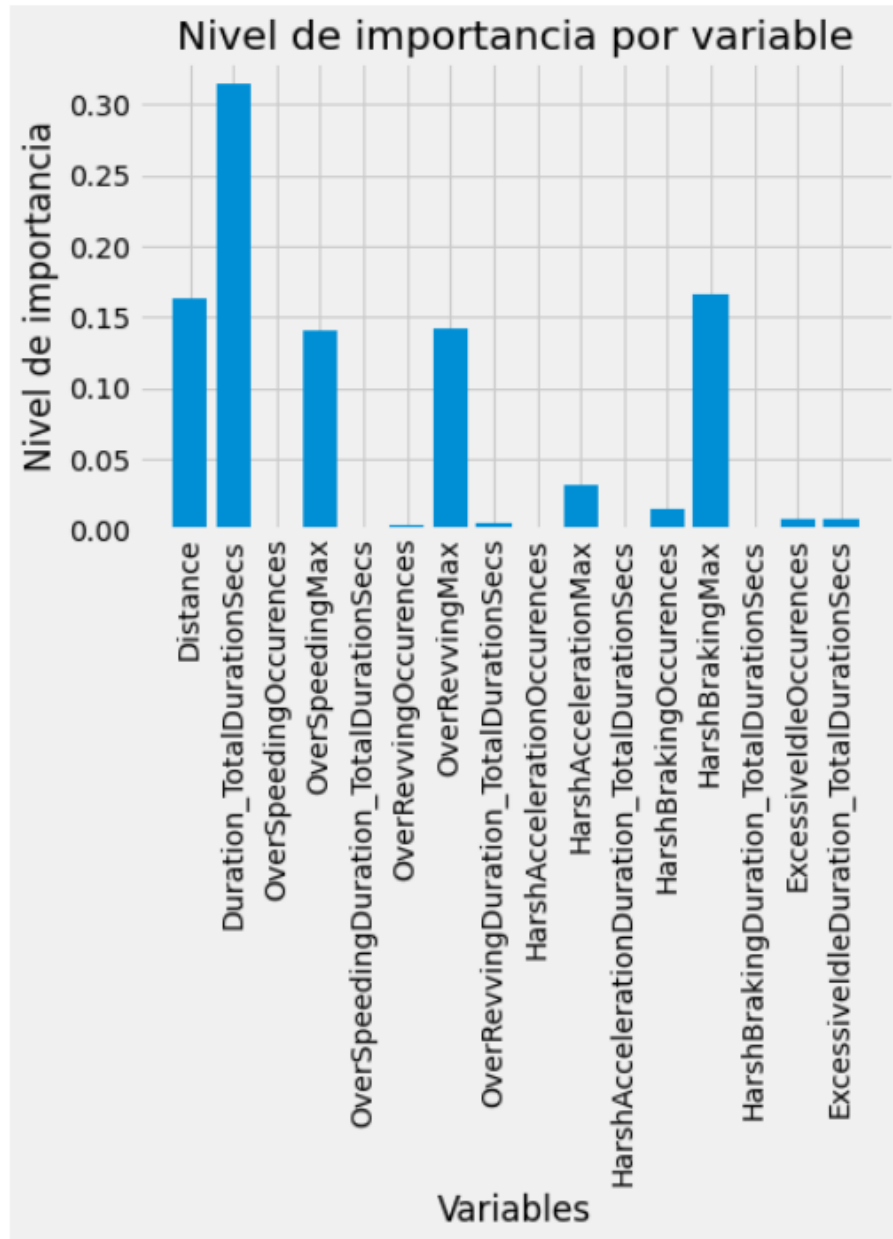


FIGURA 4.16: Diagrama de nivel de importancia por factor

4.4.4 CLASIFICACIÓN DE CONDUCTORES

Por último, se grafican los conductores entre aquellos que fueron clasificados de alto riesgo y los de bajo riesgo. Según el diagrama (Figura 4.17), aquellos viajes de los conductores que se encuentren más cercanos al 100 son aquellos que se clasifican como de alto riesgo. Se han contabilizado 12 puntos (Figura 4.18) o conductores que se han acercado y clasificado como en alto riesgo. Una comparación es realizada entre la lista de los incidentes reales registrados y los nombres de los conductores clasificados en alto riesgo y encontramos que al menos un conductor efectivamente si sufrió un accidente (Figura 4.19). Con este resultado final se puede considerar que el modelo si puede identificar cierto grado de riesgo pero será necesario para mejorar la precisión de clasificación más parámetros u observaciones registradas de accidentes para que la herramienta pueda conocer más características y aprenderlas.



FIGURA 4.17: Diagrama de clasificación de conductores

	Riesgo	Driver_Name
43189	10.0	Miranda Alor Sergio Gerardo
47026	10.0	Miguel Angel Santos Romero
51225	11.2	Delgadillo Jauregui Arturo Gabriel
67151	10.4	Fernandez Pegueros Miguel Angel
75421	12.0	Norberto Valle
112049	10.0	Norberto Valle
114134	10.0	Cesar Monroy Leal
119804	12.8	Edgar Gutierrez Arredondo
134637	11.2	Delgadillo Jauregui Arturo Gabriel
145500	10.4	Edgar Gutierrez Arredondo
146904	10.8	Fernandez Pegueros Miguel Angel
151287	10.0	Juan Miranda Grihop

FIGURA 4.18: Lista de conductores clasificados en alto riesgo

	Unnamed: 0	Driver_Name	StartDate	time_Mx
446108	159	Oscar Ortiz Ortiz	19/05/2020	2020-05-19 09:23:24-05:00
625408	3	Cesar Monroy Leal	20/04/2020	2020-04-20 10:20:30-05:00
447593	108	Demis Torres Luis Martin	20/05/2020	2020-05-20 06:45:27-05:00
122825	201	Aguilar Ortiz Juan Jesus	12/2/2020	2020-12-02 18:19:21-06:00

FIGURA 4.19: Lista de incidentes reales registrados

4.5 CONCLUSIONES DEL CAPÍTULO

Al desarrollar el clasificador y cumplir con los pasos de la metodología y las herramientas de minería de datos, se puede concluir que la implementación de este modelo representa una gran oportunidad para identificar los potenciales riesgos. Obteniendo esta información el personal administrativo de distribución y logística podrá tomar decisiones más asertivas para impactar en la mejora de la seguridad para los conductores profesionales. A continuación se enlistan los hallazgos más relevantes de este proceso de análisis.

- * La recopilación de información llevó un proceso largo debido a que los datos requirieron un trabajo de limpieza y filtración. La falta de precisión en la información limita el análisis de datos
- * La reducción de las observaciones registradas como incidente, de 77 accidentes al año a 7, disminuyó la capacidad de que los algoritmos aprendieran las características de tales eventos
- * A raíz de la reducción de información se acentuó la problemática del desbalance de clases. Aunque fueron implementadas las herramientas de SMOTE y RandomUnderSampler para aminorar el desbalance, no son suficientes los parámetros que el modelo puede replicar
- * En la fase exploratoria inicial con la implementación del algoritmo de *K-Means*, se puede hacer la suposición que es posible que haya más viajes que tengan un alto riesgo. Cabe la posibilidad que el modelo pueda ser mejorado para identificar más patrones
- * El algoritmo de *Label Propagation* tuvo un buen funcionamiento dando la oportunidad de que, en nuevos registros sin categoría, pueda hacer una buena clasificación

- * El intento por reforzar la capacidad de distinción entre clases con el aporte de un aprendizaje supervisado con el algoritmo de *Random Forest*, no mejoró la precisión del modelo. Esto se entiende que es debido a la falta de más datos o características de los viajes accidentados para poder hacer una mejor predicción
- * A pesar de que las métricas de *Precision*, *Recall* y *F1-Score* obtuvieron muy buenos resultados, con la curva ROC y AUC queda claro que es necesario mejorar la calidad de los datos porque el modelo no tiene la capacidad suficiente para identificar y clasificar las rutas con alto riesgo
- * Aunque el modelo no se encuentra en su mejor nivel, es una herramienta exploratoria bastante útil para comenzar con el proceso de mejora de registro en los datos. Esto beneficiará el análisis de la información y por consiguiente la optimización del modelo
- * La herramienta de aprendizaje semi-supervisado ha resultado ser una buena opción para el inicio de la implementación de las técnicas de *Machine Learning* y con ello tener un mejor entendimiento entre la relación del conductor y el vehículo con base en los datos
- * Analizando el resultado de los principales factores de riesgo, llama la atención que dos variables coinciden con la literatura como factores importantes: la distancia y el tiempo. Ambos por obvias razones están interconectados. Los otros tres más importantes se relacionan con el estilo de conducir: frenado brusco, exceso de velocidad y el exceso de revoluciones

CAPÍTULO 5

CONCLUSIONES

En este capítulo final se engloban los resultados y análisis de todo el proceso del estudio. Se hacen conclusiones generales del proyecto así como observaciones particulares que ayuden para mejorar el modelo desarrollado. También en este capítulo se pretende hacer una serie de recomendaciones, así como resaltar la contribución que esta investigación hace en el área de la logística y cómo aporta para tomar decisiones más asertivas para la seguridad en la distribución. Adicionalmente, se hacen algunas sugerencias para continuar con el proyecto ya que finalmente han sido sentadas las bases.

5.1 CONCLUSIONES GENERALES

Para llevar a cabo esta investigación de tesis, se hizo la propuesta de utilizar las herramientas de *Machine Learning* para, primero hacer un análisis exploratorio de los datos y así crear un modelo de clasificación binaria para los conductores del área de distribución. Con el uso de las técnicas de minería de datos se extrae conocimiento aparentemente oculto entre la información, identificando patrones y definiendo los principales factores que impactan sobre el objetivo. La propuesta concreta fue, desarrollar un modelo de aprendizaje semi-supervisado para clasificar de manera predictiva a los conductores entre alto y bajo riesgo.

Al implementar la metodología fase por fase, se logra identificar los principales factores de riesgo, hacer la clasificación y crear la herramienta que aporta una claridad en cuánto a la necesidad de hacer mejoras en el registro de la información para que sea más confiable. En resumen, se cumple con lo planeado aunque el modelo no tiene por el momento el nivel de precisión que se requiere para una buena clasificación.

Desde el inicio de la investigación fue necesario comprender la relación entre la seguridad, la distribución, los conductores, la logística y cada factor de riesgo estudiado anteriormente. Este tipo de problemática ha tomado relevancia dentro de las compañías porque permanece siendo una situación constante de riesgo para el producto, servicio y vida de los conductores. Afortunadamente esto ha aportado a la investigación una amplia gama de análisis y herramientas para encontrar la mejor manera de disminuir la cantidad de accidentes. En los últimos 15 años, las herramientas más utilizadas han sido las de inteligencia artificial porque son capaces de analizar grandes cantidades de datos y encontrar relaciones entre ellos. Con la variedad de tipos de algoritmos para desarrollar modelos basados en el aprendizaje automático, fue realizado un análisis comparativo para elegir las herramientas adecuadas.

La minería de datos y el aprendizaje automático trabajan de la mano, por lo tanto, se propuso seguir la metodología de ambos para estudiar los datos. Se llevó el proceso en 4 fases. Primera fase se determinan objetivos a seguir. En la segunda fase, fue identificado que entre los dos reportes elegidos para obtener las variables de estudio había inconsistencias lo que terminó por reducir las observaciones de la variable objetivo. Esto limitó la capacidad de aprendizaje del modelo. Para la fase 3 se realizó la implementación del análisis exploratorio con *K-Means*, seguido del uso de las técnicas de SMOTE y RandomUnderSampler para balancear las clases. Por último, en la fase 4 se hace el proceso de entrenamiento y prueba, seguido de la validación del modelo y resultados.

Los hallazgos más relevantes, analizando cada fase son que no existe por el momento una integración para el análisis entre los conductores y su relación con el vehículo. La compañía sigue considerando que deben analizarse de manera aislada y por tal motivo la información se encuentra dispersa. Al utilizar *K-Means* hubo una agrupación de 7 clústers significando los estilos de conducir identificados; algunos con gran cantidad de puntos y otros con pocos. Esto nos da un indicio de que pueden realizarse análisis más profundos. El algoritmo de LabelPropagation tuvo un buen nivel de funcionamiento logrando clasificar correctamente la mayoría de las observaciones. Las clases estimadas de este algoritmo no lograron fortalecer el algoritmo de Random Forest porque no pudo identificar correctamente los 4 incidentes en los datos de prueba.

Las métricas de Recall, Precision y F1-Score tuvieron buenos resultados indicando que el modelo generaliza bien, pero al evaluar con la curva ROC y AUC, obtenemos que desafortunadamente, con un valor de 0.49, no tiene la capacidad suficiente para diferenciar entre las clases. Entendemos que se debe a la falta de más observaciones registradas de accidentes que limitó la opción de tener más parámetros por aprender. Se extrajeron los principales factores de decisión del modelo, siendo los factores más importantes de riesgo y son: duración del viaje 31 %, frenado brusco 17 %, distancia 16 %, exceso de velocidad 14 % y exceso de revoluciones 14 %. La duración y la distancia se encuentran vinculados y confirman la literatura siendo definidos como componentes principales de accidentes y su relación con la fatiga. El resto pertenecen a la relación entre el conductor, su estilo de conducir y el vehículo.

Finalmente, a pesar de no contar con el óptimo nivel de desempeño del modelo, al comparar los resultados de la clasificación de los conductores de alto riesgo y los incidentes reales, un conductor efectivamente sí había sufrido un accidente. Por lo tanto, esta herramienta permite ver la problemática de la seguridad laboral desde otro enfoque y abre la posibilidad de priorizar la importancia de contar con un buen sistema de registro de información y toma de datos para poder desarrollar profundizar en los análisis de datos con la ayuda de la inteligencia artificial.

5.2 RECOMENDACIONES

Con la realización de este proyecto se identificaron algunos puntos relevantes para poder llevar a cabo una mejora en la captación de datos. Estas recomendaciones ayudarán para la aplicación de esta herramienta de análisis.

- * El registro de los accidentes de tráfico aún requiere de una mayor precisión para obtener resultados más confiables además de que quepa la posibilidad de ser utilizado en otros estudios
- * En el caso del reporte Driver Behaviour, es importante que los datos de los conductores sigan un mismo formato. Además, que sea imprescindible que todos ingresen al sistema al momento de conducir la ruta
- * Asegurarse que los nombres de los conductores sigan el mismo formato en todos los reportes. Evitar acentos.
- * Registrar cada incidente y cambiar las observaciones descriptivas por categorías o valores estandarizados
- * Crear un nuevo reporte dónde se estandarice la información en el mismo orden tal cómo se llevó a cabo en este estudio. Esto con el propósito de que los datos puedan ser ingresados directamente en el modelo y puedan hacerse las clasificaciones
- * Por el momento no existe una claridad de las rutas o los viajes, ya sea el caso, para el adecuado registro del comportamiento de conducir. Se interpreta que el sistema telematics registra desde que se enciende la unidad o ya sea por viaje, pero se desconocen las verdaderas rutas planeadas
- * Al conocer los principales factores de riesgo, se recomienda trabajar con un equipo de expertos tanto de seguridad como del área de distribución y planeación para proponer cambios en las rutas

- * La clasificación de los conductores puede ser una herramienta de retroalimentación para los conductores. Esto ayudará para que conozcan el riesgo al que se enfrentan si continúan manejando de la misma manera. También, puede ser un puente de comunicación para conocer el punto de vista del conductor y su sentir al conducir
- * Se confirma que la distancia y la duración del viaje impactan negativamente sobre el conductor, lo que permite una oportunidad para analizar la planeación de las rutas

5.3 CONTRIBUCIONES Y TRABAJO FUTURO

La presente tesis contribuye a las empresas y investigadores algunos puntos relevantes que pueden ser utilizados en las áreas de logística y transporte:

1. Se confirma que la distancia y la duración del viaje impactan negativamente sobre el conductor, lo que permite una oportunidad para analizar la planeación de las rutas
2. La clasificación de los conductores puede ser una herramienta de retroalimentación para los conductores. Esto ayudará para tengan conocimiento y noción del riesgo al que se enfrentan día a día si continúan manejando sin restricciones. También, puede ser un puente de comunicación para conocer el punto de vista del conductor y cuál es su sentir al conducir
3. Se recomienda utilizar la mayor cantidad posible de variables relacionadas tanto de la unidad como de los conductores para llevar a cabo análisis más profundos donde las herramientas de machine learning puedan encontrar más relaciones entre variables y patrones

4. Los pasos del modelo demuestran cómo es que una compañía puede iniciar y comparar qué tipo de datos requieren y tienen para poder hacer uso de las herramientas de análisis de datos
5. El modelo conformado por un análisis semisupervisado proporciona la oportunidad de que haya una implementación más sencilla para los análisis que comienzan a enfocarse hacia el análisis de datos
6. Se aporta una herramienta predictiva completamente adaptable ante cualquier diferencia de variables, siempre y cuando la calidad de los datos sea confiable. La herramienta es útil tanto para identificar riesgos así como para identificar a aquellas personas que se encuentran en ese perfil que puede ser perjudicial para su propio bienestar y de los demás
7. Finalmente, este tipo de análisis lleva al personal de logística a conocer más a fondo las operaciones de las compañías con lo que pueden tomar decisiones más informadas y detalladas

5.3.1 TRABAJO FUTURO

El análisis abre el camino para poder construir nuevas líneas de estudio. Como parte de un interés para la disminución de costos operativos en el transporte, es posible también, con base en el estilo de conducir, determinar cuál es aquel que es mejor para tener un óptimo consumo de combustible y cuál es aquel que hace mayor gasto. También, es posible que, a manera que haya una mayor y mejor captación de información, el modelo será cada vez mejor y aprenderá más de los datos con lo que hará mejores predicciones y resultados.

BIBLIOGRAFÍA

- ABKOWITZ, M. (2002), «Transportation Risk Management: a new paradigm», *Annual Meeting of the Transportation Research Board*, págs. 1–16.
- AGGARWAL, C. (2015), *Data Mining: The Textbook*, primera edición, Springer Cham.
- AKSJONOV, A., P. NEDOMA, V. VODOVOZOV, E. PETLENKOV y M. HERRMANN (2018), «A novel driver performance model based on machine learning», *IFAC PapersOnLine*, págs. 267–272.
- BADRA, R. D. (2020), «Métricas de clasificación», https://www.themachinelearners.com/metricas-de-clasificacion/#Curva_ROC.
- BHAVSAR, P., I. SAFRO, N. BOUAYNAYA, R. POLIKAR y D. DERA (2017), *Data analytics for intelligent transportation systems*, Elsevier Inc., USA.
- BHOLOWALIA, P. y A. KUMAR (2014), «EBK-Means: A clustering technique based on Elbow Method and K-Means in WSN», *International Journal of Computer Applications*, **105**, págs. 17–24.
- BORJA-ROBALINO, R., A.-M. GETINO y J. RODELLAR (2020), «Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning», *Revista Ibérica de Sistemas y Tecnología de Información*, págs. 184–196.
- BROWNLEE, J. (2020), «8 Tactics to combat imbalanced classes in your Machine Learning dataset», <https://machinelearningmastery.com/>

- tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.
- CAINTRA (2017), «Estadísticas de Accidentes en México», *Informe técnico*, CAINTRA, Monterrey, Nuevo León.
- CHEN, C. y Y. XIE (2019), «Machine Learning for Recognizing Driving Patterns of Drivers of Large Commercial Trucks», *SAGE Journals*, **2517**, págs. 18–27.
- CHOI, T.-M., C.-H. CHIU y H.-K. CHAN (2016), «Risk management of logistics systems», *Transportation Research Part E: Logistics and Transportation review*, **90**(2), págs. 1–6.
- CONSTANTINESCU, Z., C. MARINOIU y M. VLADOIU (2010), «Driving style analysis using data mining techniques», *International Journal of Computers, Communications and Control*, **5**, págs. 654–663.
- CORSO, C. L. (2007), «Aplicación de algoritmos de clasificación supervisada usando Weka», *Informe técnico*, Universidad Tecnológica Nacional, Facultad Regional Córdoba, Córdoba, Argentina.
- DATTA, P. (2017), «Supply network resilience: A systematic literature review and future research», *International Journal of Logistics Management*, **28**, págs. 1387–1424.
- DUNCAN, R. B. (1972), *Characteristics of Organizational Environments and Perceived Environmental Uncertainty*, tomo 17, tercera edición, Sage Publications, Inc.
- ELASSAD, Z. A., H. MOUSANNIF, H. A. MOATASSIME y A. KARKOUCH (2020), «The application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review», *Engineering Applications of Artificial Intelligence*, **87**, págs. 1–27.
- FAYYAD, U. y P. STOLORZ (1997), «Data mining and KDD: Promise and challenges», *Future Generation Computer Systems*, **13**, págs. 99–115.

- GARCÍA, L. A. M. (2014), *Logística del transporte y distribución de carga*, Ecoe Ediciones, Bogotá.
- GARG, R. (2018), «7 Types of Classification Algorithms», <https://analyticsindiamag.com/7-types-classification-algorithms/>.
- GARRO, J. A. C. (2019), *Clasificación de aceptación de campañas para una entidad financiera usando random forest con datos balanceados y datos no balanceados*, Tesis de Maestría, Universidad Ricardo Palma.
- GWAK, J., M. SHINO y A. HIRAO (2018), «Early detection of driver drowsiness utilizing Machine Learning based on physiological signals, behavioral measures and driving performance», *21st International Conference on Intelligent Transportation System*.
- HIGGS, B. y M. ABBAS (2015), «Segmentation and clustering of car-following behavior: recognition of driving patterns», *IEEE Transactions on Intelligent Transportation Systems*, págs. 81–90.
- HOWARD, M., A. DESAI, R. GRUNSTEIN, C. HUKING, J. ARMSTRONG, D. JOFFE, P. SWANN, D. CAMPBELL y R. PIERCE (2004), «Sleepiness, Sleep-disordered breathing, and accident risk factors in commercial vehicle drivers», *Accident Analysis and Prevention*, **48**, págs. 101–110.
- INEGI (2018), «Medios de transporte», <https://cuentame.inegi.org.mx/economia/>.
- IRANITALAB, A. y A. KHATTAK (2017), «Comparison of four statistical and machine learning methods for crash severity prediction», *Accident analysis and prevention*, **108**, págs. 27–36.
- ISO31000 (2018), «ISO 31000 Administración/Gestión de Riesgo - Lineamientos», *Informe técnico*, Norma Internacional.
- JAHANGIRI, M., K. ALI, S. SLAMIZAD, M. OLYAEI, S. MOOSAVI y F. AMIRI (2013), «Occupational Risk Factors in Iranian Professional Drivers and their Impacts on Traffic Accidents», *International Journal of Occupational Hygiene*, **5**, págs. 184–190.

- JAISWAL, S. (2019), «Classification algorithm in Machine Learning», <https://www.javatpoint.com/classification-algorithm-in-machine-learning>.
- JAMES, G., D. WITTEN, T. HASTIE y R. TIBSHIRANI (2021), *An Introduction to Statistical Learning: with Applications in R*, segunda edición, Springer.
- JUNIOR, J. F., E. CARVALHO, B. FERREIRA, C. DE SOUZA, YOSHIHIKO, A. PENTLAND y G. PESSIN (2017), «Driver behavior profiling: An investigation with different smartphone sensors and machine learning», *Plos One*, págs. 1–16.
- KAPLAN, S., M. A. GUVENSAN, A. G. YAVUZ y Y. KARALURT (2014), «Driver behavior analysis for safe driving: A survey», *IEEE Transactions on Intelligent Transportation Systems*, págs. 1–16.
- KUMAR, V. (2021), «What is K-means algorithm and how it works», <https://towardsmachinelearning.org/k-means/>.
- LI, F., G. LI, M. XUE, C. YU, F. XIA y N. YANG (2014), *Normalized label propagation for imbalanced scenario classification*, tomo 277, Springer, Berlin, Heidelberg.
- MARTINEZ, C. M., M. HEUCKE, F.-Y. WANG, B. GAO y D. CAO (2018), «Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey», *IEEE Transactions on Intelligent Transportation Systems*, **19**, págs. 666–676.
- MARUCHECK, A., N. GREIS, C. MENA y L. CAI (2011), «Product safety and security in the global supply chain: Issues, challenges and research opportunities», *Journal of Operations Management*, **29**(7–8), págs. 707–720.
- MEIRING, G. A. M. y H. C. MYBURGH (2015), «A review of intelligent driving style analysis systems and related artificial intelligence algorithms», *Sensors*, **15**, págs. 30 653–30 682.
- MOOSAVI, S., M. HOSSEIN, R. T. SRINIVASAN PARTHASARATHY y R. RAMNATH (2019), «Accident risk prediction based on heterogeneous sparse data: New data set and insights», Recurso libre, disponible en: <https://arxiv.org/pdf/1909.09638.pdf>.

- NAQA, I. E. y M. MURPHY (2015), *What is Machine Learning?*, Springer International Publishing Switzerland, Switzerland.
- NARASIMHAN, R. y S. TALLURI (2009), «Perspectives on risk management in supply chains», *Journal of Operations Management*, **27**(2), págs. 114–118.
- NOVAKOVIC, J. D., A. VELJOVIC, S. S. ILIC, Z. PAPIC y M. TOMOVIC (2017), «Evaluation of Classification Models in Machine Learning», *Theory and Applications of Mathematics and Computer Science*, **7**, págs. 39–46.
- ORGANIZATION, W. H. (2018), «Global Health Observatory Data», [urlhttps://www.who.int/publications/i/item/9789241565684](https://www.who.int/publications/i/item/9789241565684).
- PARHIZKAR, T., S. HOGENBOOM, J. E. VINNEM y I. B. UTNE (2020), «Data driven approach to risk management and decision support for dynamic positioning systems», *Reliability Engineering and System Safety*, **201**, págs. 1–15.
- PEDAMKAR, P. (2021), «Classification Algorithms», <https://www.educba.com/classification-algorithms/>.
- RIQUELME, J., R. RUIZ y K. GILBERT (2006), «Minería de Datos: Conceptos y tendencias», *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, págs. 11–18.
- SANZ, L. D. B. (2017), «Choferes del autotransporte de carga en México: investigaciones sobre condiciones laborales y la cadena de suministro», *Revista Transporte y Territorio*, **17**, págs. 251–266.
- SANZ, L. D. B., E. C. G. PEÑA, F. L. VILCHIS y L. R. LONA (2017), «Estudio de los efectos de las condiciones laborales de los conductores de autotransporte en la cadena de suministro en México», *Dirección y Organización*, **71**, págs. 87–98.
- SCHOONJANS, F. (2021), «ROC curve analysis», <https://www.medcalc.org/manual/roc-curves.php>.

- SCT (2016), «Acciones para mejorar la seguridad vial del transporte de carga», *Informe técnico*, Secretaría de Comunicaciones y Transportes, Ciudad de Mexico, Mexico.
- SIORDIA, O., I. M. DE DIEGO, C. CONDE, G. REYES y E. CABELLO (2010), «Driving risk classification based on experts evaluation», *IEEE Intelligent Vehicles Symposium*, págs. 1098–1103.
- STERN, H. S., D. BLOWER, M. L. COHEN, C. A. CZEISLER, D. F. DINGES, J. B. GREENHOUSE, F. GUO, R. J. HANOWSKI, N. P. HARTENBAUM, G. P. KRUEGER, M. M. MALLIS, R. F. PAIN, M. RIZZO, E. SINHA, D. S. SMALL, E. A. STUART y D. H. WEGMAN (2018), «Data and methods for studying commercial motor vehicle driver fatigue, highway safety and long-term driver health», *Accident Analysis and Prevention*, **126**, págs. 37–42.
- USECHE, S., V. GOMEZ y B. CENDALES (2017), «Measuring fatigue and its association with job stress, health and traffic accidents in professional drivers: The case of BRT operators», *EC Neurology*, págs. 103–118.
- VILKO, J. P. P. y J. M. HALLIKAS (2012), «Risk assessment in multimodal supply chains», *International Journal of Production Economics*, **140**, págs. 586–595.
- WANG, C., L. LIN, X. CHENGCHENG y L. WEITAO (2019), «Predicting future driving risk of crash-involved drivers based on a systematic machine learning framework», *International Journal of Environmental Research and Public Health*.
- WANG, W., J. XI, A. CHONG y L. LI (2017), «Driving style classification using a semisupervised Support Vector Machine», *Transactions on human-machine systems*, págs. 650–660.
- WU, P.-J. y P. CHAPIYAPHAN (2020), «Diagnosis of delivery vulnerability in a logistics system for logistics risk management», *The International Journal Of Logistics Management*, **31**(1), págs. 43–58.

- YING, X. (2019), «An overview of overfitting and its solutions», *Journal of Physics: Conference Series*, **1168**, págs. 1–6.
- YU TSENG, Y., W. L. YUE y M. A. P. TAYLOR (2005), «The role of transportation in Logistics Chain», *Proceedings of the Eastern Asia Society for Transportation Studies*, **5**(11), págs. 1657–1672.
- ZAMBRANO, B. B., J. V. BOWEN y B. V. PINARGOTE (2020), «Evaluación de factor de riesgo en choferes de transporte de carga pesada aplicando el Método LEST», *Revista de Investigación Formativa: Innovación y Aplicaciones Técnico-Tecnológicas*, **2**, págs. 38–46.
- ZHANG, Z., Q. HE, J. GAO y M. NI (2017), *A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data*, Tesis de Maestría, The State University of New York, Buffalo, New York.
- ZHONGGUO, Y., L. HONGQI, S. ALI y A. YILE (2017), «Choosing classification algorithms and its optimum parameters based on Data Set characteristics», *Journal of Computers (Taiwan)*, **28**, págs. 26–38.
- ÖZ, B., T. ÖZKAN y T. LAJUNEN (2013), «An investigation of professional drivers: Organizational safety climate, driver behaviours and performance», *Transportation Research Part F: Traffic Psychology and Behaviour*, **16**, págs. 81–91.

RESUMEN AUTOBIOGRÁFICO

Arely Ismerai Hernández González

Candidato para obtener el grado de
Maestría en Logística y Cadena de Suministro

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica

Tesis:

AUMENTO DE LA SEGURIDAD DE DISTRIBUCIÓN A TRAVÉS DE LA
DEFINICIÓN DE VARIABLES DE RIESGO PRESENTES EN EL
TRANSPORTE DE CARGA PESADA

Aquí va tu historia. Recuerda que debe incluir: lugar y fecha de nacimiento, nombre de los padres, escuelas y universidades en las que se graduó después de la preparatoria, títulos o grados obtenidos (no incluir los estudios que se están concluyendo), experiencia profesional y organizaciones profesionales a las que pertenece (no incluir lista de publicaciones).