



UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE POSGRADO

MAESTRÍA EN CIENCIAS DE DATOS

Proyecto de investigación previo la
obtención del Grado Académico de
Magíster en Ciencia de Datos.

TEMA:

MODELO PREDICTIVO DE RIESGOS LABORALES EN EL
GOBIERNO AUTÓNOMO DESCENTRALIZADO DE PICHINCHA

AUTORA:

ING. JOCELYNE NATASHA GIRALDO MUÑOZ

DIRECTOR:

ING. EDUARDO AMABLE SAMANIEGO MENA, MSc.

QUEVEDO - ECUADOR

2023-2024

CERTIFICACIÓN:

Ing. Eduardo Samaniego Mena, MSc. docente tutor del Proyecto de investigación previo a la obtención del Título Académico de Magíster en Ciencia de Datos.

CERTIFICA:

Que la **ING. JOCELYNE NATASHA GIRALDO MUÑOZ**, ha cumplido con la elaboración del Proyecto de investigación titulado, “**MODELO PREDICTIVO DE RIESGOS LABORALES EN EL GOBIERNO AUTÓNOMO DESCENTRALIZADO DE PICHINCHA**”, el mismo que se encuentra apto para su presentación y sustentación respectiva.

Quevedo, 25 de mayo del 2024



Firmado electrónicamente por:
**EDUARDO
AMABLE
SAMANIEGO
MENA**

Ing. Eduardo Amable Samaniego Mena, MSc

DIRECTOR

AUTORÍA

Yo, Jocelyne Natasha Giraldo Muñoz con C.I. 1724191059 declaro que el proyecto de investigación con el tema: **MODELO PREDICTIVO DE RIESGOS LABORALES EN EL GOBIERNO AUTÓNOMO DESCENTRALIZADO DE PICHINCHA**, se ha desarrollado de manera íntegra, respetando derechos intelectuales de las personas que han desarrollado conceptos mediante las citas en las cuales indican la autoría, y cuyos datos se detallan de manera más completa en la bibliografía. En virtud de esta declaración, me responsabilizo del contenido, autenticidad y alcance del proyecto de investigación.

Ing. Jocelyne Natasha Giraldo Muñoz

AUTORA

DEDICATORIA

Agradezco a mi familia por estar siempre ahí durante este periodo de aprendizaje y evolución. Sus palabras de ánimo, paciencia y comprensión han sido esenciales para mi crecimiento profesional.

Jocelyne Giraldo Muñoz

AGRADECIMIENTO

Quiero expresar mi más sincero agradecimiento a la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) por brindarme la valiosa oportunidad de recibir esta beca. Gracias a este respaldo, he tenido el honor de cursar la Maestría en Ciencia de Datos en la Universidad Técnica Estatal de Quevedo (UTEQ), una experiencia que ha enriquecido tanto mi vida académica como profesional de manera significativa.

Un reconocimiento especial también va dirigido hacia la UTEQ, por permitirme formar parte de su programa de Maestría en Ciencia de Datos. Su compromiso con la excelencia académica ha sido crucial en mi crecimiento personal y profesional. Asimismo, deseo expresar mi profunda gratitud al Ing. Eduardo Samaniego, MSc., quien ha sido mi guía en este proyecto investigativo. Su dedicación, orientación y conocimientos han sido esenciales para el éxito alcanzado en este proyecto.

De igual forma, quiero agradecer a mis padres, abuelos y hermano por su apoyo constante, confianza y fe puesta en mí. Sin su estímulo y respaldo inquebrantable, este logro no habría sido posible.

Por último, quiero extender mi agradecimiento a Kevin Silva por ser una fuente continua de motivación durante todo mi trayecto académico.

A cada uno de ustedes les expreso mi más sincero reconocimiento.

Sinceramente,

Jocelyne Giraldo Muñoz

PRÓLOGO

Es un honor presentar el prefacio de este proyecto basado en los datos del Gobierno Autónomo Descentralizado de Pichincha donde se utilizó metodologías avanzadas de ciencia de datos con el fin de comprender los factores que contribuyen a los accidentes laborales.

En un mundo donde la seguridad y la salud ocupacional son aspectos fundamentales, el uso de técnicas de ciencia de datos brinda una perspectiva única para abordar los desafíos existentes. A través del análisis de datos históricos y la implementación de modelos predictivos, este proyecto busca identificar tendencias de accidentes a futuros. Es relevante destacar que las predicciones generadas por el modelo resultan tras un análisis exhaustivo de datos previos. No obstante, es importante reconocer que cualquier predicción futura conlleva cierto grado de incertidumbre y debe ser interpretada con cautela.

A lo largo del proceso investigativo, he observado desafíos significativos y comprendido la importancia que tienen tanto la calidad de los datos como la elección acertada de modelos, así como una interpretación precisa resultados.

Gratitud a la Universidad Técnica Estatal de Quevedo y a todos los docentes implicados en el transcurso de este año y medio de aprendizaje de la ing. Jocelyne, les deseo expresar mi más sincero agradecimiento por su entrega y aportación. Este proyecto lo ha desarrollado por los conocimientos impartido por Uds. durante toda la maestría en Ciencias de Datos.

Ing. Kevin Silva Castro

RESUMEN

La seguridad laboral se rige como una preocupación en el Gobierno Autónomo Descentralizado de Pichincha por los riesgos laborales debido al impacto negativo, por la falta de cultura de seguridad y salud en el trabajo. La Ciencia de Datos puede ser útil para abordar este problema, ya que brinda a los investigadores herramientas avanzadas y técnicas analíticas para estudiar conjuntos de datos pequeños como grandes. A pesar de esta preocupación por los accidentes, hasta ahora se han realizado pocos análisis para identificar tendencias o patrones específicos. Por lo tanto, este proyecto se enfoca en analizar una base de datos que contiene información sobre accidentes ocurridos en el GAD de Pichincha durante el período 2015-2023 con el objetivo principal de analizar los accidentes.

Para lograr esto, se emplearon algoritmos de aprendizaje automático y técnicas de Ciencia de Datos para identificar patrones y tendencias en los accidentes laborales. Se llevó a cabo una categorización detallada de los datos con el fin de comprender mejor el comportamiento variable. Además, se desarrolló un modelo predictivo utilizando un enfoque basado en regresión lineal que muestra una disminución progresiva en la cantidad total de accidentes registrados. Asimismo, tras realizar un análisis predictivo para el año 2023, se observa una alta concordancia entre los resultados predichos y reales, lo cual respalda la precisión del modelo utilizado.

Palabras clave: Accidentes laborales, Ciencia de Datos, Tendencia, Minería de datos.

ABSTRACT

Occupational safety is governed as a concern in the Decentralized Autonomous Government of Pichincha due to the occupational risks due to the negative impact, due to the lack of safety and health culture at work. Data Science can be helpful in addressing this problem, as it provides researchers with advanced tools and analytical techniques to study both small and large data sets. Despite this concern about accidents, few analyses have been conducted so far to identify specific trends or patterns. Therefore, this project focuses on analyzing a database containing information on accidents that occurred in the Pichincha GAD during the period 2015-2023 with the main objective of analyzing accidents.

To achieve this, machine learning algorithms and data science techniques were employed to identify patterns and trends in workplace accidents. A detailed categorization of the data was carried out in order to better understand the variable behavior. In addition, a predictive model was developed using a linear regression-based approach that shows a progressive decrease in the total number of recorded accidents. Likewise, after performing a predictive analysis for the year 2023, a high concordance between the predicted and actual results is observed, which supports the accuracy of the model used.

Keywords: Workplace Accidents, Data Science, Trends, Data Mining.

ÍNDICE

TEMA:	i
CERTIFICACIÓN:	ii
AUTORÍA	iii
DEDICATORIA	iv
AGRADECIMIENTO	v
PRÓLOGO.....	vi
RESUMEN	vii
ABSTRACT.....	viii
ÍNDICE.....	ix
INTRODUCCIÓN	xiii
CAPÍTULO I	xv
MARCO CONTEXTUAL DE LA INVESTIGACIÓN	xv
1.1. UBICACIÓN Y CONTEXTUALIZACIÓN DE LA PROBLEMÁTICA	2
1.2. SITUACIÓN ACTUAL DE LA PROBLEMÁTICA	3
1.3. PROBLEMA DE INVESTIGACIÓN.....	4
1.4. DELIMITACIÓN DEL PROBLEMA	5
1.5. OBJETIVOS	5
1.5.1. Objetivo General.....	5
1.5.2. Objetivo Específicos.....	5
1.6. JUSTIFICACIÓN	5
CAPÍTULO II.....	4
MARCO TEÓRICO	4
2.1. FUNDAMENTACIÓN CONCEPTUAL.....	8
2.1.1. Salud Ocupacional.....	8
2.1.2. Importancia de la Salud Ocupacional	8
2.1.3. Riesgo Laboral.....	8
2.1.4. Accidentes de trabajo.....	9
2.1.5. Metodología KDD	9
2.1.6. Análisis predictivo	10
2.1.7. Análisis Exploratorio de Datos (EDA)	11
2.1.8. Normalización de datos	11
2.1.9. Discretización de datos.....	11
2.1.10. Técnicas de minería de datos	11
2.1.11. Modelos de aprendizaje	12
2.1.12. Técnicas de Agrupamiento (Clustering).....	12
2.1.13. Aprendizaje supervisado.....	13
2.1.14. Regresión lineal	13
2.1.15. Bosque Aleatorio (Random forest)	13

2.1.16.	Árbol Aleatorio (Random tree).....	13
2.1.17.	Análisis predictivo	14
2.2.	FUNDAMENTACIÓN TEÓRICA.....	14
2.3.	FUNDAMENTACIÓN LEGAL	17
CAPÍTULO III.....		24
METODOLOGÍA DE LA INVESTIGACIÓN		24
3.1.	TIPO DE INVESTIGACIÓN	21
3.1.1.	Investigación Descriptiva	21
3.1.2.	Investigación correlativa.....	21
3.2.	MÉTODOS UTILIZADOS EN LA INVESTIGACIÓN	22
3.2.1.	Método deductivo	22
3.3.	CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN.....	22
3.3.1.	Técnicas de investigación	22
3.3.2.	Instrumentos de la investigación	23
3.4.	ELABORACIÓN DEL MARCO TEÓRICO	25
3.5.	RECOLECCIÓN DE LA INFORMACIÓN	26
3.6.	Procesamiento y análisis	28
3.6.1.	Etapas 1. Preprocesamiento y transformación.....	29
3.6.2.	Etapas 2. Selección de datos	30
3.6.3.	Etapas 3. Minería de datos	30
3.6.4.	Etapas 4. Interpretación y análisis.....	30
CAPÍTULO IV		25
RESULTADOS		25
4.1.	Examinar el conjunto de datos en GAD de Pichincha para su limpieza, transformación y análisis.....	33
4.2.	Factores relacionados con accidentes laborales que resultan en días de reposo prolongados mediante análisis de datos.	34
4.3.	Tendencia semestral de accidentes laborales en el GAD de Pichincha.....	51
4.3.1.	Tendencia de accidentes laborales del 2024 en el GAD de Pichincha.....	55
4.3.2.	Tendencia de accidentes laborales del 2025 en el GAD de Pichincha.....	57
4.3.3.	Tendencia de accidentes laborales del 2026 en el GAD de Pichincha.....	59
CAPÍTULO V.....		56
CONCLUSIONES Y RECOMENDACIONES		56
5.1.	CONCLUSIONES	64
BIBLIOGRAFÍA		67
ANEXOS		72

ÍNDICE DE TABLAS

Tabla 1.	27
Tabla 2.	41
Tabla 3.	41
Tabla 4.	53
Tabla 5.	54

ÍNDICE DE ILUSTRACIONES

Ilustración 1.	29
Ilustración 2.	33
Ilustración 3.	34
Ilustración 4.	35
Ilustración 5.	36
Ilustración 6.	36
Ilustración 7.	37
Ilustración 8.	37
Ilustración 9.	38
Ilustración 10.	38
Ilustración 11.	39
Ilustración 12.	40
Ilustración 13.	40
Ilustración 14.	47
Ilustración 15.	48
Ilustración 16.	48
Ilustración 17.	49
Ilustración 18.	49
Ilustración 19.	50
Ilustración 20.	50
Ilustración 21.	51
Ilustración 22.	52
Ilustración 23.	52
Ilustración 24.	53
Ilustración 25.	55

Ilustración 26.	55
Ilustración 27.	56
Ilustración 28.	56
Ilustración 29.	57
Ilustración 30.	58
Ilustración 31.	58
Ilustración 32.	59
Ilustración 33.	60
Ilustración 34.	60

INTRODUCCIÓN

Los accidentes laborales se han establecido como uno de los principales problemas de salud pública a nivel mundial, debido a que al menos un trabajador pierde la vida cada tres minutos a causa de un accidente laboral según Soto y Melara (2018). Lo que genera altos costos tanto para las empresas como para el sistema de salud pública, por lo cual se considera como una problemática.

Los costos económicos y el impacto en el producto interno bruto (PIB) a nivel mundial debido a los accidentes laborales son notables. Es importante destacar que las indemnizaciones resultantes de estos incidentes según el artículo de Toro, Vega y Romero (2021) representan una proporción significativa llegando al 4% del PIB global, ocasionando que se pierda una gran cantidad de dinero y provocando consecuencias negativas en el desarrollo y crecimiento de la nación o estado.

En Ecuador, se registraron 22.405 accidentes laborales y 1.044 enfermedades profesionales durante el año 2017 según Toro, Vega y Romero (2021), lo cual repercute de forma negativa dentro de las empresas, ya que significa ausentismo laboral, discapacidad permanente e inclusive la muerte.

Así mismo, estos accidentes laborales y las enfermedades profesionales se originan dentro de las empresas y organizaciones que no han generado políticas y acciones orientadas a la prevención de riesgos laborales. Además, Candonga y Valencia (2021) indican que, en las empresas, no se generan evaluaciones constantes acerca de la presencia de este tipo de riesgos. Debido a esto, los trabajadores no pueden identificar los riesgos a los que están expuestos y se producen accidentes o enfermedades que ocasionan grandes pérdidas económicas dentro de la empresa y del país.

En base a lo mencionado anteriormente, este estudio tiene como objetivo desarrollar un modelo que pueda predecir los riesgos laborales en el contexto ecuatoriano. Específicamente, se enfoca en el cantón Pichincha utilizando datos recolectados del GAD de Pichincha desde el 2015 hasta el 2023. El modelo se creará identificando patrones y tendencias significativas, lo que servirá como base sólida para su implementación. Esta contribución es crucial ya que muestra la cantidad de accidentes por semestre, permitiendo tomar decisiones para reducir riesgos y asegurar un ambiente laboral seguro. Cabe destacar que no se consideraron factores climáticos ni pandemias en este análisis.

En virtud de lo anterior, este documento se conforma de los siguientes capítulos:

Sección I-Marco contextual de la investigación, integrado por: ubicación y contextualización de la problemática, problema general y derivados, delimitación del problema, objetivos y justificación.

Sección II-Marco teórico de la investigación, describe: fundamentos conceptuales y teóricos relativos al estudio, y los sustentos legales.

Sección III-Metodología de la investigación, dentro de este apartado se especifica: tipo y métodos de investigación, recopilación de información, instrumento de investigación, procesamiento y análisis de la información.

Sección IV-Resultados donde, se encuentra plasmados los diferentes hallazgos de la investigación cumpliendo los objetivos propuestos.

Sección V- Conclusiones y recomendaciones para determinar los diferentes enfoques que se encontraron en la investigación.

CAPÍTULO I

MARCO CONTEXTUAL DE LA INVESTIGACIÓN

1.1. UBICACIÓN Y CONTEXTUALIZACIÓN DE LA PROBLEMÁTICA

En la actualidad los índices de accidentes laborales dentro de las empresas se han incrementado, ocasionando efectos negativos en el desarrollo y crecimiento de las organizaciones. Esto se debe a que no se ha instaurado una cultura de seguridad y salud laboral. De acuerdo con datos de la Organización Internacional de Trabajo y la investigación de Espinoza y Ramos (2021) se ha identificado 2.78 millones de muertes de trabajadores en condiciones laborales, las cuales representan una pérdida económica anual del 4%. Además, Espinoza y Ramos (2021) afirman que a diario se presentan 7500 muertes de las cuales 1000 se relacionan con accidentes laborales y cada año se generan 374 millones de accidentes de trabajo y enfermedades profesionales sin riesgo de mortalidad.

Dentro de la región latinoamericana, se ha identificado Llagua y Acosta (2020) que se presenta un número de 11,1 accidentes con riesgo de muerte relacionados con actividades laborales por cada 100.000 trabajadores. Además, afirman Llagua y Acosta (2020) que, de las muertes establecidas, se conoce que el 10.7 se generan dentro del ámbito de la agricultura y 6.9 millones dentro del área de servicios. También se determina que el área de construcción incide en gran manera en la presencia de accidentes laborales.

Esta situación se observa en varias provincias del Ecuador, una de ellas es la provincia de Pichincha con una gran cantidad de accidentes laborales. En el estudio que realizaron Toro, Vega y Romero (2021) en el 2017 se identificaron 5751 accidentes de trabajo y 630 enfermedades profesionales. Esto demuestra que dentro de la provincia no existe una cultura de salud y seguridad laboral dentro de las empresas que se oriente a prevenir los riesgos laborales.

En cambio, Del Valle (2022) indica que en la provincia de Manabí se han presentado 4124 reportes de accidentes en el lugar de trabajo, de los cuales 1151 se relacionan con la ida y vuelta del lugar de trabajo, 443 con desplazamientos de la jornada laboral, 408 en otro sitio y 150 en comisión de servicios. Además, Del Valle (2022) ha identificado que el mes de enero es el que reporta mayor número de accidentes laborales; mientras que el mes de junio tiene un reporte bajo.

La investigación se llevará a cabo en el GAD de Pichincha, se caracteriza por presentar un gran número de empresas manufactureras y de servicios públicos, lo cual representa un entorno laboral dinámico. Sin embargo, dentro de estos sectores se ha identificado la presencia de accidentes laborales que perjudican el bienestar de los trabajadores y la empresa.

Actualmente el GAD de Pichincha ha presentado una cantidad anual de 16-23 accidentes laborales. Esto provoca efectos negativos en la vida de los trabajadores y en la productividad y economía de la empresa. Dentro de la organización se ha identificado un índice medio de accidentes de trabajo que han sufrido los funcionarios públicos durante los últimos años, lo cual interfiere en la ejecución de actividades y trámites. Adicionalmente, la gestión preventiva de riesgos laborales ha presentado varios retos y dificultades, debido a que no se ha generado un enfoque de prevención en áreas críticas y escasez de recursos para abordar los problemas.

1.2. SITUACIÓN ACTUAL DE LA PROBLEMÁTICA

Dentro del GAD de Pichincha, se presenta un índice alto de accidentes laborales durante los últimos años, los cuales derivan en una gran cantidad de lesiones, traumatismos y enfermedades en diversas partes del cuerpo. Las lesiones físicas son las de mayor presencia entre los trabajadores. Sin embargo, también se ha observado un aumento en

los síntomas de dolor, inflamación y limitaciones de movilidad entre los colaboradores de la institución.

La organización requiere de un sistema adecuado que contribuya a la identificación de áreas y factores de riesgo relacionados con los accidentes laborales. Con base en esto, se requiere realizar un análisis profundo de las prácticas y políticas de salud y seguridad ocupacional que se manejan dentro de la empresa. El incremento de accidentes laborales tiene una consecuencia negativa en la productividad y economía del Gobierno Autónomo Descentralizado de Pichincha, por lo que resulta indispensable analizar todos los casos de accidentes laborales correctamente para comprender las medidas que se deben aplicar para evitar su aumento.

Abordar la problemática actual que presenta la organización requiere de la aplicación de una estrategia integral que incluye el análisis de modelos predictivos que se orientan a la cultura de salud y seguridad laboral dentro de la empresa. Así como la aplicación de políticas direccionadas a la prevención y reducción de los accidentes laborales.

1.3. PROBLEMA DE INVESTIGACIÓN

¿Cómo se podría detectar de manera oportuna los accidentes laborales presentados en el Gobierno Autónomo Descentralizado de Pichincha?

Problemas derivados

¿Cuáles son las causas más comunes que influyen en los accidentes laborales en el Gobierno Autónomo Descentralizado de Pichincha?

¿Qué factores específicos están relacionados con accidentes laborales que resultan en días de reposo prolongados?

¿Es posible pronosticar, anticipar la tendencia semestral de accidentes laborales en el Gobierno Autónomo Descentralizado de Pichincha?

1.4. DELIMITACIÓN DEL PROBLEMA

Campo: Tecnología de la información y la comunicación (TIC)

Área: Tecnología de la información y la comunicación (TIC)

Línea: Ciencias de datos

Lugar: Ecuador, Los Ríos, Quevedo

Tiempo: Desde enero del 2024 hasta mayo del 2024

1.5. OBJETIVOS

1.5.1. Objetivo General

Analizar los accidentes laborales de los empleados del Gobierno Autónomo Descentralizado de Pichincha mediante modelos predictivos para identificar tendencias que contribuyan a una mejor gestión.

1.5.2. Objetivo Específicos

- Examinar el conjunto de datos utilizando minería de datos para la limpieza, transformación y el análisis de los accidentes laborales en el Gobierno Autónomo Descentralizado de Pichincha
- Determinar los factores relacionados con accidentes laborales que resultan en días de reposo prolongados mediante análisis de datos.
- Aplicar técnicas de ciencias de datos para pronosticar una tendencia semestral de accidentes laborales en el Gobierno Autónomo Descentralizado de Pichincha.

1.6. JUSTIFICACIÓN

La ejecución de la presente investigación se basa en las estadísticas y cifras de accidentes laborales que se han encontrado dentro del país y la provincia, así como dentro del GAD de Pichincha. Los accidentes laborales interfieren de forma negativa en la calidad de vida y bienestar de las personas, por lo que es necesario implementar sistemas o programas

orientados a la prevención de riesgos laborales para que se incremente la cultura de salud y seguridad ocupacional dentro de las empresas. De este modo, logrará reducir la cantidad de accidentes y mejorar su rentabilidad y productividad.

Esta investigación contribuirá a establecer un sistema de prevención de riesgos laborales dentro del Gobierno Autónomo Descentralizado de Pichincha. El modelo que se pretende analizar los factores de riesgo existentes dentro de la institución y las estrategias que se deben adoptar para prevenirlos.

La importancia práctica radica en que el modelo predictivo posibilitará que los directivos de la organización tomen decisiones adecuadas con relación a la prevención de riesgos laborales y que destinen los recursos hacia las áreas críticas que presentan un mayor riesgo. De esta forma, se logrará la reducción de costos en materia de cobertura de accidentes y enfermedades; y se optimizará el área de salud y seguridad ocupacional, generando un impacto positivo en la productividad y rentabilidad de la organización.

CAPÍTULO II

MARCO TEÓRICO

2.1.FUNDAMENTACIÓN CONCEPTUAL

2.1.1. Salud Ocupacional

La salud ocupacional hace referencia a una disciplina que se orienta a la consolidación de la perspectiva de que la salud de los trabajadores es un problema evidente y técnico, es decir pueden provocar efectos negativos en el bienestar de los trabajadores. Su objetivo principal es prevenir las lesiones, accidentes y enfermedades ocasionadas debido a las condiciones laborales que se presentan (Arango, Correa, & Luna, 2020).

2.1.2. Importancia de la Salud Ocupacional

La salud ocupacional es un elemento de importancia dentro de las empresas, debido a que se encarga de regular, controlar y prevenir los riesgos laborales existentes en el entorno, con la finalidad de evitar accidentes, lesiones o enfermedades en los trabajadores. Además, contribuye a garantizar el cumplimiento de las políticas preventivas para incrementar la seguridad y salud laboral (Ortega, Rodríguez, & Hernández, 2017).

2.1.3. Riesgo Laboral

El riesgo laboral hace referencia al grupo de elementos físicos, psicológicos, químicos, ambientales, culturales y sociales que intervienen sobre la persona durante su área de trabajo. La interrelación y consecuencias que ocasionan dichos factores y que generalmente desencadenan en una enfermedad (Ampuero, Pozo, & Delgado, 2018).

Por otra parte, de acuerdo con Llagua y Acosta (2020), los riesgos laborales conforman a combinación de la gravedad de los efectos y la probabilidad de que se presenten resultados indeseables relacionados con las condiciones laborales inadecuadas que ponen en peligro el bienestar de las personas; así como comportamientos de trabajo incorrectos que afectan la productividad y calidad de vida de los trabajadores.

2.1.4. Accidentes de trabajo

Los accidentes de trabajo hacen referencia a toda situación o acontecimiento imprevisto que interfiere en la salud de los trabajadores debido al desempeño de sus funciones. Se presentan como consecuencia de las actividades que ejecutan los trabajadores para cumplir con lo establecido por parte de la empresa u organización. (Ampuero, Pozo, & Delgado, 2018).

El accidente de trabajo es identificado como un suceso o situación que se presenta dentro del ámbito laboral en relación a las actividades y acciones que ejecutan los trabajadores, pueden aparecer en forma de lesiones o enfermedades que provocan incapacidad para laborar (Toro, Vega, & Romero, 2021).

2.1.5. Metodología KDD

La metodología KDD (*knowledge discovery in databases*, o en español descubrimiento de conocimiento en bases de datos) facilita la organización de procesos para descubrir patrones relevantes, originales, prácticos y comprensibles dentro de un conjunto de datos, lo que posibilita la toma de decisiones informadas. Este enfoque se divide en cuatro etapas: recopilación e integración de datos, limpieza, selección y transformación, minería de datos, e interpretación y evaluación.

2.1.5.1. Recopilación e integración de datos

Durante esta etapa se elige la fuente de datos adecuada para el estudio, la cual se integra en un repositorio llamado almacén de datos integrados, que es estable y no cambia con el tiempo. Esto facilita la toma de decisiones.

2.1.5.2.Limpieza, selección y transformación

En esta etapa, se lleva un proceso de depuración de los datos incorrectos identificados en la fase previa. Luego, se definen los datos incompletos y se seleccionan filas y columnas,

conocidas como atributos, que conforman una vista minable que será utilizada en la minería de datos. Finalmente, el conjunto de datos se transforma al formato necesario, siendo este resultado denominado conjunto de datos depurado.

En el proceso de limpieza de datos (en inglés, *data cleansing* o *data scrubbing*) se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en el conjunto de datos (Pham et al., 2019). A nivel de valores de atributos se gestionan los valores ausentes, los erróneos y los inconsistentes.

2.1.5.3. Minería de datos

Durante este proceso se determinan las acciones a realizar, como clasificación, agrupamiento, asociación, entre otras. Para cada tarea específica se selecciona la técnica más apropiada en base a las características del conjunto de datos y los objetivos de la investigación.

Proceso de extraer conocimiento de grandes cantidades de datos. Se ha llamado minería de datos, al análisis de archivo y bitácoras de transacciones con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones que sean útiles para la toma de decisiones (Madrid Orrego et al., 2023).

2.1.5.4. Interpretación y evaluación

Después de haber obtenido la recopilación de patrones, se procede a analizar, interpretar y evaluar los resultados obtenidos en la etapa de minería de datos. Además, puede ser necesario retroceder a etapas anteriores, ya que podría ser necesario elegir una técnica diferente en la fase de minería para mejorar la interpretación de los datos.

2.1.6. Análisis predictivo

El análisis predictivo consiste en estudiar los datos históricos y actuales para hacer predicciones sobre el futuro. Usa una combinación de técnicas avanzadas de aprendizaje

automático, matemáticas y estadísticas para analizar los datos, identificar y extrapolar las tendencias ocultas (Centeno Martín-Romero & Giménez Abad, 2020).

2.1.7. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos constituye un reto para científicos y analistas de diferentes disciplinas. Muchos de los métodos y herramientas existentes en la actualidad no permiten realizar un análisis visual efectivo de múltiples variables a la vez (SANCHEZ et al., 2018).

2.1.8. Normalización de datos

La normalización de datos consiste en modificar los datos para lograr que estén en una escala de valores equivalentes que simplifique la comparación entre ellos. La normalización es útil para varios métodos de minería de datos, que tienden a quedar sesgados por la influencia de los atributos con valores más altos, distorsionando de esta forma el resultado del modelo (Pham et al., 2019).

2.1.9. Discretización de datos

Es el proceso mediante el cual los valores de una variable continua se incluyen en contenedores, intervalos o grupos, con el objetivo de que haya un número limitado de estados posibles. Los contenedores se tratarán entonces como si fueran categorías (Pham et al., 2019).

2.1.10. Técnicas de minería de datos

La minería de datos ayuda a lidiar con grandes volúmenes de datos generados diariamente en el comercio minorista, como resultado de las ventas generadas, puede darnos una descripción del comportamiento de los productos o familias de productos en ciertos periodos (ya sean cambios estacionales, eventos únicos, situaciones) lo llamamos minería descriptiva, este tipo de minería tiene varios tipos de algoritmos que se pueden adaptar a

diferentes situaciones. Por otro lado, intenta darnos también predicciones sobre situaciones que pueden aparecer en el futuro, a raíz del inicio de las temporadas, tomándolo como punto de partida, utilizamos datos de la misma temporada de años anteriores para poder predecir lo que sucederá en algunos días del año situaciones que pueden repetirse, esto se llama minería predictiva (González, 2021).

2.1.11. Modelos de aprendizaje

El aprendizaje automático es un campo de la tecnología de la información que ha ganado mucha popularidad en los últimos años. Aunque existen muchas definiciones, una de las más comunes es: el campo de investigación que permite a las computadoras aprender sin programación especial. Esto cubre una amplia gama de técnicas que se dividen en dos tipos principales: aprendizaje supervisado (donde se tiene una variable objetivo y el programa de computadora es "entrenado" por el conjunto de datos para aplicar el resultado al nuevo conjunto de datos) y aprendizaje no supervisado (donde no existe una variable objetivo y el programa informático debe encontrar patrones y relaciones en el conjunto de datos) (Sandoval, 2017)

2.1.12. Técnicas de Agrupamiento (*Clustering*)

El agrupamiento es una tarea de aprendizaje inductiva que, a diferencia de las tareas de clasificación y regresión, no dispone de una etiqueta de clase a predecir. Puede considerarse como un problema de clasificación, pero donde no existen un conjunto de clases predefinidas, y estas se “descubren” de forma autónoma por el método o algoritmo de agrupamiento, basándose en patrones de similitud identificados en los datos.

Un proceso de agrupación puede proporcionar información útil sobre los patrones de similitud presentes en los datos como, por ejemplo, segmentación de clientes o creación de catálogos de documentos (Pham et al., 2019).

2.1.13. Aprendizaje supervisado

El aprendizaje supervisado es una técnica en el campo del aprendizaje automático en la que un algoritmo se entrena utilizando un conjunto de datos que incluye ejemplos de entrada y sus salidas correspondientes deseadas. Durante el proceso de entrenamiento, el algoritmo aprende a relacionar las entradas con las salidas al observar ejemplos etiquetados. El objetivo principal del aprendizaje supervisado es desarrollar un modelo capaz de generalizar correctamente a nuevas entradas no vistas, es decir, poder predecir la salida correcta para datos desconocidos basándose en su experiencia previa con datos etiquetados. (Sabharwal & Selman, 2020).

2.1.14. Regresión lineal

Es un método de modelado estadístico que se utiliza para explicar cómo una variable de respuesta continua se relaciona con una o más variables predictores. Puede ayudar en el análisis de datos experimentales, financieros y biológicos o en la comprensión y predicción del comportamiento de sistemas complejos (Moncho Vasallo, 2015).

2.1.15. Bosque Aleatorio (*Random forest*)

Es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado (Espinosa Zúñiga, 2020).

2.1.16. Árbol Aleatorio (*Random tree*)

Un Árbol Aleatorio es un tipo de clasificador supervisado que se deriva del algoritmo de árbol de decisión, y se destaca por su velocidad y eficiencia. Este método utiliza la técnica de embolsado, en la que se generan conjuntos aleatorios de datos para construir múltiples

árboles de decisión. Luego, el algoritmo realiza una poda en el árbol mediante un proceso llamado poda de error reducido con ajuste posterior. (Alpan & Ilgi, 2020).

2.1.17. Análisis predictivo

En el campo de la minería de datos, el análisis predictivo emplea técnicas estadísticas y de aprendizaje automático para predecir eventos futuros basándose en datos actuales e históricos. Para comprender correctamente estos análisis, es crucial contar con una gran cantidad de datos que permitan identificar correlaciones y patrones de comportamiento entre las variables. Algunas de las técnicas comúnmente utilizadas para el análisis predictivo incluyen la regresión lineal, redes neuronales y árboles de decisión (Centeno Martín-Romero, 2020).

Según Centeno Martín-Romero (2020), se destacan tres modelos relevantes para este tipo de análisis: los Modelos Descriptivos, que se centran en analizar eventos pasados para comprender el presente y planificar el futuro, los Modelos Predictivos, que buscan predecir resultados futuros o la probabilidad de ciertas situaciones; y los Modelos Prescriptivos, que proponen acciones con base en predicciones al combinar datos, reglas empresariales y modelos matemáticos con el fin de ofrecer soluciones optimizadas.

2.2.FUNDAMENTACIÓN TEÓRICA

Antes de adentrarnos en otras investigaciones relacionadas con la predicción, es fundamental destacar la importancia y variedad de enfoques que se emplean en este ámbito. A través de la utilización de diferentes estrategias de modelado predictivo y aprendizaje automático, los investigadores han abordado el reto de prever y evitar sucesos críticos que impactan en el rendimiento y la seguridad de los trabajadores en diversos entornos laborales.

Según Altujar et al. (2016), desarrollaron un modelo predictivo utilizando una técnica de aprendizaje automático llamada algoritmo de árbol de decisión ID3 para anticipar el bajo desempeño académico de los estudiantes. De manera similar, (Hussain, Dahan, Ba-Alwi, & Ribata, 2018), realizaron pronósticos sobre los resultados educativos empleando diversos algoritmos de redes bayesianas y árboles de decisión basados en información socioeconómica y demográfica de estudiantes universitarios.

En el estudio realizado por (García, 2020) la construcción de modelo predictivo para detectar las actividades cotidianas se basó en el uso de técnicas de clasificación con machine learning. El proceso experimental para detectar las Actividades de la Vida Diaria en el conjunto de datos CASAS Kyoto se dividió en varios pasos, a saber: 1) corrección e integración de errores, 2) agrupación de casos, 3) aplicación de técnicas para representar características en un subconjunto de datos, 4) entrenamiento y prueba de modelos para la clasificación, y (5) evaluación de métricas del modelo para determinar qué tecnología híbrida produce los mejores resultados en términos de tasa de éxito. (García, 2020).

Según (Moirón, 2021) su enfoque se centró en desarrollar un modelo predictivo de accidentes laborales mediante la utilización de una variedad de algoritmos. Estos algoritmos emplean hiperplanos para distinguir entre puntos etiquetados con diferentes categorías, colocando los puntos de una categoría a un lado del plano y los puntos de otra categoría al otro lado. Un hiperplano es una extensión del concepto de plano que divide un espacio tridimensional en dos partes de dimensiones arbitrarias. Por ejemplo, en un espacio unidimensional (una línea), un hiperplano sería equivalente a un punto, mientras que en un espacio bidimensional (un plano) correspondería a una línea (Moirón, 2021). Este método se considera estándar para la clasificación ya que aborda el problema de clasificación lineal en el espacio característico utilizando el principio del margen

estructural. Para tener una comprensión más profunda sobre los principios operativos y los desafíos relacionados con la optimización, primero analizamos el SVM lineal y su óptima separación por medio de hiperplanos. (Van Vaerenbergh & Santamaría, 2018).

La presente investigación de (Villada et al., 2023) se centra en examinar distintos criterios de los autores sobre la accidentalidad laboral, tratando definiciones, teorías y clasificaciones. Asimismo, se presentan cifras de accidentalidad, antecedentes y hallazgos de investigaciones relevantes en el campo. Se resaltan diversas estrategias para prevenir la accidentalidad con el fin de garantizar la seguridad en el entorno laboral, incluyendo la utilización de modelos matemáticos. Los autores concluyen que hay acuerdo en las definiciones de accidentalidad laboral, describiéndola como un suceso repentino y multifactorial que resulta en lesiones para los trabajadores, excluyendo lesiones autoinfligidas. Se destaca la utilidad de los modelos matemáticos y la optimización para abordar problemas relacionados con la prevención de riesgos laborales, respaldando decisiones y estrategias organizativas. A partir de investigaciones previas, proponen un modelo predictivo eficaz para prevenir riesgos laborales, especialmente en el sector de la construcción.

El modelo propuesto por (Villalobos, 2020), para crear un modelo predictivo de factores de riesgo laboral mediante el uso de inteligencia artificial consistió en desarrollar una prueba inicial que pudiera detectar posibles situaciones de riesgo en una planta industrial, a través del análisis de imágenes procesadas con tecnología IA. Se estableció una colaboración con la empresa Agrícola y Exportadora AGRICOM S.A. para llevar a cabo un piloto en el área de almacenamiento y despacho de su instalación en Polpaico, dado que era donde se concentraba la mayor cantidad de tareas críticas. Se siguió la metodología propuesta por la norma ISO 45.001 para identificar peligros y evaluar

riesgos como parte central del proceso para determinar los factores de riesgo laboral, como por ejemplo el tránsito de grúas sobre líneas peatonales o pallets mal colocados, entre otros aspectos. El piloto se implementó exitosamente, identificando diversos factores de riesgo mediante algoritmos que analizaban las imágenes capturadas por las cámaras instaladas en las áreas productivas, lo que permitió desarrollar planes de acción personalizados (Villalobos, 2020).

La investigación realizada por (Jesús, 2019). tiene como objetivo crear un modelo predictivo sobre la exposición al riesgo psicosocial en los empleados de una empresa del ámbito educativo. Se empleó la metodología CRISP-DM y se llevó a cabo un análisis de regresión logística binaria utilizando datos de una muestra compuesta por 487 trabajadores, entre docentes y personal administrativo. Se identificó que los factores que predicen el riesgo psicosocial son el turno laboral y el salario del empleado. Asimismo, se observó que ambas variables están relacionadas positivamente con la probabilidad de experimentar riesgo psicosocial, siendo los turnos más irregulares y los salarios más altos los que presentan una mayor exposición al mismo. Por último, se determinó que el modelo predictivo desarrollado tiene una capacidad predictiva del 73.9%. Este modelo fue entregado a la empresa con el fin de que sea utilizado por los departamentos de recursos humanos y seguridad y salud en el trabajo para implementar planes de intervención específicos y actualizar los perfiles de selección dentro de la organización.

2.3.FUNDAMENTACIÓN LEGAL

Dentro de la Constitución de la República del Ecuador (2008), se establece lo siguiente acerca de la salud laboral:

Art. 32.- La salud es un derecho que garantiza el Estado, cuya realización se vincula al ejercicio de otros derechos, entre ellos el derecho al agua, la alimentación, la educación,

la cultura física, el trabajo, la seguridad social, los ambientes sanos y otros que sustentan el buen vivir (Asamblea Nacional Constituyente de Ecuador, 2008).

Art. 33.- El trabajo es un derecho y un deber social, y un derecho económico, fuente de realización personal y base de la economía. El Estado garantizará a las personas trabajadoras el pleno respeto a su dignidad, una vida decorosa, remuneraciones y retribuciones justas y el desempeño de un trabajo saludable y libremente escogido o aceptado (Asamblea Nacional Constituyente de Ecuador, 2008).

Art. 38.- Riesgos provenientes del trabajo. - Los riesgos provenientes del trabajo son de cargo del empleador y cuando, a consecuencia de ellos, el trabajador sufre daño personal, estará en la obligación de indemnizarle de acuerdo con las disposiciones de este Código, siempre que tal beneficio no le sea concedido por el Instituto Ecuatoriano de Seguridad Social (Asamblea Nacional Constituyente de Ecuador, 2008).

La Organización Mundial de la Salud (OMS) y la Organización Internacional del Trabajo (OIT), se definiría como la promoción y mantenimiento del mayor grado de bienestar físico, mental y social de los trabajadores en todas las ocupaciones mediante la prevención de las desviaciones de la salud, control de riesgos y la adaptación del trabajo a la gente y la gente a sus puestos de trabajo, cabe señalar que la salud ocupacional no solo se encarga de velar por las condiciones físicas de los trabajadores, sino también de la parte psicológica, proporcionándoles un apoyo adicional que repercuta en su capacidad laboral a través de un equilibrio físico y emocional (Vielma, 2021).

Según el acuerdo de Cartagena 584, Instrumento Andino de Seguridad y Salud en el Trabajo en el artículo 11.- En todo lugar de trabajo se deberán tomar medidas tendientes a disminuir los riesgos laborales, en directrices sobre sistemas de gestión de la seguridad

y salud en el trabajo y su entorno como responsabilidad social y empresarial (Instrumento Andino de Seguridad y Salud en el Trabajo, 2004).

En el código del trabajo artículo 410.- Obligaciones respecto de la prevención de riesgos.

- Los empleadores están obligados a asegurar a sus trabajadores condiciones de trabajo que no presenten peligro para su salud o su vida (Codigo del Trabajo, 2012).

Mientras la Ley Orgánica de Servicio Público, LOSEP indica que en el artículo 23, literal 1.- Derechos de las servidoras y los servidores públicos. - 1) Desarrollar sus labores en un entorno adecuado y propicio, que garantice su salud, integridad, seguridad, higiene y bienestar (LOSEP, 2016).

En el Reglamento de Seguridad y Salud de los Trabajadores Decreto Ejecutivo 239, artículo 11, numeral 2.- Adoptar las medidas necesarias para la prevención de los riesgos que puedan afectar a la salud y el bienestar de los trabajadores en los lugares de trabajo de su responsabilidad (IESS, 2015).

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN

3.1. TIPO DE INVESTIGACIÓN

La presente investigación será de tipo exploratorio y descriptivo, ya que se ejecutó una revisión profunda del objeto de estudio para obtener una visión panorámica del mismo, así como de su comportamiento dentro del entorno.

3.1.1. Investigación Descriptiva

Se consideró este tipo de investigación para describir los riesgos laborales dentro del GAD de Pichincha, para plantear todas las variables y parámetros necesarios detallando las características y aspectos principales de los accidentes laborales. Además, se basa en la recolección de datos específicos acerca de incidentes pasados, sobre todo de aspectos como la gravedad de lesiones, períodos o tiempo de reposo, circunstancias de los accidentes y eficiencia de las medidas de seguridad que se han implementado. Se utilizó estadísticas descriptivas, como medidas de tendencia central y dispersión, para resumir y presentar de manera clara la información recopilada. También, la descripción se centró en plantear asociaciones entre variables como tipo de trabajo, condiciones del entorno laboral y la frecuencia de accidentes. De esta forma, se obtuvo una visión general acerca de los factores que contribuyen a que se incrementen los riesgos laborales dentro de la institución.

3.1.2. Investigación correlativa

Para encontrar las relaciones entre las variables del conjunto de datos y accidentes en GAD de Pichincha, se revisó si una o más variables influyen en otras. Esto implica analizar las variables relevantes del conjunto de datos y utilizar principios generales para comprender la incidencia de las mismas en los riesgos laborales. Aunque el objetivo principal es generar un modelo predictivo para caracterizar y obtener las tendencias de accidentes.

3.2. MÉTODOS UTILIZADOS EN LA INVESTIGACIÓN

3.2.1. Método deductivo

El enfoque deductivo se basa en principios generales para derivar conclusiones específicas. Para esto se examinan las variables pertinentes del conjunto de datos y se analiza las variables que impactan en la ocurrencia de riesgos laborales. Al aplicar técnicas de Ciencia de Datos, se busca identificar patrones y relaciones entre las variables para entender la naturaleza de los riesgos laborales. A partir de esta comprensión, se construye un modelo predictivo que utiliza los principios generales obtenidos para predecir eventos específicos relacionados con la seguridad laboral.

3.3. CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN

3.3.1. Técnicas de investigación

Dentro de la investigación sobre el modelo predictivo de riesgos laborales en el GAD de Pichincha, se aplicaron técnicas cuantitativas debido a la naturaleza del conjunto de datos disponible. Las técnicas seleccionadas son las siguientes:

- **Análisis de Regresión:** Fue una herramienta fundamental para investigar la relación entre las variables independientes y la variable dependiente de interés. Se analizó cómo pueden influir en las causas en los accidentes laborales. Además, se evaluará el impacto de estas variables en la frecuencia y gravedad de los accidentes.
- **Análisis de Supervivencia:** Esta técnica permitirá investigar qué variables influyen en la duración de los días de reposo y cómo estas variables se relacionan con los accidentes laborales. Se analizarán las características de los accidentes para determinar cómo influyen en el tiempo de recuperación de los trabajadores. Esto permitirá identificar factores de riesgo asociados con períodos de reposo prolongados.

- **Análisis de Aprendizaje Automático:** Se aplicarán diversos algoritmos de aprendizaje automático, como árboles de decisión, para desarrollar un modelo predictivo. Estos modelos utilizarán datos históricos de accidentes laborales, así como información relevante sobre condiciones laborales y factores de riesgo, para predecir la probabilidad cuantos accidentes van a ocurrir de manera semestral. Además, se realizará una validación de los modelos para garantizar su precisión y fiabilidad.
- **Análisis de Agrupamiento:** Se empleará para identificar patrones y grupos dentro de la población laboral del GAD de Pichincha. Se agruparán a los trabajadores en función de características similares, el historial de accidentes laborales, entre otros. Esto permitirá identificar subgrupos de trabajadores con un mayor riesgo de sufrir accidentes y diseñar estrategias de prevención específicas para cada grupo.

3.3.2. Instrumentos de la investigación

Durante el desarrollo de la investigación sobre el modelo predictivo de riesgos laborales en el GAD de Pichincha, se utilizó diversas herramientas y plataformas especializadas para llevar a cabo análisis avanzados de datos. A continuación, se detallan los principales instrumentos empleados:

3.3.2.1. Google Colab

Se utilizó para la limpieza de datos, manejo de valores faltantes y transformación de variables. Además, tiene un entorno desarrollo en la nube que permite ejecutar código Python. El uso de Google Colab garantizó la eficiencia y la escalabilidad en el preprocesamiento de datos, preparando así los datos para su análisis posterior de manera precisa y confiable.

En la limpieza se le cargaron las siguientes librerías para el preprocesamiento y análisis de los datos:

- **Panda:** Para la manipulación y análisis de datos.
- **Numpy:** Para operaciones numéricas y cálculos.
- **Unidecode:** Para estandarizar la codificación de caracteres y facilitar el manejo de texto.
- **Datetime:** Para trabajar con fechas y tiempos.
- **Calendar:** Para obtener información sobre calendarios y fechas.
- **Matplotlib y Seaborn:** Para visualización de datos faltantes.
- **Plotly Express:** Para creación de gráficos interactivos.
- **Google Colab:** Para el acceso y la gestión de archivos en Google Drive.
- **Scikit-learn:** Para creación y evaluación de modelos de datos (selección de características).

También se utilizó para aplicar técnicas de ciencias de datos y llevar una predicción de accidentes semestrales, para esto se emplearon las siguientes bibliotecas:

- **Train_test_split:** Para dividir los conjuntos de prueba.
- **LogisticRegression:** Para la implementación de modelos de regresión logística y lineal.
- **RandomForestRegressor:** Para la implementación de modelos de regresión basados en árboles.
- **DecisionTreeRegressor:** Para la implementación de modelos de regresión basados en árboles.
- **Confusion_matrix:** Para evaluar el rendimiento de los modelos de clasificación.
- **Mean_squared_error:** Para evaluar el rendimiento de los modelos de regresión.
- **StandardScaler:** Para estandarizar cada característica a la media y varianza unitaria.
- **SelectKBest:** Para seleccionar las mejores características en pruebas estadísticas.

3.3.2.2.Weka

Se emplearon técnicas de análisis de datos y estadísticas inferenciales disponibles en Weka para identificar relaciones entre variables. Además, facilitó la exploración de datos y la experimentación con diferentes técnicas como:

- **Evaluadores de atributos:** Se utilizó *CfsSubsetEval* para evaluar la idoneidad del subconjunto, también se ocupó *CorrelationAttributeEval* para determinar la relevación entre atributo y clase para predecir la variable objetiva.
- **Métodos:** Se utilizó *BestFirts* porque mejora el rendimiento del clasificador, en cambio *Ranker* ayuda a identificar las características más relevantes.

3.4.ELABORACIÓN DEL MARCO TEÓRICO

Este marco proporcionó la estructura conceptual esencial para entender, analizar y abordar el problema, guiando el enfoque y estableciendo una base sólida para el análisis y la discusión de los resultados. Llevando a cabo una búsqueda literatura relevante para el tema. La investigación se enfocó en fuentes académicas, incluyendo artículos, libros, tesis.

Además, los conceptos de la investigación definieron en base al análisis. Esto facilitó en la comprensión de los términos, conceptos, teorías y modelos existentes sobre riesgos laborales. Posteriormente se realizó una revisión de la literatura seleccionada, la cual abordaba análisis predictivos similares al tema de investigación de este estudio.

Posteriormente se analizaron las regulaciones, leyes normativas relacionadas con la seguridad laboral, para identificar responsabilidades, obligaciones y derechos asociados con la seguridad laboral en el ámbito gubernamental.

3.5. RECOLECCIÓN DE LA INFORMACIÓN

La obtención de datos para esta investigación se obtuvo a través de la colaboración directa con el GAD de Pichincha, que proporcionó acceso a la información sobre riesgos laborales. El conjunto de datos utilizado se preparó a partir de 8 bases de datos consta de 206 registros divididos de la siguiente manera:

- En el año 2015 existen 31 registro de riesgos laborales.
- En el año 2016 existen 37 registro de riesgos laborales.
- En el año 2017 existen 26 registro de riesgos laborales.
- En el año 2018 existen 35 registro de riesgos laborales.
- En el año 2019 existen 21 registro de riesgos laborales.
- En el año 2020 no hay registro por motivo del Covid-19.
- En el año 2021 existen 21 registro de riesgos laborales.
- En el año 2022 existen 17 registro de riesgos laborales.
- En el año 2023 existen 18 registro de riesgos laborales hasta agosto.

A partir de las bases de datos disponibles, se definieron las variables clave y los parámetros necesarios para la investigación sobre riesgos laborales. Además, se proporcionó los datos en un formato acordado, asegurando la confidencialidad y la integridad de la información. La entrega de datos se realizó de manera segura.

Una vez recibidos los datos, se llevó a cabo una validación inicial para asegurar que la información proporcionada estuviera completa y alineada con los parámetros definidos para la investigación sobre riesgos laborales. A continuación, en la Tabla 1 se detalla las variables que conforman la base de datos. Se describe el nombre proporcionado, el tipo de variables y su significado.

Tabla 1.

Descripción de variables de la base de datos

Atributo	Tipo	Descripción
Mes	Texto	El mes que ocurrió el accidente.
Nº	Texto	Número del registro del accidente.
Día	Texto	Día específico en que ocurrió el accidente.
Fecha de accidente	Fecha	Fecha exacta en que ocurrió el accidente.
Nombres	Texto	Nombre del trabajador afectado.
Edad	Texto	Edad del trabajador afectado.
Descripción del accidente	Texto	Breve descripción de como ocurrió el accidente.
Diagnóstico	Texto	Diagnostico medico relacionado con el accidente ocurrido.
Lugar accidente	Texto	Lugar específico de donde ocurrió el accidente.
Días de reposo	Numérica	Números de días que el colaborador estuvo en reposo debido al accidente.
Días de Cargo	Numérica	Número de días que la empresa cargó al trabajador debido al accidente.
Fecha de reposo	Fecha	Fecha en la que el trabajador comenzó el reposo.
Fecha de consulta	Fecha	Fecha en la que el trabajador consultó a un médico o especialista.
Lugar de trabajo	Texto	Área donde el trabajador estaba asignado al momento del accidente.

Atributo	Tipo	Descripción
Función de accidentado	Texto	Función o cargo que desempeñaba el trabajador al momento del accidente.
Tipo de incapacidad	Texto	Tipo de incapacidad que el trabajador sufrió como resultado del accidente.
Etapas del caso	Texto	Etapas en la que se encuentra el proceso relacionado con el accidente.
Responsabilidad patronal	Texto	Indicación de si la responsabilidad del accidente recae en el empleador.
Indemnización	Texto	Detalles sobre la indemnización otorgada al trabajador debido al accidente.
Reubicación	Texto	Información sobre si el trabajador fue reubicado en otro puesto de trabajo después del accidente.
Observaciones	Texto	Cualquier observación adicional relevante sobre el accidente.
Técnico responsable	Texto	Responsable encargado del registro y seguimiento del caso de accidente.

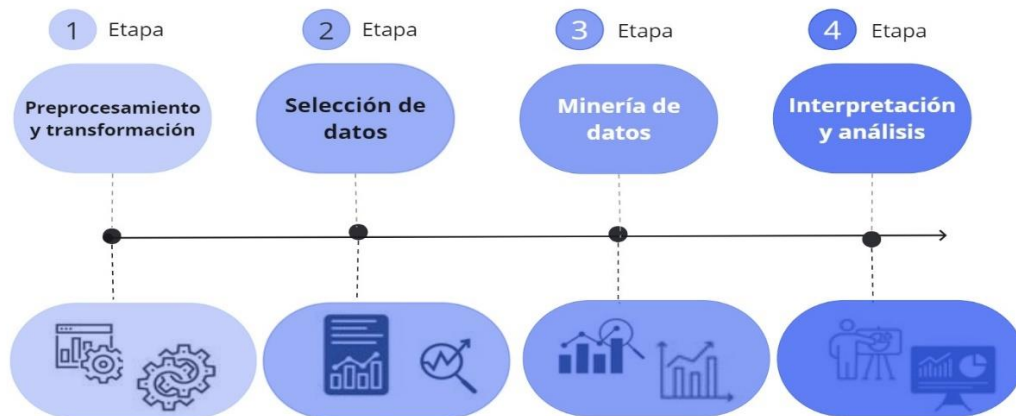
3.6. Procesamiento y análisis

Se implementó el método KDD en diversas etapas, desde la selección meticulosa de datos relevantes hasta la minería de datos utilizando técnicas avanzadas. Se buscó descubrir patrones y tendencias ocultas, asegurando que el análisis esté fundamentado en una metodología sólida.

En esta Ilustración 1, se detalló las etapas del proceso de investigación, centradas en la preparación y selección de los datos relevantes para el análisis de riesgos laborales en el GAD de Pichincha.

Ilustración 1.

Diagrama KDD



Fuente: Autor

3.6.1. Etapa 1. Preprocesamiento y transformación

En la primera fase del proceso de preparación y transformación de datos, se inicia con la selección y recopilación de información relevante para el estudio. Esta etapa implica identificar las fuentes de datos pertinentes y obtener los conjuntos de datos necesarios. Además, se realiza una exploración inicial de los datos para comprender su estructura y calidad. Durante este proceso, es posible encontrar datos incompletos, que requerirán ser eliminados en etapas posteriores. Para evaluar la integridad de los datos, se calcula el porcentaje de valores nulos en cada columna del DataFrame, lo que proporciona información crucial sobre la completitud de los datos y resalta las áreas que podrían necesitar imputación o eliminación de valores faltantes.

3.6.2. Etapa 2. Selección de datos

Una vez terminada la fase de preprocesamiento, se procedió a la selección de las 22 columnas más relevantes para la investigación (Ilustración 3). Esta etapa involucró la identificación y elección de los conjuntos de datos más significativos para el análisis de los riesgos laborales en el GAD de Pichincha. Se tomó unas decisiones informadas sobre qué variables y atributos como edad, función, diagnostico, días de reposo, entre otras, que eran las principales cumplir los objetivos de investigación, asegurando así un enfoque centrado.

3.6.3. Etapa 3. Minería de datos

Culminadas las dos etapas anteriores se procede al preprocesamiento y selección de datos, donde se aplicaron técnicas de análisis avanzado para explorar los datos y descubrir patrones, tendencias y relaciones ocultas relacionadas con los riesgos laborales en el GAD de Pichincha. Se emplearon los métodos de clasificación, regresión, árboles de decisiones, entre otros, con el fin de extraer información mediante que permitió comprender mejor los accidentes laborales, los instrumentos o herramientas utilizadas están mencionadas en la sección 3.3.2.

3.6.4. Etapa 4. Interpretación y análisis

En esta etapa se realizó un proceso analizar y predecir la cantidad de accidentes laborales GAD de Pichincha mediante técnicas de aprendizaje automático. Se comenzó con la preparación de datos, incluyendo la creación de una característica para representar los semestres y la conversión de fechas al formato adecuado. Luego, los datos se dividen en conjuntos de entrenamiento y prueba.

Se evalúan tres modelos de regresión: Regresión Lineal, *Random Forest* y Árbol de Decisión. Cada modelo se entrena con los datos de entrenamiento y se utilizan para predecir el número de accidentes en los semestres del conjunto de prueba. Posteriormente, se calculan métricas de desempeño como el Error Cuadrático Medio y el Coeficiente de Determinación.

Se repite este análisis para calcular que modelo se acopla más a la realidad para realizar la predicción de cuantos accidentes van a existir en los siguientes años, es decir que, durante esta etapa, se analizaron los patrones y relaciones para comprender mejor los accidentes laborales y su impacto en la seguridad y salud laboral de los trabajadores.

CAPÍTULO IV

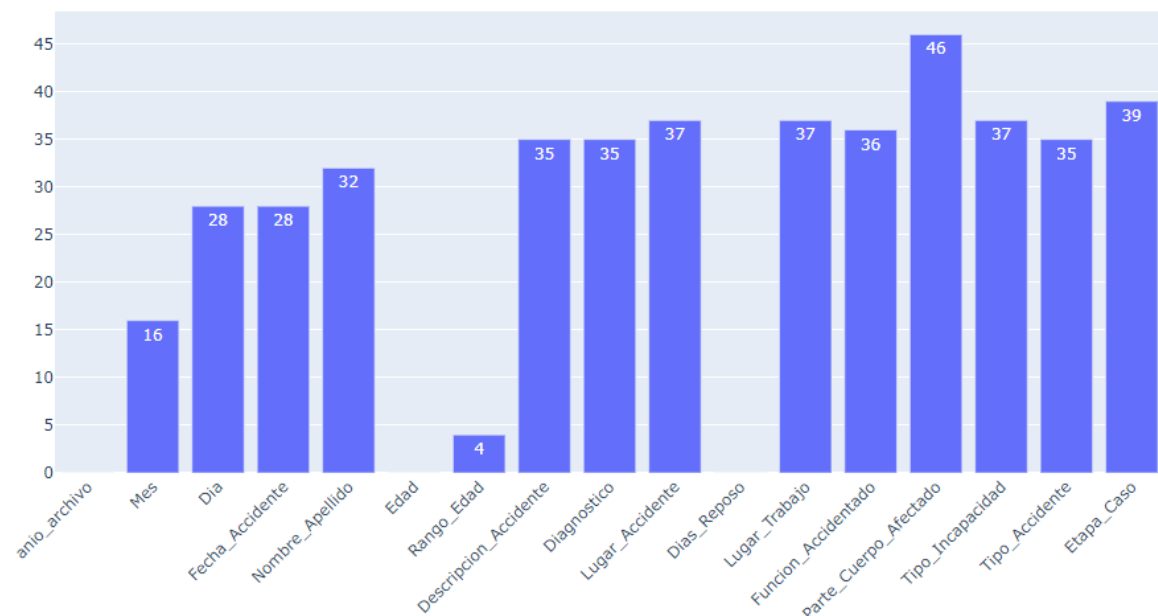
RESULTADOS

4.1.Examinar el conjunto de datos en GAD de Pichincha para su limpieza, transformación y análisis.

Usando la herramienta de Google Colab se definieron funciones para eliminar *DataFrame* y cargar nuevos conjuntos de datos desde los archivos de Excel. Se inició con una limpieza independiente por base de datos, debido a que se tenían 8 bases de datos. Posteriormente se realizó la unificación a una nueva *DataFrame*, para elaborar una gráfica de cantidades faltantes por columnas (Ilustración 2).

Ilustración 2.

Cantidad de valores Nulos por columnas



Fuente: Autor

Se realizó el proceso de transformación de los datos para identificar los posibles valores que cada variable podría tener. Se incluyeron nuevas columnas realizando acciones como unificar nombres, eliminar caracteres especiales, acentos, comas, saltos de línea y tabulaciones. Asimismo, se crearon funciones para convertir los nombres de los meses y días de la semana a sus equivalentes numéricos. Se excluyeron las entradas que no pudieron ser asociadas con conceptos como compensación, técnico responsable, observaciones, reubicación y responsabilidad patronal.

Además, se realizó un proceso de filtrado y clasificación de palabras claves para agrupar términos, este proceso se aplicó en las variables como funciones laborales, lugar de trabajo, tipo de accidentes y en otras variables relevantes. Consecutivamente los resultados generales se pueden observar en la Ilustración 3. Se logro procesar un total de 206 datos, el *DataFrame* contiene 22 columnas en total.

Ilustración 3.

Conjunto de datos después del procesamiento y transformación

```
Data columns (total 22 columns):
```

#	Column	Non-Null Count	Dtype
0	año_archivo	206 non-null	int64
1	Mes	206 non-null	object
2	Dia	206 non-null	object
3	Fecha_Accidente	206 non-null	datetime64[ns]
4	Nombre_Apellido	206 non-null	object
5	Edad	206 non-null	int64
6	Rango_Edad	206 non-null	object
7	Descripcion_Accidente	206 non-null	object
8	Diagnostico	206 non-null	object
9	Lugar_Accidente	206 non-null	object
10	Dias_Reposo	206 non-null	float64
11	Lugar_Trabajo	206 non-null	object
12	Funcion_Accidentado	206 non-null	object
13	Parte_Cuerpo_Afectado	206 non-null	object
14	Tipo_Incapacidad	206 non-null	object
15	Tipo_Accidente	206 non-null	object
16	Etapa_Caso	206 non-null	object
17	Clasificacion_Diagnostico	206 non-null	object
18	Clasificacion_Lugar_Accidente	206 non-null	object
19	Clasificacion_Lugar_Trabajo	206 non-null	object
20	Clasificacion_Parte_Cuerpo	206 non-null	object
21	Clasificacion_Funcion_Accidentado	206 non-null	object

dtypes: datetime64[ns](1), float64(1), int64(2), object(18)

Fuente: Autor

Los mapeos realizados y transformaciones de texto se estandarizaron para garantizar la consistencia. Al mismo tiempo se exporto los datos limpios a un archivo de CSV y Excel para facilitar los análisis posteriores.

4.2. Factores relacionados con accidentes laborales que resultan en días de reposo prolongados mediante análisis de datos.

Los resultados detallados de la selección de atributos revelan las características más importantes. Este enfoque permitió organizar los atributos según su relación con la

variable objetivo y evaluando cómo contribuían a disminuir el error cuadrático medio (RMSE). En los análisis se empleó el método de *Ranking* para evaluar la importancia de cada atributo en la predicción de la variable de clase, que es el número de días de reposo. En el análisis inicial de selección de atributos, se utilizó un esquema de aprendizaje que se basaba en el clasificador *GaussianProcesses*. Los atributos fueron ordenados según su relevancia para el modelo. A continuación, se presenta la ilustración 4.

Ilustración 4.

Clasificador GaussianProcesses

```
Using Wrapper Subset Evaluator
Learning scheme: weka.classifiers.functions.GaussianProcesses
Scheme options: -L 1.0 -N 0 -K weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007 -S 1
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5

Ranked attributes:
0.7747 12 Clasificacion_Diagnostico
0.0957 16 Clasificacion_Funcion_Accidentado
0.0204 3 Mes
0.0168 13 Clasificacion_Lugar_Accidente
-0.0193 6 Edad
-0.0274 1 numero
-0.043 2 anio_archivo
-0.0895 7 Rango_Edad
-0.163 5 Fecha_Accidente
-0.3619 15 Clasificacion_Parte_Cuerpo
-0.3838 11 Etapa_Caso
-0.4018 4 Dia
-0.5463 9 Tipo_Incapacidad
-0.6384 14 Clasificacion_Lugar_Trabajo
-0.7361 10 Tipo_Accidente
```

Fuente: Autor

Como se observó en la ilustración 4, el atributo más relevante en este resultado es la *Clasificacion_Diagnostico*, lo que podría indicar la gravedad del accidente. A pesar de que la *Clasificacion_Lugar_Trabajo* no es el atributo más influyente, su presencia sugiere que el lugar de trabajo podría ser un factor importante en la duración del reposo.

En este segundo análisis de selección de atributos, se utilizó el clasificador *LinearRegression*. En la ilustración 5, se puede observar que el atributo *Mes* podría estar relacionado con diferentes condiciones climáticas, y aunque el atributo *Edad* tiene una influencia negativa moderada, podría afectar la duración del descanso.

Ilustración 5.
Clasificador LinearRegression

```
Learning scheme: weka.classifiers.functions.LinearRegression
Scheme options: -S 0 -R 1.0E-8 -num-decimal-places 4
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5
```

```
Ranked attributes:
0.6516 12 Clasificacion_Diagnostico
0.0984 3 Mes
0      1 numero
0      5 Fecha_Accidente
-0.063 2 anio_archivo
-0.0682 16 Clasificacion_Funcion_Accidentado
-0.0723 6 Edad
-0.1476 7 Rango_Edad
-0.1755 15 Clasificacion_Parte_Cuerpo
-0.2581 4 Dia
-0.3705 13 Clasificacion_Lugar_Accidente
-0.6343 9 Tipo_Incapacidad
-0.8136 10 Tipo_Accidente
-0.9435 11 Etapa_Caso
-1.3168 14 Clasificacion_Lugar_Trabajo
```

Fuente: Autor

En este tercer análisis de selección de atributos, se utilizó un enfoque de aprendizaje basado en el clasificador *IBk* (k vecinos más cercanos). Tal como se muestra en la ilustración 6, una vez más, el atributo *Clasificacion_Diagnostico* sobresale como el más relevante. Aunque la influencia del atributo *Tipo_Accidente* puede variar considerablemente y afectar la recuperación.

Ilustración 6.
Clasificador IBk

```
Learning scheme: weka.classifiers.lazy.IBk
Scheme options: -K 1 -W 0 -A weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance"
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5
```

```
Ranked attributes:
0.7862 12 Clasificacion_Diagnostico
0.0513 16 Clasificacion_Funcion_Accidentado
-0.1054 3 Mes
-0.1186 7 Rango_Edad
-0.1431 13 Clasificacion_Lugar_Accidente
-0.4284 15 Clasificacion_Parte_Cuerpo
-0.5071 4 Dia
-0.6937 9 Tipo_Incapacidad
-0.7135 5 Fecha_Accidente
-0.8713 11 Etapa_Caso
-0.8821 2 anio_archivo
-1.0918 10 Tipo_Accidente
-1.5558 14 Clasificacion_Lugar_Trabajo
-5.1654 6 Edad
-7.4344 1 numero
```

Fuente: Autor

En este cuarto análisis de selección de atributos, se utilizó un enfoque de aprendizaje que se basa en el clasificador *KStar*. En la ilustración 7, se presentan los atributos ordenados y se encontró que la variable *Clasificacion_Diagnostico* se mantiene la mayor correlación.

Ilustración 7.
Clasificador KStar

```
Learning scheme: weka.classifiers.lazy.KStar
Scheme options: -B 20 -M a
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5

Ranked attributes:
0.58837 12 Clasificacion_Diagnostico
0.4486 3 Mes
0.25027 16 Clasificacion_Funcion_Accidentado
0.19633 13 Clasificacion_Lugar_Accidente
0.05301 7 Rango_Edad
0.00709 11 Etapa_Caso
0.00331 5 Fecha_Accidente
-0.14785 10 Tipo_Accidente
-0.14847 15 Clasificacion_Parte_Cuerpo
-0.15877 4 Dia
-0.17801 6 Edad
-0.22613 14 Clasificacion_Lugar_Trabajo
-0.26329 9 Tipo_Incapacidad
-0.3136 2 anio_archivo
-0.32571 1 numero
```

Fuente: Autor

En este quinto análisis de selección de atributos, se utilizó un enfoque de aprendizaje basado en el clasificador *LWL*. En la ilustración 8. Una vez más, se destaca la *Clasificacion_Diagnostico*.

Ilustración 8.
Clasificador LWL

```
Learning scheme: weka.classifiers.lazy.LWL
Scheme options: -U 0 -K -1 -A weka.core.neighboursearch.LinearNNSearch -A "weka.core.
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5

Ranked attributes:
0.8066 12 Clasificacion_Diagnostico
0.0557 16 Clasificacion_Funcion_Accidentado
-0.105 3 Mes
-0.1186 7 Rango_Edad
-0.1395 13 Clasificacion_Lugar_Accidente
-0.4091 2 anio_archivo
-0.4279 15 Clasificacion_Parte_Cuerpo
-0.4636 6 Edad
-0.507 4 Dia
-0.6961 9 Tipo_Incapacidad
-0.7434 5 Fecha_Accidente
-0.7529 1 numero
-0.8772 11 Etapa_Caso
-1.1056 10 Tipo_Accidente
-1.5581 14 Clasificacion_Lugar_Trabajo
```

Fuente: Autor

En este sexto análisis de selección de atributos, se utilizó el clasificador *DecisionTable*. Se observó que el atributo *Fecha_Accidente* tiene un impacto negativo, lo que sugiere que no está relacionado con factores estacionales (ilustración 9).

Ilustración 9.
Clasificador *DecisionTable*

```
Learning scheme: weka.classifiers.rules.DecisionTable
Scheme options: -X 1 -S weka.attributeSelection.BestFirst -D 1 -N 5
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5
```

Ranked attributes:		
0.7862	12	Clasificacion_Diagnostico
0	1	numero
0	7	Rango_Edad
0	4	Dia
0	2	anio_archivo
0	6	Edad
0	9	Tipo_Incapacidad
0	14	Clasificacion_Lugar_Trabajo
0	10	Tipo_Accidente
0	11	Etapa_Caso
-0.0574	13	Clasificacion_Lugar_Accidente
-0.2198	15	Clasificacion_Parte_Cuerpo
-0.3031	3	Mes
-0.3205	16	Clasificacion_Funcion_Accidentado
-0.7812	5	Fecha_Accidente

Fuente: Autor

En este séptimo análisis de selección de atributos, se empleó un esquema de aprendizaje basado en el clasificador *M5Rules*. Al igual que en la mayoría de los resultados previos, el atributo *Clasificacion_Diagnostico* es crucial (Ilustración 10).

Ilustración 10.
Clasificador *M5Rules*

```
Learning scheme: weka.classifiers.rules.M5Rules
Scheme options: -M 4.0 -num-decimal-places 4
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5
```

Ranked attributes:		
0.0994	12	Clasificacion_Diagnostico
0	1	numero
-0.0282	16	Clasificacion_Funcion_Accidentado
-0.0496	3	Mes
-0.063	2	anio_archivo
-0.0723	6	Edad
-0.1476	7	Rango_Edad
-0.1688	15	Clasificacion_Parte_Cuerpo
-0.2581	4	Dia
-0.4255	13	Clasificacion_Lugar_Accidente
-0.5856	14	Clasificacion_Lugar_Trabajo
-0.6104	5	Fecha_Accidente
-0.7493	11	Etapa_Caso
-0.9834	9	Tipo_Incapacidad
-1.0208	10	Tipo_Accidente

Fuente: Autor

En este octavo análisis de selección de características, se utilizó un enfoque de aprendizaje basado en el clasificador *DecisionStump*. En la ilustración 11, se observa que el atributo del Mes puede estar vinculado e impactar en la recuperación

Ilustración 11.
Clasificador DecisionStump

```
Learning scheme: weka.classifiers.trees.DecisionStump
Scheme options:
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5

Ranked attributes:
 0.78716 12 Clasificacion_Diagnostico
 0.57839  3 Mes
 0.00272  9 Tipo_Incapacidad
-0.00936  5 Fecha_Accidente
-0.11972  7 Rango_Edad
-0.16053 15 Clasificacion_Parte_Cuerpo
-0.25934  6 Edad
-0.27738 10 Tipo_Accidente
-0.28769  2 anio_archivo
-0.3121  1 numero
-0.34607 16 Clasificacion_Funcion_Accidentado
-0.44403  4 Dia
-0.64046 13 Clasificacion_Lugar_Accidente
-0.95351 11 Etapa_Caso
-1.01235 14 Clasificacion_Lugar_Trabajo
```

Fuente: Autor

En este noveno análisis de selección de atributos, se utilizó un enfoque de aprendizaje basado en el clasificador *RandomForest* (ver ilustración 12). Los atributos fueron ordenados según su relevancia para el modelo. El atributo *Clasificacion_Diagnostico* sigue siendo destacado como el más significativo en este resultado, mientras que el atributo *Clasificacion_Funcion_Accidentado* tiene una influencia moderada en la determinación del tiempo de reposo.

Ilustración 12.
Clasificador RandomForest

```
Learning scheme: weka.classifiers.trees.RandomForest
Scheme options: -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5
```

```
Ranked attributes:
0.7745 12 Clasificacion_Diagnostico
0.0855 16 Clasificacion_Funcion_Accidentado
-0.0502 13 Clasificacion_Lugar_Accidente
-0.1245 3 Mes
-0.1343 7 Rango_Edad
-0.2638 5 Fecha_Accidente
-0.3842 15 Clasificacion_Parte_Cuerpo
-0.4768 11 Etapa_Caso
-0.5036 4 Dia
-0.598 9 Tipo_Incapacidad
-0.8261 14 Clasificacion_Lugar_Trabajo
-0.8423 2 anio_archivo
-0.8956 10 Tipo_Accidente
-3.5321 6 Edad
-4.9232 1 numero
```

Fuente: Autor

En este décimo análisis de selección de atributos, se empleó un esquema de aprendizaje basado en el clasificador *RandomTree*. El atributo nuevamente con más importancia es *Clasificacion_Diagnostico*, como se observa en la ilustración 13.

Ilustración 13.
Clasificador RandomTree

```
Learning scheme: weka.classifiers.trees.RandomTree
Scheme options: -K 0 -M 1.0 -V 0.001 -S 1
Subset evaluation: RMSE
Number of folds for accuracy estimation: 5
```

```
Ranked attributes:
0.7862 12 Clasificacion_Diagnostico
0.0513 16 Clasificacion_Funcion_Accidentado
-0.1054 3 Mes
-0.1186 7 Rango_Edad
-0.1431 13 Clasificacion_Lugar_Accidente
-0.4284 15 Clasificacion_Parte_Cuerpo
-0.5071 4 Dia
-0.6937 9 Tipo_Incapacidad
-0.7135 5 Fecha_Accidente
-0.8713 11 Etapa_Caso
-0.8821 2 anio_archivo
-1.0918 10 Tipo_Accidente
-1.5558 14 Clasificacion_Lugar_Trabajo
-5.1223 6 Edad
-9.5688 1 numero
```

Fuente: Autor

Después de analizar los 10 algoritmos, se realizó una tabla con el valor cuadrático de cada variable. La información presentada en la tabla 2, resume los resultados obtenidos al aplicar diferentes enfoques de selección de atributos. Cada fila describe un método particular utilizado para identificar los atributos más significativos en la predicción de la duración de los días de reposo, mientras que las columnas indican la relevancia relativa de cada atributo según el método empleado.

Tabla 2.
Error cuadrático de las variables
Fuente: Autor

Método	Clasificación Diagnóstico	Número	Mes	Rango Edad	Día	Año Archivo	Edad	Tipo Incapacidad	Fecha Accidente	Tipo Accidente	Etapas Caso	Clasificación Función Accidentado	Clasificación Lugar Accidente	Clasificación Lugar Trabajo	Clasificación Parte Cuerpo
Gaussian Processes	0.774	0.095	0.02	-0.019	-0.361	0	0	-0.04	-0.08	0	0	-0.1	0.016	0.09	0.774

Linear Regression	0.651	0.098	0	0	-0.813	0.175	-0.072	-0.147	-0.258	0	-0.370	0.651	-0.634	-0.813	-0.943
IBk	0.786	0.051	-0.105	-0.118	-0.693	-0.882	-0.429	-0.143	-0.507	-0.259	-0.713	0.786	-0.428	-0.11	-0.696
KStar	0.588	0.448	0.250	0.196	0	-0.346	-0.1587	0	-0.263	0	0	0.053	0.196	-0.64	-0.325
LWL	0.806	0.055	-0.105	-0.118	-0.696	-0.409	-0.427	-0.463	-0.507	-0.105	-0.696	0.806	-0.139	-0.118	-0.696
Decision Table	0.786	0	0	0	-0.781	0	0	0	0	0	0	0	0	0	0

A continuación, se presenta la tabla 3 con los cinco atributos más significativos, basada en los diez algoritmos previamente realizados.

Estos atributos se han identificado como los factores más influyentes en la predicción del tiempo de recuperación, proporcionando una visión clara de los elementos clave que deben considerarse en la gestión y análisis de datos relacionados con los días de reposo.

M5Rules	0.099	0	-0.028	-0.049	-0.61	0	0	0	-0.07	-0.425	-0.168	0.099	-0.147	-0.98	-1.020
Decision Stump	0.787	0.578	0.002	-0.009	-0.119	-0.259	-0.160	-0.287	0	-0.444	-0.953	0.787	-0.640	-0.312	-0.147
Random Forest	0.774	0.085	-0.050	-0.124	-0.263	0	-0.384	-0.895	-0.50	0	-0.842	0.774	-0.050	-0.13	-0.476
RandomTree	0.786	0.051	-0.105	-0.118	-0.693	-0.428	-0.143	-0.507	-0.26	-0.871	-0.882	0.786	-0.143	-0.11	-0.693

Tabla 3.
Cinco atributos mas significativos

Algoritmo utilizado	Primer atributo	Segundo atributo	Tercer atributo	Cuarto atributo	Quinto atributo
<i>Gaussian Processes</i>	Clasificacion_Diagnostico	Clasificacion_Lugar_Trabajo	Mes	Edad	Fecha_Accidente

Algoritmo utilizado	Primer atributo	Segundo atributo	Tercer atributo	Cuarto atributo	Quinto atributo
Linear Regression	Mes	Edad	Tipo_Incapacidad	Tipo_Accidente	Fecha_Accidente
IBk	Clasificacion_Diagnostico	Tipo_Accidente	Mes	Edad	Clasificacion_Lugar_Trabajo
KStar	Mes	Clasificacion_Lugar_Trabajo	Clasificacion_Lugar_Accidente	Edad	Etapas_Caso
LWL	Clasificacion_Diagnostico	Año_Archivo	Mes	Clasificacion_Lugar_Accidente	Clasificacion_Funcion_Accidentado
DecisionTable	Clasificacion_Diagnostico	Fecha_Accidente	Clasificacion_Funcion_Accidentado	Clasificacion_Lugar_Accidente	Mes

Algoritmo utilizado	Primer atributo	Segundo atributo	Tercer atributo	Cuarto atributo	Quinto atributo
<i>M5Rules</i>	Clasificacion_ Diagnostico	Clasificacion_ Funcion_ Accidentado	Clasificacion_ Lugar_Trabajo	Mes	Año_Archivo
<i>DecisionStump</i>	Clasificacion_ Diagnostico	Mes	Tipo_Incapacidad	Fecha_Accidente	Edad
<i>RandomForest</i>	Clasificacion_ Diagnostico	Clasificacion_ Funcion_ Accidentado	Mes	Fecha_Accidente	Clasificacion_Lugar_ Accidente
<i>RandomTree</i>	Clasificacion_ Diagnostico	Mes	Clasificacion_ Funcion_	Clasificacion_Lugar_ Accidente	Tipo_Accidente

Algoritmo utilizado	Primer atributo	Segundo atributo	Tercer atributo	Cuarto atributo	Quinto atributo
--------------------------------	----------------------------	-----------------------------	------------------------	------------------------	------------------------

Accidentado

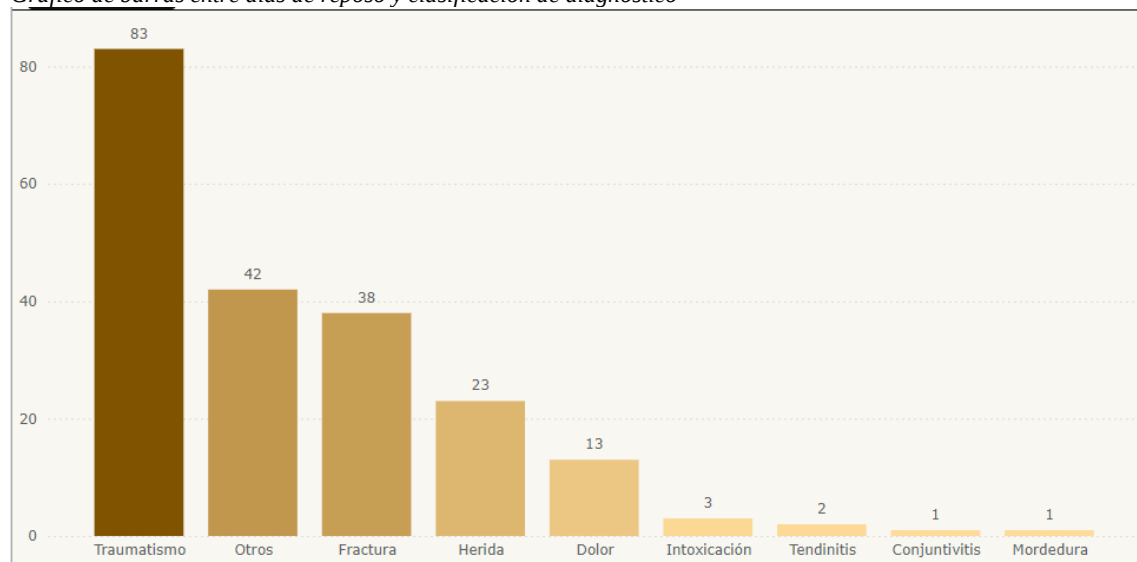
Fuente: Autor

Se analizó varios algoritmos, desde regresión lineal hasta bosques aleatorios y árboles de decisión. Esto resalta la importancia de explorar diferentes enfoques para identificar los atributos más relevantes que resultan de los días de reposo. Como se observa en la tabla 3, la clasificación del diagnóstico sigue siendo un elemento crucial, resaltando su papel fundamental en predecir el tiempo de recuperación. Así mismo, el mes se presenta como un factor importante, ya que se registran más accidentes posiblemente debido a cambios ambientales, festividades locales, vacaciones y otros aspectos. Por último, se enfatiza con la función del accidentado, lo cual sugiere que el tipo de trabajo o actividad realizada por la persona afectada tienen un impacto significativo en su proceso de recuperación.

La relación entre días de reposo y la clasificación de diagnóstico es el asociado con los accidentes laborales, para esto se realizó un rango por los días reposo para obtener un gráfico más visible. Este rango es de 0 a 3 días, de 4 a 10 días, de 11 a 20 días y así sucesivamente hasta los 100 y el ultimo rango es de más de 101 días. Como se observa en la ilustración 14, el traumatismo es la causa por la cual se dan más días de reposo con un 40.74%. Es importante mencionar que dos datos no tenían una clasificación definida, por lo que se categorizaron como no otros.

Ilustración 14.

Gráfico de barras entre días de reposo y clasificación de diagnostico

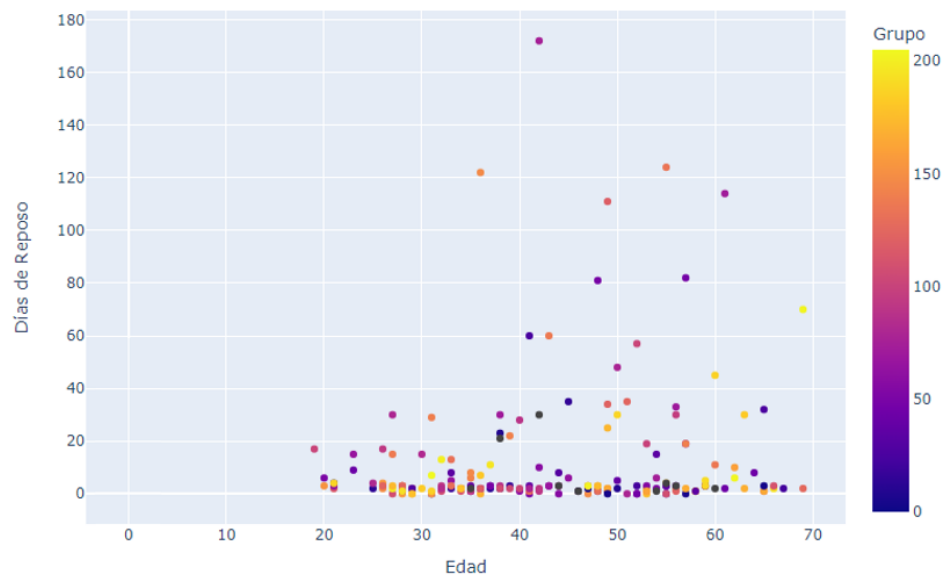


Fuente: Autor

La relación entre edad y cantidad de días de reposo tras un accidente laboral. Se observa que a medida aumenta la edad de los trabajadores tienen a experimentar periodos más prolongados de recuperación. En la ilustración 15, se muestra que un trabajador de 42 años tuvo 142 días de reposo, seguido por el trabajador de 55 años con 124 días, el de 36 años con 122 días, el de 49 años con 119 días y finalmente el trabajador de 61 años con 114 días de reposo. Este análisis numérico se ve respaldado por el gráfico de dispersión, donde se evidencia claramente cómo la edad influye en la duración de los periodos de recuperación tras un accidente laboral.

Ilustración 15.

Gráfico de dispersión entre días de reposo y edad

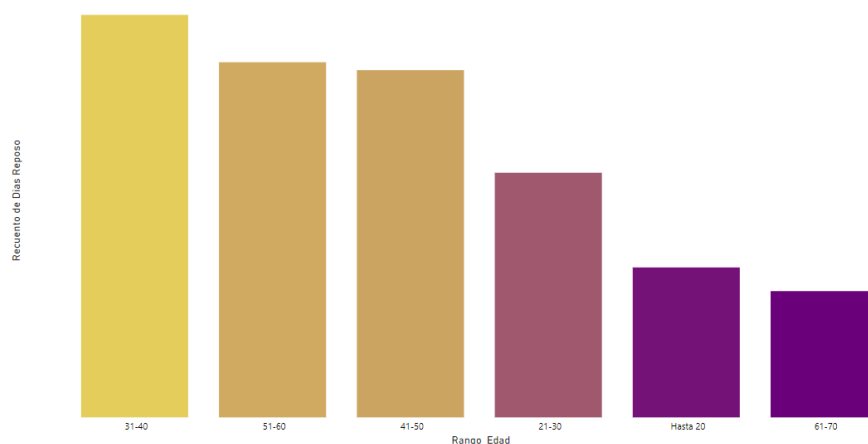


Fuente: Autor

Para complementar estos resultados se generaron gráficos de barras, el grupo de edad de 31 a 40 años se destaca con más accidentes laborales con más días de reposo registrando un total de 51 accidentes según los datos proporcionados (Ilustración 16). Esta información resalta la importancia de dirigir las medidas de prevención y capacitación hacia este grupo demográfico específico. Por otro lado, llama la atención que el grupo de edad de 61 a 70 tienen menores días de reposo.

Ilustración 16.

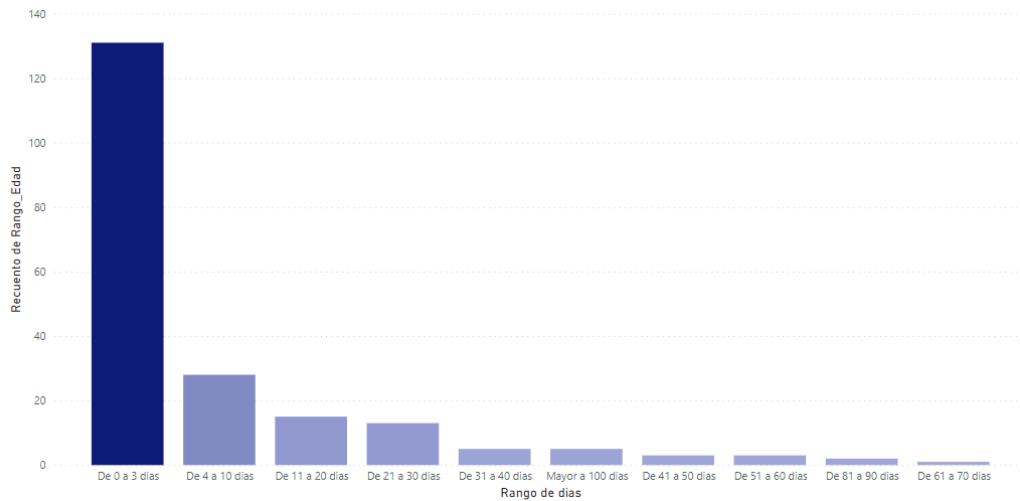
Gráfico de barras en días de reposo y rango de edad



Fuente: Autor

Comprender las razones detrás de esta tendencia cual es el mayor rango de días de reposo, se utilizó nuevamente el rango de los días de reposo. Con esto nos damos cuenta que a la mayoría de trabajadores se les otorga de 0 a 3 días de permiso por el tipo de accidente que haya sufrido (ilustración 17).

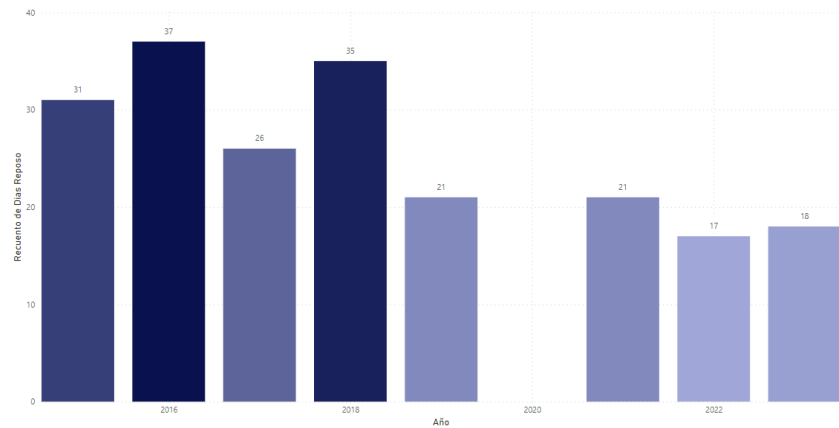
Ilustración 17.
Gráfico de barras de cantidad de días de reposo



Fuente: Autor

Otra relación es con el año, en la ilustración 18, se relevan que los años 2016 y 2018 se destacan por tener más días de reposo prolongados. Además, la ausencia de registros en el año 2020 es ocasionada por la pandemia de COVID-19.

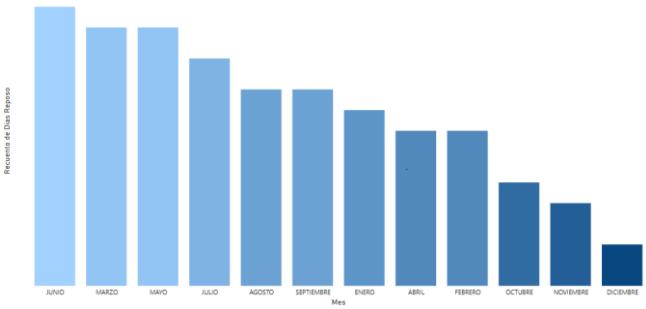
Ilustración 18.
Gráfico de barras entre días de reposo y año



Fuente: Autor

Los meses de junio, marzo y mayo destacan como periodos con mayores días de reposo, estos podrían estar relacionados con cambios climáticos o picos de actividad, como se visualiza en la ilustración 19. Además, se debería profundizar en las circunstancias que rodean estos meses para realizar una planificación de estrategias de prevención.

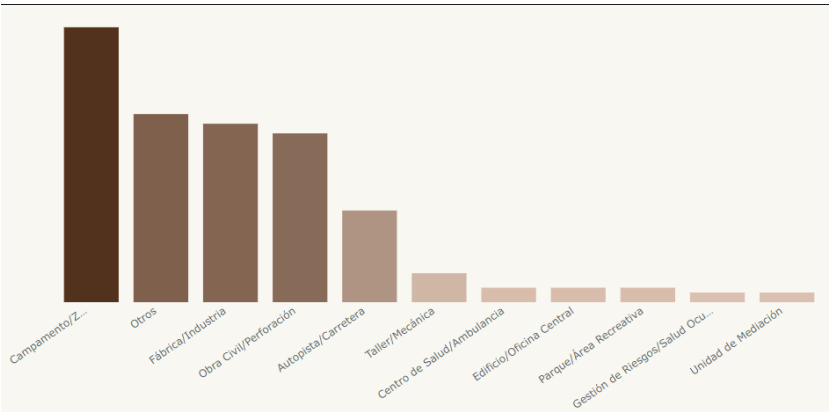
Ilustración 19.
Gráfico de barras entre días de reposo y mes



Fuente: Autor

El lugar de trabajo más frecuente relacionado con los días de reposo según los accidentes laborales que han existido en el GAD de Pichincha. Se identifica la categoría Otros porque no contaba con detalles específicos en la descripción o habían colocado únicamente las coordenadas, por esta razón se agruparon bajo una categoría genérica. Además, los lugares como campamentos, fábricas y obra civil son los lugares donde existían más riesgos laborales. (Ilustración 20)

Ilustración 20.
Gráfico de barras entre días de reposo y lugar de trabajo

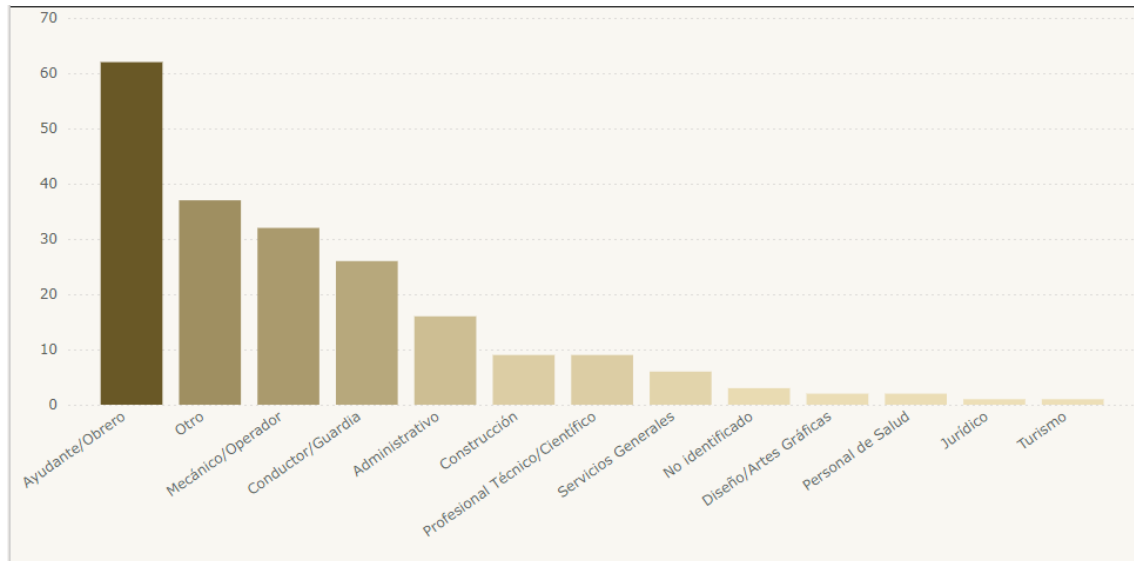


Fuente: Autor

El atributo de función del accidentado Ayudante obrero es la más afectada por ende es la que es el atributo que tiene más días de reposo por los accidentes laborales, seguida por la categoría de Otros con 37 casos que agrupaba a los trabajadores cuyas funciones no estaban especificadas en los registros como se puede visualizar en la ilustración 21.

Ilustración 21.

Gráfico de barras entre días de reposo y función del trabajador



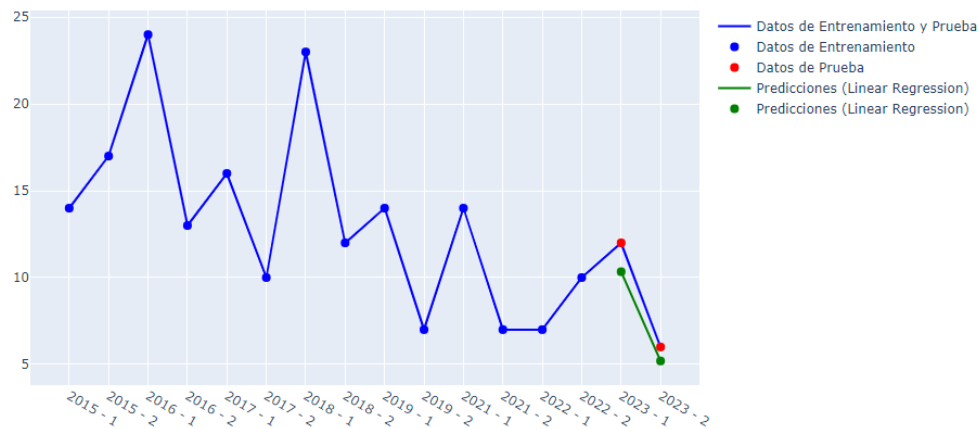
Fuente: Autor

4.3. Tendencia semestral de accidentes laborales en el GAD de Pichincha

La tendencia fue analizada considerando medidas de desempeño. Al evaluar distintos modelos como Regresión Lineal, Bosque Aleatorio y Árbol de Decisión, es crucial tener en cuenta el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2) para estimar la cantidad de accidentes de tráfico por semestre.

Al estudiar los resultados obtenidos, se destaca que la Regresión Lineal exhibe el MSE más bajo y el R^2 más alto. Estos hallazgos sugieren que la regresión lineal tiene un mejor rendimiento en predecir los datos, tal como se muestra en la ilustración 22.

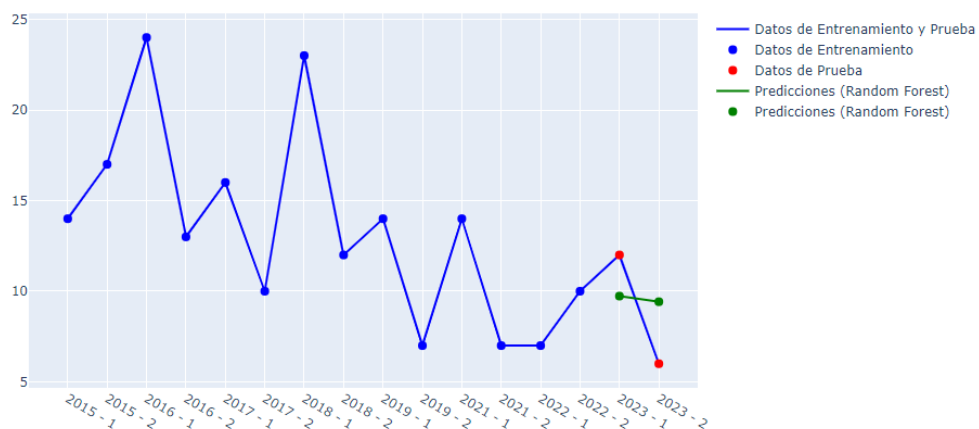
Ilustración 22.
Modelo de Regresión Lineal



Fuente: Autor

En diferencia, al analizar el modelo Bosque Aleatorio (*Random Forest*), se observa un MSE relativamente elevado en comparación con la regresión lineal, lo que podría indicar ciertas dificultades para ajustarse correctamente a los datos de prueba. Además, su R^2 cercano a cero señala una capacidad limitada para explicar la variabilidad presente en los datos (Ilustración 23).

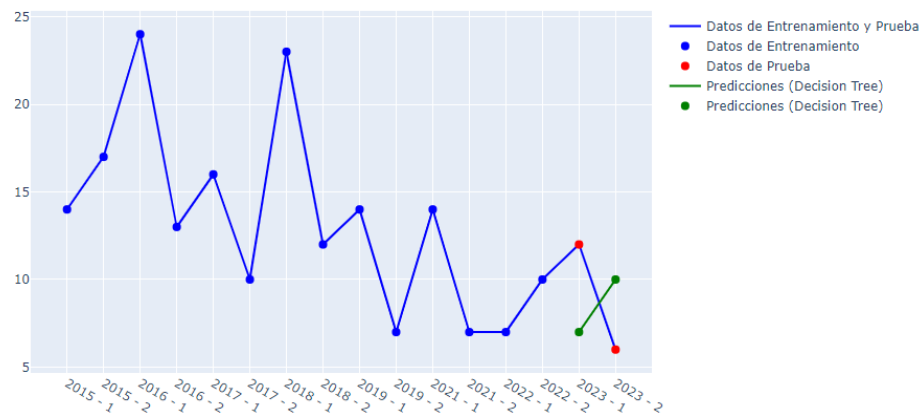
Ilustración 23.
Modelo de Bosque Aleatorio



Fuente: Autor

Por último, el modelo del Árbol de Decisión presenta el peor rendimiento, con un MSE significativamente más alto y un R^2 negativo según se puede apreciar en las ilustraciones. Estos resultados sugieren que este modelo es menos efectivo incluso que una aproximación simple que solo predice la media de los datos utilizados para entrenamiento.

Ilustración 24.
Modelo de Árbol de Decisión



Fuente: Autor

La comparación entre varios modelos de regresión, como Regresión Lineal, Bosque Aleatorio y Árbol de Decisión, muestra resultados importantes sobre su capacidad predictiva en un entorno particular. Se presentan los descubrimientos en la tabla siguiente:

Tabla 4.
Comparación de modelos

Modelo	MSE	R^2
Regresión Lineal	1.7057	0.8105
Bosque Aleatorio	8.4246	0.0639
Árbol de Decisión	20.5	-1.2778

Fuente: Autor

Se destaca la superioridad del modelo de Regresión Lineal, que presenta el MSE más bajo y el R^2 más alto. Esto señala una menor discrepancia entre las predicciones y los valores observados, así como una mejor capacidad para explicar la variabilidad en los datos.

Por otro lado, tanto el modelo Random Forest como el Árbol de Decisión muestran MSE más altos y R^2 más bajos, lo que sugiere un ajuste menos adecuado a los datos y una capacidad predictiva inferior en comparación con el modelo de Regresión Lineal.

Se utilizaron diferentes tipos de regresiones lineales, incluyendo la regresión lineal simple y la regresión lineal con características polinómicas, además de técnicas de regularización como Ridge (L2) y Lasso (L1). Los resultados indicaron que incorporar características polinómicas de grado.

En adición a los resultados previamente mencionados, se presentará en la tabla 6 las predicciones realizadas para los años 2024, 2025 y 2026.

Tabla 5.
Predicciones de los 3 años siguientes

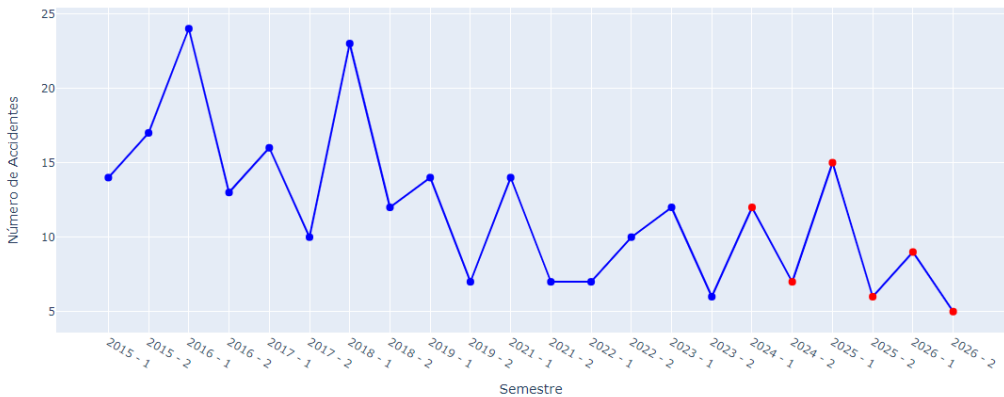
Semestre	Numero de accidente	Tipo
1	12.48	Predicciones 2024
2	7.23	Predicciones 2024
1	15.50	Predicciones 2025
2	6.25	Predicciones 2025
1	9.52	Predicciones 2026
2	5.27	Predicciones 2026

Fuente: Autor

En la ilustración 25, se muestra el número de accidentes laborales por semestre, donde los puntos azules representan los datos de entrenamiento y los puntos rojos muestran las

predicciones para los años 2024, 2025 y 2026. La línea azul conecta tanto los datos de entrenamiento como las predicciones, ofreciendo una representación visual clara de la tendencia descendente pronosticada en el número de accidentes laborales durante los próximos semestres.

Ilustración 25.
Predicción de accidentes por semestre y predicciones para 2024, 2025 y 2026

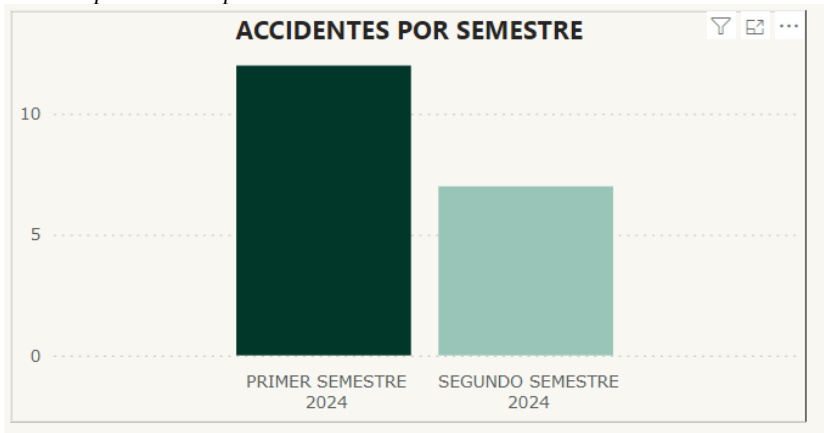


Fuente: Autor

4.3.1. Tendencia de accidentes laborales del 2024 en el GAD de Pichincha

En la ilustración 26, centrada en el primer semestre del año, se indicará que entre enero y junio se registrarán un total de 12 accidentes en el GAD de Pichincha en el primer trimestre y 7 en el segundo trimestre.

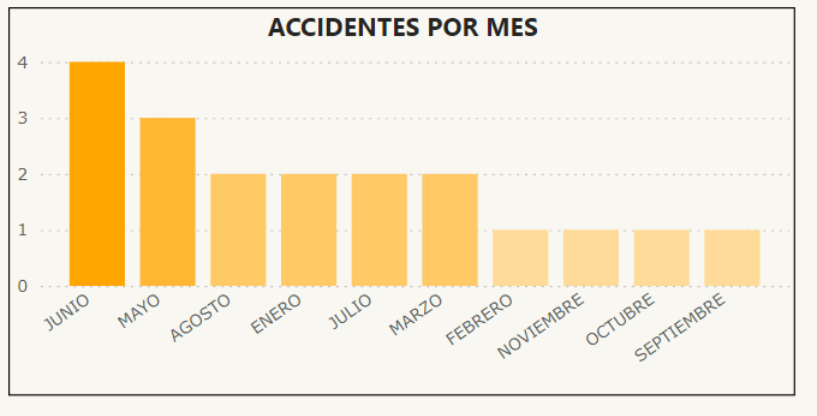
Ilustración 26.
Predicción de accidentes por semestre para 2024



Fuente: Autor

En la ilustración 27, se detallarán los meses específicos dentro de los semestres donde ocurrirán los accidentes laborales durante el año 2024. Los meses más críticos en términos de incidentes resultarán junio y mayo mostrando un aumento significativo comparado con otros meses del año.

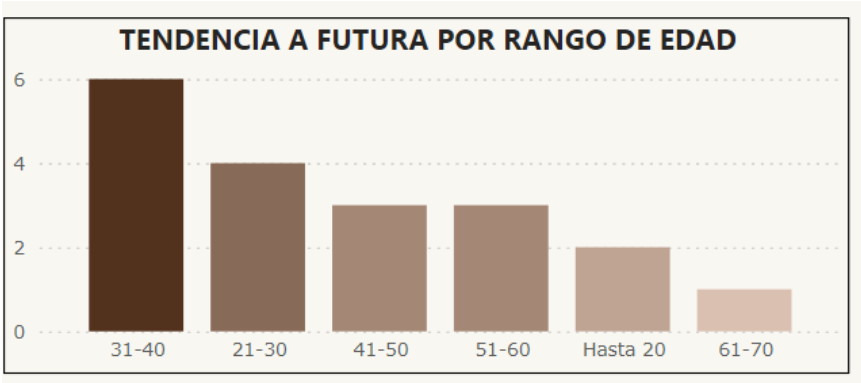
Ilustración 27.
Predicción de accidentes por mes para 2024



Fuente: Autor

La representación visual (ilustración 28) brindará una visión detallada sobre la distribución de los accidentes según el rango de edad de las personas afectadas. Se destacará una concentración importante de incidentes entre los grupos etarios de 31 a 40 años y 21 a 30 años,

Ilustración 28.
Predicción de accidentes por rango de edad para 2024



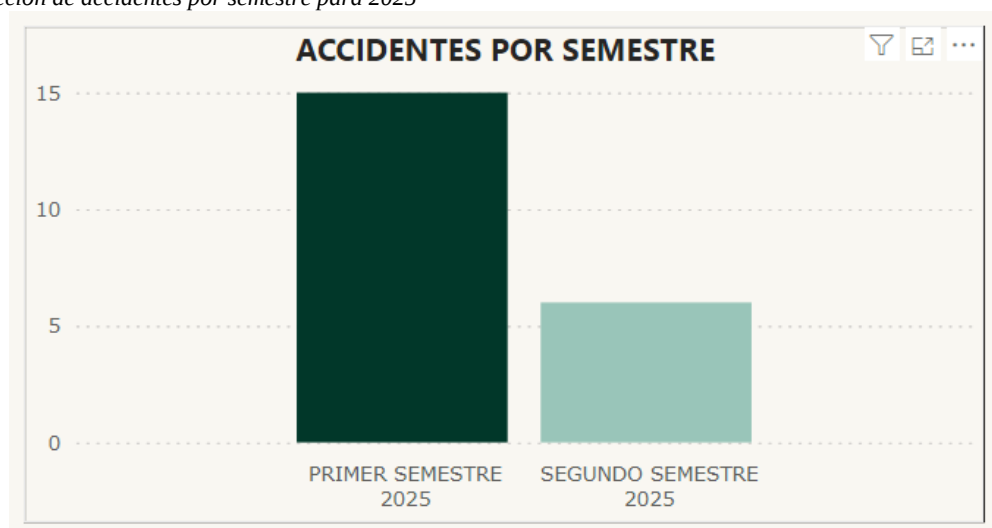
Fuente: Autor

4.3.2. Tendencia de accidentes laborales del 2025 en el GAD de Pichincha

En la ilustración 29, que se mostrará a continuación, se presentará una descripción detallada de las previsiones de accidentes laborales en el GAD de Pichincha para el año 2025. Según las estimaciones, se prevé un total de 15.50 accidentes durante el primer semestre y 6.25 durante el segundo semestre del año mencionado. Este gráfico brindará una perspectiva fundamental para identificar posibles áreas de riesgo y diseñar estrategias preventivas adecuadas.

Ilustración 29.

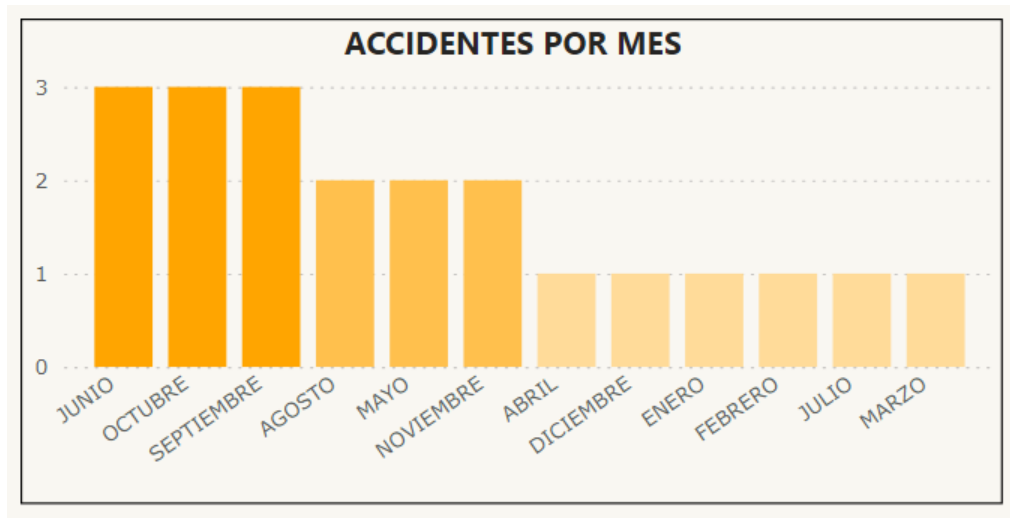
Predicción de accidentes por semestre para 2025



Fuente: Autor

En la ilustración 30, se detallarán los meses por semestre en los que se pronosticará que ocurrirán accidentes laborales en el GAD de Pichincha en el año 2025. Los meses críticos en términos de incidentes serán junio, octubre y septiembre. Esta representación temporal facilitará la comprensión de los patrones estacionales de accidentes y podrá guiar la asignación de recursos para intervenciones preventivas específicas.

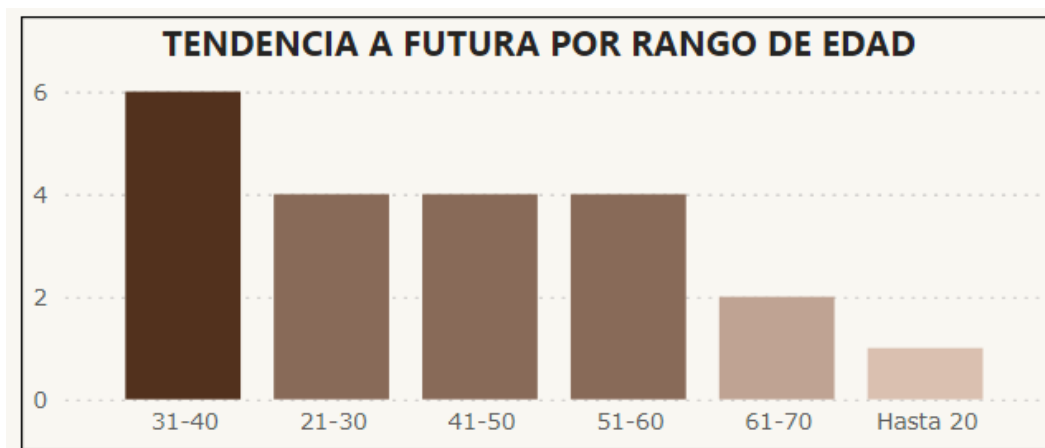
Ilustración 30.
Predicción de accidentes por mes para 2025



Fuente: Autor

Por último, en la ilustración número 31 se mostrará la franja etaria de las personas afectadas por los accidentes laborales en el GAD de Pichincha. Se observará una distribución diversa, con un número significativo de incidentes entre los grupos etarios comprendidos entre los 31 a 40 años y los 21 a 30 años.

Ilustración 31.
Predicción de accidentes por rango de edad para 2025



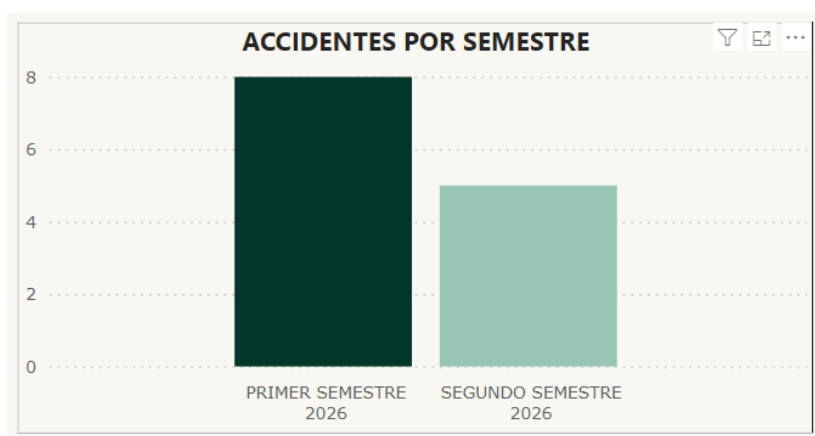
Fuente: Autor

4.3.3. Tendencia de accidentes laborales del 2026 en el GAD de Pichincha

En la ilustración 32, se mostrará una representación detallada de las proyecciones de accidentes laborales en el (GAD de Pichincha para el año 2026. Según las estimaciones, se esperará un total de 9.52 accidentes durante el primer semestre y 5.27 durante el segundo semestre del año mencionado. Esta representación visual proporcionará una valiosa perspectiva para anticipar posibles áreas de riesgo y diseñar estrategias preventivas adecuadas.

Ilustración 32.

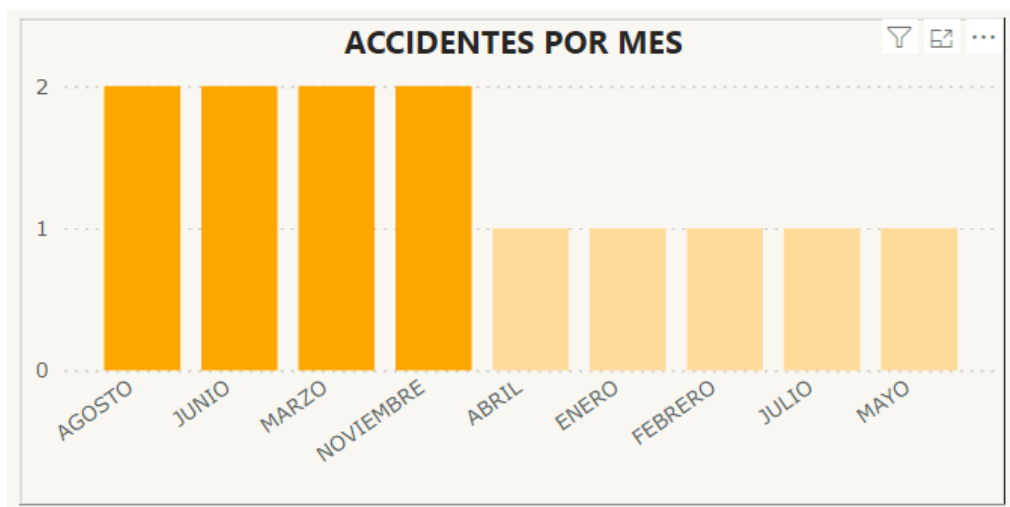
Predicción de accidentes por semestre para 2026



Fuente: Autor

En la ilustración 33, se describirán los meses por semestre en los que se pronosticarán accidentes laborales en el GAD de Pichincha para el año 2026. Se observará una distribución variada de los accidentes a lo largo del año, con concentraciones significativas en agosto, junio, marzo y noviembre.

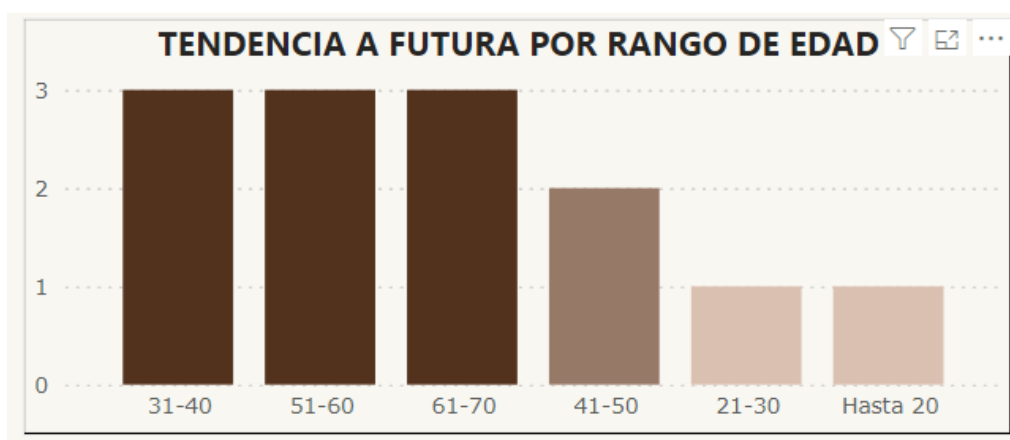
Ilustración 33.
Predicción de accidentes por mes para 2026



Fuente: Autor

Por último, en la ilustración 34, se mostrará la franja etaria de las personas afectadas por los accidentes laborales en el GAD de Pichincha. Se apreciará una diversidad en los grupos etarios afectados, con una destacada concentración de casos entre las edades comprendidas entre 31-40 años y 51-70 años. Estos resultados resaltarán la importancia de ajustar las estrategias de seguridad en entornos laborales para abordar las necesidades específicas de cada segmento demográfico.

Ilustración 34.
Predicción de accidentes por rango de edad para 2026



4.4. Discusión

La investigación actual se centró en analizar los accidentes laborales en el GAD de Pichincha, centrándose en identificar factores relacionados con días de reposo prolongados y patrones semestrales. Es fundamental vincular estos descubrimientos con estudios previos realizados por otros investigadores en el campo de la seguridad laboral.

En cuanto a las similitudes, este estudio concuerda con investigaciones como el estudio de Administración de Seguridad y Salud Ocupacional (OSHA) o el estudio longitudinal sobre la seguridad laboral y la salud de los trabajadores de diferentes edades en varios sectores industriales que resaltan la importancia de la edad como un factor determinante en la duración de los días de reposo tras un accidente laboral. La tendencia observada aquí, donde los trabajadores mayores experimentan períodos de recuperación más largos, ha sido consistentemente respaldada por la literatura existente (Smith et al., 2018; García-Pérez et al., 2020). Esta coincidencia refuerza aún más la relevancia y solidez de los resultados en este aspecto particular.

Sin embargo, los descubrimientos significativos de esta investigación no se enfocan en la utilización de técnicas específicas de análisis de datos, como `CorrelationAttributeEval`, para identificar de manera sistemática los factores más relevantes asociados a los días de descanso prolongado y la ubicación de los incidentes. A través de estos métodos, se logró resaltar aspectos cruciales como las tareas laborales más propensas a sufrir accidentes, la clasificación del diagnóstico, la fecha del incidente, la posición del trabajador afectado y el sitio del suceso. Estos resultados subrayan la importancia de abordar estos aspectos

para mejorar la prevención y gestión de accidentes laborales dentro del ámbito del GAD de Pichincha.

Es fundamental reconocer las limitaciones de este estudio. La principal restricción reside en el tamaño reducido del conjunto de datos utilizado, que constaba únicamente de 206 variables. Esta limitación podría haber impactado la capacidad para abarcar toda la complejidad de los accidentes laborales en la región y posiblemente haya sesgado los resultados obtenidos. Además, la falta de información detallada sobre ciertas variables, como la salud general de los trabajadores y las condiciones específicas del entorno laboral, pudo haber restringido la capacidad para identificar relaciones causales más complejas entre las variables analizadas.

En cuanto a las implicaciones prácticas, los resultados de este estudio podrían tener un impacto significativo en el desarrollo de políticas y estrategias relacionadas con la seguridad laboral en el GAD de Pichincha. La identificación de factores específicos vinculados a períodos prolongados de ausencia laboral y al lugar donde ocurren los accidentes podría asistir a las autoridades competentes a implementar medidas preventivas más eficaces y focalizadas. Por ejemplo, sería factible diseñar programas formativos dirigidos a grupos demográficos con mayor riesgo, así como establecer medidas adicionales para mejorar la seguridad en áreas laborales particularmente susceptibles a sufrir accidentes.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

El proceso de preparación, limpieza y transformación de los datos se realizó utilizando herramientas avanzadas como Pandas y librerías específicas de limpieza de datos, asegurando la integridad y coherencia de los registros. Además, se implementaron técnicas de visualización de datos para identificar y abordar posibles anomalías en la información recopilada, lo que contribuyó aún más a la robustez de la base de datos. Los resultados revelan que se han analizado un total de 206 registros, que incluyen 22 columnas con información importante sobre los accidentes laborales, como año, edad, puesto de trabajo, tipo de accidente y días de descanso. Es relevante destacar que se ha establecido un formato estándar para la recopilación de datos, garantizando la consistencia en la estructura de los datos cada año. Esta medida ha sido clave para asegurar la coherencia y comparabilidad de la información a lo largo del tiempo, a pesar de las posibles variaciones en las variables recopiladas en distintas bases de datos.

El análisis de los datos sobre incidentes laborales en el GAD de Pichincha ha permitido obtener un profundo entendimiento de los factores que impactan en la seguridad en el trabajo en la zona. A través del uso de técnicas avanzadas de análisis estadístico y minería de datos, se han descubierto diversos patrones y relaciones importantes entre variables clave que ponen de manifiesto la importancia de variables como la categorización del diagnóstico, la edad del personal y el mes del suceso en la determinación de la duración de los períodos de reposo. Este análisis detallado ofrece una visión más completa sobre los riesgos laborales en el área, como por ejemplos, los empleados más mayores podrían necesitar más tiempo para recuperarse después de una lesión en comparación con sus colegas más jóvenes. Reconocer esta conexión puede ser útil para adaptar programas específicos a las necesidades particulares de cada grupo demográfico. Estos resultados

podrían aportar al GAD de Pichincha y establecer bases sólidas para desarrollar políticas y estrategias preventivas más efectivas.

La tendencia semestral de accidentes laborales en el GAD de Pichincha ha mostrado que la regresión lineal presenta un Error Cuadrático Medio (MSE) más bajo y un coeficiente de determinación (R^2) más alto en comparación con el Random Forest y el Árbol de Decisión. A través de este análisis de métricas de rendimiento, se ha observado un mejor desempeño en el algoritmo para predecir accidentes laborales, lo que sugiere que podría ser una herramienta útil para pronosticar accidentes en el futuro. Es importante señalar que este modelo solo brinda una estimación del número total de accidentes por semestre y anualmente.

5.2. RECOMENDACIONES

La adopción de un sistema de normalización de datos aseguraría la coherencia e integridad de la información recopilada en todas las bases de datos. Esto implicaría establecer un conjunto uniforme de variables y campos que deben ser registrados anualmente, garantizando así que se capturen todos los aspectos relevantes de los accidentes laborales de manera consistente. Además, se propone llevar a cabo capacitaciones regulares para el personal encargado de recopilar los datos con el fin de mejorar la precisión y exhaustividad de la información registrada. También se aborda cómo optimizar el proceso de recopilación y análisis de datos. Por último, se destaca la importancia de adoptar un enfoque proactivo hacia la prevención de accidentes laborales, promoviendo una cultura segura en el entorno laboral.

Para mejorar la seguridad laboral y reducir los accidentes en el GAD de Pichincha, se sugiere implementar medidas específicas basadas en los resultados de este estudio, para

crear programas de formación y sensibilización dirigidos a los grupos demográficos identificados como más propensos a sufrir accidentes laborales, así como reforzar la supervisión y el cumplimiento de las normas de seguridad en los lugares de trabajo. También es importante mejorar la calidad de los registros de accidentes laborales y establecer un sistema de seguimiento continuo para detectar y abordar activamente posibles riesgos. Estas sugerencias, respaldadas por pruebas empíricas, pueden contribuir significativamente a proteger y cuidar tanto a los trabajadores del GAD de Pichincha como a aumentar la eficiencia y productividad en el ámbito laboral.

Se sugiere mejorar la calidad y cantidad de datos disponibles para análisis futuros, a pesar de la eficacia de la regresión lineal. Es importante considerar la inclusión de datos del año 2020, que no se tuvieron en cuenta debido a la pandemia, y realizar investigaciones adicionales para estudiar cómo fenómenos como pandemias o cambios climáticos podrían influir en la incidencia de accidentes laborales. Asimismo, se recomienda continuar monitoreando y recolectando datos con el tiempo para poder reevaluar el rendimiento de diferentes modelos y explorar nuevas técnicas de análisis de datos a medida que la base de datos crezca. Esto ayudará a obtener predicciones más precisas y relevantes que contribuyan significativamente a mejorar la seguridad laboral en el GAD de Pichincha.

BIBLIOGRAFÍA

- Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. *Procedia Computer Science* , 82, 65-71. doi:<https://doi.org/10.1016/j.procs.2016.04.010>
- Ampuero, E., Pozo, M., & Delgado, K. (2018). Administración de riesgo laboral en el Ecuador. *Digital Publisher*, 3(5), 1-10.
- Arango, J., Correa, Y., & Luna, J. (2020). La salud ocupacional y su respuesta histórica a las necesidades de salud de los trabajadores . *Revista Cbana de Salud y Trabajo*, 21(2), 14-29.
- Asamblea Nacional Constituyente de Ecuador. (2008). *Constitución de la República de Ecuador*. Quito: Lexus.
- Asamblea Nacional del Ecuador. (2008). *Código del Trabajo*. República del Ecuador.
- Asamblea Nacional del Ecuador. (2008). *Constitución de la República del Ecuador*. República del Ecuador.
- Auqui, D. (2020). Reflexiones científicas sobre la salud ocupacional y el sistema general de riesgos profesionales en Ecuador. *Polo del Conocimiento*, 5(4), 166-191. doi:10.23857/pc.v5i4.1371
- Benalcázar, J., Castro, R., & Vélez, M. (2017). Algunas circunspecciones acerca de la salud ocupacional. *Polo del Conocimiento*, 2(4). doi:10.23857/pc.v2i4.109
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. doi:<https://doi.org/10.1023/A:1010933404324>
- Candonga, J., & Samaniego, P. (2021). Percepción de riesgos laborales en el ámbito ocupacional universitario. *Revista Publicando*, 8(28), 47-58. doi:<https://doi.org/10.51528/rp.vol8.id2147>
- Cánovas, F., Alonso, F., Gomariz, F., & Oñate, F. (2017). Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. . *Computers & Geoscience*, 103, 1-11. doi:<https://doi.org/10.1016/j.cageo.2017.02.012>
- Codigo del Trabajo. (2012). *Obligaciones respecto de la prevención de riesgos- Artículo 410- Registro Oficial Suplemento 167 de 16-dic-2005*. Lexis.
- Delvalle, F. (2022). *Modelo de Gestión Administrativa en el Sistema de Gestión de Riesgo de Trabajo del IESS*. Obtenido de Universidad Estatal del Sur de Manabí: <https://repositorio.unesum.edu.ec/bitstream/53000/4039/1/Fernando%20Del%20Valle%20Mor%C3%A1n.%20tesis%20final.pdf>

- Espinosa-Zúñiga, J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, XXI (3), 1-16. doi:<https://doi.org/10.22201/fi.25940732e.2020.21.3.022>
- Espinoza, M., & Ramos, J. (2021). Análisis comparativo de la accidentabilidad laboral en Ecuador: periodo 2014 al 2019. *Digital Publisher*, 6(6). doi:<http://doi.org/10.33386/593dp.2021.6.735>
- Fernández Villafañez, S. (2022). *MÉTODOS DE REGRESIÓN Y CLASIFICACIÓN BASADOS EN ÁRBOLES*. UNIVERSIDAD DE VALLADOLID. Obtenido de <https://uvadoc.uva.es/bitstream/handle/10324/53822/TFG-I-2181.pdf?sequence=1>
- Florencia, M. (2018). *EXPLORACIÓN Y ANÁLISIS EXHAUSTIVO DE LOS MÉTODOS ESTADÍSTICOS APLICADOS A LAS SERIES DE TIEMPO*. Universidad Nacional del Litoral. Obtenido de <https://fce.unl.edu.ar/jornadasdeinvestigacion/trabajos/uploads/trabajos/45.pdf>
- García, J. (2020). *MODELO PREDICTIVO PARA LA IDENTIFICACIÓN DE ACTIVIDADES DE LA VIDA DIARIA (ADL) EN AMBIENTES INDOOR USANDO TÉCNICAS DE CLASIFICACIÓN BASADAS EN MACHINE LEARNING*. Universidad de la Costa. Obtenido de <https://repositorio.cuc.edu.co/bitstream/handle/11323/8179/Modelo%20predictivo%20para%20la%20identificaci%C3%B3n%20de%20actividades%20de%20la%20vida%20diaria%20%28ADL%29%20en%20ambientes%20INDOOR%20usando%20t%C3%A9cnicas%20de%20clasificaci%C3%B3n%20basada>
- González, R. (2021). *Técnicas de Minería de Datos Aplicadas al Análisis de Rotación de Artículos*. Universidad Católica Nuestra Señora de la Asunción. Obtenido de https://www.conacyt.gov.py/sites/default/files/Tesis-Ricardo_Gonzalez.pdf
- Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression Second Edition*. Wiley-Interscience. Obtenido de https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/8401/Chavez_Martinez_Renato.pdf?sequence=3&isAllowed=y
- Hussain, S., Dahan, N., Ba-Alwi, F., & Ribata, N. (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2), 447-459. doi:<https://doi.org/10.11591/ijeecs.v9.i2.pp447-459>
- IBM. (2021). *Modelos personalizados de suavizado exponencial*. Obtenido de <https://www.ibm.com/docs/es/spss-statistics/27.0.0?topic=modeler-custom-exponential-smoothing-models>
- IESS. (2015). *Reglamento de Seguridad y Salud de los trabajadores- Arículo 11- Decreto Ejecutivo 2393*.

- IESS. (2018). *Seguro general de riesgos del trabajo*. Obtenido de IESS: https://www.iesgob.ec/documents/10162/51889/Boletin_estadistico_2018_nov_dic.pdf
- Indstituto Nacional de Seguridad y Salud en el Trabajo. (2017). *Accidente de Trabajo*. Obtenido de INSST: <https://www.insst.es/documents/94886/789577/seguridad+guia+monitor.pdf/0b93b59d-5d99-4747-9887-b4244b8fef39?t=1605801764370>
- Instrumento Andino de Seguridad y Saud en el Trabajo. (2004). *Decisión del Acuerdo de Cartagena 584-Registro Oficial Suplemento 461 de 15-nov.-2004*. Lexis.
- Jesús, C. P. (2019). *“Modelamiento predictivo de riesgos psicosociales en trabajadores de una empresa del sector educativo” para mejorar el proceso de inducción de personal nuevo en la Entidad Pública, en Lima – Perú*. Lima- Peru: Universidad San Martin de Loyola. Obtenido de <https://repositorio.usil.edu.pe/server/api/core/bitstreams/cacd6e47-449a-4a68-aacb-ba5d501cec5b/content>
- Llagua, A., & Acosta, P. (2020). Riesgos laborales: análisis exploratorio de los indicadores más comunes en Ecuador. *Universidad Indoamérica*, 1-24.
- LOSEP. (2016). *Derechos de servidores publicos*. Quiro: Lexus.
- Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universidad Tecnológica Nacional. Obtenido de https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/8401/Chavez_Martinez_Renato.pdf?sequence=3&isAllowed=y
- Ministerio de Empleo y Seguridad Social de España. (2017). *Guía de ayuda para la valoración de las enfermedades profesionales*. Obtenido de Ministerio de Empleo y Seguridad Social de España: https://www.lmee-svmt.org/archivos/20170515_3409_ORI_GUIAVALORACIONEP.pdf
- Ministerio de Relaciones Laborales. (2015). *Programa de Seguridad y Salud Ocupacional*. Quito: MRL. Recuperado el 20 de Agosto de 2021, de <https://www.trabajo.gob.ec/seguridad-y-salud-en-el-trabajo/>
- Ministerio de Salud de Argentina. (2019). *Salud Ocupacional*. Obtenido de Ministerio de Salud de Argentina: <https://www.argentina.gob.ar/salud/ocupacional#:~:text=Funciones%20principales&text=Vigilancia%20de%20las%20condiciones%20y,enfermedades%20relacionadas%20con%20el%20trabajo>.
- Miranda, C. (2021). *Modelización de Series Temporales modelos clásicos y SARIMA*. Universidad de Granada. Obtenido de https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_MIRANDA_CHINLLI_CARLOS.pdf

- Moirón, R. (2021). *Análisis de datos: Estudio de la probabilidad de marcha de un empleado*. Universidad de Granada. Obtenido de https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_MoironSanchezRuben.pdf
- Mosquera, R., Parra, L., Ledesma, A. J., & Bonilla, H. F. (2021). Predicción de la accidentalidad laboral en la industria de pulpa y papel usando algoritmos de clasificación. *Información tecnológica*, 32(1), 133-142. doi:<https://dx.doi.org/10.4067/S0718-07642021000100133>
- One-class svm for anomaly detection. (2022). *One-class svm for anomaly detection*. Obtenido de <https://medium.com/grabngoinfo/one-class-svm-for-anomaly-detection-6c97fdd6d8af>
- Organizacion Iberoamericana de Seguridad Social. (2020). *Estrategia Iberoamericana de Seguridad y Salud en el Trabajo 2015-2020*. OISS. Obtenido de <http://www.relatos.org/documentos/SST.Iberoamerica.Doc2015.20.pdf>
- Ortega, J., Rodríguez, J., & Hernández, H. (2017). Importancia de la seguridad de los trabajadores en el cumplimiento de procesos. *Revista Academia & Derecho*, 8(14), 155-176.
- Ortuño, M. (2022). *Un modelo de regresion logistica para el analisis de los aspectos que influyen en la anulacion de polizas de seguros de automoviles*. Universidad de Granada. Obtenido de https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_ORTU%C3%91O_ROIG_MARIA.pdf
- Pérez, L., Pinzón, C., Hernández, N., González, L., Silva, C., Dugarte, J., . . . Ramírez, E. (2023). La importancia de la prevención dela enfermedad laboral en Colombia. *REDIPE*, 12(9), 1-18.
- RAE. (2022). *Factor de Riesgo*. Obtenido de RAE: <https://dpej.rae.es/lema/factor-de-riesgo#:~:text=Gral.,con%20riesgo%20de%20generar%20adicciones>.
- RAE. (2022). *Lesión*. Obtenido de RAE: <https://dpej.rae.es/lema/lesi%C3%B3n>
- RAE. (2022). *Riesgo Laboral*. Obtenido de RAE: <https://dpej.rae.es/lema/riesgo-laboral>
- Ramírez, A. V. (2012). Servicios de salud ocupacional. *Anales de la Facultad de Medicina*, 73(1). Obtenido de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1025-55832012000100012&lng=es&tlng=es
- Sánchez, J., Rodríguez, O., Sánchez, D., & Cuadrado, G. (2022). Discapacidad:definición,normativaycontexto. *Conciencia Digital*, 5(1.3), 234-247. doi:<https://doi.org/10.33262/concienciadigital.v5i1.3.2138>

- Sandoval, L. (2017). *Machine Learning algorythms for analysis and data prediction*. 2017. IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII). doi:<http://doi.org/10.1109/CONCAPAN.2017.8278511>
- Soto, L., & Melara, M. (2018). *RIESGOS LABORALES DEL PERSONAL DE ENFERMERÍA EN EL SERVICIO DEINFECTOLOGÍA DEL HOSPITAL NACIONAL DE NIÑOS BENJAMÍN BLOOM DE MAYO -SEPTIEMBRE DE 2017*. Obtenido de UNIVERSIDAD DE EL SALVADOR: <https://docs.bvsalud.org/biblioref/2021/02/1148066/237.pdf>
- Toro, J., Vega, V., & Romero, A. (2021). Los accidentes de trabajo y enfermedades profesionales y su aplicación en la justicia ordinaria. *13*(2), 357-362.
- Trabajo, M. d. (2017). *Instructivo para el cumplimineto de obligaciones de los empleados- Acuerdo Ministerial 135*.
- UESS. (2022). *Medicina del Trabajo- Noción en legislación de salud labora*.
- Van Vaerenbergh, S., & Santamaría, I. (2018). *Métodos kernel para clasificación*. Obtenido de https://gtas.unican.es/files/docencia/APS/apuntes/07_svm_kernel.pdf
- Vielma, G. J. (2021). *Todos los accidentes de trabajo y enfermedades ocupacionales son evitables*". Republica Dominicana: UNIBE.
- Villada, J., Castaño, L., & Laura, P. (2023). *Propuesta de diseño de un modelo matemático predictivo aplicado a la prevención de riesgos laborales en el sector de la construcción*. Universidad ECCI. Obtenido de <https://repositorio.ecci.edu.co/bitstream/handle/001/3311/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y>
- Villalobos, M. (2020). *Modelo predictivo de factores de riesgos laborales con uso de inteligencia artificial*. Chile: Superintendencia de Seguridad Social Santiago - Chile. Obtenido de https://www.suseso.cl/619/articles-672236_archivo_01.pdf

ANEXOS

ANEXO 1

CERTIFICADO DEL SISTEMA ANTIPLAGIO COMPILATIO

Quevedo, 25 de mayo del 2024

Ing. Byron Oviedo Bayas, PhD

Decano de Posgrado UTEQ

En su despacho,

Por medio de la presente se da a conocer que el proyecto de investigación titulado “**MODELO PREDICTIVO DE RIESGOS LABORALES EN EL GOBIERNO AUTÓNOMO DESCENTRALIZADO DE PICHINCHA**”, del estudiante **Jocelyne Giraldo Muñoz** en la **Maestría en Ciencia de Datos**, se obtuvo un 2% en la herramienta anti-similitud COMPILATIO; en consecuencia, el proyecto de investigación se encuentra listo para seguir con el proceso de defensa.

The image shows a screenshot of a plagiarism analysis certificate from the system COMPILATIO. The certificate is titled 'CERTIFICADO DE ANÁLISIS' and 'magister'. The document being analyzed is 'GIRALDO TESIS 2024'. The analysis results show a 2% similarity score, with a note that 0% similarity was found between the document and the sources mentioned, and 1% similarity was found between the document and the sources mentioned. The document is 1.26 MB in size and contains 12,460 words and 85,825 characters. The certificate was generated on 5/6/2024.

Nombre del documento:	Depositar:	Número de palabras:
GIRALDO TESIS 2024.docx	EDUARDO AMABLE SAMANIEGO MENA	12.460
ID del documento: 5cd67240bd725e0da3e05044412748c573b67d	Fecha de depósito: 5/6/2024	Número de caracteres: 85.825
Tamaño del documento original: 1,26 MB	Tipo de carga: interface	
	fecha de fin de análisis: 5/6/2024	

Particular que comunico para los fines pertinentes.



Ing. Eduardo Samaniego Mena, Msc.

Director Proyecto de Investigación

Anexo 2. Importación de librerías y carga de archivos .xlsx.

```
# Cargamos librerías
import os
import re
import numpy as np
import pandas as pd
from unidecode import unidecode
from datetime import datetime
import calendar

# Librerías para graficas
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import statsmodels.api as sm
import plotly.graph_objects as go
from scipy.cluster.hierarchy import dendrogram, linkage

# Librerías para google drive
from google.colab import drive

# Librerías para modelos de datos
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression, LinearRegression
from sklearn.metrics import confusion_matrix, classification_report, mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor

# Librerías para manejo de errores
import warnings
warnings.filterwarnings("ignore", category=UserWarning)
warnings.filterwarnings("ignore", category=DeprecationWarning)
warnings.simplefilter(action='ignore', category=FutureWarning)
```

Anexo 3. Archivos de Excel brindados por el GAD de Pichincha y archivos de unificados y resultados de predicciones.

Copia de ESTADÍSTICAS DE ACCIDENTABILIDAD 2015 nuevo.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2016.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2017.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2021.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2022 FINAL.xlsx
ESTADÍSTICA DE ACCIDENTABILIDAD 2023 CORTE AGOSTO ENV.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2015.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2019.xlsx
ESTADÍSTICAS DE ACCIDENTABILIDAD 2018.xlsx
columnas para uso final.xlsx
clean.xlsx
unificado.xlsx
dataFinal.xlsx
resultados_predicciones.xlsx

Anexo 3. Nuevo diccionario de datos para los nuevos encabezados.

```
[ ] # Crear un diccionario de datos con los nuevos encabezados
new_headers = {
    'Nº' : 'Numero',
    'amplia': 'Amplia',
    'MES' : 'Mes',
    'Día': 'Dia',
    'FECHA DEL \nACCIDENTE': 'Fecha_Accidente',
    'NOMBRE Y APELLIDOS': 'Nombre_Apellido',
    'EDAD': 'Edad',
    'CUMPLE': 'Cumple',
    'DESCRIPCIÓN DEL\n ACCIDENTE' : 'Descripcion_Accidente',
    'DIAGNÓSTICO': 'Diagnostico',
    'LUGAR\n ACCIDENTE' : 'Lugar_Accidente',
    'DIAS DE \nREPOSO': 'Dias_Reposo',
    'DIAS DE CARGO': 'Dias_Cargo',
    'FECHA DE REPOSO': 'Fecha_Reposo',
    'FECHA DE CONSULTA': 'Fecha_Consulta',
    'LUGAR DE \nTRABAJO' : 'Lugar_Trabajo',
    'FUNCIÓN DE \nACCIDENTADO' : 'Funcion_Accidentado',
    'PARTE DEL CUERPO AFECTADO': 'Parte_Cuerpo_Afectado',
    'TIPO DE \nINCAPACIDAD': 'Tipo_Incapacidad',
    'TIPO ACCIDENTE': 'Tipo_Accidente',
    'ETAPA DEL CASO': 'Etapa_Caso',
    'RESPONSABILIDAD \nPATRONAL' : 'Responsabilidad_Patronal',
    'INDEMNIZACIÓN': 'Indemnizacion',
    'REUBICACIÓN': 'Reubicacion',
    'REUBICACIÓN TEMPORAL': 'Reubicacion_Temporal',
    'REUBICACIÓN DEFINITIVA': 'Reubicacion_Definitiva',
    'CAUSALIDAD      1= Ac. Sub.      2= Con. Sub': 'Causalidad',
    'OBSERVACIONES': 'Observaciones',
    'OBSERVACIONES' : 'Observaciones',
    'TECNICO RESPO': 'Tecnico_Responsable',
    'RANGO EDAD': 'Rango_Edad',
    'año': 'anio_archivo'
}
```

Anexo 4. Diccionario para reemplazar palabras o sinónimos por un término estándar

```
reemplazos = {
    'ayudante de maquina': 'ayudante de maquinaria',
    'ayudante de maquina': 'ayudante de maquinaria',
    'operador de guillotina': 'operador de maquinaria',
    'operador de maquina': 'operador de maquinaria',
    'ayudante de mecanica': 'ayudante de maquinaria',
    'prensista-imprenta': 'operador de maquinaria',
    'chofer de volqueta': 'chofer',
    'ayudante de maquinaria excavadora': 'ayudante de maquinaria',
    'operador de plataforma': 'operador de maquinaria',
    'diseño y serigrafía': 'diseño y serigrafía',
    'operador minicargadora': 'operador de maquinaria',
    'ayudante de operador': 'ayudante de maquinaria',
    'supervisor de maquinaria pesada': 'supervisor',
    'operador de equipo pesado': 'operador de maquinaria',
    'ayudante de maquina': 'ayudante de maquinaria',
    'ayudante de trituradora': 'ayudante de maquinaria',
    'asistente administrativo 1': 'asistente administrativo',
    'asistente administrativo': 'asistente administrativo',
    'analista de talento humano': 'analista de recursos humanos',
    'tecnica agraria': 'técnico agrario',
    'administrativo/analista': 'analista administrativo',
    'tecnico': 'técnico',
    'ayudante de soldador': 'ayudante de soldadura',
    'conductor equipo liviano': 'conductor',
    'administrativo': 'asistente administrativo',
    'analista juridica': 'abogada',
    'conductor eq. liviano': 'conductor',
    'ab. ejecucion de proyectos y elaboracion de informes': 'analista administrativo',
    'ejecucion de procesos elaboracion de informes': 'analista administrativo',
    'auxiliar de ingenieria': 'auxiliar de ingeniería',
    'asistente de laboratorios de suelos': 'asistente de laboratorio',
    'operador': 'operador de maquinaria',
    'asesor economico': 'asesor económico',
    'jornalero': 'obrero',
    'guardia': 'guardia de seguridad',
    'personal tecnico': 'personal técnico',
    'auxiliar de ingenieria': 'auxiliar de ingeniería'
}
```

Anexo 5. Eliminar caracteres especiales, tildes, comas, saltos de línea y tabulaciones.

```
texto_limpio = re.sub(r'^\w\s', '', texto)
texto_limpio = re.sub(r'[ÁÁÁÁ]', 'A', texto_limpio)
texto_limpio = re.sub(r'[ÉÉÉÉ]', 'E', texto_limpio)
texto_limpio = re.sub(r'[ÍÍÍÍ]', 'I', texto_limpio)
texto_limpio = re.sub(r'[ÓÓÓÓ]', 'O', texto_limpio)
texto_limpio = re.sub(r'[ÚÚÚÚ]', 'U', texto_limpio)
texto_limpio = re.sub(r'[Ñ]', 'N', texto_limpio)
texto_limpio = re.sub(r'[.]', '', texto_limpio)
texto_limpio = re.sub(r'[\n\t]', ' ', texto_limpio)
# Eliminar espacios múltiples
texto_limpio = re.sub(r'\s+', ' ', texto_limpio).strip()
return texto_limpio
```

Anexo 6. Función para clasificar los diagnósticos.

```
def clasificar_accidente(diagnostico):
    if diagnostico is None or pd.isnull(diagnostico) or diagnostico == '' or diagnostico.strip() == '' or diagnostico == 'Sin Datos':
        return 'No identificado'
    diagnostico = diagnostico.lower()
    if 'fractura' in diagnostico:
        return 'Fractura'
    elif 'herida' in diagnostico or 'corte' in diagnostico:
        return 'Herida'
    elif 'trauma' in diagnostico or 'contusión' in diagnostico or 'esguince' in diagnostico or 'luxación' in diagnostico:
        return 'Traumatismo'
    elif 'intoxicación' in diagnostico or 'envenenamiento' in diagnostico:
        return 'Intoxicación'
    elif 'dolor' in diagnostico or 'lumbalgia' in diagnostico or 'cefalea' in diagnostico:
        return 'Dolor'
    elif 'lesión' in diagnostico:
        return 'Lesión'
    elif 'cervicalgia' in diagnostico or 'lumbalgia' in diagnostico or 'dorsalgia' in diagnostico:
        return 'Dolor de Espalda'
    elif 'conjuntivitis' in diagnostico:
        return 'Conjuntivitis'
    elif 'luxofractura' in diagnostico:
        return 'Luxofractura'
    elif 'tendinitis' in diagnostico:
        return 'Tendinitis'
    elif 'mordedura' in diagnostico:
        return 'Mordedura'
    else:
        return 'Otros'
```

Anexo 7. Función para clasificar las partes del cuerpo y así mismo se clasifico otras variables.

```
# Función para clasificar las partes del cuerpo
def clasificar_parte_cuerpo_accidente(parte_cuerpo):
    if parte_cuerpo is None or pd.isnull(parte_cuerpo) or parte_cuerpo == '' or parte_cuerpo.strip() == '' or parte_cuerpo == 'Sin Datos':
        return 'No identificado'
    parte_cuerpo = parte_cuerpo.lower()
    if 'cabeza' in parte_cuerpo or 'cráneo' in parte_cuerpo or 'ojos' in parte_cuerpo or 'ojo' in parte_cuerpo or 'nariz' in parte_cuerpo or 'boca' in parte_cuerpo or 'orejas' in parte_cuerpo or 'rostro':
        return 'Cabeza'
    elif 'cuello' in parte_cuerpo or 'cervicales' in parte_cuerpo:
        return 'Cuello'
    elif 'columna' in parte_cuerpo:
        return 'Columna'
    elif 'torax' in parte_cuerpo or 'tórax' in parte_cuerpo or 'pecho' in parte_cuerpo or 'costillas' in parte_cuerpo or 'costilla' in parte_cuerpo:
        return 'Tórax'
    elif 'abdomen' in parte_cuerpo or 'hipocondrio' in parte_cuerpo or 'pelvis' in parte_cuerpo:
        return 'Abdomen'
    elif 'espalda' in parte_cuerpo or 'dorso' in parte_cuerpo:
        return 'Espalda'
    elif 'brazo' in parte_cuerpo or 'biceps' in parte_cuerpo or 'codo' in parte_cuerpo or 'antebrazo' in parte_cuerpo or 'hombro' in parte_cuerpo or 'mano' in parte_cuerpo or 'muñeca' in parte_cuerpo:
        return 'Brazo'
    elif 'dedo' in parte_cuerpo or 'falange' in parte_cuerpo or 'meñique' in parte_cuerpo or 'pulgar' in parte_cuerpo or 'manos' in parte_cuerpo:
        return 'Mano'
    elif 'pierna' in parte_cuerpo or 'muslo' in parte_cuerpo or 'rodilla' in parte_cuerpo or 'tibia' in parte_cuerpo or 'fémur' in parte_cuerpo or 'tobillo' in parte_cuerpo or 'pies' in parte_cuerpo or 'pie' in parte_cuerpo:
        return 'Pierna'
    elif 'hueso' in parte_cuerpo or 'fractura' in parte_cuerpo or 'clavícula' in parte_cuerpo or 'fémur' in parte_cuerpo or 'húmero' in parte_cuerpo or 'cadera' in parte_cuerpo or 'húmero' in parte_cuerpo:
        return 'Hueso'
    elif 'lesión' in parte_cuerpo or 'herida' in parte_cuerpo or 'contusión' in parte_cuerpo or 'esguince' in parte_cuerpo or 'luxación' in parte_cuerpo or 'hematoma' in parte_cuerpo or 'hemorragia' in parte_cuerpo:
        return 'Lesión'
    elif 'respiratorio' in parte_cuerpo or 'pulmones' in parte_cuerpo or 'nariz' in parte_cuerpo:
        return 'Sistema respiratorio'
    elif 'miembros' in parte_cuerpo or 'múltiples' in parte_cuerpo or 'múltiples' in parte_cuerpo or 'politraumatismos' in parte_cuerpo or 'varias' in parte_cuerpo:
        return 'Lesiones múltiples'
    else:
        return 'Otros'
```

Anexo 8. Limpieza de la base de datos del año 2015, así mismo se realizó las limpiezas de las otras bases de datos.

```
[ ] # Columna Fecha de Accidente
# Reemplamos valores de texto para dejar en un formato para convertir en tipo fecha
df_2015['Fecha_Accidente'] = df_2015['Fecha_Accidente'].str.upper()
df_2015['Fecha_Accidente'] = df_2015['Fecha_Accidente'].str.replace(' DE ', '/')
df_2015['Fecha_Accidente'] = df_2015['Fecha_Accidente'].str.replace(' DEL ', '/')
# Usamos el diccionario de nombres de mes para crear reemplazar en las fechas
month_map_tuple = tuple(month_map.items())
for old, new in month_map.items():
    df_2015['Fecha_Accidente'] = df_2015['Fecha_Accidente'].str.replace(old, new, regex=False)
# Transformamos la columna a fecha
df_2015['Fecha_Accidente'] = pd.to_datetime(df_2015['Fecha_Accidente'])
```

```
[ ] # Columna Edad
df_2015['Edad'] = df_2015['Edad'].str.replace(r' años| de edad', '', regex=True)
# Transformamos la columna a entero
df_2015['Edad'].fillna(0, inplace=True)
df_2015['Edad'] = df_2015['Edad'].astype(int)
```

```
[ ] # Columna Dias de Reposo
df_2015['Dias_Reposo'] = df_2015['Dias_Reposo'].str.replace(r'(continua|días|día)', '', regex=True)
df_2015['Dias_Reposo'] = df_2015['Dias_Reposo'].str.replace(' ', '')
df_2015['Dias_Reposo'] = df_2015['Dias_Reposo'].str.replace('.', '')
df_2015['Dias_Reposo'] = df_2015['Dias_Reposo'].str.replace('(', '')
df_2015['Dias_Reposo'] = df_2015['Dias_Reposo'].str.replace(')', '')

# Transformamos la columna a entero
df_2015['Dias_Reposo'].fillna(0, inplace=True)
df_2015['Dias_Reposo'] = df_2015['Dias_Reposo'].astype(int)
```

```
[ ] # Aplicar la función a la columna 'Edad' y crear una nueva columna 'Rango_Edad'
df_2015['Rango_Edad'] = df_2015['Edad'].apply(rango_edad)
```

```
[ ] # Aplicar la función a la columna Fecha_Accidente y obtenemos el nombre del día
df_2015['Dia'] = df_2015['Fecha_Accidente'].dt.dayofweek.apply(get_dia_semana)
```

```
[ ] # Aplicar la función a la columna Fecha_Accidente y obtenemos el nombre del mes
df_2015['Mes'] = df_2015['Fecha_Accidente'].dt.month.apply(get_mes_espanol_with_null_check)
```

Anexo 9. Unificación de dataframes, limpieza y tratamiento.

```
# Unir los dataframes por coincidencia de columnas
df_unificado = pd.merge(df_2015, df_2016, on=['Numero', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Fecha_Reposo', 'Lugar_Trabajo', 'Funcion_A
df_unificado = pd.merge(df_unificado, df_2017, on=['Numero', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Fecha_Reposo', 'Lugar_Trabajo', 'Func
df_unificado = pd.merge(df_unificado, df_2018, on=['Numero', 'Amplia', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Cumple', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Dias_Cargo', 'Fech
df_unificado = pd.merge(df_unificado, df_2019, on=['Numero', 'Amplia', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Cumple', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Dias_Cargo', 'Fech
df_unificado = pd.merge(df_unificado, df_2020, on=['Numero', 'Amplia', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Cumple', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Dias_Cargo', 'Fech
df_unificado = pd.merge(df_unificado, df_2021, on=['Numero', 'Amplia', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Cumple', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Dias_Cargo', 'Fech
df_unificado = pd.merge(df_unificado, df_2022, on=['Numero', 'Amplia', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Cumple', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Dias_Cargo', 'Fech
df_2023['Numero'] = df_2023['Numero'].astype('object')
df_2023['Amplia'] = df_2023['Amplia'].astype('object')
df_2023['Responsabilidad_Patronal'] = df_2023['Responsabilidad_Patronal'].astype('object')
df_2023['Indemnizacion'] = df_2023['Indemnizacion'].astype('object')
df_2023['Reubicacion'] = df_2023['Reubicacion'].astype('object')
df_2023['Causalidad'] = df_2023['Causalidad'].astype('object')
df_2023['Observaciones'] = df_2023['Observaciones'].astype('object')
df_unificado = pd.merge(df_unificado, df_2023, on=['Numero', 'Amplia', 'Mes', 'Dia', 'Fecha_Accidente', 'Nombre_Apellido', 'Edad', 'Cumple', 'Descripcion_Accidente', 'Diagnostico', 'Lugar_Accidente', 'Dias_Reposo', 'Dias_Cargo', 'Fech
```

Anexo 10. Exportar el dataframe a un archivo.xlsx.

```
df_unificado.to_excel('unificado.xlsx', sheet_name='Hoja1', index=False)
```

Anexo 11. Valores nulos en cada columna.

	Columna	Porcentaje de Nulos
0	anio_archivo	0.00
1	Mes	6.69
2	Dia	11.72
3	Fecha_Accidente	11.72
4	Nombre_Apellido	13.39
5	Edad	0.00
6	Rango_Edad	1.67
7	Descripcion_Accidente	14.64
8	Diagnostico	14.64
9	Lugar_Accidente	15.48
10	Dias_Reposo	0.00
11	Lugar_Trabajo	15.48
12	Funcion_Accidentado	15.06
13	Parte_Cuerpo_Afectado	19.25
14	Tipo_Incapacidad	15.48
15	Tipo_Accidente	14.64
16	Etapas_Caso	16.32

Anexo 12. Clasificación de datos para el reemplazo por una variable” Sin datos”.

```
df_clean = df_clean.copy()
# Convertir la columna 'Mes' a tipo str y reemplazar NaN por 'Sin Datos'
df_clean.loc[:, 'Mes'] = df_clean['Mes'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Dia'] = df_clean['Dia'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Rango_Edad'] = df_clean['Rango_Edad'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Diagnostico'] = df_clean['Diagnostico'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Lugar_Trabajo'] = df_clean['Lugar_Trabajo'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Lugar_Accidente'] = df_clean['Lugar_Accidente'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Funcion_Accidentado'] = df_clean['Funcion_Accidentado'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Parte_Cuerpo_Afectado'] = df_clean['Parte_Cuerpo_Afectado'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Tipo_Incapacidad'] = df_clean['Tipo_Incapacidad'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Tipo_Accidente'] = df_clean['Tipo_Accidente'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Etapas_Caso'] = df_clean['Etapas_Caso'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Descripcion_Accidente'] = df_clean['Descripcion_Accidente'].astype(str).replace({'None': 'Sin Datos', 'nan': 'Sin Datos', 'NaN': 'Sin Datos'})
df_clean.loc[:, 'Edad'] = pd.to_numeric(df_clean['Edad'], errors='coerce')
df_clean.loc[:, 'Dias_Reposo'] = df_clean['Dias_Reposo'].astype(float)

# Eliminar las filas con valores nulos en 'Edad'
df_clean = df_clean.dropna(subset=['Edad'])
```

Anexo 13. Porcentaje de valores nulos en cada columna.

	Columna	Porcentaje de Nulos
0	anio_archivo	0.0
1	Mes	0.0
2	Dia	0.0
3	Fecha_Accidente	0.0
4	Nombre_Apellido	0.0
5	Edad	0.0
6	Rango_Edad	0.0
7	Descripcion_Accidente	0.0
8	Diagnostico	0.0
9	Lugar_Accidente	0.0
10	Dias_Reposo	0.0
11	Lugar_Trabajo	0.0
12	Funcion_Accidentado	0.0
13	Parte_Cuerpo_Afectado	0.0
14	Tipo_Incapacidad	0.0
15	Tipo_Accidente	0.0
16	Etapas_Caso	0.0
17	Clasificacion_Diagnostico	0.0
18	Clasificacion_Lugar_Accidente	0.0
19	Clasificacion_Lugar_Trabajo	0.0
20	Clasificacion_Parte_Cuerpo	0.0
21	Clasificacion_Funcion_Accidentado	0.0

Anexo 14. Modelos utilizados para predecir la cantidad de accidentes en los próximos años.

```
models = {  
    "Linear Regression": LinearRegression(),  
    "Random Forest": RandomForestRegressor(),  
    "Decision Tree": DecisionTreeRegressor()  
}  
  
# Iterar sobre los modelos  
for name, model in models.items():  
    # Entrenar el modelo  
    model.fit(X_entrenamiento, y_entrenamiento)  
    # Realizar predicciones  
    predicciones = model.predict(X_prueba)  
    # Calcular el error cuadrático medio  
    mse = mean_squared_error(y_prueba, predicciones)  
    # Calcular el coeficiente de determinación (R^2)  
    r2 = r2_score(y_prueba, predicciones)  
    # Imprimir el nombre del modelo y sus métricas  
    print(f"Modelo: {name}")  
    print(f"Error Cuadrático Medio (MSE): {mse}")  
    print(f"Coeficiente de Determinación (R^2): {r2}")  
    print("\n")
```

Anexo 15. Modelo de Regresión Lineal mejorado.

```
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
```

```
# Estandarización de las características  
scaler = StandardScaler()  
X_entrenamiento_scaled = scaler.fit_transform(X_entrenamiento)  
X_prueba_scaled = scaler.transform(X_prueba)  
  
# Creación de características polinomiales  
poly = PolynomialFeatures(degree=2, include_bias=False)  
X_entrenamiento_poly = poly.fit_transform(X_entrenamiento_scaled)  
X_prueba_poly = poly.transform(X_prueba_scaled)
```

```
from sklearn.linear_model import Ridge, Lasso  
from sklearn.model_selection import GridSearchCV  
  
# Definición del modelo Ridge con búsqueda de hiperparámetros  
ridge = Ridge()  
param_grid = {'alpha': [0.001, 0.01, 0.1, 1, 10, 100]}  
grid_search_ridge = GridSearchCV(ridge, param_grid, cv=5, scoring='neg_mean_squared_error')  
grid_search_ridge.fit(X_entrenamiento_poly, y_entrenamiento)  
  
# Definición del modelo Lasso con búsqueda de hiperparámetros  
lasso = Lasso()  
grid_search_lasso = GridSearchCV(lasso, param_grid, cv=5, scoring='neg_mean_squared_error')  
grid_search_lasso.fit(X_entrenamiento_poly, y_entrenamiento)  
  
# Selecciona el modelo y los hiperparámetros que mejor funcionen  
modelo_optimizado = grid_search_ridge.best_estimator_ if -grid_search_ridge.best_score_ < -grid_search_lasso.best_score_ else grid_search_lasso.best_estimator_
```

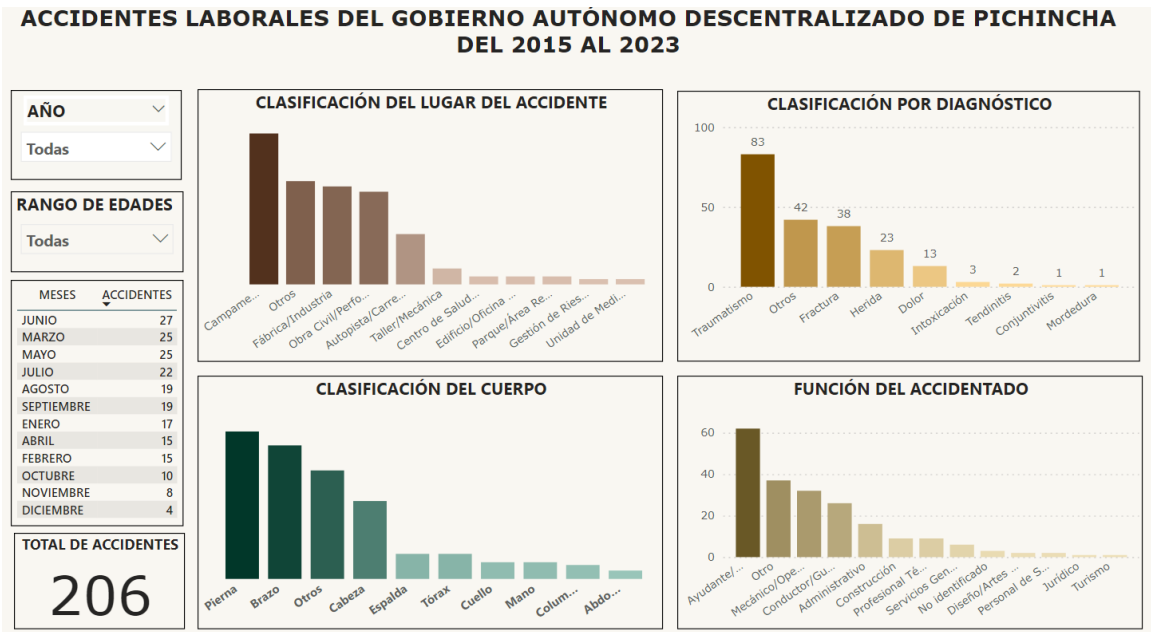
```
predicciones_optimizadas = modelo_optimizado.predict(X_prueba_poly)  
mse_optimizado = mean_squared_error(y_prueba, predicciones_optimizadas)  
r2_optimizado = r2_score(y_prueba, predicciones_optimizadas)  
  
print("Modelo Optimizado:")  
print(f"Error Cuadrático Medio (MSE): {mse_optimizado}")  
print(f"Coeficiente de Determinación (R^2): {r2_optimizado}")
```

```
Modelo Optimizado:  
Error Cuadrático Medio (MSE): 2.607329845976161  
Coeficiente de Determinación (R^2): 0.7102966837804265
```

Anexo 16. Datos semestrales desde el 2015 con datos entrenamiento y de predicciones.

	Semestre	Número de Accidentes
0	2015 - 1	14.000000
1	2015 - 2	17.000000
2	2016 - 1	24.000000
3	2016 - 2	13.000000
4	2017 - 1	16.000000
5	2017 - 2	10.000000
6	2018 - 1	23.000000
7	2018 - 2	12.000000
8	2019 - 1	14.000000
9	2019 - 2	7.000000
10	2021 - 1	14.000000
11	2021 - 2	7.000000
12	2022 - 1	7.000000
13	2022 - 2	10.000000
14	2023 - 1	12.000000
15	2023 - 2	6.000000
16	2024 - 1	12.483811
17	2024 - 2	7.233811
18	2025 - 1	15.505042
19	2025 - 2	6.255042
20	2026 - 1	9.526274
21	2026 - 2	5.276274

Anexo 17. Dashboard de los accidentes del 2015 al 2023.



Anexo 18. Dashboard de la tendencia de accidentes para 2024 al 2026.

