

03

MODELO PREDICTIVO

**PARA DETERMINAR PATRONES Y TENDENCIAS QUE
INFLUYEN EN LOS ACCIDENTES DE TRABAJO**

MODELO PREDICTIVO

PARA DETERMINAR PATRONES Y TENDENCIAS QUE INFLUYEN EN LOS ACCIDENTES DE TRABAJO

PREDICTIVE MODEL TO DETERMINE PATTERNS AND TRENDS THAT INFLUENCE WORK ACCIDENTS

Erika Zamora-Cevallos¹

E-mail: eazamorac@ube.edu.ec

ORCID: <https://orcid.org/0000-0002-2362-740X>

Leyda Zavala-Arteaga¹

E-mail: ljzavalaa@ube.edu.ec

ORCID: <https://orcid.org/0009-0004-7517-4432>

Byron Oviedo-Bayas²

E-mail: boviedo@uteq.edu.ec

ORCID: <https://orcid.org/0000-0002-5366-5917>

Mireya Stefania Zúñiga-Delgado¹

E-mail: mszuñigad@ube.edu.ec

ORCID: <https://orcid.org/0000-0002-4458-5771>

¹ Universidad Bolivariana del Ecuador. Ecuador.

² Universidad Técnica Estatal de Quevedo. Ecuador.

Cita sugerida (APA, séptima edición)

Zamora-Cevallos, E., Zavala-Arteaga, L., Oviedo-Bayas, B., & Zúñiga-Delgado, M. E. (2024). Modelo predictivo para determinar patrones y tendencias que influyen en los accidentes de trabajo. *Revista Metropolitana de Ciencias Aplicadas*, 7(S2), 29-41.

RESUMEN

La seguridad en el trabajo es preocupación de toda organización que aspire la excelencia, no solo porque garantiza un espacio de trabajo libre de accidentes, sino también, porque aumenta la confianza, el compromiso y la motivación d los empleados. El uso de herramientas avanzadas y técnicas analíticas para explorar conjuntos de datos grandes y pequeños puede ayudar a identificar estos problemas con suficiente antelación para tomar decisiones oportunas. A pesar de la preocupación por los accidentes, hasta la fecha se han realizado pocos análisis para identificar tendencias o patrones específicos, por lo que este estudio se centra en una base de datos que contiene información sobre los accidentes ocurridos entre 2015 y 2023 en una empresa del Estado. El objetivo principal es analizar las causas de los accidentes. Se utilizó algoritmos de aprendizaje automático y técnicas de ciencia de datos para identificar patrones y tendencias en accidentes laborales. Luego, los datos se clasifican en detalle para comprender mejor el comportamiento cambiante basándose en la regresión lineal. Después de analizar las previsiones se logró determinar que estas fueron muy consistentes con los resultados reales, lo que confirma la precisión del modelo utilizado.

Palabras clave:

Accidentes, ciencia de datos, modelos predictivos, minería de datos, seguridad.

ABSTRACT

Safety at work is a concern for any organization that aspires to excellence, not only because it guarantees an accident-free workspace, but also because it increases the confidence, commitment, and motivation of employees. Using advanced tools and analytical techniques to explore large and small data sets can help identify these issues early enough to make timely decisions. Despite concerns about accidents, few analyzes have been conducted to date to identify specific trends or patterns, so this study focuses on a database containing information on accidents that occurred between 2015 and 2023 at a company. of the State. The main objective is to analyze the causes of accidents. Machine learning algorithms and data science techniques were used to identify patterns and trends in workplace accidents. The data is then classified in detail to better understand the changing behavior based on linear regression. After analyzing the forecasts, it was determined that they were very consistent with the actual results, which confirms the precision of the model used.

Keywords:

Accidents, data science, predictive models, data mining, security.

INTRODUCCIÓN

Los accidentes laborales están reconocidos como uno de los mayores problemas de salud pública en todo el mundo, al menos un trabajador muere como resultado de un accidente laboral cada tres minutos. Además, las lesiones que sufren, que no ponen en peligro sus vidas, a veces resultan en una discapacidad permanente. Según Soto & Melara (2018), se conoce que cada año mueren 180 mil trabajadores y 110 millones sufren de lesiones de diferente gravedad. Lo que genera altos costos tanto para las empresas como para el sistema de salud pública, por lo cual se considera como una problemática.

Es importante destacar que las indemnizaciones resultantes de estos incidentes según el artículo de Toro et al. (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. La cobertura de salud y seguridad laboral varía en distintos países, identificando que en algunos lugares existe una escasa o amplia cobertura. Sin embargo, es necesario que se implemente una cultura de salud y seguridad laboral para prevenir el incremento de accidentes de trabajo (Ampuero et al., 2018; Arango et al., 2020).

Los accidentes de trabajo y las enfermedades profesionales se originan dentro de las empresas y organizaciones donde no se han generado políticas y acciones orientadas a la prevención de riesgos laborales. Además, Candonga & Samaniego (2021), indican que en estas 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.

Por lo tanto, los riesgos de trabajo son elementos que se encuentran con frecuencia dentro de los lugares de trabajo. Por esto es recomendable que se generen sistemas o planes de prevención de accidentes y enfermedades laborales, para evitar pérdidas económicas y productivas a las diferentes empresas.

En base a lo mencionado anteriormente, este estudio tiene como objetivo desarrollar un modelo que pueda predecir los riesgos de trabajo en el contexto ecuatoriano utilizando datos recolectados desde 2015 hasta agosto del 2023. El modelo se creó identificando patrones y tendencias

significativas, lo que sirvió como base sólida para su implementación.

MATERIALES Y MÉTODOS

La presente investigación es 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.

Para encontrar las relaciones entre las variables del conjunto de datos y accidentes, 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 estas en los riesgos laborales. Aunque el objetivo principal es generar un modelo predictivo para caracterizar y obtener las tendencias de accidentes.

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

- Análisis de Regresión
- Análisis de Supervivencia
- Análisis de Aprendizaje Automático
- Análisis de Agrupamiento

Se utilizaron diversas herramientas y plataformas especializadas para llevar a cabo análisis avanzados de datos, entre las que podemos indicar:

- Google Colab
- Weka

Luego, tocó elaborar el marco teórico, mismo que 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.

Posteriormente, se realizó la recopilación de los datos para esta investigación, esta se llevó a cabo a través de la colaboración directa de los involucrados, que facilitó el acceso a la información necesaria sobre riesgos de trabajo. El conjunto de datos utilizado se preparó a partir de 8 bases de datos en el que consta 206 registros. 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	Mes que ocurrió el accidente.
Nº	Texto	Número del registro del accidente.
Día	Texto	Día específico que ocurrió el accidente.

Atributo	Tipo	Descripción
Fecha de accidente	Fecha	Fecha exacta que ocurrió el accidente.
Nombres	Texto	Nombre del trabajador afectado.
Edad	Texto	Edad del trabajador afectado.
Descripción del accidente	Texto	Cómo ocurrió el accidente.
Diagnóstico	Texto	Diagnóstico médico 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.
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.

Luego, se realizó el procesamiento y análisis de los datos, para lo cual 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 (Figura 1).



Figura 1. Diagrama KDD, en el que se detalla se detalla las etapas del proceso de investigación, centradas en la preparación y selección de los datos relevantes para el análisis de riesgos de trabajo.

Etapa 1. Preprocesamiento y transformación

En la primera fase del proceso de preparación y transformación de datos, se inició con la selección y recopilación de información relevante para el estudio. Esta etapa implicó identificar las fuentes de datos pertinentes y obtener los conjuntos de datos necesarios. Además, se realizó una exploración inicial de los datos para comprender su estructura y calidad. Durante este proceso, se encontraron datos incompletos, que requirieron ser eliminados en etapas posteriores. Para evaluar la integridad de los datos, se calculó el porcentaje de valores nulos en cada columna del DataFrame, lo

que proporcionó información crucial sobre la completitud de los datos y resaltó las áreas que podrían necesitar imputación o eliminación de valores faltantes.

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. Esta etapa involucró la identificación y elección de los conjuntos de datos más significativos para el análisis de los riesgos de trabajo. Se tomó decisiones informadas sobre qué variables y atributos como edad, función, diagnóstico, días de reposo, entre otras, eran las principales para cumplir los objetivos de investigación, asegurando así un enfoque centrado.

Etapa 3: Minería de datos

Culminadas las dos etapas anteriores se procedió 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 de trabajo. Se emplearon los métodos de clasificación, regresión, árboles de decisiones, entre otros, con el fin de extraer información que permitió comprender mejor los accidentes laborales, los instrumentos o herramientas utilizadas.

Etapa 4. Interpretación y análisis

En esta etapa se realizó un proceso de analizar y predecir la cantidad de accidentes de trabajo 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. A continuación, se evaluaron tres modelos de regresión: Regresión Lineal, Random Forest y Árbol de Decisión. Cada modelo se entrenó con los datos de entrenamiento y se utilizan para predecir el número de accidentes en los semestres del conjunto de prueba. Posteriormente, se calcularon métricas de desempeño como el Error Cuadrático Medio y el Coeficiente de Determinación.

Se repite este análisis para calcular el modelo que 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

RESULTADOS Y DISCUSIÓN

Examinar el conjunto de datos 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. Posteriormente se realizó la unificación a una nueva DataFrame, para elaborar una gráfica de cantidades faltantes por columnas (Figura 2).

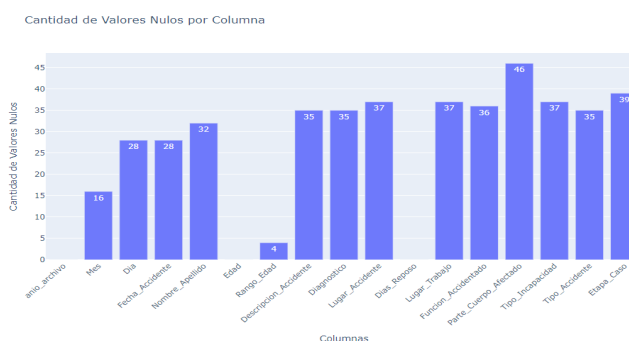


Figura 2. Cantidad de valores Nulos por columnas.

Siguiendo con la metodología se realizó el proceso de transformación de los datos para identificar los posibles valores que cada variable podría tener. 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 clave 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

```

Data columns (total 22 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   anio_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)

```

Figura 3. Conjunto de datos después del procesamiento y transformación.

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

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

Los resultados detallados de la selección de atributos revelan las características más importantes que tiene relación con el número de días de reposo. Este enfoque permitió organizar los atributos según su relación con la variable objetivo y evaluando cómo contribuyen a disminuir el error cuadrático medio (RMSE) en las predicciones. En todos 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. 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.

Para un segundo análisis de selección de atributos, se utilizó un enfoque de aprendizaje basado en el clasificador LinearRegression. En este análisis 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.

En el tercer análisis de selección de atributos, se utilizó un enfoque de aprendizaje basado en el clasificador IBk (k vecinos más cercanos). 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.

En un cuarto análisis de selección de atributos, se utilizó un enfoque de aprendizaje que se basa en el clasificador KStar. Ordenando los atributos ordenados se encontró que la variable Clasificacion_Diagnostico aún mantiene la mayor correlación.

Luego, en un quinto análisis de selección de atributos, se utilizó un enfoque de aprendizaje basado en el clasificador LWL. A continuación, se presentan los atributos ordenados según su importancia. Una vez más, se destaca la Clasificacion_Diagnostico, lo que subraya su relevancia en varios modelos.

En este sexto análisis de selección de atributos, se utilizó un enfoque de aprendizaje que se basa en el clasificador DecisionTable. Se observó que el atributo Fecha_Accidente tiene un impacto negativo, lo que sugiere que no está relacionado con factores estacionales (Figura 3).

En el 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. A pesar de su efecto negativo, se sugiere que el atributo Clasificacion_Funcion_Accidentado podría estar vinculado al diagnóstico y tener un impacto en el tiempo de reposo. Para un octavo análisis de selección de características, se utilizó un enfoque de aprendizaje basado en el clasificador DecisionStump. Como resultado se obtiene que el atributo del Mes puede estar vinculado e impactar en la recuperación.

Se analizaron 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. A continuación, se presenta la tabla 2 con los cinco aspectos más significativos, basada en los diez análisis previamente realizados.

Tabla 2. Cinco atributos más significativos.

Algoritmo utilizado	Primer atributo	Segundo atributo	Tercer atributo	Cuarto atributo	Quinto atributo
Gaussian Proces- ses	Clasificacion_ Diagnostico	Clasificacion_ Lugar_Trabajo	Mes	Edad	Fecha_Accidente
Linear Regression	Mes	Edad	Tipo_Incapacidad	Tipo_Accidente	Fecha_Accidente
IBk	Clasificacion_ Diagnostico	Tipo_Accidente	Mes	Edad	Clasificacion_Lu- gar_Trabajo
KStar	Mes	Clasificacion_ Lugar_Trabajo	Clasificacion_ Lugar_Accidente	Edad	Etapas_Caso
LWL	Clasificacion_ Diagnostico	Año_Archivo	Mes	Clasificacion_Lu- gar_Accidente	Clasificacion_Fun- cion_Accidentado
DecisionTable	Clasificacion_ Diagnostico	Fecha_Accidente	Clasificacion_Fun- cion_Accidentado	Clasificacion_Lu- gar_Accidente	Mes
M5Rules	Clasificacion_ Diagnostico	Clasificacion_Fun- cion_Accidentado	Clasificacion_ Lugar_Trabajo	Mes	Año_Archivo
DecisionStump	Clasificacion_ Diagnostico	Mes	Tipo_Incapacidad	Fecha_Accidente	Edad
RandomForest	Clasificacion_ Diagnostico	Clasificacion_Fun- cion_Accidentado	Mes	Fecha_Accidente	Clasificacion_Lu- gar_Accidente
RandomTree	Clasificacion_ Diagnostico	Mes	Clasificacion_Fun- cion_Accidentado	Clasificacion_Lu- gar_Accidente	Tipo_Accidente

Como se observa en la tabla 2, 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. En la tabla 3, se 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. La relación entre días de reposo y la clasificación de diagnóstico permite determinar que, 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 otros.

Tabla 3. Error cuadrático de las variables.

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
DecisionTable	0.786	0	0	0	-0.781	0	0	0	0	0	0	0	0	0	0
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
DecisionStump	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
RandomForest	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

El grupo de edad de 31 a 40 años se destaca con más accidentes laborales y con más días de reposo registrando un total de 51 accidentes según los datos proporcionados (Figura 4). 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.

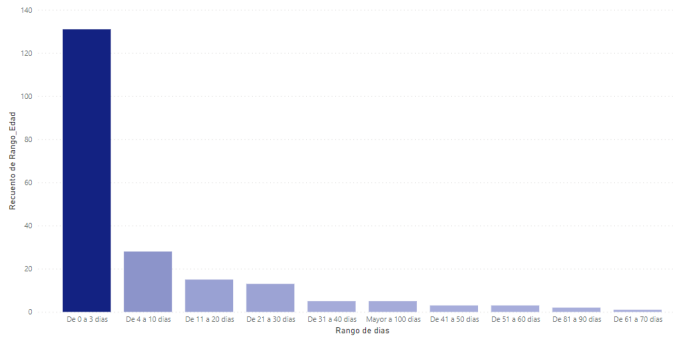


Figura 5. Gráfico de barras de cantidad de días de reposo.

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 Figura 6. Además, se debería profundizar en las circunstancias que rodean estos meses para realizar una planificación de estrategias de prevención.

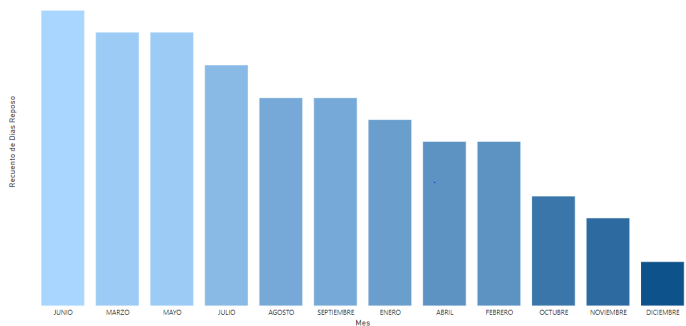


Figura 6. Gráfico de barras entre días de reposo y mes.

El lugar de trabajo más frecuente relacionado con los días de reposo según los accidentes de trabajo que han existido es otro análisis que permitió determinar 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 con mayores días de descanso. (Figura 7).

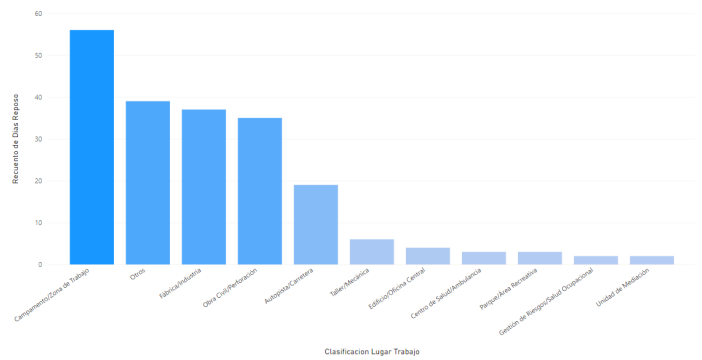


Figura 7. Gráfico de barras entre días de reposo y lugar de trabajo.

Tendencia semestral de accidentes laborales

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.

A diferencia, al analizar el modelo 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.

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 figuras. 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.

La comparación entre varios modelos de regresión muestra resultados importantes sobre su capacidad predictiva en un entorno particular. Se presentan los descubrimientos en la tabla 4.

Tabla 4. Comparación de modelos.

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

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 5 las predicciones realizadas para los años 2024, 2025 y 2026.

Tabla 5. Predicciones de los 3 años siguientes.

Semestre	Número de accidentes	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

En la figura 8 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.

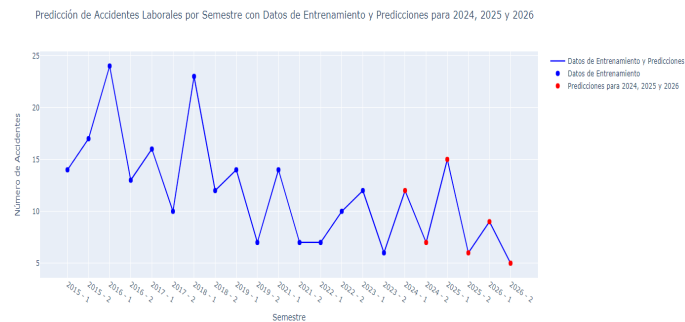


Figura 8. Predicción de accidentes por semestre y predicciones para 2024, 2025 y 2026.

Tendencia de accidentes laborales del 2024

En la Figura 9, centrado en el primer semestre del año, indica que entre enero y junio se registran un total de 12 accidentes en el primer trimestre y 7 en el segundo trimestre.

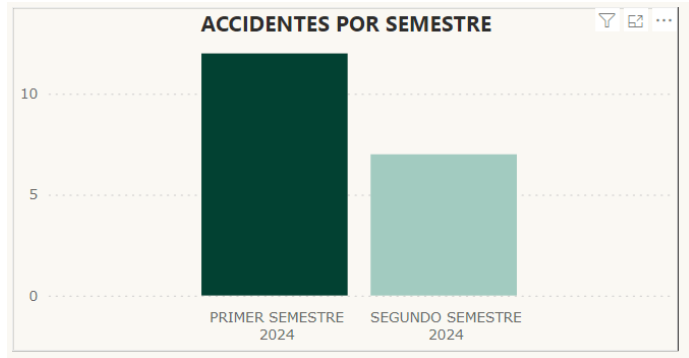


Figura 9. Predicción de accidentes por semestre para 2024.

En la Figura 10, se detallan los meses específicos dentro de los semestres donde ocurrieron los accidentes laborales durante el año 2024. Los meses más críticos en términos de incidentes resultaron ser marzo, mayo y junio, mostrando un aumento significativo comparado con otros meses del año.

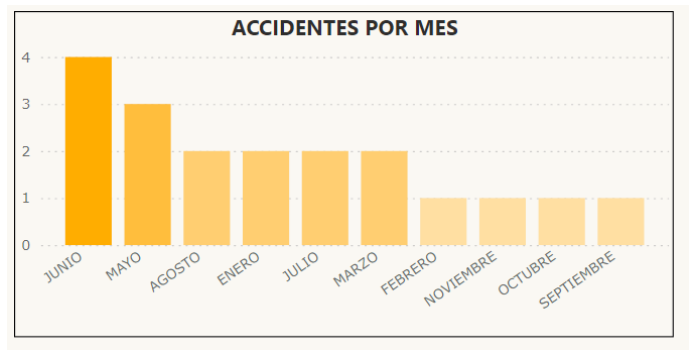


Figura 10. Predicción de accidentes por mes para 2024.

A continuación, la Figura 11 brinda una visión detallada sobre la distribución de los accidentes según el rango de edad de las personas afectadas. Se destaca una concentración importante de incidentes entre los grupos etarios de 31 a 40 años y 21 a 30 años, lo que sugiere la necesidad urgente de implementar medidas preventivas y protocolos específicos enfocados en estos segmentos particulares dentro del ámbito laboral.

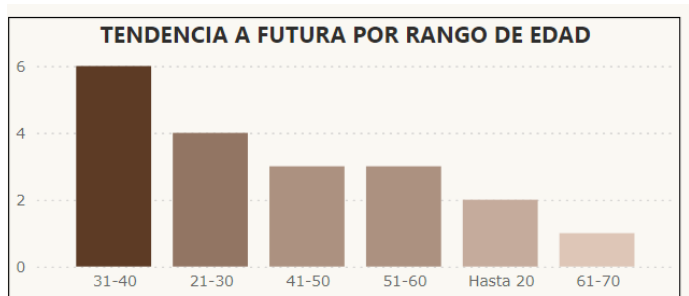


Figura 11. Predicción de accidentes por rango de edad para 2024.

Tendencia de accidentes laborales del 2025

En la Figura 12, que se muestra a continuación, se presenta una descripción detallada de las previsiones de accidentes de trabajo 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 brinda una perspectiva fundamental para identificar posibles áreas de riesgo y diseñar estrategias preventivas adecuadas.

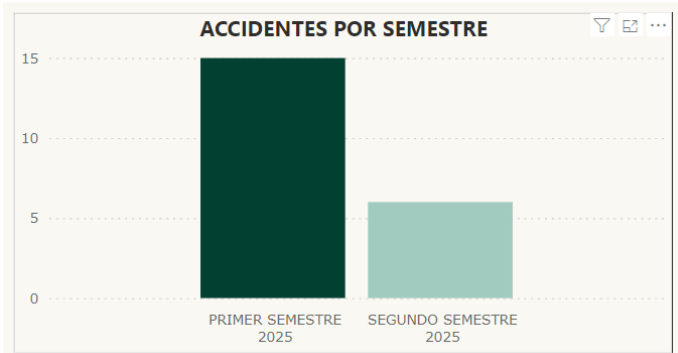


Figura 12. Predicción de accidentes por semestre para 2025.

En la Figura 13, se detallan los meses por semestre en los que se pronostica que ocurrirán accidentes de trabajo en el año 2024. Los meses críticos en términos de incidentes son junio, octubre y septiembre, seguidos de agosto. Esta representación temporal facilita la comprensión de los patrones estacionales de accidentes y puede guiar la asignación de recursos para intervenciones preventivas específicas.

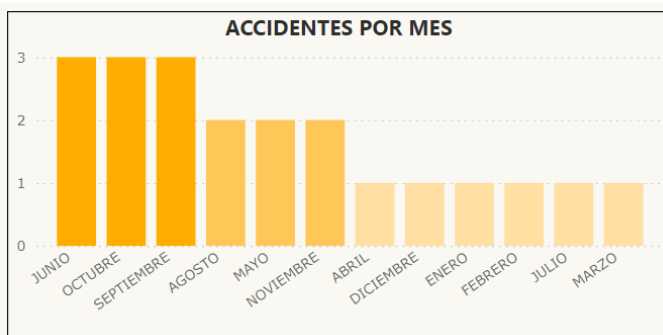


Figura 13. Predicción de accidentes por mes para 2025.

Por último, en la Figura 14 se muestra la franja etaria de las personas afectadas por los accidentes de trabajo. Se observa 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.

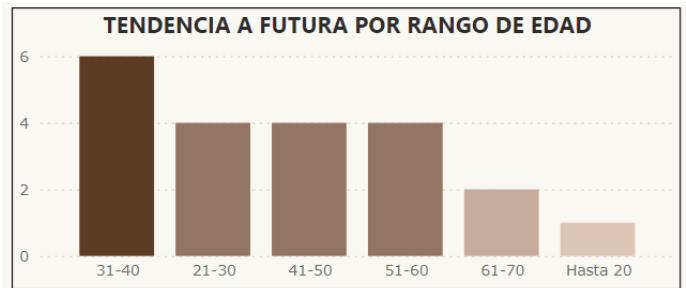


Figura 14. Predicción de accidentes por rango de edad para 2025.

Tendencia de accidentes laborales del 2026

Las proyecciones de accidentes de trabajo para el año 2026 estimadas son en un total de 9.52 accidentes durante el primer semestre y 5.27 durante el segundo semestre del año mencionado.

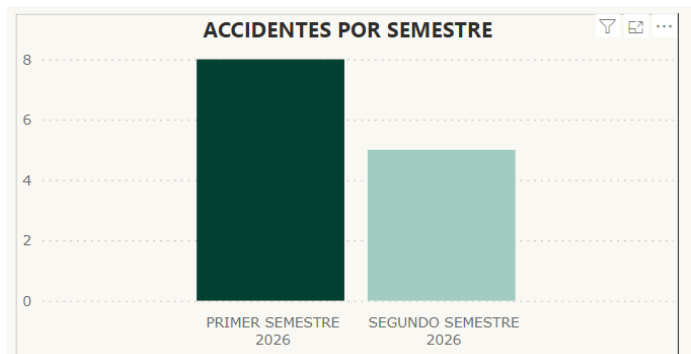


Figura 15. Predicción de accidentes por semestre para 2026.

Por último, en la Figura 16, se muestra la franja etaria de las personas afectadas por los accidentes laborales. Se aprecia 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 resaltan la importancia de ajustar las estrategias de seguridad en entornos laborales para abordar las necesidades específicas de cada segmento demográfico.

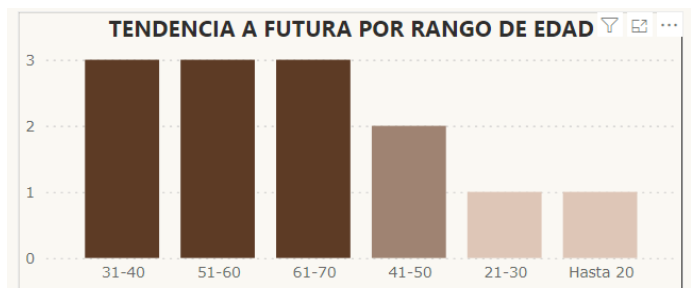


Figura 16. Predicción de accidentes por rango de edad para 2026.

Este estudio concuerda con investigaciones anteriores 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 (García, 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 Correlation Attribute Eval, 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 de trabajo en dicha Institución.

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 registros. 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 en el trabajo. 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.

CONCLUSIONES

El proceso de preparación, limpieza y transformación de los datos sobre accidentes de trabajo ha dado como resultado una base de datos sólida y detallada. 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 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 a la Institución y establecer bases sólidas para desarrollar políticas y estrategias preventivas más efectivas.

La tendencia semestral de accidentes de trabajo 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.

REFERENCIAS BIBLIOGRÁFICAS

- Ampuero, E., Pozo, M., & Delgado, K. (2018). Administración de riesgo laboral en el Ecuador. Digital Publisher, 3(5), 1-10. https://www.593dp.com/index.php/593_Digital_Publisher/article/view/66
- Arango, J., Correa, Y., & Luna, J. (2020). La salud ocupacional y su respuesta histórica a las necesidades de salud de los trabajadores. Revista Cubana de Salud y Trabajo, 21(2), 14-29. <https://revsaludtrabajo.sld.cu/index.php/revsyt/article/view/140>
- Candongá, J., & Samaniego, P. (2021). Percepción de riesgos laborales en el ámbito ocupacional universitario. Revista Publicando, 8(28), 47-58. <https://doi.org/10.51528/rp.vol8.id2147>
- 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. (Tesis de maestría). Universidad de la Costa.

Soto, L., & Melara, M. (2018). Riesgos laborales del personal de enfermería en el servicio de infectología del Hospital Nacional de niños Benjamín Bloom de mayo -septiembre de 2017. (Tesis de maestría). Universidad de El Salvador.

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. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S2218-36202021000200357