Mundo Fesc E-ISSN: 2216-0388 P- ISSN: 2216-0353 Artículo Original

Redes neuronales recurrentes y crisp-dm en análisis y predicción de accidentes de tránsito en bucaramanga, colombia

Recurrent neural networks and crisp-dm in traffic accident analysis and prediction in bucaramanga, colombia

Recibido: 18 de mayo de 2024
Aprobado: 16 de agosto 2024

Forma de citar: C. A. Mejía Rodríguez, D. A. Diaz Vergel, and L. M. Arévalo Vergel, "Redes Neuronales Recurrentes Y Crisp-Dm En Análisis Y Predicción De Accidentes De Tránsito En Bucaramanga, Colombia"., *Mundo Fesc*, vol. 14, no. 30, pp. 79–96, Sep. 2024, doi: 10.61799/2216-0388.1716.

Carlos Alberto Mejía-Rodríguez

Magíster en Big Data y Ciencia de Datos, Universidad Internacional de Valencia. Docente Investigador de la Universidad Popular del Cesar. Correo electrónico: calbertomejia@unicesar.edu.co. ORCID ID: https://orcid.org/0000-0001-5084-6010

Deider Alfonso Diaz-Vergel

Magíster en Gestión de la tecnología Educativa, Docente Asistente de la Universidad Popular del Cesar. Correo electrónico: deideradiaz@unicesar.edu.co. ORCID ID: https://orcid.org/0009-0004-9805-5276

Lina Marcela Arévalo-Vergel

Especialista en Gerencia de Proyecto. Docente Universidad Popular del Cesar (Colombia). Correo electrónico: linamarcelaarevalo@unicesar.edu.co. ORCID ID: https://orcid.org/0000-0002-7731-5444

***Autor para correspondencia:**
calbertomejia@unicesar.edu.co.

Redes neuronales recurrentes y crisp-dm en análisis y predicción de accidentes de tránsito en Bucaramanga, Colombia

Resumen

El presente trabajo tuvo como objetivo implementar redes neuronales recurrentes (RNN) utilizando la metodología CRISP-DM para analizar y predecir la gravedad, la frecuencia mensual y el número diario de personas involucradas en accidentes de tránsito en Bucaramanga, Colombia. Se utilizaron datos recopilados entre enero de 2012 y septiembre de 2023. La metodología CRISP-DM guió todas las fases del proyecto, desde la comprensión del negocio hasta el modelado y evaluación. Los modelos RNN, incluyendo Many-to-One y LSTM, demostraron una alta precisión en la clasificación de la gravedad y un rendimiento aceptable en la predicción de accidentes e involucrados. Estos hallazgos respaldan la utilidad de las RNN y la metodología CRISP-DM como herramientas para fortalecer la seguridad vial mediante decisiones basadas en evidencia.

Palabras clave: Accidentes de tránsito, CRISP-DM, LSTM, Predicción, Redes Neuronales Recurrentes.

Recurrent neural networks and crisp-dm in traffic accident analysis and prediction in Bucaramanga, Colombia

Abstract

The objective of this study was to implement recurrent neural networks (RNN) using the CRISP-DM methodology to analyze and predict the severity, monthly frequency, and daily number of people involved in traffic accidents in Bucaramanga, Colombia. Data collected between January 2012 and September 2023 was used. The CRISP-DM methodology guided all phases of the project, from business understanding to modeling and evaluation. The RNN models, including Many-to-One and LSTM, demonstrated high accuracy in classifying accident severity and acceptable performance in predicting accidents and those involved. These findings support the usefulness of RNNs and the CRISP-DM methodology as tools to enhance road safety through evidence-based decision-making.

Keywords: RISP-DM, LSTM, prediction, Recurrent Neural Networks (RNN), traffic accidents.

Introducción

La ciencia de datos ha adquirido una creciente relevancia en sectores clave como la industria, la salud y la seguridad pública. En el contexto de los accidentes de tránsito, el análisis de grandes volúmenes de datos mediante técnicas de aprendizaje automático permite identificar patrones y factores de riesgo, facilitando así la formulación de políticas públicas más eficaces.

En este sentido, los accidentes de tránsito, definidos como eventos inesperados derivados de la circulación vehicular, suelen implicar daños a personas, vehículos e infraestructuras [1]. Es por ello que, este estudio surge como una oportunidad para aplicar conocimientos teóricos y prácticos adquiridos en el ámbito del Big Data, es importante entender que la Big Data comprende la captura y almacenamiento de datos, requiriendo capacidades de gestión de datos, antes de realizar modelos estadísticos y matemáticos, útiles para el proceso de toma de decisiones[2]. Además, muestra el comportamiento, las tendencias de los usuarios, para encontrar soluciones a través del análisis de grandes volúmenes de datos, utilizando gran capacidad de cómputo[3], respondiendo de esta manera a una problemática social y de salud pública que exige enfoques innovadores y basados en evidencia.

Si bien predecir y clasificar accidentes es fundamental, los conjuntos de datos asociados son complejos, voluminosos y con múltiples variables temporales, lo que dificulta su análisis. Frente a estos retos, el aprendizaje automático ofrece soluciones prometedoras [4].

Las redes neuronales recurrentes (RNN), en particular, se han posicionado como herramientas eficaces para el procesamiento de series temporales. Estas permiten utilizar datos históricos como entrada y generar predicciones sobre eventos futuros [5]. Aplicadas al caso de Bucaramanga, permiten analizar dependencias temporales en la ocurrencia de accidentes, facilitando tanto la evaluación histórica como la predicción de su frecuencia.

Para abordar estos procesos de forma estructurada, la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) ofrece una guía práctica que abarca desde la comprensión del problema hasta la implementación de modelos predictivos [6]. Esta metodología ha demostrado su eficacia en proyectos de análisis de datos complejos, incluyendo áreas como la movilidad urbana y la gestión del riesgo [7].

Fundamentos de Machine Learning

Los modelos de Machine Learning (ML) son sistemas matemáticos que procesan datos para identificar patrones y relaciones. Los últimos avances han mejorado su comprensión.

Conceptualizados como "cajas grises", transforman una entrada vectorial en una salida vectorial, como muestra la Figura 1 [8].

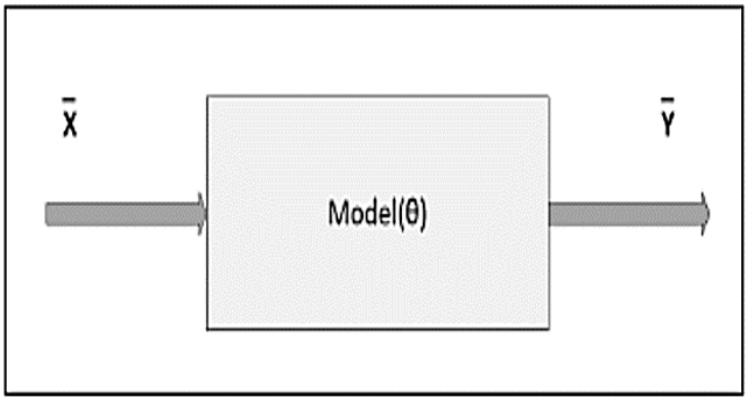

Figura 1. Esquema de un modelo genérico parametrizado con el vector ∅.
Fuente: Bonaccorso 2018 [9].

Los algoritmos de ML aproximan funciones estableciendo correspondencias entre entradas y salidas correctas mediante métodos algebraicos. Este proceso combina datos, modelos, funciones de pérdida y técnicas de optimización para aprender y generalizar estas relaciones [8].

Redes Neuronales artificiales

Las redes neuronales artificiales están compuestas por unidades básicas llamadas neuronas, que se encuentran interconectadas. Cada neurona realiza una función específica [10]. En la Figura 2, se puede observar el esquema básico de una neurona artificial.

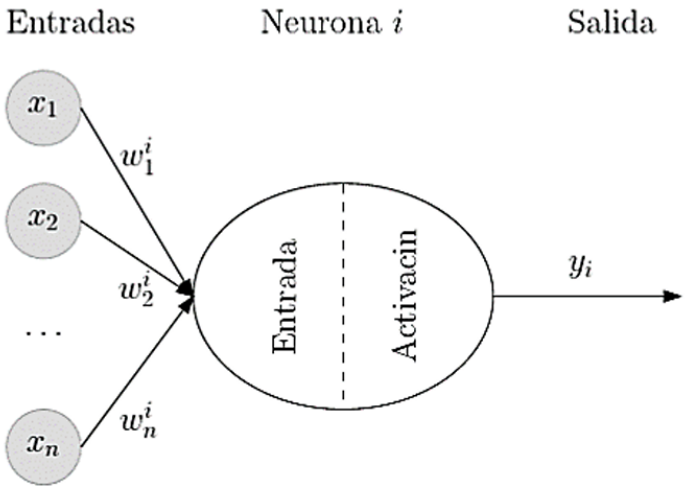

Figura 2. Esquema básico de una neurona artificial
Fuente: Casas [10].

En este esquema, el conjunto de entradas se denota como $X = \{x_1, x_2, \ldots, x_n\}$ y cada entrada tiene un peso asociado $W^i = \{w_1^i, w_2^i, \ldots, w_n^i\}$, que refleja la importancia del valor de entrada x_j que llega a la neurona i. Cada neurona combina estos valores de entrada mediante una función de entrada, y luego procesa el resultado a través de una función de activación, ajustando las entradas para generar el valor de salida y_i el cual se propaga a otras neuronas o se usa como salida final.

Cada entrada x_j tiene un peso específico w_j^i que indica su relevancia. La función de entrada fusiona las distintas entradas ponderadas por sus pesos y genera un único valor de salida. Las funciones más comunes para combinar los valores de entrada son:

La función suma ponderada:

$$z(x) = \sum_{j=1}^{n} x_j w_j^i \quad (1)$$

La función "máximo":

$$z(x) = \max\left(x_1 w_1^i, \ldots, x_n W_n^i\right) \quad (2)$$

La función "mínimo":

$$z(x) = \min\left(x_1 w_1^i, \ldots, x_n w_n^i\right) \quad (3)$$

La función lógica AND (∧) o OR (v), aplicable en el caso de entradas binarias:

$$z(x) = \left(x_1 W_1^i \wedge \ldots \wedge x_n W_n^i\right), \quad (4)$$

$$z(x) = \left(x_1 W_1^i V \ldots V x_n W_n^i\right), \quad (5)$$

La elección de la función de combinación depende del problema y los datos, pero la suma ponderada es la más comúnmente utilizada [11].

Aplicaciones del aprendizaje automático en el análisis de accidentes de tránsito

El aprendizaje automático, mediante algoritmos avanzados y técnicas de modelado, ha facilitado una comprensión más profunda de los factores que contribuyen a los accidentes de tránsito. Esta tecnología permite desarrollar estrategias de prevención más efectivas y adaptadas al contexto vial [12], resaltan los avances en la predicción de accidentes, especialmente en el análisis de factores que inciden en su frecuencia y severidad. En este campo, las redes neuronales (NN), y particularmente los enfoques de aprendizaje profundo como las redes neuronales recurrentes (RNN) y convolucionales (CNN), han demostrado ser altamente eficaces, ofreciendo precisión y eficiencia notables al predecir la gravedad de las lesiones.

Un ejemplo innovador se encuentra en el estudio "Traffic Accident Detection Using

Background Subtraction and CNN Encoder–Transformer Decoder in Video Frames", el cual propone un enfoque que combina un codificador CNN con un decodificador Transformer. Esta arquitectura permite la sustracción del fondo en secuencias de video para facilitar la detección de accidentes, alcanzando una precisión superior al 96% y mejorando en un 5% el rendimiento frente a métodos que no emplean dicha técnica [13]. Asimismo, se identifican líneas de investigación futura como la validación en conjuntos de datos más amplios y la integración con sistemas inteligentes de transporte.

Cada año, los accidentes viales provocan numerosas muertes, lo que subraya la necesidad de sistemas eficaces de detección temprana. Soluciones basadas en aprendizaje profundo han demostrado ser valiosas para identificar accidentes y enviar alertas rápidas a los servicios de emergencia. En estudios recientes se han comparado arquitecturas como Perceptrón Multicapa (MLP), CNN, DenseNet e Inception V3. Aunque MLP ofrece alta precisión, Inception V3 sobresale por su velocidad predictiva, haciéndolo ideal para aplicaciones en tiempo real [14]. Sin embargo, persisten desafíos como la escasez de datos etiquetados y limitaciones técnicas que requieren atención en futuras investigaciones.

Además, como señalan [15], la integración de Big Data con modelos de aprendizaje automático optimiza el rendimiento predictivo en la detección y análisis de accidentes de tráfico, al tiempo que habilita la identificación de patrones complejos y correlaciones no lineales. Este enfoque potencia no solo la precisión en las predicciones, sino también el diseño de estrategias preventivas y políticas públicas basadas en evidencia empírica y analítica.

Método y Materiales

Este estudio adoptó un enfoque cuantitativo, ya que no se pretendió alterar o hacer modificaciones en la variable, en este caso[16] para analizar y predecir accidentes de tránsito en Bucaramanga, Colombia, para ello se tomó en cuenta la necesidad de comprobar la importancia estadística de las variables predictoras[17].

Se utilizó el marco metodológico CRISP-DM (Cross-Industry Standard Process for Data Mining), ampliamente reconocido en proyectos de minería de datos por sus ventajas en la estandarización de procesos, reducción de costos, ahorro de tiempo, transferencia de conocimiento y reutilización de buenas prácticas [18]. La CRISP-DM proporciona un enfoque estructurado, basado en el modelo de Descubrimiento de Conocimiento en Bases de Datos (KDD) [4], y consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación, para ilustrar gráficamente ver Figura 3.

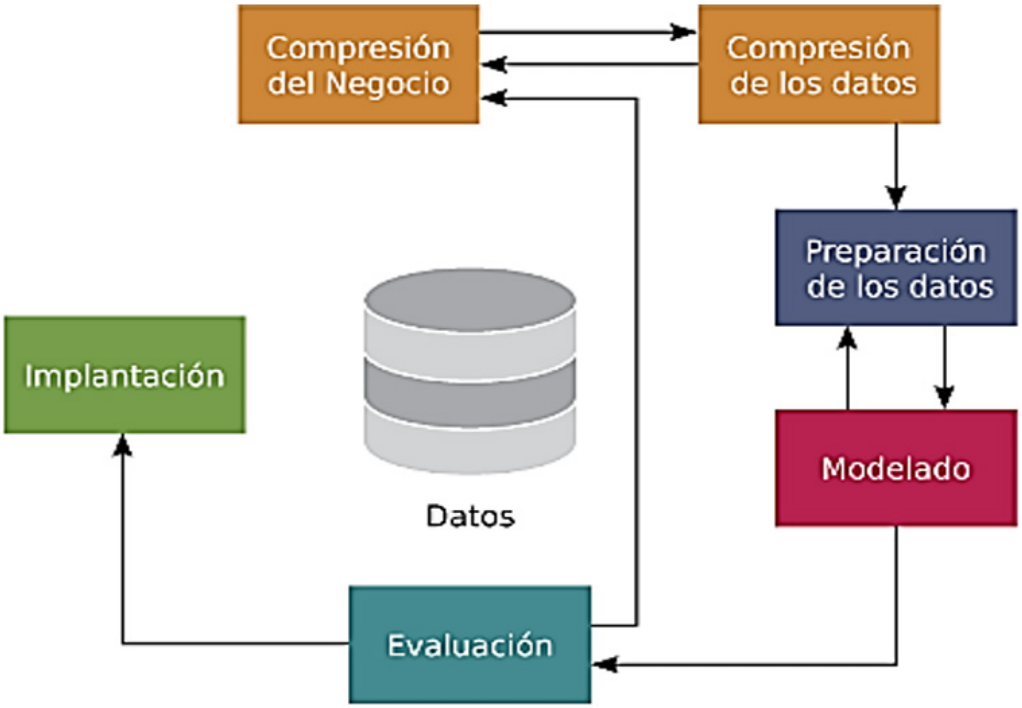

Figura 3. Flujo CRISP-DM
Fuente: Castilo 2019.[6]

Participantes

Se utilizaron registros oficiales de accidentes de tránsito ocurridos en Bucaramanga entre enero de 2012 y septiembre de 2023, sin distinción por tipo de vehículo o gravedad. Se excluyeron datos incompletos o inconsistentes.

Instrumentos

Se emplearon técnicas de análisis de datos y modelado predictivo mediante Redes Neuronales Recurrentes (RNN). Estas se eligieron por su capacidad para procesar datos secuenciales y modelar dependencias temporales, elementos clave en la predicción de eventos como los accidentes de tránsito. Este enfoque ha demostrado ser eficaz en problemas similares, donde los datos temporales requieren modelos dinámicos y adaptativos [19].

Procedimientos

Los datos fueron sometidos a un riguroso proceso de preprocesamiento, que incluyó limpieza y normalización. Posteriormente, se implementaron modelos RNN del tipo Many-to-One con celdas LSTM, siguiendo las fases de la metodología CRISP-DM. Los modelos se evaluaron tanto en tareas de clasificación de la gravedad del accidente como

en regresión del número de involucrados, demostrando su potencial como herramienta predictiva para fortalecer las estrategias de seguridad vial.

Resultados y Discusión

El desarrollo del proyecto se estructura conforme a las fases de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), permitiendo una ejecución sistemática y reproducible del análisis de datos. A continuación, se presenta la discusión de los resultados obtenidos en cada fase, iniciando con la comprensión del negocio.

Comprensión del negocio

La primera fase del proceso CRISP-DM se orienta a establecer una comprensión profunda del problema desde la perspectiva del negocio u organización involucrada. De acuerdo con [20], esta etapa implica identificar los requerimientos y expectativas de los actores clave, así como evaluar la viabilidad del proyecto con base en los datos disponibles. Además, permite traducir los objetivos del negocio en metas técnicas concretas y establecer los criterios de éxito del proyecto [21].

La fuente principal de datos corresponde a la Dirección de Tránsito de Bucaramanga, cuya misión institucional es garantizar la seguridad vial, regular el tránsito vehicular y promover el desarrollo de infraestructura vial sostenible. Los registros históricos de accidentalidad, publicados en el portal gubernamental "Datos Abiertos", constituyen la base para el análisis predictivo planteado.

El objetivo del estudio es implementar modelos predictivos basados en redes neuronales recurrentes (RNN), particularmente con arquitectura Many-to-One y celdas LSTM, para clasificar la gravedad de los accidentes y predecir el número mensual de eventos y el número diario de personas involucradas. Se espera que estos modelos alcancen una precisión aceptable y contribuyan al diseño de estrategias de prevención vial basadas en datos.

Comprensión de los datos

La fase de comprensión de los datos en el marco de trabajo CRISP-DM tiene como objetivo obtener una visión detallada del conjunto de datos disponible, identificando patrones, inconsistencias y posibles relaciones que podrían ayudar en las etapas posteriores del análisis. Este proceso implica explorar el dataset, comprender su estructura, y evaluar su calidad para detectar valores faltantes, ruidos o anomalías [22].

Para el estudio se analizó el conjunto de datos titulado "Accidentes de Tránsito acontecidos en el municipio de Bucaramanga", proporcionado por la Dirección de Tránsito de Bucaramanga y disponible en el portal "Datos Abiertos" de Colombia. Este dataset contiene 39193 registros de accidentes ocurridos entre 2012 y 2023, con variables

numéricas y categóricas relacionadas con el tipo de vehículo, gravedad del accidente, ubicación y hora, entre otras. Las principales variables numéricas incluyen el número de vehículos involucrados en cada accidente, mientras que las variables categóricas abarcan la fecha, hora, gravedad y ubicación del accidente.

Preparación de los datos

En la preparación de datos para los modelos de análisis, se abordaron desafíos como valores nulos y atípicos. Se realizó una limpieza exhaustiva, corrigiendo errores tipográficos, rellenando valores faltantes y eliminando categorías redundantes. Se aplicaron técnicas de reducción de dimensionalidad, manteniendo solo la información relevante. Se crearon nuevas columnas, como "TOTAL_INVOLUCRADOS" y "SEMANA_DEL_AÑO", y se ajustaron las variables categóricas y numéricas para mejorar la precisión. Finalmente, se exportó un dataset optimizado en formato CSV, listo para la implementación de modelos predictivos.

Modelado

En el modelado se emplean datos históricos etiquetados para entrenar modelos capaces de hacer predicciones precisas sobre nuevos datos. El proceso implica la recopilación de datos, su división en conjuntos de entrenamiento y prueba, y el ajuste de parámetros según sea necesario. Se experimenta con diferentes algoritmos utilizando bibliotecas de Python, y los resultados se documentan en un cuaderno de Google Colab para su evaluación detallada.

Las redes neuronales recurrentes (RNN) son especialmente efectivas para la predicción de series temporales debido a su capacidad para procesar información secuencial [23]. Dado que el conjunto de datos abarca una serie temporal de accidentes ocurridos entre 2012 y 2023, los modelos RNN se utilizarán para capturar las dependencias temporales y mejorar la precisión de las predicciones.

El modelo Many-to-One RNN, que procesa secuencias de datos con múltiples entradas y una sola salida, es adecuado para tareas donde toda la secuencia de datos es necesaria para predecir un solo resultado, como la gravedad de un accidente [24]. Con datos históricos de accidentes, este modelo se aplicará para predecir la gravedad de accidentes futuros.

Construcción de los Modelos

Many-to-One RNN para la predicción de la gravedad de un accidente
Para clasificar la gravedad de los accidentes, se utilizó un modelo Many-to-One RNN optimizado con la técnica de eliminación hacia atrás, seleccionando solo las columnas clave. El preprocesamiento incluyó Label Encoding para convertir las variables categóricas a formato numérico, normalización de los datos y la división del conjunto en 80% para

entrenamiento y 20% para prueba. La optimización del modelo se complementó con la técnica de "Búsqueda de Hiperparámetros" utilizando la biblioteca "keras_tuner" de Python.

La Tabla I presenta los componentes y procedimientos clave empleados.

Tabla I. Many-to-One RNN Optimizada mediante Búsqueda de Hiperparámetros.

Característica	**Valor**
Técnica de Optimización	Búsqueda de Hiperparámetros (Hyperparameter Tuning)
Número de unidades (Primera capa RNN)	Variable entre 30 y 70 (con paso de 10)
Tasa de Dropout (Primera capa)	Variable entre 0.2 y 0.4 (con paso de 0.1)
Número de unidades (Segunda capa RNN)	Variable entre 30 y 70 (con paso de 10)
Tasa de Dropout (Segunda capa)	Variable entre 0.2 y 0.4 (con paso de 0.1)
Optimizador	Adam o RMSprop
Función de pérdida	Binary Crossentropy
Métricas	Accuracy
Número de épocas	10 (durante la búsqueda de hiperparámetros)
Tamaño del batch	Automático (ajustado por el optimizador)
Hiperparámetros aplicados	Número de unidades, Tasa de Dropout, Optimizador

El rendimiento del modelo Many-to-One RNN se puede evaluar mediante la Tabla II.

Tabla II. Métricas Generales del modelo Many-to-One RNN (optimizada).

	precision	**recall**	**f1-score**	**support**	**accuracy**	**f1_score**
Clase 0	0.89	0.78	0.83	3966	0.84	0.84
Clase 1	0.80	0.90	0.85	3858	0.84	0.84
accuracy	0.85	0.84	0.84	7824	0.84	0.84
macro avg	0.85	0.84	0.84	7824	0.84	0.84
weighted avg	0.85	0.84	0.84	7824	0.84	0.84

El modelo Many-to-One RNN tiene un rendimiento sólido y equilibrado en la predicción de la gravedad de los accidentes. Si bien es un poco más preciso en predecir accidentes que solo causan daños (Clase 0), es más sensible en detectar accidentes con heridos (Clase 1). El alto F1-Score y la precisión global del 85% refuerzan la idea de que este modelo es una herramienta eficaz para predecir la gravedad de los accidentes de

tránsito, proporcionando un equilibrio adecuado entre la precisión y la capacidad de identificación en ambas clases.

Many-to-One LSTM para la predicción del número de accidentes mensuales

Para la predicción del número de accidentes mensuales, se implementó un modelo Many-to-One LSTM, que es adecuado para capturar dependencias a largo plazo en series temporales. En este modelo, se seleccionaron columnas del dataset original y otras derivadas de transformaciones previas. Las variables utilizadas fueron el AÑO del accidente, el MES_NUMERO del accidente, el TOTAL_ACCIDENTES registrados en el mes, el TOTAL_INVOLUCRADOS en accidentes durante ese mes, así como el número de accidentes DIURNOS y NOCTURNOS. Además, se incluyó el número de vehículos PARTICULARES y EMPLEADOS involucrados en los accidentes.

El modelo se configuró con 100 unidades LSTM por capa, un dropout de 0.3 para evitar el sobreajuste, y una tasa de aprendizaje de 0.001 utilizando el optimizador Adam. Además, se estableció un total de 50 épocas de entrenamiento, un tamaño de lote de 32, y un periodo de secuencia de 6 meses. Los resultados del modelo se compararon con los datos reales de accidentes mensuales, y la Figura 4 muestra la comparación entre las predicciones y los datos reales obtenidos.

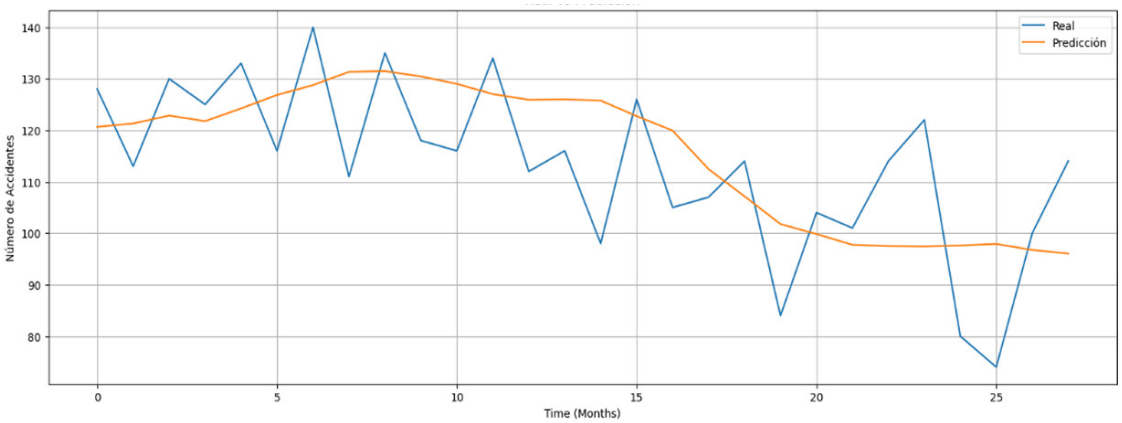

Figura 4. Número de accidentes en un mes Real vs Predicción.

El modelo Many-to-One LSTM obtuvo un MSE de 182.62 y un RMSE de 13.51, indicando una variabilidad moderada en las predicciones, con un error promedio de 13.51 accidentes mensuales. Esto refleja una precisión aceptable considerando la naturaleza impredecible de los accidentes.

Many-to-One LSTM para la predicción del número de involucrados en accidentes

La predicción del número de involucrados en accidentes diarios es esencial para el diseño de políticas de seguridad vial efectivas. En este estudio, se utilizó un modelo

Many-to-One LSTM para predecir la cantidad de personas involucradas en accidentes diarios, aprovechando datos históricos. Las columnas utilizadas y creadas incluyen la FECHA, que fue convertida al formato datetime, y el TOTAL_INVOLUCRADOS, que representa el número total de personas involucradas por día y es la variable principal para la predicción. Además, se incorporaron las variables DIADEL_ANO, DIA_DE_LA_SEMANA y PERIODO_DEL_DIA (mañana, tarde, noche), que ofrecen una visión más detallada sobre el comportamiento diario de los accidentes.

El modelo fue configurado con 50 unidades LSTM por capa y un dropout de 0.2 para evitar el sobreajuste. El optimizador Adam fue utilizado con su configuración por defecto, entrenando el modelo durante 50 épocas con un tamaño de lote de 32. El periodo de secuencia se definió en 30 días para capturar las tendencias de corto plazo. La Figura 5 ilustra la comparación visual entre las predicciones del modelo y los valores reales de los accidentes diarios.

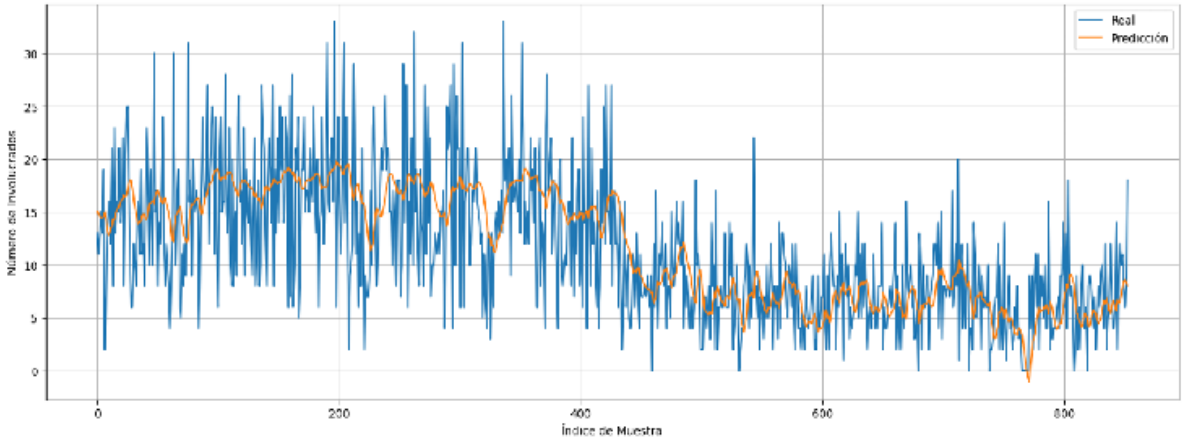

Figura 5. Número de involucrados en un accidente Real vs Predicción

El modelo Many-to-One LSTM mostró un MSE de 30.34 y un MAE de 4.28, indicando un error de predicción moderado, mientras que un R^2 de 0.33 sugiere una capacidad limitada para explicar la variabilidad en los datos de accidentes. Aunque las predicciones siguen la tendencia general de los valores reales, hay una considerable variabilidad no capturada. Esto indica que, si bien el modelo ofrece una precisión medianamente acertada, se requieren mejoras, como la incorporación de nuevas características o el ajuste de hiperparámetros, para mejorar su rendimiento. Aun con estas limitaciones, el modelo demuestra potencial en la predicción diaria del número de involucrados en accidentes.

En la evaluación de los modelos RNN para predecir accidentes de tránsito en Bucaramanga, el modelo Many-to-One Optimizado destacó con una precisión, recall, F1-score y accuracy de 0.84 en la predicción de la gravedad de los accidentes. Estos

Evaluación de los Modelos

En la evaluación de los modelos RNN para predecir accidentes de tránsito en Bucaramanga, el modelo Many-to-One Optimizado destacó con una precisión, recall, F1-score y accuracy de 0.84 en la predicción de la gravedad de los accidentes. Estos resultados sugieren que el modelo puede ser una herramienta efectiva para identificar con precisión la gravedad de los accidentes, permitiendo a las autoridades priorizar recursos y aplicar medidas preventivas de manera más eficiente.

En cuanto a la predicción del número de accidentes mensuales, el modelo arrojó un RMSE de 13.51, lo que indica un error moderado pero útil para prever la carga general de accidentes. Esta capacidad de identificar patrones en la ocurrencia de accidentes es valiosa para la planificación de recursos y la implementación de políticas de seguridad vial, aunque hay margen para mejorar su precisión.

Para la predicción del número de involucrados diarios en accidentes, el modelo Many-to-One LSTM, con un MSE de 30.34 y un R^2 de 0.33, demostró una precisión moderada. Aunque estos resultados indican la necesidad de ajustes y mejoras, el modelo tiene el potencial de contribuir significativamente a la gestión y prevención de accidentes en Bucaramanga, al proporcionar predicciones útiles que pueden guiar intervenciones más efectivas y oportunas.

Implementación de los Modelos

La fase final del proyecto, la implementación, se centra en aplicar los resultados de la minería de datos para satisfacer las necesidades de los usuarios [25]. Aunque no se contó con datos completamente actualizados para validar los modelos en un entorno en tiempo real, se logró avanzar significativamente a través de la exploración inicial y la experimentación con modelos de clasificación y regresión.

Los modelos evaluados, particularmente el Many-to-One LSTM utilizado para predecir la gravedad de los accidentes, presentaron un nivel de precisión aceptable. Por su parte, los modelos de regresión demostraron un desempeño moderado al predecir la frecuencia mensual de accidentes y el número diario de personas involucradas. Estos resultados reafirman la viabilidad de utilizar técnicas de aprendizaje automático como herramienta estratégica en la prevención de accidentes de tránsito, cumpliendo satisfactoriamente con los objetivos planteados.

A pesar de ciertas limitaciones, como la falta de datos en tiempo real, los hallazgos aportan información significativa que puede guiar decisiones futuras relacionadas con la planificación y gestión del tráfico. La implementación de estos modelos predictivos en los sistemas de gestión vial no solo optimizaría la respuesta institucional ante los incidentes, sino que también facilitaría la asignación eficiente de recursos y la formulación

de políticas públicas basadas en evidencia. Este enfoque permite anticipar escenarios críticos y diseñar estrategias proactivas para mitigar riesgos viales.

Conclusiónes

La implementación de modelos predictivos basados en redes neuronales recurrentes para el análisis de accidentes en Bucaramanga, Colombia, representa un avance notable en la incorporación de tecnologías de inteligencia artificial al ámbito de la seguridad vial. Al capturar patrones temporales y espaciales de los accidentes, estos modelos no solo validan su utilidad para predecir la gravedad y frecuencia de los incidentes, sino que también se consolidan como herramientas de apoyo para la toma de decisiones en la gestión del tráfico urbano.

La implementación de modelos predictivos basados en redes neuronales recurrentes para el análisis de accidentes en Bucaramanga, Colombia, representa un avance notable en la incorporación de tecnologías de inteligencia artificial al ámbito de la seguridad vial. La posibilidad de anticipar riesgos, identificar zonas críticas y optimizar los tiempos de respuesta ante emergencias establece una nueva dinámica preventiva en el abordaje de la siniestralidad vial.

Además, el uso de la metodología CRISP-DM aportó una estructura clara y sistemática en todas las fases del proyecto, desde la comprensión del problema hasta la evaluación de los modelos. Esta metodología permitió gestionar eficientemente los datos, asegurar la trazabilidad de los procesos analíticos y facilitar la reproducibilidad del modelo, lo cual es especialmente valioso en contextos de seguridad vial donde la precisión y la oportunidad de las predicciones pueden tener un impacto directo en la vida de las personas.

Referencias

[1] J. Cabrerizo, *Manual para la investigacion y reconstrucción de las causas de accidentes de tráfico*. Wolters Kluwer Espana, 2016. [Online]. Available: https://elibro.net/es/lc/biblioupc/titulos/55965

[2] M. Valencia-Cárdenas, J. A. Restrepo-Morales, and F. J. Día-Serna, "Big Data Analytics in the Agribusiness Supply Chain Management", *AiBi Revista de Investigación, Administración e Ingeniería*, vol. 9, no. 3, pp. 32–42, Sep. 2021, doi: 10.15649/2346030X.2583

[3] J. Quintero, L. Orjuela, J. Gordillo y A. Sánchez-Quiñones, "Análisis de implementación del Big Data en empresas y en profesionales de Contaduría Pública en Colombia. *Revista Temario Científico*, 2 (1), pp.50-60, 2022. Doi: https://doi.org/10.47212/Alinin.1.2.5

[4] F. Alhaek, W. Liang, T. M. Rajeh, M. H. Javed, and T. Li, "Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach," *Knowl Based Syst,* vol. 286, p. 111406, 2024, doi: https://doi.org/10.1016/j.knosys.2024.111406.

[5] J. Casas Roma, T. Lozano Bagén, and A. Bosch Rué, Deep Learning : *Principios y Fundamentos*. Barcelona, SPAIN: Editorial UOC, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=7025971

[6] J. A. Castilo Romero, *Big data*. IFCT128PO. IC Editorial, 2019. [Online]. Available: https://elibro.net/es/lc/biblioupc/titulos/124254

[7] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep Learning on Traffic Prediction: Methods, Analysis, and Future Directions," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 6, pp. 4927–4943, Jun. 2022, doi:10.1109/tits.2021.3054840

[8] P. D. Smith, *Hands-On Artificial Intelligence for Beginners : An Introduction to AI Concepts, Algorithms, and Their Implementation.* Birmingham, United Kingdom: Packt Publishing, Limited, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5607070

[9] G. Bonaccorso, A. Fandango & R. Shanmugamani, *Python: Advanced Guide to Artificial Intelligence : Expert Machine Learning Systems and Intelligent Agents Using Python*. Packt Publishing, Limited. http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5626921

[10] J. Casas Roma, T. Lozano Bagén, and A. Bosch Rué, *Deep Learning: Principios y Fundamentos*. Barcelona, SPAIN: Editorial UOC, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=7025971

[11]T. Zhang and Y. Zheng, *Introduction to Machine Learning*. Springer, 2020. doi: 10.1007/978-3-030-41804-7.

[12] Md. E. Shaik, Md. M. Islam, and Q. S. Hossain, "A review on neural network techniques for the prediction of road traffic accident severity," *Asian Transport Studies*, vol. 7, p. 100040, 2021, doi: https://doi.org/10.1016/j.eastsj.2021.100040

[13] Y. Zhang and Y. Sung, "Traffic Accident Detection Using Background Subtraction and CNN Encoder–Transformer Decoder in Video Frames," *Mathematics*, vol. 11, no. 13, p. 2884, 2023, doi: https://doi.org/10.3390/math11132884

[14] P. D. F, "An Overview of Different Deep Learning Techniques Used in Road Accident

Detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023, doi: https://doi.org/10.14569/IJACSA.2023.0141144

[15] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int J Inf Manage*, vol. 35, no. 2, pp. 137–144, 2015, doi:10.1016/j.ijinfomgt.2014.10.007

[16] D.C. Rojas Nieves, Y.V. Chirinos Araque, N. Barbera, M.J. Nieves Álvarez "Empleados tóxicos hasta donde son una amenaza para las organizaciones", *Mundo Fesc*, vol 13, no. 27, pp. 325-340, 2023. https://doi.org/10.61799/2216-0388.1490

[17] M. Vergel-Ortega, Z. C. Nieto-Sánchez, C. S. Gómez-Vergel, "Predictores de innovación en programas de ingeniería y postgrado utilizando estrategias basadas en plataformas digitales," *Revista. UIS Ingeniería*, vol. 20, no. 1, pp. 213-222, 2021, doi: 10.18273/revuin.v20n1-2021018

[18] W. Y. Ayele, "Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas Using a Textual Dataset," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020, doi: https://doi.org/10.14569/IJACSA.2020.0110603

[19] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol: O'Reilly Media, 2019.

[20] A. So, T. V. Joseph, R. Thas John, A. Worsley, and S. Asare, *The Data Science Workshop : A New, Interactive Approach to Learning Data Science.* Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6033288

[21] H. Wiemer, L. Drowatzky, and S. Ihlenfeldt, "Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model," *Applied Sciences*, vol. 9, no. 12, 2019, doi: https://doi.org/10.3390/app9122407

[22] D. Berrar, "Cross-Validation," in Encyclopedia of Bioinformatics and Computational Biology, Academic Press, 2018, pp. 542–545. doi:10.1016/B978-0-12-809633-8.20483-8

[23] X. Zhang, C. Zhong, J. Zhang, T. Wang, and W. W. Y. Ng, "Robust recurrent neural networks for time series forecasting," Neurocomputing, vol. 526, pp. 143–157, 2023, doi: https://doi.org/10.1016/j.neucom.2023.01.037

[24] O. Colliot, Machine Learning for Brain Disorders. in Neuromethods, Vol. 197. New

York, NY: Humana, 2023. [Online]. Available: https://doi.org/10.1007/978-1-0716-3195-9

[25] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," Data Knowl Eng, vol. 139, p. 102013, 2022, doi: https://doi.org/10.1016/j.datak.2022.102013 NY: Humana, 2023. [Online]. Available: https://doi.org/10.1007/978-1-0716-3195-9

[25] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data Knowl Eng*, vol. 139, p. 102013, 2022, doi: https://doi.org/10.1016/j.datak.2022.102013