sciendo

VOLUMEN 16, N° 4 2022
Revista de Automatización, Robótica Móvil y Sistemas Inteligentes

Un modelo distribuido de análisis de big data para accidentes de tráfico Clasificación y reconocimiento basados en núcleos SparkMlLib

Enviado: 21 de junio de 2022; aceptado el 2 de agosto de 2022

Imad El Mallahi, Jamal Riffi, Hamid Tairi, Abderrahamane Ez-Zahout, Mohamed Adnane Mahraz

DOI: 10.14313/JAMRIS/4-2022/34

Abstracto:

Este artículo se centra en el análisis de big data para la predicción de accidentes de tráfico basado en núcleos SparkMlLib. Sin embargo, las canalizaciones de aprendizaje automático de Spark proporcionan una API útil y adecuada que ayuda a crear y optimizar modelos de clasificación y predicción para la toma de decisiones en accidentes de tráfico. Los científicos de datos se han centrado recientemente en las técnicas de clasificación y predicción de accidentes de tráfico; las técnicas de análisis de datos para la extracción de características también han seguido evolucionando. El análisis de un gran volumen de datos recibidos requiere un tiempo de procesamiento considerable. En la práctica, la implementación de estos procesos en sistemas de tiempo real requiere una alta velocidad de cálculo. La velocidad de procesamiento desempeña un papel importante en el reconocimiento de accidentes de tráfico en sistemas de tiempo real. Requiere el uso de tecnologías modernas y algoritmos rápidos que aumenten la velocidad en la extracción de los parámetros de las características de los accidentes de tráfico. Los problemas de overclocking durante el procesamiento digital de accidentes de tráfico aún no se han resuelto por completo. Nuestro modelo propuesto se basa en el procesamiento avanzado del núcleo Spark MlLib. Utilizamos la API de transmisión de datos en tiempo real de Spark para recopilar continuamente datos en tiempo real de múltiples fuentes externas en forma de flujos de datos. En segundo lugar, estos flujos de datos se tratan como tablas independientes. Posteriormente, utilizamos continuamente el algoritmo de bosque aleatorio para extraer los parámetros característicos de un accidente de tráfico.

El uso de este método propuesto permite aumentar la velocidad de los procesadores. Los resultados experimentales demostraron que el método propuesto extrae con éxito las características del accidente y logra una clasificación uniforme en comparación con otros algoritmos convencionales de reconocimiento de accidentes de tráfico. Finalmente, compartimos con otros usuarios todos los accidentes detectados, junto con sus detalles, en aplicaciones en línea.

Palabras clave: Big data, aprendizaje automático, accidente de tráfico, predicción de gravedad, red neuronal convolucional

1. Introducción

La creación de enfoques de comunicación entre el vehículo, el radar y la tecnología informática es una de las Tareas más críticas en la inteligencia artificial moderna. Una de las maneras más fáciles para que un usuario ingrese información es a través de un accidente de tráfico. Por lo tanto, la tecnología de procesamiento y análisis de datos y sus herramientas se han convertido en una parte esencial de la sociedad de la información. Además, el reconocimiento de accidentes de tráfico es un aspecto esencial de la investigación en el procesamiento de datos y una técnica vital de interacción entre el vehículo, el radar y la computadora. Los datos de accidentes de tráfico contienen características semánticas y personales, así como información ambiental.

Recientemente, la predicción de la gravedad de los accidentes de tráfico se ha convertido en una preocupación importante [1-5]. El estudio que se presenta aquí busca proporcionar una herramienta de predicción para el problema de la gravedad, que constituye información importante para la logística de emergencias [6-7]. El mayor desafío es la falta de datos en tiempo real de los departamentos de seguridad vial sobre los accidentes de tráfico.

El trabajo propuesto realizó pruebas de significancia estadística sobre el impacto de la aplicación de una red neuronal multiclase y un bosque aleatorio multiclase en un conjunto de datos de accidentes de tráfico [8-12]. Algunos algoritmos de aprendizaje automático pueden ayudar en la toma de decisiones complejas que respaldan soluciones de sistemas [13-22]; además, algunos autores abordan el control de semáforos como un problema complejo en las sociedades modernas [23-27,36].

Este artículo presenta una solución eficiente: utilizar datos para la predicción de la gravedad, detectando el problema de predicción de la gravedad en accidentes de tráfico. En este artículo se propusieron dos técnicas de aprendizaje automático para la detección de la predicción de la gravedad en accidentes de tráfico. La red neuronal multiclase demostró una mayor precisión, con una predicción de severidad del 93,64 %; esto supera al algoritmo de bosque aleatorio multiclase, que alcanzó un 87,71 % de precisión. El algoritmo de bosque aleatorio combina los resultados de múltiples árboles de decisión (creados aleatoriamente) para generar el resultado final.

Conclusión: La aplicación de algoritmos de aprendizaje automático a datos de predicción de gravedad puede ayudar a los proveedores de predicción de gravedad y a las personas a prestar atención a los riesgos y cambios en el estado de los accidentes de tráfico para mejorar la calidad de vida. El sistema propuesto se aplicó a un conjunto de datos de accidentes de tráfico. Los resultados experimentales del trabajo propuesto demostraron que el uso del método de redes neuronales multiclase puede aumentar la probabilidad de precisión diagnóstica. Utilizamos Apache Spark, compatible con aprendizaje automático y otros macrodatos.

Plataformas analíticas que admiten aprendizaje automático, como Hadoop, AzureML y BigML (Fig. 1). El aprendizaje automático (DL) es una rama del aprendizaje automático que puede resolver problemas de clasificación, predicción y agrupamiento en entornos del Internet de los Vehículos (IoT).

Almacenamiento de Big Data: Para almacenar los datos entrantes en tiempo real, utilizamos un clúster especial como HDFS (Sistema de Archivos Distribuidos de Hadoop) o cualquier otra base de datos NoSQL. Para nuestros tratamientos, especificamos el procedimiento de localidad de datos llamando desde el sistema solicitante externo a cualquier función de análisis de Big Data de lam para realizar el siguiente paso (véase la Fig. 1). Para el procesamiento y el procesamiento avanzado, podemos aplicar directamente MapReduce, traduciendo todo el procesamiento de clasificación o reconocimiento a las tareas de mapeadores y reductores.

Y tenemos que llamar directamente a las funciones avanzadas preimplementadas y a las API desde el núcleo de Apache Spark, SparkMILib.

Visualización: En este caso, nuestros ojos tienden a sentirse más atraídos por lo visual que por el contenido escrito. Existen varias maneras de practicar este paso (véase la Fig. 2). Primero, trazamos las trayectorias de los vehículos o podemos trazar histogramas.

2. Trabajos relacionados

Durante las últimas diez décadas, la cuestión de la seguridad vial ha sido de interés para la sociedad y la economía.

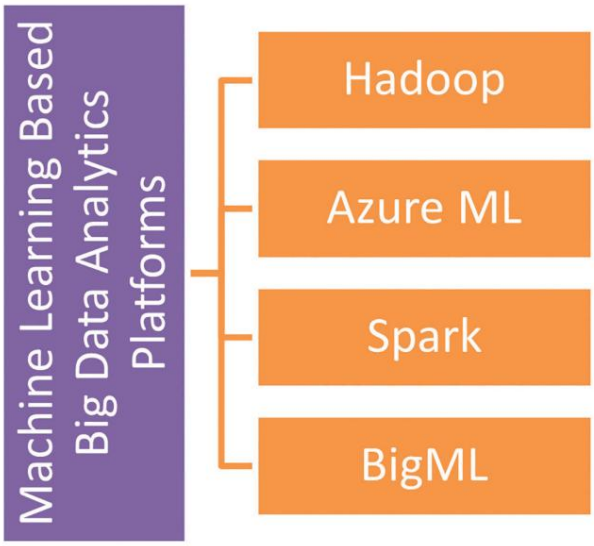

Figura 1. Plataformas de análisis de big data para el aprendizaje automático

Desarrollo a nivel mundial. Diversas soluciones en sistemas inteligentes se emplearon para la clasificación de accidentes de tránsito. Los métodos anteriores utilizaban características definidas manualmente, principalmente basadas en la combinación de imágenes de accidentes viales e información estadística [28].

Durante este año (2022), en Marruecos se instalaron radares de velocidad en las carreteras para controlar y transmitir imágenes y vídeos en tiempo real en caso de infracciones de tráfico o accidentes, lo que puede ayudar en la toma de decisiones sobre accidentes de tráfico. Los descriptores de características definidos manualmente se incorporan a los modelos tradicionales de aprendizaje automático (SVM) [29]. En Sharma et al. [30] se puede encontrar un análisis exhaustivo de los estudios sobre enfoques de visión artificial para el reconocimiento de accidentes de tráfico.

Con los avances en hardware para cámaras de velocidad, especialmente con el uso incorporado de GPU, las redes neuronales profundas (DNN) han alcanzado nuevos estándares en muchas fronteras de investigación. La principal ventaja de las DNN es que no requieren la selección manual de características, y las características se aprenden dentro de un marco de DNN. Sin embargo, las DNN requieren una gran cantidad de datos de entrenamiento, que no siempre están disponibles. Para conjuntos de datos de tamaño pequeño o moderado, el aprendizaje por transferencia puede ayudar a superar las limitaciones de tamaño del conjunto de datos [31]. Delen y Sharda [32] identificaron los predictores significativos de la gravedad de las lesiones en accidentes de tráfico utilizando una serie de redes neuronales artificiales. Alikhan y Lee [33-41] utilizaron el método heurístico de clasificación por clústeres para mejorar la precisión en la clasificación de la gravedad de los accidentes de tráfico.

3. Antecedentes

Es necesario procesar datos a gran escala generados desde el entorno de Internet de los vehículos a partir de diversas fuentes, como cámaras y sensores.

El aprendizaje automático (AA) puede utilizarse para procesar big data del Internet de los Vehículos (IoV). Las plataformas de análisis de big data compatibles con AA son necesarias para el análisis del IoV.

En esta sección, presentamos Apache Spark, compatible con ML y otras plataformas de análisis de Big Data que admiten aprendizaje automático, como Hadoop, AzureML y BigML. DL es una rama del aprendizaje automático que puede resolver problemas de clasificación, predicción y agrupamiento en entornos de Internet de los vehículos.

3.1. Apache Spark

Spark es un marco de procesamiento de big data basado en transmisión, aprendizaje automático y procesamiento de gráficos.

Figura 2. Procesamiento de Big Data

[36]. Es un marco de código abierto y fue desarrollado para superar algunas de las limitaciones de Hadoop MapReduce. Spark utiliza memoria basada en el procesamiento de grandes cantidades de datos, y es más rápido en términos de procesamiento de datos que el marco MapReduce. Como resultado, los datos se almacenan en memoria utilizando conjuntos de datos distribuidos resilientes. Además, Spark admite análisis en tiempo real. Chiroma et al. [36] presentó la biblioteca de aprendizaje automático distribuido de código abierto de Spark, MLlib. Existen varias configuraciones de aprendizaje en MLlib para mejorar la funcionalidad de manera eficiente, como la optimización, las primitivas de álgebra lineal y los métodos estadísticos subyacentes. Además, MLlib proporciona una API de alto nivel y varios lenguajes que aprovechan el rico ecosistema de Spark para simplificar el desarrollo de canalizaciones de aprendizaje automático de extremo a extremo. Chiroma et al. [36] discutieron el DL sobre Apache Spark para BDA móvil. Los autores mostraron cómo Spark puede realizar DL distribuido en MapReduce. El trabajador de Spark aprende cada partición del modelo profundo para todo el big data móvil. Luego, los parámetros utilizan el modelo profundo maestro de todos los modelos parciales mediante el promedio.

3.2. Hadoop.

Hadoop se ha consolidado como un marco importante para el procesamiento distribuido de grandes conjuntos de datos en clústeres de máquinas [36]. A lo largo de los años, se han desarrollado numerosos proyectos relacionados con Hadoop para respaldar este marco, como Hive, Pig, Tez, Zookeeper y Mahout. Mahout es uno de los marcos de álgebra lineal distribuida para el aprendizaje automático escalable.

3.3. AzureML

AzureML es una plataforma de aprendizaje automático colaborativo basada en análisis predictivo en big data, que permite un desarrollo sencillo de modelos predictivos y API.
AzureML ofrece numerosas características únicas, como la fácil operatividad, la colaboración en el control de versiones y la integración del código de usuario. Chiroma et al. [36] propusieron una técnica para AzureML basado en la nube denominada Flujo Generalizado, que permite la clasificación binaria y conjuntos de datos multiclase, y los procesa para maximizar la precisión general de la clasificación. El rendimiento de la técnica se prueba en conjuntos de datos basados en el modelo de clasificación optimizado.

Los autores utilizaron tres conjuntos de datos públicos y un conjunto de datos local para evaluar el flujo propuesto mediante la clasificación. El resultado de los conjuntos de datos públicos mostró una precisión del 97,5 %. Además, el concepto se ha vuelto indispensable en las tecnologías de big data. Por ejemplo, AzureML admite redes neuronales para regresión, clasificación de dos clases y clasificación multiclase.

3.4. BigML

BigML proporciona servicios de aprendizaje automático (ML) y análisis predictivo altamente escalables en la nube. El objetivo de BigML es ayudar a desarrollar un conjunto de servicios, dado que es fácil de usar y se integra perfectamente. BigML se ha utilizado en muchos estudios para análisis predictivo y aprendizaje automático (AA) debido a su robustez y simplicidad. Proporciona una interfaz intuitiva. Por ejemplo, un estudio sobre las características distintivas de las imágenes de huellas humanas ofrece un análisis profundo mediante BigML. La idea es aprovechar el concepto de la huella humana para la identificación personal mediante diversas reglas difusas para el análisis predictivo. Se verifica la calidad de los datos de 440 imágenes de huellas. Se han aplicado GPU para optimizar el rendimiento.

Además, Chiroma et al. [36] presentaron un análisis predictivo sobre el lugar más frecuente de dengue en Malasia para obtener alertas tempranas y concienciar a las personas que utilizan la plataforma BigML. El estudio se basa en el modelo de algoritmo de árbol de decisión, que se basa en BigML para facilitar la clasificación.

Además, Chiroma et al. [36] analizaron las características del juego y las estrategias de adquisición, retención y monetización como impulsores principales del éxito de las aplicaciones de juegos móviles.

4. Método propuesto

4.1 Conjunto de datos empleado

En este artículo, propusimos modelos de big data y aprendizaje automático basados en soluciones para el desarrollo de un sistema inteligente de predicción de accidentes de tráfico basado en diferentes fuentes de datos. La Figura 3 representa las redes de tecnología de acceso inalámbrico que conectan vehículos e Internet, así como la red heterogénea comúnmente conocida como Internet de los Vehículos. La figura muestra la representación del Internet de los Vehículos en un entorno distribuido a gran escala en términos de comunicación inalámbrica entre diversos dispositivos. El modelo del Internet de los Vehículos está integrado en la nube, equipado con un servidor informático de alto rendimiento con múltiples GPU, modelos de aprendizaje automático a gran escala y Apache Spark. En el primer pilar, este procesamiento se realiza mediante la captura en tiempo real de todos los conjuntos de datos entrantes, que se almacenan en el clúster del Sistema de Archivos Distribuido de Hadoop. Posteriormente, se recurrió al núcleo SparkMlLib para utilizar todas las funciones de análisis de LamBig Data preimplementadas. En el segundo pilar, se realizó un estudio analítico para agrupar las características importantes. En el tercer pilar, se realizó una selección de características y se generaron nuevos conjuntos de datos. En cuarto lugar, se utilizaron algoritmos de aprendizaje automático para definir las tasas de accidentes. Finalmente, las tasas de accidentes se enviaron a los vehículos. Este artículo se centra en la segunda etapa del esquema. Las fuentes de datos representadas para este sistema son: el asesor policial, las condiciones del tráfico, el radar automático, los datos del vehículo, las cámaras fijas o móviles, los datos del conductor, el clima u otros factores externos. Cada fuente del conjunto de datos puede integrarse en el sistema propuesto.

El objetivo de este artículo es triple. Primero, presentamos el conjunto de datos TRAFFIC ACCIDENTS_2019_LEEDS. Segundo, analizamos la calidad de este conjunto de datos para la tarea de clasificación de accidentes de tráfico. Tercero, ampliamos el estudio utilizando ANN, SVM y modelos de bosque aleatorio para el preentrenamiento de la tarea de clasificación de accidentes de tráfico mediante la exploración de un mayor número de modelos de aprendizaje automático.

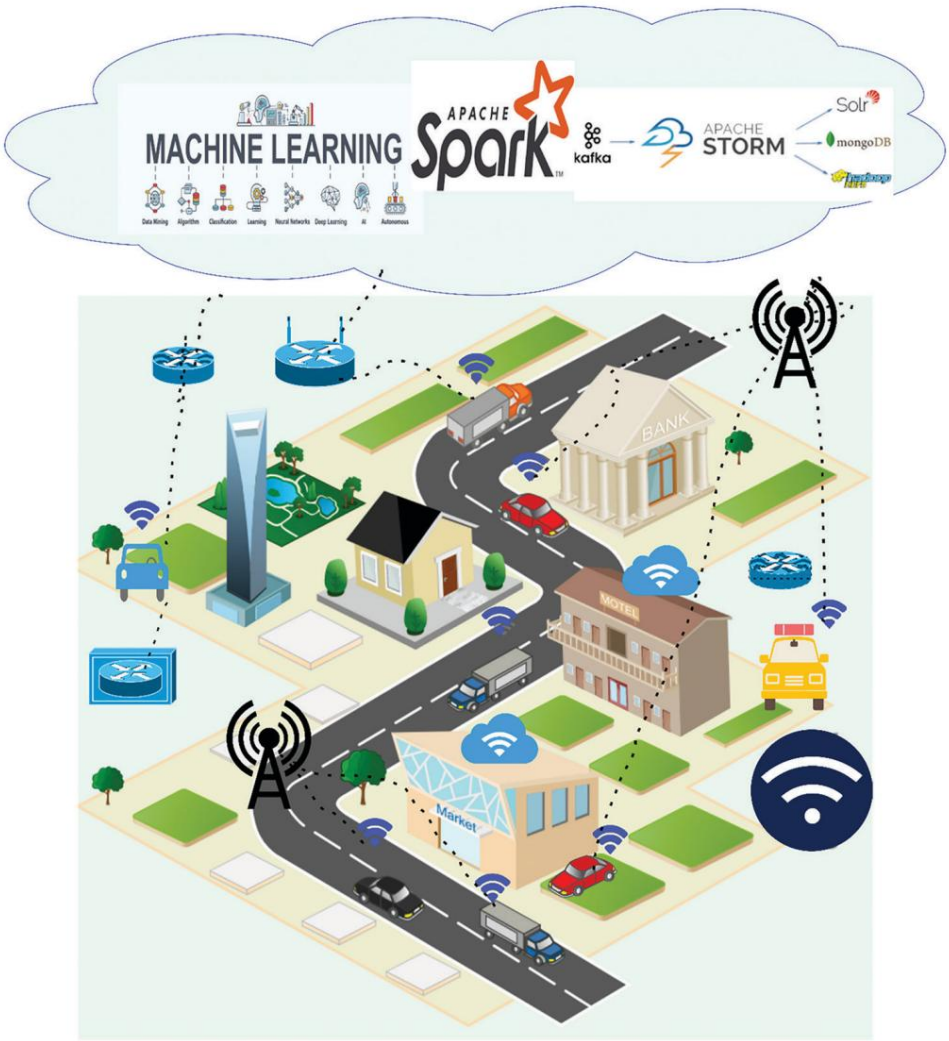

Fig. 3. Modelo de Internet de los Vehículos integrado en la nube equipado con un servidor de computación de alto rendimiento con múltiples GPU, modelos ML a gran escala y Apache

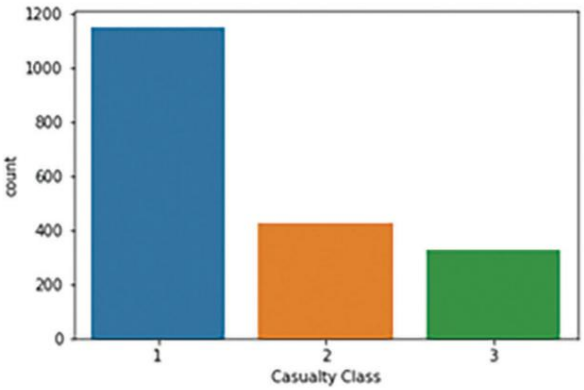

Figura 4. Distribución de la clase de víctimas para la predicción de la gravedad de los accidentes de tráfico.

En este estudio, utilizamos los datos de ACCIDENTES DE TRÁFICO_2019_LEEDS de la Oficina de Seguridad Vial del Departamento de Transporte. Las etiquetas de clasificación representan cada conjunto de datos. En esta base de datos, se registraron 1152 accidentes clasificados como de peatones, 405 como de conductores o motociclistas, y los 350 restantes como de vehículos o pasajeros (véase la Fig. 4).

Pestaña 1. Detalles completos del conjunto de datos

Tipo	Número de características
Peatonal	1152
Conductor o jinete	405
Vehículo o peatón pasajero	350
Total	1907

La Tabla 1 presenta los detalles del conjunto de datos y el número de características para peatón, vehículo o pasajero, o conductor o motociclista.

4.2 Equilibrio de la base de datos

Como se muestra en la siguiente figura (Fig. 4), la base de datos no está balanceada, porque el número de cada clase es bastante diferente (1 es peatón, 2 es vehículo o pasajero, 3 es conductor o motociclista).

Para equilibrar la base de datos, hay dos posibilidades: realizar un muestreo ascendente, o remuestrear los valores para hacer que su recuento sea igual a la etiqueta de clase con el recuento más alto, o realizar un muestreo descendente, tomando n muestras de cada etiqueta de clase donde n = número de muestras en la clase con el menor recuento.

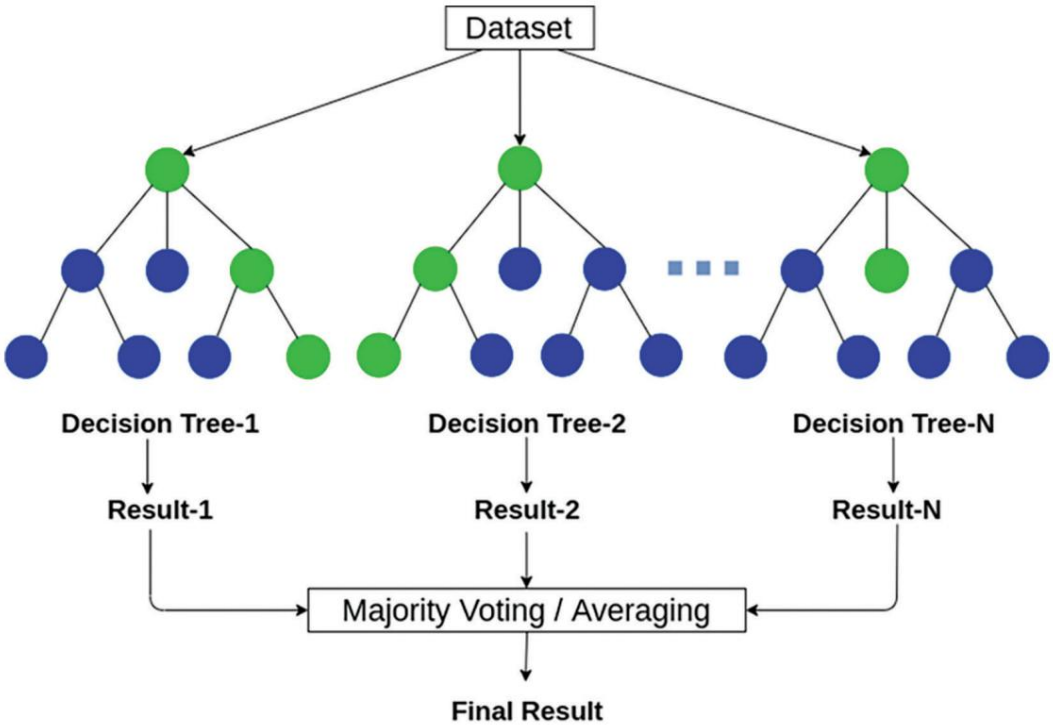

Fig. 5. Una descripción general del bosque aleatorio

Pestaña 2. Detalles del conjunto de datos después del aumento

Tipo	Número de características
Peatonal	1152
Pasajero del vehículo o del acompañante	1152
Conductor o jinete	1152
Total	3456

En este estudio, optamos por ampliar la base de datos. Obtuvimos 1152 registros para cada clase, para un total de 3456 registros después del aumento (ver Tabla 2).

Luego, dividimos la base de datos en dos partes: una de entrenamiento (Conjunto de Datos de Entrenamiento) y otra de prueba (Conjunto de Datos de Prueba). Utilizamos el 80% de la base de datos para el entrenamiento y el 20% para las pruebas: es decir, 2764 características para peatones, vehículos o pasajeros, o conductores o conductores para el conjunto de entrenamiento, y 692 características para el conjunto de prueba, como se muestra en la Tabla 2. Este procesamiento se realizó capturando en tiempo real todos los conjuntos de datos entrantes almacenados en el clúster del Sistema de Archivos Distribuido de Hadoop. Después de esto, llamamos al núcleo SparkMLLib para utilizar todas las funciones de análisis de datos LamBig preimplementadas. Utilizamos una RNA que consta de una capa de entrada. La figura 5 muestra el algoritmo de bosque aleatorio, que combina la salida de múltiples árboles de decisión (creados aleatoriamente) para generar el resultado final.

5. Resultados experimentales y discusión

5.1 Métricas de evaluación

La precisión es un criterio para evaluar los modelos de clasificación. Informalmente, la precisión se refiere a la precisión de nuestro modelo. Porcentaje de predicciones verdaderas. La definición formal de precisión es la siguiente:

$$\text{Exactitud} = \frac{TP\ TN +}{TP\ \cancel{TN}+FP\ FN} \quad (1)$$

Los determinantes son verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN).

La precisión es el porcentaje de positivos detectados con éxito en relación con todos los positivos esperados. Matemáticamente:

$$\text{Precisión} = \frac{TP}{TP\ \cancel{+}FP} \quad (2)$$

Donde TP denota Verdadero Positivo (número de predicciones positivas correctas) y FP denota Falso Positivo (cantidad de predicciones positivas mal clasificadas). La recuperación es el número total de predicciones positivas correctas en todas las muestras positivas. Matemáticamente:

$$\text{Recordatorio} = \frac{TP}{TP\ \cancel{+}FN} \quad (3)$$

Donde TP denota Verdadero Positivo (número de predicciones positivas correctas) y FN denota Falso Negativo (número de predicciones negativas incorrectas).

La puntuación F1 representa la relación simbiótica entre precisión y recuperación. Para datos no balanceados, la puntuación F1 es un estadístico de rendimiento superior a la métrica de precisión [37].

$$F_1\ 2 = \times \frac{\text{Recuperación de precisión}}{\text{Recuperación de precisión}} \quad (4)$$

5.2 Entorno experimental

Comparamos el rendimiento de nuestro modelo con el rendimiento de los enfoques ANN, SVM y RF.
Este procesamiento se realiza capturando en tiempo real todos los conjuntos de datos entrantes, que se almacenan en el clúster del Sistema de Archivos Distribuido de Hadoop. Posteriormente, se recurrió al núcleo SparkMlLib para utilizar todas las funciones de análisis de datos preimplementadas de LamBig. Se conservó la proporción de clases entre peatones, vehículos o pasajeros, y conductores o conductores, dividiendo aleatoriamente los conjuntos de datos en datos de entrenamiento y de prueba. Cada modelo que probamos se entrenó con datos de entrenamiento, mientras que su rendimiento se evaluó con datos de prueba. Para garantizar la consistencia del modelo, se realizó una validación cruzada de 10 pasos en cada uno de los modelos.

Para comparar los resultados con nuestro sistema, utilizamos los clasificadores ANN, SVM y RF del conjunto de datos TRAFFIC ACCIDENTS_2019_LEEDS. Los algoritmos se crearon utilizando el kit de herramientas Python scikit-learn y la configuración de hiperparámetros proporcionada.

La Fig. 6 representa la matriz de confusión para peatones, vehículos o pasajeros, o conductores o motociclistas utilizando el modelo de RNA. El rendimiento del modelo de RNA para el conjunto de datos de prueba se evaluó tras la fase de entrenamiento y se comparó mediante diversas medidas de rendimiento: precisión (VPP), sensibilidad o recuerdo, especificidad, área bajo la curva (AUC) y puntuación F1. La Fig. 7 también presenta la matriz de confusión para la clasificación de peatones, vehículos o pasajeros, o conductores o motociclistas utilizando el modelo de bosque aleatorio y el modelo SVM. El modelo que representa la Fig. 8 es la curva de precisión de entrenamiento y validación.

$$\text{CENTÍMETRO} = \begin{matrix} TP & FP \\ FN & TN \end{matrix} \quad (5)$$

6. Discusión

Para predecir la gravedad de los accidentes de tráfico, propusimos una solución basada en big data y modelos de aprendizaje automático. Este procesamiento se realizó capturando en tiempo real todos los conjuntos de datos entrantes almacenados en el clúster del Sistema de Archivos Distribuido de Hadoop. Posteriormente, recurrimos al núcleo SparkMlLib para utilizar todas las funciones de análisis de Big Data de Lam preimplementadas. Utilizamos clasificadores ANN, SVM y RF. Los resultados de los modelos en términos de exactitud, precisión, recuperación y puntuación F1 se calculan a partir de las matrices de confusión. En cuanto a exactitud, precisión, recuperación y puntuación F1, la Tabla 3 muestra los resultados de los modelos en el conjunto de datos.

Para este conjunto de datos, el clasificador Random Forest supera a los demás modelos en cuanto a precisión, exactitud y puntuación F1. Si bien el clasificador ANN presenta la mejor recuperación, presenta un rendimiento deficiente en los demás criterios de rendimiento para este conjunto de datos. En comparación con RF, ANN y SVM.

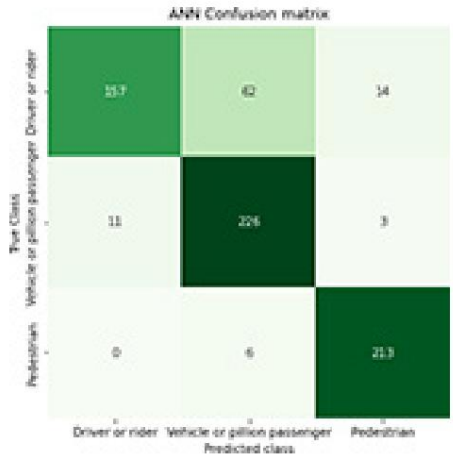

Fig. 6. Matriz de confusión para la clasificación de peatón, vehículo o pasajero, o conductor o motociclista utilizando el modelo ANN.

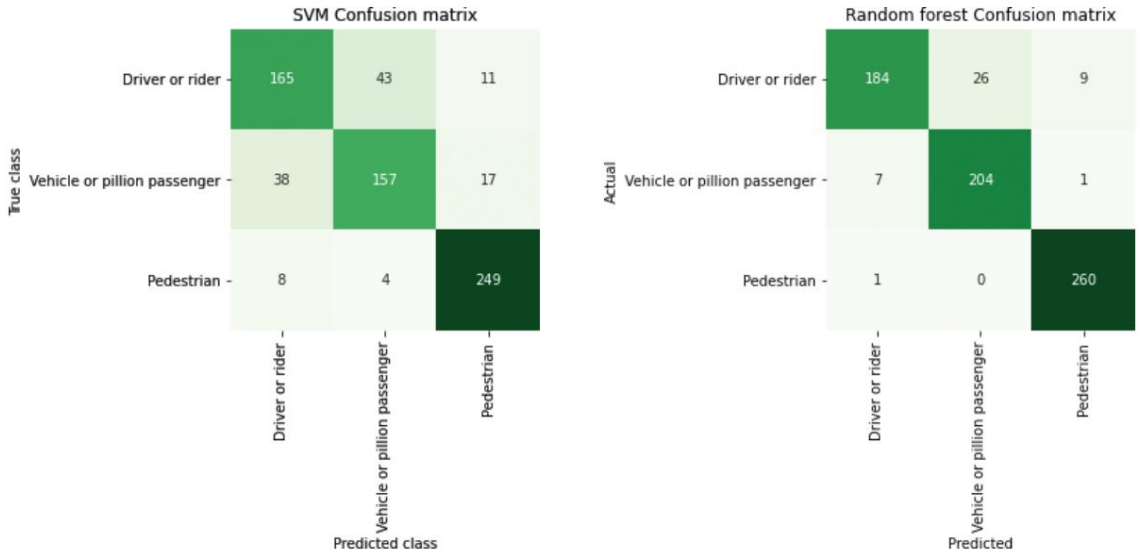

Fig. 7. Matriz de confusión para la clasificación de peatón, vehículo o pasajero, o conductor o motociclista utilizando el modelo Random Forest (a) modelo SVM y (b) modelo Random Forest.

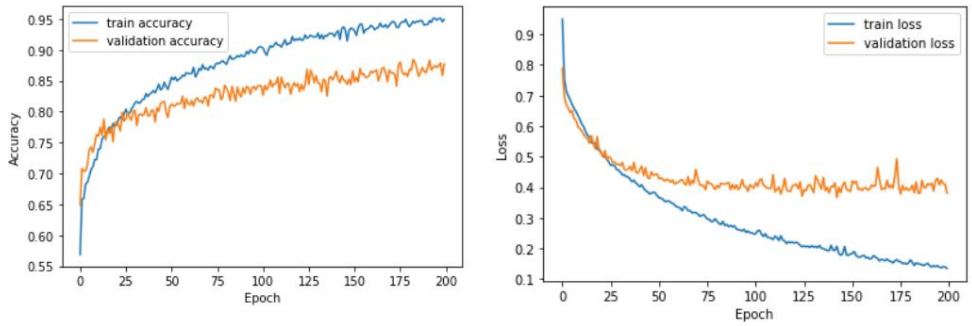

Fig. 8. Curva de precisión de entrenamiento y validación (a) Curva de precisión de entrenamiento y validación, (b) Curva de pérdida de entrenamiento y validación.

Tabla 3. Valores obtenidos para las diferentes métricas

	KNN	Bosque aleatorio	SVM	ANA
Exactitud	0,62	0.9364161849710982	0.8251445086705202	0.8771676300578035
precisión	0,38	0.9382125952919493	0.8222978546756788	0.8788867858874835
Recordar	0,27	0.9364161849710982	0.8251445086705202	0.8771676300578035
Puntuación de F1	0,28	0.9355102879655588	0.8232424765175214	0.8770074547672448

Como se muestra en la Tabla 3, el modelo con mejor rendimiento para detectar el estado fatal es el modelo Random Forest, que arrojó mejores valores de precisión (93,64 % para precisión de entrenamiento y 93,82 % para precisión de prueba).

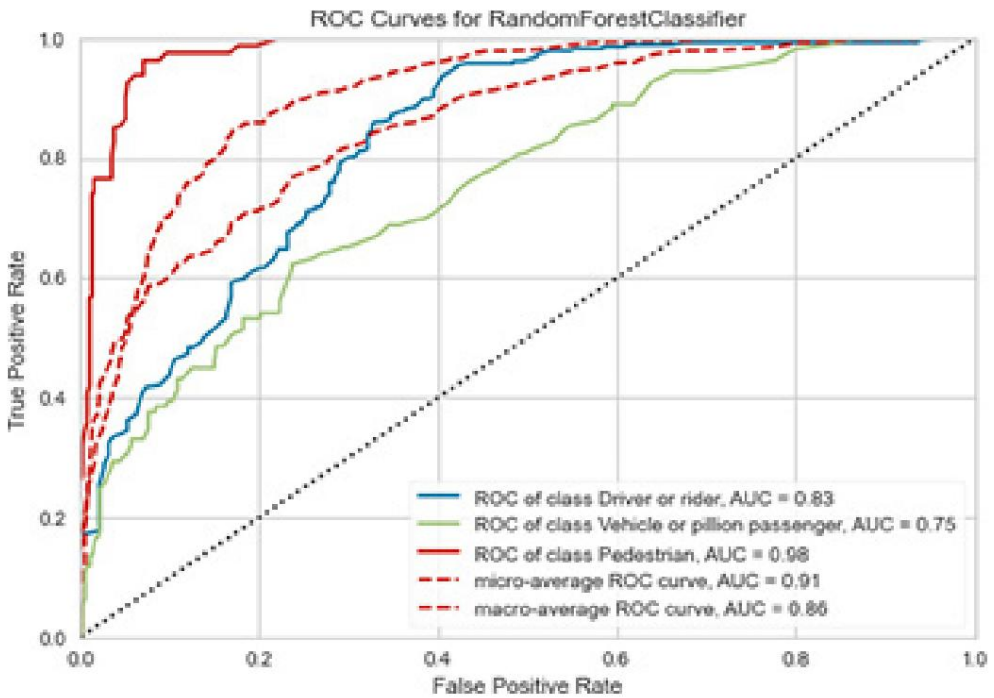

Fig. 9. Curva ROC del clasificador para el clasificador Random Forest.

Los clasificadores funcionan admirablemente. En comparación con RF, ANN y SVM son menos precisos. Sin embargo, en comparación con otros enfoques, SVM no logra una puntuación F1 ni una puntuación de recuperación satisfactorias, a pesar de que la precisión es correcta en comparación con los clasificadores RF y ANN.

Finalmente, la figura 9 representa las curvas de características operativas del receptor (ROC) para cada clase en el modelo de bosque aleatorio, mostrando la tasa de verdaderos positivos versus la tasa de falsos positivos a medida que se establece el umbral de clasificación.

varió entre 0 y 1. La curva ROC para cada modelo es un promedio de 10 curvas de la validación cruzada décupl, determinada por la regla del trapezoide.

7. Conclusión

En este trabajo, proponemos una solución basada en big data y modelos de aprendizaje automático para la predicción de accidentes de tráfico. Este procesamiento se realiza mediante la captura en

En tiempo real, todos los conjuntos de datos entrantes se almacenan en el clúster del Sistema de Archivos Distribuido de Hadoop. A continuación, recurrimos al núcleo de SparkMlLib para utilizar todas las funciones preimplementadas de análisis de LamBig Data. A continuación, nos centramos en la predicción de la gravedad de los accidentes de tráfico, lo cual representa un gran avance en la gestión de accidentes viales. Posteriormente, este número proporciona información importante para el transporte logístico de emergencia. Finalmente, para evaluar la gravedad de los accidentes de tránsito, evaluamos su impacto potencial e implementamos procedimientos efectivos de gestión de accidentes.

En este estudio, implementamos algoritmos para clasificar la gravedad de los accidentes de tránsito y presentamos la matriz de confusión para especificar: peatón, vehículo o pasajero, o conductor o motociclista, mediante Bosque Aleatorio, Máquina de Vectores de Soporte y Redes Neuronales Artificiales. Para validar este experimento, se utilizó el conjunto de datos TRAFFIC ACCIDENTS_2019_LEEDS para clasificar la predicción de gravedad de los accidentes de tránsito en tres clases: peatón, vehículo o pasajero, o conductor o motociclista. En trabajos futuros, será posible utilizar más características y encontrar las mejores para las clasificaciones de datos reales en nuestra ciudad. Nuevamente, podemos extraer estas características seleccionadas del archivo del programa; además, podemos implementar el costo para la predicción de la gravedad de los accidentes de tránsito. La importante ventaja de usar el paradigma de big data es que mejora el procesamiento de datos y establece un buen índice de seguridad vial basado en la clasificación y el reconocimiento de accidentes de tránsito. Hemos llamado directamente y en tiempo real a las funciones de Machine Learning preimplementadas para clasificar y predecir el accidente de tráfico en tiempo real.

El nuevo objetivo y reto de este trabajo reside en el procesamiento de grandes flujos de datos en tiempo real. Esto permite el uso eficaz de bibliotecas muy avanzadas y un sistema más rápido. Los resultados obtenidos se han probado para la prevención de accidentes en diferentes tipos de zonas y carreteras.

AUTORES

Imad El Mallahi* – Estudiante de doctorado en análisis de Big Data, accidentes de tráfico, inteligencia artificial, Laboratorio LISAC, Departamento de Ciencias de la Computación, Facultad de Ciencias, Universidad Sidi Mohamed Ben Abdellah de Fez, Fez, Marruecos, imade.elmallahi@
usmba.ac.ma.

Jamal Riffi, Hamid Tairi, Mohamed Adnane Mahraz – Universidad Sidi Mohammed ben Abdellah, Facultad de Ciencias Dhar el Mahraz, Universidad Sidi Mohammed ben Abdellah, Facultad de Ciencias Dhar el Mahraz, laboratorio LISAC, Fez, Marruecos.

Abderahamane Ez-Zahout – Universidad Mohamed V, Facultad de Ciencias, Equipo de Sistemas de Procesamiento Inteligente y Seguridad (IPSS), Departamento de Ciencias de la Computación, Rabat, Marruecos.

*Autor correspondiente

Referencias

[1] Lesiones por Accidentes de Tránsito. Consultado: 18 de julio de 2018. [En línea]. Disponible: http://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] F. Zong, H. Xu y H. Zhang, Predicción de la gravedad de los accidentes de tráfico: comparación de los modelos de red bayesianos y de regresión, Math. Problems Eng., n.º 23, 2013, Art. n.º 475194.

[3] Shiran, G.; Imaninasab, R.; Khayamim, R. Análisis de la gravedad de accidentes en carreteras basado en un modelo de regresión logística multinomial, técnicas de árboles de decisión y redes neuronales artificiales: Una comparación de modelos. Sustainability 2021, 13, 5670.

[4] Abdel-Aty, M. Análisis de los niveles de gravedad de las lesiones de los conductores en múltiples ubicaciones utilizando modelos probit ordenados. J. Saf. Res. 2003, 34, 597–603.

[5] Sze, NN; Wong, SC. Análisis diagnóstico del modelo logístico para la gravedad de las lesiones de peatones en accidentes de tránsito. Accid. Anal. Prev., vol. 39, 2007. 1267–1278.

[6] Savolainen, PT; Mannering, F.; Lord, D.; Quddus, MA. Análisis estadístico de la gravedad de las lesiones por accidentes de tráfico: Revisión y evaluación de alternativas metodológicas. Accid. Anal. Prev., vol. 43, 2011, 1666–1676.

[7] Moghaddam, FR; Afandizadeh, S.; Ziyadi, M. Predicción de la gravedad de accidentes mediante redes neuronales artificiales. Int. J. Civ. Eng., vol. 9, 2011, 41–48.

[8] Taamneh, M.; Alkheder, S.; Taamneh, S. Técnicas de minería de datos para modelado y predicción de accidentes de tráfico en los Emiratos Árabes Unidos. J. Seguridad en el Transporte 2017, 9, 146–166. [CrossRef]

[9] Zheng, M.; Li, T.; Zhu, R.; Chen, J.; Ma, ZF; Tang, MJ; Cui, ZQ; Wang, Z. Predicción de la gravedad de los accidentes de tráfico: una red CNN basada en un enfoque de aprendizaje profundo. IEEE Access 2019, 7, 39897–39910.

[10] Breiman, L. Bosques aleatorios. Mach. Learn., vol. 45, 2001, 5–32.

[11] Lu, Z.; Long, Z.; Xia, J.; An, C. Un modelo de bosque aleatorio para la identificación de modos de viaje basado en datos de señalización de telefonía móvil. Sustainability, vol. 11, 2019, 5950.

[12] Evans, J.; Waterson, B.; Hamilton, A. Pronóstico de las condiciones del tráfico vial mediante un algoritmo de bosque aleatorio basado en el contexto. Transp. Plan. Technol., vol. 42, 2019, 554–572.

[13] Hamad, K.; Al-Ruzouq, R.; Zeiada, W.; Abu Dabous, S.; Khalil, MA Predicción de la duración del incidente utilizando bosques aleatorios. Transp.A-Transp. Sci. vol. 16, 2020, 1269–1293. [CrossRef]

[14] Macioszek, E. Cálculo de la capacidad de entrada a rotondas: un estudio de caso basado en rotondas en Tokio, Japón, y los alrededores de Tokio.
Sostenibilidad, vol. 12, 2020, 1533.

[15] Severino, A.; Pappalardo, G.; Curto, S.; Trubia, S.; Olayode, IO. Evaluación de la seguridad de la rotonda de flores considerando la operación de vehículos autónomos.
Sustainability, vol. 13, 2021, 10120.

[16] Macioszek, E. Comparación de modelos para la estimación de intervalos críticos en rotondas. En Actas de la 13.ª Conferencia Científica y Técnica sobre Sistemas de Transporte. Teoría y Práctica (TSTP), Katowice, Polonia, 19-21 de septiembre de 2016.

[17] Karpathy, G. Toderici, S. Shetty, T. Leung, R.
Sukthankar y L. FeiFei, Clasificación de video a gran escala con redes neuronales convolucionales, en Proc. IEEE Conf. Comput. Vis. Pattern Recognit., junio de 2014, 1725-1732.

[18] S. Lawrence, CL Giles, AC Tsoi y AD Back, Reconocimiento facial: un enfoque de red neuronal convolucional, IEEE Trans. Neural Netw., vol. 8, no. 1, 1997, 98113.

[19] MG Karlaftis y EI Vlahogianni, Métodos estadísticos versus redes neuronales en la investigación del transporte: diferencias, similitudes y algunas ideas, Transp. Res. C, vol. 19, no. 3, 2011, 387399.

[20] S. Al-Ghamdi, Uso de regresión logística para estimar la influencia de los factores de accidentes en los accidentes.
Gravedad de la abolladura, Accident Anal. Prevención, vol. 34, núm. 6, 2002. 729741.

[21] M. Bédard, GH Guyatt, MJ Stones y JP
Hirdes, «La contribución independiente del conductor, el accidente y las características del vehículo a las muertes de conductores», Accid Anal. Prevention, vol. 34, n.º 6, 2002, 717727.

[22] KM Kockelman y YJ Kweon, Gravedad de las lesiones del conductor: una aplicación de modelos probit ordenados, Accident Anal. Prevention, vol. 34, n.º 3, 2002, 313321.

[23] Mohamed AbdElAziz, Khamis Ahmed, El-Mahdy Ahmed El-Mahdy, Kholoud Osama Shata Kholoud Osama Shata, Walid Gomaa Walid Gomaa, Sistema y método para la detección y notificación de accidentes durante colisiones, Número de patente: 2020/771, Fecha de presentación: 9 de junio de 2020, Lugar de presentación: Egipto.

[24] Mohamed AbdElAziz Khamis, Ahmed El-Mahdy, Kholoud Osama Shata. Un sistema y método a bordo de un vehículo para la detección de accidentes sin fijado al vehículo, Número de patente: 2020/769, Fecha de presentación: 9 de junio de 2020, Lugar de presentación: Egipto.

[25] Mohamed A. Khamis, Walid Gomaa, Aprendizaje de refuerzo multiobjetivo adaptativo con exploración híbrida para el control de señales de tráfico basado en un marco cooperativo multiagente, Aplicaciones de ingeniería de inteligencia artificial, vol. 29, marzo de 2014, 134–151 https://doi.org/10.1016/j.
engappai.2014.01.007.

[26] Mohamed AbdElAziz Khamis, Walid Gomaa, Aprendizaje por refuerzo multiagente y multiobjetivo para Mejorado el control de semáforos urbanos, Actas de la 11.ª Conferencia internacional IEEE sobre aprendizaje automático y aplicaciones (ICMLA 2012), Boca Raton, Florida, EE. UU., 12-15 de diciembre de 2012, págs. 586-591.

[27] Mohamed A. Khamis , Walid Gomaa, Hisham El-Shishiny, Sistema de control de semáforo multiobjetivo basado en la interpretación de probabilidad bayesiana, Actas de la 15.ª Conferencia sobre sistemas de transporte inteligente del IEEE (ITSC 2012), Anchorage, Alaska, EE. UU., 16-19 de septiembre de 2012, págs. 995-1000.

[28] Zhang XG. Introducción a la teoría del aprendizaje estadístico y máquinas de vectores de soporte. Acta Automat Sinica, vol. 26, 2000, 32–41.

[29] Yuan F y Cheu RL. Detección de incidentes mediante máquinas de vectores de soporte. Transp Res Part C Emerg Technol. vol. 11, 2003, 309–328.

[30] Sharma B, Katiyar VK y Kumar K. Modelo de predicción de accidentes de tráfico mediante máquinas de vectores de soporte con núcleo gaussiano. En: Pant M, Deep K, Bansal JC, et al. (eds.), Actas de la quinta conferencia internacional sobre computación blanda para la resolución de problemas, vol. 437. Singapur: Springer 2016, pp. 1-10.

[31] Flores MJ, Armingol JM y de la Escalera A.
Sistema de alerta en tiempo real para la detección de somnolencia del conductor mediante información visual. J Intell Robot Syst., vol. 59, 2010, 103–125.

[32] Delen D, Sharda R y Bessonov M. Identificación de predictores significativos de la gravedad de las lesiones en accidentes de tráfico mediante una serie de redes neuronales artificiales. Acc Anal Prevent., vol. 38, 2006, 434–444.

[33] Alikhani M, Nedaie A y Ahmadvand A.
Presentación del método heurístico de agrupamiento y clasificación para mejorar la precisión en la clasificación de la gravedad de los accidentes de tráfico en Irán. Safe Sci., vol. 60, 2013, 142–150.

[34] Lee SL. Predicción de la gravedad de los accidentes de tráfico mediante técnicas de clasificación. Adv Sci Lett., vol. 21, 2015, 3128–3131.

[35] Ez-zahout A., Un modelo distribuido de análisis de big data para la reidentificación de personas basada en la reducción de la dimensionalidad. Revista Internacional de Arquitectura de Sistemas de Alto Rendimiento. vol. 10, n.º 2, 2021, 57–63.

[36] Haruna Chiroma, Shafi'i M. Abdulhamid, Ibrahim AT Hashem, Kayode S. Adewole, Absalom E. Ezugwu, Saidu Abubakar y Liyana Shuib. Análisis de big data basado en aprendizaje profundo para el Internet de los vehículos: taxonomía, desafíos y direcciones de investigación. Hindawi, Problemas matemáticos en ingeniería, vol.

2021, ID del artículo 9022558, 20 páginas, https://doi.org/10.1155/2021/9022558

[37] Asmae Rhanizar, Zineb El Akkaoui, Un marco predictivo de la ubicación de radares de velocidad para la seguridad vial. Ciencias de la Computación y la Información, vol. 12, n.º 3, 2019. URL: https://doi.org/10.5539/jmr.v12n3p92

[38] Salma Bouaich, Mahraz, MA, Riffi, J. et al. Conteo de vehículos basado en líneas de carretera. Reconocimiento de patrones. Image Anal., vol. 31, 2021, 739–748. https://doi.org/10.1134/S1054661821040076

[39] Bouti, A., Mahraz, MA, Riffi, J. et al. Un sistema robusto para la detección y clasificación de señales de tráfico mediante la arquitectura LeNet basada en una red neuronal convolucional. Soft Comput., vol. 24, 2020, 6721. https://doi.org/10.1007/s00500-019-04307-6

[40] Bouaich, S., Mahraz, MA, Riffi, J., Tairi, H., Detección de vehículos mediante líneas de carretera, 3.ª Conferencia internacional sobre computación inteligente en ciencias de datos, ICDS, 2019, 8942305

[41] Bouaich, S., Mahraz, MA, Rifii, J., Tairi, H., Sistema de conteo de vehículos en tiempo real, Conferencia internacional sobre sistemas inteligentes y visión artificial de 2018, ISCV 2018, mayo, págs. 1–4