Ciberseguridad y aplicaciones 3 (2025) 100063

Listas de contenidos disponibles enCiencia Directa

Ciberseguridad y aplicaciones

Página de inicio de la revista:http://www.keaipublishing.com/es/revistas/ciberseguridad-y-aplicaciones/

Un nuevo enfoque para gestionar datos faltantes para mejorar el sistema de detección de intrusiones en la red

Mahjabeen Tahir[a,b,*], Azizol Abdullah[a], Nur Izura Udzir[a], Khairul Azhar Kasmiran[a]

[a] Departamento de Ciencias de la Computación y Tecnología de la Información, Universidad Putra (UPM), Serdang 43400, Selangor, Malasia
[b] Departamento de Ingeniería Informática, Universidad Sir Syed de Ingeniería y Tecnología, Karachi 75300, Sindh, Pakistán

información del artículo

Palabras clave:
Detección de intrusiones
Datos faltantes Imputación
Aprendizaje automático
Codificador automático
Aumento de gradiente
Redes neuronales

abstracto

La gestión de datos faltantes es un desafío crítico en los conjuntos de datos de sistemas de detección de intrusiones (IDS), que afecta significativamente el rendimiento de los modelos de aprendizaje profundo. Para abordar este problema, presentamos DeepLearning_Based_MissingData_Imputation (DMDI), un novedoso método diseñado para mejorar la calidad de los datos de entrada mediante la gestión eficiente de los valores faltantes. Nuestro enfoque emplea el algoritmo Random Missing Value (RMV) para simular datos faltantes, lo que permite realizar pruebas exhaustivas y comparar diversas técnicas de imputación. El método DMDI integra un autocodificador de reducción de ruido apilado con potenciación de gradiente para mejorar la precisión de la imputación. Evaluamos la eficacia de nuestro enfoque mediante tres fases experimentales: generación de datos faltantes, imputación de valores faltantes y evaluación de los modelos de imputación. Utilizando los conjuntos de datos NSL-KDD y UNSW-NB15, nuestros resultados demuestran mejoras significativas en el rendimiento de cinco clasificadores diferentes (SVM, KNN, regresión logística, árbol de decisión y bosque aleatorio) tras la imputación. En promedio, nuestro método logró mejoras de precisión de entre 0,95 y 0,97 en estos clasificadores, en comparación con los métodos de imputación de referencia. Un análisis detallado con Python 3 valida nuestros hallazgos, demostrando una mejora en el rendimiento y la robustez del modelo. Este estudio subraya la necesidad de una imputación precisa de datos faltantes para optimizar las tareas de aprendizaje profundo, en particular en sistemas de detección de anomalías. Proporciona una solución fiable para la gestión de datos faltantes en conjuntos de datos IDS.

1. Introducción

La creciente complejidad de la interconexión de redes ha provocado un aumento de ataques avanzados a la seguridad, lo que dificulta la seguridad de la información que transporta. Con el creciente uso de dispositivos, la complejidad de la seguridad de la red también ha aumentado, lo que presenta obstáculos considerables para su mantenimiento. Numerosos sistemas de detección de intrusiones (IDS) actualmente en uso se basan en métodos de detección basados en firmas, que examinan patrones de archivos o entradas de datos típicas.1Debido al creciente número de dispositivos y componentes, existe una creciente demanda de sistemas de detección de intrusiones resistentes y capaces de detectar irregularidades sutiles. A medida que crece el número de equipos, crece la demanda de sistemas robustos de detección de intrusiones capaces de identificar anomalías sutiles. Si bien un sistema de detección de intrusiones no sustituye las medidas preventivas, actúa como una última línea de defensa vital para proteger el sistema.2]. Diversos sectores, como las actividades fraudulentas [3], programas de bienestar [4], diagnóstico de fallas [5], y sistemas de intrusión y prevención [6], dependen del IDS para la recopilación y el análisis de datos con el fin de identificar actividades anormales. Sin embargo, lograr un alto nivel de precisión en la identificación de anomalías y, al mismo tiempo, minimizar las tasas de falsos positivos sigue siendo un desafío.

Los datos incompletos introducen sesgo en la calidad de la clasificación del conjunto de datos [7], especialmente en el contexto de datos desequilibrados. Abordar este problema es esencial para mejorar el rendimiento del reconocimiento de patrones. Estrategias como el análisis completo de casos o la recuperación de valores faltantes a partir de instancias observadas son eficaces para mitigar el problema de los datos faltantes, como se menciona en la referencia [8La recuperación de datos faltantes suele ofrecer mejores resultados que la eliminación. La interpolación, la imputación y la completación de matrices son tres enfoques comunes para estimar los datos faltantes.9, 10].

Existen diversas técnicas de imputación de aprendizaje automático y profundo, cada una con ventajas sobre las demás. Por ejemplo, los k vecinos más cercanos (KNN) representan un método sencillo y basado en instancias.11]. Los valores faltantes en KNN se miden considerando las k instancias más cercanas en términos de distancia [12Sin embargo, en escenarios con altas tasas de datos faltantes, emplear el método KNN se vuelve un desafío. Por el contrario, algunas metodologías ajustan estas distancias mediante métricas como la información mutua (IM) o el coeficiente relacional de Gray (GRA) para abordar estos desafíos.1, 13]

La construcción de modelos de aprendizaje automático puede ser un desafío cuando se trabaja con datos faltantes, especialmente cuando los conjuntos de datos de entrenamiento y prueba contienen valores faltantes [14]. Esto puede provocar una distorsión de la articulación.

Revisión por pares bajo responsabilidad de KeAi Communications Co., Ltd.
*Autor correspondiente.
*Dirección de correo electrónico:*engr.mahjabeen.tahir@gmail.com (M. Tahir).

https://doi.org/10.1016/j.csa.2024.100063
Recibido el 7 de marzo de 2024; Recibido en forma revisada el 12 de junio de 2024; Aceptado el 27 de junio de 2024.
Disponible en línea el 1 de julio de 2024.

distribución, que se complica aún más cuando los datos comprenden diferentes tipos, como componentes nominales, ordinales, binarios y continuos [15,16]. Aunque solo hay un número limitado de algoritmos de aprendizaje supervisado que pueden manejar valores faltantes, los modelos basados en árboles son una excepción notable [17,18El método predominante para abordar los valores faltantes generalmente implica completar los datos no disponibles para generar un conjunto de datos completo.

Se ha dedicado una amplia investigación a explorar los métodos de imputación y sus aplicaciones prácticas [16,19,20], especialmente en escenarios donde los valores faltantes contribuyen al desequilibrio de los datos. En tales casos, los modelos predictivos pueden sufrir sobreajuste en conjuntos de entrenamiento desequilibrados, lo que puede resultar en un rendimiento inferior en conjuntos de prueba o aplicaciones reales. Cuando existen valores faltantes en los datos que están altamente correlacionados con la variable objetivo, esto puede generar problemas como fugas de datos y causalidad inversa en las predicciones del modelo. Esto suele deberse a deficiencias en la recopilación de datos y puede suponer un desafío significativo durante las pruebas, incluso si el modelo ha tenido un buen rendimiento durante el entrenamiento.

1.1. Motivación

Seleccionar el modelo de imputación más adecuado es vital para abordar los problemas causados por valores faltantes, como posibles fugas de datos o desequilibrios. El uso de la Imputación Múltiple por Ecuaciones Encadenadas (MICE) [21] representa un enfoque valioso para gestionar datos faltantes. Este método implica generar varios conjuntos de datos mediante simulaciones de Monte Carlo y luego imputar los valores faltantes mediante muestreo de Gibbs. Otras técnicas, como la imputación múltiple mediante autocodificadores de reducción de ruido (MIDA), [19] y las redes de imputación generativa adversarial (GAIN) también han demostrado resultados prometedores en la gestión de la imputación de valores faltantes.

Existe un conocimiento limitado sobre la eficacia de los sistemas de detección de intrusiones (IDS) tradicionales para detectar ciberataques modernos. Esto se debe a la falta de conjuntos de datos fiables. El estudio destaca un problema importante en los conjuntos de datos de IDS: la ausencia de datos [20El progreso en ciberseguridad se ve mejorado gracias a la tecnología de Inteligencia Artificial (IA). Sin embargo, esto también crea nuevas oportunidades para que los ciberatacantes lleven a cabo ataques. Los sistemas de detección de intrusos (IDS) se han utilizado ampliamente para contrarrestar estas amenazas, pero existen desafíos con datos incompletos y desequilibrados, lo que dificulta el entrenamiento de los IDS. Los métodos de aprendizaje profundo, en particular los autocodificadores, se han convertido en un enfoque prometedor para completar los datos faltantes. El enfoque DMDI ofrece un método más preciso y aprendido para imputar valores faltantes que las técnicas tradicionales, lo que mejora el rendimiento de los IDS.

1.2. Contribución

La calidad del conjunto de datos influye en la eficacia del IDS. Los conjuntos de datos desequilibrados causan problemas a los investigadores. Este artículo presenta una técnica denominada DMDI, basada en un método de aprendizaje profundo para completar los valores ausentes en los conjuntos de datos IDS. Este enfoque no se ha aplicado anteriormente a conjuntos de datos IDS desequilibrados.

Este artículo amplía la investigación presentada en [22]. Sus principales contribuciones son las siguientes:

- **Implementación del algoritmo RMV**Utilizamos el algoritmo de Valores Aleatorios Perdidos (RMV) para inyectar sistemáticamente valores faltantes en el conjunto de datos. Esto permite evaluar y comparar de forma estructurada diferentes métodos para gestionar los datos faltantes durante el preprocesamiento.
- **Desarrollo del modelo DMDI**Presentamos el modelo de imputación de datos faltantes basada en aprendizaje profundo (DMDI), que emplea una arquitectura de autocodificador de eliminación de ruido apilado (SDA). Esta técnica de aprendizaje no supervisado se optimiza con un método de conjunto, el refuerzo de gradiente, para refinar la imputación inicial y mejorar la calidad de los datos en los conjuntos de datos IDS.
- **Evaluación del desempeño**:La eficacia del modelo DMDI se prueba mediante simulaciones exhaustivas en dos conjuntos de datos IDS de referencia, NSL-KDD y UNSW-NB15. Estas pruebas implican escenarios con pérdidas de valores aleatorios para simular condiciones reales de pérdida de datos.
- **Comparación completa**Realizamos una comparación exhaustiva de DMDI con otros modelos de aprendizaje automático destacados. Este análisis evalúa el impacto de diferentes estrategias de imputación en el rendimiento del modelo, lo que proporciona información sobre las fortalezas y debilidades de los distintos enfoques.

OrganizaciónEl documento está dividido en varias secciones, cada una de las cuales abarca un tema diferente.Sección 2Proporciona una breve descripción general de la investigación relacionada con los sistemas de detección de intrusiones (IDS), incluidos los diversos conjuntos de datos y técnicas utilizadas.Sección 3profundiza en el algoritmo RMV, así como en la metodología utilizada para imputar datos faltantes. EnSección 4El artículo examina el modelo de seguridad y el análisis funcional, complementados con experimentación en situaciones reales. Esto incluye la evaluación de la complejidad computacional del algoritmo y el análisis de las implicaciones del método propuesto para la investigación y la práctica. Además, esta sección describe las limitaciones de nuestro trabajo. Finalmente,Sección 5concluye el artículo.

2. Trabajos relacionados

En esta sección, analizaremos brevemente los datos faltantes y las técnicas de imputación. Para un análisis más detallado, consulte [23]. En este trabajo se presenta un análisis comparativo de diferentes técnicas de imputación, destacando sus ventajas y desventajas.Tabla 1.

Gestionar flujos de datos incompletos y desequilibrados es un desafío, ya que requiere imputar datos faltantes y equilibrar los datos desequilibrados, que pueden verse afectados por la deriva conceptual. Los investigadores han explorado diversos métodos para abordar estos problemas de forma independiente. En lo que respecta a los datos faltantes, existen dos metodologías comunes: el análisis completo de casos y la recuperación de valores faltantes. El análisis completo de casos implica utilizar únicamente instancias observadas, lo que puede resultar en la omisión de información importante, especialmente cuando existe una alta tasa de datos faltantes.

Usar una técnica adecuada de imputación de datos faltantes suele ser mejor que simplemente eliminar los casos incompletos. Existen varios métodos para imputar datos faltantes, incluyendo enfoques simples, múltiples, fraccionales o iterativos. Las técnicas de imputación simple, como la imputación de media o moda, así como métodos como la imputación de baraja caliente (HD) y la de baraja fría (CD), buscan reemplazar los valores faltantes utilizando datos observados o fuentes externas.24Si bien la imputación única puede reducir la incertidumbre, la imputación múltiple restablece las validez naturales, pero requiere el conocimiento de los mecanismos faltantes. Por el contrario, los métodos de imputación iterativos utilizan toda la información disponible, incluidos los datos faltantes, para imputar valores.

Los métodos de imputación estadística, como la Maximización de Expectativas (EM) y la imputación de regresión (IR), junto con las tecnologías de soft computing y aprendizaje automático, ofrecen soluciones integrales para gestionar la falta de datos. Por ejemplo, EM emplea un algoritmo iterativo útil en el aprendizaje automático y la minería de datos.25]. Mientras tanto, RI emplea imputación de regresión múltiple con métodos paramétricos y no paramétricos, cuya eficacia depende del modelado paramétrico preciso de datos incompletos.

Se emplean varias técnicas de aprendizaje automático para manejar los datos faltantes, incluidos los K vecinos más cercanos (KNN) [12], árboles de decisión [12], mapas autoorganizados (SOM) [18] y máquinas de vectores de soporte (SVM) [26]. SOM se entrena inicialmente sin datos faltantes y posteriormente se utiliza para la imputación [27]. DT demuestra versatilidad en la gestión de datos numéricos y categóricos para tareas de imputación. KNN identifica los k vecinos más cercanos de las muestras faltantes mediante métricas de distancia [12]. Métodos avanzados de imputación KNN, como la imputación KNN secuencial (SKNN) y la imputación KNN iterativa (IKNNI) [28], también se han desarrollado para proporcionar resultados más fiables al gestionar datos faltantes. También se ha propuesto el uso de métricas de distancia ponderadas por características mediante MI para mejorar la precisión de la clasificación con valores faltantes estimados.

Tabla 1
Comparación de diferentes técnicas de imputación.

Técnicas de imputación	Metodología	Ventajas	Contras
Impuación única (moda/media, HD, CD) [24]	Reemplaza los valores faltantes utilizando datos observados o fuentes externas	Reduce la incertidumbre	Puede introducir sesgo, ignora la variabilidad
Impuación Múltiple [46]	Genera múltiples conjuntos de datos imputando valores faltantes varias veces	Restaura las variaciones naturales	Requiere conocimiento de los mecanismos faltantes, es computacionalmente intensivo,
Imputación iterativa (por ejemplo, MICE) [47]	Utiliza toda la información disponible para imputar iterativamente los valores faltantes. Algoritmo iterativo	Utiliza datos completos y más precisos	computacionalmente costoso, complejo de implementar
Maximización de expectativas (EM) [25]	que estima los valores faltantes.	Útil en ML y minería de datos	Requiere un modelado paramétrico adecuado, puede ser lento.
Imputación de regresión (RI) [48]	Utiliza modelos de regresión para predecir valores faltantes	Eficaz con el modelo paramétrico correcto	Depende de la especificación precisa del modelo.
K-Vecino más cercano (KNN) [12]	Identifica los k vecinos más cercanos para imputar valores faltantes	Simple y eficaz para distintos tipos de datos.	Sensible a la métrica de distancia, computacionalmente costoso
Mapas autoorganizados (SOM) [18]	Red neuronal entrenada utilizada para imputación	Puede manejar no lineal relaciones	Requiere entrenamiento inicial sin perder datos
Árbol de decisión (DT) [12]	Utiliza estructuras de árbol para predecir valores faltantes	Maneja tanto datos numéricos como categóricos	Propenso al sobreajuste, puede ser inestable.
KNN secuencial(SKNN) [28]	Mejora KNN con imputación secuencial	Más confiable que el KNN estándar	Más intensivo computacionalmente que el KNN estándar
KNN iterativo (IKNN) [28]	Refina iterativamente los valores imputados utilizando KNN	Mejora la precisión con respecto a KNN de un solo paso	Altamente intensivo en computación
Análisis relacional de Gray (GRA) [12]	Establece asociaciones utilizando coeficientes relacionales de Gray	Puede capturar imágenes complejas relaciones	Requiere un ajuste cuidadoso de los parámetros

Del mismo modo, los investigadores [13] Se investigó la importancia de las características y las métricas GRA ponderadas por MI para recuperar valores faltantes. GRA establece asociaciones entre una observación de referencia y una observación de comparación mediante el coeficiente relacional de Gray (GRC) y el grado relacional de Gray (GRG).12]. Por el contrario, otro estudio [11] utilizaron la distancia de gris para identificar los vecinos más cercanos de los datos faltantes. Estos enfoques, en conjunto, buscan identificar los datos faltantes con mayor precisión en los conjuntos de datos, y algunos métodos muestran un mejor rendimiento que otros. Sin embargo, ninguno de estos métodos aborda eficazmente el problema del desequilibrio de datos dentro de los conjuntos de datos.

Algoritmo 1
Algoritmo RMV.

```
InjectMissingData (número de celdas, número de columnas)
    celdas por columna←núm. de celdas/núm. de columnas
    celdas adicionales←núm. celdas% núm. columnas para i en
el rango (número de columnas)hacer
  si i<células adicionalesentonces
    Agregar un valor faltante adicional a la columna actual
  terminar si
  Agregar celdas por columna con valores faltantes a la columna actual
fin para
```

3. Metodología propuesta

3.1. Inyección de datos faltantes mediante el algoritmo RMV

El algoritmo RMV sirve como una herramienta valiosa para evaluar diversos enfoques al comparar y evaluar diferentes métodos para gestionar datos faltantes en tareas de preprocesamiento y modelado de datos.22Para insertar datos faltantes aleatorios en un archivo CVS, utilizamos el algoritmo RMV. Diversos estudios han contribuido al campo de los datos faltantes [29-34]. Para implementar la función RMV, los requisitos son los siguientes: i) el número de columnas en un archivo CSV donde se deben insertar los valores faltantes, ii) la proporción de valores ausentes a incorporar, y iii) una lista de títulos de columnas a omitir del procedimiento, si corresponde.

El proceso comienza calculando la cantidad de valores ausentes en el conjunto de datos. Posteriormente, identifica las columnas del conjunto de datos que requieren la inserción de valores faltantes, colocándolos aleatoriamente en los índices de fila especificados dentro de esas columnas.4]. Finalmente, se guarda un nuevo archivo CSV con los valores faltantes agregados.

3.2. Procedimiento de implementación del RMV

Varios algoritmos, especialmente aquellos diseñados para la imputación, carecen de directrices precisas para identificar qué valores deben designarse como no disponibles. Sin embargo, su objetivo principal es llenar estos vacíos aproximándolos a la distribución de datos utilizando datos observados y presunciones específicas. Estos algoritmos aprovechan los patrones de diseño existentes y la conexión inherente a la entrada para realizar estimaciones fundamentadas sobre los valores no disponibles. Por ejemplo, para predecir valores faltantes en la imputación basada en regresión, se utilizan las relaciones entre variables, mientras que algunos modelos imputan basándose en similitudes entre observaciones, como los k vecinos más cercanos. Por otro lado, Los algoritmos de imputación basados en aprendizaje automático entrenan modelos observando datos para revelar diseños y conexiones, que luego se emplean para completar los datos faltantes.

Nuestra configuración experimental se centra en una situación en la que la inserción de valores faltantes ocurre dentro de un conjunto de datos de manera controlada [22Estos valores faltantes se cuantifican mediante un recuento total designado, almacenado en una variable denominada 'num_cells', y se distribuyen dentro de un subconjunto de columnas denominadas 'num_columns'. A modo de ejemplo, si se desea insertar 100 valores faltantes con num_cells = 100 en 4 columnas con num_columns = 4, el algoritmo distribuye aproximadamente el 44 % de los valores faltantes a cada una de las 4 columnas. Sin embargo, debido al valor restante de 1, una columna tendrá un valor faltante adicional. Las acciones subsiguientes describen el proceso de introducción de valores faltantes en el conjunto de datos e identifican qué datos se consideran faltantes. Los pasos precisos se explican enAlgoritmo 1.

3.3. Metodología de imputación

El método DMDI presenta una estrategia innovadora para gestionar la falta de datos en los conjuntos de datos utilizados en sistemas de detección de intrusiones. Normalmente, los métodos de imputación de valores faltantes se dividen en dos categorías: técnicas estadísticas y técnicas basadas en aprendizaje automático.35La técnica tradicional es la media/moda, que constituye el enfoque fundamental para la imputación de valores faltantes en los métodos estadísticos. Por otro lado, existen técnicas basadas en aprendizaje automático, como métodos como la agrupación en clústeres, el bosque aleatorio, el árbol de decisión y el modelo de k-vecinos más cercanos. La imputación incompleta de conjuntos de datos puede afectar significativamente los resultados. Se puede encontrar una descripción detallada de ambos métodos en [35Tras la imputación de los valores faltantes, es crucial evaluar los resultados de la imputación. En el método directo, los valores faltantes se reemplazan o

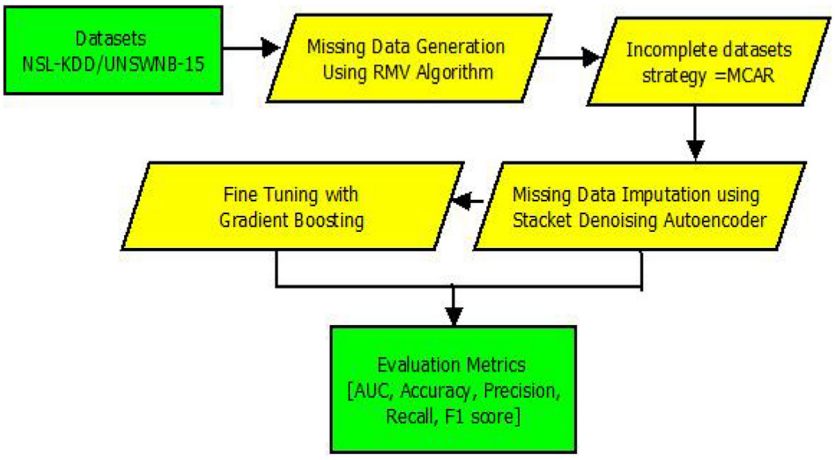

Figura 1. Metodología de nuestro Algoritmo.

Se imputan directamente con valores estimados. Estos valores estimados suelen derivarse de diversas técnicas o modelos estadísticos. Como alternativa, el enfoque de clasificación implica el entrenamiento de clasificadores seleccionados utilizando los conjuntos de datos imputados. Entre los modelos de aprendizaje automático más populares para evaluar los resultados de imputación se incluyen KNN, SVM, DT, NB y MLP [36].

Para mejorar la precisión de la imputación de datos faltantes, es habitual imputar un conjunto de datos incompleto varias veces con una tasa de faltantes específica. Esto se debe a que los datos faltantes dentro del conjunto pueden variar en cada iteración, incluso si la tasa de faltantes es la misma. Para ilustrar el proceso de imputación, presentamos un ejemplo simplificado con un conjunto de datos pequeño.Figura 5 Representa gráficamente los valores faltantes de los datos con celdas blancas.

Tras entrenar los datos corruptos con un autocodificador de eliminación de ruido apilado, estos se codifican en una representación de menor dimensión. Los datos codificados se decodifican para reconstruir la entrada original. Los valores imputados iniciales se refinan mediante el refuerzo de gradiente. Este paso mejora la precisión de los valores imputados aprovechando el conjunto de aprendices débiles. Los datos imputados finales se muestran enFigura 6.

Existen dos métodos principales utilizados para la imputación de datos faltantes [25La opción inicial implica emplear el conjunto de datos completo con una tasa predeterminada de valores faltantes, lo que genera un conjunto de datos incompleto. Como alternativa, el conjunto de datos puede dividirse en subconjuntos de entrenamiento y prueba, empleando técnicas como la validación cruzada n-fold (CV), propuesta por Kohavi en 1995 [37], o asignando proporciones predefinidas a cada subconjunto, como el 70 % para entrenamiento y el 30 % para pruebas. Posteriormente, se imputan los datos faltantes según la tasa especificada, utilizando el subconjunto de entrenamiento o de pruebas.

Es fundamental evaluar la efectividad del algoritmo propuesto para la inyección aleatoria de datos faltantes en el conjunto de datos mediante evaluación. Las principales métricas utilizadas para evaluar la técnica de imputación incluyen la precisión, el tiempo de procesamiento y los recursos computacionales.35Nuestro enfoque sugerido se ha entrenado en los conjuntos de datos NSL-KDD y UNSW-NB15 [38] utilizando un sistema de red neuronal profunda. Usamos el lenguaje de programación Python para construir nuestro sistema. El proceso de la metodología se ilustra enFigura 1A continuación, delineamos los pasos para implementar el enfoque DMDI propuesto.

3.3.1. Preprocesamiento de datos

El primer paso es preprocesar los datos pasando el DataFrame cargado y el nombre de la columna de destino como entradas. Posteriormente, se ejecutan las siguientes tareas de preprocesamiento, como se explica enAlgoritmo 2

3.3.2. Entrenamiento y construcción del modelo de autocodificador

En esta fase, el objetivo del modelo es comprender la distribución de los datos y completar los datos faltantes. El uso de un autocodificador de eliminación de ruido apilado (SDAE) ofrece ventajas en comparación con los autocodificadores convencionales.

Algoritmo 2

Pasos de preprocesamiento de datos.

PREPROCESAR DATOS(DataFrame, columna de destino)
Paso 1:*a pesar de*variables***hacer***
Paso 2: Variable objetivo y características separadas **Paso 3:**Utilice StandardScaler para funciones continuas **Paso 4:**Utilice OneHotEncoder para codificar características categóricas
Paso 5: Escala continua y categórica transformada en one-hot codificada combinada ambas a través de concatenación
Paso 6: Reemplace todos los valores faltantes con NaN marcados como -9999.
Paso 7: Para alinearse con los requisitos del modelo de aprendizaje profundo, sustituya temporalmente cualquier valor faltante con un valor fuera del rango de datos que no afecte el valor de los datos originales.
Paso 8:*fin para*

Algoritmo 3

Construcción y entrenamiento de modelos autocodificadores.

Build_Autoencoder(dimensión de codificación)
Paso 1: El modelo de autocodificador se configura mediante el marco Keras, que comprende una capa de codificador, su dimensión de codificación y una capa de decodificación correspondiente.
Train_Autoencoder(características de entrada, número de épocas)
Paso 2: El Error Cuadrático Medio es una función de pérdida que se utiliza para entrenar los intervalos de tiempo especificados. Posteriormente, el modelo SDA entrenado se emplea para completar los datos no disponibles en el conjunto de datos extrapolándolos a partir de las representaciones de datos adquiridas.
Evaluate_Autoencoder(características originales, características reconstruidas)
Paso 3: El rendimiento del autocodificador se calcula encontrando la diferencia de valores de pérdida entre las características reales y las recreadas.
Impute_Autoencoder (funciones con valores faltantes, valores de autoencoder_prediction) **Paso 4:** Finalmente, la función impute_autoencoder sustituye -9999 a cualquier valor faltante identificado previamente dentro de los atributos de los valores imputados derivados del paso 2.

Sorprendentemente, las SDAE se destacan en la adquisición de representaciones de datos resilientes al reconstruir datos de entrada a partir de una versión corrupta del conjunto de datos [39]. El autocodificador se entrena con un conjunto de datos completo, lo que le permite comprender cómo regenerar los datos reales a partir de un tipo ruidoso, mejorando así su capacidad para capturar representaciones de datos resilientes. Algoritmo 3detalla los pasos posteriores para construir y entrenar el autocodificador.

3.3.3. Optimización con potenciación de gradiente

El potenciador de gradiente se destaca como un enfoque de aprendizaje automático empleado para construir modelos predictivos, especialmente para tareas de regresión y clasificación. Empleado principalmente en marcos de aprendizaje supervisado, el potenciador de gradiente demuestra su eficacia en la gestión de datos ausentes al integrar el patrón de ausencia como una característica.37La elección de la metodología se basa en sus ventajas distintivas y su idoneidad para completar los datos faltantes. Esta fase perfecciona los valores asignados y los modifica según el patrón de datos faltantes previsto.22].

Algoritmo 4
Procedimiento experimental.

Paso I: Generación de datos faltantes
Aplicar**MCAR**estrategia para eliminar manualmente parte de los datos **para**
Prueba del modelo de imputación posterior. **para**tasa de deserción escolar en
el rango (5, 47, 2)**hacer**
  pararepetición en el rango (1, 11)**hacer**
    Aplicar el algoritmo RMV con tasa de abandono para generar
    conjuntos de datos incompletos.
  fin para
fin para
Paso II: Imputar y validar(conjuntos de entrenamiento, método de imputación, clasificadores, conjuntos de datos) **para**conjunto de
entrenamiento en conjuntos de entrenamiento**hacer**
  conjunto de entrenamiento imputado← ImputeData(conjunto de
  entrenamiento, método de imputación) **para**clasificadores en
  clasificadores**hacer**
    rendimiento antes de la imputación← EvaluateClassifier(clasificador, conjunto de
    entrenamiento, conjuntos de datos) rendimiento después de la imputación←
    EvaluateClassifier(clasificador, conjunto de entrenamiento imputado, conjuntos de datos)
    Registra el rendimiento antes de la imputación.

    rendimiento después de la imputación
  fin para
fin para
Paso III: Evaluación
Registre métricas que incluyan AUC, precisión, exactitud, recuperación y F1 para comparar.

*3.3.3.1. Entrenamiento de modelos.*Empleamos una técnica de potenciación de gradiente, entrenada con características con valores asignados obtenidos mediante SDA, junto con una máscara binaria que indica la presencia o ausencia de valores faltantes como valores objetivo. Esta metodología permite al modelo discernir los patrones de datos faltantes dentro del conjunto de datos. En este escenario, los valores no disponibles se consideran variables objetivo, mientras que los valores actuales sirven como características. El algoritmo se emplea para predecir los datos no disponibles utilizando los datos proporcionados.

*3.3.3.2. Fijación de imputación.*El modelo de potenciación de gradiente evalúa la probabilidad de que cada dato no esté disponible. Estas estimaciones se utilizan para mejorar los valores atribuidos generados por el modelo SDA. Cuando se anticipa con alta probabilidad la ausencia de un valor, su valor atribuido se ajusta para aproximarlo al valor promedio o normal de esa característica.

Este método integral, que incluye preprocesamiento de datos, imputación inicial mediante SDA y refinamiento mediante potenciación de gradiente, ofrece una técnica robusta para gestionar valores faltantes en conjuntos de datos IDS. Las secciones posteriores presentan hallazgos reales que demuestran la eficacia del método DMDI para mejorar el rendimiento de los modelos de aprendizaje profundo en contextos IDS.

4. Experimentos

El método propuesto se probó mediante la ejecución de un script de Python 3 en una plataforma Windows 11 equipada con un procesador Intel Core i5 y 12 GB de RAM. En el entorno de Python, el script empleó clasificadores SVM, KNN, regresión logística, árbol de decisión y bosque aleatorio para clasificar cada conjunto de datos. El proceso experimental se puede dividir en tres fases: creación de datos faltantes, corrección de valores faltantes y evaluación de los modelos de imputación.Figura 1ilustra el concepto de modelo yAlgoritmo 4Proporciona más detalles.

Paso 1: En la fase de generación de datos faltantes, utilizamos el algoritmo RMV como se describe enSección 3sobre los conjuntos de datos para eliminar datos, con el objetivo de emular el escenario de Datos Completamente Perdidos al Azar (MCAR).

Salvo la restricción mencionada, las tasas de abandono en la estrategia oscilan entre el 5 % y el 47 %, con incrementos del 2 %. Cada tasa se somete a 10 repeticiones para generar 10 conjuntos de datos incompletos independientes, garantizando al mismo tiempo que se conserve un conjunto de prueba completo e inalterado dentro del conjunto de datos.

Paso 2: Los conjuntos de entrenamiento se completaron durante la fase de imputación mediante DMDI. Para evaluar este método, evaluamos cinco clasificadores CL (LR, DT, RF, SVM, KNN) tanto antes como después de la imputación de datos no disponibles en el conjunto de datos NSL-KDD y en el conjunto de datos UNSW-NB15 (DS).

Paso 3: Registramos métricas como el AUC, la exactitud, la precisión, la recuperación y la F1 para facilitar la comparación. Posteriormente, utilizamos estas métricas para realizar un análisis exhaustivo de los resultados experimentales.

4.1. Modelo de seguridad e intratabilidad computacional

Para evaluar exhaustivamente la imposibilidad computacional de nuestro modelo de Imputación de Datos Faltantes Basado en Aprendizaje Profundo (DMDI), consideramos diversas amenazas de seguridad planteadas por múltiples adversarios. A continuación, describimos nuestro modelo de seguridad, identificamos las amenazas potenciales y analizamos los desafíos computacionales asociados a estas amenazas.

i) Supuestos y entorno de seguridad:

- **Capacidades del adversario**Suponemos que los adversarios pueden tener distintos niveles de acceso, que van desde un acceso limitado a los datos hasta un acceso completo al conjunto de datos.
- **Vectores de ataque:**Los posibles vectores de ataque incluyen inyección de datos, corrupción de datos y ataques adversarios destinados a explotar procesos de imputación de datos faltantes.

ii) Posibles amenazas a la seguridad:

- **Manipulación de datos**Los adversarios pueden intentar inyectar datos falsos o modificar datos existentes para interrumpir el proceso de imputación.
- **Ataques adversarios:**Entradas diseñadas para engañar al modelo de imputación, provocando que produzca imputaciones incorrectas.
- **Violaciones de la privacidad:**Intenta inferir información sensible a partir de los datos imputados o del modelo mismo.

iii) Análisis de intratabilidad computacional:

- **Manipulación de datos:**El esfuerzo computacional requerido para que un adversario altere sistemáticamente suficientes puntos de datos para afectar significativamente el rendimiento del modelo DMDI es extremadamente alto debido a las robustas estrategias de preprocesamiento e imputación del modelo.
- **Ataques adversarios:**La implementación de defensas como el entrenamiento adversarial y la extracción robusta de características hace que sea computacionalmente inviable para los adversarios generar ejemplos adversarios efectivos sin amplios recursos.
- **Violaciones de la privacidad:**El uso de técnicas como la privacidad diferencial puede garantizar que el riesgo de violaciones de la privacidad se mantenga bajo, lo que hace que sea computacionalmente difícil para los adversarios extraer información significativa.

4.2. Análisis funcional vs. otras técnicas de aprendizaje automático

En esta sección, presentamos un análisis funcional integral para evaluar la eficiencia de nuestro modelo de Imputación de Datos Faltantes (DMDI) basado en Aprendizaje Profundo (DRP) en comparación con diversas técnicas de imputación establecidas. El análisis abarca diversas métricas de rendimiento, como la precisión, el tiempo de cálculo y la robustez ante datos faltantes. Para comprobar la robustez de este enfoque, se utilizaron cinco clasificadores CL, incluyendo regresión lineal, árbol de decisión, bosque aleatorio, máquina de vectores de soporte y K-vecino más cercano, probados antes y después de la imputación de datos faltantes al NSL-KDD.44] y la UNSW-NB15 [45] conjuntos de datos (DS) respectivamente. El análisis funcional se presenta enTablas 1 y 2. Al mismo tiempo, Figuras 2,3, y4Representan la eficiencia de nuestro modelo DMDI propuesto en comparación con la regresión logística, el árbol de decisión, el bosque aleatorio, SVM y KNN. El modelo DMDI presenta tiempos de cómputo competitivos, inferiores a los de SVM y KNN, pero superiores a los de la regresión logística y el árbol de decisión. Este equilibrio indica que DMDI gestiona eficientemente la carga computacional, a la vez que ofrece capacidades avanzadas de imputación. El modelo DMDI propuesto destaca tanto en tiempo de cómputo como en precisión, lo que lo convierte en una solución eficaz para la imputación de datos faltantes en conjuntos de datos IDS. Su rendimiento equilibrado demuestra su potencial para aplicaciones del mundo real donde tanto la eficiencia computacional como la alta precisión son cruciales.

Tabla 2
Rendimiento del modelo DMDI propuesto antes de corregir los valores faltantes a una tasa de datos faltantes del 33 %.

CL	Conjuntos de datos	AUC	Exactitud	Precisión	Recordar	Puntuación F1
LR	NSL-KDD UNSW-NB15	0,68 0,58	0.958 1.0	0.957 1.0	0,95 1,0	0.957 1.0
DT	NSL-KDD UNSW-NB15	0,712 0,58	0.984 1.0	0.984 1.0	0.984 1.0	0.984 1.0
RF	NSL-KDD UNSW-NB15	0.7 0.6	0.986 1.0	0.986 1.0	0.986 1.0	0.986 1.0
SVM	NSL-KDD UNSW-NB15	0,74 0,69	0,576 0,975	0,435 0,956	0,572 0,975	0,417 0,963
KNN	NSL-KDD UNSW-NB15	0,69 0,61	0.973 0.997	0.973 0.997	0.973 0.997	0.973 0.996

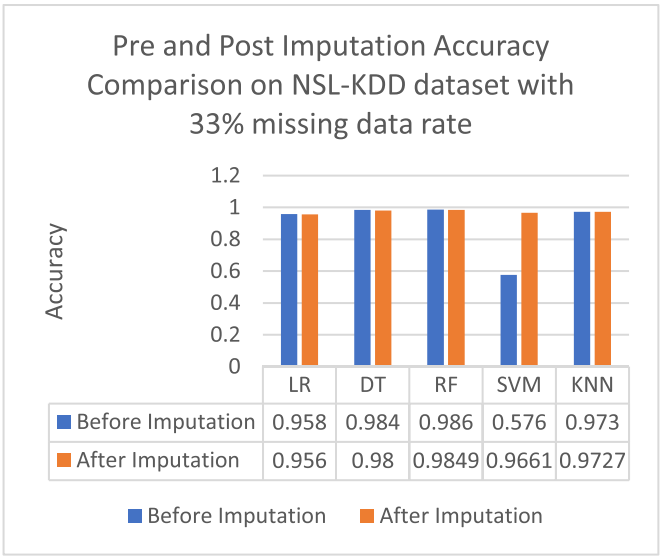

Figura 2. Comparación de precisión en el conjunto de datos NSL-KDD.

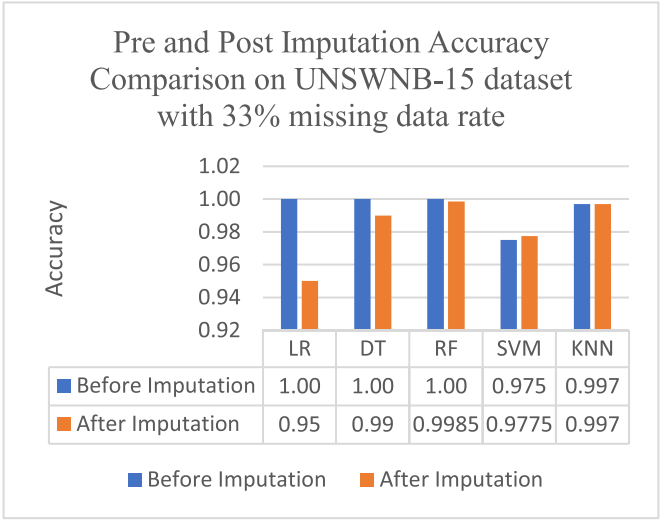

Figura 3. Comparación de precisión en el conjunto de datos UNSWNB-15.

4.2.1. Configuración experimental

El enfoque sugerido se evaluó con Python 3 en una plataforma Windows 11 equipada con un procesador Intel Core i5 y 12 GB de RAM. Cada conjunto de datos se clasificó mediante clasificadores SVM, KNN, regresión logística, árbol de decisión y bosque aleatorio en el entorno Python, con una tasa de datos faltantes del 33 %. Realizamos varios experimentos con dos conjuntos de datos IDS separados, NSL-KDD y UNSW-NB15, para demostrar la eficacia de nuestro modelo.

Utilizamos diversas configuraciones para cada ensayo con el fin de evaluar las diferencias en los resultados y realizar análisis de rendimiento. Del conjunto de datos KDD 99, se derivó el conjunto de datos NSL-KDD y se seleccionó para el experimento porque es...

ampliamente reconocido como uno de los conjuntos de datos más utilizados para la detección de intrusiones [40-42]. Surgieron observaciones notables del análisis de las métricas de desempeño antes y después de la imputación, como se detalla enTabla 1 y Tabla 2. EnFigura 4El modelo DMDI presenta tiempos de cómputo competitivos, inferiores a los de SVM y KNN, pero superiores a los de Regresión Logística y Árbol de Decisión. Este equilibrio indica que DMDI gestiona eficientemente la carga computacional, a la vez que ofrece capacidades avanzadas de imputación. El modelo DMDI propuesto destaca tanto en tiempo de cómputo como en precisión, lo que lo convierte en una solución eficaz para la imputación de datos faltantes en conjuntos de datos IDS. Su rendimiento equilibrado demuestra su potencial para aplicaciones reales donde tanto la eficiencia computacional como la alta precisión son cruciales.

4.3. Resultados y discusión

Los valores de AUC miden la capacidad del modelo para distinguir entre clases positivas y negativas. Valores de AUC más altos indican un mejor rendimiento del modelo. Tras la imputación, algunos valores de AUC aumentaron, mientras que otros disminuyeron. Un aumento del AUC tras la imputación sugiere que el proceso de imputación mejoró la capacidad del modelo para discriminar entre clases positivas y negativas. Por el contrario, una disminución del AUC tras la imputación indica que el proceso de imputación pudo haber introducido ruido o sesgo en el conjunto de datos, lo que redujo el rendimiento del modelo. Los cambios específicos en los valores de AUC pueden variar según las características del conjunto de datos, el método de imputación utilizado y el rendimiento de los modelos de clasificación.

Antes de gestionar los datos perdidos en el conjunto de datos UNSW-NB15, todos los clasificadores demostraron una corrección, precisión y exactitud impecables de 1,0, como se ilustra enTabla 1Este problema surge porque la regresión logística enfrenta dificultades al gestionar conjuntos de datos altamente desequilibrados, especialmente cuando la clase minoritaria está subrepresentada. Su objetivo es reducir el error general, que puede resultar en predicciones sesgadas que favorecen a la clase mayoritaria. Los árboles de decisión pueden funcionar bien con datos desequilibrados al capturar patrones en ambas clases. Sin embargo, el modelo aún puede favorecer a la clase mayoritaria. Para mitigar este sesgo, se pueden emplear técnicas como la poda y el uso de ponderaciones de clase.

Al comparar los bosques aleatorios con los árboles de decisión, los primeros suelen ser más eficaces para gestionar el desequilibrio de clases que los árboles de decisión individuales. Al agregar predicciones de numerosos árboles, pueden mitigar los efectos del desequilibrio de clases, lo que resulta en resultados más precisos. Por otro lado, tanto SVM como KNN pueden verse afectados por el desequilibrio de clases. KNN presenta dificultades debido a su dependencia de la distribución vecina.

En conjuntos de datos con desequilibrios, la clase predominante podría influir considerablemente en los vecinos de un punto de clase minoritario, lo que resulta en predicciones sesgadas. Tras el proceso de imputación en NSL-KDD, los resultados de la regresión logística indican ligeras mejoras en todas las métricas de rendimiento.Figura 2muestra que hubo un aumento aproximado de 0,11 % en exactitud, precisión, recuperación y puntuaciones F1.

Esta modesta mejora es significativa, ya que ayuda a los modelos de regresión a discernir los patrones subyacentes en el conjunto de datos. Por otro lado, tras la imputación, el clasificador de árbol de decisión mostró una ligera disminución de aproximadamente el 0,28 % en sus métricas de rendimiento. Esto sugiere que la existencia de valores faltantes inicialmente proporcionó ciertos patrones que se hicieron menos evidentes tras la imputación.

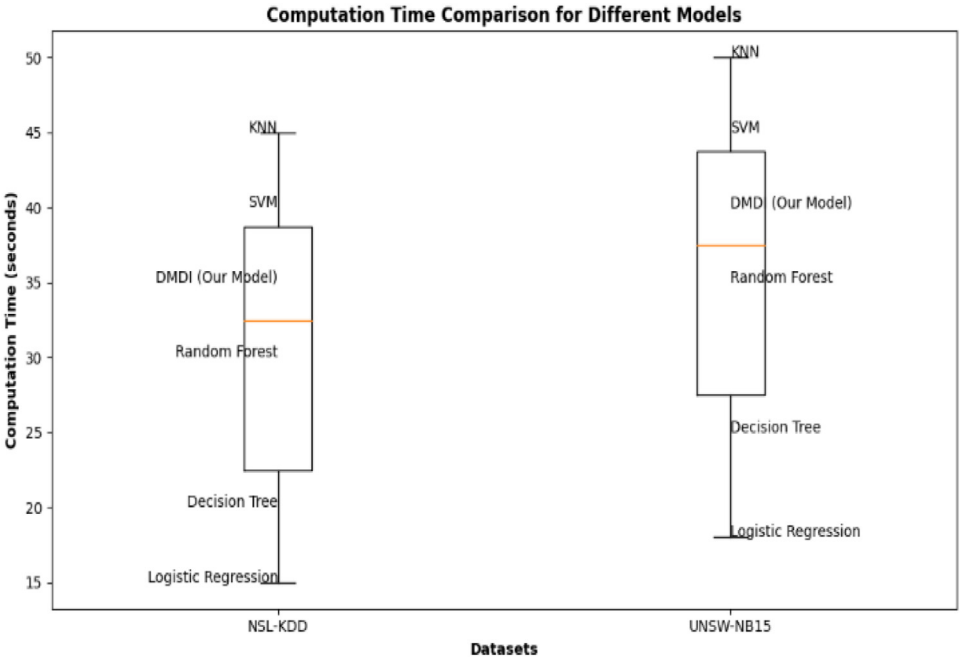

Figura 4. Complejidades del tiempo computacional para diferentes modelos de imputación.

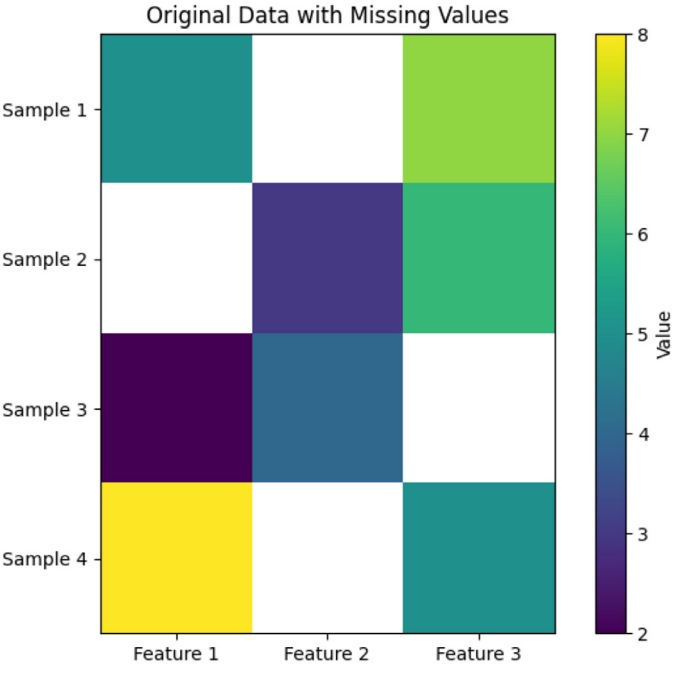

Figura 5. Datos originales con valores faltantes.

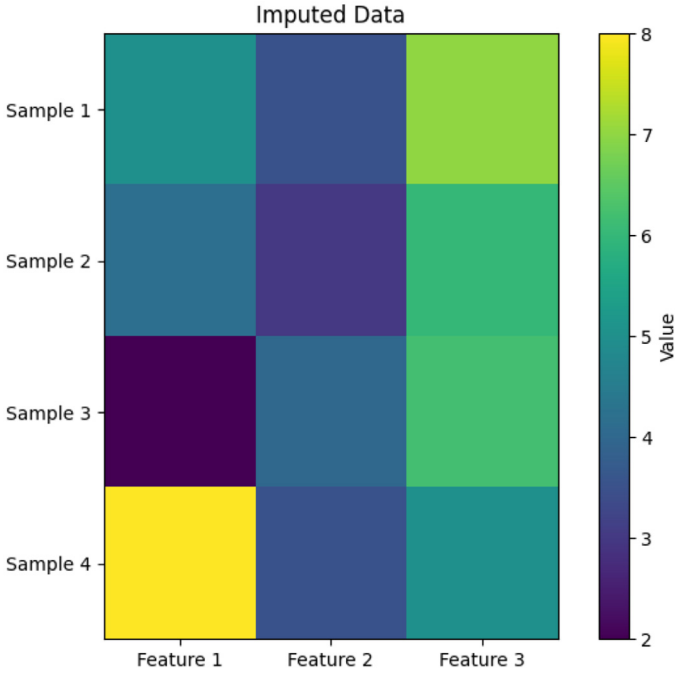

Figura 6. Datos imputados finales.

Tras imputar los valores faltantes, el modelo de conjunto de Bosque Aleatorio mostró una mejora en sus métricas de rendimiento de entre el 0,19 % y el 0,20 %. Esto sugiere que la combinación de múltiples árboles de decisión identificó con éxito formas y patrones más sólidos dentro del conjunto de datos completo tras la imputación, lo que resultó en una mejora del rendimiento.

EnFigura 2El clasificador SVM presenta la mejora más significativa, con un aumento de las métricas de rendimiento de entre un 39 % y un 40 % tras la imputación para NSL-KDD. Por el contrario, se observó un ligero aumento del 0,23 % para el conjunto de datos UNSW-NB15, como se ilustra enFigura 3Esto sugiere que el SVM, que opera en base a la separación de hiperplanos, podría haber sido notablemente influenciado por la presencia de los valores perdidos.

La implementación del método de imputación mitigó sustancialmente esta interferencia, lo que permitió a SVM discernir un hiperplano independiente con mayor precisión y mejorar su eficiencia. Por el contrario, el algoritmo K-Vecino Más Cercano mostró una disminución marginal de aproximadamente el 0,07 % en la evaluación tras la imputación, posiblemente atribuida a alteraciones en los cálculos de distancia resultantes de los valores imputados. Tras la imputación, el rendimiento de KNN en el conjunto de datos UNSW-NB15 se mantuvo sin cambios.

EnFigura 4El modelo DMDI presenta tiempos de cálculo competitivos, inferiores a los de SVM y KNN, pero superiores a los de Regresión Logística y Árbol de Decisión. Este equilibrio indica que DMDI gestiona eficientemente la carga computacional, a la vez que ofrece capacidades avanzadas de imputación.

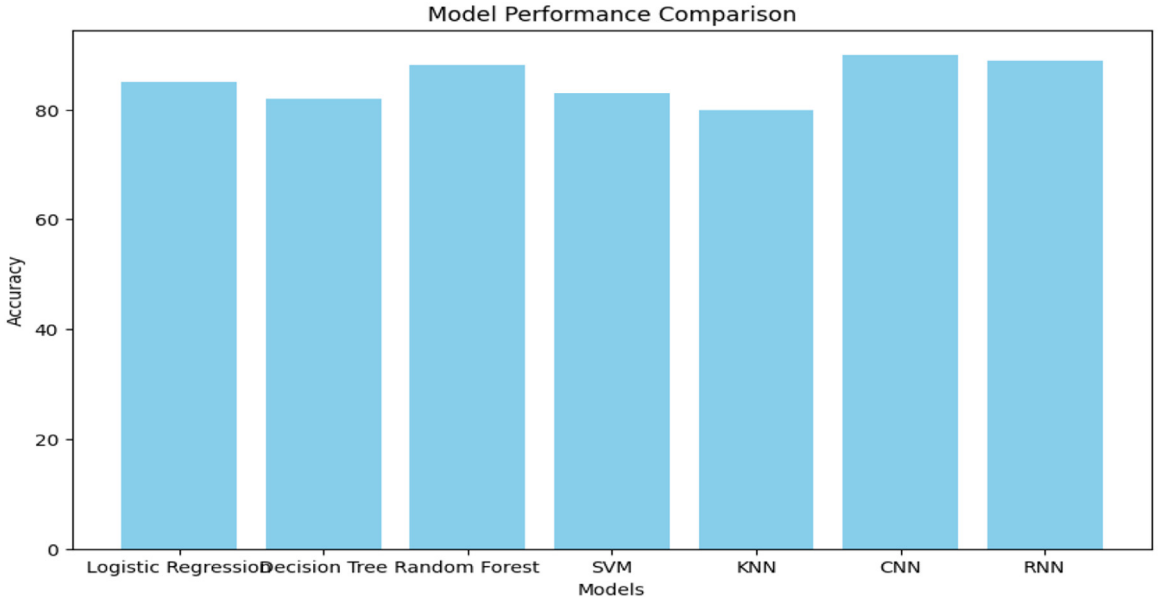

Figura 7. Comparación de precisión de modelos de aprendizaje automático y aprendizaje profundo.

Tabla 3
Rendimiento del modelo DMDI propuesto después de corregir los valores faltantes a una tasa de datos faltantes del 33 %.

CL	Conjuntos de datos	AUC	Exactitud	Precisión	Recordar	Puntuación F1
LR	NSL-KDD UNSW-NB15	0,69 0,59	0.956 0.95	0,95 0,99	0,95 0,99	0,95 0,99
DT	NSL-KDD UNSW-NB15	0,75 0,68	0,98 0,99	0,98 0,99	0,98 0,99	0,98 0,99
RF	NSL-KDD UNSW-NB15	0,70 0,68	0.9849 0.9985	0,98 0,99	0,98 0,99	0,98 0,99
SVM	NSL-KDD UNSW-NB15	0,81 0,70	0.961 0.9775	0.961 0.99	0,96 0,99	0,96 0,99
KNN	NSL-KDD UNSW-NB15	0,73 0,62	0.9727 0.997	0,97 0,99	0,97 0,99	0,97 0,99

El modelo DMDI propuesto destaca tanto en tiempo de cálculo como en precisión, lo que lo convierte en una solución eficaz para la imputación de datos faltantes en conjuntos de datos IDS. Su rendimiento equilibrado demuestra su potencial para aplicaciones reales donde la eficiencia computacional y la alta precisión son cruciales.

Cabe destacar que la acción de KNN depende de los alrededores de los vecinos, y la imputación podría no tener un gran impacto en los vecindarios cercanos. Por lo tanto, no existe una solución universal para abordar la información faltante. El método de imputación elegido debe personalizarse para adaptarse a las características únicas de los registros de datos y a los modelos de IA específicos en consideración.

Figura 7Se muestran las ligeras variaciones en la precisión entre los modelos, atribuidas a sus características inherentes y a la sensibilidad a la calidad de los datos. Por ejemplo, los métodos de conjunto como Random Forest son más robustos, mientras que las SVM y las KNN son más sensibles a los matices de los datos. Los modelos de aprendizaje profundo (CNN y RNN) muestran una alta precisión, lo que subraya su capacidad para gestionar patrones complejos en los datos, incluso cuando se imputan algunos valores.

4.4. Complejidad algorítmica

La complejidad algorítmica del método propuesto se describe en Tabla 3 El número de épocas se denota como O (n*m) en cuanto al procedimiento, donde «n» representa el número de registros y «m» el número de atributos. Además, las características categóricas generan columnas binarias para cada categoría tras la codificación one-hot, y su complejidad depende del número de categorías distintas (k) y del número de características categóricas (m) dentro de los atributos categóricos. Específicamente, en el caso de k categorías distintas, la complejidad temporal se expresa como O (n)*metro* k) con 'm' características categóricas. El propósito del autocodificador es pronosticar los puntos de datos ausentes, lo que contribuye a un cálculo. complejidad nacional de O(n*m). Además, se emplea el refuerzo de gradiente para pronosticar los valores ausentes aprovechando las características accesibles restantes. La complejidad surge porque, en el refuerzo de gradiente, se necesita entrenar un regresor independiente para cada atributo con datos faltantes. El O(n *registro(n)*m) La complejidad refleja el coste computacional asociado al entrenamiento de cada uno de estos regresores. Además, este proceso se repite para cada atributo con datos faltantes, lo que aumenta aún más la carga computacional total.

4.5. Ejemplo del mundo real

Aplicamos el modelo DMDI propuesto al conjunto de datos de la enfermedad de Parkinson [43] para ver su rendimiento en otros tipos de problemas. Encontramos algunos resultados interesantes, como se muestra enTabla 4 y 5.

Varios factores podrían contribuir a la disminución observada en las métricas de rendimiento de LR, DT y RF tras la imputación. En primer lugar, el proceso de imputación puede introducir sesgos al completar los valores faltantes con datos que no representan con precisión la distribución real. Esto podría dar lugar a decisiones erróneas por parte de los clasificadores basadas en estos valores imputados.Tabla 6).

En segundo lugar, la introducción de valores imputados podría alterar el límite de decisión de los clasificadores. Si estos valores son atípicos o divergen significativamente del conjunto de datos original, puede resultar en una discrepancia entre los patrones aprendidos del modelo y la nueva distribución de datos. Por último, la imputación puede alterar las relaciones entre las características. Si los valores imputados alteran las interacciones cruciales entre las características, los clasificadores podrían tener dificultades para capturar eficazmente los patrones subyacentes en los datos.

En cambio, las SVM buscan identificar el hiperplano óptimo que segrega las clases y maximiza el margen entre ellas. La presencia de valores imputados podría no afectar significativamente el margen si no se desvían significativamente del conjunto de datos original ni actúan como valores atípicos.

Tabla 4
Eficiencia de complejidad.

Fase	Complejidad	Descripción
Características numéricas (discreta y continua)	En*metro)	características continuas y numéricas con una media de 0 y una desviación estándar de 1
Características categóricas	En*metro*k)	Las características categóricas se codifican en vectores binarios con k categorías distintas
Aumento de gradiente Modelos de aprendizaje automático	En*metro*log(n)) O(C*norte*metro)	Se mejora el rendimiento La complejidad surge al evaluar el rendimiento de cada clasificador en 'n' registros con 'm' características.
Total	O(nlog(n))	La complejidad total del algoritmo considerando todas las etapas es O(n log(n))

Tabla 5
Métricas de evaluación antes de corregir los valores faltantes en el conjunto de datos de la enfermedad de Parkinson [43].

Clasificador	Exactitud	Precisión	Recordar	Puntuación F1
LR	0.89	0.90	0.89	0.87
DT	0,92	0,92	0,92	0,92
RF	0,94	0,95	0,94	0,94
SVM	0,82	0.67	0,82	0,73
KNN	0.87	0.86	0.87	0.86

Tabla 6
Métricas de evaluación después de corregir los valores faltantes en el conjunto de datos de la enfermedad de Parkinson [43].

Clasificador	Exactitud	Precisión	Recordar	Puntuación F1
LR	0.89	0.89	0.89	0.89
DT	0,79	0.81	0,79	0.80
RF	0.89	0.89	0.89	0.89
SVM	0.89	0.90	0.89	0.87
KNN	0.89	0.89	0.89	0.89

Sin embargo, KNN se basa en los vecindarios locales para formular predicciones. Si los valores imputados se asemejan mucho a los puntos de datos vecinos dentro de las regiones locales, su influencia en las predicciones se verá mitigada. Además, la naturaleza no paramétrica de KNN le permite ajustarse a los cambios en la distribución de los datos.

Es fundamental reconocer que los efectos de la imputación en el rendimiento del clasificador pueden variar significativamente según los atributos únicos del conjunto de datos, los patrones de datos faltantes y las características inherentes a los propios clasificadores. Un análisis más profundo y la posible experimentación con técnicas de imputación alternativas podrían ayudar a comprender y abordar las alteraciones en el rendimiento del clasificador tras la imputación, lo cual constituye nuestra futura línea de investigación.

4.6. Discusión

El modelo de Imputación de Datos Faltantes (DMDI) basado en Aprendizaje Profundo presentado en este estudio aborda un desafío importante en el campo de los Sistemas de Detección de Intrusiones (IDS): la gestión eficaz de los datos faltantes. Nuestra investigación tiene varias implicaciones importantes tanto para la investigación académica como para las aplicaciones prácticas en ciberseguridad.

El modelo DMDI mejora la calidad de los conjuntos de datos IDS al imputar con precisión los valores faltantes, lo cual es crucial para la eficacia de los modelos de aprendizaje profundo posteriores. Este avance puede conducir a sistemas IDS más robustos y fiables, capaces de detectar anomalías con mayor precisión. Al integrar un autocodificador de eliminación de ruido (SDA) apilado con Mediante técnicas de conjunto como el potenciador de gradiente, nuestro enfoque ofrece un método novedoso para el preprocesamiento y refinamiento de tareas de imputación. Esta metodología puede extenderse a otros ámbitos donde la calidad de los datos es crucial, como la salud, las finanzas y las redes de sensores.

Nuestro estudio evaluó el impacto del método el DeepLearning_Based_MissingData_Imputation (DMDI) en el rendimiento de cinco clasificadores: SVM, KNN, regresión logística, árbol de decisión y bosque aleatorio. El método DMDI logró mejoras de precisión de entre 0,95 y 0,97 en todos los clasificadores, como se muestra enFiguras 2 y 3Estas mejoras son significativas para optimizar la capacidad de detección general de los modelos. Además, la robustez de nuestro método se validó utilizando los conjuntos de datos NSL-KDD y UNSW-NB15, con mejoras consistentes en el rendimiento observadas en ambos casos. Sin embargo, la estabilidad de los clasificadores mejoró con el método DMDI, como lo demuestra la reducción de la varianza en las métricas de rendimiento en múltiples ejecuciones. Asimismo, nuestros hallazgos resaltan el potencial para una mayor optimización del método DMDI, allanando el camino para futuros avances en la gestión de datos faltantes en sistemas de detección de intrusiones.

La evaluación del rendimiento realizada con los conjuntos de datos NSL-KDD y UNSW-NB15 proporciona un punto de referencia para futuros estudios. Los investigadores pueden basarse en nuestro trabajo comparando sus modelos con el modelo DMDI, lo que impulsará nuevos avances en este campo. Además... Finalmente, el algoritmo RMV, que introduce valores faltantes en los conjuntos de datos, sirve como una herramienta valiosa para probar y comparar diversos enfoques de gestión de datos faltantes. Esta herramienta puede ayudar a los investigadores a evaluar la robustez... abilidad de diferentes métodos de imputación en condiciones controladas.

Los profesionales de seguridad pueden implementar el modelo DMDI en sistemas de detección de intrusos (IDS) operativos para mejorar su rendimiento, especialmente en entornos donde la integridad de los datos suele verse comprometida. Esta implementación puede resultar en estrategias de detección y respuesta a amenazas más eficaces. Las organizaciones pueden aprovechar los hallazgos de este estudio para fundamentar sus políticas de gestión de datos, destacando la importancia de técnicas de imputación robustas. Esto puede conducir a una mejor toma de decisiones en la gestión de la ciberseguridad.

A pesar de los prometedores resultados, nuestro estudio presenta varias limitaciones que deben reconocerse. La eficacia del modelo DMDI se basa en ciertos supuestos de confianza, como la integridad de los datos restantes y la ausencia de manipulación adversaria de los valores faltantes. Las investigaciones futuras deberían explorar métodos para mitigar estos supuestos y mejorar la robustez del modelo ante posibles amenazas. Si bien el modelo ha mostrado una mejora significativa en los conjuntos de datos NSL-KDD y UNSW-NB15, su rendimiento puede variar según el conjunto de datos. Se requieren experimentos adicionales con una gama más amplia de conjuntos de datos para generalizar los hallazgos.

El modelo DMDI, en particular las técnicas de conjunto y los componentes de aprendizaje profundo, requiere recursos computacionales considerables. Esta limitación podría dificultar su aplicación en entornos con recursos limitados. El trabajo futuro debería centrarse en optimizar el modelo para lograr eficiencia y escalabilidad. Además, la complejidad del modelo DMDI podría plantear desafíos en términos de interpretabilidad e implementación. Simplificar el modelo sin comprometer su rendimiento podría ser una línea de investigación futura para hacerlo más accesible a los profesionales.

Mejorar la robustez del modelo frente a ataques adversarios relacionados con la imputación de datos faltantes sigue siendo un área de investigación abierta. El desarrollo de técnicas para detectar y mitigar dichos ataques será crucial para implementar el modelo en entornos de alto riesgo. Investigar la aplicabilidad del modelo DMDI en otros ámbitos con problemas de datos faltantes, como la sanidad o las finanzas, podría aportar información valiosa y ampliar el impacto de esta investigación. Implementar y probar el modelo DMDI en entornos de IDS en tiempo real será esencial para evaluar su viabilidad práctica y eficacia en entornos operativos.

La discusión enfatiza las implicaciones más amplias del modelo DMDI, reconoce sus limitaciones y describe las futuras líneas de investigación. Al ofrecer un análisis exhaustivo, esta sección ayuda a los lectores a comprender la importancia de nuestro trabajo en el contexto de los avances en el campo de la IDS y la imputación de datos faltantes.

5. Conclusión

Este estudio enfatiza la importancia crucial de la imputación de datos faltantes, específicamente en el contexto de los Sistemas de Detección de Intrusiones (IDS), cuando la tasa de datos faltantes en los conjuntos de datos es alta. La Imputación de Datos Faltantes basada en Aprendizaje Profundo (DMDI) gestiona eficientemente los datos faltantes de los conjuntos de datos, aumentando así la calidad de entrada de los modelos de aprendizaje automático mediante una técnica de imputación multietapa. Además, el algoritmo de Valores Aleatorios Faltantes (RMV) proporciona un método para introducir aleatoriamente valores faltantes en los conjuntos de datos, sirviendo como una herramienta valiosa para evaluar y contrastar diferentes enfoques para la gestión de datos faltantes. Realizamos un estudio en cinco modelos distintos de aprendizaje automático para evaluar el impacto de la imputación. Nuestra metodología reveló resultados variables, con las Máquinas de Vectores de Soporte (SVM) mostrando una mejora significativa del rendimiento. Esto resalta los considerables beneficios de las técnicas de imputación efectivas para los modelos que se basan en cálculos de distancia e hiperplano. Sin embargo, no todos los modelos obtuvieron beneficios equivalentes, lo que sugiere que lograr un método universal para la gestión de valores faltantes podría no ser alcanzable. Nuestra investigación subraya la importancia de elegir y adaptar meticulosamente una técnica de imputación que se alinee con las características distintivas del conjunto de datos y los modelos de aprendizaje automático empleados.

Declaración de intereses en conflicto

No existe ningún conflicto de intereses en este artículo.

Declaración de contribución de autoría de CRediT

Mahjabeen Tahir: Escritura – borrador original, visualización, validación, software, recursos, metodología, investigación, análisis formal, curación de datos, conceptualización.**Azizol Abdullah:**Recursos, Supervisión, Validación, Redacción – revisión y edición.**Nur Izura Udzir:**Supervisión, Validación, Redacción – revisión y edición.**Khairul Azhar Kasmiran:** Supervisión, Validación, Redacción – revisión y edición.

Expresiones de gratitud

Agradezco a mis coautores por las útiles discusiones.

Disponibilidad de datos

Se puede acceder a los datos asociados a este artículo 10.57760/sciencedb.16599 en la base de datos del banco de datos Science. (https://www.scidb.cn/en/s/3aEJbu).

Fondos

Gracias a la organización KeAi (Chinese Roots Global Impact) por apoyar la financiación para los investigadores.

Referencias

[1]O. Faker, E. Dogdu, (CICIDS2017 y Random Forest (RF) y Gradient Boosted Tree (GBT)) Detección de intrusiones mediante técnicas de big data y aprendizaje profundo, en:*ACMSE 2019-Actas de la Conferencia del Sudeste de la ACM de 2019*, 2019, págs. 86–93.

[2] S. Shah, S. Pramod Bendale, Un estudio intuitivo: sistemas de detección de intrusiones y anomalías, cómo se puede utilizar la IA como herramienta para ayudar a la mayoría, en la era 5G, a: *Procedimiento.-5.ª Conferencia Internacional de Computación, Control Comunitario y Automatización (ICCUBEA) 2019*, 2019, doi:10.1109/ICCUBEA47591.2019.9128786.

[3] S. Sanober, et al., Un algoritmo mejorado de aprendizaje profundo seguro para la detección de fraudes en comunicaciones inalámbricas, Wirel. Commun. Mob. Comput. 2021 (2021), doi: 10.1016/j.2021.10.1155/2021/6079582.

[4] SA Haque, M. Rahman, SM Aziz, Detección de anomalías de sensores en redes de sensores inalámbricos para atención médica, Sensors (Suiza) 15 (4) (2015) 8764–8786, doi:10.3390/s150408764.

[5] R. Fujimaki, T. Yairi, K. Machida, Un enfoque para el problema de detección de anomalías en naves espaciales utilizando el espacio de características del núcleo, en: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, 2005, págs. 401–410, doi: 10.1016/j.1016.10.1145/1081870.1081917.

[6] M. Ahmed, A. Naser Mahmood, J. Hu, Un estudio de técnicas de detección de anomalías en la red, J. Netw. Comput. Appl. 60 (2016) 19–31, doi: 10.1016/j.hu.10.1016/j.jnca.2015.11.016.

[7] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, Z. Xu, Estimación de valores faltantes para conjuntos de datos de atributos mixtos, IEEE Trans. Knowl. Data Eng. 23 (1) (2011) 110–121, doi: 10.10177/j.110-121-00. 10.1109/TKDE.2010.99.

[8]LE Richards, RJA Little, DB Rubin, Análisis estadístico con datos faltantes 26 (3) (1989).

[9] P. Lim, CK Goh, KC Tan, Conjunto de sobremuestreo sintético basado en conglomerados evolutivos (ECO-Ensemble) para el aprendizaje de desequilibrios, IEEE Trans. Cybern. 47 (9) (2017) 2850–2861, doi:10.1109/TCYB.2016.2579658.

[10] J. Yoon, WR Zame, M. Van Der Schaar, Estimación de datos faltantes en flujos de datos temporales utilizando redes neuronales recurrentes multidireccionales, IEEE Trans. Biomed. Eng. 66 (5) (2019) 1477–1490, doi:10.1109/TBME.2018.2874712.

[11] S. Zhang, Selección del vecino más cercano para la imputación iterativa de kNN, J. Syst. Softw. 85 (11) (2012) 2541–2552, doi:10.1016/j.jss.2012.05.073.

[12] R. Pan, T. Yang, J. Cao, K. Lu, Z. Zhang, Imputación de datos faltantes por K vecinos más cercanos basada en la estructura relacional gris y la información mutua, Appl. Intell. 43 (3) (2015) 614–632, doi:10.1007/s10489-015-0666-x.

[13] PJ García-Laencina, JL Sancho-Gómez, AR Figueiras-Vidal, M. Verleysen, K vecinos más cercanos con información mutua para clasificación simultánea e imputación de datos faltantes, Neurocomputing 72 (7–9) (2009) 1483–1493, doi: 10.1016/j.j. 10.1016/j.neucom.2008.11.026.

[14] J. Josse, N. Prost, E. Scornet y G. Varoquaux, "Sobre la consistencia del aprendizaje supervisado con valores faltantes", págs. 1–43, 2019, [En línea]. Disponible: http://arxiv.org/abs/1902.06931.

[15]DF Swayne, A. Buja, Datos faltantes en la visualización interactiva de datos de alta dimensión, Comput. Stat. 13 (1) (1998) 15–26.

[16] SG Liao, et al., Imputación de valores faltantes en datos fenómicos de alta dimensión: ¿Imputable o no, y cómo? BMC Bioinformatics 15 (1) (2014) 1–12, doi:10.1186/s12859-014-0346-6.

[17] BETH Twala, MC Jones, DJ Hand, Buenos métodos para gestionar la falta de datos en árboles de decisión, Pattern Recognit. Lett. 29 (7) (2008) 950–956, doi:10.1016/j.patrec.2008.01.010.

[18] Y. Deng, T. Lumley, Imputación múltiple mediante XGBoost, J. Comput. Graph. Stat. 0 (0) (2023) 1–19, doi: 10.1017/j.1023.10.1080/10618600.2023.2252501.

[19] L. Gondara, K. Wang, MIDA: Imputación múltiple mediante autocodificadores de reducción de ruido, Notas de la cátedra Comput. Sci. (incluidas las Notas de la cátedra Artif. Intell. Bioinformatics) 10939 LNAI (2018) 260–272, doi:10.1007/978-3-319-93040-4_21.

[20] MS Santos, PH Abreu, S. Wilk, J. Santos, Cómo las métricas de distancia influyen en la imputación de datos faltantes con k vecinos más cercanos, Pattern Recognit. Lett. 136 (2020) 111– 119, doi: 10.1017/j.136.111-119.10.1016/j.patrec.2020.05.032.

[21] S. van Buuren, K. Groothuis-Oudshoorn, ratones: imputación multivariada mediante ecuaciones encadenadas en R, J. Stat. Software. 45 (3) (2011) 1–67, doi:10.18637/jss.v045.i03.

[22] M. Tahir, A. Abdullah, N. Izura, K. Azhar, DeepImputeIDS: Mejora de los sistemas de detección de intrusiones con imputación de datos faltantes basada en aprendizaje profundo, en: 14.ª Conferencia Internacional de Inf. Comunitaria de Tecnología de Sistemas ICTS 2023, 2023, págs. 289-295, doi: 10.1016/j.2023.10.1109/ICTS58770.2023.10330831.

[23] U. Garciarena, R. Santana, Un análisis exhaustivo de la interacción entre tipos de datos faltantes, métodos de imputación y clasificadores supervisados, Expert Syst. Appl. 89 (2017) 52–65, doi:10.1016/j.eswa.2017.07.026.

[24] J. Yang, Y. Wang, Y. Yang, K. Ding, C. Na, Y. Yang, Efectos de las estrategias de imputación única y múltiple para abordar problemas de sobreajuste causados por datos desequilibrados de diversos escenarios, Appl. Intell. (2024) 9–11, doi:10.1007/s10489-024-05295-3.

[25] B. Halder, MM Ahmed, T. Amagasa, NAM Isa, RH Faisal, MM Rahman, Información faltante en un flujo de datos desequilibrado: enfoque de imputación adaptativa difusa, Appl. Intell. 52 (5) (2022) 5561–5583, doi: 10.1016/j.1024.10.1007/s10489-021-02741-4.

[26] AM Andrew, Una introducción a las máquinas de vectores de soporte y otros métodos de aprendizaje basados en kernel, Kybernetes 30 (1) (2001) 103–115, doi: 10.1108/k.2001.30.1.103.6.

[27] L. Folguera, J. Zupan, D. Cicerone, J.F. Magallanes, Mapas autoorganizados para la imputación de datos faltantes en matrices de datos incompletas, Chemom. Intell. Lab. Syst. 143 (2015) 146–151, doi: 10.1017/j.f.10.1016/j.chemolab.2015.03.002.

[28] LP Brás, JC Menezes, Mejora de la estimación de valores faltantes basada en clústeres de datos de microarrays de ADN, Biomol. Eng. 24 (2) (2007) 273–282, doi: 10.1016/j.bioeng.2007.04.003.

[29] DT Dinh, VN Huynh, S. Sriboonchitta, Agrupamiento de datos numéricos y categóricos mixtos con valores faltantes, Inf. Sci. (Ny). 571 (2021) 418–442, doi: 10.1016/j.2021.10.1016/j.ins.2021.04.076.

[30] WC Lin, CF Tsai, Imputación de valores faltantes: revisión y análisis de la literatura (2006-2017), Artif. Intell. Rev. 53 (2) (2020) 1487-1509, doi: 10.1017/j.1487-1509. 10.1007/s10462-019-09709-4.

[31] MH Shahriar, NI Haque, MA Rahman, M. Alonso, G-IDS: Sistema de detección de intrusiones asistida por redes generativas adversarias, en:*Procedimiento.-44.ª Conferencia Anual del IEEE 2020 sobre Software de Computación y Aplicaciones, COMPSAC 2020*, 2020, págs. 376–385, doi:10.1109/COMPSAC48688.2020.0-218.

[32] TN Dao, H. Lee, Extracción de características probabilísticas basada en autocodificadores apilados para la detección de intrusiones en la red del dispositivo, IEEE Inter. Things J 9 (16) (2022) 14438–14451, doi:10.1109/JIOT.2021.3078292.

[33] H. Zhang, CQ Wu, S. Gao, Z. Wang, Y. Xu, Y. Liu, Un esquema eficaz basado en aprendizaje profundo para la detección de intrusiones en la red, en:*Procedimiento.-Conferencia Internacional de Reconocimiento de Patrones*, 2018-agosto de 2018, págs. 682–687, doi:10.1109/ICPR.2018.8546162.

[34] C. Wang, CT Butts, JR Hipp, R. Jose, CM Lakon, Imputación múltiple para datos de bordes faltantes: un método de evaluación predictiva con aplicación a Add Health, Soc. Networks 45 (2016) 89–98, doi:10.1016/j.socnet.2015.12.003.

[35] SDD Anton, S. Sinha, H. Dieter Schotten, Detección de intrusiones basada en anomalías en datos industriales con SVM y bosques aleatorios, en: 2019 27th Int. Software de Conferencia, Redes Informáticas de Telecomunicaciones, SoftCOM 2019, 2019, págs. 1–6, doi:10.23919/SOFTCOM.2019.8903672.

[36] J. Gu, S. Lu, Un enfoque eficaz de detección de intrusiones mediante SVM con incrustación de características bayesianas ingenuas, Comput. Secur. 103 (2021) 102158, doi:10.1016/j.cose.2020.102158.
[37] C. Atik, RA Kut, R. Yilmaz, D. Birant, Cadenas de máquinas de vectores de soporte con una novedosa votación de torneo, Electron 12 (11) (2023) 1–16, doi:10.3390/electrónica12112485.
[38] MA Ferrag, L. Maglaras, S. Moschoyiannis, H. Janicke, Aprendizaje profundo para la detección de intrusiones en ciberseguridad: Enfoques, conjuntos de datos y estudio comparativo, J. Inf. Secur. Appl. 50 (2020) 102419, doi: 10.102419 10.1016/j.jisa.2019.102419.
[39] P. Dai, J. Luo, K. Zhao, H. Xing, X. Wu, Autocodificador de eliminación de ruido apilado para la reconstrucción de datos de tráfico faltantes mediante computación de borde móvil, Neural Comput. Appl. 35 (19) (2023) 14259–14274, doi:10.1007/s00521-023-08475-3.
[40] N. Abedzadeh, M. Jacobs y A. Definition, "Un estudio sobre técnicas para conjuntos de datos de sistemas de detección de intrusiones desequilibrados", vol. 17, n.º 1, págs. 9-18, 2023.
[41] S. Choudhary, N. Kesswani, Análisis de los conjuntos de datos KDD-Cup'99, NSL-KDD y UNSW-NB15 mediante aprendizaje profundo en IoT, Procedia Comput. Sci. 167 (2019) (2020) 1561– 1573, doi: 10.1017/j.1561-1573.10.1016/j.procs.2020.03.367.
[42] S. Bhatia, A. Jain, P. Li, R. Kumar, B. Hooi, MStream: Detección rápida de anomalías en transmisiones de múltiples aspectos, en: *Conferencia Web 2021-Actas de la Conferencia Mundial de la Web (WWW) 2021*, 2, 2021, págs. 3371–3382, doi:10.1145/3442381.3450023.
[43] V. Ukani, Conjunto de datos sobre la enfermedad de Parkinson, Kaggle (2020) [En línea]. Disponible: https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set .
[44] Conjunto de datos UNBISCX NSL-KDD 2009, Instituto Canadiense de Ciberseguridad, 2009 [En línea]. Disponible:https://www.unb.ca/cic/datasets/nsl.html.
[45] UNSW, "El conjunto de datos UNSW-NB15", 2015. [En línea]. Disponible: https://research.unsw.edu.au/projects/unsw-nb15-dataset.
[46]AD Woods, D. Gerasimova, B. Van Dusen, J. Nissen, S. Bainter, A. Uzdavines, . . . MM Elsherif, Mejores prácticas para abordar datos faltantes a través de imputación múltiple, Infant and Child Develop. 33 (1) (2024) e2407.
[47]G. Chhabra, V. Vashisht, J. Ranjan, Una comparación de múltiples métodos de imputación para datos con valores faltantes, Indian J. Sci. Technol. (2017).
[48]M. Templ, A. Kowarik, P. Filzmoser, Imputación iterativa de regresión escalonada utilizando métodos estándar y robustos, Comput. Statist. Data Anal. 55 (10) (2011) 2793–2806.