Primera entrega - Reporte de la primera etapa

- Análisis del problema: identificación clara del problema, objetivos y árbol de problemas

  La empresa enfrenta el desafío de identificar actividades de atletismo atípicas que puedan comprometer la equidad de sus competencias deportivas. Aunque los usuarios registran sus actividades con dispositivos confiables, pueden surgir anomalías debido a errores técnicos, uso inadecuado o intentos de fraude, como simular actividad física o alterar la modalidad registrada. Estas irregularidades afectan la validez de los retos y premios

  Objetivos

  Objetivo General

  Desarrollar un sistema automatizado para la detección de actividades de corrida anómalas mediante técnicas de aprendizaje no supervisado utilizando datos de los últimos 5 meses (10/2024 - 02/2025) y empleando métricas cómo velocidad, distancia, elevación, tiempo y frecuencia cardiaca.

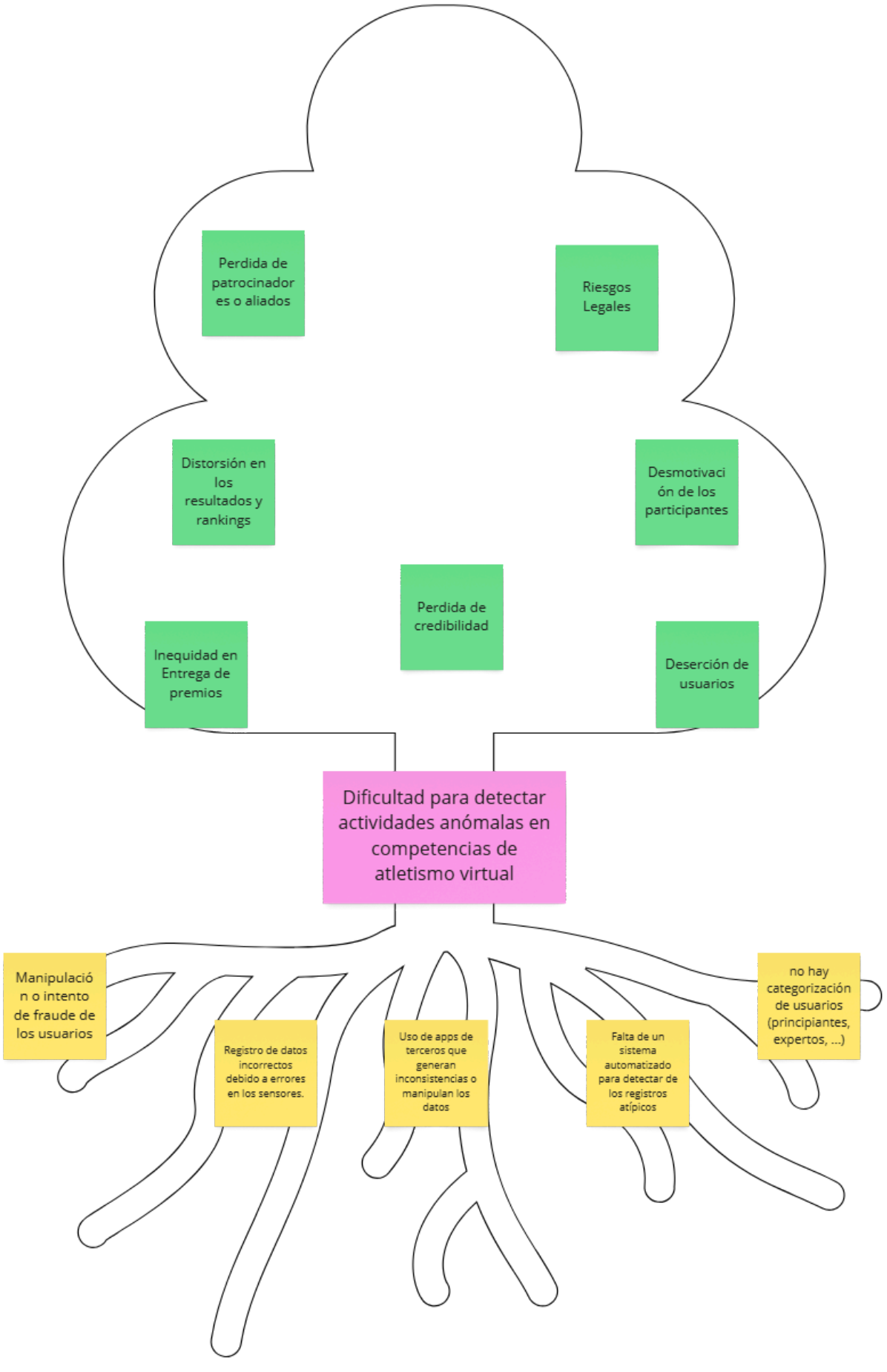
  Objetivos Específicos

  - Implementar al menos tres modelos no supervisados para la detección de anomalías.

  - Evaluar la calidad de los modelos seleccionados mediante métricas apropiadas para cada modelo o visualizaciones para validación cualitativa.

  - Seleccionar un modelo para su posterior análisis y optimización durante la fase final del desarrollo.

  - Analizar los datos obtenidos cómo atípicos para observar posibles patrones de comportamiento.


Perdida de patrocinadores o aliados
Riesgos Legales
Distorsión en los resultados y rankings
Desmotivación de los participantes
Perdida de credibilidad
Inequidad en Entrega de premios
Deserción de usuarios
Dificultad para detectar actividades anómalas en competencias de atletismo virtual
Manipulación o intento de fraude de los usuarios
Registro de datos incorrectos debido a errores en los sensores.
Uso de apps de terceros que generan inconsistencias o manipulan los datos
Falta de un sistema automatizado para detectar de los registros atípicos
no hay categorización de usuarios (principiantes, expertos, ...)

- Estado del arte: revisión bibliográfica y análisis de soluciones existentes

  La detección de fraudes representa una de las áreas más críticas y complejas dentro del campo del aprendizaje automático. A medida que los estafadores perfeccionan continuamente sus métodos para evitar ser detectados, las organizaciones se ven en la necesidad de implementar mecanismos cada vez más sofisticados para protegerse de pérdidas económicas y daños a su reputación. De acuerdo con un informe de LexisNexis, el impacto global del fraude registró un incremento del 32 % en 2022 [1], alcanzando un total de 42.700 millones de dólares.
  En este contexto, el aprendizaje automático ofrece herramientas poderosas para combatir el fraude. Esta disciplina, parte de la inteligencia artificial, permite que los sistemas aprendan a partir de los datos para hacer predicciones o tomar decisiones sin ser programados explícitamente. Gracias a su capacidad para procesar grandes volúmenes de datos, identificar patrones inusuales y segmentar información según sus características, el aprendizaje automático resulta especialmente útil en la detección de actividades fraudulentas. Además, su naturaleza adaptable permite que los modelos se ajusten a nuevas circunstancias y mejoren su eficacia con el tiempo.
  La necesidad de soluciones basadas en aprendizaje automático para la detección de fraude se extiende a múltiples sectores, incluyendo las finanzas, la salud, la educación, los videojuegos, el comercio electrónico, y muchos otros. En cada uno de estos contextos, la detección temprana y precisa de comportamientos anómalos es esencial para garantizar la seguridad, la confianza de los usuarios y la integridad de los sistemas.

  En el artículo [2], los autores presentan una revisión de algoritmos de detección de anomalías aplicados a técnicas de clustering, comparando métodos como K-Means, DBSCAN, HDBSCAN, OPTICS, Local Outlier Factor y Mean Shift. La evaluación se realizó utilizando un conjunto de datos de actividades de atletismo registradas mediante dispositivos *wearables*, con el objetivo de identificar registros anómalos. Definen una anomalía como “una actividad que se desvía significativamente del patrón de comportamiento esperado o normal, determinado a partir de variables como distancia, duración y ritmo”. Para comparar el desempeño de los algoritmos, emplean métricas como precisión, sensibilidad y F1-Score, destacando que DBSCAN obtuvo los mejores resultados en las tres métricas. El modelo propuesto logró detectar patrones anómalos de carrera, incluyendo casos de alto rendimiento, bajo rendimiento y otro tipo de actividades físicas.

  La detección de anomalías en redes informáticas se ha convertido en un área crítica en el contexto de la ciberseguridad. Este artículo evalúa y compara distintos métodos de machine learning aplicados a la detección de anomalías en tráfico de red. Se analizaron los modelos One-Class SVM, DBSCAN, Isolation Forest, NL y LSTM. Utilizaron un dataset real proveniente de tráfico de red de una competencia de defensa cibernética y lo procesaron para aplicar los modelos y evaluarlos con métricas que tienen en cuenta verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos donde concluyeron que Ningún modelo alcanzó un nivel de rendimiento suficiente para aplicaciones críticas en producción sin ajustes adicionales, sin embargo, DBSCAN obtuvo la mejor exactitud promedio entre los

métodos evaluados, Isolation Forest y LSTM mostraron buen balance entre exactitud y eficiencia y todos los modelos presentaron cierta tendencia a generar falsos positivos[3]

De igual forma, se lleva a cabo la detección de anomalías en conjuntos de datos educativos. Para ello, los autores de [4] proponen comparar los algoritmos K-means, DBSCAN, One-Class SVM, Isolation Forest y HBMO con el objetivo de identificar a los estudiantes en riesgo de fracaso escolar. Para ello, disponen de las calificaciones de las asignaturas cursadas durante el año, los promedios trimestrales, así como las evaluaciones extraordinarias y ordinarias. Estos datos son sometidos a un preprocesamiento que incluye el manejo de valores nulos y la aplicación de PCA. Dado que no cuentan con atípicos etiquetados, definen como valores atípicos a aquellos estudiantes que han sido designados para repetir curso o aquellos que son promovidos automáticamente debido a la imposibilidad de repetir. Posteriormente, implementan y evalúan los modelos utilizando métricas como Accuracy, Precision, Recall y F1 Score. A partir de los resultados, concluyen que Isolation Forest ofrece un mejor equilibrio entre Precision y Recall, lo que lo convierte en el algoritmo seleccionado. Esto se debe a que el algoritmo maneja de manera más eficaz datos de alta dimensionalidad y tiene un enfoque de partición superior. Por otro lado, observan que K-means es ideal para conjuntos de datos simples y bien separados, mientras que DBSCAN, One-Class SVM e Isolation Forest son más adecuados para problemas más complejos. Finalmente, no aconsejan el uso de HBMO debido a resultados deficientes y su alto coste computacional, y señalan que LOF no es consistente en todas las dimensiones.

El artículo [5] propone un enfoque automatizado y basado en datos para la detección de intrusiones y anomalías en redes del Internet de las Cosas (IoT). Se aplicaron técnicas como la selección de características basada en información mutua y el algoritmo SMOTE para balancear las clases minoritarias. Además, utilizaron AutoML para seleccionar automáticamente el algoritmo de clasificación más adecuado y ajustar sus hiperparámetros, optimizando el rendimiento sin intervención manual. Los modelos utilizados fueron KNN, SVM, Tree y NN. El enfoque propuesto demuestra que la combinación de técnicas de preprocesamiento de datos y AutoML puede mejorar significativamente la detección de intrusiones y anomalías en redes IoT.

- Planeación: tareas y cronograma

- ☐ Realizar un análisis univariado y multivariado de los registros de actividades para identificar el comportamiento de las variables.
- ☐ Implementar la limpieza del dataset y la transformación de las variables.
- ☐ Implementar y evaluar modelos de aprendizaje no supervisado para la detección de actividades atípicas.
- ☐ Seleccionar uno de los modelos para identificar el proceso de optimización más adecuado
- ☐ Ejecutar el proceso de optimización al modelo seleccionado
- ☐ Evaluar y analizar los resultados del modelo optimizado

Actividad	Duración Estimada	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8
1. Análisis univariado y multivariado de los registros de actividades	1 semana								
2. Limpieza y transformación de las variables	1 semana								
3. Implementación y evaluación de modelos de aprendizaje no supervisado	2 semanas								
4. Selección del modelo para optimización	1 semana								
5. Optimización del modelo seleccionado	2 semanas								
6. Evaluación y análisis de los resultados del modelo optimizado	1 semana								

Referencias

[1] https://risk.lexisnexis.com/-/media/files/financial%20services/research/lnrs-true_cost_of_fraud-study-retail_and_ecommerce-north%20america2-nxr16806-00-0225-en-us.pdf

[2] I. Ursul and A. Pereymybida, "Unsupervised Detection of Anomalous Running Patterns Using Cluster Analysis," *2023 IEEE 13th International Conference on Electronics and Information Technologies (ELIT)*, Lviv, Ukraine, 2023, pp. 148-152, doi: 10.1109/ELIT61488.2023.10310751.

[3] GAJDA, Jakub; KWIECIEŃ, Joanna; CHMIEL, Wojciech. Machine learning methods for anomaly detection in computer networks. En 2022 26th International Conference on Methods and Models in Automation and Robotics (MMAR). IEEE, 2022. p. 276-281.

[4] I. Arbizu Castillón, *Comparación de algoritmos de detección de outliers para prevenir el fracaso escolar*, Trabajo Fin de Grado, Universidad Pública de Navarra, Escuela Técnica Superior de Ingeniería Agronómica y Biociencias, 2024. [En línea]. Disponible en: https://academica-e.unavarra.es/handle/2454/51498

[5] XU, Hao, et al. A data-driven approach for intrusion and anomaly detection using automated machine learning for the Internet of Things. Soft Computing, 2023, vol. 27, no 19, p. 14469-14481.