Sistema automatizado para la detección de actividades de corrida anómalas mediante técnicas de aprendizaje no supervisado

Laura Isabel Chaparro Navia, Fabian Ortiz Collazos y Ricardo Alonso Chicangana Vidal
lauraichaparro11@gmail.com, fabianortiz.iet@gmail.com, rchicangana@gmail.com
Tutor: Milton Sarria Paja

Resumen - Este trabajo presenta el desarrollo de un sistema automatizado para la detección de actividades de atletismo anómalas utilizando técnicas de aprendizaje no supervisado. Se emplearon datos recolectados durante cinco meses que incluyen variables como velocidad, distancia, elevación, tiempo y frecuencia cardiaca. Se implementaron y evaluaron distintos modelos de detección de anomalías, destacando algoritmos como DBSCAN, Isolation Forest y K-Means. Los resultados evidencian el potencial del aprendizaje automático para identificar comportamientos atípicos y mejorar la equidad en competencias deportivas.

Índice de Términos - Anomalías deportivas, Aprendizaje no supervisado, Detección de fraudes, Modelos de clustering.

I. INTRODUCCIÓN

La detección de actividades deportivas anómalas representa un desafío creciente para las plataformas que organizan competencias virtuales de atletismo. Aunque se utilizan dispositivos generalmente confiables para registrar datos como velocidad, distancia, elevación y frecuencia cardíaca, persisten problemas relacionados con registros atípicos, ya sea por errores técnicos, mal uso o fraudes intencionales. Estas irregularidades comprometen la equidad de las competencias y la validez de los resultados.

Este proyecto busca desarrollar un sistema automatizado de detección de anomalías mediante técnicas de aprendizaje no supervisado, con el fin de identificar comportamientos inusuales en actividades de corrida registradas por los usuarios. La importancia de esta solución radica en su capacidad para preservar la integridad de los retos deportivos, prevenir ventajas injustas y fortalecer la confianza de los participantes en la plataforma.

El enfoque propuesto se fundamenta en una revisión del estado del arte sobre métodos de detección de anomalías aplicados en distintos contextos, como la ciberseguridad, el sector educativo y redes IoT. Diversos estudios destacan la efectividad de algoritmos como DBSCAN, Isolation Forest y K-Means para la identificación de patrones atípicos en datos no etiquetados. A partir de esta base teórica, el presente informe describe el proceso de análisis e implementación de modelos aplicados a un conjunto de datos reales de actividades deportivas.

Estado del Arte

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.

Xu et al. [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.

Tabla I. Comparativa de los resultados del estado del arte

Referencia	Contexto de aplicación	Algoritmos evaluados	Técnicas adicionales	Métricas utilizadas	Conclusiones principales
[2] Ursul & Pereymybida	Running con wearables	K-Means, DBSCAN, HDBSCAN, OPTICS, LOF, Mean Shift	N/A	Precisión, Sensibilidad, F1-Score	DBSCAN mejor en precisión, sensibilidad y F1.
[3] Gajda et al.	Redes informáticas	One-Class SVM, DBSCAN, Isolation Forest, NL, LSTM	Preprocesamiento de tráfico real	TP, TN, FP, FN, Exactitud	DBSCAN con mejor exactitud. Isolation Forest y LSTM balancean precisión y eficiencia.
[4] Arbizu Castillón	Educación (riesgo escolar)	K-Means, DBSCAN, One-Class SVM, Isolation Forest, HBMO	PCA, manejo de nulos	Accuracy, Precision, Recall, F1 Score	Isolation Forest balancea Precision y Recall. K-means útil en datos simples.
[5] Xu et al.	Redes IoT	KNN, SVM, Árboles, Redes Neuronales (NN)	SMOTE, selección de características, AutoML	No especificadas (rendimiento general)	AutoML mejora rendimiento. Técnicas de preprocesamiento aumentan eficacia.

II. Materiales y métodos

Marco teórico y metodología

Swetro es una plataforma que permite a los usuarios participar en retos deportivos utilizando dispositivos inteligentes. A medida que la tecnología avanza, también lo hacen las oportunidades de fraude y los errores técnicos. Por tanto, es fundamental implementar mecanismos automáticos de detección de registros anómalos para garantizar la integridad de las competencias.

La detección de anomalías en datos deportivos puede abordarse mediante técnicas de machine learning no supervisado, ya que generalmente no se dispone de etiquetas que identifiquen registros fraudulentos. Entre estas técnicas destacan los métodos de reducción de dimensionalidad como PCA (Análisis de Componentes Principales), y los algoritmos de clustering como DBSCAN, que permiten identificar patrones inusuales sin requerir etiquetas previas.

Enfoque y técnicas utilizadas

Los datos, obtenidos de diversas fuentes como celulares y smartwatches, se procesan y almacenan en una base de datos. Allí se aplican reglas de negocio que ayudan a identificar registros potencialmente irregulares.

Estas pautas incluyen:

- Si la duración de la actividad es menor a 5 minutos.
- Si es una actividad de ciclismo y la elevación ganada supera los 16 metros por minuto (1,000 metros por hora).
- Si es una actividad de running y la elevación ganada supera los 8 metros por minuto (500 metros por hora).
- Si es una actividad de running con un ritmo promedio menor a 3.5 minutos por kilómetro (3:30

min/km).

- Si la actividad no registra distancia recorrida.

Para evitar fugas de información la columna que contiene estas etiquetas no es considerada para el modelo

Filtrado del dataset

Se trabajó exclusivamente con registros de actividades tipo Running, ya que esta categoría contenía la mayor cantidad de datos útiles. Se eliminaron las variables con valores únicos, contenidas en otra variable y las ligadas directamente con el usuario que realizó la actividad

A nivel de registros se eliminaron aquellos que carecían de sentido común teniendo en cuenta la actividad de estudio (running), por ejemplo

- Duraciones menores o iguales a 0 o mayores a 100000 segundos
- Distancias menores o iguales a 0
- Elevaciones mayores a 100000 metros

Exploración visual y estadística

Se analizaron distribuciones y relaciones entre variables como velocidad, duración, elevación, entre otras. En algunos casos se esperaban relaciones más claras que no fueron reflejadas por los datos reales, como es el caso de la duración de la actividad con respecto a la distancia recorrida.

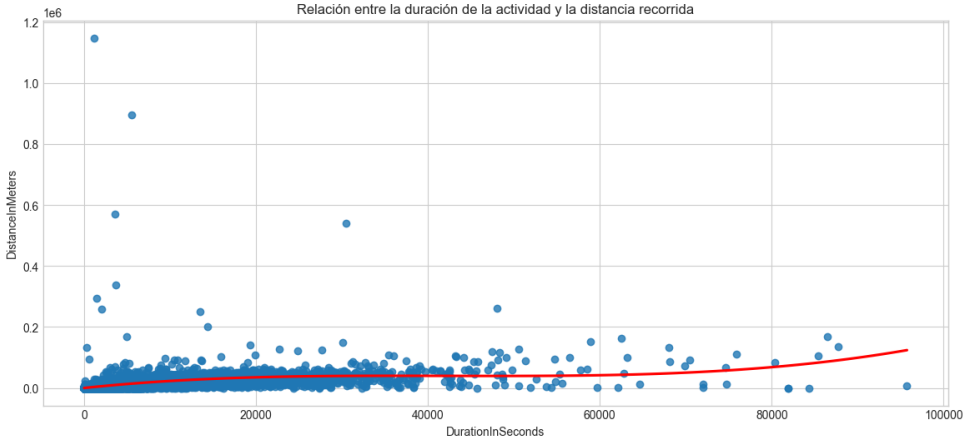

Fig. 1. Relación entre Duración y Distancia

Además, se tenían variables categóricas con valores dominantes que podrían generar un desbalance en los resultados del modelo.

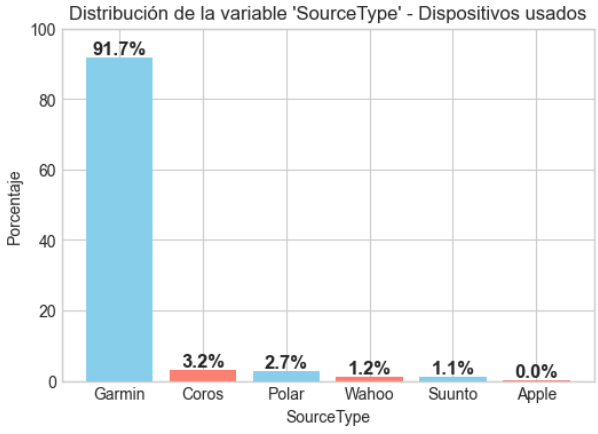

Fig. 2. Marcas de dispositivos que registraron actividades

Preprocesamiento

Esta fase incluyó el manejo de valores faltantes, conversión de formatos de fecha, transformación de unidades y estandarización de las variables para asegurar la consistencia en el análisis.

Las elevaciones ganadas y perdidas tenían una correlación muy alta, por esta razón se creó una variable llamada tipo de terreno así:

- Flat: Si TotalElevationGainInMeters y TotalElevationLossInMeters son 0 o nulos
- Uphill: Si TotalElevationGainInMeters es mayor a 0 y TotalElevationLossInMeters es 0
- Downhill: Si TotalElevationGainInMeters es 0 y TotalElevationLossInMeters es mayor a 0
- Mixed: Si TotalElevationGainInMeters y TotalElevationLossInMeters son mayores a 0

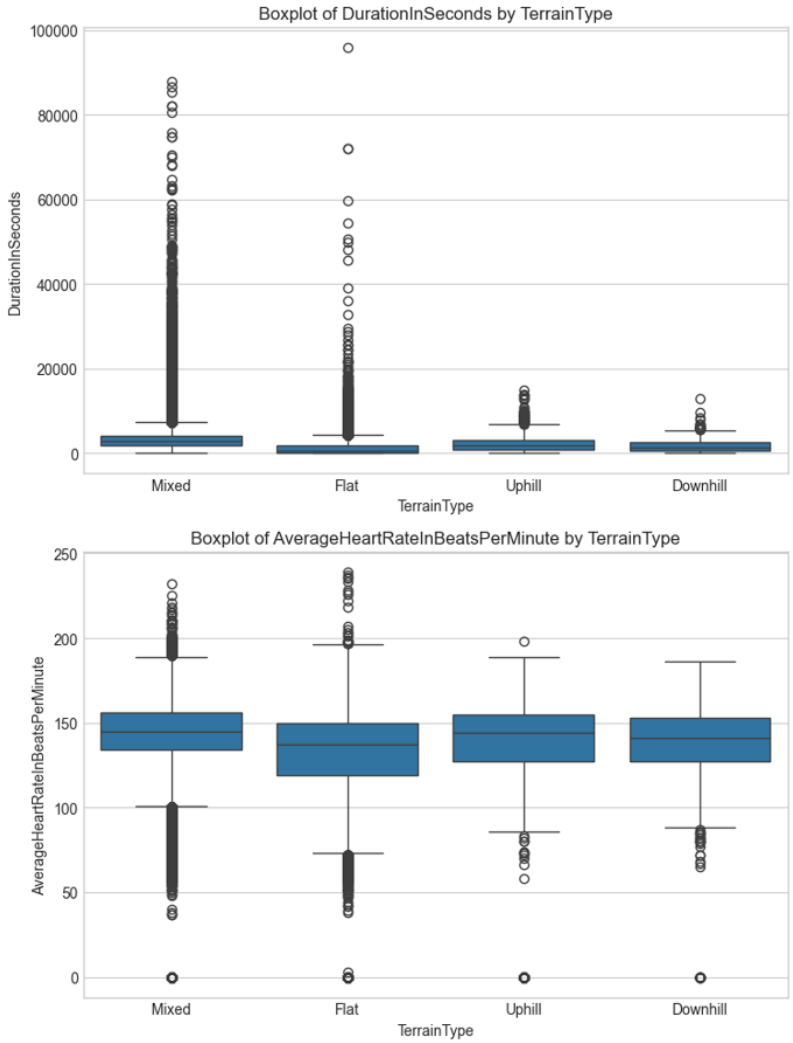

Fig. 3. Resultados del feature engineering

Para el proceso de imputación de datos se usó KNN y se realizó el proceso con algunas de las variables más importantes, no usando todas las disponibles. Luego de eso se volvieron a analizar las variables para validar que no se hubiera afectado el comportamiento de los datos entendido en el análisis inicial.

Después de todo el procesamiento se mantenían valores atípicos (outliers) en variables como DistanceInMeters, AverageSpeedInMetersPerSecond y AverageHeartRateInBeatsPerMinute, lo que indica posibles errores de medición o actividades sospechosas, que son justamente el foco de estudio.

Reducción de dimensionalidad con PCA

Se aplicó PCA para preservar la mayor parte de la varianza con menos dimensiones, facilitando la visualización del clustering. Para la técnica de PCA, se optó por las componentes que explicaban entre el 90% y el 95% de la varianza

Herramientas y Modelos

El desarrollo y análisis se llevó a cabo utilizando Python en un entorno Jupyter Notebook, empleando diversas librerías:

- pandas, numpy para el manejo de datos.
- matplotlib, seaborn para visualización.
- scikit-learn: StandardScaler, PCA, DBSCAN, silhouette_samples, silhouette_score, KMeans, IsolationForest, OneClassSVM, etc.

Se utilizaron dos modelos de clustering: DBSCAN y K-Means y dos de clasificación binaria: SVM e Isolation Forest

Criterios de Evaluación o Métricas

Coeficiente de Silueta (Silhouette Score): Valor promedio obtenido: 0.485, lo que indica una separación moderada entre clusters.

Visualización de silueta: El gráfico de silueta permitió evaluar la cohesión interna de los clusters y la separación entre ellos.

Etiquetas de cluster: Se observaron varios clusters significativos y una proporción notable de puntos etiquetados como ruido, lo cual es deseable en detección de anomalías.

III. RESULTADOS

Se trabajó con un conjunto de datos de 369.320 observaciones correspondientes a actividades de corrida registradas entre octubre de 2024 y febrero de 2025. A partir de este conjunto de datos, se aplicaron técnicas de preprocesamiento y estandarización y posteriormente, se implementaron cuatro modelos para la detección de anomalías: k-means, DBSCAN, SVM e Isolation Forest. A continuación, se presentan los resultados obtenidos con cada uno:

K-means

Para este modelo, se evaluaron distintas cantidades de clusters utilizando las métricas Silhouette Score y Calinski-Harabasz Index. Si bien la métrica de Silhouette sugirió que el valor óptimo podría ser 2 clusters, la visualización correspondiente mostró que uno de los grupos presentaba un coeficiente de Silhouette negativo, lo cual indicaba una mala cohesión interna o una posible asignación incorrecta. Por tanto, este resultado se consideró poco confiable. En contraste, el índice de Calinski-Harabasz indicó que una partición en 13 clusters ofrecía una mejor separación entre los grupos, por lo que se seleccionó esta configuración para el modelo final. En la Figura 4 se puede observar la representación gráfica del coeficiente de Silhouette para el caso de 2 clusters, donde se evidencia la presencia de valores negativos en uno de los grupos.

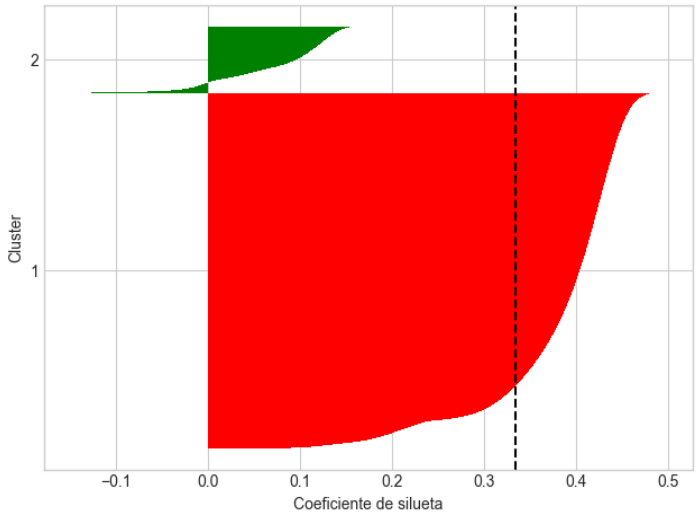

Fig. 4. gráfica de Silhouette para k-means

Para este modelo, los únicos parámetros configurados fueron n_clusters = 13, según lo determinado por el índice de Calinski-Harabasz, y random_state = 42 para garantizar la reproducibilidad de los resultados. Con el fin de representar gráficamente la distribución de los clusters generados, se utilizó PCA (Análisis de Componentes Principales) para reducir la dimensionalidad de los datos a dos componentes principales, obteniendo lo siguiente:

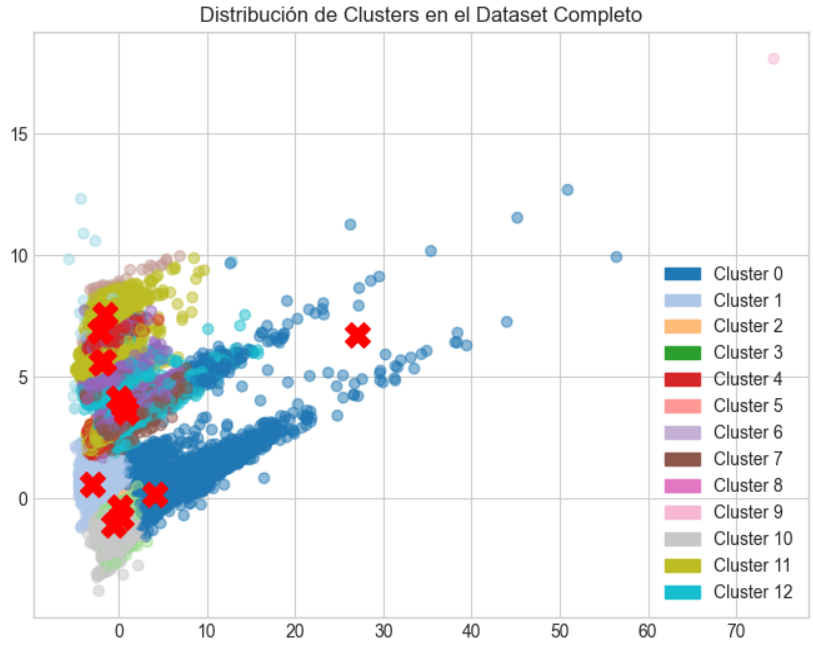

Fig. 5. gráfica de distribución de clusters para k-means

Para identificar anomalías, se calcularon las distancias de cada punto al centroide de su cluster. Se consideraron outliers aquellos puntos cuya distancia excedía el percentil 95, es decir, el 5% más alejado en cada grupo. Este enfoque permitió identificar 18.381 anomalías.

DBSCAN

En este modelo se utilizaron los parámetros eps = 0.4, min_samples = 9 y distance = manhattan. Aunque esta configuración no ofreció un resultado óptimo en términos de separación global, permitió minimizar la

cantidad de clusters con coeficientes de Silhouette negativos. La Figura 6 muestra la distribución de los coeficientes de Silhouette bajo esta configuración.

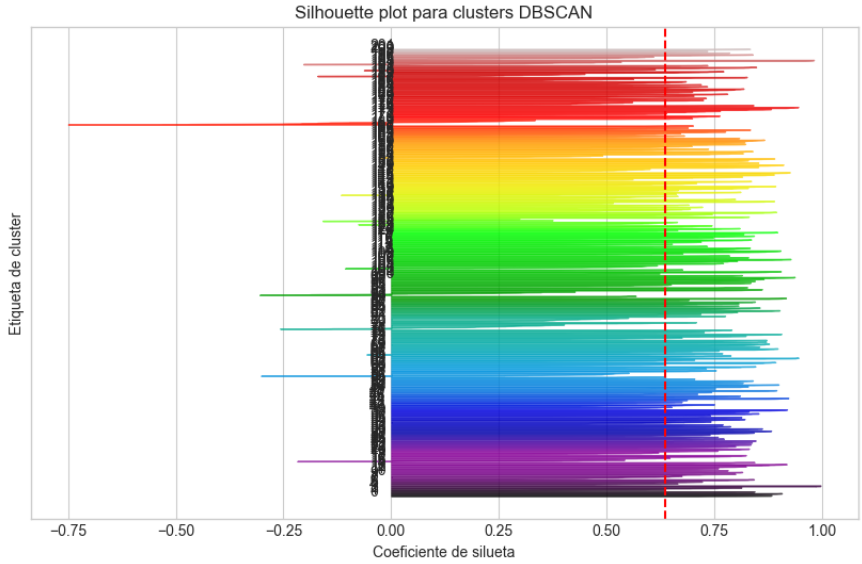

Fig. 6. gráfica de Silhouette para DBSCAN

Al igual que en el modelo k-means, se empleó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos a dos componentes principales y así visualizar gráficamente la distribución de los clusters obtenidos. El resultado de esta proyección se muestra a continuación:

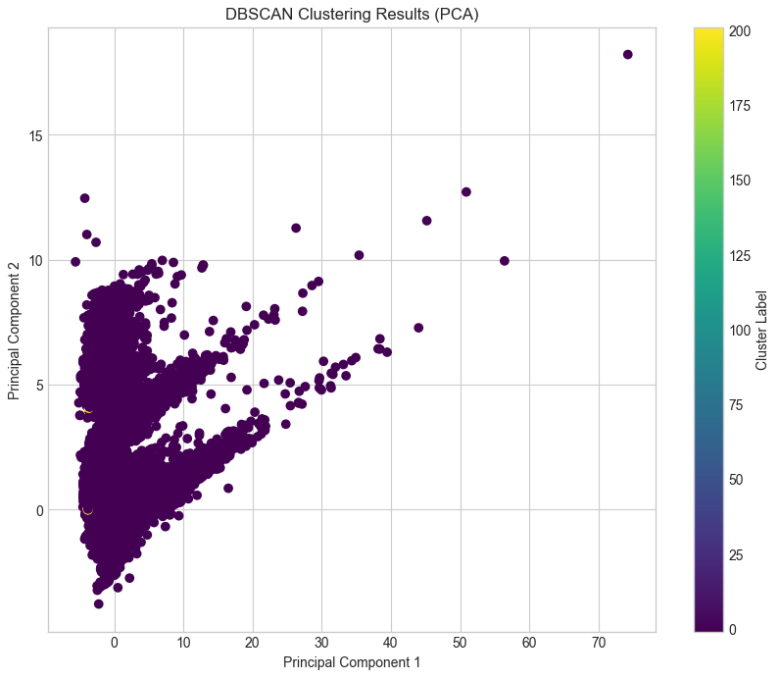

Fig. 7. gráfica de distribución de clusters para DBSCAN

El modelo generó un total de 202 clusters. En el caso de DBSCAN, las observaciones que no se agrupan en ningún cluster denso se etiquetan automáticamente con el valor -1, lo cual indica que han sido clasificadas como ruido o anomalías según la densidad local. Con base en este criterio, se identificaron 364.881 actividades como anómalas.

Para comparar el desempeño de los modelos de clustering aplicados previamente, se implementaron dos modelos de clasificación para detección de anomalías: Support Vector Machine (SVM) en su variante de una clase (One-Class SVM) e Isolation Forest. Aunque estos modelos pertenecen al enfoque supervisado, pueden emplearse en contextos no supervisados cuando se entrenan exclusivamente con datos que se asumen como normales. Posteriormente, las observaciones que no se ajustan al patrón aprendido son clasificadas como anómalas. Esta estrategia permite observar cómo varían los resultados frente a los obtenidos con métodos de agrupamiento no supervisado como k-means y DBSCAN.

SVM

Para este modelo se utilizó la variante One-Class SVM, diseñada para detectar observaciones que se desvían significativamente del patrón general del conjunto de entrenamiento. El modelo fue configurado con los parámetros kernel = rbf, gamma = auto y nu = 0.02.

Una vez entrenado, el modelo clasificó las observaciones como normales (etiqueta 1) o anómalas (etiqueta -1). Con base en esta clasificación, se consideraron como anomalías aquellas observaciones etiquetadas con -1, lo que permitió identificar un total de 7.347 actividades anómalas.

La Figura 8 muestra la distribución de estas clasificaciones en el espacio reducido mediante PCA, donde puede observarse la separación entre las observaciones consideradas normales y las catalogadas como anómalas.

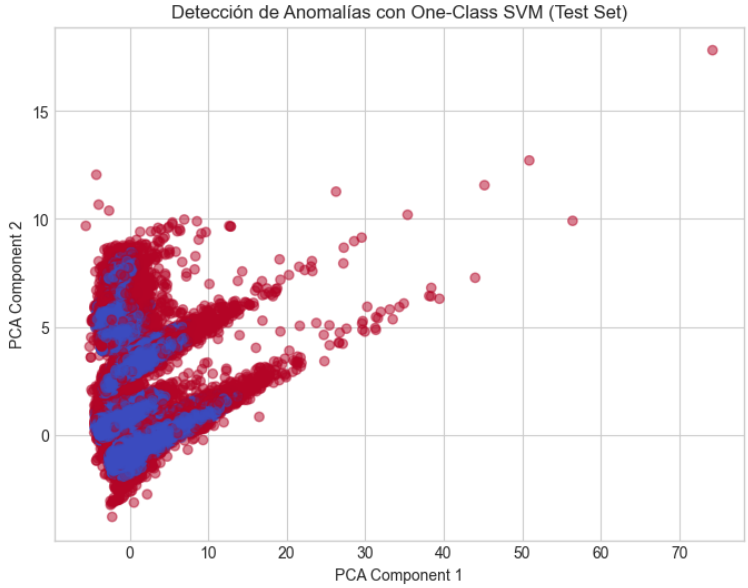

Fig. 8. proyección en dos dimensiones mediante PCA de las clasificaciones generadas por el modelo One-Class SVM

Isolation Forest

Se aplicó el modelo Isolation Forest, configurando el parámetro de contaminación en 0.2, lo que indica que el modelo asume que aproximadamente el 20 % de las observaciones son anómalas. Además, se establecieron los parámetros n_estimators = 100 y random_state = 42 para definir la cantidad de árboles utilizados y asegurar la reproducibilidad de los resultados, respectivamente.

Las predicciones generadas por el modelo clasifican las observaciones como normales (etiqueta 1) o anómalas (etiqueta -1). Bajo esta configuración, se identificaron un total de 73.521 actividades anómalas.

Al igual que en los modelo anteriores en la Figura 9 muestra la distribución de estas clasificaciones en el espacio reducido mediante PCA:

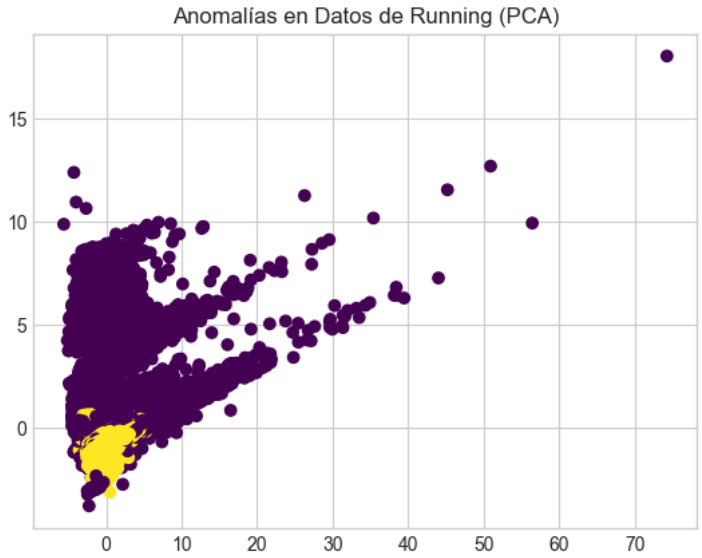

Fig. 9. proyección en dos dimensiones mediante PCA de las clasificaciones generadas por el modelo Isolation Forest

En la Tabla II se enseña un resumen de los clusters y cantidad de anomalías encontradas por cada uno de los modelos:

Tabla II. Resumen de los resultados obtenidos por los modelos

Modelo	Clusters	Cantidad de anomalías detectadas
k-means	13	18.381
DBSCAN	202	364.881
SVM	2	7.347
Isolation Forest Model	2	73.521

IV. Discusión

De acuerdo con los resultados presentados en la sección anterior, DBSCAN fue el modelo que detectó la mayor cantidad de anomalías, clasificando aproximadamente el 98 % de las observaciones como atípicas. Este comportamiento sugiere una configuración inadecuada de sus hiperparámetros, posiblemente debido a un valor de eps demasiado bajo, una elección poco adecuada de min_samples, o al uso de un tipo de distancia que no se ajusta bien a la naturaleza del conjunto de datos.

En consecuencia, es necesario reajustar estos parámetros para lograr una detección más precisa y representativa del comportamiento real de las actividades. Además, al proyectar los resultados mediante PCA, se observa que la distribución de los clusters resultantes no es visualmente clara ni diferenciada.

En contraste, el modelo One-Class SVM mostró un comportamiento más conservador, detectando 7.347 anomalías, lo que representa una fracción considerablemente menor del conjunto de datos. Su desempeño está altamente influenciado por el valor del parámetro nu, que en este caso fue configurado en 0.02, restringiendo intencionalmente la proporción de observaciones que el modelo puede considerar como atípicas. Además, este modelo solo permite una clasificación binaria (normal o anómalo), lo que impide capturar posibles subgrupos de comportamiento anómalo dentro de los datos y puede reducir su capacidad para representar la complejidad del conjunto de actividades. A pesar de lo anterior en los resultados gráficos con PCA se pudo observar una distribución un poco más clara que en el modelo anterior.

Por otro lado, Isolation Forest mostró un desempeño más equilibrado en comparación con otros modelos, de hecho al proyectar los resultados mediante PCA, el modelo Isolation Forest mostró la separación más clara entre las clases (normales y anómalas) en comparación con los demás modelos evaluados, sin embargo, una de sus principales limitaciones es que requiere definir a priori la proporción de anomalías esperadas mediante el parámetro de contaminación. Esta necesidad implica asumir un valor sin conocer realmente la distribución

de las anomalías, lo que puede afectar la objetividad del modelo. En contextos donde no se tiene información clara sobre la tasa real de anomalías, esta configuración puede conducir a una subestimación o sobreestimación del número de casos anómalos, limitando la capacidad del modelo para adaptarse de forma dinámica a las características reales del conjunto de datos.

Finalmente, en el caso de k-means, el modelo permitió identificar anomalías a partir de la distancia de cada observación al centroide de su respectivo cluster, considerando como atípicas aquellas que se encontraban por encima del percentil 95 de dicha distancia. Esta estrategia resultó útil para detectar observaciones que se alejaban significativamente del comportamiento típico dentro de cada grupo, identificando un total de 18.381 actividades anómalas. No obstante, al analizar la distribución de los clusters mediante PCA, se observa que algunos grupos no presentan una conformación claramente definida, con solapamientos visibles entre ciertos clusters.

Es importante tener en cuenta que k-means no fue diseñado específicamente para la detección de anomalías, por lo que su efectividad depende en gran medida de una adecuada elección del número de clusters. En este caso, se utilizaron 13 clusters, seleccionados con base en el índice de Calinski-Harabasz, dado que la métrica de Silhouette no arrojó un resultado concluyente.

A pesar de estas limitaciones, k-means se posiciona como el modelo que ofrece resultados más coherentes entre los algoritmos evaluados. A diferencia de DBSCAN, que clasificó como anómalas casi todas las observaciones, y de Isolation Forest, que requiere definir previamente la proporción esperada de anomalías. Por lo tanto, se analizará en mayor detalle los resultados entregados por este modelo:

Al analizar los clusters generados por K-Means, se observa que los clusters 0 y 8 presentan altos valores de distancia (Figura 10), destacando especialmente el cluster 0, donde se identifican posibles valores atípicos. Además, tanto en el cluster 0 como en el 12 se evidencian valores inusuales en la métrica de duración (Figura 11).

Además, en la Figura 12 se observó que los registros atípicos se presentaban principalmente en terrenos flat y mixed, con mayor porcentaje en mixed. En cuanto a los dispositivos, Garmin fue el que registró la mayor cantidad de datos atípicos (Figura 13). La hora con mayor presencia de anomalías fue las entre las 6:00 y 7:00 AM (Figura 14). Al analizar la combinación de terreno y dispositivo, se identificó que la mayor cantidad de registros atípicos correspondía a Garmin en terreno Mixed.

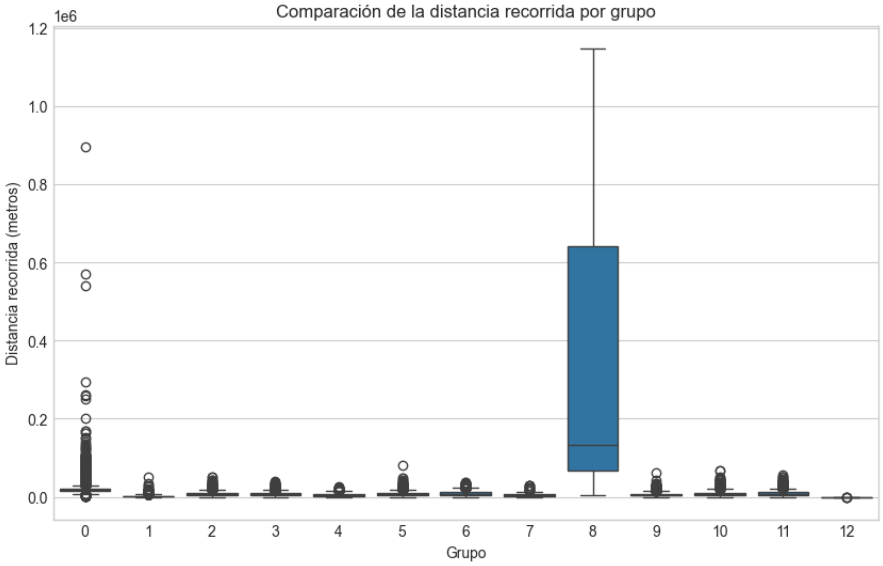

Fig. 10. Comparación de la distancia recorrida por cluster

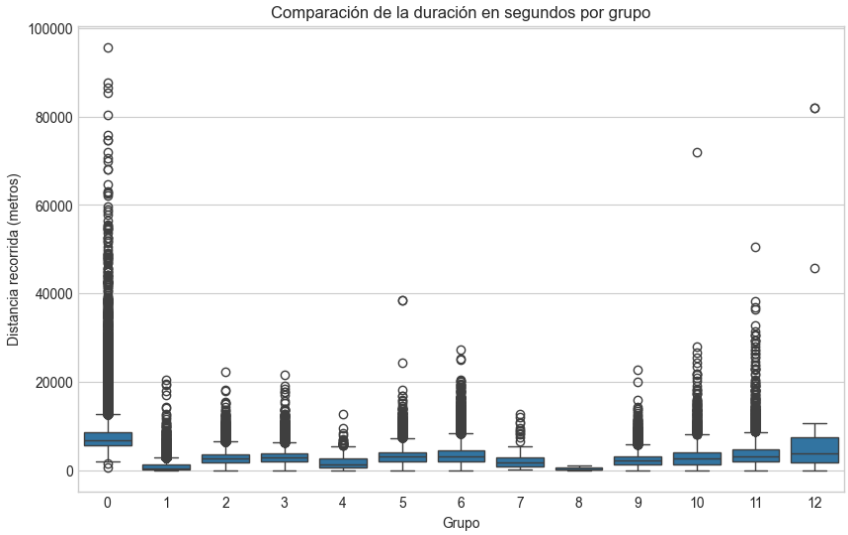

Fig. 11. Comparación de la duración recorrida por cluster

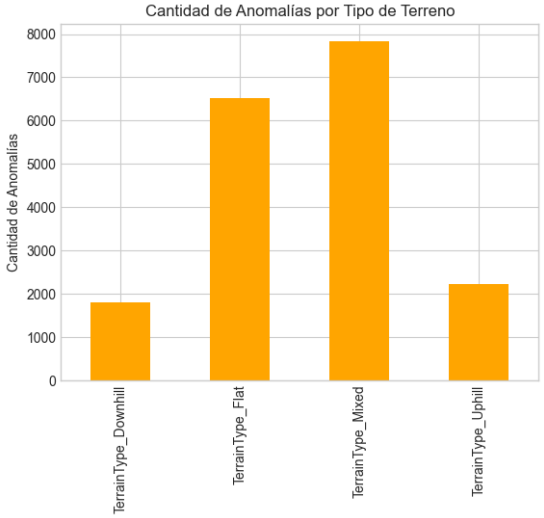

Fig. 12. Cantidad de anomalías por tipo de terreno

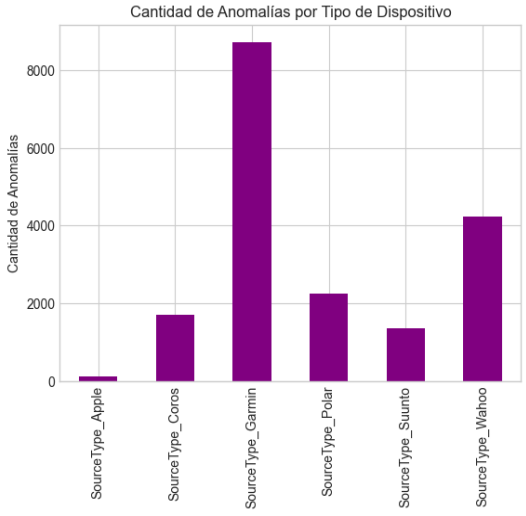

Fig. 13. Cantidad de anomalías por marca del dispositivo

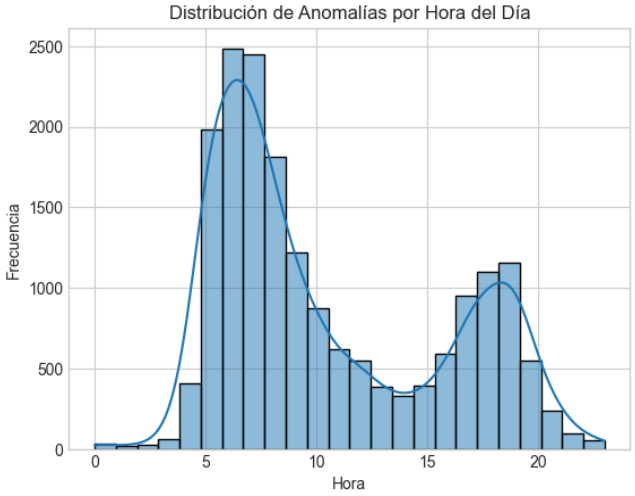

Fig. 14. Distribución de anomalías por hora del día

En general, los resultados obtenidos evidencian que la configuración de los modelos tiene un impacto significativo en la detección de anomalías, y que no todos los enfoques aplicados lograron representar de forma adecuada la complejidad y variabilidad del conjunto de datos.

Estos hallazgos reflejan la necesidad de revisar y afinar la configuración de los algoritmos, ya sea mediante ajustes en los hiperparámetros, el uso de técnicas de validación más robustas o incluso la integración de métodos híbridos que combinen diferentes enfoques.

Por lo tanto, se concluye que si bien los resultados actuales son un punto de partida útil, aún es necesario un proceso de optimización y validación más profundo para garantizar que las detecciones sean precisas, relevantes y aplicables en contextos reales de evaluación de actividades deportivas.

V. Conclusiones

El presente estudio desarrolló e implementó un sistema automatizado para la detección de actividades anómalas en retos virtuales de atletismo, utilizando técnicas de aprendizaje no supervisado sobre un conjunto de datos reales recolectados durante cinco meses. Se evaluaron modelos como DBSCAN, K-Means, One-Class SVM e Isolation Forest, con el objetivo de identificar comportamientos atípicos que pudieran comprometer la equidad de las competencias.

K-means y DBSCAN ofrecen la ventaja de identificar patrones sin requerir una estimación previa de la proporción de datos atípicos. Al tratarse de algoritmos de clustering, permiten además descubrir distintos grupos de comportamiento, a diferencia de modelos como SVM e Isolation Forest, que se limitan a una clasificación binaria entre observaciones normales y anómalas.

No obstante, entre estos dos enfoques, k-means fue seleccionado como el modelo más apropiado para un análisis posterior, ya que, si bien requiere definir previamente el número de clusters, logró una detección más equilibrada de anomalías en comparación con DBSCAN, que clasificó cerca del 98 % de los registros como atípicos. Cabe señalar que parte de estas diferencias puede estar influenciada por la configuración de los hiperparámetros, lo que resalta la importancia de ajustar cuidadosamente cada modelo según la naturaleza de los datos.

El estudio confirma el cumplimiento del objetivo principal: demostrar la viabilidad de un sistema automatizado basado en técnicas de aprendizaje automático para detectar registros sospechosos, fraudes o errores técnicos en actividades deportivas digitales. Además, se evidenció que una correcta etapa de preprocesamiento, junto con la reducción de dimensionalidad mediante PCA, mejora significativamente el desempeño de los modelos aplicados.

Como recomendaciones futuras, se sugiere:

- Encontrar y optimizar los parametros más adecuados para cada uno de los modelos, para encontrar resultados que puedan capturar de mejor manera el comportamiento de los datos.
- Integrar estos modelos en tiempo real dentro de plataformas como Swetro, para la evaluación automática de nuevas actividades.
- Explorar enfoques híbridos que combinen aprendizaje no supervisado con reglas heurísticas o retroalimentación humana para reducir los falsos positivos.
- Aplicar metodologías AutoML para optimizar parámetros críticos como eps, min_samples o la tasa de contaminación.
- Validar los resultados utilizando conjuntos de datos con etiquetas verificadas de fraudes conocidos, con el fin de afinar la precisión y la robustez del sistema.
- En resumen, este trabajo aporta una solución práctica y escalable para fortalecer la transparencia y la integridad en entornos de competencia deportiva digital, y sienta las bases para futuras investigaciones en detección automática de anomalías en otros dominios.

ANEXOS

Link del video de presentación: https://www.youtube.com/watch?v=UdXmoOBOUD4

REFERENCIAS

[1] LexisNexis, "True Cost of Fraud Study – 2022", [en línea]. Disponible en: [URL]

[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.