[213017] Comparación de metodologías de detección de valores atípicos en la variable VO_2 de la prueba cardiopulmonar de ejercicio (CPET).

Johan Sebastián Barrera Virgüez [a,d] , María Mónica Diaz Sarmiento [a,d] , Santiago Andrés Luque Ochoa[a,d], María Juliana Serrano Vargas [a,d]

Ana María Beltrán Cortés[b,d], Manuel Augusto Cárdenas Romero[c,d]

[a]*Estudiante de Ingeniería Industrial*
[b]*Profesora, Directora del Trabajo de Grado, Departamento de Ingeniería Industrial*
[c]*Profesor, Co-director del Trabajo de Grado, Departamento de Ciencias Fisiológicas - Facultad de Medicina*
[d]*Pontificia Universidad Javeriana, Bogotá, Colombia*

Resumen de diseño en Ingeniería

This document seeks to compare outlier detecting methods in the variable of oxygen consumption (VO_2) in the cardiopulmonary exercise test. With the selected method, outliers were detected and the maximum VO_2 value was calculated to determine if it changed with detection, which would imply a different interpretation of the test for experts.
Cardiopulmonary exercise testing (CPET) assesses the function of a patient's cardiorespiratory system and its functional capacity. Analyzing CPET results is challenging for medical staff because they are voluminous, contain a lot of information, and come in a raw format that makes direct analysis difficult. In addition, it presents a high variability that depends on the patient, the variables measured and the test execution process; all these factors generate data that depart from the usual pattern and that can alter the analysis of results. For this reason, it is necessary to preprocess the data obtained and detect outliers in order to facilitate the reading of the results and increase the reliability of the interpretation.

To develop the degree work, 407 CPET were used, obtained from the research project "Ejercicio físico en sujetos sanos no entrenados nativos a altitud moderada en comparación con crónicamente aclimatados "financed by Minciencias (Id: 120356934972) and the PUJ, collected between the year 2016 and 2018. The project seeks to make a comparison of six non-parametric outlier detection methods in the data set of the VO_2 variable of the CPET to identify the most appropriate method, through performance metrics, and to evaluate the impact of the applied methodologies. Additionally, it was created a desktop app that allow detection to be applied with the selected method, calculate an impact variable before and after detection, and also generate a graphical representation of the VO_2 variable against time before and after removing its atypical values.

The comparison between detection methods showed that the Isolation Forest was the one that obtained the best score in the F_b metric. By applying the method, a significant change in the value of the VO_2 maximum was identified before and after the detection. These results were presented to three expert evaluators in CPET through the desktop app where they gave us feedback and presented proposals for improvement. The comments were positive, because the evaluators mentioned that the app facilitates the interpretation of the test and reduces the subjectivity of the outlier detection.

Palabras claves: Valores atípicos, Prueba Cardiopulmonar de Ejercicio, Métodos de detección de atípicos, Preprocesamiento de datos

1. Justificación y planteamiento del problema

La prueba cardiopulmonar de ejercicio (CPET por sus siglas en inglés) es un procedimiento médico cuyo objetivo general es imponer un estrés sustancial en los sistemas orgánicos involucrados en la respuesta integral al ejercicio (respiratorio, cardiovascular, hematológico y muscular), para

observar el patrón de la respuesta y derivar conclusiones clínicas [1]. Esta prueba permite estudiar y determinar las posibles limitantes cardiorrespiratorias en pacientes con intolerancia al ejercicio [2]. En principio es una prueba no invasiva y se compone de varias etapas: evaluación del estado clínico, preparación, administración e interpretación de resultados [3].

Según la universidad Johns Hopkins, julio de 2021 cerró con 198 millones de casos confirmados de COVID-19 y más de cuatro millones de muertes a nivel mundial. El virus SARS-COV-2 afecta varios sistemas orgánicos incluyendo el respiratorio, el cardiovascular, y en algunos casos, el neurológico [4].

Si bien el virus permanece pocas semanas en el organismo, la resolución de la infección no implica la recuperación completa del paciente. Un estudio publicado en BMC Medicine muestra que de los pacientes infectados que estuvieron hospitalizados, el 63.9% presentaban sintomatología en los seis meses siguientes al egreso e indica que estos síntomas en su mayoría son respiratorios, sistémicos y neurológicos. Al conjunto de secuelas posteriores a la infección por SARS-COV-2 se les conoce como COVID largo [5]. En estos casos es necesario determinar el tipo de compromiso para definir las alternativas de tratamiento, ya que son muy amplias; por ejemplo, cuando el sistema respiratorio es el principalmente afectado, el tratamiento puede ir desde terapia respiratoria, con o sin oxígeno, hasta trasplante de pulmón por fibrosis pulmonar [6].

Más de una cuarta parte de los sujetos que superaron la fase aguda del COVID-19 presentan ineficiencia ventilatoria durante el ejercicio asociada a una lenta recuperación de la frecuencia cardíaca; en estos pacientes la CPET permite evaluar el verdadero impacto del COVID-19 y detectar alteraciones ventilatorias y cardiovasculares, lo cual representa una nueva indicación de la prueba [7].

Por todo lo anterior, se ha registrado un aumento en la demanda de pruebas CPET en individuos que están en proceso de recuperación del SARS-COV-2 [8]. La rápida adopción de la CPET requiere una estandarización y un control de calidad rigurosos que garanticen la validez y reproducibilidad de los resultados para la toma de decisiones clínicas, por ejemplo, para definir el pronóstico entre alternativas de tratamiento quirúrgicas y no quirúrgicas [9].

Una limitación para la estandarización de la CPET radica en que los datos arrojados por la prueba son voluminosos, integran mucha información y vienen en un formato en bruto que dificulta el análisis directo [10]. En particular, los datos obtenidos en la CPET máxima incremental (ya sea escalonada o en rampa), son probablemente el conjunto de resultados más difícil de interpretar en un laboratorio de función pulmonar [11]; de hecho, existe la posibilidad de una interpretación inconsistente, y a veces inexacta, de los resultados de la prueba [12]. Esto representa un obstáculo para la rápida adopción de la CPET en momentos de alta demanda, especialmente teniendo en cuenta que la Asociación Torácica Americana en conjunto con el Colegio Americano de Médicos del Tórax (ATS/ACCP) mencionan que:

> "un alto nivel de comprensión de la CPET es útil en un amplio espectro de entornos clínicos. Su impacto se puede apreciar en todas las fases de la toma de decisiones clínicas, incluido el diagnóstico, la evaluación de la gravedad, la progresión de la enfermedad, el pronóstico y la respuesta al tratamiento" [3].

Existen diferentes enfoques para interpretar la CPET. El más recomendado es el enfoque integrador, que enfatiza en las interrelaciones entre variables, los fenómenos de tendencia y los patrones de respuesta de las variables clave para identificar una alteración clínica; este enfoque permite una evaluación más exacta al considerar la relación de diferentes medidas [3].

En la CPET se mide, en cada respiración, más de 100 variables entre primarias y derivadas, pero de acuerdo con el enfoque integrador es necesario identificar las variables clave de la prueba para la interpretación. La principal variable es el consumo de oxígeno (VO_2), que cuantifica la demanda de oxígeno celular en un momento determinado (reposo o ejercicio) y hace referencia al volumen de oxígeno utilizado por el organismo en un periodo de tiempo determinado, expresado en mililitros o litros por minuto [3] . Las otras variables clave son la frecuencia cardiaca (HR), la saturación arterial de la hemoglobina con oxígeno (SaO_2) y la ventilación total espirada (VE).

La interrelación gráfica de las variables recomendada para la interpretación de la CPET se presenta en la Tabla 1 en la cual se evidencia que el VO_2 hace parte de todas las gráficas sugeridas para interpretación de resultados.

Eje Y	**Eje X**
VO_2	VO_2 o V_{CO2}
V_E	VO_2
V_T y f_R	VO_2
HR y O_2	VO_2
V_{CO2}	VO_2
V_E/VO_2 y V_E/V_{CO2}	VO_2
P_{ETo2} y P_{Etco2}	VO_2
Pa_{O2}, $P_{(A-a)O2}$, y SaO_2	VO_2
Pa_{CO2} y V_D/V_T	VO_2
$[La^-]$ o HCO_3^-	VO_2

Tabla 1: Interrelaciones gráficas sugeridas de variables para interpretación de la prueba de ejercicio cardiopulmonar. Tomado y traducido de ("ATS/ACCP Statement on Cardiopulmonary Exercise Testing.," 2003) [3]

Para arribar a la generación de información a partir de los datos generados por la CPET, y evitar las posibles consecuencias de una interpretación sesgada por la presencia de datos anómalos, los datos deben someterse a un proceso de preprocesamiento antes de ser interpretados [3]. Esta etapa se encuentra después de la selección de los datos, y antes de las etapas de transformación, minería, e interpretación [13]. El preprocesamiento de datos, aunque no tiene establecido un orden inmutable, se suele componer de las siguientes fases: limpieza de datos, donde se corrigen los datos que pueden ser incorrectos; manejo de datos faltantes, con diferentes abordajes; y finalmente, detección de atípicos o datos anómalos [14].

Adicionalmente, los datos generados en la CPET se pueden ver afectados tanto por el proceso de recolección como por las limitaciones y restricciones técnicas de los dispositivos utilizados [12]. Son múltiples los factores que pueden contribuir a la variabilidad de los datos incluyendo: cambios en la medicación del paciente, motivación del individuo al momento de realizar la prueba y hora del día en que se aplica, procedimientos de la prueba, errores en el equipo, calibración, y de igual forma las instrucciones e inducción del personal médico al paciente [3]. Del mismo modo, una de las causas principales de la aparición de valores atípicos se debe a errores en ejecución [15]. Para el caso de la CPET estos errores se presentan en eventos como pérdidas del registro por la caída de la máscara, o introducción de artificios como cuando el paciente habla o tose durante la prueba, o fisuras en los conductos utilizados.

Es necesario proponer soluciones al problema de adopción del enfoque integrador; la presencia de valores atípicos se presenta como un obstáculo para su adopción dada la amplia gama de problemas asociados a los datos biomédicos: ruido, valores faltantes y alta dimensionalidad [16].Una posible solución es preprocesar y transformar los datos antes de analizarlos [17] con el fin de obtener resultados confiables y comprensibles.

De acuerdo con lo anterior, dos de los pasos de preprocesamiento más importantes para el análisis de datos son la detección de valores atípicos y el manejo del ruido. Los primeros, son aquellos registros que se alejan significativamente del comportamiento general de los datos [18], se encuentran ubicados de una manera distante con relación al patrón de datos [19], no presentan un comportamiento usual [20] o provienen de una distribución distinta a la de los demás datos.

El manejo del ruido incluye todos aquellos casos donde existen valores con formatos incorrectos (como cuando un valor que debe ser numérico es representado por texto), los valores son erróneos (por ejemplo, incluir valores negativos para indicar la edad), o cuando no existen valores registrados para un atributo [21]. A su vez, el ruido que se puede presentar en los datos generados por la CPET máxima es un obstáculo a la hora de interpretarlos, pues hace que la visualización de todas las respiraciones para una determinada variable (por ejemplo, consumo de oxígeno) sea confusa [22]. En la CPET máxima diferenciar un patrón de respiración caótico del ruido normal de respiración a respiración puede resultar complejo si los datos graficados no se suavizan adecuadamente [23]. En consecuencia, no tener en cuenta la detección de valores atípicos y el tratamiento del ruido de la prueba puede afectar los resultados del análisis [24].

La Tabla 2 presenta cuatro gráficas referentes a los datos de la variable VO_2, provenientes de pruebas CPET máximas en rampa en individuos jóvenes aparentemente sanos. Se observa una alta variabilidad en los datos e incluso puntos extremos alejados del patrón de la gráfica (los que consideraremos atípicos en este trabajo). Estas observaciones varían de acuerdo con el individuo, con el momento de la prueba y con los eventuales problemas en la ejecución.

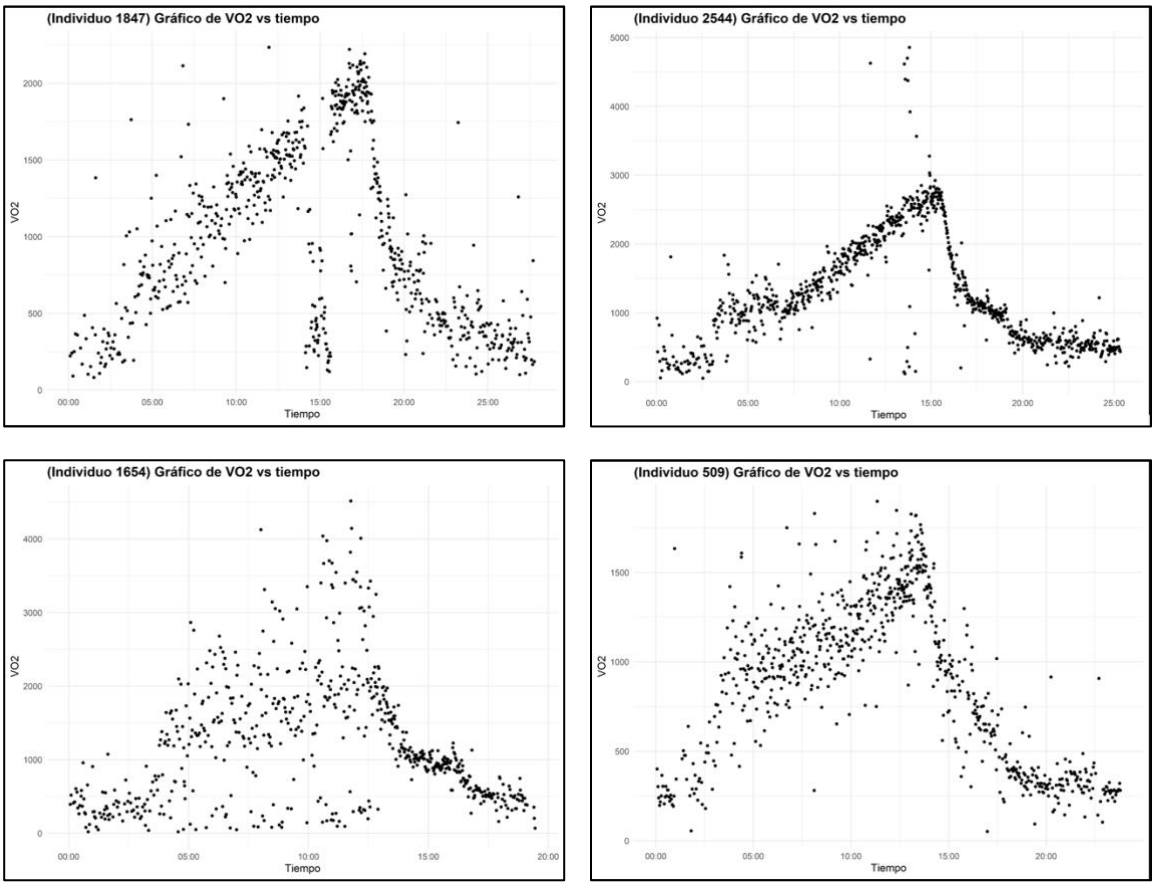

Tabla 2: Gráficas construidas con datos del proyecto de investigación "Ejercicio físico en sujetos sanos no entrenados nativos a altitud moderada en comparación con crónicamente aclimatados"

La optimización y mejoramiento de la visualización de los datos de la CPET máxima aumenta la fiabilidad, interpretabilidad y reproducibilidad de los resultados [22] por lo que es necesario contar con un despliegue gráfico capaz de mostrar las interrelaciones existentes entre las distintas variables medidas durante una CPET máxima [3], e incluso, el despliegue gráfico de algunas variables se considera mandatorio para diagnosticar en forma rápida y confiable [22].

La prueba CPET permite evaluar la función del sistema cardiorrespiratorio de un paciente y su capacidad funcional. Dentro de las aplicaciones más recientes de la CPET está la detección de COVID largo en pacientes que manifiestan fatiga prolongada o insuficiencia al ejercicio. En cuanto a la interpretación de la prueba, su confiabilidad puede incrementarse si se emplean técnicas estadísticas de detección y tratamiento de valores atípicos asociados a la variabilidad intrínseca de la respiración.

Para que el preprocesamiento de los datos generados por la CPET se haga de manera apropiada es necesario evaluar y comparar diferentes métodos de detección de valores atípicos en los datos de la variable VO_2 y elegir el de mejor desempeño. Con este trabajo se busca responder la siguiente pregunta de investigación:

¿Qué método de detección de valores atípicos es el más apropiado para ser aplicado en el conjunto de datos de la variable VO_2 que arroja la CPET?

2. Antecedentes

La revisión bibliográfica se llevó a cabo durante los meses de junio, julio, agosto, y septiembre de 2021 en la base de datos Scopus. Los términos identificados como más relevantes para el desarrollo de la presente investigación fueron: “Outlier”, “CPET”, “COVID”, “medical test”, y “Outlier detection method”. La búsqueda se limitó a artículos en inglés y español y un resumen de los resultados se presenta en la Tabla 3.

Búsqueda	Ecuación de búsqueda	Número de resultados
#1	Outlier AND cpet	1
#2	Outlier AND detection AND method	10.187
#3	Anomaly AND detection AND cpet	0
#4	Outlier AND medical AND test	467
#5	Cpet	2.032
#6	Cpet AND Covid	17

Tabla 3. Ecuaciones de búsqueda

Las fuentes revisadas se agruparon en cuatro secciones: detección de atípicos, métricas de desempeño para la comparación de metodologías de detección de atípicos, aplicativos y software que soportan el proceso diagnóstico en el sector de la salud, y aplicativos para interpretación de CPET.

Detección de atípicos

El preprocesamiento o preparación de datos es un aspecto importante que se debe realizar antes de analizar conjuntos de información de cualquier naturaleza [19], esto se debe a que los datos del mundo real suelen tener una alta variabilidad asociada al fenómeno medido y es usual que estén incompletos,

sean ruidosos y presenten inconsistencias que dificulten la detección de patrones útiles para tareas de regresión, clasificación y/o agrupamiento [25].

El tipo de preprocesamiento aplicado varía de acuerdo con el conjunto de información que se va a analizar. Por su forma de obtención, particularmente los datos clínicos presentan problemas de disponibilidad de datos, valores atípicos y modelos de representación complejos [19]; por consiguiente, previo al análisis de datos se debe realizar el respectivo preprocesamiento [17].

Dentro del preprocesamiento, la detección de registros atípicos es una de las tareas fundamentales del análisis de datos [26]. Según Abir Smiti [21] los métodos para la detección de valores atípicos se pueden clasificar en las siguientes categorías: métodos estadísticos, donde una observación es atípica si se desvía significativamente de una distribución estándar; métodos basados en distancia, los cuales asumen que un valor es atípico si la proximidad del objeto a sus vecinos se desvía significativamente de la proximidad de la mayoría de los otros objetos a sus vecinos; métodos basados en clústeres, que asumen que los datos regulares (datos no atípicos) pertenecen a grupos grandes y densos, mientras que los atípicos pertenecen a grupos pequeños, dispersos o a ninguno [25]; métodos basados en densidad, en donde un valor es atípico si su densidad local difiere de la de los demás datos ubicados en esa región [21] y, finalmente, métodos basados en aislamiento, que aíslan cada punto de un conjunto de datos de los demás, realizando particiones aleatorias para identificar valores atípicos. [27].

A continuación, se presentan algunas generalidades de los tipos de métodos mencionados anteriormente, con ejemplos de algunos algoritmos que responden a sus características.

Métodos estadísticos

Los métodos estadísticos para la detección de valores atípicos pueden ser categorizados en dos grupos: paramétricos y no paramétricos. Los métodos paramétricos requieren que se cumplan supuestos distribucionales sobre los datos, mientras que los métodos no paramétricos liberan a los datos de supuestos sobre su distribución de procedencia [26].

Dentro de los métodos paramétricos, el de la puntuación z es probablemente el más utilizado a la hora de detectar atípicos. El valor z de una observación representa la cantidad de desviaciones estándar que esta se aleja de la media de los datos. El método del valor z se aplica sobre datos provenientes de una distribución normal (supuesto de normalidad) [28] . La idea central del método es calcular el valor z para cada dato: si el valor absoluto es mayor a 3, el dato se considera atípico. Sin embargo, esta metodología presenta graves problemas al ser aplicada en conjuntos de datos pequeños o cuyo comportamiento no se aproxime al de la campana de Gauss [29].

Por otra parte, dentro de los métodos no paramétricos, uno de los más implementados para detectar atípicos es el del *Rango Intercuartílico* (IQR por sus siglas en inglés). Este método procede de la siguiente manera: sobre los datos se hallan los tres cuartiles (Q1, Q2 y Q3), se calcula el IQR como Q3-Q1 y se definen las fronteras de datos interiores como Q1-1.5IQR y Q3+1.5IQR. La regla de decisión indica que todo valor que esté fuera del intervalo [Q1-1.5IQR ; Q3+1.5IQR] se considera atípico [30].

Otro método no paramétrico empleado en la detección de valores atípicos es la *Estimación de Densidad Kernel* (KDE por sus siglas en inglés). El objetivo de KDE es estimar una función de densidad para los datos analizados sin importar la distribución de la cual provienen. Esta técnica se ha utilizado en conjunto con otras obteniendo los métodos *Local Kernel Density Estimation using*

Volcano Kernel function [31], *Kernel Density Estimation for Outlier detection* [32], y *Local Density Estimate using Kernel functions* [33]. Se han planteado variaciones en cuanto a la implementación de KDE en la detección de atípicos evaluando su desempeño en datos tanto reales como sintéticos (simulados) y se destaca su desempeño en conjuntos de datos que contienen valores atípicos distribuidos en diferentes densidades [34].

Por otro lado, se ha planteado un método que combina características de estimaciones de densidad de Kernel con las nociones de K vecinos más cercanos (KNNs), vecinos compartidos más cercanos (SNNs por sus siglas en inglés) y vecinos invertidos (RNNs) [35]. Este método fue usado para detectar atípicos en cinco bases de datos médicas: sobre la hepatitis, cardiotocográficos (relativos a la frecuencia cardiaca y actividad fetal), arrítmicos, sobre enfermedades del corazón y sobre hipotiroidismo. Estas bases de datos contenían información acerca de pacientes que padecían dichas complicaciones: edad, género, peso, estatura, además de variables relevantes según cada patología: resultados de exámenes médicos, historial de tabaquismo, presión arterial y presencia de dolor, entre otras. Los autores del método destacaron la estabilidad y robustez de combinar distintos algoritmos para detectar atípicos frente a las variaciones en el parámetro *k* que representa la cantidad de vecinos más cercanos considerados.

Métodos basados en distancias

El enfoque basado en distancia se introdujo como una solución para eliminar una de las desventajas que tienen los métodos estadísticos: el requerimiento de información histórica sobre los parámetros del conjunto de datos, como por ejemplo la distribución de probabilidad [36].

El método *Nearest Neighbor Distance Based Outlier Detection Technique* (NDot) se basa en la noción de vecino más cercano. Este método introduce el parámetro *factor de vecino más cercano* (NNF por sus siglas en inglés) para medir el grado de atipicidad de un punto con respecto a cada uno de sus vecinos; este factor es la razón entre la distancia del punto y el vecino, y la distancia promedio a los *k* vecinos más cercanos. NDot calcula el NNF de cada punto con respecto a todos sus *k* vecinos más cercanos. Si el NNF del punto con respecto a la mayoría de sus vecinos supera un umbral predefinido, entonces el punto se declara un valor atípico potencial [37].

Por otro lado, el algoritmo *Factor de valores atípicos basado en la distancia local* (LDOF por sus siglas en inglés), es un método en el cual, para cada dato, se calcula cuánto se desvía el factor LDOF de su vecindario. Este factor es la relación entre la distancia media de los k vecinos más cercanos y la distancia interior de los k vecinos más cercanos. La distancia interior es la distancia promedio por pares del conjunto de k vecinos más cercanos. Es usual emplear el umbral igual a 1 para discernir la naturaleza de un punto: los datos cuyo factor LDOF es mayor a 1 se consideran valores atípicos, de lo contrario se consideran datos regulares. Otra implementación de este método para detectar valores atípicos es la técnica "Top-n", que clasifica como atípicos los n objetos con los LDOF más altos de la base [38].

Métodos basados en clústeres

El método *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) agrupa los datos obtenidos de una muestra y los categoriza de la siguiente manera: los datos regulares formarán grupos grandes y densos, mientras que los datos atípicos formarán grupos más pequeños y alejados, o simplemente no pertenecerán a ninguno de los otros grupos. El objetivo del método es reconocer las regiones con mayor densidad por medio de la cantidad de objetos alrededor de un punto [39].

Este enfoque ha sido utilizado en la fase de detección de atípicos dentro del modelo híbrido de diagnóstico temprano de diabetes e hipertensión compuesto por los métodos DBSCAN, SMOTE (balanceo) y Random Forest (clasificación). La técnica de detección se aplicó en tres bases de datos diferentes y se identificaron, en promedio, 24 datos atípicos en cada conjunto de datos. Las bases eran de carácter multidimensional y contenían datos como la edad del paciente, algunas mediciones antropométricas como la circunferencia de la cintura y de la cadera, y variables cardiovasculares. Este estudio utilizó esta metodología para detección de atípicos debido a que en la literatura revisada por los autores se encontró que su uso mejoraba en un 10% la exactitud de otros clasificadores [39].

El método *Optimized Deep Clustering* (ODC) detecta valores atípicos con base en una versión modificada del método clásico de clustering, *kmeans*. Luego de ejecutar *kmeans*, se calcula para cada centroide tanto su distancia absoluta como su distancia promedio a todos los datos de la base. Aquellos registros cuya distancia al centroide más cercano sea mayor a la distancia promedio entre los centroides y cada dato, se consideran atípicos y se remueven de la base. Este enfoque mejora el ejercicio de clustering [40].

La técnica ODC se utilizó en un estudio realizado por varias universidades australianas para detectar valores atípicos asociados al comportamiento inadecuado del tráfico de red. El estudio comparó la exactitud para detectar atípicos de seis métodos, obteniendo el método ODC el mayor valor (97.5%) [40].

Métodos basados en densidad

Algunos métodos basados en densidad utilizan el enfoque *Local Outlier Factor* (LOF), que se refiere a la razón entre la densidad local de un objeto, que es igual al recíproco de la distancia promedio entre el objeto y sus k vecinos más cercanos, y el promedio de las densidades de sus vecinos más cercanos. Un punto que tiene un LOF alto es considerado atípico mientras que un LOF bajo se clasifica como dato regular. El método *Influenced Outliers* (INFLO) es una versión mejorada de LOF, en donde la detección de valores atípicos se basa en la relación simétrica con sus vecinos. Este método usa los vecinos más cercanos de un punto en el conjunto de datos, y los puntos del conjunto de datos que tienen el mismo vecino [41]. Un objeto con un INFLO alto es considerado atípico [36].

Un estudio comparó el desempeño de LOF e INFLO sobre el mismo set de datos que contiene en total 1700 registros, de los cuales siete son atípicos y los demás forman cuatro clústeres de tamaños 200 y 500 aproximadamente. En los resultados se encontró que LOF identifica correctamente los siete valores atípicos que tiene la base de datos, mientras que INFLO solo detecta cuatro de estos valores; los tres restantes los agrupa en un clúster pequeño debido a su ubicación en el espacio, por lo cual no los reconoce como atípicos [42].

Métodos basados en aislamiento

El método *Isolation Forest* (iForest) se basa en la construcción de un conjunto de árboles aleatorios de aislamiento para un conjunto de datos. El término aislamiento significa realizar particiones recursivas y aleatorias sobre el conjunto de datos hasta que cada punto quede separado de los demás. Gráficamente, se puede representar al iForest como un árbol de decisión, en donde cada nodo representa una partición y desde el cual nacen dos nodos que simbolizan los datos que quedan a un lado u otro de dicha partición. De esta manera se producen caminos más cortos para los datos atípicos ya que, al alejarse de la distribución regular de los datos, son más susceptibles a ser aislados en pocas particiones [27].

La detección de valores atípicos mediante este método es un proceso de dos etapas. En la primera etapa, conocida como entrenamiento, se construyen árboles de aislamiento utilizando submuestras del conjunto de datos. Para la segunda etapa se pasan los datos a través de los árboles construidos para así obtener la longitud del camino de cada dato, es decir, la cantidad de particiones requeridas para aislarlo. Con los diferentes árboles generados, se promedia la longitud del camino de cada dato y los que tengan una longitud menor a un percentil determinado serán considerados como valores atípicos. [27]

Se realizó un experimento con cuatro conjuntos de datos provenientes del repositorio de aprendizaje automático de la UCI (University of California, Irvine), los cuales superan los 10000 registros. Los conjuntos de datos provenían de distintas fuentes: registros simulados de intrusión a redes militares, ubicación de árboles en zonas boscosas, y ubicaciones de aterrizaje espacial. En cada uno de ellos se seleccionaron las variables continuas para evaluar el rendimiento del algoritmo de detección de atípicos basado en iForest, siendo el tiempo de recolección una de las variables de interés. Para la etapa de entrenamiento, cada conjunto de datos se dividió en ventanas móviles, las cuales representaban el mismo intervalo de tiempo y la misma cantidad de datos. La funcionalidad del método radica en que divide los datos en intervalos donde se espera un comportamiento similar, facilitando la identificación de atípicos. Se probó el algoritmo con diferentes tamaños de ventana para seleccionar la más adecuada evaluando el área bajo la curva (AUC), donde el tamaño con mejor desempeño es distinto para cada conjunto de datos. Esto implica que no existe una aproximación teórica al cálculo del tamaño óptimo de la ventana de tiempo.

En la etapa de prueba, el algoritmo examinó cada punto en la ventana para determinar si era atípico o no en función de la puntuación de anomalía. En los resultados se encontró que el desempeño del iForest para la detección de valores atípicos fue alto, con un AUC entre 0.86 y 0.98 para los cuatro conjuntos de datos [43].

Métricas de desempeño para la comparación de metodologías de detección de atípicos

Para hacer una comparación de metodologías de detección de atípicos, es necesario seleccionar las métricas con las que se va a evaluar el desempeño de los algoritmos. A continuación, se presentan las tres principales métricas para este fin que se hallaron en la revisión de literatura [32]: *Sensibilidad*, *F-score y* el área bajo la curva ROC (*AUC).*

Las métricas mencionadas tienen un enfoque supervisado en el cual se deben tener rotulados los datos como atípicos/no atípicos y, una vez ejecutado el algoritmo de detección, se procede a construir la matriz de confusión con base en los siguientes recuentos:

- Cantidad de atípicos bien detectados (verdaderos positivos, *TP* por sus siglas en inglés),
- Cantidad de datos regulares identificados como atípicos por el algoritmo (falsos positivos, *FP*),
- Cantidad de datos regulares etiquetados como tal (verdaderos negativos, *TF*), y
- Cantidad de atípicos clasificados como datos regulares (falsos negativos, *FN*).

Recall (R) o Sensibilidad

La sensibilidad de un método de detección de valores atípicos se define como el cociente entre la cantidad de verdaderos valores atípicos detectados (TP) y la cantidad de valores atípicos presentes en

los datos. Este valor cuantifica el poder de detección que tiene el algoritmo. En fórmulas, la sensibilidad se calcula como en la Ecuación 1.

$$R = \frac{TP}{TP + FN}$$

Ecuación 1. Sensibilidad tomada de "Clinical Tests: Sensitivity and specificity" Continuing Education in Anesthesia, Critical Care & Pain | Volume 8 Number 6 2008 p. 221[44]

La métrica de sensibilidad fue implementada en el estudio realizado por Kornel Chrominski y Magdalena Tkacz en el 2010 en el cual se compararon distintos métodos para detectar valores atípicos en datos biomédicos. Los datos empleados corresponden a conteos de glóbulos rojos sanguíneos de 80 personas. Estos datos fueron perturbados mediante el reemplazo del 10% de los registros por valores aleatorios dentro de un intervalo determinado, para así obtener dos bases con registros artificiales que, a su vez, resultaban ser los valores atípicos por detectar [45].

En dicho experimento se aplicaron cuatro métodos de detección: método del IQR y las pruebas estadísticas de Grubbs, Hampel y Dixon con distintos niveles de significancia. En resumen, la sensibilidad para los distintos métodos osciló entre 0.25 y 1, siendo el test de Hampel el que logró identificar todos los atípicos de la base. Adicionalmente, para cada método se midieron los tiempos de ejecución y se cuantificó la complejidad de implementación [45].

F beta

El criterio de evaluación *F* beta, denotado F_β, usa la relación existente entre Precisión (P) y Sensibilidad. La precisión se define como la razón entre los verdaderos positivos (TP) y la cantidad de datos que el algoritmo identifica como atípicos (TP+FP). El F_β es una métrica para evaluar algoritmos binarios de detección de atípicos [46] y para su cálculo se usa la Ecuación 2.

Usar la precisión y sensibilidad como coeficientes para la métrica F_β es importante porque como métricas individuales no alcanzan a explicar completamente el desempeño de un método [47]. Al tenerlas en cuenta la métrica evaluaría tanto la calidad de la detección, como la cantidad de valores detectados; dos atributos cruciales en una clasificación binaria.

$$F_\beta = \frac{(1 + \beta^2) \cdot Precisión \cdot Sensibilidad}{(\beta^2 \cdot Precisión) + Sensibilidad}$$

Ecuación 2. F. Tomada de " A Surrogate Loss Function for Optimization of Fβ Score in Binary Classification with Imbalanced Data " (Lee et al.) [46]

Esta métrica toma valores en el intervalo (0, 1], y denota éxito en la clasificación cuando se acerca a 1. El valor de β ajusta el peso relativo que se desea dar entre la precisión y la sensibilidad. Si el valor de β aumenta, la métrica se orienta hacia la sensibilidad, mientras que, si se reduce, hacia la precisión. Por tanto, si el valor de β aumenta, el algoritmo en cuestión recibe una penalización mayor cuando incurre en falsos negativos, mientras que, si disminuye, la recibe por la presencia de falsos positivos en la predicción. Además, es menester aclarar que la definición de un valor de β no es un problema trivial y debe ser objeto de discusión por parte de los investigadores. En diagnósticos médicos, es necesario hacer un balance entre los "costos" de los dos tipos de clasificación errónea: clasificar a un paciente enfermo como sano y clasificar a un paciente sano como enfermo. Incluso en entornos financieros donde, aparentemente, el costo de aumentar la precisión o la sensibilidad es fácil de definir, esta no es una tarea sencilla [47]. Por tanto, es común en la literatura utilizar $\beta = 2$ o $\beta =$

0.5 para ejemplificar situaciones en donde se preferiría dar más influencia a la precisión o a la sensibilidad, respectivamente, sobre el valor final del F_β. Esto se puede evidenciar en el trabajo de Nancy Chinchor [48].

En la práctica, $\beta = 1$ se ha convertido en el valor más común para el parámetro β [49]. Esto se debe a que este valor de β iguala el F_β a la media armónica entre precisión y sensibilidad, otorgándole la misma importancia a ambas mediciones. En este caso, la métrica se conoce simplemente como *F1*, *F-score*, o *F1-score* y se calcula mediante la expresión dada en la Ecuación 3.

$$F = \frac{2 \times P \times R}{P + R}$$

Ecuación 3. Tomada de F: Una medida interpretable de la medida F Machine Learning 110(3)[50]3

Un ejemplo de comparación de métodos de detección de atípicos usando F1 como métrica de desempeño se realizó en China en el 2019, donde se aplicaron cuatro algoritmos basados en el vecino más cercano: Local Outlier Factor (LOF), Influenced Outlierness (INFLO), Relative Density-based Outlier Score (RDOS), Local Distance-based Outlier detection factor (LDOF), y uno creado por los autores y denominado Outlier Detection Algorithm based on Density and Distance double Parameters Outlier factor Scores (ODA-DDPOS), a 13 conjuntos de datos pertenecientes al repositorio de *machine learning* de la Universidad de California, Irvine. Para determinar el algoritmo más efectivo para cada conjunto de datos, se analizó el comportamiento del F1 al variar el número de vecinos, *k*, de 1 a 100 [51] y se concluyó que el algoritmo ODA-DDPOS tuvo el mejor desempeño en la detección de atípicos, con un F1 máximo de 0.8 con 13 vecinos, seguido, en orden, por LOF, RDOS, INFLO y LDOF.

Ahora, utilizar la precisión y la sensibilidad, implica no solo reconocer la cantidad de errores que comete el método de detección, sino también el tipo de errores que comete. Eliminar un dato normal puede generar un cambio mucho menor en el cálculo del VO_2 máximo real que aquel que puede generar el no eliminar un dato atípico. Es por esto que la sensibilidad debe contar con una influencia mayor en la escogencia del mejor método.

En el caso particular de la prueba CPET, el valor escogido para β es 2. Los datos recogidos en la prueba CPET son utilizados para calcular distintos indicadores sobre el estado cardiopulmonar del individuo. Uno de los indicadores más importante es el VO_2 máximo, el cual se halla como el máximo de los promedios móviles en ventanas de 20 segundos dentro de un rango de 2 minutos antes de eliminar la carga de la prueba. Al tratarse de un promedio, el VO_2 máximo es especialmente susceptible a la presencia de valores atípicos; un solo dato atípico puede cambiar el valor del VO_2 máximo calculado, cambiando el diagnóstico que pueda emitir el especialista que interprete la prueba. Por esta razón, y teniendo en cuenta que una CPET produce gran cantidad de registros, se considera que la sensibilidad debe tener una influencia mayor a la de la precisión en la métrica de desempeño; es decir, por la naturaleza de nuestros datos, resulta aceptable la pérdida de algunos datos normales a cambio de un aumento en la detección de datos atípicos.

Área bajo la curva ROC (*AUC por sus siglas en inglés)*

Para poder definir la métrica AUC es necesario introducir el concepto de *Receiver Operating Characteristic (ROC).* La curva ROC representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos detectados por el algoritmo cuando el punto de corte para las probabilidades de que un punto sea atípico varía. El área debajo de la curva ROC, AUC, es una métrica independiente del punto de corte que cuantifica el desempeño global del método de detección. El AUC toma valores

entre 0 y 1, siendo 0 un bajo desempeño y 1 un desempeño perfecto. Un ejemplo de se encuentra en el estudio realizado por Xu, Li, Liu y Yao, quienes aplicaron diez métodos de detección de atípicos sobre nueve sets de datos referentes a condiciones médicas. El método con mejor desempeño fue kNN con un AUC promedio de 68.67% [52].

Operaciones con matrices de confusión

La matriz de confusión es una herramienta que resume los conteos de la clasificación de las predicciones y los valores verdaderos de un modelo o algoritmo. Varias métricas de desempeño como exactitud, precisión y sensibilidad, usadas para calificar el desempeño de los modelos, se derivan de la matriz de confusión, y los algoritmos o modelos de clasificación generalmente se optimizan para mejorar el rendimiento de estas métricas [53].

Para describir el rendimiento general de un modelo cuando se tienen distintos conjuntos de datos o varias iteraciones, es preferible tener una sola matriz de confusión o un único valor de la métrica en lugar de múltiples valores o matrices [54]. Capodeci (2020) propone un enfoque donde se evalúan tres métodos de clasificación, para definir el plan de mantenimiento del motor de los aviones, clasificando las 817 piezas de repuesto en "eficiente", "reparado" y "reemplazado". En este enfoque se suman las matrices de confusión de las piezas por modelo, con el fin de obtener una matriz global, de la cual se calculan las métricas precisión, sensibilidad y f-score. De acuerdo con las métricas calculadas se encontró que el método Naive Bayes obtuvo el mejor desempeño.

Otro ejemplo de operaciones con matrices de confusión para evaluar una matriz global lo encontramos en el estudio realizado en el 2013 por Walter Dirk, de la Ohio State University, en el que se midió la capacidad de respuesta del cerebro ante estímulos visuales y se estableció la información mutua entre las respuestas del experimento con datos previamente validados del verdadero comportamiento del cerebro ante los estímulos visuales [55].

En dicho experimento participaron 18 personas, que debieron observar 360 imágenes de las cuales la mitad eran a color y las otras se componían de dibujos lineales. Los participantes debían clasificar las imágenes en una de las seis categorías presentadas y las respuestas fueron registradas en matrices de confusión: una por cada individuo, y separadas entre las clasificaciones en el formato a color y las del formato de dibujos lineales. Las matrices obtenidas se promediaron entre los participantes y se encontró que la precisión de la clasificación fue mayor para las imágenes a color que para los dibujos lineales, con un 77.3% y 66.2% respectivamente.

Aplicativos y software que soportan el proceso diagnóstico en el sector de la salud

Los aplicativos que soportan los procesos diagnósticos en el sector de la salud se han convertido en una herramienta para mejorar los servicios y la experiencia presentada al paciente, así como para facilitar el trabajo del personal médico, representando la unión entre innovación y tecnología. Para este propósito en el 2010, el gobierno de Estados Unidos creó el proyecto de aplicaciones médicas sustituibles y de tecnología reutilizable (SMART Health IT por sus siglas en inglés) con el fin de facilitar el desarrollo de aplicaciones médicas de todo tipo, desde el diagnóstico de enfermedades específicas, información sobre medicamentos y recordatorios, maternidad y embarazo, hasta dieta, nutrición y salud mental [56].

SMART Health IT es una plataforma tecnológica abierta que permite a los innovadores crear aplicaciones que se ejecuten sin problemas y de forma segura en todo el sistema de atención médica (SMART, s.f.); esta plataforma cuenta con bibliotecas disponibles para HTML5, iOS y Python, además tiene una biblioteca con enlaces a las tiendas en línea donde se pueden comprar las aplicaciones móviles [56].

Por medio de esta plataforma en 2016 la Universidad de Vanderbilt creó SMART Precision Cancer Medicine (PCM), una aplicación que busca presentar en tiempo real información de salud genómica a nivel poblacional a oncólogos y sus pacientes, como un componente de la práctica clínica para soportar procesos diagnósticos [57]. El aplicativo se diseñó para el sistema operativo iOS, específicamente para iPad y iPad Mini. PCM incluye la información demográfica del paciente (nombre, género, edad, número de historia clínica), el diagnóstico de cáncer primario y los resultados del perfil molecular. La aplicación contiene gráficos interactivos del paralelo entre las mutaciones genéticas del paciente con respecto a otros pacientes que han sido evaluados en el centro médico de la Universidad de Vanderbilt, así como enlaces externos para la consulta de terminología específica en páginas confiables.

Aplicativos para interpretación de CPET

Otros autores ya han propuesto soluciones para el complejo análisis de los datos de CPET. Robert Ross y David Corey (2007), del Departamento de Medicina de Baylor College, crearon un software para la interpretación de CPET llamado XINT, el cual usa un enfoque integrador para leer la prueba en conformidad con las recomendaciones de la ATS/ACCP [3]. XINT crea un informe interpretativo de la prueba utilizando información demográfica del paciente y el valor de las variables requeridas: consumo máximo de oxígeno (VO_2 max), umbral anaeróbico (AT), capacidad vital forzada (FVC), ventilación máxima alcanzada (VEmax), entre otras. El reporte depende de la cantidad de variables ingresadas.

Del mismo modo, Oxynet es una aplicación web que busca determinar de manera automática y mediante el entrenamiento de una red neuronal, los umbrales ventilatorios 1 y 2 (VT1 y VT2) obtenidos por CPET y muy útiles para evaluar la capacidad funcional de una persona [58]. El VT1 y VT2 se determinan gráficamente y corresponden a instantes durante la prueba donde se presentan cambios significativos en el consumo de oxígeno en relación con otras variables como la ventilación. El proyecto se basa en la colaboración de diversos laboratorios y expertos alrededor del mundo para recolectar y analizar conjuntos de datos resultantes de CPETs. La evaluación de la precisión en la detección de los umbrales se hace a través de la opinión de expertos. El proceso es realmente sencillo: el usuario carga a la aplicación web un archivo en formato Excel exportado del equipo utilizado para la CPET y la aplicación despliega tanto gráficas relevantes como los valores de los umbrales.

3. Objetivos

Objetivo general: *Identificar el método más adecuado para detectar valores atípicos en el conjunto de datos de la variable VO_2 que arroja la prueba cardiopulmonar de ejercicio (CPET).*

Objetivos específicos:

- Construir la vista minable de los datos de la variable VO_2.
- Aplicar los métodos de detección de valores atípicos en los datos de la variable VO_2.
- Medir el impacto de las metodologías de detección en términos de la métrica de comparación seleccionada.
- Construir un aplicativo para la representación gráfica de la variable VO_2 antes y después de tratar sus valores atípicos.
- Evaluar el desempeño del aplicativo construido.

4. Metodología y resultados

A continuación, se presenta la metodología y los resultados de cada uno de los 5 objetivos específicos diseñados para alcanzar el objetivo general del trabajo de grado.

Objetivo específico 1

Procedencia de los datos

El trabajo se desarrolla sobre los datos de 407 pruebas cardiopulmonares de ejercicio, CPET, del proyecto *"Ejercicio Físico en Sujetos Sanos no Entrenados Nativos a Altitud Moderada en Comparación con Crónicamente Aclimatados"* financiado por Minciencias (id:120356934972), tomadas entre el 2016 y el 2018, de individuos aparentemente sanos con edades entre los 18 y los 25 años. Cada prueba consta de 68 variables y un promedio de 781.55 filas que corresponden a la medición de cada uno de los 68 atributos en cada respiración del paciente. Dichas observaciones están divididas en las cuatro fases de la prueba CPET: reposo, calentamiento, ejercicio de intensidad progresiva y recuperación.

Preliminarmente se excluyeron los registros de 7 participantes pues sus pruebas no satisfacen uno o más criterios de una CPET exitosa: tener una relación de intercambio respiratorio (RER por sus siglas en inglés) mayor a 1.1, una frecuencia cardíaca esperada superior al 85% y finalmente un lactato en sangre durante el reposo superior a 8 mmol/l [59].

Con respecto a las variables de interés, en cada base se seleccionaron las columnas de tiempo y VO_2. Así, en este trabajo se emplearon 400 bases de datos que fueron almacenadas en una carpeta en Google Drive, la cual se vinculó con Google Colab, plataforma en donde se realizó todo el proyecto usando Python como lenguaje de programación.

Al revisar las 400 gráficas de VO_2 vs tiempo, se observó que el patrón respiratorio de todos los sujetos tiene una estructura similar. Este patrón corresponde al comportamiento esperado en cada una de las etapas de la prueba. Inicialmente durante la etapa de reposo se presenta un valor del VO_2 constante; luego al entrar en la etapa del calentamiento este valor de VO_2 aumenta y nuevamente se mantiene constante. Durante la etapa de ejercicio, el VO_2 aumenta de forma lineal cuando la tasa de trabajo del cicloergómetro se incrementa en un patrón de rampa continuo [60]. Finalmente, la etapa de recuperación debería presentar un comportamiento exponencial decreciente si se ha logrado llegar a un esfuerzo máximo durante la prueba [61].

Exploración descriptiva

El análisis descriptivo y de la calidad de los datos se hizo a través de estadígrafos de tendencia central, de dispersión y de análisis de frecuencia de valores faltantes. El valor promedio del VO_2 en las pruebas es de 1228.19 ml/min y tiene una desviación estándar de 295.82 ml/min. Con la biblioteca Pandas se determinó la cantidad de NA's (valores vacíos) por prueba, que son espacios en blanco o con contenido no numérico en la columna de VO_2. El 49.5% de las pruebas contienen campos vacíos (NA) en la variable VO_2, y en promedio hay 1.75 NAs por prueba. Aquella con mayor cantidad de valores vacíos es la del individuo 0422 con 85.

Etiquetado de valores atípicos

La identificación de valores atípicos en la prueba se hizo de forma manual observando el gráfico de dispersión VO_2 vs tiempo de cada prueba. Siguiendo las indicaciones de un profesional de la salud experto en CPET, el criterio de identificación consistió en marcar como atípicos los puntos que se

encontraban distantes de sus vecinos en relación con el comportamiento de la gráfica. Una vez finalizado el etiquetado se trazó el diagrama de dispersión para cada prueba con los valores atípicos marcados en color rojo, como se muestra en las gráficas de la Tabla 4. Los diagramas de dispersión con el etiquetado inicial fueron revisados por dos expertos en pruebas CPET, que agregaron y/o eliminaron valores marcados como atípicos.

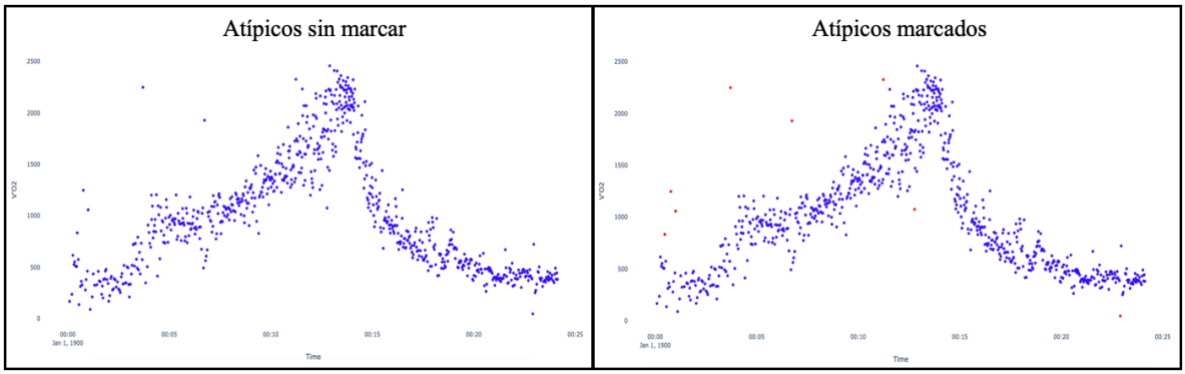

Tabla 4. Prueba 0024 sin etiquetado de valores atípicos (izquierda) y con etiquetado de atípicos (derecha)

Posterior al etiquetado definitivo de los valores atípicos en cada una de las pruebas válidas (400), fue posible identificar que todas tienen presencia de valores atípicos. La prueba con menor cantidad de atípicos contiene 2 puntos identificados de esta forma, y aquella con mayor cantidad de atípicos tiene 186. En promedio las pruebas tienen 10.82 valores atípicos y un porcentaje de contaminación, es decir la proporción de atípicos con respecto al total de registros, promedio de 1.46%.

Para cada prueba se agregó una columna binaria con el etiquetado de cada punto. Si un punto era identificado como atípico la etiqueta tomaba el valor de 1, de lo contrario, el valor de 0. Esta es una de nuestras variables de interés puesto que las predicciones de los métodos se contrastarán con el etiquetado manual de los datos. De esta forma será posible medir el desempeño de cada método en la identificación de valores atípicos.

Objetivo Específico 2

Se seleccionaron 6 métodos para la detección de valores atípicos. No se incluyeron métodos paramétricos pues la variable VO_2 no cumplió con el supuesto de normalidad para ningún individuo. Esto se comprobó mediante la prueba Shapiro-Wilk de normalidad sobre la variable.

Para implementar los métodos y, posteriormente, evaluar su desempeño en la identificación de atípicos, las 400 bases de datos se dividieron aleatoriamente en dos grupos: *train* con 280 archivos, y *test* con 120 archivos.

Selección de parámetros
Dependiendo del método, este puede requerir uno o más parámetros para ser implementado en Python. Los métodos LOF y KNN tienen dos parámetros: contaminación (porcentaje de valores atípicos presentes en los datos) y número de vecinos más cercanos (k); los algoritmos DBSCAN y LDOF usan únicamente el número de vecinos más cercanos (k); mientras que Isolation Forest depende del parámetro de contaminación y finalmente la Regresión Kernel tiene como parámetro el ancho de banda (B).

En cada caso, la estimación de parámetros se hizo sobre los individuos de train probando distintos valores para cada parámetro. Para los métodos que requieren dos parámetros, se probaron las

combinaciones posibles dentro del rango de valores establecidos. Por ejemplo, para el método kNN se probaron 16 valores distintos para el parámetro k y cuatro para el parámetro de contaminación, resultando en 64 combinaciones distintas.

En la Tabla 5 se resume el nombre de los métodos usados, parámetros y los valores que se probaron para cada parámetro. Los valores se seleccionaron mediante una evaluación de las métricas de desempeño (precisión, sensibilidad, F-score y F_β) calculadas sobre los datos de train, el objetivo era maximizarlas. Por ejemplo, en el caso del LOF se aplicó el método con valores de k desde 2 a 30 vecinos, y se calcularon las cuatro métricas mencionadas anteriormente, se encontró que con los valores de k más cercanos a 2 y con los más cercanos a 30 el valor de las métricas disminuye, llegando así a un rango de 10 a 25 en el cual se alcanza el máximo desempeño.

Método	**Parámetro**	**Valores**
LOF	Contaminación	0.01, 0.0146, 0.02, 0.03
	Número de vecinos (k)	[10,26)
DBSCAN	Número de vecinos (k)	[2,42)
KNN	Contaminación	0.01, 0.0146, 0.02, 0.03
	Número de vecinos (k)	[1,41)
Isolation Forest	Contaminación	0.01, 0.0146, 0.02, 0.03
LDOF	Número de vecinos (k)	[10,42) en saltos de 2
Regresión Kernel	Ancho de banda (B)	[5,85) en saltos de 5

Tabla 5, métodos, parámetros y sus valores para train.

Para cada valor de parámetro, se aplicaron los distintos métodos de detección sobre los datos de train, se construyeron las respectivas matrices de confusión y se calcularon las métricas precisión, sensibilidad, f-score y f_β con un valor de β igual a 2, los resultados se encuentran presentados en tablas por método en el anexo 2.

Finalmente, la métrica seleccionada para escoger el valor de los parámetros de cada método es F_β. Los valores de los parámetros escogidos para cada método se presentan en la Tabla 6.

Método	**Contaminación**	**K**	**B**
LOF	0.02	18	-
DBSCAN	-	8	-
KNN	0.02	3	-
Isolation Forest	0.03	-	-
LDOF	-	30	-
Regresión Kernel	-	-	20

Tabla 6. Parámetros de los métodos seleccionados.

Implementación de métodos

Para la aplicación de Isolation Forest sobre los individuos tanto de train como de test se utilizaron ventanas móviles de 20 datos. Esto se debe a que en intervalos entre 15 y 20 respiraciones se espera una ventana con una baja variabilidad en la variable de consumo de oxígen; por el comportamiento de la prueba y debido al aumento de la carga y a la fatiga del paciente, el consumo de oxígeno aumente progresivamente [3]. Es de esperar que los valores de la variable VO_2 varíen en gran medida en ventanas de tiempo extensas. Posteriormente, se realizaron particiones aleatorias sobre las ventanas de datos teniendo en cuenta únicamente sus respectivos valores de VO_2; aquellos puntos que quedaron aislados después de pocas particiones aleatorias fueron considerados atípicos, siguiendo el procedimiento indicado en la explicación del método.

En cuanto a la Regresión Kernel, su aplicación sobre los datos implicó la implementación del método de BoxPlot sobre los residuales. Después de variar el valor del ancho de banda (B) dentro del rango indicado en la Tabla 5, se calcularon los residuales entre el valor del punto y el valor de la regresión para dicho punto. Posteriormente, se construyó el BoxPlot de estos residuales y los datos no incluidos dentro del rango [Q1 - IQR*3, Q3+IQR*3] fueron considerados atípicos. Inicialmente se utilizó 1.5 como factor multiplicativo del IQR, pero, al observar que las métricas indicaban un bajo desempeño, se optó por un factor multiplicativo de 3. Este ajuste aumentó la precisión del método al disminuir la cantidad de puntos regulares etiquetados como atípicos. El acercamiento explicado es una propuesta metodológica por parte del grupo con el fin de adaptar el método estadístico enfocado a la detección de valores atípicos.

Una vez estimados los parámetros, se aplicaron los 6 algoritmos sobre los 120 individuos del conjunto de test y se obtuvo una clasificación de dato atípico/regular para cada uno de los registros de la prueba. Con base en esta clasificación y, empleando como variable de referencia la de etiquetado manual de atípicos creada en el objetivo 1, se construyó una matriz de confusión agregada para cada método; es decir, se sumaron casilla a casilla las 120 matrices de confusión individuales. Los resultados de las matrices de confusión agregadas se presentan en las Tablas 7, 8, 9, 10, 11, y 12.

LOF	Positivos Reales	Negativos Reales
Positivos Predichos	966	920
Negativos Predichos	376	89011

Tabla 7. Matriz de confusión agregada para el método LOF sobre los individuos en test

DBSCAN	Positivos Reales	Negativos Reales
Positivos Predichos	978	737
Negativos Predichos	364	89194

Tabla 8. Matriz de confusión agregada para el método DBSCAN sobre los individuos en test

KNN	Positivos Reales	Negativos Reales
Positivos Predichos	1044	839
Negativos Predichos	298	89092

Tabla 9. Matriz de confusión agregada para el método kNN sobre los individuos en test

Isolation Forest	Positivos Reales	Negativos Reales
Positivos Predichos	1262	1533
Negativos Predichos	80	88398

Tabla 10. Matriz de confusión agregada para el método Isolation Forest sobre los individuos en test

LDOF	Positivos Reales	Negativos Reales
Positivos Predichos	843	897
Negativos Predichos	499	89034

Tabla 11. Matriz de confusión agregada para el método LDOF sobre los individuos en test

Regresión Kernel	Positivos Reales	Negativos Reales
Positivos Predichos	1094	707
Negativos Predichos	248	89224

Tabla 12. Matriz de confusión agregada para el método Regresión Kernel sobre los individuos en test

Objetivo Específico 3

De acuerdo con las matrices de confusión agregadas para cada algoritmo sobre los datos de test se calcularon las métricas: sensibilidad, precisión, F-score y F_β con β igual a 2, los resultados se muestran en la Tabla 13.

	LOF	**DBSCAN**	**KNN**	**Isolation Forest**	**LDOF**	**Regresión Kernel**
Precisión	0,5122	0,5703	0,5544	0,4522	0,4845	0,6074
Sensibilidad	0,7198	0,7288	0,7779	0,9419	0,6282	0,8152
F-score	0,5985	0,6398	0,6474	0,6111	0,5470	0,6962
F_β	0,6658	0,6904	0,7199	0,7742	0,5930	07630

Tabla 13. Métricas de desempeño métodos en test

El método Isolation Forest obtuvo el valor de F_β más alto, 0.7742, situándose como el método de detección de valores atípicos seleccionado para preprocesar la prueba CPET. Este método también registró el mayor valor de sensibilidad y el menor de precisión, 0.9419 y 0.4522 respectivamente, lo que significa que detecta de manera adecuada los valores atípicos, pero en su detección marca valores normales como atípicos también. El valor de F_β obtenido por la Regresión Kernel fue igual a 0.7630 con el que es considerado el segundo método más adecuado para la prueba CPET, cabe resaltar que este método presenta un mejor balance entre sensibilidad y precisión, ambos con puntajes altos. Finalmente, el método KNN obtuvo un valor de F_β de 0.7199, situándose en el tercer lugar.

VO_2 Máximo

Para medir el impacto de implementar el método Isolation Forest en la detección de atípicos de la CPET se calculó el VO_2 máximo para las 120 pruebas de test antes y después de eliminar los atípicos identificados por el método. Para obtener el indicador se debe promediar la variable VO_2 en todas las ventanas móviles de 20 segundos que estén dentro de los dos últimos minutos en los que se aplica carga a la prueba; el mayor de dichos promedios corresponde al VO_2 máximo. Es un indicador importante dentro de los que se pueden calcular sobre los datos de una CPET, ayuda al personal médico a tomar decisiones sobre el tratamiento quirúrgico y no quirúrgico más apropiado, estimar la

probabilidad de morbilidad y mortalidad postquirúrgica, clasificar a los pacientes para la atención postoperatoria requerida, identificar nuevas comorbilidades, predice el resultado postoperatorio en pacientes quirúrgicos, demostrando que tiene una alta utilidad clínica predictiva [9].

Como análisis preliminar del cambio en el VO_2 máximo que tiene la detección de atípicos se hizo una exploración descriptiva de los valores calculados. el promedio del VO_2 máximo con valores atípicos es 2499.55 ml/min, con una desviación estándar de 625.55, mientras que el promedio sin atípicos es 2473.73 ml/min con una desviación estándar de 627.27. En la Tabla 14 se presentan los histogramas del VO_2 máximo antes de eliminar atípicos, después de eliminarlos y de la diferencia entre VO_2 máximo, respectivamente.

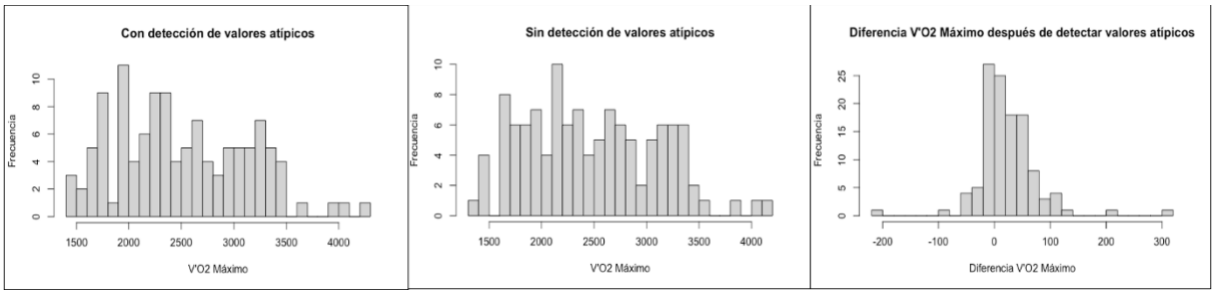

Tabla 14. Histogramas del VO2 máximo con valores atípicos, sin valores atípicos y diferencias.

Se construyó un boxplot pareado para visualizar la diferencia del VO_2 máximo con y si valores atípicos. Este gráfico se puede observar en la imagen 1, con líneas que conectan los datos en las dos instancias. Se pueden evidenciar diferencias en el valor de la medida siguiendo la trayectoria de la línea en los puntos, donde la gran mayoría no son rectas y tienen cierta inclinación, representando el cambio. El 67.52% de los valores de VO_2 máximo disminuyó al aplicar la detección y eliminar los valores atípicos, el 14.53% aumentó, y el 17.95% no presentó ningún cambio.

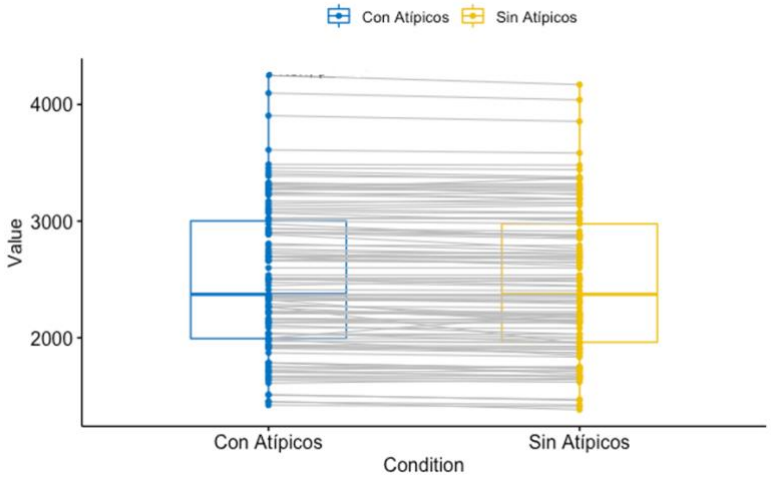

Imagen 1. Boxplot pareado del VO_2 máximo

Adicionalmente, se corrieron pruebas de normalidad Shapiro-Wilk tanto para los valores del VO_2 máximo antes y después como para su diferencia, donde se concluyó que ninguna de las tres variables cumple el supuesto de normalidad, pues sus p-values (todos menores a la significancia de 5%) fueron, respectivamente, 0.011, 0.012 y 0.012. Así, no fue posible comparar el VO_2 máximo antes y después de la detección mediante pruebas paramétricas como la t pareada. Con lo anterior se procedió a implementar el test de Wilcoxon pareado, una prueba no paramétrica que compara la mediana de las diferencias de dos muestras dependientes [62]; es decir, es apropiada para evaluar a la misma

población bajo dos condiciones diferentes, y se usó para establecer si existe diferencia significativa en el valor del VO_2 antes y después de la detección de atípicos. La hipótesis nula de Wilcoxon establece que la mediana de las diferencias entre dos poblaciones es igual a cero.

Se aplicó la prueba en Rstudio y se obtuvo un p-value de $1.991x10^{-9}$, menor a la significancia del 5%, por lo que se rechaza la hipótesis nula y se concluye que sí hay una diferencia significativa en los valores del VO_2 máximo con y sin valores atípicos. Si bien variaciones en el VO_2 máximo puede que no sean críticas dependiendo del estado de salud del paciente, de acuerdo con un análisis realizado por la Fundación de Investigación Cardiovascular (CRF por sus siglas en inglés) con 102980 participantes, el incremento de 1-MET en el VO_2 máximo, que corresponde a 3.5 ml O_2/kg/min, se asocia con una reducción de riesgo del 13% al 15% de mortalidad para todas las causas y de enfermedades cardiovasculares. De igual forma la capacidad cardiorrespiratoria, generalmente expresada en consumo máximo de oxígeno (VO_2 max), se reconoce como un predictor de mortalidad más fuerte que los factores de riesgo establecidos, como la hipertensión arterial, la diabetes y el tabaquismo, tanto en personas sanas como en aquellas que presentan alguna patología [63].

Objetivo Específico 4

El aplicativo se diseñó y construyó con el fin de proporcionar una herramienta funcional a los médicos que les permita aplicar el método de detección de atípicos Isolation Forest a los datos de una prueba CPET máxima de cualquier paciente y, de paso, soportar su proceso diagnóstico. Así mismo, dentro de las funciones del aplicativo están la generación de gráficas de VO_2 antes y después de la detección, y el cálculo del VO_2 máximo con y sin valores atípicos.

La interfaz del aplicativo se construyó en Python con la biblioteca *Tkinter*. En la imagen 2 se ven las tres secciones que componen la interfaz de la herramienta: a la izquierda se encuentra el logo de la Pontifica Universidad Javeriana con el nombre del aplicativo; la sección derecha se compone de tres botones: uno para seleccionar el archivo de los datos de la prueba (este archivo se obtiene del dispositivo en el que se efectúa la CPET y debe estar en formato de Excel), otro que pone en marcha el método de detección de valores atípicos en los datos y un botón que realiza el cálculo de VO_2 máximo con y sin atípicos. Por último, en la sección inferior hay un botón inicial que despliega dos botones que hacen referencia a las dos opciones de gráfico, con y sin valores atípicos marcados.

Imagen 2. Interfaz del aplicativo

Los gráficos se programaron para desplegarse en un formato interactivo con el fin de facilitar la lectura e interpretación por parte del médico tratante. Para ello se utilizó la librería "plotly express" que permite obtener gráficos en formato html en los cuales se puede interactuar, obteniendo, por ejemplo, la proyección en cada eje del punto seleccionado y la detección de valores atípicos en rojo como se muestra en la imagen 3.

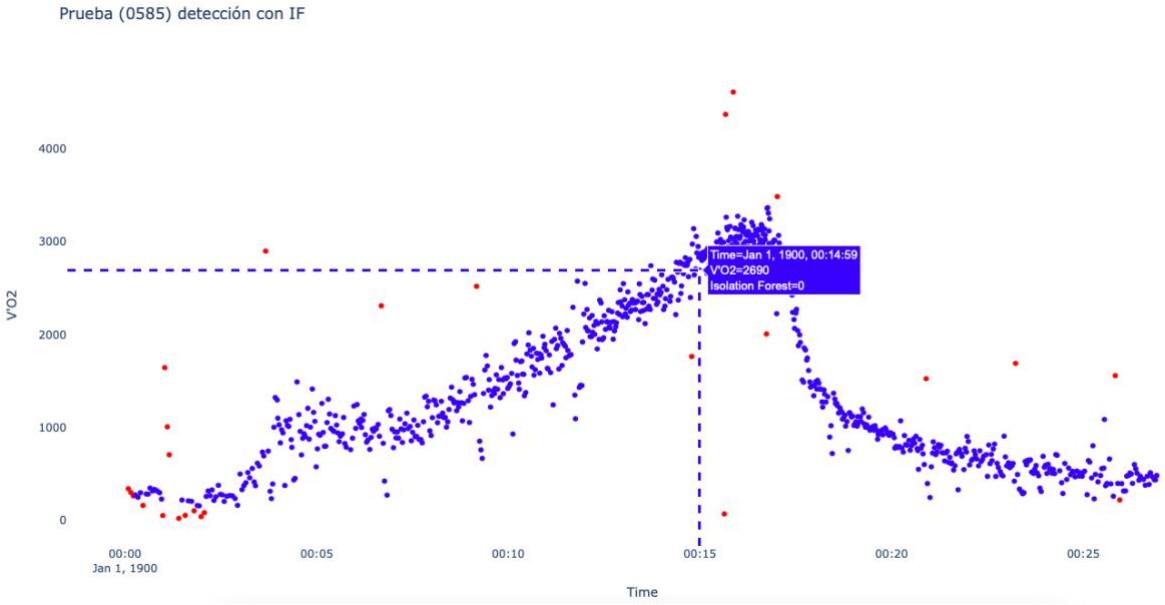

Imagen 3. Gráfico de VO_2 vs Tiempo con detección de atípicos por el método Isolation Forest en la CPET del paciente 0585.

Este despliegue interactivo de los gráficos cuenta con una barra de herramientas que se presenta en la imagen 4. Esta permite la interacción con el usuario, en donde podrá descargar el gráfico en formato png, hacer zoom a una sección específica, desplazarse por el gráfico, seleccionar un recuadro determinado para análisis, seleccionar una figura personalizada, ampliar o reducir el tamaño de la imagen y redimensionar el gráfico después de algunos cambios.

Imagen 4. Barra de herramientas que acompaña la interacción del gráfico para cada CPET.

Contar con un gráfico interactivo que permita a un experto en CPET profundizar en el análisis de la prueba es importante porque le permitirá validar distintos comportamientos y momentos críticos del desempeño de cada paciente, permitiendo así identificar visualmente los umbrales ventilatorios que podrán apoyar la toma de decisiones del experto.

Adicionalmente, se construyó un manual de uso del aplicativo para el usuario (ver anexo 4), el cual se divide en tres secciones. Inicialmente se presenta un resumen general del aplicativo y se enuncia su objetivo, aclarando que es una herramienta para el despliegue visual del comportamiento de la variable VO_2 con respecto al tiempo de la prueba y que puede usarse como apoyo al profesional durante el proceso diagnóstico. En segunda instancia se encuentran los requerimientos necesarios para el adecuado funcionamiento del aplicativo y, por último, se explica el paso a paso que se debe seguir para aplicar la detección de valores atípicos sobre un archivo de datos de una prueba CPET máxima, el posterior cálculo del valor de VO_2 máximo y la creación de gráficas antes y después de eliminar los registros etiquetados como atípicos por el método.

Objetivo Específico 5

La evaluación del aplicativo se desarrolló bajo la norma ISO 9126, un estándar internacional para la evaluación de la calidad de productos de software, el cual establece que cualquier componente calidad de un software puede ser descrito en seis características: funcionalidad, confiabilidad, usabilidad, eficiencia, portabilidad y mantenibilidad [64].

En la Tabla 15 se presenta el cuestionario diseñado para evaluar el desempeño del aplicativo. Este se compone de afirmaciones que exploran y evalúan cada una de las primeras cinco características establecidas en la norma ISO 9126. No se abordó en la evaluación la característica de mantenibilidad pues hace referencia a los esfuerzos necesarios para corregir o realizar modificaciones en el software, lo cual no está dentro del alcance de este trabajo. Para la evaluación de esta característica se requiere una implementación real del aplicativo y un monitoreo constante de su evolución y errores presentados.

La escala que se utilizó para la calificación de cada afirmación fue la escala de Likert, un instrumento psicométrico donde el encuestado indica su nivel de acuerdo o desacuerdo sobre una afirmación [65]. Para efectos de nuestra evaluación 1 significa totalmente en desacuerdo y 5 totalmente de acuerdo.

Característica	Afirmación	1	2	3	4	5
Usabilidad	El manual de uso explica de forma adecuada cómo usar el aplicativo y da un entendimiento completo de sus funciones.					
	La interfaz del aplicativo es visualmente clara, sus botones son fáciles de ubicar y siguen una secuencia lógica.					
	El diseño del aplicativo es intuitivo y fácil de utilizar.					
Funcionabilidad	Las funciones que tiene el aplicativo ejecutan de forma adecuada las tareas que fueron explicadas en el manual de uso.					
	El aplicativo permite cargar fácilmente el archivo a ejecutar.					
	El valor del VO_2 máximo con y sin atípicos es legible y corresponde a valores coherentes derivados de una prueba CPET.					
	Las gráficas de VO_2 vs Tiempo se pueden leer y analizar de forma clara, con un formato adecuado para la lectura de la prueba CPET.					
Confiabilidad	El aplicativo protege la información que el usuario comparte con él ante posibles filtraciones.					
	Ante una falla, el aplicativo restablece su nivel operación y recobra los datos que había recibido.					
Eficiencia	El tiempo de respuesta del aplicativo es rápido.					
	La calidad del producto que entrega el aplicativo es adecuada con respecto al tiempo que toma ejecutarlo.					
Portabilidad	El aplicativo es una herramienta que cumple su objetivo (detección de valores atípicos en la CPET) y no conozco de otro software o herramienta que cumpla con la misma función.					

Tabla 15. Formato de evaluación del aplicativo

Para la evaluación del aplicativo, se contactaron 3 médicos dentro de la Universidad Javeriana que tienen experiencia en la lectura de pruebas cardiopulmonares de ejercicio y que son ajenos al desarrollo de este trabajo. Una vez cada médico leyó el manual, se le solicitó que siguiera las instrucciones allí indicadas para leer el archivo de una prueba CPET, aplicar el algoritmo de detección de atípicos, calcular el VO_2 máximo y desplegar las gráficas del VO_2 versus tiempo. Luego, se les solicitó evaluar el formato con base en su experiencia de uso.

El resumen de los resultados se muestra en la Tabla 16, donde se realizó un promedio de la evaluación de las afirmaciones por característica para cada evaluador, obteniendo así tres puntajes generales por característica. Posteriormente para obtener una calificación global se promediaron los puntajes

generales de los evaluadores. La escala de Likert permite medir directamente factores que constituyen actitudes o creencias, por lo tanto, es común sumar o promediar las valoraciones obtenidas de estos ítems para su evaluación [66].

Característica	Evaluador 1	Evaluador 2	Evaluador 3	Puntaje global
Usabilidad	4.3	5	5	4.7
Funcionabilidad	5	4.5	5	4.8
Confiabilidad	5	3	Respuesta= No aplica	2.6
Eficiencia	5	4.5	4	4.5
Portabilidad	4	5	5	4.6

Tabla 16. Resumen evaluación del aplicativo

Los resultados obtenidos muestran que el desempeño del aplicativo fue óptimo ya que los puntajes de usabilidad, funcionabilidad y portabilidad superan el valor 4.5. El puntaje global del criterio de confiabilidad es bajo debido a que la respuesta del tercer evaluador fue "No aplica" ya que, según su criterio, no le es posible saber si el aplicativo puede proteger la información compartida ni tampoco si recupera los datos ante alguna falla.

5. Limitaciones, conclusiones y recomendaciones.

5.1 Limitaciones

- Los datos sobre los cuales se hizo la selección del método más adecuado para detectar atípicos son obtenidos del proyecto de investigación "Ejercicio físico en sujetos sanos no entrenados nativos a altitud moderada en comparación con crónicamente aclimatados" financiado por Colciencias y la PUJ, y recolectados entre el año 2016 y el 2018, lo cual podría limitar los resultados de desempeño del algoritmo al ser obtenidos de pacientes entre 18 y 25 años sin patologías y considerados sanos. No fue posible verificar si el método Isolation Forest es el más adecuado para identificar atípicos en pacientes de otros rangos de edad y características de salud.
- El aplicativo se usa de manera local en el computador y solo se puede acceder a él mediante el código de Python, lo que limita el acceso del personal médico a la herramienta.
- El aplicativo está diseñado para reconocer las variables solo si las columnas tienen el nombre "Time" y "V'O2".

5.2 Conclusiones

Después de recuperar las 407 pruebas CPET, se procedió con el análisis descriptivo y preparación de los datos, donde se encontró que el promedio de VO_2 en los individuos que tomaron la prueba es de 1228.19 ml/min con una desviación estándar de 295.82 ml/min. Se realizó una limpieza de los datos, identificando y eliminando datos erróneos y faltantes, presentes en el 49.5% de las pruebas. Posteriormente, bajo la supervisión de un médico experto en pruebas cardiopulmonares de ejercicio, se etiquetaron los datos bajo una clasificación binaria, siendo 0 no atípico y 1 atípico, en las 400 pruebas que cumplieron con el requisito de prueba máxima.

Con el fin de determinar la naturaleza de los métodos de detección de valores atípicos a implementar, se aplicó la prueba de normalidad Shapiro Wilk sobre el conjunto de datos, donde se encontró que no siguen una distribución normal. Se seleccionaron seis métodos no paramétricos, LOF, DBSCAN, KNN, Isolation Forest, LDOF y Regresion Kernel, que fueron evaluados

inicialmente en 280 pruebas de "train" para estimar los parámetros correspondientes a cada método (número de vecinos, contaminación y ancho de banda), de acuerdo con la métrica F_β. Finalmente, con los parámetros definidos se aplicaron los métodos de detección en las 120 pruebas de "test", construyendo seis matrices de confusión agregadas, donde el método que identificó correctamente la mayor cantidad de atípicos fue el Isolation Forest con 1262 TP.

La métrica seleccionada fue F_β con $\beta=2$, priorizando la sensibilidad sobre la precisión en la detección. El método que presentó un mejor desempeño fue el Isolation Forest, con un F_β de 0.7742, siendo el más sensible y el menos preciso. De igual forma, se recomienda no descartar la Regresión Kernel debido a que estuvo 1.12 puntos porcentuales por debajo del Isolation Forest, en cuanto a la métrica seleccionada, con un buen desempeño. Es importante seguir evaluando esta métrica para futuros estudios de detección en pruebas cardiopulmonares de ejercicio.

De acuerdo con la prueba Wilcoxon se determinó que sí existe una diferencia significativa en el VO_2 máximo con y sin valores atípicos. Estas variaciones pueden ser importantes para el diagnóstico ya que esta medida es una forma de representar la capacidad cardiorrespiratoria de los pacientes y se considera como un predictor de mortalidad. Su variación, en especial el incremento, puede reducir la mortalidad y las enfermedades cardiovasculares. Es de suma importancia calcular de la manera más exacta el VO_2 máximo para evitar resultados erróneos o que no se asemejen a la realidad del paciente.

Se construyó un aplicativo de escritorio capaz de aplicar el método de detección Isolation Forest sobre los resultados de una prueba CPET máxima cargados en formato Excel. Además, el aplicativo cuenta con la función de calcular el VO_2 máximo antes y después de eliminar los datos atípicos posibilitando la comparación de estos valores. Dada la importancia de los elementos gráficos al momento de interpretar una CPET, el aplicativo permite al usuario observar gráficamente el comportamiento de la variable VO_2 a través del tiempo; en la gráfica los datos atípicos se identifican con color rojo y los normales con color azul. Otro aspecto a destacar del aplicativo es que no es necesario contar con una conexión a internet para hacer uso del mismo.

El uso del aplicativo tuvo resultados favorables, con un puntaje promedio de 4.24 entre las cinco características definidas. Esto se debe a la gran utilidad que encontraron los evaluadores sobre el aplicativo, explicando que facilitaba el análisis de la prueba y reduce la subjetividad de la detección de valores atípicos. Esta herramienta unifica la detección de valores atípicos y el cálculo de una variable de alto impacto en la evaluación de la CPET con una interfaz amigable con el usuario y de fácil entendimiento, que aplicando las recomendaciones realizadas por los evaluadores puede convertirse en un aplicativo de uso clínico profesional.

5.3 Recomendaciones

- Por parte de los evaluadores del aplicativo se recomienda obtener el cociente entre el VO_2 máximo y el peso del paciente para hacer un análisis más preciso de esta medida. De igual manera presentar las gráficas de dispersión de VO_2 vs tiempo con y sin valores atípicos en un formato por promedio de cinco a ocho respiraciones, para obtener un gráfico que facilite la lectura de la prueba.

6. Anexos

Número del anexo	Nombre	Desarrollo	Tipo de archivo	Enlace
1	Notebook de Google Colab	Propio	Notebook en Google Colab	https://colab.research.google.com/drive/1lNaHFjcMsU2r55OZzT2uSDmjqCPXweeZ?usp=sharing
2	Métricas de desempeño de los métodos en test	Propio	Excel	https://docs.google.com/spreadsheets/d/1tORfeBZH0lVLqK090b2ulU1RV0vRINOl/edit?usp=sharing&ouid=113642708350744420309&rtpof=true&sd=true
3	Código de diseño del aplicativo	Propio	.text	https://drive.google.com/file/d/1YO9VOlHyiIrcQUVoMFrWMku-mKE3N4pA/view?usp=sharing
4	Manual de uso	Propio	Word	https://docs.google.com/document/d/1f1TwZuICVHSzVSRt7XH588jyvt4hIkwl/edit?usp=sharing&ouid=113642708350744420309&rtpof=true&sd=true
5	Evaluación de aplicativo	Propio - evaluadores	PDF	https://drive.google.com/file/d/1UIARikqF3qXZFUdIWzaLvECI4_E0z4dF/view?usp=sharing

7. Referencias

1. Porszasz J, Stringer W, Casaburi R. Equipment, measurements and quality control. ERS Monograph. 2018;2018:59–81.
2. de Boer E, Petrache I, Mohning MP. Cardiopulmonary Exercise Testing. JAMA - Journal of the American Medical Association. 2022;327:1284–5.
3. Weisman IM, Weisman IM, Marciniuk D, Martinez FJ, Sciurba F, Sue D, et al. ATS/ACCP Statement on cardiopulmonary exercise testing. Am J Respir Crit Care Med. 2003;167:211–77.
4. McQuaid C, Brady M, Deane R. SARS-CoV-2: is there neuroinvasion? Fluids and Barriers of the CNS. BioMed Central Ltd; 2021.
5. Romero-Duarte Á, Rivera-Izquierdo M, Guerrero-Fernández de Alba I, Pérez-Contreras M, Fernández-Martínez NF, Ruiz-Montero R, et al. Sequelae, persistent symptomatology and outcomes after COVID-19 hospitalization: the ANCOHVID multicentre 6-month follow-up study. BMC Medicine. BioMed Central Ltd; 2021;19.
6. Ali RMM, Ghonimy MBI. Post-COVID-19 pneumonia lung fibrosis: a worrisome sequelae in surviving patients. Egyptian Journal of Radiology and Nuclear Medicine. Springer Science and Business Media Deutschland GmbH; 2021;52.
7. Dorelli G, Braggio M, Gabbiani D, Busti F, Caminati M, Senna G, et al. Importance of Cardiopulmonary Exercise Testing amongst Subjects Recovering from COVID-19. Diagnostics [Internet]. 2021;11:507. Available from: https://www.mdpi.com/2075-4418/11/3/507

8. Faghy MA, Sylvester KP, Cooper BG, Hull JH. Cardiopulmonary exercise testing in the COVID-19 endemic phase. British Journal of Anaesthesia [Internet]. Elsevier Ltd; 2020;125:447–9. Available from: https://doi.org/10.1016/j.bja.2020.06.006
9. Levett DZH, Jack S, Swart M, Carlisle J, Wilson J, Snowden C, et al. Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation. British Journal of Anaesthesia. 2018;120.
10. Peterková A, Michaľčonok G. PREPROCESSING RAW DATA IN CLINICAL MEDICINE FOR A DATA MINING PURPOSE.
11. Ross RM, Corry DB. Software for interpreting cardiopulmonary exercise tests. BMC Pulmonary Medicine. 2007;7.
12. Inbar O, Inbar O, Reuveny R, Segel MJ, Greenspan H, Scheinowitz M. A Machine Learning Approach to the Interpretation of Cardiopulmonary Exercise Tests: Development and Validation. Pulmonary Medicine. Hindawi Limited; 2021;2021.
13. Ahmad P, Qamar S, Qasim Afser Rizvi S. Techniques of Data Mining In Healthcare: A Review. International Journal of Computer Applications. Foundation of Computer Science; 2015;120:38–50.
14. Larose DT. Discovering Knowledge in Data : An Introduction to Data Mining [Internet]. Hoboken, N.J.: Wiley-Interscience; 2005. Available from: https://login.ezproxy.javeriana.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=127312&lang=es&site=eds-live
15. Kuppusamy M, Kannan Kaliyaperumal S, Kannan SK. Comparison of Methods for detecting Outliers Comparative analysis of community discovery methods in social networks View project Comparison of methods for detecting outliers. 2013; Available from: http://www.ijser.org
16. Wang S, Celebi ME, Zhang Y-D, Yu X, Lu S, Yao X, et al. Advances in Data Preprocessing for Biomedical Data Fusion: An Overview of the Methods, Challenges, and Prospects. Information Fusion. Elsevier BV; 2021;76:376–421.
17. Lin JH, Haug PJ. Data preparation framework for preprocessing clinical data in data mining. AMIA . Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006;489–93.
18. ur Rehman A, Belhaouari SB. Unsupervised outlier detection in multidimensional data. Journal of Big Data. Springer Science and Business Media Deutschland GmbH; 2021;8.
19. Kwak SK, Kim JH. Statistical data preparation: Management of missing values and outliers. Korean Journal of Anesthesiology. 2017;70:407–11.
20. Jaward M, Wang H. A Parametric and Non-Parametric Approach for High-Accurate Outlier Detection. JOURNAL OF INFORMATION SCIENCE AND ENGINEERING. 2020;36:441–65.
21. Smiti A. A critical overview of outlier detection methods. Computer Science Review. Elsevier Ireland Ltd; 2020.
22. Dumitrescu D, Rosenkranz S. Graphical data display for clinical cardiopulmonary exercise testing. Ann Am Thorac Soc. American Thoracic Society; 2017. p. S12–21.
23. Neder JA, Phillips DB, Marillier M, Bernard AC, Berton DC, O'Donnell DE. Clinical Interpretation of Cardiopulmonary Exercise Testing: Current Pitfalls and Limitations. Frontiers in Physiology. Frontiers Media S.A.; 2021;12.
24. Abdallah J, Astal AL. Comparison of Methods for Detecting Outliers in Medical Data.
25. Tallón-Ballesteros AJ, Riquelme JC. Deleting or keeping outliers for classifier training? 2014 6th World Congress on Nature and Biologically Inspired Computing, NaBIC 2014. IEEE; 2014;281–6.
26. Kulczycki P, Franus K. Methodically unified procedures for a conditional approach to outlier detection, clustering, and classification. Information Sciences. Elsevier Inc.; 2021;560:504–27.
27. Liu FT, Ting KM, Zhou ZH. Isolation forest. Proceedings - IEEE International Conference on Data Mining, ICDM. 2008;413–22.
28. Anusha PV, Anuradha C, Chandra Murty PSR, Kiran CS. Detecting outliers in high dimensional data sets using Z-score methodology. International Journal of Innovative Technology

and Exploring Engineering. Blue Eyes Intelligence Engineering and Sciences Publication; 2019;9:48–53.
29. Gradshteyn IS, Ryzhik IM, Johnson NL, Kotz S. Maximum Z Scores and Outliers. Academic Press; 1988.
30. Pupovac V, Petrovecki M. Summarizing and presenting numerical data. 2011.
31. Hu W, Gao J, Li B, Wu O, Du J, Maybank S. Anomaly Detection Using Local Kernel Density Estimation and Context-Based Regression. IEEE Transactions on Knowledge and Data Engineering. 2020;32:218–33.
32. Latecki LJ, Lazarevic A, Pokrajac D. Outlier detection with kernel density functions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2007;4571 LNAI:61–75.
33. Tang B, He H. A local density-based approach for outlier detection. Neurocomputing. Elsevier B.V.; 2017;241:171–80.
34. Smiti A. A critical overview of outlier detection methods. Computer Science Review. Elsevier Ireland Ltd; 2020.
35. IEEE Computational Intelligence Society, International Neural Network Society, Institute of Electrical and Electronics Engineers, IEEE World Congress on Computational Intelligence (2020 : Online). An Outlier Detection Algorithm based on KNN-kernel Density Estimation.
36. Mandhare H, Idate S. Distance Based Outlier Detection and Density Based Outlier Detection Techniques. International Conference on Intelligent Computing and Control Systems. 2017;931–5.
37. Hubballi N, Patra BK, Nandi S. N DoT : Nearest Neighbor Distance Based Outlier. 2011;36–42.
38. Zhang K, Hutter M, Jin H. A new local distance-based outlier detection approach for scattered real-world data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009;5476 LNAI:813–22.
39. Ijaz MF, Alfian G, Syafrudin M, Rhee J. Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest. Applied Sciences (Switzerland). MDPI AG; 2018;8.
40. Roselin AG, Nanda P, Nepal S, He X. Intelligent Anomaly Detection for Large Network Traffic with Optimized Deep Clustering (ODC) Algorithm. IEEE Access. Institute of Electrical and Electronics Engineers Inc.; 2021;9:47243–51.
41. Papadias D, Tao Y. Reverse Nearest Neighbor Query 1. 1998;4:2434–8.
42. Yuan Y, Zhang Y, Cao H, Yao R. New local density definition based on minimum hyper sphere for outlier mining algorithm using in industrial databases. 26th Chinese Control and Decision Conference, CCDC 2014. IEEE; 2014;5182–6.
43. Ding Z, Fei M. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window [Internet]. IFAC Proceedings Volumes (IFAC-PapersOnline). IFAC; 2013. Available from: http://dx.doi.org/10.3182/20130902-3-CN-3020.00044
44. Lalkhen AG, McCluskey A. Clinical tests: Sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care and Pain. Oxford University Press; 2008;8:221–3.
45. CHROMIŃSKI K, TKACZ M. COMPARISON OF OUTLIER DETECTION METHODS IN BIOMEDICAL DATA. JournalJof medical informatices and technologies. 2010.
46. Lee N, Yang H, Yoo H. A surrogate loss function for optimization of F_β score in binary classification with imbalanced data. 2021;1–17. Available from: http://arxiv.org/abs/2104.01459
47. Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms (and classification and information retrieval systems). Statistics and Computing. 2018;28:1–13.
48. Chinchor N. MUC-4 EVALUATION METRIC S.
49. Berthold MR, Feelders A, Krempl G, editors. Advances in Intelligent Data Analysis XVIII [Internet]. Cham: Springer International Publishing; 2020. Available from: http://link.springer.com/10.1007/978-3-030-44584-3

50. Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. Machine Learning [Internet]. Springer US; 2021;110:451–6. Available from: https://doi.org/10.1007/s10994-021-05964-1
51. Zhou H, Liu H, Zhang Y, Zhang Y. An outlier detection algorithm based on an integrated outlier factor. Intelligent Data Analysis. 2019;23:975–90.
52. Xu X, Liu H, Li L, Yao M. A comparison of outlier detection techniques for high-dimensional data. International Journal of Computational Intelligence Systems. 2018;11:652–62.
53. Badithela A, Wongpiromsarn T, Murray RM. Leveraging Classification Metrics for Quantitative System-Level Analysis with Temporal Logic Specifications. Proceedings of the IEEE Conference on Decision and Control. IEEE; 2021;2021-Decem:564–71.
54. Capodieci A, Caricato A, Carlucci AP, Ficarella A, Mainetti L, Vergallo C. Using different machine learning approaches to evaluate performance on spare parts request for aircraft engines. E3S Web of Conferences. 2020;197.
55. Walther DB. Using confusion matrices to estimate mutual information between two categorical measurements. Proceedings - 2013 3rd International Workshop on Pattern Recognition in Neuroimaging, PRNI 2013. IEEE; 2013;220–4.
56. Mandl KD, Gottlieb D, Ellis A. Beyond One-Off Integrations: A Commercial, Substitutable, Reusable, Standards-Based, Electronic Health Record-Connected App. J Med Internet Res. 2019;21:e12902.
57. Warner JL, Rioth MJ, Mandl KD, Mandel JC, Kreda DA, Kohane IS, et al. SMART precision cancer medicine: A FHIR-based app to provide genomic information at the point of care. Journal of the American Medical Informatics Association. 2016;23:701–10.
58. Zignoli A, Fornasiero A, Rota P, Muollo V, Peyré-Tartaruga LA, Low DA, et al. Oxynet: A collective intelligence that detects ventilatory thresholds in cardiopulmonary exercise tests. European Journal of Sport Science. Taylor and Francis Ltd.; 2021;
59. Midgley AW, McNaughton LR, Polman R, Marchant D. Criteria for Determination of Maximal Oxygen Uptake. Sports Medicine [Internet]. New Zealand; 2007;37:1019–28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18027991
60. Wasserman K, Kathy S, Hansen J, Sun X-G, Sue DY, Whipp BJ, et al. Principles of Exercise Testing and Interpretation Including Pathophysiology and Clinical Applications Fifth Edition. 2012.
61. Ichikawa Y, Maeda T, Takahashi T, Ashikaga K, Tanaka S, Sumi Y, et al. Changes in oxygen uptake kinetics after exercise caused by differences in loading pattern and exercise intensity. ESC Heart Failure. Wiley-Blackwell; 2020;7:1109–17.
62. Ghadhban GA, Rasheed HA. Robust tests for the mean difference in paired data using Jackknife resampling technique. Iraqi Journal of Science. 2021;62:3081–90.
63. Lee D chul, Artero EG, Sui X, Blair SN. Mortality trends in the general population: the importance of cardiorespiratory fitness. J Psychopharmacol. 2010;24:27–35.
64. Abud Figueroa María. Calidad en la Industria del Software . La Norma ISO-9126. Calidad en la Industria del Software La Norma ISO-9126 [Internet]. 2012;255. Available from: http://www.monografias.com/trabajos5/%0Ahttp://www.monografias.com/trabajos5/%0Ajavier8a.com/itc/bd1/Normas iso 9126.pdf
65. Matas A. Diseño del formato de escalas tipo Likert: Un estado de la cuestión. Revista Electronica de Investigacion Educativa. 2018;20:38–47.
66. León-Mantero C, Casas-Rosal JC, Pedrosa-Jesús C, Maz-Machado A. Measuring attitude towards mathematics using Likert scale surveys: The weighted average. PLoS ONE. 2020;15:1–15.