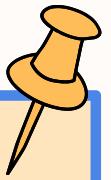
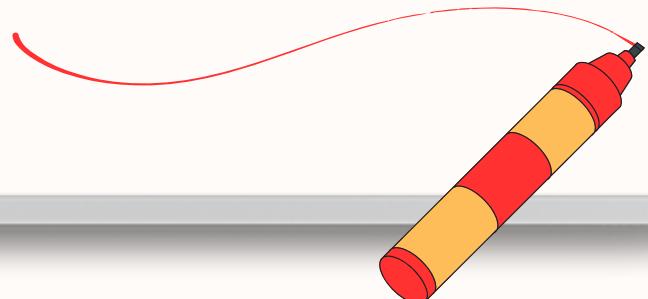


# ENGENHARIA E CIÊNCIA DE DADOS

## ENTREGA: 31 DE OUTUBRO DE 2025

### PROJETO 1

### ENGENHARIA DE ATRIBUTOS



**TRABALHO REALIZADO POR:**

Beatriz Martins - 2023210538; uc2023210538@student.uc.pt  
Laura Barreto - 2023214375; uc2023214375@student.uc.pt

# Projeto 1 - EA

## Introdução

Para este trabalho, o problema proposto é o reconhecimento de 16 atividades físicas humanas. O dataset disponibilizado tem dados de 15 participantes com 5 devices cada um, cada um com 3 sensores. Esses sensores, acelerômetro, giroscópio e magnetômetro, medem a aceleração, velocidade angular e variação do campo magnético, respectivamente, em diferentes partes do corpo.

## Análise e tratamento de outliers

Nesta etapa, iremos mostrar vários métodos na deteção de outliers nomeadamente o **interquartil**, o **Z-score**, **K-means** e **DBSCAN**.

### 1. Boxplots

Primeiramente, calculámos os módulos do acelerômetro, giroscópio e magnetômetro com base nas coordenadas x, y e z de cada um e obtivemos os seguintes boxplots destes módulos em cada atividade para todos os devices.

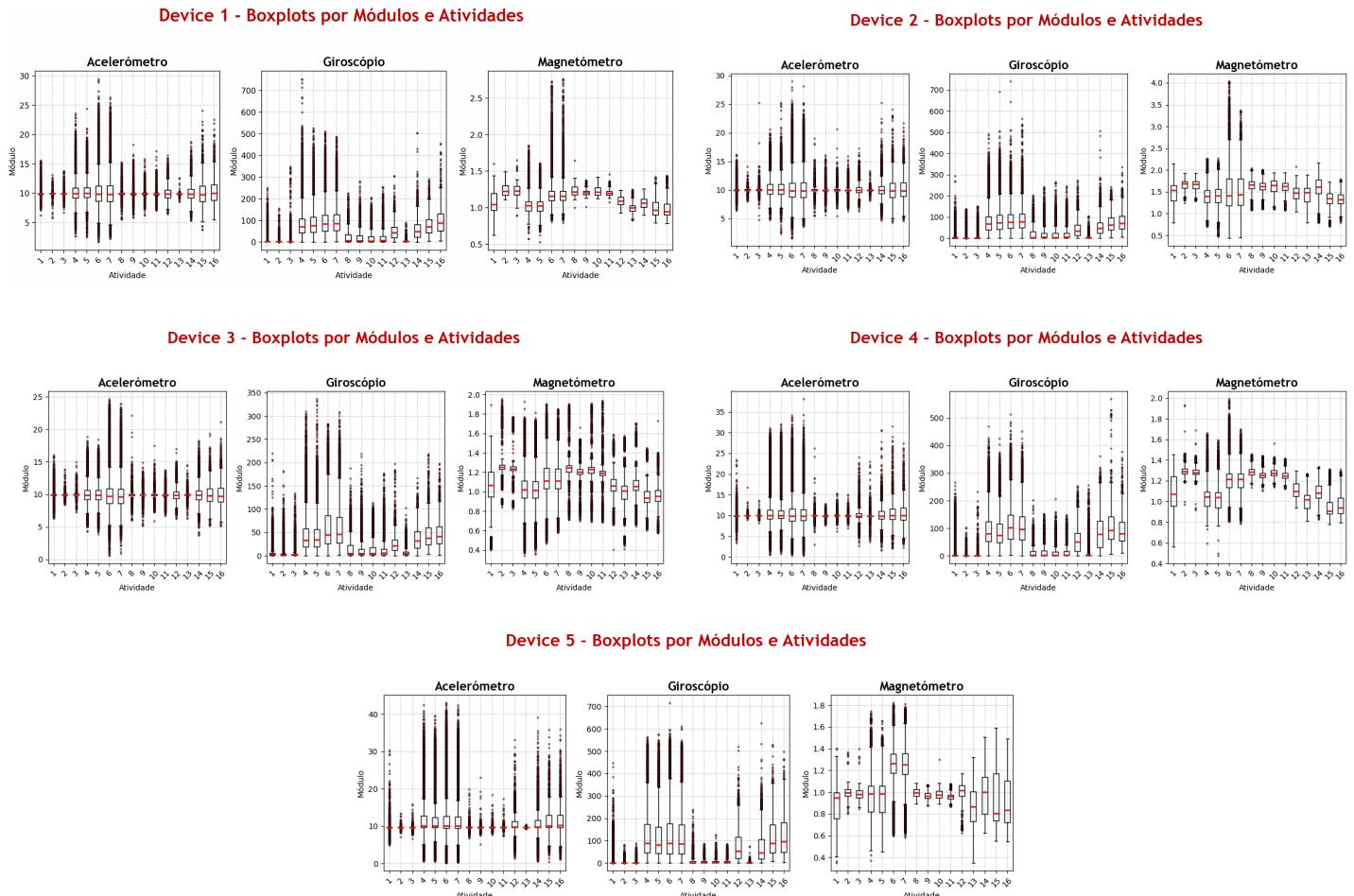


Figura 1: Distribuições dos módulos das 16 atividades para cada device

De forma geral, a aplicação do método **Interquartil (IQR)** para deteção de outliers (utilizando os limites  $Q3 + 1.5 \times IQR$  e  $Q1 - 1.5 \times IQR$ ) revelou distribuições com um número considerável de outliers. Observou-se ainda que o módulo do giroscópio apresenta, de modo consistente, médias mais baixas nas diferentes atividades, atingindo mesmo valores nulos em várias delas.

## 2. Densidade de Outliers por Módulo e Atividade

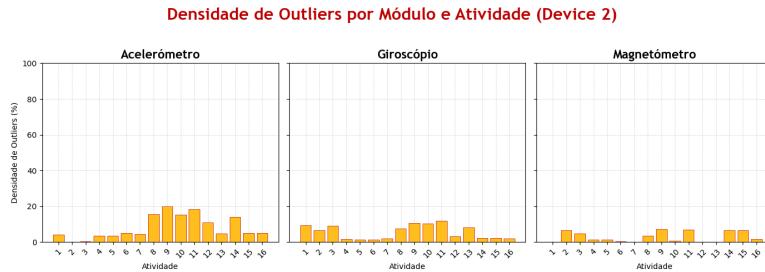


Figura 2: Densidade (%) de outliers nas atividades dos três módulos no device 2

Ao analisar a densidade de outliers no dispositivo 2, constata-se que esta não é tão elevada quanto os boxplots anteriores faziam parecer. Para além disso, a aceleração possui valores mais elevados, no entanto, estes nunca ultrapassam 20% dos dados da atividade. A atividade 9 (Sit → Stand) é a que apresenta uma maior percentagem de outliers no módulo do acelerômetro.

## 3. Deteção de outliers com o Z-score

Ao usar o **Z-score** para a deteção de outliers com  $k$  igual a 3, 3.5 e 4, obtivemos os seguintes gráficos.

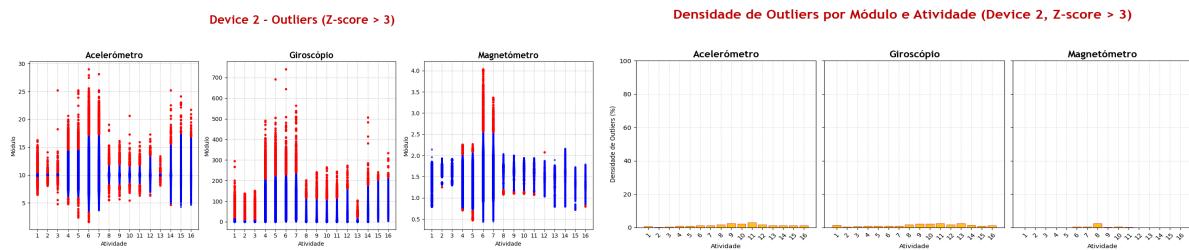


Figura 3: Outliers e a sua densidade detetados pelo z-score com  $k = 3$  no device 2

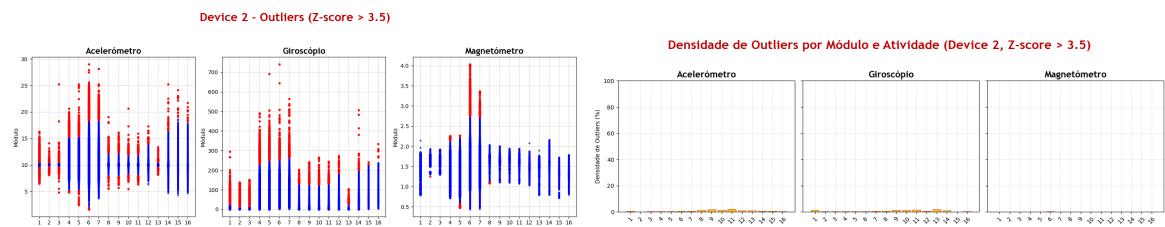
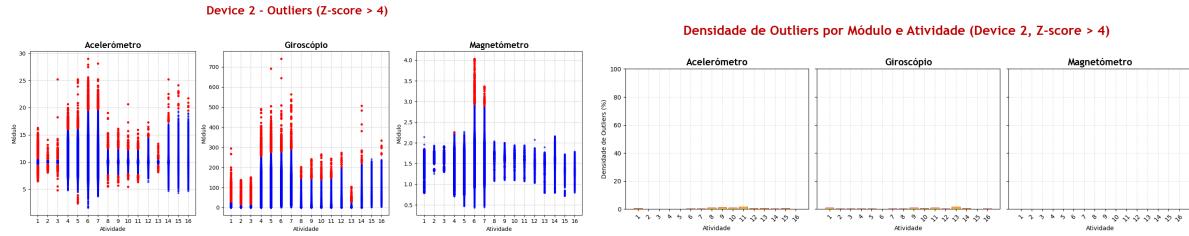


Figura 4: Outliers e a sua densidade detetados pelo z-score com  $k = 3.5$  no device 2

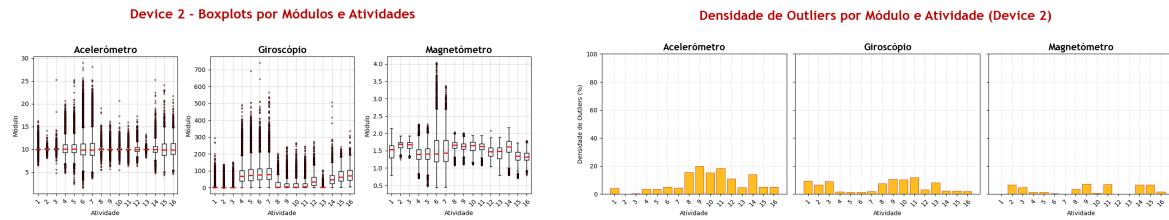


**Figura 5:** Outliers e a sua densidade detetados pelo z-score com  $k = 4$  no device 2

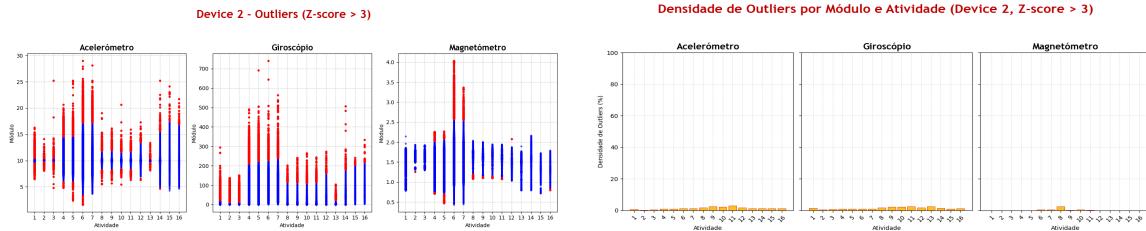
Podemos observar que existem alguns outliers nos extremos das distribuições das atividades, sendo as atividades 2 (Sit) e 3 (Sit and Talk) as que apresentam ter maior quantidade deles ao longo de toda a sua distribuição. Também podemos perceber que, com um  $k$  maior, obtemos uma menor densidade de outliers, porque o nosso limiar de decisão aumenta, tendo assim uma maior tolerância para aceitar pontos como não outliers.

#### 4. Comparação dos resultados obtidos no IQR e Z-score

Em resposta ao exercício 3.5, abaixo comparámos e discutimos os resultados obtidos nas duas abordagens: **IQR** e **Z-Score**.



**Figura 6:** Outliers e a sua densidade detetados pelo IQR com  $k = 1.5$  no device 2

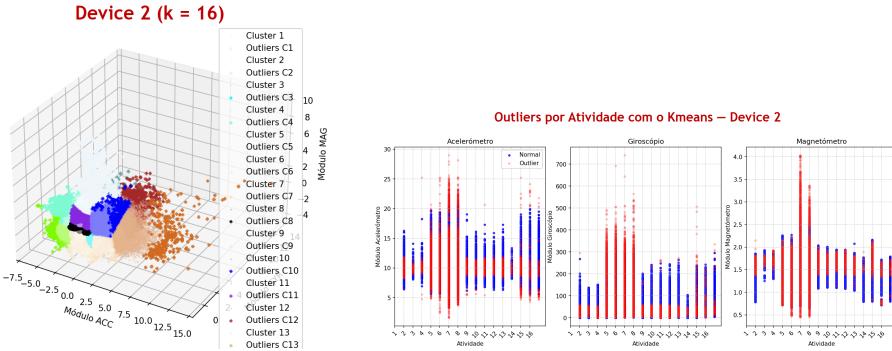


**Figura 7:** Outliers e a sua densidade detetados pelo Z-score com  $k = 3$  no device 2

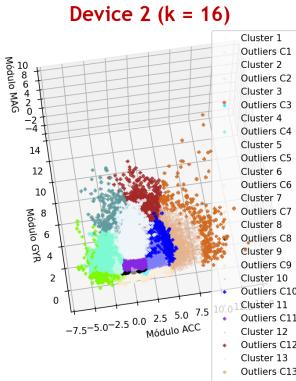
Observa-se que o método baseado no **Z-score** com  $k = 3$  identifica menos outliers comparativamente ao boxplot que usa o método **Interquartil** ( $k = 1.5$ ). Isto deve-se ao facto de o **Z-score** avaliar a distância normalizada de cada ponto em relação à média. Desta forma, com  $k = 3$ , apenas os valores mais extremos são assinalados como outliers. Em contrapartida, o **IQR** é mais sensível a variações acima do terceiro quartil e abaixo do primeiro quartil, assinalando um maior número de pontos como outliers.

## 5. Deteção de outliers com o K-means

Utilizámos o **K-means** com 16 clusters (de forma a tentar representar as 16 atividades) e dentro de cada um, aplicámos o método **IQR** com  $k = 2$  para detetar os outliers. Assim obtivemos bastantes outliers no meio da distribuição de cada atividade.

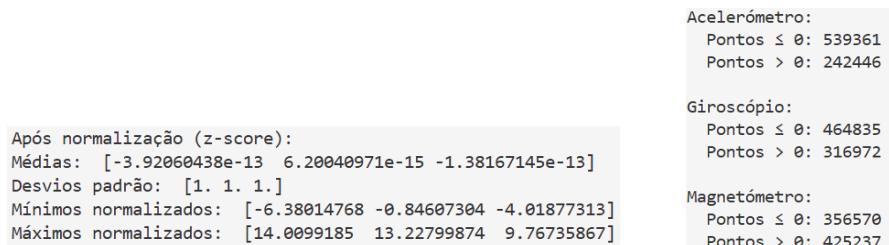


**Figura 8:** Outliers e a sua densidade detectados pelo k-means com 16 clusters no device 2

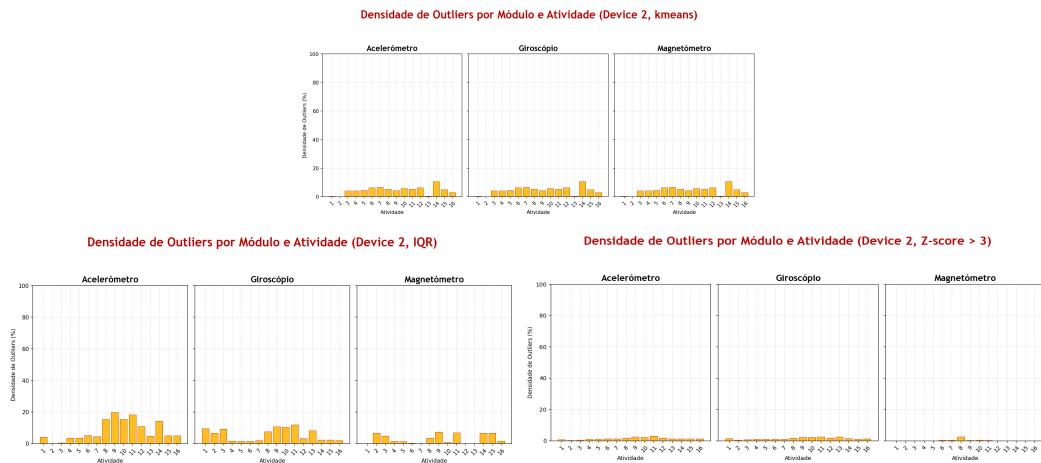


**Figura 9:** Outliers no device 2 detectados pelo k means com 16 clusters evidenciando o eixo do Giroscópio

Apesar de termos normalizado os módulos dos três eixos (através de **Z-score**), estes não aparecem estar totalmente normalizados, em particular, o módulo do giroscópio, cujos valores variam entre -0.85 e 13 (figura 10). Embora a média e o desvio padrão sejam aproximadamente 0 e 1, respetivamente (figura 10), observa-se uma grande concentração de pontos abaixo de zero (464 835 amostras), enquanto acima de zero existem apenas 316 972 pontos, com valores muito mais elevados. Esta assimetria faz com que o gráfico pareça enviesado: os outliers positivos, por serem numericamente elevados, compensam a massa de pontos negativos, resultando numa média próxima de zero, mas numa distribuição visualmente assimétrica.



**Figura 10:** Informações sobre os módulos normalizados de aceleração, giroscópio e magnetômetro

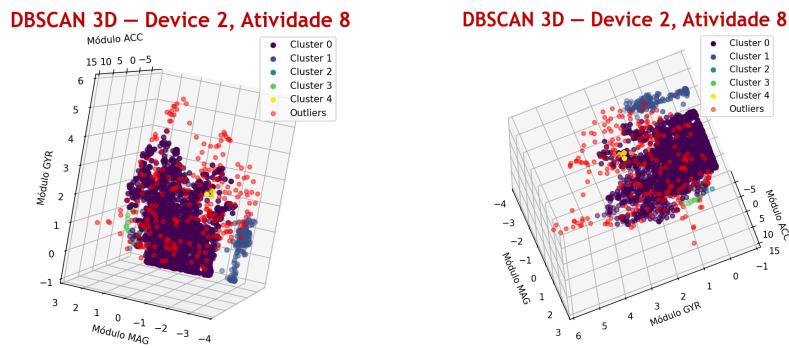


**Figura 11:** Comparação da densidade de outliers nas 3 abordagens: IQR, Z-score e K-means

Podemos perceber que, usando o **K-means**, temos valores intermédios de densidade de outliers, comparando com as duas abordagens anteriores. Para além disso, obtemos menor densidade de outliers nas primeiras duas atividades em todos os módulos, talvez porque a variância dessas atividades seja mais reduzida.

## 6. Deteção de outliers com o DBSCAN

Usando o método de densidades **DBSCAN** no device 2 e atividade 8, com os parâmetros eps e min samples a 0.5 e 5, respectivamente, de forma a detetar outliers, obtivemos o seguinte gráfico.



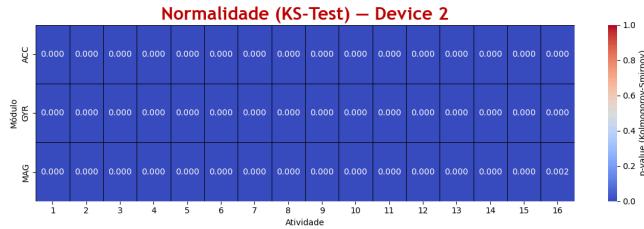
**Figura 12:** Resultados do método DBSCAN

## Extração de informação característica

Nesta etapa, determinámos a significância estatística das médias dos módulos nas diversas atividades e também foram extraídas as 110 features que o artigo “BodyNets 2011” (Zhang & Sawchuk) sugere e utilizados os métodos de **PCA**, **Fisher** e **ReliefF**

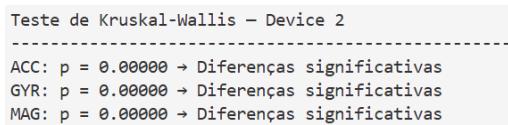
## 1. Significância estatística das médias dos módulos

O teste de **Kolmogorov–Smirnov**, aplicado aos módulos ACC, GYR e MAG em cada uma das atividades no dispositivo 2, revelou p-values próximos de zero (menores que 0.05), indicando que as distribuições não seguem uma distribuição normal.



**Figura 13:** P-values do teste Kolmogorov-Smirnov de cada atividade

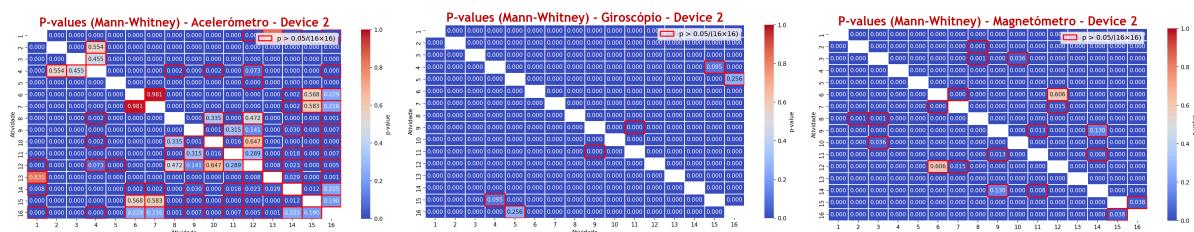
Assim, para comparar as médias entre atividades, foi utilizado o teste **não paramétrico e não emparelhado de Kruskal-Wallis**. Observando a Figura 14, verifica-se que os p-values dos três módulos são aproximadamente zero (menores que 0.05), o que confirma que existe pelo menos uma atividade cuja média difere significativamente das restantes, embora o teste não indique qual. Para identificar quais atividades diferem entre si, foi posteriormente aplicado o teste de **Mann-Whitney U**.



**Figura 14:** Resultado do teste não paramétrico Kruskal nos três módulos do device 2

O teste de **Mann–Whitney U** foi utilizado para avaliar a significância das médias de cada módulo em todas as combinações possíveis entre pares de atividades, identificando quais não apresentam diferenças estatisticamente significativas. Trata-se de um teste **não paramétrico e não emparelhado**, uma vez que a quantidade de amostras difere de atividade para atividade.

Como foram realizados  $16 \times 16$  testes, aplicou-se a **correção de Bonferroni**, que ajusta o nível de significância para 0.05 dividido pelo número total de testes. Assim, as células com p-valor superior a esse limiar são assinaladas a vermelho, conforme indicado na legenda dos gráficos abaixo.



**Figura 15:** Resultado do teste não paramétrico e não emparelhado Mann–Whitney U nos três módulos do device 2

O módulo do acelerômetro apresenta mais combinações de duas atividades com p-values acima do limiar, indicando menor capacidade de distinção entre essas atividades. Por exemplo, as médias deste módulo nas atividades 6 (Climb Stair) e 7 (Climb Stair and

Talk) não diferem significativamente, tendo o p-value mais elevado de 0.98. O Giroscópio, em particular, parece ser o módulo que melhor distingue as atividades, apresentando p-values baixos na maioria dos pares.

## 2. Redução de dimensionalidade através do PCA

Através dos dados originais, extraímos 110 features que o artigo “BodyNets 2011” sugere usando janelas deslizantes de 5 segundos e overlap de 50% nos dados de cada participante (excluindo janelas com mais do que uma atividade e menos de 10 amostras). Obtivemos uma matriz de features com 30688 linhas e 110 features, como é possível observar na figura 17.

```
Matriz de features Final: (30688, 110)
Labels Finais: (30688,)
```

Figura 16: Matriz de features extraídas e o seu vetor de labels com as atividades de cada janela

Através do método do **PCA**, concluímos que são necessárias 22 componentes principais para explicar pelo menos 75% da variabilidade da matriz com as 110 features.

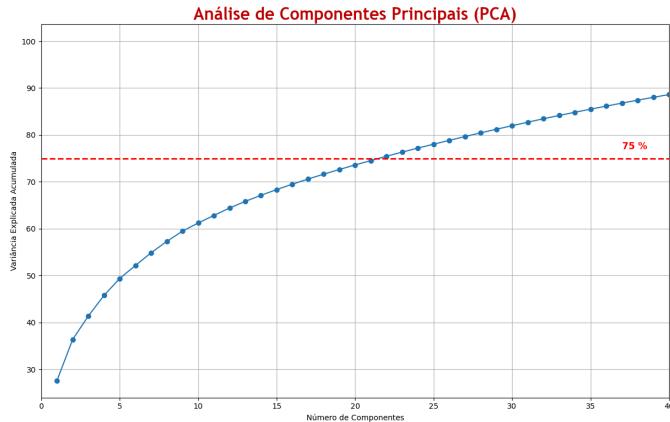


Figura 17: Aplicação do método PCA à matriz de features

### 2.1. Cálculo de um instante do PCA

Cada instante do conjunto de dados pode ser projetado num novo espaço reduzido. Para isso, **multiplica-se a linha correspondente a esse instante** (no conjunto de dados original normalizado) **pelas colunas da matriz dos vetores próprios**. O resultado dessa operação são as novas features comprimidas. Assim, ao selecionar um instante específico, basta extrair a sua linha da matriz original e realizar essa multiplicação para obter o vetor de componentes principais que representa esse instante no espaço reduzido.

## 2.2. Vantagens e Desvantagens da aplicação do PCA

Esta abordagem reduz significativamente a dimensionalidade, preservando a maior parte da variância dos dados. Além de economizar memória e tempo de processamento, atenua o ruído (ao descartar componentes com pouca variância) e mantém uma estrutura de componentes ordenadas decrescentemente, pelo valor de variância explicada, o que facilita a escolha do valor das k componentes principais. Para além disso, é uma técnica rápida e determinística.

No entanto, mesmo tendo muitas vantagens também tem limitações. Primeiramente, o **PCA** só representa relações lineares, levando a que padrões não lineares ou mais complexos não sejam bem captados. Para além disso, há perda de interpretabilidade, pois as componentes são combinações lineares de features, dificultando a sua interpretação. Outras duas limitações relacionam-se com o facto de ser um método não supervisionado (as componentes que mais variância explicam podem não ser as mais discriminantes para a classificação das atividades) e ser sensível a outliers, o que faz com que estes tenham uma enorme influência na covariância e nos vetores próprios. A normalização também é um problema nesta abordagem, porque sem ela, features com escalas maiores dominam, sendo obrigatório o uso de **Z-score**.

## 3. Fisher Score e ReliefF

O **Fisher Score** avalia cada feature pela sua capacidade de distinguir atividades. O objetivo deste método é maximizar a diferença entre as médias das atividades e minimizar a variância dentro de cada uma delas. Quanto maior o score da feature, mais discriminativa é a feature.

O método **ReliefF** escolhe uma amostra aleatória e encontra os seus vizinhos mais próximos da mesma classe (nearest hits) e de outras classes (nearest misses), calculando as distâncias com base em todas as features. Em seguida, avalia cada feature individualmente: se nessa dimensão a amostra estiver perto dos vizinhos da mesma classe e distante dos vizinhos de outras classes, o score dessa feature aumenta, já que essa feature ajuda a distinguir melhor as classes. Foram utilizados 10 vizinhos neste método.

Ao aplicar **Fisher Score** e **ReliefF**, obtivemos os resultados da figura 18.

Top Features segundo Fisher Score:

```
01. gyr_y_rms - Score = 10351.4353
02. gyr_y_std - Score = 8085.8412
03. gyr_y_iqr - Score = 6864.7055
04. AVH - Score = 6275.1373
05. gyr_z_mean_cross_rate - Score = 4193.9364
06. acc_y_std - Score = 3681.8874
07. gyr_y_mean_cross_rate - Score = 3253.3481
08. gyr_y_var - Score = 2539.8382
09. acc_y_rms - Score = 2446.0236
10. gyr_z_spec_entropy - Score = 2442.5634
```

Top Features segundo ReliefF:

```
01. corr_gyr_x_gyr_z - Score = 0.0744
02. corr_gyr_x_gyr_y - Score = 0.0738
03. corr_gyr_y_gyr_z - Score = 0.0643
04. corr_acc_x_acc_z - Score = 0.0551
05. acc_x_rms - Score = 0.0494
06. gyr_x_zero_cross_rate - Score = 0.0493
07. acc_x_mean_cross_rate - Score = 0.0467
08. corr_acc_z_gyr_y - Score = 0.0439
09. EVA3 - Score = 0.0437
10. acc_x_mean - Score = 0.0436
```

Figura 18: Top 10 features consoante o método Fisher e Relief

De acordo com **Fisher Score**, as features com maior score, ou seja que permitem distinguir melhor as atividades são **gyr\_y\_rms** (root mean square do eixo y do giroscópio),

`gyr_y_std` (desvio padrão do eixo y do giroscópio) e `gyr_y_iqr` (Interquartil do eixo y do giroscópio).

Já no **ReliefF**, as features que melhor dividem as atividades são `corr_gyr_x_gyr_z` (correlação entre os eixos x e z do giroscópio), `corr_gyr_x_gyr_y` (correlação entre os eixos x e y do giroscópio) e `corr_gyr_y_gyr_z` (correlação entre os eixos y e z do giroscópio).

Decidimos ainda representar graficamente as três melhores features de ambos os métodos num gráfico de três dimensões para verificar qual das duas abordagens consegue distinguir melhor as atividades. No entanto observando os gráficos concluímos que não parecem diferir significativamente.

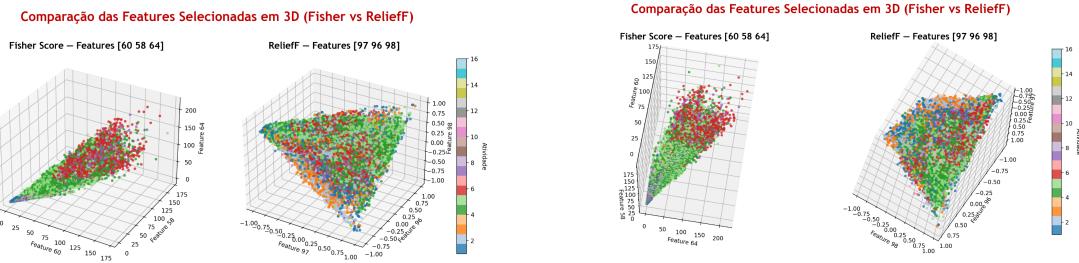


Figura 19: Gráficos de comparação das features selecionadas por Fisher Score e ReliefF

### 3.1. Cálculo de um instante do Fisher e Score ReliefF

Após a aplicação dos métodos **Fisher Score** e **ReliefF**, obtêm-se listas ordenadas das features mais relevantes de acordo com o seu poder discriminativo entre classes. As 10 melhores features correspondem às colunas do feature set cujos índices apresentam os valores de score mais elevados.

Assim, as features relativas a esta seleção são obtidas extraíndo as colunas do feature set que correspondem aos índices das 10 melhores features. Cada linha do feature set representa uma amostra (ou instante temporal). Deste modo, para um instante específico, o vetor reduzido de features é formado pelos valores das 10 features selecionadas nessa linha.

### 3.2. Vantagens e Desvantagens da aplicação do Fisher Score

A abordagem Fisher Score é simples e rápida, porque os cálculos são diretos e é eficiente. Para além disso, mantém a interpretabilidade, ou seja o resultado é fácil de entender (feature com maior score é mais discriminativa). O Fisher Score funciona bastante bem quando as classes têm médias bem distintas e é fácil de aplicar, não necessitando do uso de parâmetros complexos.

No entanto, uma abordagem deste tipo também tem as suas limitações. Para além de apenas analisar cada feature isoladamente, ignorando as relações entre elas, é sensível a correlações, ou seja, se duas features são muito correlacionadas, podem ambas receber scores mais altos, mesmo não adicionando nova informação. Este método tem ainda o problema de assumir que a distribuição é normal, tendo dificuldade em fronteiras de decisão

não lineares e de poder falhar caso as classes sejam desbalanceadas, uma vez que as classes maiores acabam por ter mais peso nos cálculos.

### **3.3. Vantagens e Desvantagens da aplicação do ReliefF**

Já a abordagem ReliefF, considera interações entre features, o que valoriza bastante aquelas que funcionam melhor em conjunto. Para além disso, não assume a linearidade, o que funciona bem com relações e fronteiras não lineares e identifica features úteis mesmo com ruído nos dados.

Todavia, o ReliefF é mais lento porque precisa de calcular distâncias entre amostras exigindo maior custo computacional em datasets maiores e os seus resultados podem variar se os parâmetros escolhidos para a abordagem não forem os melhores. Outras limitações são a sua sensibilidade à escala, caso as features não sejam normalizadas antes da sua aplicação e ainda uma menor interpretabilidade, uma vez que o seu score final não é tão intuitivo quanto ao score de Fisher.

## **Conclusão**

Concluindo, através deste projeto foi-nos possível compreender melhor todos os métodos de redução de dimensionalidade, seleção de features e deteção de outliers, analisando como cada um deles contribui para a simplificação, interpretação e melhoria do desempenho na análise de dados.