UNIVERSIDADE FEDERAL DO PIAUÍ
SISTEMAS DE INFORMAÇÃO
DISCIPLINA: TÓPICOS ESPECIAIS EM COMPUTAÇÃO
DOCENTE: ALAN RAFAEL FERREIRA DOS SANTOS

TRABALHO I

Cristina de Moura Sousa
Daniel Rodrigues de Sousa

Picos-PI, 17 de setembro de 2025

Sumário

1 Introdução

Esta atividade prática tem como objetivo aplicar técnicas de análise de dados em um conjunto de informações de alta dimensionalidade, utilizando a base UCI HAR Dataset, amplamente empregada em estudos de reconhecimento de atividades humanas. O desenvolvimento ocorre na plataforma Google Colab, que integra código, visualizações e documentação. O processo envolve a preparação e o pré-processamento dos dados, incluindo normalização e separação entre atributos e rótulos, garantindo a qualidade das análises subsequentes.

Em seguida, são aplicadas técnicas de redução de dimensionalidade e seleção de características, com foco na Análise de Componentes Principais (PCA), Recursive Feature Elimination (RFE) e Lasso. Cada método é avaliado quanto à capacidade de simplificar o conjunto de dados sem comprometer a representatividade das informações. As visualizações permitem identificar padrões e tendências entre as atividades humanas registradas, contribuindo para a compreensão dos impactos de cada abordagem.

2 Desenvolvimento

Nesta seção são descritas as etapas de preparação e pré-processamento dos dados, a aplicação do PCA para redução de dimensionalidade e os métodos de seleção de características RFE e Lasso. Também são apresentadas visualizações como gráficos de dispersão, variância explicada, desempenho do modelo e coeficientes resultantes. Por fim, realiza-se a análise de padrões e tendências observados no conjunto de dados.

2.1 Preparação dos Dados

Nesta seção, realiza-se o download do dataset UCI HAR e o carregamento dos arquivos em Python, incluindo *X_train.txt*, *y_train.txt*, *X_test.txt*, *y_test.txt*, *features.txt* e *activity_labels.txt*. Após o carregamento, os conjuntos de treino e teste são unidos em *X* (atributos) e *y* (rótulos). As colunas de *X* recebem nomes genéricos, enquanto *y* é convertido para rótulos descritivos das atividades (*WALKING*, *WALKING_UPSTAIRS*, *WALKING_DOWNSTAIRS*, *SITTING*, *STANDING*, *LAYING*). Em seguida, identificam-se o número total de amostras, o número de atributos e as classes distintas (Listagem 1) , além da distribuição de amostras por classe (Listagem 2). Por fim, realiza-se a verificação de valores faltantes para assegurar a integridade dos dados.

```
numero_amostras = 10299
numero_atributos = 561
numero_classes = 6
classes = ['WALKING', 'WALKING_UPSTAIRS',
           'WALKING_DOWNSTAIRS',
           'SITTING', 'STANDING', 'LAYING']
```

Listing 1: Identificação do conjunto de dados

```
Shape X_train: (7352, 561)
Shape y_train: (7352, 1)
Shape X_test: (2947, 561)
Shape y_test: (2947, 1)
```

```
Shape features: (561, 2)
Shape activity_labels: (6, 2)
```

Listing 2: Carregamento dos dados

2.2 Pré-processamento

Nesta seção, os dados são organizados em atributos (*X_data*) e rótulos (*y_data*) para posterior análise. Aplica-se a normalização utilizando StandardScaler (Listagem 3), procedimento para métodos como PCA. São exibidas as médias e desvios padrão das variáveis antes da normalização, permitindo observar a dispersão original dos dados. Após a normalização, as médias aproximam-se de zero e os desvios padrão aproximam-se de um, garantindo a padronização. As estatísticas de algumas features são apresentadas para confirmar a aplicação correta do escalonamento. Este processo garante que os dados estejam em escala compatível, evitando vieses. A padronização facilita a aplicação de PCA e a interpretação dos resultados. Os conjuntos normalizados são mantidos organizados para as etapas seguintes.

```
Dados normalizados com StandardScaler
Media antes da normalizacao: -0.5086
Desvio padrao antes da normalizacao: 0.5294
Media apos normalizacao: -0.0000
Desvio padrao apos normalizacao: 1.0000

Estatisticas de algumas features apos normalizacao:
Feature 0: media = -0.000000, std = 1.000000
Feature 1: media = 0.000000, std = 1.000000
Feature 2: media = -0.000000, std = 1.000000
Feature 3: media = -0.000000, std = 1.000000
Feature 4: media = -0.000000, std = 1.000000
```

Listing 3: Normalização dos dados com StandardScaler

2.3 Redução de Dimensionalidade com PCA

Nesta seção, aplica-se o PCA para reduzir os 561 atributos a 2 componentes principais (Listagem 4), permitindo visualizar a distribuição das amostras no gráfico de dispersão 2D da Figura 2, onde cada ponto é colorido de acordo com a classe. Calcula-se a variância explicada por cada componente e a variância total retida pelos dois primeiros. Em seguida, o PCA é aplicado sem restrição de componentes para analisar a contribuição individual de cada um. A variância acumulada é calculada para determinar o número mínimo de componentes necessários para explicar pelo menos 90% da informação (Listagem 5). Os resultados incluem o número de componentes que atingem o limiar de 90% e a variância acumulada correspondente.

```
Variancia explicada pelo PC1: 0.5074 (50.74%)
Variancia explicada pelo PC2: 0.0624 (6.24%)
Variancia total explicada: 0.5698 (56.98%)
```

Listing 4: PCA com 2 Componentes

```
Numero de componentes necessarios para 90% da variancia: 65
Variancia explicada com 65 componentes: 0.9005
```

Listing 5: Análise de Componentes Principais para 90% da Variância

2.4 Seleção de Características

Nesta seção, realiza-se a seleção de características utilizando RFE e Lasso. O RFE aplica um classificador para reduzir progressivamente os atributos e identificar os mais relevantes. O Lasso utiliza regularização L1 para zerar coeficientes irrelevantes e destacar variáveis importantes. Por fim, compara-se o conjunto de atributos selecionados pelos dois métodos para avaliar consistência e relevância.

2.4.1 Recursive Feature Elimination (RFE)

Nesta etapa, aplica-se a técnica de RFE (Recursive Feature Elimination) para selecionar os atributos mais relevantes do conjunto de dados. Os dados são divididos em treino (80%) e teste (20%) com estratificação, garantindo representatividade das classes. São utilizados os classificadores Logistic Regression e Random Forest (versão leve com 20 estimadores (Listagem 7)). Testam-se diferentes números de atributos selecionados: 10, 30 e 50, avaliando o impacto na acurácia do modelo (Listagem 6). As acurácias obtidas são registradas e comparadas em gráficos, permitindo visualizar a relação entre número de features e desempenho. O subconjunto mais relevante de atributos é identificado combinando Random Forest + RFE com 30 features.

```
--- Logistic Regression ---
  10 features  Acuracia = 0.9175
  30 features  Acuracia = 0.9587
  50 features  Acuracia = 0.9709

--- Random Forest ---
  10 features  Acuracia = 0.9447
  30 features  Acuracia = 0.9762
  50 features  Acuracia = 0.9782
```

Listing 6: RFE Simplificado

```
Selecionadas: 30
Primeiras 20: [  9  37  40  41  49  50  52  53  56  65  69  70  73
   74  75 129 139 179 201 202]
```

Listing 7: Features mais relevantes (RandomForest + RFE)

2.4.2 Lasso (L1 Regularization)

Nesta etapa, aplicou-se a regularização L1 (Lasso) por meio de um modelo de Regressão Logística multiclasse. Foram testados diferentes valores de C (0.001, 0.01, 0.1, 1, 10), avaliando-se a acurácia no conjunto de teste e o número de atributos não zerados para cada valor. O melhor

resultado foi obtido com $C = 1$, alcançando acurácia de 0.9869 e selecionando 357 atributos (Listagem 8). Em seguida, analisaram-se os coeficientes do modelo, identificando quais variáveis foram eliminadas pela regularização e listando as primeiras 20 mantidas (Listagem 9). Para verificar a consistência da seleção, comparou-se o conjunto de atributos do Lasso com o obtido pelo RFE (Listagem 10). Foram encontradas 25 variáveis em comum, 5 exclusivas do RFE e 332 exclusivas do Lasso. Essa análise revelou uma sobreposição de 6.91%, indicando que o Lasso retém um número significativamente maior de variáveis.

```
C = 0.001: Acuracia = 0.7922, Features nao-zero = 16
C = 0.01: Acuracia = 0.9456, Features nao-zero = 124
C = 0.1: Acuracia = 0.9791, Features nao-zero = 246
C = 1: Acuracia = 0.9869, Features nao-zero = 357
C = 10: Acuracia = 0.9869, Features nao-zero = 497

Melhor C: 1
Melhor acuracia: 0.9869
Features selecionadas: 357
```

Listing 8: Resultados do LASSO (Regularização L1)

```
Features selecionadas pelo LASSO: 357
Indices das primeiras 20 features: [0, 1, 2, 3, 4, 6, 7, 9, 10, 11,
   12, 13, 14, 16, 17, 18, 19, 21, 22, 23]
```

Listing 9: Features selecionadas pelo LASSO

```
Features selecionadas pelo RFE: 30
Features selecionadas pelo LASSO: 357
Features em comum: 25
Features apenas no RFE: 5
Features apenas no LASSO: 332
Percentual de sobreposicao: 6.91%
```

Listing 10: Comparação entre RFE e LASSO

Na Figura 1, os gráficos comparam os métodos de seleção de atributos RFE e Lasso, destacando diferenças significativas na quantidade e sobreposição de features escolhidas. O diagrama de Venn mostra que o Lasso selecionou 357 atributos, dos quais 332 são exclusivos, enquanto o RFE escolheu apenas 30, com 5 exclusivos. Apenas 25 atributos foram comuns aos dois métodos, indicando baixa concordância entre eles e pode representar um núcleo relevante para o modelo. O gráfico de barras reforça essa diferença, evidenciando o caráter mais seletivo do RFE em contraste com a abordagem mais abrangente do Lasso. Essa divergência decorre dos critérios distintos de cada técnica: o RFE elimina atributos com base em desempenho iterativo, enquanto o Lasso penaliza coeficientes via regularização.

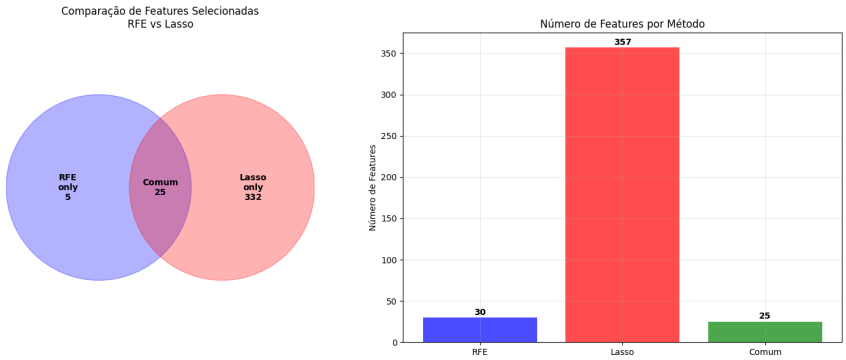

Figure 1: Diagrama de Venn e Gráfico de Barras

2.5 Vizualizações

- **Gráfico de dispersão em 2D com PCA**

No gráfico da Figura 2, a componente principal 1 (PC1) representa 50,7% da variância dos dados, enquanto a componente 2 (PC2) explica 6,2%, totalizando mais de 56% da variabilidade original capturada em apenas duas dimensões. Observa-se que as atividades dinâmicas como *WALKING*, *WALKING_UPSTAIRS* e *WALKING_DOWNSTAIRS* formam agrupamentos relativamente próximos, mas ainda distintos, indicando similaridade nos padrões de movimento. Já as atividades estáticas como *SITTING*, *STANDING* e *LAYING* estão mais concentradas em regiões específicas e bem separadas entre si, evidenciando diferenças claras nas características dos sinais capturados. A concentração das bolinhas em regiões delimitadas do gráfico indica que o PCA conseguiu preservar a estrutura dos dados e destacar os padrões relevantes para cada classe.

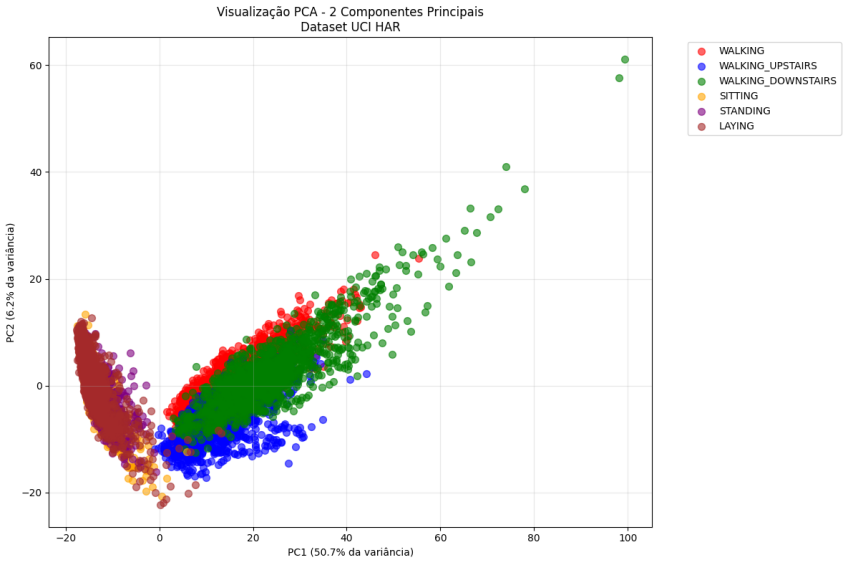

Figure 2: Gráfico Dispersão em 2D com PCA

- **Gráfico da variância explicada acumulada (scree plot)**

Na Figura 3, a esquerda dele apresenta o gráfico da variância explicada por cada um dos primeiros 50 componentes principais após aplicação do PCA. O eixo X representa os com-

ponentes numerados de 0 a 50, enquanto o eixo Y mostra a proporção de variância explicada, que varia de 0 a aproximadamente 0,55. A curva decai rapidamente, indicando que os primeiros componentes concentram a maior parte da informação do conjunto de dados. O primeiro componente, por exemplo, explica mais de 50% da variância, enquanto os seguintes contribuem progressivamente menos. Isso revela que há alta redundância entre os 561 atributos originais, permitindo uma compressão eficiente. A partir do décimo componente, a variância explicada por componente se estabiliza, tornando os demais menos relevantes individualmente.

A direita mostra o gráfico da variância explicada acumulada pelos primeiros 100 componentes. O eixo X representa o número de componentes utilizados, e o eixo Y indica a soma acumulada da variância explicada, variando de 0,5 a 1,0. A curva vermelha crescente demonstra que, à medida que mais componentes são adicionados, mais variância é preservada, embora com ganhos decrescentes. Duas linhas pontilhadas verdes destacam pontos críticos: uma vertical em 65 componentes e outra horizontal em 90% de variância acumulada. A interseção dessas linhas indica que 65 componentes são suficientes para manter 90% da informação original. Juntos, os gráficos evidenciam a eficácia do PCA na redução dimensional com mínima perda de conteúdo relevante.

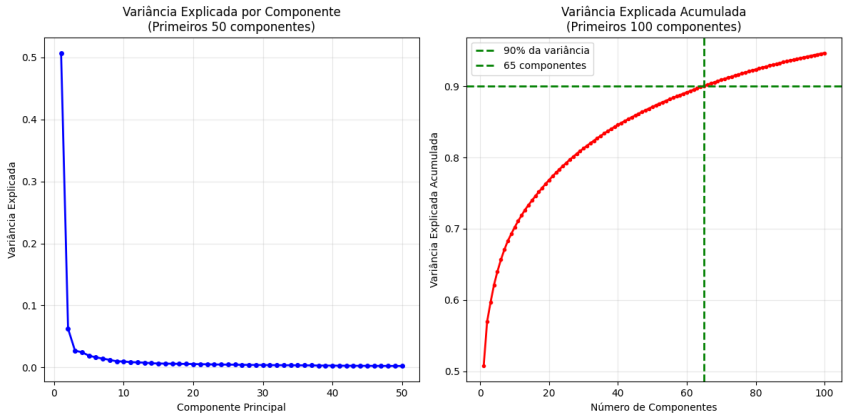

Figure 3: Gráfico das Variâncias

• Gráfico comparando a performance do modelo ao variar o número de atributos no RFE

Foi aplicado o método de seleção de atributos Recursive Feature Elimination (RFE) utilizando os classificadores Logistic Regression e Random Forest. A acurácia de ambos os modelos aumentou com o número de atributos, sendo o Random Forest superior em todas as configurações. Com 50 atributos, o Random Forest atingiu cerca de 98% de acurácia, enquanto o Logistic Regression alcançou aproximadamente 97%. A análise comparativa, no gráfico da Figura 4, evidencia que o Random Forest apresenta melhor desempenho na tarefa de classificação com diferentes subconjuntos de atributos.

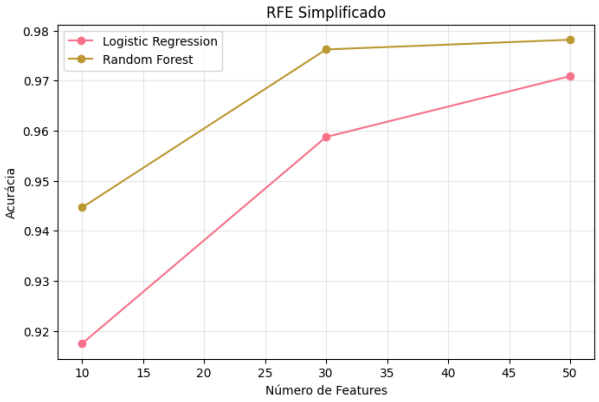

Figure 4: Gráfico RFE Simplificado

- **Visualização dos coeficientes do Lasso (quais atributos permaneceram)**

Na Figura 5, observa-se que o ajuste do parâmetro C influencia diretamente tanto o número de atributos selecionados quanto o desempenho do modelo. Os gráficos evidenciam o impacto da regularização L1 (Lasso) na seleção de atributos e na performance do modelo. No gráfico à esquerda, observa-se que, conforme C aumenta (de 10^{-3} a 10^{1}, em escala logarítmica), o número de atributos com coeficientes diferentes de zero também cresce. Isso indica que valores menores de C impõem maior penalização, resultando em modelos mais esparsos, enquanto valores maiores permitem que mais atributos sejam mantidos. A linha azul com marcadores mostra essa tendência de forma contínua, revelando o controle direto que C exerce sobre a complexidade do modelo.

O gráfico central mostra a acurácia do modelo em função de C, também em escala logarítmica. A linha vermelha revela que a acurácia aumenta com C, estabilizando-se em valores mais altos, o que sugere que modelos menos penalizados conseguem capturar melhor os padrões dos dados. Já o gráfico à direita apresenta os coeficientes não nulos para $C = 1$, distribuídos por índice de atributo e coloridos por classe. Algumas regiões apresentam alta densidade de pontos, mostrando atributos próximos com coeficientes significativos, enquanto outras áreas têm poucos ou nenhum ponto, indicando atributos zerados pelo Lasso. A variação nos sinais dos coeficientes evidencia a direção da influência de cada atributo nas predições por classe.

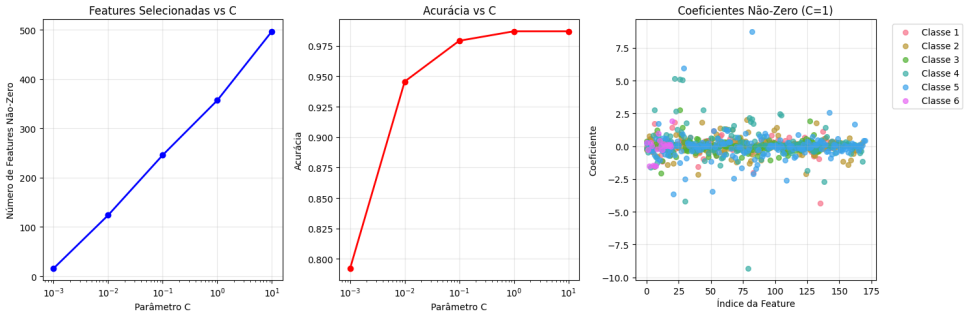

Figure 5: Coeficientes do Lasso (Quais Atributos Permaneceram)

2.6 Análise de Padrões e Tendências

A análise de padrões no conjunto de dados UCI HAR evidencia que o PCA facilita a visualização das atividades, embora com apenas duas componentes seja possível explicar apenas 56,98% da variância total. As atividades estáticas, como SITTING, STANDING e LAYING, tendem a se agrupar, enquanto as dinâmicas, representadas pelas variações de WALKING, formam clusters distintos. No entanto, há sobreposição entre algumas classes, indicando que duas componentes não são suficientes para separação completa. Apesar de reduzir efetivamente a dimensionalidade e preservar a variância máxima, o PCA perde a interpretabilidade dos atributos originais e nem sempre gera componentes discriminativos para todas as classes. Assim, ele é mais adequado para exploração visual do que para análise detalhada de features.

A comparação entre RFE e Lasso mostra baixa concordância na seleção de atributos, com apenas 7,0% de sobreposição. O RFE mantém a interpretabilidade das features originais e avalia a importância de cada atributo para a classificação, sendo flexível quanto ao classificador, mas é computacionalmente mais caro e depende do modelo base. Por outro lado, o Lasso realiza seleção automática de features e evita overfitting com regularização L1, sendo eficiente computacionalmente, embora possa eliminar atributos correlacionados importantes e seja sensível à escala dos dados. No contexto da base HAR, o RFE selecionou 30 features, enquanto o Lasso manteve 357, sendo 25 em comum, mostrando que a escolha do método depende dos objetivos: visualização e redução de dimensionalidade com PCA, ou interpretabilidade e performance de classificação com RFE e Lasso.

3 Conclusão

Este trabalho realizou uma análise técnica e comparativa de métodos de seleção de atributos e redução de dimensionalidade aplicados ao dataset UCI HAR. A aplicação do PCA permitiu reduzir os 561 atributos originais para dois componentes principais, proporcionando uma visualização clara da separação entre classes, e posteriormente definiu-se 65 componentes como ponto ótimo para preservar 90% da variância. Os métodos de seleção RFE e Lasso foram empregados para identificar os atributos mais relevantes, sendo o Lasso mais abrangente e o RFE mais seletivo. A acurácia dos modelos foi avaliada em função do número de atributos e do parâmetro de regularização C, evidenciando que o ajuste adequado desses parâmetros impacta diretamente na performance.

A análise dos coeficientes do Lasso mostrou quais atributos foram mantidos e sua contribuição por classe, reforçando a capacidade do modelo de realizar seleção automática e interpretável. A comparação entre RFE e Lasso revelou baixa interseção entre os atributos selecionados, indicando que cada método prioriza características distintas. A visualização dos dados e dos coeficientes permitiu compreender a estrutura interna do modelo e a importância relativa dos atributos. Conclui-se que a integração entre PCA, RFE e Lasso oferece uma abordagem robusta para otimização de modelos preditivos, reduzindo a dimensionalidade, melhorando o desempenho e facilitando a interpretação dos resultados.

4 Anexos

Link do Repositório do GitHub: TC-data-II