Análise exploratória da influência dos hábitos estudantis na performance acadêmica

1st José Ferreira Lessa
Dept. of Teleinformatics Engineering
Universidade Federal do Ceara
Fortaleza, Brazil
josefflessa@alu.ufc.br

2rd Nataniel Marques Viana Neto
Dept. of Teleinformatics Engineering
Universidade Federal do Ceara
Fortaleza, Brazil
natanielmarques@alu.ufc.br

3nd Matheus Rocha Gomes da Silva
Dept. of Teleinformatics Engineering
Universidade Federal do Ceara
Fortaleza, Brazil
theusrocha2004@alu.ufc.br

4th Victor Guedes Alves Texeira
Dept. of Teleinformatics Engineering
Universidade Federal do Ceara
Fortaleza, Brazil
victorguedes80@alu.ufc.br

***Abstract*—Compreender os fatores que influenciam no sucesso acadêmico é uma etapa essencial na construção de medidas para melhorar a educação e o desempenho dos estudantes. Nesse artigo, será analisado, de forma detalhada, um conjunto de dados sintéticos chamado "Student Habits vs Academic Performance", no qual possui informações sobre hábitos de vida, estudo e tecnológicos de 1000 estudantes. O principal objetivo é identificar padrões e correlações entre variáveis como horas de estudo, uso de redes sociais, horas de sono e saúde mental, e seu impacto na pontuação final nos exames escolares.**

***Index Terms*—desempenho acadêmico, análise exploratória de dados, hábitos de estudantes, dataset, variáveis**

I. Introdução

O desempenho acadêmico é um fator crucial para o sucesso de instituições de ensino e para o sucesso profissional dos estudantes. O estudo dos diversos fatores que influenciam esse sucesso já possui um longo histórico, com diversas conclusões e avanços. Fatores, como por exemplo, horas de sono e horas de estudo já foram comprovados que claramente influenciam positivamente no desempenho.

No entanto, com o aumento da tecnologia e a crescente presença dela na vida dos estudantes, novos fatores que estão diretamente ligados aos estudantes apareceram, como as plataformas de streaming e a ascensão de redes sociais como Instagram e TikTok. Tais fatores, além de muitos outros, acabaram moldando o estilo de vida moderno, prejudicando a saúde mental, a capacidade de concentração e o aprendizado em si.

Sendo assim, este trabalho se propõe explorar essas múltiplas facetas, através de um dataset que combina variáveis comportamentais, de saúde e tecnologias, oferecendo uma oportunidade única para uma análise integrada. O objetivo dessa análise exploratória de dados é desvendar padrões ocultos, identificar as variáveis mais influentes e formular hipóteses sobre as interações que moldam o desempenho dos estudantes nos exames.

O estudo se inicia adquirindo as principais informações do dataset, organizando os dados, de forma a tratar quaisquer imperfeições e deixar o conjunto de dados pronto para uma investigação mais aprofundada. Em seguida, foi feito uma observação completa dos preditores presentes no dataset, realizando análises univariadas, bivariadas e multivariadas, com a implementação da PCA (Principal Component Analysis). Por fim, foi feito uma discussão sobre as conclusões e implicações dos achados.

II. Metodologia

A. Análise Inicial

Visualizando inicialmente o dataset, foi visto que ele consiste em 1000 observações, cada uma representando um estudante único. Além disso, o dataset possui 16 variáveis. A variável alvo desse estudo é `exam_score`, que é uma medida quantitativa do desempenho do estudante no teste. As 14 variáveis preditores, retirando a variável identificadora (`student_id`), são:

- **Demográficas:** `age, gender`
- **Hábitos de Estudo e Lazer:** `study_hours_per_day, social_media_hours, netflix_hours`
- **Comportamentais:** `part_time_job, attendance_percentage, extracurricular_participation`
- **Saúde e Bem-estar:** `sleep_hours, diet_quality, exercise_frequency, mental_health_rating`
- **Contextuais:** `parental_education_level, internet_quality`

Além disso, considerando que o número de observações em um dataset e a quantidade de variáveis preditoras são denominados, respectivamente, por N e D, neste caso temos N = 1000 e D = 14.

Em seguida, como forma de organizar melhor as variáveis preditoras, foi feito um dicionário de dados, classificando as variáveis de acordo com suas características em quantitativas (discretas ou contínuas) ou qualitativas (ordinais ou nominais).

Ao observar os dados, também foi constatado que a variável preditora `parental_education_level` possuia valores nulos em 91 linhas. Logo, como forma de tratar esses dados faltantes, foi feito preenchimento desses espaços nulos pela moda da variável preditora. A moda consiste basicamente no resultado que mais se repetiu para aquela determinada variável no dataset. No caso da variável `parental_education_level`, a moda era 'High School'.

Por conseguinte, como forma de garantir a limpeza dos dados, foi feito uma verificação de linhas duplicadas. No entanto, o dataset não possui linhas repetidas e todos os dados estavam dispostos corretamente.

Depois de ter organizado o dataset, realizamos a discretização da variável alvo, `exam_score`, que é contínua. Dessa forma, fica mais fácil investigar como a distribuição de cada variável preditora se comporta dentro de cada grupo de desempenho. Com isso, criamos 4 classes de desempenho, inspiradas no sistema de notas utilizado na Universidade Federal do Ceará (UFC), sendo divididas da seguinte forma:

- **Reprovado:** `exam_score` < 40
- **Recuperação:** $40 \leq$ `exam_score` < 70
- **Bom:** $70 \leq$ `exam_score` < 90
- **Excelente:** `exam_score` ≥ 90

Considerando que o número de classes da variável-alvo é denominado L, nesse caso, temos que L = 4.

Em seguida, foi feito uma visualização dessa distribuição de classes, como se pode ver abaixo:

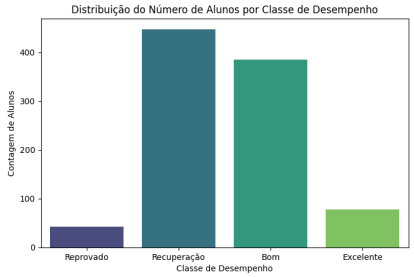

Fig. 1: Distribuição das classes de desempenho

Com base no gráfico acima, podemos ver que a maioria dos alunos possui o desempenho 'Recuperação'. Além disso, percebe-se que a quantidade de alunos com desempenho 'Recuperação' ou 'Reprovado' é minimamente maior que a quantidade de alunos com desempenho 'Excelente' ou 'Bom', sendo assim, uma informação importante, e que, deve ser levada em conta posteriormente.

Por fim, encerrando essa etapa inicial da análise dos dados, através da função `.describe()`, conseguimos obter diversas estatísticas descritivas de todas as variáveis preditoras do dataset, como média, mediana, valor máximo e mínimo, dentre outros. Através disso, foi possível obter informações interessantes, como, por exemplo, que há mais estudantes do gênero feminino no dataset (com outliers) e que a média de horas de estudo por dia é, de aproximadamente, 3,55 horas. Tais informações serão analisadas de forma mais aprofundada em seguida.

B. Análise Univariada

A análise univariada tem como objetivo compreender o comportamento de cada variável presente em um dataset, permitindo identificar distribuições, assimetrias, presença de outliers e padrões iniciais relevantes para análises importantes que podem implementar melhorias, estudar efeitos, saber os comportamentos e muitas outras coisas de fenômenos do nosso mundo. Fizemos duas formas de análise univariada, uma para variáveis quantitativas e outra para as qualitativas, tanto em sua forma incondicional quanto condicional às classes da variável alvo, `exam_score`.

1) Variáveis Quantitativas (Análise Incondicional): Inicialmente, foram gerados histogramas com curvas de densidade e boxplots para cada uma das 8 variáveis quantitativas. Com isso, foi possível identificar a forma da distribuição (simétrica, assimétrica ou multimodal) com os histogramas e também suas concentrações e limites com os boxplots.

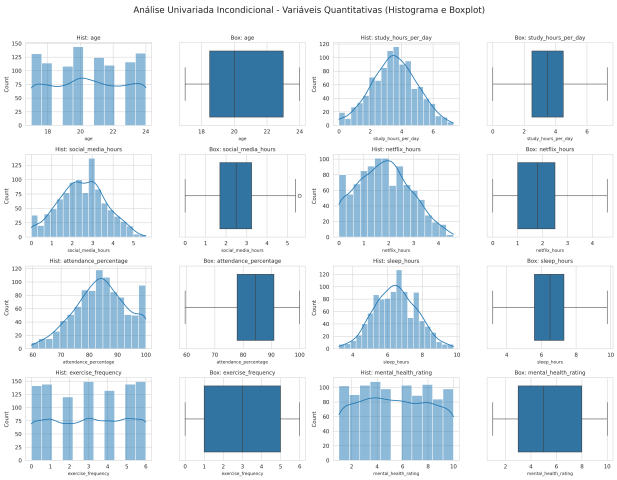

Fig. 2: Gráficos da análise univariada incondicional quantitativa.

Também foram calculadas estatísticas descritivas dessas variáveis, como média, mediana, desvio padrão e assimetria (skewness), possibilitando uma compreensão ainda maior e mais detalhada do comportamento delas, fizemos essas estatísticas após a remoção dos outliers.

TABLE I: Estatísticas Descritivas para Variáveis Quantitativas

Variável	Contagem	Mínimo	Máximo	Média	Mediana	Desvio Padrão	Assimetria
age	979.0	17.0	24.0	20.501532	20.0	2.309567	0.004524
study_hours_per_day	979.0	0.0	7.3	3.520327	3.5	1.439092	-0.053408
social_media_hours	979.0	0.0	5.6	2.483146	2.5	1.149425	0.000016
netflix_hours	979.0	0.0	4.6	1.803984	1.8	1.057254	0.168719
attendance_percentage	979.0	59.5	100.0	84.127681	84.3	9.314441	-0.187444
sleep_hours	979.0	3.2	9.8	6.463841	6.5	1.220036	0.055033
exercise_frequency	979.0	0.0	6.0	3.036772	3.0	2.025572	-0.026471
mental_health_rating	979.0	1.0	10.0	5.436159	5.0	2.853757	0.037740

Com os gráficos e esses resultados, observou-se que, incondicionalmente, determinadas variáveis apresentaram assimetria positiva, evidenciando concentrações em faixas in-

feriores (ex: *netflix_hours*) e outras com assimetria negativa (ex: *attendance_percentage*). Em outras, a distribuição se mostrou simétrica, evidenciando uma maior homogeneidade entre os estudantes (ex: *social_media_hours*). Percebemos uma concentração entre 2 e 4 horas de estudo da variável *study_hours_per_day*, uma concentração entre 6 e 8 horas de sono da variável *sleep_hours*. Em relação às variáveis *exercise_frequency*, *mental_health_rating* e *age* vemos uma frequência mais uniformizada em todos os valores, mostrando que os dados se encontram mais distribuídos e não apresentam essa forma de ”montanha” como as outras variáveis. Agora condicionalmente

Além disso, a presença de valores extremos (outliers) foi identificada principalmente em variáveis relacionadas a hábitos comportamentais, mas foram previamente tratadas com o método de Tukey (usando intervalo interquartil) com funções advindas do pré-processamento, resíduos desses, ocorrem pelo recálculo do intervalo interquartil.

2) Variáveis Qualitativas (Análise Incondicional): Para as variáveis qualitativas, foram gerados gráficos de barras que continham a frequência absoluta de cada categoria observada ao menos uma vez. Essa visualização permitiu identificar algumas predominâncias das nossas variáveis categóricas.

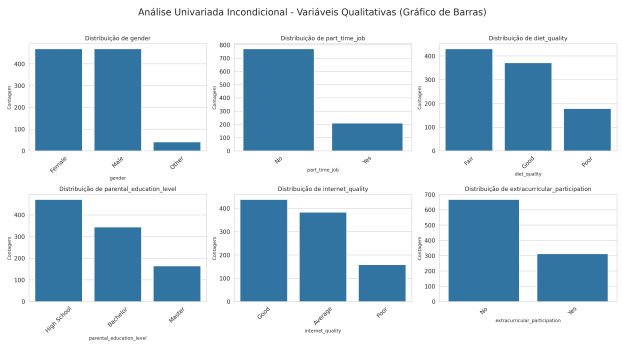

Fig. 3: Gráficos da análise univariada incondicional qualitativa.

Variáveis como *diet_quality* e *internet_quality*, se mostraram mais equilibradas, apesar de ainda sim, terem tendências como a maioria tendo uma dieta ’Fair’ e uma internet ’Good’, evidenciando diversidade nos perfis estudantis. Também observamos distribuições mais desequilibradas como *part_time_job*, com a maioria não tendo esse emprego de meio período e também na variável *extracurricular_participation*, com a maioria não participando de atividades extracurriculares, evidenciando um foco maior dos estudantes apenas com os estudos. Percebemos também uma igualdade entre pessoas do gênero feminino e masculino (após a remoção dos outliers), com uma baixa representatividade de pessoas com outro gênero. Além disso a variável *parental_education_level* mostra que os alunos tendem a ter pais com somente o ensino médio concluído. Por fim, fizemos uma descrição estatísticas dessas frequências no Notebook da análise univariada, com dados mais numéricos, mas os gráficos acima representam bem as informações que descobrimos.

3) Análise Condicional por Classe de Desempenho (Quantitativas): Em seguida, fizemos a análise de forma condicional à variável-alvo (`exam_score`) discretizada em classes de desempenho (`performance_class`), assim como falamos anteriormente, permitindo avaliar como cada variável se distribui em diferentes faixas de performance (Reprovado, Recuperação, Bom e Excelente). Começamos realizando uma descrição estatística para as variáveis de forma condicional, ou seja, em relação às classes de desempenho.

TABLE II: Estatísticas Descritivas Condicionais - Classe Reprovado

Variável	Contagem	Média	Desvio Padrão	Assimetria
age	42.0	20.55	2.14	-0.11
study_hours_per_day	42.0	1.03	0.74	0.36
social_media_hours	42.0	3.02	0.88	-0.33
netflix_hours	42.0	2.27	0.95	0.37
attendance_percentage	42.0	84.20	9.01	-0.27
sleep_hours	42.0	6.29	1.32	-0.11
exercise_frequency	42.0	2.52	1.85	0.02
mental_health_rating	42.0	3.64	2.08	0.78

TABLE III: Estatísticas Descritivas Condicionais - Classe Recuperação

Variável	Contagem	Média	Desvio Padrão	Assimetria
age	442.0	20.51	2.32	0.00
study_hours_per_day	442.0	2.73	1.04	-0.13
social_media_hours	442.0	2.62	1.15	-0.08
netflix_hours	442.0	1.95	1.06	0.04
attendance_percentage	442.0	83.18	9.37	-0.18
sleep_hours	442.0	6.32	1.24	0.11
exercise_frequency	442.0	2.85	2.01	0.08
mental_health_rating	442.0	4.70	2.74	0.35

TABLE IV: Estatísticas Descritivas Condicionais - Classe Bom

Variável	Contagem	Média	Desvio Padrão	Assimetria
age	376.0	20.44	2.27	0.04
study_hours_per_day	376.0	4.09	0.90	-0.04
social_media_hours	376.0	2.34	1.15	0.20
netflix_hours	376.0	1.67	1.02	0.30
attendance_percentage	376.0	84.43	9.17	-0.15
sleep_hours	376.0	6.56	1.17	0.10
exercise_frequency	376.0	3.07	2.03	-0.01
mental_health_rating	376.0	6.00	2.79	-0.19

TABLE V: Estatísticas Descritivas Condicionais - Classe Excelente

Variável	Contagem	Média	Desvio Padrão	Assimetria
age	119.0	20.64	2.48	-0.09
study_hours_per_day	119.0	5.52	0.90	-0.37
social_media_hours	119.0	2.24	1.12	-0.17
netflix_hours	119.0	1.53	1.09	0.34
attendance_percentage	119.0	86.66	9.24	-0.33
sleep_hours	119.0	6.75	1.19	-0.06
exercise_frequency	119.0	3.81	1.97	-0.58
mental_health_rating	119.0	7.01	2.56	-0.73

Além dessas tabelas, realizamos a plotagens de vários gráficos. Para variáveis quantitativas, foram utilizados boxplots e histogramas, possibilitando identificar bem as variações de mediana e amplitudes interquartílicas assim como fizemos na análise incondicional.

Notou-se, como destaque dos nossos gráficos, que estudantes classificados como *Excelente* tendem a apresentar

maiores valores em *study_hours_per_day*, ao passo que os grupos *Reprovado* e *Recuperação* têm valores menores nessa variável.

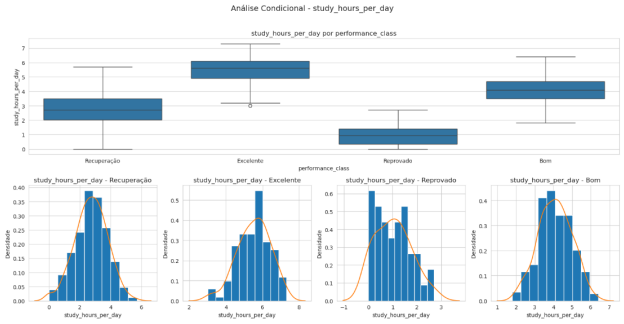

Fig. 4: Boxplot e Histograma de *study_hours_per_day* por classe de desempenho (outros gráficos se encontram no Notebook da análise univariada)

Observando as tabelas e os gráficos (todos eles estão comentados e apresentados no Notebook da análise univariada, devido a grande quantidade deles foi decidido não mostrá-los neste artigo), notamos que os estudantes dos grupos *Reprovado* e *Recuperação* possuem, no geral, uma maior mediana e média de horas em redes sociais e na Netflix, e uma menor mediana e média nas variáveis que falam sobre saúde mental, frequência de exercícios, horas dormidas e frequência nas aulas. Enquanto os estudantes dos grupos *Bom* e *Excelente* apresentam um perfil oposto. A variável que fala sobre a idade apresentou uma distribuição mais homogênea mostrando que idade não é um fator tão importante na performance dos estudantes (mas não significa que certas tendências não ocorram), os outros gráficos se encontram no Notebook da análise univariada.

4) Análise Condicional por Classe de Desempenho (Qualitativas): Para as variáveis qualitativas, foram utilizados gráficos de proporção (countplots), permitindo visualizar a predominância de determinadas categorias em cada classe de desempenho, assim como fizemos na análise incondicional com os barplots, só que dessa vez temos a performance no eixo x ao invés dos valores.

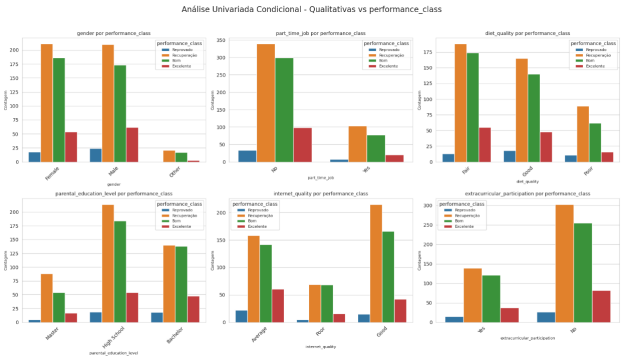

Fig. 5: Gráficos da análise univariada condicional qualitativa.

Analisando os gráficos abaixo, vemos que o gênero não é um fator de muita importância no desempenho. A variável de emprego de meio período tem mais estudantes bons e excelentes para aqueles que não têm esse emprego do que os que têm, mostrando que o emprego pode trazer um certo desfoque dos estudos. Na qualidade da dieta temos uma proporção maior de reprovados no grupo de pessoas com qualidade ruim do que nos outros, possivelmente indicando que uma dieta ruim tende a uma performance pior. No nível educacional dos pais, podemos perceber que os alunos que têm pais com bacharelado possuem uma proporção maior de alunos bons e excelentes quando comparado aos que têm mestrado e ensino médio. A qualidade da internet não aparece afetar a performance quanto as outras variáveis. E por fim participar de atividades extracurriculares parece estar associado a um desempenho melhor.

C. Análise bi-variada

A análise bivariada tem como objetivo explorar e compreender as relações entre pares de variáveis em um dataset, permitindo identificar associações, padrões conjuntos, tendências e possíveis dependências. Essa abordagem é importante para detectar interações entre variáveis, avaliar como mudanças em uma variável se refletem em outra e gerar insights relevantes para modelagem, interpretação de dados e tomada de decisão.

No presente trabalho, realizamos a análise bivariada considerando dois tipos de variáveis:

- Quantitativas: foram gerados gráficos de dispersão (*scatter plots*) entre todas as variáveis numéricas, coloridos de acordo com a variável categórica `performance_class`, permitindo visualizar como os valores de `exam_score` e demais métricas se distribuem entre as classes de desempenho. Além disso, geramos um *heatmap* de correlação global, contemplando todas as variáveis quantitativas.
- Qualitativas: avaliamos a associação entre categorias por meio de *heatmaps* de tabelas cruzadas (*crosstabs*) e *histogramas* de frequência. Os pares que incluíam a variável `performance_class` foram plotados separadamente nos histogramas, destacando a relação de cada variável qualitativa com o desempenho dos alunos.

Dessa forma, conseguimos obter uma visão detalhada das interações presentes no dataset, identificando padrões conjuntos, dependências e possíveis efeitos de classes de desempenho sobre variáveis quantitativas e qualitativas.

1) Variáveis Quantitativas: Através do heatmap acima, podemos ver que as relações entre as variáveis preditoras são absurdamente baixas (a mais forte entre elas tem um percentual de míseros 5%), o que pode indicar que a remoção de variáveis no futuro é inviável, já que cada uma delas é quase que totalmente independente das demais. Da mesma forma, para a variável alvo (`exam_score`) essa falta de relação não muda tanto, exceto para o preditor `study_hours_per_day`, que tem uma correlação de 82% com a variável alvo, seguida por `mental_health_rating`, com relação de 32% com `exam_score`, indicando que a percepção da própria saúde

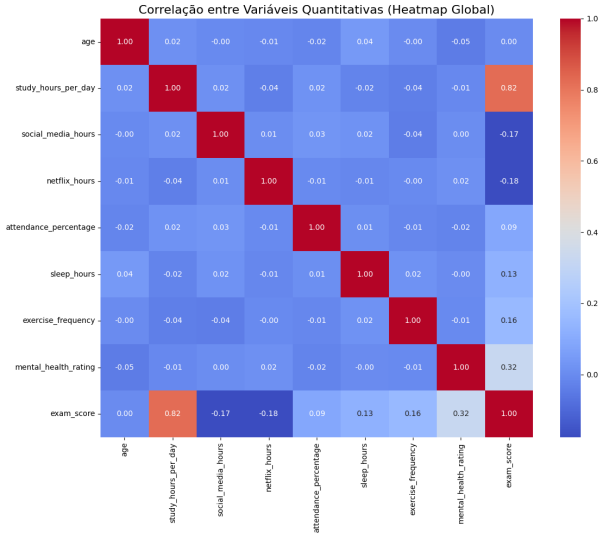

Fig. 6: Gráfico de Análise Bivariada Quantitativa

mental também tem uma relação (apesar de bem mais fraca) com o desempenho acadêmico.

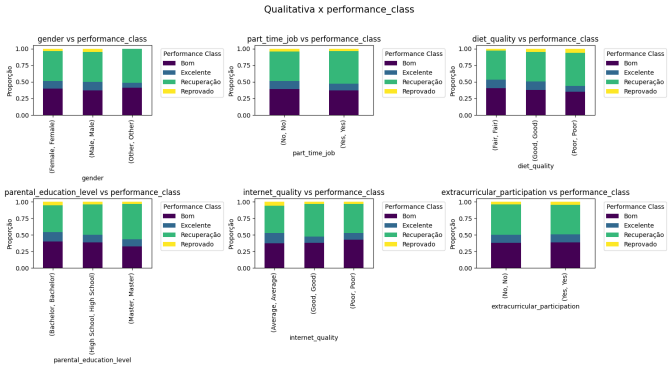

Fig. 7: Gráfico de Análise Bivariada Qualitativa

2) Variáveis Qualitativas: Através desses histogramas relacionando a frequência de cada classe de desempenho com os preditores qualitativos, podemos ver que nenhum apresentou uma frequência muito diferente entre si (alguns apresentam maior ou menor proporção de determinada classe, mas nada muito expressivo ou diferente dos demais). Através disso, podemos ver que a relação entre desempenho e essas variáveis também é tão baixa quanto a vista com o heatmap anterior.

D. Análise Multivariada/PCA

O principal objetivo da análise multivariada é examinar simultaneamente múltiplas variáveis pertencentes a um mesmo conjunto de dados, de modo a proporcionar uma compreensão mais abrangente e realista de fenômenos complexos.

Para essa análise, foram aplicados os conceitos de *Principal Component Analysis* (PCA), que permite reduzir a dimensionalidade do conjunto de dados ao transformar os preditores em *componentes principais*. Esses componentes são combinações lineares das variáveis originais, ortogonais entre si (isto é, independentes) e preservam a maior parte da variabilidade presente nos dados, permitindo representar a informação original de forma compacta e eficiente.

Antes de aplicar efetivamente os passos do PCA realizamos uma etapa de pré-processamento. Nessa etapa, as variáveis categóricas foram transformadas em *dummy variables*, que criam uma variável binária para cada valor possível, com cada dummy assumindo 1 quando o dado corresponde ao valor e 0 caso contrário. Essa transformação é fundamental para que técnicas como o PCA possam ser aplicadas, pois o PCA requer variáveis numéricas contínuas. Com as *dummy variables* determinadas via `get_dummies`, todas as informações das categorias são preservadas e incorporadas ao conjunto de dados de forma adequada para a análise multivariada. Em seguida, os dados foram centralizados (subtraindo-se a média) e escalonados (dividindo-se pelo desvio padrão), de forma a padronizar as variáveis. Esse procedimento garante que todas as variáveis contribuam de maneira equitativa para a análise, evitando que variáveis com magnitudes maiores dominem a determinação dos componentes principais.

Com os dados pré-processados, a matriz de covariância foi calculada utilizando o método `.cov`, enquanto os autovalores e autovetores foram obtidos por meio da função `linalg.eig`. É imprescindível ordenar os autovalores e seus respectivos autovetores de forma decrescente, uma vez que o maior autovalor indica a direção de maior variância nos dados. Dessa forma, ao selecionar os primeiros componentes principais, busca-se capturar a maior parte da variabilidade do conjunto de dados com o menor número possível de dimensões.

Os componentes principais (*principal components - PCs*) são obtidos projetando os dados originais no espaço definido pelos autovetores (*loadings*) da matriz de covariância. Isso é feito multiplicando cada vetor de características (*features*) pelos autovetores correspondentes. Na prática, essa operação transforma as variáveis originais em novas variáveis ortogonais, chamadas scores dos componentes principais, que representam combinações lineares das variáveis originais. Cada componente principal captura uma fração da variabilidade total dos dados, sendo que os primeiros componentes retêm a maior parte da informação, permitindo a redução da dimensionalidade sem perda significativa de informação.

Os dados foram projetados nos dois primeiros componentes principais para facilitar a visualização em duas dimensões, utilizando a coluna `performance_class` como rótulo (*label*). A plotagem resultante apresentou uma nuvem de pontos bastante próxima, indicando que, nesses dois componentes, as classes possuem grande sobreposição e não estão claramente separadas. Isso sugere que a maior parte da variabilidade dos dados (capturada pelos dois primeiros componentes) não está diretamente relacionada com a distinção entre as classes de performance.

O *scree plot* indica que apenas os dois primeiros componentes principais não capturam variabilidade suficiente para representar adequadamente os dados, uma vez que a análise da variância explicada por cada componente mostra que seria

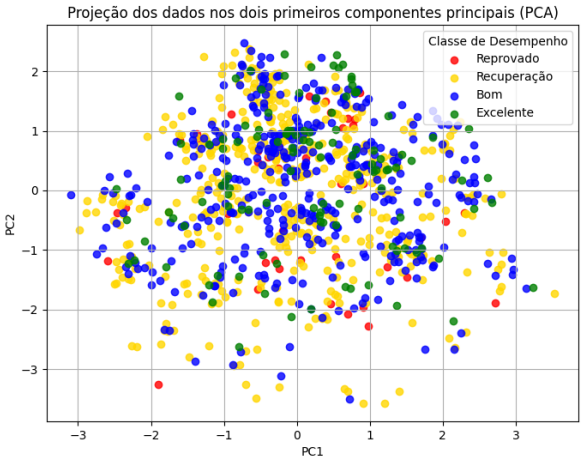

Fig. 8: Gráfico de Dispersão das Observações Projetadas

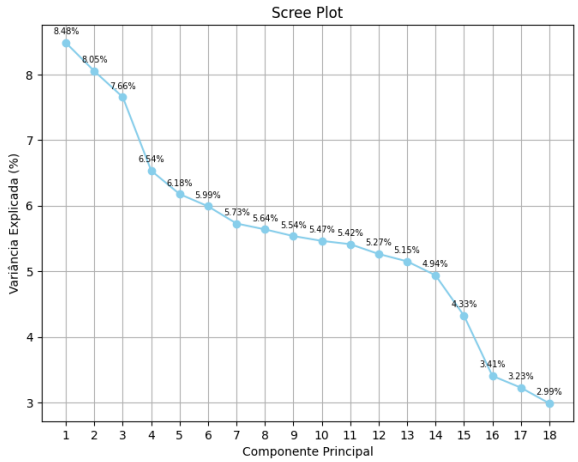

Fig. 9: Gráfico da Variância Explicada por Componente Principal

necessário considerar mais componentes para alcançar uma boa representatividade do conjunto.

III. Resultados e Discussão

A análise exploratória dos dados sobre os hábitos de 1000 estudantes revelou que o desempenho acadêmico, medido pelo *exam_score*, é um fenômeno multifacetado, influenciado por uma combinação de fatores de estudo, comportamentais e de bem-estar.

Os principais achados da etapa univariada indicam que:

- Há variáveis com forte assimetria e dispersão, sugerindo a necessidade de padronização em etapas posteriores.
- Padrões diferenciais entre classes de desempenho emergem, especialmente, de hábitos ruins que influenciam negativamente no desempenho (muitas horas em redes sociais, poucas horas de sono, presença nas aulas baixa, hábitos alimentares e físicos ruins e etc) e hábitos bons que influenciam positivamente no desempenho (muitas horas de estudo, alimentação saudável, exercícios físicos, muitas horas de sono e etc).

A variável com a correlação mais expressiva e positiva com o desempenho acadêmico foi **study_hours_per_day** (horas de estudo por dia), apresentando uma forte correlação de 82%. Isso sugere que a dedicação diária aos estudos é o preditor mais significativo do sucesso nos exames.

Por outro lado, o uso de redes sociais (*social_media_hours*) e o tempo gasto em plataformas de streaming (*netflix_hours*) apresentaram correlações negativas, embora mais fracas, com o desempenho acadêmico. Isso sugere que, embora não sejam os fatores mais determinantes, o tempo excessivo dedicado a essas atividades de lazer pode ter um impacto prejudicial nas notas.

As análises condicionais, que segmentaram os alunos em classes de desempenho (Reprovado, Recuperação, Bom e Excelente), reforçaram essas observações. Estudantes com desempenho ”Excelente” tendem a apresentar, em média, mais horas de estudo, melhor saúde mental e menor tempo em redes sociais e Netflix, em comparação com os grupos de ”Reprovado” e ”Recuperação”. Curiosamente, a análise bivariada mostrou que a maioria das outras variáveis preditoras, como dieta e qualidade da internet, possuem uma relação muito baixa entre si e com a nota final, indicando que são fatores quase independentes.

A Análise de Componentes Principais (PCA) indicou que a variabilidade dos dados é distribuída por múltiplos componentes, sem que os dois primeiros consigam, sozinhos, separar claramente as classes de desempenho. Isso sugere que não existe uma única ”fórmula para o sucesso”, mas sim uma complexa interação entre diversos hábitos que moldam o resultado acadêmico de cada estudante.

Em suma, os resultados apontam para a importância de cultivar uma rotina de estudos consistente e de priorizar a saúde mental. Embora o lazer digital seja parte da vida moderna, o equilíbrio é fundamental para não comprometer o desempenho acadêmico.

Referências

[1] M. R. G. da Silva, et al., “Códigos e notebooks utilizados na análise exploratória da performance acadêmica,” GitHub repository, 2025. [Online]. Available: https://github.com/MRGdSFS/HW1----ICA-2025.2

[2] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[3] BBC News Brasil, ”Por que seu cérebro precisa que você saia de casa e vá para a natureza,” BBC, Oct. 8, 2018. [Online]. Available: https://www.bbc.com/portuguese/vert-fut-45765704.

[4] Ministério da Saúde, ”Como a atividade física protege o cérebro,” Governo do Brasil, Nov. 23, 2022. [Online]. Available: https://www.gov.br/saude/pt-br/assuntos/saude-brasil/eu-quero-me-exercitar/noticias/2022/como-a-atividade-fisica-protege-o-cerebro.

[5] M. Mulas, ”Predictive Modelling process”, apresentação de slides, Universidade Federal do Ceará, Fortaleza, Brazil, 2025.

[6] M. Mulas, ”Data Analysis and pre-processing”, apresentação de slides, Universidade Federal do Ceará, Fortaleza, Brazil, 2025.

[7] M. Mulas, ”Data pre-processing”, apresentação de slides, Universidade Federal do Ceará, Fortaleza, Brazil, 2025.