Relatório Final

Danilo Carrara Sanches Angelim - danilo.carrarac@gmail.com
Diogo Martins Almeida Faria - diogofaria073@gmail.com
Igor Siqueira Souza - pessoal.siqueira@gmail.com
Luiza Jamarino - luijamarino@yahoo.com.br

Primeira Entrega - Relatório da Base de Dados

Este relatório apresenta os resultados da análise exploratória realizada no banco de dados **StudentPerformance.csv**, com o objetivo de compreender o desempenho acadêmico dos alunos e identificar fatores que influenciam nesse desempenho.

Descrição do Banco de Dados

O conjunto de dados utilizado nesta análise contém informações sobre 1.000 estudantes, abrangendo variáveis relacionadas ao desempenho acadêmico e características sociodemográficas.

Variáveis de Desempenho

- → Matemática (math_score): Nota obtida pelos alunos em testes de matemática.
- → Leitura (reading_score): Nota obtida pelos alunos em testes de leitura.
- → Escrita (writing_score): Nota obtida pelos alunos em testes de escrita.

Variáveis Sociodemográficas

- → Gênero (gender): Sexo do aluno (masculino ou feminino).
- → Raça/Etnia (race_ethinicity): Grupo racial ou étnico ao qual o aluno pertence.
- → Nível de Educação dos Pais (parental_level_of_education): Nível de escolaridade dos pais ou responsáveis.
- → Almoço (lunch): Tipo de almoço recebido pelo aluno (gratuito ou padrão).
- → Curso de Preparação para o Teste (test_preparation_course): Indica se o aluno participou ou não de um curso preparatório para os testes.

Variáveis Criadas - Agregadora de Dados

Nota Final - final_score_avg: Calculada como a média aritmética das notas de matemática, leitura e escrita. A fórmula utilizada foi:

final_score_avg = (math_score + reading_score + writing_score) / 3

Análise Exploratória

Exploração Básica (entendimento dos contextos)

Este conjunto de dados contém informações sobre 1000 estudantes, incluindo suas características demográficas e desempenho acadêmico.

Colunas:

- 1. **gênero:** O gênero do estudante (Masculino/Feminino).
- 2. raça_etnia: O grupo racial ou étnico ao qual o estudante pertence.
- 3. **nível_de_escolaridade_dos_pais:** O nível de escolaridade dos pais ou responsáveis.
- 4. **almoço:** O tipo de almoço que o estudante recebe (padrão ou gratuito/reduzido).
- 5. **curso_preparatório:** Se o estudante completou um curso preparatório para o teste (sim ou não).
- 6. **nota_matematica:** A nota do estudante na prova de matemática.
- 7. **nota_leitura:** A nota do estudante na prova de leitura.
- 8. **nota_redação:** A nota do estudante na prova de redação.
- 9. **nota_final_média:** A média das notas do estudante em matemática, leitura e redação.

Informações básicas sobre os dados

O conjunto de dados avaliado contém as pontuações de 1000 alunos em diferentes 3 disciplinas, incluindo matemática, leitura e escrita, além de uma média final dessas pontuações.

	math_score	reading_score	writing_score	final_score_avg
count	1000.00000	1000.000000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000	67.762000
std	15.16308	14.600192	15.195657	14.258354
min	0.00000	17.000000	10.000000	9.000000
25%	57.00000	59.000000	57.750000	58.000000
50%	66.00000	70.000000	69.000000	68.000000
75%	77.00000	79.000000	79.000000	78.000000
max	100.00000	100.000000	100.000000	100.000000

- math_score: Representa que a pontuação média em matemática é de 66,09, com uma variação de 15,16 pontos (desvio padrão). As pontuações variam entre 0 e 100. Observando os quartis, 25% dos alunos marcaram até 57 pontos, 50% alcançaram até 66 pontos (mediana) e 75% chegaram a 77 pontos.
- **reading_score**: A média das pontuações em leitura é de 69,17, com um desvio padrão de 14,60. As pontuações variam de 17 a 100. Os quartis mostram que 25% dos alunos têm pontuação de até 59, a mediana é de 70, e 75% dos alunos alcançam até 79 pontos.
- writing_score: A média para escrita é de 68,05, com um desvio padrão de 15,20. As pontuações variam de 10 a 100. Nos quartis, vemos que 25% dos alunos pontuaram até 57,75, a mediana é de 69, e 75% têm até 79 pontos.
- **final_score_avg**: A média final das pontuações combinadas é de 67,76, com um desvio padrão de 14,26. As pontuações variam de 9 a 100. Nos quartis, 25% dos alunos têm uma média final até 58, a mediana é de 68 e 75% têm uma média até 78.

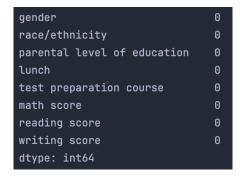
De uma maneira geral, podemos perceber que as pontuações são bem distribuídas ao redor das médias, algumas pontuações mais baixas indicam alunos com possíveis dificuldades, porém a maioria dos alunos obtiveram pontuações médias para alta, o que representa um bom desempenho geral.

Desta forma, torna-se importante entender o que implica nos resultados baixo e alto para que seja possível tomar ações adequadas para melhoria dos alunos com baixa performance.

Segunda Entrega - Relatório da Análise e Treinamento

Pré-processamento

O Banco escolhido, possuía uma boa estrutura e não existiam dados faltantes. A verificação da existência de dados faltantes foi realizada através do comando: df.isna().sum()



Ao verificar os dados de forma mais ampla, foi possível notar que existiam colunas que possuiam caracteres especiais e/ou espaço entre os nomes, o que poderia gerar algum problema ou induzir a um erro durante as etapas posteriores. A imagem a seguir apresenta o formato nativo do banco.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

Afim de resolver o problema, foi construído um método responsável por converter o formato nativo para um formato sem espaços e sem caracteres especiais, tipo: "/".

O que retornou um dataframe com os seguintes padrões de títulos das colunas:

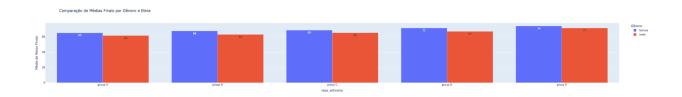
	gender	race_ethnicity	parental_level_of_education	lunch	test_preparation_course	math_score	reading_score	writing_score	final_score_avg
0	female	group B	bachelor's degree	standard	none	72	72	74	73.0
1	female	group C	some college	standard	completed	69	90	88	82.0
2	female	group B	master's degree	standard	none	90	95	93	93.0
3	male	group A	associate's degree	free/reduced	none	47	57	44	49.0
4	male	group C	some college	standard	none	76	78	75	76.0
995	female	group E	master's degree	standard	completed	88	99	95	94.0
996	male	group C	high school	free/reduced	none	62	55	55	57.0
997	female	group C	high school	free/reduced	completed	59	71	65	65.0
998	female	group D	some college	standard	completed	68	78	77	74.0
999	female	group D	some college	free/reduced	none	77	86	86	83.0
1000 rd	1000 rows x 9 columns								

Análise Exploratória

Exploração Básica (entendimento dos contextos - Gráfico)

Entendo a Distribuição dos Dados - Média Final por Gênero e Etnia

A fim de entender o comportamento dos dados, foi realizada uma análise gráfica com o intuito de entender, principalmente as distribuições de notas.



Em geral, as pessoas do gênero feminino apresentaram médias finais superiores em todos os grupos étnicos, com uma diferença que varia de 3 a 5 pontos em comparação com as pessoas do gênero masculino.

O grupo E se destaca por apresentar as maiores médias tanto para pessoas do gênero feminino quanto para pessoas do gênero masculino, enquanto o grupo A apresenta as menores médias para ambos os gêneros. Como conclusão da análise, fica evidente que o gênero feminino tende a ter desempenho superior nas médias finais, independentemente do grupo étnico.

gender	race_ethnicity	final_score_avg		
female	group A	65.083333		
female	group B	67.548077		
female	group C	68.588889		
female	group D	71.441860		
female	group E	74.014493		
male	group A	61.509434		
male	group B	62.988372		
male	group C	65.223022		
male	group D	66.984962		
male	group E	71.450704		

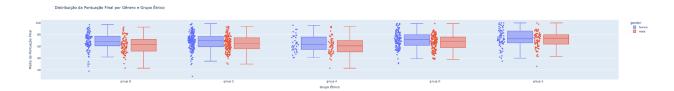
Entendo a Distribuição dos Dados - Notas de Matemática x Média Final por Gênero e Etnia

As notas finais variam amplamente, com a maioria das notas concentradas entre 50 e 80 pontos. Isso sugere que a maioria dos alunos têm desempenho moderado, com poucas pontuações muito baixas ou muito altas. Este comportamento fica evidente através do gráfico apresentado a seguir:

Além disso, analisando o gráfico, é possível observar uma maior frequência de notas de matemática próximas ou iguais a média final, com um ligeiro aumento na faixa de 60 a 80, grupo onde se concentram os alunos com a maioria obtendo notas que acompanham a média final.

Entendo a Distribuição dos Dados - Boxplot por Grupo

O gráfico boxplot, indica que as alunas tendem a apresentar uma maior consistência nas notas finais, com distribuições menos dispersas, enquanto os alunos frequentemente apresentam uma maior variabilidade, com alguns valores extremos tanto altos quanto baixos.

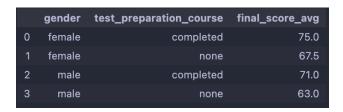


- Grupo A: A mediana das notas finais para ambos os gêneros está próxima, mas as alunas apresentam uma distribuição levemente mais concentrada em torno da mediana, enquanto os alunos possuem uma amplitude de notas um pouco maior. Ambos os gêneros possuem alguns valores atípicos em níveis mais baixos.
- Grupo B: A mediana para as alunas é ligeiramente superior à dos alunos, e as alunas apresentam uma menor variabilidade nas pontuações, o que indica maior consistência nas notas. Os alunos têm uma amplitude maior, com notas mais espalhadas, o que pode indicar uma variabilidade de desempenho.
- **Grupo C**: As alunas também apresentam uma mediana ligeiramente superior e uma distribuição mais compacta em relação aos alunos, que apresentam uma dispersão mais ampla de notas.

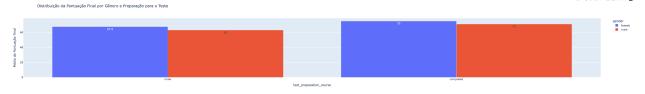
- **Grupo D**: Embora a mediana seja próxima entre os gêneros, as alunas novamente mostram uma distribuição mais compacta, com menor variabilidade. Os alunos apresentam uma maior dispersão, indicando algumas notas mais altas e mais baixas do que as alunas.
- Grupo E: Neste grupo, as alunas têm a maior mediana entre todos os grupos e gêneros, com uma distribuição bastante compacta. Os alunos, embora tenham uma média elevada, apresentam maior variabilidade, com algumas pontuações atípicas em níveis baixos.

Entendo a Pontuação Final com Relação ao Preparo para o Teste

Alunas que completaram o curso preparatório alcançaram uma média de 75.0, enquanto aquelas que não participaram tiveram uma média de 67.5. Para os alunos, a média final daqueles que completaram o curso foi de 71.0, enquanto os que não participaram tiveram uma média de 63.0. Desta forma torna-se evidente que as pessoas que fizeram o curso preparatório obtiveram melhores resultados do que as que não fizeram.

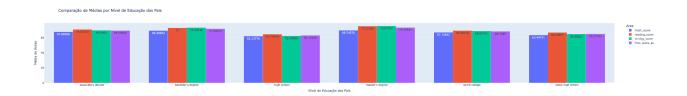


- → Para as alunas, a média aumentou em 7.5 pontos após completar o curso (de 67.5 para 75.0).
- → Para os alunos, a média aumentou em 8.0 pontos (de 63.0 para 71.0) com a conclusão do curso.



Entendendo a Influência do Nível de Conhecimento dos Pais com as Notas dos Alunos

Há uma correlação positiva entre o nível de educação dos pais e o desempenho acadêmico dos alunos, especialmente em leitura e escrita. Os alunos cujos pais possuem nível de educação superior (bacharelado e mestrado) tendem a ter melhores médias finais, o que sugere que um maior nível educacional dos pais pode estar associado a um desempenho acadêmico superior dos filhos. Esses dados podem ser usados para direcionar políticas educacionais e programas de suporte para alunos cujos pais possuem menor nível de escolaridade.



De uma maneira, geral, podemos chegar às seguintes conclusões para cada um dos níveis de educação atingido pelos pais:

- 1. **Associate's Degree** (Grau Associado): Alunos cujos pais têm grau de associado apresentam uma média final de aproximadamente 69.6, com notas equilibradas em todas as disciplinas.
- 2. **Bachelor's Degree** (Graduação): Alunos com pais que possuem graduação apresentam as melhores médias entre os grupos, especialmente em leitura e escrita, o que sugere que o nível educacional superior dos pais pode ter um impacto positivo no desempenho acadêmico.
- High School (Ensino Médio): Esse grupo apresenta as menores médias, indicando que pais com apenas o ensino médio podem influenciar em um desempenho acadêmico mais modesto dos alunos.
- 4. **Master's Degree** (Mestrado): Alunos com pais que possuem mestrado têm um desempenho muito forte, com a média final sendo a mais alta entre os grupos, indicando uma forte influência do nível avançado de escolaridade dos pais.
- 5. **Some College** (Algum Nível Superior): Esse grupo apresenta um desempenho intermediário, próximo ao grupo com grau de associado, o que sugere que algum nível de educação superior pode ter um impacto positivo nas notas dos alunos.
- 6. Some High School (Algum Nível de Ensino Médio): Alunos com pais que completaram apenas parte do ensino médio apresentam médias superiores ao grupo de ensino médio completo, mas ainda inferiores aos demais níveis de escolaridade superior.

Entendo os Outliers

Para entender o comportamento de outliers, foi realizado uma análise com base no seguinte código:

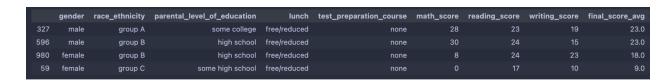
```
from scipy.stats import zscore
from numpy import abs

z = abs(zscore(df['final_score_avg']))

df_outliers = df[z >= 3.0]

df_outliers.sort_values(by='final_score_avg', ascending=False)
```

Este código tem como objetivo identificar e isolar os outliers na coluna final_score_avg do dataframe df (aqueles com Z-score absoluto maior ou igual a 3), e os ordena de forma decrescente. Esses valores representam pontuações finais que se desviam significativamente da média, indicando que podem ser casos atípicos ou anômalos.



Realizando uma análise mais detalhadas dos outliers, podemos chegar em uma análise de agrupamento, as seguintes conclusões:

Gênero e Etnia:

- Entre os quatro outliers, temos uma distribuição equilibrada de gênero (dois homens e duas mulheres).
- A etnia varia entre os grupos A, B e C, mostrando que o baixo desempenho está presente em diferentes grupos étnicos.

Nível de Educação dos Pais:

- O nível de educação dos pais é variado: temos registros com pais que concluíram apenas o ensino médio e outros com "some college" (algum nível de ensino superior).
- Esses dados sugerem que o nível de escolaridade dos pais pode não ser o principal fator isolado influenciando as notas muito baixas.

Tipo de Alimentação (Lunch):

Todos os registros possuem o tipo de alimentação "free/reduced" (gratuita ou reduzida). Este fator
pode indicar um possível aspecto socioeconômico associado ao baixo desempenho, pois alunos
com acesso a alimentação gratuita ou reduzida geralmente pertencem a famílias de baixa renda.

Participação no Curso de Preparação para o Teste:

 Nenhum dos alunos outliers participou do curso de preparação para o teste (test_preparation_course está marcado como "none"). Isso sugere que a falta de preparação formal pode estar contribuindo para o desempenho abaixo da média.

Notas Individuais em Matemática, Leitura e Escrita:

- As notas desses alunos s\u00e3o consistentemente baixas em todas as \u00e1reas (matem\u00e1tica, leitura e escrita), com alguns registros mostrando notas extremamente baixas, como 0 em matem\u00e1tica e apenas 9 de m\u00e9dia final.
- Essas pontuações reforçam que esses alunos podem ter dificuldades generalizadas, afetando todas as disciplinas.

Treinamento do Modelo

A primeira etapa do processo foi a de organizar os dados para serem "inputados" ao modelo. Este processo se inicia através da transformação de variáveis categóricas em variáveis dummy, ou seja, converte as categorias das colunas especificadas (gender, lunch, parental_level_of_education, race_ethnicity e test_preparation_course) em colunas binárias. Cada categoria única dentro dessas colunas é convertida em uma nova coluna com valores 0 ou 1, indicando a presença ou ausência da categoria em cada linha.

```
# Transform features in dummies
df_dummies = pd.get_dummies(df, columns=['gender', 'lunch', 'parental_level_of_education', 'race_ethnicity', 'test_preparation_course'])
```

Separação de Dados - Treino e Teste

```
# Split target and features

X = df_dummies.drop(columns=['final_score_avg'])

y = df_dummies['final_score_avg']
```

Onde o código implementado separa os dados em **features** (variáveis independentes) e **target** (variável dependente) para preparação de um modelo de machine learning.

- X contém todas as colunas do DataFrame df_dummies, exceto final_score_avg, sendo assim as variáveis independentes (ou preditoras).
- y contém apenas a coluna final_score_avg, que é a variável alvo (target ou label) que o modelo de machine learning tentará prever.

Iniciando o Modelo de Regressão Linear

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
lr_predict_result = model.fit(X_train, y_train)
print(f'Coeficientes: {lr_predict_result.coef_}')
```

O código apresentado a seguir, instancia o modelo de Regressão Linear provido pela lib Scikit-learn. Este processo acontece em 3 etapas, onde:

Instanciar o modelo LinearRegression, que será utilizado para ajustar o modelo aos dados.
 O model conterá todos os métodos e atributos necessários para treinar o modelo e realizar

previsões.

- 2. Execução método fit() do objeto model para treinar o modelo de regressão linear. Passamos duas variáveis:
 - X_train: um conjunto de dados de entrada que contém as variáveis independentes (features), que influenciam a variável dependente.
 - y_train: um vetor que contém os valores da variável dependente (target), que estamos tentando prever.
- 3. Exibimos os coeficientes do modelo ajustado utilizando a propriedade coef_. Esses coeficientes são fundamentais, pois indicam a magnitude e a direção (positiva ou negativa) da influência de cada variável independente na variável dependente. Cada coeficiente corresponde a uma variável em X_train, permitindo a interpretação da importância relativa de cada feature no modelo.

Analisando Métricas

 R2 - Score: O R2score calculado foi de 0.9996941579270728, o que fornece uma indicação clara de que o modelo se ajusta muitos aos dados, ou seja, ele mede a adequação do modelo em relação aos dados observados. O que indica que o modelo de Regressão Linear é uma excelente escolha como algoritmo. Esta analise foi realizada através do seguinte comanado:

```
from sklearn.metrics import r2_score

# Calculate the R2 score

r2 = r2_score(y_test, model.predict(X_test))

print(f' Verificando o resultado de R2: {r2}')
```

• **RMSE**: O Resultado obtido foi 0.26266230106771016. A Análise de RMSE, indica sempre que quanto menor o resultado, melhor o modelo.

```
# Raiz do Erro Quadrático Médio (RMSE - Root Mean Squared Error)

from sklearn.metrics import root_mean_squared_error

rmse = root_mean_squared_error(y_test, y_pred=model.predict(X_test))

print("Raiz do Erro Quadrático Médio (RMSE):", rmse)
```

 MSE: O Resultado obtido foi 0.0689914844021844. A Análise de MSE, indica sempre que quanto menor o resultado, melhor o modelo.

```
# Calculo do Errp Quadrático Médio (MSE - Mean Squared Error)

from sklearn.metrics import mean.squared.error
mse = mean_squared_error(y_test, y_pred=model.predict(X_test))

print(f' Verificando o resultado de MSE: {mse}')
```

 MAPE: O Resultado obtido foi 0.3485412701296795. Desta forma, podemos entender que quanto menor o valor, melhor. Um MAPE menor indica que as previsões do modelo estão mais próximas dos valores reais, o que significa que o modelo de Regressão Linar apresenta uma boa assertividade.

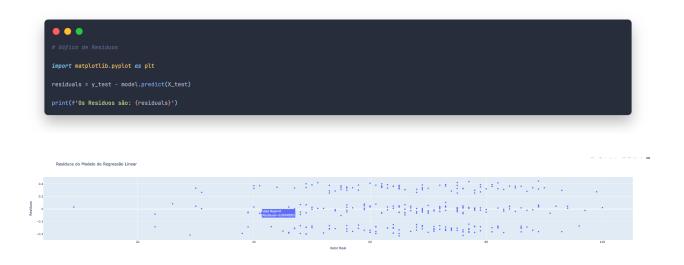
```
# Erro Médio Absoluto (MAFE - Mean Absolute Percentage Error)

def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 188

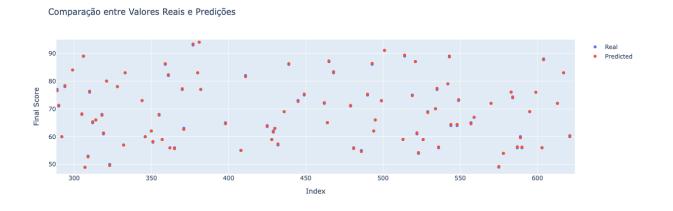
mape = mean_absolute_percentage_error(y_test, y_pred=model.predict(X_test))

print(f'O Resultado foi:', mape)
```

Gráfico de Resíduos: Os resíduos são calculados e representam a diferença entre os valores reais e
os valores previstos. Os resultados apresentados ficam em torno de 0 com algumas variâncias
entre - 0.4 e 0.4 desta forma, em uma análise mais fria, entende-se que quanto mais próximo de 0
melhor. O que sugere que o modelo de regressão é adequado, capturando bem a relação linear
entre as variáveis independentes.

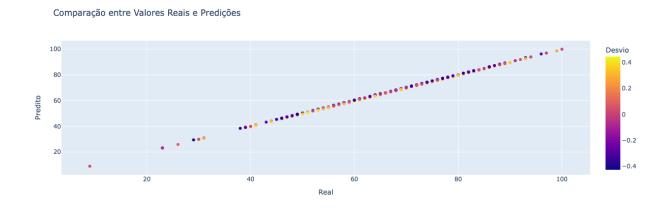


• Gráfico de Aderência - Predito x Real: O Gráfico tem como objetivo tornar visual as previsões do modelo em comparação com o seu valor real. Desta forma é possível identificar que para a maioria dos casos, os valores previstos estão muito próximos dos valores reais, indicando que o modelo é, em geral, preciso. Por exemplo, para o primeiro conjunto de dados, o valor real é 87.0, e o valor previsto é 87.021805, uma diferença de apenas 0.021805.



O Código utilizado para gerar o gráfico pode ser visualizado a seguir:

Além do gráfico de comparação, foi criado um gráfico para representar o desvio gerado para cada uma das previsões, o que deixa ainda mais claro que os erros acontecem de forma bem esporádica com erros bem pequenos.



O Código utilizado para gerar o gráfico foi:

```
df_predicted_results = pd.DataFrame({'Real': y_test, 'Predicted': y_pred, 'Desvio': y_test -
y_pred})

fig = px.scatter(df_predicted_results, x='Real', y='Predicted', color='Desvio', title='
Comparação entre Valores Reais e Predições', labels={'Real': 'Real', 'Predicted': 'Predito',
    'Desvio': 'Desvio'})

fig.show()
```

Considerações Finais

A análise exploratória realizada sobre o banco de dados *StudentPerformance.csv* proporcionou insights valiosos acerca do desempenho acadêmico dos estudantes, evidenciando a influência de variáveis sociodemográficas e de preparação para testes. Os dados revelaram que, em geral, as notas dos alunos estão distribuídas de forma que a maioria apresenta desempenho médio a alto, embora algumas discrepâncias notáveis indiquem áreas que requerem atenção.

Observamos que tanto o gênero quanto a etnia impactam as médias finais dos alunos, com o gênero feminino demonstrando resultados consistentemente superiores. Além disso, a análise das notas em matemática, leitura e escrita revelou que a maioria dos alunos se concentra em uma faixa de desempenho moderado. Os resultados também destacaram a importância do nível educacional dos pais, que se correlaciona positivamente com o desempenho acadêmico dos filhos.

O objetivo do projeto é prever uma variável contínua: a nota final média dos alunos. Nesse contexto, a escolha da regressão linear se mostra adequada, uma vez que esse modelo permite a análise da relação entre variáveis independentes (características dos alunos) e uma variável dependente contínua (nota final).

Os resultados obtidos nas previsões, quando comparados com os dados originais, são bastante satisfatórios, apresentando um baixo desvio nas respostas do modelo.

Membros do Grupo

- → Diogo Martins Almeida Faria diogofaria073@gmail.com
- → Igor Siqueira Souza pessoal.siqueira@gmail.com