1 Metodologia

O SPAECE (Sistema Permanente de Avaliação da Educação Básica do Ceará) é uma iniciativa do Governo do Estado do Ceará, através da Secretaria da Educação (SEDUC), para avaliar o desempenho dos estudantes das escolas públicas estaduais. O objetivo do SPAECE é monitorar a qualidade da educação e fornecer informações que possam subsidiar políticas públicas e estratégias pedagógicas para a melhoria do ensino. O SPAECE avalia estudantes de diversas etapas da educação básica, incluindo os anos finais do Ensino Fundamental (5° e 9° anos) e o 3° ano do Ensino Médio. As avaliações focam principalmente nas áreas de Língua Portuguesa e Matemática, buscando verificar as competências e habilidades dos alunos nessas disciplinas.

Bibliografia (https://repositorio.ufc.br/handle/riufc/39903)
(https://ojs.unifor.br/rh/article/view/6036)

Na edição de 2018 do SPAECE, os resultados para os alunos do 3° ano do Ensino Médio foram analisados considerando os níveis de proficiência em Língua Portuguesa e Matemática. Os dados são apresentados em níveis de desempenho que vão de "Muito Crítico" a "Desejável".

- **Língua Portuguesa**:
    - **Desempenho Geral**: A maioria dos alunos se encontrava nos níveis "Crítico" e "Intermediário", com uma porcentagem menor atingindo os níveis mais altos de proficiência ("Adequado" e "Desejável").
    - **Distribuição de Níveis**: Uma pequena fração dos alunos atingiu o nível "Desejável", que representa uma proficiência adequada para a conclusão do Ensino Médio e preparação para o ensino superior ou mercado de trabalho.
- **Matemática**:
    - **Desempenho Geral**: Similar à Língua Portuguesa, a maioria dos estudantes ficou nos níveis "Crítico" e "Intermediário".
    - **Distribuição de Níveis**: Poucos alunos alcançaram os níveis "Adequado" e "Desejável", indicando que a Matemática é uma área com maiores desafios para os alunos do 3° ano do Ensino Médio.
-

1.1 Proposta

Os resultados do SPAECE 2018 apontam para a necessidade de intervenções mais robustas e direcionadas no ensino médio, especialmente nas áreas de Língua Portuguesa e Matemática. Este experimento tem como objetivo aplicar técnicas de mineração de dados e Inteligência Artificial para correlacionar os dados coletados e encontrar padrões de associações, além de observar o desempenho de diferentes algoritmos e abordagens. Os resultados obtidos podem auxiliar as escolas e a Secretaria da Educação a desenvolver ações específicas para melhorar a qualidade de ensino, como:

- Formação Continuada de Professores: Investir na capacitação dos professores para melhorar as práticas pedagógicas.
- Programas de Reforço Escolar: Implementar programas de reforço e recuperação para alunos com dificuldades de aprendizagem.
- Apoio e Monitoramento: Fortalecer o acompanhamento pedagógico e psicológico dos estudantes para identificar e intervir precocemente em dificuldades.
- Incentivo ao Uso de Tecnologias: Utilizar ferramentas tecnológicas e metodologias inovadoras que possam tornar o aprendizado mais atraente e eficiente.

A abordagem do projeto pode ser simplificada no diagrama abaixo e melhor detalhada nos tópicos posteriores.

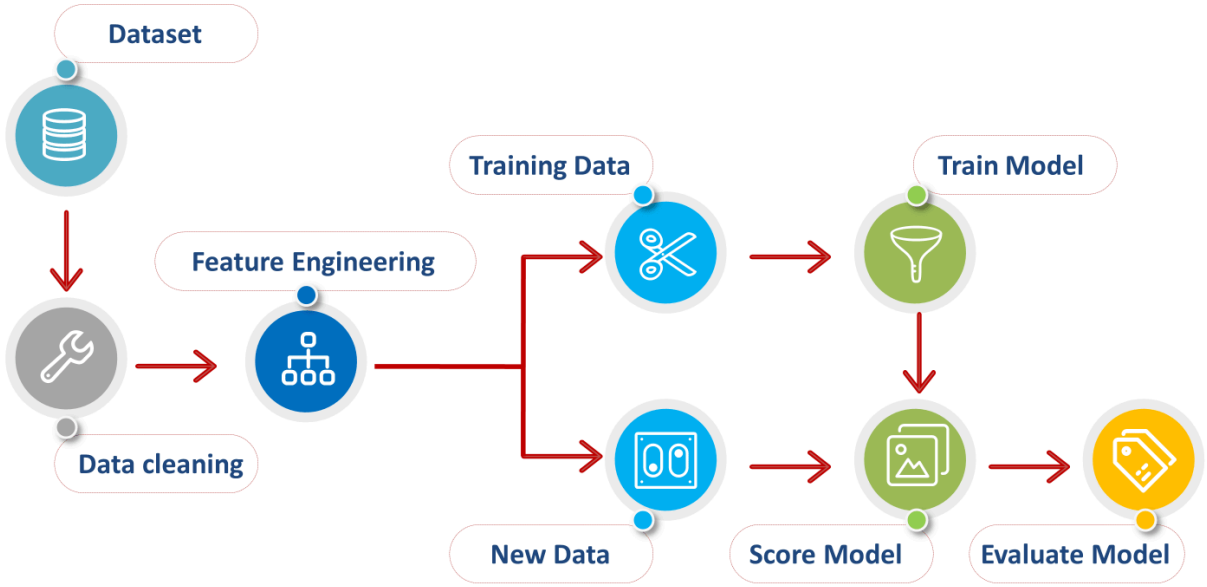

Diagrama do Processo de Aprendizagem de Máquina
Fonte: https://medium.datadriveninvestor.com/data-preprocessing-3cd01eefd438

O diagrama acima representa o fluxo de um processo de aprendizado de máquina aplicado ao contexto educacional descrito. A seguir, uma descrição dos principais passos do processo:

1. **Dataset**: Coleta de dados dos alunos, incluindo resultados de avaliações, informações socioeconômicas, entre outros.
2. **Data Cleaning**: Limpeza dos dados para remover inconsistências e valores ausentes.
3. **Feature Engineering**: Seleção e transformação das variáveis que serão usadas no modelo.
4. **Training Data**: Divisão dos dados em conjuntos de treino e teste.
5. **Train Model**: Treinamento do modelo utilizando algoritmos de aprendizado de máquina.
6. **Score Model**: Avaliação do desempenho do modelo nos dados de teste.
7. **Evaluate Model**: Avaliação final do modelo, considerando métricas de desempenho e validação.

A utilização de técnicas avançadas de análise de dados permitirá identificar padrões e informações valiosas que podem orientar intervenções mais eficazes e direcionadas para melhorar a educação no estado do Ceará.

1.1.1 Dataset

O SPAECE (Sistema Permanente de Avaliação da Educação Básica do Ceará) coleta uma ampla gama de dados por meio de seus questionários, com o objetivo de entender melhor os fatores que influenciam o desempenho dos alunos e, consequentemente, melhorar a qualidade da educação no estado. Esses dados são geralmente divididos em duas categorias principais: dados contextuais socioeconômicos e dados das avaliações de desempenho acadêmico.

1.1.1.1 Dados Contextuais Socioeconômicos

Os dados contextuais socioeconômicos são coletados para obter um panorama abrangente sobre o ambiente em que os alunos estão inseridos. Esses dados incluem informações sobre:

Perfil do Aluno:

- Idade
- Sexo
- Raça/etnia

Condições Socioeconômicas:

- Renda familiar
- Nível de escolaridade dos pais ou responsáveis
- Ocupação dos pais ou responsáveis
- Número de pessoas na residência
- Condições de moradia (tipo de habitação, posse de bens, acesso a serviços básicos como água e eletricidade)

Aspectos Educacionais:

- Tipo de escola (urbana ou rural)
- Tipo de transporte utilizado para chegar à escola
- Frequência de faltas escolares
- Participação em atividades extracurriculares (esportes, música, etc.)

Ambiente de Estudo:

- Disponibilidade de um lugar adequado para estudar
- Acesso a materiais escolares e livros
- Uso de tecnologias como computadores, tablets e internet para fins educacionais

1.1.1.2 Dados das Avaliações

Os dados das avaliações são obtidos através de provas padronizadas aplicadas aos alunos. Esses dados incluem:

Desempenho Acadêmico:

- Notas obtidas nas avaliações de Língua Portuguesa e Matemática
- Níveis de proficiência, que variam de "Muito Crítico" a "Desejável"
- Comparação de desempenho entre diferentes escolas e regiões

Distribuição de Desempenho:

- Percentual de alunos em cada nível de proficiência
- Identificação de padrões de desempenho em diferentes grupos demográficos e socioeconômicos

Progressão Escolar:

- Taxas de aprovação e reprovação
- Evolução do desempenho ao longo dos anos

1.1.1.3 Importância dos Dados Coletados

A coleta desses dados é fundamental para:

- **Diagnóstico Preciso**: Entender as variáveis que impactam o desempenho dos alunos e identificar áreas críticas que necessitam de intervenção.
- **Planejamento de Políticas Públicas**: Subsidiar a criação e implementação de políticas educacionais mais eficazes e direcionadas.
- **Apoio às Escolas**: Fornecer informações detalhadas que ajudam as escolas a adaptar suas práticas pedagógicas às necessidades específicas dos alunos.
- **Desenvolvimento de Programas de Intervenção**: Criar programas específicos de apoio e reforço escolar, além de iniciativas para melhorar as condições socioeconômicas que influenciam negativamente o aprendizado.

1.1.1.4 Utilização dos Dados na Prática

Os dados coletados pelo SPAECE são utilizados de várias maneiras:

- **Análise de Desempenho**: Identificar tendências de desempenho e áreas de dificuldade comuns entre os alunos.
- **Correlação de Fatores Socioeconômicos e Desempenho**: Estudar como diferentes fatores socioeconômicos influenciam o desempenho acadêmico.

- **Desenvolvimento de Intervenções**: Basear intervenções específicas em dados concretos, como programas de reforço escolar ou iniciativas de formação continuada para professores.
- **Monitoramento e Avaliação**: Avaliar a eficácia das políticas e programas implementados, ajustando estratégias conforme necessário para garantir melhorias contínuas.

1.1.1.5 Dados Aplicados no Projeto

O conjunto de dados aplicado nos testes consiste em 54.500 instâncias, sendo cada uma delas correspondente a um aluno; e 18 atributos, 14 atributos categóricos textuais correspondentes às perguntas de questionário socioeconômico e 4 atributos numéricos correspondentes aos escores e níveis de proficiência obtidos nas avaliações de Matemática e Português. Dados estes, que foram coletados na edição de 2018 do exame. O dataset foi denominado de “SPAECE_2018” e estruturado no formato “.csv” e não possuía dados faltantes ou inválidos. Exemplificação do processo de codificação categórica das pontuações:

1. **Prova de Matemática**:
    - Número total de questões: 40
    - Cada questão correta vale 1 ponto.
    - Aluno A acertou 30 questões.
    - Pontuação bruta do Aluno A: 30 pontos.
2. **Escalonamento**:
    - Pontuação bruta de 30 é convertida para uma escala de 0 a 500.
    - Após o escalonamento, a pontuação do Aluno A pode ser, por exemplo, 300 pontos.
3. **Nível de Proficiência**:
    - A pontuação de 300 pontos pode ser classificada como "Intermediário" conforme a definição dos níveis de proficiência do SPAECE.

Dados numéricos são úteis para predições usando regressão e dados categóricos podem ser aplicados em problemas de classificação. Podemos também transformar os dados numéricos em codificação de categorias e vice-versa. Por exemplo**:**

Níveis de Pontuação:

- **Muito Crítico**: 0 a 125 pontos
- **Crítico**: 126 a 200 pontos
- **Intermediário**: 201 a 275 pontos
- **Adequado**: 276 a 350 pontos
- **Desejável**: 351 a 500 pontos

Assim, dividimos os dados numéricos em 5 categorias diferentes, fazendo a possível a classificação dos mesmos. É possível aplicar o mesmo tipo de codificação para os dados numéricos de proficiência, criando assim variações do dataset original que melhor se aplicam a diferentes aplicações e modelos de aprendizagem de máquina. O ajuste e codificação do dataset será melhor abordado no tópico posterior: “1.1.4.2 Tipos de Codificações Aplicadas”.

1.1.2 Limpeza de dados

A limpeza de dados é uma etapa crucial no processo de análise de dados, especialmente em projetos de aprendizado de máquina e mineração de dados, como no contexto do SPAECE. Essa fase assegura que os dados utilizados são de alta qualidade, livres de erros e inconsistências, e adequados para a análise e aplicação subsequentes. O processamento dos dados foi simplificado nas seguintes etapas:

1. **Verificação de Integridade:** Inicialmente, procedeu-se à verificação da integridade dos dados coletados, garantindo a presença de todos os registros esperados e a ausência de duplicações ou perdas significativas de informações. Esta verificação foi conduzida por meio de planilhas, considerando que o conjunto de dados original está no formato ".csv", bem como através de datagramas gerados pelo pacote Python Pandas. Dado que cada entrada corresponde a um aluno e possui um identificador único (nome), esta etapa do processamento foi simplificada.
2. **Verificação e Tratamento de Valores Ausentes:** Valores ausentes são comuns em grandes conjuntos de dados e podem ocorrer devido a diversos motivos, como falhas na coleta de dados ou respostas incompletas dos questionários. No conjunto de dados abordado não foram identificados valores ausentes.
3. **Correção de Erros e Padronização dos Dados:** Os dados podem conter erros, como valores fora do intervalo esperado, formatação incorreta ou entradas duplicadas. Neste caso, foram encontrados erros na notação numérica de ponto flutuante dos atributos de Pontuação e Proficiência, uma possuía notação de ponto flutuante com "." (notação americana) e outra em ponto flutuante com "," (notação numérica nacional). Este erro de formatação e padronização dos dados pode impactar negativamente na interpretação dos dados, no treinamento e predição dos modelos. A formatação foi corrigida usando Python.

```
# Formatando colunas que estão com float separados por virgula
gnetDf['PROFICIENCIA EM PORTUGUES'] = gnetDf['PROFICIENCIA EM PORTUGUES'].str.replace(',', '.').astype(float)
gnetDf['PROFICIENCIA EM MATEMATICA'] = gnetDf['PROFICIENCIA EM MATEMATICA'].str.replace(',', '.').astype(float)
```

Bibliografia

https://cs.brown.edu/courses/cs227/archives/2017/papers/data-cleaning-IEEE.pdf#page=5
https://link.springer.com/book/10.1007/978-3-319-10247-4
https://ieeexplore.ieee.org/abstract/document/7944072

1.1.2.1 Remoção de Outliers com Orange Data Mining

Outliers são valores que se desviam significativamente do restante dos dados em um conjunto de dados. Eles podem surgir por várias razões, incluindo erros de coleta de dados, variabilidade natural ou fenômenos extremos. A detecção e o tratamento de outliers são etapas essenciais na limpeza de dados, pois os outliers podem afetar negativamente a análise e os resultados dos modelos de aprendizado de máquina. Para tratar possíveis *outliers* no conjunto de dados, foi aplicada a ferramenta Orange Data Mining. Esta é uma plataforma de código aberto desenvolvida para análise de dados, aprendizado de máquina, e mineração de dados. É conhecida por sua interface gráfica amigável, que permite aos usuários construir fluxos de trabalho de análise de dados de maneira intuitiva e visual, sem necessidade de programação e que oferece várias opções para a detecção e tratamento de outliers, uma das quais é o uso do método Isolation Forest que baseia-se em árvores para detectar outliers. Diferentemente de outros algoritmos de detecção de outliers que modelam a distribuição dos dados, o Isolation Forest explicitamente isola observações. A premissa é que outliers são poucos e diferentes, portanto, mais fáceis de isolar. Abaixo o diagrama de Widgets do Orange:

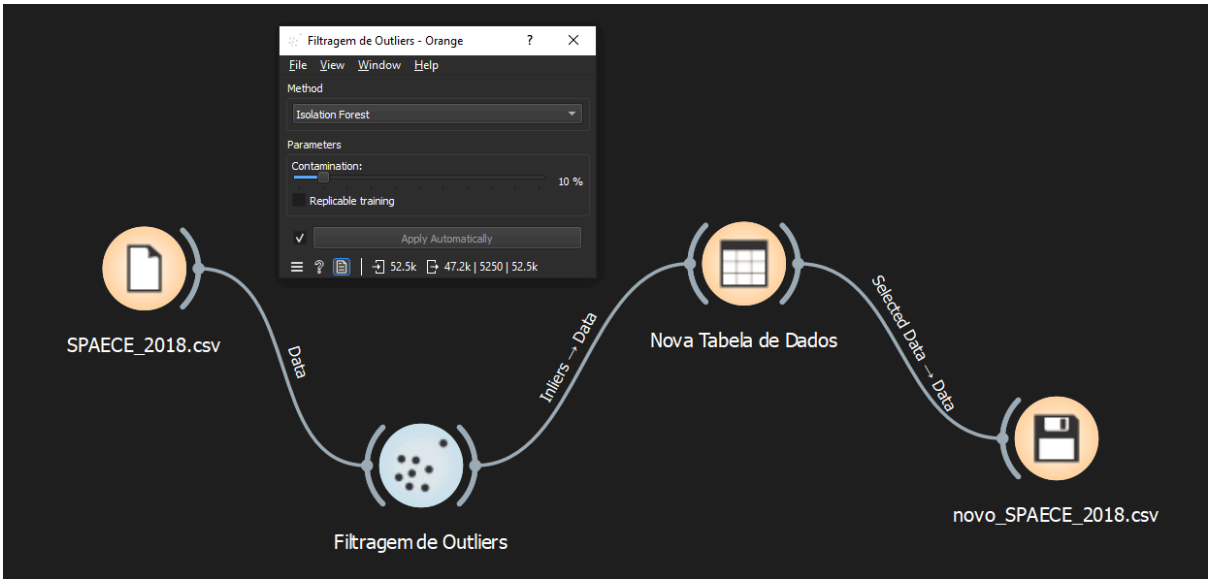

Figura 1: Diagrama de Filtragem de *Outliers* com *Orange Data Mining*
Fonte: Elaborado pelo Autor

Bibliografia: https://dl.acm.org/doi/abs/10.1145/1541880.1541882
https://link.springer.com/chapter/10.1007/978-3-319-47578-3_8
https://ieeexplore.ieee.org/abstract/document/4781136

Parâmetros na Configuração

1. **Detection Method: Isolation Forest**:
    - Método de detecção de outliers selecionado é o Isolation Forest.
2. **Contamination: 10%**:
    - Este parâmetro define a proporção de outliers esperada no conjunto de dados. No caso da configuração, 10% dos dados são considerados como outliers.
    - O parâmetro "contamination" ajuda o modelo a calibrar o corte entre pontos normais e anômalos.
3. **Replicable Training: No**:
    - Quando desativado, o treinamento não é replicável. Isso significa que, a cada execução, os resultados podem variar ligeiramente devido à natureza aleatória da construção das árvores.
    - Se ativado, o treinamento seria replicável, ou seja, os resultados seriam consistentes em execuções repetidas, útil para fins de reprodução dos resultados.

Como resultado obteve-se um novo conjunto de dados com 47.200 instâncias e livre de *outliers*. Este novo conjunto de dados foi salvo como “novo_SPAECE.csv”.

1.1.2.2 Visualização de Dados

A exploração e visualização dos dados é uma etapa fundamental para entender o comportamento das variáveis e suas relações. Utilizamos diferentes tipos de gráficos e análises estatísticas para identificar padrões, tendências e possíveis outliers. Através do “*pyplot*” pode-se plotar gráficos e visualizar informações como distribuição de dados e possíveis correlações de atributos. O “*pyplot”* é um módulo dentro da biblioteca “*matplotlib”* que oferece uma interface de programação semelhante ao MATLAB para plotagem de gráficos em Python. Ele fornece uma maneira conveniente de criar figuras, eixos, gráficos de linha, de dispersão, histogramas, barras, entre outros tipos de visualizações. Alguns dos gráficos gerados e dados extraídos:

Bibliografia:
https://www.researchgate.net/profile/Dr-Subhendu-Pani/publication/337146539_IJITEE/links/5dc70b124585151435fb427e/IJITEE.pdf

https://books.google.com.br/books?hl=pt-BR&lr=&id=G99YDwAAQBAJ&oi=fnd&pg=PP1&dq=Matplotlib+for+Data+Visualization:+A+Practical+Guide+for+Real-world+Data&ots=tz67wv8NZg&sig=hXHUV3qJyLSgUQlkxbfpjpRwizo&redir_esc=y#v=onepage&q&f=false

https://d1wqtxts1xzle7.cloudfront.net/65736020/ijatcse391012021-libre.pdf?1613836209=&response-content-disposition=inline%3B+filename%3DComparative_Analysis_of_Data_Visualizati.pdf&Expires=1718721477&Signature=IX2KvIbNO6nEcjHmDOyxdhUOm7Gyq~Unll4PKnof6W8St4WYeiXQtuJPv6RzF29WPaPjSARL4lQ4r4BaHaCld1CxvPaJwSIUYyTUKwpB7QCwXkbHeFN9JVslfR~Xb2pkrbZzVvERmJg5xitWS-Wa5g8TdXObqzUaqVCqjQykdD1g5xTjFrSLx2kZ5DP0nvS9HxOKKs1jkm3UpRKEQpCPkzYmFbEAaEqgID1W4H08axqF2-4YwdCGOHyKhL-NeZhDfP96GG9yjE4TNvcMZK2PCpaNngXID9Eq54nnPKv23dVwcTd-dUeCb-JIZkfDr7YgMT5MZOaNTQMlHCXA8bRhlA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

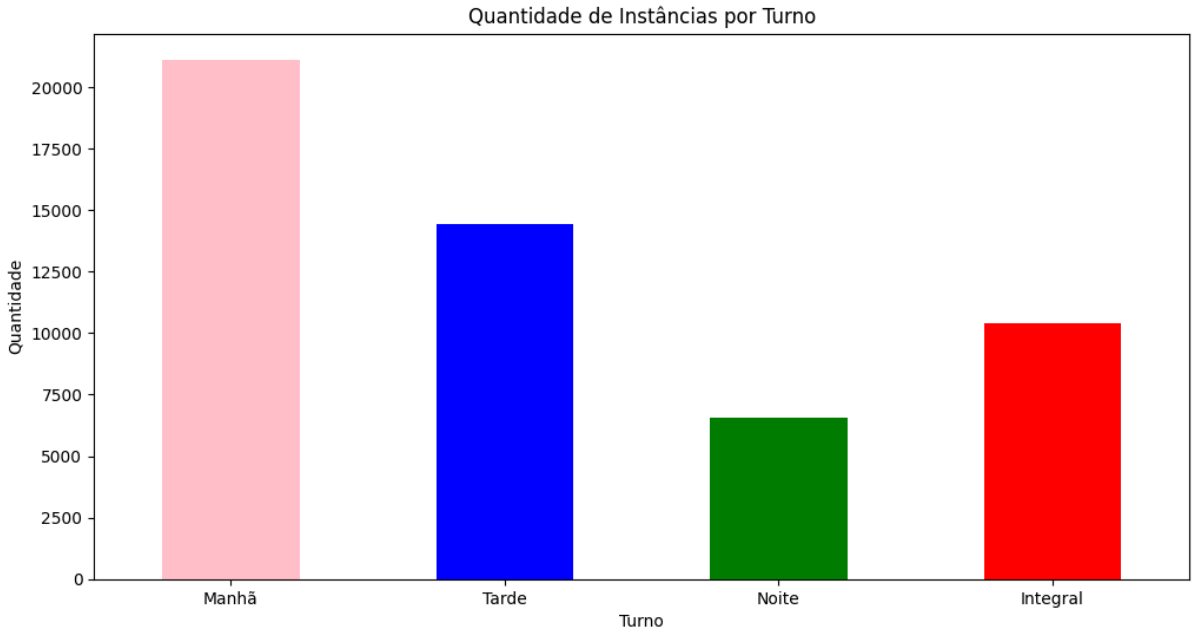

Gráfico 1: Gráfico de Barras da Quantidade de Instâncias por Turno Escolar
Fonte: Elaborado pelo Autor

No gráfico 1 podemos observar a distribuição das instâncias (eixo y) nos grupos definidos pelos turnos escolares dos alunos (eixo x). Sendo a maioria dos alunos, estudantes do turno diurno.

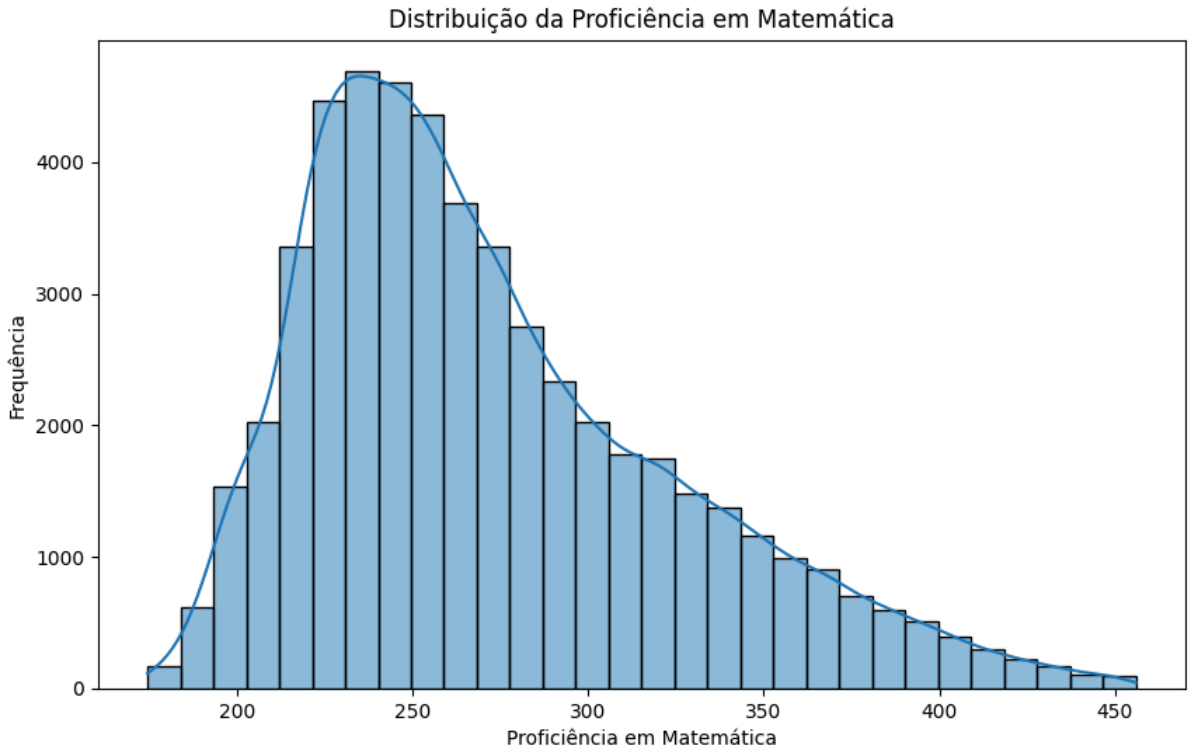

Gráfico 2: Histograma de Distribuição da Proficiência em Matemática.
Fonte: Elaborado pelo Autor

O Gráfico 2 representa os diferentes valores de proficiência em matemática no eixo x e o número de alunos que possuem determinada proficiência no eixo y. É notável uma distribuição assimétrica à direita, ou seja, a cauda se estende mais para a direita. Apontando que a maioria dos alunos possui uma proficiência em matemática entre 200 e 300 pontos. O valor mais frequente (pico do histograma), ou moda, está em torno de 250 pontos e há uma variação considerável na proficiência, com uma cauda que se estende até cerca de 450 pontos, indicando uma diversidade significativa no desempenho dos alunos. A cauda da direita se estende para valores mais altos de proficiência, mas com menor frequência. Isso indica que, embora existam alunos com alta proficiência, eles são menos comuns. Com essas informações é possível observar:

Desempenho Médio: a maioria dos alunos tem uma proficiência em matemática concentrada em torno de 250 pontos, o que pode ser interpretado como a média geral dos alunos. A assimetria à direita sugere que há mais alunos com proficiência abaixo da média, mas com uma presença notável de alunos com desempenho acima da média.

Variabilidade: A presença de uma cauda longa indica que há uma variabilidade significativa no desempenho, com alguns alunos apresentando proficiência muito alta. A dispersão larga sugere que diferentes fatores podem influenciar o desempenho dos alunos.

Na tentativa de encontrar o fator mais impactante no desempenho dos alunos, o Gráfico 3 foi gerado.

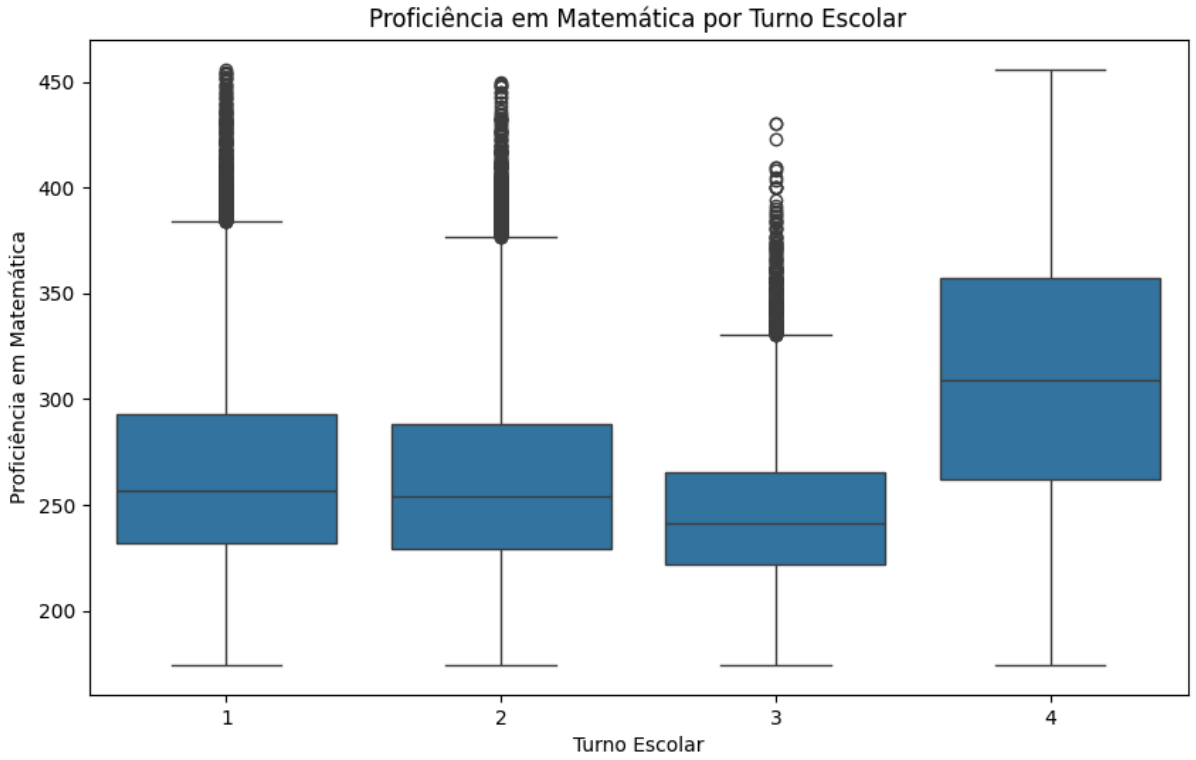

Gráfico 3: Boxplot de Distribuição da Proficiência em Matemática por Turno Escolar.
Fonte: Elaborado pelo Autor

Onde o eixo ‘x’ representa os diferentes turnos escolares (1: Manhã, 2: Tarde, 3: Noite, 4: Integral) e o eixo ‘y’ representa os valores de proficiência em matemática dos alunos. Observa-se a seguinte distribuição:

Turno 1 (Manhã):

- A mediana está em torno de 275.
- O intervalo interquartil (IQR) é aproximadamente de 250 a 300.
- Muitos outliers acima de 300.
- A dispersão dos dados é maior em comparação com outros turnos, indicando uma variabilidade significativa na proficiência dos alunos.

Turno 2 (Tarde):

- A mediana está ligeiramente abaixo de 275.
- O IQR é similar ao do turno da manhã, mas a dispersão parece um pouco menor.
- Também possui muitos outliers acima de 300, mas menos do que o turno da manhã.

Turno 3 (Noite):

- A mediana está abaixo de 250.

- O IQR vai de aproximadamente 225 a 275.
- Menos dispersão comparada aos turnos da manhã e da tarde, mas ainda com outliers significativos.

Turno 4 (Integral):

- A mediana está acima de 300, a maior entre todos os turnos.
- O IQR vai de aproximadamente 275 a 350.
- Menor número de outliers e menos dispersão, indicando uma maior consistência na proficiência dos alunos nesse turno.

Estas informações proporcionam as seguintes interpretações:

Consistência e Desempenho:

- Alunos do turno integral (4) têm, em média, uma proficiência maior em matemática comparada aos alunos dos outros turnos.
- A consistência é maior no turno integral, sugerindo que os alunos nesse turno têm desempenho mais uniforme.

Variabilidade:

- Os turnos da manhã (1) e tarde (2) mostram grande variabilidade nos resultados dos alunos, o que pode indicar diferenças significativas no ensino ou nas condições de aprendizagem.
- O turno da noite (3) tem menor desempenho mediano e uma menor variabilidade interna, mas ainda apresenta outliers significativos.

Impacto do Turno:

- O turno escolar parece ter um impacto significativo na proficiência em matemática dos alunos.
- Programas integrais (turno 4) podem oferecer um ambiente mais favorável para a aprendizagem de matemática, possivelmente devido a mais tempo de estudo, recursos ou métodos de ensino.

O que sugere que o turno escolar influencia significativamente a proficiência dos alunos em matemática. Alunos em turnos integrais tendem a ter melhores resultados, com menor variabilidade e menos outliers negativos. No Gráfico 4 podemos confirmar que a situação se repete quando observamos os dados de Pontuação em Matemática e os dados de Proficiência. Relações como esta são observadas pelo algoritmo de Aprendizagem de Máquina.

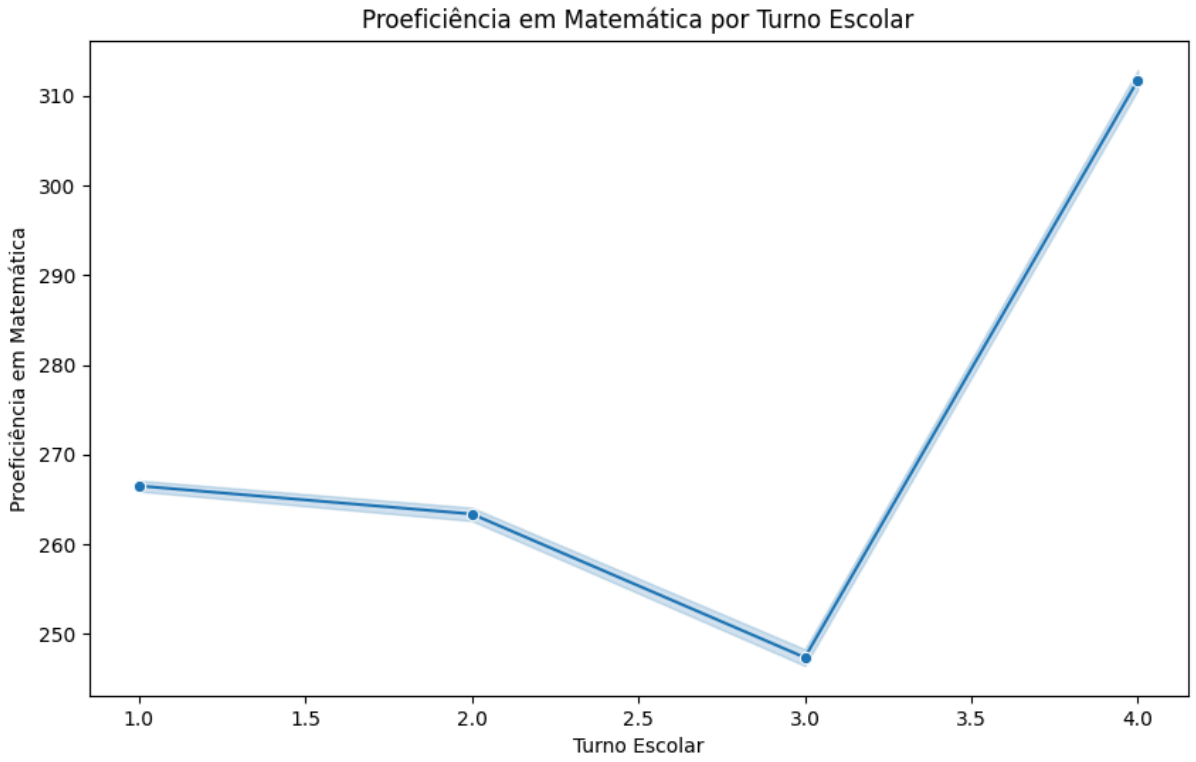

Gráfico 4: Relação Proficiência em Matemática x Turno Escolar
Fonte: Elaborado pelo Autor

No Gráfico 5 tem-se a pontuação direta dos alunos em matemática como eixo ‘x’ e a medida de proficiência em matemática dos alunos (derivada da pontuação) como eixo ‘y’. O gráfico mostra uma relação aproximadamente linear entre a pontuação direta e a proficiência. Isso indica que conforme a pontuação dos alunos aumenta, sua proficiência também aumenta de maneira proporcional. Dada a natureza dos dados, onde a proficiência é derivada diretamente da pontuação, é esperado observar uma forte correlação linear. Este é um exemplo claro de multicolinearidade, onde duas variáveis estão intimamente relacionadas.

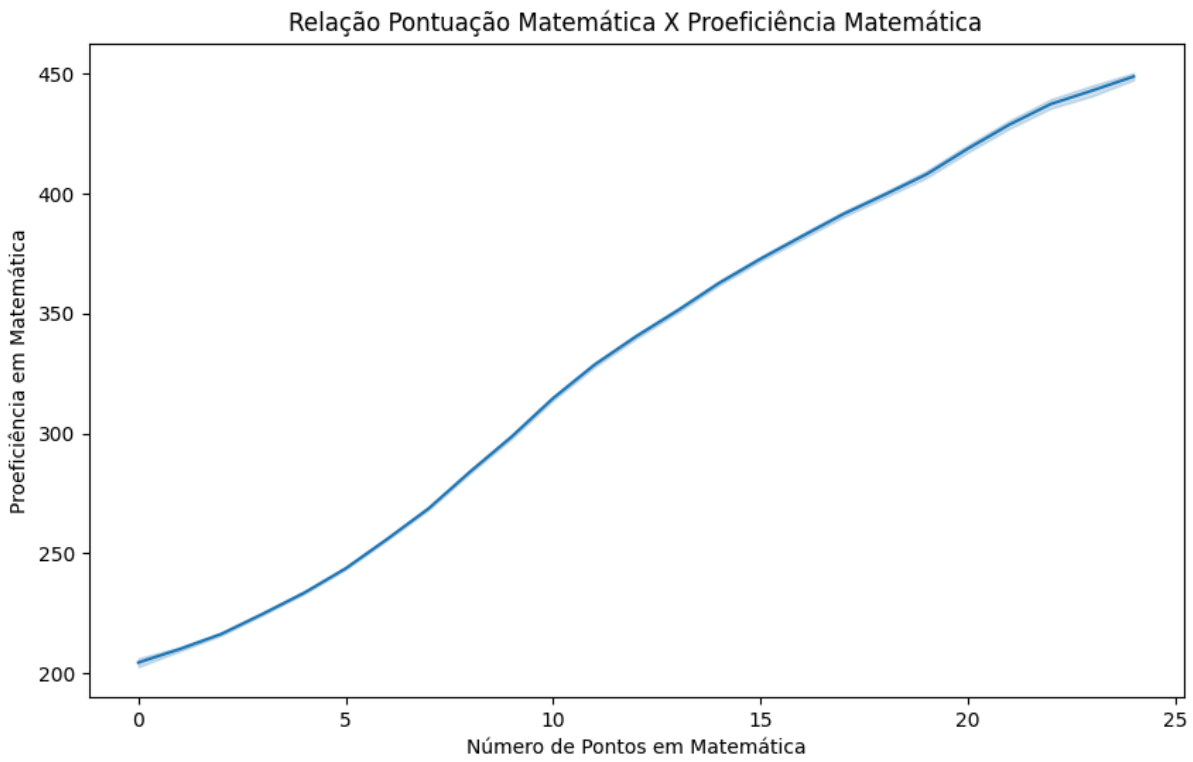

Gráfico 5: Relação de Proficiência em Matemática e Pontuação em Matemática
Fonte: Elaborado pelo Autor

Como a proficiência é calculada diretamente a partir da pontuação, a forte correlação não é surpreendente. Contudo, esta relação pode introduzir viés ao utilizar esses dados em modelos de aprendizado de máquina, pois o modelo pode "aprender" esta relação artificialmente inflada. Então a presença de uma alta correlação entre pontuação e proficiência pode resultar em problemas de multicolinearidade, especialmente em modelos de regressão linear ou em técnicas que assumem independência das features. Isso pode levar a coeficientes instáveis e a dificuldades na interpretação dos resultados. A avaliação do modelo deve considerar o potencial viés introduzido por essa correlação. É crucial utilizar técnicas de validação cruzada ou desconsiderar um dos atributos para garantir que o modelo não esteja super ajustado aos dados devido à alta correlação entre essas variáveis. O mesmo cenário é válido para a relação Pontuação em Português e Proficiência em Português

1.1.2.3 Matriz de Correlação

Para observar com maior precisão a correlação entre os atributos, foi criada uma matriz de correlação. A matriz de correlação é uma representação tabular que mostra as relações lineares entre todas as variáveis em um dataset. Cada célula na matriz contém o coeficiente de correlação entre duas variáveis. O coeficiente de correlação é uma medida estatística que indica o grau de relação linear entre duas variáveis. Ele varia de -1 a 1, onde:

- 1 indica uma correlação positiva perfeita: à medida que uma variável aumenta, a outra também aumenta de forma proporcional.
- -1 indica uma correlação negativa perfeita: à medida que uma variável aumenta, a outra diminui de forma proporcional.
- 0 indica ausência de correlação linear: não há relação linear entre as variáveis.

A matriz de correlação é frequentemente visualizada como um heatmap, onde cores diferentes representam diferentes níveis de correlação. Cores mais claras indicam correlação mais forte (positiva ou negativa), enquanto cores mais escuras indicam correlação mais fraca ou ausência de correlação. A análise da matriz de correlação pode fornecer insights valiosos, como:

- **Identificação de variáveis fortemente correlacionadas:** variáveis altamente correlacionadas podem fornecer informações redundantes e podem ser candidatas para redução de dimensionalidade.
- **Identificação de relações interessantes:** correlações entre variáveis podem revelar padrões ou relações importantes no dataset.
- **Seleção de variáveis para modelagem:** variáveis altamente correlacionadas podem ser eliminadas para simplificar modelos e evitar multicolinearidade.

No contexto de um projeto de análise de dados ou modelagem preditiva, a matriz de correlação é uma ferramenta poderosa para explorar as relações entre as variáveis e tomar decisões informadas sobre o processo de modelagem. A Matriz de Correlação resultante foi a seguinte:

Bibliografia: https://psycnet.apa.org/record/1980-08757-001
https://onlinelibrary.wiley.com/doi/full/10.1002/imt2.43

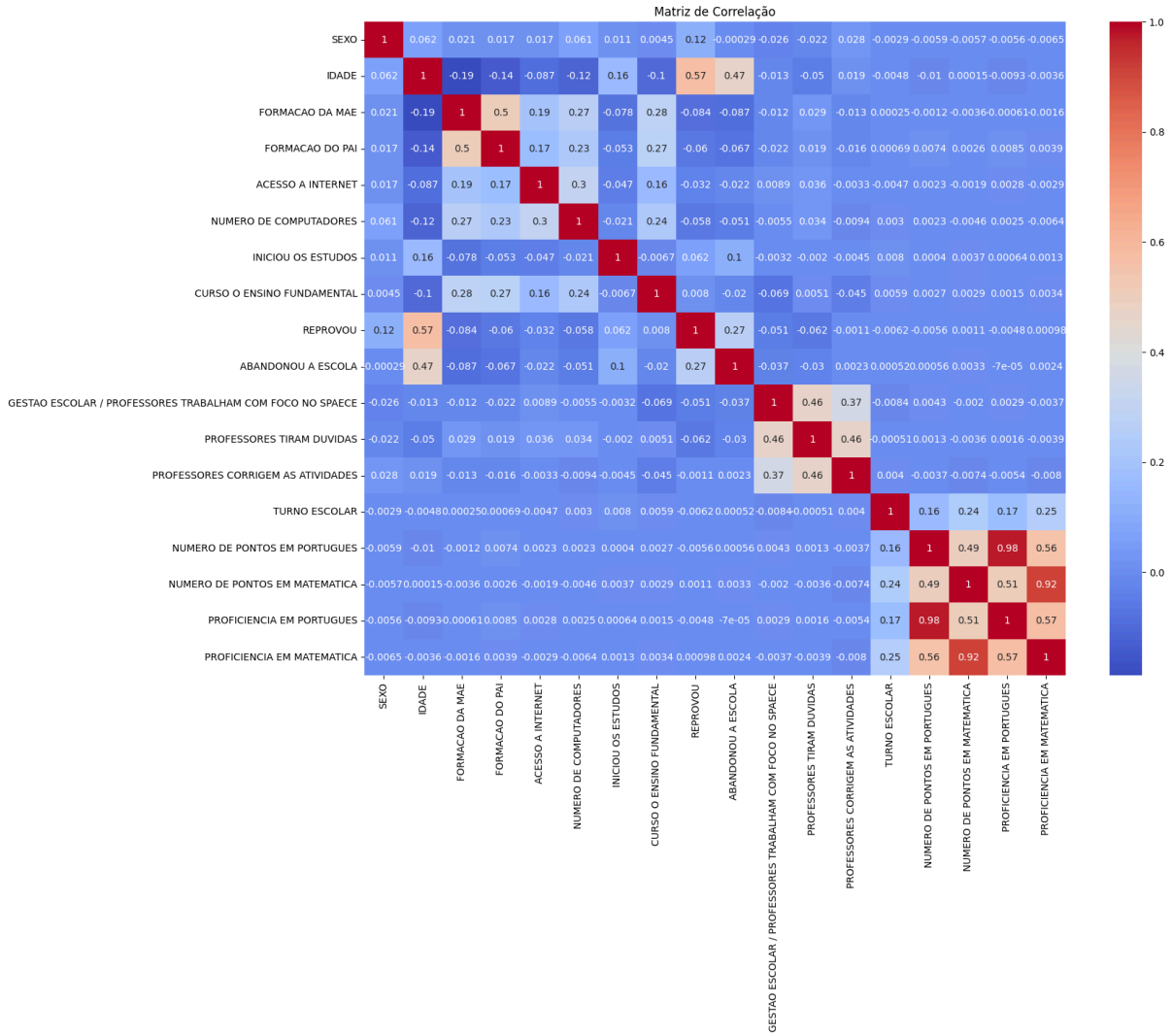

Figura 2: Matriz de Correlação de Atributos
Fonte: Elaborado pelo Autor

1.1.3 Seleção de Features

Conjuntos de dados frequentemente contêm muitos atributos, alguns dos quais podem não contribuir significativamente para a predição do modelo. A seleção de features ajuda a reduzir o número de atributos, simplificando assim a complexidade do modelo e melhorando o desempenho computacional. Ao remover atributos irrelevantes ou redundantes, a seleção de features pode melhorar a precisão, a generalização e a interpretabilidade dos modelos de ML. Modelos treinados com um conjunto de features mais relevantes tendem a ter um desempenho melhor em novos dados. A seleção de features também pode reduzir o overfitting, fenômeno no qual o modelo se ajusta muito bem aos dados de treinamento, mas falha em generalizar para novos dados. Reduzir a dimensionalidade dos dados pode ajudar a mitigar esse problema. Os métodos de Seleção de Features aplicados no projeto foram os seguintes:

Bibliografia: https://cdn.aaai.org/ICML/2003/ICML03-111.pdf
https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web
https://ieeexplore.ieee.org/abstract/document/7160458
https://www.jmlr.org/papers/volume10/tuv09a/tuv09a.pdf

1.1.3.3 Seleção com Árvores de Decisão

A árvore de decisão é uma técnica popular para seleção de features, especialmente em problemas de classificação e regressão. Ela funciona dividindo o conjunto de dados em subconjuntos menores com base nas features mais importantes. O critério de divisão pode ser medido usando diferentes métodos, como ganho de informação, índice Gini ou erro quadrático médio.

- **Regressão:** Para problemas de regressão, as árvores de decisão são usadas para identificar as features mais importantes para prever a variável alvo. As features são selecionadas com base em sua capacidade de reduzir a variabilidade na variável de resposta.
- **Classificação:** Para problemas de classificação, as árvores de decisão são usadas para classificar os dados com base nas features mais discriminativas. As features são selecionadas com base em sua capacidade de separar as diferentes classes no conjunto de dados.

1.1.3.4 Seleção com Features Automáticas do Sklearn

O SKlearn oferece várias técnicas para seleção automática de features, incluindo SelectKBest, SelectPercentile, Recursive Feature Elimination (RFE) e SelectFromModel. Essas técnicas utilizam diferentes critérios para selecionar as melhores features, como pontuação de teste estatístico, importância de features em modelos de aprendizado de máquina ou recursão sobre subconjuntos de features.

- **Regressão:** Para problemas de regressão, essas técnicas podem ser usadas para identificar as features mais importantes com base em sua contribuição para a precisão do modelo de regressão.
- **Classificação:** Para problemas de classificação, essas técnicas podem ser usadas para selecionar as features mais discriminativas com base em sua capacidade de separar as classes no conjunto de dados.

1.1.4 Dados de Treino

Os dados de Teste e Treino foram distribuídos de acordo com o Diagrama 2, novamente montado com o auxílio da ferramenta Orange Datamining. Onde o conjunto de dados foi dividido na proporção 10% para testes e 90% para treino, aplicando *data sampling,* que é o processo de selecionar uma parte dos dados de um conjunto maior de dados para realizar análises. A amostragem aleatória é o método pelo qual cada observação do conjunto de dados tem a mesma probabilidade de ser selecionada. Não há qualquer regra ou padrão específico na escolha dos dados; eles são escolhidos de maneira aleatória.

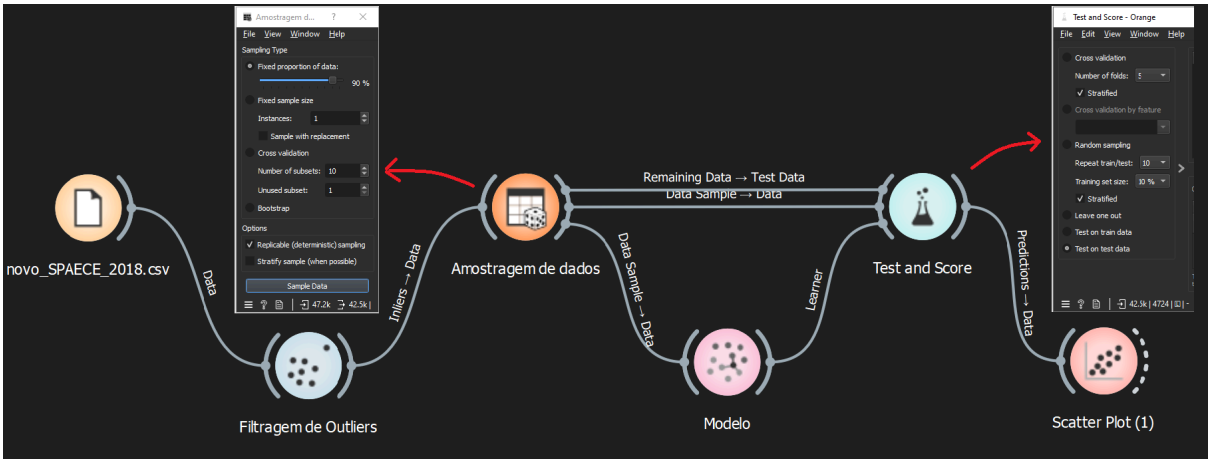

Figura 2: Diagrama de Distribuição de Dados de Treino e Teste com Orange Datamining
Fonte: Elaborado pelo Autor

A opção "Deterministic" refere-se à consistência na amostragem. Quando a amostragem é determinística, ela produz o mesmo subconjunto de dados todas as vezes que o procedimento é executado, dada uma mesma semente de aleatoriedade. Isso é útil para garantir a reprodutibilidade dos resultados. Com isso podemos criar diferentes datasets com diferentes proporções de dados para testes e treinos de modelos.

1.1.4.2 Tipos de Codificações Aplicadas

Ao aplicar modelos de Machine Learning (ML) a problemas de classificação, regressão e associação, é essencial entender como lidar com diferentes tipos de dados: categóricos, textuais e numéricos. Cada tipo de dado requer uma abordagem específica para pré-processamento e codificação. No conjunto de dados aplicado foram encontrados 2 tipos de dados: categóricos (algumas respostas textuais dos questionários) e numéricos (notas, pontuações de proficiência e algumas respostas numéricas dos questionários). Então foram aplicadas as seguintes codificações:

Dados Categóricos: dados categóricos são valores discretos que representam categorias ou classes. Exemplos incluem cores ("vermelho", "verde", "azul"), estados civis ("solteiro", "casado", "divorciado"), e tipos de produtos.

1 - Label Encoding:

- **Descrição**: Converte cada categoria em um número inteiro único.
- **Aplicação**: Simples de implementar e pode ser usado quando as categorias têm uma ordem implícita.
- **Uso**: Frequentemente usado em árvores de decisão e florestas aleatórias.

2 - One-Hot Encoding:

- **Descrição**: Cria uma nova coluna binária (0 ou 1) para cada categoria.
- **Aplicação**: Evita atribuir ordens às categorias. É especialmente útil quando não há hierarquia entre as categorias.
- **Uso**: Usado em regressão linear, redes neurais e qualquer modelo que não interprete bem valores ordenados.

Dados Numéricos: dados numéricos são valores quantitativos que podem ser contínuos ou discretos. Exemplos incluem idade, salário, e temperatura.

3 - Discretização (Binning):

- **Descrição**: Converte dados contínuos em categorias discretas.
- **Aplicação**: Pode ser usado para transformar variáveis contínuas em categóricas.
- **Uso**: Usado em árvores de decisão e quando a relação entre os valores numéricos e a target variável é não-linear.

Aplicações para Problemas de ML

Classificação

- **Dados Categóricos**: One-Hot Encoding, Target Encoding.
- **Dados Textuais**: TF-IDF, Word Embeddings.
- **Dados Numéricos**: Normalização, Padronização.

Regressão

- **Dados Categóricos**: One-Hot Encoding, Target Encoding.
- **Dados Textuais**: TF-IDF, Word Embeddings.
- **Dados Numéricos**: Padronização, Discretização.

Associação

- **Dados Categóricos**: One-Hot Encoding.
- **Dados Textuais**: Bag of Words, TF-IDF.
- **Dados Numéricos**: Discretização

Na Figura 3 podemos observar um exemplo de codificação Label Encoding usando o *Widget Edit Domain* do *Orange Data Mining* e na Figura 4 pode-se observar o exemplo de codificação usando o mapeamento com datagramas do *Python Pandas*

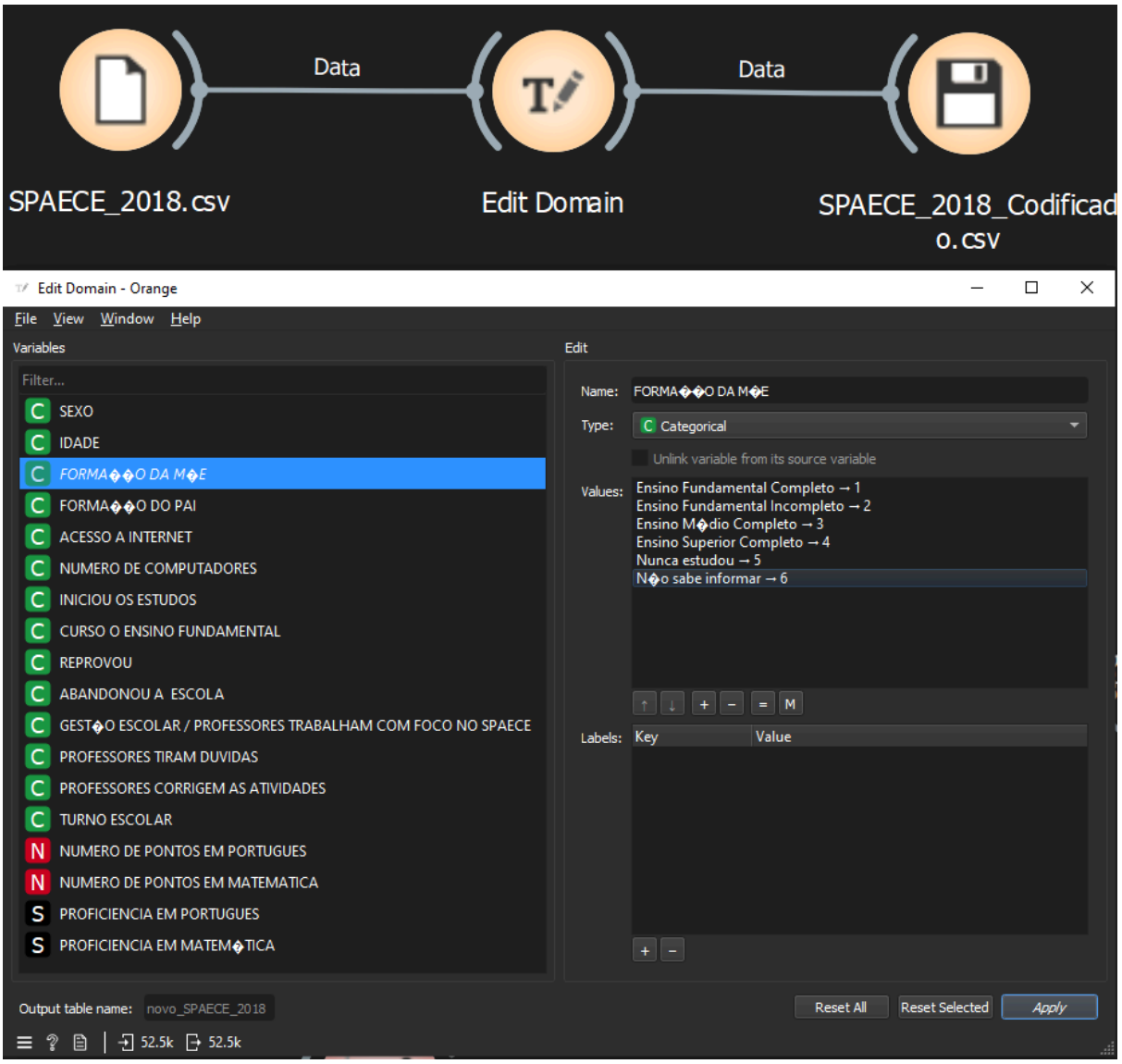

Figura 3: Diagrama e Widget de Edição de Domínios do Orange Data Mining
Fonte: Elaborado pelo Autor

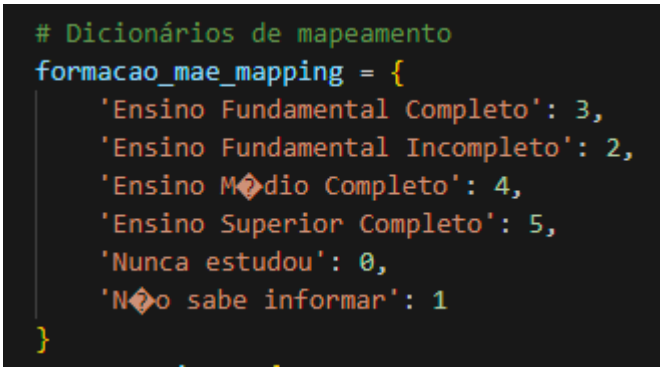

Figura 4: Exemplo de Código para Mapeamento de Categorias com Python e Pandas
Fonte: Elaborado pelo Autor

Com isso, a partir do dataset “novo_SPAECE_2018.csv” processado foram criadas variações com diferentes codificações para diferentes aplicações. Sendo eles:

“SPAECE_2018_codificado.csv": *dataset* codificado em *Label Encoding* e com os valores numéricos das notas e proficiências em formato *float*. Aplicável em problemas de Regressão.

"CLASS.csv”: *dataset* codificado em *Label Enconding* com os valores das notas e proficiências distribuídas em 5 grupos ou classes (sendo 1 = pontuação mínima / 5 = pontuação máxima). Aplicável em problemas de Classificação.

“Apriori.csv": *dataset* com colunas codificadas em *One-Hot Encoding* (binários bool) para aplicar em algoritmos de associação Apriori.

“PCY_binário.csv": é o *dataset* "Apriori.csv" no formato *One-Hot Encoding* numérico (0-1) para aplicar em algoritmos de Associação PCY, FP-Growth e ECLAT.

1.1.5 Modelos de Aprendizagem de Máquina

O conjunto de dados SPAECE contém informações variadas sobre os alunos, incluindo características demográficas, socioeconômicas e acadêmicas. Esta etapa do projeto testa como o conjunto de dados SPAECE pode ser utilizado para resolver diversos problemas de machine learning, desde previsão de valores contínuos (regressão), classificação de categorias e descoberta de padrões (associação). As etapas anteriores de preparação de dados, seleção de features, modelagem e avaliação são cruciais para obter modelos eficazes. Exemplos de abordagens:

- **Regressão:** Prever a proficiência em matemática com base em características demográficas e socioeconômicas dos alunos.
- **Classificação** Prever se um aluno irá abandonar a escola (sim ou não) com base em suas características e histórico acadêmico.
- **Associação** Descobrir regras associativas entre diferentes características dos alunos, como a relação entre a formação dos pais e o desempenho acadêmico.

Bibliografia: https://www.mdpi.com/2071-1050/14/15/9056
https://link.springer.com/chapter/10.1007/978-3-319-01854-6_26
https://link.springer.com/article/10.1007/s11269-017-1807-2
https://www.sciencedirect.com/science/article/abs/pii/S0360544220303467
https://www.sciencedirect.com/science/article/abs/pii/S1364032115001884
https://books.google.com.br/books?hl=pt-BR&lr=&id=Us4YE8lJVYMC&oi=fnd&pg=PP2&dq=regression+analysis&ots=WWGnhxShYk&sig=xwrFuajiEyBMhEEwvPOUybRnhj4&redir_esc=y#v=onepage&q=regression%20analysis&f=false
https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d4058d9f3f66c53ddea776c974fbd740afd994b4
https://link.springer.com/chapter/10.1007/3-540-46027-6_3
https://ieeexplore.ieee.org/abstract/document/7079081

1.1.5.1 Modelos comparados

Regressão apenas:

1. **Regressão Linear:** Modelo simples e interpretável que assume uma relação linear entre as features e a target. Funciona bem quando há uma relação linear ou quase linear entre as variáveis.

Regressão e Classificação:

2. **Árvores de Decisão (Decision Trees):** Divide os dados em subconjuntos com base em valores de features, criando uma estrutura de árvore. Capaz de capturar relações não lineares e interações entre features. Fácil de interpretar e visualizar, não requer muita preparação de dados, pode capturar relações não lineares. Pode se tornar complexa e sobre ajustada facilmente, especialmente com muitos dados.

3. **K-Nearest Neighbors Regression (KNN):** Prediz o valor de um ponto com base nos valores dos k pontos mais próximos. Simples e eficaz para pequenos conjuntos de dados com relações não lineares. ão escala bem com grandes datasets, sensível à escolha dos hiperparâmetros. Computacionalmente intensivo para grandes datasets, sensível ao ruído nos dados.

4. **Support Vector Machine (SVM):** Utilizando um conjunto de funções de base para criar um hiperplano de regressão. Bom para problemas de alta dimensionalidade e quando a relação entre as variáveis não é linear. Não escala bem com grandes datasets, sensível à escolha dos hiperparâmetros.

5. **Multilayer Perceptron (MLP):** Rede neural feedforward com uma ou mais camadas ocultas. Utiliza backpropagation para ajustar os pesos durante o treinamento. Capaz de capturar relações complexas e não lineares. Funciona bem com grandes conjuntos de dados e é altamente flexível devido ao ajuste dos hiperparâmetros, como número de camadas e neurônios por camada. Requer mais dados e poder computacional, mais difícil de interpretar e ajustar, sensível à escolha dos hiperparâmetros e à arquitetura da rede.

Classificação Apenas:

6. **Naive Bayes:** Calcula a probabilidade de uma instância pertencer a uma determinada classe com base na probabilidade condicional das features. A principal suposição é que as features são independentes entre si, o que significa que a presença de uma feature não afeta a presença de outra.Simples e rápido de treinar, eficaz para grandes datasets, funciona bem com dados categóricos. Assume independência entre as features, o que pode não ser verdadeiro.

Associação Apenas:

7. **Apriori:** Este algoritmo é baseado na propriedade "anti monotonia" da regra de associação, que afirma que se um itemset é frequente, então todos os seus subconjuntos são frequentes. O Apriori utiliza um processo iterativo para encontrar os *itemsets* frequentes e gerar regras de associação.
8. **FP-Growth (Frequent Pattern Growth):** Este algoritmo utiliza uma estrutura de árvore (FP-tree) para armazenar informações de frequência de *itemsets* de forma compacta. Ele é mais eficiente do que o Apriori, especialmente em grandes bases de dados, pois evita a geração de candidatos desnecessários.

1.1.5.1 Problemas de Regressão

Problemas de regressão envolvem prever um valor contínuo baseado em uma ou mais variáveis independentes. Ao contrário dos problemas de classificação, onde a saída é uma classe ou categoria, em problemas de regressão, a saída é um valor numérico. Modelos de regressão são usados em diversas áreas, incluindo economia, finanças, ciências naturais, engenharia e muitas outras, para prever valores como preços, temperaturas, pontuações, etc. No contexto do dataset do SPAECE 2018, um exemplo de problema de regressão seria prever a "PROFICIENCIA EM MATEMATICA" dos estudantes com base em várias características como sexo, idade, formação dos pais, acesso à internet, e outras variáveis presentes no dataset.

1.1.5.1.1 Métricas para Avaliação de Modelos de Regressão

Erro Médio Absoluto (Mean Absolute Error - MAE):O MAE é a média da diferença absoluta entre os valores previstos pelo modelo e os valores observados. É calculado pela fórmula:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Onde y_i são os valores observados, $\widehat{y}$ são os valores previstos pelo modelo e nnn é o número total de observações. O MAE mede a magnitude média dos erros em uma escala similar aos dados originais, sendo menos sensível a outliers comparado ao MSE.

Erro Quadrático Médio (Mean Squared Error - MSE):

O MSE é a média dos quadrados dos erros entre os valores previstos e os valores observados. É calculado pela fórmula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

O MSE penaliza erros grandes mais do que o MAE, pois eleva os erros ao quadrado. É frequentemente usado em problemas onde erros maiores são críticos.

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

O RMSE é a raiz quadrada do MSE e é uma das métricas mais comuns para avaliação de modelos de regressão. É calculado pela fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

O RMSE fornece uma interpretação na mesma escala das variáveis dependentes, o que facilita a compreensão do quão bem o modelo está performando.

Coeficiente de Determinação (R²):

O R² é uma medida estatística que indica a proporção da variância dos dados que é explicada pelo modelo. É uma medida de quão bem os pontos de dados se ajustam à linha de regressão ajustada. Valores de R^2 variam de 0 a 1, sendo 1 indicativo de um ajuste perfeito.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Onde $\bar{y}$ é a média dos valores observados y_i. Um valor de R² próximo a 1 indica que o modelo explica bem a variabilidade dos dados, enquanto valores próximos a 0 indicam que o modelo não explica bem a variabilidade dos dados.

Erro Absoluto Percentual Médio (Mean Absolute Percentage Error - MAPE):

O MAPE é uma métrica útil para problemas onde a magnitude do erro é significativa em relação ao valor real. É calculado pela fórmula:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{|y_i|}\right) \times 100$$

O MAPE expressa o erro médio como uma porcentagem do valor real, sendo útil em contextos onde a escala dos dados é importante.

1.1.5.2 Problemas de Classificação

Problemas de classificação envolvem a previsão de categorias ou rótulos para novas observações com base em um conjunto de dados de treinamento. Ao contrário dos problemas de regressão, que prevêem valores contínuos, os problemas de classificação prevêem rótulos discretos. No contexto do dataset do SPAECE 2018, podemos formular problemas de classificação para predizer uma categoria baseada em outras features. Por exemplo:

Previsão de Performance Acadêmica: Classificar os alunos em categorias de desempenho com base em suas características e respostas ao questionário.
Previsão de Abandono Escolar: Classificar se um aluno tem probabilidade de abandonar a escola com base nas características demográficas e educacionais.

1.1.5.1.2 Métricas para Avaliação de Modelos de Classificação

Ao avaliar modelos de classificação, é importante escolher métricas de desempenho apropriadas para entender como o modelo está se saindo em diferentes aspectos. As métricas e métodos de avaliação que iremos aplicar serão:

Acurácia (Accuracy): A acurácia é a proporção de exemplos classificados corretamente pelo modelo em relação ao total de exemplos. Sendo uma métrica geral que indica a proporção de previsões corretas do modelo. Calculada por:

$$(TP + TN)/(TP + TN + FP + FN)$$

Precisão (Precision): A precisão é a proporção de verdadeiros positivos (TP) em relação a todos os exemplos classificados como positivos pelo modelo (verdadeiros positivos mais falsos positivos). É calculada pela fórmula:

$$TP/(TP + FP)$$

Revocação (Recall ou Sensibilidade): A revocação é a proporção de verdadeiros positivos (TP) em relação a todos os exemplos que são realmente positivos (verdadeiros positivos mais falsos negativos). A revocação mede a capacidade do modelo em identificar corretamente exemplos positivos. É calculada pela fórmula:

$$TP/(TP + FN)$$

O F1-Score: É a média harmônica da precisão e da revocação e fornece um único número que representa o balanceamento entre essas duas métricas. Sendo útil quando há desequilíbrio entre as classes. É calculado pela fórmula:

$$F1 = 2 \times \frac{\text{Preciso} \times \text{Revocao}}{\text{Preciso} + \text{Revocao}}$$

Matriz de Confusão: A matriz de confusão é uma tabela que mostra a frequência de classificações corretas e incorretas feitas pelo modelo. É uma ferramenta fundamental para entender o desempenho do modelo em cada classe.

1.1.5.3 Problemas de Associação

A mineração de dados com algoritmos de associação é uma técnica utilizada para descobrir relações interessantes, padrões ou regras de associação entre conjuntos de itens em grandes bases de dados. Esse processo é especialmente útil em contextos como análise de mercado, cestas de compras, detecção de fraudes, e muitas outras áreas onde a identificação de padrões ocultos pode fornecer dados importantes. Algoritmos de associação, como Apriori e FP-Growth, são ferramentas poderosas para descobrir padrões ocultos em grandes conjuntos de dados. No contexto do dataset SPAECE 2018, essas técnicas podem ser usadas para explorar e identificar associações entre características dos alunos e suas proficiências acadêmicas, fornecendo insights valiosos para educadores e formuladores de políticas.

1.1.5.3.1 Interpretação de Resultados

Os resultados incluirão regras como:

- **Support:** A proporção de transações que contém o itemset.
- **Confidence:** A confiança da regra, ou seja, a proporção de transações que contém o antecedente e o consequente.
- **Lift:** A razão de confiança para a regra em comparação com a confiança esperada se os antecedentes e os consequentes fossem independentes.

Exemplo:

```
###    antecedents  /  consequents / support / confidence / lift
### 0  {IDADE=17} / {PROFICIENCIA=3} / 0.15   /   0.80   /  1.33
```

Esta regra poderia ser interpretada como: "Se a idade do aluno é 17, há uma confiança de 80% de que a proficiência é 3, e essa relação é 1.33 vezes mais provável do que se esses eventos fossem independentes".