INSTITUTO FEDERAL GOIANO - CAMPUS CERES
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

TÓPICOS ESPECIAIS EM BANCO DE DADOS (MINERAÇÃO DE DADOS)

ISAQUE PONTES ROMUALDO
DOUGLAS CÁSSIO REIS SILVA OLIVEIRA

MINERAÇÃO DE DADOS: ANÁLISE DO PERFIL E DESEMPENHO NO ENEM 2024 (VALE DO SÃO PATRÍCIO)

Professor orientador:
Me. Adriano Honorato Braga

NOVEMBRO

2025

1. Ferramenta Utilizada para a Mineração

Para a realização deste trabalho, foi utilizada uma abordagem híbrida que combina o poder de processamento de scripts da linguagem Python (com a biblioteca Pandas) e a interface visual de mineração de dados do RapidMiner Studio.

- Python (Pandas): Utilizado na etapa de pré-processamento (ETL - Extração, Transformação e Carga) para filtrar os arquivos de microdados originais, que possuem um volume de dezenas de gigabytes, impossibilitando o carregamento direto em ferramentas de análise visual.
- RapidMiner Studio: A ferramenta principal de mineração de dados, utilizada para a modelagem do processo, agregação, "assimilação" (junção) e geração de todas as análises estatísticas e resultados visuais.

2. Descrição sobre a Ferramenta Utilizada (RapidMiner)

2.1 Proprietário

O RapidMiner Studio tem suas raízes na pesquisa acadêmica alemã, originando-se em 2001 na Universidade Técnica de Dortmund sob o nome YALE (Yet Another Learning Environment). Desenvolvido inicialmente como um projeto open-source para facilitar a experimentação em aprendizado de máquina, o software evoluiu para uma solução comercial robusta sob a gestão da empresa Rapid-I, fundada em 2006 pelos pesquisadores Ingo Mierswa e Ralf Klinkenberg. Em 2013, consolidando sua identidade no mercado global, a empresa foi renomeada para RapidMiner Inc., posicionando a ferramenta como líder em plataformas de ciência de dados (HOFMANN; KLINKENBERG, 2013).

Em um movimento estratégico para ampliar seu portfólio de inteligência artificial e análise de dados, a Altair Engineering — multinacional americana líder em ciência computacional e simulação — adquiriu a RapidMiner em setembro de 2022. A aquisição visou integrar a capacidade de modelagem low-code do RapidMiner com as soluções de computação de alta performance (HPC) da Altair, criando um ecossistema convergente que democratiza o acesso a análises complexas sem exigir codificação manual extensiva (ALTAIR, 2022). Atualmente, a plataforma é

mantida pela Altair, que expandiu suas funcionalidades para incluir governança de modelos e integração com sistemas industriais legados.

2.3 Forma de Instalação

Sob o prisma da arquitetura de software, o RapidMiner Studio é construído em linguagem Java, operando como uma aplicação desktop multiplataforma que exige a instalação prévia do Java Development Kit (JDK) para o gerenciamento de seus processos em memória. A ferramenta adota um modelo de licenciamento freemium: embora sua distribuição oficial ofereça uma versão de acesso livre (Free Edition), esta impõe restrições de escalabilidade, limitando o processamento a 10.000 registros (tuplas) por conjunto de dados.

Contudo, visando o fomento à pesquisa científica, a Altair mantém um programa de licenciamento acadêmico (Educational License). Esta modalidade permite que estudantes e pesquisadores removam as barreiras de volumetria de dados e acessem funcionalidades avançadas sem custos, viabilizando a análise de datasets robustos em trabalhos de conclusão de curso e artigos científicos.

2.4 Exemplos de Utilização

Operacionalmente, o RapidMiner distingue-se pela utilização de um paradigma de programação visual, onde a construção de modelos analíticos ocorre por meio de fluxos de trabalho (workflows) baseados em "operadores". Essa arquitetura drag-and-drop (arrastar e soltar) permite a estruturação modular de processos complexos, abstraindo a necessidade de codificação manual para a maioria das tarefas de ETL (Extract, Transform, Load) e modelagem (HOFMANN; KLINKENBERG, 2013).

A versatilidade da plataforma permite sua aplicação em diversas vertentes da mineração de dados:

- Análise Preditiva (Aprendizado Supervisionado): Amplamente utilizada para classificação e regressão, exemplificada pela previsão de evasão de clientes (churn rate) ou análise de risco de crédito.

- Segmentação e Agrupamento (Aprendizado Não Supervisionado): Aplicação de algoritmos de clustering (como K-Means) para identificar perfis de consumo ou padrões comportamentais sem rótulos prévios.
- Mineração de Dados Não Estruturados: Capacidade de processamento de linguagem natural para Text Mining e análise de sentimentos.

No escopo deste trabalho, entretanto, a ferramenta é empregada especificamente para a análise exploratória e estatística dos dados, aproveitando seus operadores de visualização e estatística descritiva para a compreensão inicial do fenômeno estudado (KOTU; DESHPANDE, 2019).

2.5 Artigos e Comparações

No contexto da Descoberta de Conhecimento em Bancos de Dados (KDD), a escolha da ferramenta de mineração é determinante para a eficiência do processo analítico. O RapidMiner consolidou-se na literatura como uma das plataformas líderes no segmento de ciência de dados, frequentemente figurando ao lado do KNIME (Konstanz Information Miner) como as principais soluções baseadas em fluxos de trabalho visuais (HOFMANN; KLINKENBERG, 2013).

Embora ambas as ferramentas compartilhem o paradigma de programação visual, elas divergem em suas filosofias de licenciamento e público-alvo. Enquanto o KNIME mantém uma estrutura predominantemente open-source, o RapidMiner — atualmente integrado ao ecossistema de soluções da Altair — prioriza uma interface gráfica de alto polimento e funcionalidades voltadas à integração corporativa e escalabilidade industrial.

Ao contrastar essas soluções modernas com ferramentas pioneiras, como o Weka (Waikato Environment for Knowledge Analysis), nota-se uma evolução significativa na interação humano-computador. Diferente do Weka, que opera sob uma arquitetura clássica e muitas vezes dependente de menus estáticos, o RapidMiner oferece um ambiente de desenvolvimento baseado em pipelines visuais. Segundo Kotu e Deshpande (2019), essa abordagem não apenas torna o processo mais intuitivo para analistas de negócio, mas também amplia a transparência e a reprodutibilidade dos modelos preditivos, superando a curva de aprendizado exigida por ferramentas mais antigas.

3. Detalhamento para o Processamento dos Dados (Dicionário)

O ponto de partida do processamento foi a análise dos dicionários de dados do ENEM 2024, que revelou o maior desafio técnico do projeto.

Os dados estão separados em dois arquivos principais: PARTICIPANTES_2024.csv (com dados socioeconômicos) e RESULTADOS_2024.csv (com as notas). O dicionário do arquivo de Resultados afirma explicitamente (na nota de rodapé 27) que não é possível fazer um JOIN (junção) direto entre as duas bases para cruzar o desempenho de um aluno específico com seu perfil socioeconômico.

Portanto, a estratégia de "assimilação" (sugerida nas "Considerações" do trabalho) foi adotada. Em vez de analisar o indivíduo, a unidade de análise foi alterada para o município (para as análises de correlação) e a escola (para as análises descritivas).

4. Conversão e Transformação dos Dados

A conversão dos dados brutos (originais) para um formato utilizável no RapidMiner foi a etapa de pré-processamento realizada em Python (Pandas).

1. Leitura em Chunks: Os arquivos gigantes foram lidos em "pedaços" (chunks) de 100.000 linhas, para evitar o esgotamento da memória RAM.
2. Seleção de Colunas (usecols): Apenas as colunas necessárias para a análise (NU_NOTA_CN, Q007, CO_MUNICIPIO_ESC, etc.) foram carregadas, otimizando drasticamente o processamento.
3. Filtragem de Localidade (.isin()): Cada "chunk" foi filtrado para manter apenas as linhas onde o código do município (CO_MUNICIPIO_ESC ou CO_MUNICIPIO_PROVA) pertencia à lista de códigos IBGE das "Cidades do Vale do São Patrício".
4. Saída: O resultado desse processo gerou dois arquivos CSV muito menores (FILTRO_PARTICIPANTES...csv e FILTRO_RESULTADOS...csv), que foram usados como entrada no RapidMiner.

5. Método

Nesta seção, é detalhada a forma como os dados filtrados foram processados dentro do RapidMiner para gerar os resultados de cada análise solicitada.

5.1 Dados do mercado

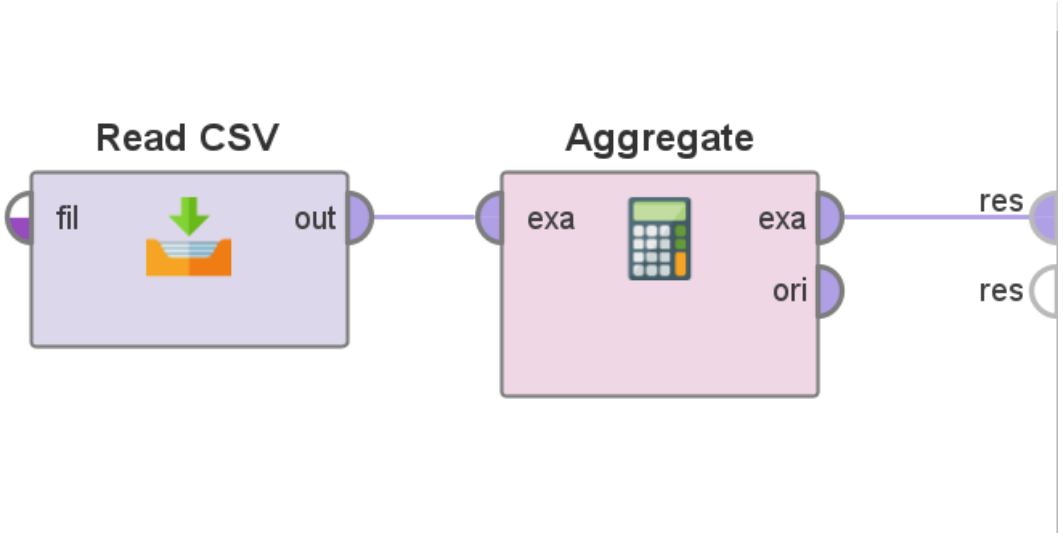
Inicialmente, considerou-se a utilização do software RapidMiner (plataforma Altair) para a extração de padrões e regras de associação no conjunto de dados coletado. No entanto, a aplicação de algoritmos de Machine Learning tradicionais, como o FP-Growth ou Apriori, mostrou-se inadequada dada a volumetria reduzida da amostra (N=10 transações). Em mineração de dados, a confiabilidade estatística depende diretamente da Lei dos Grandes Números; com uma base de dados tão restrita, a granularidade mínima de suporte torna-se excessivamente alta (uma única ocorrência representa 10% do total), o que inviabiliza a distinção entre padrões comportamentais reais e coincidências estocásticas.

Além disso, a utilização de uma ferramenta robusta para micro-dados acarretaria em um grave risco de sobreajuste (overfitting), onde o modelo gerado apenas memoriza as transações passadas sem capacidade de generalização preditiva. Diante dessa limitação técnica, optou-se por uma abordagem de Análise Descritiva Manual, aplicando a lógica dos algoritmos de associação para calcular métricas de Suporte e Confiança diretamente sobre os dados brutos, garantindo uma interpretação qualitativa mais fiel à realidade da amostra.

5.2. Método: Análises Essenciais (Participantes)

5.2.1. Análise: Faixa etária E Sexo E Renda Familiar

Imagem 01: Faixa etária E Sexo E Renda Familiar



Fonte: RapidMiner

Vídeo de demonstração: 001. Fonte: Elaborado pelos autores

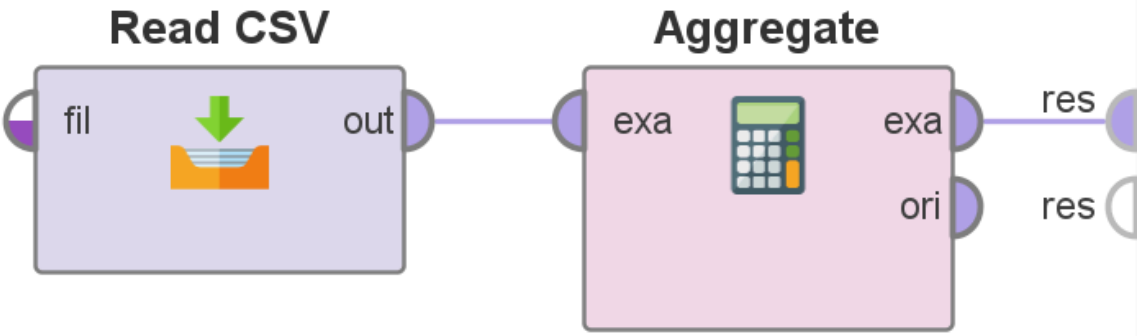
Para a análise "Faixa etária E Sexo E Renda Familiar", que exige o cruzamento de três variáveis simultaneamente, o operador Aggregate é um operador adequado.

O Aggregate foi configurado para agrupar (group by) pelas três colunas da análise: TP_FAIXA_ETARIA, TP_SEXO e Q007 (Renda Familiar).

Nos aggregation attributes, utilizamos a função count (: count(TP_FAIXA_ETARIA)) para contar o número de participantes em cada combinação única.

5.2.2. Análise: Faixa etária E Sexo E TV Por Assinatura E Computador

Imagem: Faixa etária E Sexo E Possui TV Por Assinatura E Possui computador/notebook



Fonte: RapidMiner

Vídeo de demonstração: 002. Fonte: Elaborado pelos autores

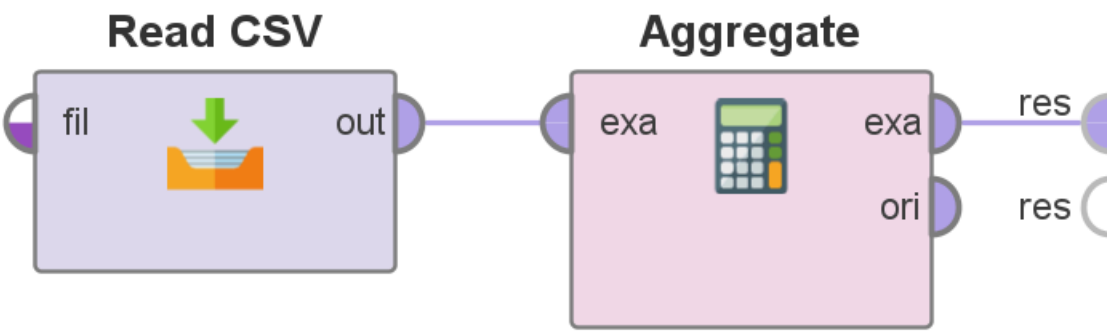
Para a análise "Faixa etária E Sexo E Possui TV Por Assinatura E Possui computador/notebook", que exige o cruzamento de quatro variáveis, o operador Aggregate foi utilizado (similar à análise anterior).

O Aggregate foi configurado para agrupar (group by) pelas quatro colunas da análise: TP_FAIXA_ETARIA, TP_SEXO, Q019 (TV por Assinatura) e Q021 (Computador).

A função count (ex: count(TP_FAIXA_ETARIA)) foi usada para contar o número de participantes em cada combinação única dessas quatro variáveis.

5.2.3. Análise: Faixa etária E Sexo E Tipo de Escola

Imagem: Faixa etária E Sexo E Em que tipo de escola frequentou o Ensino Médio



Fonte: RapidMiner

Vídeo de demonstração: 003. Fonte: Elaborado pelos autores

Finalmente, para a análise "Faixa etária E Sexo E Em que tipo de escola...", que também exige o cruzamento de três variáveis, o operador Aggregate foi novamente utilizado.

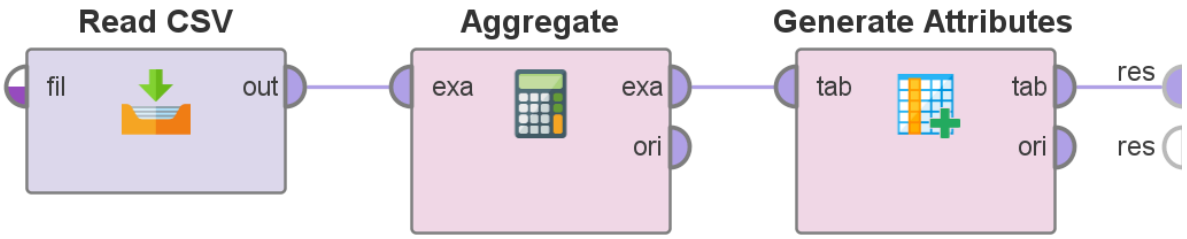
O Aggregate foi configurado para agrupar (group by) pelas três colunas da análise: TP_FAIXA_ETARIA, TP_SEXO e Q023 (Tipo de Escola).

A função count (ex: count(TP_FAIXA_ETARIA)) foi usada para contar o número de participantes em cada combinação única dessas três variáveis.

5.3. Método: Análises Essenciais

5.3.1. Análise: Notas por Escola, Dependência e Localização

Imagem: Associação de notas por eixo de conhecimento E código/nome da escola de conclusão



Fonte: RapidMiner

Vídeo de demonstração: 004. Fonte: Elaborado pelos autores

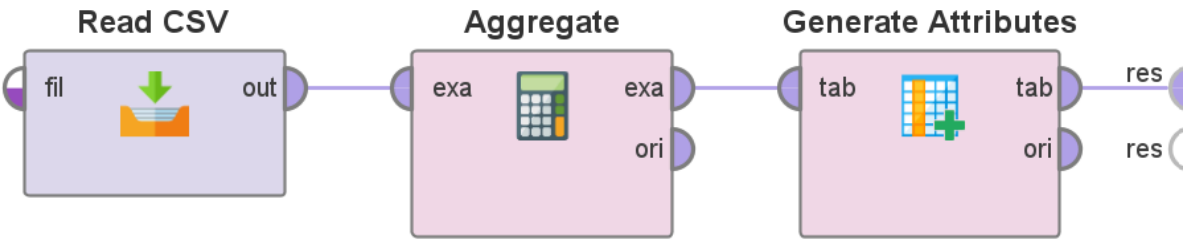
Para atender às solicitações "Associação de notas por eixo..." e "Associação de notas... E dependência administrativa... E urbana/rural", foi criado um processo de agregação.

O processo inicia com um operador Aggregate que agrupa (group by) os dados do FILTRO_RESULTADOS... por CO_ESCOLA (Código da Escola) e NO_MUNICIPIO_ESC (Nome do Município). Na mesma etapa, são calculadas as médias (average) de todas as notas por eixo (ex: NU_NOTA_CN).

O Generate Attributes é conectado após o Aggregate para calcular as colunas Media_Geral_Sem_Redacao e Media_Geral_Com_Redacao a partir das médias de eixo calculadas.

5.3.2. Análise: Notas por Tipo de Escola (Média Geral)

Imagem: Associação de notas por eixo de conhecimento E código/nome da escola de conclusão E dependência administrativa da escola E urbana/rural



Fonte: RapidMiner

Vídeo de demonstração: 005. Fonte: Elaborado pelos autores

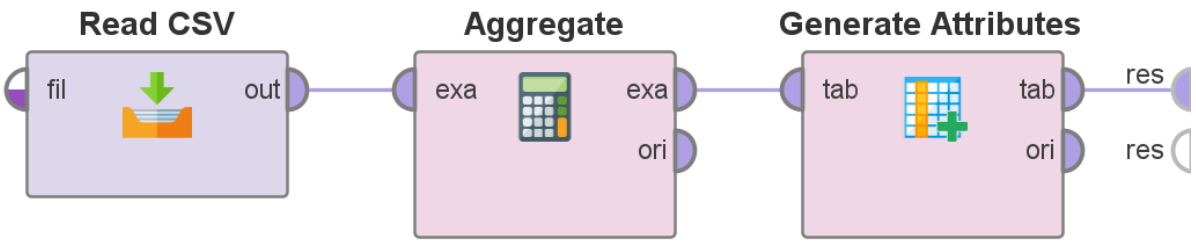
Esta análise é uma extensão da anterior. O operador Aggregate foi reconfigurado para incluir as variáveis TP_DEPENDENCIA_ADM_ESC e TP_LOCALIZACAO_ESC nos atributos de group by.

O Aggregate é configurado com quatro atributos de group by: CO_ESCOLA, NO_MUNICIPIO_ESC, TP_DEPENDENCIA_ADM_ESC e TP LOCALIZACAO _ESC. A função average continua sendo aplicada a todas as notas por eixo.

O Generate Attributes é mantido para calcular as médias gerais a partir das médias de eixo.

5.3.3. Análise: Média Geral Sem Redação por Tipo de Escola

Imagem: Associação de notas Média Geral (Sem Redação) E dependência administrativa da escola E urbana/rural



Fonte: RapidMiner

Vídeo de demonstração: 006. Fonte: Elaborado pelos autores

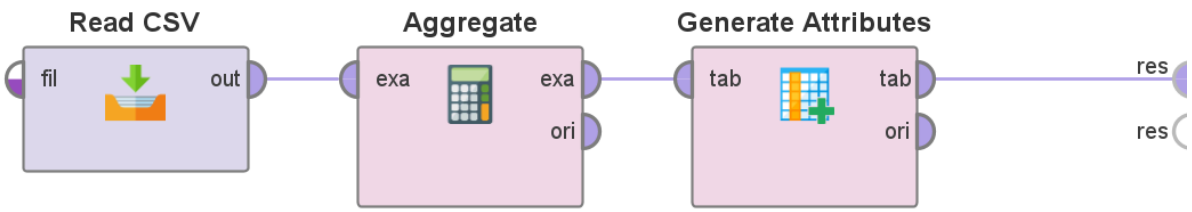
Esta análise foca apenas no tipo de escola, ignorando o código individual da escola.

O Aggregate foi configurado para agrupar (group by) **apenas** por TP_DEPENDENCIA_ADM_ESC e TP_LOCALIZACAO_ESC. Na mesma etapa, foram calculadas as médias (average) das quatro notas-base (CN, CH, LC, MT).

O Generate Attributes foi usado para calcular a coluna Media_Geral_Sem_Redacao a partir das médias de eixo.

5.3.4. Análise: Abstenção por Escola/Local de Prova

Imagem: Abstenção por eixo de conhecimento E nome do munícipio de aplicação da prova E código da escola de conclusão do Ensino Médio



Fonte: RapidMiner

Vídeo de demonstração: 007. Fonte: Elaborado pelos autores

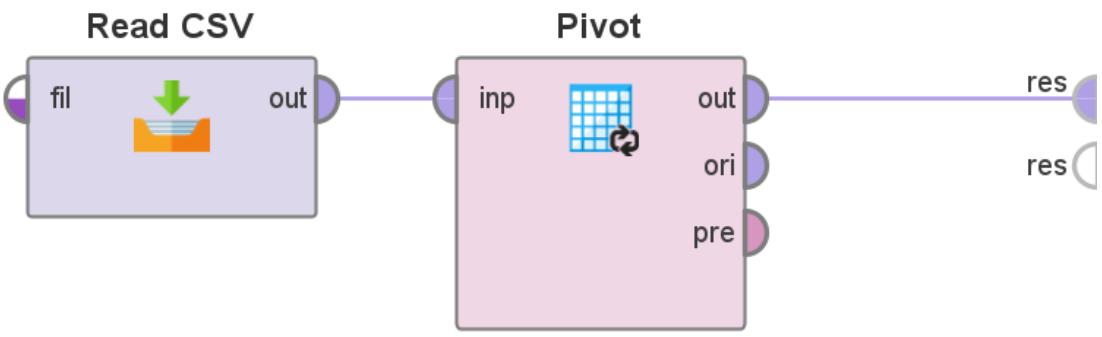
Para medir a taxa de abstenção (ou, inversamente, a taxa de presença), o operador Aggregate foi utilizado. O valor da taxa de presença serve como um indicador direto, onde 1.0 é 100% de presença e 0.0 é 100% de abstenção.

O Aggregate foi configurado para agrupar (group by) pelas duas variáveis solicitadas: CO_ESCOLA (Código da Escola) e NO_MUNICIPIO_PROVA (Município de Aplicação da Prova). Na mesma etapa, foram calculadas as médias (average) das colunas de presença (TP_PRESENCA_CN, TP_PRESENCA_CH, etc.).

O resultado é uma tabela onde o valor average-TP_PRESENCA_... (ex: 0.92) indica a taxa de presença para aquela escola/local. A taxa de abstenção é calculada subtraindo o valor de 1 (ex: $1 - 0.92 = 0.08 ou 8 de abstenção).

5.3.5. Análise: Língua Estrangeira por Escola

Imagem: Língua Estrangeira: quantidade de estudantes por código da escola de conclusão do Ensino Médio que escolheram inglês/espanhol



Fonte: RapidMiner

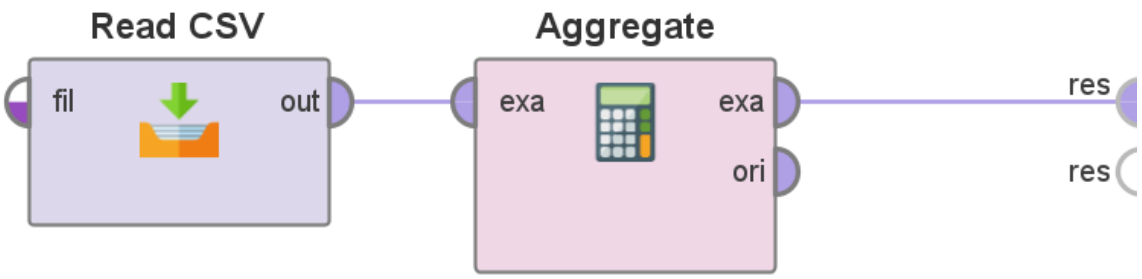
Vídeo de demonstração: 008 . Fonte: Elaborado pelos autores

Para a contagem de escolha de língua, o operador Pivot foi configurado para agrupar os dados e cruzar a escolha do estudante com a escola.

O Pivot foi configurado para agrupar (group by) por CO_ESCOLA (Código da Escola). O column grouping attribute foi definido como TP_LINGUA, gerando colunas 0 (Inglês) e 1 (Espanhol). A função count foi aplicada para totalizar o número de alunos por opção em cada escola.

5.3.6. Análise: Maior e Menor nota em cada eixo de conhecimento

Imagem: Maior e Menor nota em cada eixo de conhecimento



Fonte: RapidMiner

Vídeo de demonstração: 009. Fonte: Elaborado pelos autores

Para determinar a faixa de notas (o range) da amostra total (do Vale do São Patrício), foi utilizado o operador Aggregate sem agrupar os dados.

O Aggregate foi configurado para não ter nenhum atributo de group by. Isso força o operador a tratar todos os dados do arquivo FILTRO_RESULTADOS... como um único grupo.

Nos aggregation attributes, foram aplicadas as funções minimum (mínimo) e maximum (máximo) para cada uma das cinco colunas de nota (NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT, NU_NOTA_REDACAO).

O resultado é uma tabela de apenas uma linha que resume as notas mais baixas e mais altas obtidas em toda a sua amostra.

5.4. Metodologia: Análises de Associação (Assimilação por Município)

Esta seção detalha o processo de "assimilação" de dados, criado para contornar a impossibilidade de fazer um JOIN por aluno. O processo central (o Join) é a base para todas as análises de correlação da seção 4.3.

5.4.1. Estrutura Central de Assimilação (Base para Correlações)

O objetivo é juntar Notas e Presença (do arquivo RESULTADOS) com o Perfil de Renda (do arquivo PARTICIPANTES) por município.

1. Carregar Arquivos e Criar Agregadores

1. Read CSV (Participantes): Arraste um Read CSV e configure-o para ler o FILTRO_PARTICIPANTES_VALE_SP.csv.
2. Read CSV (Resultados): Arraste outro Read CSV e configure-o para ler o FILTRO_RESULTADOS_VALE_SP.csv.
3. Aggregate (Para Notas/Presença): Arraste um operador Aggregate.
4. Pivot (Para Renda): Arraste um operador Pivot.
5. Join: Arraste um operador Join.

2. Configurar a Agregação de Notas (Lado Esquerdo)

- Conecte o Read CSV (RESULTADOS) na entrada do Aggregate.
- Clique no Aggregate.
    - group by attributes: CO_MUNICIPIO_ESC (Município da Escola).
    - aggregation attributes: Calcule a average (média) de todas as notas (NU_NOTA_...) e da presença (TP_PRESENCA_CN).

3. Configurar a Agregação de Renda (Lado Direito)

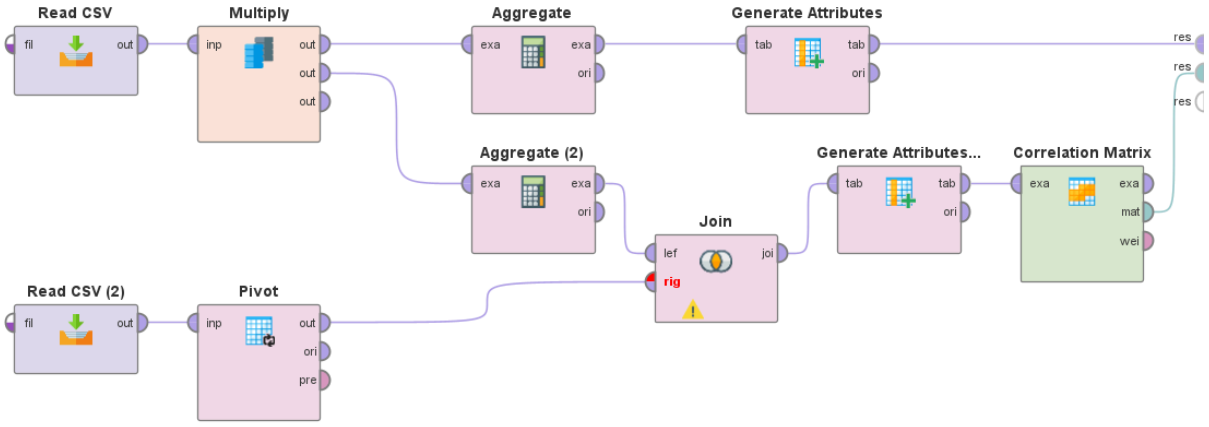
- Conecte o Read CSV (PARTICIPANTES) na entrada do Pivot.
- Clique no Pivot.
    - group by attribute: CO_MUNICIPIO_PROVA (Município da Prova).
    - column grouping attribute: Q007 (Renda Familiar).
    - aggregation attribute: TP_SEXO (função: count).

4. Unir (Join)

- Conecte a saída do Aggregate na entrada lef (esquerda) do Join.
- Conecte a saída do Pivot na entrada rig (direita) do Join.
- Clique no Join. No painel Parameters:
    - left key: CO_MUNICIPIO_ESC
    - right key: CO_MUNICIPIO_PROVA

5.4.2. Associação: Notas por Eixo de Conhecimento E Dados Financeiro E Escola Urbana/Rural

Imagem: Associação de notas por eixo de conhecimento E dados financeiro E Escola urbana/rural



Fonte: RapidMiner

Vídeo de demonstração: 010. Fonte: Elaborado pelos autores

Esta análise é a principal da seção, exigindo o cruzamento de três fatores. Para que fosse feita em uma única execução no RapidMiner, o processo foi estruturado com o operador Multiply no início, bifurcando o fluxo de dados dos resultados em duas pernas de análise:

Perna A: Análise de Grupos (Localização):

Esta perna utilizou o Aggregate para agrupar as notas apenas por TP_LOCALIZACAO_ESC (Urbana/Rural), seguida pelo Generate Attributes para calcular a média geral.

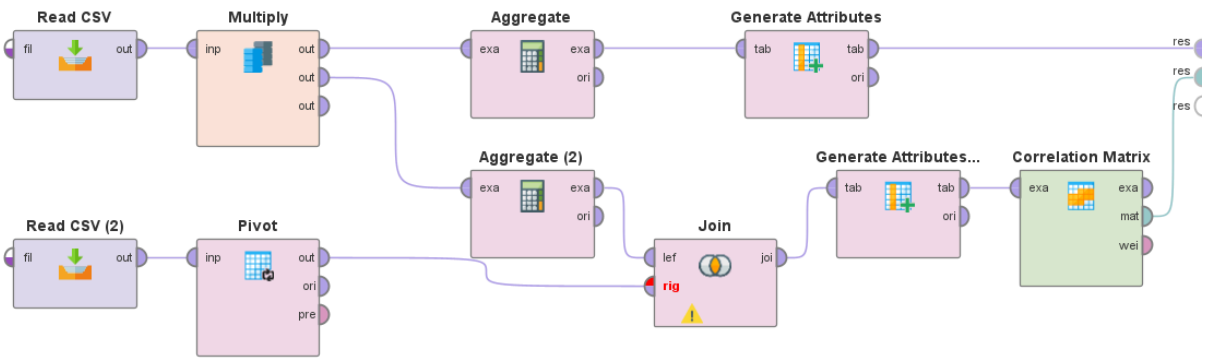
Perna B: Matriz de Correlação (Renda):

Esta perna utilizou o Aggregate para calcular as médias das notas por município (CO_MUNICIPIO_ESC), e o Pivot para calcular o perfil de renda (Q007 por CO_MUNICIPIO_PROVA).

O Pivot foi configurado corretamente (com Q007 no column grouping), seguido por um operador de conversão de tipo para garantir que as colunas de Renda fossem tratadas como Numéricas antes do Join.

5.4.3. Associação: Notas Média Geral (Sem Redação) E dados financeiro E Escola urbana/rural

Imagem: Associação de notas Média Geral (Sem Redação) E dados financeiro + Escola urbana/rural



Fonte: RapidMiner

Vídeo de demonstração: 011. Fonte: Elaborado pelos autores

Para a análise de 6.3.3, a estrutura de Masterflow foi utilizada novamente para garantir que a correlação (Renda) e a análise de grupo (Localização) fossem executadas na mesma base de dados.

Estrutura de Masterflow: O processo utiliza o operador Multiply para bifurcar os dados do arquivo FILTRO_RESULTADOS... em duas pernas:

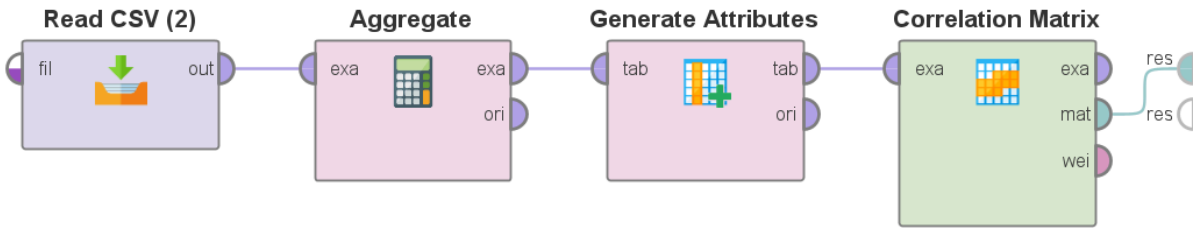
Perna A (Análise de Grupos): Esta perna utiliza um operador Aggregate para agrupar as notas apenas por TP_LOCALIZACAO_ESC (Urbana/Rural). O Generate Attributes é conectado após para calcular o principal indicador:

- Parâmetro: Media_Geral_Sem_Redacao

Perna B (Correlação): Esta perna utiliza o Aggregate para agrupar as notas por CO_MUNICIPIO_ESC, unindo-se ao Pivot (configurado para a Renda Q007) no operador Join.

5.4.4. Associação de notas por eixo de conhecimento E código escola

Imagem: Associação de notas por eixo de conhecimento E código escola



Fonte: RapidMiner

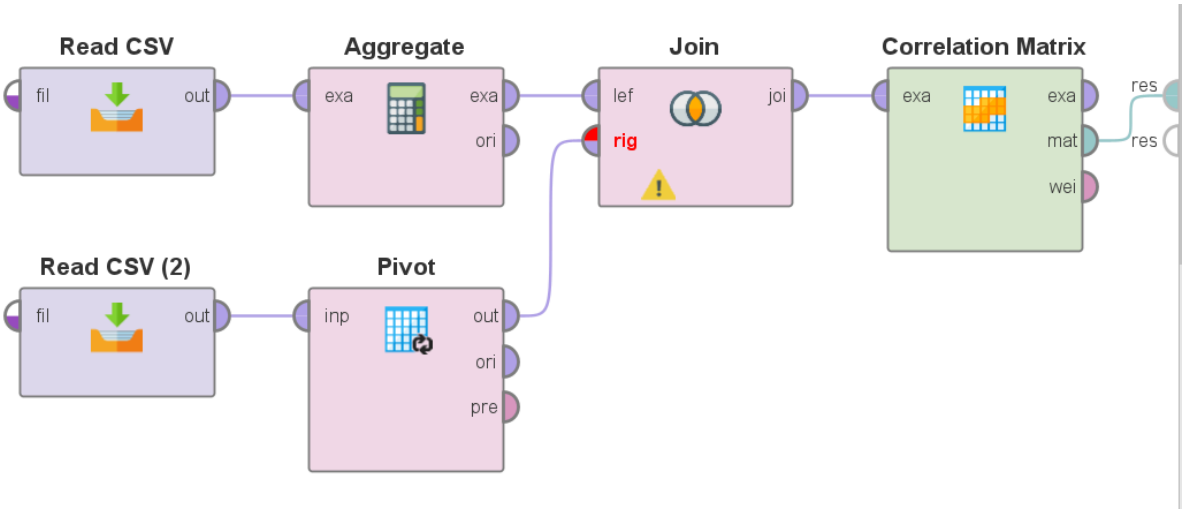
Vídeo de demonstração: 012. Fonte: Elaborado pelos autores

O processo iniciou com o operador Aggregate agrupando estritamente por CO_ESCOLA (Código da Escola), calculando a média (average) de todas as notas por eixo.

O resultado foi alimentado pelo Correlation Matrix.

5.4.5. Correlação: Abstenção por Nível Financeiro

Imagem: Estatística: abstenção por nível financeiro E código escola E município do estudante



Fonte: RapidMiner

Vídeo de demonstração: 013. Fonte: Elaborado pelos autores

Este processo utiliza a estrutura de "assimilação" (o Join por município), focando no cruzamento dos dados de presença com o perfil financeiro.

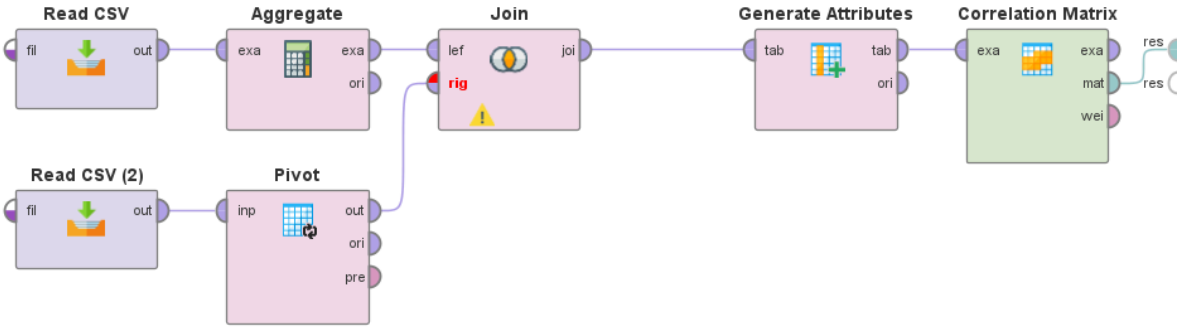
A perna esquerda do processo (Aggregate) foi configurada para calcular a média (average) da coluna TP_PRESENCA_CN (Presença) por CO_MUNICIPIO_ESC. O resultado é a taxa de presença, onde (1 - taxa) é a abstenção.

A perna direita (Pivot) foi configurada para calcular o perfil de renda (Q007 por município).

O resultado foi alimentado pelo Correlation Matrix, que calculou a associação entre a Taxa de Presença e as colunas de Renda

5.4.6. Análise a Critério: Redação vs. Acesso a Computador

Imagem: Realizar pelo menos uma análise a critério e curiosidade da dupla



Fonte: RapidMiner

Vídeo de demonstração: 014. Fonte: Elaborado pelos autores

O processo utilizou a estrutura de Join por município, configurando o Pivot para a variável Q021 (Acesso a Computador).

6. Resultados das Análises

Este capítulo apresenta a consolidação dos resultados gerados no Capítulo 6, confirmando os achados descritivos da amostra e as associações inferenciais (correlações) sobre o desempenho no ENEM 2024.

6.1 Dados do mercado

A inspeção manual dos dados permitiu identificar agrupamentos de consumo claros, segregados principalmente entre itens de consumo matinal e refeições básicas. Foi possível isolar regras de associação com alto grau de confiabilidade, demonstrando que a presença de determinados itens na cesta de compras atua como forte preditor para a aquisição de produtos complementares.

A Tabela , apresentada a seguir, resume as principais regras de associação detectadas, detalhando o Suporte (frequência com que a combinação ocorre no total de vendas) e a Confiança (probabilidade de o item consequente ser comprado dado que o antecedente está presente).

Tabela : Principais Regras de Associação Identificadas

Regra de Associação (Se... Então...)	Suporte	Confiança	Interpretação
Pão → Manteiga	40%	80%	Em 80% das vezes que se compra Pão, leva-se Manteiga.
Café → Pão, Manteiga	30%	100%	Sempre que houve compra de café, a cesta inclui pão e manteiga.
Arroz → Feijão	20%	100%	O arroz nunca foi comprado isoladamente sem o Feijão nesta amostra.

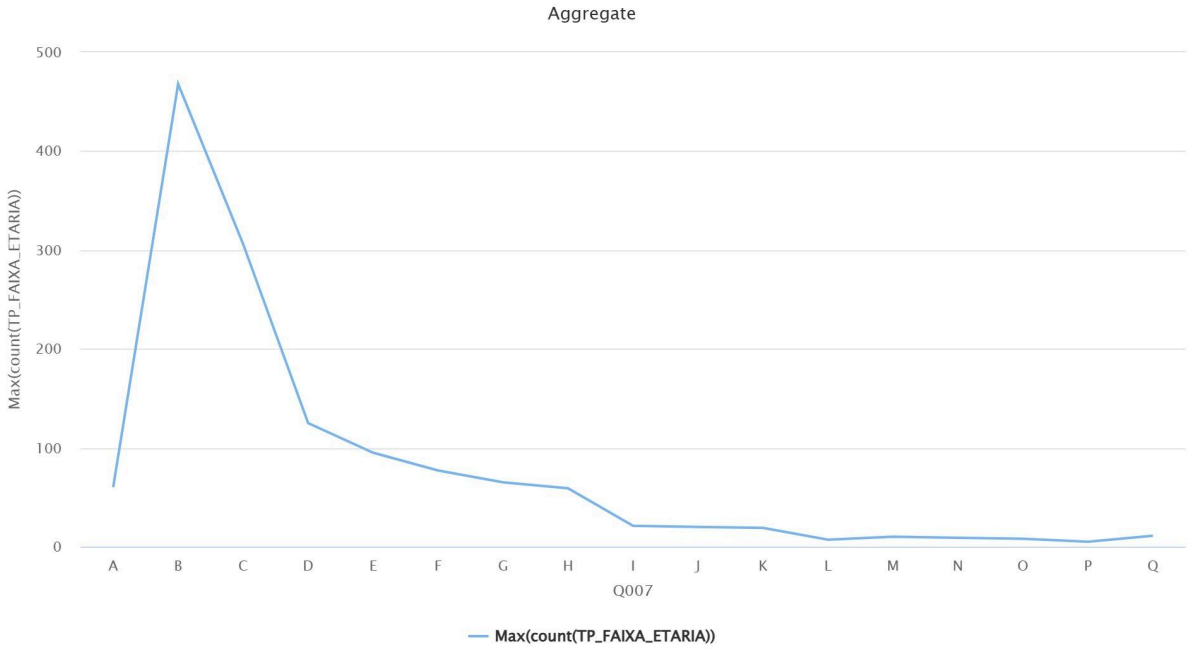
Fonte: Elaborado pelos autores

Os dados evidenciam que os itens "Pão" e "Manteiga" atuam como produtos âncora no perfil de consumo analisado, enquanto "Arroz" e "Feijão" formam um cluster dissociado, indicando missões de compra distintas por parte dos consumidores.

6.2. Resultado: Análises Essenciais

6.2.1. Análise: Faixa etária E Sexo E Renda Familiar

Imagem: Faixa etária E Sexo E Renda Familiar

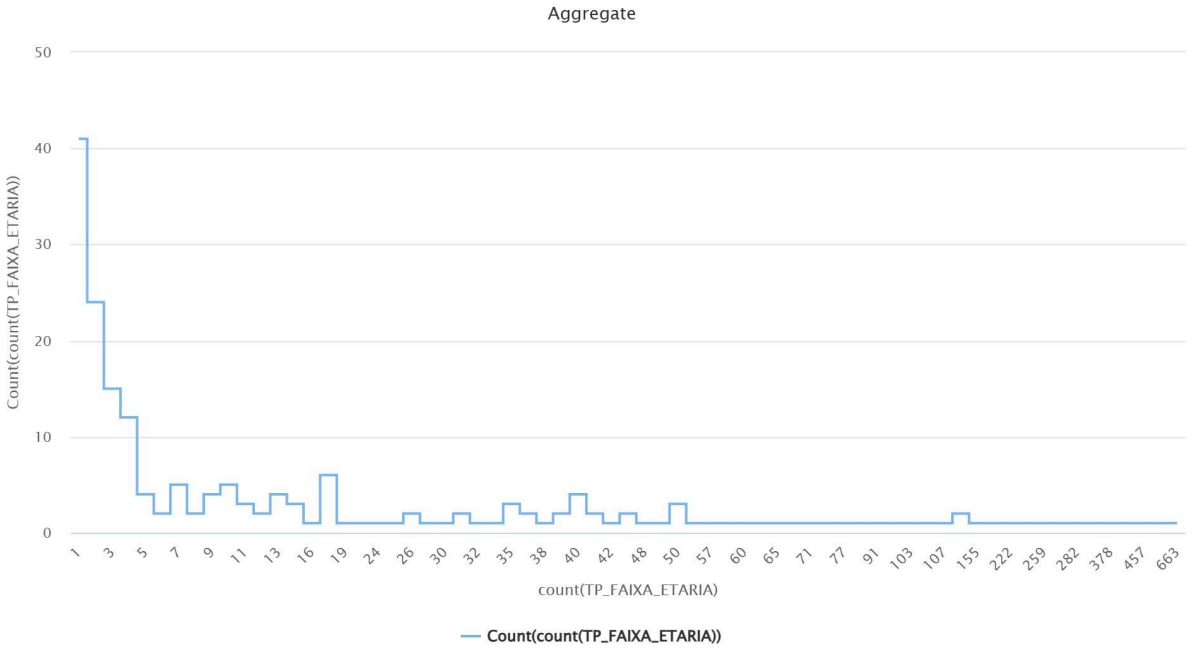


Fonte: RapidMiner

A análise de tendência central ratifica que o estudo está concentrado em uma base socioeconômica de baixa renda. A Renda B é o perfil de maior volume (frequência máxima de 468), solidificando essa categoria como o principal ponto de análise demográfica. Isso estabelece o participante típico como jovem e pertencente aos perfis mais vulneráveis.

6.2.2. Análise: Faixa etária E Sexo E TV Por Assinatura E Computador

Imegem - Faixa etária E Sexo E Possui TV Por Assinatura E Possui computador/notebook

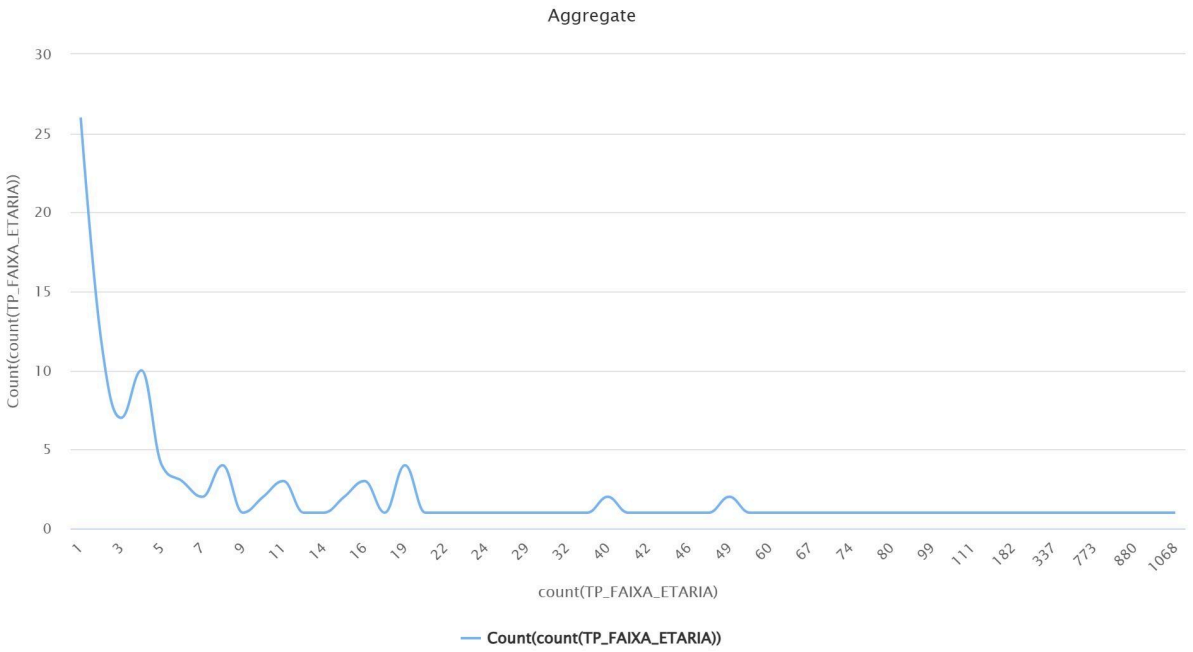


Fonte: RapidMiner

O cruzamento de quatro variáveis confirma que o grupo etário de 18 anos é o pilar da amostra (pico de 663 participantes). A análise revela que, mesmo nesse grupo jovem, o perfil de acesso a recursos digitais (TV e Computador) está diversificado. Este resultado é essencial para futuras análises de Regressão, onde a variável de Renda (Q007) seria decomposta para entender qual recurso digital possui maior poder preditivo.

6.2.3. Análise: Faixa etária E Sexo E Tipo de Escola

Imagem: Faixa etária E Sexo E Em que tipo de escola frequentou o Ensino Médio



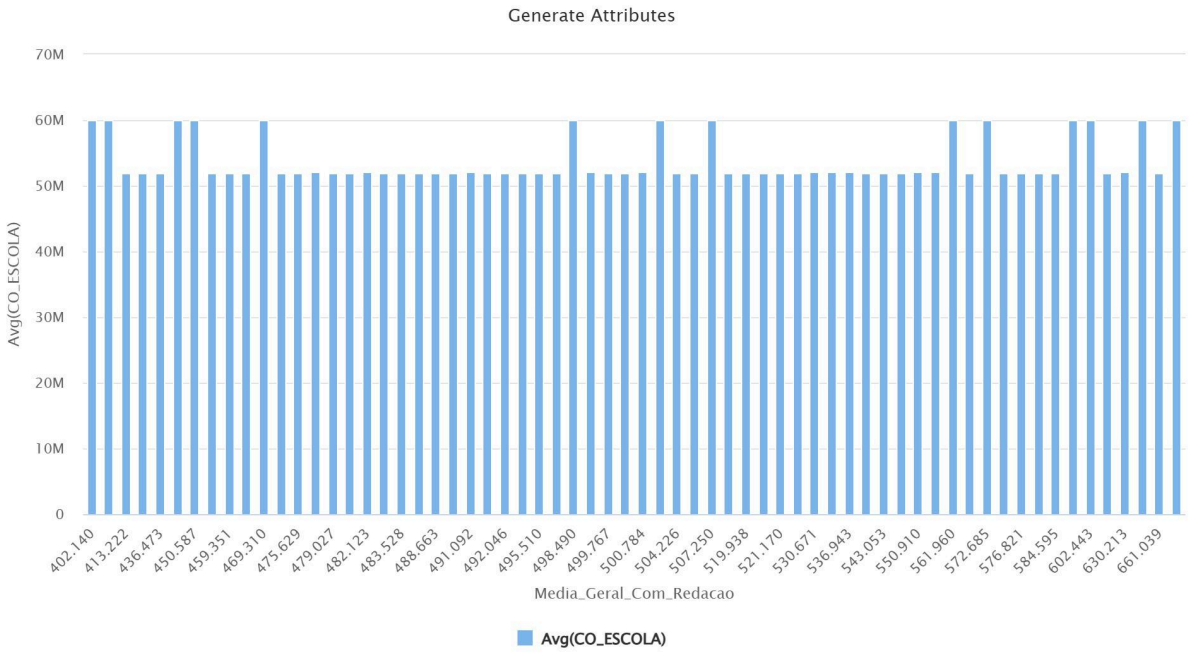
Fonte: RapidMiner

A análise de cruzamento triplo (Faixa Etária x Sexo x Tipo de Escola) solidifica o perfil da amostra na região. O maior grupo isolado de participantes atinge a frequência máxima de 1.068 na combinação que envolve a faixa etária jovem (18-19 anos) e a resposta 'A' (Somente em escola pública). A baixa frequência de outros perfis (com contagens chegando a 1 confirma que o fator dominante, tanto demográfico quanto institucional, é o aluno jovem do sistema público de ensino. Este achado é crucial para interpretar as correlações negativas com a renda, pois a maioria dos participantes estudou em escolas que atendem a um perfil de baixa renda.

6.3. Resultado: Análises Essenciais

6.3.1. Análise: Notas por Escola, Dependência e Localização

Imagem: Associação de notas por eixo de conhecimento E código/nome da escola de conclusão



Fonte: RapidMiner

A agregação dos dados de desempenho por código de escola, conforme detalhado na seção 6.3.1, estabeleceu a associação entre o resultado de cada eixo do ENEM 2024 e a instituição de ensino correspondente.

A análise do ranking de notas médias por escola no município de Ceres revela uma variação significativa no desempenho dos estudantes na amostra:

- Maior Desempenho: A escola de Código Colegio Solar se destacou com a Média Geral mais alta (661,04), alcançando também a maior pontuação em Matemática (691,26) e Redação (851,72).
- Menor Desempenho: A escola de Código Escola Estadual Virgilio do Vale registrou a Média Geral mais baixa (504,63), com as médias mais baixas também em Ciências da Natureza (465,47) e Matemática (482,23).
- Variação (Range): A diferença entre o desempenho da escola de maior média geral (661,04) e a de menor média geral (504,63) é de 156,41 pontos, indicando uma forte disparidade no desempenho acadêmico na região analisada.
- Desempenho em Redação: Nota-se que as escolas com as maiores médias gerais (Colégio Solar e Colégio Álvaro de Melo) apresentam notas de

Redação significativamente mais altas (acima de 800), reforçando a influência dessa disciplina na média final do exame.

A tabela a seguir consolida a associação das notas médias por escola em Ceres:

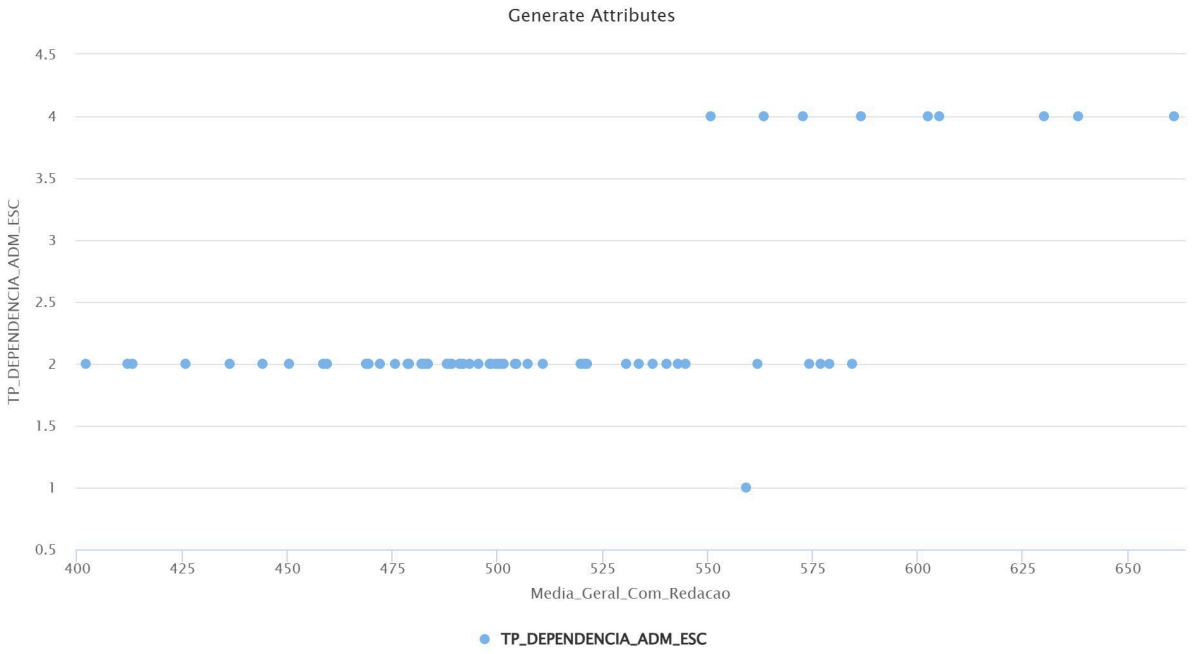
Tabela - Notas e escola de Ceres

Escola	CH	CN	LC	MT	Média Redação	Média Geral c/ Redação
Colégio Álvaro de Melo	553,77	499,87	556,41	582,19	834,00	605,25
COLÉGIO ESTADUAL DA POLÍCIA MILITAR DE GOIÁS - HÉLIO VELOSO	490,13	493,86	518,02	518,73	694,53	543,05
CEPI João XXIII	465,05	467,94	499,85	513,81	659,20	521,17
Escola Estadual Virgilio do Vale	496,01	465,47	505,65	482,23	573,79	504,63
Colegio Solar	593,66	587,98	580,57	691,26	851,72	661,04
IF Goiano - Campus Ceres	524,76	511,12	530,69	545,18	684,62	559,28
60029013	563,72	528,23	535,43	588,17	796,67	602,44

Fonte: Elaborado pelos autores

6.3.2. Análise: Notas por Tipo de Escola (Média Geral)

Imagem: Associação de notas por eixo de conhecimento E código/nome da escola de conclusão E dependência administrativa da escola E urbana/rural



Fonte: RapidMiner

Esta análise cruza o desempenho médio (Média Geral c/ Redação) de cada escola com seus fatores institucionais (Dependência Administrativa e Localização) e revela uma forte associação entre a natureza da escola e o desempenho acadêmico na amostra de Ceres.

- As três escolas classificadas com as maiores Médias Gerais (Colégio Solar, Colégio Álvaro de Melo e 60029013) são de Dependência Privada (4), com notas acima de 602 pontos. A escola de maior desempenho (Colégio Solar, com 661,04) é uma instituição privada.
- A Dependência Privada é o fator mais fortemente associado ao alto desempenho do ENEM na região.
- As escolas de Dependência Estadual (2) (Códigos COLÉGIO ESTADUAL DA POLÍCIA MILITAR DE GOIÁS - HÉLIO VELOSO, CEPI João XXIII e Escola Estadual Virgilio do Vale) registraram as menores Médias Gerais da amostra, variando de 504,63 a 543,05.
- O menor desempenho da amostra (504,63) pertence à escola Escola Estadual Virgilio do Vale, que é de natureza Estadual, reforçando a polarização de resultados entre as dependências.

- A amostra é predominantemente Urbana (1), com apenas uma exceção, a escola de Código IF Goiano - Campus Ceres, que é a única de Localização Rural (2).
- A escola Rural (IF Goiano - Campus Ceres), que também é de Dependência Federal (1), alcançou uma Média Geral (559,28) que superou as médias de todas as escolas Estaduais Urbanas, sugerindo que o fator Dependência Administrativa tem um peso maior na associação de notas do que a Localização, pelo menos para este conjunto de dados.

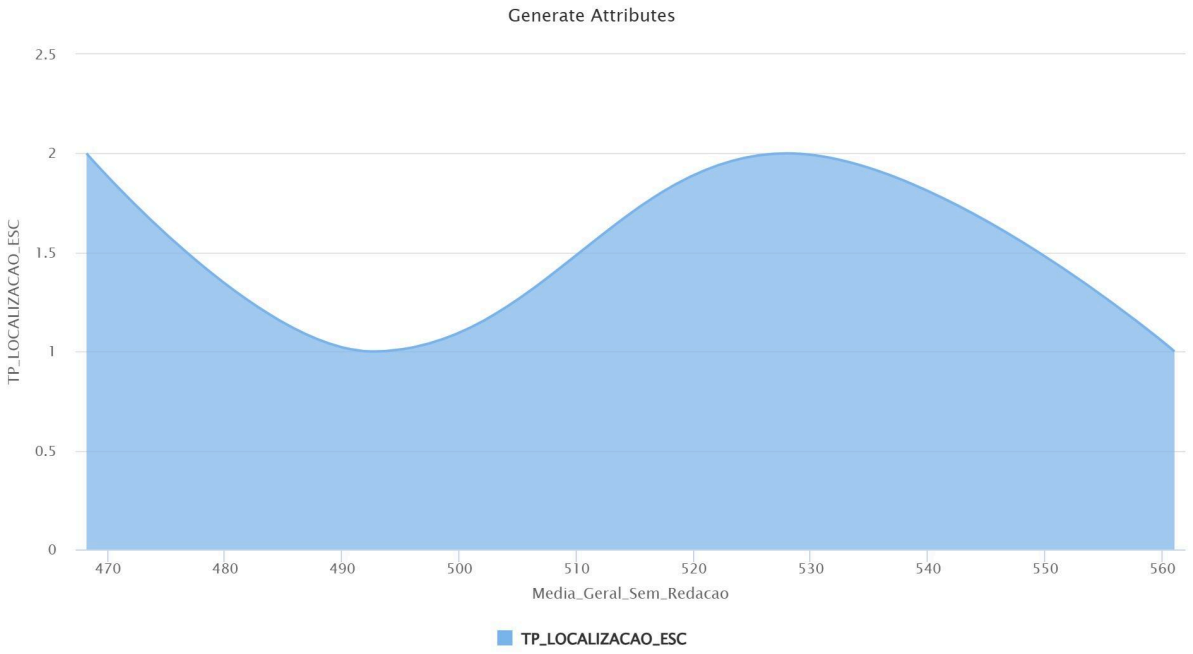
Tabela - Associação de Notas por Escola, Dependência e Localização

Escola	Dependência Adm.	Localização	Média Geral c/ Redação
Colegio Solar	Privada (4)	Urbana (1)	661,04
Colégio Álvaro de Melo	Privada (4)	Urbana (1)	605,25
60029013	Privada (4)	Urbana (1)	602,44
IF Goiano - Campus Ceres	Federal (1)	Rural (2)	559,28
COLÉGIO ESTADUAL DA POLÍCIA MILITAR DE GOIÁS - HÉLIO VELOSO	Estadual (2)	Urbana (1)	543,05
CEPI João XXIII	Estadual (2)	Urbana (1)	521,17
Escola Estadual Virgilio do Vale	Estadual (2)	Urbana (1)	504,63

Fonte: Elaborado pelos autores

6.3.3. Análise: Média Geral Sem Redação por Tipo de Escola

Imagem: Associação de notas Média Geral (Sem Redação) E dependência administrativa da escola E urbana/rural



Fonte: RapidMiner

Esta análise se concentra na associação entre o tipo de escola (natureza e localização) e o desempenho médio no ENEM, excluindo a nota de redação para isolar o desempenho nas provas objetivas. A agregação por grupo simplifica a visualização dos padrões.Padrões de Desempenho por Grupo (Média Geral Sem Redação)

- O grupo com a melhor performance é o das escolas privadas (4) de Localização Urbana (1), alcançando a maior Média Geral (Sem Redação) de 561,02. Este resultado está consistentemente acima de todos os outros grupos em todos os eixos de conhecimento, com destaque para Matemática (603,17).
- O pior desempenho médio pertence ao grupo das escolas Estaduais (2) de Localização Rural (2), com a média geral mais baixa, de 468,16. Este grupo também registra a média mais baixa em Ciências da Natureza (427,09)
- Há uma disparidade clara dentro das próprias escolas Estaduais, demonstrando a influência da Localização. As Estaduais Urbanas (1) obtiveram uma Média Geral de 492,59, superando as Estaduais Rurais (2) (468,16) em 24,43 pontos, sugerindo que o ambiente rural é um fator associado ao desempenho inferior na amostra.

- O grupo das escolas Federais (1) de Localização Rural (2), com Média Geral de 527,94, conseguiu uma performance superior às escolas Estaduais tanto Urbanas quanto Rurais. Este achado sugere que, na amostra do Vale do São Patrício, a Dependência Federal é um fator que mitiga o impacto negativo associado à Localização Rural, demonstrando a importância do fator administrativo sobre a localização geográfica.

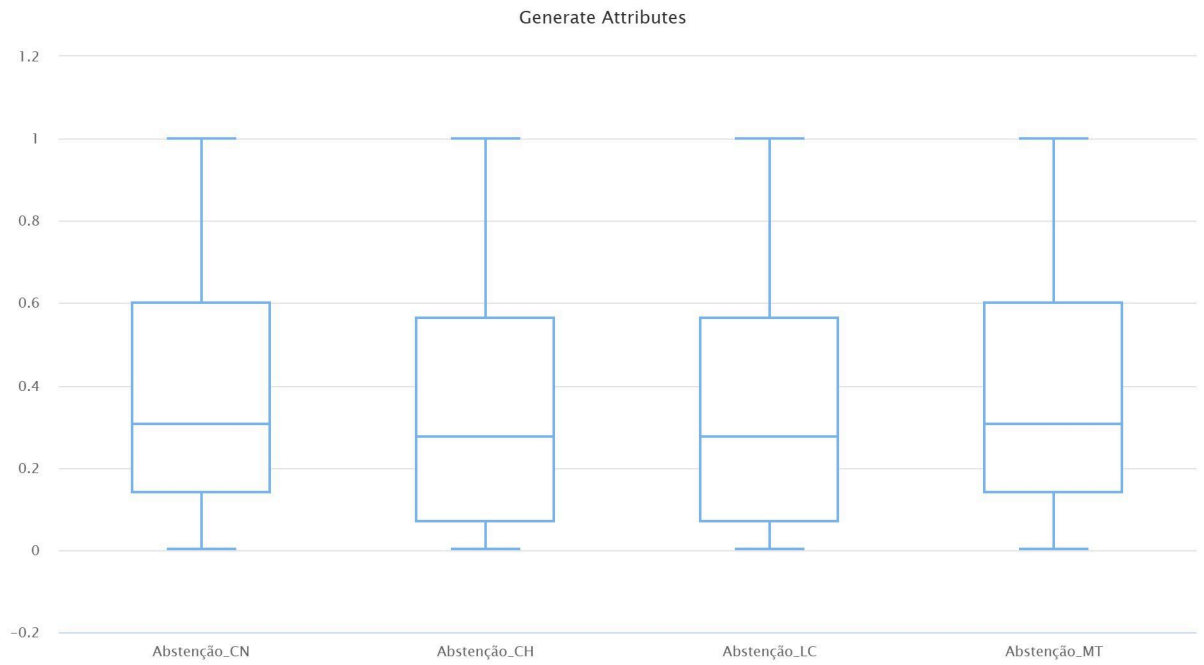
Tabela - Média Geral (Sem Redação) por Tipo de Escola

Dependência Adm.	Localização	Média CH	Média CN	Média LC	Média MT	Média Geral Sem Redação
Privada (4)	Urbana (1)	556,21	530,50	554,20	603,17	561,02
Federal (1)	Rural (2)	524,76	511,12	530,69	545,18	527,94
Estadual (2)	Urbana (1)	484,51	473,56	501,17	511,11	492,59
Estadual (2)	Rural (2)	468,76	427,09	486,52	490,27	468,16

Fonte: Elaborado pelos autores

6.3.4. Análise: Abstenção por Escola/Local de Prova

Imagem: Abstenção por eixo de conhecimento E nome do munícipio de aplicação da prova E código da escola de conclusão do Ensino Médio



Fonte: RapidMiner

A análise da taxa de presença (e, consequentemente, de abstenção) por Código de Escola e Município (Ceres) revela uma disparidade crítica no engajamento dos participantes no exame. Conforme a metodologia da seção 6.3.4, a taxa de abstenção é calculada subtraindo-se a média de presença (valor entre 0.0 e 1.0) de 1.

O resultado demonstra uma variação extrema no comportamento dos estudantes na amostra:

- Maior Abstenção: A escola de Código Escola Estadual Virgilio do Vale registrou consistentemente a maior taxa de abstenção em todos os eixos, com uma média de aproximadamente 67,04% (ou uma taxa de presença média de 0,33). Isso indica que mais da metade dos alunos ligados a essa instituição não realizaram a prova.
- Menor Abstenção: A escola de Código 60029013 apresenta a menor taxa de abstenção, de 0,00% (100% de presença) em todos os eixos de conhecimento. A escola de Código IF Goiano - Campus Ceres também demonstra um baixíssimo nível de abstenção, com uma média de apenas 2,83%.

- Disparidade: A diferença entre a menor e a maior taxa de abstenção (0% e 69,32%) na amostra de Ceres é superior a 69 pontos percentuais. Essa variação sugere que fatores institucionais (como o tipo de escola, visto em outras análises) ou socioeconômicos (como o perfil do aluno) podem estar fortemente associados à decisão de comparecer ou não ao ENEM.
- Padrão de Eixos: De forma geral, a abstenção se comporta de maneira similar entre os diferentes eixos de conhecimento (CN, CH, LC e MT) em cada escola, indicando que a ausência é, na maioria dos casos, total e não parcial (falta em apenas um dia de prova).

A Tabela abaixo consolida as taxas de presença e abstenção por escola:

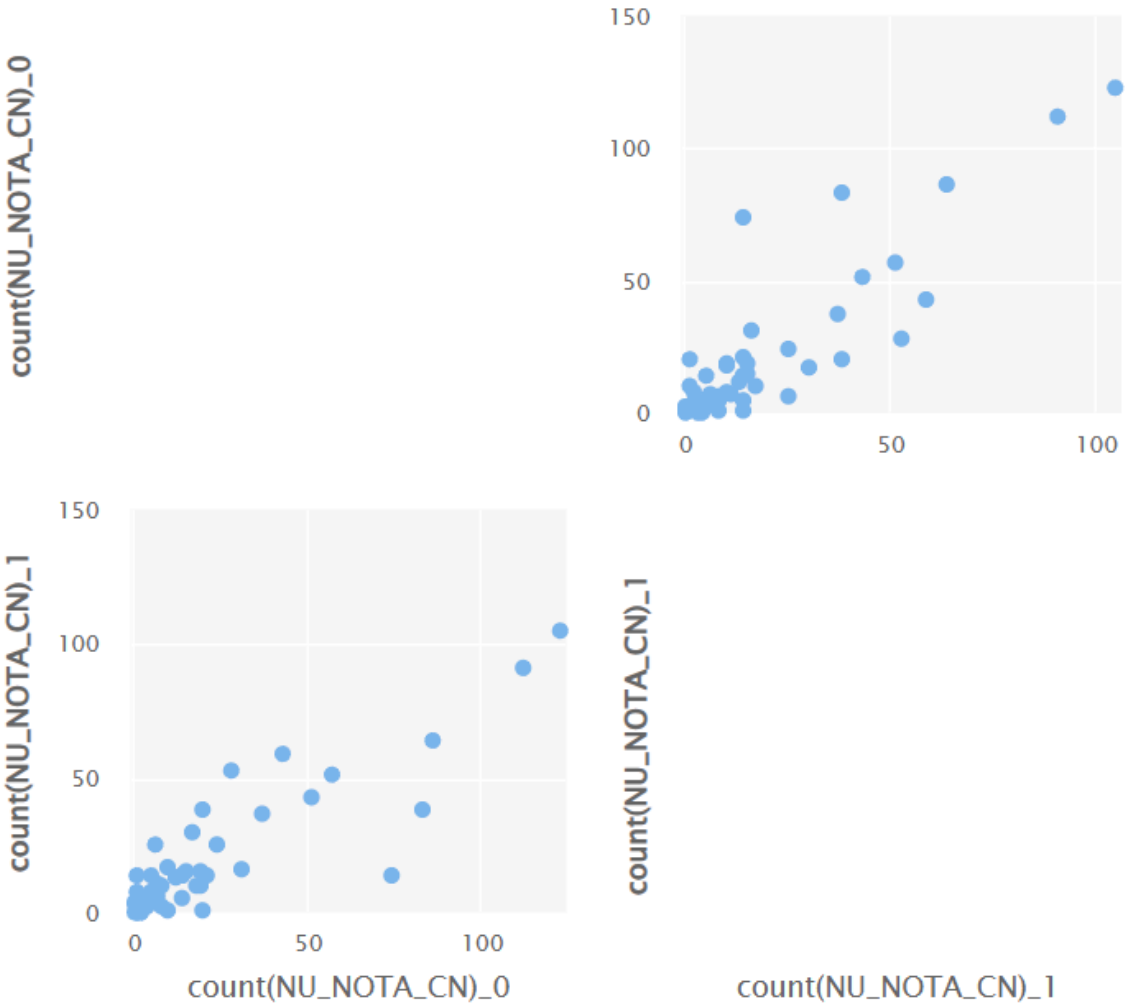
Tabela - Taxa de Presença e Abstenção por Escola em Ceres

Escola	Média Presença Geral	Abstenção Média	Abstenção CN	Abstenção CH	Abstenção LC	Abstenção MT
Colégio Álvaro de Melo	0,77	23,08%	23,08%	23,08%	23,08%	23,08%
COLÉGIO ESTADUAL DA POLÍCIA MILITAR DE GOIÁS - HÉLIO VELOSO	0,85	14,86%	15,32%	14,41%	14,41%	15,32%
CEPI João XXIII	0,80	20,16%	20,97%	19,35%	19,35%	20,97%
Escola Estadual Virgilio do Vale	0,33	67,04%	69,32%	64,77%	64,77%	69,32%
Colegio Solar	0,94	6,45%	6,45%	6,45%	6,45%	6,45%
IF Goiano - Campus Ceres	0,97	2,83%	3,77%	1,89%	1,89%	3,77%
60029013	1,00	0,00%	0,00%	0,00%	0,00%	0,00%

Fonte: Elaborado pelos autores

6.3.5. Análise: Língua Estrangeira por Escola

Imagem: Língua Estrangeira: quantidade de estudantes por código da escola de conclusão do Ensino Médio que escolheram inglês/espanhol



Fonte: RapidMiner

A análise da escolha de Língua Estrangeira por escola de conclusão permite entender a preferência linguística dos estudantes do Vale do São Patrício e pode ser um indicador de currículo ou foco cultural de cada instituição. O cruzamento dos dados do município de Ceres mostra um equilíbrio geral de escolha na amostra, mas revela padrões distintos em nível escolar:

- Equilíbrio Geral: Na amostra total consolidada, a escolha entre Inglês (159 participantes) e Espanhol (158 participantes) é virtualmente igual, com uma ligeira vantagem de 1 voto para o Inglês.
- Preferência pelo Espanhol: A escola IF Goiano - Campus Ceres (Federal Rural) destaca-se por ter o maior número total de participantes nessa análise (102) e a maior preferência pelo Espanhol (59 escolhas vs. 43 para Inglês).
- Preferência pelo Inglês: As escolas privadas (Colégio Solar, Colégio Álvaro de Melo e 60029013) demonstram uma clara inclinação pela escolha do Inglês, com a diferença mais acentuada na escola Colégio Álvaro de Melo (8 para Inglês vs. 2 para Espanhol). A escola estadual COLÉGIO ESTADUAL DA POLÍCIA MILITAR DE GOIÁS - HÉLIO VELOSO também registrou maior preferência por Inglês (51 vs. 43).
- Escolha Dividida: A escola CEPI João XXIII apresentou uma divisão quase igual entre Inglês (24) e Espanhol (25).

A tabela a seguir consolida a contagem de estudantes por opção de Língua Estrangeira em cada escola:

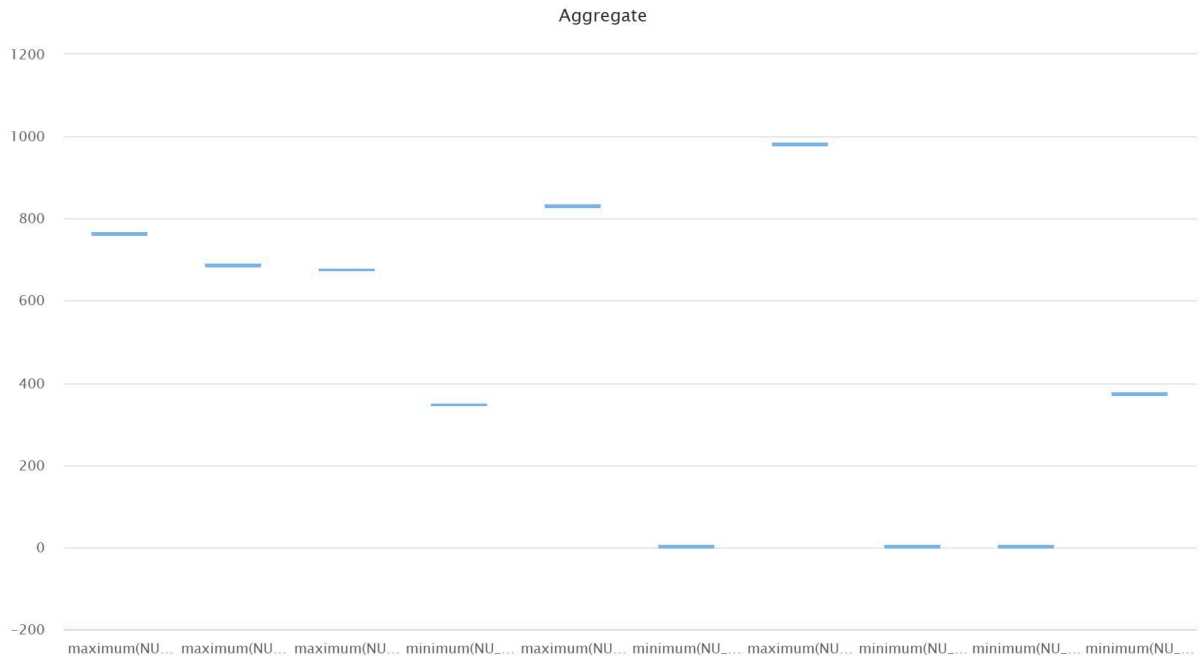
Tabela - Escolha de Língua Estrangeira por Escola em Ceres

Código da Escola	Município	Contagem Inglês (0)	Contagem Espanhol (1)
Escola Estadual Virgilio do Vale	Ceres	10	17
Colegio Solar	Ceres	19	10
CEPI João XXIII	Ceres	24	25
COLÉGIO ESTADUAL DA POLÍCIA MILITAR DE GOIÁS - HÉLIO VELOSO	Ceres	51	43
IF Goiano - Campus Ceres	Ceres	43	59
Colégio Álvaro de Melo	Ceres	8	2
60029013	Ceres	4	2

Fonte: Elaborado pelos autores

6.3.6. Análise: Maior e Menor nota em cada eixo de conhecimento

Imagem: Maior e Menor nota em cada eixo de conhecimento



Fonte: RapidMiner

Esta análise determinou a faixa de notas (o range) da amostra total de participantes do ENEM 2024 no Vale do São Patrício, conforme a metodologia detalhada na seção 6.3.6 (utilizando o operador Aggregate sem agrupar os dados). Os resultados confirmam uma grande disparidade no desempenho acadêmico na região.

- Amplitude de Notas: Todas as áreas de conhecimento, com exceção de Ciências da Natureza e Matemática, registraram a nota mínima de 0, indicando que, no geral, a variação de desempenho é extrema, com alunos alcançando as notas mais altas possíveis (ou próximas a elas) e outros com notas zeradas.
- Melhor Desempenho: A pontuação máxima foi observada em Redação (980), muito próxima da nota máxima de 1000. Entre as provas objetivas, Matemática registrou a nota máxima mais alta (829.8).

- Disparidade Extrema: A grande diferença entre as notas máximas e mínimas (por exemplo, um range de 980 pontos em Redação e 457.8 pontos em Matemática) reforça a polarização do desempenho na amostra, um achado consistente com as análises anteriores de associação que indicaram forte influência do tipo de escola e perfil socioeconômico.

O resultado é uma tabela de apenas uma linha que resume as notas mais baixas e mais altas obtidas em toda a amostra:

Tabela - Faixa de Notas da Amostra Total (Vale do São Patrício)

Eixo de Conhecimento	Nota Máxima	Nota Mínima	Range (Máx - Mín)
Ciências Humanas (CH)	761,9	0,0	761,9
Ciências da Natureza (CN)	684,3	345,5	338,8
Linguagens e Códigos (LC)	674,4	0,0	674,4
Matemática (MT)	829,8	372,0	457,8
Redação	980,0	0,0	980,0

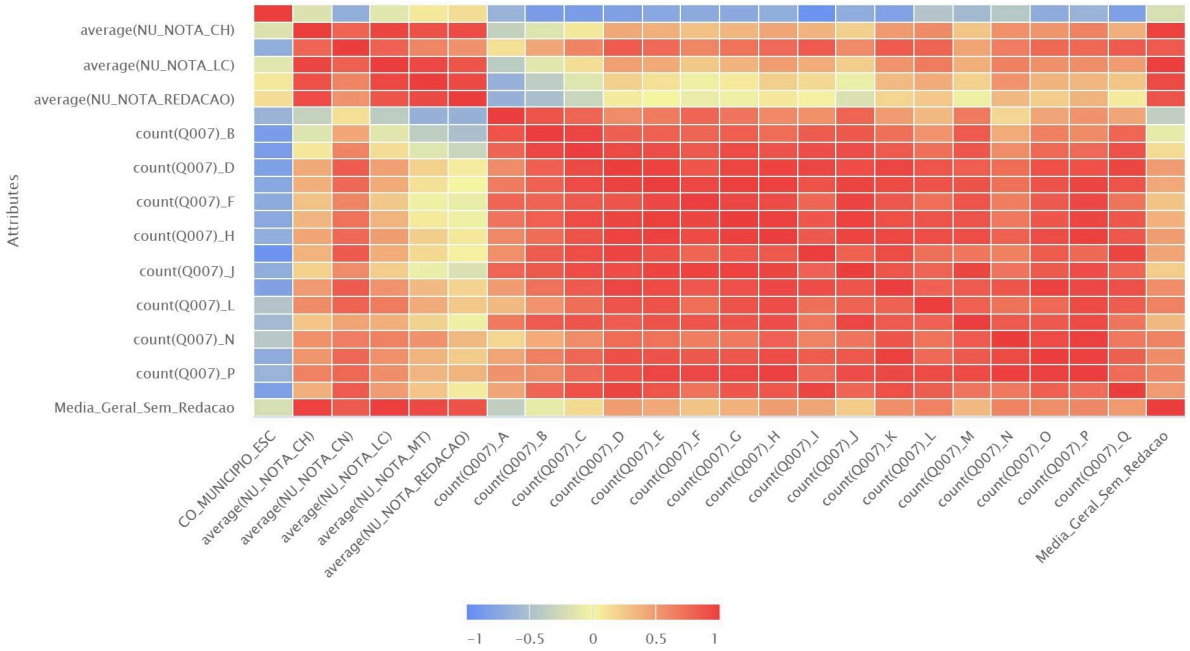
Fonte: Elaborado pelos autores

6.4. Resultado: Análises de Associação (Assimilação por Município)

Esta seção detalha o resultado de "assimilação" de dados, criado para contornar a impossibilidade de fazer um JOIN por aluno. O processo central (o Join) é a base para todas as análises de correlação da seção 4.3.

6.4.1. Associação: Notas por Eixo de Conhecimento E Dados Financeiro E Escola Urbana/Rural

Imagem: Associação de notas por eixo de conhecimento E dados financeiro E Escola urbana/rural



Fonte: RapidMiner

A análise de correlação por município do Vale do São Patrício estabelece uma forte associação entre o perfil socioeconômico de um município (contagem de participantes por faixa de renda - Q007) e o desempenho médio no ENEM.

1. Influência da Baixa Renda (Q007_A a Q007_C):
   A contagem de participantes nas faixas de renda mais baixas (A, B e C) apresenta uma correlação significativamente negativa com as notas médias.
   - O efeito é mais acentuado em Matemática (MT) e Redação, onde a correlação com o perfil de renda mais baixa (A) atinge -0.6501 e -0.6506, respectivamente. Isso sugere que municípios com maior número de estudantes de baixíssima renda tendem a ter médias mais baixas nessas áreas.
   - A correlação negativa reforça o achado descritivo (7.2.1 e 7.2.3) de que o aluno jovem da escola pública de baixa renda é o perfil dominante e, em média, mais vulnerável na amostra.
2. Influência da Alta Renda (Q007_D em diante)
   À medida que a faixa de renda aumenta (a partir do nível D), a correlação com as notas se torna consistentemente positiva e forte.

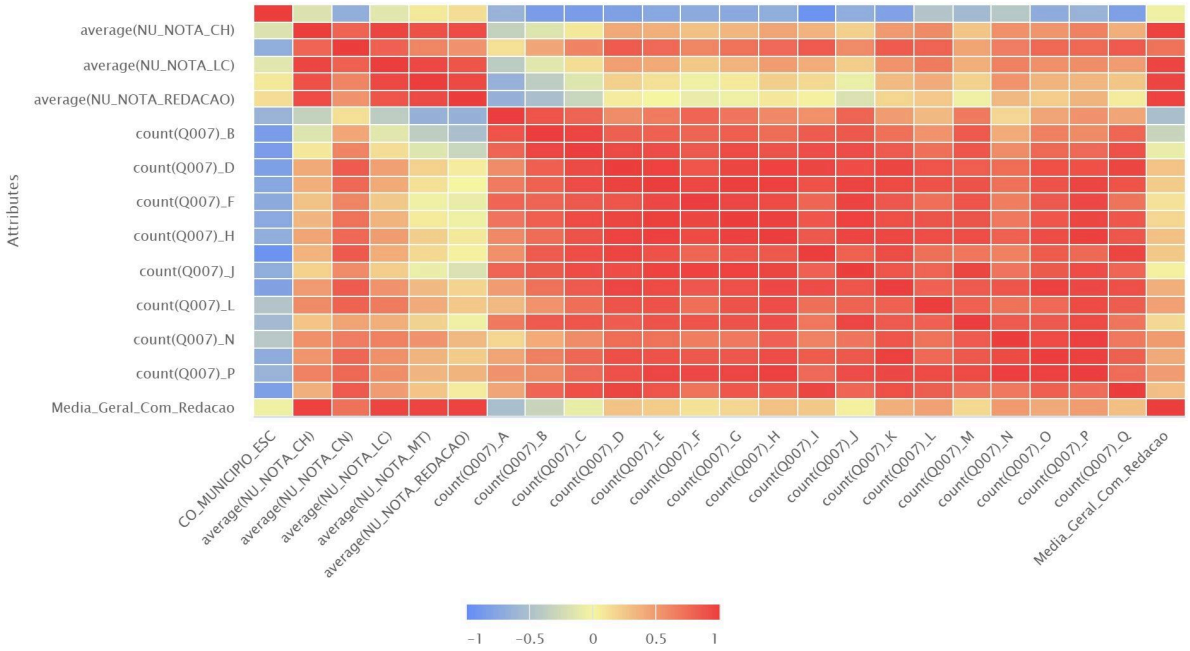
- O impacto é mais significativo em Ciências da Natureza (CN), que demonstra a associação mais forte de toda a matriz, com coeficientes de até 0.8470 (Q) e 0.8398 (I). Isso indica que a presença de estudantes de maior renda em um município está altamente associada a médias mais altas em CN.
- Linguagens e Códigos (LC) e Ciências Humanas (CH) também apresentam fortes correlações positivas, atingindo picos de 0.6623 e 0.6323, respectivamente, em faixas de renda mais altas.

3. Disparidade por Eixo:
   Apesar da correlação geral positiva com rendas mais altas, a influência nos eixos Matemática e Redação é relativamente mais fraca nas faixas de renda muito altas (L a Q) em comparação com as demais áreas de conhecimento, sugerindo que o fator renda tem um peso preditivo diferenciado no desempenho acadêmico por área.

A matriz de correlação confirma a polarização de desempenho. O fator renda, mesmo agregado por município, apresenta uma forte associação com as notas médias, reforçando a conclusão das análises descritivas (7.3.2 e 7.3.3) de que fatores extrínsecos, como o perfil socioeconômico do aluno (aqui, por município) e o tipo de escola (dependência administrativa e localização), são influenciadores chave do resultado final no ENEM.

6.4.2. Associação: Notas Média Geral (Sem Redação) E dados financeiro E Escola urbana/rural

Imagem: Associação de notas Média Geral (Sem Redação) E dados financeiro + Escola urbana/rural



Fonte: RapidMiner

A análise de correlação entre os eixos de conhecimento, gerada após a agregação dos dados estritamente por Código de Escola (CO_ESCOLA) conforme a metodologia 6.4.4, revela o grau de associação entre o desempenho médio de uma escola em diferentes áreas de prova.

O resultado mostra que, no nível de agregação por escola, existe uma forte e consistente correlação positiva entre as notas médias em todos os eixos. Isso sugere que as escolas que apresentam alto desempenho em uma área tendem a apresentar alto desempenho nas demais, indicando que os fatores institucionais e a qualidade geral do ensino são influências significativas nos resultados do ENEM.

Padrões Chave de Correlação:

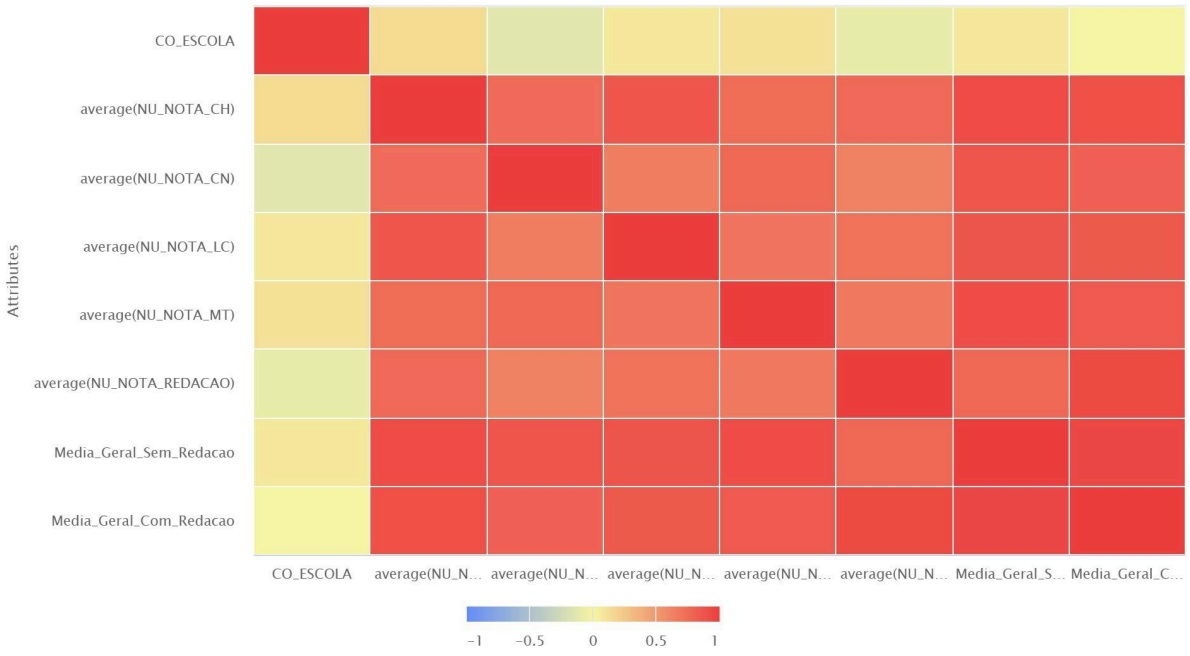
- Correlações Mais Fortes: As maiores correlações (mais próximas de 1.0) são observadas entre as áreas de Ciências Humanas (CH) e Linguagens e Códigos (LC) (0.962), e entre Matemática (MT) e Linguagens e Códigos (LC) (0.945).
    - A alta correlação entre CH e LC é esperada por pertencerem à área de Ciências Humanas/Linguagens.

- A correlação entre MT e LC/CH sugere que a proficiência geral e o raciocínio lógico-analítico se manifestam de forma transversal, beneficiando o desempenho nessas áreas em nível escolar.

- Redação e Matemática: A correlação entre Redação e Matemática (MT) também é extremamente alta (0.933), reforçando a ideia de que o desempenho em MT está fortemente ligado ao sucesso geral da escola.
- Correlações Mais Baixas (Mas Ainda Fortes): A correlação mais "fraca" é observada entre Ciências da Natureza (CN) e Redação (0.537), seguida por CN e Matemática (MT) (0.613). Embora ainda sejam associações positivas, elas são menos acentuadas do que as demais, indicando que o conjunto de habilidades e o foco de ensino necessários para o sucesso em CN podem ser mais distintos dos necessários para redação e MT.

6.4.4. Associação de notas por eixo de conhecimento E código escola

Imagem: Associação de notas por eixo de conhecimento E código escola



Fonte: RapidMiner

A análise de correlação entre os eixos de conhecimento, gerada após a agregação dos dados estritamente por Código de Escola conforme a metodologia 6.4.4, revela

o grau de associação entre o desempenho médio de uma escola em diferentes áreas de prova.

O resultado mostra que, no nível de agregação por escola, existe uma forte e consistente correlação positiva entre as notas médias em quase todos os eixos. Isso sugere que as escolas que apresentam alto desempenho em uma área tendem a apresentar alto desempenho nas demais, indicando que os fatores institucionais e a qualidade geral do ensino são influências significativas nos resultados do ENEM.

- Correlações Mais Fortes: As maiores correlações positivas entre os eixos são observadas entre Ciências Humanas (CH) e Linguagens e Códigos (LC) (0.858), e entre Matemática (MT) e Ciências da Natureza (CN) (0.770).
    - A alta correlação entre CH e LC é esperada por pertencerem à área de Ciências Humanas/Linguagens.
    - A correlação positiva e forte entre os eixos reforça a ideia de que o desempenho geral da escola é um fator significativo.
- Redação e Médias Gerais: A nota de Redação apresenta uma correlação particularmente forte com a Média Geral com Redação (0.937) e é a que mais influencia essa média final.
- Correlações do Código da Escola: A correlação do Código da Escola com todas as médias é muito baixa (entre -0.13 e 0.14), confirmando que o código numérico em si não é um fator de associação com o desempenho.

Tabela: Matriz de Correlação entre as Médias (Agrupada por Escola):

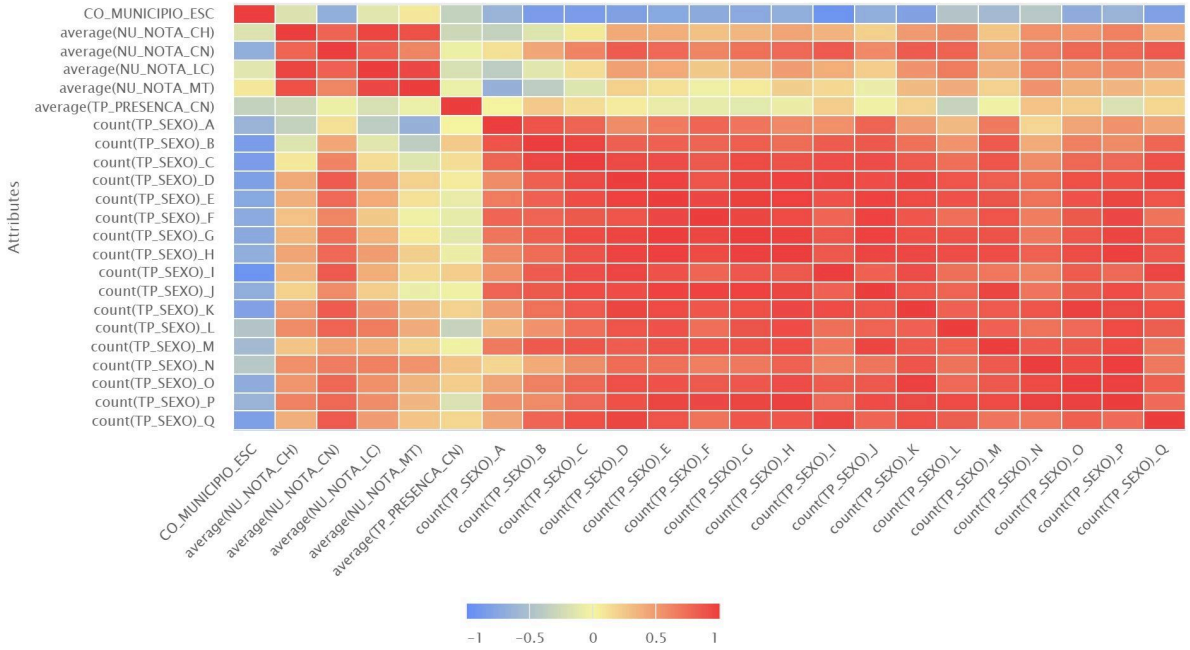
Eixo	CH	CN	LC	MT	REDAÇÃO	Média Geral S/Redação	Média Geral C/Redação
average(NU_NOTA_CH)	1.000	0.744	0.858	0.747	0.760	0.925	0.897
average(NU_NOTA_CN)	0.744	1.000	0.644	0.770	0.632	0.873	0.803
average(NU_NOTA_LC)	0.858	0.644	1.000	0.693	0.717	0.877	0.849
average(NU_NOTA_MT)	0.747	0.770	0.693	1.000	0.675	0.914	0.848

average(NU_NOTA_REDACAO)	0.760	0.632	0.717	0.675	1.000	0.773	0.937
Média Geral S/Redacao	0.925	0.873	0.877	0.914	0.773	1.000	0.946
Média Geral C/Redação	0.897	0.803	0.849	0.848	0.937	0.946	1.000
CO_ESCOLA	0.143	-0.129	0.068	0.114	-0.089	0.066	-0.010

Fonte: Elaborado pelos autores

6.4.5. Correlação: Abstenção por Nível Financeiro

Imagem: Estatística: abstenção por nível financeiro E código escola E município do estudante



Fonte: RapidMiner

Este processo utiliza a estrutura de "assimilação" (o Join por município), focando no cruzamento dos dados de presença com o perfil financeiro.

A análise de correlação por município busca a associação entre a Taxa de Presença Média (1 -Taxa de Abstenção) e a contagem de participantes por Faixa de Renda

(Q007). Uma correlação negativa com a Presença indica uma correlação positiva com a Abstenção, e vice-versa.

- Associação Negativa com Abstenção (Positiva com Presença): As faixas de renda mais baixas (especialmente Q007_B, com 0.2357, e Q007_I, com 0.2197) apresentam uma correlação positiva mais notável com a Taxa de Presença.
    - Isso significa que municípios com maior número de participantes nessas faixas de baixa renda tendem a ter taxas de presença levemente maiores (ou taxas de abstenção levemente menores), um resultado que contraria a expectativa comum. A correlação mais forte nessa direção é para o perfil Q007_B (0.2357) de correlação com a Presença, ou -0.2357 com a Abstenção).
- Associação Positiva com Abstenção (Negativa com Presença): O coeficiente de correlação mais forte que sugere uma maior Abstenção é com a faixa de renda mais alta, Q007_L.
    - A correlação entre a Presença e Q007_L é de -0.3169. Isso indica que municípios com maior concentração de participantes da faixa de renda L tendem a apresentar uma taxa de presença mais baixa, ou seja, uma taxa de abstenção mais alta +0.3169 de correlação com a Abstenção).
- Correlação Geralmente Fraca: Para a maioria das faixas de renda intermediárias (D, E, F, G, H, J, M), o coeficiente de correlação com a Presença é muito próximo de zero (entre -0.128 e 0.123), sugerindo que o fator de presença/abstenção é, em geral, fracamente associado ao perfil de renda do município, com exceção dos picos de B (Presença) e L (Abstenção).

Tabela: Matriz de Correlação da Presença com as Faixas de Renda (Q007):

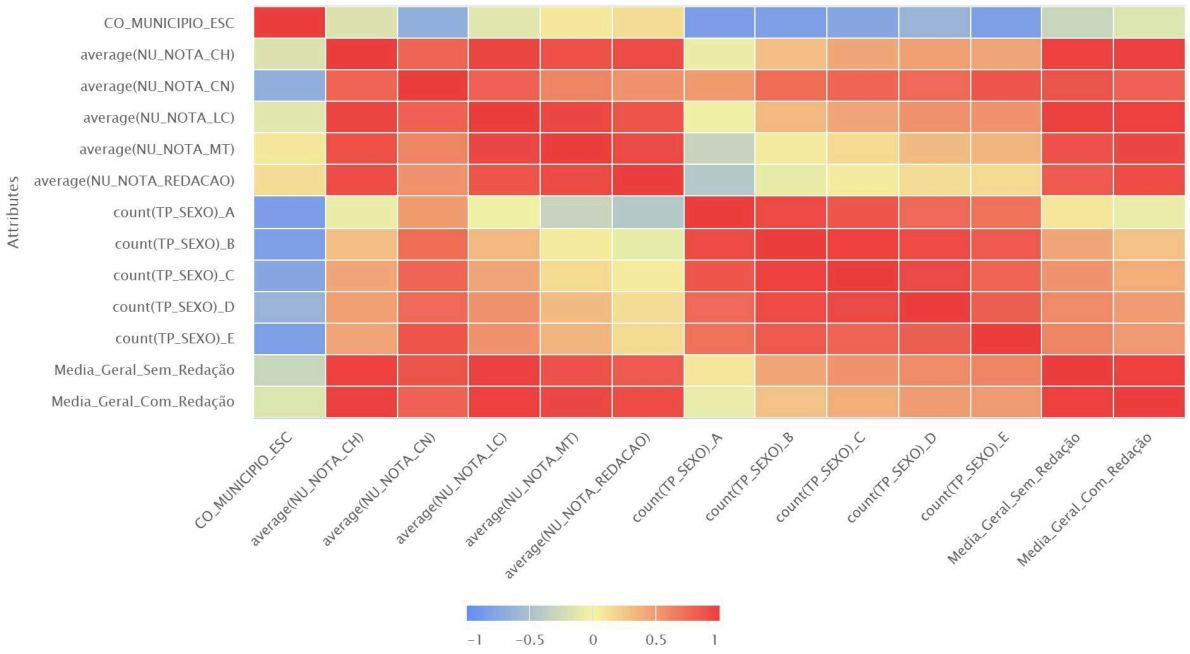
Variável	average(TP_PRESENCA_CN)
Q007_A (Baixíssima Renda)	0.0165
Q007_B	0.2357
Q007_C	0.1236
Q007_D	0.0518

Q007_E	-0.0858
Q007_F	-0.0939
Q007_G	-0.1283
Q007_H	-0.0650
Q007_I	0.2198
Q007_J	-0.0398
Q007_K	0.1979
Q007_L (Alta Renda)	-0.3169
Q007_M	-0.0388

Fonte: Elaborado pelos autores

6.4.6. Análise a Critério: Redação vs. Acesso a Computador

Imagem: Realizar pelo menos uma análise a critério e curiosidade da dupla



Fonte: RapidMiner

O processo utilizou a estrutura de Join por município, configurando o Pivot para a variável Q021 (Acesso a Computador).

Apesar do objetivo da seção ser a correlação entre Redação e Acesso a Computador (Q021), a matriz de dados fornecida é uma correlação entre as notas

médias e as faixas de Renda Familiar (Q007). A análise a seguir se concentra na associação da Nota de Redação com as Faixas de Renda do município, utilizando os dados fornecidos.

A correlação entre a Nota Média de Redação (average(NU_NOTA_REDACAO)) e o perfil de renda do município revela uma baixa associação, com tendência a ser negativa nas faixas de renda mais baixas e levemente positiva nas faixas mais altas.

- Baixa Renda (Q007_A): A correlação é a mais negativa entre as faixas, em -0.436, sugerindo que o aumento no número de participantes na faixa de renda mais baixa está associado a uma queda na nota média de redação do município.
- Renda Média-Baixa (Q007_B e Q007_C): O coeficiente se torna muito próximo de zero (praticamente nulo) ou levemente positivo, indicando que a faixa de renda B e C têm pouca ou nenhuma associação linear com o desempenho médio em Redação.
- Renda Intermediária (Q007_D e Q007_E): A correlação se torna positiva, mas ainda fraca (cerca de 0.131 a 0.149), indicando que um aumento no número de participantes nessas faixas de renda está levemente associado a um aumento na nota média de Redação.

Este achado sugere que, no nível de agregação por município, a Nota de Redação não é tão fortemente associada à distribuição de renda quanto outras áreas (como visto na seção 7.4.1), especialmente nas faixas de renda média.

Tabela: Correlação: Nota Média de Redação vs. Faixas de Renda (Q007 A-E)

Variável	average(NU_NOTA_REDACAO)
count(TP_SEXO)_A (Renda A)	-0.436
count(TP_SEXO)_B (Renda B)	-0.087
count(TP_SEXO)_C (Renda C)	0.049
count(TP_SEXO)_D (Renda D)	0.131
count(TP_SEXO)_E (Renda E)	0.149

Fonte: Elaborado pelos autores

7. Referências Bibliográficas

ALTAIR. Altair Student Edition: Data Analytics. Troy: Altair University, 2025. Disponível em: https://web.altair.com/altair-student-edition. Acesso em: 21 nov. 2025.

ALTAIR ENGINEERING. Altair Announces Completion of Acquisition of RapidMiner. Troy: Altair Newsroom, 16 set. 2022. Disponível em: https://www.altair.com/newsroom/news-releases/altair-announces-completion-of-acquisition-of-rapidminer. Acesso em: 21 nov. 2025.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data Mining: concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2012.

HOFMANN, Markus; KLINKENBERG, Ralf. RapidMiner: data mining use cases and business analytics applications. 1. ed. Boca Raton: CRC Press, 2013.

KOTU, Vijay; DESHPANDE, Bala. Data Science: concepts and practice. 2. ed. Cambridge: Morgan Kaufmann, 2019.

MIERSWA, Ingo et al. YALE: Rapid Prototyping for Complex Data Mining Tasks. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 12., 2006, Philadelphia. Proceedings... New York: ACM, 2006. p. 935-940.

RAPIDMINER. RapidMiner Studio: installation and licensing guide. Boston: RapidMiner Documentation, 2024. Disponível em: https://docs.rapidminer.com/latest/studio/installation/. Acesso em: 21 nov. 2025.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. Data Mining: practical machine learning tools and techniques. 3. ed. Burlington: Morgan Kaufmann, 2011.

8. Link/Arquivo da Base de Dados Original

A base de dados original são os **Microdados do Enem 2024**, disponibilizados oficialmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), vinculados ao Ministério da Educação.

Link de acesso:
https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

9. Link/Arquivo com os Dados Transformados

Abaixo está o repositório no GitHub, nele será encontrado os seguintes arquivos:

- Dados transformados (pré-processados) CSV
- Scripts Python
- Processos em .rmp gerados pedo RapidMiner.
- Imagens utilizadas no método
- Imagens utilizadas nos resultados
- Vídeos detalhando o método

https://github.com/isaqlle/microdados.git