RELATÓRIO DE INICIAÇÃO CIENTÍFICA

TÍTULO: ANÁLISE DE DADOS ABERTOS DO ENSINO SUPERIOR BRASILEIRO: UM ESTUDO SOBRE O PERFIL DO FINANCIAMENTO ESTUDANTIL VIA MINERAÇÃO DE DADOS

AUTOR: LUCAS FERREIRA DIAS

ORIENTADOR: ANDRÉ LUIS SCHWERZ

INSTITUIÇÃO: UTFPR - CAMPUS CAMPO MOURÃO

RESUMO

O Fundo de Financiamento Estudantil (FIES) é a principal política pública de financiamento reembolsável para o ensino superior no Brasil. Este trabalho de Iniciação Científica teve como objetivo aplicar técnicas de Ciência de Dados para auditar, higienizar e analisar os microdados do FIES referentes ao triênio 2019-2021, abrangendo todas as áreas do conhecimento. Desenvolveu-se um pipeline de Engenharia de Dados (ETL) automatizado em Python capaz de processar mais de 2,1 milhões de registros, resolvendo inconsistências de formatação, codificação e integrando as bases com o Censo da Educação Superior (INEP). Para mensurar a demanda real, criou-se a métrica de "Candidatos Únicos por Prioridade", eliminando duplicidades estatísticas. Os resultados revelam uma ociosidade estrutural de vagas em nível nacional e confirmam a hipótese da "seletividade financeira": candidatos de baixa renda, mesmo com notas inferiores, apresentam taxas de contratação superiores devido à garantia do subsídio governamental, enquanto candidatos de renda intermediária com alto desempenho acadêmico enfrentam barreiras de crédito, evidenciando que a gestão de risco assumiu protagonismo sobre o mérito acadêmico.
Palavras-chave: FIES. Mineração de Dados. Políticas Públicas. Seletividade Financeira. Engenharia de Dados.

1. INTRODUÇÃO

O Fundo de Financiamento Estudantil (FIES) passou por uma reestruturação profunda em 2018 (o "Novo Fies"), que prometia sustentabilidade fiscal e foco na meritocracia. No entanto, dados recentes indicam um paradoxo: enquanto o discurso oficial aponta para a "sobra de vagas" por suposta falta de qualificação dos candidatos, a realidade social sugere uma demanda reprimida por ensino superior.

A disponibilização dos microdados pelo Ministério da Educação (MEC) permite auditar essa contradição. Contudo, os dados brutos apresentam desafios de "Big Data", duplicidade de registros, inconsistências de formatação e falta de integração com o Censo da Educação

Superior que historicamente dificultaram diagnósticos precisos sobre a eficácia do programa.

Nesse contexto, este projeto de Iniciação Científica não se limitou a higienizar a base de dados (processo de ETL), mas utilizou a Mineração de Dados para investigar a hipótese da **"Seletividade Financeira"**. O estudo parte da premissa de que a ociosidade de vagas no triênio 2019-2021 não decorre apenas da falta de nota (mérito acadêmico), mas de barreiras de solvência (capacidade de pagamento e fiador) impostas pelo novo desenho da política pública.

Ao estruturar um pipeline de dados capaz de identificar a "Demanda Real" (Candidatos Únicos) e cruzar o desempenho acadêmico (Notas do ENEM) com a condição socioeconômica (Renda), este trabalho busca evidenciar se o FIES atual prioriza o aluno com melhor desempenho pedagógico ou o aluno com menor risco financeiro para o agente bancário.

2. OBJETIVOS

2.1. Objetivo Geral

Realizar uma Análise Exploratória de Dados (AED) abrangente sobre o cenário do FIES no Brasil, investigando o fenômeno da **ociosidade estrutural** e da **seletividade financeira** no Novo Fies (2019-2021), analisando se os gargalos de acesso ao ensino superior são motivados por insuficiência de mérito acadêmico ou por barreiras de solvência financeira e burocrática.

2.2. Objetivos Específicos

- **Engenharia de Dados:** Implementar um pipeline ETL (Extract, Transform, Load) automatizado em Python para a ingestão, limpeza e padronização dos microdados do MEC.
- **Integridade:** Sanear a base de dados, removendo duplicatas e corrigindo inconsistências de tipagem através de algoritmos de validação de integridade onde Total de Linhas = Total de Linhas Únicas.
- **Classificação CINE:** Enriquecer os dados integrando-os com o Censo da Educação Superior, utilizando algoritmos de "autocura" para classificar cursos com códigos ausentes e criando as colunas codigo_cine e nome_area_cine.
- **Regras de Negócio:** Implementar a segmentação de candidatos por Modalidade (1, 2, 3, P-FIES ou Eliminado) baseada no salário mínimo vigente de cada ano.
- **Análise de Eficiência:** Mensurar a demanda real utilizando a métrica de "Candidatos Únicos por Prioridade" e analisar o funil de conversão por Grandes Regiões.

3. MATERIAIS E MÉTODOS

A metodologia foi estruturada em um fluxo de maturação da informação, dividido em módulos sequenciais de processamento auditável, conforme detalhado a seguir.

3.1. Ferramentas e Tecnologias

O projeto foi desenvolvido utilizando a linguagem **Python** e as seguintes bibliotecas e ferramentas:

- **Pandas:** Para manipulação de DataFrames, leitura de arquivos CSV de grande porte e operações vetoriais.
- **OS e Shutil:** Para manipulação programática do sistema de arquivos e organização de diretórios.
- **Git:** Para versionamento do código.
- **Ambiente:** VS Code integrado com Jupyter Notebooks para prototipagem.

3.2. Arquitetura do Pipeline ETL e Ingestão (Pipeline Inicial)

A primeira etapa técnica consistiu na resolução da heterogeneidade dos dados brutos baixados do Portal de Dados Abertos.

- **Padronização de Nomenclatura:** Script de renomeação automática para o formato ano_tipo_semestre.csv.
- **Tratamento de Encoding:** Leitura forçada em Latin-1 para preservar a acentuação e conversão de separadores decimais (vírgula para ponto) para viabilizar cálculos numéricos.
- **Deduplicação:** Aplicação de algoritmos drop_duplicates para remover redundâncias exatas geradas na origem.

3.3. Normalização e Validação de Integridade (Módulo 1)

Nesta etapa, os dados foram preparados para operações relacionais.

- **Sanitização via Regex:** Aplicação de Expressões Regulares para limpar cabeçalhos, removendo espaços extras e convertendo para snake_case (ex: "Renda mensal" -> "renda_mensal").
- **Sufixação:** Adição automática de sufixos _inscricao e _ofertas para evitar colisão de nomes durante o cruzamento das bases.

Validação de Integridade (Chaves Primárias):
Para garantir a unicidade, foram definidas Chaves Primárias (PK) compostas.

- Inscrições: id_estudante + opcoes_cursos.
- Ofertas: codigo_mantenedora + local + curso + turno.
  Executou-se um algoritmo de verificação a cada etapa do processo que comparou:

Total de Linhas vs. Total de Linhas Únicas (Agrupadas pela PK). O teste confirmou a integridade de 100% da base de inscrições e detectou/corrigiu linhas vazias espúrias na base de ofertas de 2020.1.

3.4. Enriquecimento e Classificação CINE (Módulo 1.3 e Módulo 2)

A classificação baseada apenas em nomes de cursos é imprecisa. Realizou-se a integração com os Microdados do Censo da Educação Superior (INEP 2016-2024).

A. Construção do Dataset Mestre:

Unificação de 9 anos de Censo em um dicionário histórico, selecionando apenas colunas estáveis (CO_CURSO, NO_CINE_AREA_GERAL) para evitar erros de *Schema Drift* detectados no Censo de 2020.

B. Autocura de Dados (Data Healing - Módulo 2):

Após o cruzamento inicial (Left Join), cerca de 25.000 registros de inscrições permaneceram sem código CINE (NaN). Desenvolveu-se um algoritmo de imputação baseado em conhecimento prévio:

1. Criou-se uma tabela auxiliar mapeando "Nome do Curso" -> "Área CINE" a partir dos registros válidos.
2. Aplicou-se esse dicionário aos registros órfãos.
   Resultado: Foram criadas e populadas as colunas codigo_cine_area_geral e nome_cine_area_geral. A base de inscrições atingiu **100% de preenchimento** (0 NaNs) e a de ofertas 99,9%, garantindo que nenhuma área do conhecimento fosse excluída da análise.

3.5. Regras de Negócio: Modalidades e Regionalização (Módulo 3)

Para a análise socioeconômica, implementou-se a classificação oficial do FIES baseada no **Salário Mínimo (SM)** nominal de cada ano (2019: R$998; 2020: R$1.045; 2021: R$1.100). Foi criada a coluna modalidade_fies com as seguintes categorias:

- **Modalidade 1:** Renda per capita até 3 SM.
- **Modalidade 2:** Renda entre 3 e 5 SM (Regiões N/NE/CO).
- **Modalidade 3 (P-FIES):** Renda entre 3 e 5 SM (Regiões S/SE/CO).
- **Eliminado:** Candidatos com renda nula, superior a 5 SM ou inconsistentes (P-FIES em curso sem oferta P-FIES).

Adicionalmente, criou-se a coluna regiao (Norte, Sul, etc.) baseada na UF, para permitir análises macro regionais em vez de estaduais.

3.6. Consolidação (Módulo 4)

Os dados semestrais foram concatenados (pd.concat) e ordenados cronologicamente, gerando os arquivos mestres finais: inscritos_limpo.csv (2.197.234 registros) e ofertas_limpo.csv (143.622 registros).

4. ANÁLISE EXPLORATÓRIA: O FUNIL E A DEMANDA REAL

A compreensão da dinâmica do FIES exige superar a análise superficial do número bruto de inscrições, que frequentemente infla a percepção de demanda devido à duplicidade de tentativas por um mesmo indivíduo. Para mensurar a eficiência real da política pública, este trabalho desenvolveu métricas de unicidade e taxas de conversão baseadas no pipeline de contratação.

4.1. Metodologia: Candidatos Únicos por Prioridade

O FIES permite que um mesmo estudante se inscreva em múltiplas opções de curso ou participe de diferentes semestres. Para evitar distorções estatísticas, implementou-se o algoritmo de **"Candidatos Únicos por Prioridade"**.

O método agrupa os registros pelo CPF (**id_estudante**) e condensa o histórico do candidato em um único status final, respeitando a hierarquia de sucesso do programa:

1. **Contratada** (Maior prioridade: o aluno obteve o financiamento).
2. **Pré-Selecionado** (O aluno foi chamado, mas não contratou).
3. **Não Contratado** (Desfechos de insucesso burocrático).
4. **Lista de Espera** (O aluno não foi chamado a priori).

Essa abordagem permitiu identificar **1.109.893 indivíduos únicos** no triênio 2019-2021, purificando a base de dados de mais de 2,1 milhões de registros brutos.

4.2. O Funil de Conversão e a Ociosidade Estrutural

A análise do funil de conversão (Vagas -> Candidatos Únicos -> Aptos com Nota -> Contratados) revelou uma **ociosidade estrutural** sistêmica. Ao contrário do senso comum de que vagas sobram por falta de mérito acadêmico, os dados processados indicam o oposto.

Definiu-se a métrica de **"Candidatos Aptos"** como aqueles cuja média no ENEM foi igual ou superior à nota de corte do grupo de preferência (Média maior ou igual à Nota de Corte). A análise das taxas de conversão calculadas (vide *Tabela de Taxas Regionais*) demonstra que o número de **Candidatos/Inscritos Aptos** é frequentemente superior ao número de Vagas Ocupadas.

Imagem 1 de Funil de Seleção do Fies:

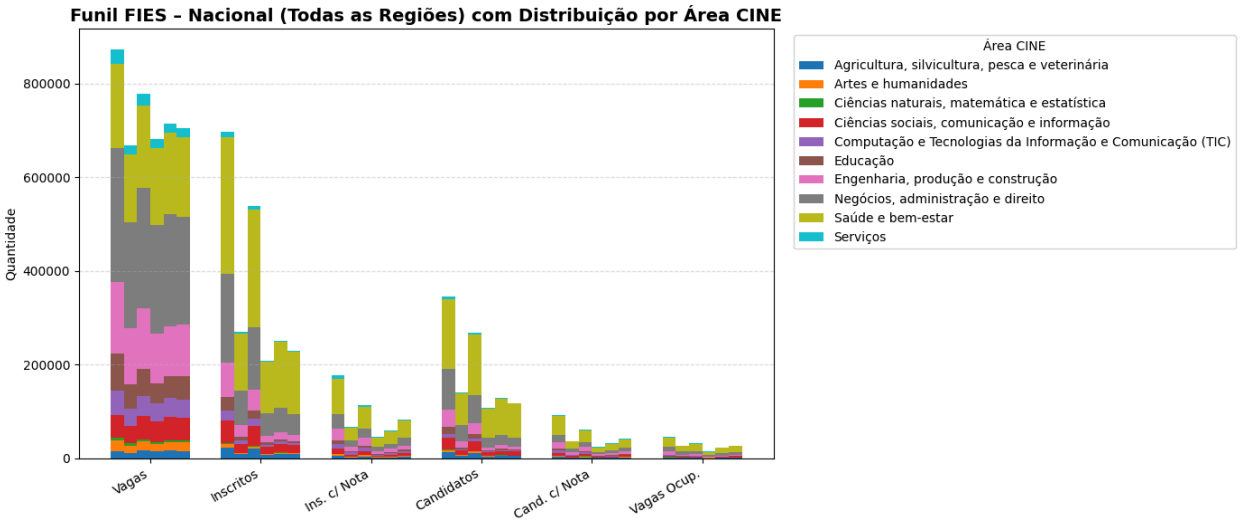
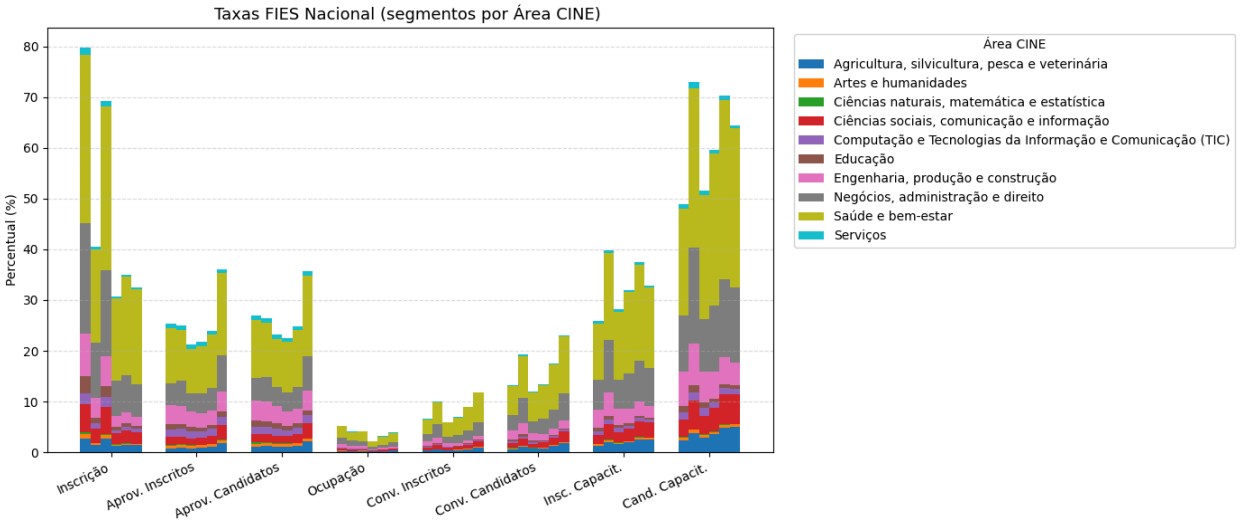


Imagem 2 de Taxas Regionais:



Observa-se visualmente que a quebra do funil não ocorre na etapa de qualificação (inscritos com notas acima da nota de corte), mas na etapa de efetivação do contrato. A **taxa_candidatos_capacitados** (Contratos / Candidatos Aptos) ou mesmo a **taxa_inscritos_capacitados** evidencia que uma parcela significativa de estudantes elegíveis e com mérito acadêmico suficiente é barrada antes da matrícula. Isso sugere que o gargalo do FIES 2019-2021 deslocou-se do desempenho pedagógico para a viabilidade burocrática e financeira.

5. EM BUSCA DAS EXPLICAÇÕES: A MATRIZ DE SELETIVIDADE (MÓDULO 5)

Para investigar as causas da retenção de candidatos aptos, desenvolveu-se uma análise matricial cruzando as duas variáveis determinantes do acesso: **Renda Familiar per Capita** e **Desempenho Acadêmico Relativo**.

5.1. Metodologia do "Gap de Nota" e Níveis de Desempenho

A nota bruta do ENEM, isoladamente, não reflete a competitividade do candidato em relação ao curso desejado. Por isso, criou-se a métrica matemática de **Gap de Nota**:

Gap = Média ENEM do Candidato - Nota de Corte do Grupo

Com base no Gap, os candidatos foram segmentados em 6 níveis de desempenho, variando de **"Muito Inferior"** (Gap menor ou igual á -150, sem chance matemática de aprovação) a **"Muito Superior"** (Gap ou igual 150, aprovados com ampla folga). Simultaneamente, a renda foi estratificada em 6 faixas, de 0 a 3 salários mínimos ou mais.

5.2. O Paradoxo do Mérito: Dois Muros Distintos

A aplicação dos *Heatmaps* (mapas de calor) sobre essa matriz 6x6 revelou que o insucesso no FIES possui duas naturezas distintas, dependendo da classe social do candidato:

Imagem 3 sobre Heatmap taxa de contratados:

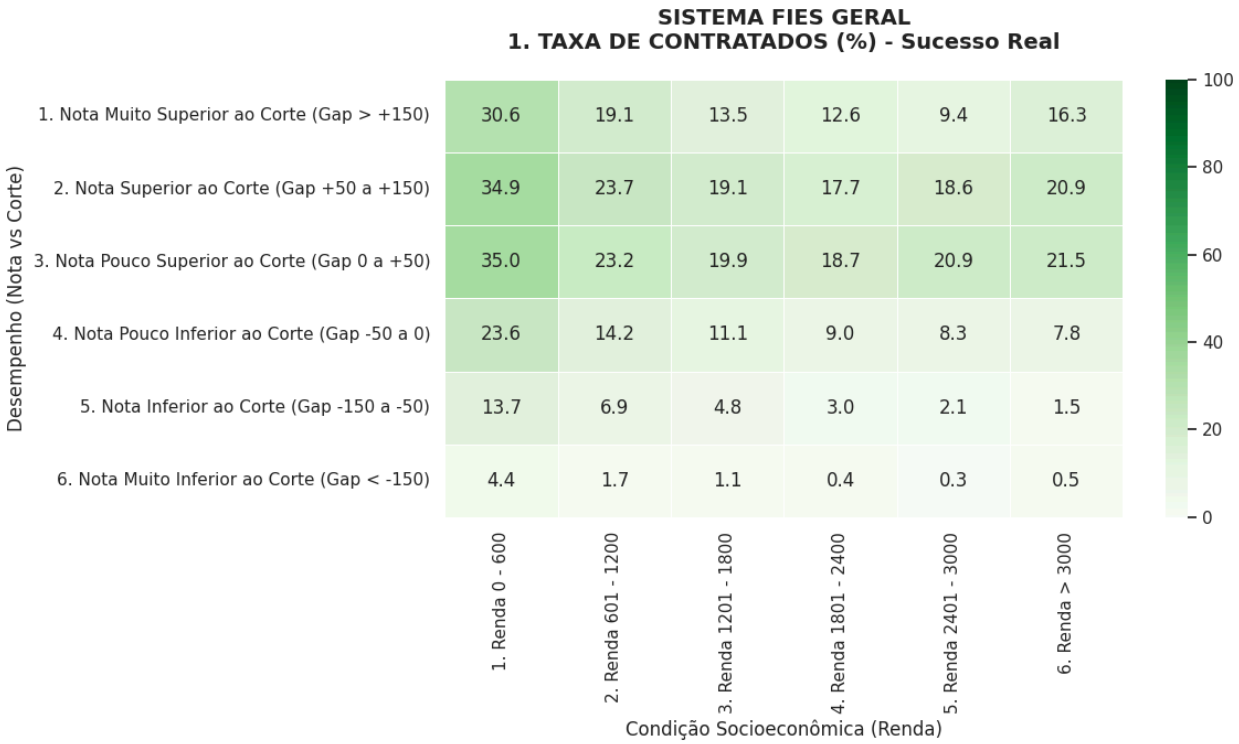


Imagem 4 sobre Heatmap taxa de lista de espera (não selecionados a priori):

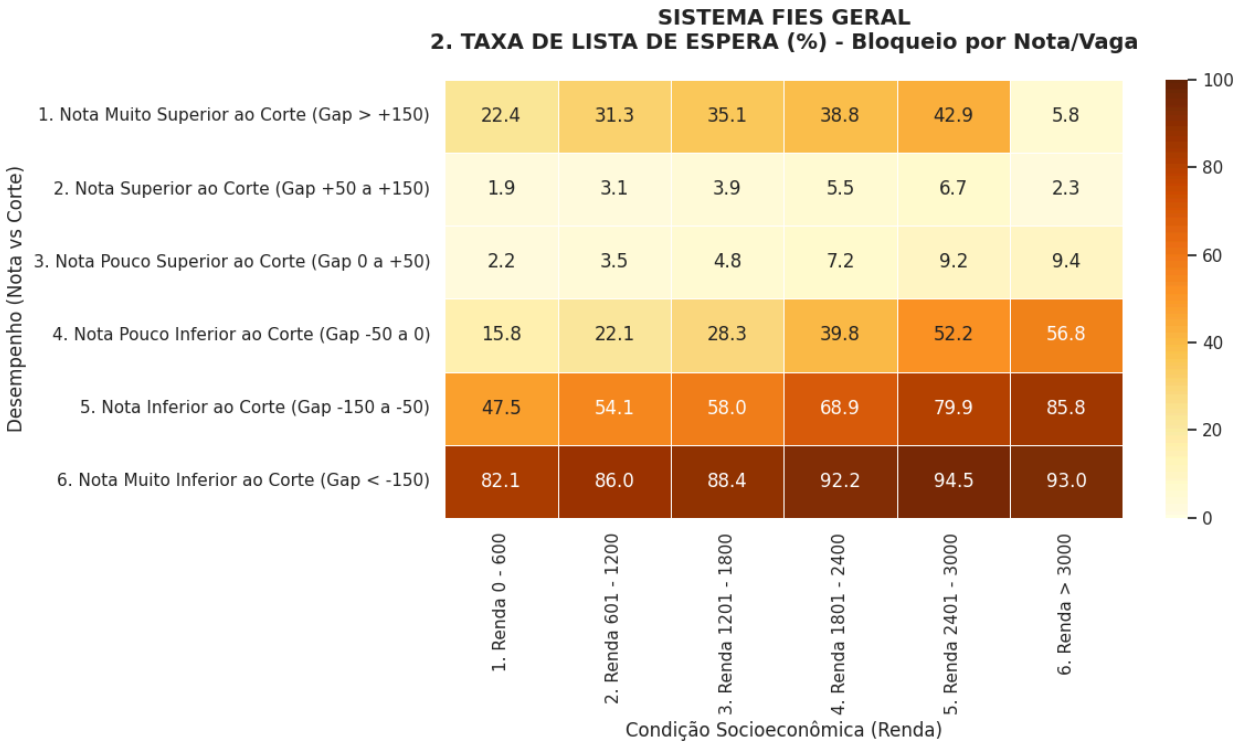
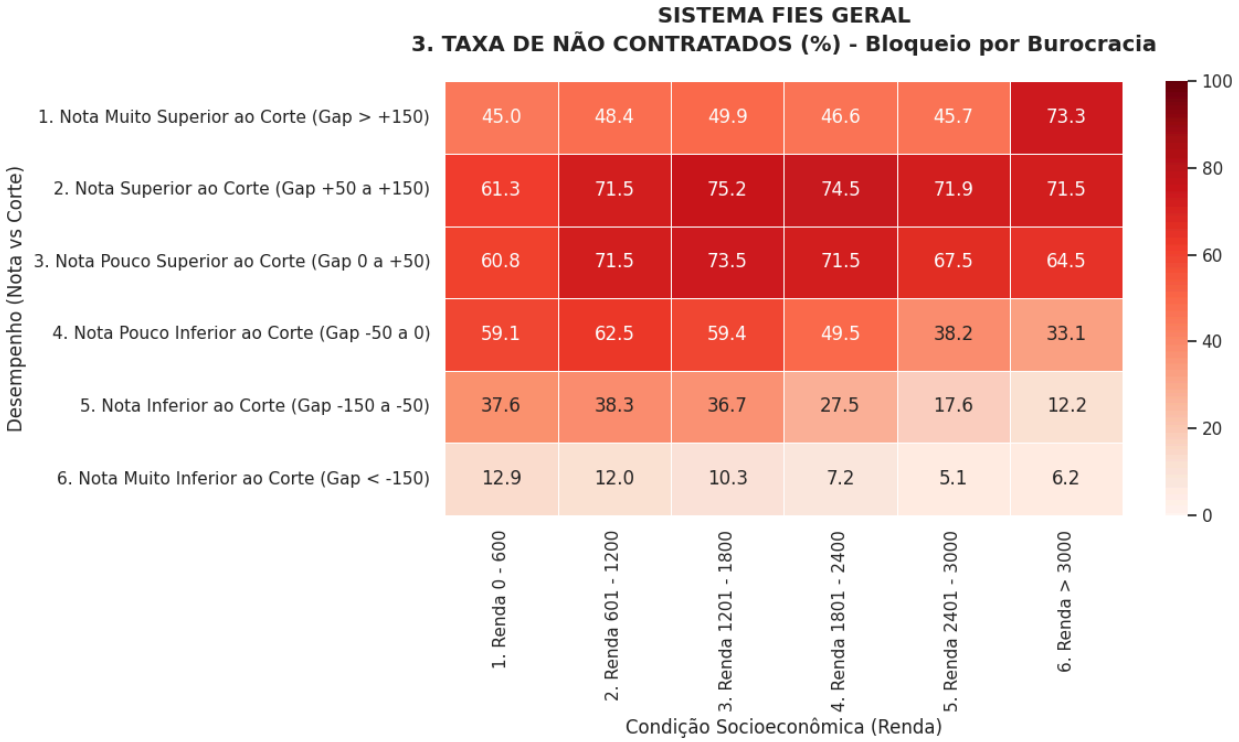


Imagem 4 sobre Heatmap taxa de não contratados (selecionados mas não efetivaram):



A. O "Muro do Banco" (Renda Alta + Nota Alta)

Os dados evidenciam um fenômeno contra-intuitivo: candidatos nas faixas de renda mais altas

do programa (acima de R$ 2.400 per capita), mesmo possuindo um **Gap Muito Superior (> +150 pontos)**, apresentam as maiores taxas de status **"Pré-Selecionado (Não Contratou)"** (manchas vermelhas no gráfico 3).

Esses alunos superam a barreira acadêmica (são chamados pelo MEC), mas esbarram na barreira financeira. Como o percentual de financiamento para essa faixa é reduzido (frequentemente cobrindo apenas 50% a 60% da mensalidade), a exigência de coparticipação financeira e de fiadores com renda compatível torna o contrato inviável, resultando na "expulsão" do candidato qualificado.

B. A "Repescagem Social" (Renda Baixa + Nota Inferior)

Em contrapartida, observa-se uma alta taxa de contratação (manchas verdes no gráfico 1) entre candidatos de **Baixíssima Renda (0-600 reais)**, inclusive aqueles com **Gap Negativo** (notas inferiores ao corte inicial).

Isso ocorre devido à dinâmica de rodagem da lista: à medida que os candidatos de renda mais alta e nota superior desistem por entraves financeiros, a lista roda até alcançar candidatos de renda baixa com notas menores. Para este grupo, o Fundo Garantidor (FG-Fies) cobre até 100% da mensalidade e flexibiliza as exigências de fiança.

5.3. Síntese da Análise Matricial

O sistema opera sob uma lógica de solvência: **o** candidato "ideal" para o FIES atual não é necessariamente o de maior nota, mas aquele cujo perfil financeiro garante **a assinatura do contrato.** O subsídio integral do governo para os mais pobres remove o risco de inadimplência bancária, facilitando a entrada destes em detrimento da classe média, que representa um "risco de crédito" para as instituições financeiras.

6. DISCUSSÃO

Os resultados empíricos obtidos através da mineração de dados confirmam, com evidência quantitativa, as hipóteses sobre a mudança de paradigma do financiamento estudantil no Brasil recente. A análise dos dados de 2019-2021 dialoga diretamente com a avaliação de desenho da política pública.

6.1. A Renda como Filtro Real

A análise matricial corrobora que o programa sofreu uma reorientação estrutural. Conforme detalhado por Meneguin e Bezerra (2022), o redesenho do FIES em 2018 (Lei nº 13.530/2017) teve como objetivo central a sustentabilidade fiscal, impondo travas orçamentárias rígidas para evitar o crescimento exponencial da dívida pública observado até 2016.

Os autores destacam que a União limitou sua participação no fundo garantidor a um teto global de R$ 3 bilhões. Os dados processados nesta pesquisa demonstram que essa "trava fiscal" teórica se traduziu, na prática, em uma barreira de entrada. A nota do ENEM tornou-se apenas um pré-requisito; a viabilidade do contrato (risco vs. garantia do fundo) assumiu o papel de critério decisivo, priorizando a solvência do sistema em detrimento da pura expansão do acesso.

6.2. Gestão de Risco em Detrimento do Mérito

Os resultados do Módulo 5 evidenciam que o FIES atua hoje primordialmente como um instrumento de gestão de risco. Meneguin e Bezerra (2022) explicam que, no novo modelo, as Instituições de Ensino Superior tornaram-se cotistas obrigatórias do Fundo Garantidor (FG-FIES), que possui natureza privada, compartilhando o risco da inadimplência diretamente com a União.

Isso cria um incentivo econômico para mitigar riscos: o sistema favorece candidatos com subsídio integral (risco absorvido pelo governo), em detrimento da classe média. Para o agente financeiro e para as mantenedoras, o aluno de baixa renda com 100% de cobertura é um ativo seguro, enquanto o estudante que necessita de coparticipação representa um risco de crédito que o novo desenho do programa busca evitar para manter a sustentabilidade do fundo.

6.3. O Muro Burocrático e a Ociosidade

A ociosidade estrutural detectada no funil de conversão deste estudo reflete a severa retração na execução da política descrita na literatura. Segundo Meneguin e Bezerra (2022), no primeiro ano de vigência do Novo FIES (2018), embora a dotação inicial fosse de aproximadamente R$ 19 bilhões, o montante efetivamente executado foi de apenas R$ 1,3 bilhão, representando menos de 7% do previsto.

Essa subexecução orçamentária valida a hipótese de que as vagas permanecem ociosas não por ausência de demanda qualificada ou mérito acadêmico, mas devido aos mecanismos de contenção de despesas e controle de risco desenhados na nova estrutura do programa. Existem vagas e existem alunos aptos, mas as travas institucionais financeiras impedem a efetivação do contrato.

6.4. A Lógica da Solvência

Por fim, o fenômeno da "Repescagem Social" observado nos dados alinha-se à necessidade de preenchimento de vagas com segurança financeira. O sistema prioriza candidatos que possuem garantia de solvência (subsídio integral via FG-FIES), independentemente de estarem em posições inferiores na classificação acadêmica. Isso ocorre para evitar o prejuízo da vaga ociosa sem, contudo, expor o fundo garantidor aos riscos de inadimplência associados aos contratos de coparticipação parcial, garantindo a continuidade da política de maneira

financeiramente sustentável, conforme preconizado pelo novo desenho do FIES.

7. CONCLUSÃO

O presente trabalho cumpriu seu objetivo duplo de entregar uma infraestrutura de Engenharia de Dados robusta e uma análise crítica do FIES. Sob a ótica técnica, o pipeline de ETL garantiu a integridade e higienização de mais de 2,1 milhões de registros, permitindo uma visão inédita baseada na métrica de "Candidatos Únicos", superando as limitações dos dados brutos disponibilizados pelo governo.

Analiticamente, conclui-se que o FIES (2019-2021) operou sob a lógica do "Funil de Solvência". Os dados confirmam que o redesenho da política pública, focado na sustentabilidade fiscal e na mitigação de riscos conforme descrito por Meneguin e Bezerra (2022), gerou uma barreira de entrada seletiva. O sistema favorece a contratação de candidatos de baixíssima renda (protegidos pelo subsídio integral do fundo garantidor) e penaliza a classe média baixa, cujo risco de crédito e exigência de coparticipação resultam em indeferimentos, mesmo diante de alto desempenho pedagógico.

Portanto, a ociosidade estrutural de vagas detectada neste estudo não decorre da falta de mérito acadêmico dos estudantes brasileiros, mas sim de um desenho institucional que prioriza a segurança orçamentária e a gestão de risco bancário. O "gap" de nota deixou de ser o principal obstáculo; o verdadeiro muro para o acesso ao ensino superior tornou-se a capacidade de solvência exigida pelas novas regras do FIES.

REFERÊNCIAS

MENEGUIN, Fernando B.; BEZERRA, Felipe Portela. **A evolução do FIES: uma avaliação de desenho sobre mudanças e continuidade do programa**. *Revista de Administração Pública*, [S.l.], 2022.