Análise de Dados do PROUNI 2018 Utilizando Modelos LLM e Data Analyst

LUCAS DE SOUSA PEREIRA, RAYANE BEZERRA DA SILVA E RODRIGO PEDROZA GADELHA RODRIGUES, Universidade Federal de Campina Grande, Brasil

Este artigo apresenta uma análise de dados do PROUNI 2018, com foco na distribuição de vagas, relação entre notas de corte e quantidade de vagas, e a correlação entre mensalidades e notas de corte. Utilizamos o Data Analyst da OpenAI e realizamos experimentos para avaliar a eficácia de modelos LLM nessa tarefa, complementados por análises descritivas, diagnósticas e preditivas. Os resultados revelam a capacidade do modelo em fornecer insights descritivos e preditivos com alta acurácia, mas resultados medianos com análise diagnóstica.

Palavras-chave: Análise de Dados, LLM, PROUNI, Notas de Corte, Educação.

1 INTRODUÇÃO

No cenário atual, a análise de dados desempenha um papel crucial em diversas áreas, especialmente no campo da educação. Com o avanço da Inteligência Artificial (IA), novas ferramentas de análise têm surgido, e uma das inovações mais notáveis são os Grandes Modelos de Linguagem (LLMs), como o GPT-4, desenvolvido pela OpenAI. Esses modelos, capazes de compreender e gerar linguagem natural, estão revolucionando a forma como os dados são interpretados e utilizados em diversos contextos, incluindo o setor educacional. Ferramentas como o ChatGPT Data Analyst utilizam esses modelos para automatizar tarefas analíticas, oferecendo uma maneira eficiente de processar grandes volumes de dados e gerar insights valiosos.

Este trabalho tem como objetivo reproduzir o experimento de conclusão de curso de Beatriz Andrade de Miranda que, por sua vez, consiste em avaliar a aplicação do ChatGPT Data Analyst na análise de dados utilizando um experimento estruturado para medir o desempenho em diferentes níveis de complexidade de análise, abordando categorias discritivas, diagnósticas e preditivas. Essa abordagem nos permite avaliar não apenas a precisão das respostas fornecidas pela ferramenta, mas também suas limitações em cenários práticos.

Os resultados deste estudo contribuem para uma melhor compreensão do papel das LLMs na automatização de análises de dados educacionais, além disso, ampliam o estudo realizado por Beatriz Miranda e contribuem para os resultados encontrados. Dessa forma, o trabalho fornece uma base sólida para compreender a capacidade da ferramenta, suas limitações e o comportamento diante de um cenário real de análise de dados.

2 METODOLOGIA

Nesta seção, descrevemos o processo metodológico empregado para avaliar a eficácia da ferramenta Data Analyst do ChatGPT na análise do conjunto de dados do PROUNI 2018. O estudo foi estruturado de forma a gerar perguntas analíticas em três níveis de dificuldade: fácil, médio e difícil, com o objetivo de testar as capacidades da ferramenta em diferentes contextos de complexidade. Para cada tipo de análise (descritiva, diagnóstica e preditiva), foram formuladas perguntas específicas em cada um desses níveis.

Author's address: Lucas de Sousa Pereira, Rayane Bezerra da Silva e Rodrigo Pedroza Gadelha Rodrigues, lucas.pereira@ccc.ufcg.edu.br,rayane.silva@ccc.ufcg.edu.brerodrigo.rodrigues@ccc.ufcg.edu.br, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brasil, 58429-900.

Um aspecto central dessa avaliação é a interação com a ferramenta através de comandos específicos, conhecidos como "prompts". Esses prompts são utilizados para orientar o modelo na execução de tarefas analíticas, que incluem desde a preparação de dados até a geração de insights diagnósticos e preditivos.

Para garantir a eficácia da ferramenta, adotamos uma abordagem estruturada, que inclui o pré-processamento dos dados, seguido de diferentes análises sobre o conjunto de dados educacionais, abordando tanto aspectos quantitativos (número de vagas, mensalidades) quanto qualitativos (notas de corte).

2.1 Conjunto de Dados

O conjunto de dados utilizado neste estudo contém informações detalhadas sobre os cursos oferecidos pelo PROUNI em 2018. Ele inclui variáveis como o nome das instituições de ensino, cursos, turnos, mensalidades, número de bolsas oferecidas (integrais e parciais), e notas de corte para diferentes categorias de cotas e ampla concorrência.

O dataset original foi processado com informações organizadas em múltiplas colunas relevantes para a análise. Cada registro no conjunto de dados corresponde a um curso oferecido por uma instituição em um determinado turno e grau de ensino, com atributos relacionados à oferta de vagas, distribuição de bolsas e as respectivas notas de corte exigidas. A diversidade dos dados permite uma ampla exploração das relações entre essas variáveis, possibilitando uma análise abrangente.

2.2 Pré-processamento dos Dados

Para garantir a clareza e a coerência dos resultados, foi realizado um pré-processamento cuidadoso dos dados. Inicialmente, os dados que não eram relevantes para o escopo do estudo foram removidos, como possíveis registros duplicados ou incompletos. Além disso, as colunas relacionadas a informações não utilizadas nas análises principais, como identificadores genéricos, foram filtradas. Os principais passos de pré-processamento incluíram limpeza de valores ausentes, conversão dos tipos de dados e criação de variáveis adicionais.

2.3 Análise Descritiva

Na análise descritiva, buscamos explorar características como o tipo de curso, turno, mensalidades e a distribuição de bolsas, sendo possível identificar padrões relevantes que auxiliam na compreensão do panorama educacional e no funcionamento da ferramenta Data Analyst. Esse primeiro nível de análise permite um entendimento preliminar das informações antes de avançar para abordagens mais sofisticadas, como análises preditivas e identificação de correlações.

Para estruturar as perguntas e análises, foram definidos três níveis de complexidade, cada um com suas particularidades. Perguntas fáceis envolvem cálculos diretos e descritivos, como sumarização de frequências, médias, valores mínimos e máximos. Perguntas de nível médio exigem um processamento adicional e a criação de métricas derivadas, como percentuais, distribuições e cálculos de médias ajustadas. Já perguntas difíceis envolvem análises mais complexas, como a identificação de correlações entre variáveis e a aplicação de métricas de dispersão e assimetria.

A seguir, as perguntas geradas com base no dataset estão organizadas conforme o nível de dificuldade, permitindo uma análise incremental das informações:

2.3.1 Perguntas fáceis.

(1) Qual é a quantidade de cursos disponíveis por grau (Bacharelado, Licenciatura, etc.)?
(2) Qual é o turno mais comum entre os cursos disponíveis?
(3) Qual é o valor mínimo e máximo de mensalidade para cursos no estado de São Paulo?

2.3.2 Perguntas médias.

(1) Qual é a proporção de bolsas integrais ofertadas por cotas e por ampla concorrência?
(2) Qual é a distribuição média de mensalidades por turno (matutino, vespertino, noturno)?
(3) Qual é a distribuição de cursos por cidade para os 5 estados com o maior número de cursos?

2.3.3 Perguntas difíceis.

(1) Qual é a correlação entre o valor das mensalidades e as notas dos cursos (integral ampla, integral cotas, parcial ampla, parcial cotas)?
(2) Qual é a diferença de oferta de bolsas integrais e parciais ao longo das universidades? Considere a média de bolsas integrais e parciais por universidade e avalie a dispersão (desvio padrão).
(3) Como a proporção de bolsas integrais em relação ao número total de bolsas mudou ao longo das diferentes universidades e estados?

2.4 Análise Diagnóstica

Na análise diagnóstica, investigamos as relações entre diferentes variáveis, como a correlação entre as mensalidades e as notas de corte. Esta análise foi conduzida utilizando coeficientes de correlação, com o objetivo de entender como o custo dos cursos impacta as notas mínimas necessárias para obter uma bolsa.

Para formular perguntas simples, foram utilizadas métricas básicas, como percentual, frequência e percentil. Perguntas de dificuldade média envolveram os testes de hipótese, a definição de um índice que exige mais processamento e o coeficiente de variação. Para perguntas difíceis, recorreram a métodos estatísticos mais complexos, incluindo análise de variância, teste de normalidade, homogeneidade das variâncias e teste não paramétrico. Em seguida, as perguntas estão organizadas por nível de dificuldade.

2.4.1 Perguntas fáceis.

(1) Qual é a distribuição de vagas por região no PROUNI 2018?
(2) Qual a relação entre o valor médio da mensalidade e a quantidade de vagas oferecidas?
(3) Qual a nota de corte média dos cursos de Medicina em comparação aos cursos de Direito?

2.4.2 Perguntas médias.

(1) Existe uma correlação significativa entre a nota de corte e o valor da mensalidade dos cursos oferecidos no PROUNI 2018?
(2) Qual o impacto das regiões na quantidade de vagas e notas de corte dos cursos de Engenharia?
(3) Qual a relação entre a quantidade de vagas oferecidas e a forma de ingresso nas diferentes instituições?

2.4.3 Perguntas difíceis.

(1) A nota de corte influencia de forma significativa a distribuição de vagas por instituição no PROUNI
(2) Há uma diferença estatisticamente significativa nas notas entre estudantes que frequentam o turno integral e aqueles que frequentam o turno noturno?
(3) O curso (Medicina, Enfermagem, Psicologia, etc.) tem alguma influência estatisticamente significativa nas notas ou no fato de os alunos receberem algum tipo de bolsa?

2.5 Análise Preditiva

Para a análise preditiva, utilizamos modelos de regressão linear e clustering para prever a relação entre o número de vagas e as notas de corte. A regressão linear foi aplicada para identificar tendências, enquanto o algoritmo de clustering (KMeans) foi empregado para agrupar as instituições com base em características similares, como mensalidade e nota de corte. Essas previsões foram validadas usando métricas como o coeficiente de determinação (R^2).

Ao formular as perguntas fáceis, pretendia-se que a ferramenta utilizasse modelos simples de regressão em suas respostas, como a Regressão Linear. Para as perguntas médias, esperava-se obter modelos mais robustos, como Floresta Aleatória e Clustering. Em relação às perguntas difíceis, eram esperados modelos mais elaborados, incluindo aplicações avançadas de Regressão e Classificação, bem como Redes Neurais. Segue a classificação das perguntas por nível de dificuldade.

2.5.1 Perguntas fáceis.

(1) A relação entre mensalidade e vagas oferecidas por curso é significativa?
(2) É possível prever se uma instituição terá um número elevado de vagas com base na região e no turno dos cursos?
(3) Com base na distribuição de vagas e na mensalidade, podemos categorizar as instituições em grupos (baixo, médio e alto) de acessibilidade?

2.5.2 Perguntas médias.

(1) Dada a relação entre a quantidade de vagas oferecidas e as notas de corte, é possível determinar a probabilidade de uma instituição ser considerada de alta concorrência?
(2) Há padrões de similaridade entre instituições de diferentes regiões em relação à quantidade de vagas e ao valor da mensalidade?
(3) Dos cursos, é possível prever o tipo de bolsa mais ofertada (integral ou parcial)?

2.5.3 Perguntas difíceis.

(1) É possível prever a relação entre a nota de corte e a mensalidade de cursos, e identificar os cursos que têm uma maior relação custo-benefício para os alunos?
(2) Através de uma análise avançada, é possível identificar se o tipo de bolsa (integral ou parcial), o turno (integral, matutino, noturno), e a localização do campus influenciam no desempenho acadêmico dos estudantes? Defina três abordagens para resolver essa questão e siga a mais eficaz.
(3) Através de análise avançada, é possível identificar padrões que influenciam a nota parcial de alunos em diferentes cursos e universidades, considerando variáveis como cidade, mensalidade e tipo de bolsa (cotas ou ampla)?

3 RESULTADOS

Esta seção apresenta os resultados obtidos em nossos experimentos.

3.1 Análise Descritiva

Durante a Análise Descritiva, obteve-se 100.0% de aproveitamento, empregando dados consistentes e as métricas solicitadas. Foram exploradas variáveis como a distribuição de vagas por curso, turnos e mensalidades dos cursos oferecidos no PROUNI 2018. A ferramenta apresentou resultados precisos para questões simples, utilizando técnicas como cálculos de média, frequência e valores máximos e mínimos para descrever o panorama geral dos dados. Essa

análise preliminar permitiu identificar padrões importantes, fornecendo uma base sólida para as etapas subsequentes de análise. A tabela a seguir resume as respostas obtidas durante a Análise Descritiva:

ANÁLISE DESCRITIVA				
Nível		Pergunta	Resposta	
			Correta	Incorreta
FÁCIL		1	X	
		2	X	
		3	X	
MÉDIO		1	X	
		2	X	
		3	X	
DIFÍCIL		1	X	
		2	X	
		3	X	

Fig. 1. Resultados da Análise Descritiva

Conforme observado na tabela, a ferramenta foi capaz de responder corretamente a todas as perguntas. Mostrando ser uma ferramenta útil e adequada para a análise descritiva.

3.2 Análise Diagnóstica

Durante a Análise Diagnóstica, teve-se 77.7% de aproveitamento, empregando dados adequados e as métricas requisitadas. Foram investigadas as relações entre as variáveis, como por exemplo, mensalidades dos cursos e as notas de corte exigidas para bolsas integrais e parciais, utilizando os dados do PROUNI 2018. A ferramenta indicou na resolução das questões a utilização do coeficiente de Pearson, dentre outras técnicas. A tabela a seguir resume as respostas obtidas durante a Análise Diagnóstica:

ANÁLISE DIAGNÓSTICA				
Nível		Pergunta	Resposta	
			Correta	Incorreta
FÁCIL		1	X	
		2	X	
		3	X	
MÉDIO		1	X	
		2		X
		3		X
DIFÍCIL		1	X	
		2	X	
		3	X	

Fig. 2. Resultados da Análise Diagnóstica

Conforme observado na tabela, a ferramenta foi capaz de responder corretamente às perguntas de nível fácil e difícil na maioria dos casos. No entanto, em questões de nível médio, houve maior incidência de respostas incorretas, particularmente em perguntas que exigiam análise mais aprofundada das variáveis envolvidas. As perguntas incorretas envolveram inconsistências ao aplicar agrupamento ao invés de metrificar o impacto(aplicar correlação) das variáveis, sugerindo que a ferramenta enfrenta limitações ao lidar com análises de complexidade intermediária.

3.3 Análise Preditiva

Durante a Análise Preditiva, teve-se 88.8% de aproveitamento, empregando dados adequados e as métricas requisitadas.Na Análise Preditiva, foram utilizados modelos de regressão linear e clustering para prever relaações, como por exemplo, a relação entre o número de vagas e as notas de corte dos cursos oferecidos no PROUNI 2018. Esses modelos foram aplicados com o objetivo de identificar padrões e tendências, como a influência das mensalidades nas notas de corte e a segmentação de instituições com base em características similares, como faixa de mensalidade e notas de corte. A tabela abaixo resume as respostas obtidas durante a Análise Preditiva:

ANÁLISE PREDITIVA				
Nível		Pergunta	Resposta	
			Correta	Incorreta
FÁCIL		1	X	
		2	X	
		3	X	
MÉDIO		1	X	
		2	X	
		3	X	
DIFÍCIL		1	X	
		2		X
		3	X	

Fig. 3. Resultados da Análise Preditiva

Os experimentos preditivos mostraram uma alta taxa de respostas corretas, especialmente nas perguntas de nível fácil e médio, com todos os modelos preditivos retornando resultados satisfatórios. No entanto, em questões de nível difícil, houve uma resposta incorreta em uma das perguntas, onde o modelo apresentou dificuldades em prever corretamente a relação entre variáveis mais complexas. A ferramenta não incluiu em consideração o tipo de bolsa e isto levou a uma resolução inconsistente da questão, gerando previsões imprecisas.

4 DISCUSSÃO

A partir do experimento realizado com o Data Analyst do ChatGPT, podemos discutir o seu desempenho. Foram realizadas 27 perguntas distribuídas pelas principais categorias de análise de dados. A Figura 4 apresenta a sumarização de todas as respostas e problemas identificados.

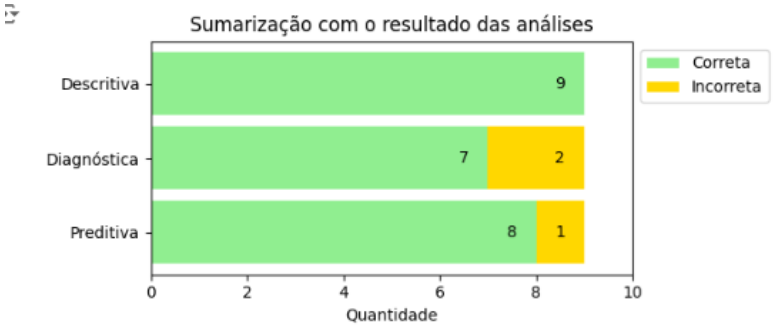

Fig. 4. Comparativo entre os Resultados da Análise Descritiva, Diagnóstica e Preditiva

Podemos observar que a ferramenta apresentou excelentes resultados na análise descritiva e preditiva, mas resultados menos satisfatórios na análise diagnóstica. A ferramenta demonstrou ter grande potencial, entretanto precisa de ajustes em casos que envolvem uma quantidade de variáveis ou complexidade maior. Foi possível perceber que, indicando o erro, ela conseguia resolvê-lo, mas apenas com as instruções iniciais. Em alguns casos, a ferramenta falhou na resolução da questão.

5 CONCLUSÃO

Os resultados indicam que a utilização de LLMs, como o Data Analyst, pode ser uma ferramenta poderosa para automatizar análises de dados, permitindo uma compreensão mais rápida e precisa de dados complexos. No entanto, desafios ainda permanecem em questões mais avançadas de análise diagnóstica e no tratamento de grandes volumes de dados.

AGRADECIMENTOS

Gostaríamos de agradecer ao Professor Cláudio Campelo e a Cientista da Computação Beatriz Miranda por fornecer acompanhamentos e materiais necessários para a realização deste estudo.

REFERÊNCIAS

[1] Beatriz A. de Miranda, Claudio E. C. Campelo. *How effective is an LLM-based Data Analysis Automation Tool? A Case Study with ChatGPT's Data Analyst.* Federal University of Campina Grande (UFCG), 2023.

[2] Brasil.IO. *Cursos PROUNI - 2018.* Disponível em: https://brasil.io/dataset/cursos-prouni/cursos/. Acesso em: 10/09/2024.