
**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE**
**CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA**
**CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**BEATRIZ ANDRADE DE MIRANDA**

# QUÃO EFICAZ É UMA FERRAMENTA DE AUTOMAÇÃO DE ANÁLISE DE DADOS BASEADA EM LLM? UM ESTUDO DE CASO COM O DATA ANALYST DO CHATGPT

**CAMPINA GRANDE - PB**

**2024**

**BEATRIZ ANDRADE DE MIRANDA**

# QUÃO EFICAZ É UMA FERRAMENTA DE AUTOMAÇÃO DE ANÁLISE DE DADOS BASEADA EM LLM? UM ESTUDO DE CASO COM O DATA ANALYST DO CHATGPT

**Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.**

**Orientador : Cláudio Elízio Calazans Campelo**

**CAMPINA GRANDE - PB**

**2024**

**BEATRIZ ANDRADE DE MIRANDA**

# QUÃO EFICAZ É UMA FERRAMENTA DE AUTOMAÇÃO DE ANÁLISE DE DADOS BASEADA EM LLM? UM ESTUDO DE CASO COM O DATA ANALYST DO CHATGPT

**Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.**

## BANCA EXAMINADORA:

**Cláudio Elízio Calazans Campelo**

**Orientador – UASC/CEEI/UFCG**

**Maxwell Guimarães de Oliveira**

**Examinador – UASC/CEEI/UFCG**

**Francisco Vilar Brasileiro**

**Professor da Disciplina TCC – UASC/CEEI/UFCG**

**Trabalho aprovado em: 15 de MAIO de 2024.**

**CAMPINA GRANDE - PB**

# RESUMO

Este artigo investiga a aplicação eficaz dos Grandes Modelos de Linguagem (LLMs), em particular o Generative Pre-trained Transformer (ChatGPT) da OpenAI, na análise de dados. A relevância desta pesquisa surge com a crescente adoção de ferramentas de Inteligência Artificial (IA) em processos analíticos, exigindo uma avaliação meticulosa de suas capacidades e limitações para aprimorar a tomada de decisões e a eficiência operacional. Utilizando o Data Analyst do ChatGPT como estudo de caso, este trabalho implementa um experimento estruturado com 36 perguntas distribuídas em análises Descritiva, Diagnóstica, Preditiva e Prescritiva, para mensurar sua eficácia. Os resultados apontam uma eficiência geral de 86,11%, com destaque para o desempenho em análises descritivas e diagnósticas, enquanto enfrenta desafios nas categorias mais complexas, como as preditivas e prescritivas. Apesar das limitações técnicas, tais como restrições no processamento de dados e falhas operacionais, o estudo destaca o potencial significativo do Data Analyst em auxiliar analistas de dados, estabelecendo um marco importante para futuras melhorias e pesquisas na aplicação prática de LLMs na análise de dados.

# HOW EFFECTIVE IS AN LLM-BASED DATA ANALYSIS AUTOMATION TOOL? A CASE STUDY WITH CHATGPT'S DATA ANALYST


## ABSTRACT

This paper investigates the effective application of Large Language Models (LLMs), specifically the OpenAI's Generative Pre-trained Transformer (ChatGPT), in data analysis. The relevance of this research emerges with the growing adoption of Artificial Intelligence (AI) tools in analytical processes, necessitating a meticulous evaluation of their capabilities and limitations to enhance decision-making and operational efficiency. Using the ChatGPT's Data Analyst as a case study, this work implements a structured experiment with 36 questions distributed across Descriptive, Diagnostic, Predictive, and Prescriptive analyses to measure its effectiveness. The results indicate an overall efficiency of 86,11%, with notable performance in descriptive and diagnostic analyses, while facing challenges in more complex categories, such as predictive and prescriptive. Despite technical limitations, such as data processing constraints and operational failures, the study underscores the significant potential of the Data Analyst in assisting data analysts, establishing an important milestone for future improvements and research in the practical application of LLMs in data analysis.

# Quão eficaz é uma ferramenta de Automação de Análise de Dados baseada em LLM? Um Estudo de Caso com o Data Analyst do ChatGPT

**Trabalho de Conclusão de Curso**

Beatriz Andrade de Miranda
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil
beatriz.miranda@ccc.ufcg.edu.br

Claudio E. C. Campelo
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil
campelo@dsc.ufcg.edu.br

## ABSTRACT

This paper investigates the effective application of Large Language Models (LLMs), specifically the OpenAI's Generative Pre-trained Transformer (ChatGPT), in data analysis. The relevance of this research emerges with the growing adoption of Artificial Intelligence (AI) tools in analytical processes, necessitating a meticulous evaluation of their capabilities and limitations to enhance decision-making and operational efficiency. Using the ChatGPT's Data Analyst as a case study, this work implements a structured experiment with 36 questions distributed across Descriptive, Diagnostic, Predictive, and Prescriptive analyses to measure its effectiveness. The results indicate an overall efficiency of 86,11%, with notable performance in descriptive and diagnostic analyses, while facing challenges in more complex categories, such as predictive and prescriptive. Despite technical limitations, such as data processing constraints and operational failures, the study underscores the significant potential of the Data Analyst in assisting data analysts, establishing an important milestone for future improvements and research in the practical application of LLMs in data analysis.

## RESUMO

Este artigo investiga a aplicação eficaz dos Grandes Modelos de Linguagem (LLMs), em particular o Generative Pre-trained Transformer (ChatGPT) da OpenAI, na análise de dados. A relevância desta pesquisa surge com a crescente adoção de ferramentas de Inteligência Artificial (IA) em processos analíticos, exigindo uma avaliação meticulosa de suas capacidades e limitações para aprimorar a tomada de decisões e a eficiência operacional. Utilizando o Data Analyst do ChatGPT como estudo de caso, este trabalho implementa um experimento estruturado com 36 perguntas distribuídas em análises Descritiva, Diagnóstica, Preditiva e Prescritiva, para mensurar sua eficácia. Os resultados apontam uma eficiência geral de 86,11%, com destaque para o desempenho em análises descritivas e diagnósticas, enquanto enfrenta desafios nas categorias mais complexas, como as preditivas e prescritivas. Apesar das limitações técnicas, tais como restrições no processamento de dados e falhas operacionais, o estudo destaca o potencial significativo do Data Analyst em auxiliar analistas de dados, estabelecendo um marco importante para futuras melhorias e pesquisas na aplicação prática de LLMs na análise de dados.

## PALAVRAS-CHAVE

Grandes Modelos de Linguagem, LLMs, Automação de Análise de Dados, ChatGPT Data Analyst, Estudo de Caso

## 1 INTRODUÇÃO

No cenário atual, caracterizado por avanços tecnológicos marcantes, a Inteligência Artificial (IA) tem sido uma catalisadora de transformações em diversos setores, influenciando desde operações industriais até interações sociais cotidianas [11, 18]. Dentre as inovações mais significativas na IA, estão os Grandes Modelos de Linguagem (LLMs). Como exemplo, citam-se os modelos comerciais da família GPT [3, 20], desenvolvidos pela OpenAI[1] e o Gemini [21] do Google[2], além de modelos de código aberto como o Mistral [13] e LLaMA [22, 23]. Esses modelos estão revolucionando a comunicação entre humanos e máquinas, demonstrando uma capacidade excepcional de compreender e produzir linguagem humana, posicionando-se na vanguarda das inovações em IA [5, 17]. Este artigo explora o impacto desses modelos na análise de dados, com ênfase no Data Analyst do ChatGPT[3], uma ferramenta desenvolvida especificamente para aprimorar tarefas analíticas.

O Data Analyst é baseado na arquitetura do GPT-4 [3], que utiliza redes neurais do tipo transformador [24]. Essas redes processam palavras em paralelo e identificam relações de longo alcance no texto por meio de um mecanismo de atenção que avalia a importância de cada palavra, independentemente de sua posição. Demonstrando um potencial revolucionário para a análise de dados, conforme evidenciado em estudos recentes [10, 27], o Data Analyst emprega essa arquitetura e recebe treinamento especializado em análise de dados. Quando confrontado com dados e um problema específico, a ferramenta executa os passos necessários dentro de seu ambiente interno, utilizando Python, para fornecer respostas analíticas detalhadas e o código correspondente.

Não obstante a ampla utilização do GPT em diversas áreas, sua aplicação na análise de dados é particularmente significativa. O



[1] OpenAI - https://openai.com/
[2] Google - https://www.google.com/
[3] Data Analyst do ChatGPT - https://chatgpt.com/g/g-HMNcP6w7d-data-analyst

GPT atua não apenas como anotador de dados [9] e assistente na preparação de dados [12], mas também tem sido comparado a analistas humanos quanto à eficiência [7]. Contudo, esses estudos focam primariamente na API do GPT e não abordam de forma abrangente o ecossistema de ferramentas associadas que expandem suas funcionalidades. Esta observação revela uma oportunidade de pesquisa ainda não explorada, especialmente no que diz respeito à avaliação do Data Analyst do ChatGPT. Este trabalho busca preencher essa lacuna na literatura, examinando o Data Analyst com um estudo de caso específico na análise de dados, o que proporciona uma visão mais detalhada de seu impacto e potencial transformador.

Embora os modelos abertos tenham se apresentado bastante competitivos, superando os modelos comerciais em diversos *benchmarks*, ainda há uma escassez de ferramentas abertas equivalentes ao Data Analyst do ChatGPT. Essa ausência no ecossistema de código aberto gera otimismo quanto à possibilidade de emergência de uma ferramenta similar que possa ser desenvolvida e aprimorada pela comunidade. Assim, através de uma análise detalhada das capacidades atuais do Data Analyst, buscamos fomentar o potencial de expansão de tecnologias similares e abertas.

Este estudo avalia a capacidade do Data Analyst do ChatGPT em lidar com as quatro principais categorias de análise de dados: Descritiva, Diagnóstica, Preditiva e Prescritiva [26]. Através de um experimento controlado, composto por 36 questões de diferentes níveis de dificuldade - fácil, médio e difícil -, medimos o desempenho da ferramenta ao responder a um conjunto diversificado de perguntas utilizando uma base de dados tabular, contendo dados históricos de desempenho acadêmico de alunos. Este método visa avaliar não apenas a eficácia do modelo em fornecer análises acuradas e úteis, mas também avalia suas habilidades frente a questões de complexidade variada. Além de medir a eficácia, este trabalho identifica as limitações dessa ferramenta em um contexto prático, explorando especificamente seu impacto na análise de dados, um campo onde precisão e capacidade de interpretação são essenciais [8].

Ao melhor de nosso conhecimento, este é o primeiro trabalho a conduzir um estudo de caso desta natureza, explorando o desempenho da ferramenta e destacando informações relevantes sobre suas funcionalidades e desafios operacionais. Embora o ChatGPT[4] mostre competência notável em análises descritivas e diagnósticas, enfrenta dificuldades em tarefas prescritivas e preditivas, que exigem um nível mais elevado de inferência e manipulação de dados complexos [4]. As limitações incluem restrições técnicas, como a capacidade recomendada de processar até 10MB de dados, suscetibilidade a alucinações, e falhas operacionais que podem exigir a reinicialização das sessões. Este estudo é uma contribuição valiosa para a literatura, pois fundamenta futuras pesquisas e oferece perspectivas importantes para profissionais e pesquisadores sobre o uso de LLMs na análise de dados, explorando capacidades analíticas e desafios técnicos, e ampliando a compreensão teórica enquanto investiga aplicações práticas em cenários reais.

As próximas seções deste artigo estão organizadas da seguinte maneira: a Seção 2 oferece uma revisão dos trabalhos relacionados, estabelecendo o contexto para a pesquisa. Em seguida, a Seção 3 detalha os dados e métodos empregados no estudo, fornecendo uma base sólida para a compreensão das técnicas utilizadas. A análise dos resultados é apresentada na Seção 4, onde são discutidas as implicações dos achados. Finalmente, a Seção 5 conclui o estudo e explora possíveis direções para pesquisas futuras, delineando o potencial impacto e as aplicações práticas do trabalho realizado.

[4]ChatGPT - https://chat.openai.com/

## 2 TRABALHOS RELACIONADOS

Esta seção explora as tendências atuais e os desafios associados ao uso de LLMs em análise de dados. Ao detalhar aplicações práticas e limitações, proporcionamos uma base robusta para discutir o papel do Data Analyst do ChatGPT nesse panorama dinâmico.

### 2.1 Aplicações de LLMs na Análise de Dados

A especialização dos LLMs para tarefas específicas de domínio evidencia seu potencial de adaptação a nichos específicos, destacando a versatilidade dos modelos e sua capacidade de serem otimizados para atender a necessidades particulares. Os estudos de Ding et al. [9] e Cheng et al. [7] são essenciais para compreender a eficácia dos LLMs em tarefas de análise de dados. Eles avaliam a precisão e eficiência do GPT-3 e GPT-4 em anotar e analisar dados, respectivamente. Ambos os estudos demonstraram eficácia em suas análises, sendo diretamente relevantes para nosso estudo de caso ao fornecer um contexto sobre o desempenho esperado de um LLM em operações analíticas semelhantes às realizadas pelo Data Analyst do ChatGPT.

Adicionalmente, conforme proposto por Sharma et al. [19], os LLMs podem ser adaptados para automatizar transformações de dados em setores específicos, como o energético. O modelo foi capaz de realizar transformações complexas de dados, reduzindo significativamente o tempo e o esforço comparados a métodos manuais. A relevância deste estudo para o nosso trabalho reside na demonstração prática da capacidade dos LLMs de manipular e transformar dados eficientemente em um cenário de dados específico e real, um aspecto crucial no processo de análise de dados.

### 2.2 LLMs na Automação de Tarefas Analíticas

A pesquisa sobre a automação de tarefas analíticas por meio de LLMs é explorada por Nasseri et al. [16] e por Jaimovitch-López et al. [12]. Esses trabalhos destacam como os LLMs podem simplificar e automatizar a preparação e manipulação de dados, respectivamente. Ambos obteram resultados eficazes em uma variedade de tarefas cruciais no fluxo de trabalho analítico. Dessa forma, o alinhamento com nosso estudo surge ao avaliarmos como o Data Analyst do ChatGPT pode facilitar processos semelhantes, aumentando a eficiência e a precisão das análises.

Adicionalmente, a abordagem proposta por Zhang et al. [28] explora uma avaliação comparativa de diferentes agentes de ciência de dados, incluindo LLMs, e fornece uma visão importante sobre o desempenho dessas tecnologias em uma ampla gama de tarefas analíticas. Este estudo é crucial para possíveis pesquisas futuras, fornecendo um benchmark que pode ser usado para determinar a posição do Data Analyst em relação a outros agentes de ciência de dados.

## 2.3 Desafios e Perspectivas Futuras dos LLMs

Os desafios e limitações dos LLMs, além das perspectivas futuras, são explorados em estudos como os de Kasetty et al. [14] e Liu et al. [15], que examinam a capacidade dos LLMs em realizar raciocínio interventivo e causal, áreas que representam desafios significativos para as tecnologias de IA atuais. Este trabalho se relaciona com esses estudos ao destacar as limitações do Data Analyst do ChatGPT em tarefas de análise preditiva e prescritiva, que requerem raciocínio avançado.

Ademais, Achiam et al. [3] destacam avanços na arquitetura do modelo GPT mais recente, ressaltando melhorias na precisão e na profundidade da compreensão contextual. Em paralelo, Bubeck et al. [6] demonstram que o GPT-4 possui uma inteligência mais abrangente em comparação a modelos anteriores, exibindo desempenho próximo ao humano em diversos domínios, incluindo programação. Portanto, esses estudos são essenciais para compreender as tendências de desenvolvimento que podem influenciar futuras implementações de LLMs em análises de dados, fornecendo um contexto tecnológico para discutir a evolução potencial do Data Analyst.

# 3 METODOLOGIA

Esta seção delineia a metodologia empregada para avaliar a eficácia do Data Analyst do ChatGPT na resolução de problemas específicos de análise de dados. Um elemento crucial dessa avaliação é o entendimento e a utilização de "prompts". Um prompt é um comando textual enviado ao modelo, que serve para instruí-lo ou solicitar uma resposta específica. No contexto dos LLMs, os prompts são usados como comandos ou perguntas que direcionam o modelo quanto ao tipo de informação ou processamento requerido. Neste estudo, empregaremos um prompt de pré-processamento seguido de um prompt com uma questão específica de análise de dados para examinar como o Data Analyst interpreta e processa essas entradas, visando gerar respostas úteis e precisas.

A Seção 3.1 apresenta o conjunto de dados utilizado em nossos experimentos. Em seguida, a Seção 3.2 apresenta a estratégia adotada para o pré-processamento dos dados. As Seções 3.3, 3.4, 3.5 e 3.6 discutem as estratégias para condução das análises descritiva, diagnóstica, preditiva e prescritiva, respectivamente. Por fim, a Seção 3.7 apresenta as métricas utilizadas nas avaliações experimentais.

## 3.1 Conjunto de dados

O conjunto de dados utilizado neste estudo abrange informações acadêmicas dos estudantes de Ciência da Computação da Universidade Federal de Campina Grande (UFCG), originárias do sistema de Controle Acadêmico — uma plataforma de acompanhamento do percurso acadêmico dos alunos na universidade. Esses dados foram disponibilizados para pesquisa, garantindo o anonimato dos estudantes. Cada registro no conjunto de dados representa uma matrícula efetuada pelo aluno, incluindo detalhes sobre as disciplinas cursadas, como notas, carga horária e tipo de matrícula, além de dados pessoais do aluno, como o período e a idade de ingresso no curso.

O conjunto original possui 37.9 MB, com 150.703 registros distribuídos em 34 colunas. No entanto, considerando a recomendação de que a ferramenta analítica otimiza análises em conjuntos de até 10MB, foi realizada uma amostra estratificada. A seleção foi baseada na coluna que identifica o setor das disciplinas, garantindo que a amostra represente todos os setores. Esta amostra estratificada representa 20% do total dos dados, resultando em um arquivo de 8.2 MB com 30.130 linhas e mantendo as 34 colunas originais.

## 3.2 Pré-processamento dos dados

Após a criação da amostra do conjunto de dados, foi essencial realizar o pré-processamento no conjunto de dados para obter resultados mais claros e coerentes. Um dos primeiros passos foi delimitar o período de ingresso dos estudantes entre os anos letivos de 2006.1 a 2019.2, abrangendo 14 anos de curso. Esse intervalo foi escolhido porque contém dados mais completos, facilitando uma análise detalhada das informações. Os dados do semestre de 2020.1 não foram incluídos, pois este período foi cancelado devido à pandemia do SARS-CoV-2.

Posteriormente, as matrículas classificadas como "Dispensa" foram removidas, por representarem situações em que o aluno é isento de cursar uma disciplina específica devido a conhecimentos prévios ou créditos equivalentes adquiridos. Essas matrículas não são úteis para as análises propostas, uma vez que resultam em aprovação automática do estudante sem avaliação por nota. Além disso, as matrículas categorizadas como "Em curso" também foram excluídas. Como o aluno ainda não concluiu a disciplina, não possui nota final e sua situação acadêmica pode variar entre aprovado, reprovado por nota ou por falta, o que compromete a integridade dos dados analisados. Por fim, no prompt de pré-processamento é solicitado que o dataframe pós-processamento tenha um nome definido, para maior coesão.

Assim, estabeleceu-se o fluxo para cada sessão, onde o primeiro passo consiste em carregar a amostra do conjunto de dados, inserir o prompt padrão de pré-processamento e fazer uma pergunta sobre a análise dos dados ao Data Analyst. Cada pergunta é abordada em uma nova sessão de chat para garantir que as respostas não influenciem as subsequentes. O processo de pré-processamento provou ser consistente em todas as perguntas, produzindo conjuntos de dados filtrados idênticos.

Portanto, este é o prompt padrão de pré-processamento:
"Analise os dados e limpe as colunas da seguinte forma:
periodo_ingresso: de 2006.1 até 2019.2
tipo_matricula: remover "Dispensa"
situacao: remover "Em curso"
o novo dataframe após a limpeza deve se chamar df_filtrado"

## 3.3 Análise Descritiva

A análise de dados descritiva visa apresentar aspectos fundamentais dos dados, proporcionando uma visão clara dos eventos ocorridos, para identificar padrões e tendências. Geralmente, constitui o primeiro passo na análise de dados, proporcionando uma compreensão inicial das características e distribuições dos dados antes de avançar para análises mais complexas.

Para cada nível de dificuldade das perguntas, foram definidos critérios específicos. Em perguntas fáceis, o foco é a compreensão dos dados por meio de cálculos simples, como sumarização, média, mínimo e máximo. Em perguntas de nível médio, as métricas exigem processamentos adicionais, envolvendo a criação de novas

colunas, percentuais, distribuição e cálculos de médias máximas e mínimas. Em perguntas difíceis, as métricas abrangem processos mais complexos, como correlação, entropia e assimetria. A seguir, as perguntas estão agrupadas por nível de dificuldade.

#### 3.3.1 *Perguntas fáceis.*

(1) Qual é a quantidade de alunos por tipo de ingresso?
(2) Qual é a carga horária média cursada no primeiro período?
(3) Qual é a idade de evasão máxima e mínima?

#### 3.3.2 *Perguntas médias.*

(1) Qual é a proporção de alunos evadidos por ano de ingresso?
(2) Qual é a distribuição de notas pelo período da matrícula?
(3) Considere que a média geral do aluno é definida como a média de todas as notas não nulas nas disciplinas que o aluno cursou. Qual foi o aluno com menor quantidade de períodos cursados e maior média geral evadido como graduado?

#### 3.3.3 *Perguntas difíceis.*

(1) Considere que a média geral do aluno é definida como a média de todas as notas não nulas nas disciplinas que o aluno cursou. Além disso, considere que a cada ano existem dois períodos, o primeiro período do ano é caracterizado por terminar com ".1" e o segundo período do ano termina com ".2", por exemplo, no ano de 2015 existem os períodos 2015.1 e 2015.2. A partir destas definições, qual é a correlação entre a média geral dos alunos e ingressar em cada periodo do ano (primeiro e segundo)?
(2) Como a entropia da distribuição de alunos por setor acadêmico mudou ao longo dos últimos 5 períodos registrados?
(3) Considere que a média geral do aluno é definida como a média de todas as notas não nulas nas disciplinas que o aluno cursou. Qual é o grau de assimetria na distribuição das médias gerais dos alunos e como isso afeta o desempenho acadêmico geral?

## 3.4 Análise Diagnóstica

A análise de dados diagnóstica normalmente acontece após a análise descritiva e busca compreender as causas subjacentes dos eventos observados. Ela emprega técnicas para explorar padrões de dados e identificar os fatores que influenciaram resultados específicos.

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, recorreu-se 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.

#### 3.4.1 *Perguntas fáceis.*

(1) Qual é a taxa de aprovação por disciplina?
(2) Como a frequência de evasão mudou ao longo dos períodos?
(3) Para os alunos que entraram por ação afirmativa, qual é o percentil 70 da idade de ingresso?

#### 3.4.2 *Perguntas médias.*

(1) Determine se a forma de ingresso tem um impacto significativo nas taxas de graduação e evasão
(2) Considere que o Índice de Dificuldade da Disciplina é calculado como a média da diferença entre a média de notas da disciplina e a média geral de todas as disciplinas. Qual é o índice de dificuldade da disciplina Cálculo Diferencial e Integral I?
(3) Quão consistentes são as notas dos alunos ao longo do tempo?

#### 3.4.3 *Perguntas difíceis.*

(1) O tipo de ingresso influencia de forma estatisticamente significativa na quantidade de períodos até a graduação?
(2) Existe uma diferença estatisticamente significativa nas notas entre os alunos de matrícula Normal e Extra Curricular?
(3) Considere que a média geral do aluno é definida como a média de todas as notas não nulas nas disciplinas que o aluno cursou. A ação afirmativa, a forma de ingresso e o sexo do aluno tem alguma influência significativa na sua média geral?

## 3.5 Análise Preditiva

A análise preditiva de dados emprega modelos estatísticos e algoritmos de aprendizado de máquina para prever eventos futuros a partir de padrões derivados de dados históricos e atuais. Neste estudo, utilizou-se a técnica Chain of Thoughts, conforme proposto por Wei et al. [25], uma abordagem de prompt de raciocínio, desenvolvida para aprimorar a capacidade de resposta dos LLMs. Essa técnica foi aplicada para solicitar à ferramenta a recomendação de três soluções para a questão apresentada, selecionando a mais adequada. Para perguntas difíceis, a ferramenta é instruída a realizar uma análise avançada, visando escolher a solução mais completa.

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.

#### 3.5.1 *Perguntas fáceis.*

(1) Considere que a média geral do aluno é definida como a média de todas as notas não nulas nas disciplinas que o aluno cursou. A idade de ingresso e idade de evasão do aluno tem alguma influência significativa na sua média geral? Defina 3 opções de como solucionar essa questão e siga a melhor.
(2) Qual é a probabilidade da forma de saída do aluno ser graduado versus evadido, baseando-se na forma de ingresso, período de ingresso e situação acadêmica? Defina 3 opções de como solucionar essa questão e siga a melhor.
(3) É possível classificar alunos em categorias de desempenho acadêmico (por exemplo, alto, médio, baixo) com base em suas médias finais e carga horária de disciplinas? Defina 3 opções de como solucionar essa questão e siga a melhor.

#### 3.5.2 *Perguntas médias.*

(1) Considere que a média geral do aluno é definida como a média de todas as notas não nulas nas disciplinas que o aluno cursou. Qual é a probabilidade de um aluno com a média geral abaixo de 7.0 e com mais de 3 reprovações, ser aprovado na próxima disciplina? Defina 3 opções de como solucionar essa questão e siga a melhor.
(2) É possível identificar padrões de similaridade entre alunos com forma de saída evadido e graduado, considerando a forma de ingresso e a carga horária das disciplinas? Defina 3 opções de como solucionar essa questão e siga a melhor.
(3) É possível determinar a forma de saída de um aluno com base em características como o número de créditos cursados, tipo de matrícula e situação das disciplinas? Defina 3 opções de como solucionar essa questão e siga a melhor.

#### 3.5.3 *Perguntas difíceis.*

(1) Por meio de uma análise avançada, podemos obter o desempenho de um aluno em PROGRAMAÇÃO II, baseando-se no seu desempenho nas disciplinas de PROGRAMAÇÃO l e LABORATÓRIO DE PROGRAMAÇÃO I? Defina 3 opções de como solucionar essa questão e siga a melhor.
(2) Por meio de uma análise avançada, é possível prever a forma de ingresso com base no período de ingresso, sexo, ação afirmativa e idade de ingresso? Defina 3 opções de como solucionar essa questão e siga a melhor.
(3) Por meio de uma análise avançada, podemos obter a trajetória acadêmica de um aluno ao longo do tempo, utilizando a sua sequência de notas e a sua situação nas disciplinas (como Aprovado e Reprovado), para antecipar a possibilidade de uma situação de trancamento no futuro? Defina 3 opções de como solucionar essa questão e siga a melhor.

## 3.6 Análise Prescritiva

A análise prescritiva busca antecipar o que pode acontecer no futuro, permitindo recomendar ações específicas que influenciem positivamente os resultados. Nesta análise, a ferramenta deve recomendar três soluções para a questão apresentada e selecionar a mais adequada, utilizando a técnica Chain of Thoughts [25]. Para as perguntas difíceis, ela é orientada a realizar uma análise avançada, a fim de escolher a solução mais completa.

Ao definir as perguntas fáceis, a ferramenta deve empregar modelos de IA simples e análises estatísticas. Para as perguntas médias, busca-se modelos mais robustos, como Análises Temporais. Por fim, nas perguntas difíceis, utilizar modelos de Aprendizado de Máquina Profundo, como Redes Neurais Artificiais. Em seguida, a classificação das perguntas por nível de dificuldade.

#### 3.6.1 *Perguntas fáceis.*

(1) Com base no histórico de reprovações por falta, quantos casos ocorrerão no próximo período da disciplina de LABORATÓRIO DE PROGRAMAÇÃO I? Defina 3 opções de como solucionar essa questão e siga a melhor.
(2) Quais são as variáveis que mais impactam na diferenciação entre alunos aprovados e reprovados? Defina 3 opções de como solucionar essa questão e siga a melhor.
(3) Qual é o desempenho do limite inferior a 10% dos alunos em disciplinas do setor de matemática? Defina 3 opções de como solucionar essa questão e siga a melhor.

#### 3.6.2 *Perguntas médias.*

(1) De que forma o perfil do aluno que cursou disciplinas do tipo extra-curricular, nunca reprovou e nunca trancou o curso, influencia significativamente o seu sucesso acadêmico? Defina 3 opções de como solucionar essa questão e siga a melhor.
(2) Qual é a previsão de taxa de graduação para o próximo ano baseada em tendências passadas? Defina 3 opções de como solucionar essa questão e siga a melhor.
(3) É possível determinar as tendências de evasão de alunos ao longo do tempo, baseando-nos em padrões históricos de desempenho acadêmico e tipos de matrícula? Defina 3 opções de como solucionar essa questão e siga a melhor.

#### 3.6.3 *Perguntas difíceis.*

(1) Por meio de uma análise avançada, com base na redução de dimensionalidade das características acadêmicas dos alunos, como podemos prever quem são os alunos em risco de evasão? Defina 3 opções de como solucionar essa questão e siga a melhor.
(2) Por meio de uma análise avançada, é possível gerar novos perfis de alunos que maximizem a probabilidade de graduação, baseado nas características de alunos previamente graduados? Defina 3 opções de como solucionar essa questão e siga a melhor.
(3) Por meio de uma análise avançada, é possível identificar os momentos críticos na trajetória acadêmica de um aluno, como períodos onde o risco de evasão é maior, baseando-se em sequências de notas e situações acadêmicas? Defina 3 opções de como solucionar essa questão e siga a melhor.

## 3.7 Métricas

As métricas são analisadas após cada pergunta. Avaliamos a resposta do Data Analyst, considerando tanto o conteúdo textual quanto o código gerado. Classificamos a resposta como "Certa" se ela atende ao solicitado ou "Errada" se falha em satisfazer a questão proposta de forma adequada.

Além da análise da resposta, identificamos problemas potenciais, como alertas, que são mensagens emitidas durante a compilação ou execução. Esses alertas indicam práticas potencialmente problemáticas, mas não impedem a execução. É importante observar que as versões utilizadas internamente pela ferramenta são o Python 3.11.8 e o Pandas 1.5.3, enquanto as versões mais atuais são Python 3.12.3 [2] e Pandas 2.2.2 [1], o que pode resultar em problemas no código gerado.

Outro problema frequente é a intercorrência, que interrompe o fluxo de resposta. Isso pode ser causado por falhas técnicas, como problemas na conexão, ou por um erro fatal que elimina a aba de digitação, exigindo a abertura de um novo chat e o reinício da análise. Também pode ser devido a um erro de processamento que pausa temporariamente a análise para recálculo e, em seguida, a retoma.

# 4 RESULTADOS

Esta seção apresenta os resultados obtidos em nossos experimentos.

## 4.1 Análise Descritiva

Durante a Análise Descritiva, todas as perguntas foram respondidas corretamente, utilizando os dados adequados e aplicando as métricas solicitadas. Não houve problemas de alertas ou intercorrências. Na questão de nível médio 2 foi gerado um gráfico de caixas (boxplot) para fornecer suporte à análise. A Tabela 1 apresenta os resultados obtidos na Análise Descritiva.

| Nível | Pergunta | Resposta | | Problema | |
|---|---|---|---|---|---|
| | | Correta | Incorreta | Alerta | Intercorrência |
| Fácil | 1 | X | | | |
| | 2 | X | | | |
| | 3 | X | | | |
| Médio | 1 | X | | | |
| | 2 | X | | | |
| | 3 | X | | | |
| Difícil | 1 | X | | | |
| | 2 | X | | | |
| | 3 | X | | | |

**Tabela 1: Resultados da Análise Descritiva**

## 4.2 Análise Diagnóstica

Durante a Análise Diagnóstica, todas as perguntas foram respondidas corretamente, empregando dados adequados e as métricas requisitadas. Contudo, uma intercorrência por falha na conexão ocorreu na pergunta de nível difícil 3, exigindo a reexecução de parte da análise, o que não afetou os resultados finais.

Adicionalmente, foram registrados quatro alertas. Os dois primeiros, referentes às perguntas de nível médio 1 e 3, geraram o alerta "SettingWithCopyWarning". Esse aviso, da biblioteca Pandas, ocorre ao tentar modificar um conjunto de dados que é uma cópia de uma fatia de um DataFrame original, sem alterar este último. Embora este alerta não tenha causado erros diretos nas análises, ele é significativo, pois alterações não refletidas no DataFrame original podem resultar em dados inesperados.

O terceiro alerta surgiu na questão difícil 2, durante a tentativa de aplicação do teste de T de Student, o qual requer normalidade e homogeneidade de variâncias. Inicialmente, o teste de normalidade Shapiro-Wilk, executado pela biblioteca SciPy, emitiu o aviso "UserWarning: p-value may not be accurate for N > 5000", indicando que para grandes tamanhos de amostra o valor-p pode não refletir com precisão a normalidade dos dados. Apesar disso, a ferramenta prosseguiu para o teste de homogeneidade de variâncias, que também não foi satisfeito, impedindo a aplicação do teste de T de Student. Embora este alerta não afete os resultados diretamente, merece atenção.

Na pergunta de nível difícil 3, foram gerados gráficos de caixas (boxplot) para apoiar a análise. Durante essa visualização, ocorreu o último alerta, "UserWarning: FixedFormatter should only be used together with FixedLocator", emitido pelo Matplotlib. Este aviso indica que as posições dos traços devem ser definidas antes de formatar os rótulos. Apesar disso, este alerta não impacta o resultado da visualização.

A Tabela 2 apresenta os resultados obtidos na Análise Diagnóstica.

| Nível | Pergunta | Resposta | | Problema | |
|---|---|---|---|---|---|
| | | Correta | Incorreta | Alerta | Intercorrência |
| Fácil | 1 | X | | | |
| | 2 | X | | | |
| | 3 | X | | | |
| Médio | 1 | X | | X | |
| | 2 | X | | | |
| | 3 | X | | X | |
| Difícil | 1 | X | | | |
| | 2 | X | | X | |
| | 3 | X | | X | X |

**Tabela 2: Resultados da Análise Diagnóstica**

## 4.3 Análise Preditiva

Na Análise Preditiva, a ferramenta acertou sete questões, errou duas, apresentou alertas em quatro e intercorrências em duas. Nas questões fáceis e médias, optou frequentemente por abordagens diretas e uma modelagem de dados simplificada que, embora elementares, satisfizeram os requisitos. Com a devida especificação nas perguntas complexas, observou-se uma transição dos modelos utilizados de regressões simples para a utilização de Floresta Aleatória, permitindo análises mais sofisticadas.

Na primeira resposta incorreta, ao abordar a pergunta média 3, utilizou-se uma Árvore Aleatória que obteve apenas 54.5% de precisão. Durante a execução, o modelo gerou um "UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples" da biblioteca scikit-learn, sinalizando múltiplos erros e confirmando o desempenho insatisfatório. Para tentar corrigir, a ferramenta expandiu o conjunto de características, mas ao aplicar técnicas de balanceamento de classes, descobriu-se que a biblioteca "imblearn" não era compatível com o ambiente. Mesmo após aumentar o conjunto de variáveis e retreinar o modelo, a precisão continuou inalterada. Posteriormente, uma intercorrência levou ao encerramento da sessão. Devido à baixa

precisão e à falha em resolver o problema, a questão foi marcada como errada.

A outra questão marcada como errada foi a difícil 3. Durante a preparação para a modelagem, o modelo alucinou, gerando resultados incorretos, realizando cálculos em colunas inexistentes nos dados originais e que não foram previamente criadas por ele. Esse erro ilustra uma limitação significativa da ferramenta. Durante esses processamentos equivocados, ocorreram ainda problemas de intercorrência por recalculo.

Quanto aos demais alertas, o primeiro ocorreu na pergunta fácil 2: um FutureWarning da biblioteca scikit-learn, alertando que o parâmetro sparse foi renomeado para sparse_output. O segundo, na pergunta fácil 3, um "SettingWithCopyWarning" do Matplotlib, referente à modificação de um conjunto de dados que é uma cópia de uma fatia de um DataFrame sem alterar o original. O terceiro alerta, na pergunta média 2, "WARNING: matplotlib.legend: No artists with labels found to put in legend", ocorre ao tentar adicionar uma legenda a um gráfico sem elementos gráficos com rótulos. Nenhum desses alertas afetou as análises.

A Tabela 3 apresenta os resultados obtidos na Análise Preditiva.

| Nível | Pergunta | Resposta | | Problema | |
|---|---|---|---|---|---|
| | | Correta | Incorreta | Alerta | Intercorrência |
| Fácil | 1 | X | | | |
| | 2 | X | | X | |
| | 3 | X | | X | |
| Médio | 1 | X | | | |
| | 2 | X | | X | |
| | 3 | | X | X | X |
| Difícil | 1 | X | | | |
| | 2 | X | | | |
| | 3 | | X | | X |

**Tabela 3: Resultados da Análise Preditiva**

## 4.4 Análise Prescritiva

Na Análise Prescritiva, houve seis respostas corretas, três incorretas, dois alertas e quatro intercorrências. Nas questões de dificuldade fácil e média, a ferramenta escolheu consistentemente opções diretas e modelagens de dados simplificadas que, embora elementares, cumpriram os requisitos. Nas perguntas difíceis desta seção, a exigência de análises avançadas levou à adoção de modelos de maior complexidade, como Autoencoder, Modelo de Markov Oculto (HMM) e Floresta Aleatória, demonstrando um aumento significativo na sofisticação em relação à Análise Preditiva.

A questão média 1 foi classificada como incorreta. Durante a sua resolução, a ferramenta inicialmente aplicou Regressão Linear, alcançando um coeficiente de determinação de apenas 0.047, indicando uma explicação muito baixa da variância. Seguiu-se uma tentativa com Floresta Aleatória, que resultou em possível *overfitting*, com o coeficiente caindo para -0.215. Posteriormente, tentou-se ajustar os hiperparâmetros via validação cruzada, mas problemas técnicos interromperam o processo repetidamente. Devido à criação de um modelo insatisfatório que não superou os desafios, esta questão foi definida como errada.

A questão média 2 foi resolvida com sucesso, embora tenham ocorrido algumas intercorrências durante o processamento dos dados que exigiram recalculagens devido a erros. Na execução do modelo ARIMA para análise de séries temporais, a ferramenta gerou alertas da biblioteca statsmodels. Dois "ValueWarning" foram emitidos porque o modelo não pôde reconhecer o índice do DataFrame como um índice de data e hora, o que é esperado dado o formato de nosso conjunto de dados. Um "FutureWarning" também foi apresentado, indicando que essa condição será tratada como erro em futuras versões da biblioteca. Além disso, um "SettingWithCopyWarning" do Matplotlib foi observado. Apesar desses alertas, nenhum deles comprometeu as análises.

Um aspecto crucial dos resultados é a limitação da ferramenta em relação às bibliotecas disponíveis. A questão difícil 1 foi classificada como incorreta. Durante a preparação dos dados, um "SettingWithCopyWarning" do Matplotlib foi emitido, mas não afetou a análise. Na tentativa de implementar o Autoencoder, houve uma intercorrência, e foi relatado que o TensorFlow e o Keras não estavam instalados no ambiente atual. Um código base sugerido, ao ser executado em outro ambiente, resultou no erro "Failed to convert a NumPy array to a Tensor". Uma situação semelhante ocorreu com a questão difícil 3, onde, após o tratamento de dados ausentes e o *Feature Engineering*, uma intercorrência foi reportada durante a modelagem, devido à falta da biblioteca hmmlearn no ambiente da ferramenta.

A Tabela 4 apresenta os resultados obtidos na Análise Prescritiva.

| Nível | Pergunta | Resposta | | Problema | |
|---|---|---|---|---|---|
| | | Correta | Incorreta | Alerta | Intercorrência |
| Fácil | 1 | X | | | |
| | 2 | X | | | |
| | 3 | X | | | |
| Médio | 1 | | X | | X |
| | 2 | X | | X | X |
| | 3 | X | | | |
| Difícil | 1 | | X | X | X |
| | 2 | X | | | |
| | 3 | | X | | X |

**Tabela 4: Resultados da Análise Prescritiva**

### 4.5 Discussão

A partir do experimento realizado com o Data Analyst do ChatGPT, podemos discutir o seu desempenho. Foram realizadas 36 perguntas distribuídas pelas principais categorias de análise de dados. A Figura 1 apresenta a sumarização de todas as respostas e problemas identificados.

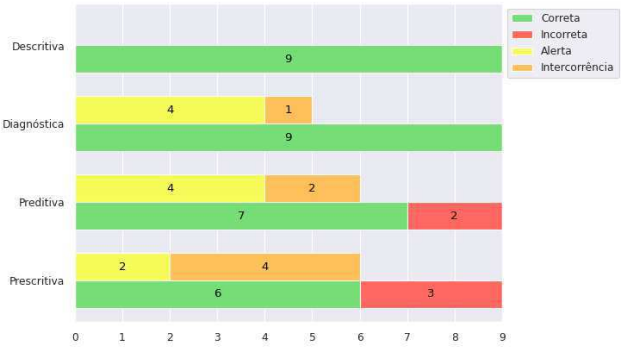


**Figura 1: Sumarização com o resultado das análises**

Podemos observar que a ferramenta apresentou excelentes resultados nas análises descritiva e diagnóstica, acertando todas as perguntas dessas categorias. No entanto, nas análises preditiva e prescritiva, ocorreram algumas respostas incorretas. Além disso, uma quantidade significativa de alertas e intercorrências foi registrada nas análises diagnóstica, preditiva e prescritiva.

Dessa forma, a partir da acurácia de 86,11%, a ferramenta demonstrou ser eficaz, principalmente nas análises descritivas e diagnósticas. Entretanto, as taxas de respostas incorretas (13,89%), alertas (27,78%) e intercorrências (19,44%) indicam desafios em termos de confiabilidade e estabilidade operacional, especialmente nas análises preditivas e prescritivas, por serem mais complexas e demandarem mais poder computacional.

As principais limitações identificadas incluem a restrição de processamento de dados a 10MB e a falta de suporte para bibliotecas robustas como hmmlearn, imblearn, Keras e Tensorflow. A ferramenta também se mostrou suscetível a alucinações e falhas operacionais que podem requerer reinicialização das sessões, comprometendo a eficiência e a eficácia do processo analítico.

## 5 CONCLUSÃO

O estudo avaliou o Data Analyst do ChatGPT, uma ferramenta baseada em LLM que auxilia na análise de dados, abordando as quatro categorias de análise: Descritiva, Diagnóstica, Preditiva e Prescritiva. Constatou-se que, embora os LLMs ofereçam um potencial transformador na automação de análise de dados, enfrentam barreiras técnicas significativas, como a integração com bibliotecas avançadas e o gerenciamento de grandes volumes de dados. O estudo contribui com perspectivas valiosas para a literatura, enriquecendo o conhecimento teórico e prático sobre a aplicação dos LLMs.

Para trabalhos futuros, planejamos realizar o ajuste fino do modelo GPT-4 com o objetivo de minimizar erros e alertas. Também pretendemos aplicar a metodologia deste estudo a outros conjuntos de dados para verificar a consistência dos resultados. Além disso, exploraremos a propensão do modelo a produzir alucinações em diferentes cenários. Por fim, utilizaremos o paradigma de avaliação proposto por Zhang et al. [28] para avaliar onde o Data Analyst do ChatGPT se posiciona em relação a outros agentes de ciência de dados. Essas investigações são fundamentais para ampliar nossa compreensão sobre a aplicabilidade e as limitações dos LLMs na automação da análise de dados.

## REFERÊNCIAS


[1] [n. d.]. Pandas Release Notes Documentation. https://pandas.pydata.org/docs/whatsnew/index.html. Acesso em: 08 de maio de 2024.

[2] [n. d.]. Python Release Notes Documentation. https://www.python.org/downloads/source/. Acesso em: 08 de maio de 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2024). https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023). https://doi.org/10.48550/arXiv.2303.12712 arXiv:2303.12712

[7] Liying Cheng, Xingxuan Li, and Lidong Bing. 2023. Is GPT-4 a Good Data Analyst? *Journal of Artificial Intelligence Research* Findings of the Association for Computational Linguistics: EMNLP 2023 (2023), 9496—-9514. https://doi.org/10.18653/v1/2023.findings-emnlp.637

[8] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

[9] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? 11173–11195. https://doi.org/10.18653/v1/2023.acl-long.626

[10] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (Eds.), Vol. 206. PMLR, 5549–5581. https://proceedings.mlr.press/v206/hegselmann23a.html

[11] Ihab Ilyas and Xu Chu. 2019. *Machine learning and probabilistic data cleaning*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3310205.3310213

[12] Gonzalo Jaimovitch-López, Cèsar Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, and María José Ramírez-Quintana. 2022. Can language models automate data wrangling? *Machine Learning* 112 (2022), 2053—-2082. https://doi.org/10.1007/s10994-022-06259-9

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. (2023). https://doi.org/10.48550/arXiv.2310.06825 arXiv:2310.06825

[14] Tejas Kasetty, Divyat Mahajan, Gintare Karolina Dziugaite, Alexandre Drouin, and Dhanya Sridhar. 2024. Evaluating Interventional Reasoning Capabilities of Large Language Models. *arXiv preprint arXiv:2404.05545* (2024). https://doi.org/10.48550/arXiv.2404.05545

[15] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking

Advanced Quantitative Reasoning with Data. (2024). https://doi.org/10.48550/arXiv.2402.17644 arXiv:2402.17644

[16] Mehran Nasseri, Patrick Brandtner, Robert Zimmermann, Taha Falatouri, Farzaneh Darbanian, and Tobechi Obinwanne. 2023. Applications of Large Language Models (LLMs) in Business Analytics – Exemplary Use Cases in Data Preparation Tasks. 14059 (2023), 182–198. https://doi.org/10.1007/978-3-031-48057-7_12

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[18] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2017. Data Management Challenges in Production Machine Learning. (2017), 1723–1726. https://doi.org/10.1145/3035918.3054782

[19] Ankita Sharma, Xuanmao Li, Hong Guan, Guoxin Sun, Liang Zhang, Lanjun Wang, Kesheng Wu, Lei Cao, Erkang Zhu, Alexander Sim, Teresa Wu, and Jia Zou. 2023. Automatic Data Transformation Using Large Language Model - An Experimental Study on Building Energy Data. (12 2023), 1824–1834. https://doi.org/10.1109/BigData59044.2023.10386931

[20] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. *CoRR* abs/1908.09203 (2019). https://doi.org/10.48550/arXiv.1908.09203 arXiv:1908.09203

[21] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2024. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2024). https://doi.org/10.48550/arXiv.2312.11805 arXiv:2312.11805

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971

[23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://doi.org/10.48550/arXiv.2307.09288 arXiv:2307.09288

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[26] Ahmet Yalçın, Hüseyin Selçuk Kılıç, and Dursun Delen. 2022. The use of multi-criteria decision-making methods in business analytics: A comprehensive literature review Multi-criteria decision making (MCDM) Multi-attribute decision-making (MADM) Multi-objective decision-making (MODM). *Technological Forecasting and Social Change* 174 (01 2022), 121193. https://doi.org/10.1016/j.techfore.2021.121193

[27] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. 2023. Large Language Models as Data Preprocessors. (2023). https://doi.org/10.48550/arXiv.2308.16361 arXiv:2308.16361

[28] Yuge Zhang, Qiyang Jiang, Xingyu Han, Nan Chen, Yuqing Yang, and Kan Ren. 2024. Benchmarking Data Science Agents. *arXiv e-prints*, Article arXiv:2402.17168 (feb 2024), arXiv:2402.17168 pages. https://doi.org/10.48550/arXiv.2402.17168 arXiv:2402.17168