



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ  
CAMPUS PEDRO II**

**TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS**

**YARA BEATRIZ FERREIRA GOMES**

**APIs de visão computacional em plataformas de *machine learning* como serviço  
(MLaaS): uma análise comparativa**

**PEDRO II**

**2025**

YARA BEATRIZ FERREIRA GOMES

APIs de visão computacional em plataformas de *machine learning* como serviço (MLaaS): uma análise comparativa

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Pedro II.

Orientador: Prof. Dr. Manuel Gonçalves da Silva Neto.

PEDRO II

2025

## APIs de visão computacional em plataformas de *machine learning* como serviço (MLaaS): uma análise comparativa

Yara Beatriz Ferreira Gomes<sup>1</sup>

Prof. Dr. Manuel Gonçalves da Silva Neto<sup>2</sup>

### RESUMO

A computação em nuvem tem impulsionado o uso de *Machine Learning as a Service (MLaaS)*, oferecendo escalabilidade e acesso simplificado a recursos de inteligência artificial. Dentro desse contexto, este trabalho realiza uma análise comparativa das principais APIs de visão computacional oferecidas pelas plataformas *Amazon Web Services (AWS)*, *Microsoft Azure*, *Google Cloud Platform (GCP)*, *Google DeepMind (Gemini)* e *OpenAI (ChatGPT)*. A pesquisa adota uma abordagem documental e exploratória, considerando critérios técnicos como funcionalidades, custo-benefício, usabilidade e integração com o ecossistema das plataformas. Os resultados indicam que as APIs baseadas em nuvem (*AWS*, *Azure* e *GCP*) apresentam robustez e alto desempenho, sendo adequadas a ambientes corporativos, enquanto os modelos generativos (*Gemini* e *ChatGPT*) se destacam pela flexibilidade e prototipação rápida. Conclui-se que a escolha da solução de *MLaaS* deve considerar o nível de complexidade da aplicação, o custo operacional e a experiência do usuário.

**Palavras-chave:** aprendizado de máquina; computação em nuvem; amazon aws, microsoft azure, google cloud.

### ABSTRACT

*Cloud computing* has driven the use of *Machine Learning as a Service (MLaaS)*, providing scalability and simplified access to artificial intelligence resources. In this context, this study performs a comparative analysis of major computer vision APIs offered by *Amazon Web Services (AWS)*, *Microsoft Azure*, *Google Cloud Platform (GCP)*, *Google DeepMind (Gemini)*, and *OpenAI (ChatGPT)*. The research follows a documental and exploratory approach, focusing on technical criteria such as functionality, cost-effectiveness, usability, and ecosystem integration. Results show that cloud-based APIs (*AWS*, *Azure*, and *GCP*) offer robustness and high performance, making them suitable for enterprise environments, while generative models (*Gemini* and *ChatGPT*) stand out for flexibility and rapid prototyping. It is concluded that the choice of an *MLaaS* vision solution should consider the application's complexity, operational cost, and user experience.

**Keywords:** machine learning; cloud computing; amazon aws, microsoft azure, google cloud.

Data da aprovação: 11/08/2025

## 1 INTRODUÇÃO

A computação em nuvem ou *Cloud Computing* é um modelo de computação que fornece acesso sob demanda a conjuntos compartilhados de recursos de computação, como

<sup>1</sup> Graduanda em Tecnologia em Análise e Desenvolvimento de Sistemas. Instituto Federal de Educação, Ciência e Tecnologia do Piauí – Campus Pedro II. caped.2023123tads0037@aluno.ifpi.edu.br.

<sup>2</sup> Dr. em Engenharia de Teleinformática. IFPI - Campus Pedro II. manuel@ifpi.edu.br.

servidores, armazenamento, redes e aplicativos, de forma escalável e a partir de um provedor (Freitas, 2023). O termo "*nuvem*" é utilizado como uma metáfora para agregar diversos serviços e recursos integrados entre si e disponíveis aos usuários (Shilpashree; Patil; Parvathi, 2018).

A computação em nuvem (*Cloud Computing*) tem servido de base e impulsionado o uso de *Machine Learning* à medida que a inteligência artificial transforma diversos setores, ao oferecer escalabilidade, acesso sob demanda a recursos computacionais de alto desempenho, como: GPUs (*Graphics Processing Units* ou Unidades de Processamento Gráfico), além de flexibilidade, agilidade e redução de custos, fatores essenciais para viabilizar o desenvolvimento e a execução de modelos de aprendizado de máquina (ML JOURNEY, 2024). A *Data Bridge Market Research* (Data Bridge Market Research, 2021) prevê que o mercado de aprendizagem automática como serviço ou *Machine Learning as a Service (MLaaS)* continuará crescendo nos próximos anos, atingindo um valor de 38,81 mil milhões de dólares até o final de 2028.

A aprendizagem automática como serviço (*MLaaS*) é considerada uma subcategoria dos serviços de computação em nuvem, esse serviço trata-se de um conjunto de funcionalidades que oferece uma ampla variedade de ferramentas e componentes de aprendizagem automática para realizar operações com eficiência e eficácia (Data Bridge Market Research, 2021).

Com base nisso, o *MLaaS* torna a infraestrutura de *Machine Learning* acessível, permitindo que organizações utilizem ferramentas de ML diretamente na nuvem sem mover dados, além de simplificar a incorporação de aprendizado de máquina em aplicativos com soluções prontas e modelos rápidos de criar, garantindo escalabilidade, integração com outros serviços em nuvem e permitindo que profissionais sem especialização técnica utilizem essas tecnologias (Data Bridge Market Research, 2021).

Além disso, as *APIs* (*Application Programming Interfaces* ou Interfaces de Programação de Aplicações) desempenham papel fundamental nesse ecossistema, pois são responsáveis por viabilizar a comunicação e integração entre os diversos serviços e recursos oferecidos pelas plataformas em nuvem. Conforme apontado por (Engineering Brasil, 2025) as *APIs* facilitam o consumo escalável e flexível dos serviços de *MLaaS*, permitindo que desenvolvedores criem aplicações otimizadas e integradas, essenciais para aproveitar plenamente as vantagens da computação em nuvem.

A escolha adequada de uma plataforma de *Machine Learning como serviço (MLaaS)* pode impactar diretamente no desempenho e custo das aplicações, tornando relevante a análise comparativa de diferentes soluções disponíveis no mercado (Synnex, 2024). Tendo em vista isso, realizou-se o levantamento das principais *APIs* disponíveis na *Amazon Web Service*

(AWS)<sup>3</sup>, Microsoft Azure<sup>4</sup>, Google (Google Cloud Platform<sup>5</sup> e Google DeepMind<sup>6</sup>) e OpenAI (ChatGPT)<sup>7</sup> por representarem provedoras globais de serviços em nuvem. Segundo Mufti *et al.* (2021), essas plataformas se destacam pela robustez dos recursos oferecidos, presença consolidada no mercado e estratégias competitivas voltadas à flexibilidade, inovação e variedade de serviços (Mufti; Mittal; Gupta, 2021). Neste contexto, o presente artigo tem como objetivo realizar uma análise comparativa das *APIs* voltadas para *MLaaS* disponíveis pela AWS, Azure, GCP, Google DeepMind e OpenAI. Empregou-se análise documental nas *APIs* disponíveis em plataforma apoiadas por uma revisão bibliográfica da literatura.

Este trabalho está organizado da seguinte forma. A Seção 1 apresenta uma introdução e conceitualização sobre a temática, problemática e os objetivos do estudo. A Seção 2 trata da fundamentação teórica, com as definições e características das abordagens relacionadas. Na Seção 3, descreve-se a metodologia adotada para a realização da pesquisa. A Seção 4 expõe os resultados obtidos, enquanto a Seção 5 apresenta as conclusões e recomendações para trabalhos futuros.

## 1.1 OBJETIVO GERAL

O presente artigo tem como objetivo prover uma análise comparativa das *APIs* relacionados ao *MLaaS* das plataformas Google (Cloud e DeepMind), Microsoft Azure, Amazon AWS e OpenAI (ChatGPT), incluindo uma descrição dos serviços, custo e recomendação de uso.

## 1.2 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo geral, os seguintes objetivos específicos foram definidos:

- Levantaram-se os principais serviços *MLaaS* das plataformas selecionadas;
- Extraíram-se informações sobre usabilidade, custo, utilidade e abrangência das documentações de cada *API*;
- Comparou-se as vantagens e desvantagens baseados nos critérios de usabilidade, abrangência da documentação e custo financeiro das soluções disponíveis;

## 2 FUNDAMENTAÇÃO TEÓRICA

<sup>3</sup> AWS (Amazon Web Services): Plataforma de computação em nuvem da Amazon.

<sup>4</sup> Microsoft Azure: Plataforma de computação em nuvem da Microsoft.

<sup>5</sup> Google Cloud Platform (GCP): Plataforma de computação em nuvem do Google.

<sup>6</sup> Google DeepMind: Laboratório de pesquisa em inteligência artificial do Google.

<sup>7</sup> OpenAI (ChatGPT): Organização de pesquisa em inteligência artificial, desenvolvedora do modelo Chat.

Esta seção apresenta conceitos fundamentais para o desenvolvimento deste trabalho de pesquisa.

## 2.1 APIS DE DESENVOLVIMENTO

Uma *API* (*Application Programming Interface* ou Interfaces de Programação de Aplicações) é uma interface de programação de *software* e web que permite a interação entre diferentes aplicações por meio de solicitações, possibilitando o compartilhamento de dados de forma segura e eficiente. Ela define regras e protocolos que facilitam o uso de funcionalidades de um sistema por outro, promovendo a integração entre softwares (Escola Britânica de Artes Criativas e Tecnologia, 2023).

As APIs são fundamentais para a integração e comunicação entre sistemas distribuídos, papel que se torna relevante no contexto da computação em nuvem. Conforme destacado no artigo “Transformação digital: A influência da computação em nuvem por meio de APIs” (Engineering Brasil, 2025), as APIs facilitam o desenvolvimento e gerenciamento de aplicações em ambientes cloud, impulsionando a transformação digital ao permitir que serviços sejam consumidos de forma escalável, flexível e eficiente.

Atualmente existem aplicações que utilizam APIs de *Machine Learning* fornecidas por plataformas em nuvem, como o *ChatGPT*<sup>8</sup> da *OpenAI* e o *Gemini*<sup>9</sup> da *Google*. Ambas utilizam APIs para treinamento como geração de texto, compreensão de linguagem natural, síntese de fala e integração com outras aplicações por meio de endpoints *REST* (Haponik, 2024). Essas ferramentas fazem uma demonstração de como os serviços de *MLaaS* pode ser utilizado para oferecer soluções inteligentes com uma alta escalabilidade e acessibilidade, servindo de referência prática de como as APIs impulsionam a adoção de *IA* em diferentes setores.

## 2.2 COMPUTAÇÃO EM NUVEM

A computação em nuvem é um modelo de computação que fornece acesso sob demanda a conjuntos compartilhados de recursos de computação, como servidores, armazenamento, redes e aplicativos pela *internet* sob demanda e de forma escalável, a partir de um provedor (Freitas, 2023). O termo “*nuvem*” serve como uma metáfora para a *Internet*, onde vários serviços e recursos estão disponíveis, exemplos notáveis de serviços de computação em nuvem incluem grandes empresas como *Google*, *Amazon* e *Microsoft*, que fornecem uma variedade de soluções baseadas em nuvem para os usuários (Shilpashree; Patil; Parvathi, 2018).

A computação em nuvem transformou a gestão de recursos ao fornecer serviços escaláveis e sob demanda pela Internet, sendo assim, surgindo como evolução dos modelos cliente-servidor,

---

<sup>8</sup> ChatGPT: Modelo de linguagem baseado em inteligência artificial, desenvolvido pela OpenAI, voltado à compreensão e geração de textos em linguagem natural.

<sup>9</sup> Gemini: Modelo de inteligência artificial multimodal desenvolvido pelo Google DeepMind.

ela permite o uso de infraestrutura de *TI* sem necessidade de recursos locais, o que trouxe uma flexibilidade, redução de custos e inovação, tornando-se essencial em diversos setores, além de garantir a escalabilidade necessária para a competitividade das empresas (Araujo, 2019; Freitas 2023). Com base nisso, dentro da computação em nuvem, é oferecido alguns tipos de nuvens para o uso, dentre eles, foram elencados os principais tipos oferecidos:

- **Nuvem Pública:** Pertence a provedores de serviços de nuvem terceirizados e são administradas por eles, fornecendo recursos de computação, tais como servidores e armazenamento pela Internet, com ela todo o *hardware*, *software* e outras infraestruturas de suporte são de propriedade e gerenciadas pelo provedor de serviços de nuvem os quais podem ser acessados e gerenciados por uma conta usando um navegador da *Web* (Microsoft, 2024; Freitas, 2023; Araujo, 2019).
- **Nuvem Privada:** Pode ser fornecida por um provedor de serviços ou ser *on premise*, ou seja, instalada no data center local da empresa, com acesso restrito aos usuários autorizados dessa nuvem privada, podendo ser local ou remota e administrada tanto pela própria empresa quanto por terceiros, adotando políticas de acesso aos serviços por meio de técnicas de gerenciamento de redes, configurações dos provedores de serviços e utilização de tecnologias de autenticação e autorização (Araujo, 2019; Freitas 2023).
- **Nuvem Híbrida:** É a combinação das nuvens públicas e privadas ligadas por uma tecnologia que permite que dados e aplicativos sejam compartilhados entre elas, sendo assim, permitindo que os dados e os aplicativos se movam entre nuvens privadas e públicas. Uma nuvem híbrida oferece à sua empresa flexibilidade, opções de implantação e ajuda a otimizar sua infraestrutura, segurança e conformidade existentes (Microsoft, 2024; Freitas, 2023; Araujo, 2019).

Nesse contexto, a computação em nuvem também oferece alguns modelos de serviços para o uso, dentre eles, foram elencados os principais serviços oferecidos: **Software como Serviço (SaaS)**, o qual oferece aplicativos de *software* prontos para uso pela *Internet*, permitindo que os usuários acessem o *software* sem a necessidade de instalação ou gerenciamento, um exemplo seria o *Google Colab* desenvolvido pela *Google*; **Infraestrutura como Serviço (IaaS)** que oferece recursos de infraestrutura escaláveis sob demanda, permitindo que as empresas gerenciem e personalizem seus ambientes de *TI*, um exemplo disso é o *Google Compute Engine*, que disponibiliza máquinas virtuais configuráveis para executar cargas de trabalho em escala, com flexibilidade e alta disponibilidade; **Plataforma como Serviço (PaaS)** o qual fornece uma plataforma que simplifica o desenvolvimento de aplicativos, eliminando a necessidade de os usuários gerenciarem a infraestrutura subjacente, ele oferece suporte ao desenvolvimento de *software* e microsserviços personalizados. Esses modelos podem ser utilizados tanto de maneira

isolada quanto combinada, um exemplo disso é o *Google App Engine*, que permite criar e hospedar aplicações web diretamente na infraestrutura do *Google*, com escalabilidade automática, suporte a várias linguagens de programação e integração com outros serviços da nuvem. (Younis *et al.*, 2024; Microsoft, 2024; Freitas, 2023; Araujo, 2019).

Com base nisso, a computação em nuvem vem se consolidando como um alicerce para a evolução de tecnologias emergentes, especialmente o ***Machine Learning as a Service (MLaaS)***, ao disponibilizar recursos escaláveis, flexíveis e acessíveis sob demanda, viabilizando a adoção de soluções de inteligência artificial por diferentes perfis organizacionais (Freitas, 2023). O MLaaS destaca-se como um modelo estratégico por reunir ferramentas de aprendizado de máquina acessíveis por meio de *APIs*, eliminando a necessidade de infraestrutura local robusta e conhecimento técnico aprofundado (Pletcher, 2023).

Sob essa perspectiva, as *APIs* cumprem papel fundamental ao operar o consumo desses serviços, promovendo a integração entre sistemas e acelerando o desenvolvimento de aplicações inteligentes (Engineering Brasil, 2025). Assim, plataformas como *Amazon Web Services (AWS)*, *Microsoft Azure*, *Google Cloud Platform (GCP)*, *Gemini (Google DeepMind)* e *ChatGPT (OpenAI)* consolidam-se como vetores dessa transformação digital ao oferecerem ecossistemas robustos e interfaces programáveis que ampliam o potencial analítico e preditivo das aplicações modernas (Mufti; Mittal; Gupta, 2021).

## 2.3 MACHINE LEARNING

O *Machine learning* ou Aprendizado de Máquina é um subconjunto da inteligência artificial que permite que um sistema aprenda e melhore de maneira autônoma usando redes neurais e aprendizado profundo, sem ter sido programado explicitamente para isso, ao ser alimentado com grandes quantidades de dados, o aprendizado de máquina permite também que os sistemas de computador se ajustem e se aperfeiçoem continuamente à medida que acumulam "*experiências*". Assim, o desempenho desses sistemas pode ser melhorado, fornecendo conjuntos de dados maiores e variados para processamento (Google, 2024; Janiesch; Zschech; Heinrich, 2021).

Com a crescente demanda por *Machine learning*, surgiram os serviços de *MLaaS*, que permitem às empresas utilizarem essa tecnologia sem a necessidade de um cientista de dados (Synnex, 2024). O *Machine Learning as Service (MLaaS)* refere-se à variedade de ferramentas de *Machine Learning* oferecidas como serviços por provedores de computação em nuvem, semelhante aos modelos de *SaaS* e o *PaaS*. Usar *Machine learning* como serviço significa obter acesso imediato a ferramentas eficientes pela Internet, sem a necessidade de investimento financeiro ou conhecimento técnico avançado (Pletcher, 2023). Com isso, é possível gerenciar grandes volumes

de dados e aproveitar o potencial da computação em nuvem utilizando ferramentas simples fornecidas pelos provedores de nuvem, pagando apenas pelo que for realmente necessário. Existem serviços prontos para uso que permitem empregar uma tecnologia de aprendizado de máquina eficaz por meio de uma chamada de *API* e poucos cliques (Pletcher, 2023).

Operações de *Machine Learning* ou *Machine Learning Operations (MLOps)* constituem um conjunto de práticas que automatizam e simplificam os fluxos de trabalho e a implantação de modelos (Amazon, 2024; Kreuzberger; Kuhl; Hirschl, 2023). Essa abordagem integra *Machine learning (ML)* com *DevOps* para promover a colaboração entre desenvolvedores e acelerar etapas como desenvolvimento, implantação, escalonamento, monitoramento e re-treinamento de modelos de *IA*. Com conceitos de integração e entrega contínuas (*CI/CD*), o *MLOps* permite às empresas automatizar a implantação e manutenção de modelos de *ML* em ambientes de produção, garantindo que os modelos estejam sempre atualizados e funcionando conforme o esperado (Moutaoakal; Baña, 2023).

Nesse cenário, o uso de *APIs* específicas para *Machine learning* integrado à nuvem se mostra essencial, pois, além de facilitar o acesso a grandes volumes de dados e reduzir gastos com infraestrutura, contribui para a padronização dos processos e a reutilização dos modelos de serviços em diferentes aplicações. As *APIs* favorecem a interoperabilidade entre sistemas, simplificando a integração de soluções de *ML* com ambientes de produção já existentes com baixo esforço de desenvolvimento. Essa abordagem agiliza o desenvolvimento e a implantação, permitindo o uso de modelos prontos via chamadas de *API* padronizadas, com suporte dos provedores de nuvem (Synnex, 2024; Pletcher, 2023; Amazon, 2024).

## 2.4 COMPUTAÇÃO EM NUVEM E MACHINE LEARNING

O aprendizado de máquina baseado em nuvem vem sendo essencial em setores como comércio eletrônico, bancos e saúde, devido à sua escalabilidade, custo-benefício e acessibilidade, já a computação em nuvem fornece o poder de processamento necessário para operações de treinamento e inferência, possibilitando que modelos de *ML* lidem com grandes volumes de dados sem perda de desempenho (Ravindranathan; Vadivu; Rajagopalan, 2024). Nesse contexto, o *Machine Learning as a Service (MLaaS)* surge como uma alternativa prática, aproveitando a computação distribuída e paralela para simplificar a execução de sistemas de *ML* (Ravindranathan; Vadivu; Rajagopalan, 2024; Noshiri; Khorramfar; Halabi, 2021). Com uma interface amigável, o *MLaaS* facilita a consulta a classificadores treinados em conjuntos de dados carregados, superando desafios relacionados ao armazenamento de dados e ao treinamento de modelos (Noshiri; Khorramfar; Halabi, 2021).

Empresas vem adotando o *MLaaS* para reduzir custos com infraestrutura local e minimizar

a necessidade de especialistas técnicos, com o modelo de preços flexível das soluções em nuvem as empresas conseguem ter um controle de gastos, alocando recursos conforme o consumo real (Noshiri; Khorramfar; Halabi, 2021).

Em decorrência disso, a acessibilidade do *ML* baseado em nuvem facilita o trabalho colaborativo entre equipes distribuídas globalmente, impulsionando a produtividade e a inovação ao mesmo tempo em que os serviços gerenciados simplificam a manutenção e permitem que as organizações se concentrem no desenvolvimento e implantação de seus modelos (Noshiri; Khorramfar; Halabi, 2021).

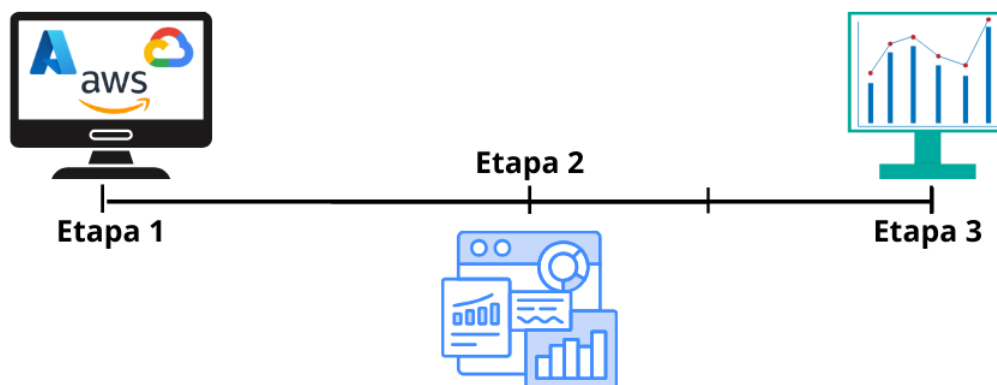
De forma complementar, as *APIs* desempenham um papel bem estratégico nesse ecossistema, pois são elas que possibilitam o acesso simplificado e padronizado aos serviços de Machine Learning oferecidos pelas plataformas em nuvem. Conforme destaca Engineering Brasil (2025), as *APIs* promovem o consumo flexível e escalável de serviços de *MLaaS*, permitindo que desenvolvedores criem aplicações otimizadas com agilidade e integração o (Engineering Brasil, 2025). Essa abordagem reduz a complexidade técnica envolvida no uso de modelos de aprendizado de máquina, ao mesmo tempo em que promove a portabilidade entre diferentes plataformas.

### 3 METODOLOGIA

Esta seção apresenta os materiais e métodos empregados nesta pesquisa.

A metodologia deste trabalho caracteriza-se como exploratória e documental. Com objetivo de prover um panorama comparativo sobre as funcionalidades e custo-benefício das *APIs* de *MLaaS* das plataformas *AWS*, *Azure*, *Google (DeepMind e GCP)* e *OpenAI (ChatGPT)* para desenvolvedores de aplicações de *Machine learning*. Este estudo está estruturado em três etapas: levantamento das principais *APIs*, síntese das informações e comparação entre os serviços e *APIs* disponíveis. A Figura 1 apresenta uma representação em alto nível das etapas.

Figura 1– Representação do fluxo lógico para a realização do projeto



- **Etapa 1:** levantaram-se os serviços que cada plataforma disponibiliza e uma descrição sobre as recomendações de uso presentes nas documentações oficiais.
- **Etapa 2:** Realizou-se uma análise documental sobre as funcionalidades de cada API e do nível de documentação disponibilizada pelos provedores de serviços *MLaaS*, destacando suas vantagens e limitações.
- **Etapa 3:** Realizou-se um comparativo dos resultados obtidos a partir das documentações, incluindo análise e síntese dos dados. Os dados coletados nas documentações foram analisados a fim de apresentar o custo-benefício de cada provedor em relação às vantagens e limitações de cada plataforma. Avaliou-se aqui itens como: disponibilidade e nível de cobertura da documentação, presença de casos de uso dos serviços, custo financeiro e a presença ou não de literatura disponível sobre as funcionalidades e serviços.

Acrescenta-se que para avaliação dos níveis de cobertura das documentações empregou-se critérios relacionados as Principais Funcionalidades; Integração com o Ecossistema; Facilidade de uso; Desempenho e Escalabilidade; Modelo de Precificação; SLA e Conformidade; Suporte e Comunidade e Casos de Uso. Metodologias similares foram empregadas com sucesso na avaliação de serviços relacionados a *APIs*, conforme apresentado por Mikael Nilton (Araujo, 2019)

## 4 RESULTADOS

Nesta seção, são apresentados os resultados obtidos a partir das documentações das *APIs* de *Machine Learning* oferecidas pelas plataformas e serviços *Amazon Web Services (AWS)*, *Microsoft Azure*, *Google Cloud Platform (GCP)*, *Google DeepMind (Gemini)* e *OpenAI (ChatGPT)*. Essa análise foi conduzidos com o intuito de avaliar o custo-benefício de cada plataforma conforme apresentado na Seção de Metodologia 3.

### 4.1 PLATAFORMAS DE MACHINE LEARNING E SUAS APIS

A Tabela 1 apresenta os links oficiais das documentações utilizadas para cada provedor *MLaaS*. Para cada provedor, levantou-se as *APIs* e suas funcionalidades.

Tabela 1 - Links da documentação utilizada de cada provedor *MLaaS*.

| Provedor                     | Link Documentação                                                                                                               |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| <i>Amazon AWS</i>            | <a href="https://docs.aws.amazon.com/?nc2=h ql doc do">https://docs.aws.amazon.com/?nc2=h ql doc do</a>                         |
| <i>Microsoft Azure</i>       | <a href="https://learn.microsoft.com/pt-br/azure/?product=popular">https://learn.microsoft.com/pt-br/azure/?product=popular</a> |
| <i>Google Cloud (GCP)</i>    | <a href="https://cloud.google.com/docs?hl=pt-br">https://cloud.google.com/docs?hl=pt-br</a>                                     |
| <i>Google Cloud (Gemini)</i> | <a href="https://ai.google.dev/gemini-api/docs?hl=pt-br">https://ai.google.dev/gemini-api/docs?hl=pt-br</a>                     |
| <i>OpenAI (ChatGPT)</i>      | <a href="https://platform.openai.com/docs/models/chatgpt-">https://platform.openai.com/docs/models/chatgpt-</a>                 |

A Tabela 2, apresenta uma visão geral das *APIs* de *Machine Learning* oferecidas pelas plataformas selecionadas. Agrupou-se por similaridade dos serviços oferecidos para fins de facilitar o processo de decisão por parte dos desenvolvedores. É importante ressaltar que neste trabalho, as plataformas *Google DeepMind* e *OpenAI* foram representadas pelos seus produtos principais *Gemini* e *ChatGPT* respectivamente, dando ênfase as *APIs* de desenvolvimento disponibilizadas por estas.

Percebe-se que serviços relacionados à visão computacional estão presentes em 3 das 5 plataformas avaliadas, destacando-se o *Amazon Rekognition (AWS)*, *Azure AI Vision (Azure)* e *Cloud Vision AI (GCP)*, evidenciando a importância dessa funcionalidade para análise de imagens e vídeos. Em seguida, serviços para análise de texto, como *Amazon Comprehend*, *Text Analytics* e *Cloud Natural Language*, também aparecem em 3 plataformas, mostrando a relevância do processamento de linguagem natural para extração de sentimentos, entidades e tópicos, serviços de conversão de texto e fala, incluindo transcrição e síntese vocal, estão igualmente presentes em 3 plataformas, indicando a demanda por interfaces de voz naturais.

Observa-se ainda a presença de *APIs* específicas para algumas plataformas, como o *Amazon Kendra da AWS*, voltado para busca semântica em bases corporativas, a *Face API* do *Azure*, que proporciona reconhecimento facial avançado com análise detalhada de emoções e o *Video Intelligence API* do *GCP*, que realiza análise detalhada de vídeos, também se destaca como serviço exclusivo. Acrescenta-se que, os modelos multimodais *Gemini* e *ChatGPT* oferecem funcionalidades integradas de geração e compreensão de linguagem natural, análise de imagens, sons e documentos estruturados, acessados por meio de interfaces unificadas e adaptáveis a diferentes contextos de uso. Os serviços de detecção de anomalias e previsão (*forecasting*) aparecem em 3 plataformas, reforçando seu papel em monitoramento e análise de dados temporais, enquanto sistemas de recomendação personalizados estão disponíveis nas três principais plataformas de nuvem (*AWS*, *Azure* e *GCP*).

Tabela 2 - *APIs* de *Machine Learning* em *AWS*, *Azure*, *GCP*, *Gemini* e *ChatGPT*

| Solução (Plataforma)                                                                                 | Descrição                                                                                                                                                     |
|------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Amazon Comprehend (AWS)</i> , <i>Text Analytics (Azure)</i> e <i>Cloud Natural Language (GCP)</i> | Usados para análise de textos para extrair sentimentos, entidades e tópicos relevantes, sendo utilizado em análises de opinião e processamento de documentos. |
| <i>Amazon Rekognition (AWS)</i> , <i>Azure AI Vision (Azure)</i> e <i>Cloud Vision AI (GCP)</i>      | Oferecem recursos de visão computacional, como reconhecimento facial, detecção de objetos e análise de vídeos.                                                |
| <i>Amazon Lex (AWS)</i> , <i>Language Understanding-</i>                                             | São voltados para a criação de chatbots e                                                                                                                     |

|                                                                                  |                                                                                                                                                                                                                                 |
|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>LUIS (Azure)</i>                                                              | assistentes virtuais com compreensão de linguagem natural, facilitando o atendimento automatizado em diferentes canais.                                                                                                         |
| <i>Amazon Polly (AWS), Speech (Azure) e Cloud Speech-to-Text (GCP)</i>           | Convertem fala em texto, o que tornam úteis para transcrição de chamadas, reuniões e conteúdo audiovisual.                                                                                                                      |
| <i>Amazon Transcribe (AWS), Speech (Azure) e Cloud Text-to-Speech (GCP)</i>      | Convertem texto em fala com vozes naturais, usados em aplicações de acessibilidade, audiobooks e interfaces faladas.                                                                                                            |
| <i>Amazon Translate (AWS), Translator (Azure) e Cloud Translation (GCP)</i>      | Realizam traduções automáticas entre vários idiomas, facilitando a comunicação em ambientes multilíngues.                                                                                                                       |
| <i>Amazon Textract (AWS), Form Recognizer (Azure) e Document AI (GCP)</i>        | Usados para a extração de texto e informações estruturadas de documentos digitalizados, como formulários e contratos.                                                                                                           |
| <i>Amazon Forecast (AWS), Anomaly Detector (Azure) e AI Forecasting (GCP)</i>    | Detectam anomalias e realizam previsões com base em dados temporais, o que auxilia no monitoramento e análise de demanda e comportamento.                                                                                       |
| <i>Amazon Personalize (AWS), Personalizer (Azure) e Recommendations AI (GCP)</i> | Usados na criação de sistemas de recomendação personalizados em tempo real.                                                                                                                                                     |
| <i>Amazon Kendra (AWS)</i>                                                       | Oferece uma busca inteligente baseada em linguagem natural, e é capaz de responder a perguntas de forma precisa dentro de bases de conhecimento corporativas.                                                                   |
| <i>Face API (Azure)</i>                                                          | Oferece reconhecimento facial avançado, com análise de atributos e emoções, sendo específico em relação ao AI Vision.                                                                                                           |
| <i>Video Intelligence API (GCP)</i>                                              | Permite a análise detalhada de conteúdos de vídeo, incluindo detecção de objetos, reconhecimento de texto e segmentação de cenas.                                                                                               |
| <i>ChatGPT-4o (OpenAI)</i>                                                       | Oferece geração e compreensão de linguagem natural, suporte a múltiplos idiomas, interpretação de imagens e sons, recursos para criação de assistentes virtuais, redação de textos, respostas contextuais e integração via API. |
| <i>Gemini (Google DeepMind)</i>                                                  | Oferece geração e compreensão de linguagem natural, análise de imagens, áudio e código; utilizado em tarefas como assistência virtual, produção de conteúdo, análise de dados e automação de processos.                         |

Fonte: Elaborado pelo autor

Conforme apresentado na Tabela 2, as *APIs* foram agrupadas por similaridade no que se refere as suas principais funcionalidades. É possível verificar a existência majoritária de serviços compatíveis entre as plataformas apesar da presença de serviços específicos e singulares por plataforma como o *Amazon Kendra*, voltado para buscas inteligentes baseadas em *NLP*.

Percebe-se também que as *APIs* da *AWS* são voltadas para atividades genéricas com foco em

automação de processos, análise de dados e aprimoramento da experiência do usuário. Seus serviços permitem incorporar funcionalidades de inteligência artificial sem a necessidade de construir modelos complexos do zero ou possuir conhecimento aprofundado em *IA*. Entre os destaques estão *APIs* como o *Amazon Comprehend*, *Rekognition*, *Polly*, *Transcribe*, *Translate* e *Textract*, que oferecem recursos prontos para uso em aplicações de análise textual, visão computacional, conversão de fala e texto, tradução e extração de dados estruturados.

As *APIs* da *Azure*, por sua vez, se destaca pela oferta de soluções voltadas à automação, análise avançada e melhoria da experiência do usuário. A integração nativa com o ecossistema *Microsoft* (como o *Azure Machine Learning*, *Power BI* e *Microsoft 365*) torna essas soluções acessíveis e eficientes para empresas que já utilizam essas plataformas. *APIs* como o *Text Analytics*, *AI Vision*, *Speech*, *Translator* e *Form Recognizer* proporcionam funcionalidades robustas para interpretação de texto, visão computacional, reconhecimento de fala e tradução automática, alinhadas à demanda de negócios por eficiência operacional e inteligência aplicada. Já o *Face API* oferece reconhecimento avançado de análise de atributos e emoções.

As *APIs* do *GCP* (*Google Cloud Platform*) oferecem alto desempenho, escalabilidade e integração nativa com outras ferramentas do ecossistema *Google*. Seus serviços abrangem funcionalidades avançadas de processamento de linguagem natural, visão computacional, tradução automática e recomendação personalizada. Entre os principais recursos estão o *Cloud Natural Language*, *Vision AI*, *Speech-to-Text*, *Text-to-Speech*, *Translation*, *Document AI* e *Recommendations AI*. Além disso, o *GCP* se destaca por soluções especializadas como a *Video Intelligence API*, que permite análise detalhada de vídeos, sendo utilizada em mídia, entretenimento e segurança. A integração com ferramentas como *BigQuery* e *Firebase* torna o *GCP* uma opção poderosa para empresas que já operam com produtos *Google* e desejam construir aplicações inteligentes com base em dados.

Por outro lado, os modelos *Gemini* e *ChatGPT*, apresentam soluções avançadas de *IA* com um conjunto de *APIs* integrado, ambos são projetados para atender desde usuários com conhecimento técnico básico até desenvolvedores com experiência. Esses modelos oferecem interface multimodais e flexíveis capazes de lidar com tarefas de geração e compreensão de *NLP*, análise de imagens, vídeos, áudios e documentos estruturados. A integração dessas *APIs* em soluções corporativas ou aplicações pessoais permite a automação de processos, o suporte ao atendimento virtual, a criação de conteúdos personalizados e a realização de análises complexas. Além disso, o suporte a múltiplos idiomas e a capacidade de adaptação a diferentes domínios tornam essas soluções altamente acessíveis e versáteis em diversos contextos.

O modelo *Gemini* conta com três versões distintas. Para facilitar a compreensão, a Tabela 3 apresenta uma visão geral de cada uma delas:

Tabela 3 - Versões do modelo Gemini

| <b>Modelos Gemini (Google DeepMind)</b> | <b>Descrição</b>                                                                                                                                                                                                                                         |
|-----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>2.5 Pro</i>                          | Aceita entradas de áudio, imagens, vídeo, texto e PDF, e saída apenas de texto. Oferece resolução de problemas complexos, análise de grandes bancos de dados, o que o torna ideal para programação avançada, raciocínio lógico e compreensão multimodal. |
| <i>2.5 Flash</i>                        | Aceita entradas multimodais, com saída apenas de texto. Possui processamento rápido e eficiente, adequado para tarefas de baixa latência e alto volume com alguma necessidade de raciocínio.                                                             |
| <i>2.5 Flash-Lite</i>                   | É uma variante econômica, com suporte às mesmas entradas, porém otimizada para cenários em tempo real, mantendo alta velocidade e custo reduzido.                                                                                                        |

Fonte: Elaborada pelo autor

Ambas as versões do modelo *Gemini* oferecem suporte a entrada de áudio, textos, imagens e *PDFs*, com saída apenas de texto. Ainda assim, cada versão apresenta especificidades que a qualificam para usos diferenciados conforme a complexidade e o desempenho exigido. A versão *2.5 Pro* é a versão avançada, sendo indicada para tarefas complexas que exigem raciocínio sofisticado, como programação, análise de grandes volumes de dados e compreensão multimodal aprofundada. Já as versões *2.5 Flash* e *2.5 Flash-Lite* foram otimizadas para cenários que demandam respostas rápidas, com baixo nível de latência e alto volume de requisições, sendo ideais para aplicações em tempo real. A versão *Flash-Lite*, em especial, apresenta uma proposta econômica, mantendo boa performance para tarefas com baixa exigência.

Além das diferenças funcionais, existe também uma hierarquia de custo entre os modelos, sendo a *2.5 Pro* a opção robusta e cara, enquanto a *2.5 Flash-Lite* representa uma alternativa acessível. Dessa forma, a escolha entre as versões depende diretamente das necessidades específicas da aplicação, considerando fatores como complexidade das tarefas, volume de requisições e restrições orçamentárias.

#### 4.1.1 APIs Comuns nas Plataformas e Serviços

As plataformas e serviços escolhidos para análise neste trabalho (*AWS*, *Azure*, *Google (GCP, Gemini)* e *OpenAI (ChatGPT)*) oferecem APIs com funcionalidades em comum. Dentre elas as APIs de visão computacional, processamento de linguagem natural (*NLP*) e reconhecimento de fala. A Tabela 4 apresenta as APIs que se repetem entre as três plataformas e nos dois modelos de *IA*, evidenciando os serviços populares e amplamente utilizados nas implementações de *Machine Learning*.

Tabela 4 - APIs comuns de Machine Learning em AWS, Azure e GCP

| <b>Categoria</b>                         | <b>AWS</b>         | <b>Azure</b>           | <b>GCP</b>                    | <b>Gemini</b> | <b>ChatGPT</b> |
|------------------------------------------|--------------------|------------------------|-------------------------------|---------------|----------------|
| Visão Computacional                      | <i>Rekognition</i> | <i>Azure AI Vision</i> | <i>Cloud Vision AI</i>        | ✓             | ✓              |
| Processamento de linguagem Natural - NLP | <i>Comprehend</i>  | <i>Text Analytics</i>  | <i>Cloud Natural Language</i> | ✓             | ✓              |
| Reconhecimento de Fala                   | <i>Transcribe</i>  | <i>Speech</i>          | <i>Speech-to-Text</i>         | ✓             | ✓              |
| Síntese de Fala                          | <i>Polly</i>       | <i>Speech</i>          | <i>Text-to-Speech</i>         | ✓             | ✓              |
| Tradução Automática                      | <i>Translate</i>   | <i>Translator Text</i> | <i>Cloud Translation</i>      | ✓             | ✓              |
| Reconhecimento de OCR                    | <i>Textract</i>    | <i>Form Recognizer</i> | <i>Document AI</i>            | ✓             | ✓              |

Fonte: Elaborado pelo autor

Como apresentado na Tabela 4 fica claro uma convergência entre os serviços oferecidos pelas três plataformas de computação em nuvem e os modelos de IA oferecidos pelo *Google DeepMind* e *OpenAI* avaliadas nesse trabalho. Isso demonstra que, apesar das diferenças de implementação e integração, há um núcleo de funcionalidades que são consideradas indispensáveis para aplicações de *Machine Learning*, como visão computacional e processamento de linguagem natural. Essa semelhança também pode indicar um grau de consolidação do mercado de *MLaaS*, no qual os provedores buscam atender às principais demandas dos usuários de forma padronizada. Nota-se que, essa padronização facilita a portabilidade de soluções entre diferentes provedores, ao mesmo tempo em que desafia as plataformas a se diferenciarem não apenas pelo número de serviços, mas pela qualidade, integração com outros serviços nativos e experiência do desenvolvedor.

No caso do *Gemini* e do *ChatGPT*, os símbolos de verificação utilizados na Tabela 4 referem-se à capacidade desses modelos de executarem as respectivas tarefas, ainda que não disponibilizem APIs especializadas e independentes como ocorre nas plataformas de computação em nuvem *AWS*, *Azure* e *GCP*. Eles indicam também que essas funcionalidades podem ser acessadas por meio da interface dos modelos.

#### 4.1.2 APIs para Visão Computacional

Com base nas APIs comuns listadas na Tabela 4, foram selecionadas as soluções de visão computacional *Amazon Rekognition* (*AWS*), *Azure Computer Vision* (*Microsoft*) e *Cloud Vision API* (*Google*), além dos modelos multimodais *Gemini* (*Google DeepMind*) e *ChatGPT-4o* (*OpenAI*) para um melhor entendimento teórico com base na documentação. A inclusão do *Gemini* e *ChatGPT-4o* se justifica pela sua capacidade de interpretar imagens de forma contextual, mesmo

não sendo APIs especializadas. A seleção considerou a ampla adoção no mercado e a oferta de funcionalidades comparáveis, possibilitando uma análise técnica entre as abordagens utilizadas nas plataformas de computação em nuvem e os modelos multimodais.

- Amazon Rekognition
  - O Amazon Rekognition é um serviço da *AWS* o qual visa a integração de análises de imagem e vídeo em aplicativos, sem a necessidade de experiência prévia em *machine learning*. Utilizando aprendizado profundo ele é capaz de identificar objetos, cenas, rostos, celebridades e aplicação de filtros para incluir ou excluir rótulos. A documentação apresenta esse serviço como escalável, capaz de processar diversas imagens, tornando-o adequado para grandes volumes de dados visuais, também oferece recursos para avaliar propriedades como nitidez, brilho, contraste, qualidade e conteúdo das imagens. Ele pode ser integrado com outras soluções da *AWS*, como *Amazon S3* e *AWS Lambda*, permitindo assim o processamento direto de imagens armazenadas na nuvem, com o modelo de cobrança baseado no uso, sem taxas mínimas. Ele conta com opções para redução de custos em larga escala, além de um nível gratuito inicial. Esse serviço possibilita a personalização por intermédio de adaptadores treinados com imagens de exemplo, o que permite ajustar a precisão para diferentes necessidades, mesmo sem conhecimento técnico em inteligência artificial (Amazon Web Services, 2024).
- Microsoft Azure Computer Vision
  - O Azure Computer Vision é um serviço da *Microsoft* que permite a integração de análises de imagens e vídeos em aplicativos, sem a necessidade de experiência prévia em *machine learning*. O serviço utiliza técnicas de aprendizado profundo para identificar objetos, cenas, textos impressos e manuscritos, rostos e marcas, além de gerar descrições automáticas do conteúdo visual. De acordo com a documentação, apresenta escalabilidade, o que possibilita o processamento de grandes volumes de dados visuais, tanto em tempo real quanto em lote. Essa característica permite sua aplicação em contextos como segurança, acessibilidade digital, automação de processos e gerenciamento de conteúdo. O *Azure Computer Vision* também oferece funcionalidades para a avaliação de propriedades visuais das imagens, como nitidez, contraste, cor dominante e qualidade, apoiando a análise técnica e estética dos arquivos (Microsoft Azure, 2024).
  - Esse serviço, integrado a outras soluções do ecossistema *Microsoft*, como *Azure Blob Storage*, *Azure Functions* e *Power Automate* pode permitir o desenvolvimento de *pipelines* de processamento automatizado, nos quais imagens podem ser

analisadas diretamente a partir do armazenamento em nuvem. O modelo de cobrança é baseada no uso, sem taxas mínimas e com a oferta de uma camada gratuita voltada à experimentação e prototipagem. Além das funcionalidades pré-configuradas, a plataforma também dispõe do *Custom Vision*, serviço complementar que permite o treinamento de modelos personalizados com imagens fornecidas pelo usuário. Essa abordagem possibilita a adaptação a detecção de elementos visuais a diferentes contextos e necessidades, inclusive por usuários sem formação técnica em inteligência artificial (Microsoft Azure, 2024).

- Google Cloud Vision API
  - A Google Cloud Vision API é um serviço da *Google Cloud Platform (GCP)* que permite a integração de análises de imagem em aplicações sem exigir experiência prévia em *machine learning*. Utilizando modelos de aprendizado profundo, o serviço oferece detecção de objetos, cenas, textos (impressos e manuscritos), rostos, logotipos e marcos geográficos. A documentação destaca sua escalabilidade e integração nativa com outras soluções do *GCP*, como o *Cloud Storage* e o *Cloud Functions*, viabilizando fluxos automatizados de análise visual diretamente a partir do armazenamento em nuvem. A plataforma também dispõe de recursos para avaliar propriedades visuais, como cor dominante, brilho e nitidez, além da geração automática de rótulos descritivos e classificação semântica do conteúdo visual (Google, 2024).
  - A *Cloud Vision API* adota um modelo de cobrança baseado no uso, com tarifação separada para cada funcionalidade como *OCR*, detecção de rostos e moderação de conteúdo, o que pode requerer um planejamento complexo. Também conta com uma camada gratuita com limite mensal, voltada à prototipagem e o desenvolvimento inicial. Para demandas específicas, a plataforma oferece o serviço complementar *AutoML Vision*, que permite o treinamento de modelos personalizados com dados do usuário, embora essa funcionalidade requeira familiaridade técnica aprofundada e conjuntos de dados estruturados. Dessa forma, o serviço pode ser aplicado a diferentes contextos que envolvam visão computacional em nuvem, como acessibilidade digital, comércio eletrônico, segurança e gerenciamento de conteúdo visual (Google, 2024).
- Gemini
  - O Gemini é uma família de modelos de inteligência artificial desenvolvida pela *Google DeepMind*, voltada para o processamento de informações multimodais, capaz de interpretar e gerar conteúdos em diferentes formatos, como texto, imagem, áudio e código. O modelo conta com interfaces como, o *Google AI Studio* e o assistente *Gemini*, com integração ao ecossistema da *Google Cloud Platform* para aplicações empresariais. Ele foi projetado para realizar tarefas de linguagem natural e raciocínio

avançado, além de oferecer suporte a entradas visuais e auditivas em suas versões recentes (Google, 2025).

- Entre suas capacidades estão o reconhecimento de elementos visuais em imagens, como objetos, cenas, texto e gráficos, a interpretação de imagens fornecidas pelo usuário e a geração de respostas contextualizadas em linguagem natural com base na análise visual. Segundo a documentação, o *Gemini* se destaca em contextos interativos, como assistentes virtuais, suporte ao usuário e análise de dados visuais, com raciocínio contextual. Seu acesso via prompts multimodais dispensa a configuração de múltiplas *APIs* e permitindo integração direta via interfaces conversacionais ou programáticas, o que o torna uma solução flexível e eficaz para tarefas visuais integradas (Google, 2025).
- ChatGPT
  - O *ChatGPT* é uma família de modelos de linguagem desenvolvida pela *OpenAI*, com sua versão recente denominada *ChatGPT-4o*. Conforme a documentação, o modelo 4o oferece suporte a entrada e saída de texto, imagens, áudios e vídeo. Sua capacidade multimodal permite compreender e gerar conteúdos em diferentes formatos de forma natural e fluida, sem a necessidade de converter dados ou utilizar *APIs* separadas (OpenAI, 2025).
  - Ele disponibiliza uma interface unificada acessada por meio de *API REST*, viabilizando o desenvolvimento de aplicações que combinam texto, imagens e áudio em um único fluxo conversacional. Dentre suas principais funcionalidades estão: a descrição de imagens, interpretação de gráficos, reconhecimento de fala, síntese de voz e resposta contextual avançada, o que pode torná-lo um recurso com ampla aplicabilidade em assistentes virtuais, sistemas de suporte ao cliente, produção de conteúdo e acessibilidade digital. Seu uso é simplificado, com autenticação via chave de *API* e ampla documentação para integração com diferentes linguagens e frameworks (OpenAI, 2025).

4.2 COMPARAÇÃO ENTRE OS SERVIÇOS E APIS

Com base nas *APIs* comuns apresentadas na Tabela 4 e nos modelos de *IA*, foram elencadas as *APIs* de Visão Computacional de cada plataforma para uma análise aprofundada e técnica:

Tabela 5 - Comparação dos serviços de visão computacional para *ML* nas plataformas *AWS*, *Azure* e *GCP* e os modelos de *IA Gemini* e *ChatGPT*, com base na documentação oficial.

| <b>Critério</b>              | <i>Amazon Rekognition</i> | <i>Azure Vision</i>        | <i>Cloud Vision API</i>   | <i>Gemini</i>               | <i>ChatGPT</i>    |
|------------------------------|---------------------------|----------------------------|---------------------------|-----------------------------|-------------------|
| <b>Principais Funcionali</b> | (i) Detecção de objetos,  | (i) <i>OCR</i> (reconhecim | (i) Rotulagem de imagens; | (i) Entrada e interpretação | (i) Interpretação |

|                                     |                                                                                                                                                                                                           |                                                                                                                                                    |                                                                                                                      |                                                                                                        |                                                                                                                                                                                 |
|-------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>dades</b>                        | rostos, cenas, textos ( <i>OCR</i> ); (ii) Moderação de conteúdo; (iii) Comparação facial.                                                                                                                | ento óptico de caracteres); (ii) Análise de imagens; (iii) Identificação facial.                                                                   | (ii) Detecção de rostos, <i>OCR</i> ; (iii) Detecção de logotipos e propriedades.                                    | de imagens; (ii) Descrição visual; (iii) Localização e objetos, <i>OCR</i> .                           | o de imagens, gráficos e capturas de tela; (ii) Geração de descrições, respostas a perguntas sobre imagens, leitura de textos; (iii) Interpretação de dados visuais.            |
| <b>Integração com o Ecossistema</b> | <i>AWS S3, Lambda, IAM, Amazon SageMaker</i> via <i>APIs</i> e <i>SDKs</i> documentados.                                                                                                                  | <i>Azure Blob Storage, Cognitive Services, Logic Apps, Power Automate</i> e <i>Azure Functions</i> . <i>APIs REST</i> e <i>SDKs</i> oficiais.      | <i>Google Cloud Storage, AutoML Vision</i> e <i>Vertex AI</i> . <i>APIs REST</i> e <i>client libraries</i> oficiais. | <i>Google AI Studio</i> , suporte a <i>REST API</i> , biblioteca para <i>Python</i> e <i>Node.js</i> . | Disponível via <i>OpenAI API</i> e <i>ChatGPT</i> (com plugins e visão ativada); integração via <i>SDKs</i> ( <i>Python</i> , <i>Node.js</i> etc.) e plataforma <i>OpenAI</i> . |
| <b>Facilidade de Uso</b>            | Documentação detalhada com tutoriais e exemplos de chamadas <i>REST</i> ; em decorrência disso a <i>API</i> permite troca de arquivos via <i>JSON</i> , facilitando a integração com diversas linguagens. | Documentação detalhada com tutoriais e exemplos de chamadas <i>REST</i> ; oferece <i>SDKs</i> para múltiplas linguagens, facilitando a integração. | Documentação oficial com tutoriais, exemplos de código e <i>SDKs</i> atualizados.                                    | Documentação oficial com exemplos de uso, interface de teste no <i>AI Studio</i> .                     | Documentação completa e exemplos práticos, com interface simples, além de <i>API REST</i> .                                                                                     |
| <b>Desempenho e Funcionalidade</b>  | Suporte a chamadas síncronas e assíncronas, escalabilidade automática conforme carga.                                                                                                                     | Suporte a chamadas síncronas, processamento em lote e escalabilidade gerenciada.                                                                   | Suporte a chamadas síncronas e assíncronas; escalabilidade automática e suporte a processamento em lote.             | Suporte a chamadas assíncronas e <i>backend</i> escalável na infraestrutura <i>Google</i> .            | Processamento rápido de imagens com respostas multimodais; escalável via <i>API</i> da <i>OpenAI</i> ; desempenho                                                               |

|                                                                                     |                                                                                                                                                                           |                                                                                                                                               |                                                                                                                                                   |                                                                                                                                                                                          |                                                                                                                                                                            |
|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                     |                                                                                                                                                                           |                                                                                                                                               |                                                                                                                                                   |                                                                                                                                                                                          | otimizado no modelo <i>GPT-4o</i> .                                                                                                                                        |
| <b>Modelo de Precificação</b>                                                       | Cobrança por tipo de análise e volume mensal, com plano gratuito de até 5.000 imagens/mês por 12 meses.                                                                   | Cobrança por análise e uso mensal, com 5.000 transações/mês gratuitas na <i>API</i> de <i>OCR</i> dentro do <i>Free Tier</i> .                | Cobrança por análise e volume mensal, com 1.000 unidades gratuitas por mês sem limite de tempo.                                                   | Cobrança por milhão de tokens, com modelos variando em custo e desempenho; há uso gratuito diário.                                                                                       | Cobrança por milhão de <i>tokens</i> , com preços variando conforme o conteúdo visual; inclui camada gratuita diária.                                                      |
| <b>SLA (Service Level Agreement ou Acordo de Nivel de Serviço) e Confiabilidade</b> | <i>SLA</i> oficial de 99,9%, com redundância regional.                                                                                                                    | <i>SLA</i> de 99,9%, disponível em múltiplas regiões com alta disponibilidade.                                                                | <i>SLA</i> de 99,9%, infraestrutura global com replicação e redundância.                                                                          | Não informado na documentação.                                                                                                                                                           | Não informado na documentação.                                                                                                                                             |
| <b>Segurança e Conformidade</b>                                                     | Suporte a criptografia em trânsito e repouso, controle de acesso via <i>IAM</i> .<br><b>Conformidades:</b> <i>HIPAA</i> , <i>GDPR</i> , <i>SOC 2</i> , <i>ISO 27001</i> . | Criptografia, <i>RBAC</i> , <i>logs</i> de auditoria.<br><b>Conformidades:</b> <i>GDPR</i> , <i>HIPAA</i> , <i>ISO 27001</i> , <i>SOC 2</i> . | Controle via <i>IAM</i> , criptografia padrão, auditoria.<br><b>Conformidades:</b> <i>GDPR</i> , <i>HIPAA</i> , <i>ISO 27001</i> , <i>SOC 2</i> . | Autenticação via <i>API Key</i> , integrada aos mecanismos de segurança do <i>Google Cloud</i> .<br><b>Conformidades:</b> <i>GDPR</i> , <i>HIPAA</i> , <i>ISO 27001</i> , <i>SOC 2</i> . | Criptografia em trânsito, autenticação por <i>API Key</i> .<br><b>Conformidades:</b> <i>GDPR</i> , <i>HIPAA</i> (via <i>Azure BAA</i> ), <i>SOC 2</i> , <i>ISO 27001</i> . |
| <b>Suporte e Comunidade</b>                                                         | Suporte oficial <i>AWS</i> , fóruns da <i>AWS</i> e ampla comunidade.                                                                                                     | Suporte <i>Microsoft</i> e fóruns oficiais, documentação ativa.                                                                               | Suporte <i>Google Cloud</i> e comunidade ativa em fóruns e <i>Stack Overflow</i> .                                                                | Suporte oficial, <i>Google AI Studio</i> e canais da comunidade <i>Google</i> .                                                                                                          | Suporte oficial, fóruns de desenvolvedores, suporte via plataforma <i>OpenAI</i> , comunidade ativa em <i>GitHub</i> e <i>Stack Overflow</i> .                             |
| <b>Atualizações e Ciclos de Vida</b>                                                | Atualizações pontuais (média de uma vez ao ano); <i>changelog</i> (registro de                                                                                            | Atualizações frequentes (a cada 3–6 meses); <i>changelog</i> ativo e                                                                          | Atualizações periódicas (4–8 meses); <i>changelog</i> público disponível.                                                                         | Versões <i>Gemini 1.5</i> e <i>2.5</i> documentadas com melhorias constantes.                                                                                                            | Atualizações frequentes; com <i>changelog</i> disponível na plataforma.                                                                                                    |

|                                     | mudança) disponível.                                       | roadmap publicado.                                                         |                                                                                |                                                                       |                                                                                                                                 |
|-------------------------------------|------------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------------------------------------|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| <b>Casos de Uso na Documentação</b> | Exemplos em segurança, e-commerce e moderação de conteúdo. | Exemplos em <i>OCR</i> de documentos, acessibilidade e análise de imagens. | Exemplos em classificação de imagens, extração de texto e detecção de objetos. | <i>OCR</i> , perguntas sobre imagens, descrição de conteúdos visuais. | Interpretação de gráficos, leitura de textos, análise de imagens de interface, suporte educacional, geração de conteúdo visual. |

Fonte: Elaborado pelo autor

#### 4.2.1 Principais Funcionalidades

As plataformas analisadas neste trabalho oferecem funcionalidades similares no que se refere a tarefas de visão computacional, como detecção de objetos, *OCR*, reconhecimento facial, moderação de conteúdo e rotulagem de imagens, apresentados na Tabela 5. No entanto, há diferenças na forma como essas funcionalidades são organizadas e disponibilizadas. O *Amazon Rekognition*, por exemplo, apresenta uma estrutura modular, com cada funcionalidade tratada de forma separada, o que permite controle técnico elevado, mas exige uma configuração detalhada. O *Azure Computer Vision*, por sua vez, adota uma abordagem unificada, integrando múltiplas capacidades em chamadas simples, o que favorece a usabilidade. Já *Google Cloud Vision API* destaca-se pela ênfase em enriquecimento semântico, com recursos como detecção de landmarks, logotipos e propriedades visuais da imagem.

Os modelos generativos, como *Gemini* e *ChatGPT-4o*, que operam de forma abstrata, permitindo interpretar imagens por meio de linguagem natural, o que os torna úteis para aplicações baseadas em interação, geração de descrições e acessibilidade, ainda que com controle técnico limitado sobre tarefas específicas.

#### 4.2.2 Integração com o Ecossistema

A integração com o ecossistema da nuvem tem papel fundamental na adoção e na eficiência das *APIs* de visão computacional, como apresentado na Tabela 5, o *Amazon Rekognition* integrado a outros serviços da *AWS*, como *S3*, *Lambda* e *SageMaker*, permite a automação de fluxos complexos e arquiteturas escaláveis baseadas em eventos; o *Azure Computer Vision* oferece integração direta com ferramentas como *Power BI*, *Power Apps* e *Logic Apps*, o que facilita o uso em contextos empresariais e soluções *low-code/no-code*; já o *Google Cloud Vision API* integra-se com serviços como *BigQuery*, *AutoML Vision* e *Vertex AI*, favorecendo assim fluxos orientados a dados e inteligência artificial. No caso dos modelos generativos, a integração é voltada a aplicações

finais como o *Gemini* que se conecta ao ecossistema do *Google AI Studio* e serviços da *Google Cloud*, enquanto o *ChatGPT-4o* que se destaca pela acessibilidade via API REST e interface direta no *ChatGPT*, embora com a integração nativa reduzida aos serviços de nuvem da *AWS*, *Azure* e *GCP*.

#### 4.2.3 Facilidade de Uso

A facilidade de uso das APIs de visão computacional varia conforme a linguagem suportada, interface e nível de abstração técnica (todas as linguagens suportadas por cada provedor estão presentes na Tabela 6).

O *Amazon Rekognition*, oferece uma ampla compatibilidade com linguagens e integração profunda com ecossistema *AWS*. Apesar da documentação apresentar cobertura abrangente, a curva de aprendizado pode ser acentuada para iniciantes, uma vez que a utilização eficiente demanda conhecimento prévio sobre gerenciamento de identidade e acesso (*IAM*), políticas de segurança e arquitetura da nuvem *AWS*. Isso torna o *Amazon Rekognition* recomendado para desenvolvedores com experiência na *AWS* que beneficiam da integração direta com serviços como *S3*, *Lambda* e *SageMaker*.

Em contraste, o *Azure Computer Vision* se destaca pela usabilidade, focando na acessibilidade, com um portal visual intuitivo (*Azure Portal*) e um ambiente de teste (*Playground*) o qual permite a experimentação interativa com os serviços antes do código em si. A documentação é organizada, e conta com guias passo a passo. Além disso, conta com integração com ferramentas *low-code/no-code* do ecossistema *Microsoft* como, *Power Apps*, *Logics* e *Power BI*. Essa abordagem visual e multilinguagem torna o *Azure Computer Vision* uma solução equilibrada entre potência e simplicidade, adequada tanto para desenvolvedores quanto para usuários com pouca especialização.

O *Google Cloud Vision API* já oferece uma cobertura de linguagens e integração com *AutoML*, permitindo assim a personalização do código, mas para isso requer um domínio de autenticação e configuração do *GCP*, ou seja, o uso por desenvolvedores iniciantes que ainda não utilizaram a plataforma pode ser complicado. Apesar da interface visual ser acessível e a documentação conter exemplos práticos, a fase inicial do projeto, ativação e gerenciamentos de credenciais do serviço pode ser uma barreira. Mas em contrapartida, para os desenvolvedores familiarizados com o ecossistema da *Google Cloud*, o *Vision API* destaca-se como uma solução escalável e eficiente, especialmente se integrada com outras ferramentas.

Já as soluções baseadas em modelos de IA, como *Gemini* e *ChatGPT-4o*, focam na simplicidade e acessibilidade no uso, contando com interfaces projetadas para facilitar a interação por meio da linguagem natural. O *Gemini*, por meio do *Google AI Studio*, oferece uma experiência interativa com suporte a entrada multimodal, permitindo prototipação rápida e testes diretos. No

entanto, para a sua utilização requer uma compreensão básica sobre autenticação via chave de *API* e estrutura de prompts, o que pode representar uma barreira para usuários iniciantes. Já o *ChatGPT-4o*, proporciona uma interface direta no *ChatGPT* e integração simplificada via *API REST*, mesmo para quem não possui conhecimento técnico aprofundado. Apesar disso, ambos os modelos operam com alto nível de abstração e não oferece controle detalhado de tarefas visuais específicas, como ocorre nas *APIs* oferecidas pelas plataformas de nuvem *Amazon Web Service*, *Azure* e *Google Cloud Platform*. Ainda assim, destacam-se em aplicações que demandam agilidade, automação textual, acessibilidade e experiências conversacionais com entrada visual.

A seguir, a Tabela 6 apresenta as linguagens suportadas e interfaces de acesso por cada plataforma/modelo:

Tabela 6– Linguagens de programação e interfaces suportadas pelas APIs

| Plataforma                      | Linguagem de Programação                                                  | Interfaces de Acesso             |
|---------------------------------|---------------------------------------------------------------------------|----------------------------------|
| <b>Amazon Rekognition (AWS)</b> | Python (boto3), Java, JavaScript (Node.js), Go, C++, Ruby, .NET (C#), PHP | SDKs, AWSCLI, REST API           |
| <b>Azure Computer Vision</b>    | Python, C#, Java, JavaScript, TypeScript, Go, Node.js                     | SDKs, Azure Portal, REST API     |
| <b>Google Cloud Vision API</b>  | Python, Java, Go, Node.js, Ruby, C#, PHP                                  | SDKs, GCP Console, REST API      |
| <b>Gemini (Google DeepMind)</b> | Python, Node.js                                                           | Google AI Studio, REST (OpenAPI) |
| <b>ChatGPT (OpenAI)</b>         | Python, Node.js, cURL                                                     | Google AI Studio, REST (OpenAPI) |

Fonte: Elaborado pelo autor

Com base na Tabela 6 observa-se a presença de linguagens comuns entre as *APIs* das plataformas em nuvem e os modelos de *IA* como *Python* e *Node.js*; as linguagens *Java*, *Go* e *C#* estão presentes apenas nas três *APIs* de visão computacional (*AWS*, *Azure* e *GCP*); já as linguagens *PHP* e *Ruby* estão presentes somente nas *APIs*, *Amazon Rekognition* e *Vision API*, o *JavaScript* está presente nas *APIs*, *Amazon Rekognition* e *Azure Computer Vision*. Mesmo que haja linguagens comuns entre as *APIs* das plataformas e os modelos de *IA*, há também algumas linguagens presentes apenas em uma plataforma/modelo, como *C++* e *.NET* que estão presentes apenas no *Amazon Rekognition* e *cURL* presente no *ChatGPT-4o*.

Com base nisso, é possível averiguar que o *Amazon Rekognition* da *AWS* destaca-se pela sua ampla variedade de linguagens suportadas, seguido do *Azure Computer Vision* e *Vision API*. Os Modelos de *IA*, por sua vez, apresentam um suporte enxuto, voltado a linguagens populares e de fácil integração com *APIs REST*, o que está alinhado com seu foco em acessibilidade e prototipação rápida.

4.2.4 Desempenho e Escalabilidade

Com base na tabela 5, em relação ao desempenho e à escalabilidade, observa-se que as APIs de visão computacional da *Amazon Rekognition*, *Azure Computer Vision* e *Google Cloud Vision API*, foram projetadas para operar em ambientes com alta demanda, oferecendo suporte nativo a processamento em lote, chamadas assíncronas e escalabilidade automática conforme o volume de requisições. Essas soluções aproveitam a infraestrutura das grandes nuvens públicas, garantindo desempenho consistente mesmo em cargas elevadas, com baixa latência e alta tolerância a falhas. Os modelos generativos, como *Gemini* e *ChatGPT-4o*, por sua vez, são otimizados para interações em tempo real e prototipação rápida, com desempenho em tarefas multimodais e capacidade contextual. No entanto, eles não são a escolha adequada para processamento massivo e contínuo de grandes volumes de imagens, pois priorizam a compreensão semântica e a resposta contextualizada, o que pode gerar um elevado custo computacional e previsibilidade de performance limitada em escala industrial. Assim, enquanto as plataformas *AWS*, *Azure* e *GCP* são indicadas para cenários de uso intensivo e integração em pipelines robustos, os modelos generativos destacam-se pela flexibilidade e resposta inteligente em interações específicas.

4.2.5 Modelo de Precificação

No que se refere à precificação, os serviços de *AWS*, *Azure* e *GCP* seguem o modelo tradicional de cobrança por transação ou tipo de análise, com planos gratuitos limitados (tempo ou volume). Já o *Gemini* e o *ChatGPT* utilizam um modelo baseado em tokens, com diferentes níveis de custos e desempenho conforme o modelo utilizado. Segue a Tabela 7 com uma visão detalhada dos preços atuais de cada Plataforma e Modelo:

Tabela 7– Linguagens de programação e interfaces suportadas pelas APIs

| Plataformas                    | Descrição                                                                                                                                                                                                                                                             |
|--------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Amazon Rekognition</i>      | <ul style="list-style-type: none"><li>• Gratuito: 5.000 imagens/mês (por 12 meses).</li><li>• Detecção facial/rótulos: US\$ 1,00 por 1.000 imagens.</li><li>• Vídeo: US\$ 0,10 por minuto.</li></ul> <p>Exemplo: 10.000 imagens com detecção facial = US\$ 10,00.</p> |
| <i>Azure Computer Vision</i>   | <ul style="list-style-type: none"><li>• Gratuito: 5.000 transações OCR/mês.</li><li>• OCR: US\$ 1,00 por 1.000.</li><li>• Análise de imagem: US\$1,50 por 1.000.</li></ul> <p>Exemplo: 20.000 transações (OCR + análise) = US\$ 25,00.</p>                            |
| <i>Google Cloud Vision API</i> | <ul style="list-style-type: none"><li>• Gratuito: 1.000 unidades/mês.</li><li>• OCR: US\$ 1,50 por 1.000.</li></ul>                                                                                                                                                   |

|                       |                                                                                                                                                                                                                                                                                                                                                                                       |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                       | <ul style="list-style-type: none"> <li>• Rótulos: US\$ 1,50 por 1.000.</li> <li>• Logotipos/landmarks: US\$ 2,00 por 1.000.</li> </ul> <p>Exemplo: 7.000 unidades = US\$ 10,50.</p>                                                                                                                                                                                                   |
| <b><i>Gemini</i></b>  | <ul style="list-style-type: none"> <li>• Gratuito diário (limite variável).</li> <li>• Flash: US\$ 0,35 (entrada), US\$ 1,05 (saída)/milhão de tokens.</li> <li>• Pro: \$3,50 (entrada), \$10,50 (saída)/milhão de tokens.</li> <li>• <i>Flash-Lite</i>: US\$ 0,075 (entrada), US\$ 0,225 (saída)/milhão de tokens.</li> </ul> <p>Exemplo: 10 imagens descritas = US\$ 0,20–0,30.</p> |
| <b><i>ChatGPT</i></b> | <ul style="list-style-type: none"> <li>• Gratuito via <i>ChatGPT</i>.</li> <li>• API: US\$ 5,00 (entrada), US\$ 5,00 (saída)/milhão de tokens.</li> <li>• Imagem: US\$ 0,00765–0,02 por imagem.</li> </ul> <p>Exemplo: 100 imagens + 10k tokens = US\$ 3–4.</p>                                                                                                                       |

Fonte: Elaborado pelo autor

Conforme apresentado na Tabela 7, as plataformas de computação em nuvem *Amazon Rekognition*, *Azure Computer Vision* e *Google Cloud Vision API*, apresentam modelos de cobrança previsíveis, baseados em volume de requisições e tipo de análise executada, sendo adequados para aplicações com controle operacional definido e uso contínuo. Em contrapartida, os modelos baseados em tokens, como *Gemini* e *ChatGPT-4o*, embora acessíveis para prototipação e testes pontuais, exigem atenção ao consumo acumulado de entrada e saída, especialmente em contextos multimodais ou com interações complexas. O custo-benefício, portanto, varia conforme o perfil do projeto: soluções baseadas em nuvem oferecem controle aprimorado e estabilidade financeira, enquanto os modelos de *IA* generativa destacam-se pela flexibilidade, rapidez de integração e experiência interativa.

#### 4.2.6 Service Level Agreement (SLA) / Acordo de Nível de Serviço

Como apresentado na Tabela 5, as plataformas *AWS*, *Azure* e *GCP* oferecem *SLAs* formais de 99,9%, o que indica um nível elevado de disponibilidade e estabilidade para aplicações com requisitos de continuidade operacional. Esses compromissos de serviço são sustentados por infraestrutura com suporte a escalabilidade e redundância. Já os modelos generativos, como *Gemini* e *ChatGPT-4o*, não possuem *SLA* explícito em suas versões públicas, o que os torna indicados para ambientes de prototipação e uso com baixa criticidade, onde a tolerância a falhas e instabilidades é elevada.

#### 4.2.7 Segurança e Conformidade

Com base na Tabela 5, pode-se notar que as soluções avaliadas adotam práticas consolidadas de segurança, como criptografia em trânsito e em repouso, autenticação por chave de API e controle de acesso. As APIs *Amazon Rekognition*, *Azure Computer Vision* e *Google Cloud Vision API* oferecem mecanismo avançados como, *IAM (Identity and Access Management)*, *RBAC (Role-Based Access Control)*, suporte a auditoria e políticas de segurança detalhadas. Por outro lado, os modelos generativos, por operarem com um nível adequado de segurança, apresentam flexibilidade de configuração limitada por parte do usuário, mas conta com uma conexão segura por *HTTPS*.

Acrescenta-se que tanto as plataformas em nuvem quanto os modelos generativos seguem padrões reconhecidos de conformidade aplicados na área de visão computacional, assegurando a proteção dos dados processados. Dentre esses, destaca-se a *HIPAA (Health Insurance Portability and Accountability Act)*, que regula o tratamento de dados sensíveis no setor de saúde, assegurando a confidencialidade e integridade das informações. O *GDPR (General Data Protection Regulation)* estabelece diretrizes para a coleta, armazenamento e uso de dados pessoais na União Europeia, exigindo transparência e consentimento do usuário. Já a norma *ISO 27001 (International Organization for Standardization)* define requisitos para sistemas de gestão de segurança da informação, com foco na prevenção de vazamentos e na mitigação de riscos. O *SOC 2 (Service Organization Control 2)* avalia critérios relacionados à segurança, disponibilidade, integridade de processamento, confidencialidade e privacidade dos serviços prestados.

#### 4.2.8 Suporte e Comunidade

Conforme apresentado na Tabela 5, ambas as plataformas em nuvem (*AWS*, *Azure* e *GCP*) e os modelos generativo, possuem suporte oficial por meio de suas próprias documentações, portais e canais institucionais. Há participação ativa de usuários em fóruns, tutoriais e repositórios abertos. O *Google Cloud* e o *ChatGPT-4o* estão presentes no *Stack Overflow*, onde desenvolvedores compartilham soluções, boas práticas e resolvem dúvidas de implementação. A *Amazon Web Services (AWS)* mantém fóruns dedicados e uma base de conhecimento técnico organizada. Já a *Microsoft* oferece integração com o suporte técnico do *Azure* além de conteúdo disponibilizado por meio da plataforma *Microsoft Learn*. No caso dos modelos generativos, como *Gemini* e *ChatGPT-4o*, embora o suporte formal seja limitado em comparação às plataformas em nuvem, observa-se a existência de materiais produzidos pela comunidade, com destaque para tutoriais, bibliotecas auxiliares e projetos abertos.

#### 4.2.9 Atualização e Ciclos de Vida

As atualizações apresentam variações entre as plataformas em nuvem (*AWS*, *Azure* e *GCP*) e modelos generativos, conforme a Tabela 5. As *APIs* de visão computacional mantêm ciclos regulares: *Azure Computer Vision* recebe atualizações em intervalos aproximados de 3 a 6 meses, o *Google Cloud Vision* varia entre 4 e 8 meses, já o *Amazon Rekognition* apresenta um intervalo médio de atualização anual, tornando assim o ciclo lento em relação aos demais.

No caso dos modelos generativos, como *ChatGPT* e *Gemini*, os ciclos de atualização tendem a ser frequentes e acelerados, refletindo o ritmo contínuo de evolução da inteligência artificial generativa. Esses modelos passam por melhorias incrementais aproximadamente mensais, com versões aprimoradas, ajustes na capacidade de contexto, multimodalidade e otimizações de desempenho. Essa frequência elevada de atualizações pode ser relevante para aplicações que demandam inovação constante, mas também exige atenção quanto à reprodutibilidade e à estabilidade em ambientes produtivos.

#### 4.2.10 Casos de Uso

Com base na Tabela 5, observa-se que os casos de uso recorrentes refletem o posicionamento e o público-alvo de cada solução. *APIs* de visão computacional (*Amazon Rekognition*, *Azure Computer Vision* e *Google Cloud Vision API*), concentram-se em aplicações empresariais e operacionais, o que inclui controle de acesso, automação de processos, moderação de conteúdo, *OCR* para documentos e análise de imagens em escala. Esses serviços são utilizados em setores como segurança, varejo, saúde e governo, com foco em precisão, robustez e integração a sistemas existentes.

Por outro lado, os modelos generativos multimodais, como *Gemini* e *ChatGPT-4o*, têm seus casos de uso voltados à interação com usuários finais, educação, acessibilidade e geração de descrições com base em imagens. Essas soluções são aplicadas em interfaces conversacionais, assistentes virtuais, tutoriais visuais, criação de conteúdo e aplicações que exigem interpretação contextual de imagens combinadas com linguagem natural.

Contudo, enquanto as *APIs* da *AWS*, *Azure* e *GCP* enfatizam confiabilidade e controle técnico, os modelos generativos apresentam foco na experiência do usuário, criatividade e flexibilidade na construção de soluções interativas.

## 5 CONSIDERAÇÕES

Dadas as informações obtidas através da análise desenvolvida ao longo deste trabalho, foi possível apresentar um panorama do ecossistema atual de soluções de *Machine Learning as a Service (MLaaS)*, evidenciando tanto a diversidade funcional das *APIs* disponíveis quanto os critérios que influenciam sua adoção em diferentes contextos. A comparação entre plataformas em nuvem (*AWS*, *Azure* e *GCP*) e modelos generativos multimodais (*Gemini* e *ChatGPT*) revelou um

cenário marcado pela coexistência de abordagens complementares.

As APIs das plataformas em nuvem (*AWS*, *Azure* e *GCP*) demonstraram maturidade em termos de documentação, integração com ambientes corporativos, políticas de segurança e garantia de desempenho, sendo recomendados para aplicações com altos requisitos técnicos e operacionais. Por outro lado, os modelos baseados em *IA* generativa apresentaram um avanço relevante em flexibilidade, acessibilidade e capacidade multimodal, tornando-se úteis para prototipação rápida, atendimento virtual e geração de conteúdo contextualizado. Essas observações indicam que para a escolha da *API* adequada depende não tanto da quantidade de funcionalidades ofertadas, mas sim do perfil do projeto, recursos da equipe e o escopo da aplicação.

Nesse sentido, os critérios de avaliação aqui adotados como funcionalidades, integração, escalabilidade, desempenho, custo, facilidade de uso, suporte e atualizações, foram essenciais para estruturar uma análise comparativa clara, sistemática e tecnicamente fundamentada.

O critério de funcionalidades permitiu identificar o núcleo de serviços recorrentes como visão computacional, *NLP* e tradução, bem como as particularidades que diferenciam as plataformas, como a busca semântica da *AWS* (*Amazon Kendra*) ou o *Face API* da *Azure*.

A integração com o ecossistema revelou a capacidade das plataformas de se conectarem a serviços nativos como *S3* e *Lambda* na *AWS*, *Logic Apps* e *Power BI* na *Azure*, e *Vertex AI* no *GCP*, o que impacta diretamente a viabilidade de adoção em ambientes corporativos.

A análise da facilidade de uso evidenciou que as *APIS* das plataformas em nuvem (*AWS*, *Azure* e *GCP*), apesar da ampla documentação, exigem familiaridade com conceitos como autenticação, configuração de *SDKs* e gestão de permissões, enquanto os modelos generativos se destacam pela interface simples e linguagem natural, mesmo oferecendo controle técnico limitado.

O critério de desempenho e escalabilidade demonstrou que as soluções das plataformas em nuvem (*AWS*, *Azure* e *GCP*) são adequadas para operações com grandes volumes e processos contínuos, com suporte a chamadas assíncronas e processamento em lote, ao passo que os modelos generativos funcionam melhor para interações pontuais com raciocínio contextual.

No aspecto do modelo de precificação, identificou-se que as plataformas em nuvem trabalham com cobrança por tipo de análise, enquanto os modelos generativos adotam tarifação por tokens, o que requer atenção ao uso multimodal e ao volume de entrada e saída.

Já os critérios de *SLA* e conformidade mostraram-se decisivos para o uso corporativo: *AWS*, *Azure* e *GCP* apresentam garantias de 99,9% de disponibilidade e atendem normas como *GDPR* e *HIPAA*; os modelos generativos, apesar de seguros, não possuem *SLA* público documentado.

No que diz respeito os critérios de suporte, atualizações e comunidade revelaram o diferencial das plataformas em nuvem (*AWS*, *Azure* e *GCP*) em termos de base técnica, exemplos práticos, fóruns e *SDKs* para múltiplas linguagens, enquanto *Gemini* e *ChatGPT* apresenta integração simplificada, mas dependem de conteúdo da comunidade para tarefas específicas.

Esses critérios permitiram destacar não apenas as capacidades operacionais de cada solução, mas também sua adaptabilidade a diferentes realidades de uso, evidenciando padrões, lacunas e oportunidades estratégicas. Portanto, o estudo propõe não apenas um levantamento descritivo, mas também uma reflexão crítica orientada por parâmetros objetivos, que pode contribuir para a adoção consciente, eficiente e alinhada às reais necessidades de desenvolvedores, empresas e usuários finais no cenário em constante evolução do *MLaaS*.

## 6 CONCLUSÃO

Este trabalho teve como objetivo apresentar uma análise comparativa entre as principais APIs de *Machine Learning como Serviço (MLaaS)* disponibilizadas pelas plataformas *Amazon Web Services (AWS)*, *Microsoft Azure*, *Google Cloud Platform (GCP)*, *Google DeepMind (Gemini)* e *OpenAI (ChatGPT)*. A partir de uma metodologia baseada em análise documental e considerando critérios técnicos como funcionalidades, integração, desempenho, custo-benefício e usabilidade, foi possível identificar as particularidades de cada solução e os contextos adequados para sua adoção.

Os resultados demonstraram que, embora haja um núcleo comum de funcionalidades como, visão computacional, processamento de linguagem natural (*NLP*), tradução automática e *OCR*, cada plataforma adota estratégias distintas quanto à arquitetura, modelo de precificação, suporte e integração com ecossistemas nativos. As APIs das plataformas em nuvem (*AWS*, *Azure* e *GCP*) mostraram-se adequadas a ambientes corporativos, enquanto os modelos generativos multimodais apresentaram flexibilidade e acessibilidade, especialmente em aplicações voltadas à interação com o usuário.

Diante dos resultados obtidos, conclui-se que o cenário atual de *Machine Learning como Serviço* é dinâmico, diverso e em constante evolução. Este trabalho buscou contribuir com uma análise estruturada e comparativa entre diferentes soluções, destacando que a escolha da API ideal não se limita a recursos disponíveis, mas envolve fatores como perfil da aplicação, exigências técnicas, orçamento e integração com outros serviços. Cabe ressaltar que este estudo apresenta limitações, tais como o foco exclusivo na análise documental e a ausência de testes práticos que avaliem o desempenho das APIs em contextos reais, o que pode restringir a aplicabilidade dos resultados. Espera-se que os resultados aqui obtidos possam contribuir para tomada de decisão na escolha da plataforma ou serviços a serem empregados por parte dos desenvolvedores.

## 7 TRABALHOS FUTUROS

Tendo em vista o caráter não exaustivo das buscas aqui realizadas, existem lacunas em aberto para realização de estudos futuros, dentre eles, comparativos envolvendo estudos de caso práticos com os modelos nas plataformas e serviços; existem ainda possibilidades de ampliação dos

comparativos ao adicionar análises de mercado e percepções de usuários ou desenvolvedores de aplicações *MLaaS*.

## REFERÊNCIAS

AMAZON. **O que é MLOps?** 2024. Disponível em: <https://aws.amazon.com/pt/what-is/mlops/>. Acesso em: 29 mar. 2025.

AMAZON WEB SERVICES. **What is Amazon Rekognition?** 2024. Disponível em: <https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html>. Acesso em: 28 mar. 2025.

ARAUJO, M. N. de. **Computação em nuvem: uma análise comparativa das plataformas disponibilizadas por Amazon, Google e Microsoft.** 2019. Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, Natal, 2019. Disponível em: <https://memoria.ifrn.edu.br/handle/1044/2524>. Acesso em: 8 abr. 2025.

DATA BRIDGE MARKET RESEARCH. **Mercado global de machine learning como serviço (MLaaS): tendências da indústria e previsões para 2028.** 2021. Disponível em: <https://www.databridgemarketresearch.com/pt/reports/global-machine-learning-service-mlaas-market>. Acesso em: 6 mai. 2025.

ENGINEERING BRASIL. **Transformação digital: a influência da computação em nuvem por meio de APIs.** 2025. Disponível em: <https://blog.engdb.com.br/computacao-em-nuvem-para-desenvolvimento-api-e-transformacao-digital/>. Acesso em: 9 jun. 2025.

ESCOLA BRITÂNICA DE ARTES CRIATIVAS E TECNOLOGIA. **O que é uma API: para que serve e como utilizar.** 2023. Disponível em: <https://ebaonline.com.br/blog/o-que-e-uma-api-seo>. Acesso em: 3 jun. 2025.

FREITAS, LEANDSON DE OLIVEIRA. **Computação em nuvem: uma breve revisão bibliográfica.** 2023. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal Rural do Semi-Árido, Mossoró, 2023. Disponível em: <https://repositorio.ufersa.edu.br/items/5846175d-de61-42ad-8d38-7812a647ae9c>. Acesso em: 10 jan. 2025.

GOOGLE. **O que é machine learning (ML)?** 2024. Disponível em: <https://cloud.google.com/learn/what-is-machine-learning?hl=pt-BR>. Acesso em: 2 jan. 2025.

GOOGLE. **Recursos da API Cloud Vision.** 2024. Disponível em: <https://cloud.google.com/vision/docs/features-list?hl=pt-br>. Acesso em: 6 mai. 2025.

GOOGLE. **Documentação da API Gemini.** 2025. Disponível em: <https://ai.google.dev/gemini-api/docs?hl=ptbr>. Acesso em: 3 fev. 2025.

HAPONIK, A. **Google Gemini API vs. OpenAI API: main differences.** 2024. Disponível em: <https://addepto.com/blog/google-gemini-api-vs-open-ai-api-main-differences/>. Acesso em: 3 fev.

2025.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. **Machine learning and deep learning**. *Electronic Markets*, v. 31, p. 685–695, 2021. DOI: 10.1007/s12525-021-00475-2. Disponível em: <https://link.springer.com/article/10.1007/s12525-021-00475-2>. Acesso em: 9 fev. 2025.

KREUZBERGER, D.; KUHL, N.; HIRSCHL, S. **Machine learning operations (MLOps): overview, definition, and architecture**. *IEEE Access*, v. 11, p. 31866–31879, 2023. DOI: 10.1109/ACCESS.2023.3257413. Disponível em: <https://ieeexplore.ieee.org/document/10069189>. Acesso em: 8 fev. 2025.

MICROSOFT. **O que é computação em nuvem?** 2024. Disponível em: <https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-cloud-computing/>. Acesso em: 4 jan. 2025.

MICROSOFT. **Visão geral do serviço de Visão Computacional do Azure**. 2024. Disponível em: <https://learn.microsoft.com/pt-br/azure/ai-services/computer-vision/overview>. Acesso em: 6 mai. 2025.

ML JOURNEY. **Machine learning in cloud computing**. 2025. Disponível em: <https://mljourney.com/machine-learning-in-cloud-computing/>. Acesso em: 10 mai. 2025.

MOUTAOUAKAL, W. E.; BAÏNA, K. **Comparative experimentation of MLOps power on Microsoft Azure, Amazon Web Services, and Google Cloud Platform**. In: 2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech). Anais [...]. [S.l.]: IEEE, 2023. p. 1–8. Acesso em: 15 jan. 2025.

MUFTI, T.; MITTAL, P.; GUPTA, B. **A review on Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) services**. *International Journal of Computer Applications*, 2021. DOI: 10.4108/eai.27-2-2020.2303255. Disponível em: <https://doi.org/10.4108/eai.27-2-2020.2303255>. Acesso em: 14 jun. 2025.

NOSHIRI, N.; KHORRAMFAR, M.; HALABI, T. **Machine learning-as-a-service performance evaluation on multi-class datasets**. In: 2021 IEEE International Conference on Smart Internet of Things (SmartIoT). Anais [...]. [S.l.]: IEEE, 2021. p. 332–336. Acesso em: 14 jun. 2025.

OPENAI. **ChatGPT-4o**. 2025. Disponível em: <https://platform.openai.com/docs/models/chatgpt-4o-latest>. Acesso em: 14 mai. 2025.

PLETCHER, S. **What is machine learning as a service (MLaaS)?** 2023. Disponível em: <https://www.pluralsight.com/resources/blog/cloud/what-is-machine-learning-as-a-service-mlaas>. Acesso em: 16 mar. 2025.

RAVINDRANATHAN, M. K.; VADIVU, D. S.; RAJAGOPALAN, N. **Cloud-driven machine learning with AWS: a comprehensive review of services**. In: 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE). Anais [...]. [S.l.]: IEEE, 2024. p. 1–8. Acesso em: 14 jun. 2025.

SHILPASHREE, S.; PATIL, R. R.; PARVATHI, C. **Cloud computing: an overview**. *International Journal of Engineering and Technology*, v. 7, p. 1–6, 2018. Disponível em: [https://www.researchgate.net/publication/328774881\\_Cloud\\_computing\\_an\\_overview](https://www.researchgate.net/publication/328774881_Cloud_computing_an_overview). Acesso em: 15 jun. 2025.

TD SYNnex. **Machine learning as a service: como funciona e quais as vantagens?** 2024. Disponível em: <https://blog-pt.lac.tdsynnex.com/machine-learning-as-a-service-como-funciona-e-quais-as-vantagens>. Acesso em: 20 jan. 2025.

YOUNIS, R. et al. **A comprehensive analysis of cloud service models: IaaS, PaaS, and SaaS in the context of emerging technologies and trends**. In: 2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE). Anais [...]. [S.l.]: IEEE, 2024. p. 1–6. Acesso em: 30 jan. 2025.