


FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE DADOS – UM ESTUDO COMPARATIVO

ARTIFICIAL INTELLIGENCE TOOLS FOR DATA PREDICTION – A COMPARATIVE STUDY

HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL PARA LA PREDICCIÓN DE DATOS: UN ESTUDIO COMPARATIVO

 <https://doi.org/10.56238/arev7n6-335>

Data de submissão: 28/05/2025

Data de publicação: 28/06/2025

Gabriel Henrique Lenz

Curso de Bacharelado em Sistemas de Informação – Universidade Federal de Santa Maria (UFSM), Campus Frederico Westphalen
E-mail: gabrielxlens@gmail.com

Evandro Preuss

Departamento de Tecnologia da Informação (DTecInf), Universidade Federal de Santa Maria (UFSM), Campus Frederico Westphalen
E-mail: evandro.preuss@ufsm.br

Roberto Franciscatto

Departamento de Tecnologia da Informação (DTecInf), Universidade Federal de Santa Maria (UFSM), Campus Frederico Westphalen
E-mail: roberto.franciscatto@ufsm.br

Solange Pertile

Departamento de Tecnologia da Informação (DTecInf), Universidade Federal de Santa Maria (UFSM), Campus Frederico Westphalen
E-mail: solange.pertile@ufsm.br

RESUMO

Com a grande evolução das Inteligências Artificiais, diferentes organizações necessitam se adaptar a essa nova era. Uma problemática enfrentada, são os desafios de implementar essa tecnologia dentro do contexto de negócio. Esse artigo realiza a exploração de diferentes ferramentas de Inteligência Artificial que prometem abstrair a dificuldade da criação de diferentes Modelos de IA, realizando uma análise de forma prática das principais ferramentas gratuitas existentes no mercado atualmente.

Palavras-chave: Análise Preditiva. Machine Learning.

ABSTRACT

With the rapid advancement of Artificial Intelligence, different organizations need to adapt to this new era. One of the challenges faced is the difficulty of implementing this technology within a business context. This article explores various AI tools that aim to simplify the creation of different AI models, providing a practical analysis of the main free tools currently available on the market.

Keywords: Predictive Analysis. Machine Learning.

RESUMEN

Con la gran evolución de la Inteligencia Artificial, diversas organizaciones necesitan adaptarse a esta nueva era. Un problema al que se enfrentan es la dificultad de implementar esta tecnología en el contexto empresarial. Este artículo explora diferentes herramientas de Inteligencia Artificial que prometen simplificar la creación de diferentes modelos de IA, realizando un análisis práctico de las principales herramientas gratuitas disponibles actualmente en el mercado.

Palabras clave: Análisis predictivo. Aprendizaje automático.

1 INTRODUÇÃO

Em um mundo onde os dados se tornaram um dos ativos mais valiosos para empresas e organizações, a capacidade de prever comportamentos, tendências e resultados futuros, tornou-se essencial para a tomada de decisões estratégicas. Nesse contexto, a Inteligência Artificial (IA) tem desempenhado um papel cada vez mais relevante, especialmente através de ferramentas voltadas à análise preditiva de dados.

As ferramentas de IA permitem que usuários de diferentes níveis técnicos, desde cientistas de dados até analistas de negócios, construam modelos de machine learning capazes de identificar padrões e realizar previsões com base em grandes volumes de dados. Com o avanço das plataformas de AutoML (Aprendizado de Máquina Automatizado) e interfaces cada vez mais intuitivas, o acesso a esse tipo de tecnologia está se tornando mais democrático.

Neste artigo, vamos explorar e comparar algumas das principais ferramentas de IA disponíveis no mercado para predição de dados. Serão analisadas plataformas como BigML, Azure Machine Learning, DataRobot, Orange e Google Data Assistant, destacando seus diferenciais, pontos fortes e principais aplicações. O objetivo é oferecer uma visão clara e prática sobre como essas soluções podem ser aplicadas em diferentes contextos, ajudando na escolha da ferramenta mais adequada às eventuais necessidades.

2 REFERENCIAL TEÓRICO

Apresenta-se, nesta seção, um breve referencial teórico sobre as áreas envolvidas neste trabalho, apresentando conceitos de análise de dados, inteligência artificial, algoritmos de machine learning e métricas de avaliação.

2.1 ANÁLISE DE DADOS

Segundo Géron,(2021), a análise de dados pode ser separada em quatro tipos:

- **Análise Descritiva:** Seu principal objetivo é compreender e sumarizar eventos com técnicas estatísticas simples, como medidas de tendência central (média, moda e mediana) medidas de dispersão (amplitude, variância e desvio padrão) e visualizações (GÉRON, 2021). Geralmente constitui o primeiro passo na análise de dados, proporcionando uma compreensão inicial das características e distribuições dos dados (MIRANDA, 2024);
- **Análise Diagnóstica:** O seu objetivo é investigar causas de um evento observado dentro da análise descritiva, através de correlações, segmentações e comparações (GÉRON, 2021), ou seja, tenta identificar as causas subjacentes aos dados observados (CIO, 2022);

- **Análise Preditiva:** O objetivo desse tipo de análise é prever eventos futuros com base em eventos passados, utilizando tanto técnicas de modelagem estatística como aprendizado de máquina, como modelos supervisionados, não supervisionados e semi supervisionados (GÉRON, 2021);
- **Análise Prescritiva:** Por fim, nesse tipo de análise, é indicado ações a serem tomadas baseado nos resultados de análises preditivas (GÉRON, 2021), buscando antecipar o que pode acontecer no futuro, permitindo recomendar ações específicas que influenciem positivamente os resultados (MIRANDA, 2024).

2.2 ALGORITMOS DE MACHINE LEARNING

Algoritmos de Machine Learning (ou aprendizado de máquina) podem ser definidos como a “ciência da programação de computadores de modo que eles possam aprender com os dados” (GÉRON, 2021). Os algoritmos de Machine Learning têm a capacidade de aprender automaticamente padrões dentro dos dados, tornando possível a criação de algoritmos muito extensos para serem feitos apenas por humanos.

Existem diferentes tipos de Sistemas de Aprendizado de Máquina, como:

- **Aprendizado supervisionado**, não supervisionado, semi supervisionado e aprendizado por reforço (sendo ou não treinados com supervisão humana);
- **Se tem conhecimento**, incrementado gradativamente em tempo real (aprendizado online e aprendizado em batch);
- **Se funcionam comparando** novos pontos de dados com pontos de dados conhecidos, ou se detectam padrões em dados de treinamento e criam um modelo preditivo (aprendizado baseado em instâncias e aprendizado baseado em modelos).

2.2.1 Aprendizado Supervisionado

Dentro de um aprendizado supervisionado, o conjunto de dados contém colunas com características (features), como idade, estado civil e renda. Além de uma coluna com a variável alvo, essa podendo ser, por exemplo, se o usuário em questão “Pagou” ou “Não Pagou” (0 ou 1). O aprendizado supervisionado é baseado no treinamento de uma amostra de dados com a classificação correta já atribuída, (PAIXÃO, SANTOS, ARAUJO, RIBEIRO, MORAES, RIBEIRO, 2020).

2.2.2 Aprendizado Não Supervisionado

Diferente do aprendizado supervisionado, o aprendizado não supervisionado não possui uma coluna com a variável alvo. É usado para identificar novos padrões e detectar anomalias, tentando dar sentido aos dados por conta própria, encontrando recursos e padrões (SILVA, 2023). Esse tipo de algoritmo é utilizado em problemas de:

- Clusterização;
- Detecção de anomalias;
- Visualização e redução de dimensionalidade;
- Aprendizado de regras por associação.

2.2.3 Aprendizado Semi Supervisionado

Mistura o aprendizado supervisionado junto do aprendizado não supervisionado. Um exemplo clássico é um aplicativo de armazenamento de imagens. O algoritmo não supervisionado irá separar pessoas (Clusterização), porém, como não sabe o nome delas, cabe ao usuário inserir um nome para cada uma dessas pessoas (Aprendizado não supervisionado). Segundo Matsubara (2004), o aprendizado supervisionado requer uma quantidade expressiva de exemplos rotulados, entretanto, obter esses exemplos nem sempre é uma tarefa trivial, muitas vezes requerendo um trabalho manual, lento e altamente custoso. Submeter um especialista para rotular 1000 exemplos não seria uma tarefa factível, porém, solicitar a esse especialista que rotule 20 ou 30 exemplos, é possível utilizar algoritmos de aprendizado de máquina para realizar a previsão de qual rótulo os dados restantes se adequam.

2.2.4 Aprendizado por Reforço

Segundo Géron (2020), o sistema de aprendizado, chamado de agente no contexto de aprendizado por reforço, pode assistir o ambiente, selecionar e executar ações além de obter recompensas ou penalidades em troca de suas ações. O algoritmo deve aprender sozinho qual é a melhor estratégia, chamada de política, para obter o maior número de recompensas ao longo do tempo. Uma política define qual ação o agente deve escolher quando está em determinada situação. O aprendizado por reforço ocorre por meio de uma interação direta entre o agente e o ambiente, ou seja, é necessária uma exaustiva exploração do espaço de estados-ações, normalmente demorada (BIANCHI, COSTA, 2004).

Por exemplo, em um jogo como o “Super Mario”, é possível definir recompensas e penalidades para que o sistema “aprenda” a passar de fase, por exemplo:

- Coletar moedas: 5 pontos;

- Derrotar inimigos: 20 pontos;
- Morrer: -50 pontos.

2.3 MÉTRICAS DE AVALIAÇÃO

As métricas de avaliação dentro de um algoritmo de Machine Learning servem para avaliar um modelo e garantir que está apto a seguir para o próximo passo.

2.3.1 Acurácia

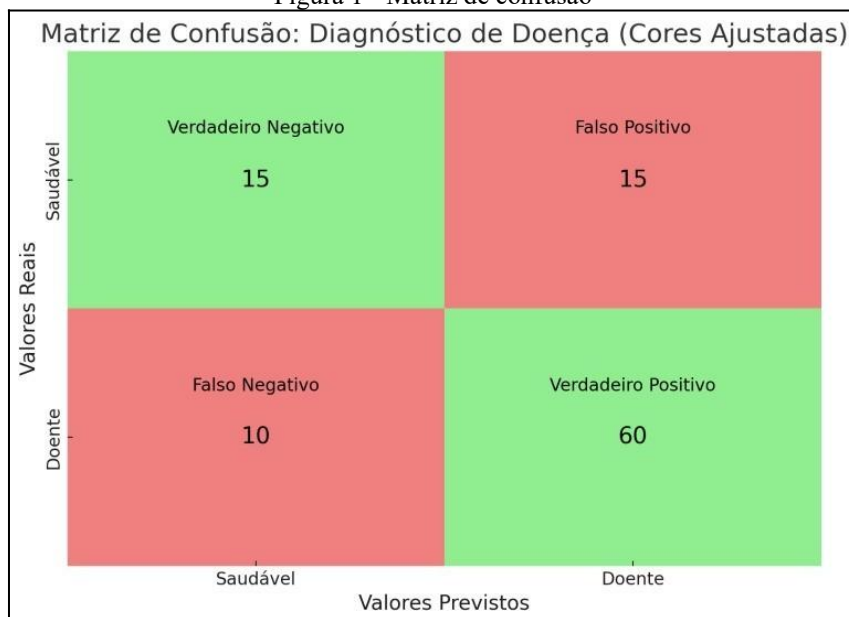
A acurácia é uma métrica muito utilizada para a avaliação de modelos de Machine Learning. Ela funciona medindo a proporção de exemplos corretamente classificados pelo modelo. Porém, a acurácia pode ser enganosa, principalmente em dados desbalanceados (ex: 90% da base de dados é pertencente a uma classe, enquanto somente 10% pertencem à outra). Caso existam nove classes "1" e uma classe "0", em um cenário onde o modelo seja treinado e alcance 90% de acurácia, mesmo sendo uma ótima avaliação, o mesmo resultado seria alcançado em um modelo que apenas indicasse que todas as classes são pertencentes a "1", ou seja, mesmo com um ótimo desempenho na acurácia, o modelo é ineficiente. Como apontam Castro e Braga (2011), em problemas desbalanceados, a acurácia pode induzir interpretações equivocadas sobre a real capacidade do classificador em identificar corretamente a classe minoritária.

2.3.2 Matriz de Confusão

A matriz de confusão é uma tabela que mostra a quantidade de exemplos classificados corretamente e incorretamente em cada uma das classes. Na imagem abaixo é possível ver a matriz de confusão, nela é possível identificar:

- TP (True Positive): Verdadeiros Positivos, valores que foram classificados como positivos e realmente são positivos;
- TN (True Negative): Verdadeiros Negativos, valores que foram classificados como negativos e realmente são negativos;
- FP (False Positive): Falsos Positivos, valores que foram classificados como positivos, mas na verdade são negativos;
- FN (False Negative): Falsos Negativos, valores que foram classificados como negativos, mas na verdade são positivos.

Figura 1 - Matriz de confusão



Fonte: (Datageeks, 2024)

A partir dos resultados da Matriz de Confusão, é possível realizar o cálculo de novas métricas, que serão vistas a seguir.

2.3.3 Precisão

A precisão mede a proporção de exemplos classificados como positivos e que realmente são positivos, sua fórmula da precisão pode ser dada por:

$$\text{Precisão} = \text{TP} / (\text{TP} + \text{FP})$$

2.3.4 Revocação

A revocação mede a proporção de exemplos positivos que foram corretamente identificados pelo modelo, sua fórmula pode ser dada por:

$$\text{Revocação} = \text{TP} / (\text{TP} + \text{FN})$$

2.3.5 F1-Score

Geralmente, tanto a Precisão, quanto a Revocação, são combinadas para formar a F1- Score, que nada mais é que uma média harmônica (dá mais peso para um valor mais baixo) da Precisão e Revocação. Sua fórmula pode ser dada por:

$$\text{F1} = 2 / ((1/\text{Revocação}) + (1/\text{Precisão}))$$

2.3.6 ROC AUC

A curva ROC representa a taxa de verdadeiros positivos (TP) em relação à taxa de falsos positivos (FP).

2.4 OVERFITTING E UNDERFITTING

Um problema que deve ter sua devida atenção são os casos de Overfitting e Underfitting.

O Overfitting é extremamente comum quando certos cuidados não são tomados. Ele ocorre quando o modelo de Machine Learning se ajusta excessivamente aos dados de treino, não sendo capaz de generalizar para dados que não foram vistos anteriormente. Basicamente, quando ocorre um Overfitting, é o mesmo que dizer que o modelo “decorou” os dados de treino, tendo uma ótima performance predizendo os dados de treino (muitas vezes, 100% de acerto), porém, tem uma péssima performance em dados novos. Existem vários motivos para que o Overfitting ocorra, como uma baixa quantidade de dados ou um modelo muito complexo. Segundo Jabbar, Khan (2015), em determinado momento durante o treinamento, o modelo não melhora mais a sua capacidade de resolver problemas, mas sim começa a aprender alguma regularidade aleatória contida no conjunto de treinamento.

O Underfitting ocorre quando o modelo não consegue se ajustar nem aos dados de treinamento, tendo uma performance extremamente ruim tanto nos dados de treino quanto nos dados de teste. O Underfitting pode ocorrer quando o modelo é muito simples, possui instâncias de dados insuficientes, ou até mesmo erros de rotulagem (GÉRON, 2020). É o oposto do Overfitting, ocorrendo quando o modelo é incapaz de capturar a variabilidade dos dados (JABBAR, KHAN, 2015).

3 TRABALHOS RELACIONADOS

Nessa seção apresentam-se alguns trabalhos relacionados à ferramenta proposta. Foram selecionados três trabalhos que possuem relação com ferramentas para análise ou predição de dados.

3.1 FERRAMENTAS PARA ANÁLISE DE DADOS

O trabalho de Gallego, Cruz e Alves (2023) apresenta uma fundamentação teórica que define uma análise de dados como o processo de extrair insights corporativos valiosos a partir de informações coletadas de diversas fontes dentro de uma empresa. Destacam a análise a partir de quatro formas utilizadas:

1. **Análise Descritiva:** Consiste em descrever os dados observados e compreender o que aconteceu (Rosa, 2018);

2. Análise Diagnóstica: Tem o objetivo de identificar as causas subjacentes aos dados observados (CIO, 2022);
3. Análise Preditiva: Busca prever eventos futuros com base na compreensão das dinâmicas subjacentes dos dados (Rosa, 2018);
4. Análise Prescritiva: Dados e informações são usados para definir metas e objetivos, permitindo a empresa a tomar decisões com base em cenários hipotéticos (Cortex, 2022).

Também incluiu ferramentas populares para a análise de dados, como:

1. Python;
2. Power BI;
3. Linguagem R;
4. Data Studio.

Foi realizada uma pesquisa Quantitativa/Qualitativa (mostrada abaixo):

Figura 2 - Exemplo de Pesquisa Quali/Quanti

Pesquisa de campo para coleta de dados utilizando 61 pessoas

1. Você sabe o que é Análise de Dados?
2. Você utiliza ou já utilizou alguma ferramenta/software para análise de dados?
() sim () não
3. Se sim, responder a pergunta a seguir
4. Qual das ferramentas/software abaixo você já utilizou?
() *Excel* ,
() *Python*,
() *Microsoft Power Bi*,
() *Linguagem R*,
() *Google Data Studio*,
() outro (especificar) _____
5. Cite 3 vantagens e 3 desvantagens na ferramenta/software que você já utilizou.

Fonte: (Gallego, 2023)

Como resultado, 63% afirmam saber o que é a análise de dados, em contrapartida, apenas 38% afirmam já ter utilizado alguma ferramenta ou software para análise de dados. Já as ferramentas utilizadas contam com 55% para o Excel, 32% Power BI, 5,88 Google Data Studio, 3% Python e 3% responderam outro. Quanto a quinta questão, sobre as 3 vantagens, a maioria respondeu que a facilidade é a principal vantagem dessas ferramentas, enquanto nas desvantagens, 46% acreditam que a complexidade é um problema, enquanto 39% responderam que não tem, já uma baixa parcela (15%) relatou o fato de serem pagas.

3.2 AVALIAÇÃO E ANÁLISE DE FERRAMENTAS BUSINESS INTELLIGENCE PARA VISUALIZAÇÃO DE DADOS

No estudo de Lousa, Pedrosa e Bernardino (2019) foi apresentada uma análise comparativa entre Power BI, Tableau, Sisense e Qlikview. Dentre os critérios de avaliação das ferramentas, estão:

1. Categoria do Software: Como aplicação Desktop ou plataforma Web;
2. Estrutura de Visualização: Detalhando qual o tipo de visualização predominante no Software, como gráficos, mapas, dashboards etc.;
3. Sistema Operacional: Sistema em que opera, como Linux e Windows;
4. Licença: Diferencia entre licença comercial e open-source;
5. Escalabilidade: Se trata da capacidade que as ferramentas têm de se ligarem a grandes volumes de dados;
6. Extensibilidade: Está relacionado com os mecanismos que vão além das funcionalidades das plataformas, como plugins;
7. Versão mais recente e data: Apresenta a última versão lançada até o presente.

3.3 FERRAMENTAS OPEN-SOURCE DE QUALIDADE DE DADOS PARA LICITAÇÕES PÚBLICAS: UMA ANÁLISE COMPARATIVA

O trabalho considera oito ferramentas open-source que consideram diferentes dimensões de qualidade de dados, elegendo a ferramenta Great Expectations (GE) como a mais adequada para o problema analisado. Entre as oito ferramentas analisadas estão:

1. Aggregate Profiler;
2. Apache Griffin;
3. Great Expectations;
4. MobyDQ;
5. OpenRefine & Metric;
6. PyDeequ;
7. Talend Open Studio.

No trabalho é realizada uma comparação de funcionalidades, sendo categorizadas como atendido, não atendido, parcialmente atendido e customização de indicadores disponível. Já entre os requisitos comparados estão:

1. Formatação de tabelas;
2. Restrições sobre valores;

3. Intervalo de valores;
4. Casamento de padrões em strings;
5. Dados em formato de data ou JSON;
6. Funções de agregação de dados;
7. Operações multicolumna;
8. Funções relacionadas a distribuições de probabilidade;
9. Funções relacionadas a arquivos;
10. Possibilidade de customização de indicadores.

4 ESTUDO COMPARATIVO DE FERRAMENTAS DE IA PARA PREDIÇÃO DE DADOS

Nesta seção apresenta-se o estudo comparativo das principais ferramentas de Inteligência Artificial voltadas para análise e predição de dados existentes atualmente no mercado.

4.1 METODOLOGIA

Este artigo adota uma abordagem quantitativa e exploratória com o objetivo de comparar ferramentas de Inteligência Artificial (IA) voltadas para a predição de dados, avaliando seu desempenho, aplicabilidade e facilidade de uso. A metodologia foi estruturada em quatro etapas principais: seleção das ferramentas, definição dos critérios de avaliação, construção do conjunto de dados e realização dos testes comparativos.

4.1.1 Seleção das Ferramentas

Foram selecionadas cinco ferramentas de IA amplamente utilizadas no contexto de predição de dados, levando em consideração principalmente: sua popularidade, documentação, acesso gratuito ou por plano educacional, e aplicabilidade em diferentes cenários de análise preditiva. As ferramentas selecionadas foram:

- BigML (<https://bigml.com/>).
- Azure Machine Learning (<https://azure.microsoft.com/pt-br/products/machine-learning>).
- DataRobot (<https://www.datarobot.com/>).
- Orange (<https://orangedatamining.com/>).
- Google Data Assistant (<https://colab.research.google.com/>).

4.1.2 Conjunto de Dados

Para assegurar uma comparação similar entre as ferramentas (explorando seus recursos) foi utilizado um conjunto de dados padronizado e que pudesse explorar os principais recursos usuais presentes em softwares de IA voltados para a predição.

4.1.3 Critérios de Avaliação

As ferramentas foram avaliadas de acordo com os seguintes critérios (não exaustivos):

- Principais recursos: funcionalidades centrais da ferramenta aplicadas à predição de dados.
- Pontos Positivos: vantagens observadas no uso da ferramenta.
- Limitações: restrições ou dificuldades que podem impactar sua adoção.
- Tipo de Software: classificação quanto à natureza do software (aberto, proprietário, biblioteca, etc.).
- Preço/Licença: modelo de custo e licenciamento para uso individual ou institucional.
- Suporte e documentação: disponibilidade de materiais de apoio, tutoriais e fóruns.

4.2 FERRAMENTAS ESTUDADAS NESTA PESQUISA

O estudo das ferramentas ocorreu a partir de informações disponibilizadas no próprio site, assim como validação prática do conteúdo gratuito disponibilizado pelas ferramentas.

4.2.1 BigML

BigML é uma plataforma completa de Machine Learning que tem como principal objetivo simplificar a criação e implementação de modelos preditivos. Com uma abordagem intuitiva e acessível, o BigML permite que profissionais de diferentes áreas possam explorar o potencial da inteligência artificial sem a necessidade de conhecimentos técnicos avançados.

Um dos principais diferenciais do BigML é a facilidade de uso. Através de uma interface web amigável ou de sua API REST, usuários podem acessar, criar e automatizar modelos em poucos cliques. A plataforma ainda conta com recursos de visualização interativa e explicações das previsões, tornando os modelos interpretáveis e úteis para a tomada de decisões.

Todos os modelos podem ser exportados em formatos compatíveis com as principais linguagens de programação, como Python, Java, Node.js, entre outras. Isso permite que sejam integrados facilmente a sistemas, aplicativos ou dispositivos.

A plataforma oferece uma ampla variedade de algoritmos de aprendizado supervisionado, como classificação, regressão e séries temporais, além de técnicas de aprendizado não supervisionado, como

análise de clusters, detecção de anomalias, modelagem de tópicos, descoberta de associações e análise de componentes principais (PCA).

A colaboração é outro ponto forte da plataforma. Com recursos para gerenciamento de projetos e equipes, é possível compartilhar e controlar o acesso a recursos com diferentes permissões, facilitando a integração do Machine Learning em ambientes corporativos.

Em termos de implantação, o BigML é flexível. Pode ser utilizado na nuvem pública, em nuvens privadas ou em servidores próprios da empresa. A plataforma se adapta à necessidade de cada organização, desde pequenos projetos até implementações em larga escala.

A segurança e privacidade dos dados são garantidas por conexões criptografadas e painéis privados para cada usuário. Além disso, há opções de implantação privada para atender requisitos regulatórios rigorosos.

O BigML oferece contas gratuitas com acesso completo à plataforma, porém com limitações em tamanho de dados e tarefas paralelas. Para empresas e equipes maiores, existem planos pagos com mais recursos e suporte corporativo.

4.2.1.1 Pontos positivos e Limitações da ferramenta:

Pontos Positivos:

- Interface amigável baseada na web;
- Modelagem preditiva automatizada com suporte a classificação, regressão, clustering e séries temporais;
- APIs para integração com outras aplicações;
- Implementação de workflows completos de machine learning sem necessidade de código.

Limitações:

- Não é possível saber ao certo como o modelo está sendo treinado, já que ele não fornece informações sobre como é feita a tunagem de hiperparâmetros, o que dá a impressão de uma “caixa-preta” do modelo;
- Baixa adoção da comunidade: Conteúdos de explicação sobre o uso da ferramenta mostram-se limitados;
- Valor elevado: O plano mais barato custa U\$30,00 mensais, porém, para um único usuário e com baixo armazenamento. Planos para equipes são mais caros, se estendendo até dez mil dólares mensais (podendo ultrapassar esse valor dependendo da equipe).

4.2.1.2 Conclusão sobre a Ferramenta

Em resumo, o BigML é uma solução prática e versátil que torna o Machine Learning acessível, colaborativo e aplicável a diversos setores da indústria, como saúde, finanças, energia, transporte, telecomunicações, entre outros, porém, não proporciona todo o potencial que um projeto de Machine Learning pode entregar.

4.2.2 Azure Machine Learning

O Azure Machine Learning é uma plataforma da Microsoft que permite a criação, implantação e gerenciamento de modelos de aprendizado de máquina na nuvem, facilitando o trabalho em equipe, já que conta com:

- Notebooks compartilhados;
- Ambientes prontos;
- MLOps para implantar e monitorar modelos;
- Funciona em nuvem.

Tem compatibilidade com outras ferramentas da Microsoft, além de REST APIs. Fornece também uma experiência visual, com visualização de métricas, um designer sem código (arrastar e soltar), ML automatizado e rotulagem de dados com interface simples.

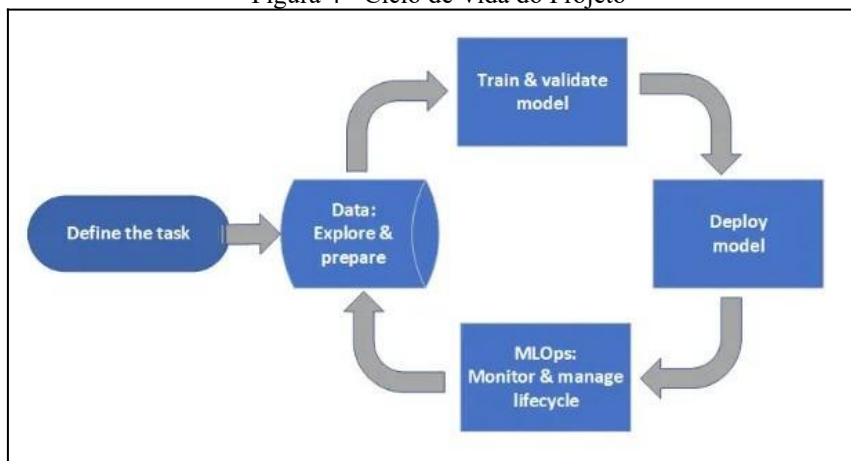
Figura 3 - Tela Inicial - Azure Machine Learning



Fonte: (Azure Microsoft, 2024)

Fornece suporte para criar apps com LLMs como OpenAI, Meta e outros, incluindo um catálogo de modelos prontos e um Prompt Flow (Ferramenta que simplifica o ciclo de desenvolvimento de aplicativos de IA alimentados por LLMs). Além da segurança corporativa do Azure, ele fornece suporte a todo fluxo de trabalho que um projeto de ML necessita.

Figura 4 - Ciclo de Vida do Projeto



Fonte: (Azure Microsoft, 2024)

Além de possuir um AutoML (escolhe o melhor algoritmo e prepara os dados de forma automática) fazendo o mesmo ajuste automático de hiperparâmetros e visualização dos resultados, também fornece a opção de o usuário usar os próprios scripts e frameworks como Scikit-Learn, XGBoost, LightGBM entre outros.

O Azure Machine Learning também fornece a opção de treinamento distribuído e paralelo.

4.2.2.1 Pontos Positivos e Limitações da ferramenta

Pontos Positivos:

- Integração com o ecossistema Microsoft;
- Interface gráfica e notebooks integrados;
- Automated ML (AutoML);
- Gestão de ciclo de vida de modelos;
- Permite uma boa escalabilidade;
- Segurança e governança corporativa. Limitações:
- Curva de aprendizado demorada, o que pode ser complexo para iniciantes;
- Custo elevado;
- Dependência da nuvem Azure;

- A interface gráfica fornecida é útil, porém, menos flexível do que o código puro em tarefas mais complexas.

4.2.2.2 Conclusão sobre a Ferramenta

A Azure Machine Learning é uma das ferramentas estudadas que fornece mais funcionalidades, porém, é paga. Diferente de outras ferramentas, ela requer um conhecimento mais elevado sobre algoritmos, configurações de workspace, pipelines e deploy.

4.2.3 DataRobot

O DataRobot é uma plataforma Web de Machine Learning automatizado que permite a criação de modelos preditivos, abstraindo a necessidade de conhecimentos avançados em programação. É uma ferramenta completa no quesito de automação do ciclo de vida de um projeto de Machine Learning, dando suporte desde a preparação dos dados até a implantação do modelo.

4.2.3.1 Pontos Positivos e Limitações da ferramenta

Pontos Positivos:

- Automatiza todo o processo de Machine Learning;
- Interface amigável (sem código);
- Gera e compara diversos modelos automaticamente;
- Funcionalidades de explicabilidade e interpretação;
- Deploy e monitoramento de modelos com poucos cliques;
- Suporte a dados tabulares, séries temporais e imagens. Limitações:
- Custo alto, voltado principalmente para grandes empresas;
- Plataforma fechada e proprietária;
- Pouca flexibilidade para customizações avançadas;
- Baixa adoção na comunidade (pouco conteúdo para pesquisar);
- Não substitui o conhecimento em dados ou negócios.
- Outras empresas, como a Azure, oferecem serviço de AutoML com preços mais competitivos.

4.2.3.2 Conclusão sobre a Ferramenta

O DataRobot fornece uma demonstração de 15 dias para novos usuários, porém, é necessária a criação com um e-mail corporativo, o que dificulta o acesso para usuários que querem apenas testar suas aplicabilidades.

4.2.4 Orange

A Orange é um software de código aberto que pode ser utilizado tanto para visualização, mineração e análise de dados. Foi desenvolvida em um meio acadêmico, na universidade de Ljubljana (localizada em Ljubljana, capital da Eslovênia).

A Orange é uma ferramenta bem simples. Possui uma interface baseada em blocos para a criação de pipelines de análise de dados e algoritmos de Machine Learning. A ferramenta dispõe opções para a entrada de dados, transformação, visualização, modelos supervisionados, avaliação dos modelos e também modelos não supervisionados.

Figura 5 – Entrada de Dados (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

As opções de dados tratam de importar e exportar os dados para o programa, além de customizações de como serão tratados enquanto rodam.

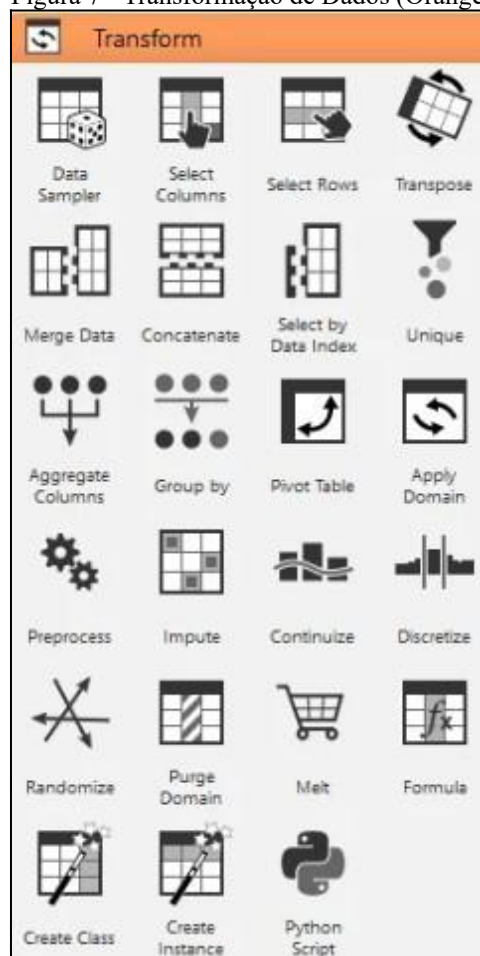
Figura 6 – Tipos de Dados (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

Fornece opções completas de transformação nos dados, abrangendo desde opções simples de selecionar linhas e colunas, até a criação de Scripts em Python, para eventuais necessidades.

Figura 7 - Transformação de Dados (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

A parte de visualização de dados conta com alguns gráficos simples, como histogramas, gráfico de linhas e de dispersão, assim como gráficos mais específicos como uma projeção linear.

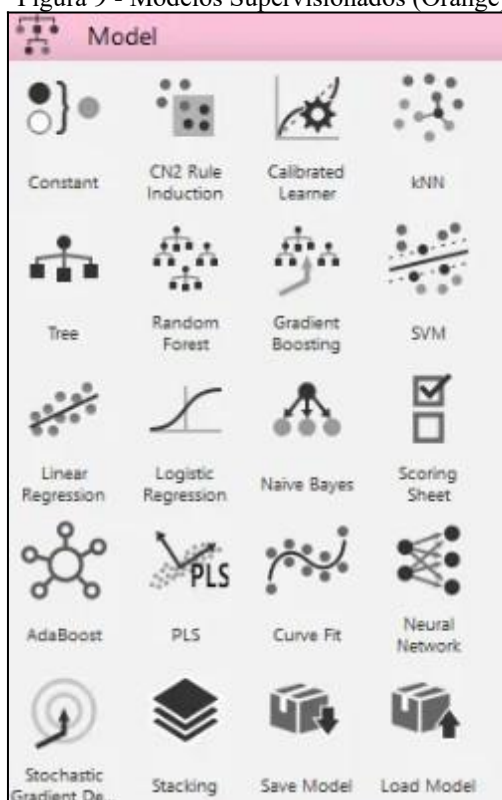
Figura 8 - Visualização de Dados (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

Fornece também diferentes modelos supervisionados, como KNN, Árvore de Decisão, Florestas Aleatórias, Gradient Boosting, SVM, Regressão Linear e Logística. Redes Neurais, Stacking de modelos, o fluxo de salvar e carregar um modelo também estão presentes nessa parte da ferramenta.

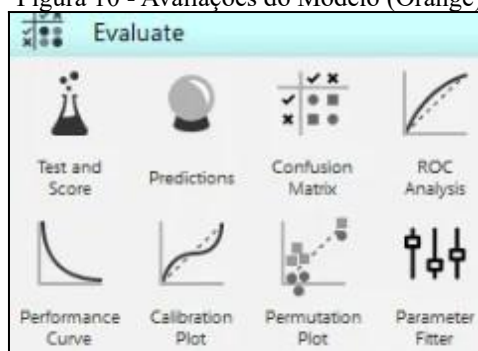
Figura 9 - Modelos Supervisionados (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

Nas opções de avaliações do modelo, temos algumas métricas comuns, como a curva ROC, matriz de confusão, e também a possibilidade de realizar as predições.

Figura 10 - Avaliações do Modelo (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

Por fim, a ferramenta também disponibiliza a utilização de algoritmos não supervisionados, como k-Means e PCA. Na imagem abaixo estão listadas todas as opções.

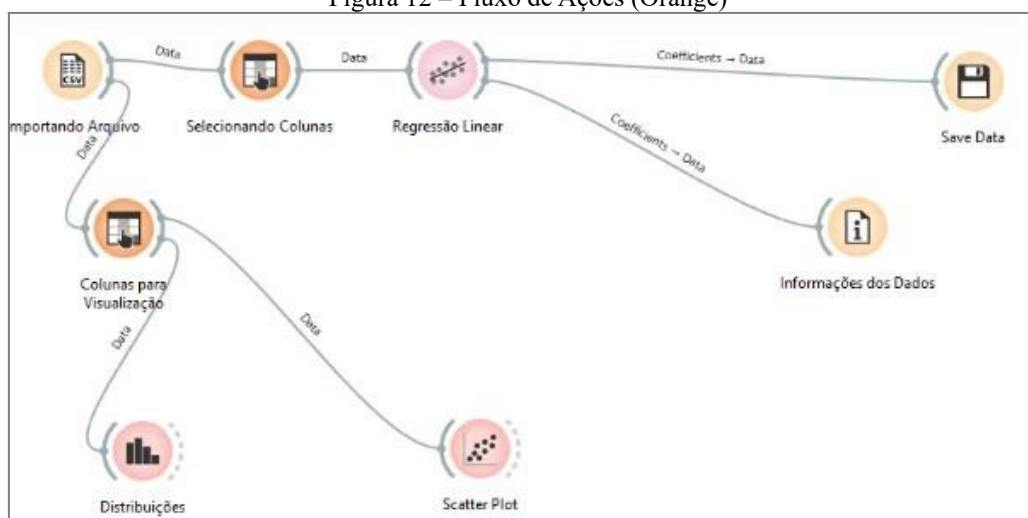
Figura 11 – Modelos de Algoritmos não-supervisionados



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

Com essas opções, é possível montar um fluxo de ações tanto para analisar os dados quanto criar modelos. Abaixo é possível ver esse fluxo acontecendo:

Figura 12 – Fluxo de Ações (Orange)



Fonte: (BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA, [s.d.])

Inicialmente, os dados são importados. A partir desses dados, são selecionadas colunas para a visualização dos dados e também para a criação de um algoritmo de regressão linear.

4.2.4.1 Pontos Positivos e Limitações da ferramenta

Pontos Positivos:

- Gratuito e open source;
- Interface 100% visual (arrastar e soltar);
- Ótima ferramenta para ensino e aprendizado;
- Boa variedade de algoritmos, além da possibilidade de integração com Python.

Limitações:

- Não é viável para produção (não possui suporte para deployment);
- Limitada para projetos complexos ou com grandes volumes de dados;
- Menor performance comparado a ferramentas mais robustas.

4.2.4.2 Conclusão sobre a Ferramenta

Orange é uma boa opção para aprender, prototipar e visualizar modelos de ML mas não é indicada para uso em produção ou pipelines complexos.

4.2.5 Google Data Assistant

O Google Data Assistant é uma ferramenta totalmente gratuita e pode ser utilizada através de um notebook no Google Colab. Dentro do notebook, basta carregar o conjunto de dados e escrever um prompt do que deseja que seja feito, como:

- “Visualize tendências de vendas”;
- “Treine um modelo de Regressão Linear”.

Após isso, o Data Science Agent, atualmente apoiado pelo modelo Gemini, insere no notebook todo o código necessário, desde a importação de bibliotecas, limpeza de dados, visualizações e execução de algoritmos, até mesmo, corrige os erros que acontecem em tempo real.

Por fim, o notebook completo é entregue ao usuário, podendo personalizar os trechos de código e avaliar o código feito 100% por Inteligência Artificial.

Durante os testes junto a ferramenta, foi possível notar que algumas técnicas utilizadas para a criação dos modelos podem não ser as melhores possíveis, para isso, faz-se necessário ser claro no prompt, quanto ao que se espera do código que será criado, uma vez que algumas decisões tomadas pela inteligência artificial podem não ser tão eficazes, além de lentas em alguns casos.

4.2.5.1 Pontos Positivos e Limitações da ferramenta

Pontos Positivos:

- É uma grande aliada na economia de tempo, permitindo ao usuário focar apenas na interpretação do código e em possíveis melhorias, não necessitando começar um código do zero;
- O Google Colab permite o compartilhamento de notebooks, o que significa que toda a equipe pode ter acesso a esses notebooks executáveis de forma online;
- Permite ajustes manuais diretamente no código, fornecendo algo que muitas vezes falta nas ferramentas, a flexibilidade.

Limitações:

- Algumas vezes, o algoritmo sofre de “erros sequenciais”, ficando nesse loop por um bom tempo até achar uma solução que conserte o problema;
- Só funciona dentro do Google Colab, porém, o notebook ou seu código pode ser importado para outras ferramentas;
- No seu estado atual, o algoritmo tem uma certa limitação com análises estatísticas mais complexas;
- Como os dados são treinados em uma ferramenta do Google, não em um ambiente totalmente local, existe uma preocupação quanto a dados sensíveis.

4.2.5.2 Conclusão sobre a Ferramenta

O Google Data Assistant é uma ótima ferramenta para agilizar processos manuais, porém, necessita do conhecimento tanto estatístico quanto de algoritmos de Machine Learning para que o seu uso seja de fato efetivo.

4.3 RESUMO COMPARATIVO DAS FERRAMENTAS ANALISADAS

Tabela 1 – Estudo Comparativo

Nome	Site Oficial	Objetivo	Software	Licença
BigML	https://bigml.com/	Aprendizado de máquina simplificado para todos	Plataforma Web	Comercial; Versão Trial limitada.
Azure Machine Learning	https://azure.microsoft.com/pt-br/products/machine-learning	Usar um serviço de IA de nível empresarial para o ciclo de vida de ML (aprendizado de máquina) de ponta a ponta	Plataforma Web	Comercial, Versão Trial limitada.
DataRobot	https://www.datarobot	Plataforma de Ciência de	Plataforma	Comercial;

	com/	Dados e Aprendizado de Máquina	Web	Versão Demo limitada.
Orange	https://orangedatamining.com/	Aprendizado de máquina e visualização de dados de código aberto.	Aplicação Desktop	Software gratuito
Google Data Assistant	https://colab.research.google.com/	Assistente de dados com tecnologia de IA	Plataforma Web diretamente no Google Colab	Software gratuito

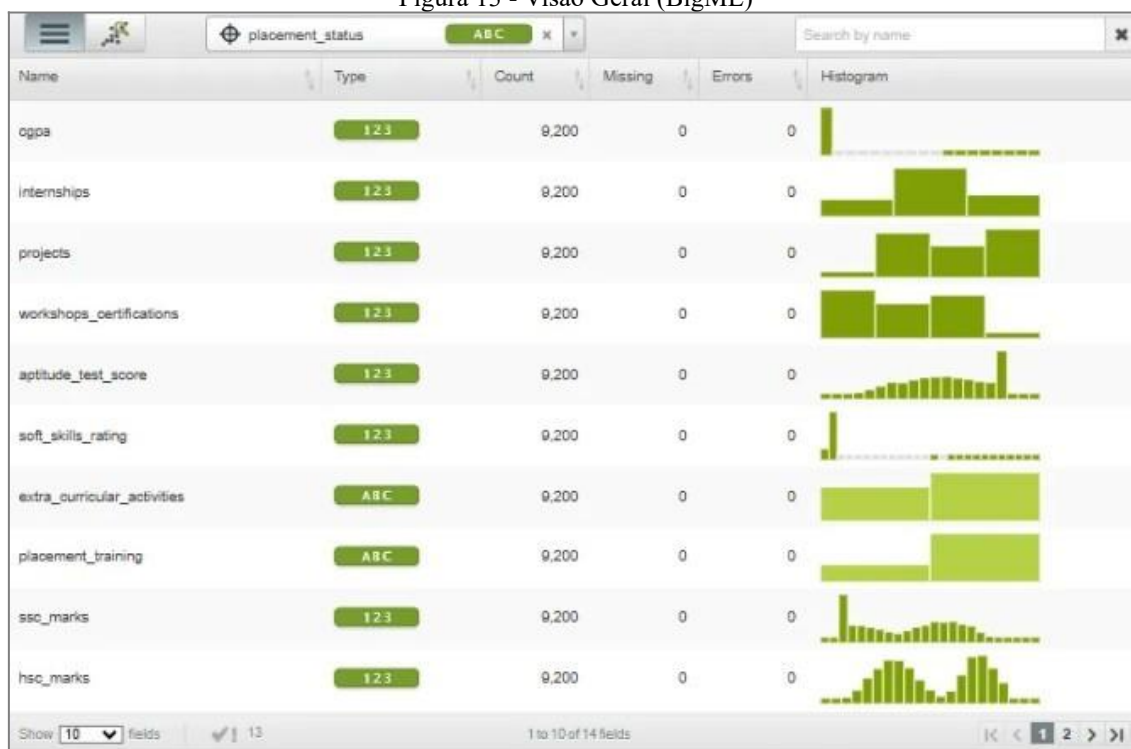
Fonte: autores

5 ESTUDO DE CASO

O BigML foi escolhido para ser explorado com dados fictícios, já que oferece uma versão gratuita limitada mais acessível em relação ao DataRobot e Azure Machine Learning, contando com uma abstração a respeito da criação de modelos de Machine Learning, que o Google Data Assistant não proporciona, além da escalabilidade da aplicação, que não é visto no Orange. Nesta etapa o fluxo da criação de um modelo no BigML é explorado.

Inicialmente é carregado o arquivo dentro da ferramenta, após isso, é necessário transformar esse arquivo em um dataset. Nessa transformação, além do próprio BigML classificar automaticamente a variável alvo, ele também mostra uma representação visual dos dados, mostrando seu tipo, quantidade de linhas, valores faltantes, erros e a distribuição dos dados representados por um histograma simples:

Figura 13 - Visão Geral (BigML)



Fonte: (Bigml.com, [s.d.])

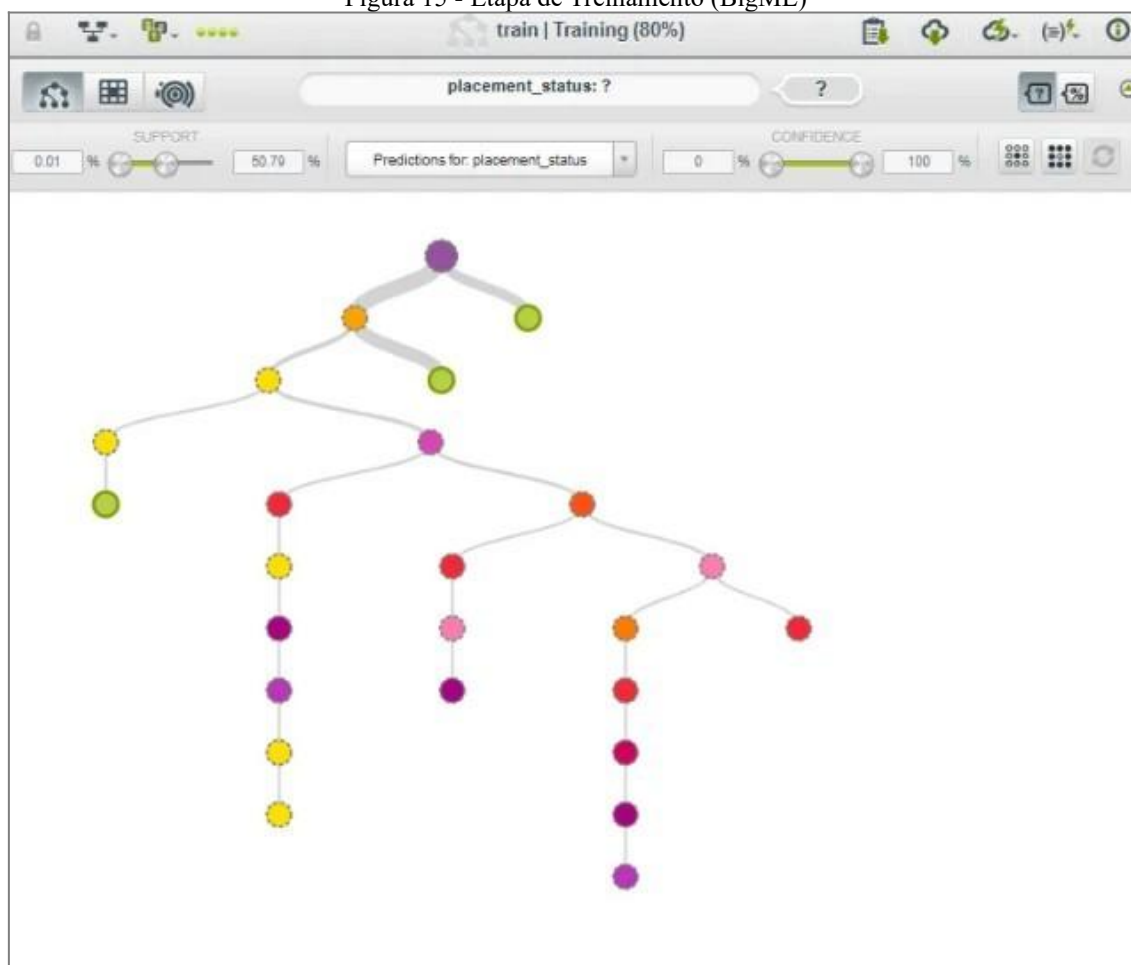
Após essa etapa, é possível escolher entre os algoritmos supervisionados e não supervisionados. Também é possível realizar o split dos dados se necessário. Para exemplificar, foi inicialmente feito um split dos dados de treino e teste e depois criado um modelo supervisionado a partir a opção “MODEL”. Na sequência, foi criada uma árvore de decisão.

Figura 14 - Seleção de Algoritmos (BigML)



Fonte: (Bigml.com, [s.d.])

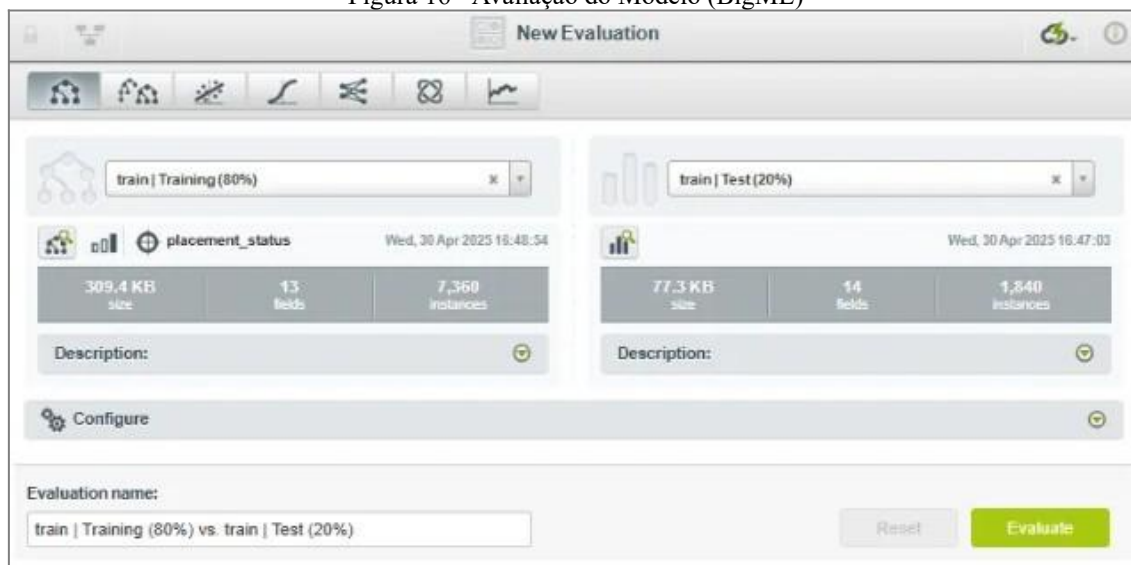
Figura 15 - Etapa de Treinamento (BigML)



Fonte: (Bigml.com, [s.d.])

Com o algoritmo criado, é possível realizar uma avaliação do modelo com os dados de teste separados anteriormente.

Figura 16 - Avaliação do Modelo (BigML)



Fonte: (Bigml.com, [s.d.])

Após a avaliação, é fornecido as métricas de desempenho do modelo, como a acurácia, precisão, recall, curva ROC, e curva Precision-Recall.

Figura 17 - Métricas de Desempenho do Modelo (BigML)



Fonte: (Bigml.com, [s.d.])

Figura 18 - Métricas de Desempenho do Modelo - Parte 2 (BigML)



Fonte: (Bigml.com, [s.d.])

Figura 19 - Métricas de Desempenho do Modelo - Parte 3 (BigML)



Fonte: (Bigml.com, [s.d.])

6 CONSIDERAÇÕES FINAIS

As avaliações realizadas nas cinco ferramentas propostas evidenciou que cada uma delas caracteriza-se por um perfil distinto, tanto em termos de capacidade, customização, acessibilidade, escalabilidade e custos.

A escolha de cada ferramenta pode variar a partir das necessidades do projeto em questão. Enquanto a Orange, por exemplo, tem um grande valor na prototipação ou meios acadêmicos, o BigML realiza o trabalho de um Cientista de Dados abstraindo o usuário da complexidade dos modelos, assim como o DataRobot.

Já o Azure Machine Learning também oferece muito do que promete o BigML e DataRobot, por exemplo, porém, pode exigir mais do usuário no momento de sua implementação. Por fim, o Google Data Assistant oferece uma proposta completamente diferente das demais opções, criando algoritmos para o que lhe for pedido, porém, não trabalhando diretamente com a parte de implementação do modelo.

Devido a fácil usabilidade e acessibilidade do BigML, tanto para cadastro quanto utilização da ferramenta, mesmo que de forma limitada, a ferramenta foi escolhida para o estudo de caso em questão, como forma de explorar as suas principais funcionalidades e entender como a ferramenta pode ser útil para projetos que não contam com uma equipe com conhecimento avançado em aplicações de algoritmos de Machine Learning.

Também foi possível de observar durante este estudo, que ferramentas com esse propósito podem abstrair até certo nível a complexidade da aplicação de algoritmos de Machine Learning, porém, não substituem um profissional qualificado que entenda as limitações das ferramentas e saiba interpretar os resultados que as mesmas produzem.

Por fim, destaca-se que a constante evolução das tecnologias de Inteligência Artificial exige avaliações periódicas das ferramentas disponíveis, visto que atualizações, mudanças nos modelos de licenciamento e novas funcionalidades podem alterar significativamente sua aplicabilidade. Assim, estudos comparativos como este não apenas subsidiam decisões mais informadas, como também contribuem para o desenvolvimento de estratégias alinhadas às inovações tecnológicas e às necessidades reais de formação de novos profissionais na área de ciência de dados.

REFERÊNCIAS

- AZURE MACHINE LEARNING. Disponível em: <https://azure.microsoft.com/pt-br/products/machine-learning>. Acesso em: 10 jun. 2025.
- BIANCHI, R. A. C.; COSTA, A. H. R. Uso de heurísticas para a aceleração do aprendizado por reforço. [S.l.: s.n.], 2004.
- BIGML. Disponível em: <https://bigml.com/>. Acesso em: 10 jun. 2025.
- BIOINFORMATICS LABORATORY; UNIVERSITY OF LJUBLJANA. Orange data mining. Disponível em: <https://orangedatamining.com/>. Acesso em: 10 jun. 2025.
- CASTRO, C. L.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. [S.l.: s.n.], 2011.
- CIO. O que é análise de dados? Análise e gerenciamento de dados para decisões. IT Fórum, 9 jun. 2022. Disponível em: <https://itforum.com.br/o-que-e-analise-de-dados-analise-e-gerenciamento-de-dados-para-decisoes/>. Acesso em: 10 jun. 2025.
- CORTEX, I. O que é análise de dados? Saiba tudo sobre o tema! Cortex, 2022. Disponível em: <https://www.cortex-intelligence.com/blog/inteligencia-de-mercado/o-que-e-analise-de-dados>. Acesso em: 10 jun. 2025.
- DATAROBOT. Disponível em: <https://www.datarobot.com/>. Acesso em: 10 jun. 2025.
- GALLEGO, C. F.; CRUZ, J. V. da; ALVES, T. S. Ferramentas para análise de dados. [S.l.: s.n.], 2023.
- GÉRON, A. Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow. Rio de Janeiro: Alta Books, 2021.
- GOMES, P. C. T. Matriz de confusão: diagnóstico de doença (cores ajustadas). Datageeks, 19 out. 2024. Disponível em: <https://www.datageeks.com.br/matriz-de-confusao/>. Acesso em: 10 jun. 2025.
- GOOGLE COLAB. Disponível em: <https://colab.research.google.com/>. Acesso em: 10 jun. 2025.
- JABBAR, H. K.; KHAN, D. R. R. Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). In: STEPHEN, J.; VASAVI, H. R. A. (org.). [S.l.: s.n.], 2015.
- MATSUBARA, E. T. O algoritmo de aprendizado semi-supervisionado co-training e sua aplicação na rotulação de documentos. [S.l.: s.n.], 2004.
- MIRANDA, B. A.; CAMPELO, C. E. C. Quão eficaz é uma ferramenta de automação de análise de dados baseada em LLM? Um estudo de caso com o Data Analyst do ChatGPT. Campina Grande: Universidade Federal de Campina Grande, 2024.

OLIVEIRA, Gabriel P. et al. Ferramentas open-source de qualidade de dados para licitações públicas: uma análise comparativa. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBB), 37., 2022, Búzios. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 116-127. ISSN 2763-8979. DOI: <https://doi.org/10.5753/sbbd.2022.224351>.

PAIXÃO, G. M. M.; SANTOS, B. C.; ARAÚJO, R. M. de; RIBEIRO, M. H.; MORAES, J. L. de; RIBEIRO, A. L. Machine learning na medicina: revisão e aplicabilidade. [S.l.: s.n.], 2020.

PEDROSA, I.; LOUSA, A.; BERNARDINO, J. Evaluation and analysis of business intelligence data visualization tools. [S.l.: s.n.], 2019.

ROSA, C. S.; BERTARINI, D. P. L. Estudo sobre as técnicas e métodos de análise de dados no contexto de Big Data. Uberlândia: Universidade Federal de Uberlândia, 2018.

SILVA, D. M. A.; SOARES, P. L. B. Aplicação de aprendizagem não supervisionada no problema da diversidade máxima em grafos. Fortaleza: Universidade Federal do Ceará, 2023.