Refas
11 anos
Revista Fatec Zona Sul
1ª. Edição Outubro 2014
ISSN 2359-182X

ÉTICA E RESPONSABILIDADE NO USO DAS TÉCNICAS DE MINERAÇÃO DE DADOS

ETHICS AND RESPONSIBILITY IN THE USE OF DATA MINING TECHNIQUES

ÉTICA Y RESPONSABILIDAD EN EL USO DE TÉCNICAS DE MINERÍA DE DATOS

Kevin Silva Estrela[1]
Isabelle Vicente Oliveira[2]
Adriana Marroni Zaniol Palombo[3]

Artigo recebido em agosto de 2024
Artigo aceito em junho de 2025

DOI: 10.26853/Refas_ISSN-2359-182X_v12n01_01

RESUMO

O presente artigo objetiva debater o uso e o tratamento de dados coletados através de técnicas de mineração. A criação de um projeto de análise descritiva e preditiva que usará dados coletados da região de Carapicuíba, demonstra as etapas do processo de mineração, desde a coleta, o tratamento, a interpretação e a avaliação dos resultados, e por fim, a ética no uso dessas informações. As bases de dados coletadas para o artigo foram retiradas da plataforma Data MPE Brasil, um serviço de Produção e Disseminação de informações do Sebrae. As bases de dados referem-se ao crescimento de empresas divididos por setores na região de Carapicuíba entre 2014 e 2024. A técnica de mineração escolhida para o projeto foi a criação de um modelo de Machine Learning de Regressão Linear. O resultado da Regressão Linear foi a previsão, para determinar o possível crescimento dos setores até a próxima década. A pesquisa determinou o crescimento e a diminuição para diversos setores da região. Ao final, foi constatado a necessidade de um tratamento adequado e responsável aos dados para uso em análises estatísticas, garantindo que os resultados sejam éticos, precisos e imparciais.

Palavras-chave: Mineração de Dados; Ética; Tecnologia.

ABSTRACT

This article aims to discuss the use and treatment of data collected through mining techniques. The creation of a descriptive and predictive analysis project that will use data collected from the Carapicuíba region, demonstrates the stages of the mining process, from collection, treatment, interpretation and evaluation of final results, and finally, ethics in use of this information. The databases collected for the

[1] Técnico em Desenvolvimento de Sistemas e atualmente graduando em Análise e Desenvolvimento de Sistemas. E-mail: kevin.estrela@hotmail.com. Lattes: http://lattes.cnpq.br/0699327665384077. OrcId: 0009-0008-9495-3633.
[2] Técnica em Desenvolvimento de Sistemas e atualmente graduanda em Análise e Desenvolvimento de Sistemas. E-mail: isabelleoliveira767@gmail.com. Lattes: http://lattes.cnpq.br/6104444838699017. OrcId: 0009-0006-2420-4181.
[3] Doutora em Linguística Aplicada e Professora de Língua Inglesa da FATEC Carapicuíba. E-mail: a.marroni71@gmail.com. Lattes: http://lattes.cnpq.br/0081924780798399. OrcId: 0009-0004-0305-3236.

article were taken from the Data MPE Brasil platform, a Sebrae information production and dissemination service. The databases refer to the growth of companies divided by sectors in the Carapicuíba region between 2014 and 2024. The mining technique chosen for the project was the creation of a Linear Regression Machine Learning model. The result of the Linear Regression was the forecast, to determine the possible growth of the sectors until the next decade. The research determined growth and decline for several sectors in the region. In the end, it was verified the need for adequate and responsible treatment of data for use in statistical analyses, ensuring that the results are ethical, accurate and impartial.

Keywords: Data Mining; Ethic; Technology.

RESUMEN

Este artículo tiene como objetivo discutir el uso y tratamiento de los datos recopilados mediante técnicas de minería. La creación de un proyecto de análisis descriptivo y predictivo que utilizará datos recolectados en la región de Carapicuíba, demuestra las etapas del proceso minero, desde la recolección, tratamiento, interpretación y evaluación de los resultados finales, y finalmente, la ética en el uso de esta información. Las bases de datos recopiladas para el artículo fueron tomadas de la plataforma Data MPE Brasil, servicio de producción y difusión de información del Sebrae. Las bases de datos se refieren al crecimiento de las empresas divididas por sectores en la región de Carapicuíba entre 2014 y 2024. La técnica de minería elegida para el proyecto fue la creación de un modelo de Machine Learning de Regresión Lineal. Se proyectó el resultado de la Regresión Lineal, para determinar el posible crecimiento de los sectores hasta la próxima década. La investigación determinó crecimiento y caída de varios sectores de la región. Al final, se constató la necesidad de un tratamiento adecuado y responsable de los datos para su uso en los análisis estadísticos, asegurando que los resultados sean éticos, precisos e imparciales.

Palabras clave: Minería de datos; Principio moral; Tecnología.

1 INTRODUÇÃO

No contexto contemporâneo, a quantidade de dados gerados diariamente cresce de forma exponencial. Desta maneira, a capacidade de extrair informações importantes de conjuntos massivos de dados tornou-se muito valiosa. Ao longo das décadas foram criados diversos métodos para analisar e interpretar esses conjuntos, obtendo assim conhecimento extremamente importante para diversos setores poderem se desenvolver. A fim de suprir a necessidade de obter esse conhecimento, foi criada a chamada Mineração de dados.

Também chamada de *Data Mining*, a Mineração de Dados consiste em técnicas de extração de informações em conjuntos de dados brutos. Tais técnicas permitem encontrar informações úteis como padrões e previsões nesses conjuntos. Empresas, governos e instituições de pesquisa utilizam-no para tomar decisões mais inteligentes, otimizar processos e identificar tendências.

A Mineração se baseia em extrair dados brutos de alguma fonte, como bancos de dados locais, bases públicas na internet, bibliotecas virtuais, entre outras. A forma de trabalhar com essas bases se trata de organizar, limpar, transformar e interpretar os dados, para assim obter os resultados desejados.

Essas técnicas se tornaram muito populares e comuns, resultando em seu uso em praticamente todos os lugares, como na medicina para diagnosticar doenças, em marketing para

recomendar produtos que se encaixam com os perfis dos clientes, no mercado financeiro para prever o possível crescimento de ações, em bancos para detectar fraudes nas contas, entre diversas outras aplicações.

À medida que a mineração de dados se torna mais comum, surgem questões éticas que requerem atenção. A necessidade de transparência quanto ao uso e o tratamento que esses dados irão receber é de extrema importância. Ao tratar de dados pessoais, é essencial garantir que os resultados da mineração sejam imparciais e precisos, e que não demonstrem uma situação irreal.

Nesse sentido, surge uma questão importante a ser discutida, que é: Como garantir que os algoritmos de mineração de dados sejam transparentes, auditáveis e responsáveis pelos seus resultados? Como evitar preconceitos e discriminação algorítmica?

Dentro desse contexto, o artigo busca analisar o cenário da utilização de dados de maneira prática, e verificar quais medidas precisam ser tomadas para garantir a ética e a integridade dos resultados.

Para analisar o uso prático da Mineração de Dados, foi elaborado um projeto de programação que realiza o Processo KDD (*knowledge discovery in database*). Este processo é utilizado de forma a organizar a pesquisa e a exploração de informações de um projeto por meio de etapas. O projeto está dividido em cada uma das etapas do KDD e demonstra a eficiência da mineração de dados no mundo real.

O objetivo do projeto é verificar o crescimento de estabelecimentos divididos por setor econômico na região de Carapicuíba nos últimos anos, e prever o possível crescimento deles até a próxima década. As bases de dados são retiradas da plataforma Data MPE Brasil, um serviço de disseminação e produção de informações do Sebrae, que por sua vez agrega os dados disponibilizados pela Receita Federal (RF). As análises de dados utilizadas são a Descritiva, para observar a quantidade de estabelecimentos, e a Preditiva, para prever o possível crescimento por meio de um algoritmo de Regressão Linear em Machine Learning.

Ao final, é realizada a análise dos resultados, e debatido como as medidas tomadas no projeto influenciaram na conclusão, verificando a ética e a responsabilidade das informações.

2 REFERENCIAL TEÓRICO

Neste tópico serão apresentados os principais conceitos referentes a mineração, a técnica de regressão linear e a ética da coleta e o armazenamento de dados.

2.1 Mineração de Dados

Na era atual, a expressão "dados são o novo petróleo" tem se tornado cada vez mais presente na linguagem corporativa e no discurso público sobre tecnologia e inovação. Em um mundo movido por informações, a quantidade de dados gerados diariamente é incalculável. No entanto, assim como o petróleo, os dados brutos por si têm pouco valor, é a capacidade de extrair, refinar e utilizar esses dados de maneira eficaz que realmente os transforma em ativos valiosos. É aqui que entra a mineração de dados.

Segundo TOTVS (2022), a mineração de dados ou data mining é um processo de exploração de grandes conjuntos de dados com o objetivo de descobrir padrões, tendências e informações valiosas que existem nos dados. A mineração de dados é uma área multidisciplinar

que utiliza estatísticas, algoritmos, inteligência artificial e machine learning e ao contrário do que o nome sugere, a mineração de dados não extrai novos dados, mas sim, se aprofunda nas informações já existentes de forma a descobrir padrões e tendências e através dessas descobertas, empresas de todos os tipos conseguir resolver problemas, analisar o impacto futuro das decisões de negócios e aumentar suas margens de lucro.

2.1.1 Processo KDD

Quando se menciona a mineração de dados, é crucial considerar que a mineração é apenas uma parte de um todo. Essa prática faz parte do processo da Extração de conhecimento conhecida como KDD (Knowledge Discovery in Databases). O KDD assim como os seus processos objetivam gerar conhecimento.

De acordo com Devmidia (2017), O KDD é composto por cinco fases:

a) Processamento (Preprocessing): No processamento os dados passam por uma revisão para garantir qualidade. Isso inclui limpeza, correção ou remoção de informações inconsistentes, identificação de dados ausentes ou incompletos, e detecção de anomalias (outliers);
b) Transformação (Transformation): Na etapa de transformação, técnicas como normalização, agregação, criação de novas variáveis e redução de dados são aplicadas. Isso permite que os dados estejam prontos para serem usados em modelos analíticos;
c) Mineração de Dados (Data Mining):Na etapa, modelos são construídos ou técnicas de mineração de dados são aplicadas para atingir diferentes objetivos, como verificar hipóteses ou descobrir padrões de forma autônoma. A mineração pode ser preditiva ou descritiva e envolve interatividade, com modelos sendo refinados conforme necessário;
d) Interpretação e Avaliação (Interpretation / Evaluation): Nesta fase, o desempenho do modelo é avaliado, geralmente usando dados não utilizados durante o treinamento ou mineração. A validação é feita por meio de medidas estatísticas ou pela revisão de especialistas de negócio.

2.2 Análise de dados

De acordo com Zendesk (2023), Análise de dados é o processo de aplicação de técnicas estatísticas e lógicas para avaliar informações obtidas a partir de determinados processos. O principal objetivo da prática é extrair informações úteis a partir dos dados. A partir destas informações, é possível tomar decisões mais assertivas e orientadas para resultados.

As duas técnicas usadas no projeto:

a) Análise descritiva: Como o próprio nome diz, a análise descritiva é um dos tipos de análise de dados baseado em fatos. Isso significa que, na prática, este tipo de avaliação de dados é feito a partir de resultados obtidos. São exemplos de análise de dados descritiva:
    a) relatórios;
    b) segmentação e controle de clientes;
    c) análises de negócio;
    d) aplicação de métricas;
    e) avaliação de resultados.

Um dos principais usos para a análise descritiva é orientar a construção de estratégias.

b) Análise preditiva: O mais popular dos tipos de análise de dados é justamente o modelo preditivo. Como o nome diz, sua essência está na previsão de cenários futuros com base na análise de padrões revelados pela base de dados.

2.2.1 Regressão linear

Segundo AWS (2023), A regressão linear é uma técnica de análise de dados que prevê o valor de dados desconhecidos usando outro valor de dados relacionado e conhecido. Ele modela matematicamente a variável desconhecida ou dependente e a variável conhecida ou independente como uma equação linear.

Os modelos de regressão linear são relativamente simples e fornecem uma fórmula matemática fácil de interpretar para gerar previsões. A regressão linear é uma técnica estatística estabelecida e se aplica facilmente a softwares e à computação. As empresas a utilizam para converter dados brutos de forma confiável e previsível em *Business Intelligence* e insights práticos. Muitos métodos de ciência de dados, como machine learning e inteligência artificial, usam a regressão linear para resolver problemas complexos.

Em sua essência, uma técnica de regressão linear simples tenta traçar um gráfico de linhas entre duas variáveis de dados, x e y. Como variável independente, x é plotada ao longo do eixo horizontal. Variáveis independentes também são chamadas de variáveis explicativas ou variáveis preditoras. A variável dependente, y, é plotada no eixo vertical.

No machine learning, programas de computador, chamados de algoritmos, analisam grandes conjuntos de dados e trabalham regressivamente a partir desses dados para calcular a equação de regressão linear. Os cientistas de dados primeiro treinam o algoritmo em conjuntos de dados conhecidos ou rotulados e depois o utilizam para prever valores desconhecidos.

2.3 Métricas de avaliação

Uma parte importante ao criar modelos de Machine Learning é avaliar a qualidade dos resultados em relação a tarefa atribuída.

Mario (2018), lista as métricas utilizadas para classificação em Machine Learning:

a) Acurácia (Accuracy/Taxa de Acerto): É o número de acertos (positivos) divido pelo número total de exemplos. Ela deve ser usada em dados com a mesma proporção de exemplos para cada classe, e quando as penalidades de acerto e erro para cada classe forem as mesmas. Em problemas com classes desproporcionais, ela causa uma falsa impressão de bom desempenho. Por exemplo, num dataset em que 80% dos exemplos pertençam a uma classe, só de classificar todos os exemplos naquela classe como positivos já se atinge uma precisão de 80%, mesmo que todos os exemplos da outra classe estejam classificados incorretamente.
b) Precisão (Precision): Número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe (positivos verdadeiros), dividido pela soma entre este número, e o número de exemplos classificados nesta classe, mas que pertencem a outras (falsos positivos);
c) Recall: Número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe, dividido pela quantidade total de exemplos que

pertencem a esta classe, mesmo que sejam classificados em outra. No caso binário, positivos verdadeiros divididos por total de positivos;

d) F1 Score: O F1 Score é uma média harmônica entre precisão e recall. Ela é muito boa quando você possui um dataset com classes desproporcionais. Em geral, quanto maior o F1 score, melhor.

Se tratando especificamente de modelos de regressão, são usadas métricas diferentes para a avaliação.

De acordo com Junior (2021), as métricas usadas em modelos de regressão são:

a) R^2 (R-squared): A métrica R^2, também conhecida como R-dois ou coeficiente de determinação, representa o percentual da variância dos dados que é explicado pelo modelo. Os resultados variam de 0 a 1, geralmente também são expressos em termos percentuais, ou seja, variando entre 0% e 100%. Quanto maior é o valor de R^2, mais explicativo é o modelo em relação aos dados previstos;

b) Erro Médio Absoluto: O erro médio absoluto (MAE — do inglês Mean Absolute Error), mede a média da diferença entre o valor real com o predito. Mas por haver valores positivos e negativos, é adicionado um módulo entre a diferença dos valores. Além disso, está métrica não é afetada por valores discrepantes — os denominados outliers;

c) Erro Quadrático Médio: O erro quadrático médio (MSE — do inglês Mean Squared Error) é uma métrica que calcula a média de diferença entre o valor predito com o real, como a métrica MAE. Entretanto, ao invés de usar o módulo do resultado entre o valor de y e ŷ, nesta métrica a diferença é elevada ao quadrado. Desta maneira penalizando valores que sejam muito diferentes entre o previsto e o real. Portanto, quanto maior é o valor de MSE, significa que o modelo não performou bem em relação as previsões;

d) Raiz do Erro Quadrático Médio: A raiz do erro quadrático médio (RMSE — do inglês, Root Mean Squared Error) é basicamente o mesmo cálculo de MSE, contendo ainda a mesma ideia de penalização entre diferenças grandes do valor previsto e o real. Porém, para lidar com o problema da diferença entre unidades, é aplicada a raiz quadrática como demonstrado na equação 6. Assim a unidade fica na mesma escala que o dado original, resultando em uma melhor interpretabilidade do resultado da métrica;

e) Erro Percentual Absoluto Médio: O erro percentual absoluto médio (MAPE — do inglês Mean Absolute Percentual Error) é uma métrica que mostra a porcentagem de erro em relação aos valores reais. Então se o resultado de MAPE for igual a 40% significa que o modelo faz previsões que em média a diferença entre o valor previsto e o real equivale a 40% do valor real tanto para mais quanto para menos.

2.4 Ética na mineração

A mineração de dados oferece uma variedade de oportunidades para o mundo empresarial, desempenhando um papel fundamental tanto na descoberta de informações quanto na tomada de decisão e no impulsionamento de novas tecnologias. No entanto, umas das preocupações relacionadas a esse campo são os desafios éticos enfrentados não só na mineração de dados, mas nos processos que vem antes e depois como coleta, armazenamento e uso de dados, cada uma dessas etapas apresenta suas próprias questões éticas que precisam ser consideradas e gerenciadas.

Segundo Awari (2023), as etapas dos processos de mineração são divididas em:

a) Coleta de Dados: A coleta de dados marca o início de todo o processo de mineração. Embora os dados usados neste artigo sejam de acesso públicos e já estavam prontos para uso, é crucial reconhecer que o cenário pode ser significativamente diferente quando empresas se preparam para realizar sua própria coleta de dados. As empresas recorrem a uma variedade de fontes e métodos para coletar dados, como lojas online, pesquisa de mercado, interação com o cliente e muito mais. É importante que esses métodos de coleta sejam transparentes e éticos, para garantir isso é de extrema importância que as empresas sejam claras sobre como, onde e por que os dados estão sendo coletados também deve haver uma comunicação aberta sobre o uso dos dados, permitindo que os indivíduos entendam o propósito da coleta.
   Além disso, é crucial garantir que os participantes compreendam completamente como seus dados serão utilizados e tenham a oportunidade de dar consentimento informado, para isso o consentimento deve ser obtido de maneira clara e inequívoca, com os participantes cientes de seus direitos e do uso que será feito de seus dados. Deve-se evitar o uso de jargões técnicos ou termos complicados que possam confundir os participantes;
b) Armazenamento: O armazenamento de dados levanta questões e preocupações sobre segurança e privacidade e sobre a necessidade daqueles dados serem armazenados. As empresas são responsáveis por proteger os dados sensíveis de acessos não autorizados e garantir que estejam em conformidade com regulamentações de privacidade de dados. Ou seja, a regra é clara, quanto mais dados as empresas tiverem em seu porte, maior é a responsabilidade da empresa perante aqueles dados. Por isso, é essencial ter uma compreensão clara dos objetivos da mineração de dados e estabelecer metas realistas para evitar a coleta de dados desnecessários. Depois de definir quais dados são realmente necessários, é importante ao lidar com dados sensíveis, utilizar medidas para anonimizar e protege as informações pessoais. Considerado o uso de técnicas de mascaramento, bloqueio ou remoção de identificadores pessoais. Além disso, é importante considerar por quanto tempo os dados devem ser retidos e como devem ser descartados de forma segura quando não forem mais necessários;
c) Uso de dados: Mesmo tomado cuidado com todas as etapas é importante ter em mente que ainda sim os resultados não são imunes a erros de ética e responsabilidade por exemplo: durante a análise de dados, há a possibilidade de o viés e a interpretação seletiva distorcerem os resultados e levar a conclusões errôneas. Por isso é fundamental garantir que os algoritmos e técnicas de análise sejam imparciais e transparentes, e que os resultados sejam interpretados com responsabilidade e contexto adequado;

2.5 Discriminação algorítmica e viés nos modelos de mineração de dados

No texto anterior, foi destacada a questão do viés e das interpretações equivocadas, particularmente no contexto do viés algorítmico, também conhecido como discriminação algorítmica. Este é um tema de considerável relevância e complexidade, sobretudo nas áreas da inteligência artificial e da mineração de dados. A discriminação algorítmica ocorre quando algoritmos ou modelos de mineração de dados produzem resultados discriminatórios direcionados a determinados grupos de pessoas. Esses problemas frequentemente têm suas raízes na coleta de dados, onde bases de dados tendenciosas introduzem preconceitos que refletem os vieses institucionais da sociedade, influenciando assim os resultados de modelos de previsão, como aqueles empregados no sistema de justiça criminal.

Além disso, é importante destacar os impactos do viés nos dados, que se manifestar de diversas formas. Primeiramente, os resultados dos modelos podem ser imprecisos se os dados utilizados estiverem enviesados. Da mesma forma, as interpretações dos resultados tendem a ser enganosas se o viés nos dados não for corretamente considerado. Por fim, e talvez mais preocupante, o viés nos dados pode levar a decisões discriminatórias contra certos grupos de pessoas, perpetuando desigualdades já existentes.

Para mitigar esses problemas, existem diversas técnicas que são empregadas. Isso inclui uma coleta de dados mais cuidadosa, garantindo que os dados sejam representativos e precisos. Além disso, é essencial realizar uma avaliação constante dos modelos, monitorando e ajustando-os para reduzir o viés. Por fim, promover a transparência e explicabilidade dos algoritmos é crucial para identificar e corrigir possíveis discriminações, permitindo uma abordagem mais ética e justa na utilização da inteligência artificial e da mineração de dados.

A transparência e explicabilidade dos algoritmos de inteligência artificial e mineração de dados se referem à capacidade de compreender e explicar o funcionamento dos algoritmos. Isso inclui entender como os dados são processados, quais técnicas de análise são utilizadas. Também é de extrema importância que os usuários e as pessoas envolvidas nesse processo de extração de conhecimento possam compreender e contextualizar as conclusões obtidas pelos algoritmos, não só para ética ser empregada como também para que a tomada de decisões que é um dos principais objetivos da mineração de dados, possa ser realizada de forma adequada.

2.6 Responsabilidade na mineração de dados

Até o presente momento, este artigo tem focado na discussão sobre ética na mineração de dados. No entanto, é fundamental abordar também a responsabilidade associada a essa prática. Para compreender a distinção entre ética e responsabilidade, é essencial reconhecer que a ética fornece um quadro moral para uma determinada escolha, enquanto a responsabilidade se refere às ações específicas que uma pessoa ou entidade deve executar com base em suas obrigações para com os outros. As responsabilidades na mineração de dados são:

a) Responsabilidade dos Pesquisadores e Profissionais de Dados: Os pesquisadores e profissionais de dados têm a responsabilidade de garantir que o processo de gerenciamento de dados que inclui coleta, armazenamento, processamento, manutenção, análise e segurança dos dados seja conduzido de maneira ética e responsável. Isso inclui garantir a integridade dos dados coletados, respeitar a privacidade e os direitos dos participantes e utilizar métodos de análise que sejam justos e imparciais. Além disso, os profissionais de dados devem ser transparentes sobre suas práticas e resultados, promovendo a confiabilidade e a credibilidade em suas pesquisas.
b) Papel das Empresas na Proteção dos Dados do Usuário: Com base na LGPD que é a lei nº 13.709/2018, também conhecida como Lei Geral de Proteção de Dados Pessoais, as empresas que coletam e utilizam dados dos usuários têm uma responsabilidade na proteção dessas informações. Isso inclui garantir a segurança e a privacidade dos dados, implementar medidas de segurança cibernética robustas como políticas de segurança da informação, treinamento para funcionários, criptografia, controle de acesso, monitoramento de acesso, backup e recuperação de dados, auditorias e avaliações de risco e por fim segurança física. Além disso, as empresas devem adotar políticas de transparência e consentimento informado, garantindo que os usuários tenham controle sobre suas informações pessoais.
c) Regulamentações e Padrões Éticos Relevantes: As regulamentações e padrões éticos, como o GDPR e a LGPD, desempenham um papel crucial na promoção da

responsabilidade na mineração de dados. Essas leis estabelecem diretrizes claras para o processo de gestão de dados garantindo que os direitos dos usuários sejam protegidos. Além disso, essas regulamentações incentivam a adoção de práticas éticas e transparentes por parte das empresas e organizações que lidam com dados.

3 MÉTODO

Para demonstrar o uso da mineração de dados, foi elaborado um projeto prático de análise preditiva com dados da região de Carapicuíba. O projeto é um notebook escrito na linguagem de programação Python, que realiza as etapas do processo KDD. As bases de dados usadas foram retiradas do Data MPE Brasil, um serviço de disseminação e produção de informações do Sebrae, que por sua vez utiliza os dados disponibilizados pela Receita Federal (RF).

As tecnologias usadas no projeto são:

a) Google Colab: Plataforma de programação em nuvem.
b) Linguagem de programação Python e as respectivas bibliotecas:
    a) Pandas: Manipulação e tratamento de dados.
    b) Matplotlib: Criação de gráficos 2D e 3D.
    c) Seaborn: Criação de gráficos 2D.
    d) Scikit-learn: Métodos de Machine Learning.

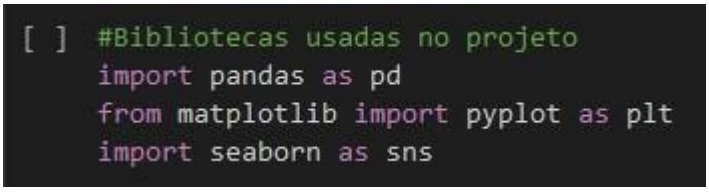
A divisão das bases de dados se dar por:

a) Quantidade de estabelecimentos em Carapicuíba no ano de 2024 separados por setor.
b) Diversas bases de dados registrando o aumento de estabelecimentos em Carapicuíba no período de 2014 até 2024.

3.1 Seleção

A parte inicial consiste em importar as bibliotecas (Figura 1).

Figura 1 - Importação das bibliotecas Pandas, Matplotlib e Seaborn.



Fonte: Elaborada pelos autores

Após a importação, é criada uma variável para receber os dados do arquivo. Usando o método “read_cvs”, o algoritmo transforma os dados contidos no arquivo que antes estavam no formato *Comma Separeted Values* (CSV) para um DataFrame, uma estrutura que organiza os

valores em linhas e colunas, semelhante a uma planilha. Em seguida, o DataFrame é impresso no terminal do Colab para visualizar os dados, observado na Figura 2.

Figura 2 - DataFrame da distribuição de estabelecimentos por setor em Carapicuíba 2024

```
[ ] distribuicao = pd.read_csv('/content/distribuicao-dos-estabelecimentos-ativos-todos-os-anos-todos-os-estabelecimentos-1.csv')
    distribuicao
```

	Grande Sector ID	Grande Sector	Division ID	Division	Establishments
0	1	Agricultura	A01	Agricultura, Pecuária E Serviços Relacionados	65
1	1	Agricultura	A03	Pesca E Aqüicultura	4
2	2	Indústria	B08	Extração De Minerais Não Metálicos	1
3	2	Indústria	C10	Fabricação De Produtos Alimentícios	577
4	2	Indústria	C11	Fabricação De Bebidas	1
...	...	...	...	...	...
69	4	Serviços	S94	Atividades De Organizações Associativas	793
70	4	Serviços	S95	Reparação E Manutenção De Equipamentos De Info...	974
71	4	Serviços	S96	Outras Atividades De Serviços Pessoais	2667
72	4	Serviços	T97	Serviços Domésticos	434
73	5	Administração pública	O84	Administração Pública, Defesa E Seguridade Social	12

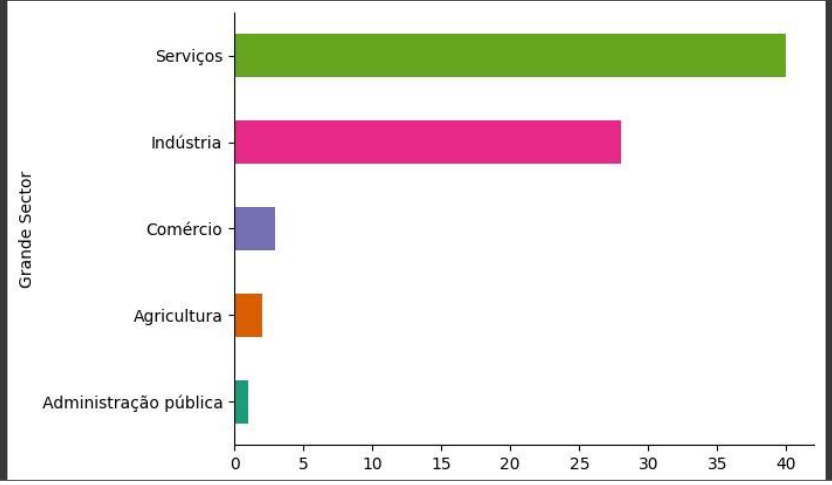
74 rows × 5 columns

Fonte: Elaborada pelos autores

As colunas do DataFrame estão separadas em: ID do Setor, Nome do Setor, ID da Divisão, Divisão do Setor e a quantidade de estabelecimentos. Observando apenas a tabela, não é possível determinar quaisquer informações relevantes sobre a base de dados. Então, para analisar melhor, foi criado um gráfico de barras que ilustra a quantidade de estabelecimentos por setor (Figura 3).

Figura 3 - Gráfico de barras referente a quantidade de estabelecimentos por setor - Carapicuíba 2024

```
[ ] distribuicao.groupby('Grande Sector').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
    plt.gca().spines[['top', 'right',]].set_visible(False)
```



Fonte: Elaborada pelos autores

É possível notar que o maior setor da região de Carapicuíba é o de Serviços, em seguida bem próximo o de Indústria.

Os próximos Datasets informam o crescimento dos estabelecimentos por setor na última década. No Data MPE Brasil eles estão divididos por ano, o de 2024 se refere ao crescimento em comparação a 2023, de 2023 ao de 2022, e assim por diante. Para melhor utilização desses dados foi usado o método 'concat' para concatenar os Datasets em um único Dataframe, como pode ser observado na Figura 4.

Figura 4 - Datasets do crescimento de setores entre 2014 e 2024 em Carapicuíba

```
[ ] aumento_2024 = pd.read_csv('/content/taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2023 = pd.read_csv('/content/2023_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2022 = pd.read_csv('/content/2022_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2021 = pd.read_csv('/content/2021_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2020 = pd.read_csv('/content/2020_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2019 = pd.read_csv('/content/2019_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2018 = pd.read_csv('/content/2018_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2017 = pd.read_csv('/content/2017_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2016 = pd.read_csv('/content/2016_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2015 = pd.read_csv('/content/2015_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')

[ ] aumento_2014 = pd.read_csv('/content/2014_taxa-de-crescimento-dos-estabelecimentos-todos-os-estabelecimentos-1.csv')
```

Fonte: Elaborada pelos autores

A concatenação dos Datasets é feita da seguinte maneira (Figura 5).

Figura 5 - Concatenação dos Datasets

```
[ ] df = pd.concat([aumento_2024, aumento_2023, aumento_2022, aumento_2021, aumento_2020,
                    aumento_2019, aumento_2018, aumento_2017, aumento_2016, aumento_2015, aumento_2014])

    df.reset_index(drop=True, inplace=True)
    df
```

Fonte: Elaborada pelos autores

Agora que os datasets estão juntos em um só, é preciso salvar tudo em um único arquivo para que não seja necessário fazer esse procedimento novamente, demonstrado na Figura 6.

Figura 6 - DataFrame do crescimento de estabelecimentos de 2014 a 2024.

```
[ ] df.to_csv('taxa_de_crescimento_dos_setores_ao_longo_da_decada.csv', index=False)

[ ] taxa_crescimento = pd.read_csv('/content/taxa_de_crescimento_dos_setores_ao_longo_da_decada.csv')
    taxa_crescimento
```

	index	Grande Sector ID	Grande Sector	Section ID	Section	Division ID	Division	Status Year	Establishments	Establishments Growth	Establishments Growth Value	Establishment Prev Year	Share
0	0	1	Agricultura	A	Agricultura, Pecuária, Produção Florestal, Pes...	A01	Agricultura, Pecuária E Serviços Relacionados	2024	4	-0.777778	-14.0	18	0.001591
1	1	2	Indústria	C	Indústrias De Transformação	C10	Fabricação De Produtos Alimentícios	2024	47	-0.587719	-67.0	114	0.018695
2	2	2	Indústria	C	Indústrias De Transformação	C13	Fabricação De Produtos Têxteis	2024	6	-0.785714	-22.0	28	0.002387
3	3	2	Indústria	C	Indústrias De Transformação	C14	Confecção De Artigos Do Vestuário E Acessórios	2024	20	-0.823009	-93.0	113	0.007955
4	4	2	Indústria	C	Indústrias De Transformação	C16	Fabricação De Produtos De Madeira	2024	6	-0.454545	-5.0	11	0.002387
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Fonte: Elaborada pelos autores

3.2 Pré-processamento

Nesta etapa é feita uma filtragem inicial para corrigir nos campos do DataFrame. É possível observar que existem campos com valores ausentes denominados *NaN*. Estes valores tendem a interferir na coleta de informações, o que significa que eles precisam ser modificados. Utilizando uma função de filtragem do Pandas chamada “fillna”, os valores foram substituídos por 0, como é mostrado na Figura 7.

Figura 7 - Filtragem dos valores nulos

```
[ ] #Inplace=True altera o DataFrame original
    taxa_crescimento.fillna(0, inplace=True)
    taxa_crescimento
```

Fonte: Elaborada pelos autores

3.3 Transformação

Nesta etapa é realizada a transformação dos dados de forma a padronizar e organizar as informações. Para saber quais mudanças precisam ser feitas, é necessário observar as informações do DataFrame indicado na Figura 8.

Figura 8 - Informações técnicas do DataFrame

```
[ ] # Método que retorna as características do DataFrame
    taxa_crescimento.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 623 entries, 0 to 622
    Data columns (total 13 columns):
     #   Column                     Non-Null Count  Dtype
    ---  ------                     --------------  -----
     0   index                      623 non-null    int64
     1   Grande Sector ID           623 non-null    int64
     2   Grande Sector              623 non-null    object
     3   Section ID                 623 non-null    object
     4   Section                    623 non-null    object
     5   Division ID                623 non-null    object
     6   Division                   623 non-null    object
     7   Status Year                623 non-null    int64
     8   Establishments             623 non-null    int64
     9   Establishments Growth      623 non-null    float64
     10  Establishments Growth Value  623 non-null  float64
     11  Establishment Prev Year    623 non-null    int64
     12  Share                      623 non-null    float64
    dtypes: float64(3), int64(5), object(5)
    memory usage: 63.4+ KB
```

Fonte: Elaborada pelos autores

Observando o DataFrame, a coluna 'Establishments Growth Value' contém apenas valores inteiros, mas seu tipo está como 'float64'. Converter o tipo dessa coluna deixará o DataFrame mais leve e os dados estarão padronizados (Figura 9).

Figura 9 - Conversão de tipo da coluna do DataFrame

```
[ ] # Astype permite alterar o tipo dos dados
    taxa_crescimento['Establishments Growth Value'] = taxa_crescimento['Establishments Growth Value'].astype(int)
    taxa_crescimento.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 623 entries, 0 to 622
    Data columns (total 13 columns):
     #   Column                      Non-Null Count  Dtype
    ---  ------                      --------------  -----
     0   index                       623 non-null    int64
     1   Grande Sector ID            623 non-null    int64
     2   Grande Sector               623 non-null    object
     3   Section ID                  623 non-null    object
     4   Section                     623 non-null    object
     5   Division ID                 623 non-null    object
     6   Division                    623 non-null    object
     7   Status Year                 623 non-null    int64
     8   Establishments              623 non-null    int64
     9   Establishments Growth       623 non-null    float64
     10  Establishments Growth Value  623 non-null    int64
     11  Establishment Prev Year     623 non-null    int64
     12  Share                       623 non-null    float64
    dtypes: float64(2), int64(6), object(5)
    memory usage: 63.4+ KB
```

Fonte: Elaborada pelos autores

3.4 Mineração de dados

Agora são utilizadas técnicas de Mineração para obter insights a partir do conjunto de dados. A técnica a ser usada será a Regressão Linear para prever o crescimento dos estabelecimentos na próxima década.

3.4.1 Regressão linear

A Regressão Linear é um algoritmo que prevê resultados futuros baseados em dados anteriores. O cálculo feito no projeto utiliza os registros do crescimento dos setores na última década e realiza a previsão para o ano de 2034, demonstrado na Figura 10.

Figura 10 - Importação de bibliotecas, divisão de recursos e treinamento do modelo de Regressão Linear.

```
[ ] # Bibliotecas usadas para Machine Learning
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LinearRegression

[ ] # Dividindo os dados em recursos (X) e alvo (y)
    X = taxa_crescimento[['Status Year', 'Establishments']]
    y = taxa_crescimento['Establishments Growth Value']

[ ] # Dividindo os dados em conjuntos de treinamento e teste
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[ ] # Criando e treinando o modelo
    model = LinearRegression()
    model.fit(X_train, y_train)
```

Fonte: Elaborada pelos autores

Agora que os dados foram gerados, é criado um DataFrame contendo os nomes dos setores, o possível valor de crescimento e o ano da previsão (Figura 11, 12 e 13).

Figura 11 - Função que cria uma coluna com os valores previstos pelo modelo

```
[ ]  # Criando uma função que adiciona uma Coluna com os valores gerados pelo Modelo
     def adicionar_coluna_previsao(taxa_crescimento, model):
         # Fazendo previsões
         predicted_values = model.predict(taxa_crescimento[['Status Year', 'Establishments']])
         # Adicionando os valores previstos ao DataFrame
         taxa_crescimento['Predicted Values'] = predicted_values
         return taxa_crescimento

     # Adicionando os valores previstos ao DataFrame original
     taxa_crescimento = adicionar_coluna_previsao(taxa_crescimento, model)

     # Visualizando o DataFrame com os valores previstos
     taxa_crescimento
```

Fonte: Elaborada pelos autores

Figura 12 - Criação de um novo DataFrame

```
[ ]  # Criando um novo DataFrame com os valores gerados pelo modelo e adicionando o ano
     aumento_decada = taxa_crescimento[['Grande Sector', 'Section', 'Division' ,'Predicted Values']]

     ano = []
     for i in range(taxa_crescimento.shape[0]):
       ano.append('2034')

     aumento_decada['Status Year'] = ano
     aumento_decada
```

Fonte: Elaborada pelos autores

Figura 13 - DataFrame com os valores previstos para a próxima década

	Grande Sector	Section	Division	Predicted Values	Status Year
0	Agricultura	Agricultura, Pecuária, Produção Florestal, Pes...	Agricultura, Pecuária E Serviços Relacionados	-35.846962	2034
1	Indústria	Indústrias De Transformação	Fabricação De Produtos Alimentícios	-28.619160	2034
2	Indústria	Indústrias De Transformação	Fabricação De Produtos Têxteis	-35.510785	2034
3	Indústria	Indústrias De Transformação	Confecção De Artigos Do Vestuário E Acessórios	-33.157547	2034
4	Indústria	Indústrias De Transformação	Fabricação De Produtos De Madeira	-35.510785	2034
...	...	...	...	...	...
618	Serviços	Artes, Cultura, Esporte E Recreação	Atividades Esportivas E De Recreação E Lazer	21.341827	2034
619	Serviços	Outras Atividades De Serviços	Atividades De Organizações Associativas	22.518446	2034
620	Serviços	Outras Atividades De Serviços	Reparação E Manutenção De Equipamentos De Info...	23.358888	2034
621	Serviços	Outras Atividades De Serviços	Outras Atividades De Serviços Pessoais	27.729187	2034
622	Serviços	Serviços Domésticos	Serviços Domésticos	21.005651	2034

623 rows × 5 columns

Fonte: Elaborada pelos autores

Os resultados gerados pelo modelo mostram um grande aumento e/ou diminuição para os setores na próxima década.

3.5 Interpretação e avaliação

Nesta etapa é avaliado o modelo. Os dados previstos têm a possibilidade de não serem precisos e acabar gerando resultados incorretos. A avaliação usará técnicas matemáticas importadas da própria biblioteca de Machine Learning e de uma para cálculos estatísticos avançados (Figura 16). As métricas usadas são:

a) Erro Quadrático Médio (MSE - Mean Squared Error);
b) Raiz do Erro Quadrático Médio (RMSE);
c) Erro Absoluto Médio (MAE - Mean Absolute Error);
d) R^2 (R-squared);
e) Coeficientes de regressão (Coef);
f) Valores P (P-Values).

Figura 16 – Algoritmo de avaliação do modelo de Regressão Linear

```
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import statsmodels.api as sm

# Fazendo previsões no conjunto de teste
y_pred = model.predict(X_test)

# Calculando as métricas de avaliação
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

y_with_const = sm.add_constant(y_test)
model_sm = sm.OLS(y_pred, y_with_const).fit()

print("MSE:", mse)
print("RMSE:", rmse)
print("MAE:", mae)
print("R2: ", r2)
print(f"Coeficientes de regressão: {model.coef_}")
print("Valores p:")
print(model_sm.summary())
```

Fonte: Elaborada pelos autores

Figura 17 – Matriz de valores estatísticos do Modelo

```
MSE: 1775.6495725119967
RMSE: 42.13845716814982
MAE: 25.53895138279311
R2:  0.2709358498087939
Coeficientes de regressão: [-5.73568775  0.16808842]
Valores p:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.344
Model:                            OLS   Adj. R-squared:                  0.339
Method:                 Least Squares   F-statistic:                     64.63
Date:                Sat, 08 Jun 2024   Prob (F-statistic):           6.35e-13
Time:                        18:32:29   Log-Likelihood:                -556.11
No. Observations:                 125   AIC:                             1116.
Df Residuals:                     123   BIC:                             1122.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -3.8695      1.930     -2.005      0.047      -7.690      -0.049
Establishments Growth Value      0.3040      0.038      8.039      0.000       0.229       0.379
==============================================================================
Omnibus:                       27.977   Durbin-Watson:                   2.226
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               47.489
Skew:                           1.022   Prob(JB):                     4.87e-11
Kurtosis:                       5.222   Cond. No.                         52.8
==============================================================================
```

Fonte: Elaborada pelos autores

Como observado na Figura 17, as métricas mostraram que o modelo não está bem ajustado e está gerando resultados imprecisos. O R-squared por exemplo, deu o resultado de 0.27, o que significa que o modelo tem 27% de precisão. Isso pode ocorrer devido a qualidade dos dados do Dataset, os parâmetros usados no treinamento e/ou na previsão. Ainda assim, apesar de sua baixa precisão, os resultados indicados pelo modelo podem ser considerados, ao menos em um cenário mais amplo para analisar o contexto do ambiente.

4 RESULTADOS E DISCUSSÃO

Os resultados obtidos durante da pesquisa e do projeto constataram os seguintes pontos.

4.1 Projeto

A elaboração de um projeto prático resultou em duas vertentes, o resultado do projeto em si, e a análise da ética no contexto da pesquisa.

4.1.1 Resultado do projeto

Os resultados da análise exploratória indicaram que o principal setor econômico da região de Carapicuíba é o de serviços, o que envolve diversas categorias de emprego, como serviços domésticos, atividades de organizações associativas, arte, recriação, e outras atividades pessoais que se enquadram na categoria. Esses dados demonstram que a região tem uma forte

economia ligada a empresas pequenas, desde microempreendedores até associações beneficentes e de caráter humanitário. Outra informação relevante é a de que a região tem grandes setores ligados a indústria. A fama da região por si só está ligada a grandes empresas se estabelecerem próximas de Alphaville, considerada uma área mais nobre e com uma população de maior poder aquisitivo. Tais empresas são de marcas famosas e consolidadas, o que explica a grande influência econômica mesmo existindo em menor quantidade em comparação ao setor de serviços.

4.1.2 Resultado do uso da Mineraçâo de Dados e KDD

O processo KDD demonstrou ser uma ferramenta muito eficiente quanto a organização e manutenção do projeto. Cada etapa resultou em uma melhora significa do algoritmo, o que aumenta as chances de resultados corretos serem obtidos. Entretanto, é notável como ocorreram problemas significativos na previsão do modelo, o que gerou informações não verídicas. São diversos fatores que contribuem para isso ocorrer, desde a seleção dos dados na hora do treinamento, a forma de utilizar as métricas de avaliação, entre outras. O que esse resultado demonstra é que é necessário muito cuidado na hora de realizar análises estatísticas. O modelo de regressão aqui tinha o objetivo de prever o crescimento dos setores até a próxima década, dessa maneira, os gestores da região poderiam tomar medidas com relação a esses resultados. Porém, em vista de sua falta de precisão, levar em consideração tais resultados pode não condizer com a realidade.

4.2 Ética na mineração

As bases de dados usadas no projeto foram coletadas de órgãos especializados que fizeram o tratamento adequado com os dados para não violar nenhum princípio moral e legal. Entretanto, quando se está falando de coleta de informações em larga escala, nem sempre o procedimento correto é realizado. No caso dos dados de empresas em Carapicuíba, as informações não comprometem a integridade pessoal de nenhuma empresa, visto que não foram informados dados pessoais de nenhum tipo. Porém, caso um dado pessoal esteja presente na pesquisa, o mal uso dele por qualquer intenção, pode acarretar problemas tanto para o dono desses dados quanto a quem os está manipulando.

Outra questão é no cuidado ao utilizar modelos de mineração, visto que seus resultados podem não precisos, o que acabaria gerando interpretações incorretas de uma situação. Como mencionado anteriormente, o viés nos dados pode levar a decisões discriminatórias contra certos grupos de pessoas, perpetuando desigualdades já existentes. Isso inclui garantir a integridade dos dados coletados e utilizar métodos de análise que sejam justos e imparciais. Além disso, os profissionais de dados devem ser transparentes sobre suas práticas e resultados.

5 CONSIDERAÇÕES FINAIS

Ao final da presente pesquisa constatou-se que para garantir resultados precisos e imparciais, é necessário tomar muito cuidado na hora de manipular dados. Desde o início do processo de coleta, é preciso se certificar de que não houve discriminação ou qualquer fator que leve a um viés ou a uma tendencia preconceituosa na mineração. O resultado disso pode

prejudicar a sociedade como um todo, e, portanto, é essencial implementar medidas que evitem esses problemas.

O projeto elaborado utilizando o processo KDD trouxe informações significativamente relevantes sobre a região de Carapicuíba, o que traz um viés informativo para a população. A maior preocupação se deve ao uso da regressão linear imprecisa, que não deve ser levado como verdade absoluta para o futuro.

Em geral, foi constatada a necessidade do bom planejamento de projetos de análise de dados para garantir que a ética e a responsabilidade estejam sendo empregadas.

6 REFERÊNCIAS

AWARI. **Ética na mineração de dados: considerações e melhores práticas**, 2023. Disponível em https://awari.com.br/etica-na-mineracao-de-dados-consideracoes-e-melhores-praticas/?utm_source=blog&utm_campaign=projeto+blog&utm_medium=%C3%89tica%20na%20minera%C3%A7%C3%A3o%20de%20dados:%20considera%C3%A7%C3%B5es%20e%20melhores%20pr%C3%A1ticas#:~:text=A%20%C3%A9tica%20na%20minera%C3%A7%C3%A3o%20de%20dados%20exige%20que%20as%20organiza%C3%A7%C3%B5es,de%20medidas%20corretivas%2C%20quando%20necess%C3%A1rio. Acesso em 7 maio. 2024.

AWS. **O que é regressão linear?** 2023. Disponível em https://aws.amazon.com/pt/what-is/linear-regression/. Acesso em 21 maio 2024.

DEVMEDIA. **Descoberta de conhecimento utilizando o processo KDD**, 2017. Disponível em https://www.devmedia.com.br/descoberta-de-conhecimento-utilizando-o-processo-kdd/38709. Acesso em 7 maio. 2024.

MARIO Filho. As Métricas Mais Populares para Avaliar Modelos de Machine Learning, 2018. Disponível em https://mariofilho.com/as-metricas-mais-populares-para-avaliar-modelos-de-machine-learning/#precis%C3%A3o-precision. Acesso em 21 maio 2024.

OLIVEIRA, Clébio de. **Métricas para Regressão: Entendendo as métricas R², MAE, MAPE, MSE e RMSE**. Medium, 2021. Disponível em https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70. Acesso em 21 maio 2024.

RIBEIRO, Débora. **Ética**. Dicionário Online de Português, 2024. Disponível em https://www.dicio.com.br/etica/. Acesso em 21 maio. 2024.

SHASHKO, Daniel. **As 10 principais técnicas de mineração de dados.** Astera, 2024. Disponível em https://www.astera.com/pt/type/blog/top-10-data-mining-techniques/. Acesso em 7 maio. 2024.

TOTVS. **Mineração de dados: o que é, importância e ferramentas**, 2022. Disponível em https://www.totvs.com/blog/negocios/mineracao-de-dados/. Acesso em 7 maio. 2024.

ZENDESK. **4 tipos de análise de dados para criar estratégias certeiras**, 2023. Disponível em https://www.zendesk.com.br/blog/tipos-analise-de-dados/#section-3. Acesso em 21 maio 2024.