DISSERTAÇÃO DE MESTRADO PROFISSIONAL

Análise de integridade de dados e desempenho em cursos online utilizando métodos de aprendizado de máquina

Flávio Garcia Praciano

Brasília, Julho de 2023

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

Análise de integridade de dados e desempenho em cursos online utilizando métodos de aprendizado de máquina

Flávio Garcia Praciano

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Prof., Daniel Alves da Silva, Dr, FT/UnB _____

Prof.ª, Edna Dias Canedo, Dra, FT/UnB _____

Prof., Gilmar dos Santos Marques, Dr, UPIS _____

FICHA CATALOGRÁFICA

PRACIANO, FLÁVIO
Análise de integridade de dados e desempenho em cursos online utilizando métodos de aprendizado de máquina [Distrito Federal] 2023.
xvi, 86 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2023).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica
1. Aprendizagem móvel
2. Mineração de dados
3. Aprendizado de maquina
4. Dados abertos
5. Inteligência artificial
I. ENE/FT/UnB
II. Título (série)
PUBLICAÇÃO: PPEE.MP.058

REFERÊNCIA BIBLIOGRÁFICA

PRACIANO, FLÁVIO (2023). *Análise de integridade de dados e desempenho em cursos online utilizando métodos de aprendizado de máquina* . Dissertação de Mestrado Profissional, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 86 p.

CESSÃO DE DIREITOS

Flávio Garcia Praciano
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

Aos que buscam descobrir soluções mais inteligentes e eficientes com a utilização da tecnologia.

"A ciência sem a religião é manca,
religião sem a ciência é cega".

Albert Einstein

AGRADECIMENTOS

Agradeço primeiramente a Deus, reconhecendo sua generosidade como a fonte primordial da minha vida, posses e existência. Toda glória e honra pertencem a ele, especialmente nestes últimos meses, durante os quais sua orientação paciente, sabedoria, perseverança e, acima de tudo, humildade, me conduziram adiante. A concretização deste trabalho assume uma importância inestimável para meu desenvolvimento profissional. Assim, expresso meus mais sinceros agradecimentos e compartilho um sentimento único de triunfo, superação, realização e dever cumprido, mesmo reconhecendo que há um longo percurso à frente.

Quero externar minha profunda gratidão à Universidade de Brasília (UnB) e ao Programa de Pós-Graduação Profissional em Engenharia Elétrica (PPEE) pela oportunidade concedida.

Minha gratidão também se estende ao Laboratório de Tecnologias da Tomada de Decisão - LATITUDE/UnB.

Agradeço imensamente ao meu orientador, Professor Dr. Daniel Alves da Silva, e ao meu coorientador, Dr. Rafael Timóteo de Sousa Júnior, pela disponibilidade em compartilhar seus conhecimentos, orientações e pelo apoio cordial que me guiaram ao longo de toda essa jornada, sobretudo durante os desafiantes e tumultuados anos de 2020 a 2022.

Aos Professores Dr. Fábio Lúcio Lopes de Mendonça e Dr. Georges Daniel Amvame Nze, estendo minha gratidão pelo acompanhamento, apoio, orientação e preciosos ensinamentos que enriqueceram este trabalho.

À minha mãe, Dalva Lúcia Garcia Souza, que, apesar das adversidades da vida, sempre me guiou com amor, dedicação e carinho, mesmo diante de limitações financeiras, jamais deixou de me apoiar nas escolhas e desafios do mundo contemporâneo.

Ao meu pai, Marconde Praciano Souza (in memoriam), que, mesmo ausente, sempre me orientou a trilhar o caminho justo e a priorizar a dedicação aos estudos.

Aos meus filhos, Mestre Bruno Justino Garcia Praciano e Victor Hugo Justino Garcia Praciano, expresso minha gratidão pelo apoio e incentivo constantes, que evidenciam que tudo é alcançável.

À minha namorada, Laurinda dos Santos Souza, agradeço pela dedicação, pela companhia nos momentos compartilhados e pela compreensão nos momentos mais desafiadores que vivenciei a seu lado.

Também desejo agradecer às minhas irmãs, Adriana Garcia Praciano, Renata Garcia Praciano e Patrícia Garcia Praciano, pelo apoio, paciência e compreensão demonstrados.

Por último, manifesto minha gratidão aos colegas de projeto na Escola Nacional de Administração Pública (Enap)/UnB e na Procuradoria Geral da Fazenda Nacional (PGFN)/UnB, cuja contribuição, de várias formas, enriqueceu minha bagagem de conhecimento e culminou nesta dissertação.

Durante todo o processo de desenvolvimento deste trabalho, pude contar com o apoio fundamental da Enap (TED nº 83/2016 - Integração de Tecnologias e Métodos Aplicados à Prática de Escola Virtual da Administração Pública Federal) e da Procuradoria-Geral da Fazenda Nacional (TED nº 01/2021 - Pesquisas e Inovação Tecnológica Aplicadas às Temáticas da Informação e das Comunicações no Domínio da PGFN).

RESUMO

Este trabalho trata de uma pesquisa voltada para a análise da integridade de dados e do desempenho em cursos online, utilizando métodos de aprendizado de máquina. O objetivo principal é desenvolver uma ferramenta para prever o número de alunos que concluíram o curso e identificar possíveis casos de abandono ou desistência. Para alcançar esse propósito, foram empregados algoritmos de aprendizado de máquina supervisionados, como *Support Vector Machine* (SVM) e Redes Neurais Artificiais (ANNs), possibilitando uma análise detalhada e preditiva dos dados. A abordagem adotada nesta pesquisa foi bibliográfica e qualitativa, explorando informações de cursos online em bases abertas e utilizando técnicas de análise de dados. Os resultados obtidos por meio desses métodos de aprendizado de máquina permitiram a identificação de padrões e tendências nos dados, proporcionando uma compreensão mais profunda da integridade dos registros e do desempenho dos alunos. Com uma visão mais precisa do perfil dos alunos e dos desafios enfrentados, torna-se viável implementar estratégias proativas para aumentar a taxa de conclusão, aprimorar a oferta de cursos e proporcionar uma experiência de aprendizado mais satisfatória e eficaz. Os resultados apresentados têm relevância significativa, fornecendo uma clara contribuição para a tomada de decisões estratégicas por parte dos gestores educacionais e de recursos humanos. Essa abordagem culmina em uma melhor qualidade e efetividade da aprendizagem móvel em benefício dos alunos.

Palavras-chave: Aprendizagem móvel, Mineração de dados, Aprendizado de maquina, Dados abertos, Inteligência artificial.

ABSTRACT

This study focuses on research aimed at analyzing data integrity and performance in online courses using machine learning methods. The main objective is to develop a tool to predict the number of students who have completed the course and identify potential cases of abandonment or withdrawal. To achieve this goal, we used supervised machine learning algorithms such as SVM and ANNs, allowing a detailed and predictive data analysis. The approach taken in this research was bibliographic and qualitative, exploring information from online courses in open databases and using data analysis techniques. The results obtained through these machine learning methods enabled the identification of patterns and trends in the data, providing a deeper understanding of the integrity of the records and the performance of the students. With a clearer view of the students' profile and the challenges faced, it is possible to implement proactive strategies to increase the completion rate, improve the course offerings, and provide a more satisfying and effective learning experience. The presented results are highly relevant and offer a clear contribution to strategic decision-making by educational and human resources managers. This approach leads to better quality and effectiveness of mobile learning for the benefit of students.

Keywords: Mobile learning, Data mining, Machine learning, Open data, Artificial intelligence.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

/ - divisão

= - igual

μ - mu

- - negativo

+ - positivo

σ - sigma

LISTA DE SIGLAS

ABNT Associação Brasileira de Normas Técnicas.

ANNs Redes Neurais Artificiais.

APIs *Application Programming Interface*.

BI *Business Intelligence*.

CSV *Comma Separated Values*.

EAD Eduacação a Distância.

EDA *Exploratory Data Analysis*.

Enap Escola Nacional de Administração Pública.

EV.G Escola Virtual de Governo.

FGV Fundação Getúlio Vargas.

IA Inteligência Artificial.

ILB Instituto Legislativo Brasileiro.

KPIs *Key Performance Indicators*.

LAI Lei de Acesso à Informação.

MC Ministério da Economia.

ML *Machine Learning*.

MLP *Multi-Layer Perceptron*.

MP Medida Provisória.

PDF *Portable Document Format*.

PGFN Procuradoria Geral da Fazenda Nacional.

PPEE Programa de Pós-Graduação Profissional em Engenharia Elétrica.

RBF *Radial Basis Function*.

SNNs *Stuttgart Neural Network Simulator*.

SVM *Support Vector Machine*.

SVN *Subversion*.

TED Termo de Execução Descentralizada.

TI Tecnologia da Informação.

UF Unidade Federativa.

UnB Universidade de Brasília.

1 INTRODUÇÃO

A Escola Nacional de Administração Pública (Enap), por meio da Escola Virtual de Governo Escola Virtual de Governo (EV.G), oferece à comunidade painéis de controle que proporcionam a visualização gráfica e legível de dados. Esses painéis permitem aos usuários conhecerem o perfil dos participantes e as principais características dos cursos oferecidos pela EV.G.

Dessa forma, os *Dashboards* disponibilizados podem auxiliar, entre outras possibilidades, na tomada de decisão por parte dos gestores responsáveis pela capacitação de pessoal em órgãos da administração pública. A EV.G expandiu seu catálogo de 18 para 330 cursos online, elevando o número de formandos de 38.000 alunos em 2013 para 69.000 alunos em 2014 e 145.000 em 2015.

Em 2020, alcançou aproximadamente mais de 1,6 milhões de alunos. Já em 2021, ultrapassou a marca de 1,6 milhão de alunos, totalizando mais de 5,5 milhões de alunos matriculados até o final de 2021, quando o estudo foi conduzido.

Assim o presente trabalho visa definir uma abordagem inovadora voltada à análise da integridade de dados e ao aprimoramento do desempenho em cursos online, empregando para isso técnicas avançadas de aprendizado de máquina.

A pesquisa tem como foco principal o ambiente da Escola Virtual de Governo (EV.G), uma iniciativa de destaque no âmbito do Governo Federal, cujo propósito é a centralização de informações relacionadas aos programas de capacitação disponibilizados.

Dentro desse contexto, são abrangidos elementos como categorias temáticas dos cursos, histórico de cursos realizados, demanda por capacitação, perfis dos participantes, instituições usuárias dos cursos e o número de funcionários impactados pelas formações oferecidas.

Este estudo visa não somente explorar, mas também revolucionar a forma como tais dados são tratados, através da aplicação de abordagens de vanguarda no campo da análise de dados e aprendizado de máquina.

Ao tornar tais informações de conhecimento público, incentiva-se, por parte do governo, a análise destes dados para atender a uma variedade de necessidades, promovendo, dessa forma, o exercício do controle social. Atualmente, a base de dados da EV.G contabiliza aproximadamente 9,3 milhões de registros, abrangendo o intervalo entre 2006 e julho de 2023.

A consolidação desta base proporciona uma fonte abastada para a análise e compreensão dos cursos de capacitação, permitindo uma perspectiva abrangente e embasada que sustenta as decisões administrativas e estratégicas. (2).

No âmbito deste estudo, surge a proposta intrigante de empregar técnicas avançadas de aprendizado de máquina para a avaliação abrangente da integridade dos dados e do desempenho dos cursos online.

Através da aplicação meticulosa de algoritmos de regressão e da adoção de estratégias sofisticadas de análise preditiva, busca-se discernir de maneira aprofundada os padrões subjacentes, tendências emergentes e fatores determinantes que exercem influência sobre o êxito ou a desistência dos cursos.

O propósito primordial desse empreendimento é destilar *Insights* perspicazes, destinados a enriquecer substancialmente a eficácia, qualidade e a própria tomada de decisão no contexto multifacetado da esfera educacional *online* (3).

Durante a migração da plataforma EmNumeros para uma versão atualizada por meio do *Power BI*, deparou-se com desafios que exigiram a superação de obstáculos, como a necessidade de atualização manual dos dados nas representações visuais devido à utilização da versão gratuita da ferramenta.

A nova abordagem adotada demonstrou ausência de registros relacionados a abandono ou cursos não finalizados. Para antecipar quantitativamente o número de estudantes que concluíram os cursos, desenvolveu-se uma ferramenta fundamentada em aprendizado de máquina, selecionando-se um algoritmo de regressão como base. Além disso, foram investigadas Redes Neurais Artificiais (ANNs) em conjunto com *Support Vector Machine* (SVM) com duas camadas ocultas.

Os desfechos do processo de treinamento foram classificados em categorias tais como "abandonado", "concluído", "fracassado", "interrompido"e "incompleto". Posteriormente, numa etapa subsequente, os cursos foram agrupados em "aprovados"(representando cursos finalizados com êxito) e uma outra categoria englobando situações de não aprovação, como abandono e fracasso.

1.1 MOTIVAÇÃO

A jornada de pesquisa foi impulsionada pela vontade de explorar profundamente a integridade dos dados e o desempenho vibrante dos cursos online fornecidos por meio da notável iniciativa conhecida como EV.G, um empreendimento inspirado pelo Governo Federal.

No âmago desse trabalho está a aspiração por transparência cristalina e participação social, por meio da divulgação aberta de detalhes sobre esses cursos, acompanhada pela aplicação destemida de métodos de aprendizado de máquina para desvelar padrões complexos e fatores influentes que moldam o sucesso ou a desistência.

A EV.G abriga um tesouro de informações abrangentes sobre a panóplia de cursos de aprimoramento, abraçando disciplinas variadas, histórico de cursos ministrados, demanda efervescente, tapeçaria dos aprendizes, participação de agências colaboradoras e magnitude dos indivíduos aperfeiçoados.

O ato de tornar pública essa riqueza de dados incita a vigilância colaborativa e análise multifacetada, moldada segundo inúmeras necessidades, pavimentando o caminho para decisões administrativas e estratégicas solidamente embasadas (2).

O cerne desta empreitada acadêmica é a adoção destemida de métodos de aprendizado de máquina, entre eles a aplicação sofisticada de algoritmos de regressão e o emprego refinado de técnicas de previsão. Nossa almejada recompensa é a descoberta de *insights* profundos que tenderão a aprimorar o dinamismo, excelência e fundamentação de escolhas na esfera dos cursos online.

Tudo isso é dirigido à causa de aprimorar o compromisso distintivo da EV.G com a clareza e a dispensa de informações de valor inestimável para o público em geral.

1.2 OBJETIVO GERAL

O objetivo geral deste trabalho concentra-se na implementação de algoritmos de aprendizado de máquina, particularmente aqueles relacionados à regressão. O foco é comparar a eficácia do SVM com a de uma Rede Neural Artificial de duas camadas ocultas, conhecida como ANNs.

1.2.1 Objetivos Específicos

- Investigar e avaliar o desempenho do algoritmo *Rondon Forest*, por meio da realização de experimentos que utilizam diversas quantidades de estimadores (4, 8, 16, 32, 64, 128, 256). O propósito é determinar a configuração ideal.
- Identificar e analisar os dados para propor uma ferramenta que possa prever o número de alunos que concluirão o curso.
- Interpretar os dados e desenvolver um novo tipo de análise visual para compreender o comportamento dos alunos em alguns cursos.
- Modelar a métrica de precisão de cada algoritmo para avaliação, considerando apenas duas variáveis de interesse: Aprovado e Falho.
- Definir a etapa em que o treinamento será realizado, incluindo a fusão das classes Aprovado em duas categorias: uma com o status completo e outra com o rótulo de Abandono para alunos não aprovados.
- Realizar o treinamento do modelo de aprendizagem de máquina utilizando um conjunto variado de classificadores.
- Avaliar a consistência dos dados dos cursos e propor uma ferramenta para garantir a qualidade e confiabilidade das informações.

1.3 PRINCIPAIS CONTRIBUIÇÕES

Artigo completo - 17ª Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI)/ 2022

- **Referência**: Praciano, F. G., Praciano, B. J., de Mendonça, F. L., Gallindo, E. L., da Silva, D. A., Duarte, F. C., & de Sousa, R. T. (2022, June). *Integrity of Training Data for Federal Civil Employees in Brazil. In 2022 17th Iberian Conference on Information Systems and* Technologies (CISTI) (pp. 1-5). IEEE.
- **Link**: https://ieeexplore.ieee.org/abstract/document/9820275

A EV.G vinculada à Enap é uma instituição que tem como objetivo proporcionar capacitação para servidores públicos nas três esferas do Governo, além de oferecer formação profissional para cargos de

carreira. Seu propósito é a discussão sobre a formação de servidores públicos e outros assuntos correlatos. A pesquisa bibliográfica qualitativa foi conduzida para fins descritivos, utilizando um painel EmNumeros construído com a ferramenta *Tableau* e o programa de mineração de dados em *Python*.

No decorrer deste artigo, foi ressaltada a importância do EV.G como um instrumento de valorização do serviço público. O sistema desempenha um papel crucial na valorização e no aprimoramento do conhecimento dos profissionais que atuam nas três esferas de governo.

Seu papel fundamental na condução das atividades dos servidores públicos representa um avanço nos serviços prestados à comunidade, resultando em uma melhoria na qualidade dos serviços disponibilizados pela Administração Pública.

Evidências numéricas apresentadas no artigo sustentam a afirmação de que o EV.G surge como uma ferramenta confiável para validar a formação dos funcionários públicos.

Essa abordagem visa capacitar servidores públicos proativos e responsáveis em suas funções dentro das estruturas administrativas públicas.(4).

1.4 ORGANIZAÇÃO DO TRABALHO

A estrutura da organização do trabalho segue o seguinte padrão. O Capítulo 2 aborda os referenciais teóricos que orientam o desenvolvimento da dissertação, juntamente com uma breve revisão da literatura, na qual são elencados os trabalhos correlatos que fundamentam a proposta deste estudo. No Capítulo 3, apresenta-se o modelo proposto, com uma descrição detalhada da estrutura e do funcionamento da integridade de dados, acompanhada de seus respectivos problemas e soluções. O Capítulo 4 engloba a exposição dos resultados e experimentos realizados. Assim, definem-se as conclusões e os trabalhos futuros no Capítulo 5. Por fim, são listadas as referências bibliográficas.

2 REFERENCIAL TEÓRICO

A fundamentação teórica deste estudo explora o papel primordial da geração de dados dentro de projetos de investigação e análise de informações. Essa etapa constitui a base essencial para embasar decisões informadas por parte dos pesquisadores.

Nesse estágio, emerge o procedimento de colheita e obtenção de dados pertinentes provenientes de diversas fontes, que abrangem desde bases de dados até arquivos em formato *Comma Separated Values* (CSV) e interfaces de programação *Application Programming Interface* (APIs). Mediante uma eficaz administração dos dados, torna-se possível obter um conjunto abrangente e representativo.

Este conjunto de dados desempenhará um papel fundamental ao servir como alicerce para as análises estatísticas e os algoritmos de aprendizado de máquina empregados no âmbito deste estudo dissertativo.

2.1 TÉCNICAS AVANÇADAS DE APRENDIZADO DE MÁQUINA E INTELIGÊNCIA ARTIFICIAL

Com o rápido avanço tecnológico e a ampla disponibilidade de dados, tem havido um impulso significativo no desenvolvimento de técnicas e algoritmos avançados no campo do aprendizado de máquina e inteligência artificial. Essas abordagens têm sido extensivamente exploradas em diversos setores, oferecendo oportunidades promissoras para análise e tomada de decisões em dados.

Nesse contexto, destacam-se diversas ferramentas que desempenham um papel fundamental na dissertação em questão. A garantia de dados é essencial para obter as informações necessárias para as análises, enquanto o *Z-score* e a análise do *Z-score* permitem uma compreensão profunda da distribuição dos dados e identificação de valores atípicos.

Os SVM são algoritmos poderosos para classificação e regressão, enquanto as ANNs têm a capacidade de aprender padrões complexos e realizar tarefas de classificação e previsão. O processamento de máquina e o aprendizado profundo são abordagens mais amplas que englobam uma variedade de algoritmos e técnicas, permitindo a análise e interpretação dos dados de forma mais avançada.

A inteligência artificial, como campo multidisciplinar, abrange essas técnicas e algoritmos para o desenvolvimento de sistemas capazes de realizar tarefas inteligentes.

Enfim, para a interpretação e compreensão dos dados, destacam-se o *boxplot* e o diagrama de dispersão, que fornecem informações sobre a distribuição dos dados e possíveis correlações entre variáveis.

2.1.1 *Z-score*

Uma estimativa do número de desvios padrão por qual um ponto está distante da média de seu conjunto de dados é realizado. O *Z-score* recebe a denominação de pontuação padrão e tem a capacidade de ser

representada em um gráfico de distribuição comum. No geral, esse escore mensura os desvios quando um ponto de dados se encontra abaixo ou acima da média populacional.

O uso do *Z-score* se destina a comparar pontos de dados entre diversos conjuntos, buscando identificar correlações. A pontuação resultante pode ser zero, positiva ou negativa. Uma pontuação igual a zero indica que o ponto é médio, portanto, idêntico à média. Um valor negativo evidencia o quanto um ponto se encontra abaixo da média na curva de distribuição. Já um valor positivo demonstra o quanto um ponto está acima da média. (5).

2.1.2 Análise do *Z-Score*

O conceito foi adaptado ao mundo dos negócios e financiamentos pelo Dr. Edward Altman, que utilizou-o como método preditivo para a falência de empresas. O cálculo desenvolvido por ele recebe o nome de Altman *Z-score*.

O Dr. Altman detém uma reputação de alcance internacional como especialista em falências corporativas, títulos de alto rendimento, dificuldades de endividamento e análise de risco de crédito. Em 1984, ele recebeu o título de *Laureate da Hautes Etudes Commerciales Foundation*, em Paris, em reconhecimento às suas contribuições acumuladas no desenvolvimento de modelos de previsão de dificuldades corporativas e procedimentos para a reabilitação financeira de empresas.

Sendo nomeado Professor Honorário pela Universidade de Buenos Aires em 1996 e recebendo o título de "Doutor Honorário"pela *Universidade de Lund* (Suécia) em 2011, bem como pela Escola de Economia de Varsóvia (Polônia) em 2015, o Professor Altman também desempenhou o papel de conselheiro para a *Centrale dei Bilanci* na Itália e aconselhou diversos bancos centrais estrangeiros.

O papel do professor Altman se estende ainda à presidência do Conselho Consultivo Acadêmico da *Turnaround Management Association*. Suas conquistas incluem a entrada no *Fixed Income Analysts Society Hall of Fame* em 2001, a posição de Presidente da *Financial Management Association* em 2003, tornando-se um *FMA Fellow* em 2004, e sendo um dos pioneiros a ser incluído no *Hall da Fama da Turnaround Management Association* em 2008.

Além disso, em 2005, o Prof. Altman foi distinguido como uma das “100 Pessoas Mais Influentes em Finanças” pela revista *Treasury & Risk Management*. (6).

Basicamente, um *Z-score* representa o número de desvios padrão em relação à média de um dado ponto de informação. O termo utilizado para referir-se a um *Z-score* é "pontuação padrão", e este pode ser convenientemente plotado em um gráfico de dispersão usual. Para a utilização de um *Z-score*, faz-se necessário estar ciente da média μ e também do desvio padrão da população σ.

Os *Z-scores* representam uma abordagem para contrastar os resultados de um teste com uma população "comum". Os resultados de testes ou estudos apresentam várias unidades e resultados potenciais (7).

A equação essencial da pontuação z para um exemplo é:

$$z = (x - \mu)/\sigma \tag{2.1}$$

Por exemplo, pode-se considerar o cenário em que uma avaliação de 190 é obtida por um indivíduo em um teste. O referido teste apresenta uma média (μ) de 150 e um desvio padrão (σ) de 25. Em uma situação esperada de ocorrência comum, a pontuação z correspondente a esse indivíduo seria calculada.

$$z = (x - \mu)/\sigma \tag{2.2}$$

$$= 190 - 150/25 = 1.6 \tag{2.3}$$

O *Z-score* revela o número de desvios padrão acima da média da pontuação de uma pessoa. Neste caso, a pontuação dessa pessoa está 1,6 desvios padrão acima da média.

A equação do *Z-score* é calculada utilizando o erro padrão da média, especialmente quando lidando com vários exemplos e é necessário representar o desvio padrão desses exemplos. Portanto, a equação de Z é usada quando se deseja considerar o erro padrão.

$$z = (x - \mu)/(\mu/\sigma) \tag{2.4}$$

O *Z-score* revela o número de erros padrão que há entre o meio do exemplo e o meio da população.

Qual é a fórmula?

As Figuras 2.1 e 2.2 a seguir descrevem como a fórmula e a equação do Altman *Z-score* são calculadas.

Fonte: Profit.co

Figura 2.1: Formula *Z-score*

A equação Altman Z-score é calculada assim:

Fonte: Profit.co

Figura 2.2: Cálculo Altman *Z-score*

2.1.3 Cálculo do *Z-score*

O *Z-score* é conhecido por representar, em termos numéricos, a relação entre um valor e a média de um conjunto de valores.

O cálculo do *Z-score* baseia-se nos desvios padrão em relação à média. Por exemplo, um *Z-score* de 0 indica que a pontuação do dado é similar à pontuação média. Um *Z-score* de 1,0 indicaria um valor desviado da média por um desvio padrão.

Os *Z-scores* podem ser tanto negativos quanto positivos. Um valor positivo indica que a pontuação está acima da média, enquanto um valor negativo indica que a pontuação está abaixo da média.

O funcionamento do *Z-score* revela aos *traders* e estatísticos se uma pontuação é específica ou anormal para um conjunto de dados referenciado anteriormente. Essa medida também possibilita que os analistas ajustem as pontuações de alguns conjuntos de dados para produzir pontuações que possam ser comparadas umas com as outras com maior precisão.

Uma das principais diferenças entre o *Z-score* e o Desvio Padrão é que o Desvio Padrão reflete a oscilação dentro de um conjunto de dados definido. Ele é calculado começando pela diferença entre cada ponto de dados e a média. Essas diferenças são elevadas ao quadrado e somadas para calcular a variância.

O Desvio Padrão é então obtido como a raiz quadrada da variância.

Por outro lado, o *Z-score* é a medida de quantos desvios padrão um ponto de dados está afastado da média. Para pontos de dados abaixo da média, o *Z-score* é negativo.

Em grandes conjuntos de dados, a maioria dos valores (99%) apresenta um *Z-score* entre -3 e 3, o que significa que estão dentro de 3 desvios padrão acima e abaixo da média (8).

Na estatística, o *Z-score* (ou escore padrão) é utilizado para comparar médias de conjuntos de dados distintos que têm distribuição uniforme.

Ele indica quantos desvios padrão uma observação está acima ou abaixo da média. O *Z-score* é empregado em pesquisas que utilizam análise estatística, permitindo a comparação de valores de observações de diferentes distribuições normais.

Quando os itens de diferentes conjuntos de dados são transformados em *Z-scores*, eles podem ser comparados diretamente. A seguir, é apresentada uma demonstração de como calcular um *Z-score* (9).

Passo 1

A fórmula utilizada para o cálculo do *Z-Score* (ou escore padrão) é a seguinte:

$$z = (x - \mu)/\sigma \tag{2.5}$$

Passo 2

As variáveis presentes na fórmula do *Z-Score* são: z = *Z-Score*, x = escore bruto ou observação a ser padronizada, μ = média da população e σ = desvio padrão da população.

Passo 3

Modelo de cálculo do *Z-Score*: Uma observação de 14,75; uma média populacional de 12,2; e um desvio padrão de 1,75. Neste caso, o *Z-Score* é calculado como:

$$z = (14,75 - 12,2)/1,75 \;\; \textit{Z-Score} = 1,46. \tag{2.6}$$

2.1.4 Máquinas vetoriais de suporte - SVM

Uma máquina de vetor de suporte *Subversion* (SVN) constitui um algoritmo de aprendizado supervisionado apto a ser empregado em tarefas de classificação e regressão. O algoritmo opera como um classificador discriminativo, delineando um limiar decisório que otimiza a margem interclasses. Trata-se de um método de aprendizado supervisionado.

O algoritmo SVN é também autodenominado classificador de margem máxima. Sua interpretação ocorre em duas fases. Primeiramente, o algoritmo identifica um hiperplano que efetua a separação entre ambas as categorias. Posteriormente, detecta os vetores de suporte, isto é, os pontos de dados mais afins ao hiperplano.

O algoritmo SVN possui diversas vantagens. Revela-se altamente eficaz em espaços de dimensões ele-

vadas. Conforme sua formulação, demonstra resistência ao sobreajuste. Além disso, possibilita a resolução de problemas não lineares e, por fim, apresenta eficiência computacional.

O SVN representa um algoritmo de aprendizado de máquina, com aplicabilidade em diversas tarefas, a exemplo de regressão e classificação. O método de vetor de suporte carrega consigo algumas desvantagens. Não se mostra invariante à escala, tradução ou rotação, tampouco robusto a interferências. Em âmbito de regressão, o SVN viabiliza a previsão de resultados contínuos, como valores de ativos.

Da mesma forma, emprega-se em classificação para antecipar a inclusão ou não de uma instância em uma classe específica. Além disso, encontra aplicação em outras modalidades de análise de dados.

O SVN emerge como uma ferramenta notável, uma vez que se adapta a uma gama variada de dados, englobando desde estruturas lineares até as não lineares. O método SVN apresenta diversas vantagens sobre alternativas algorítmicas, tais como sua menor susceptibilidade ao sobreajuste e habilidade para lidar com conjuntos de dados de alta dimensão.

Há três categorias de SVM: Máquinas Vetoriais de Suporte, Máquinas Vetoriais de Suporte Linear e Máquinas Vetoriais de Suporte Não-linear.

- **As máquinas de vetores de suporte** são um tipo de algoritmo de aprendizado supervisionado utilizado para tarefas de classificação e regressão. O algoritmo determina o ótimo hiperplano capaz de separar os pontos de dados em duas categorias distintas. O uso das máquinas de vetores de suporte estende-se à resolução de questões envolvendo classificação, regressão e também categorização.
- **As máquinas de vetores de suporte lineares** representam uma variante das máquinas de vetores de suporte, empregada quando os dados apresentam separação linear. Essa situação ocorre quando os dados podem ser particionados em duas classes por meio de uma única linha. Além de sua aplicação na classificação, as máquinas de vetores de suporte são empregadas na disposição de informações na tela de um computador.
- **As máquinas de vetores de suporte não lineares** constituem um subtipo das máquinas de vetores de suporte, adotado nos cenários em que os dados não podem ser linearmente separados. Ou seja, a divisão dos dados em duas categorias distintas não é possível por meio de uma única linha. Tais máquinas de vetores de suporte são frequentemente empregadas nas áreas de análise de dados e processamento de informações na indústria de software SVN (10).

2.1.5 Rede neural artificial - ANN

As redes neurais, também referidas como ANNs ou *Stuttgart Neural Network Simulator* (SNNs), fazem parte do âmbito do aprendizado de máquina, desempenhando um papel essencial nas operações dos algoritmos de inteligência artificial.

Com inspiração no funcionamento do cérebro humano, essas redes consistem em unidades interconectadas, chamadas neurônios artificiais ou nodos, os quais conduzem o processamento de dados e a geração de saídas correspondentes. A aplicação das redes neurais abarca várias esferas, abrangendo o reconhecimento de padrões, o processamento de linguagem natural, a computação e outras aplicações pertinentes.

As redes neurais artificiais ANNs são constituídas por camadas de nó, incorporando uma camada de entrada, uma ou múltiplas camadas ocultas e uma camada de saída. Cada nodo, ou neurônio artificial, estabelece conexões com outros e carrega consigo um peso e um limiar correlacionados.

Quando a saída de qualquer nodo individual ultrapassa o valor do limiar especificado, esse nodo é ativado, enviando dados à próxima camada da rede. Caso contrário, nenhum dado é transmitido à próxima camada da rede. No material apresentado, a Figura 2.3 delineia uma "Rede Neural Profunda".

Fonte: IBM

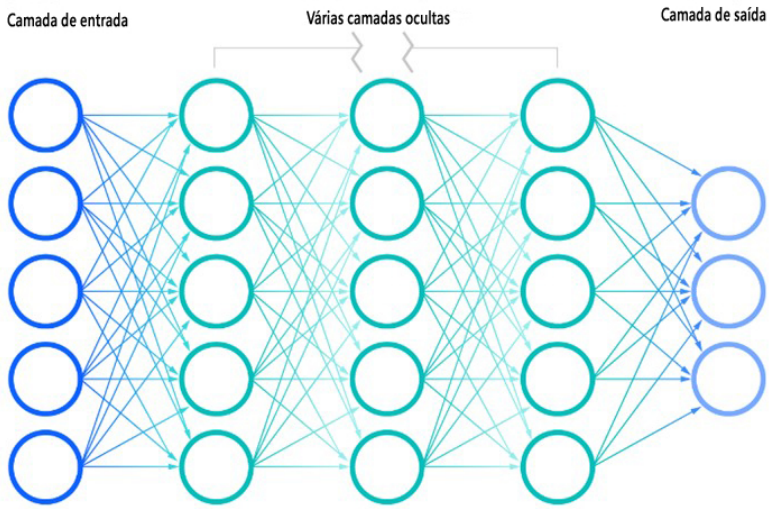

Figura 2.3: Rede Neural Profunda

As redes neurais possuem dados de treinamento que são utilizados para compulsionar e aprimorar a precisão delas ao longo do tempo. No entanto, uma vez que os algoritmos de aprendizado são ajustados para atingir alta precisão, eles se tornam poderosos mecanismos no campo da ciência da computação e da inteligência artificial, permitindo a classificação e o agrupamento de dados em alta velocidade.

Tarefas como reconhecimento de fala ou identificação de imagens podem ser executadas em minutos, ao invés de horas, quando comparadas com a identificação manual realizada por especialistas humanos. Um exemplo proeminente de rede neural é o algoritmo de busca do *Google* (11).

2.1.6 Aprendizado de máquina

A divisão da Inteligência Artificial (IA) e da ciência da computação conhecida como *Machine learning* engloba a utilização de dados e algoritmos para replicar a maneira como os seres humanos aprendem, aprimorando gradualmente sua precisão.

Machine learning representa uma parte crescente e fundamental no campo da ciência de dados. Por meio da aplicação de métodos estatísticos, os algoritmos passam por treinamento para realizar classificações ou previsões, desvendando *insights* essenciais em projetos de mineração de dados.

Estes *insights* subsequentemente orientam a tomada de decisões em aplicações e empreendimentos, oti-

mizando o alcance das principais métricas de crescimento. À medida que o fenômeno da *Big Data* continua sua expansão e evolução, a demanda no mercado por profissionais de ciência de dados aumentará, requerendo que estes auxiliem na identificação dos problemas de negócios mais pertinentes e, posteriormente, utilizem os dados para solucioná-los (12).

2.1.7 Deep learning

O *Deep Learning*, também conhecido como Aprendizado Profundo, representa uma subárea do campo de *Machine Learning* (ML) que aborda as Redes Neurais Artificiais. Essa disciplina tem como objetivo emular o funcionamento cerebral por meio da máquina que está aprendendo. Os registros científicos iniciais sobre a tentativa de criar neurônios artificiais datam da década de 1950.

Com o avanço da era digital, houve um substancial aumento na capacidade de processamento dos computadores. No contexto das Redes Neurais, que seguem um paradigma de conexões ponderadas, a habilidade de desenvolver modelos inteligentes está diretamente relacionada à rede de conexões entre um número considerável de neurônios artificiais.

Esse conceito se alinha com a noção de *Big Data*, uma vez que esse tipo de arquitetura demonstra um desempenho mais eficaz ao lidar com grandes volumes de dados de entrada.

A área de ML engloba o estudo científico de algoritmos e modelos estatísticos que sistemas computacionais utilizam para desempenhar tarefas específicas sem depender de instruções diretas. Baseando-se em padrões e inferências, o ML envolve a criação de algoritmos que possam adaptar-se e modificar-se autonomamente, visando produzir resultados desejados. Em outras palavras, é a capacidade de treinar algoritmos para tomar decisões por conta própria, sem intervenção humana.

O objetivo subjacente ao Aprendizado de Máquina é permitir que as máquinas aprendam de forma autônoma a partir dos dados recebidos, possibilitando previsões futuras. Um exemplo prático é a previsão de tempestades: dados de tempestades passadas são coletados juntamente com as condições climáticas simultâneas ao longo de um período específico.

Esses dados são então comparados com as condições em tempo real para prever a ocorrência de uma tempestade específica, resultando em um aprimoramento contínuo dos dados de entrada (13). Esse processo é ilustrado na Figura 2.4.

Fonte: Quality Magazine

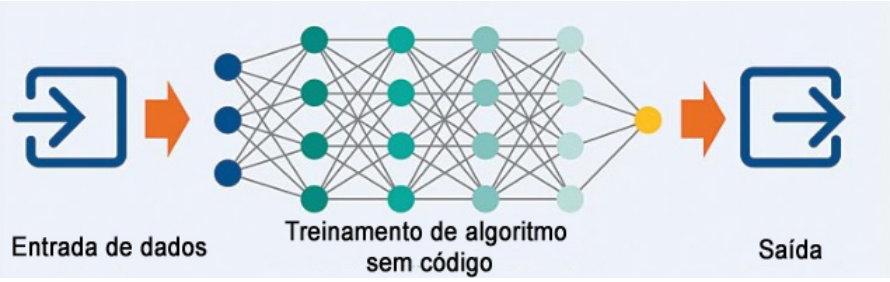

Figura 2.4: Aprendizado de máquina

2.1.8 Inteligência artificial

O termo IA teve sua origem em 1956 durante a famosa reunião de *Dartmouth*. Os participantes daquela reunião incluíram *Allen Newell, Herbert Simon, Marvin Minsky, Oliver Selfridge e John McCarthy*. Ao invés de construir sistemas baseados em números, eles se dedicaram a desenvolver sistemas que manipulassem símbolos. Essa abordagem demonstrou ser poderosa e desempenhou um papel fundamental em muitos empreendimentos posteriores (14).

Desde o seu surgimento, a pesquisa em IA enfrentou um desafio significativo, como exemplificado pela questão colocada por *Minsky* em seu livro *Semantic Information Processing* há quase trinta anos. Ao longo desse período, várias correntes de pensamento dentro do campo da IA exploraram maneiras de instilar comportamentos inteligentes em máquinas (15).

A área de pesquisa conhecida como IA tem como objetivo a criação de máquinas capazes de executar tarefas que tradicionalmente eram reservadas apenas aos seres humanos. A meta subjacente é desenvolver sistemas que possam aprender, raciocinar, perceber, compreender e se adaptar a novas situações (14) (16).

A pesquisa em IA engloba diversas disciplinas, incluindo ciência da computação, matemática, estatística e engenharia. Alguns dos principais campos de estudo dentro da IA abrangem o aprendizado de máquina, o processamento de linguagem natural, a visão computacional, a robótica e os sistemas de recomendação (17).

Dentre as aplicações concretas da IA, destacam-se áreas como o reconhecimento de fala e de imagens, os sistemas de recomendação em plataformas de compras online, os *chatbots* utilizados no atendimento ao cliente, os veículos autônomos, os jogos eletrônicos, entre outras aplicações (18).

A IA detém o potencial para transformar diversos setores, incluindo saúde, finanças, transporte e manufatura. No entanto, é fundamental considerar as implicações éticas e sociais associadas ao uso da IA, como a preservação da privacidade dos dados, a existência de possíveis preconceitos algorítmicos e o impacto sobre o mercado de trabalho (19).

Somente recentemente, a IA atraiu um interesse crescente devido ao surgimento de aplicações práticas no âmbito comercial. Um fator crucial para essa transição bem-sucedida da academia para a indústria foi o notável avanço tecnológico em capacidade computacional ocorrido nas duas últimas décadas (20).

Um sistema de IA não se restringe à capacidade de armazenar e manipular dados, mas também abrange a habilidade de adquirir, representar e manipular conhecimento. Essa manipulação envolve a capacidade de deduzir ou inferir novos conhecimentos a partir das informações existentes. Uma das ideias mais impactantes provenientes da pesquisa em IA é a separação entre fatos e regras de conhecimento declarativo e os algoritmos de tomada de decisão (21).

O desenvolvimento de um sistema de IA resume-se em adquirir e codificar regras e fatos adequados para um domínio específico. Esse processo de codificação é conhecido como engenharia de conhecimento. Essa abordagem influenciou profundamente a maneira como os cientistas enfrentam problemas e as técnicas de engenharia empregadas na criação de sistemas inteligentes.

O designer de um sistema de IA enfrenta desafios essenciais: a aquisição, a representação e a manipulação do conhecimento. Para orientar esses processos, é necessária uma estratégia de controle ou uma

máquina de inferência que determine quais conhecimentos devem ser acessados, que deduções devem ser feitas e a sequência de passos a ser seguida. A Figura 2.5 ilustra a interconexão dos componentes em um sistema clássico de IA (14).

Fonte: Schutzer, 1987

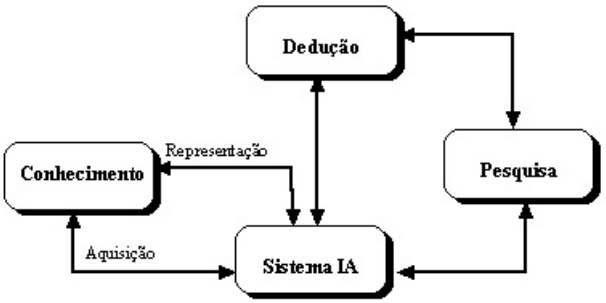

Figura 2.5: Uma visão conceitual dos sistemas de Inteligência Artificial

2.1.9 *Boxplot*

O *boxplot*, também referido como diagrama de caixa, é uma ferramenta gráfica poderosa amplamente empregue para a visualização e análise da distribuição de conjuntos de dados. Ele é construído com base em parâmetros estatísticos cruciais, tais como valores mínimo e máximo, primeiro e terceiro quartil, mediana e *outliers*, todos derivados dos dados subjacentes.

A principal finalidade do *boxplot* consiste em proporcionar uma compreensão rápida e sucinta das medidas estatísticas essenciais dos conjuntos de dados. Isso inclui a identificação da tendência central (representada pela mediana), a dispersão dos valores (refletida pela amplitude da caixa) e a detecção de pontos atípicos *outliers*, que podem sinalizar a existência de dados incomuns ou erros de medição.

Ao comparar o *boxplot* com outras representações gráficas, como o histograma, nota-se que ele traz vantagens significativas. Enquanto o histograma apresenta a distribuição completa dos dados e é útil para analisar a forma da distribuição, o *boxplot* concentra-se nas principais estatísticas resumidas, facilitando a detecção de valores discrepantes e a identificação de padrões gerais.

Por meio da utilização do *boxplot*, os analistas de dados conseguem obter uma visão pronta das características fundamentais dos conjuntos de dados. Isso possibilita a identificação de possíveis assimetrias e discrepâncias, além de destacar grupos ou subconjuntos que demandam investigação adicional.

De maneira resumida, o *boxplot* constitui uma ferramenta valiosa para a análise exploratória de dados. Ele proporciona uma compreensão aprofundada das propriedades estatísticas essenciais e realça aspetos relevantes da distribuição. Quando combinado com outras técnicas de análise, o *boxplot* torna-se parte integrante do conjunto de ferramentas dos analistas de dados, que buscam obter *insights* valiosos a partir dos dados disponíveis, como ilustrado na figura 2.6 (22) (23).

Fonte: *Statplace*

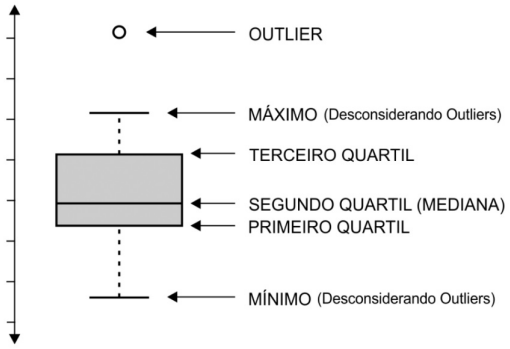

Figura 2.6: *Boxplot* e seu formato

2.1.10 Diagrama de dispersão

O Diagrama de dispersão, também conhecido como Gráfico de dispersão, faz parte das ferramentas incorporadas à área de qualidade. Ele se caracteriza como um gráfico com eixos verticais e horizontais que correlacionam a causa e o efeito.

Assim, a partir dos dados correlacionados, é possível compreender no referido Diagrama se há uma relação de causa e efeito entre as variáveis em questão (24). Isso é ilustrado na Figura 2.7 apresentada a seguir.

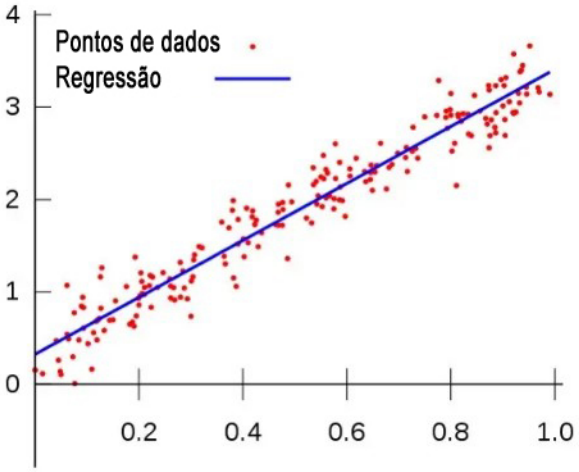

Figura 2.7: Diagrama de dispersão

2.1.10.1 Explorando Diagramas e Histogramas

O Diagrama é utilizado principalmente para verificar se há uma verdadeira relação entre as duas variáveis e se existe viabilidade de uma relação de causa e efeito. Além disso, o nível de intensidade do relacionamento entre as duas variáveis pode ser mensurado, determinando se é forte ou fraco.

O histograma organiza as referências de forma que a visualização da distribuição de um conjunto de dados seja possível, assim como a percepção da localização do valor central e da dispersão dos dados ao redor desse valor. Os eixos do histograma são os seguintes:

- **O eixo horizontal**, dividido em intervalos pequenos, apresenta os valores expressos por uma variável de interesse.
- **O eixo vertical**, cuja área harmoniza-se com o número de observações na amostra que possuem valores dentro do intervalo correspondente a frequência (25).

2.1.11 Histograma

O histograma é mais do que um gráfico de barras que apresenta a distribuição de dados (distribuição de frequência). Pode ser percebido como um indicador da variabilidade de um processo (dispersão) (25).

2.1.11.1 Processos com o Uso do Histograma

Usa-se o Histograma para:

• Condensar uma variedade de dados graficamente (população muito grande);

• Confrontar os resultados de um processo com as especificações;

• Aferir o número de produto não conforme;

• Comunicar informações e intensificar equipe de melhoria;

• Assessorar o processo de tomada de decisão.

2.1.11.2 Partes de um histograma

Um histograma é composto por três elementos: classes, amplitude e frequência.

No que diz respeito às classes, elas constituem barras que indicam valores estatísticos, refletindo tanto os valores mínimos quanto os máximos (conhecidos como limites de classe).

Quanto à amplitude, essa característica representa o tamanho de cada uma das classes (barras).

A frequência, por sua vez, representa a ocorrência da variação nos conjuntos de dados.

As partes do histograma são ilustradas no exemplo de análise de dados referente aos alunos que abandonaram o curso em um determinado período.

As partes do histograma podem ser observadas na figura 2.8 da análise de dados (26) (27).

Fonte: Significados Educação

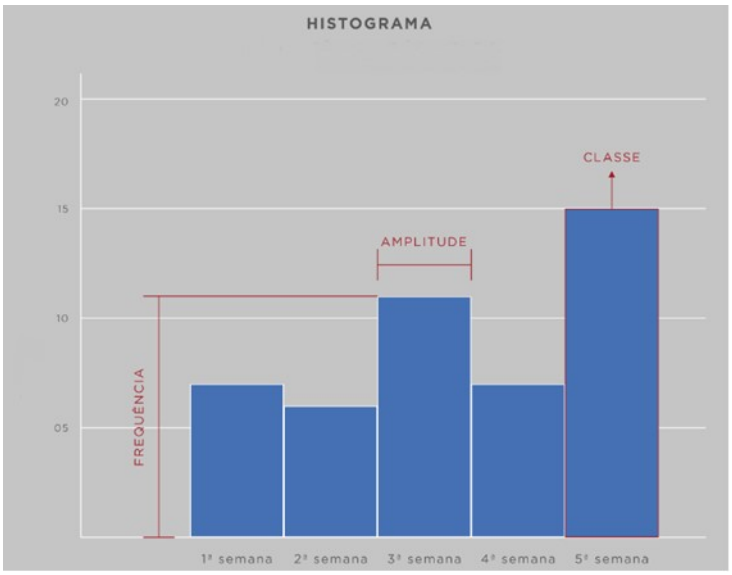

Figura 2.8: Histograma e seus elementos

2.1.11.3 Os seis tipos de histograma e suas formas de apresentação

Há seis tipos de histograma, os quais são categorizados de acordo com a maneira pela qual as barras são apresentadas: simétrico, assimétrico, despenhadeiro, dois picos, achatado e pico isolado (25).

- Simétrico: Este estilo de histograma revela a frequência mais elevada no centro, enquanto as mais baixas estão situadas nas extremidades. Sua aplicação é comum na representação de médias de dados obtidos, que são empregadas para estabelecer simetrias com outros detalhes provenientes da pesquisa, conforme ilustrado na Figura 2.9 subsequente.

Fonte: Significados Educação

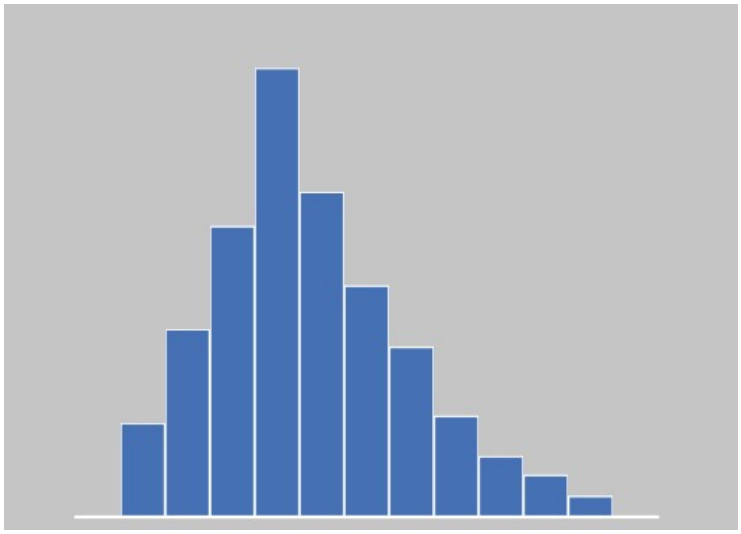

Figura 2.9: Histograma Assimétrico

- Assimétrico: Conforme ilustrado na Figura 2.10, observa-se que o objeto em questão exibe assimetria ao apresentar somente um pico. No entanto, o significado desse padrão fica evidente ao se analisar sua representação. Geralmente, ele indica uma situação na qual está sujeito a um único limite de especificação e é controlado durante todo o processo.

Fonte: Significados Educação

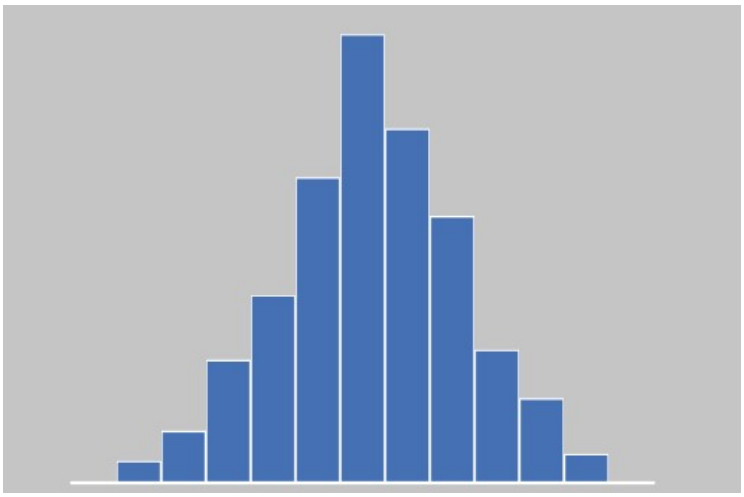

Figura 2.10: Histograma Simétrico

- Despenhadeiro: Conforme o próprio nome sugere, a situação assemelha-se a um declive. As características lembram as de um barranco. Isso acontece quando ocorre a eliminação de dados. Como resultado, verifica-se uma interrupção na representação, criando uma sutil percepção de que o histograma está incompleto. A média dos valores situa-se fora do centro da faixa de especificações, como ilustrado no exemplo presente na Figura 2.11.

Fonte: Significados Educação

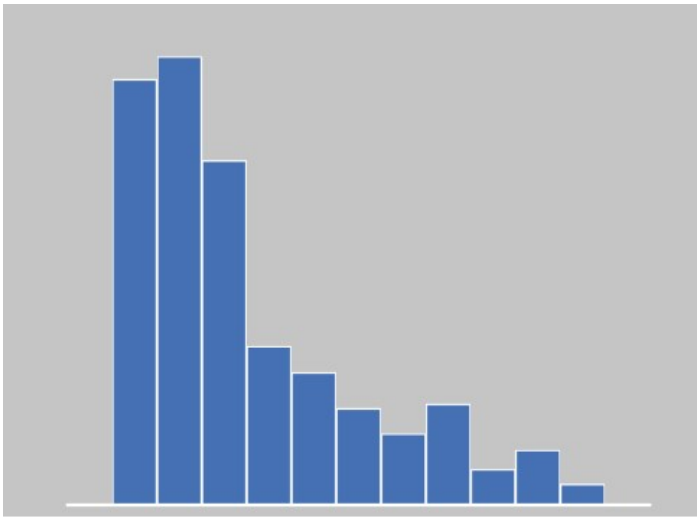

Figura 2.11: Histograma Despenhadeiro

- Dois picos: Conforme ilustrado na Figura 2.12, a característica distintiva consiste na presença de

duas frequências mais altas em relação às restantes. Esse fenômeno manifesta-se consistentemente em situações que envolvem a combinação de dados diversos. A fim de facilitar a compreensão, essa combinação refere-se, por exemplo, à coleta de dados obtidos sob condições significativamente distintas, seja em termos de operador, equipamento ou matéria-prima.

Fonte: Significados Educação

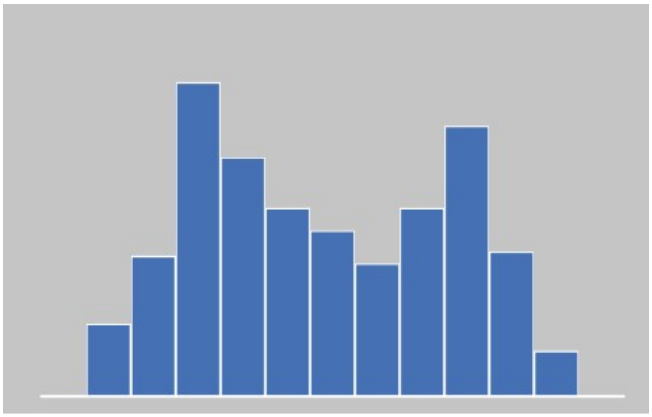

Figura 2.12: Histograma Dois picos

- Achatado: Neste padrão, as frequências estão em proximidade umas das outras, em níveis altamente comparáveis. Isso ocorre quando ocorre a combinação de distribuições com médias diversas, ele é também reconhecido como "platô". Isso é ilustrado na Figura 2.13.

Fonte: Significados Educação

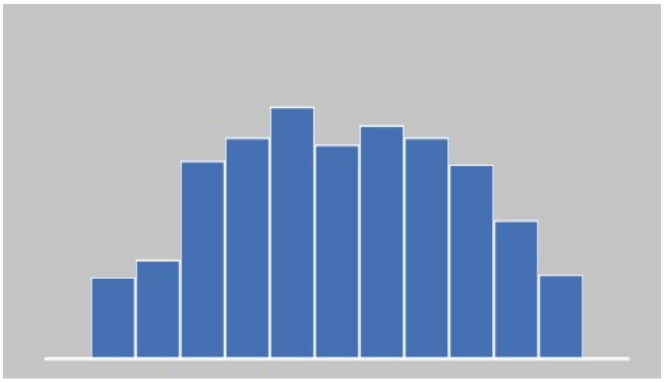

Figura 2.13: Histograma Achatado

- Pico Isolado: Este perfil se manifesta quando ocorre uma leve incorporação de dados de uma distribuição distinta, como ocorre em situações de anormalidade no processo, erros de medição ou a inclusão de dados provenientes de um processo diferente. Tal ilustração é apresentada na Figura 2.14.

Fonte: Significados Educação

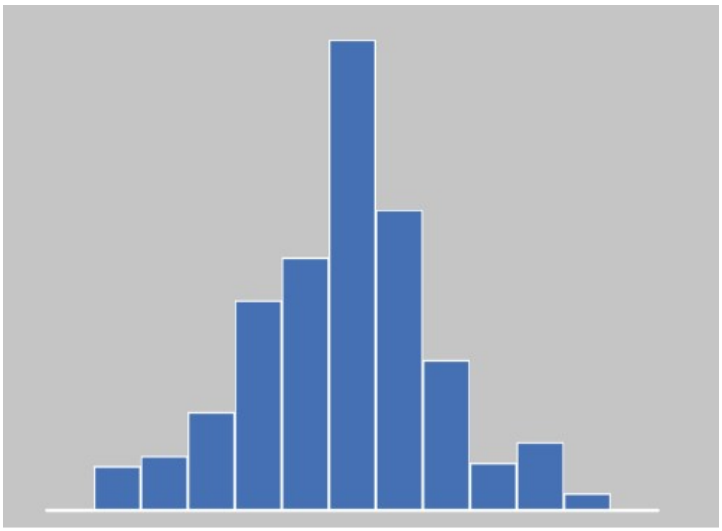

Figura 2.14: Histograma Pico isolado

No histograma, é possível observar melhor a média e o desvio padrão. Já no *Boxplot* se percebe um pouco melhor as medidas de quartis, mediana, amplitude, além de identificar muito bem os *outliers*.

No *Boxplot* a parte central do gráfico contém os valores que estão entre o primeiro quartil e o terceiro quartil. As hastes inferiores e superiores se estendem, respectivamente, do primeiro quartil até o menor valor, limite inferior, e do terceiro quartil até o maior valor (23).

2.1.12 Aprendizagem móvel

A aprendizagem móvel, referida como ML, representa uma abordagem de educação a distância em que dispositivos móveis, como *smartphones* e *tablets*, são empregados como recursos para simplificar o processo de ensino e aprendizagem. O destaque atribuído a essa forma de educação deriva da crescente aceitação de dispositivos móveis e da ampla disponibilidade de acesso à internet (28).

2.1.12.1 Arquitetura pedagógica de aprendizagem móvel na integração interdisciplinar

A arquitetura pedagógica de aprendizagem móvel desempenha um papel relevante na promoção da integração entre os grupos disciplinares por meio da utilização de aplicativos controlados. Nesse contexto, emerge o desafio de gerenciar de forma efetiva o tempo e as informações durante o uso dessas ferramentas tecnológicas.

Ao adotar a arquitetura pedagógica de aprendizagem móvel, os educadores enfrentam a tarefa de otimizar o tempo disponível e administrar as informações de maneira eficiente. A utilização de aplicativos controlados possibilita a interação entre estudantes de diferentes disciplinas, ampliando as oportunidades de aprendizagem colaborativa e troca de conhecimentos.

No entanto, é necessário enfrentar o desafio de gerir adequadamente o tempo de utilização dessas

ferramentas e lidar com o fluxo contínuo de informações geradas. A complexidade desse processo requer uma abordagem cuidadosa e estratégias de gerenciamento eficazes por parte dos educadores.

Dessa forma, a arquitetura pedagógica de aprendizagem móvel revela-se como um componente crucial para promover a integração e a interação entre os grupos disciplinares por meio de aplicativos controlados. A capacidade de gerenciar de forma efetiva o tempo e as informações durante o uso dessas ferramentas é fundamental para garantir uma experiência de aprendizagem enriquecedora e produtiva para os alunos, conforme demostrado na Figura 2.14 (29).

Fonte: ResearchGate GmbH.

Figura 2.15: Arquitetura pedagógica de aprendizagem móvel

2.1.12.2 Características

A aprendizagem móvel possui características singulares que a distinguem de outras modalidades de aprendizagem online. A característica primordial consiste na portabilidade dos dispositivos móveis, o que proporciona aos estudantes o acesso ao conteúdo educacional em qualquer local e a qualquer hora. Essa flexibilidade temporal e espacial possibilita que os alunos se envolvam no processo de aprendizagem de acordo com suas próprias conveniências e disponibilidades.

2.1.12.3 Vantagens

Uma das vantagens da aprendizagem móvel é a personalização do aprendizado. São oferecidas aos estudantes uma variedade de opções e recursos por meio de dispositivos móveis, permitindo que se adaptem de acordo com suas preferências individuais. Diferentes modalidades de aprendizado, como leitura de texto, visualização de vídeos, participação em fóruns de discussão ou realização de exercícios interativos, estão ao alcance, promovendo, assim, uma experiência de aprendizagem mais envolvente e eficaz.

Com a chegada dos dispositivos móveis, uma ampla gama de oportunidades se abriu para a aprendizagem. Nesse contexto, fica evidente a importância de planejar e desenvolver atividades educacionais que incorporem o uso desses dispositivos.

A aprendizagem móvel proporciona aos alunos a chance de construir e aprimorar seus conhecimentos, desde que tenham acesso a *smartphones* ou *tablets* e estejam conectados à internet. Essa abordagem pedagógica apresenta o potencial de transformar o processo educacional, garantindo uma experiência de aprendizagem mais flexível e acessível (30).

2.1.12.4 Benefícios

Além disso, a interação entre os alunos e os professores, bem como a colaboração entre os estudantes, é incentivada pela aprendizagem móvel. Através de aplicativos e plataformas móveis, os participantes podem engajar-se em discussões online, partilhar ideias, realizar atividades colaborativas e receber um retorno imediato. Essa dinâmica promove a construção coletiva do conhecimento e estimula o envolvimento ativo dos estudantes no processo educativo.

Um outro benefício crucial da aprendizagem móvel consiste no acesso aos recursos multimídia. Os dispositivos móveis suportam variados formatos de mídia, como áudio, vídeo, imagens e animações. Isto capacita os alunos a experienciarem uma aprendizagem mais rica e diversificada, simplificando a compreensão de conceitos complexos e fomentando o interesse dos estudantes.

No contexto da Eduacação a Distância (EAD), a incorporação da aprendizagem móvel exige a escolha de uma plataforma ou aplicativo apropriado que seja capaz de satisfazer as necessidades educativas específicas. Além disso, é necessário o desenvolvimento de conteúdo educativo otimizado para dispositivos móveis, levando em consideração os recursos e limitações destes aparelhos.

Fornecer apoio técnico e pedagógico aos alunos é igualmente importante para assegurar uma experiência de aprendizagem móvel eficaz.

Em resumo, a aprendizagem móvel representa uma abordagem de educação a distância que se utiliza de dispositivos móveis como instrumentos para fomentar uma aprendizagem flexível, personalizada e interativa.

Esta modalidade educativa oferece benefícios consideráveis, como o acesso flexível ao conteúdo, a personalização da aprendizagem, a interação entre os participantes e a utilização de recursos multimídia. A introdução da aprendizagem móvel no âmbito da EAD requer a seleção cuidadosa de plataformas e aplicativos, o desenvolvimento de conteúdo otimizado e um contínuo suporte aos alunos (31).

2.1.13 Pandas

O Pandas é uma biblioteca voltada para Ciência de Dados de código aberto, construída com base na linguagem *Python*. Ela oferece uma abordagem ágil e flexível, fornecendo estruturas robustas para a manipulação de dados relacionais ou rotulados, tudo de forma acessível e intuitiva.

A despeito de sua nomenclatura estar vinculada ao mamífero da família dos ursos, da mesma forma

que o nome *Python* é equivocadamente associado à espécie de serpente, o termo *Pandas* tem origem no conceito de *Panel Data*, um termo do inglês que está relacionado ao campo de estudo da econometria.

De maneira abrangente, o Pandas desempenha um leque diversificado de funções e processos. Isso abrange desde a depuração e o aprimoramento de dados até a exploração analítica de dados *Exploratory Data Analysis* (EDA), respaldo em tarefas de Aprendizado de Máquina, bem como a execução de consultas em bancos de dados relacionais, visualização de dados, *web scraping*, entre outras atividades.

Adicionalmente, ele se integra de maneira eficiente com diversas outras bibliotecas amplamente empregadas na área de Ciência de Dados, como *Numpy, Scikit-Learn, Seaborn, Altair, Matplotlib, Plotly, Scipy*, dentre outras (32)

2.1.13.1 Funcionamento do Pandas

Dentro do pacote Pandas, existem dois objetos primários significativos: as Series e os *DataFrames*. A fim de compreender essas estruturas de maneira mais aprofundada, é empregado um conjunto de dados denominado *Iris* como exemplo. Esse conjunto apresenta diversas informações relacionadas às características de espécies de flores Íris.

1 - Séries

As Séries são elementos do tipo *array* unidimensional, acompanhados por um eixo de rótulos, também chamado de índice, desempenhando o papel de identificador único para cada entrada. No conjunto de dados *Iris*, um exemplo ilustrativo de Séries pode ser encontrado ao isolar uma das variáveis para exibição, como, por exemplo, o comprimento da pétala *PetalLengthCm*. Ao fazer isso, é possível visualizar o formato apropriado, que se apresenta da seguinte maneira demonstrado na Figura 2.16 (32).

Fonte: Alura

```
0      1.4
1      1.4
2      1.3
3      1.5
4      1.4
      ...
145    5.2
146    5.0
147    5.2
148    5.4
149    5.1
Name: PetalLengthCm, Length: 150, dtype: float64
```

Figura 2.16: Series, *Pandas*

A coluna de números à esquerda dos espaços corresponde ao índice, enquanto os dados são exibidos à direita. Concluindo a exibição, uma breve descrição é fornecida, contendo informações sobre o nome, formato e tipo de dados contidos na Series.

2 - *DataFrames*

Os *DataFrames* são objetos bidimensionais de tamanho variável, apresentando um formato semelhante ao de uma tabela na qual os dados se organizam em linhas e colunas. Além disso, enquanto uma Series pode ser concebida como uma única coluna de dados, o *DataFrame* representa a reunião de várias Series distintas sob um índice comum. A estrutura característica do *DataFrame* é ilustrada na Figura 2.17 (32).

Fonte: Alura

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

Figura 2.17: *DataFrame*, Pandas

É possível trabalhar com a criação de cada uma dessas estruturas, utilizando os métodos do Pandas *pandas.DataFrame* e *pandas.Series*, em relação a estruturas nativas do *Python*, como listas, *arrays* e dicionários. Além disso, é viável lidar com a leitura e escrita de diversos tipos de arquivos de dados, tais como: CSV, Planilhas do *Excel, Parquet, SQL, HTML, JSON, XML* e muitos outros.

3 - Utilização

No dia a dia de um cientista de dados, observa-se uma forte predileção pelo uso do *Pandas* em conjunto com notebooks interativos em *Python*, notadamente nos arquivos .ipynb. Exemplos proeminentes desses *notebooks* incluem o renomado *Jupyter Notebook* e sua contraparte hospedada no *Google Colab*.

Essa prática conjunta exerce um papel de destaque, visando aprimorar a apresentação meticulosa do código e seus resultados, explorando, assim, a praticidade característica desse ambiente interativo. Nesse contexto, a habilidade de elaborar e executar código convive em harmonia, proporcionando ao cientista de dados a capacidade de observar quase que instantaneamente os resultados à medida que o código é

refinado. Essa interação fluida entre desenvolvimento de código e visualização é vividamente ilustrada por meio da representação visual na Figura 2.18 (32).

Fonte: Alura

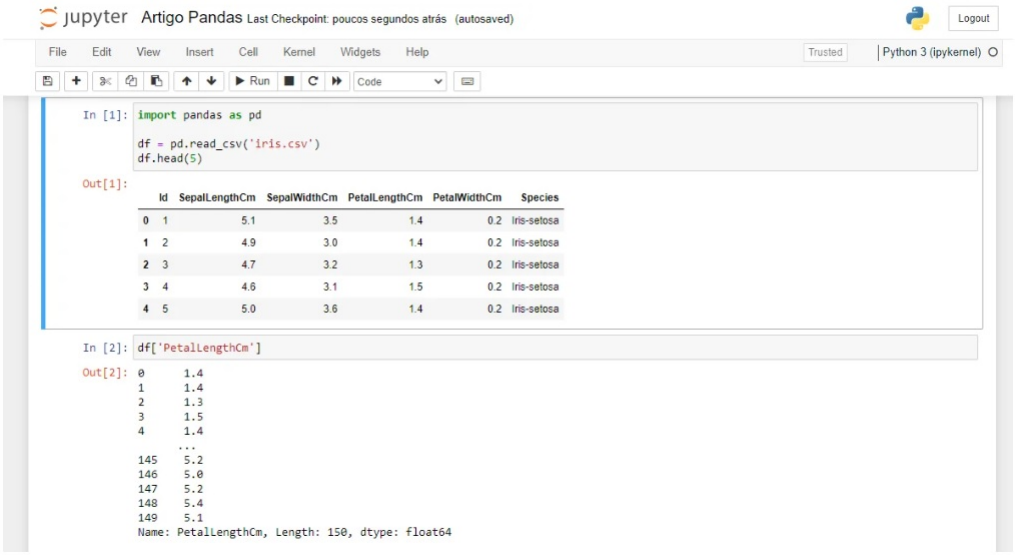

Figura 2.18: Linha de código, Pandas

Além das opções oferecidas pelos *Jupyter Notebooks*, é igualmente viável realizar operações por meio de *scripts Python* convencionais (arquivos.py). O contraste fundamental reside na maneira como os resultados provenientes dos trechos de código são apresentados no terminal. Nele, tais resultados são exibidos de forma contígua, sequencial e em formato *raw* ou bruto.

O subsequente exemplo ilustra como essa saída seria representada, se estivéssemos utilizando um *script* correspondente, no ambiente do terminal conforme demonstrado na Figura 2.19 (32).

Fonte: Alura

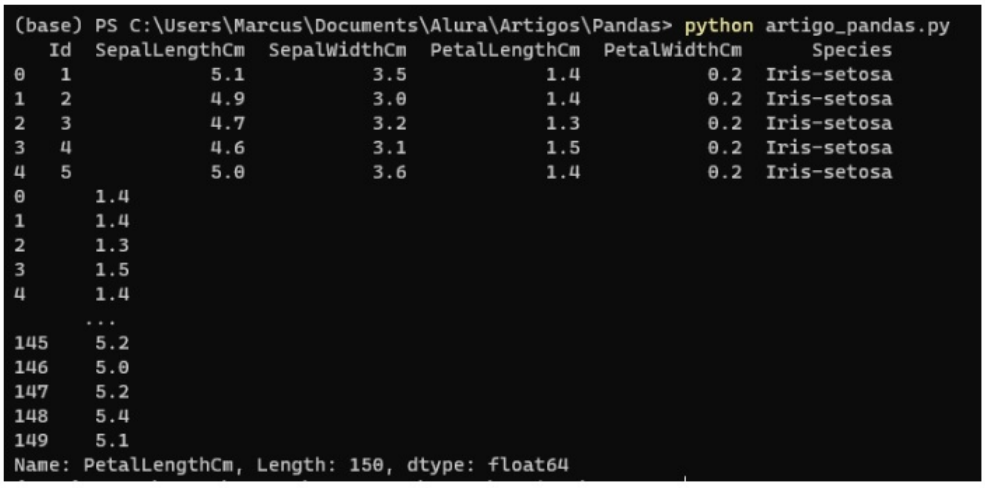

Figura 2.19: Linha de código, Terminal, Pandas

2.2 TRABALHOS RELACIONADOS

O presente estudo teve como objetivo realizar uma análise das informações de produção de uma grande empresa de mineração do Brasil. Essa análise visou demonstrar, por meio de ferramentas estatísticas, as informações que foram inseridas manualmente nos sistemas de produção. Essas inserções manuais estavam diretamente relacionadas aos resultados dos indicadores-chave de desempenho *Key Performance Indicators* (KPIs) da organização.

Além disso, o estudo investigou como esses resultados influenciam as decisões estratégicas baseadas nos referidos indicadores.

Para alcançar esses objetivos, os pesquisadores contrastaram evidências obtidas a partir das informações inseridas manualmente pelos operadores com os dados provenientes de um computador de bordo recentemente instalado nos equipamentos móveis da mina. Os resultados obtidos na análise confirmaram de maneira satisfatória a pergunta de pesquisa e a hipótese formulada pelos autores do estudo.

Ficou claro que a inserção manual de dados possui uma integridade questionável devido à possibilidade de erros no processo de coleta e armazenamento dos dados.

Utilizando a base de dados disponibilizada pela empresa, o estudo teve como propósito comparar duas amostras de informações de produção. Uma dessas amostras foi coletada manualmente, enquanto a outra foi obtida por meio de um sistema especializado que realiza a entrada de informações por meio de telemetria durante as operações.

Ambas as amostras foram suscetíveis a manipulações excessivas de dados, problemas de digitação e até mesmo questões relacionadas à fonte de aquisição dos dados (de (33).

A Arquitetura da Informação para Processos de Negócio é caracterizada pelo seu elemento imprescindível, que é a administração dos metadados de negócio. Ela estabelece um percurso para a gestão dos dados e das informações dos processos, e, nesse contexto, para a implementação da governança de dados.

Segundo a fonte citada, organizações que não possuem uma boa gestão de seus dados frequentemente também não gerenciam adequadamente seus metadados. A solução para esse desafio reside no fato de que a administração de metadados frequentemente serve como ponto inicial para aprimorar a gestão dos dados como um todo (34).

Os dados são elaborados, alimentados, relidos, analisados e transformados ao longo das atividades e procedimentos, e instituem-se em uma fonte de conhecimento essencial para o desenvolvimento das organizações. Neste caso, a Arquitetura da Informação para Processos de Negócio é apresentada como uma metodologia de Arquitetura da Informação que parte dos processos de negócio para obter os conhecimentos mais importantes dos mesmos e representá-los.

Isso é feito por meio de metadados de negócio, os quais são manifestados e descritos por atributos, configurando requisitos de informação. Em resumo, nota-se que a Arquitetura da Informação para Processos de Negócio deve desempenhar o papel de uma teoria na construção dos sistemas de informação que fornecem suporte ao negócio. Dessa forma, ela fornece os insumos para a gestão dos dados e conhecimentos para a governança de dados por meio de seus resultados (35).

O processo de tornar os governos transparentes para a sociedade por meio da divulgação de informações com o propósito de promover a responsabilização é potencializado quando Tecnologias da Informação e Comunicação são empregadas. O propósito do presente artigo consiste em identificar obstáculos à adoção das estratégias relacionadas à transparência, de acordo com a Política de Governança Digital que foi recentemente lançada pelo Governo Federal (36).

Objetiva-se também realizar uma revisão da literatura sobre a técnica de mineração de dados nas bases de dados abrangendo o Literatura Latino-Americana e do Caribe em Ciências da Saúde, *Scientific Eletronic Library Online* e alguns livros sobre o tema, buscando utilizar uma coleta que minere dados do período de 1999 a 2008.

Como critérios de exclusão, foram utilizados os descritores: indústria mineira, minas, mineralogia; foram excluídos artigos que não esclareciam o método e as tarefas relacionadas à mineração de dados. Observou-se que o volume de dados armazenados é gigantesco e continua crescendo exponencialmente.

Nas últimas décadas, em que a maior parte das operações e atividades das entidades privadas e públicas são registradas computacionalmente e se empilham em grandes bases de dados, a metodologia da mineração de dados – *Data Mining* – é uma das opções mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar as instituições nas tomadas de decisões mais rápidas ou, até mesmo, a atingir um maior grau de confiança.

Na esfera da saúde, em especial o público, a aplicação está sendo aceita como uma forma de acelerar a busca de entendimento. Todavia, para atender esse novo contexto, a informática em saúde vem apreendendo essas metodologias da ciência da computação para realizar seus estudos (37).

A expressão integridade da pesquisa vem sendo utilizada para demarcar um campo particular no interior da ética profissional do cientista, entendida como a esfera total dos deveres éticos, a que o cientista está submetido ao realizar suas atividades propriamente científicas.

Pretendo aqui explorar, em linhas gerais, o conceito de integridade da pesquisa e em seguida, esboçar um balanço de como se vem lidando no mundo com as questões relativas à integridade da pesquisa, visto que o pesquisador deve sempre visar a contribuição para a construção coletiva da ciência tal como um patrimônio coletivo, deve abster-se de agir, intencionalmente ou por negligência, de modo a impedir ou prejudicar o trabalho coletivo de construção da ciência e a apropriação coletiva de seus resultados (38).

A análise de dados é uma das etapas mais importantes em qualquer processo e determinado dentro de uma empresa ou órgão público. Mas para que tais informações sejam importantes e garantam retornos práticos para o negócio, é fundamental que a organização ou órgão público ateste a qualidade dos dados. Por isso, deve-se consultar a origem dos fatos, verificar a composição dos elementos, avaliar a consistência das informações disponíveis, entre outros procedimentos (39).

No contexto, o Governo Federal, por meio da EV.G, concentra um conjunto de informações sobre os treinamentos realizados, que podem despertar o interesse do público, como áreas temáticas, cursos realizados, demanda e perfil do público para os treinamentos ofertados, órgãos que utilizam esses treinamentos, número de funcionários treinados e outros.

Assim, quanto os esforços para tornar essas informações publicamente disponíveis, incentiva-se o con-

trole social e a análise das informações de acordo com as mais variadas necessidades. Até o momento, essa base unificada armazena cerca de 4,3 milhões de cadastros realizados entre os anos de 2006 e setembro de 2021.

Por meio de um portal, desde 2017, a EV.G adota a cultura de transparência ativa sobre o serviço prestado, disponibilizando publicamente suas informações sem a necessidade de solicitação prévia dos interessados (40) (41).

A adoção de princípios e diretrizes em ações de integridade de dados eletrônicos nas diferentes esferas da Fundação Getúlio Vargas (FGV), faz-se necessário, assim como, a compreensão de práticas relacionadas à integridade de dados durante todo o seu ciclo de vida nas atividades em laboratório de pesquisa científica experimental, seja em caráter orientativo ou visando a promoção da melhoria contínua.

O trabalho versou em um estudo exploratório na população constituída de treze unidades técnico-científicas e três escritórios técnicos da FGV que possuem laboratórios com atividades de ensaio, plataformas tecnológicas e/ou pesquisas experimentais básica ou aplicada. Verificou-se que a população estudada possui práticas que apoiam o processo de implementação de integridade de dados eletrônicos, mas com grau de adoção baixo em sua maioria, principalmente em medidas de controles técnicos (42).

A integridade dos dados refere-se à confiabilidade e consistência das informações ao longo do ciclo de vida útil delas. A finalidade é preservar as informações para que nada seja comprometido ou perdido. Devido à sua magnitude, a integridade de dados tornou-se o ponto principal de muitas elucidações de segurança.

O artigo disponibilizou os tipos de integridade de dados que as empresas e os órgãos públicos devem dar atenção especial. São mencionados dois tipos de integridade de dados, um deles subdividido em mais quatro categorias:

1. Integridade física: Caracterizada pela precisão dos dados à medida que são armazenados e recuperados. Geralmente, essa integridade é afetada por situações como desastres naturais ou ataques de hackers, que limitam as funções do banco de dados.

2. Integridade lógica: Mantém os dados imutáveis, pois são usados de diferentes modos em um banco de dados relacional, prevenindo erros humanos futuros ou ataques de hackers. Dentro da integridade lógica, há quatro tipos adicionais de identidade lógica:

- Integridade da entidade: Garante que as informações contidas nas tabelas, colunas e linhas do banco de dados sejam precisas e relevantes.
- Integridade referencial: Declara uma série de processos para garantir que os dados armazenados sejam usados de forma uniforme, seguindo regras específicas no banco de dados, como a criação de chaves de acesso para permitir apenas alterações, adições e exclusões adequadas. Inclui restrições para evitar dados repetidos e garantir a precisão dos dados inseridos.
- Integridade de domínio: Representa processos que garantem a exatidão de cada parte dos dados em um domínio, estabelecendo limites para os valores que cada coluna pode conter. Por exemplo, definindo o número de casas decimais para valores monetários.

- Integridade definida pelo usuário: Os usuários podem criar regras específicas para proteger ainda mais os dados e garantir sua preservação.

Os principais fatores que podem influenciar a integridade dos dados armazenados são: erro humano, inserção incorreta, duplicação ou exclusão de dados pelos usuários e o não cumprimento das regras de segurança estabelecidas pelos órgãos. Erros de transferência também podem ocorrer quando os dados não são transferidos adequadamente de um banco de dados para outro. Além disso, a presença de *bugs* e vírus pode comprometer os dados, assim como falhas inesperadas no *hardware* dos computadores ou servidores (43).

Na competência do Estado, a informação é, de fato, um dever da Administração Pública e um direito sancionado do cidadão. De fato, no Estado Democrático de Direito, toda e qualquer ação da Administração deve se submeter ao processo vasto de justificação e fundamentação perante sociedade (44).

O Senado Federal ampliou suas políticas de transparência e acesso à informação, sendo destacada a criação da Secretaria da Transparência, responsável pela divulgação de informações públicas de forma democrática e inclusiva. Em 2011, o Senado assumiu a Lei de Acesso à Informação (LAI), Lei nº 12.527, que cumpre o mandamento constitucional de garantir aos cidadãos o direito de receber informações dos órgãos públicos.

A LAI promove a publicidade como princípio geral, tornando exceção as práticas sigilosas, e esforços do Instituto Legislativo Brasileiro (ILB), da Consultoria Legislativa e do Núcleo de Estudos e Pesquisas do Senado, sem custos adicionais para os cofres públicos, possibilitaram a elaboração de uma cartilha didática para explicar os principais aspectos da Lei de Acesso à Informação ao público (41).

3 METODOLOGIA, SOLUÇÃO E PROPOSTA

Esta seção detalha os métodos para verificar a integridade dos dados e os estende à abordagem de aprendizado de máquina. Conforme ilustrado na Figura 3.1, a análise é dividida em quatro blocos funcionais.

Fonte: Elaborado pelo autor (2022)

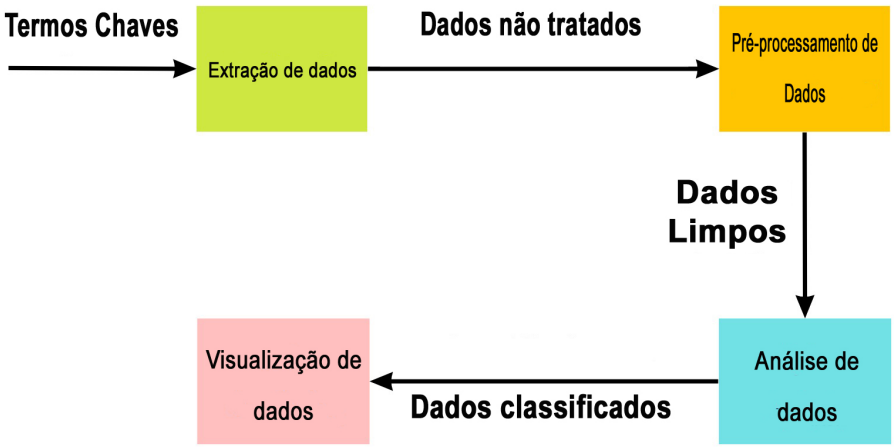

Figura 3.1: Diagrama de analise do EmNumeros

3.1 EXTRAÇÃO DE DADOS

Em 2013, a EV.G contava com um catálogo de 18 cursos a distância, capacitando um total de 38 mil alunos. Nos anos subsequentes, ocorreu um crescimento exponencial, com matrículas de 69 mil alunos em 2014 e 145 mil em 2015.

Os esforços de expansão foram continuados e, em 2020, a plataforma alcançou aproximadamente mais de 1 milhão e 600 mil alunos. No ano seguinte, em 2021, foi alcançado um marco ainda maior, com mais de 1,6 milhão de alunos matriculados. Ao somar todas as inscrições ao longo dos anos, a EV.G atingiu um impressionante total de mais de 5,5 milhões de alunos inscritos, representando uma amostra significativa do público atendido.

A trajetória de crescimento evidencia a relevância dos cursos a distância oferecidos pela EV.G no cenário educacional. A capacitação de milhões de alunos demonstra o impacto positivo que a plataforma teve na formação e na aquisição de novos conhecimentos para um amplo público, representando uma amostra representativa dos beneficiados.

Além disso, é relevante mencionar que a EV.G possui uma estrutura organizada para fornecer informações relevantes sobre o perfil dos alunos matriculados. O portal apresenta painéis com filtros que possibilitam análises detalhadas, incluindo informações sobre as cidades de origem dos alunos. Destaca-se a presença significativa de matrículas provenientes das capitais das regiões Nordeste, Sudeste e Sul.

A plataforma também procura identificar a frequência das matrículas, ou seja, o percentual de alunos que frequentam cursos da EV.G mais de uma vez. Essa abordagem tem o objetivo de compreender o grau de satisfação e interesse dos estudantes em retornar para aprimorar seus conhecimentos e habilidades, utilizando uma amostra representativa do comportamento das matrículas.

A reestruturação dos painéis proporciona uma experiência mais dinâmica e intuitiva aos usuários, permitindo uma visualização clara das matrículas e dos cursos realizados no âmbito da EV.G. Com base no *feedback* recebido, a plataforma está em constante busca pela melhoria de seus serviços, fornecendo informações visualmente atrativas que são de interesse tanto para a gestão da EV.G quanto para qualquer indivíduo interessado em se beneficiar dos cursos oferecidos, como demonstrado na Figura 3.2.

Fonte: Painel EmNumeros

Figura 3.2: Números de inscritos na EV.G

3.1.1 Pré-processamento de dados

O pré-processamento dos dados foi realizado utilizando *Pandas*, uma biblioteca de *software* criada para a linguagem *Python* para manipulação e análise de dados, no qual os dados inconsistentes e os *outliers* encontrados em alguns cursos foram eliminados. É importante ressaltar que as informações pessoais contidas no conjunto de dados foram excluídas. No que diz respeito aos *outliers*, adotou-se a abordagem do *Boxplot*, que permite visualizar de forma eficaz as faixas medianas e interquartis. A Figura 3.3 ilustra um exemplo dos *outliers* referentes à idade dos alunos matriculados.

A remoção dos *outliers* foi conduzida utilizando a métrica *Z-Score*, a qual elimina observações que se encontram acima da média. Essa medida foi empregada visando evitar o impacto excessivo dessas condições atípicas na estimativa (45).

Fonte: Elaborado pelo autor (2022)

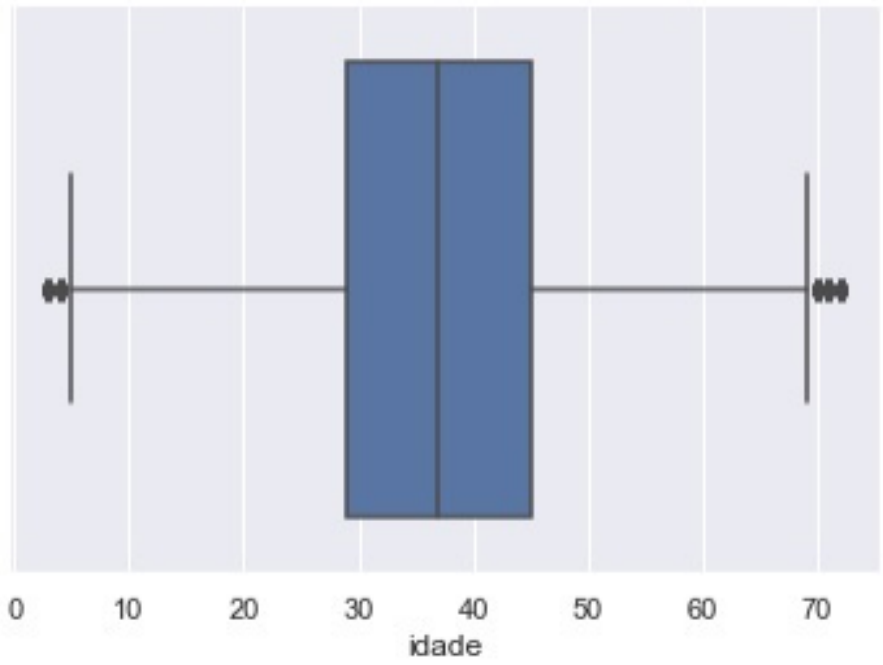

Figura 3.3: *Boxplot* com os dados antes do pré-processamento de dados

3.2 ANÁLISE DE DADOS

No estudo, uma ferramenta foi desenvolvida pela equipe para prever o número de alunos que concluem o curso. Os pesquisadores aplicaram algoritmos de aprendizado de máquina supervisionados após realizar a rotação dos dados. Eles deram preferência a algoritmos de regressão e optaram por usar o Algoritmo da *Random Forest* para análise. Para determinar a quantidade adequada de estimadores, um cálculo foi realizado com base em uma matriz que continha os valores [4, 8, 16, 32, 64, 128, 256] (46).

Além do algoritmo de aprendizado de máquina mencionado anteriormente, os pesquisadores também compararam as SVN e as ANNs com duas camadas ocultas. Eles obtiveram os resultados do treinamento para as quatro classes registradas na primeira interação: Abandono, Completo, Falha e Travado não Concluído. Na segunda etapa do treinamento, as classes foram agrupadas em apenas duas categorias: "Aprovado", que incluiu o status de conclusão, e "Não Aprovado", que englobou abandono, falha, bloqueio e não preenchido.

A abordagem das ANNs neste estudo consistiu em quatro camadas, com a camada de entrada contendo cinco perceptrons. A estrutura proposta das ANNs, como mostrada na Figura 3.4, foi classificada como binária (Aprovado/Falho), utilizando uma única camada de saída com um perceptron. Se a saída fosse

menor que 0,5, o aluno era considerado aprovado; caso contrário, era considerado como falho.

Cada camada recebeu uma matriz de características organizadas em colunas (por exemplo, Idade do Estudante, Duração do Curso, Sexo do Estudante e Carga Horária do Curso), que foram rotuladas como I = 1, 2, 3, 4. Cada uma dessas características possuía várias entradas, sendo representada como um vetor (1).

Figura 3.4: Fonte: Tiago P. Oliveira, Jamil S. Barbar e Alexandre S. Soares (1)

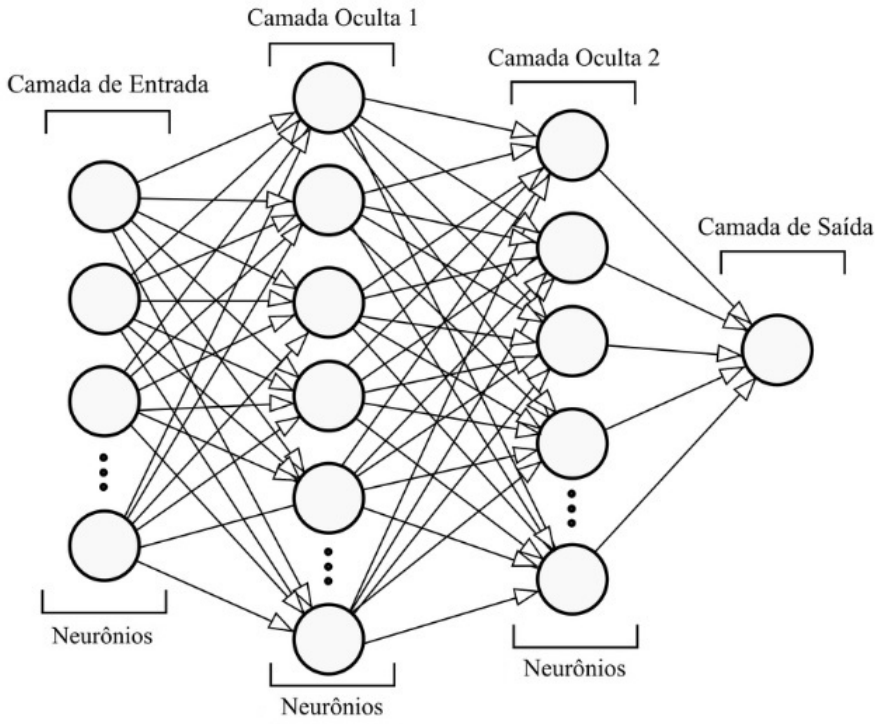

Figura 3.5: Arquitetura ANN

3.3 VISUALIZAÇÃO DE DADOS

Os dados obtidos do site Emnumeros foram empregados na criação de gráficos para auxiliar os responsáveis por cada curso a acompanhar as métricas e tomar medidas no momento apropriado. Uma vez que o portal mencionado oferece análises visuais, foi necessário desenvolver uma forma adicional de análise visual e uma maneira de compreender o comportamento dos alunos em certos cursos.

A iniciativa conhecida como EV.G é coordenada pela Enap e visa centralizar a oferta de cursos de capacitação profissional a distância para servidores públicos de todo o país, além de cidadãos brasileiros e estrangeiros interessados.

A EV.G possui um sistema próprio de gestão acadêmica disponível em <www.escolavirtual.gov.br>, por meio do qual os interessados podem identificar as capacitações disponíveis, inscrever-se para participar e realizar as atividades relacionadas para completar os cursos. A partir desse ambiente, os alunos também podem acessar seus certificados de conclusão de curso, e os interessados podem verificar a validade dos

documentos emitidos pela EV.G.

Dentro desse contexto, a EV.G centraliza diversas informações sobre as capacitações realizadas, que podem despertar interesse público, como áreas temáticas e cursos oferecidos, demanda e perfil do público-alvo, principais órgãos que utilizam essas capacitações, quantidade de servidores capacitados, entre outros. Portanto, ao dedicar esforços para disponibilizar essas informações ao público, promove-se não apenas o controle social, mas também a análise das informações de acordo com diversas necessidades.

3.3.1 Problema

Durante o processo de migração do site do EmNumeros para uma nova versão desenvolvida com a ferramenta *Power BI*, foram discernidos diversos problemas. Um dos principais desafios encontrados residia na necessidade de atualizar manualmente os dados nas visualizações construídas nos painéis, em função da utilização da versão gratuita no projeto de desenvolvimento.

Adicionalmente, a nova abordagem não lograva identificar eventuais casos de abandono ou cursos não concluídos por parte dos estudantes, o que revelava uma limitação no controle destes aspectos.

Para tratar da estimativa do número de estudantes que finalizariam o curso, foi sugerida a implementação de uma ferramenta empregando algoritmos de aprendizado de máquina, com a seleção de um algoritmo de regressão para tal finalidade.

Além disso, foram exploradas ANNs com SVM apresentando duas camadas ocultas. Os desfechos do treinamento foram incorporados às quatro categorias definidas na primeira iteração, com classificações como "abandonado", "completo", "fracassado", "bloqueado"e "incompleto".

Na fase subsequente de treinamento, optou-se por agrupar as aulas em apenas dois conjuntos: "aprovados", englobando os cursos concluídos com sucesso, e outra categoria que abrangia situações não aprovadas, tais como abandono, fracasso, bloqueio e ausência de preenchimento.

Entretanto, vale ressaltar que os dados vinculados às formações conduzidas na EV.G suscitam interesses diversos, desde pesquisadores que almejam aprofundar suas investigações sobre aprendizado contínuo em serviço, até órgãos da Administração Pública que disponibilizaram tais informações para embasar decisões administrativas.

3.3.2 Solução

As visualizações do portal, que englobam os dados do sistema legado *WebCef* e os dados do sistema acadêmico atual (Secretaria Virtual), foram construídas por meio da utilização da ferramenta *Business Intelligence* (BI) conhecida como *Tableau*.

A partir de sua concepção original, o portal foi desenvolvido com três painéis principais: 1 - Indicadores gerais das inscrições efetuadas, 2 -A companhamento da evolução dos cursos oferecidos pela entidade EV.G e 3 - Perfil detalhado dos estudantes inscritos.

Ao longo dos três anos de existência do projeto EmNumeros, a experiência adquirida no uso dos referidos painéis, juntamente com a remodelação das bases de dados e os avanços contínuos nas capacidades das

ferramentas de visualização, trouxeram à tona a necessidade de reformular tanto a maneira de apresentar os dados quanto as próprias informações disponibilizadas.

Dentro desse contexto, as seções subsequentes oferecem uma análise das visualizações originais e das novas visualizações desenvolvidas, oferecendo uma exploração detalhada dos painéis atualmente expostos no portal da EV.G intitulado EmNumeros.

Cada painel é equipado com um conjunto de filtros que conferem aos usuários a flexibilidade necessária para conduzir suas próprias análises. Isso permite a aplicação de filtros aos gráficos de acordo com o ano de inscrição (no período de 2006 a 2021), a esfera governamental (federal, estadual ou municipal), a esfera de poder (Executivo, Judiciário ou Legislativo), tema, curso e tipo de oferta.

A reestruturação dos painéis resultou na reorganização das informações, empregando abordagens de visualização mais dinâmicas para representar as matrículas e os cursos administrados pela EV.G.

Adicionalmente, com base no *feedback* obtido sobre as informações previamente divulgadas, o portal agora está apto a fornecer melhorias nos serviços oferecidos, apresentando informações de maneira visualmente acessível tanto para a administração da EV.G quanto para qualquer cidadão interessado. A Figura 3.5, exibida abaixo, ilustra a página principal do portal EmNumeros da EV.G, apresentando os pilares temáticos e suas descrições correspondentes.

Portal EV.G EmNumeros

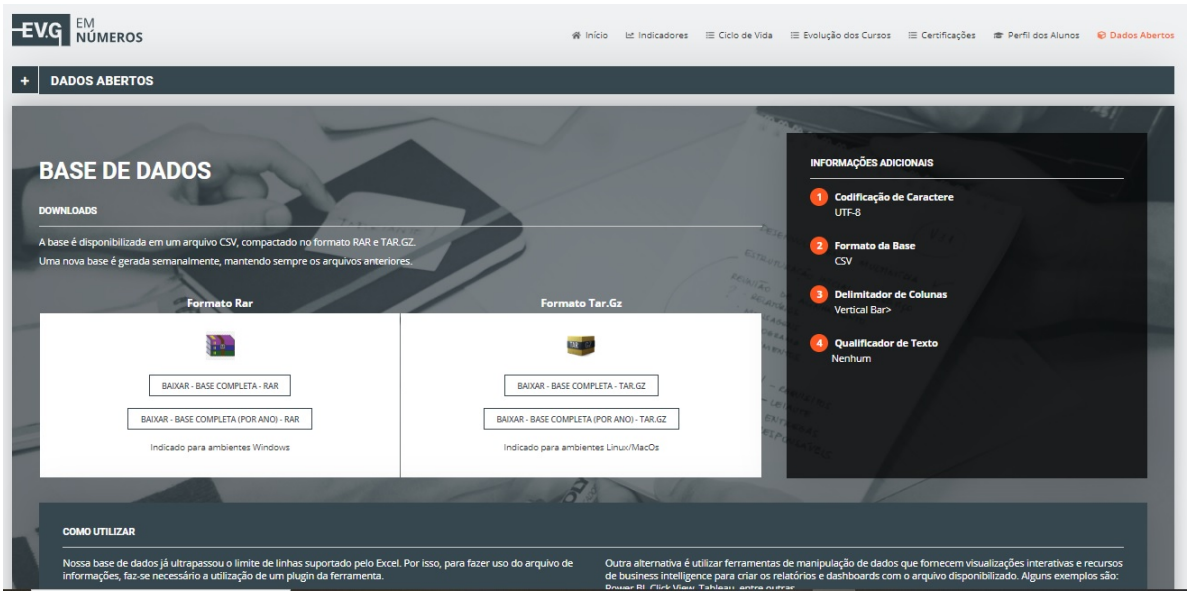

Figura 3.6: Base de dados abertos EmNumeros

A Enap, por meio da iniciativa EV.G, oferece à comunidade painéis que apresentam dados de forma visual e compreensível, possibilitando aos usuários entender o perfil dos inscritos e as principais características dos cursos oferecidos pela EV.G Esses *dashboards* têm o potencial de auxiliar os gestores responsáveis pela formação de pessoal da administração pública em seus processos decisórios, entre outras possibilidades.

Além disso, pesquisadores acadêmicos também podem se beneficiar dos dados fornecidos para apro-

fundar seus estudos no campo da educação continuada, principalmente no que diz respeito aos cursos de curta duração oferecidos pela EV.G

Atualmente, os painéis estão hospedados no endereço https://emnumeros.escolavirtual.gov.br/ , e foram construídos com a ferramenta *Tableau*. O *Tableau* é um *software* de BI que oferece aos usuários a capacidade de coletar, organizar, analisar, compartilhar e apresentar informações.

Essa ferramenta permite a criação de painéis de diversas formas, utilizando diversos tipos de gráficos embutidos, como tabelas, mapas, barras, entre outros. Além disso, a ferramenta disponibiliza filtros interativos que auxiliam o usuário na seleção das informações de forma mais detalhada.

O site EmNumeros disponibiliza quatro painéis distintos com informações diversas sobre as matrículas nos cursos oferecidos pela Enap, tanto no período pré-EV.G (2006-2017), que abrange o período anterior à criação formal do EV.G, quanto no período EV.G (2018-2021), que apresenta dados de matrículas em cursos desenvolvidos no âmbito institucional da EV.G.

O site EmNumeros foi desenvolvido através da compilação de informações relacionadas às inscrições nos cursos realizados por meio da plataforma EV.G. Os conteúdos do site foram organizados em diferentes seções, com foco nos seguintes aspectos:

- Indicadores: Nesta seção, são apresentadas informações gerais sobre as inscrições e os inscritos nos cursos. As inscrições são categorizadas por critérios como região, tema do curso, e outros.
- Evolução dos Cursos: Esta área destaca a trajetória de crescimento dos cursos ao longo do tempo, demonstrando o volume de inscrições realizadas. Além disso, ela destaca os cursos mais populares e suas respectivas temáticas.
- Certificações Avançadas: A ênfase nesta seção está nas certificações avançadas. Estas englobam um conjunto de certificações que os usuários podem adquirir ao concluírem uma série de cursos relacionados.
- Perfil dos Alunos: O objetivo desta parte é criar perfis dos participantes, consolidando informações sobre a faixa etária, situação de emprego, local de origem e outros dados pertinentes relativos aos inscritos nos cursos oferecidos.

Esses painéis proporcionam uma visão ampla e minuciosa das informações ligadas aos cursos disponibilizados pela plataforma EV.G. Isso permite uma análise mais profunda e fundamentada tanto para os gestores encarregados da capacitação quanto para os pesquisadores interessados no campo da educação contínua.

3.3.2.1 Indicadores

O painel Indicadores, como mostrado na Figura 3.6, apresenta a ele um amplo leque de informações relacionadas aos inscritos nos cursos oferecidos pela EV.G e suas inscrições. A visualização proporciona uma perspectiva abrangente por meio da apresentação de diversos cartões que consolidam os dados relevantes.

As informações exibidas incluem o total de inscrições realizadas durante o período selecionado, a contagem das inscrições feitas no atual ano e o número de indivíduos que matricularam-se em pelo menos um curso na EV.G. Esses cartões, por sua vez, concedem uma visão geral dos indicadores-chave, permitindo uma análise mais precisa e facilitando a compreensão dos dados associados à audiência e ao envolvimento com os cursos disponibilizados pela EV.G.

Fonte: Painel Indicadores EmNumeros - EV.G

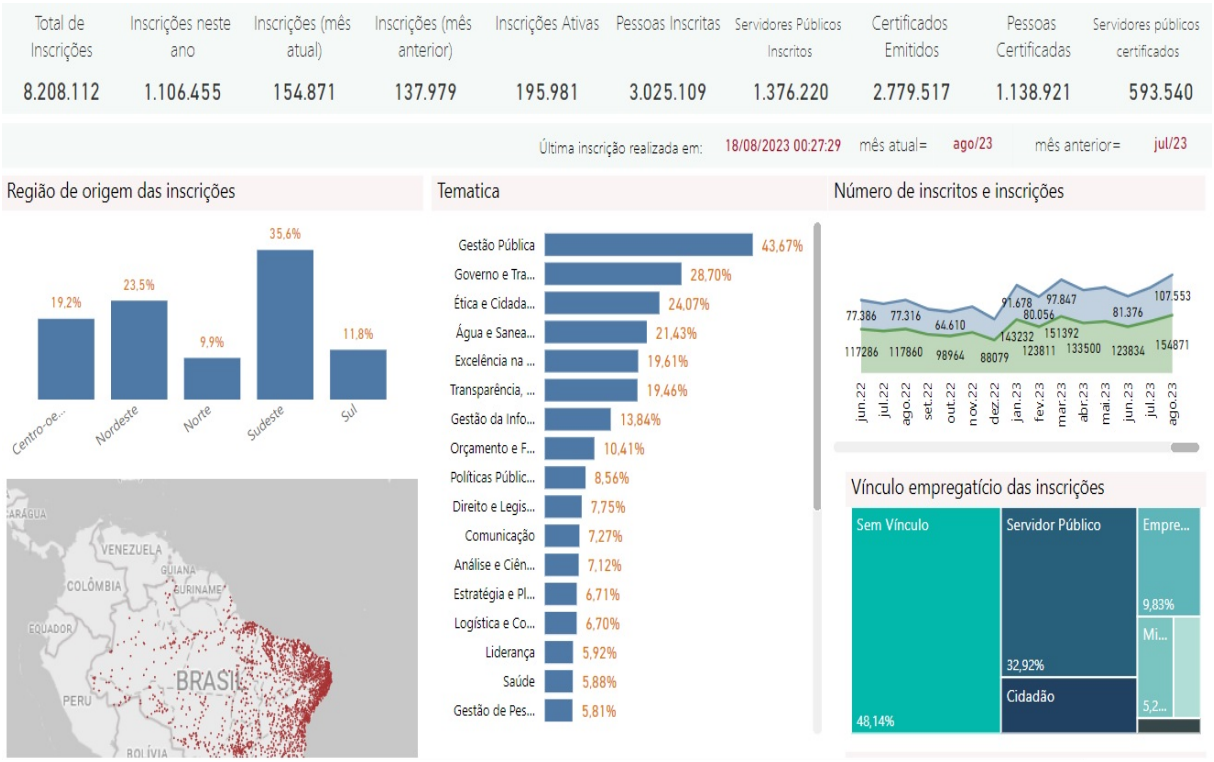

Figura 3.7: Indicadores EmNumeros

O painel Indicadores também fornece uma gama de informações detalhadas a respeito das inscrições, como quantidade de inscrições por região, tema, tipo de vínculo empregatício e status de matrícula. Essas informações possibilitam uma rápida identificação do desempenho dos cursos, consolidando o número de inscritos que concluíram, desistiram ou estão em andamento.

Dentro desse mesmo painel, encontra-se um mapa que apresenta as cidades de origem dos inscritos, com base na cidade de nascimento fornecida, e mostra o número de inscrições originárias de cada localidade. Adicionalmente, é exibido um gráfico que ilustra o número de inscrições por pessoa, permitindo identificar a frequência com que os indivíduos se inscrevem nos cursos oferecidos.

Além das funcionalidades mencionadas, o painel oferece filtros que permitem uma seleção mais detalhada dos dados, incluindo ano de inscrição, período (pré-EV.G ou EV.G), esfera de poder (Executivo, Judiciário e Legislativo), esfera de governo (estadual, municipal e federal), Unidade Federativa (UF) de nascimento, órgão de origem, tema do curso e entidade responsável pelo desenvolvimento dos cursos (conteudista).

Essa diversidade de informações e recursos interativos presentes no painel Indicadores proporciona

uma análise aprofundada dos dados, permitindo uma compreensão mais precisa e embasada para os gestores encarregados da capacitação, bem como para os pesquisadores interessados no campo da educação continuada.

3.3.2.2 Evolução dos Cursos

Conforme mencionado anteriormente, a segunda seção do site EmNumeros, que é representada na Figura 3.7, apresenta uma visualização gráfica da evolução do número de inscrições por curso e temática. Essa visualização abrange o período desde o início dos dados em 2006 até o momento atual. A referida seção é identificada como o painel "Evolução dos Cursos"dentro da plataforma EV.G no site EmNumeros.

Fonte: Painel Indicadores EmNumeros - EV.G

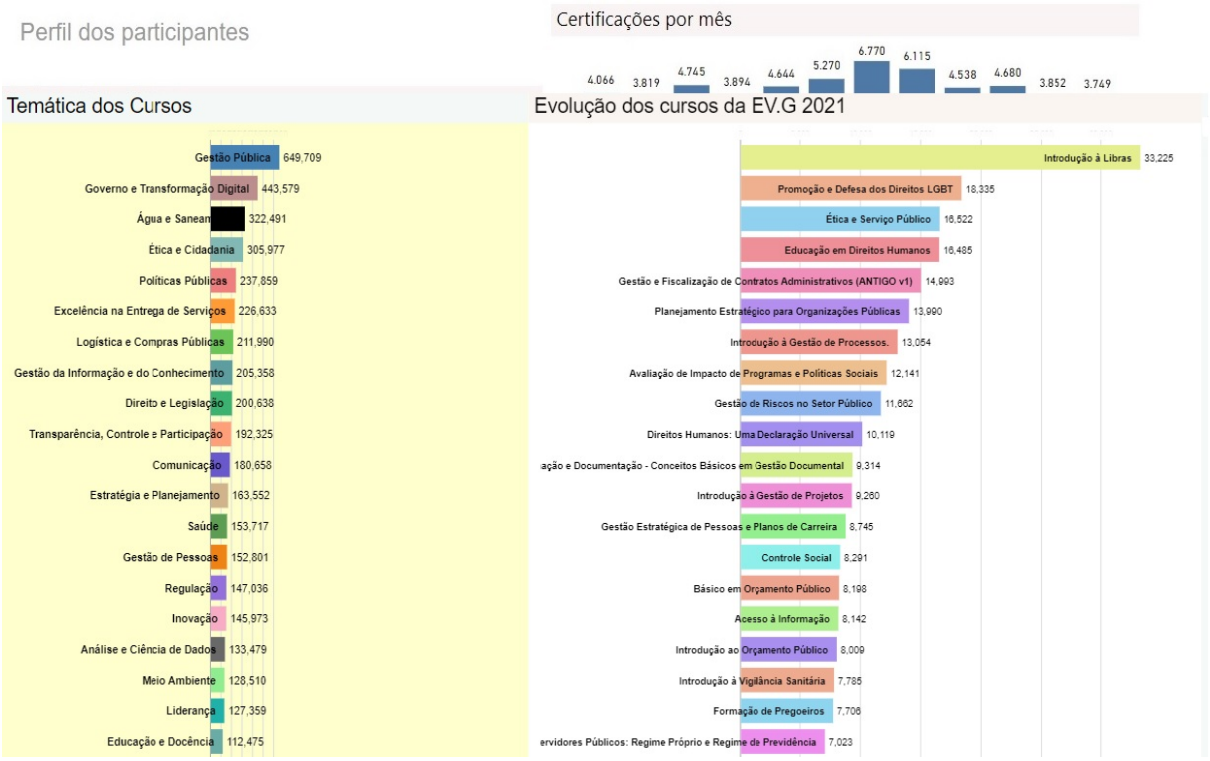

Figura 3.8: Indicadores, Evolução dos Cursos

No lado esquerdo do painel, há um gráfico de barras que organiza as temáticas relacionadas aos cursos, exibindo o número de inscrições e sua representação percentual em relação ao total de inscrições. A representação visual presente no gráfico permite que se compreenda imediatamente a distribuição das inscrições de acordo com as diversas temáticas abordadas.

No lado direito, encontra-se outro gráfico de barras que ilustra a evolução dos cursos ao longo dos anos. Ele destaca, em ordem decrescente, os cursos que tiveram o maior número de inscrições a cada ano no intervalo de 2006 a 2021.

Ambos os gráficos são coloridos de acordo com as temáticas correspondentes, conforme indicado na legenda de cores que está presente no painel. Essa abordagem visual torna mais fácil a rápida e clara

identificação das áreas temáticas mais populares ao longo do tempo. Adicionalmente, o painel oferece um filtro que possibilita a seleção de um ano específico.

Também possui um sistema de controle de reprodução *play/pause* que permite uma interação dinâmica com a análise da evolução dos cursos, em termos do número de inscrições realizadas.

Os recursos visuais e interativos do painel de "Evolução dos Cursos"proporcionam uma compreensão abrangente e detalhada do crescimento e da popularidade dos cursos ao longo dos anos. Isso permite uma análise aprofundada das tendências e padrões de inscrições. Essas informações são valiosas tanto para os gestores encarregados de tomar decisões quanto para os pesquisadores que têm interesse em investigar o campo da educação continuada oferecida pela instituição EV.G.

3.3.2.3 Certificações Avançadas

Na EV.G, as certificações avançadas são constituídas por conjuntos de cursos interligados que se complementam, culminando em uma formação sólida sobre um tópico específico. Ao completar com sucesso todos os cursos que compõem uma certificação avançada, o usuário não somente obtém certificações individuais para cada curso, mas também é conferido um certificado abarcando a carga horária total da certificação avançada escolhida.

O propósito dessas certificações avançadas é proporcionar aos usuários uma expertise aprofundada em um campo de conhecimento específico, abordando diversas facetas e tópicos relacionados. Ao finalizar o conjunto completo de cursos, o usuário adquire uma qualificação mais ampla, reconhecida por meio da certificação avançada, que realça sua capacitação e maestria na respectiva área.

A abordagem de certificações avançadas na EV.G enaltece a aprendizagem contínua e a busca por conhecimento especializado, motivando os usuários a expandirem suas competências de maneira holística e a se destacarem no mercado de trabalho. A emissão de certificados individuais para cada curso, assim como um certificado abrangente para a certificação avançada, oferece um reconhecimento oficial do empenho e comprometimento do usuário em aprimorar suas habilidades e conhecimentos em uma área determinada.

A seção de Certificações Avançadas, ilustrada na Figura 3.8, proporciona uma visão geral dos dados pertinentes às certificações avançadas na EV.G. Nesse segmento, os usuários têm a possibilidade de observar as certificações conquistadas, organizadas por mês no ano escolhido e região de origem do inscrito, entre outras opções de filtragem disponíveis.

Fonte: Painel Indicadores EmNumeros - EV.G

Figura 3.9: Indicadores, Certificações Avançadas

Nesta área, são apresentados os quantitativos de inscritos por certificação, agrupados segundo faixa etária e certificações completas. Esta análise proporciona uma compreensão mais minuciosa do perfil dos inscritos em relação às certificações avançadas, destacando a distribuição por faixa etária e o número de certificações completas.

Tal como nos painéis anteriores, o painel de Certificações Avançadas oferece opções de filtragem que permitem uma seleção mais precisa dos dados, possibilitando análises personalizadas feitas pelos próprios utilizadores da plataforma. Estes filtros abrangem o ano em que a certificação foi obtida, bem como as esferas de poder e entidades governamentais às quais os inscritos estão afiliados.

Os recursos de filtragem e visualização disponíveis no painel de Certificações Avançadas viabilizam uma análise mais aprofundada e personalizada dos dados, ajudando tanto os gestores encarregados da EV.G quanto os utilizadores da plataforma a compreender o panorama das certificações avançadas e a identificar tendências e padrões relevantes relacionados com a formação dos utilizadores.

3.3.2.4 Perfil dos Alunos

Na seção denominada "Perfil dos Alunos", conforme ilustrado na Figura 3.9, são exibidos vários indicadores relacionados aos estudantes matriculados nos cursos. Esses indicadores englobam o total de inscrições, tanto ao longo do tempo quanto no ano vigente e nos meses anteriores e atuais. Também são fornecidos dados sobre o número de indivíduos atendidos, juntamente com a quantidade total de funcioná-

rios públicos que se inscreveram nos cursos disponibilizados pela EV.G. Adicionalmente, é apresentada a quantidade de cursos concluídos.

Esses indicativos proporcionam uma visão holística do perfil dos estudantes, destacando tanto o fluxo de inscrições ao longo do período, quanto a extensão do alcance da EV.G no que diz respeito às pessoas atendidas e aos funcionários públicos capacitados. Essas informações são de importância crucial para a avaliação do impacto e abrangência dos programas de capacitação promovidos pela EV.G, oferecendo informações pertinentes sobre o contingente de participantes e o grau de conclusão dos cursos.

Por meio da análise desses indicadores, gestores e pesquisadores têm acesso a dados valiosos relacionados à aceitação dos estudantes, à procura por oportunidades de formação e ao desempenho dos programas de educação contínua. Essa análise minuciosa do perfil dos estudantes contribui para a tomada de decisões estratégicas, o refinamento dos cursos e a otimização dos benefícios proporcionados pela EV.G no âmbito da Administração Pública.

Fonte: Painel Indicadores EmNumeros - EV.G

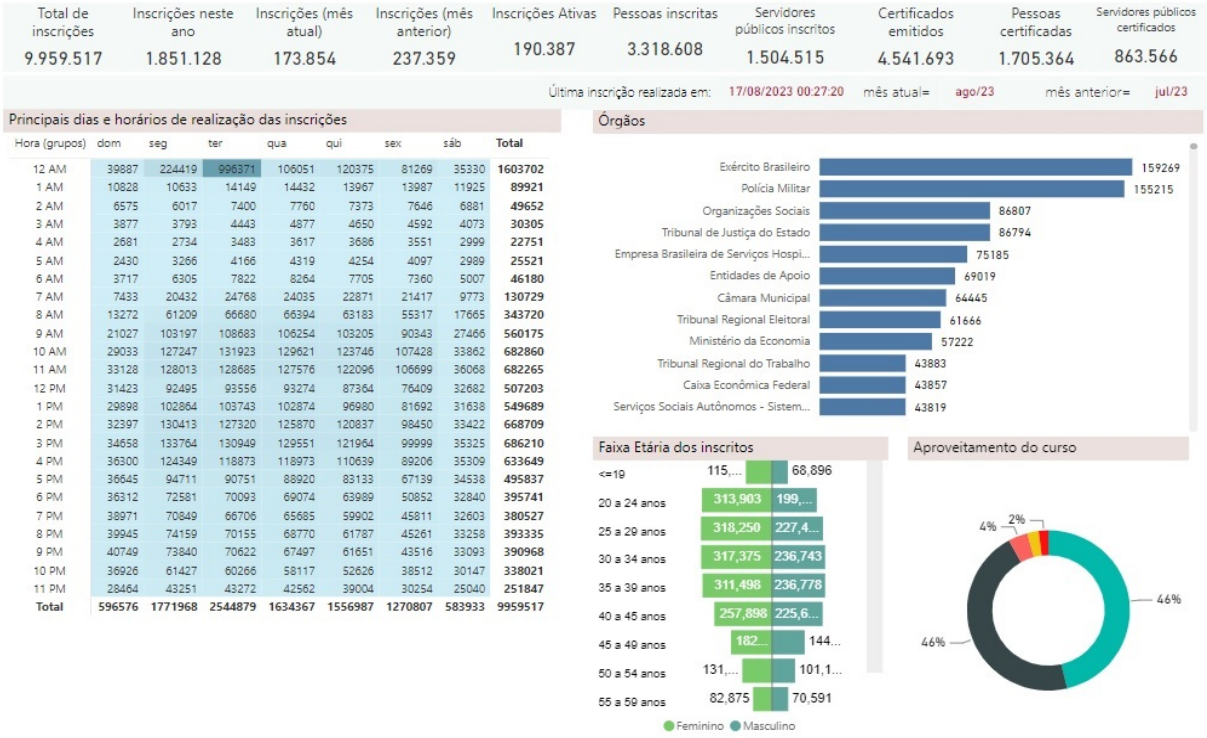

Figura 3.10: Indicadores, Perfil dos alunos

Na área do Perfil dos Alunos, encontra-se igualmente uma tabela que realça os principais dias e horários nos quais as inscrições são efetuadas. Isso viabiliza a identificação dos momentos de maior procura pelos cursos disponibilizados pela EV.G. Adicionalmente, apresenta-se uma pirâmide etária dos inscritos, oferecendo informações acerca da distribuição de idade e gênero dos inscritos.

Além disso, é possível determinar os órgãos públicos com maior quantidade de inscrições nos cursos da EV.G, bem como analisar a utilização dos cursos ao longo dos anos selecionados. Estes dados providenciam informações valiosas sobre o envolvimento de várias entidades governamentais e o nível de

aproveitamento das capacitações propostas pela EV.G em cada ano.

No painel do Perfil dos Alunos, encontram-se disponíveis diversos filtros de seleção para conduzir análises mais específicas. Estes filtros incorporam o período de inscrição (Pré-EV.G: 2006-2017 e EV.G: 2018-2021), o ano de inscrição, a esfera de competência, a esfera governamental, a UF de nascimento, o órgão público, a temática do curso, o curso propriamente dito, a situação da oferta, a categoria de situação e o responsável pelo conteúdo dos cursos.

A utilização destes filtros possibilita uma análise minuciosa e personalizada dos dados, permitindo que os gestores, investigadores e utilizadores da plataforma explorem as informações de acordo com as suas necessidades particulares. Esta flexibilidade na filtragem contribui para a obtenção de dados pertinentes acerca da participação nos cursos, da distribuição demográfica dos inscritos, do desempenho das entidades públicas e de outros aspetos ligados à capacitação na EV.G.

3.3.3 Comparação dos algoritmos

Conforme mencionado na Seção 3, três algoritmos distintos foram empregados para efeitos de comparação. A análise da consistência dos dados de um curso específico foi realizada, com foco nas características selecionadas do conjunto de dados: sexo, carga horária do curso, idade do estudante e duração do curso. A escolha destes atributos resultou de uma Análise de Componentes Principais, que proporcionou *insights* sobre suas interações.

A etapa subsequente envolveu o treinamento do modelo de aprendizado de máquina utilizando vários classificadores distintos. A divisão do conjunto de dados seguiu a proporção de 70% para treinamento, 20% para testes e 10% para validação.

A avaliação dos algoritmos foi conduzida com base na métrica de precisão, uma vez que a problemática contava com apenas duas variáveis objetivas: Aprovado e Reprovado. A escolha pela métrica de precisão se deu em virtude de sua objetividade nesse contexto. Também foram empregadas medições de precisão para a avaliação dos algoritmos, as quais ofereceram uma avaliação dos pontos de dados relevantes.

É crucial evitar a classificação equivocada de alunos que efetivamente foram reprovados como aprovados, embora o modelo sugerido os tenha previsto como aprovados. A Equação II formaliza essa métrica, definindo-a como a proporção de verdadeiros positivos sobre a soma de verdadeiros positivos e falsos positivos.

Uma vez que a configuração do ambiente de teste foi estabelecida, a métrica escolhida pôde ser calculada para os três algoritmos sob avaliação. A performance de cada um destes é apresentada na Tabela 3.1, onde ocorre uma comparação entre um algoritmo baseado em conjunto *Random Forest*, um algoritmo linear conhecido como SVM e um algoritmo multi-camada referido como ANNs.

É notável, conforme se observa na Tabela 3.1, que ANNs se destaca como a solução mais eficiente quando a tarefa envolve a estimativa de uma saída com base em duas classes esperadas.

Fonte: Elaborado pelo autor (2022)

	Random Forest	Máquinas de Vetores de Suporte	Rede Neural Artificial
Precisão	0,73	0,83	0,91

Tabela 3.1: Resultados das métricas para o proposto

4 RESULTADOS

No estudo em referência, conforme indicado na seção três, foram utilizados três algoritmos diferentes com o propósito de realizar uma análise comparativa. O objetivo subjacente à aplicação desses algoritmos era efetuar uma avaliação abrangente das conclusões obtidas por meio deles. Nesse contexto, uma criteriosa seleção de atributos foi conduzida com base no conjunto de dados disponível.

Os atributos escolhidos incluíram o gênero, carga horária do curso, idade do aluno e duração do curso, os quais foram identificados após terem sido submetidos a uma análise de componentes principais. Essa abordagem possibilitou a visualização das interações intrínsecas entre essas características.

Após essa seleção criteriosa de atributos, prosseguiu-se para a etapa de treinamento do modelo de aprendizado de máquina. A diversidade de classificadores empregados enriqueceu a análise, possibilitando uma abordagem multifacetada das características do conjunto de dados e do desempenho dos algoritmos perante diferentes cenários.

A alocação dos dados para essa finalidade foi realizada de modo a destinar 70% para o treinamento do modelo, 20% para a fase de testes e reservar os restantes 10% para a validação, garantindo, desse modo, uma avaliação robusta e confiável das capacidades preditivas dos algoritmos em exame.

4.1 ANÁLISE DO PERFIL DE ALUNOS

A análise do perfil dos alunos desempenha um papel fundamental na compreensão e no aprimoramento do sistema educacional da Enap. Ao examinar as características como idade, gênero, nível socioeconômico e histórico acadêmico dos estudantes, obtêm-se conhecimentos valiosos sobre as necessidades e desafios enfrentados pelos alunos (47).

Por meio dessa análise, torna-se possível a identificação de possíveis fatores que exercem influência sobre o desempenho e engajamento dos estudantes, além das taxas de evasão e conclusão dos cursos. A compreensão do perfil dos alunos também auxilia na detecção de disparidades e desigualdades educacionais, contribuindo positivamente para a formulação de políticas e estratégias de inclusão e equidade.

Através dessa análise, obtém-se um entendimento mais aprofundado das características individuais e coletivas dos alunos, fornecendo informações valiosas para a melhoria da qualidade da educação e promoção de uma formação mais abrangente e inclusiva.

Portanto, a análise do perfil dos alunos desempenha um papel fundamental no desenvolvimento de estratégias eficazes para o aprimoramento do sistema educacional da Enap, assim como na promoção do sucesso acadêmico e pessoal de todos os estudantes (48).

4.1.1 Número de inscritos por conteudistas

A análise exploratória dos números quantitativos inscritos pelos conteudistas no painel EmNumeros, conforme demonstrado na Figura 4.1, revela informações importantes sobre a distribuição e participação dos conteudistas no sistema.

Ao examinar os dados, é possível identificar a quantidade de inscritos por cada conteudista, fornecendo *insights* sobre o engajamento e envolvimento dos produtores de conteúdo. Essa análise permite uma visão abrangente da contribuição de cada conteudista para o sistema, identificando possíveis disparidades ou tendências.

As informações auxiliam na formulação para a avaliação da experiência e diversidade dos conteúdos oferecidos, bem como no direcionamento de estratégias de incentivo e suporte aos conteudistas.

Fonte: Elaborado pelo autor (2022)

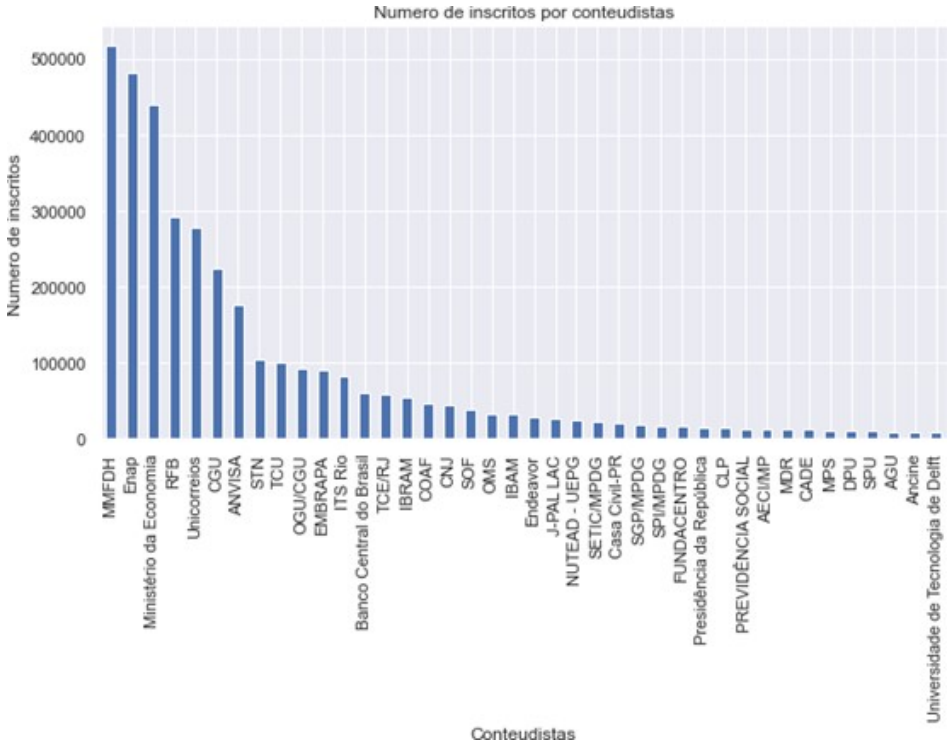

Figura 4.1: Número de inscritos por conteudistas

4.1.2 Número de inscritos por curso

A análise exploratória dos quantitativos de inscritos por curso no painel em números, conforme demonstrado nas Figuras 4.2 e 4.3, fornece visões valiosas sobre a demanda e popularidade dos cursos oferecidos. Ao examinar esses dados, podes-se identificar a quantidade de inscritos em cada curso, permitindo uma compreensão abrangente do interesse dos alunos em relação às diferentes opções de formação.

Essas informações auxiliam na identificação dos cursos com alta demanda, bem como cursos que podem precisar de mais divulgação ou ajustes para atrair um maior número de participantes. Com base

nessas informações, é possível tomar decisões sobre o planejamento e a alocação de recursos, buscando maximizar a eficiência e eficácia da oferta de cursos.

Fonte: Elaborado pelo autor (2022)

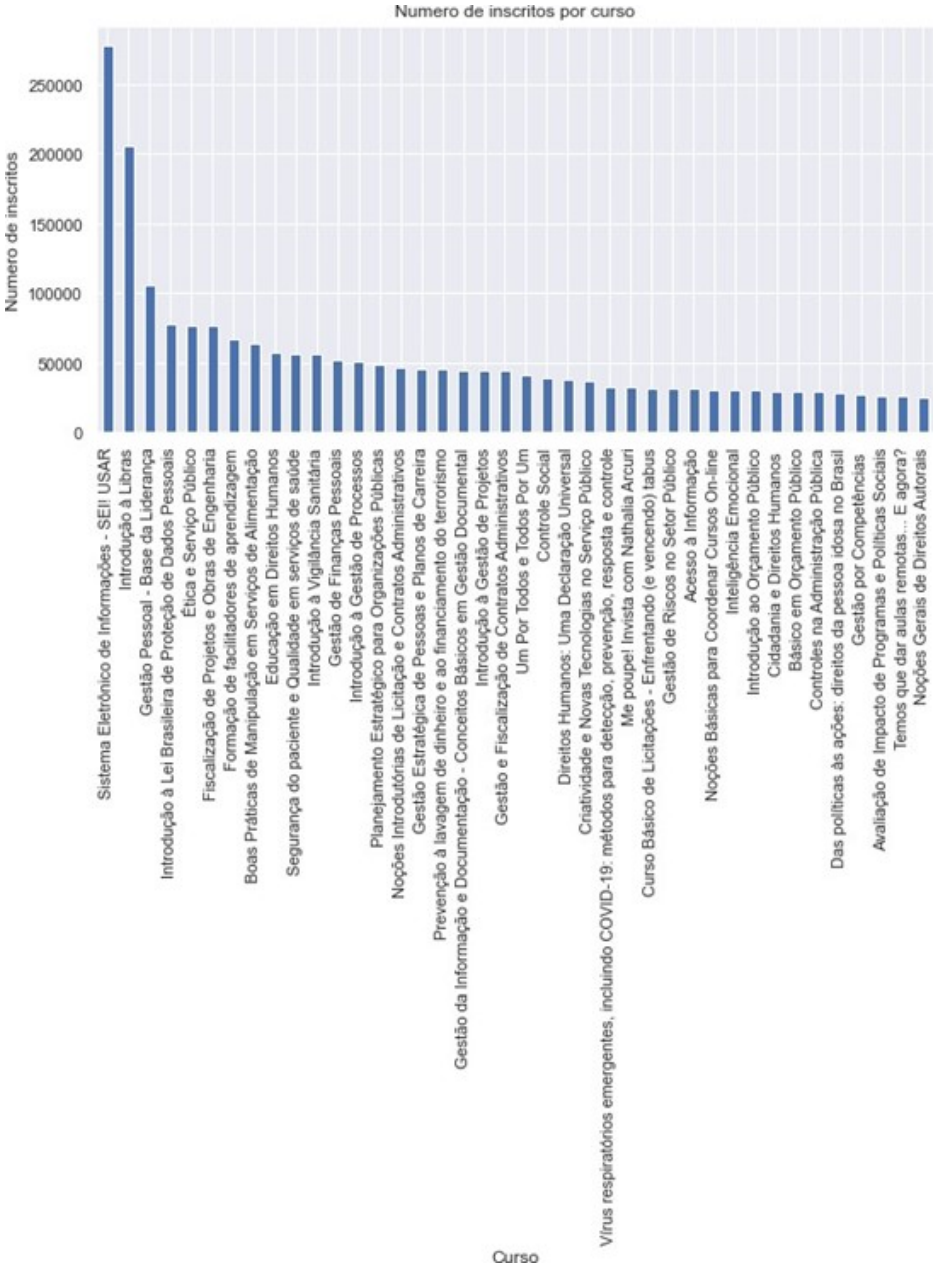

Figura 4.2: Número de inscritos por curso 1.

Fonte: Elaborado pelo autor (2022)

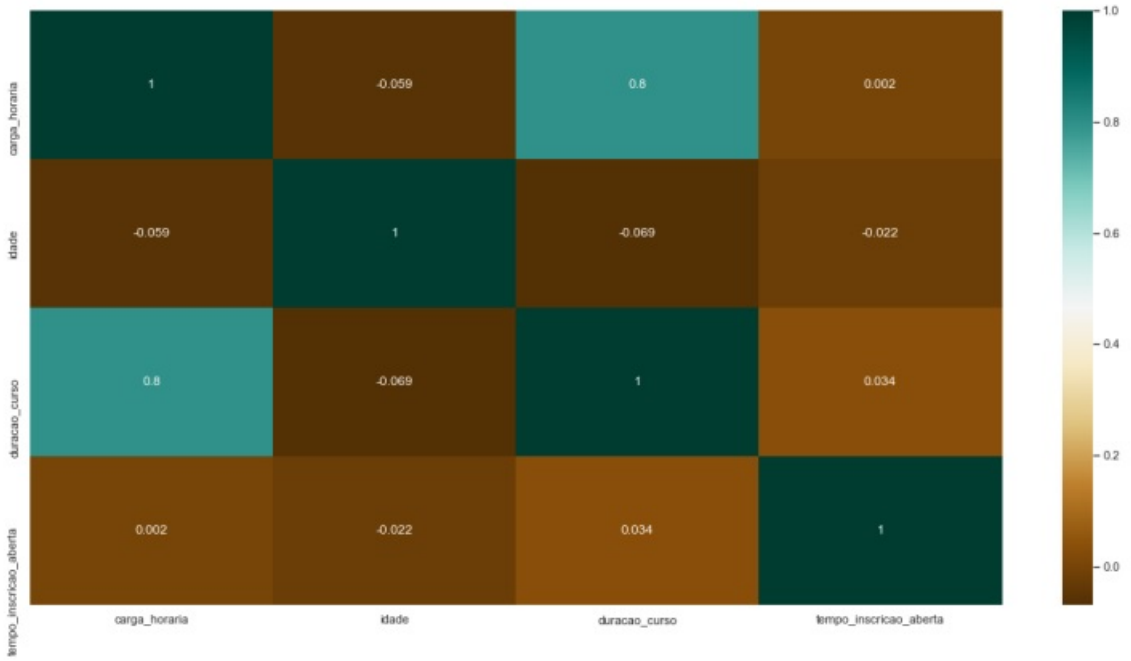

Figura 4.3: Número de inscritos por curso 2.

4.1.3 Carga horária x tempo de inscrição aberta

Na EV.G, uma certificação avançada refere-se a um conjunto de cursos, de assuntos relacionados, que se complementam resultando em uma formação consolidada acerca de um determinado tema. Ao concluir com êxito o conjunto dos cursos que compõem uma certificação avançada, o usuário, além de receber uma certificação separada para cada um dos cursos, recebe também um certificado com a carga horária total da certificação avançada escolhida.

Em relação à influência que esses dois conceitos podem ter, é possível que em alguns casos a carga horária influencie o tempo de inscrição. Por exemplo, se um curso tiver uma carga horária muito extensa, a instituição pode limitar o período de inscrição para que os interessados tenham tempo suficiente para concluir todas as atividades dentro do prazo estabelecido.

Em resumo, a carga horária e o tempo de inscrição são conceitos diferentes que podem ter alguma relação em certos casos, mas geralmente não estão diretamente relacionados.

Foi necessário criar medidas para que se obtivesse o resultado desejado. De início, foi criada uma medida para contar o número de inscrições pela carga horária verso tempo de inscrição aberta, em seguida, foi criada uma medida para dividir o total de inscrições pelo número total de inscrições por situação de inscrição. Além disso, é usado como eixo o ano da matrícula de inscrição e, como legenda, a descrição da situação da inscrição, conforme demonstrado na Tabela 4.1.

Fonte: Elaborado pelo autor (2022)

	Carga horária	Idade	Duração do curso	Tempo de inscrição aberta
Carga horária	1.000.000	-0.059068	0.796833	0.002014
Idade	-0.059068	1.000.000	-0.068661	-0.021604
Duração do curso	0.796833	-0.068661	1.000.000	0.034360
Tempo de inscrição aberta	0.002014	-0.021604	0.034360	1.000.000

Tabela 4.1: Carga horária x tempo de inscrição aberta

4.1.4 Número de desistentes por estado

A análise exploratória dos números de inscrições por região, conforme apresentada na Figura 4.4, desempenha um papel fundamental na compreensão do aproveitamento dos cursos. Essa análise permite a identificação imediata do número de inscrições concluídas, desistentes e em andamento, consolidando informações essenciais para a compreensão do panorama educacional.

Ao analisar esses dados, é possível avaliar o progresso dos alunos, identificar as taxas de resistência e entender o status atual dos cursos em cada região. Essa informação auxilia e possibilita uma tomada de decisão embasada e direcionada, permitindo o desenvolvimento de estratégias eficazes para promover a conclusão dos cursos e garantir uma melhor qualidade educacional em todo o país.

Fonte: Elaborado pelo autor (2022)

Figura 4.4: Desistentes por estado

4.1.5 Número de desistentes por município

Uma análise exploratória dos números quantitativos de desistentes por município no painel EmNumeros, conforme ilustrado na Figura 4.5, oferece perspectivas relevantes sobre as taxas de evasão em diferentes

localidades. Ao examinar esses dados, torna-se possível identificar os municípios com as maiores taxas de desistência, ressaltando possíveis desafios socioeconômicos ou educacionais específicos nessas áreas.

Essa análise possibilita uma compreensão mais aprofundada do fenômeno da evasão, inspirando a implementação de medidas direcionadas para a redução da evasão escolar e a promoção da conclusão dos cursos. Baseando-se nessas informações, podem ser desenvolvidas estratégias e políticas específicas para cada município, visando a melhoria da qualidade da educação e a garantia de oportunidades educacionais mais equitativas em todo o país.

Fonte: Elaborado pelo autor (2022)

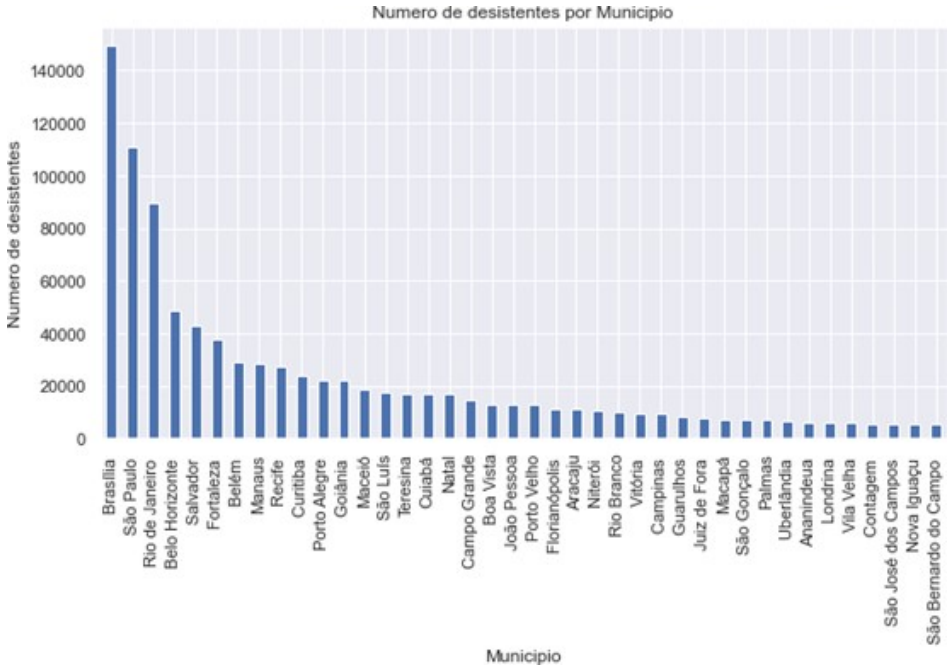

Figura 4.5: Desistentes por municipio

4.1.6 Numero de reprovados por estado

A análise exploratória dos dados quantitativos relacionados aos índices de reprovação por estado, apresentada no painel intitulado EmNumeros, é ilustrada por meio da Figura 4.6. Por meio dessa análise, são observadas as discrepâncias na distribuição das taxas de reprovação entre os diferentes estados. Alguns estados demonstram taxas de reprovação significativamente mais altas do que outros, o que indica a possível existência de desafios educacionais ou socioeconômicos específicos nessas regiões.

Adicionalmente, essa análise pode revelar tendências temporais, como variações sazonais ou mudanças ao longo dos anos. Essas informações são cruciais para auxiliar na formulação de políticas educacionais direcionadas, intervenções pontuais e na alocação de recursos com o propósito de reduzir os índices de reprovação e promover a equidade educacional em todo o país.

Fonte: Elaborado pelo autor (2022)

Figura 4.6: Numero de reprovados por estado

4.1.7 Número de reprovados por município

Uma análise exploratória dos números de reprovados por município no painel em números, conforme demonstrado na Figura 4.7, revela informações importantes sobre as taxas de reprovação em diferentes localidades. Ao examinar esses dados, podemos identificar os municípios com maiores índices de reprovação, indicando possíveis desafios pedagógicos ou de contexto específico dessas regiões.

Essas informações permitem compreender as disparidades na distribuição das reprovações e orientar a formulação de políticas e intervenções educativas direcionadas. Com base nessas informações, podem ser elaboradas estratégias personalizadas para cada município, visando melhorar os índices de aprovação e promover a equidade no ensino.

Fonte: Elaborado pelo autor (2022)

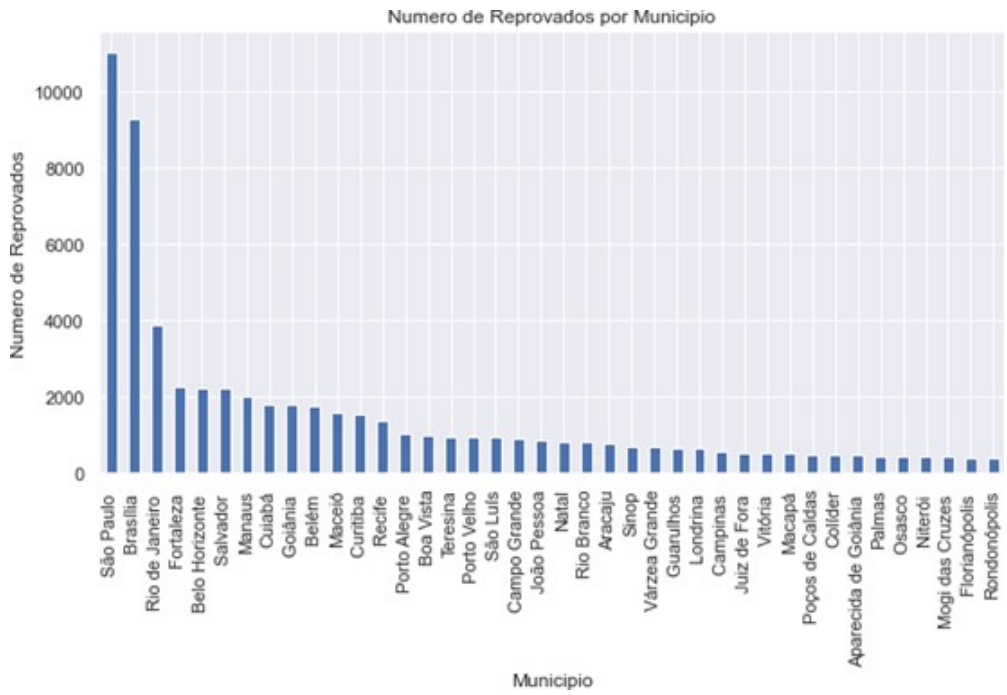

Figura 4.7: Número de reprovados por município

4.1.8 Número de aprovados por estado

Uma análise exploratória dos números de aprovados por estado no painel em números, conforme demonstrado na Figura 4.8, disponibiliza informações relevantes sobre o desempenho dos estudantes em cada região. Ao analisar esses dados, podemos identificar os estados com maiores índices de aprovação, indicando possíveis pontos fortes e boas práticas educacionais nessas áreas.

Essas informações auxiliam a compreensão das disparidades a compreender as disparidades na distribuição das aprovações e na indicação de estratégias eficazes que podem ser replicadas em outros estados. Com base nessas informações, podem ser tomadas políticas educacionais direcionadas para melhorar o desempenho dos alunos e promover a excelência acadêmica em todo o país.

Fonte: Elaborado pelo autor (2022)

Figura 4.8: Número de aprovados por estado

4.1.9 Numero de aprovados por município

A análise exploratória dos números de aprovados por município no painel EmNumeros, conforme demonstrado na Figura 4.9, oferece informações valiosas sobre o desempenho dos estudantes em nível local. Ao examinar esses dados, e possível identificar os municípios com maiores índices de aprovação, evidenciando possíveis fatores de sucesso educacional nessas regiões.

Essas informações permitem compreender as disparidades na distribuição das aprovações, com base nessas informações, a análise exploratória desses números quantitativos é essencial para orientar a tomada de decisões educacionais direcionadas.

Fonte: Elaborado pelo autor (2022)

Figura 4.9: Número de aprovados por município

4.1.10 Número de trancamentos por estado

A análise exploratória dos quantitativos de trancamentos por região no painel EmNumeros, conforme representado na Figura 4.10, apresenta dados significativos sobre os motivos e tendências de interrupção dos cursos em cada área geográfica. Ao examinar esses dados, é possível identificar as regiões com maiores índices de trancamento, indicando possíveis desafios acadêmicos, pessoais ou contextuais enfrentados pelos estudantes nessas localidades.

Essa análise permite a compreensão das disparidades na distribuição dos trancamentos e direcionar esforços para o desenvolvimento de medidas de suporte e intervenções específicas em cada região. Essas informações podem ser implementadas como estratégias de apoio, como programas de orientação acadêmica e suporte social, visando a redução dos trancamentos e a promoção da conclusão dos cursos.

Fonte: Elaborado pelo autor (2022)

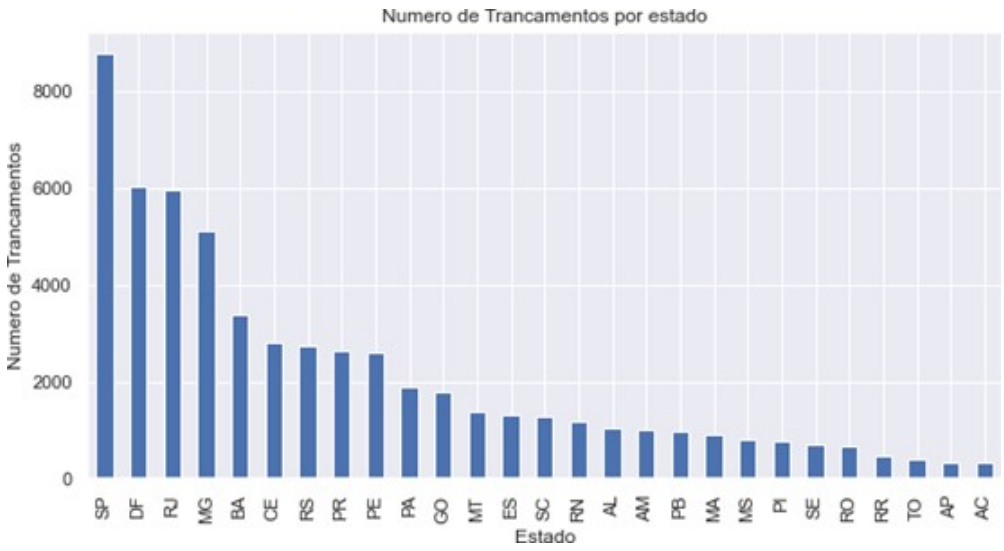

Figura 4.10: Número de trancamento por estado

4.1.11 Número de trancamentos por município

A análise exploratória dos quantitativos de trancamentos por região no painel EmNumeros, conforme representado na figura 4.11, apresenta dados significativos sobre os motivos e tendências de interrupção dos cursos em cada área geográfica. Ao examinar esses dados, pode-se identificar as regiões com maiores índices de trancamento, indicando possíveis desafios acadêmicos, pessoais ou contextuais enfrentados pelos estudantes nessas localidades.

Essa análise permite compreender as disparidades na distribuição dos trancamentos e direcionar esforços para o desenvolvimento de medidas de suporte e intervenções específicas em cada região. Baseado nesses conhecimentos, podem ser implementadas estratégias de apoio, como programas de orientação acadêmica e suporte social, visando a redução dos trancamentos e a promoção da conclusão dos cursos.

Fonte: Elaborado pelo autor (2022)

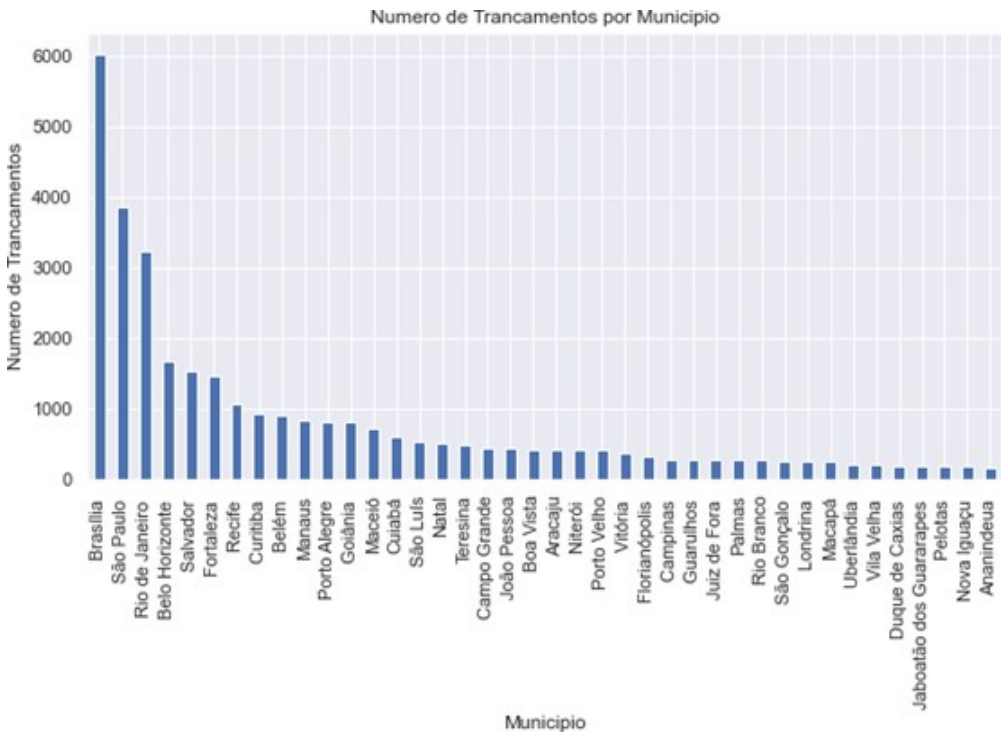

Figura 4.11: Número de trancamento por município

4.1.12 Agrupamento concluído, reprovado, trancado e não concluído

Durante a análise dos dados, fez-se necessário ajustar os rótulos da situação de matrícula e do sexo dos estudantes devido a uma concentração inconsistente e particularmente relevante. Para corrigir essa discrepância, adotou-se uma abordagem de substituição de valores utilizando a linguagem de programação *Python*.

Quanto ao atributo 'sexo', os rótulos 'feminino', 'masculino' e 'outro' foram substituídos pelos valores 0, 1 e 2, respectivamente. Em relação à situação de registro, os rótulos 'Desistente', 'Concluído', 'Reprovado', 'Trancado' e 'Não Concluído' foram trocados pelos valores 0, 1, 0, 0 e 0, respectivamente.

Após a aplicação dessas substituições, realizou-se uma etapa de agrupamento dos dados com base na situação de registro. Essa etapa permitiu a contabilização da quantidade de ocorrências em cada uma das categorias de situação de matrícula, proporcionando uma visão abrangente da distribuição dos estudantes em diferentes estados.

Essa correção na codificação dos atributos revela-se essencial para garantir a consistência e a confiabilidade dos dados, além de facilitar as análises estatísticas e a compreensão dos resultados obtidos.

Na primeira coluna, denominada 'sit_matricula', os rótulos foram modificados para uma codificação do tipo *one-hot*, em que os valores 'Desistente', 'Concluído', 'Reprovado', 'Trancado' e 'Não Concluído' foram substituídos pelos valores binários 0, 1, 0, 0 e 0, respectivamente. Essa codificação demonstra utilidade quando se almeja representar categorias como variáveis binárias em modelos de aprendizado de

máquina.

Já na segunda coluna, 'sexo', a modificação ocorreu através da função *replace*. Os valores originais 'feminino', 'masculino' e 'outro' foram substituídos pelos valores numéricos 0, 1 e 2, respectivamente. Essa alteração viabiliza uma representação mais adequada das informações concernentes ao sexo dos indivíduos, simplificando análises posteriores.

Após essas transformações, efetuou-se o agrupamento dos dados com base na coluna 'sit_matricula', calculando-se o número de ocorrências em cada grupo. Os resultados desse agrupamento foram dispostos em formato tabular, exibindo a contagem de registros para cada valor presente na coluna 'sit_matricula'. Os grupos foram identificados pelos valores 0 e 1.

Após a incorporação dessas modificações, o *DataFrame* resultante agora possui as seguintes colunas: 'sexo', 'carga_horaria', 'conteudista', 'tematica', 'idade', 'duracao_curso', 'tempo_inscricao_aberta' e 'sit_matricula'.

A coluna 'sit_matricula' apresenta os valores 0 e 1, correspondentes a grupos com contagens de registros de 1.799.208 e 1.501.611, respectivamente. Essas transformações possibilitam uma representação mais apropriada e coesa dos dados, viabilizando análises mais precisas e interpretações mais robustas, como ilustrado na Tabela 4.2.

Fonte: Elaborado pelo autor (2022)

sit mat.	sexo	carga horaria	conteudista	tematica	idade	duracao curso	tep insc aberta
0	1799208	1799208	1799208	1799208	1799208	1799208	1799208
1	1501611	1501611	1501611	1501611	1501611	1501611	1501611

Tabela 4.2: Distribuição dos Registros por Situação de Matrícula

4.1.13 Agrupamento, situação, conteudista, temática e tempo inscrição

No decorrer da análise dos dados, efetuou-se o processamento necessário para preparar os conjuntos de dados. O conjunto de rótulos, denominado *labels*, originou-se a partir da coluna "sit_matricula"do *dataframe*, sendo convertido em um *array numpy*. Esta coleção de rótulos engloba uma série de valores representados como '0, 1, 0, ..., 0, 1, 0', os quais refletem a situação de matrícula dos estudantes.

A coleção de rótulos abrange uma sequência de valores, tais como 0, 1, 0, ..., 0, 1, 0, os quais indicam o estado de matrícula dos discentes.

Ademais, o conjunto de características foi formado a partir do quadro de dados original, com a exclusão das colunas "sit_matricula", "tempo_inscricao_aberta", "conteudista"e "tematica", que não desempenham um papel relevante na análise. As características remanescentes incluem gênero, carga_horaria, idade e duracao_curso; a função *head*() foi empregada para exibir as primeiras cinco linhas desse conjunto, proporcionando uma visão panorâmica dos dados.

Com o propósito de simplificar a manipulação e viabilizar a aplicação de técnicas de análise de dados,

as *features* (características) foram transformadas em um *array numpy*. Essa conversão permite a organização eficiente das informações e viabiliza a realização de diversas operações e procedimentos de maneira ágil e otimizada. A transformação em um *array numpy* maximiza a utilização plena das capacidades e funcionalidades desta biblioteca, conferindo maior praticidade e eficácia à análise de dados.

A continuidade entre as Tabelas 4.3 e 4.4 demonstra sua relevância ao se obter uma visão completa e abrangente dos dados, o que possibilita a identificação de padrões, tendências e potenciais fatores influentes na resistência dos alunos. A realização dessa análise exploratória desempenha um papel crucial ao fundamentar as tomadas de decisão e estratégias de intervenção, as quais têm o objetivo de reduzir a taxa de desistência e aprimorar a qualidade da educação.

Fonte: Elaborado pelo autor (2022)

	sexo	carga horária	sit matrícula	conteudista
0	1	20	0	SPI/MPDG
1	1	20	1	CNJ
2	1	20	0	COAF
3	1	40	0	Ministério da Economia
4	1	20	0	Ministério da Economia

Tabela 4.3: Agrupamento, situação, conteudista, temática e tempo inscrição aberta_1

Fonte: Elaborado pelo autor (2022)

	Temática	idade	duração do curso	tempo de inscrição aberta
0	Gestão de Políticas Públicas	60	21	31
1	Gestão Estratégica	39	30	34
2	Auditoria e Controle	39	20	34
3	Gestão de Pessoas	39	50	32
4	Governança e Gestão de Riscos	39	40	33

Tabela 4.4: Agrupamento, situação, conteudista, temática e tempo inscrição aberta_2

4.1.14 Situação matrícula

Durante a análise dos dados, foram executadas etapas de preparação e formatação dos conjuntos de características. Inicialmente, foram removidas as colunas irrelevantes para a análise, tais como 'sit_matricula', 'tempo_inscricao_aberta', 'conteudista' e 'temática' do *dataframe* original, gerando assim o conjunto de características denominado *features*.

Após a eliminação das colunas, o conjunto de características passou a englobar as seguintes colunas remanescentes: [lista das colunas remanescentes]. Com o intuito de proporcionar uma visão preliminar desses atributos, um trecho do conjunto de dados é apresentado através do comando *head*()".

Posteriormente, o conjunto de características foi convertido em um *array numpy*, viabilizando, desse

modo, sua manipulação e aplicação em técnicas de análise de dados e aprendizado de máquina, conforme exemplificado na Tabela 4.5.

Essas etapas de preparação e formatação dos dados desempenham um papel crucial na extração de *insights* e na realização de análises mais aprofundadas do conjunto de características, visando a uma compreensão mais abrangente dos padrões e das relações presentes nos dados.

Fonte: Elaborado pelo autor (2022)

	sexo	carga horaria	idade	duracao curso
0	1	20	60	21
1	1	20	39	30
2	1	20	39	20
3	1	40	39	50
4	1	20	39	40

Tabela 4.5: Situação matrícula

4.1.15 Conjuntos de treinamento e teste

Para a realização da divisão dos dados em conjuntos de treinamento e teste, utilizou-se a biblioteca *Scikit-learn*. A proporção adotada para a divisão foi de 70% para o conjunto de treinamento, 20% para a validação e 10% para o conjunto de teste.

Inicialmente, o conjunto de dados foi dividido em conjuntos de treinamento e teste por meio da função *train_test_split* da biblioteca *Scikit-learn*. Os parâmetros fornecidos à função foram as características *features* e os rótulos *labels*. O tamanho do conjunto de teste foi definido como 1 menos a proporção de treinamento, garantindo, assim, que o conjunto de treinamento representasse 70% do tamanho original do conjunto de dados. As variáveis *train_features, test_features, train_labels e test_labels* foram utilizadas para armazenar os conjuntos de treinamento e teste, respectivamente.

Posteriormente, o conjunto de teste passou por uma segunda divisão, resultando nos conjuntos de validação e teste, novamente por meio da função *train_test_split*. Os parâmetros utilizados incluíram as características e rótulos do conjunto de teste, juntamente com a proporção do conjunto de teste em relação à soma dos tamanhos dos conjuntos de teste e validação. As variáveis "x_val", "x_test", "y_val"e "y_test"foram empregadas para representar os conjuntos de validação e teste, conforme demonstrado na Tabela 4.6.

A estratégia de divisão dos dados em conjuntos de treinamento, validação e teste desempenha um papel fundamental na avaliação do desempenho dos modelos de aprendizado de máquina, permitindo ajustes e validações adequadas ao longo do processo de desenvolvimento.

Fonte: Elaborado pelo autor (2022)

(2310573, 4)
(990246, 4)
(330082, 4)
(330082,)
(660164, 4)
(660164,)

Tabela 4.6: Escala, situação matrícula

4.2 CLASSIFICADOR *RANDOM FOREST*

No contexto do aprendizado de máquina, um estimador é comumente identificado como um modelo ou algoritmo utilizado para prever resultados baseados nos recursos de entrada. Exemplos de estimadores no campo do aprendizado de máquina abrangem a regressão linear, as árvores de decisão e as máquinas de vetores de suporte.

A avaliação dos estimadores pode ser conduzida com base em diversos critérios, tais como viés, variância e erro quadrático médio. Em termos gerais, o objetivo é que os estimadores apresentem imparcialidade, o que implica que seu valor esperado seja equivalente ao valor real do parâmetro em estimação.

Além disso, espera-se que os estimadores possuam baixa variância, indicando que não são suscetíveis a pequenas flutuações nos dados. Por último, busca-se que os estimadores apresentem um erro quadrático médio reduzido, que se configura como uma medida da precisão global do estimador. Para efetuar essa comparação, a pesquisa utilizou os estimadores 4, 8, 16, 32, 64, 128 e 256 (49), conforme ilustrado na Tabela 4.7.

Fonte: Elaborado pelo autor (2022)

Número de Estimadores	Precisão	Recall	F1-Score
4	0.72	0.60	0.70
8	0.72	0.60	0.70
16	0.73	0.60	0.70
32	0.73	0.60	0.70
64	0.73	0.60	0.70
128	0.73	0.60	0.70
256	0.73	0.60	0.70

Tabela 4.7: Estimadores 4, 8, 16, 32, 63, 128 e 256

4.2.1 Estimador 4

No contexto dos estimadores de quarta ordem, trata-se de estimativas que envolvem o quarto momento da distribuição de probabilidade da população. O quarto momento de uma distribuição de probabilidade representa uma medida da forma da distribuição e é definido como a média da quarta potência dos desvios em relação à média.

4.2.2 Estimador 8

Os estimadores de oitava ordem são aqueles que incorporam o oitavo momento da distribuição de probabilidade da população. O oitavo momento representa uma medida da forma da distribuição e é definido como a média da potência da oitava ordem dos desvios em relação à média.

4.2.3 Estimador 16

Os estimadores de ordem decimal, bem como os de quarto e oitavo, estão condicionados à distribuição de probabilidade da população e são empregados para efetuar inferências a respeito de uma população com base em uma amostra representativa.

4.2.4 Estimador 32

Em resumo, o cálculo dos estimadores de treinamento é uma prática pouco comum e pode requerer o uso de algoritmos especializados e técnicas numéricas avançadas. É importante ter em mente que, na maioria dos casos, os estimadores de ordem inferior, como o quarto e o oitavo, são suficientes para realizar inferências precisas sobre a população a partir de uma amostra.

4.2.5 Estimador 64

O cálculo dos estimadores de sexagésima quarta ordem pode se revelar extremamente complexo, sendo uma ocorrência muito rara na prática estatística. Em geral, esses estimadores são utilizados em contextos bastante específicos, exigindo o emprego de algoritmos especializados e técnicas numéricas avançadas para seu cálculo.

4.2.6 Estimador 128

Em resumo, o cálculo dos estimadores de centésima vigésima oitava ordem é extremamente raro e complexo, sendo seu uso limitado a contextos muito específicos. Na maioria das situações, estimadores de ordens inferiores (como os de quarta, oitava, décima sexta, trigésima quarta e sessenta e quarta ordem) demonstra ser suficiente para a realização de inferências precisas sobre a população a partir de uma amostra.

4.2.7 Estimador 256

O cálculo dos estimadores de duzentos quinquagésimos sextos pedidos pode tornar-se extremamente complexo, sendo excepcionalmente raro na prática estatística. Em geral, tais estimadores são empregados em contextos altamente específicos e calculados, podendo exigir a utilização de algoritmos especializados e técnicas numéricas avançadas, conforme a necessidade.

4.3 CLASSIFICADOR DE MÁQUINA DE VETORES DE SUPORTE (SVM)

O Classificador de SVM é amplamente reconhecido como um tipo popular de algoritmo de aprendizado supervisionado utilizado em tarefas de classificação e regressão. Ele se baseia no conceito de encontrar o hiperplano ótimo que realiza a separação entre diferentes classes de pontos de dados em um espaço de alta dimensão.

Nas tarefas de classificação, o algoritmo SVM se empenha em identificar o hiperplano que maximiza a margem, a qual representa a distância entre o hiperplano e os pontos de dados mais próximos de cada classe. Esses pontos de dados mais próximos do hiperplano são conhecidos como vetores de suporte, os quais determinam a localização e orientação do próprio hiperplano.

O objetivo do algoritmo SVM é determinar o hiperplano que não apenas separa as classes, mas também maximiza a margem entre elas. Essa abordagem contribui para aprimorar a capacidade de generalização do modelo e reduzir o risco de *overfitting*.

No contexto das tarefas de regressão, os SVMs são empregados para encontrar uma função que melhor se ajuste aos dados, minimizando o erro associado. O algoritmo se esforça para encontrar o hiperplano que atravessa o maior número possível de pontos de dados, ao mesmo tempo em que minimiza a distância entre o hiperplano e os pontos de dados em questão.

Os SVMs possuem a capacidade de lidar tanto com dados lineares quanto não lineares, utilizando diferentes funções de *kernel* que transformam os dados em um espaço de dimensão superior. Algumas funções de *kernel* amplamente reconhecidas incluem o *kernel* linear, o *kernel* polinomial e o *kernel* da função de Base Radial *Radial Basis Function* (RBF).

Os SVMs oferecem inúmeras vantagens em relação a outros algoritmos de classificação, se destacando pela capacidade de generalização eficaz, robustez em relação a ruídos e valores discrepantes, além da aptidão para lidar com dados de alta dimensão. No entanto, é importante observar que eles podem ser sensíveis à escolha da função de *kernel* e de seus parâmetros, podendo também apresentar custos computacionais significativos, especialmente ao lidar com conjuntos de dados extensos (50).

4.4 CLASSIFICADOR DE ÁRVORE DE DECISÃO

Realiza-se uma análise da aplicação do algoritmo *Decision Tree* para a classificação de dados. O classificador *Decision Tree* é empregado para executar a tarefa de classificação em um conjunto de dados.

Este conjunto é composto por recursos de treinamento *train_features* e pelos rótulos associados a esses recursos *train_labels*. O processo de treinamento do modelo é conduzido por meio do método *fit*, utilizando os dados de treinamento.

Em seguida, previsões são feitas em um conjunto de validação (x_val) por meio do método *predict*. As métricas de avaliação adotadas para mensurar o desempenho do classificador incluem precisão, *recall* e pontuação f1 para cada classe relevante, além da acurácia global do modelo. Estes resultados estão apresentados na Tabela 4.8.

O classificador de árvore de decisão é classificado como um algoritmo de aprendizado supervisionado utilizado para a classificação. Ele constrói uma árvore de decisão baseada nos dados de treinamento, criando, assim, uma estrutura hierárquica composta por nós que representam decisões baseadas nos valores dos recursos.

O nó-raiz da árvore de decisão corresponde ao conjunto total dos dados de treinamento e é fragmentado em subconjuntos menores conforme os valores dos recursos presentes em cada nó interno. Os nós-folha refletem a decisão final ou classificação dos dados de entrada.

O algoritmo da árvore de decisão opera ao selecionar o recurso que proporciona o máximo ganho de informação em cada nó. O ganho de informação mensura a redução na entropia (ou seja, o grau de aleatoriedade ou incerteza) após a divisão dos dados com base em um determinado recurso. O recurso que proporciona o ganho de informação máximo é eleito como critério para a divisão.

A progressão do algoritmo de árvore de decisão envolve a contínua subdivisão dos dados em conjuntos menores a cada nó interno. Essa subdivisão é guiada pelo recurso escolhido e seus respectivos valores, até que os nós-folha sejam purificados ou quase purificados (ou seja, todos os pontos de dados em um nó-folha pertencem à mesma classe). Nesse estágio, a árvore de decisão adquire o conhecimento dos padrões nos dados de treinamento e pode, então, categorizar novos pontos de dados.

O classificador de árvore de decisão apresenta múltiplos benefícios, como sua simplicidade e interpretabilidade. A estrutura da árvore pode ser visualizada e compreendida por seres humanos. Além disso, o classificador é capaz de lidar tanto com recursos categóricos quanto numéricos, podendo capturar relações não-lineares entre os recursos e a variável-alvo.

No entanto, as árvores de decisão podem ser suscetíveis ao *overfitting*, especialmente quando a árvore se torna excessivamente profunda ou quando os dados de treinamento apresentam ruídos. Técnicas de regularização, como a poda ou a restrição da profundidade máxima da árvore, são empregadas para lidar com o *overfitting*.

O *overfitting* é um problema recorrente no campo do aprendizado de máquina. Ele ocorre quando um modelo é excessivamente complexo e se ajusta de maneira muito precisa aos dados de treinamento. Consequentemente, o modelo não consegue generalizar eficazmente para novos e não-vistos dados, resultando em desempenho e precisão insatisfatórios (51).

Fonte: Elaborado pelo autor (2022)

	precisão	lembrar	f1-score	apoiar
0	0,62	0,55	0.58	349503
1	0,52	0,53	0,53	291246
2	0,00	0,18	0,01	369
3	0,00	0,05	0,00	103
4	0,70	0,57	0,63	18943
precisão			0,54	660164
média macro	0.37	0.38	0.35	660164
média ponderada	0.58	0.54	0.56	660164

Tabela 4.8: Classificador de árvore de decisão

4.5 *PERCEPTRON* DE MÚLTIPLAS CAMADAS (MLP)

O *Multi-Layer Perceptron* (MLP) é um tipo de rede neural artificial utilizado em tarefas de aprendizado supervisionado, tais como classificação, regressão e reconhecimento de padrões. O MLP consiste em várias camadas de nós ou neurônios interconectados, em que cada neurônio recebe entrada da camada anterior e produz saída para a camada seguinte.

A camada de entrada do MLP representa os atributos dos dados de entrada, enquanto a camada de saída representa a variável de destino prevista. As camadas ocultas, localizadas entre as camadas de entrada e saída, realizam transformações não lineares nos dados de entrada para extrair recursos e padrões úteis.

O MLP emprega o algoritmo de retropropagação para treinamento do modelo, o qual envolve ajustar os pesos e desvios dos neurônios para minimizar a discrepância entre a saída prevista e os resultados reais dos dados de treinamento. O algoritmo de retropropagação utiliza o método do gradiente descendente para identificar os pesos e desvios ideais que minimizam a função de perda.

O MLP é capaz de lidar tanto com relações lineares quanto não lineares entre as variáveis de entrada e saída, tornando-o um algoritmo de aprendizado de máquina poderoso e flexível. Além disso, ele pode aprender padrões e relações complexas nos dados, tornando-se adequado para diversas aplicações, como reconhecimento de imagem, identificação de fala e processamento de linguagem natural.

No entanto, o MLP pode ser computacionalmente custoso e requerir uma grande quantidade de dados de treinamento para obter um desempenho satisfatório. Além disso, ele é propenso a *overfitting* se o modelo for excessivamente complexo ou se os dados de treinamento contiverem ruído. Técnicas de regularização, como *dropout* e decaimento de peso, podem ser aplicadas para prevenir o *overfitting* e aprimorar a capacidade de generalização.

Durante a execução do código, um aviso denominado *Visible Deprecation Warning* é exibido. Esse aviso indica que a criação de um *ndarray* a partir de sequências aninhadas irregulares está obsoleta, ou seja, uma lista ou tupla de listas, tuplas ou *ndarray* com comprimentos ou formas diferentes. Caso essa

seja a intenção, é necessário especificar *dtype=object* ao criar o *ndarray*.

No trecho em questão, uma figura de dimensões 10 por 6 polegadas é criada utilizando a função plt.figure *figsize*= 10,6. Posteriormente, um gráfico de dispersão é plotado utilizando a função *plt.scatter*, no qual os valores de x são gerados com base no número de exemplos de treinamento usando a função *np.arangecom*, representado por *train_features.shape*[0].

Os valores de y correspondem aos elementos da lista acc_val. O parâmetro *alpha* define a transparência dos pontos, s define o tamanho dos pontos e *label* atribui o rótulo 'mu' ao conjunto de pontos no gráfico.

O título do gráfico é estabelecido como 'Perda para cada ponto de dados de treinamento' utilizando a função *plt.title*, e os rótulos dos eixos x e y são definidos como 'Dados de treinamento' e 'Perda', respectivamente, por meio das funções *plt.title*. Finalmente, o gráfico é exibido na tela utilizando a função *plt.show*().

Essa visualização foi gerada de acordo com o código fornecido e é apresentada na Figura 4.12. Através dessa imagem, é possível analisar a variação da perda durante o treinamento do modelo e compreender seu comportamento em relação aos dados de treinamento.

Fonte: Elaborado pelo autor (2022)

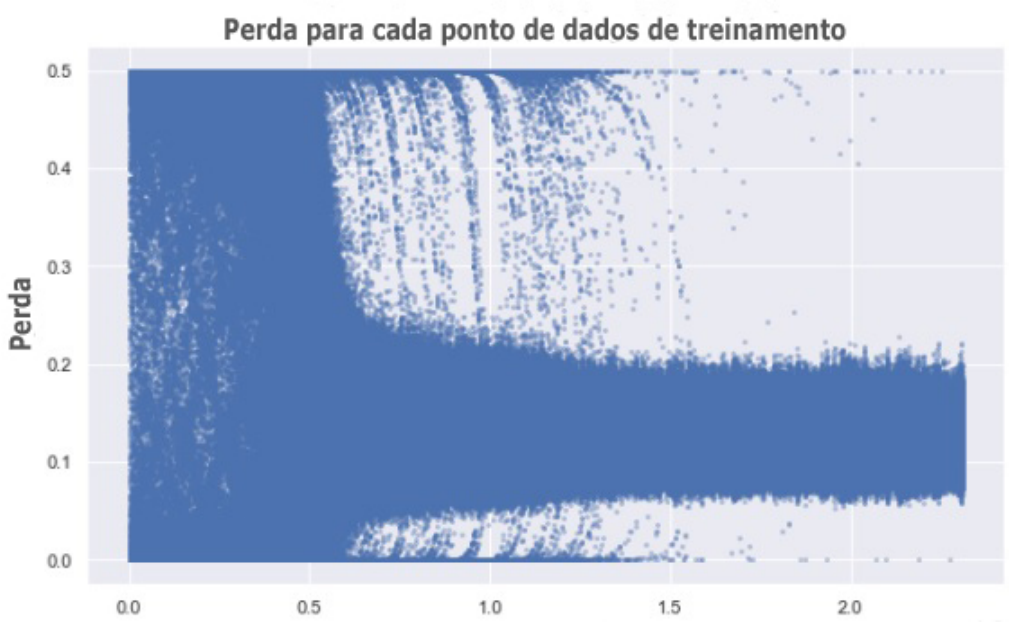

Figura 4.12: Treinamento de dados

No comando destacado, uma nova figura semelhante é gerada, com as mesmas dimensões da Figura 4.12 (ou seja, 10 por 6 polegadas), utilizando a função *plt.figure figsize*= 10,6. Em seguida, um gráfico de dispersão é plotado através da função *plt.scatter*.

Nesse gráfico, os valores de coordenada x são gerados com base no tamanho da lista acc_avg_val, utilizando a função np.arange. Os valores de coordenada y correspondem aos próprios elementos contidos

na lista acc_avg_val. Os pontos no gráfico são identificados com o rótulo 'mu', atribuído por meio do parâmetro *label*.

A legenda do gráfico é configurada como *Average Loss by epoch* com o auxílio da função *plt.title*. Adicionalmente, os eixos x e y são devidamente rotulados como *Training data* e *Loss*, respectivamente, através das funções *plt.xlabel e plt.ylabel*. Por fim, o gráfico é exibido na tela utilizando a função *plt.show*(). A visualização completa dessa ação é representada na Figura 4.13.

Fonte: Elaborado pelo autor (2022)

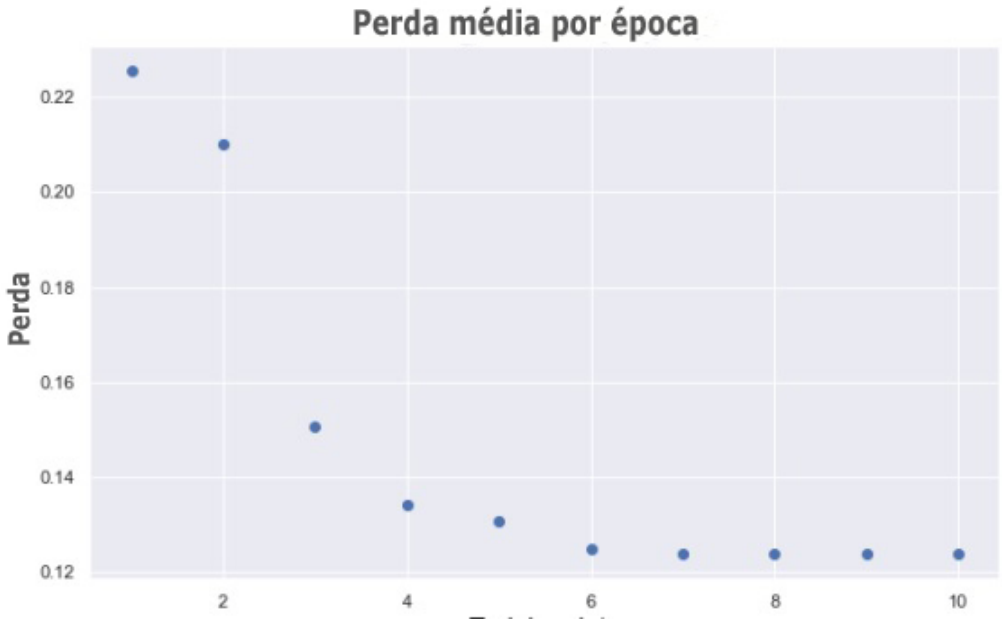

Figura 4.13: Treinamento de dados 2

O controle e a visualização, exemplificados na Figura 4.14, demonstram a representação gráfica da matriz de confusão de um modelo de classificação. A matriz de confusão é uma ferramenta frequentemente utilizada para avaliar o desempenho de algoritmos de classificação.

No trecho de código apresentado, primeiramente é elaborado um dicionário denominado "dict_liveque", o qual associa os valores das classes (0 e 1) aos seus respectivos rótulos *Failed e Approved*. Em seguida, a matriz de confusão é gerada por meio da função *confusion_matrix*, sendo alimentada com as classes reais "y_vale"e as classes previstas *predictions*.

Posteriormente, a matriz de confusão é transformada em um *DataFrame* do *pandas*, denominado "df_cm". Neste *DataFrame*, os rótulos das linhas e colunas são atribuídos com base nas informações contidas no dicionário *dict_live*. Em sequência, uma figura de dimensões 7 por 7 polegadas é criada utilizando o comando *plt.figure (figsize*=(7,7))".

A representação gráfica da matriz de confusão é obtida por meio de um mapa de calor, utilizando a função *sns.heatmap* da biblioteca *Seaborn*. A opção *annot=True* possibilita a inclusão dos valores numéricos no interior das células, e o esquema de cores é configurado como *Blues*. O argumento "fmt='g'"estabelece

o formato dos valores numéricos.

Por fim, os rótulos dos eixos x e y são definidos para indicar as classes previstas e as classes reais, respectivamente. A visualização da matriz de confusão é exibida por meio do comando *plt.show*()".

Fonte: Elaborado pelo autor (2022)

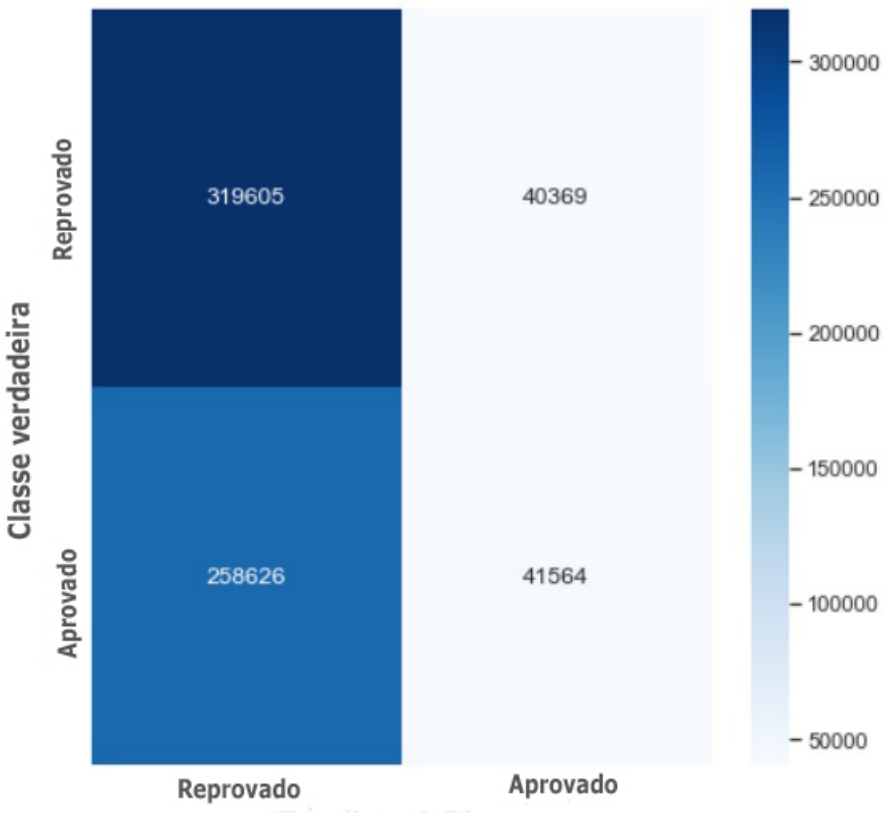

Figura 4.14: Classe prevista

A Tabela 4.9 é a representação resultante do relatório de classificação gerado a partir da execução do comando *print classification_report predictions.round*(), y_val. O relatório de classificação exibe diversas métricas de avaliação do desempenho de um modelo de classificação. Cada uma das classes (0 e 1) possui suas próprias métricas avaliativas:

- A precisão é calculada como a proporção de verdadeiros positivos em relação a todas as previsões positivas feitas pelo modelo.
- O *recall* é definido como a proporção de verdadeiros positivos em relação a todos os valores verdadeiramente positivos presentes nos dados de validação.
- O F1-score é uma métrica combinada que calcula a média harmônica entre a precisão e o *recall*.
- O suporte refere-se ao número de instâncias de cada classe nos dados de validação.

Adicionalmente, o relatório também apresenta métricas como a acurácia *precision*, a média ponderada *weighted avg* e a média ponderada para todas as classes *macro avg*. A acurácia avalia a taxa geral de previsões corretas do modelo em relação a todas as amostras. A média ponderada leva em consideração o peso de cada classe com base em sua frequência, proporcionando uma métrica agregada que lida com desequilíbrios entre as classes.

Analisando o exemplo apresentado, é possível constatar que o modelo demonstra uma precisão considerável para a classe 0 (0.89), indicando que a maioria das previsões corretas é adequadamente classificada como *Failed*. No entanto, a classe 1 (aprovação) possui uma precisão notavelmente baixa (0.14), sugerindo que ocorrem muitas previsões incorretas nessa classe.

O *recall* para ambas as classes é relativamente baixo, o que sugere que o modelo enfrenta dificuldades em identificar de forma precisa os casos positivos. O *F1-score*, que combina precisão e *recall*, também exibe valores reduzidos para ambas as classes.

Globalmente, a acurácia geral do modelo é de aproximadamente 0.55, indicando que cerca de 55% das amostras são corretamente classificadas.

Fonte: Elaborado pelo autor (2022)

	precision	recall	f1-score	support
0.0	0,89	0,55	0.68	578231
1.0	0,14	0,51	0,22	81933
acuracy			0,55	660164
macro avg	0.51	0.53	0.45	660164
weighted avg	0.79	0.55	0.62	660164

Tabela 4.9: Matriz de confusão

5 CONCLUSÃO

Em conclusão, o presente trabalho explora a aquisição de dados do portal EmNumeros e analisa aspectos relevantes no âmbito dos cursos oferecidos pela instituição denominada EV.G. Os objetivos propostos foram parcialmente alcançados, possibilitando a visualização das cidades de origem dos alunos matriculados nos cursos da EV.G nas regiões Nordeste, Sudeste e Sul, com destaque para o considerável número de alunos provenientes das capitais.

Adicionalmente, identificou-se a frequência das matrículas e viabilizou-se uma análise visual mais dinâmica das inscrições e dos cursos realizados.

No processo de preparação dos dados, utilizou-se a biblioteca Pandas para eliminar informações inconsistentes e valores atípicos presentes em alguns cursos. Essa etapa revelou-se crucial para assegurar a qualidade dos dados sob análise. A avaliação dos dados empregou algoritmos de aprendizado de máquina supervisionado, como o Algoritmo de *Random Forest*, para comparar o desempenho com máquinas de suporte vetorial SVM e redes neurais artificiais ANNs com duas camadas ocultas.

Esses modelos possibilitaram a previsão do número de alunos que concluíram os cursos, contribuindo para uma compreensão mais aprofundada do perfil dos estudantes.

A representação visual dos dados por meio de gráficos proporcionou uma análise mais acessível e clara para os administradores dos cursos e outras partes interessadas. No entanto, enfrentaram-se desafios durante a migração do site EmNumeros para uma nova versão, incluindo a atualização manual dos dados nas visualizações e a limitação na detecção de casos de desistência ou cursos não concluídos pelos alunos. Essas dificuldades indicam a necessidade de melhorias futuras na gestão e no controle desses aspectos.

De modo resumido, o trabalho contribuiu para uma análise mais profunda dos cursos fornecidos pela instituição EV.G, fornecendo informações valiosas sobre os alunos matriculados e suas características. Apesar dos desafios enfrentados, as descobertas e metodologias estabelecem uma base sólida para pesquisas futuras e aprimoramentos no campo da educação à distância e formação profissional.

A divulgação das informações e a aplicação de técnicas de análise de dados podem impulsionar o controle social e embasar decisões administrativas em áreas mais impactadas.

5.1 TRABALHOS FUTUROS

No contexto das perspectivas futuras deste trabalho, a intenção é desenvolver um pipeline automatizado de dados por meio da ferramenta *Airflow*. A implementação desse *Pipeline* automatizado permitirá a administração contínua e sistemática dos dados presentes no portal EmNumeros, bem como aprimorará a eficiência e o desempenho das expectativas estabelecidas. Essa automação também desempenhará o papel de suporte ao coordenador do curso, fornecendo informações valiosas para a previsão do sucesso dos alunos e embasando decisões fundamentadas.

A utilização do *Airflow* como ferramenta de automação possibilitará o estabelecimento de um fluxo contínuo e consistente de transmissão de dados, garantindo assim uma atualização constante das informações disponíveis. Essa abordagem viabilizará a atualização em tempo real dos resultados das expectativas, disponibilizando análises sempre atualizadas para auxiliar no planejamento e nas decisões relacionadas aos cursos oferecidos pela EV.G.

Ademais, a automação do processo de captação de dados também reduzirá a necessidade de intervenções manuais, permitindo uma análise mais ágil e precisa.

Portanto, como parte das etapas futuras, espera-se a criação de um *Pipeline* de dados automatizado através da utilização do *Airflow*. Isso contribuirá de maneira significativa para a melhoria contínua da eficácia no sucesso dos alunos nos cursos oferecidos pela EV.G.

Essa iniciativa representa um avanço de grande relevância no uso eficiente dos dados presentes no portal EmNumeros, proporcionando suporte planejado para elevar a qualidade do ensino e promover uma formação mais abrangente e inclusiva.

REFERÊNCIAS BIBLIOGRÁFICAS

1 OLIVEIRA, T.; BARBAR, J.; SOARES, A. Predição do tráfego de rede de computadores usando redes neurais tradicionais e de aprendizagem profunda. *Revista de Informática Teórica e Aplicada - RITA*, v. 22, p. 10–28, 06 2015.

2 PúBLICA, E. N. de A. Organizado por escola nacional de administração pública – enap. In: . [s.n.], 2019. v. 1. Acesso em: 17 de março de 2022. Disponível em: <https://www.escolavirtual.gov.br/>.

3 ENAP/UNB, P. Emnumeros. In: . [s.n.], 2012. v. 1, p. 1–4. Acesso: em 17 de março de 2022. Disponível em: <https://emnumeros.escolavirtual.gov.br/indicadores/>.

4 PRACIANO, G. F.; PRACIANO, B. J. G.; MENDONCA, L. L. F.; GALLINDO, E. L.; DUARTE, J. F. C. M.; SOUSA, J. R. T. Integrity of training data for federal civil employees in brazil.

5 PRIFIT.CO. Zscore. In: . [s.n.], 2022. v. 1, n. 1, p. 1. Acesso em: 01 de setembro de 2022. Disponível em: <https://www.profit.co/blog/kpis-library/finance/z-score/#/>.

6 BUSINESS, L. N. S. S. of. Biografica edward altman. In: . [s.n.], 2022. v. 1, n. 1, p. 1. Acesso em: 10 de setembro de 2022. Disponível em: <https://www.stern.nyu.edu/faculty/bio/edward-altman/>.

7 TEAM, D. S. O que é um zscore? In: . [s.n.], 2020. v. 1, n. 1, p. 1. Acesso em: 01 de setembro de 2022. Disponível em: <https://datascience.eu/pt/matematica-e-estatistica/o-que-e-um-z-score/>.

8 RETORNO, E. M. Zscore: saiba o que é e como funciona. In: . [s.n.], 2022. Acesso em: 04 de janeiro de 2023. Disponível em: <https://maisretorno.com/portal/termos/z/z-score/>.

9 BRASIL eHow. Como calcular o zscore em estatística. In: . [s.n.], 20/11/2021. v. 1, n. 1, p. 1. Disponível em: <https://www.ehow.com.br/calcular-zescore-estatistica-como_24508/>.

10 DEFINIRTEC, p. V. Máquina vetorial de suporte (svm). In: . [s.n.], 2023. p. 1. Acesso em: 06 de abril de 2022. Disponível em: <https://definirtec.com/maquina-vetorial-de-suporte-svm/>.

11 EDUCATION, I. C. O que são redes neurais? In: . [s.n.], 2020. p. 1. Acesso em: 03 de abril de 2022. Disponível em: <https://www.ibm.com/br-pt/cloud/learn/neural-networks/>.

12 EDUCATION, I. C. O que é machine learning? In: . [s.n.], 2020. p. 1. Acesso em: 05 de abril de 2022. Disponível em: <https://www.ibm.com/br-pt/cloud/learn/machine-learning>.

13 SOSHACE, p. M. V. Deep learning x machine learning: visão geral e comparação. In: . [s.n.], 2019. v. 1, n. 1, p. 1. Acesso em: 04 de janeiro de 2023. Disponível em: <https://soshace.com/deep-learning-vs-machine-learning-overview-comparison/>.

14 RUSSELL, S.; NORVIG, P. Inteligência artificial: uma abordagem moderna. *Campus,*, 2004.

15 SIMON; SCHUSTER (Ed.). *Sociedade da mente*. [S.l.: s.n.].

16 PRESS, M. (Ed.). *Aprendizado profundo*. [S.l.: s.n.].

17 VISãO computacional: algoritmos e aplicações. [S.l.: s.n.].

18 GEOFFREY, G. E. e. M. A.-r. e. o. H. Redes neurais profundas para modelagem acústica no reconhecimento de fala: as visões compartilhadas de quatro grupos de pesquisa. *revista de processamento de sinal IEEE*, v. 29.

19 JUSTIçA e Machine Learning. fairmlbook. org.

20 DEPUTADOS, C. dos. *Cidades Inteligentes: Uma abordagem humana e sustentável*. [S.l.]: Edições Câmara, 2021.

21 FREEBASE: um banco de dados de grafos criado colaborativamente para estruturar o conhecimento humano. In: PROCEEDINGS of the 2008 ACM SIGMOD international conference on Management of data. [S.l.: s.n.].

22 TUKEY, J. W. *Exploratory Data Analysis*. [S.l.]: Pearson; 1ª edição (11 janeiro 1977), 1977. v. 1.

23 RODRIGUES, L. boxplot. In: . [s.n.], 2021. p. 1. Acesso em: 04 de Janeiro de 2023. Disponível em: <https://www.voitto.com.br/blog/artigo/boxplot/>.

24 COUTINHO, T. Diagrama de dispersão. In: . [s.n.], 2020. p. 1. Acesso em: 02 de abril de 2022. Disponível em: <https://www.voitto.com.br/blog/artigo/diagrama-de-dispersao/>.

25 COUTINHO, T. O que e um histograma. In: . [s.n.], 2018. p. 1. Acesso em: 07 de abril de 2022. Disponível em: <https://www.voitto.com.br/blog/artigo/o-que-e-um-histograma/>.

26 HUMANO., S. descubra e entenda diversos temas do conhecimento. Quais as partes de um histograma. In: . [s.n.], 2021. p. 1. Acesso em: 04 de Janeiro de 2023. Disponível em: <https://www.significados.com.br/histograma/>.

27 EDUCAçãO, S. O que é um histograma? In: *https://www.significados.com.br/histograma/ Acesso em: 04 de janeiro de 2022*. [S.l.: s.n.], 2011–2023.

28 IBERDROLA, S. Mobile learning: bem-vindos à nova realidade nas salas de aula. In: . [s.n.], 2023. v. 1, n. 1, p. 1. Acesso em: 17 de julho 2023. Disponível em: <https://www.iberdrola.com/talentos/o-que-e-m-learning-e-vantagens#:~:text=O%20Mobile%20learning%20ou%20m,adaptando%20o%20mesmo%20a%20metodologia.>

29 GMBH., R. Arquitetura pedagógica de aprendizagem móvel. In: . [s.n.], 2008–2023. v. 1, n. 1, p. 1. Acesso em: 17 de julho 2023. Disponível em: <https://www.researchgate.net/figure/Figura-9-Arquitetura-pedagogica-de-aprendizagem-movel-E-valido-ressaltar-que-os-grupos_fig8_309887927>.

30 SONEGO, A. H. S.; BEHAR, P. A. M-learning: o uso de dispositivos móveis por uma geração conectada. *Educação*, PUC-RS, v. 42, n. 3, p. 514–524, 2019.

31 FIA, B. S. Mobile learning: Conceito, tendência, como funciona e vantagens. In: . [s.n.], 2021. v. 1, n. 1, p. 1. Acesso em: 18 de julho 2023. Disponível em: <https://fia.com.br/blog/mobile-learning-conceito-tendencia-como-funciona-e-vantagens/>.

32 ALURA. O que é pandas? In: . [s.n.], 2023. Acesso em: 28 de julho de 2023. Disponível em: <https://www.alura.com.br/artigos/pandas-o-que-e-para-que-serve-como-instalar>.

33 SCHRÖDER, L. de C.; MARQUES, M. A. M.; SILVA, W. de A.; RAMOS, N. C. S.; SELEME, R.; DROZDA, F. O.; CASTRO, A. de et al. Análise da integridade de dados para construção de kpi´ s na produção: estudo de caso em uma empresa de mineração. *Brazilian Journal of Development*, v. 5, n. 8, p. 12283–12301, 2019.

34 ENAP., E. N. de A. P. Governança de dados - [curso online]. In: . [s.n.], 2021. Acesso em: 23 de março de 2022. Disponível em: <https://www.escolavirtual.gov.br/curso/270>.

35 BRANDT, M. B.; VIDOTTI, S. A. B. G. Arquitetura da informação para processos de negócio: um caminho para a governança de dados. *Informação & Sociedade: Estudos*, v. 30, n. 4, p. 1–16, 2020.

36 LUCIANO, E. M.; WIEDENHOFT, G. C.; SANTOS, F. P. dos. Barreiras para a ampliação de transparência na administração pública brasileira: Questões estruturais e culturais ou falta de estratégia e governança? *Administração Pública e Gestão Social*, 2018.

37 GALVÃO, N. D.; MARIN, H. d. F. Técnica de mineração de dados: uma revisão da literatura. *Acta Paulista de Enfermagem*, SciELO Brasil, v. 22, p. 686–690, 2009.

38 SANTOS, L. H. L. Sobre a integridade ética da pesquisa. fundação de amparo à pesquisa do estado de são paulo. In: . [s.n.], 2011. Acesso em: 04 de janeiro de 2022. Disponível em: <http://www.fapesp.br/6566/>.

39 (2017), A. B. D. A. Qualidade de dados. In: . [s.n.]. Acesso em: 04 de janeiro de 2022. Disponível em: <https://blog.gs1br.org/qualidade-de-dados-o-que-e-e-qual-a-importancia-para-negocios/>.

40 PORTAL Em Numeros. In: . [s.n.], 2019. v. 1, n. 1. Acesso em: 04 de junho de 2023. Disponível em: <https://emnumeros.escolavirtual.gov.br/indicadores/>.

41 LEI nº 12.527, de 18 de novembro de 2011 - Lei de Acesso a Informação LAI. In: . [s.n.], 2011. Acesso em: 04 de março de 2022. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/12527>.

42 CARVALHO, B. B. et al. *Boas práticas de integridade de dados em registros eletrônicos gerados em sistemas computadorizados de equipamentos analíticos: uma proposta para os laboratórios de pesquisa científica experimental da Fiocruz*. Tese (Doutorado), 2021.

43 NEGóCIOS, I. de. O que é integridade de dados e por que ela é importante? In: . São Paulo: [s.n.], 2020. Acesso em: 07 de janeiro de 2022. Disponível em: <https://blog.in1.com.br/o-que-%C3%A9-integridade-de-dados-e-por-que-ela-%C3%A9-importante/>.

44 SOARES TARCISO DAL MASO JARDIM, T. B. V. H. Fabiana de M. Lei de acesso à informação no brasil. In: . Minas Gerais: [s.n.], 2013. v. 1, p. 1–44. Acesso em: 10 de setembro de 2022. Disponível em: <https://www12.senado.leg.br/transparencia/arquivos/sobre/cartilha-lai//>.

45 GHOSH, D.; VOGT, A. Outliers: An evaluation of methodologies. In: *Joint statistical meetings*. [S.l.: s.n.], 2012. v. 2012.

46 BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, Elsevier, v. 114, p. 24–31, 2016.

47 PúBLICA, E. N. de A. Análise do perfil dos alunos na enap. In: . Brasília: [s.n.], 2023. Acesso em: 01 de agosto de 2023. Disponível em: <https://www.enap.gov.br/analise-perfil-alunos>.

48 ENAP. Escola nacional de administração pública - enap. In: *https://www.enap.gov.br/pt/ Acesso em: 10 de março de 2022.* [S.l.: s.n.], 2019. v. 1.

49 SCIKIT-LEARN. Classificador de random forest. In: . [s.n.], 2007–2023. v. 1, n. 1, p. 1. Acesso em: 04 de abril de 2022. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.

50 SCIKIT-LEARN. Máquinas de vetores de suporte (svms). In: . [s.n.], 2007–2023. v. 1, n. 1, p. 1. Acesso em: 02 de abril de 2022. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>.

51 SCIKIT-LEARN. Classificador de árvore de decisão. In: . [s.n.], 2007–2023. v. 1, n. 1, p. 1. Acesso em: 02 de abril de 2022. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier>.

APÊNDICES

I A - CÓDIGO PYTHON

I.1 ANÁLISE DO PERFIL DE ALUNOS

1.1 Número de inscritos por conteudistas

```
df.conteudista.value_counts().nlargest(40).plot(kind='bar', figsize=(10,5))
plt.title('Numero de inscritos por conteudistas')
plt.ylabel('Numero de inscritos')
plt.xlabel('Conteudistas');
```

1.2 Número de inscritos por curso

```
df.nome\_curso.value\_counts().nlargest(40).plot(kind='bar', figsize=(10,5))
plt.title('Numero de inscritos por curso')
plt.ylabel('Numero de inscritos')
plt.xlabel('Curso')
plt.figure(figsize=(20,10))
c= df.corr()
sns.heatmap(c,cmap="BrBG",annot=True)
c
```

1.3 Número de desistentes por estado

```
df_desistente = df[df['sit_matricula'] == 'Desistente']
```

```
df_desistente.uf_pessoa.value_counts().nlargest(40).plot(kind='bar',
,figsize=(10,5))
plt.title('Numero de desistentes por estado')
plt.ylabel('Numero de desistentes')
plt.xlabel('Estado');
```

1.4 Número de desistentes por município

```
df_desistente.municipio_pessoa.value_counts().nlargest(40).plot(kind='bar',figsize
    =(10,5))
plt.title('Numero de desistentes por Municipio')
plt.ylabel('Numero de desistentes')
plt.xlabel('Municipio');
```

1.5 Número de Reprovados por Estado

```
df_reprovado= df[df['sit_matricula'] == 'Reprovado'
```

```
df_reprovado.uf_pessoa.value_counts().nlargest(40).plot(kind='bar'
,figsize=(10,5))
plt.title('Numero de Reprovados por estado')
plt.ylabel('Numero de Reprovados')
plt.xlabel('Estado');
```

1.6 Número de Reprovados por município

```
df_reprovado.municipio_pessoa.value_counts().nlargest(40).plot(kind='bar'
,figsize=(10,5))
plt.title('Numero de Reprovados por Municipio')
plt.ylabel('Numero de Reprovados')
plt.xlabel('Municipio');
```

1.7 Número de aprovados por estado

```
df_concluido = df[df['sit_matricula'] == 'Concluida']
```

```
df_concluido.uf_pessoa.value_counts().nlargest(40).plot(kind='bar',
figsize=(10,5))
plt.title('Numero de Aprovados por estado')
plt.ylabel('Numero de Aprovados')
plt.xlabel('Estado');
```

1.8 Número de Reprovados por município

```
df_concluido.municipio_pessoa.value_counts().nlargest(40).plot(kind='bar',figsize
    =(10,5))
plt.title('Numero de Aprovados por Municipio')
plt.ylabel('Numero de Aprovados')
plt.xlabel('Municipio');
```

1.9 Número de trancamentos por estado

```
df_trancada = df[df['sit_matricula'] == 'Trancada']
```

```
df_trancada.uf_pessoa.value_counts().nlargest(40).plot(kind='bar'
,figsize=(10,5))
plt.title('Numero de Trancamentos por estado')
plt.ylabel('Numero de Trancamentos')
plt.xlabel('Estado');
```

1.10 Número de trancamentos por município

```
df_trancada.municipio_pessoa.value_counts().nlargest(40).plot(kind='bar',figsize
    =(10,5))
plt.title('Numero de Trancamentos por Municipio')
plt.ylabel('Numero de Trancamentos')
plt.xlabel('Municipio');
```

I.1.1 Agrupamento, concluído, reprovado, trancado e não concluído

```
df['sexo'].replace({'feminino':0, 'masculino':1, 'outro':2},inplace=True)
```

```
'#It was necessary to change the label of enrollment situation ue to the one,hot
    encoding.
'#' 'Desistente': 0,'Concluida': 1,'Reprovado': 2, 'Trancada': 3,'N o , Conclu do
    ': 4" '

df['sit_matricula'].replace({'Desistente' : 0,'Concluida': 1, 'Reprovado': 0,'
    Trancada': 0,'N o  Conclu do': 0},inplace=True)
```

```
df.groupby(['sit_matricula']).count()
```

I.1.2 Agrupamento, situação, conteudista, temática e tempo inscrição aberta

```
df.to_parquet('df.parquet.gzip',
              compression='gzip')

df = pd.read_parquet('df.parquet.gzip')

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

df.head()
```

I.1.3 Situação matrícula

```
labels = np.array(df['sit_matricula'])

labels

array([0, 1, 0, , 0, 1, 0])

features = df.drop(columns=['sit_matricula'])
features = features.drop(columns=['tempo_inscricao_aberta'])
features = features.drop(columns=['conteudista'])
features = features.drop(columns=['tematica'])

features.head()

features = np.array(features)
'# Using Skicit-learn to split data into training and testing sets
# Split the data into training and testing sets'

train_ratio = 0.70
validation_ratio = 0.20
test_ratio = 0.10

'# train is now 70% of the entire data set
# the _junk suffix means that we drop that variable completely'

train_features, test_features, train_labels, test_labels =train_test_split(
    features, labels, test_size=1 - train_ratio)

print(train_features.shape)
print(test_features.shape)
print(x_test.shape)
print(y_test.shape)
print(x_val.shape)
print(y_val.shape)
```

I.1.4 Classificador Rondom Forest

```
n_estimators = [4,8, 16, 32,64,128,256]
predictions = []*len(n_estimators)
accuracy = []
for x in n_estimators:
    rf = RandomForestClassifier(n_estimators = x, random_state = 42)
    '# Train the model on training data'
    rf.fit(train_features, train_labels);
    predictions = rf.predict(x_val)
    print(x)
    print(classification_report(predictions.round(), y_val))
    print('\n')
```

I.2 CLASSIFICADORES DE MÁQUINAS DE VETORES DE SUPORTE (SVM)

```
clf = make_pipeline(StandardScaler(), SVC(kernel='linear', gamma='auto'))

clf.fit(train_features, train_labels)

predictions = clf.predict(x_val)

 print(classification_report(predictions.round(), y_val))
```

I.3 CLASSIFICADOR DE ÁRVORE DE DECISÃO

```
clf = tree.DecisionTreeClassifier()
clf.fit(train_features, train_labels)

DecisionTreeClassifier()

predictions = clf.predict(x_val)

print(classification_report(predictions.round(), y_val))
```

I.4 PERCEPTRON DE VÁRIAS CAMADAS (MLP)

```
def sigmoid_act(x, der=False):
    import numpy as np

    if (der==True) : #derivative of the sigmoid
        f = 1/(1+ np.exp(- x))*(1-1/(1+ np.exp(- x)))
    else : # sigmoid
        f = 1/(1+ np.exp(- x))

    return f

# We may employ the Rectifier Linear Unit (ReLU)
def ReLU_act(x, der=False):
import numpy as np

     if (der == True): # the derivative of the ReLU is the Heaviside Theta
        f = np.heaviside(x, 1)
    else :
        f = np.maximum(x, 0)
    return f
```

```
 '''
Artificial Neural Network Class
'''

class ANN:
     import numpy as np '# linear algebra'
     np.random.seed(10)


'''
Initialize the ANN;
HiddenLayer vector : will contain the Layers info
w, b, phi = (empty) arrays that will contain all the w, b and activation
,functions for all the Layers
mu = cost function
eta = a standard learning rate initialization. It can be modified by the
,set_learning_rate method

'''

def __init__(self) :
    self.HiddenLayer = []
    self.w = []
    self.b = []
    self.phi = []
    self.mu = []
```

```
    self.eta = 1 '#set up the proper Learning Rate!!'

'''
add method: to add layers to the network
'''

def add(self, lay = (4, 'ReLU') ):
    self.HiddenLayer.append(lay)

'''
FeedForward method: as explained before.
'''

@staticmethod
def FeedForward(w, b, phi, x):
    return phi(np.dot(w, x) + b)

'''
BackPropagation algorithm implementing the Gradient Descent
'''

def BackPropagation(self, x, z, Y, w, b, phi):
    self.delta = []

'# We initialize ausiliar w and b that are used only inside the
,backpropagation algorithm once called'

    self.W = []
    self.B = []

'# We start computing the LAST error, the one for the OutPut Layer'
self.delta.append( (z[len(z)-1] - Y) * phi[len(z)-1](z[len(z)-1],
,der=True) )

'''Now we BACKpropagate'''
'# We thus compute from next-to-last to first'

    for i in range(0, len(z)-1):
        self.delta.append( np.dot( self.delta[i], w[len(z)- 1 - i] ) *
,phi[len(z)- 2 - i](z[len(z)- 2 - i], der=True) )

'# We have the error array ordered from last to first; we flip it to
,order it from first to last'

     self.delta = np.flip(self.delta, 0)

'# Now we define the delta as the error divided by the number of
,training samples'

     self.delta = self.delta/self.X.shape[0]

'''GRADIENT DESCENT'''
```

```
79
80 '# We start from the first layer that is special, since it is connected
81 ,to the Input Layer'
82
83     self.W.append( w[0] - self.eta * np.kron(self.delta[0], x).reshape(
84 ,len(z[0]), x.shape[0] ) )
85     self.B.append( b[0] - self.eta * self.delta[0] )
86
87 '# We now descend for all the other Hidden Layers + OutPut Layer'
88     for i in range(1, len(z)):
89         self.W.append( w[i] - self.eta * np.kron(self.delta[i], z[i-1]).
90 ,reshape(len(z[i]), len(z[i-1])) )
91         self.B.append( b[i] - self.eta * self.delta[i] )
92
93 '# We return the descended parameters w, b'
94     return np.array(self.W), np.array(self.B)
95
96 '''
97 Fit method: it calls FeedForward and Backpropagation methods
98 '''
99
100     def Fit(self, X_train, Y_train):
101         print('Start fitting...')
102
103         '''
104         Input layer
105         '''
106
107         self.X = X_train
108         self.Y = Y_train
109
110 '''
111 We now initialize the Network by retrieving the Hidden Layers and
112 ,concatenating them
113 '''
114
115         print('Model recap: \n')
116         print('You are fitting an ANN with the following amount of layers: ',
117 ,len(self.HiddenLayer))
118
119     for i in range(0, len(self.HiddenLayer)) :
120         print('Layer ', i+1)
121         print('Number of neurons: ', self.HiddenLayer[i][0])
122         if i==0:
123
124 '# We now try to use the He et al. Initialization from ArXiv:
125 ,1502.01852'
126
127     self.w.append( np.random.randn(self.HiddenLayer[i][0] , self.X.
128 ,shape[1])/np.sqrt(2/self.X.shape[1]) )
129
130     self.b.append( np.random.randn(self.HiddenLayer[i][0])/np.
```

```
,sqrt(2/self.X.shape[1]))

'# Old initialization'

'#self.w.append(2 * np.random.rand(self.HiddenLayer[i][0] , self.
,X.shape[1]) - 0.5)

#self.b.append(np.random.rand(self.HiddenLayer[i][0]))

# Initialize the Activation function'

    for act in Activation_function.list_act():
        if self.HiddenLayer[i][1] == act :
            self.phi.append(Activation_function.get_act(act))
            print('\tActivation: ', act)

else :

'# We now try to use the He et al. Initialization from ArXiv:
,1502.01852'

    self.w.append( np.random.randn(self.HiddenLayer[i][0] , self.
,HiddenLayer[i-1][0] )/np.sqrt(2/self.HiddenLayer[i-1][0]))

    self.b.append( np.random.randn(self.HiddenLayer[i][0])/np.
,sqrt(2/self.HiddenLayer[i-1][0]))

'# Old initialization

#self.w.append(2*np.random.rand(self.HiddenLayer[i][0] , self.
,HiddenLayer[i-1][0] ) - 0.5)

#self.b.append(np.random.rand(self.HiddenLayer[i][0]))

# Initialize the Activation function'

    for act in Activation_function.list_act():
        if self.HiddenLayer[i][1] == act :
            self.phi.append(Activation_function.get_act(act))

            print('\tActivation: ', act)

'''
Now we start the Loop over the training dataset
'''

    for I in range(0, self.X.shape[0]): # loop over the training set

'''
Now we start the feed forward
'''

```

```
183         self.z = []
184 
185         self.z.append( self.FeedForward(self.w[0], self.b[0], self.phi[0],
186 ,self.X[I]) ) # First layers
187 
188         for i in range(1, len(self.HiddenLayer)): #Looping over layers
189 self.z.append( self.FeedForward(self.w[i], self.b[i], self.
190 ,phi[i], self.z[i-1] ) )
191 
192     '''
193     Here we backpropagate
194     '''
195 
196     self.w, self.b = self.BackPropagation(self.X[I], self.z, self.
197 ,Y[I], self.w, self.b, self.phi)
198 
199     '''
200     Compute cost function
201     '''
202 
203     self.mu.append(
204         (1/2) * np.dot(self.z[len(self.z)-1] - self.Y[I], self.
205 ,z[len(self.z)-1] - self.Y[I])
206 )
207 
208     print('Fit done. \n')
209 
210 '''
211 predict method
212 '''
213 
214 def predict(self, X_test):
215 
216     print('Starting predictions...')
217 
218     self.pred = []
219     self.XX = X_test
220     for I in range(0, self.XX.shape[0]): # loop over the training set
221 
222     '''
223     Now we start the feed forward
224     '''
225 
226     self.z = []
227 
228     self.z.append(self.FeedForward(self.w[0] , self.b[0], self.phi[0],
229 ,self.XX[I])) #First layer
230 
231     for i in range(1, len(self.HiddenLayer)) : # loop over the layers
232 self.z.append( self.FeedForward(self.w[i] , self.b[i], self.
233 ,phi[i], self.z[i-1]))
234 
```

```
'# Append the prediction;
# We now need a binary classifier; we this apply an Heaviside Theta, and we set to
      0.5 the threshold
# if y < 0.5 the output is zero, otherwise is zero'

    self.pred.append( np.heaviside( self.z[-1] - 0.5, 1)[0] ) '# NB:
,self.z[-1] is the last element of the self.z list'

    print('Predictions done. \n')

    return np.array(self.pred)

'''
We need a method to retrieve the accuracy for each training data to follow
,the learning of the ANN
'''

def get_accuracy(self):
    return np.array(self.mu)

'# This is the averaged version'

def get_avg_accuracy(self):
    import math
    self.batch_loss = []
    for i in range(0, 10):
        self.loss_avg = 0

'# To set the batch in 10 element/batch we use math.ceil method
# int(math.ceil((self.X.shape[0]-10) / 10.0)) - 1'
    for m in range(0, (int(math.ceil((self.X.shape[0]-10) / 10.0))
,)-1):

'#self.loss_avg += self.mu[60*i+m]/60'
        self.loss_avg += self.mu[(int(math.ceil((self.X.shape[0]-10) /
,10.0)) )*i + m]/(int(math.ceil((self.X.shape[0]-10) / 10.0)) )
        self.batch_loss.append(self.loss_avg)
    return np.array(self.batch_loss)

'''
Method to set the learning rate
'''

def set_learning_rate(self, et=1):
    self.eta = et

'''
layers class
'''

class layers :

```

```
'''
Layer method: used to call standar layers to add.
Easily generalizable to more general layers (Pooling and Convolutional
,layers)

'''
def layer(p=4, activation = 'ReLU'):
    return (p, activation)

'''
Activation functions class
'''

class Activation_function(ANN):
    import numpy as np
    def __init__(self) :
        super().__init__()

'''
Define the sigmoid activator; we ask if we want the sigmoid or its derivative
'''

def sigmoid_act(x, der=False):
    if (der==True) : #derivative of the sigmoid
        f = 1/(1+ np.exp(- x))*(1-1/(1+ np.exp(- x)))
    else : # sigmoid
        f = 1/(1+ np.exp(- x))
return f

'''
Define the Rectifier Linear Unit (ReLU)
'''

def ReLU_act(x, der=False):
    if (der == True): # the derivative of the ReLU is the Heaviside Theta
        f = np.heaviside(x, 1)
    else :
        f = np.maximum(x, 0)
    return f

def list_act():
    return ['sigmoid', 'ReLU']

def get_act(string = 'ReLU'):
    if string == 'ReLU':
        return ReLU_act
    elif string == 'sigmoid':
return sigmoid_act
else :
    return sigmoid_act
```

```
model = ANN()
model.add(layers.layer(8, 'ReLU'))
model.add(layers.layer(4, 'ReLU'))
model.add(layers.layer(1, 'sigmoid'))
model.set_learning_rate(0.8)
model.Fit(train_features, train_labels)
acc_val = model.get_accuracy()
acc_avg_val = model.get_avg_accuracy()
predictions = model.predict(x_val)

Start fitting
Model recap:
You are fitting an ANN with the following amount of layers: 3

Layer 1
Number of neurons: 8
        Activation: ReLU

Layer 2
Number of neurons: 4
       Activation: ReLU

Layer 3
Number of neurons: 1
       Activation: sigmoid


/usr/local/lib/python3.9/site-packages/numpy/core/_asarray.py:102:
VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences
(which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths
or shapes) is deprecated. If you meant to do this, you must specify
'dtype=object' when creating the ndarray.
return array(a, dtype, copy=False, order=order)

<ipython-input-17-f21382bb3a90>:71: VisibleDeprecationWarning: Creating an
ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-
    or ndarrays with different lengths or shapes) is deprecated. If you meant to do
     this, you must specify 'dtype=object' when creating the ndarray.
   return np.array(self.W), np.array(self.B)

Fit done.

Starting predictions...
Predictions done.
```

```
plt.figure(figsize=(10,6))
plt.scatter(np.arange(1, train_features.shape[0]+1) acc_val, alpha=0.3, s=4,
,label='mu')
plt.title('Loss for each training data point', fontsize=20)
plt.xlabel('Training data', fontsize=16)
```

```
plt.ylabel('Loss', fontsize=16)
plt.show()

plt.figure(figsize=(10,6))
plt.scatter(np.arange(1, len(acc_avg_val)+1), acc_avg_val, label='mu')
plt.title('Averege Loss by epoch', fontsize=20)
plt.xlabel('Training data', fontsize=16)
plt.ylabel('Loss', fontsize=16)
plt.show()
```

```
# Plot the confusion matrix

dict_live = {
    0 : 'Failed',
    1 : 'Approved'
}
cm = confusion_matrix(y_val, predictions)

df_cm = pd.DataFrame(cm, index  =   [dict_live[i] for i in range(0,2)], columns =
    [dict_live[i] for i in range(0,2)])
plt.figure(figsize = (7,7))
sns.heatmap(df_cm, annot=True, cmap=plt.cm.Blues, fmt='g')
plt.xlabel('"Predicted Class"', fontsize=18)
plt.ylabel('"True Class"', fontsize=18)
plt.show()
```

```
    print(classification_report(predictions.round(), y_val))
```