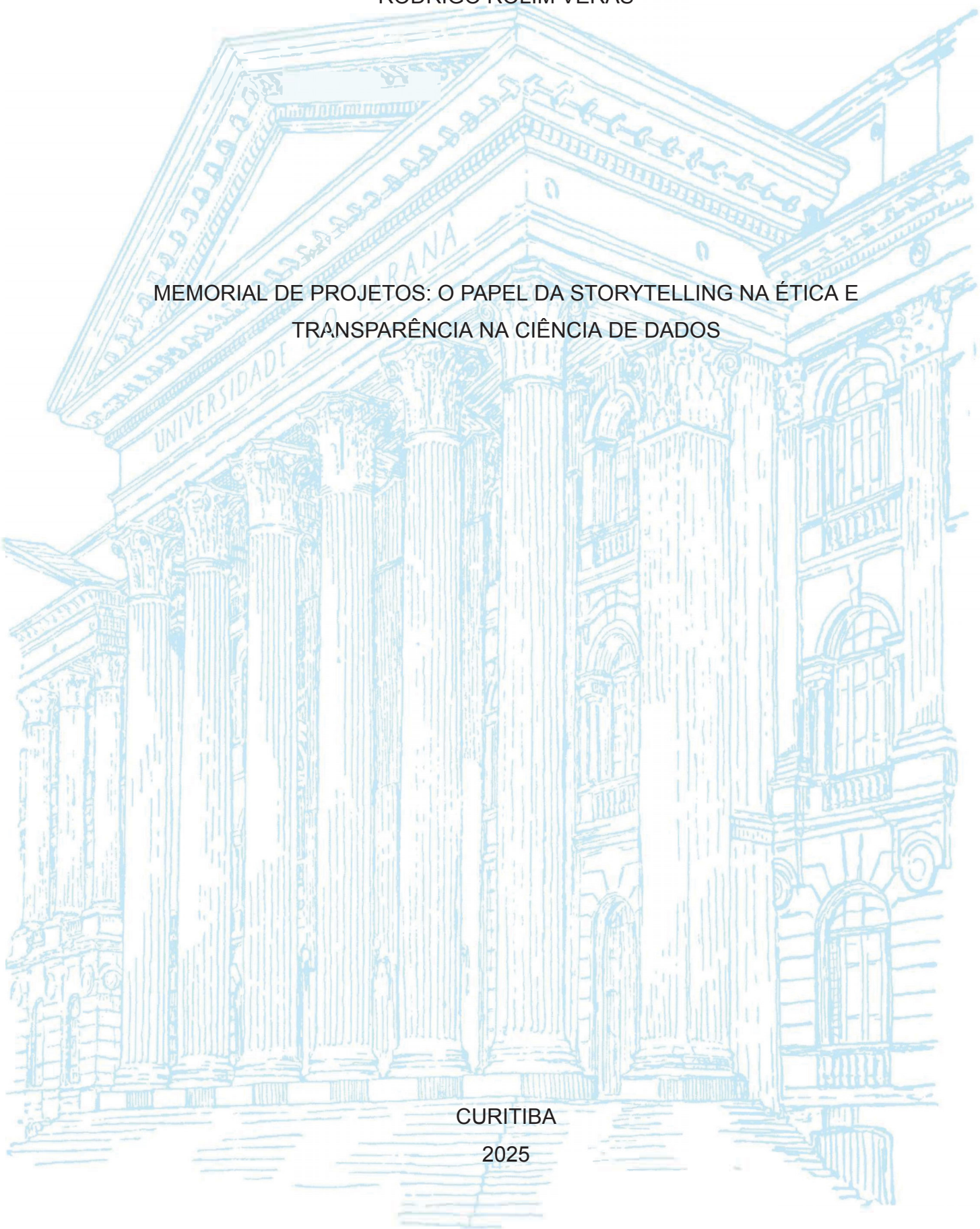
UNIVERSIDADE FEDERAL DO PARANÁ

RODRIGO ROLIM VERAS



MEMORIAL DE PROJETOS: O PAPEL DA STORYTELLING NA ÉTICA E TRANSPARÊNCIA NA CIÊNCIA DE DADOS

CURITIBA

2025

RODRIGO ROLIM VERAS

MEMORIAL DE PROJETOS: O PAPEL DA STORYTELLING NA ÉTICA E TRANSPARÊNCIA NA CIÊNCIA DE DADOS

Memorial de Projetos apresentado ao curso de Especialização em Inteligência Artificial Aplicada, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Especialista em Inteligência Artificial Aplicada.

Orientadora: Profa. Dra. Rafaela Mantovani Fontana

CURITIBA

2025

MINISTÉRIO DA EDUCAÇÃO
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
CURSO DE PÓS-GRADUAÇÃO INTELIGÊNCIA ARTIFICIAL APLICADA - 40001016399E1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação Inteligência Artificial Aplicada da Universidade Federal do Paraná foram convocados para realizar a arguição da Monografia de Especialização de **RODRIGO ROLIM VERAS**, intitulada: **MEMORIAL DE PROJETOS: O PAPEL DA STORYTELLING NA ÉTICA E TRANSPARÊNCIA NA CIÊNCIA DE DADOS**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua aprovação no rito de defesa.

A outorga do título de especialista está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 09 de Julho de 2025.

RAFAELA MANTOVANI FONTANA
Presidente da Banca Examinadora

JAIME WOJCIECHOWSKI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

RESUMO

O presente parecer técnico discute o papel do *data storytelling* sob a perspectiva da ética e da transparência na ciência de dados. A narrativa baseada em dados é uma estratégia fundamental para comunicar resultados de forma clara e acessível, mas envolve escolhas que podem influenciar a interpretação do público. Ao longo do texto, são apresentados conceitos teóricos como os pilares de um *data storytelling* ético, exemplos práticos de usos éticos e antiéticos, e as consequências da má aplicação do *data storytelling*. O trabalho destaca a importância de práticas responsáveis na construção de narrativas com dados, promovendo a confiança, a clareza e a integridade na comunicação de informações analíticas.

Palavras-chave: ciência de dados; ética; storytelling; transparência; visualização de dados.

ABSTRACT

This technical report discusses the role of data storytelling from the perspective of ethics and transparency in data science. Data-driven narratives are a key strategy for communicating analytical results in a clear and accessible way, but they involve choices that may influence public interpretation. This paper presents theoretical concepts like the three ethical pillars of the data story telling, real-world examples of ethical and unethical storytelling practices, and the potential consequences of misleading data narratives. It emphasizes the importance of responsible storytelling practices to promote trust, clarity, and integrity in data communication.

Keywords: data science; ethics; storytelling; transparency; data visualization.

SUMÁRIO

1. PARECER TÉCNICO

A ciência de dados busca extrair conhecimentos a partir dos dados. Construir uma narrativa ética vai além de apenas descobrir padrões e realizar previsões, envolve sobretudo comunicação não enviesada dos resultados. Toda narrativa envolve escolhas: o que mostrar, o que omitir e o que enfatizar. É nesse ponto que surgem desafios éticos e riscos de manipulação dos resultados (Nussbaumer, 2015; O'neil, 2016).

Storytelling, de acordo com a Alura (2022), é o processo ou a arte de contar uma história memorável e acionável com o objetivo de transmitir conhecimento, capturar o interesse ou influenciar comportamentos, emoções e tomadas de decisões do público-alvo.

Ainda de acordo com a Alura (2022), construir uma história memorável envolve elementos como o desenvolvimento da narrativa, a construção de personagens, o estabelecimento de cenários envolventes e o uso estratégico de conflitos e resoluções.

O *Storytelling*, na modernidade, é aplicado em diversas áreas, como no marketing, na liderança, no entretenimento, na educação, na ciência de dados e em muitas outras áreas.

Segundo Pessoa (2024), no contexto da ciência de dados, o storytelling, também conhecido como *data storytelling*, é a prática de combinar dados com os elementos da narração para comunicar *insights* de maneira eficiente e ao mesmo tempo envolvente e impactante para o público-alvo.

Segundo a Microsoft (s.d.), a *data storytelling* é fundamentada por três pilares essenciais: narrativa, recursos visuais e dados. Desenvolver uma narração de dados capaz de captar a atenção do público-alvo envolve a combinação inteligentes desses três pilares.

O processo envolve a combinação entre visualização de dados com gráficos e tabelas, elementos narrativos e contextos de forma coerente aos dados, ou seja, a narração não pode distorcer a verdade dos fatos, o que poderia induzir o público-alvo ao erro e a desinformação. A narração precisa está comprometida com a ética e a transparência, e precisa enfatizar a precisão, a honestidade e a clareza (Nussbaumer, 2015; Consulting club, s.d.).

Segundo Consulting Club (s.d.), o criador de uma narrativa precisa ter muita responsabilidade com os dados e o seu público-alvo. Omitir dados importantes ou forjar outros que lhe convém são práticas que ferem a moral e não condizem com a ética inerente ao *data storytelling*. Além de causar danos ao público-alvo, mina a credibilidade do próprio narrador.

Conforme Consulting Club (s.d.), as práticas ideais do *data storytelling* não devem nunca estar associadas com a manipulação dos dados, e sim com práticas honestas e verdadeiras que criem uma história envolvente e baseada em estatísticas reais.

De acordo com Consulting Club (s.d.), os sete princípios norteadores para um *data storytelling* ético são:

a) **precisão e honestidade:** evitar selecionar ou deturpar os dados para apoiar um ponto de vista específico. Os dados devem ser corretos e reais, sem manipulação ou distorção. Os arquitetos das narrativas precisam ser transparentes com público-alvo quanto às limitações ou incertezas dos dados;

b) **clareza e simplicidade:** usar de forma apropriada gráficos, cores, legendas, fontes entre outros elementos visuais para criar visualizações claras, simples e diretas, amenizando a complexidade dos dados, e tornando-os mais acessíveis ao público-alvo;

c) **justiça e objetividade:** criar uma narrativa justa e imparcial, sem vieses e sem favoritismos. Esse princípio garante que as narrativas não favoreçam a um ponto de vista específico e nem induzam o público-alvo a fazer julgamentos errôneos e tendenciosos;

d) **privacidade e confiança:** não exibir dados sensíveis e confidenciais nas narrações. A narrativa não pode violar o direito à privacidade e nem expor dados confidenciais sem consentimento. As narrações devem ser feitas respeitando as regulamentações de proteção aos dados;

e) **inclusão e acessibilidade:** o *data storytelling* em sua estrutura deve considerar a diversidade cultural e linguística do público-alvo, bem como considerar pessoas com deficiências como também parte do público-alvo.

Se um profissional de *data storytelling* seguir os princípios anteriores, certamente as suas narrações serão eticamente mais respaldadas, dando maior ao profissional maior credibilidade.

Um exemplo prático de *data storytelling* ético ocorreu em 2023, quando a ONG Nyakara, que atua no desenvolvimento educacional em Uganda, integrou dados de presença e desempenho escolar com relatos de estudantes, transmitindo de forma clara e transparente o impacto de suas ações (Surveycto, 2023).

Em seu relatório de impacto de 2023, a organização combinou indicadores quantitativos com relatos pessoais de alunos atendidos pelo programa, ilustrando como o programa educacional da ONG transformou suas vidas. Ao integrar dados objetivos com narrativas humanas, a ONG ofereceu uma narrativa clara e honesta dos resultados alcançados, respeitando o contexto e evitando a manipulação de dados (Surveycto, 2023).

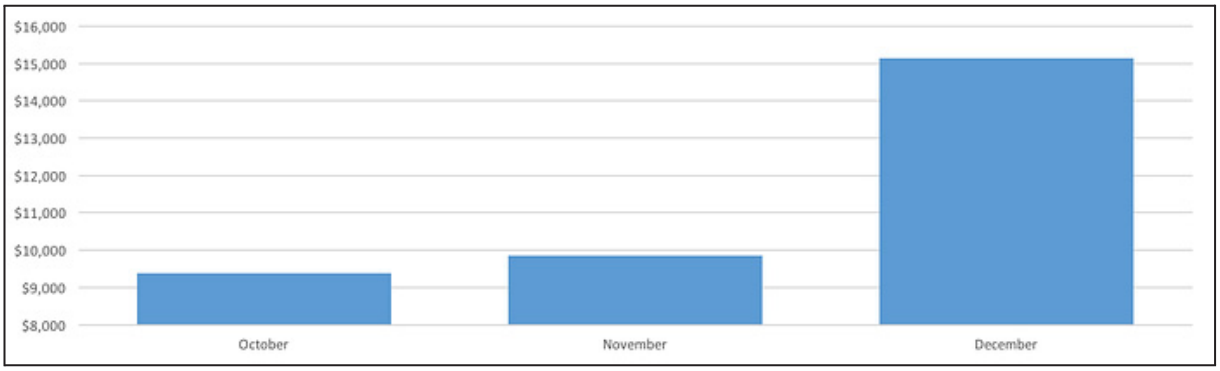
Essa foi certamente uma abordagem que reforçou os princípios da transparência e da confiança, o que evitou interpretações equivocadas e permitiu que o público-alvo entendesse não apenas o que aconteceu, mas também por que e para quem os resultados são relevantes (Surveycto, 2023).

Ao integrar dados precisos com narrativas humanas, a organização conseguiu os seguintes resultados positivos: tornou clara a relação entre os números e os impactos na vida real; educou os doadores e *stakeholders* sobre o funcionamento do programa; e aumentou a confiança e a credibilidade da ONG, resultando no fim em maior engajamento e apoio financeiro (Surveycto, 2023).

Um exemplo de um *data storytelling* antiético é apresentado por Outlier AI (2017), em um artigo onde o autor cria deliberadamente uma narrativa falsa com dados manipulados.

No gráfico de barras presente no artigo Outlier AI (2017) e exibido a seguir, o autor propositalmente trunca a escala (inicia a base do gráfico com um valor diferente de zero) amplificando visualmente uma salto de crescimento de 1.150% do produto D; ele também fez uma seleção seletiva dos dados (*cherry picking*) que colaboram com o que o autor deseja confirmar (rápido crescimento do produto D) e traz conclusões sem nenhuma evidência lógica (*jumping to conclusions*).

GRÁFICO 1 - Lucro nos meses de outubro, novembro e dezembro com a base truncada



FONTE: Outlier AI (2017)

E quais são as consequências de se propagar uma narrativa (*data* storytelling) antiético? As consequências podem ser graves, tanto no nível social quanto organizacional. Alguns impactos a seguir:

a) **desinformação do público:** o autor O'Neil (2026) mostra como narrativas com dados distorcidos podem influenciar negativamente uma sociedade, manipulando negativamente a opinião pública e até mesmo reforçar desigualdades;

b) **perda de credibilidade:** a autora Nussbaumer (2015) discorre sobre a relevância de visualizações e narrativas com dados honestas e como apresentações tendenciosas, ainda que visualmente atraentes, geram desconfiança e enfraquecem a relação com o público-alvo, investidores e parceiros;

c) **riscos legais e éticos:** em áreas reguladas (saúde, finanças, governo e etc), distorcer ou exibir dados sensíveis e confidenciais podem violar leis de transparência, LGPD ou normas de *compliance* (Dignum, 2019);

d) **tomada de decisão ruim ou ineficaz:** se os gestores de uma instituição basearem suas decisões em gráficos ou indicadores manipulados pode resultar em ações ineficazes ou prejudiciais, ocasionado desperdício de recursos (Few, 2012);

e) **degradação da cultura de dados:** se a equipe percebe que os dados são usados para reforçar narrativas convenientes e não o que os dados de fato dizem (a verdade), a equipe pode perder o engajamento com a cultura

analítica, pois a percepção de dados tendenciosos pode minar a confiança entre analistas e tomadores decisão (Patil; Mason, 2015).

O *data storytelling* é de fato uma ferramenta muito poderosa e essencial na comunicação de resultados em ciência de dados. O uso dessa ferramenta deve ser usada não só de forma inteligente, mas também ética. Narrativas que distorcem ou omitem informações comprometem a transparência e a honestidade, gerando desinformação e afetando negativamente a tomada de decisão.

Portanto, contar histórias com dados não deve ser apenas uma questão de clareza, impacto e elegância, mas também de compromisso com a verdade, a precisão e a confiança do público-alvo.

REFERÊNCIAS

ALURA. **O que é storytelling?** Como usar essa técnica no marketing, nas vendas e em apresentações. *Alura*, [S. l.], 21 fev. 2022. Disponível em: https://www.alura.com.br/artigos/storytelling. Acesso em: 30 jun. 2025.

CONSULTING CLUB. ***Storytelling com dados: qual é a sua história?*** Consulting Club, [S. l.], [s. d.]. Disponível em: https://www.consultingclub.com.br/post/storytelling-com-dados-qual-%C3%A9-a-sua-hist%C3%B3ria. Acesso em: 30 jun. 2025.

DIGNUM, Virginia. ***Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*****.** Springer, 2019.

FEW, Stephen. ***Show Me the Numbers: Designing Tables and Graphs to Enlighten*****.** Analytics Press, 2012.

MICROSOFT. ***What is data storytelling?*** Microsoft Power BI, [S. l.], [s. d.]. Disponível em: https://www.microsoft.com/en-us/power-platform/products/power-bi/topics/data-storytelling. Acesso em: 30 jun. 2025.

NUSSBAUMER KNAFLIC, Cole. ***Storytelling with Data: A Data Visualization Guide for Business Professionals*****.** Wiley, 2015.

OUTLIER AI. ***How to properly tell a story with data — and common pitfalls to avoid*****.** Medium, 29 maio 2017. Disponível em: https://medium.com/data-science/how-to-properly-tell-a-story-with-data-and-common-pitfalls-to-avoid-317d8817e0c9. Acesso em: 1 jul. 2025.

O'NEIL, Cathy. ***Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*****.** Crown Publishing Group, 2016.

PATIL, DJ; HAMMERBREAKER, Hilary Mason. ***Data-Driven: Creating a Data Culture*****. O'Reilly Media,** 2015.

PESSOA, Mariana. ***Data Storytelling: o que é, qual a importância e mais!*** Conversion, [S. l.], 13 jun. 2024. Disponível em: https://www.conversion.com.br/blog/data-storytelling-o-que-e-qual-a-importancia-e-mais/. Acesso em: 30 jun. 2025.

SURVEYCTO. ***Drive impact and engagement with data storytelling*****.** Surveycto, 2023. Disponível em: https://www.surveycto.com/analysis-reporting/data-storytelling-impact/. Acesso em: 30 jun. 2025.

APÊNDICE 1 – INTRODUÇÃO À INTELIGÊNCIA ARTIFICIAL

A – ENUNCIADO

1 ChatGPT

a) **(6,25 pontos)** Pergunte ao ChatGPT o que é Inteligência Artificial e cole aqui o resultado.
b) **(6,25 pontos)** Dada essa resposta do ChatGPT, classifique usando as 4 abordagens vistas em sala. Explique o porquê.
c) **(6,25 pontos)** Pesquise sobre o funcionamento do ChatGPT (sem perguntar ao próprio ChatGPT) e escreva um texto contendo no máximo 5 parágrafos. Cite as referências.
d) **(6,25 pontos)** Entendendo o que é o ChatGPT, classifique o próprio ChatGPT usando as 4 abordagens vistas em sala. Explique o porquê.

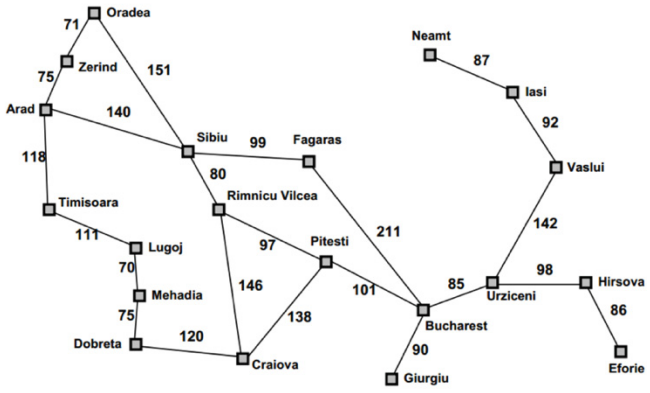
2 Busca Heurística

Realize uma busca utilizando o algoritmo A* para encontrar o melhor caminho para chegar a **Bucharest** partindo de **Lugoj**. Construa a árvore de busca criada pela execução do algoritmo apresentando os valores de *f(n)*, *g(n)* e *h(n)* para cada nó. Utilize a heurística de distância em linha reta, que pode ser observada na tabela abaixo.

Essa tarefa pode ser feita em uma **ferramenta de desenho**, ou até mesmo no **papel**, desde que seja digitalizada (foto) e convertida para PDF.

a) **(25 pontos)** Apresente a árvore final, contendo os valores, da mesma forma que foi apresentado na disciplina e nas práticas. Use o formato de árvore, não será permitido um formato em blocos, planilha, ou qualquer outra representação.

NÃO É NECESSÁRIO IMPLEMENTAR O ALGORITMO.



Arad	366	**Mehadia**	241
Bucareste	0	**Neamt**	234
Craiova	160	**Oradea**	380
Drobeta	242	**Pitesti**	100
Eforie	161	**Rimnicu Vilcea**	193
Fagaras	176	**Sibiu**	253
Giurgiu	77	**Timisoara**	329
Hirsova	151	**Urziceni**	80
Iasi	226	**Vaslui**	199
Lugoj	244	**Zerind**	374

Figura 3.22 Valores de *hDLR* — distâncias em linha reta para Bucareste.

3 Lógica

Verificar se o argumento lógico é válido.

Se as uvas caem, então a raposa as come
Se a raposa as come, então estão maduras
As uvas estão verdes ou caem

Logo

A raposa come as uvas se e somente se as uvas caem

Deve ser apresentada uma prova, no mesmo formato mostrado nos conteúdos de aula e nas práticas.

Dicas:

1. Transformar as afirmações para lógica:

p: as uvas caem
q: a raposa come as uvas
r: as uvas estão maduras

2. Transformar as três primeiras sentenças para formar a base de conhecimento

R1: $p \rightarrow q$
R2: $q \rightarrow r$
R3: $\neg r \vee p$

3. Aplicar equivalências e regras de inferência para se obter o resultado esperado. Isto é, com essas três primeiras sentenças devemos derivar $q \leftrightarrow p$. Cuidado com a ordem em que as fórmulas são geradas.

Equivalência Implicação: $(\alpha \rightarrow \beta)$ equivale a $(\neg\alpha \vee \beta)$

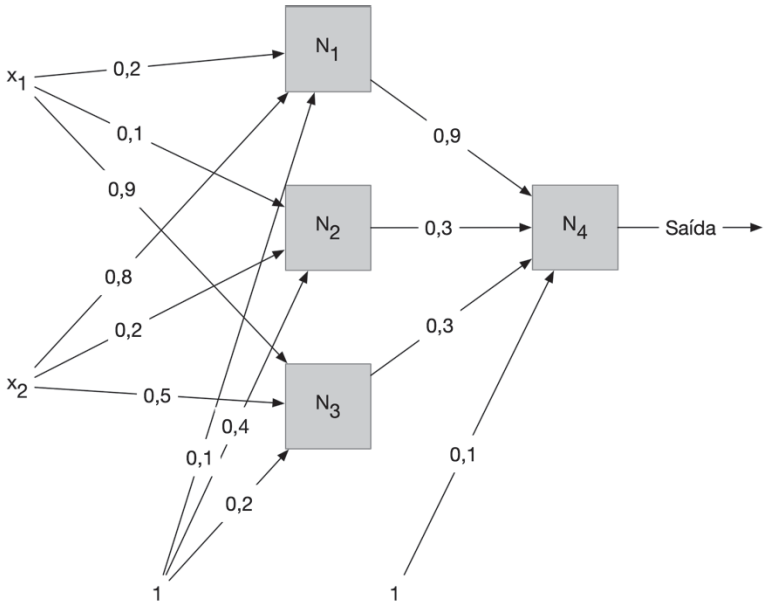
Silogismo Hipotético: $\alpha \rightarrow \beta, \beta \rightarrow \gamma \vdash \alpha \rightarrow \gamma$

Conjunção: $\alpha, \beta \vdash \alpha \wedge \beta$
Equivalência Bicondicional: $(\alpha \leftrightarrow \beta)$ equivale a $(\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha)$

a) **(25 pontos)** Deve-se mostrar todos os passos e regras aplicadas, **no mesmo formato apresentado nas aulas e nas práticas**. As equivalências e regras necessárias estão descritas acima e no material.

4 Redes Neurais Artificiais

Seja a RNA da figura abaixo.



Os neurônios N_1, N_2 e N_3 possuem função de ativação linear. Já N_4 possui função de ativação tangente hiperbólica (pesquise a fórmula e aplique).

Dada a entrada x_1=-3, x_2=1, dê:

a) **(6,25 pontos)** Valor de saída do neurônio N1
b) **(6,25 pontos)** Valor de saída do neurônio N2
c) **(6,25 pontos)** Valor de saída do neurônio N3
d) **(6,25 pontos)** Valor de saída da rede como um todo

B – RESOLUÇÃO

Questão a: CHATGPT

Inteligência Artificial (IA), segundo o ChatGPT, refere-se à capacidade de uma máquina ou programa de computador de realizar tarefas que normalmente exigiria inteligência humana. Essas tarefas incluem aprendizado, raciocínio, percepção, reconhecimento de padrões, compreensão da linguagem natural e

tomada de decisões. O objetivo da Inteligência Artificial é criar sistemas que possam executar funções complexas de forma autônoma, imitando ou replicando a inteligência humana.

Existem dois tipos principais de IA: IA fraca (ou estreita) e IA forte (ou geral). A seguir, os dois tipos.

a) **IA fraca (estreita):** Refere-se a sistemas especializados em realizar tarefas específicas sem possuir a capacidade de generalizar para diferentes domínios. Exemplos incluem assistentes virtuais, reconhecimento de voz, carros autônomos e sistemas de recomendação;

b) **IA forte (geral):** Representa um nível mais avançado de inteligência, comparável à inteligência humana em termos de habilidades cognitivas e adaptação a diferentes contextos. Até o momento, a IA forte permanece mais uma aspiração do que uma realidade prática.

As técnicas comuns em inteligência artificial incluem aprendizado de máquina (*machine learning*), redes neurais, processamento de linguagem natural, visão computacional e algoritmos de otimização. O campo da inteligência artificial continua a evoluir rapidamente com aplicações em diversas áreas como saúde, finanças, transporte, manufatura e entretenimento.

Questão b

A resposta fornecida acima pelo ChatGPT, de acordo com as quatro abordagens clássicas da IA, é categorizável como uma abordagem “agir como seres humanos”, pois o texto da resposta ressaltou a imitação de habilidades inerentemente humanas ao tentar definir o conceito de inteligência artificial.

Questão c: COMO O CHATGPT FUNCIONA

ChatGPT Sigla para Generative Pré-Trained Transformer (Transformador pré-treinado generativo, em tradução livre) é um sistema ou um modelo de linguagem baseado em inteligência artificial mais especificamente com as técnicas

de aprendizado de máquinas, redes neurais, aprendizado por reforço com feedback humano (sigla em inglês RHLF) e processamento de linguagem natural usados com o foco em diálogos virtuais.

A seguir vamos falar sobre cada uma das partes do algoritmo GPT. O ChatGPT usa como fonte de dados para manter os seus diálogos virtuais e o contexto das mesmas um conjunto de informações disponíveis e acessíveis principalmente através da internet.

O generativo pré-treinado, o **GP de GPT**, é uma técnica onde ChatGPT recebe uma grande quantidade de regras de linguagens e dados não rotulados. O GPT é então deixado sem supervisão para aprender sozinho sobre as regras e os relacionamentos que governam as linguagens.

A letra **T de GPT**, *Transformer Architecture*, refere-se à arquitetura do modelo, a qual é baseada na arquitetura Transformer. Agora, todos os dados pré-treinados são usados para criar uma rede neural de aprendizagem profunda, permitindo ao ChatGPT aprender padrões e relacionamentos de textos, dando ao ChatGPT a habilidade de criar respostas como humanos e de predizer qual texto vem em seguida em uma dada sentença.

O Transformer é capaz de ler cada palavra de uma sentença e comparar cada palavra com as outras palavras da mesma sentença. Isso permite que ele direcione a sua atenção à palavra mais importante da sentença. Esse processo é conhecido como self-attention. Com isso, é importante ressaltar que o ChatGPT não entende o que é dito e o que ele responde.

No início, a rede neural do ChatGPT não estava inteiramente segura para ser lançada ao público, pois ela foi treinada sumariamente através de dados da internet aberta sem nenhuma orientação.

Para que o ChatGPT pudesse dar respostas coerentes, sensatas e seguras ao público, foi otimizado com uma técnica chamada aprendizado por reforço com feedback humano (sigla em inglês RHLF). Basicamente são demonstrados dados que guiam a rede neural em como responder adequadamente às solicitações humanas. O RHLF, mesmo não sendo uma aprendizado supervisionado puro, permite ao GPT ser efetivamente fine-tuned.

Outra técnica igualmente utilizada é o natural language processing (NLP), que é o processo de fazer um inteligência artificial a entender as regras e a sintaxe de uma linguagem.

Referências

https://zapier.com/blog/how-does-chatgpt-work/
https://investnews.com.br/guias/chatgpt/

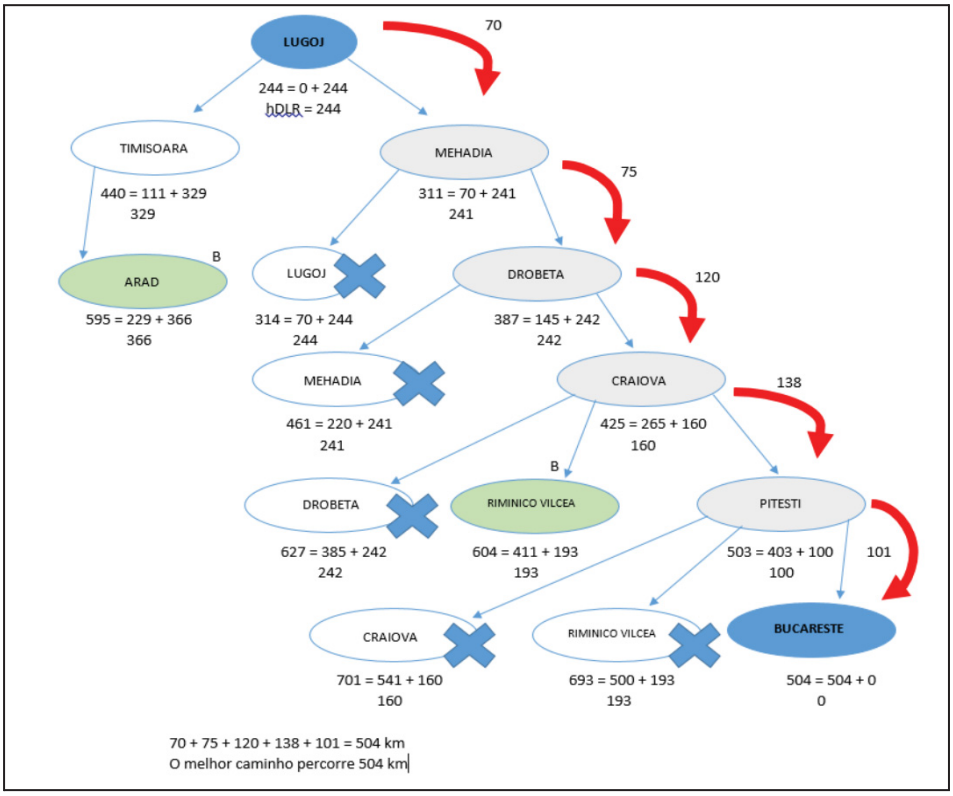
Questão d: ABORDAGEM DO CHATGPT

O ChatGPT é classificável com a abordagem clássica **Pensar racionalmente**, pois fornece respostas com base em padrões aprendidos a partir de uma vasta gama de dados. Ele não possui emoções, experiências pessoais ou compreensão profunda das nuances humanas. Ele não entende o que é perguntado e o que ele responde.

Questão 2: BUSCA HEURÍSTICA

A seguinte imagem demonstra o melhor caminho encontrado para chegar a cidade de Bucareste partindo da cidade de Lugoj, por meio da busca heurística **A***.

FIGURA 1 - BUSCA HEURÍSTICA DE LUGOJ ATÉ BUCARESTE



Fonte: O autor (2024)

Questão 3: LÓGICA

Verificação da validade do argumento através da lógica proposicional. Primeiramente determinar a base de conhecimento (BC) convertendo o argumento aristotélico para um argumento proposicional

.

R1: p → q

R2: q → r

R3: ¬r ∨ p

BC:

R1: p → q

R2: q → r

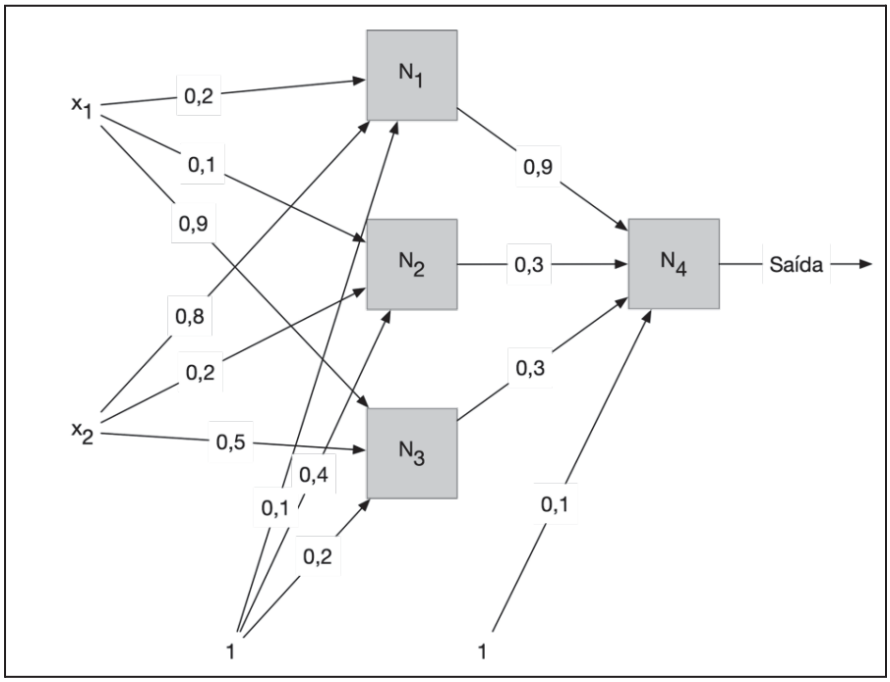
R3: ¬r ∨ p

R4: r → p		Equivalência da implicação em
R5: q → p	R3	
R6: (q → p) ^ (p → q)		Silogismo hipotético em R2 e R4
R7: q ↔ p		Conjunção em R5 e R1
		Equivalência bicondicional em R6

Questão 4: REDES NEURAIS ARTIFICIAIS

Segue abaixo a Figura 2 que contém a representação de uma simples rede neural.

FIGURA 2 - REPRESENTAÇÃO DE UMA SIMPLES RNA



Fonte: Trabalho de introdução à IA (2024)

A seguir o valor de saída de cada um dos neurônios representados na figura acima.

- **Questão a:** Valor de saída do neurônio **N1 = 0,3, função linear - onde fa(u) = -3*0,2+1*0,8+1*0,1 = 0,3;**
- **Questão b:** Valor de saída do neurônio **N2 = 0.3, função linear - onde fa(u) = -3*0,1+1*0,2+1*0,4 = 0,3;**
- **Questão c:** Valor de saída do neurônio **N3 = -2, função linear - onde fa(u) = -3*0,9+1*0,5+1*0,2 = -2;**
- **Questão d:** Valor de saída da rede como um todo: **Saída = -0,139, função hiperbólica - onde fh(u) = 0,3*0,9+0,3*0,3-2*0,3+1*0,1 = -0,14.** Função pela linguagem R **TangH= (EXP(-0,14)-EXP(-(-0,14)))/(EXP(-0,14)+EXP(-(-0,14))) = -0,139.**

APÊNDICE 2 – LINGUAGEM DE PROGRAMAÇÃO APLICADA

A – ENUNCIADO

Nome da base de dados do exercício: *precos_carros_brasil.csv*

Informações sobre a base de dados:

Dados dos preços médios dos carros brasileiros, das mais diversas marcas, no ano de 2021, de acordo com dados extraídos da tabela FIPE (Fundação Instituto de Pesquisas Econômicas). A base original foi extraída do site Kaggle (Acesse aqui a base original). A mesma foi adaptada para ser utilizada no presente exercício.

Observação: As variáveis *fuel*, *gear* e *engine_size* foram extraídas dos valores da coluna *model*, pois na base de dados original não há coluna dedicada a esses valores. Como alguns valores do modelo não contêm as informações do tamanho do motor, este conjunto de dados não contém todos os dados originais da tabela FIPE.

Metadados:

Nome do campo	Descrição
year_of_reference	O preço médio corresponde a um mês de ano de referência
month_of_reference	O preço médio corresponde a um mês de referência, ou seja, a FIPE atualiza sua tabela mensalmente
fipe_code	Código único da FIPE
authentication	Código de autenticação único para consulta no site da FIPE
brand	Marca do carro

model	Modelo do carro
fuel	Tipo de combustível do carro
gear	Tipo de engrenagem do carro
engine_size	Tamanho do motor em centímetros cúbicos
year_model	Ano do modelo do carro. Pode não corresponder ao ano de fabricação
avg_price	Preço médio do carro, em reais

Atenção: ao fazer o download da base de dados, selecione o formato .csv. É o formato que será considerado correto na resolução do exercício.

1 Análise Exploratória dos dados

A partir da base de dados **precos_carros_brasil.csv**, execute as seguintes tarefas:

a. Carregue a base de dados media_precos_carros_brasil.csv
b. Verifique se há valores faltantes nos dados. Caso haja, escolha uma tratativa para resolver o problema de valores faltantes
c. Verifique se há dados duplicados nos dados
d. Crie duas categorias, para separar colunas numéricas e categóricas. Imprima o resumo de informações das variáveis numéricas e categóricas (estatística descritiva dos dados)
e. Imprima a contagem de valores por modelo (model) e marca do carro (brand)
f. Dê um breve explicação (máximo de quatro linhas) sobre os principais resultados encontrados na Análise Exploratória dos dados

2 Visualização dos dados

A partir da base de dados **precos_carros_brasil.csv,** execute as seguintes tarefas:

a. Gere um gráfico da distribuição da quantidade de carros por marca
b. Gere um gráfico da distribuição da quantidade de carros por tipo de engrenagem do carro

c. Gere um gráfico da evolução da média de preço dos carros ao longo dos meses de 2022 (variável de tempo no eixo X)
d. Gere um gráfico da distribuição da média de preço dos carros por marca e tipo de engrenagem
e. Dê uma breve explicação (máximo de quatro linhas) sobre os resultados gerados no item d
f. Gere um gráfico da distribuição da média de preço dos carros por marca e tipo de combustível
g. Dê uma breve explicação (máximo de quatro linhas) sobre os resultados gerados no item f

3 Aplicação de modelos de machine learning para prever o preço médio dos carros

A partir da base de dados **precos_carros_brasil.csv**, execute as seguintes tarefas:

a. Escolha as variáveis **numéricas** (modelos de Regressão) para serem as variáveis independentes do modelo.A variável target é **avg_price. Observação:** caso julgue necessário, faça a transformação de variáveis categóricas em variáveis numéricas para inputar no modelo. Indique **quais variáveis** foram transformadas e **como** foram transformadas
b. Crie partições contendo 75% dos dados para treino e 25% para teste
c. Treine modelos RandomForest (biblioteca RandomForestRegressor) e XGBoost (biblioteca XGBRegressor) para predição dos preços dos carros. **Observação**: caso julgue necessário, mude os parâmetros dos modelos e rode novos modelos. Indique quais parâmetros foram inputados e indique o treinamento de cada modelo
d. Grave os valores preditos em variáveis criadas
e. Realize a análise de importância das variáveis para estimar a variável target, **para cada modelo treinado**
f. Dê uma breve explicação (máximo de quatro linhas) sobre os resultados encontrados na análise de importância de variáveis
g. Escolha o melhor modelo com base nas métricas de avaliação MSE, MAE e R^2
h. Dê uma breve explicação (máximo de quatro linhas) sobre qual modelo gerou o melhor resultado e a métrica de avaliação utilizada

B - RESOLUÇÃO

1.f. Dê um breve explicação (máximo de quatro linhas) sobre os principais resultados encontrados na Análise Exploratória dos dados

No geral, o preço médio de um carro no Brasil custa aproximadamente R$ 52,756,91. Um dos modelos mais vendidos é o Palio Week. Adv/Adv TRYON 1.8 mpi Flex, com câmbio manual e motor 1.6. O menor preço médio de carros foi de R$ 6647,00, e o maior preço médio é R$ 979358,00.

2.e. Dê uma breve explicação (máximo de quatro linhas) sobre os resultados gerados no item d

O preço médio do carro com câmbio automático é relativamente mais alto quando comparado aos de câmbio manual para todas as marcas, com exceção da Renault. Volkswagen e Fiat lideram as marcas que possuem maior preço médio para carros de câmbio automático, ao passo que lideram também as marcas com menor preço médio dos carros com câmbio manual.

2.g. Dê uma breve explicação (máximo de quatro linhas) sobre os resultados gerados no item f:

1. O preço médio do carro à diesel é muito mais elevado para todas as marcas, sendo a VolksWagen a marca com maior preço médio para esta categoria;
2. O preço médio do carro à álcool lidera com os menores valores;
3. Os carros à gasolina com menor preço médio são das marcas Fiat e Renault.

3.f. Dê uma breve explicação (máximo de quatro linhas) sobre os resultados encontrados na análise de importância de variáveis

As variáveis que tiveram maior relevância na medida da performance dos modelos treinados foram o tamanho do motor (engine_size) e ano do modelo (year_model) respectivamente. O mês de referência foi o que obteve menor relevância em todos os modelos treinados.

3.g. Escolha o melhor modelo com base nas métricas de avaliação MSE, MAE e R2

O melhor modelo com base nessas medidas de acurácia foi o modelo RandomForest sem parâmetros.

3.h. Dê uma breve explicação (máximo de quatro linhas) sobre qual modelo gerou o melhor resultado e a métrica de avaliação utilizada

O modelo que apresentou os melhores resultado foi o RandomForest sem parâmetros para medir a sua acurácia foi utilizado as medidas de acurácia MSE

(mean squared error) com o valor 12165425.33231638, MAE (mean absolute error) com o valor 12165425.33231638 e R2_score com 0.9954656414129976 de precisão.

APÊNDICE 3 – LINGUAGEM R

A – ENUNCIADO

1 Pesquisa com Dados de Satélite (Satellite)

O banco de dados consiste nos valores multiespectrais de pixels em vizinhanças 3x3 em uma imagem de satélite, e na classificação associada ao pixel central em cada vizinhança. O objetivo é prever esta classificação, dados os valores multiespectrais.

Um quadro de imagens do Satélite Landsat com MSS (*Multispectral Scanner System*) consiste em quatro imagens digitais da mesma cena em diferentes bandas espectrais. Duas delas estão na região visível (correspondendo aproximadamente às regiões verde e vermelha do espectro visível) e duas no infravermelho (próximo). Cada pixel é uma palavra binária de 8 bits, com 0 correspondendo a preto e 255 a branco. A resolução espacial de um pixel é de cerca de 80m x 80m. Cada imagem contém 2340 x 3380 desses pixels. O banco de dados é uma subárea (minúscula) de uma cena, consistindo de 82 x 100 pixels. Cada linha de dados corresponde a uma vizinhança quadrada de pixels 3x3 completamente contida dentro da subárea 82x100. Cada linha contém os valores de pixel nas quatro bandas espectrais (convertidas em ASCII) de cada um dos 9 pixels na vizinhança de 3x3 e um número indicando o rótulo de classificação do pixel central.

As classes são: solo vermelho, colheita de algodão, solo cinza, solo cinza úmido, restolho de vegetação, solo cinza muito úmido.

Os dados estão em ordem aleatória e certas linhas de dados foram removidas, portanto você não pode reconstruir a imagem original desse conjunto de dados. Em cada linha de dados, os quatro valores espectrais para o pixel superior esquerdo são dados primeiro, seguidos pelos quatro valores espectrais para o pixel superior central e, em seguida, para o pixel superior direito, e assim por diante, com os pixels lidos em sequência, da esquerda para a direita e de cima para baixo. Assim, os quatro valores espectrais para o pixel central são dados pelos atributos 17, 18, 19 e 20. Se você quiser, pode usar apenas esses quatro atributos, ignorando os outros. Isso evita o problema que surge quando uma vizinhança 3x3 atravessa um limite.

O banco de dados se encontra no pacote **mlbench** e é completo (não possui dados faltantes).

Tarefas:

1. Carregue a base de dados Satellite
2. Crie partições contendo 80% para treino e 20% para teste
3. Treine modelos RandomForest, SVM e RNA para predição destes dados.
4. Escolha o melhor modelo com base em suas matrizes de confusão.
5. Indique qual modelo dá o melhor o resultado e a métrica utilizada

2 Estimativa de Volumes de Árvores

Modelos de aprendizado de máquina são bastante usados na área da engenharia florestal (mensuração florestal) para, por exemplo, estimar o volume de madeira de árvores sem ser necessário abatê-las.

O processo é feito pela coleta de dados (dados observados) através do abate de algumas árvores, onde sua altura, diâmetro na altura do peito (dap), etc, são medidos de forma exata. Com estes dados, treina-se um modelo de AM que pode estimar o volume de outras árvores da população.

Os modelos, chamados alométricos, são usados na área há muitos anos e são baseados em regressão (linear ou não) para encontrar uma equação que descreve os dados. Por exemplo, o modelo de Spurr é dado por:

$$\textbf{Volume} = b0 + b1 * dap^2 * Ht$$

Onde dap é o diâmetro na altura do peito (1,3metros), Ht é a altura total. Tem-se vários modelos alométricos, cada um com uma determinada característica, parâmetros, etc. Um modelo de regressão envolve aplicar os dados observados e encontrar b0 e b1 no modelo apresentado, gerando assim uma equação que pode ser usada para prever o volume de outras árvores.

Dado o arquivo **Volumes.csv**, que contém os dados de observação, escolha um modelo de aprendizado de máquina com a melhor estimativa, a partir da estatística de correlação.

Tarefas

1. Carregar o arquivo Volumes.csv (http://www.razer.net.br/datasets/Volumes.csv)
2. Eliminar a coluna NR, que só apresenta um número sequencial
3. Criar partição de dados: treinamento 80%, teste 20%
4. Usando o pacote "caret", treinar os modelos: Random Forest (rf), SVM (svmRadial), Redes Neurais (neuralnet) e o modelo alométrico de SPURR

    - O modelo alométrico é dado por: Volume = b0 + b1 * dap^2 * Ht

      alom <- nls(VOL ~ b0 + b1*DAP*DAP*HT, dados, start=list(b0=0.5, b1=0.5))

5. Efetue as predições nos dados de teste
6. Crie suas próprias funções (UDF) e calcule as seguintes métricas entre a predição e os dados observados

    - Coeficiente de determinação: R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

onde y_i é o valor observado, $\hat{y}_i$ é o valor predito e $\bar{y}$ é a média dos valores y_i observados. Quanto mais perto de 1 melhor é o modelo;

- Erro padrão da estimativa: S_{yx}

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

esta métrica indica erro, portanto quanto mais perto de 0 melhor é o modelo;

- Syx%

$$S_{yx}\% = \frac{S_{yx}}{\bar{y}} * 100$$

esta métrica indica porcentagem de erro, portanto quanto mais perto de 0 melhor é o modelo;

7. Escolha o melhor modelo.

B – RESOLUÇÃO

Questão 1 - Pesquisa com Dados de Satélite (Satellite).

A seguir a matriz de confusão e outras métricas de performance que o modelo Random Forest obteve.

FIGURA 3 - MATRIZ DE CONFUSÃO PARA O MODELO RANDOM FOREST

```
## Confusion Matrix and Statistics
##
##                      Reference
## Prediction            red soil cotton crop grey soil damp grey soil
##   red soil                 300           1         3              1
##   cotton crop                0         137         1              0
##   grey soil                  3           0       263             25
##   damp grey soil             0           0         2             78
##   vegetation stubble         3           0         1              0
##   very damp grey soil        0           2         1             21
##                      Reference
## Prediction            vegetation stubble very damp grey soil
##   red soil                             4                   0
##   cotton crop                          1                   1
##   grey soil                            0                   3
##   damp grey soil                       1                  20
##   vegetation stubble                 128                   4
##   very damp grey soil                  7                 273
##
## Overall Statistics
##
##                Accuracy : 0.9182
##                  95% CI : (0.9019, 0.9326)
##     No Information Rate : 0.2383
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8987
##
##  Mcnemar's Test P-Value : NA
```

FONTE: O autor (2024)

FIGURA 4 - MATRIZ DE CONFUSÃO PARA O MODELO SVM

```
## Confusion Matrix and Statistics
##
##                      Reference
## Prediction            red soil cotton crop grey soil damp grey soil
##   red soil                 303           0         2              0
##   cotton crop                0         138         2              2
##   grey soil                  2           0       261             27
##   damp grey soil             0           1         5             74
##   vegetation stubble         1           0         0              1
##   very damp grey soil        0           1         1             21
##                      Reference
## Prediction            vegetation stubble very damp grey soil
##   red soil                              5                   0
##   cotton crop                           2                   2
##   grey soil                             0                   7
##   damp grey soil                        1                  21
##   vegetation stubble                  126                   3
##   very damp grey soil                   7                 268
##
## Overall Statistics
##
##                Accuracy : 0.9112
##                  95% CI : (0.8943, 0.9262)
##     No Information Rate : 0.2383
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8901
##
##  Mcnemar's Test P-Value : NA
```

FONTE: Autoria própria (2024)

FIGURA 5 - MATRIZ DE CONFUSÃO DO MODELO RNA

```
## Confusion Matrix and Statistics
##
##                      Reference
## Prediction            red soil cotton crop grey soil damp grey soil
##   red soil                 289          132         3              1
##   cotton crop                6            0         0              0
##   grey soil                  3            5       244            104
##   damp grey soil             0            0         0              0
##   vegetation stubble         7            3        12             13
##   very damp grey soil        1            0        12              7
##                      Reference
## Prediction            vegetation stubble very damp grey soil
##   red soil                           31                    1
##   cotton crop                         1                    0
##   grey soil                          13                  267
##   damp grey soil                      0                    0
##   vegetation stubble                 91                   15
##   very damp grey soil                 5                   18
##
## Overall Statistics
##
##                Accuracy : 0.5
##                  95% CI : (0.4723, 0.5277)
##     No Information Rate : 0.2383
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3672
```

FONTE: O autor (2024)

4. Escolha o melhor modelo com base em suas matrizes de confusão.

Random Forest

Os valores de kappa e acurácia do modelo *Random Forest* foram respectivamente os seguintes:

a) **acurácia:** 0.91;

b) **kappa:** 0.89

De acordo com os valores das medidas de acurácia e kappa, o modelo Random Forest teve um bom desempenho em realizar predições sobre dados desconhecidos.

SVM

Os valores de kappa e acurácia do modelo SVM foram respectivamente os seguintes:

a) **acurácia:** 0.91
b) **kappa:** 0.89

De acordo com os valores das medidas de acurácia e kappa, o modelo SVM também teve um bom desempenho em realizar predições sobre dados desconhecidos.

RNA

Os valores de kappa e acurácia do modelo RNA foram respectivamente os seguintes:

a) **acurácia:** 0.5
b) **kappa:** 0.36

De acordo com os valores das medidas de acurácia e kappa, o modelo RNA não teve um bom desempenho em realizar predições sobre dados desconhecidos.

5. Indique qual modelo dá o melhor o resultado e a métrica utilizada.

O melhor modelo foi random forest com acurácia de 0.918 e kappa de 0.8987. A métrica utilizada foram a acurácia e a kappa.

6. Resumo dos resultados

QUADRO 1 - DESEMPENHOS DOS MODELOS OBTIDOS

Modelo	Coeficiente	Erro padrão	Erro padrão em porcentagem

Random Forest	0.82	0.13	10.42
SVM	0.62	0.19	15.13
NNET	-1.06	0.46	35.57
SPURR	0.77	0.15	11.77

FONTE: O autor (2024)

7. Escolha o melhor modelo.

Com base nos resultados das métricas, o modelo que se saiu melhor foi o Random Forest, com R2 igual 0.8223603, Erro padrão estimado de 0.1376052 e Erro padrão de estimativa em porcentagem de 10.42195.

APÊNDICE 4 – ESTATÍSTICA APLICADA I

A – ENUNCIADO

1) Gráficos e tabelas

(15 pontos) Elaborar os gráficos box-plot e histograma das variáveis “age” (idade da esposa) e “husage” (idade do marido) e comparar os resultados

(15 pontos) Elaborar a tabela de frequencias das variáveis “age” (idade da esposa) e “husage” (idade do marido) e comparar os resultados

2) Medidas de posição e dispersão

(15 pontos) Calcular a média, mediana e moda das variáveis “age” (idade da esposa) e “husage” (idade do marido) e comparar os resultados

(15 pontos) Calcular a variância, desvio padrão e coeficiente de variação das variáveis “age” (idade da esposa) e “husage” (idade do marido) e comparar os resultados

3) Testes paramétricos ou não paramétricos

(40 pontos) Testar se as médias (se você escolher o teste paramétrico) ou as medianas (se você escolher o teste não paramétrico) das variáveis “age” (idade da esposa) e “husage” (idade do marido) são iguais, construir os intervalos de confiança e comparar os resultados.

Obs:

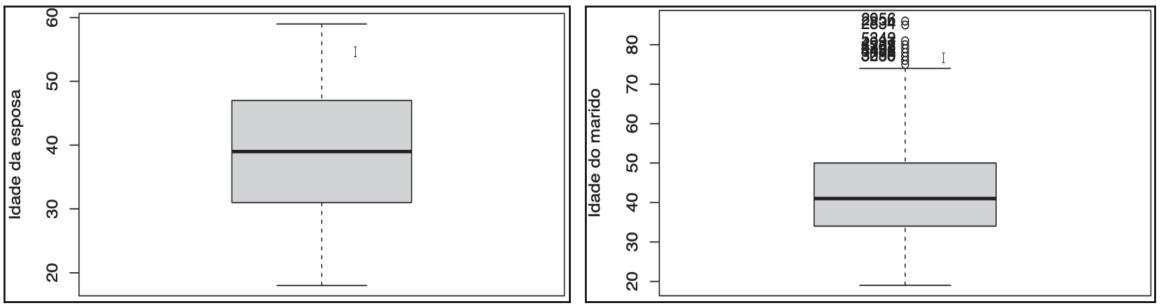
Você deve fazer os testes necessários (e mostra-los no documento pdf) para saber se você deve usar o unpaired test (paramétrico) ou o teste U de Mann-Whitney (não paramétrico), justifique sua resposta sobre a escolha.

Lembre-se de que os intervalos de confiança já são mostrados nos resultados dos testes citados no item 1 acima.

B – RESOLUÇÃO

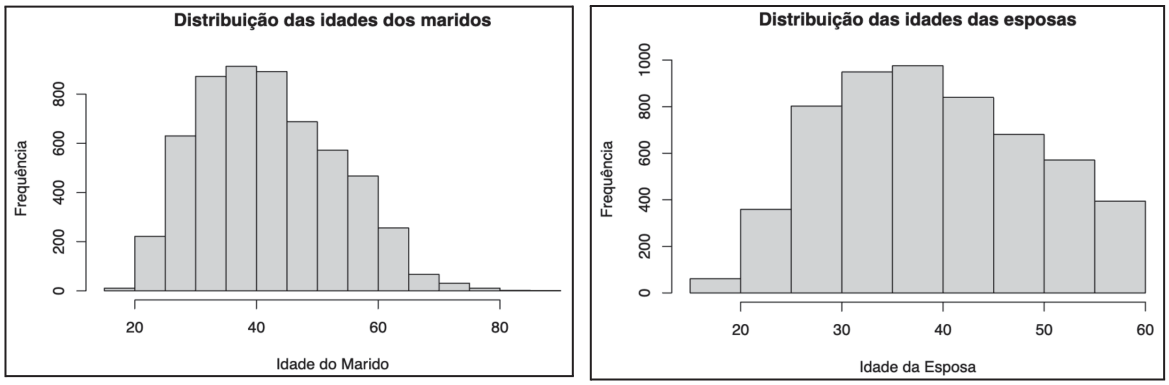
Questão 1 - Gráficos e tabelas

GRÁFICO 2 - BOX PLOTS DAS IDADES DOS MARIDOS E ESPOSAS



FONTE: O autor (2024)

GRÁFICO 3 - HISTOGRAMAS DAS IDADES DOS MARIDOS E DAS ESPOSAS



FONTE: O autor (2024)

A média de idade dos maridos foi de 42 anos e o das esposas de 39 anos. A média das idades dos maridos foi 7.68% maior que a média de idade das esposas.

A moda da idade dos maridos foi de 44 anos. A moda da idade das esposas foi 37 anos. A moda da idade dos maridos foi 18.92% maior que o das esposas.

A mediana das idades dos maridos é de 39 anos enquanto o das esposas é de 41 anos. A idade média dos maridos é 7.68% maior que a média de idade das esposas.

Todos esses resultados foram extraídos de uma amostra contendo 201 maridos e 207 esposas.

De acordo com os gráficos de box plots e histogramas e os resultados estatísticos, a idade dos maridos possui muitos outliers, indicando maridos mais velhos do que o esperado pela distribuição das idades.

Portanto, os maridos são um pouco mais velhos que as esposas na amostra, com frequências de idades mais concentradas entre 38 e 43 anos, chegando inclusive até os 80 anos com baixas frequências. Já as esposas são mais jovens com idades mais concentradas entre 35 e 38 anos e um máximo de 60 anos.

2. Medidas de dispersão e posição

a) **média das idades das esposas:** 39.43 anos;
b) **média das idades dos maridos:** 42.46.

Portanto, a média das idades dos maridos é 7.68% maior que a média das idades das esposas. Sendo que a amostra contém 201 maridos e 217 esposas.

a) **mediana das idade das esposas:** 39 anos;
b) **mediana da idade dos maridos:** 41 anos.

Portanto, a mediana das idades dos maridos é 5.13% maior que a mediana das idades das esposas. Sendo que a amostra contém 201 maridos e 217 esposas.

a) **moda das idade das esposas:** 37 anos;
b) **moda da idade dos maridos:** 44 anos.

A moda das idades dos maridos é de 44 anos, com 201 pessoas no total. A moda das idades das esposas é de 37 anos, com 217 pessoas no total. Portanto, a moda das idades dos maridos é maior que a moda das idades das esposas em 18,92 %.

3. Teste de hipóteses

Primeiramente, verificar se as amostras são independentes, se há normalidade dos dados e homogeneidade das variâncias entre grupos.

a) **premissa 1:** As duas amostras são independentes ? Sim, pois os grupos de maridos e esposas não estão relacionados. Não se trata de uma amostra ou grupos emparelhados;

b) **premissa 2:** Os dados de cada amostra ou grupo possuem distribuição normal? Para determinar a resposta foi usado os testes de normalidade de Kolmogorov-smirnov, por conta do tamanho da amostra, com os seguintes testes de hipóteses:

   - **hipótese nula (H0):** os dados são normalmente distribuídos;
   - **hipótese alternativa (Ha):** os dados não são normalmente distribuídos.

c) **premissa 3:** As duas amostras ou grupos possuem homogeneidade das variâncias ? Para determinar a resposta, os testes de hipóteses são:

   - **hipótese nula (H0):** As variâncias são estatisticamente iguais (homogêneas);
   - **hipótese alternativa (Ha):** As variâncias não são estatisticamente iguais (homogênea).

A seguir, o teste de normalidade Kolmogorov-Smirnov para as idades dos maridos:

FIGURA 6 - KOLMOGOROV-SMIRNOV PARA AS IDADES DOS MARIDOS

```
Exact one-sample Kolmogorov-Smirnov

data: unique(ages)
D = 0.31783, p-value = 0.000002446
alternative hypothesis: two-sided
```

FONTE: O autor (2024)

Como o p-value é menor do que 0.5, então o grupo das idades dos maridos não possui distribuição normal, rejeitando H0, hipótese nula, da premissa 2. Realizar o mesmo teste de normalidade para as idades das esposas.

FIGURA 7 - KOLMOGOROV-SMIRNOV PARA AS IDADES DAS ESPOSAS

```
Exact one-sample Kolmogorov-Smirnov test

data: unique(ages)
D = 0.13748, p-value = 0.3713
alternative hypothesis: two-sided
```

FONTE: O autor (2024)

O p-value é maior que 0.05 (p-value = 0.3713), logo o grupo idades das esposas possui distribuição normal.

Como o grupo das idades dos maridos não possui uma distribuição normal, foi verificado, a título de curiosidade, a premissa 3, ou seja, se os grupos possuem homogeneidade nas suas variâncias. Foi usado o teste F para verificar a homogeneidade das variâncias:

FIGURA 8 - RESULTADO DO TESTE F DE FISHER

```
F test to compare two variances

data: ages by group
F = 1.2638, num df = 5633, denom df = 5633, p-value < 0.00000000000000022
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.199526 1.331617
sample estimates:
ratio of variances
         1.263847
```

FONTE: O autor (2024)

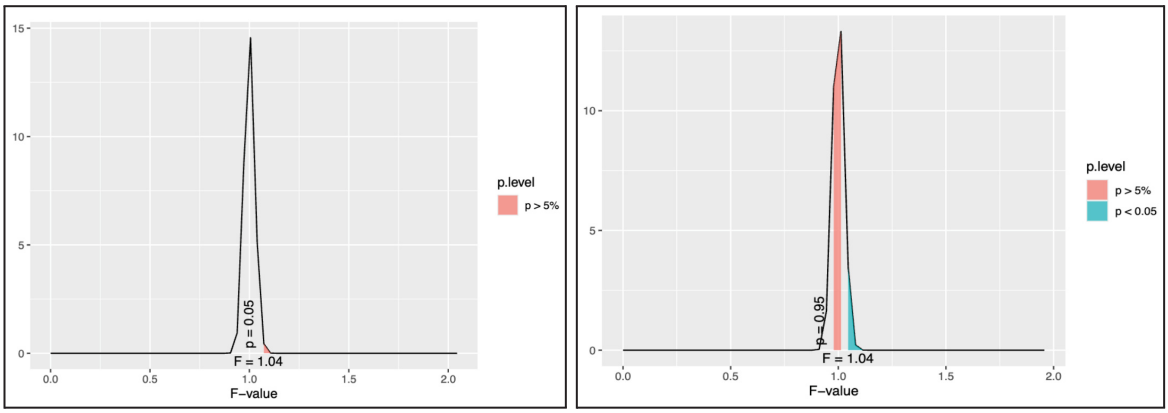
O valor obtido da estatística F de Fisher foi 1,04481. Para analisar a outra cauda da distribuição F, o recíproco deste valor é calculado a seguir:

$$\frac{1}{1{,}04481} \approx 0,95711$$

Essa inversão é necessária porque a distribuição F não é simétrica, como a normal. A F de Fisher depende dos graus de liberdade do numerador e do denominador, e sua forma é simétrica, ou seja, os valores críticos da cauda superior e inferior não são espelhados.

Os dois código a seguir vão construir e exibir os gráficos de distribuições F de Fisher, com os graus de liberdade 5633 no numerador e denominador, e marca a estatística F = 1.04481, incluindo as duas caudas (superior e inferior).

GRÁFICO 4 - DISTRIBUIÇÃO F DE FISHER CAUDA INFERIOR E CAUSA SUPERIOR



FONTE: Autoria própria (2024)

O teste F de Fisher obteve valor crítico entre 0.9571118 e 1.04481 (região de não rejeição de H0, hipótese nula), os valore acima de 1.04481 e abaixo de 0.9571118 estão na região chamada de rejeição de H0 (área em azul do gráfico X). O valor da estatística F de Fisher obtida foi de 1.2638. Como esse valor se encontra na região de rejeição de H0, então a afirmação da hipótese nula da premissa 3 de que as variâncias são estatisticamente iguais não é verdadeira.

Sabendo que os grupos de idades de ambos esposas e maridos não estão normalmente distribuídos e nem possuem variâncias homogêneas, o próximo passo foi determinar se a mediana das idades dos maridos e esposas são iguais. A nova premissa, portanto, é:

a) **hipótese nula (H0):** A mediana das idades dos maridos e esposas são iguais;

b) **hipótese alternativa (Ha):** Mediana das idades dos maridos e esposas são diferentes.

Uma vez que os grupos das idades dos maridos e das esposas não são normalmente distribuídos e nem possuem variâncias homogêneas, o teste U de Mann-Whitney, foi usado para verificar a premissa acima.

Teste U de Mann-Whitney

O que precisa ser testado: se a idade mediana das esposas difere da idade mediana dos maridos.

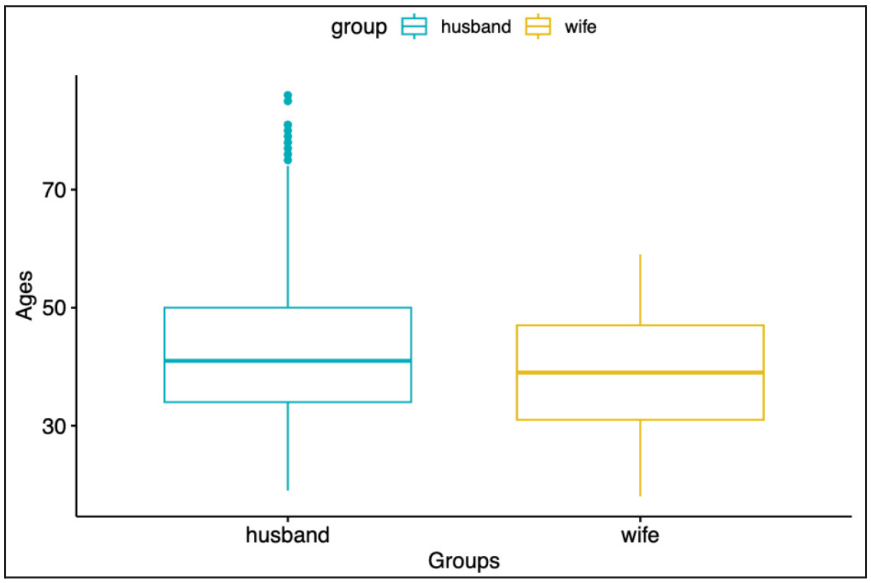
FIGURA 9 - SUMÁRIO ESTATÍSTICO DAS IDADES DE AMBOS, MARIDOS E ESPOSAS

```
A tibble: 2 x 4
  group   count  median  IQR
  <chr>   <int>  <dbl> <dbl>
1 husband 5634    41    16
2 wife    5634    39    16
```

FONTE: O autor (2024)

Visualizar os mesmos dados usando o gráfico de box-plots, plotando a coluna ages pela coluna group.

GRÁFICO 5 - *BOX PLOT* DAS IDADES DE MARIDOS E ESPOSAS



FONTE: O autor (2024)

Em seguida, a realização do teste U de Mann-Whitney propriamente dito, com o objetivo de atestar H0, a hipótese nula, da premissa estabelecida, isto é, se a mediana das idades das esposas é igual a mediana das idades dos maridos, sendo a alternativa a negação da primeira.

O teste é sempre feito com relação à disposição do vetor de dados, sempre do último para o primeiro. No caso, vetor husband contra wife.

FIGURA 10 - RESULTADO DO TESTE U DE MANN-WHITNEY

```
Wilcoxon rank sum test with continuity correction

data: ages by group
W = 18122044, p-value < 0.00000000000000022
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  2.000033 3.000024
sample estimates:
difference in location 2.999966
```

FONTE: O autor (2024)

O resultado do p-value obtido no teste foi menor do que 0.00000000000000022, que foi menor que o nível de sindicância de 0.05. Portanto, a mediana das idades dos maridos é estatisticamente diferente da mediana das idades das esposas (hipótese nula rejeitada). O intervalo de confiança da diferença entre as medianas ficou entre 2.000033 e 3.000024, com uma mediana de 2.999966.

Por fim, sendo as medianas diferentes, para determinar se a mediana das idades dos maridos é menor ou maior que a mediana das idades das esposas, as seguintes hipóteses foram estabelecidas e posteriormente testadas.

a) **hipótese nula (H0):** A mediana das idades dos maridos não é estatisticamente maior que a mediana das idades das esposas;

b) **hipótese alternativa (Ha):** A mediana das idades dos maridos é estatisticamente maior que a mediana das idades das esposas.

Para testar as hipótese foi utilizado o teste de wilcoxon, por meio da seguinte função em R:

FIGURA 11 - RESULTADO DO TESTE DE WILCOXON

```
Wilcoxon rank sum test with continuity correction

data: ages by group
W = 18122044, p-value < 0.00000000000000022
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 2.000046  Inf
sample estimates:
difference in location 2.999966
```

FONTE: O autor (2024)

Como o p-value foi menor do que 0.05, a mediana das idades dos maridos é estatisticamente maior que a mediana das idades das esposas. Portanto, a H0 (hipótese nula) foi rejeitada.

O intervalo de confiança para a diferença entre as medianas das idade obtido é valor maior que 2.000045 com uma mediana de 2.999966.

APÊNDICE 5 – ESTATÍSTICA APLICADA II

A – ENUNCIADO

Regressões Ridge, Lasso e ElasticNet

(100 pontos) Fazer as regressões Ridge, Lasso e ElasticNet com a variável dependente "lwage" (salário-hora da esposa em logaritmo neperiano) e todas as demais variáveis da base de dados são variáveis explicativas (todas essas variáveis tentam explicar o salário-hora da esposa). No pdf você deve colocar a rotina utilizada, mostrar em uma tabela as estatísticas dos modelos (RMSE e R^2) e concluir qual o melhor modelo entre os três, e mostrar o resultado da predição com intervalos de confiança para os seguintes valores:

husage = 40 (anos – idade do marido)
husunion = 0 (marido não possui união estável)
husearns = 600 (US$ renda do marido por semana)
huseduc = 13 (anos de estudo do marido)
husblck = 1 (o marido é preto)
hushisp = 0 (o marido não é hispânico)
hushrs = 40 (horas semanais de trabalho do marido)
kidge6 = 1 (possui filhos maiores de 6 anos)
age = 38 (anos – idade da esposa)
black = 0 (a esposa não é preta)
educ = 13 (anos de estudo da esposa)
hispanic = 1 (a esposa é hispânica)
union = 0 (esposa não possui união estável)
exper = 18 (anos de experiência de trabalho da esposa)
kidlt6 = 1 (possui filhos menores de 6 anos)

obs: lembre-se de que a variável dependente "lwage" já está em logarítmo, portanto voçê não precisa aplicar o logaritmo nela para fazer as regressões, mas é necessário aplicar o antilog para obter o resultado da predição.

B – RESOLUÇÃO

REGRESSÃO RIDGE

A primeira técnica de regressão linear usada para predizer o salário-base da esposa, representada pelos dados de entrada fornecidos é a técnica Regressão

Ridge. O melhor valor de lambda encontrado foi de 0.02511886, os valores de lambdas testados foram entre 10^{-3} até 10^{2}, com um passo de 0.1.

O trecho de código a seguir serve para visualizar o resultado (os valores) da estimativa (coeficientes) obtida com a execução do código anterior.

FIGURA 12 - VALORES DO COEFICIENTES OBTIDOS PARA CARACTERÍSTICA

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                     s0
## age       0.027779476
## black    -0.039857345
## hispanic -0.099943165
## educ      0.113081600
## union     0.136666522
## exper    -0.003398478
## kidlt6    0.139370028
## husage    0.012459717
## husunion  0.010031392
## husearns  0.075000829
## huseduc   0.026357981
## husblck  -0.001368182
## hushisp   0.025480993
## hushrs   -0.038698088
## kidge6    0.046727535
## earns     0.705145581
```

FONTE: O autor (2024)

Usando o modelo obtido pela execução da técnica Regressão Ridge, foi realizada a predição e avaliação nos dados de treinamento.

As métricas dos resultados das predições a partir da base de treinamento, portanto, foram:

a) **RMSE:** 0.55954;

b) **R^2:** 0.6867629.

Em seguida, os valores das predições e avaliação das predições nos dados de teste:

a) **RMSE:** 0.5375671;

b) **R^2:** 0.7008188.

Comparando as métricas das predições sobre as bases de treinamento e teste, notou-se que o tamanho dos erros e os R2 das predições base de treinamento e teste foram muito parecidos, descartando a hipótese de overfitting e underfitting. Além disso, o poder explicativo do modelo obtido foi considerado mediano, cerca de 70% para ambas as bases de dados, treinamento e teste, o que é um valor razoável.

Após a obtenção do modelo, foi realizada a predição sobre os dados de entrada previamente sentenciados.

Como os valores do dataset são padronizados, os dados de entrada também foram padronizados. Logo em seguida, é realizada a predição sobre os dados de entrada definidos.

O valor predito do salário por hora em dólar, com base na entrada de dados, foi de US$2,60, aproximadamente 2 dólares e 60 centavos. O intervalo de confiança para a entrada de dados usada foi entre US$2.58 (inferior) e US$2.63 (superior).

REGRESSÃO LASSO

A segunda técnica de regressão linear usada para predizer o salário-base da esposa a partir dos dados de entrada fornecidos, foi o modelo Regressão Lasso. O objetivo desse modelo é reduzir os coeficientes não significativos a zero.

Para a Regressão Lasso, a função de perda é alterada para minimizar a complexidade do modelo, restringindo a soma dos valores absolutos dos coeficientes do modelo, também chamado de restrição L1-Norm.

A restrição L1-Norm força alguns valores dos pesos a serem zero, a fim de permitir que outros coeficientes tenham valores mais distantes de zero.

Foi atribuído 1 para o parâmetro alpha para obter os valores de lambda para técnica Regressão Lasso. O melhor valor de lambda foi de 0.005011872.

Após treinar o modelo Regressão Linear com penalização lasso, os coeficientes obtidos com o modelo foram

FIGURA 13 - COEFICIENTES DAS VARIÁVEIS EXPLANATÓRIAS

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##
##                   s0
## age       0.0213240699
## black    -0.0135502681
## educ      0.1087336474
```

```
## hispanic -0.0618895379
## union     0.1194517245
## exper     .
## kidlt6    0.1101588639
## husage    0.0054993663
## husunion  0.0007771211
## husearns  0.0692821772
## huseduc   0.0211136534
## husblck   .
## hushisp   .
## hushrs   -0.0313711355
## kidge6    0.0287371650
## earns     0.7232692738
```

FONTE: O autor (2024)

Foram realizadas predições sobre a base de dados de treinamento, usando o modelo de regressão lasso. Os resultados obtidos foram, então, usados para avaliar a performance do modelo através das métricas RMSE e R2. Que foram:

a) **RMSE:** 0.5594957;

b) **R²:** 0.6868125.

Em seguida, as predições realizadas sobre a base de dados de teste, seguido da avaliação dos resultados, usando igualmente as métricas RMSE e R2. Os resultados foram:

a) **RMSE:** 0.5340756;

b) **R²:** 0.7046925.

Tanto os RMSE (raiz do erro quadrático médio) quanto os R² de ambos os resultados das predições realizadas anteriormente resultaram em valores muito próximos entre si.

Obtido o modelo regressão lasso devidamente já treinado com os dados de treinamento, o modelo foi então usado para predizer o resultado da entrada de dados definida na seção Entrada de dados, ou seja, usado para determinar o salário-hora da esposa com base na entrada de dados.

O salário-hora da esposa, definida pelos dados de entrada fornecidos na seção Dados de entrada, foi de aproximadamente US$2.60. Com um intervalo de confiança entre US$2,58 e US$2,63.

REGRESSÃO *ELASTICNET*

O próximo modelo de regressão linear que foi usado para predizer o salário-hora da esposa, por meio da entrada de dados apresentada na seção Entrada de dados, foi o modelo de *Regressão elasticnet*.

O treinamento do modelo foi realizado por meio da técnica cross-validation, com *10-folders*, 5 repetições e uma busca aleatória pelos componentes das amostras de treinamento. Os melhores valores de lambda e alpha obtidos foram 0.1161441 e 0.02841894 respectivamente.

Na etapa seguinte, realizou-se o treinamento do modelo com os dados de treinamento, a aplicação do modelo sobre os dados de teste e a avaliação de seu desempenho em ambas as bases, visando analisar a capacidade preditiva e generalização do modelo.

O desempenho do modelo na etapa de treinamento é representado pelos seguintes valores de métricas resultantes.

a) **RMSE:** 0.5597914;
b) **R²:** 0.6864813.

Por sua vez, o desempenho do modelo na etapa de teste seguiu com os seguintes valores de RMSE e R².

a) **RMSE:** 0.5370545;
b) **R²:** 0.7013892.

O salário-hora da esposa previsto pelo modelo com base nas características informadas na seção Entrada de dados foi de aproximadamente US$174.49 com um intervalo de confiança de aproximadamente entre US$2.30 e US$2.35.

Conclusão

De acordo com os resultados de R2 e RMSE de cada um dos modelos usados até aqui, o modelo que obteve melhor performance nas predições foi o modelo *Regressão Lasso*.

QUADRO 2 - MÉTRICAS DAS PREDIÇÕES DE CADA MODELO NA BASE DE TREINO

Métricas	**ElasticNet**	**Ridge**	**Lasso**
RMSE	0.534	0.700	0.530
R2	0.7045	0.537	0.7046

FONTE: autoria própria (2024)

QUADRO 3 - MÉTRICAS DAS PREDIÇÕES DE CADA MODELO NA BASE DE TESTE

Métricas	**ElasticNet**	**Ridge**	**Lasso**
RMSE	0.534	0.700	0.530
R2	0.7045	0.537	0.7046

FONTE: Autoria própria (2024)

APÊNDICE 6 – ARQUITETURA DE DADOS

A – ENUNCIADO

1 Construção de Características: Identificador automático de idioma

O problema consiste em criar um modelo de reconhecimento de padrões que dado um texto de entrada, o programa consegue classificar o texto e indicar a língua em que o texto foi escrito.

Parta do exemplo (notebook produzido no Colab) que foi disponibilidade e crie as funções para calcular as diferentes características para o problema da identificação da língua do texto de entrada.

Nessa atividade é para "construir características".

Meta: a acurácia deverá ser maior ou igual a 70%.

Essa tarefa pode ser feita no Colab (Google) ou no Jupiter, em que deverá exportar o notebook e imprimir o notebook para o formato PDF. Envie no UFPR Virtual os dois arquivos.

2 Melhore uma base de dados ruim

Escolha uma base de dados pública para problemas de classificação, disponível ou com origem na UCI Machine Learning.

Use o mínimo de intervenção para rodar a SVM e obtenha a matriz de confusão dessa base.

O trabalho começa aqui, escolha as diferentes tarefas discutidas ao longo da disciplina, para melhorar essa base de dados, até que consiga efetivamente melhorar o resultado.

Considerando a acurácia para bases de dados balanceadas ou quase balanceadas, se o percentual da acurácia original estiver em até 85%, a meta será obter 5%. Para bases com mais de 90% de acurácia, a meta será obter a melhora em pelo menos 2 pontos percentuais (92% ou mais).

Nessa atividade deverá ser entregue o script aplicado (o notebook e o PDF correspondente).

B – RESOLUÇÃO

PROCESSAMENTO DE RECONHECIMENTO DE PADRÕES

O objetivo do presente trabalho é demonstrar como é o processo de construção de atributos na área da inteligência artificial e como ele é fundamental para o reconhecimento de padrões.

A linguagem de programação utilizada para construir tal algoritmo capaz de determinar o idioma de um texto de entrada foi o *python*.

Para atingir esse objetivo, primeiramente foi definido um conjunto de amostras de textos em três linguagens diferentes, previamente conhecidas.

As amostras de textos precisam ser primeiramente transformadas em conjunto de padrões em cada amostra.

Um padrão é conjunto de características, geralmente representada por um vetor e um conjunto de padrões no formato de tabela, onde cada linha é um padrão e cada coluna, uma característica. Geralmente a última coluna é a coluna classe.

FIGURA 14 - CARACTERÍSTICAS ORGANIZADA EM TABELA

	0	1
0	Estou indo para o trabalho agora.	português
1	Do you speak English?	inglês
2	¡Que tengas un buen día!	espanhol
3	I love to read books.	inglês
4	O trânsito está terrível hoje.	português
...	...	...
87	Me encanta leer libros.	espanhol
88	Voy al parque todos los días.	espanhol
89	Where is the nearest restaurant?	inglês
90	I need to buy some groceries.	inglês
91	Quero aprender a tocar violão.	português

92 rows × 2 columns

FONTE: O autor (2024)

CONSTRUÇÃO DE ATRIBUTOS

Nesta etapa, foi realizada uma análise detalhada de cada frase ou sentença das amostras, com o objetivo de identificar características ou padrões recorrentes

em cada idioma representado. O propósito foi compreender quais padrões são representativos de cada idioma presente nos conjuntos de frases analisados.

O processo definido acima é conhecido como vetorização, existem diversos tipos de vetorização como TF-IDF, o n-gramas ou *embedding*.

Após a definição dos padrões relevantes, foram desenvolvidas funções em *python* para detectar esses padrões nas amostras de entrada. Os atributos construídos em linguagem *python* foram:

a) ***hasStopWords*:** verifica a presença de stopwords mais comuns do inglês, espanhol e português. A função retorna 1 se possui algum *stopword* do inglês; 2 se possui algum *stopword* do português e 3 se possui algum da língua espanhola;

b) ***hasAccentMark*:** verificar a presença de sinal de acentuação nas palavras;

c) ***countChar*:** conta a quantidade de caracteres presentes em uma dada sentença;

d) ***hasLatinChar*:** detectar se uma dada sentença possui caracteres latinos;

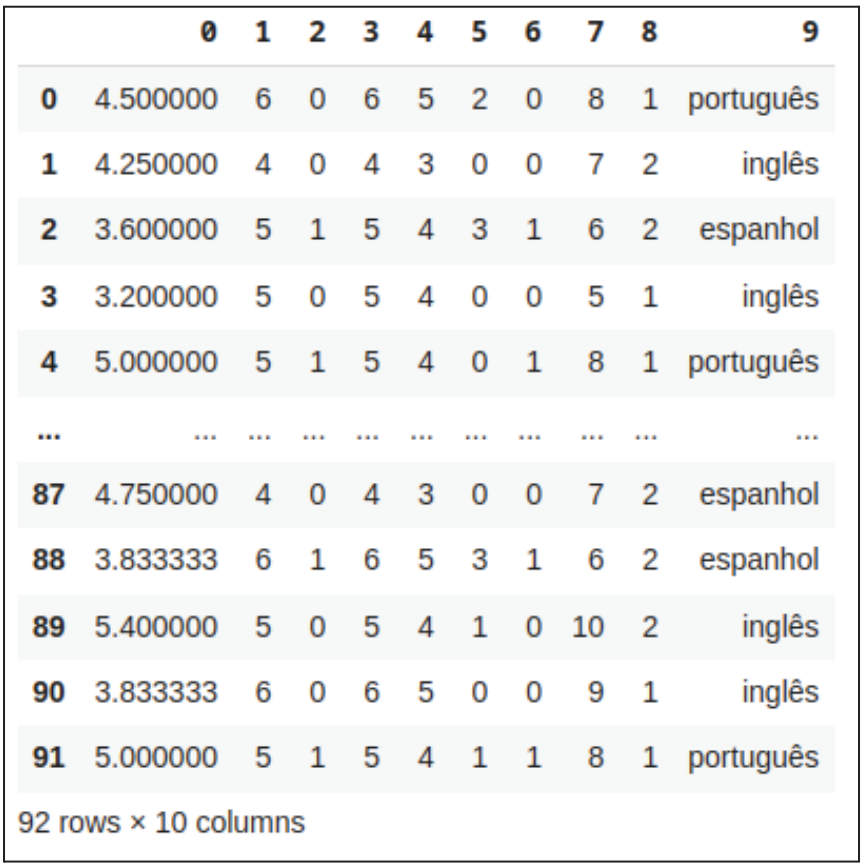
e) ***countWords:*** determina a quantidade de palavras presente em uma dada sentença;

f) ***countSpace*:** determina a quantidade de espaços presentes em uma dada sentença;

g) ***maxWordLength*:** determina o tamanho da maior palavra da frase;

h) ***minWordLength*:** determina o tamanho da menor palavra da frase.

FIGURA 15 - ALGUMAS EXEMPLOS DE CARACTERÍSTICAS EXTRAÍDAS PELOS EXTRATORES



	0	1	2	3	4	5	6	7	8	9
0	4.500000	6	0	6	5	2	0	8	1	português
1	4.250000	4	0	4	3	0	0	7	2	inglês
2	3.600000	5	1	5	4	3	1	6	2	espanhol
3	3.200000	5	0	5	4	0	0	5	1	inglês
4	5.000000	5	1	5	4	0	1	8	1	português
...	...	...	...	...	...	...	...	...	...	...
87	4.750000	4	0	4	3	0	0	7	2	espanhol
88	3.833333	6	1	6	5	3	1	6	2	espanhol
89	5.400000	5	0	5	4	1	0	10	2	inglês
90	3.833333	6	0	6	5	0	0	9	1	inglês
91	5.000000	5	1	5	4	1	1	8	1	português

92 rows × 10 columns

FONTE: O autor (2024)

TREINAMENTO DO MODELO SVM

As características extraídas pelas funções anteriores foram utilizadas como vetores de entrada para o treinamento de um modelo de classificação baseado em SVM (*Support Vector Machine*), com o objetivo de identificar automaticamente o idioma dos conjuntos de frases das amostras.

A seguir o resultado de desempenho do modelo SVM obtido durante a etapa de treinamento.

FIGURA 16 - RESULTADO DA VALIDAÇÃO CRUZADA E TREINAMENTO

```
Acurácia nos dados de treinamento: 72.46%
[[13  4  5]
 [ 1 19  3]
 [ 3  3 18]]
              precision    recall  f1-score   support

    espanhol       0.76      0.59      0.67        22
      inglês       0.73      0.83      0.78        23
   português       0.69      0.75      0.72        24

    accuracy                           0.72        69
   macro avg       0.73      0.72      0.72        69
weighted avg       0.73      0.72      0.72        69

métricas mais confiáveis
[[3 3 2]
 [1 5 1]
 [2 2 4]]
              precision    recall  f1-score   support

    espanhol       0.50      0.38      0.43         8
      inglês       0.50      0.71      0.59         7
   português       0.57      0.50      0.53         8

    accuracy                           0.52        23
   macro avg       0.52      0.53      0.52        23
weighted avg       0.52      0.52      0.51        23
```

FONTE: O autor (2024)

RESULTADOS

O modelo SVM treinado para identificação de idiomas obteve 72,46% de acurácia nos dados de treinamento e 52% nos dados de validação. O desempenho foi razoável para inglês e português, com bons índices de recall, especialmente no inglês.

APÊNDICE 7 – APRENDIZADO DE MÁQUINA

A – ENUNCIADO

Para cada uma das tarefas abaixo (Classificação, Regressão etc.) e cada base de dados (Veículo, Diabetes etc.), fazer os experimentos com todas as técnicas solicitadas (KNN, RNA etc.) e preencher os quadros com as estatísticas solicitadas, bem como os resultados pedidos em cada experimento.

B – RESOLUÇÃO

No presente apêndice, foram apresentados quatro experimentos de inteligência artificial. Cada experimento envolveu a elaboração, treinamento, teste e avaliação, usando métricas apropriadas, de modelos de classificação, regressão, agrupamento e regras de associação respectivamente.

Para medir apropriadamente os desempenhos dos modelos treinados, foram usados as seguintes métricas: R^2 (coeficiente de pearson ao quadrado); Syx (erro padrão residual); e MAE (erro absoluto médio).

EXPERIMENTO DE CLASSIFICAÇÃO

As bases de dados usadas nesse experimento de classificação foram os dados locais de veículos e diagnósticos de diabetes, ambos em formato csv.

Os algoritmos usados para realizar o treinamento, teste e avaliação do modelos subsequentes foram: KNN, RNA com e sem hiperparâmetros (*size* e *decay*) e SVM com *hold-out* e depois com *cross-validation* e usando combinações de parâmetros C e *sigma*, *Random Forest* com *hold-out* e, depois, com *cross-validation*.

BASE DE DADOS VEÍCULOS

Resultado geral

QUADRO 4 - TÉCNICAS, PARÂMETROS E MATRIZES DE CONFUSÃO RESULTANTES

Técnicas	Parâmetros	Matriz de confusão
KNN	k = 1	*Reference*

		Prediction bus opel saab van bus 38 0 0 0 opel 0 34 0 0 saab 0 0 47 0 van 0 0 0 51
RNA - Hold out	size = 5 decay = 0.1	*Reference* *Prediction* bus opel saab van bus 19 12 6 2 opel 0 0 1 0 saab 18 20 36 0 van 1 2 4 49
RNA - CV	size = 3 decay = 0.1	*Reference* *Prediction* bus opel saab van bus 35 0 0 22 opel 0 30 44 9 saab 0 0 0 0 van 3 4 3 20
SVM - Hold out	sigma = 0.06307613 C = 1	*Reference* *Prediction* bus opel saab van bus 37 0 1 0 opel 0 21 19 1 saab 0 12 27 0 van 1 1 0 50
SVM - CV	sigma = 0.06307613 C = 1	*Reference* *Prediction* bus opel saab van bus 37 0 1 0 opel 0 21 19 1 saab 0 12 27 0 van 1 1 0 50
SVM - Melhor C e sigma	sigma = 0.015 C = 50	*Reference* *Prediction* bus opel saab van bus 38 0 1 1 opel 0 24 8 1 saab 0 10 38 0 van 0 0 0 49
RF - Hold-out	mtry = 2	*Reference* *Prediction* bus opel saab van bus 38 0 2 0 opel 0 15 19 0 saab 0 17 24 1 van 0 2 2 50
RF - CV	mtry = 10	*Reference* *Prediction* bus opel saab van bus 38 0 1 0 opel 0 14 19 0 saab 0 17 26 1 van 0 3 1 50
RF - Mtry	mtry = 5	*Reference* *Prediction* bus opel saab van bus 38 0 3 0 opel 0 15 20 0 saab 0 17 22 1

		van 0 2 2 50

FONTE: Autoria própria (2024)

BASE DE DADOS DIABETES

Resultado geral

QUADRO 5 - MODELOS COM OS MELHORES DESEMPENHOS E SEUS PARÂMETROS

Técnicas	Parâmetros	Matriz de confusão
KNN	k = 9	*Reference* *Prediction* neg pos neg 87 22 pos 13 31
RNA - Hold out	size = 3 decay = 0.1	*Reference* *Prediction* neg pos neg 81 27 pos 19 26
RNA - CV	size = 5 decay = 0.1	*Reference* *Prediction* neg pos neg 84 18 pos 16 35
RNA - melhor size e decay	size = 11 decay = 0.4	*Reference* *Prediction* neg pos neg 85 23 pos 15 30
SVM - Hold out	sigma = 0.1379132 C = 0.5	*Reference* *Prediction* neg pos neg 90 22 pos 10 31
SVM - CV	sigma = 0.1379132 C = 0.25	*Reference* *Prediction* neg pos neg 90 22 pos 10 31
SVM - Melhor C e sigma	sigma = 0.01 C = 1	*Reference* *Prediction* neg pos neg 90 21 pos 10 32
Random Forest - Hold-out	mtry = 2	*Reference* *Prediction* neg pos neg 91 21 pos 9 32
Random Forest - CV e melhor mtry	mtry = 5	*Reference* *Prediction* neg pos neg 91 17

		pos 9 36

FONTE: Autoria própria (2024)

2. EXPERIMENTO DE REGRESSÃO

QUADRO 6 - RESULTADOS PARA A BASE DE DADOS ADMISSÃO

técnica	**parâmetros**	**R2**	**Syx**	**RMSE**	**MAE**
SVM - CV	sigma = 0.20 e C = 1	0.84	1.46e-05	0.052	0.038
SVM - hold-out	sigma = 0.203 e C = 0.5	0.827	4.98e-06	0.054	0.040
RF - CV	mtry = 2	0.813	0.010	0.057	0.041
RNA - CV	size = 3 e decay = 1e-04	0.812	0.012	0.057	0.042
RNA - Hold-out	size = 5 e decay = 1e-04	0.808	0.011	0.057	0.042
RF - hold out	mtry = 2	0.807	0.009	0.058	0.041
KNN	k = 9	0.757	0.026	0.065	0.050

FONTE: O autor (2024)

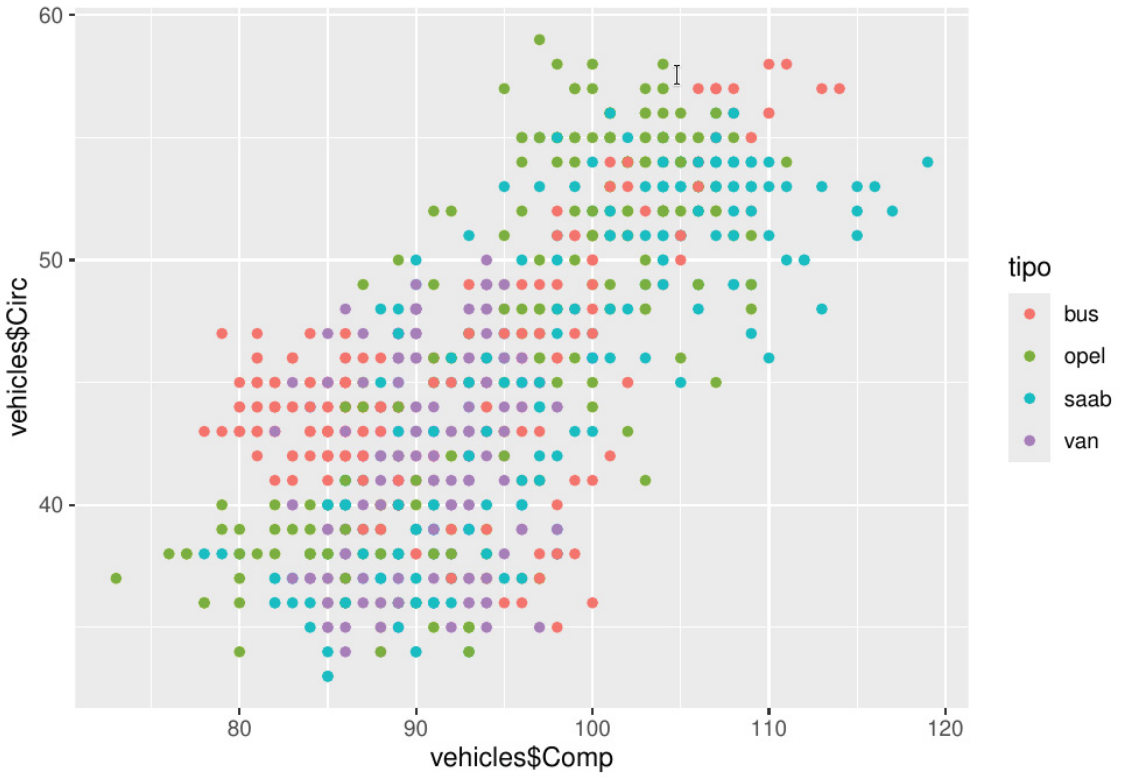
QUADRO 7 - RESULTADOS PARA A BASE DE DADOS BIOMASSA

técnica	**parâmetros**	**R2**	**Syx**	**RMSE**	**MAE**
KNN	K = 1	0.757	0.026	0.065	0.050
RF hold-out	mtry = 2	0.647	1611280	1411	246.2
RF CV	mtry = 2	0.642	1580555	1421	250.6
SVM hold-out	sigma = 1.045 C = 1	0.084	4682912	2274	488.6
SVM CV	sigma = 1.045 C = 1	0.084	4682912	2274	371.8
RNA hold-out	size = 1 decay = 1e-04	-0.042	0.109	2426	0.042
RNA CV	size = 1 decay = 1e-04	-0.042	14820664	2426	488.6

FONTE: Autoria própria (2024)

3. EXPERIMENTO DE AGRUPAMENTO

GRÁFICO 5 - GRÁFICO DE PONTOS DOS TIPOS DE VEÍCULOS EM RELAÇÃO A CIRCUNFERÊNCIA DO VEÍCULO



FONTE: O autor (2024)

Após a execução do **K-Means**, os resultados, ou seja, quantos veículos de cada tipo foram agrupados em cada cluster gerado pelo **K-Means** foram exibidos no seguinte resultado.

FIGURA 27 - GRUPOS GERADOS PELO K-MEANS

```
##       bus opel saab van
## 1      11   34   42   0
## 2       0   19   21  47
## 3      48   16   19  61
## 4       4   23   23  42
## 5      19    0    0   0
## 6       7   43   34   0
## 7       5   29   26   3
## 8      25   18   17   3
## 9      39   24   34   7
## 10     60    6    1  36
```

FONTE: O autor (2024)

4. EXPERIMENTO DE REGRAS DE ASSOCIAÇÃO

Regras de associação são técnicas de mineração de dados utilizadas para descobrir relações ou padrões frequentes entre itens em grandes conjuntos de dados, especialmente em transações de mercado.

Foi realizado um experimento de regras de associação sobre o conjunto de dados de musculação com informações coletadas sobre treinos, exercícios, desempenho físico e características corporais de praticantes de musculação. As regras foram geradas com uma configuração de suporte e confiança.

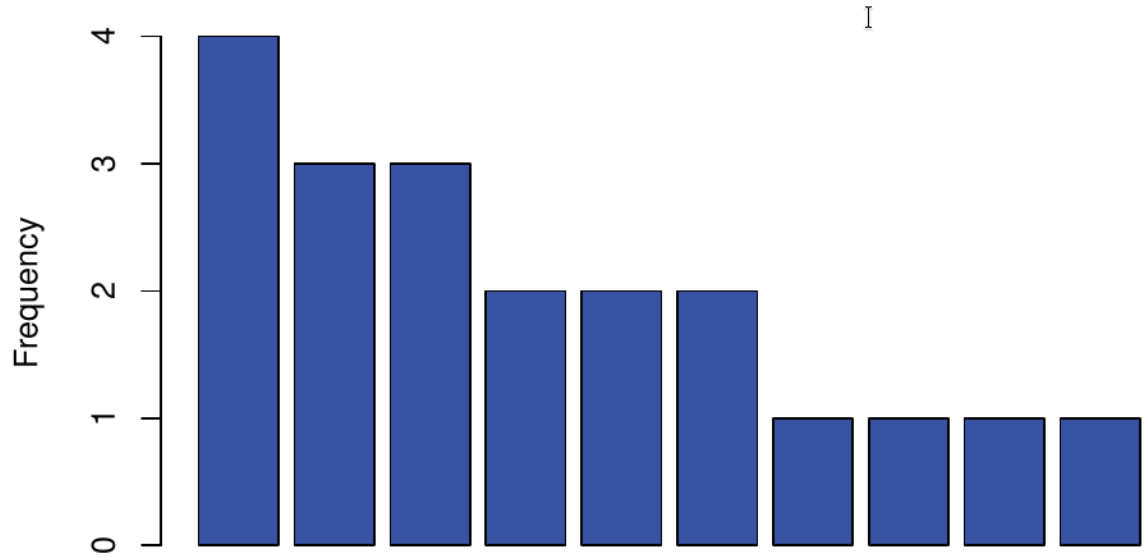
FIGURA 28 - ALGUMAS TRANSAÇÕES OBTIDAS

```
## transactionID   items
## 1               LegPress;Gemeos;Agachamento
## 2               Gemeos;LegPress;Afundo;Agachamento
## 3               Adutor;Agachamento;LegPress;Adutor
```

FONTE: O autor (2024)

Para visualizar a frequência das 10 primeiras transações, foi elaborado um gráfico de barras com as frequências.

GRÁFICO 7 - FREQUÊNCIA DAS 10 PRIMEIRAS TRANSAÇÕES



FONTE: O autor (2024)

E, por fim, para obter as regras de associações sobre as transações foi definido um suporte de 0.001 e uma confiança de 0.7

Nenhuma regra foi obtida, as possíveis razões para esse resultado podem ser a quantidade de dados da base que é muito pequena; e a confiança pode estar muito alta dado a pequena quantidade de dados na base.

APÊNDICE 8 – DEEP LEARNING

A – ENUNCIADO

1 Classificação de Imagens (CNN)

Implementar o exemplo de classificação de objetos usando a base de dados CIFAR10 e a arquitetura CNN vista no curso.

2 Detector de SPAM (RNN)

Implementar o detector de spam visto em sala, usando a base de dados SMS Spam e arquitetura de RNN vista no curso.

3 Gerador de Dígitos Fake (GAN)

Implementar o gerador de dígitos *fake* usando a base de dados MNIST e arquitetura GAN vista no curso.

4 Tradutor de Textos (Transformer)

Implementar o tradutor de texto do português para o inglês, usando a base de dados e a arquitetura Transformer vista no curso.

B – RESOLUÇÃO

Questão 1 - Classificação de Imagens (CNN)

Foi implementado um exemplo de classificação de objetos usando a base de dados **CIFAR10** e a arquitetura CNN (*Convolutional Neural Network*). A linguagem de programação utilizada para implementar o modelo foi o Python.

Relatório sobre a arquitetura da rede

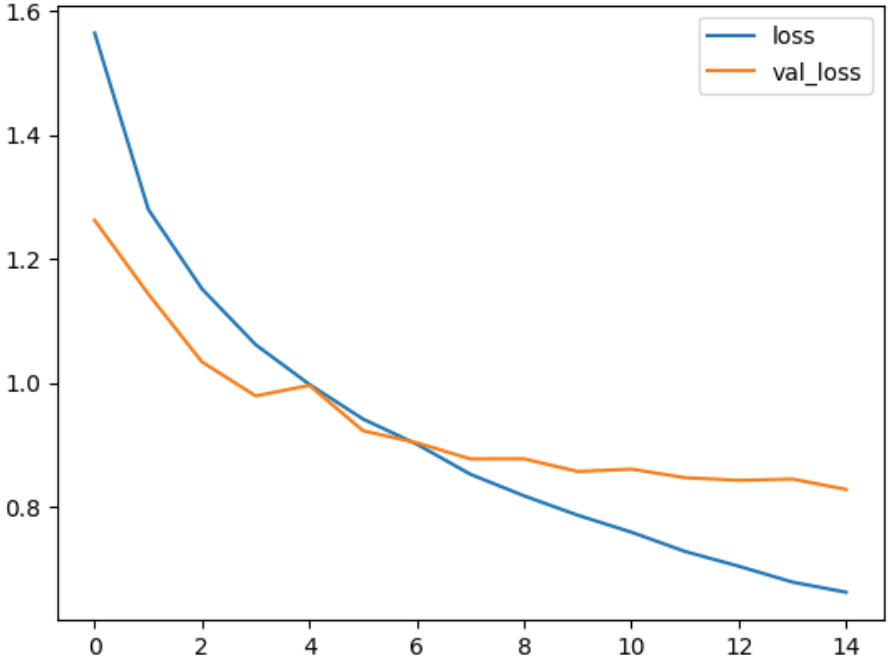
FIGURA 29 - ARQUITETURA DA REDE CNN OBTIDA

Layer (type)	Output shape	Param #
input_layer (InputLayer)	(None, 32, 32, 3)	0
conv2d (Conv2D)	(None, 15, 15, 32)	896
conv2d_1 (Conv2D)	(None, 7, 7, 64)	18,496
conv2d_2 (Conv2D)	(None, 3, 3, 128)	73,856
flatten (Flatten)	(None, 1152)	0
dropout (Dropout)	(None, 1152)	0
dense (Dense)	(None, 1024)	1,180,672
dropout_1 (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 10)	10,250

Total params: 1,284,170 (4.90 MB)
Trainable params: 1,284,170 (4.90 MB)
Non-trainable params: 0 (0.00 B)

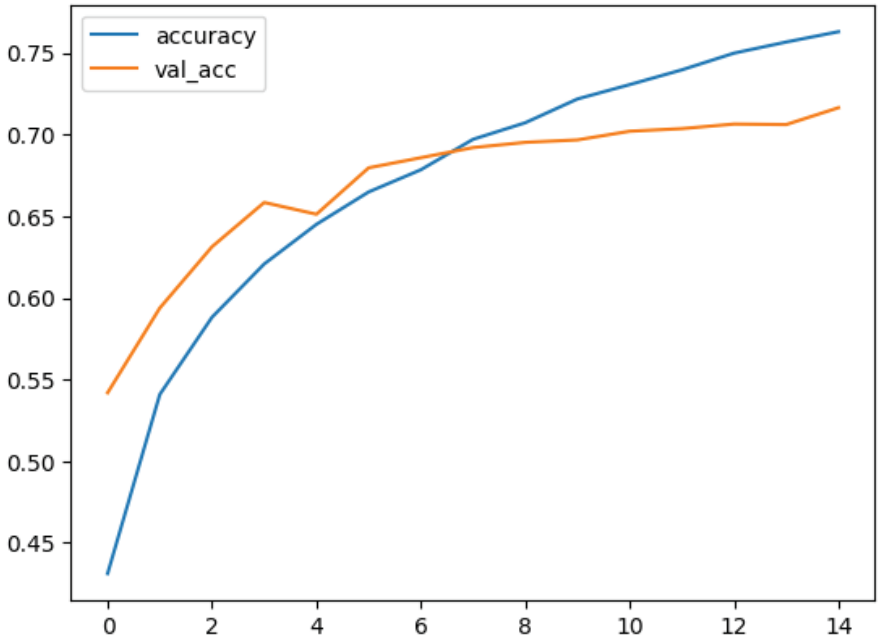
FONTE: O autor (2025)

GRÁFICO 8 - EVOLUÇÃO DA FUNÇÃO DE PERDA DURANTE O TREINAMENTO DA CNN



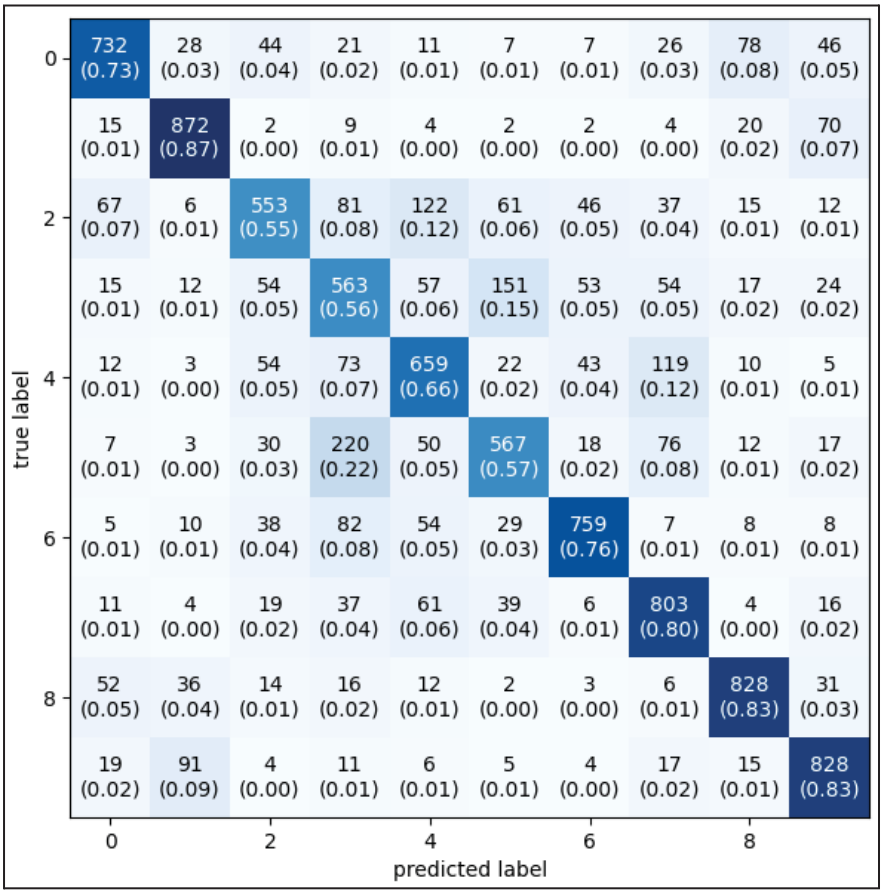
FONTE: O autor(2024)

GRÁFICO 9 - EVOLUÇÃO DA ACURÁCIA DURANTE O TREINAMENTO DA CNN



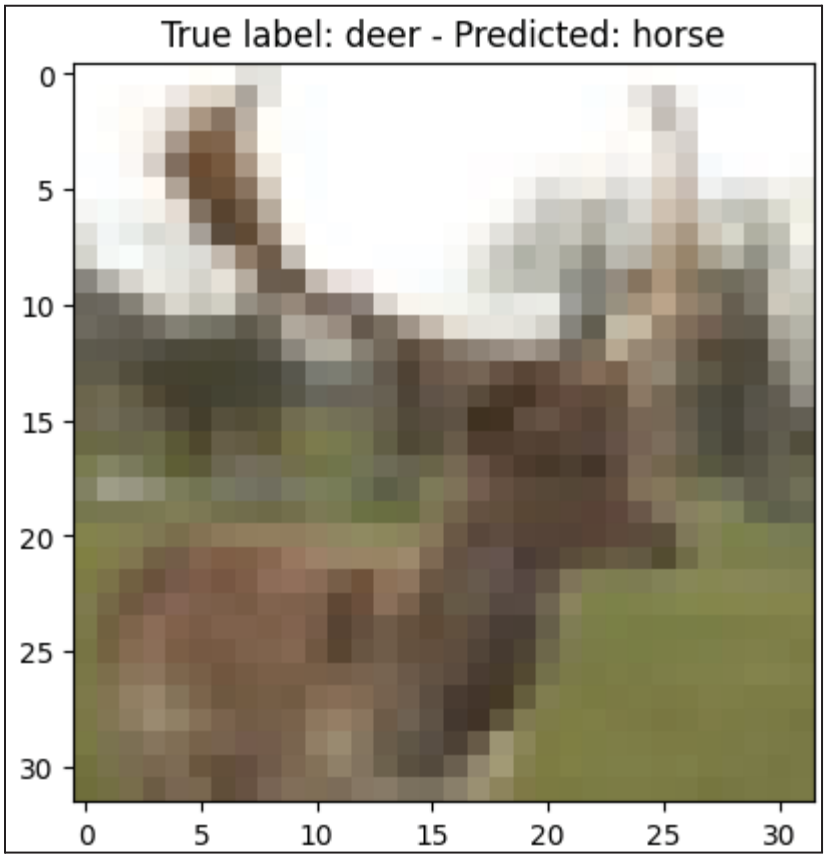
FONTE: Autoria própria (2024)

FIGURA 30 - MATRIZ DE CONFUSÃO DA PREDIÇÕES REALIZADAS PELA CNN



FONTE: O autor(2024)

FIGURA 31 - UM EXEMPLO DE CLASSIFICAÇÃO INCORRETA



FONTE: Autoria própria (2024)

Questão 2 - Detector de SPAM (RNN)

Foi implementado um modelo para detectar quais mensagens são spams, usando a base de dados SMS Spam e a arquitetura de RNN (Redes neurais recorrentes). A linguagem de programação utilizada para implementar o modelo foi Python.

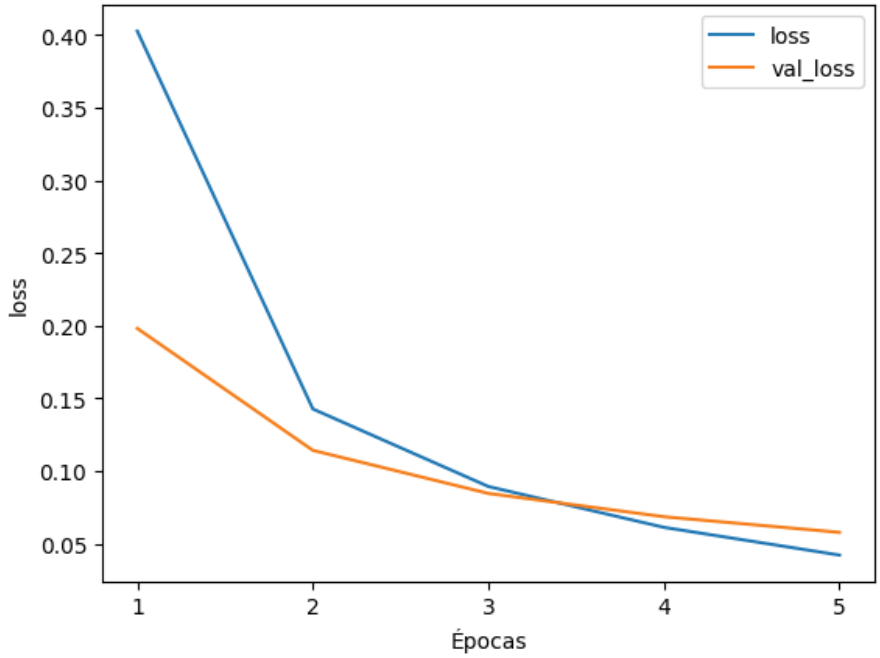
FIGURA 32 - ARQUITETURA DA REDE RNN OBTIDA

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 162)	0
embedding (Embedding)	(None, 162, 20)	145,720
lstm (LSTM)	(None, 5)	520
dense (Dense)	(None, 1)	6

```
Total params: 146,246 (571.27 KB)
Trainable params: 146,246 (571.27 KB)
Non-trainable params: 0 (0.00 B)
```

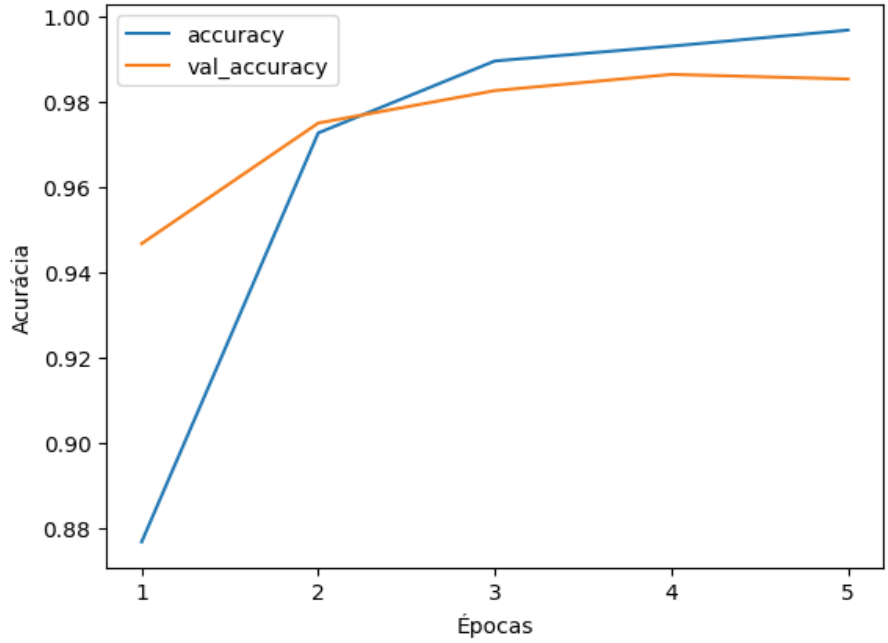
FONTE: O autor (2025)

GRÁFICO 10 - EVOLUÇÃO DA FUNÇÃO DE PERDA DURANTE O TREINAMENTO DA RNN



FONTE: O autor (2025)

GRÁFICO 11 - EVOLUÇÃO DA ACURÁCIA DURANTE O TREINAMENTO DA CNN



FONTE: O autor(2025)

Predição de um texto novo

FIGURA 33 - TEXTO TRANSFORMADO EM UM VETOR

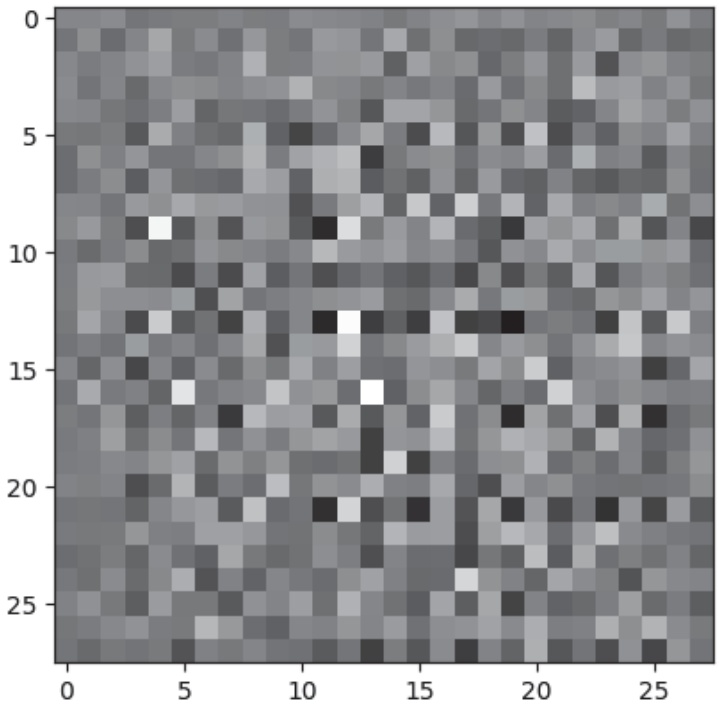
```
texto = "Is yuour car dirty? Discovery our new product. Free for all. Click the
link"

Resultado:

[[0.08851228]]
OK
```

FONTE: O autor (2025)

Questão 3 - Gerador de Dígitos Fake (GAN)

Foi implementado um modelo GAN (Generative Adversarial Network) para gerar e detectar dígitos *fakes*. GAN é um tipo de rede neural profunda criada para gerar dados novos e realistas com base em um conjunto de dados de treino. O mesmo é composto de um gerador que vai gerar dígitos falsos que se pareçam com dígitos reais; e um discriminador que tenta determinar se um dígito de entrada é real (oriundo do dataset) ou falso (criado pelo gerador).

FIGURA 34 - IMAGEM DE TESTE GERADA PELO GERADOR



FONTE: O autor(2024)

FIGURA 35 - DÍGITOS MANUSCRITOS GERADOS PELA REDE GAN TREINADA

FONTE: O autor (2024)

Questão 4 - Tradutor de Textos (Transformer)

Foi implementado um modelo Transformer em linguagem de programação Python para realizar a tradução automática de textos em português para inglês.

Foi utilizada uma base de dados que contém textos em português com suas respectivas versões em inglês. O modelo Transformer foi treinado com essa base para em seguida ser utilizado para tradução automática de textos em português para inglês.

QUADRO 8 - SENTENÇA TRADUZIDA PARA O INGLÊS PELO O MODELO OBTIDO

Sentença	**Tradução da sentença**
acima, eu vi um céu azul e um sol vermelho	b'most , i saw a blue sky and a red sun .'

FONTE: O autor (2024)

APÊNDICE 9 – BIG DATA

A – ENUNCIADO

Enviar um arquivo PDF contendo uma descrição breve (2 páginas) sobre a implementação de uma aplicação ou estudo de caso envolvendo Big Data e suas ferramentas (NoSQL e NewSQL). Caracterize os dados e Vs envolvidos, além da modelagem necessária dependendo dos modelos de dados empregados.

B – RESOLUÇÃO

Foi desenvolvido um passo a passo de como montar uma arquitetura big data para sugestões de listas de compras.

INTRODUÇÃO

Uma empresa chamada LIN Technologies possui uma aplicação muito simples e trivial para o mercado: uma aplicação para usuário criar listas de compras de supermercado. O nome dessa aplicação é shoplist. A empresa, desde o início dessa aplicação, armazena em nuvem todas as listas de compras de todos os usuários dessa aplicação.

A empresa deseja usar esses dados para treinar modelos de regras de associação e usar o modelo resultante para sugerir listas de compras, com itens apresentando seus preços médios para um usuário em tempo real à medida que o mesmo vai adicionando itens na sua lista de compras. Também deve calcular, com base nos preços médios dos itens, qual poderá ser o preço médio total ao comprar todos os itens da lista sugeridos.

Dado a grande quantidade de dados, será necessário elaborar uma arquitetura de big data de tal forma que melhor atenda à essa explicada demanda da empresa LIN Technologies.

Nas próximas seções será descrita a arquitetura de big data que será usada e descrições dos dados a partir dos conceitos dos Vs usados, seguindo uma abordagem que envolve a coleta, armazenamento, processamento e análise dos dados.

CARACTERÍSTICAS DOS DADOS

a) **volume:** A aplicação shoplist armazena grandes volumes de dados provenientes de listas de compras para supermercado criadas pelos usuários da aplicação, contendo informações como nome do item e quantidade do item a partir de alguma unidade de medida (como, und, kg, cm etc);

b) **velocidade:** O fluxo de dados não é contínuo, não exigindo que a aplicação processe informações em tempo real. O fluxo de dados e o processo são realizados em batch para sugerir produtos (itens) e quantidade desses produtos aos usuários da aplicação shoplist;

c) **variedade:** Os dados são gerados a partir de diferentes fontes e em dois diferentes formatos: lista de compras com o nome, a quantidade e o preço de cada item da lista de compras (semi estruturados), informações dos usuários, como nome e endereço (estruturados); informações sobre o supermercado, com o nome e posição geográfica do mesmo (semi estruturado).

COLETA DE DADOS

As fontes de dados são as listas de compras feitas por diversos usuários que usam o aplicativo shoplist para efetuar compras em um supermercado. O modelo de banco de dados selecionado para armazenar essas informações relativas a cada compra é o MongoDB, um banco orientado a documentos que oferece maior flexibilidade em armazenar informações das listas de compras. No entanto, para processamento dos dados será utilizado um banco de dados do tipo colunas.

Na tabela a seguir, um exemplo de lista de compras realizada por um determinado cliente.

QUADRO 9 - EXEMPLO DE LISTA DE COMPRAS REALIZADAS POR UM CLIENTE

SupermercadoID: 340394		
posSupermercado: [-4.222651246225341, -44.79355330567182]		
ClienteID: 120392		
Data hora: 12/07/2024 13:30:00		
Endereço Cliente: Trav. João alves 324A cohab 1		
Ovos	12 und	15,00
Pão	6 und	08,00
Arroz	1 kg	05,00
Feijão	2 kg	03,50
Manteiga	0.5 kg	31,00
Macarrão	0.2 kg	02,35
Total: 83,85		

FONTE: O autor (2025)

ARMAZENAMENTO DE DADOS

Será configurado um Data Lake para armazenar dados brutos (estruturados e semiestruturados) e Data Warehouse para dados mais estruturados irá armazenar os dados já processados e limpos.

Será armazenado dados de todas as listas de compras de todos os usuários; identificação e localização dos supermercados onde cada lista criada foi usada para realizar as compras; horários das compras; e dados sobre os clientes. Isso permite um armazenamento flexível de dados estruturados e não estruturados.

Para criar o Data Lake usar os buckets S3 da amazon para armazenar os dados não-estruturados e semiestruturados, dados em formatos como JSON, CSV, logs e etc. No aplicativo shoplist os dados serão armazenados no formato de documentos chave-valor, nesse caso, o banco usado pode ser MongoDB/Cassandra, um banco orientado a documentos ou orientado a colunas. Exemplo de uma lista de compra registrada no MongoDB:

FIGURA 36 - EXEMPLO DE ESTRUTURA DE DADO DE UMA COMPRA REALIZADA EM UMA SUPERMERCADO


```
{
  "SupermercadoID": 340394,
  "posSupermercado": [-4.222651246225341, -44.79355330567182],
  "ClienteID": 120392,
  "DataHora": "2024-07-12T13:30:00",
  "endereco": "Trav. João alves 324A cohab 1",
  "Itens": [
     {
       "Produto": "Ovos", "Quantidade": "12 und", "Preco": 15.00
     },
     {
       "Produto": "Pão", "Quantidade": "6 und", "Preco": 8.00
     },
     { "Produto": "Arroz", "Quantidade": "1 kg", "Preco": 5.00 },
     { "Produto": "Feijão", "Quantidade": "2 kg", "Preco": 3.50 },
     {
        "Produto": "Manteiga",
        "Quantidade": "0.5 kg",
        "Preco": 31.00
     },
     {
        "Produto": "Macarrão",
        "Quantidade": "0.2 kg",
        "Preco": 2.35
      }
    ],
    "Total": 83.85
}
```


FONTE: O autor (2025)

Para construir o Data Warehouse, será criado tabelas que organizem os dados transacionais e comportamentais. Os dados processados são armazenados no Data Warehouse como Amazon redshift, Google BigQuery ou Snowflake.

Por fim, os pipelines de coletas de dados devem ser apropriadamente conectados ao S3 e Redshift (ou BigQuery ou Snowflake).

Os dados coletados das listas de compras criadas brutas precisam ser limpos e transformados para garantir que estão padronizados. A seguir o exemplo de tabela fatos obtida a partir da reconfiguração de listas de compras armazenadas no Data Lake:

QUADRO 10 - TABELA FATO COM AS CHAVES ESTRANGEIRAS E VALORES QUANTITATIVOS

Transacao	Cliente	Supermercado	Data	Produto	Quantidade	Valor
10001	120392	340394	2024-07-12 13:30:00	1	12 und	15,00
10001	120392	340394	2024-07-12 13:30:00	2	6 und	8,00
10001	120392	340394	2024-07-12 13:30:00	3	1 kg	05,00
10001	120392	340394	2024-07-12 13:30:00	4	2 kg	03,50

FONTE: Autoria própria (2025)

Essa tabela armazena os dados transacionais, como ID de transações, IDs de clientes, IDs de supermercados, data da compra, produtos comprados, quantidades e valores.

E para finalizar a definição do *star schema*, a tabela a seguir armazena informações descritivas sobre a dimensão do produto, contendo os dados id, nome, categoria e unidade do produto, sendo um exemplo das tabelas dimensões que serão elaboradas.

QUADRO 11 - TABELA DIMENSÃO SOBRE PRODUTO

Produto	Nome	Categoria	Unidade
1	Ovos	Alimentos	und
2	Pão	Alimentos	und
3	Feijão	Alimentos	kg
4	Arroz	Alimentos	kg
5	Manteiga	Lacticínios	kg

FONTE: Autoria própria (2025)

Outras tabelas dimensão poderiam ser construídas a partir do processamento e tratamento dos dados brutos oriundos do Data Lake.

ANÁLISE DE DADOS

Criar um ML (Machine Learning) para recomendações de produtos para a lista de compras com base no padrão de procurar de produtos dos clientes do aplicativo shoplist. Três estratégias podem ser usadas.

A primeira estratégia pode ser a filtragem colaborativa, o algoritmo recomenda itens para a lista de compras do cliente com base nas preferências de clientes similares. Por exemplo, se o Cliente X comprar arroz e feijão, e o Cliente Y também comprar arroz, o sistema sugere feijão como item na lista do cliente Y.

Segunda estratégia pode ser regras de associação, *market basket analysis*, usado para identificar padrões repetitivos nos itens das listas de compras. Por exemplo, se muitos clientes compram pão e manteiga juntos, o *shoplist* pode recomendar um se o cliente adicionar o outro na sua lista.

Por fim, podem ser usados algoritmos como redes neurais, *decision tree* ou modelos de regressão podem ser usados para prever produtos que os clientes desejariam comprar com base em padrões históricos de listas de compras.

FERRAMENTAS E TECNOLOGIAS

a) **Hadoop:** Para o armazenamento de grandes volumes de dados;

b) **Spark (com MLlib):** Para processamento de dados em tempo real e execução de algoritmos de recomendação;

c) **Kafka:** Para ingestão de dados em tempo real;

d) **Scikit-learn ou TensorFlow:** Para modelagem de algoritmos de machine learning;

e) **NoSQL (MongoDB/Cassandra):** Para armazenamento de transações;

f) **Tableau ou Power BI:** Para visualização e monitoramento de resultados.

CONSIDERAÇÕES FINAIS

E como serão feitas as recomendações de itens para as listas de compras dos clientes do shoplist ? As recomendações serão feitas em interfaces diretas com os clientes, escrevendo sugestões de autocomplete semelhante ao copilot ou ao gemini, onde o algoritmo escreve sugestões de itens para a lista de compras do cliente.

APÊNDICE 10 – VISÃO COMPUTACIONAL

A – ENUNCIADO

1) **Extração de Características**

Os bancos de imagens fornecidos são conjuntos de imagens de 250x250 pixels de imuno-histoquímica (biópsia) de câncer de mama. No total são 4 classes (0, 1+, 2+ e 3+) que estão divididas em diretórios. O objetivo é classificar as imagens nas categorias correspondentes. Uma base de imagens será utilizada para o treinamento e outra para o teste do treino.

As imagens fornecidas são recortes de uma imagem maior do tipo WSI *(Whole Slide Imaging*) disponibilizada pela Universidade de Warwick (link). A nomenclatura das imagens segue o padrão XX_HER_YYYY.png, onde XX é o número do paciente e YYYY é o número da imagem recortada. Separe a base de treino em 80% para treino e 20% para validação. **Separe por pacientes (XX), não utilize a separação randômica! Pois, imagens do mesmo paciente não podem estar na base de treino e de validação, pois isso pode gerar um viés.** No caso da CNN VGG16 remova a última camada de classificação e armazene os valores da penúltima camada como um vetor de características. Após o treinamento, os modelos treinados devem ser validados na base de teste.

Tarefas:

a) Carregue a base de dados de **Treino.**
b) Crie partições contendo 80% para treino e 20% para validação (atenção aos pacientes).
c) Extraia características utilizando LBP e a CNN VGG16 (gerando um csv para cada extrator).
d) Treine modelos Random Forest, SVM e RNA para predição dos dados extraídos.
e) Carregue a base de **Teste** e execute a tarefa 3 nesta base.
f) Aplique os modelos treinados nos dados de treino
g) Calcule as métricas de Sensibilidade, Especificidade e F1-Score com base em suas matrizes de confusão.
h) Indique qual modelo dá o melhor o resultado e a métrica utilizada

2) Redes Neurais

Utilize as duas bases do exercício anterior para treinar as Redes Neurais Convolucionais VGG16 e a Resnet50. Utilize os pesos pré-treinados (*Transfer Learning*), refaça as camadas *Fully Connected* para o problema de 4 classes. Compare os treinos de 15 épocas com e sem *Data Augmentation*. Tanto a VGG16 quanto a Resnet50 têm como camada de entrada uma imagem 224x224x3, ou seja, uma imagem de 224x224 pixels coloridos (3 canais de cores). Portanto, será necessário fazer uma transformação de 250x250x3 para 224x224x3. Ao fazer o *Data Augmentation* **cuidado** para não alterar demais as cores das imagens e atrapalhar na classificação.

Tarefas:

a) Utilize a base de dados de **Treino** já separadas em treino e validação do exercício anterior
b) Treine modelos VGG16 e Resnet50 adaptadas com e sem *Data Augmentation*
c) Aplique os modelos treinados nas imagens da base de **Teste**
d) Calcule as métricas de Sensibilidade, Especificidade e F1-Score com base em suas matrizes de confusão.
e) Indique qual modelo dá o melhor o resultado e a métrica utilizada

B – RESOLUÇÃO

Questão 1 - Extração de características

Os resultados do desempenho de diferentes modelos de aprendizado de máquina e redes neurais na classificação de imagens de biópsias de câncer de mama, usando LBP (Local Binary Patterns) e VGG16 como métodos de extração de características. As métricas avaliadas são:

a) **Acurácia:** Proporção de previsões corretas;
b) **Sensibilidade *(Recall)*:** Capacidade de detectar corretamente os casos positivos;
c) **Especificidade:** Capacidade de evitar falsos positivos;
d) ***F1-Score*:** Média harmônica entre Precision e Sensitivity (melhor para classes desbalanceadas).

QUADRO 12 - SVM, *RANDOM FOREST* E RNA COM LBP

Modelo	Acurácia	Sensibilidade	Especificidade	F1-Score
SVM - LBP	68.9%	68.6%	89.6%	69%
Random Forest - LBP	68.9%	68.6%	89.6%	69%
RNA - LBP	68.9%	68.6%	89.6%	69%

FONTE: O autor (2025)

QUADRO 13 - SVM, *RANDOM FOREST* E RNA COM VGG16

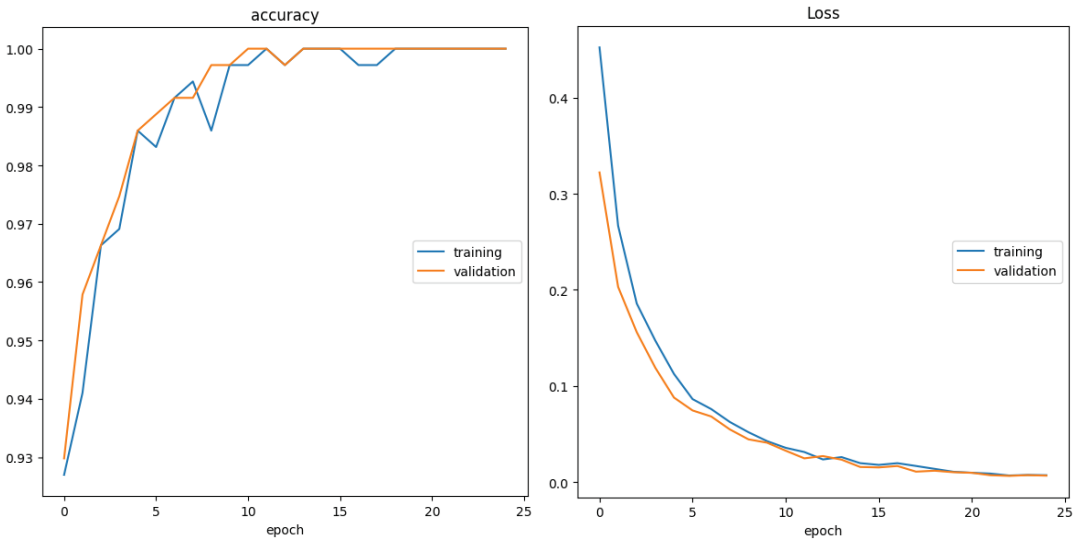
Modelo	Acurácia	Sensibilidade	Especificidade	F1-Score
SVM - VGG16	70.5%	70.3%	90.2%	70.2%
Random Forest - VGG16	58.8%	58.5%	86.3%	57.8%
RNA - VGG16	63.8%	63.5%	87.9%	63.5%

FONTE: O autor (2025)

Questão 2 - Redes Neurais

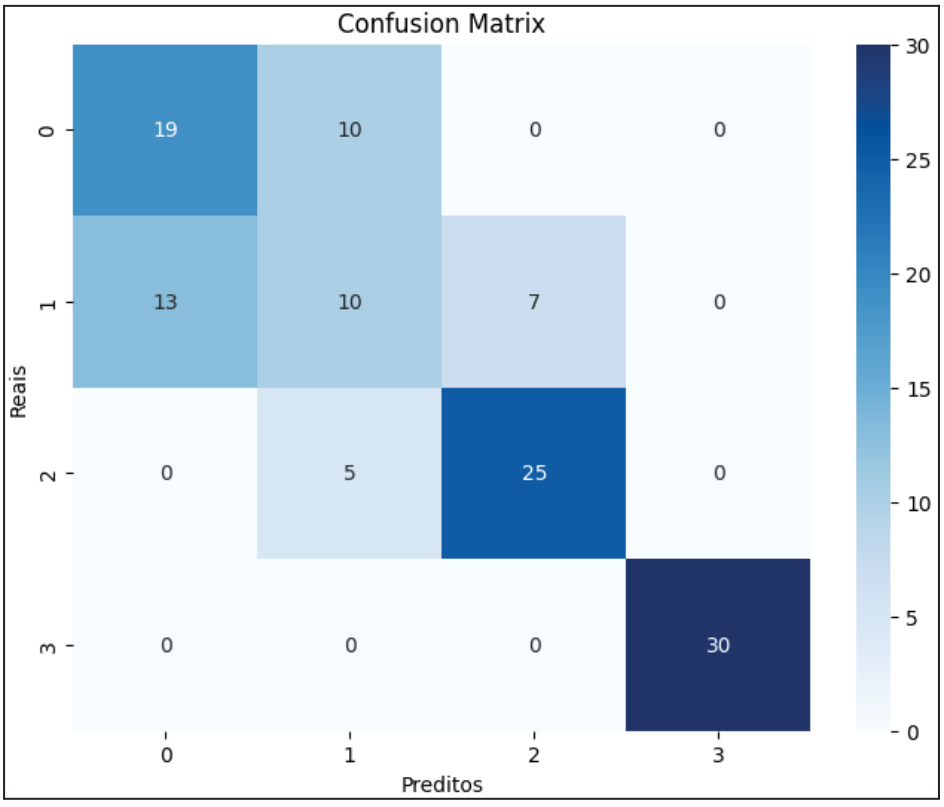
Resultados modelo VGG16 com *data augmentation*

GRÁFICO 12 - EVOLUÇÃO DA ACURÁCIA E *LOSS* DURANTE O TREINAMENTO



FONTE: O autor (2025)

FIGURA 37 - MATRIZ DE CONFUSÃO DO VGG16 COM *DATA AUGMENTATION*



FONTE: O autor (2025)

Métricas de desempenho

a) **Acurácia:** 0.70;

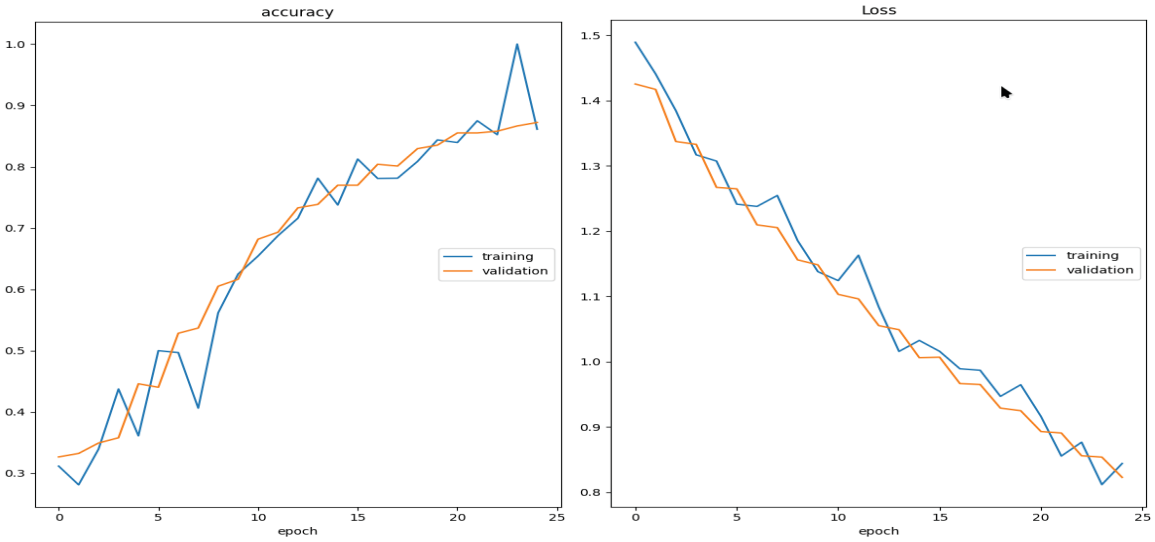
b) **Sensibilidade:** 0.70;

c) **Especificidade:** 0.90;

d) ***F1-Score*:** 0.69;

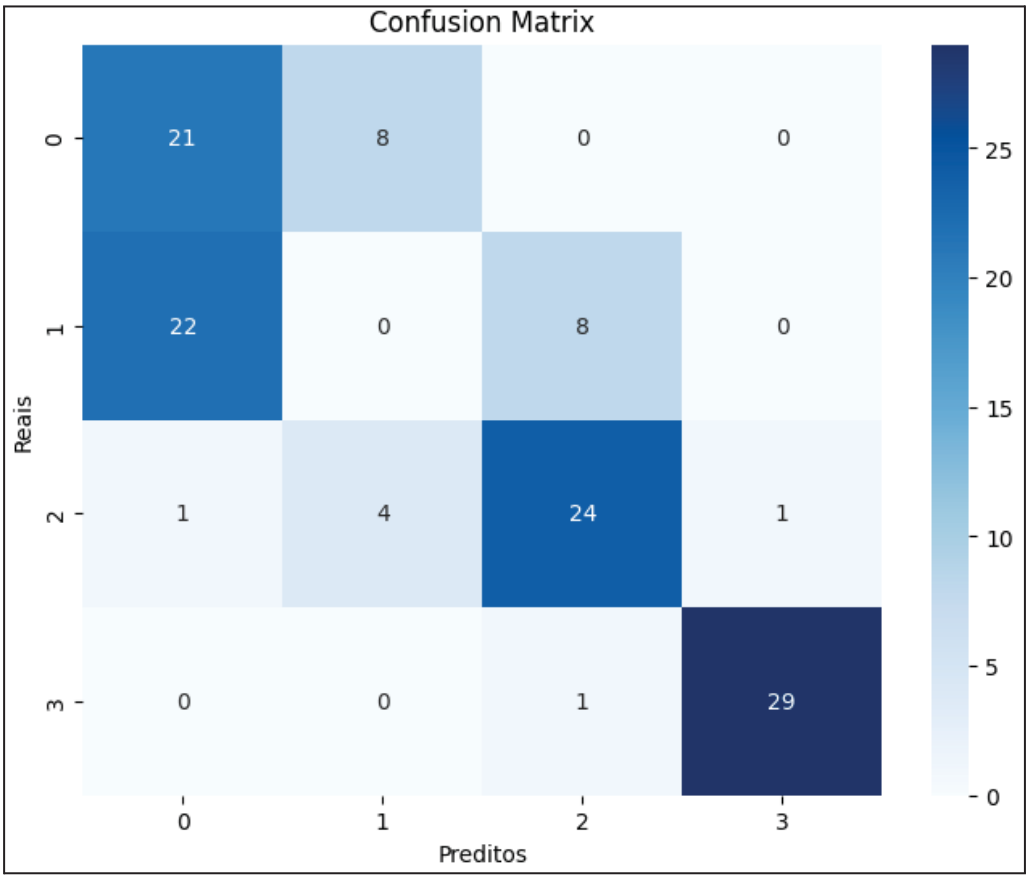
Resultados do modelo VGG16 sem augmentation

GRÁFICO 13 - EVOLUÇÃO DA ACURÁCIA DE *LOSS* DURANTE O TREINAMENTO DO VGG16 SEM *AUGMENTATION*



FONTE: O autor (2025)

FIGURA 38 - MATRIZ DE CONFUSÃO DO VGG16 SEM *DATA AUGMENTATION*



FONTE: O autor (2025)

Valores das métricas obtidas:

a) **Acurácia:** 0.62;

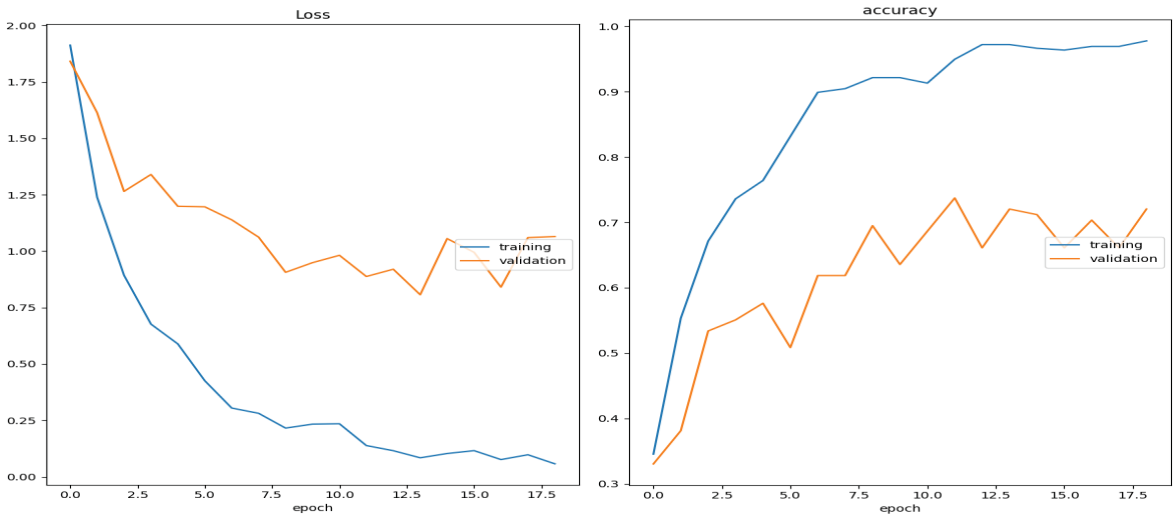
b) **Sensibilidade:** 0.62;

c) **Especificidade:** 0.87;

d) ***F1-Score*:** 0.57.

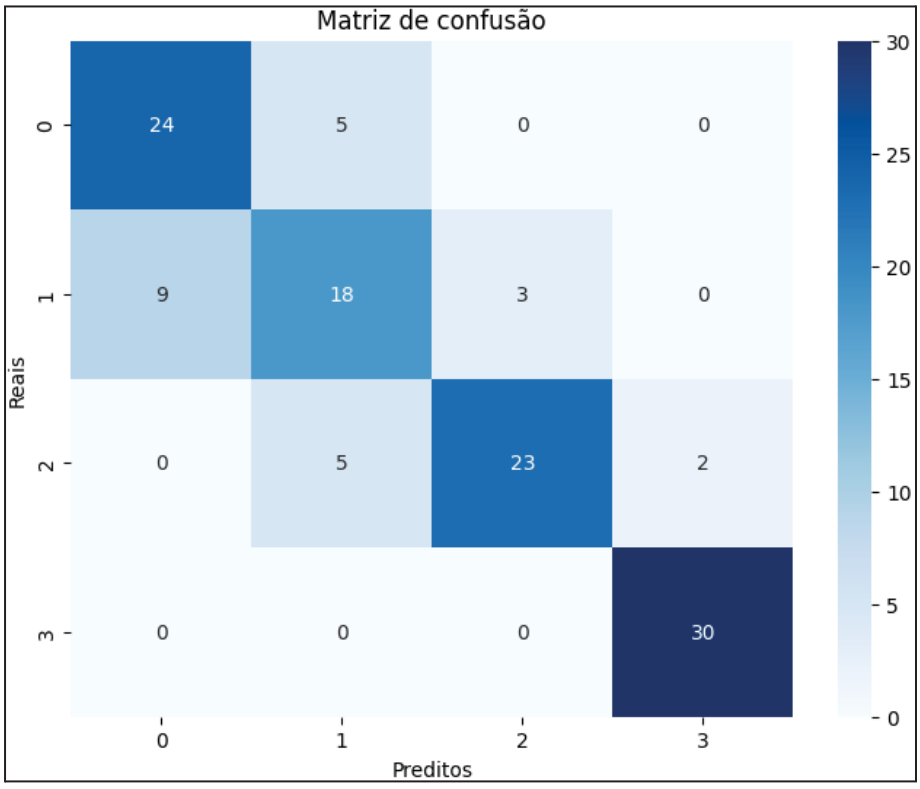
Resultados do modelo ResNet50 com data augmentation

GRÁFICO 14 - EVOLUÇÃO DA ACURÁCIA E DO *LOSS* (CUSTO) DURANTE O TREINAMENTO DO MODELO



FONTE: O autor (2025)

FIGURA 39 - MATRIZ DE CONFUSÃO RESNET50 COM *DATA AUGEMENTATION*



FONTE: O autor (2025)

Valores das métricas obtidas:

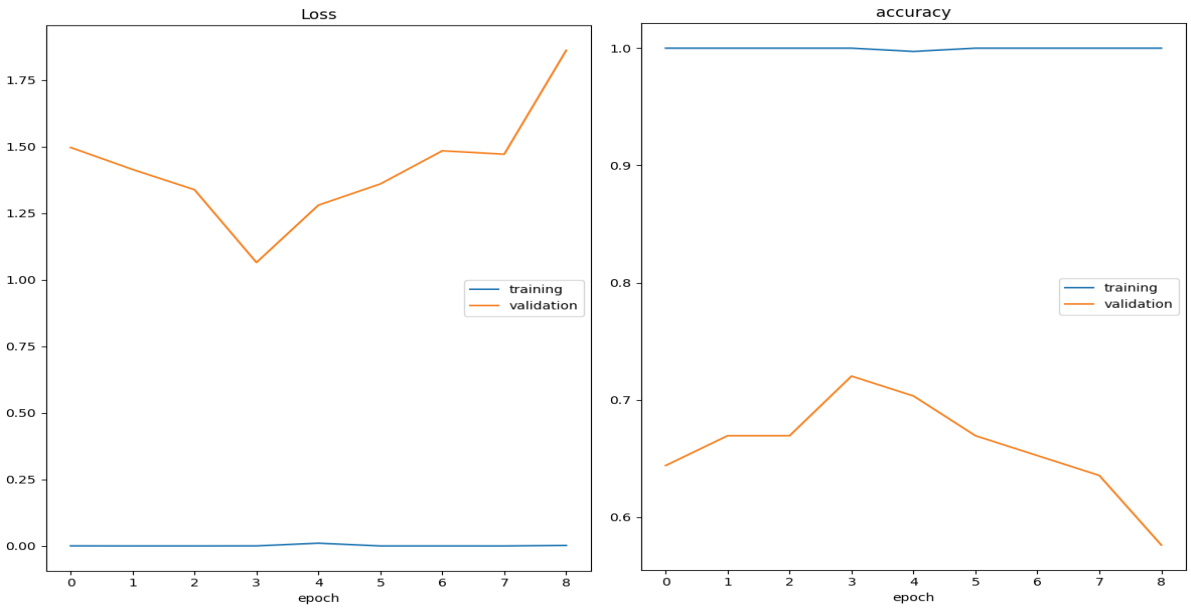
a) **Acurácia:** 0.79;

b) **Sensibilidade:** 0.79;

c) **Especificidade:** 0.93;

d) **F1-Score:** 0.79.

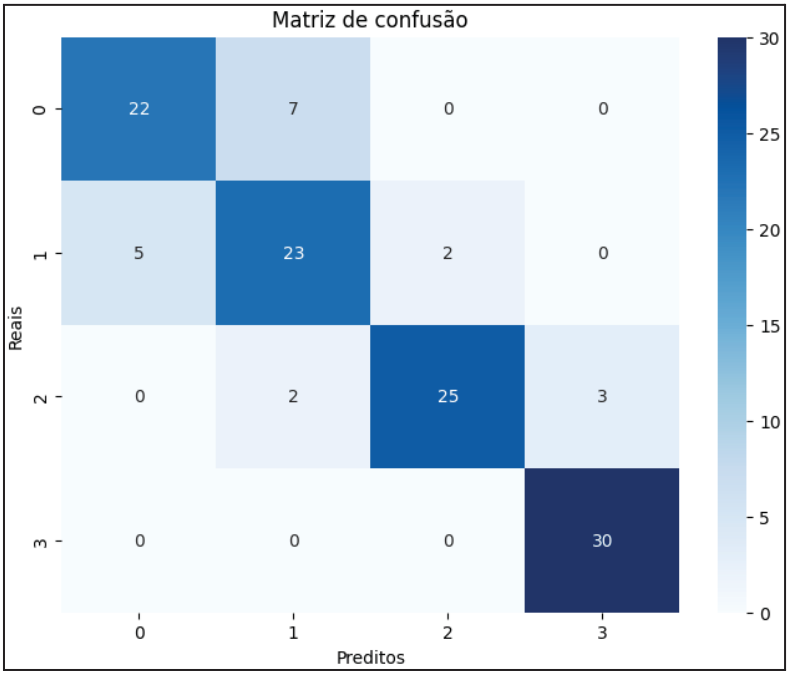
GRÁFICO 15 - EVOLUÇÃO DA ACURÁCIA E DO LOSS (CUSTO) DURANTE O TREINAMENTO DO MODELO



FONTE: O autor (2025)

Resultados do modelo ResNet50 sem *data augmentation*

FIGURA 40 - MATRIZ DE CONFUSÃO DO RESNET50 SEM AUGMENTATION.



FONTE: O autor (2025)

Valores das métricas obtidas:

a) **Acurácia:** 0.84;

b) **Sensibilidade:** 0.83;

c) **Especificidade:** 0.94;

d) **F1-Score:** 0.83.

Interpretação dos resultados

O Melhor modelo entre todos: Resnet50 sem *data augmentation* (84% de acurácia). Obteve uma sensibilidade boa de 83% demonstrando boa capacidade de identificar corretamente os casos positivos e uma especificidade excelente, de 94%, demonstrando uma excelente capacidade em identificar os casos negativos.

O modelo VGG16 com e sem *data augmentation* também obtiveram bons desempenho (79% de acurácia ambos).

Por outro lado, todos os modelos que utilizaram LBP obtiveram desempenhos medianos, inferiores aos modelos baseados exclusivamente em redes neurais convolucionais.

APÊNDICE 11 – ASPECTOS FILOSÓFICOS E ÉTICOS DA IA

A – ENUNCIADO

Título do Trabalho: "Estudo de Caso: Implicações Éticas do Uso do ChatGPT"

Trabalho em Grupo: O trabalho deverá ser realizado em grupo de alunos de no máximo seis (06) integrantes.

Objetivo do Trabalho: Investigar as implicações éticas do uso do ChatGPT em diferentes contextos e propor soluções responsáveis para lidar com esses dilemas.
Parâmetros para elaboração do Trabalho:

1. Relevância Ética: O trabalho deve abordar questões éticas significativas relacionadas ao uso da inteligência artificial, especialmente no contexto do ChatGPT. Os alunos devem identificar dilemas éticos relevantes e explorar como esses dilemas afetam diferentes partes interessadas, como usuários, desenvolvedores e a sociedade em geral.
2. Análise Crítica: Os alunos devem realizar uma análise crítica das implicações éticas do uso do ChatGPT em estudos de caso específicos. Eles devem examinar como o algoritmo pode influenciar a disseminação de informações, a privacidade dos usuários e a tomada de decisões éticas. Além disso, devem considerar possíveis vieses algorítmicos, discriminação e questões de responsabilidade.
3. Soluções Responsáveis: Além de identificar os desafios éticos, os alunos devem propor soluções responsáveis e éticas para lidar com esses dilemas. Isso pode incluir sugestões para políticas, regulamentações ou práticas de design que promovam o uso responsável da inteligência artificial. Eles devem considerar como essas soluções podem equilibrar os interesses de diferentes partes interessadas e promover valores éticos fundamentais, como transparência, justiça e privacidade.
4. Colaboração e Discussão: O trabalho deve envolver discussões em grupo e colaboração entre os alunos. Eles devem compartilhar ideias, debater diferentes pontos de vista e chegar a conclusões informadas através do diálogo e da reflexão mútua. O estudo de caso do ChatGPT pode servir como um ponto de partida para essas discussões, incentivando os alunos a aplicar conceitos éticos e legais aprendidos ao analisar um caso concreto.
5. Limite de Palavras: O trabalho terá um limite de 6 a 10 páginas teria aproximadamente entre 1500 e 3000 palavras.
6. Estruturação Adequada: O trabalho siga uma estrutura adequada, incluindo introdução, desenvolvimento e conclusão. Cada seção deve ocupar uma parte proporcional do total de páginas, com a introdução e a conclusão ocupando menos espaço do que o desenvolvimento.
7. Controle de Informações: Evitar incluir informações desnecessárias que possam aumentar o comprimento do trabalho sem contribuir significativamente para o conteúdo. Concentre-se em informações relevantes, argumentos sólidos e evidências importantes para apoiar sua análise.

8. Síntese e Clareza: O trabalho deverá ser conciso e claro em sua escrita. Evite repetições desnecessárias e redundâncias. Sintetize suas ideias e argumentos de forma eficaz para transmitir suas mensagens de maneira sucinta.

9. Formatação Adequada: O trabalho deverá ser apresentado nas normas da ABNT de acordo com as diretrizes fornecidas, incluindo margens, espaçamento, tamanho da fonte e estilo de citação. Deve-se seguir o seguinte template de arquivo: hfps://bibliotecas.ufpr.br/wp-content/uploads/2022/03/template-artigo-de-periodico.docx

B – RESOLUÇÃO

ESTUDO DE CASO: IMPLICAÇÕES ÉTICAS DO USO DA IA E DO CHATGPT

RESUMO

Uma discussão sobre a evolução e impacto da inteligência artificial (IA), destacando questões éticas, de privacidade e transparência. Abordar dilemas éticos como viés algorítmico, vigilância, responsabilidade e exclusão social. Exemplifica como tecnologias como o ChatGPT, podem reproduzir preconceitos e desinformação. A necessidade de regulamentação, transparência e educação pública é enfatizada para mitigar riscos e promover um uso justo e ético da IA. O impacto da IA em empregos e seu uso em contextos militares e de controle social.

Palavras-chave: inteligência artificial; ética, privacidade; transparência; viés algorítmico; desinformação; chatGPT; regulamentação; inclusão social; empregos; contexto militar.

ABSTRACT

Discussion about the evolution of artificial intelligence (AI), focusing on ethical issues such as algorithmic bias, privacy, surveillance, transparency, responsibility, and social exclusion. Using ChatGPT as an example, it highlights the potential for AI to reproduce biases and disseminate misinformation. The need for regulation, transparency, and public education is emphasized to mitigate risks and promote ethical AI use. The impact of AI on jobs and its use in military and social control contexts is also discussed.

Keywords: artificial intelligence; ethics; privacy; transparency; algorithmic bias; misinformation; chatGPT; regulation; social inclusion; jobs; military context.

INTRODUÇÃO

Uma das grandes inovações da humanidade, que mais vem impactando a mesma nos dias atuais, é a inteligência artificial (IA). Essa tecnologia nem é tão nova assim, teve seu início por volta da década de 50, porém é a partir do advento de uma IA mais avançada, com um marco significativo no ano de 2023, com o surgimento das IA generativas, sobretudo do ChatGPT, é que se tem notado o quanto a IA impacta de forma profunda a sociedade como um todo.

A crescente adoção da inteligência artificial tem provocado questões como responsabilidade ética, privacidade, transparência, justiça e isonomia dos usuários e de todos aqueles que são de certa forma impactados direta ou indiretamente pelas IAs.

É muito importante que sejam elaboradas discussões que resultem em um consenso evolutivo de como a IA deve evoluir e funcionar para as partes interessadas de maneira ética, responsável, transparente e justa para todas as pessoas sem nenhuma forma de discriminação.

A ÉTICA

Na prática, a ética é um conjunto de normas e valores morais coletivos que orientam as relações interpessoais dentro de uma sociedade ou grupo de pessoas de forma a assegurar a igualdade, a justiça e o bem-estar social.

A palavra ética vem do grego éthos que significa costume. Esses costumes ou valores não são imutáveis e nem possuem um consenso definitivo e mudam ao longo do tempo e espaço.

INTELIGÊNCIA ARTIFICIAL

A IA pode ser entendida como um campo da ciência da computação que visa a construção de máquinas capazes de pensar e agir de forma semelhante aos humanos. Ou melhor, é uma tecnologia capaz de aprender e tomar decisões semelhantes aos humanos.

Tecnologias como o ChatGPT vem gerando fortes impactos na sociedade e muitas discussões sobre as questões éticas e filosóficas envolvidas nessas tecnologias.

Questões como privacidade dos dados, vigilância, transparência, justiça, equidade, responsabilidade e etc.

DESENVOLVIMENTO

Neste ponto do trabalho, será discutido quais os dilemas éticos do uso das IAs e como eles afetam as diferentes partes interessadas, usuários e desenvolvedores em geral; quais as implicações éticas do uso das IAs, usando o ChatGPT como caso de uso; como os algoritmos de IA podem influenciar a disseminação de informações e nas tomadas de decisões éticas.

DILEMAS ÉTICOS

A criação, o uso e o controle das tecnologias de IAs são os três principais pontos dos quais derivam as questões mais complicadas e complexas prático-filosóficas relativas ao uso dessas tecnologias. As IAs podem trazer muitos benefícios para a sociedade, porém também podem trazer problemas de ordem ética e moral se não forem desenvolvidas com compromisso ético e cívico. Não é meramente uma questão de avaliar a IA em si, mas sobretudo de considerar os humanos conectados a ela, tanto na programação quanto na área judicial.

Algumas questões a considerar são a privacidade, a justiça, a responsabilidade, a autonomia e a transparência.

O primeiro dilema ético relacionado ao uso das IAs é o viés algorítmico, que é

quando o algoritmo aprende e replica preconceitos e discriminações pré-existentes na sociedade, sobretudo quando o algorítmo, como o GPT do ChatGPT, é treinado usando base de dados que possuem historicamente vieses de preconceito e discriminação, perpetuando-os.

O ChatGPT pode reproduzir preconceitos e discriminações. A própria OpenAI, a criadora do ChatGPT, alerta sobre a possibilidade do ChatGPT produzir conteúdo enviesado, produzindo informações tendenciosas ou maliciosas.

Esse tipo de viés pode promover decisões injustas e desiguais, o que é uma preocupação para áreas como na justiça e no processo de recrutamento, por exemplo. Podendo levar a exclusão e discriminação de pessoas de diferentes etnias, gêneros, cores, credo entre outros aspectos pessoais e culturais. É necessário pensar e desenvolver técnicas que evitem ou minimizem esses vieses.

Tecnologias de IA capazes de fazer o reconhecimento facial em locais públicos por meio de câmeras de segurança estão cada vez mais perto da realidade das pessoas. Além disso, as IAs podem ser usadas para coletar e processar uma quantidade massiva de dados pessoais, o que pode levar à violação da privacidade das pessoas. A vigilância excessiva pode promover um ambiente autoritário e opressivo, sobretudo em regimes totalitaristas. O reconhecimento facial pode ser impreciso e levar a julgamentos precipitados e a condenações injustas.

Essas tecnologias até podem ajudar a prevenir e combater o crime, melhorando a segurança, mas levantam preocupações relativas à privacidade dos cidadãos e ao uso indevido dos dados coletados dos mesmos. É necessário considerar os riscos e benefícios no uso das IAs na segurança e encontrar soluções que garantam a segurança pública sem comprometer a privacidade e os direitos dos indivíduos.

A OpenAI diz que se compromete a proteger e a garantir a privacidade dos dados dos usuários do ChatGPT. A mesma adota algumas medidas para salvaguardar os dados dos seus usuários, como encriptação de dados; autenticação e monitoramento.

As chamadas IAs fortes e generativas são aquelas que não precisam de intervenção humana para tomarem decisões. São, desse modo, também conhecidas como IAs autónomas. Com isso, surge a questão de quem é a responsabilidade das decisões tomadas por essas tecnologias. Por exemplo, se o ChatGPT cometer plágio na geração de um texto, quem deve responder civilmente por esse crime: a OpenAI, o usuário que solicitou o texto ou a própria tecnologia ? Essas questões exigem uma abordagem ética e cuidadosa por parte dos desenvolvedores de tecnologias autônomas, e o ChatGPT precisa proteger os direitos autorais de obras.

O ser humano tem medo e desconfiança daquilo que não conhece muito bem. Da mesma forma, algoritmos de IAs que não são claros no seu funcionamento podem levar à desconfiança dos usuários. Os sistemas de IA são geralmente complexos e difíceis de entender, e isso gera desconfiança e incertezas nos

usuários. A transparência e a explicabilidade são essenciais para garantir a confiança e aceitação do público.

Por exemplo, um algoritmo de aprendizado em máquina construído para decidir quando conceder ou não empréstimos para determinadas pessoas, e esse sistema nega um pedido de empréstimo para uma determinada pessoa e não deixa claro para o mesma o motivo da recusa, gerando desconfiança e incertezas para com essa tecnologia. Por outro lado, se a entidade responsável pelo software exibe todos os parâmetros desse algoritmo pode facilitar fraudes. Disso vem o dilema ético da empresa: como equilibrar clareza com a garantia de proteção e privacidade dos dados, equidade e tratamento justo pela tecnologia.

É necessário construir softwares de IA com transparência, mais compreensíveis e explicáveis, mas que assegurem a privacidade e segurança dos usuários.

As IAs podem gerar exclusão social se não for utilizada de maneira justa e equitativa. Ao favorecer alguns grupos em detrimento de outros, as IAs podem agravar ou mesmo gerar desigualdades sociais. Por exemplo, o ChatGPT pode ser usado para realizar seleção de candidatos a empregos. Os dados históricos de contratação usado para treinar o chatGPT podem possuir vieses, que refletem a discriminação e exclusão de grupo sociais específicos, como mulheres, pessoas de cor, deficientes e etc, causando a replicação e perpetuação dessas discriminações e exclusões.

Por outro lado, se o ChatGPT for treinado para dar preferência por grupos historicamente marginalizados e discriminados, ele pode fazer uma discriminação inversa, excluíndo pessoas que não pertencem a esses grupos.

O desafio ético está em equilibrar a necessidade de corrigir, no algoritmo, as desigualdades históricas e assegurar a diversidade e a inclusão, com a necessidade evitar a exclusão de pessoas que não pertencem aos grupos favorecidos pelos ajustes nos algoritmos.

Um dos maiores medos dos humanos quanto às novas tecnologias é a possibilidade das IAs tomarem os empregos hoje realizados por humanos. De fato, as IAs, como o ChatGPT, podem substituir empregos que antes eram realizados apenas por humanos. Por exemplo, o ChatGPT pode ser utilizado para escrever resenhas para jornais ou revistas eletrônicas, tornando o redator uma posição desnecessária para uma empresa do setor jornalístico e de entretenimento.

Embora os chamados robôs possam aumentar a eficiência e a produtividade, aumentando os lucros das empresas, eles também podem causar o aumento do desemprego, elevando as desigualdades sociais e econômicas, resultando em problemas sociais e políticos.

Nesse sentido, há o dilema de equilibrar os benefícios potenciais das tecnologias de IA com as consequências negativas para os trabalhadores e para a sociedade em geral.

Um uso muito polêmico para IAs é o uso para fins militares. Por exemplo, embora drones autônomos possam fazer ataques cirúrgicos em pontos específicos com precisão, reduzindo efeitos colaterais, eles não possuem supervisão humana, desse modo, se um drone autônomo cometer um erro de cálculo, por exemplo, torna-se difícil de determinar de quem a responsabilidade pelo erro, isso reduz a responsabilidade dos militares o que pode estimular conflitos armados.

O ChatGPT vem entrando no campo militar também, sendo usado na análise massiva de grande quantidade de dados, fornecendo informações úteis para determinar estratégias militares; no treinamento de soldados em ambientes virtuais simulando diferentes situações de emergência ou combate; na tradução instantânea de comunicações multinacionais; na criação e disseminação de desinformação ou propaganda falsas; entre outros casos.

É necessário discutir limites éticos no uso da IA em contextos militares, com o objetivo de rastrear e atribuir corretamente as responsabilidades das consequências do uso da IA nesses contextos.

Outro fim polêmico, e considerado por muitos como indevido, de IAs é usá-las para controle, monitoramento e dominação das pessoas, principalmente em regimes autoritários. É importante considerar os riscos de uso indevido da IA e estabelecer normas e regulamentos que reforcem a democracia e a liberdade individual e buscar um equilíbrio entre segurança pública e os direitos humanos e individuais da sociedade.

DIREITO AUTORAL E PROPRIEDADE INTELECTUAL

Atualmente, o ChatGPT consegue escrever livros infantis, poemas, obras de artes, que inclusive ganham competições de arte, ou produzir artigos acadêmicos

levando a uma série de questões sobre a autoria, a originalidade e a propriedade intelectual, que ainda carecem de respostas.

Há inúmeras maneiras de abordar de forma ética as consequências do uso do ChatGPT no cotidiano global. Por exemplo, as questões éticas do uso do ChatGPT em sala de aulas, ou uso do mesmo para reforço e disseminação do racismo algorítmico e da desinformação.

O ChatGPT vem influenciando a distribuição e o controle do conhecimento, à medida que o algoritmo foi treinado com milhares de dados da internet considerados de domínio público. Desse modo, o ChatGPT é capaz de, por meio de obras de outros autores, produzir artes cujas autorias são temas de discussão, afinal de quem é tais obras produzidas pelo ChatGPT a partir de dados históricos contendo obras de autores humanos ? Essas e outras questões são importantes de serem levantadas para determinar uma forma justa, ética e responsável de usar as IAs gerando benefícios para a sociedade em geral.

INFLUÊNCIA DOS ALGORITMOS NA DISSEMINAÇÃO DE INFORMAÇÕES

Muitos pesquisadores estão muito preocupados com o avanço dos recursos de geração de linguagem avançados como as IAs generativas, pois acreditam que esses recursos podem se tornar uma fonte significativa de desinformação. O medo é que esses recursos poderiam criar e espalhar fake news altamente críveis rapidamente. Os pesquisadores enfatizam a importância de desenvolver estratégias para mitigar esses riscos, incluindo melhores métodos de detecção de desinformação e normas éticas para o uso das IAs.

Esse problema específico é de duas ordens: primeiro a vulnerabilidade das pessoas de serem expostas a fake news e informações falsas; e em segundo quando IAs, como ChatGPT, acabam parando em mãos erradas, em pessoas mal intencionadas que usam as IAs para produzir desinformação.

Portanto, é preciso instruir as pessoas sobre o que é desinformação e quais os riscos da mesma para a sociedade. Por exemplo, a Finlândia já introduziu uma disciplina nas escolas sobre como identificar e combater a desinformação. Uma outra forma de abordar é criar técnicas para identificar quando um usuário está tentando fazer mal uso da IA.

RESPONSABILIDADE DA IA

É notória a rápida evolução das tecnologias de IAs, como o ChatGPT, e como essas tecnologias vêm impactando na sociedade trazendo avanços nos mais variados setores da economia e da sociedade. Porém, com os avanços, trazem inúmeros desafios que precisam ser discutidos e revistos. À medida que o uso das IAs tornam-se mais comuns, surge o desafio de determinar de quem é a responsabilidade por danos oriundos do uso das IAs.

Responsabilidade, em um sentido mais amplo, é a obrigação legal de assumir e prestar contas por ações, decisões ou impactos resultantes de uma atividade específica. Responsabilizar uma IA é um desafio devido a sua natureza autônoma e complexa dos softwares, esses softwares não precisam de intervenção humana para tomar decisões. Desse modo a responsabilidade pode recair tanto para os desenvolvedores quanto para os usuários finais, que precisam compreender e administrar os riscos associados ao uso desses softwares.

Por exemplo, o uso indiscriminado e perigoso do ChatGPT envolve, sobretudo, a criação e o espalhamento de informações e notícias falsas; e os problemas de plágios de dados, textos e obras artísticas em geral.

Os termos de uso do ChatGPT da OpenAI é deixar claro a responsabilidade dos usuários ao usar o ChatGPT: “a utilização da plataforma não deve ser feita para disseminar informações falsas ou gerar informações que possam de qualquer forma prejudicar terceiros ao publicá-las”.

Determinar a responsabilidade das ações de uma AI é complexo devido a sua natureza descentralizada e autônoma, que torna sombreada a linha de responsabilidade.

CONSIDERAÇÕES FINAIS

A inteligência artificial (IA) traz significativos benefícios, mas também levanta questões éticas e de responsabilidade. Uma solução competente envolve implementar padrões éticos rigorosos no desenvolvimento de IA, garantindo transparência, justiça e privacidade; educar o público sobre IA e desinformação, promovendo habilidades de identificação e combate a fake news; criar regulamentações que definam claramente responsabilidades legais para

desenvolvedores e usuários de IA; assegurar que algoritmos sejam treinados para evitar viés e promover a equidade social, sem discriminação inversa.

Dessa forma, é possível maximizar os benefícios da IA, minimizando riscos e impactos negativos.

REFERÊNCIAS

UNICEP. **Inteligência artificial e ética: conheça o impacto ético da IA na sociedade.** UNICEP, 2024. Disponível em: https://www.unicep.edu.br/post/intelig%C3%Aancia-artificial-e-%C3%A9tica-conhe%C3%A7a-o-impacto-%C3%A9tico-da-ia-na-sociedade#:~:text=Em%20resumo%2C%20a%20privacidade%20e,privacidade%20dos%20indiv%C3%ADduos%20%C3%A9%20fundamental. Acesso em: 28 jun. 2024.

UPDATE OR DIE. **Dilemas éticos da IA.** Update or Die, 23 abr. 2024. Disponível em: https://www.updateordie.com/2024/04/23/dilemas-eticos-da-ia/. Acesso em: 28 jun. 2024.

SERPRO. **ChatGPT: o que pode dar errado? Serviço Federal de Processamento de Dados**, 2023. Disponível em: https://www.serpro.gov.br/menu/noticias/noticias-2023/chatgpt-o-que-pode-dar-errado. Acesso em: 28 jun. 2024.

DIDÁTICA TECH. **Segurança e privacidade no uso do ChatGPT**. Didática Tech, 2024. Disponível em: https://didatica.tech/seguranca-e-privacidade-no-uso-do-chatgpt/. Acesso em: 28 jun. 2024.

ESTADÃO. **Como o ChatGPT pode ser usado como ferramenta militar**. Estadão, 2024. Disponível em: https://www.estadao.com.br/link/cultura-digital/como-o-chatgpt-pode-ser-usado-como-ferramenta-militar/. Acesso em: 28 jun. 2024.

OUTRAS PALAVRAS. **ChatGPT e a disputa pelo controle do conhecimento**. Outras Palavras, 2024. Disponível em: https://outraspalavras.net/outrasmidias/chatgpt-e-a-disputa-pelo-controle-do-conhecimento/. Acesso em: 28 jun. 2024.

FOLHA DE S.PAULO. **ChatGPT será maior espalhador de desinformação que já existiu, diz pesquisador**. Folha de S.Paulo, 3 fev. 2023. Disponível em: https://www1.folha.uol.com.br/tec/2023/02/chatgpt-sera-maior-espalhador-de-desinformacao-que-ja-existiu-diz-pesquisador.shtml. Acesso em: 28 jun. 2024.

CLICKSIGN. **Responsabilidade por danos decorrentes de sistemas de IA**. Clicksign, 2024. Disponível em: https://www.clicksign.com/blog/responsabilidade-por-danos-decorrentes-de-sistemas-de-ia. Acesso em: 29 jun. 2024.

APÊNDICE 12 – GESTÃO DE PROJETOS DE IA

A – ENUNCIADO

1 Objetivo

Individualmente, ler e resumir – seguindo o *template* fornecido – **um** dos artigos abaixo:

AHMAD, L.; ABDELRAZEK, M.; ARORA, C.; BANO, M; GRUNDY, J. Requirements practices and gaps when engineering human-centered Artificial Intelligence systems. Applied Soft Computing. 143. 2023. DOI https://doi.org/10.1016/j.asoc.2023.110421

NAZIR, R.; BUCAIONI, A.; PELLICCIONE, P.; Architecting ML-enabled systems: Challenges, best practices, and design decisions. The Journal of Systems & Software. 207. 2024. DOI https://doi.org/10.1016/j.jss.2023.111860

SERBAN, A.; BLOM, K.; HOOS, H.; VISSER, J. Software engineering practices for machine learning – Adoption, effects, and team assessment. The Journal of Systems & Software. 209. 2024. DOI https://doi.org/10.1016/j.jss.2023.111907

STEIDL, M.; FELDERER, M.; RAMLER, R. The pipeline for continuous development of artificial intelligence models – Current state of research and practice. The Journal of Systems & Software. 199. 2023. DOI https://doi.org/10.1016/j.jss.2023.111615

XIN, D.; WU, E. Y.; LEE, D. J.; SALEHI, N.; PARAMESWARAN, A. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. In CHI Conference on Human Factors in Computing Systems (CHI'21), Maio 8-13, 2021, Yokohama, Japão. DOI https://doi.org/10.1145/3411764.3445306

2 Orientações adicionais

Escolha o artigo que for mais interessante para você. Utilize tradutores e o Chat GPT para entender o conteúdo dos artigos – caso precise, mas escreva o resumo em língua portuguesa e nas suas palavras.

Não esqueça de preencher, no trabalho, os campos relativos ao seu nome e ao artigo escolhido.

No *template*, você deverá responder às seguintes questões:

- Qual o objetivo do estudo descrito pelo artigo?
- Qual o problema/oportunidade/situação que levou a necessidade de realização deste estudo?
- Qual a metodologia que os autores usaram para obter e analisar as informações do estudo?
- Quais os principais resultados obtidos pelo estudo?

Responda cada questão utilizando o espaço fornecido no *template*, sem alteração do tamanho da fonte (Times New Roman, 10), nem alteração do espaçamento entre linhas (1.0).

Não altere as questões do template.

Utilize o editor de textos de sua preferência para preencher as respostas, mas entregue o trabalho em PDF.

B – RESOLUÇÃO

Qual o objetivo do estudo descrito pelo artigo?

O objetivo do artigo selecionado está dividido em responder duas buscas: primeiro, mapear quais as práticas, pesquisas e guidelines de engenharia de requisitos voltados para inteligência artificial centrados no ser humano, em seus valores, sentimentos, emoções, culturas e etc, que são atualmente adotados no mercado e na indústria, destacando práticas não centradas em IA mais humanas; e segundo, é identificar as lacunas entre as práticas industriais e as pesquisas acadêmicas sobre requisitos funcionais de IA.

A pesquisa usa como referência guideline de grandes empresas como google, apple, amazon e etc. Para realizar a elaboração de IAs mais voltadas para o lado humano do usuário.

Atualmente a maioria das empresas, ao construírem sistemas de IA, não dão muita importância aos aspectos humanos. A pesquisa busca deixar clara a importância de se atender em primeira mão as necessidades e valores humanos na elaboração de sistemas de IA. Ignorar esses aspectos pode levar a severas consequências danosas tanto físicas quanto mentais.

Qual o problema/oportunidade/situação que levou à necessidade de realização desse estudo?

O estudo foi motivado por desafios na engenharia de requisitos (RE) para sistemas de IA centrados no ser humano, um campo relativamente novo e com poucas práticas consolidadas. Grandes empresas como Google, Microsoft e Apple publicaram guidelines para auxiliar no desenvolvimento de IA centrada no usuário ou

humano, mas a adoção de práticas de requisitos alinhadas a esses princípios é pouco compreendida e aplicada na indústria. O objetivo é abordar as lacunas entre guidelines da indústria e práticas acadêmicas para criar sistemas mais alinhados às necessidades e valores humanos.

Qual a metodologia que os autores usaram para obter e analisar as informações do estudo?

Para realizar os autores primeiramente conduziram uma revisão sistemática da literatura (SLR) para mapear pesquisas e guidelines em engenharia de requisitos para IA centrada no ser humano. Em seguida, desenvolveram uma pesquisa direcionada a profissionais da indústria para identificar quais práticas frequentes e ferramentas de levantamento e documentação de requisitos estão sendo usadas e quais são as lacunas entre essas diretrizes e o que realmente é praticado no mercado. Os resultados da pesquisa foram analisados para destacar as principais abordagens centradas no ser humano que devem ser integradas à requisitos funcionais para IA.

Quais os principais resultados obtidos pelo estudo?

Identificação de lacunas entre as práticas atuais de engenharia de requisitos (RE) para IA centrada no ser humano e as guidelines da indústria, destacando que a maioria das ferramentas e práticas atuais de engenharia de requisitos não são adequadas para sistemas de IA.

Foi verificado que os profissionais utilizam principalmente ferramentas como UML e Microsoft Office para representar e documentar requisitos funcionais, o que, segundo o artigo, poderia limitar a qualidade dos requisitos para IA.

Identificação de seis áreas essenciais para a prática de engenharia de requisitos centradas no humano, incluindo necessidades do usuário, modelo, dados, controle do usuário, explicabilidade e gerenciamento de erros, com sugestões para melhorias nesses aspectos.

APÊNDICE 13 – FRAMEWORKS DE INTELIGÊNCIA ARTIFICIAL

A – ENUNCIADO

1 Classificação (RNA)

Implementar o exemplo de Classificação usando a base de dados Fashion MNIST e a arquitetura RNA vista na aula **FRA - Aula 10 - 2.4 Resolução de exercício de RNA - Classificação**. Além disso, fazer uma breve explicação dos seguintes resultados:

- Gráficos de perda e de acurácia;
- Imagem gerada na seção "**Mostrar algumas classificações erradas"**, apresentada na aula prática.
  Informações:

- **Base de dados:** Fashion MNIST Dataset
- **Descrição**: Um dataset de imagens de roupas, onde o objetivo é classificar o tipo de vestuário. É semelhante ao famoso dataset MNIST, mas com peças de vestuário em vez de dígitos.
- **Tamanho**: 70.000 amostras, 784 features (28x28 pixels).
- **Importação do dataset**: Copiar código abaixo.

```
data = tf.keras.datasets.fashion_mnist
(x_train, y_train), (x_test, y_test) = fashion_mnist.load_data()
```

2 Regressão (RNA)

Implementar o exemplo de Classificação usando a base de dados Wine Dataset e a arquitetura RNA vista na aula **FRA - Aula 12 - 2.5 Resolução de exercício de RNA - Regressão**. Além disso, fazer uma breve explicação dos seguintes resultados:

- Gráficos de avaliação do modelo (loss);
- Métricas de avaliação do modelo (pelo menos uma entre MAE, MSE, R^2).

Informações:

- **Base de dados:** Wine Quality
- **Descrição**: O objetivo deste dataset prever a qualidade dos vinhos com base em suas características químicas. A variável target (y) neste exemplo será o score de qualidade do vinho, que varia de 0 (pior qualidade) a 10 (melhor qualidade)
- **Tamanho**: 1599 amostras, 12 features.
- **Importação**: Copiar código abaixo.

```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv"
data = pd.read_csv(url, delimiter=';')
```

Dica 1. Para facilitar o trabalho, renomeie o nome das colunas para português, dessa forma:

```
data.columns = [
    'acidez_fixa',            # fixed acidity
    'acidez_volatil',         # volatile acidity
    'acido_citrico',          # citric acid
    'acucar_residual',        # residual sugar
    'cloretos',               # chlorides
    'dioxido_de_enxofre_livre', # free sulfur dioxide
    'dioxido_de_enxofre_total', # total sulfur dioxide
    'densidade',              # density
    'pH',                     # pH
    'sulfatos',               # sulphates
    'alcool',                 # alcohol
    'score_qualidade_vinho'              # quality
]
```

Dica 2. Separe os dados (x e y) de tal forma que a última coluna (índice -1), chamada score_qualidade_vinho, seja a variável target (y)

3 Sistemas de Recomendação

Implementar o exemplo de Sistemas de Recomendação usando a base de dados Base_livos.csv e a arquitetura vista na aula **FRA - Aula 22 - 4.3 Resolução do Exercício de Sistemas de Recomendação**. Além disso, fazer uma breve explicação dos seguintes resultados:

- Gráficos de avaliação do modelo (loss);
- Exemplo de recomendação de livro para determinado Usuário.

Informações:

- **Base de dados:** Base_livros.csv
- **Descrição**: Esse conjunto de dados contém informações sobre avaliações de livros (Notas), nomes de livros (Titulo), ISBN e identificação do usuário (ID_usuario)
- **Importação:** Base de dados disponível no Moodle (UFPR Virtual), chamada Base_livros (formato .csv).

4 Deepdream

Implementar o exemplo de implementação mínima de Deepdream usando uma imagem de um felino - retirada do site Wikipedia - e a arquitetura Deepdream vista na aula **FRA - Aula 23 - Prática Deepdream**. Além disso, fazer uma breve explicação dos seguintes resultados:

- Imagem onírica obtida por *Main Loop*;
- Imagem onírica obtida ao levar o modelo até uma oitava;
- Diferenças entre imagens oníricas obtidas com *Main Loop* e levando o modelo até a oitava.

Informações:

- **Base de dados:** https://commons.wikimedia.org/wiki/File:Felis_catus-cat_on_snow.jpg
- **Importação da imagem**: Copiar código abaixo.

```
url = "https://commons.wikimedia.org/wiki/Special:FilePath/Felis_catus-cat_on_snow.jpg"
```

```
Dica: Para exibir a imagem utilizando display (display.html) use o link https://commons.wikimedia.org/wiki/File:Felis_catus-cat_on_snow.jpg
```

B – RESOLUÇÃO

INTRODUÇÃO

Os seguintes modelos foram treinados: RNA classificação, RNA regressão, DeepDream e um sistema de recomendações.

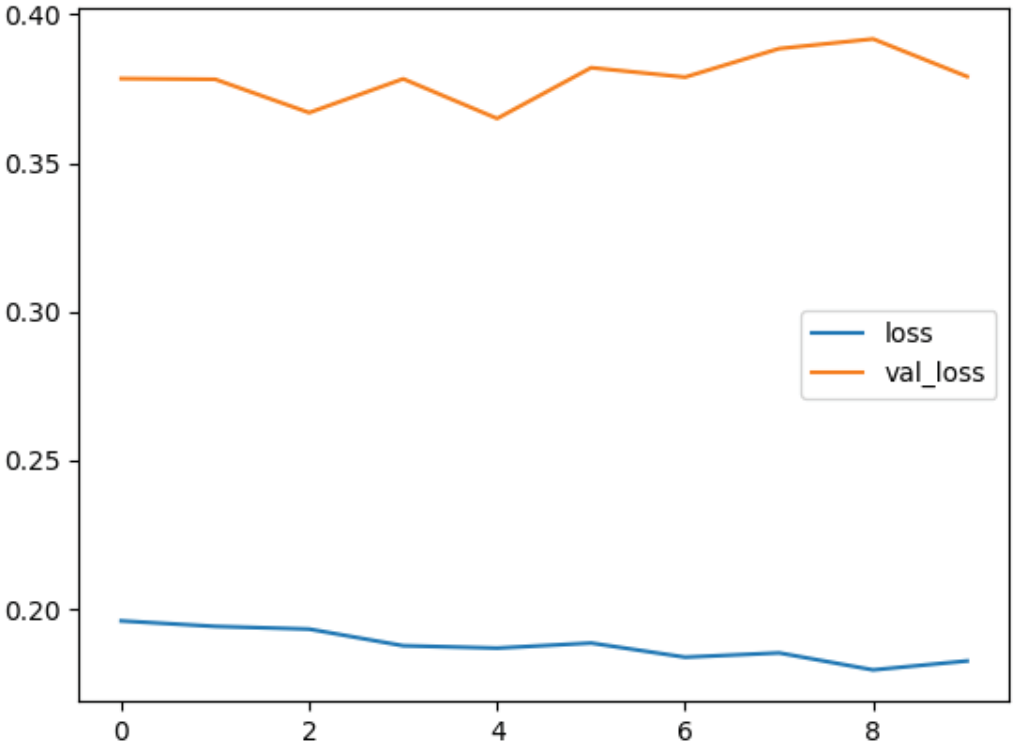
Questão 1 - RNA DE CLASSIFICAÇÃO

A base de dados fashion minist foi utilizada para treinar um modelo RNA de classificação. *Fashion minist* é uma base de dados de imagens de dígitos manuscritos, é um alternativa mais desafiadora à clássica base *minist*.

O objetivo do modelo é tentar classificar corretamente um dígito manuscrito qualquer, de 1 a 9. A seguir os passos executados em linguagem Python para construir, treinar e testar o modelo.

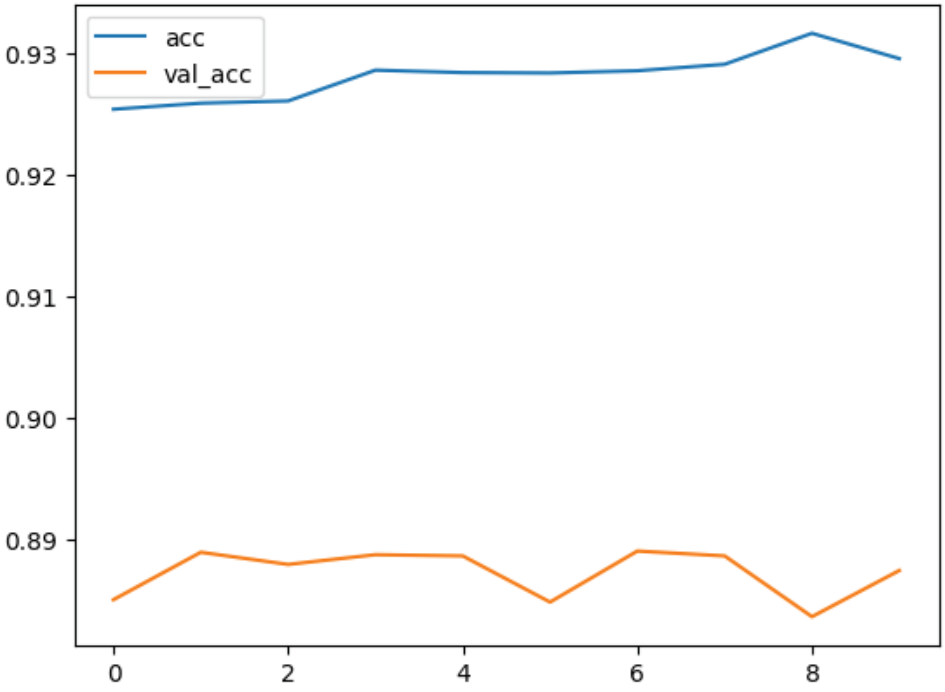
GRÁFICOS DE ACURÁCIA E PERDA

GRÁFICO 16 - EVOLUÇÃO DA FUNÇÃO DE PERDA AO LONGO DO TREINAMENTO



FONTE: O autor (2025)

GRÁFICO 17 -EVOLUÇÃO DA ACURÁCIA DURANTE O TREINAMENTO DO MODELO



FONTE: O autor (2025)

O modelo obtido teve uma acurácia de 88% com relação aos dados preditos e uma perda de 37%.

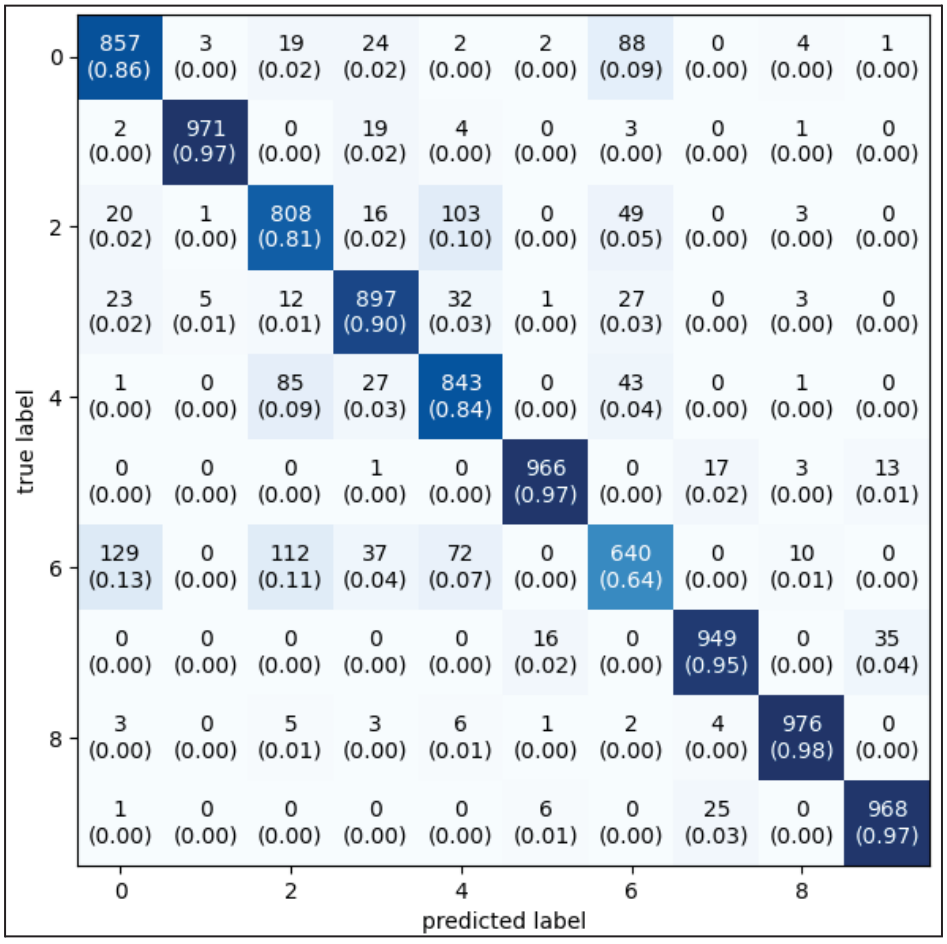
O modelo obtido foi então usado para realizar predições sobre a base de dados de teste.

FIGURA 41 - ALGUNS VALORES PREDITOS

```
[9 2 1 ... 8 1 5]
```

FONTE: O autor (2025)

FIGURA 42 - MATRIZ DE CONFUSÃO



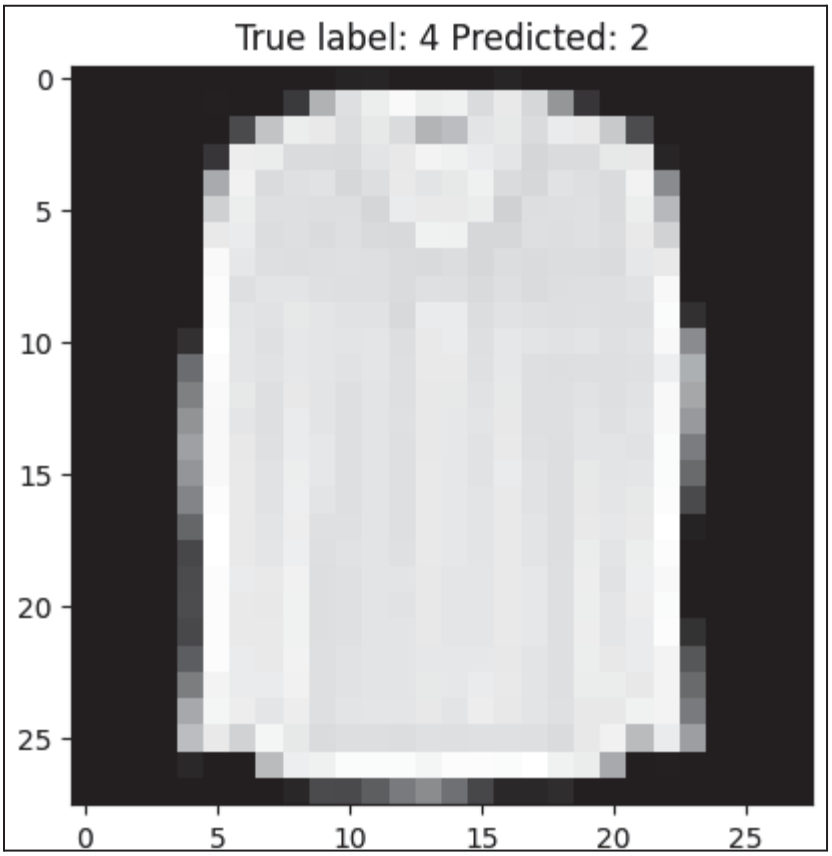
FONTE: O autor (2025)

FIGURA 43 - UM VALOR PREDITO INCORRETAMENTE PELO MODELO

```
Text(0.5, 1.0, 'True label: 4 Predicted: 2')
```

FONTE: O autor (2025)

FIGURA 44 - IMAGEM DE UM DÍGITO PREVISTO INCORRETAMENTE PELO MODELO



FONTE: Autoria própria (2025)

INTERPRETAÇÃO FINAL

Durante as 10 épocas, o modelo continuou a demonstrar alta precisão de treinamento, subindo ligeiramente de aproximadamente 92,6% para 93%, enquanto a precisão de validação permaneceu relativamente estável entre 88,5% e 88,9% aproximadamente. A perda de treinamento diminuiu marginalmente de 0,19 para 0,18, sugerindo uma melhora no ajuste ao conjunto de treinamento. Entretanto, a perda de validação oscilou em torno de 0,36 a 0,39, sem melhorias significativas, indicando que o modelo não conseguiu generalizar melhor.

Questão 2 - RNA DE REGRESSÃO

A base de dados *wine quality*, uma base de dados de vinhos vermelhos, foi utilizada para treinar um modelo RNA de regressão. *Wine quality* é uma base de dados com informações relevantes para determinar a qualidade de vinhos vermelhos.

O objetivo do modelo é tentar determinar o nível de qualidade de vinhos vermelhos com base em propriedades específicas de vinhos. A seguir os passos executados em linguagem Python para construir, treinar e testar o modelo.

A base de dados *wine quality* contém 12 colunas, sendo uma a que contém o nível de qualidade dos vinhos, as colunas são as seguintes:

a) ***fixed acidity*:** ácidos que não evaporam facilmente (como o ácido tartárico);

b) ***volatile acidity*:** Ácidos que evaporam, como o ácido acético (presente no vinagre);

c) ***citric acid*:** natural dos cítricos, adiciona frescor e acidez ao vinho;

d) ***residual sugar*:** Açúcar que sobra após a fermentação;

e) ***chlorides*:** Representam a quantidade de sal (cloreto de sódio, principalmente);

f) ***free sulfur dioxide*:** Representam a quantidade de sal (cloreto de sódio, principalmente);

g) ***total sulfur dioxide***: Soma da forma livre + ligada de SO_2;

h) ***density***: Medida da massa por volume (g/mL);

i) ***pH*:** Escala de acidez (quanto menor, mais ácido);

j) ***sulphates*:** Compostos que contribuem com a preservação e o sabor;

k) ***alcohol*:** Percentual de álcool no vinho;

l) ***quality*:** Avaliação sensorial do vinho feita por especialistas (tipicamente de 0 a 10).

RENOMEAR O NOME DAS COLUNAS PARA PORTUGUÊS

FIGURA 45 - COLUNAS DE DADOS TRADUZIDAS PARA O PORTUGUÊS

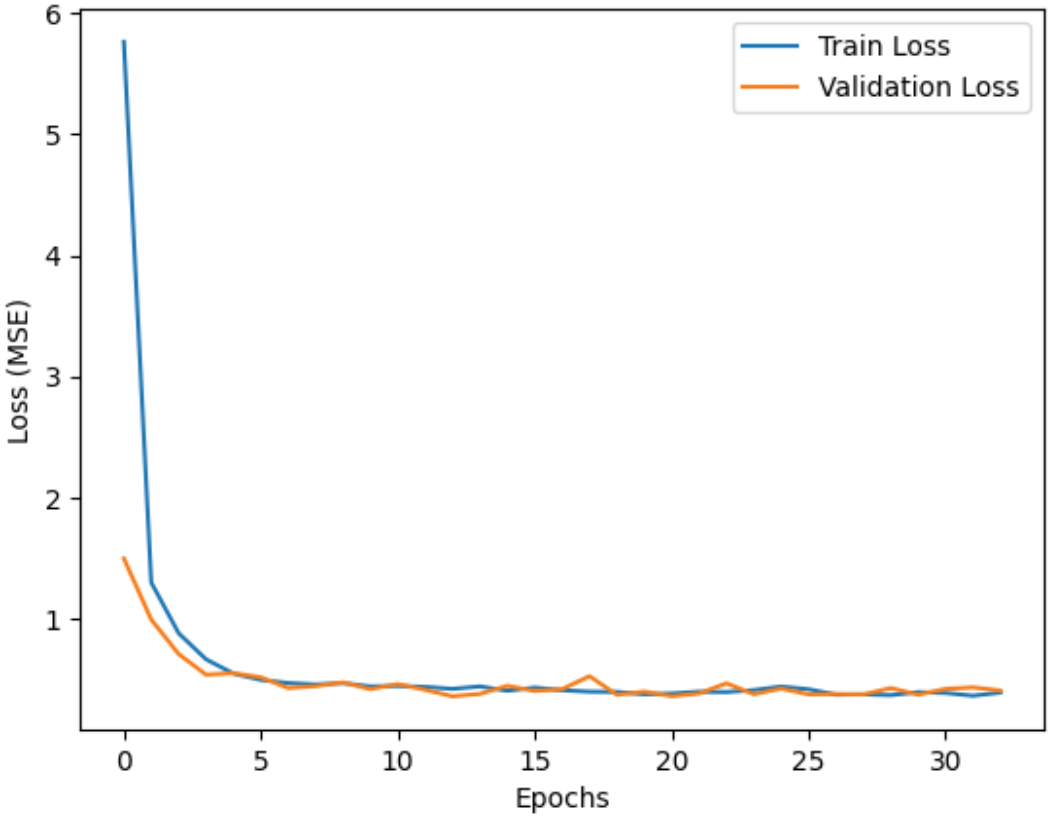
```
data.columns = [
 'acidez_fixa', # fixed acidity
 'acidez_volatil', # volatile acidity
 'acido_citrico', # citric acid
 'acucar_residual', # residual sugar
 'cloretos', # chlorides
 'dioxido_de_enxofre_livre', # free sulfur dioxide
 'dioxido_de_enxofre_total', # total sulfur dioxide
 'densidade', # density
 'pH', # pH
 'sulfatos', # sulphates
```

```
 'alcool', # alcohol
 'score_qualidade_vinho' # quality
]
```

FONTE: O autor (2025)

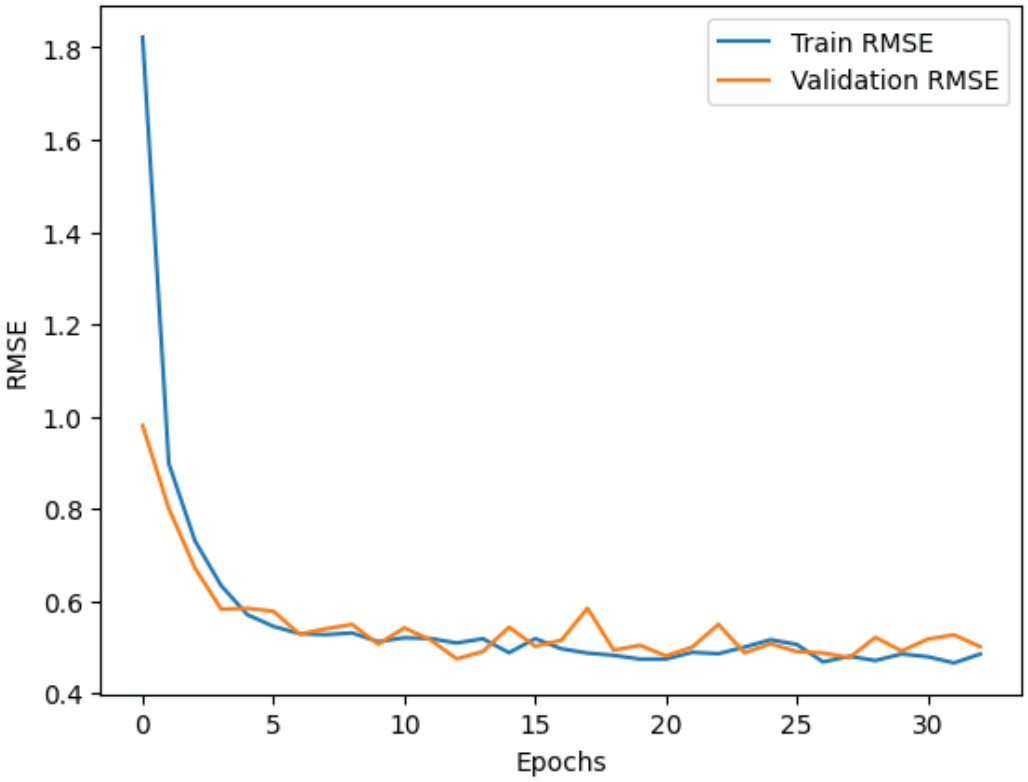
GRÁFICOS DE AVALIAÇÃO DO MODELO (*LOSS*)

GRÁFICO 18 - EVOLUÇÃO DA FUNÇÃO DE PERDA (LOSS) DURANTE O TREINO DO MODELO



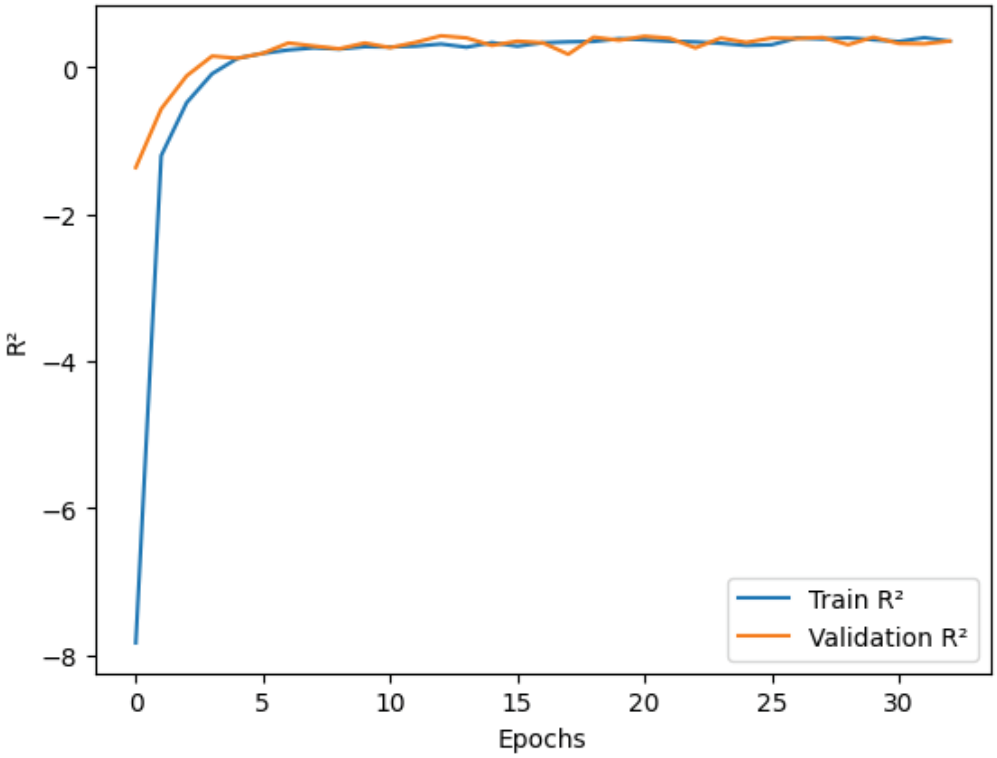
FONTE: O autor (2025)

GRÁFICO 19 - EVOLUÇÃO DO RMSE DURANTE O TREINO DO MODELO



FONTE: O autor (2025)

GRÁFICO 20 - EVOLUÇÃO DE R² DURANTE O TREINO DO MODELO



FONTE: O autor (2025)

Avaliação de desempenho do modelo nos testes

a) **MSE:** 0.36;

b) **RMSE:** 0.60;

c) **R²:** 0.45.

CONCLUSÃO

O modelo começa com uma loss inicial alta (11.2 no treino e 1.49 na validação), acompanhado de métricas R² muito negativas, indicando que o modelo estava significativamente distante de uma boa predição. Nas primeiras epochs, há uma queda acentuada no loss, indicando que o modelo está aprendendo rapidamente. O R² também melhora, embora ainda fique negativo ou apenas ligeiramente positivo no início. Após cerca de 15 epochs, o progresso diminui e as métricas começam a oscilar. O loss de validação fica em torno de 0.36 – 0.43, e o R² na validação estabiliza em valores positivos moderados (~0.30 – 0.40), sugerindo que o modelo está se aproximando do ajuste ótimo para os dados de validação.

Pode ainda ser necessários alguns ajustes para melhorar ainda mais as métricas resultantes do modelo.

Questão 3 - SISTEMA DE RECOMENDAÇÕES

Foi Implementado um Sistemas de Recomendação usando a base de dados Base_livos.csv. Os desempenhos durante o treino do modelo e a etapa de testes foram avaliados usando métricas.

FIGURA 46 - QUANTIDADE DE VALORES NULOS E INCONSISTENTES NA BASE

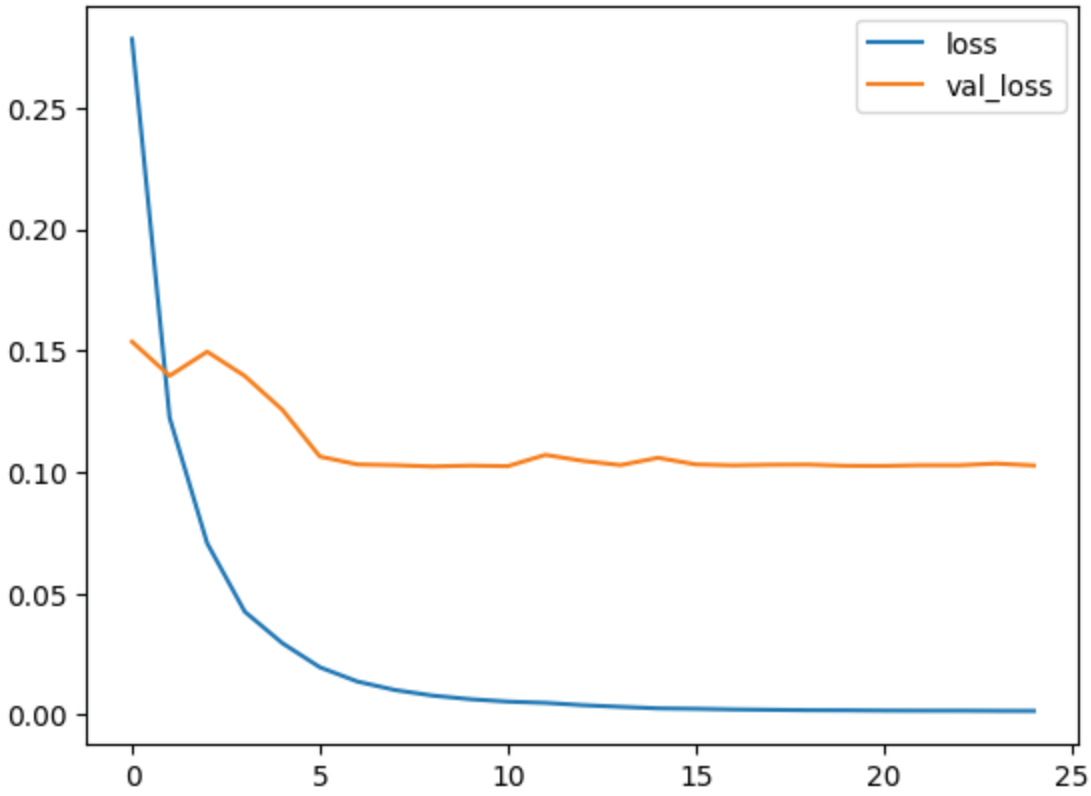
```
(antes) usuário nulos: 20225
(antes) livros nulos: 20190
(antes) autores nulos: 20191
(antes) anos nulos: 20215
(antes) editoras nulas: 20215
(depois) usuário nulos: 0
(depois) livros nulos: 0
(depois) autores nulos: 0
(depois) anos nulos: 0
(depois) editoras nulas: 0
```

FONTE: O autor (2025)

Os valores nulos e inconsistentes detectados na base de dados foram então devidamente tratados.

GRÁFICOS DE AVALIAÇÃO DO MODELO (*LOSS*)

GRÁFICO 21 - EVOLUÇÃO DA FUNÇÃO DE PERDA (LOSS) DURANTE O TREINO DO MODELO



FONTE: O autor (2025)

O valor da função de perda (Loss) na fase de teste do modelo obtido foi de 0.1027035340666771.

Resultado da avaliação de desempenho do modelo nos testes

a) **MAE:** 0.2754;

b) **RMSE:** 0.3203;

c) **R^2:** -0.0396.

QUADRO 14 - ALGUMAS PREDIÇÕES REALIZADAS PELO MODELO OBTIDO

Usuário	Livro	Autor	Nota_real	Nota_prevista
3557	50398	10161	0.4	0.494546
5003	37191	18231	0.2	0.478656
613	75241	23751	0.4	0.583592
8207	88721	20287	0.4	0.552087
1119	2804	19435	0.6	0.612701

FONTE: O autor (2025)

CONCLUSÃO

Os valores de MAE e RMSE são relativamente baixos, o que indica que, em média, o modelo não está errando muito na escala em que as notas foram normalizadas, porém o R^2 negativo é um sinal claro de que o modelo está performando pior do que uma simples média (baseline). Isso quer dizer que ele está tendo dificuldade em aprender um padrão real nos dados.

Questão 4 - DEEPDREAM

Uma base de dados contendo diversas imagens de gatos foi usada para construir e treinar um modelo de IA do tipo DeepDream.

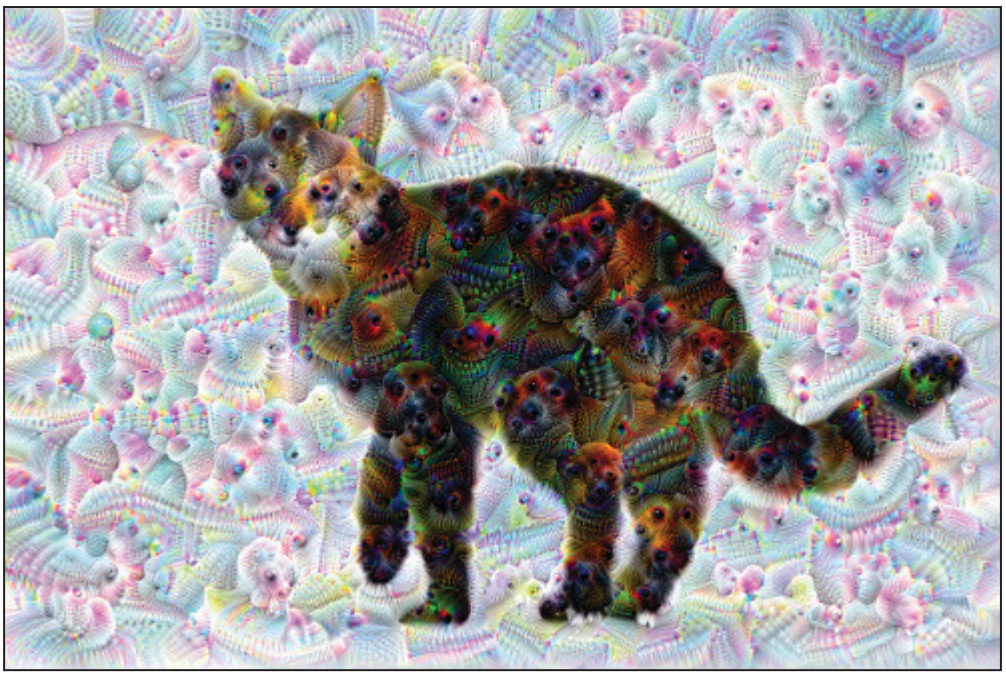
IMAGEM ONÍRICA OBTIDA PELA MAIN LOOP

FIGURA 47 - IMAGEM ONÍRICA OBTIDA POR MAIN LOOP

FONTE: O autor (2025)

IMAGEM ONÍRICA OBTIDA AO LEVAR O MODELO ATÉ UMA OITAVA

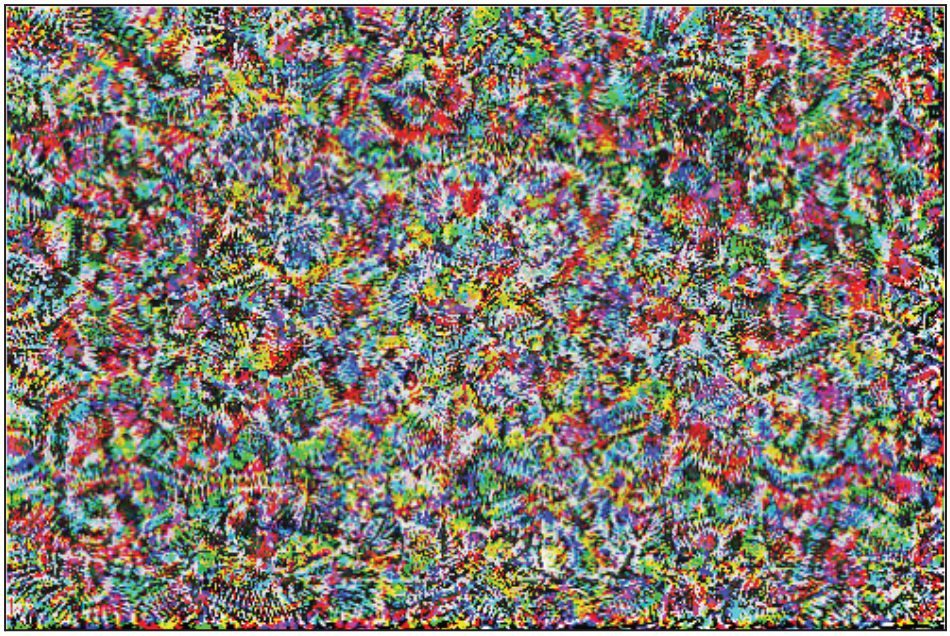
FIGURA 48 - IMAGEM ONÍRICA OBTIDA AO LEVAR O MODELO ATÉ UMA OITAVA



FONTE: O autor (2025)

DIFERENÇAS ENTRE IMAGENS ONÍRICAS OBTIDAS COM MAIN LOOP E ELEVANDO O MODELO ATÉ A OITAVA

FIGURA 49 - DIFERENÇA DIRETA ENTRE PIXELS DAS DUAS IMAGENS



FONTE: O autor (2025)

APÊNDICE 14 – VISUALIZAÇÃO DE DADOS E STORYTELLING

A – ENUNCIADO

Escolha um conjunto de dados brutos (ou uma visualização de dados que você acredite que possa ser melhorada) e faça uma visualização desses dados (de acordo com os dados escolhidos e com a ferramenta de sua escolha)

Desenvolva uma narrativa/storytelling para essa visualização de dados considerando os conceitos e informações que foram discutidas nesta disciplina. Não esqueça de deixar claro para seu possível público alvo qual **o objetivo dessa visualização de dados, o que esses dados significam, quais possíveis ações podem ser feitas com base neles.**

Entregue em um PDF:

- O **conjunto de dados brutos** (**ou uma visualização de dados** que você acredite que possa ser **melhorada**);

- Explicação do **contexto e o publico-alvo** da visualização de dados e do storytelling que será desenvolvido;

- A **visualização desses dados** (de acordo com os dados escolhidos e com a ferramenta de sua escolha) **explicando a escolha do tipo de visualização e da ferramenta usada**; (**50 pontos**)

B – RESOLUÇÃO

O declínio do cinema e a nova era do entretenimento

1. O contexto

Durante décadas, o cinema foi um dos principais meios de entretenimento no mundo. No entanto, com o avanço dos serviços de streaming e a crescente digitalização da sociedade, muitos se perguntam: as pessoas ainda estão indo ao cinema? Se sim, o que as motiva a continuar comprando ingressos?

Para responder a essas perguntas, analisamos os dados de venda de ingressos e comportamento dos clientes em cinemas, buscando padrões que expliquem quem ainda frequenta as salas de exibição e por quê.

2. Público-alvo

Gerentes de cinema, profissionais de marketing, distribuidores de filmes e investidores do setor cinematográfico.

3. Objetivo

Descobrir quais fatores influenciam as vendas e como os cinemas podem melhorar a experiência dos clientes e maximizar seus lucros.

4. Escolha da Visualização e Ferramenta

Para analisar os dados e suas relações, utilizamos um Gráfico de Colunas Empilhadas (Stacked Column Chart), gráfico de rosca, treemaps e gráficos de colunas clusterizadas no Power BI.

A escolha desses gráficos é porque eles facilitam a comparação de categorias; a fácil interpretação dos dados; e visualização de tendências e proporções. Por sua vez, o Power BI foi escolhido devido a:

- Facilidade na manipulação de dados – Permite criar medidas DAX para calcular fidelização;
- Interatividade – O usuário pode filtrar por gênero do filme, faixa etária, entre outros.
- Boa visualização para análise de tendências – Gráficos dinâmicos facilitam a interpretação.

2. Desenvolvimento

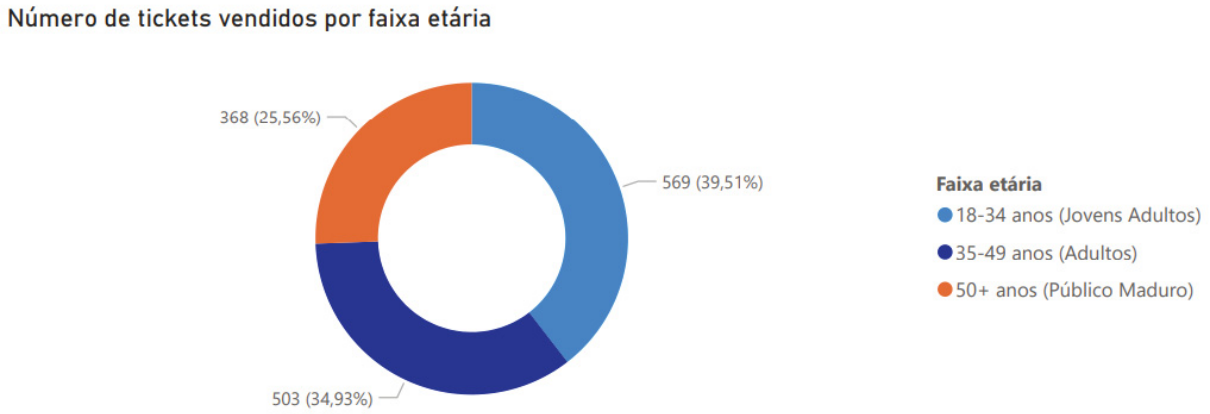
Ao examinar os dados de vendas de ingressos e comportamento do público, identificamos tendências importantes que ajudam a contar a história da atual situação dos cinemas.

2.1 Quem ainda vai ao cinema?

O cinema sempre foi um ponto de encontro: amigos combinam sessões, casais geralmente escolhem filmes românticos e as famílias levam as crianças para entretê-las. Mas os números contam um novo fato.

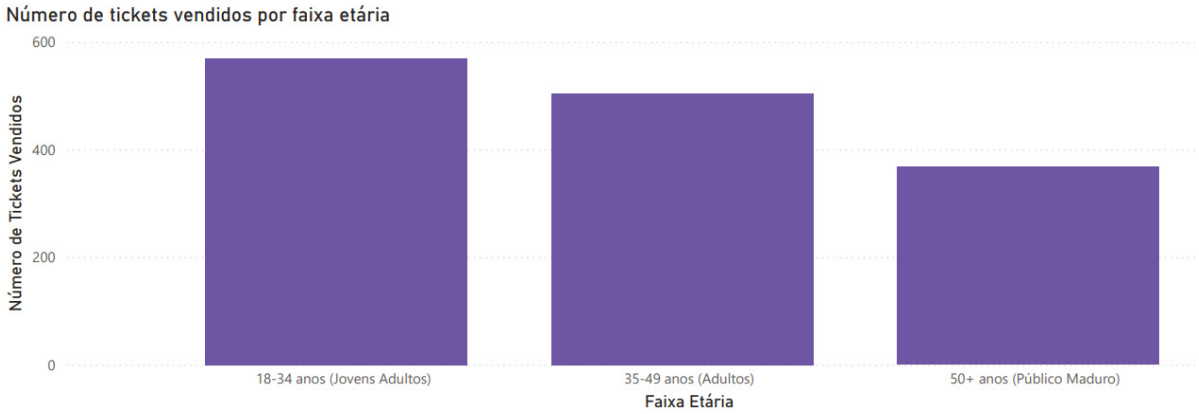
O público jovem adulto (18-34 anos) ainda é o maior frequentador, mas há uma queda significativa a partir dos 35 anos. Vide os gráficos a seguir. O tempo livre se tornou mais escasso? Ou os hábitos de consumo de entretenimento mudaram nessa faixa etária?

GRÁFICO 22 - NÚMERO DE TICKETS VENDIDOS POR FAIXA ETÁRIA



FONTE: O autor (2025)

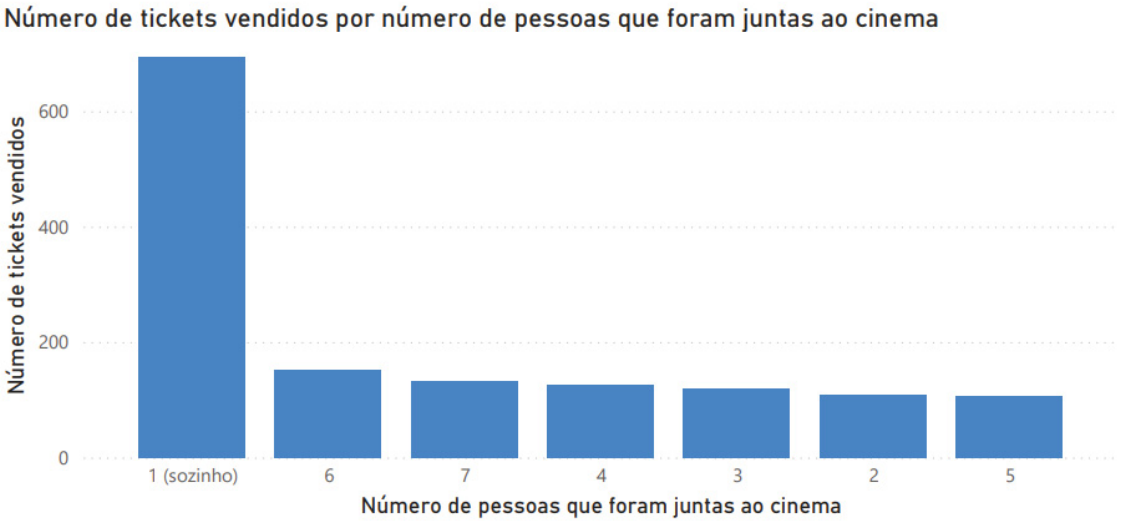
GRÁFICO 23 - NÚMERO DE TICKETS VENDIDOS POR FAIXA ETÁRIA



FONTE: O autor (2025)

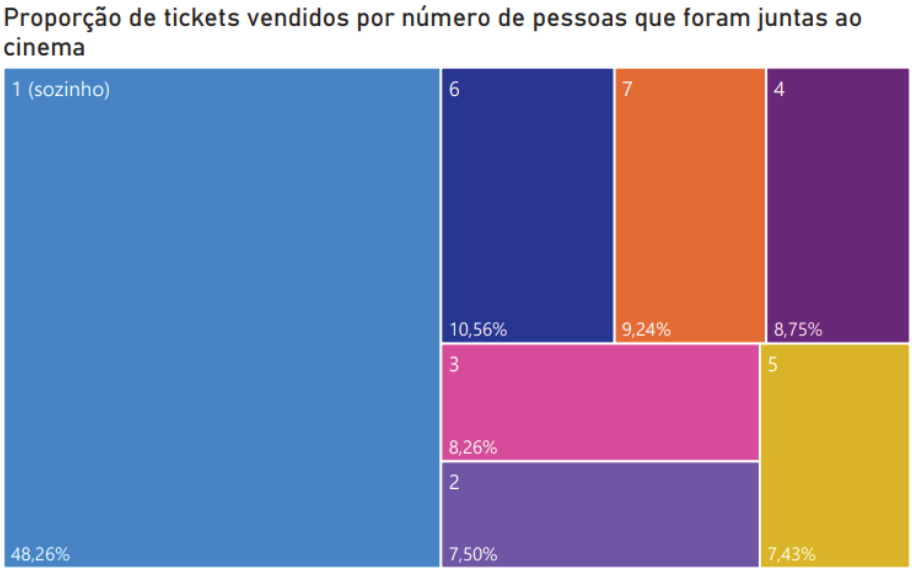
Outro ponto interessante: as pessoas estão indo mais ao cinema sozinhas. Enquanto grupos familiares diminuem, cresce a tendência de espectadores individuais. Talvez seja a conveniência dos streamings tornando as sessões em família menos comuns? Ou será que o cinema se tornou um refúgio pessoal para quem busca uma experiência imersiva? Observe os gráficos a seguir.

GRÁFICO 24 - NÚMERO DE TICKETS VENDIDOS POR NÚMERO DE PESSOAS QUE FORAM JUNTAS AO CINEMA



FONTE: O autor (2025)

GRÁFICO 25 - PROPORÇÃO DE TICKETS VENDIDOS POR NÚMERO DE PESSOAS QUE FORAM JUNTAS AO CINEMA



FONTE: O autor (2025)

Oportunidade para o cinema

a) **ofertas personalizadas:** para atrair o público acima de 35 anos, como sessões especiais, combos ou vantagens exclusivas;

b) **iniciativas para grupos:** promoções para amigos e famílias podem incentivar esse público a voltar;

c) **experiência solo aprimorada:** pacotes que valorizem o conforto e a experiência de quem escolhe assistir sozinho.

O cinema está mudando e os dados mostram que entender o público é essencial. A pergunta agora é: como transformar esses novos hábitos em novas oportunidades?

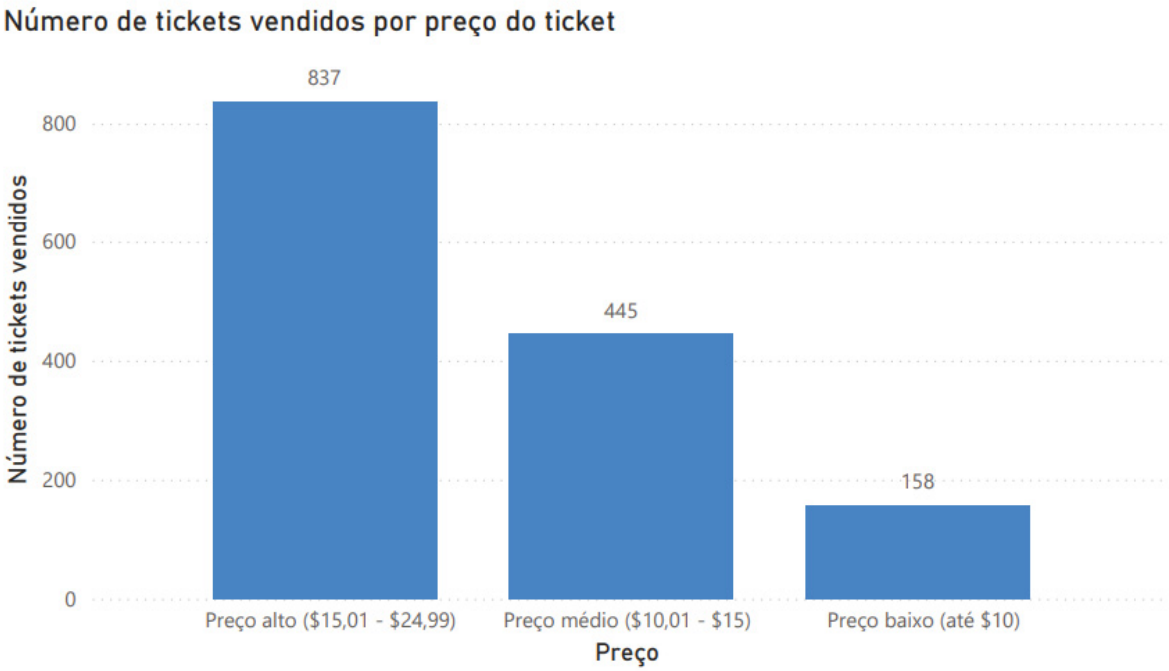
O impacto do preço do ingresso

Quando analisamos a relação entre preço do ingresso e taxa de recompra, vemos que, embora os ingressos mais caros atraiam um público disposto a pagar mais por uma experiência mais premium, a fidelização pode ser um desafio. Cinemas que conseguem equilibrar o preço com a experiência oferecida, seja em qualidade de som, imagem ou conforto, têm uma chance maior de reter seu público.

De acordo com os dados, a taxa dos que voltaria ao cinema dos que não voltariam é praticamente a mesma, mostrando que o preço pode não influenciar muito na fidelização do cliente.

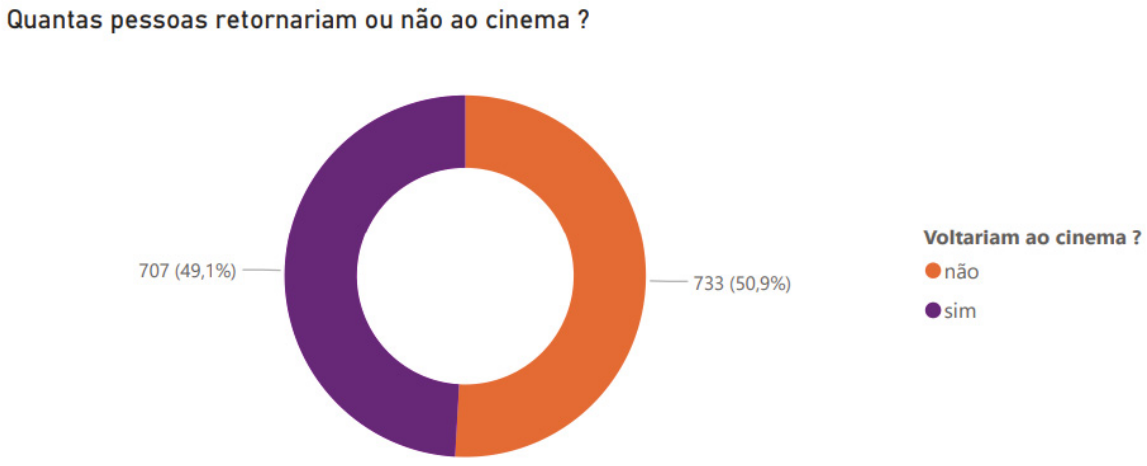
O número de pessoas que compram ingressos mais caros é maior, demonstrando que as pessoas realmente estão dispostas a pagar mais caro por uma experiência melhor, conforme pode ser observado também na proporção entre tipos de assentos escolhidos, onde o mais caro (VIP) teve a maior aderência pelo público, porém o assento do tipo standard ainda é uma escolha preferida entre o cinéfilos. Observe os gráficos a seguir.

GRÁFICO 26 - NÚMERO DE TICKETS VENDIDOS POR PREÇO DE TICKET



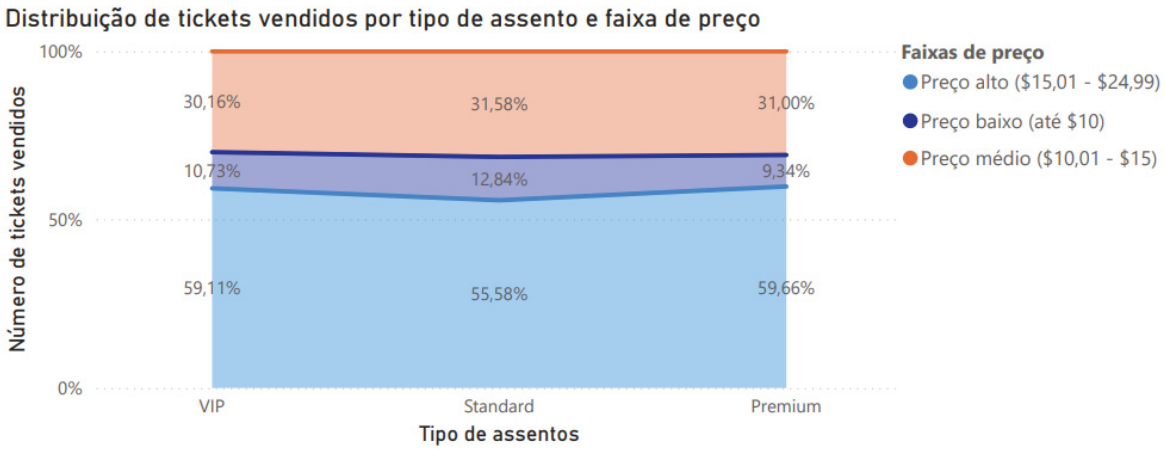
FONTE: O autor (2025)

GRÁFICO 27 - QUANTAS PESSOAS RETORNARIAM AO CINEMA ?



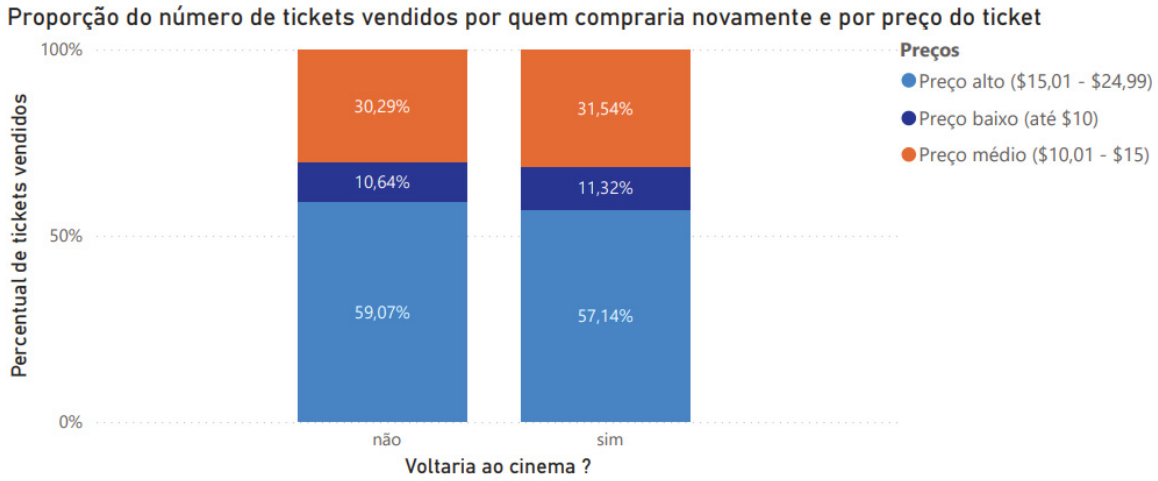
FONTE: O autor (2025)

GRÁFICO 28 - DISTRIBUIÇÃO DE TICKET VENDIDOS POR TIPO DE ASSENTO E FAIXA DE PREÇO



FONTE: O autor (2025)

GRÁFICO 29 - PROPORÇÃO DO NÚMERO DE TICKETS VENDIDOS POR QUEM COMPRARIA NOVAMENTE E POR PREÇO DO TICKET



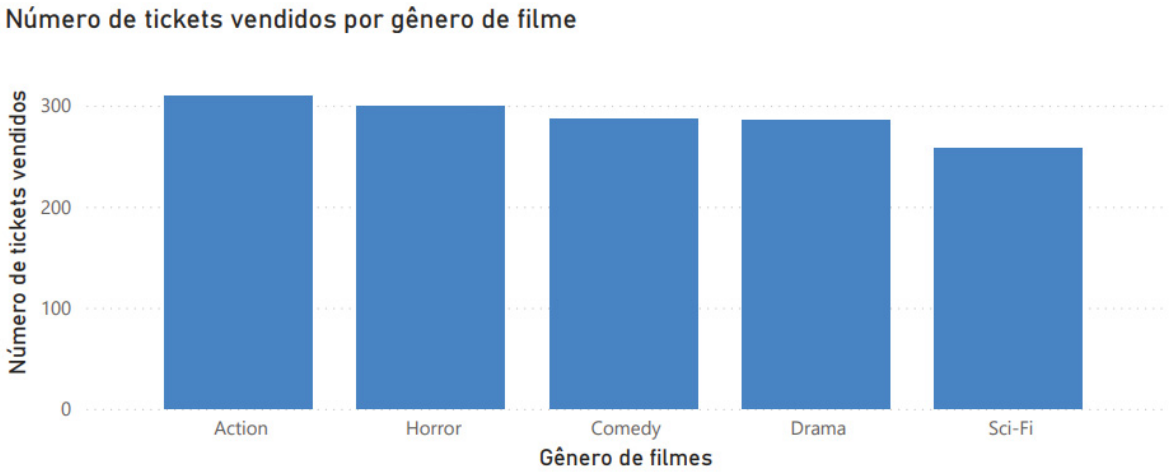
FONTE: O autor (2025)

O gênero de filme mais assistido

Ação e Terror lideram as vendas de ingressos, mostrando que grandes produções cinematográficas com efeitos visuais e sonoros impactantes ainda têm forte apelo no cinema, devido a experiência imersiva dentro das salas de cinema.

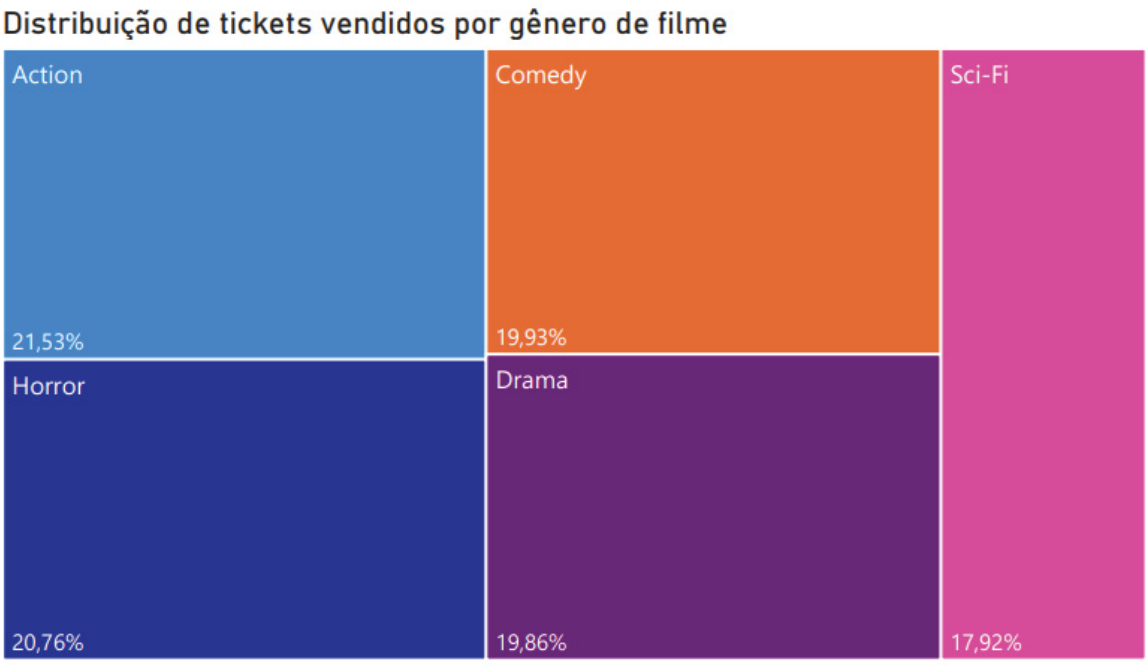
Já ficção científica são menos vistos no cinema, indicando que esse tipo de gênero de filme atrai menos público às telonas do cinema. Observe os gráficos a seguir.

GRÁFICO 30 - NÚMERO DE TICKETS VENDIDOS POR GÊNERO DE FILME



FONTE: O autor (2025)

GRÁFICO 31 - DISTRIBUIÇÃO DE TICKETS VENDIDOS POR GÊNERO DE FILME



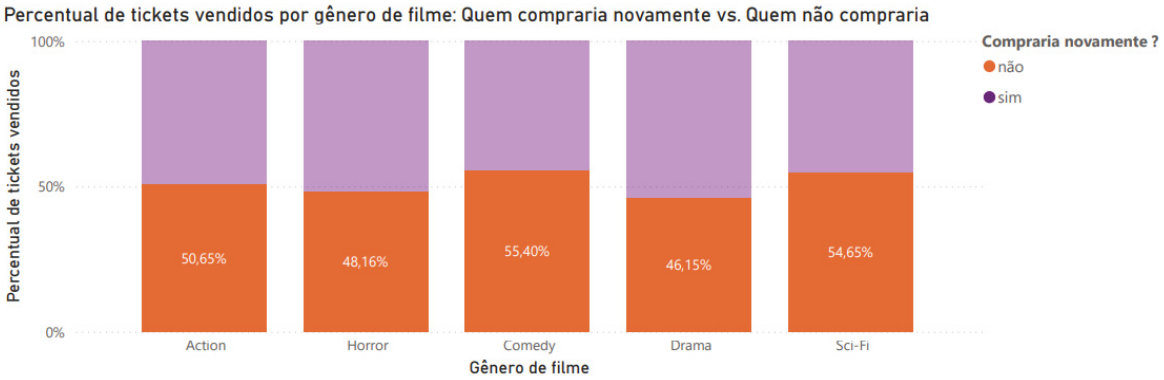
FONTE: O autor (2025)

Influência do gênero na fidelização do cliente

Filmes de ação, terror e drama são os que vendem mais bilhetes e apresentam altas taxas de recompra. Isso ocorre porque as produções de ação e

terror oferecem uma experiência imersiva, repleta de efeitos visuais impactantes e cenas que se beneficiam da tela grande e do som de alta qualidade do cinema, e filmes de drama tendem a fidelizar mais o público porque criam uma forte conexão emocional com os espectadores. Observe os gráficos a seguir.

GRÁFICO 32 - PERCENTUAL DE TICKETS VENDIDOS POR GÊNERO DE FILME: QUEM COMPRARIA NOVAMENTE VS. QUEM NÃO COMPRARIA



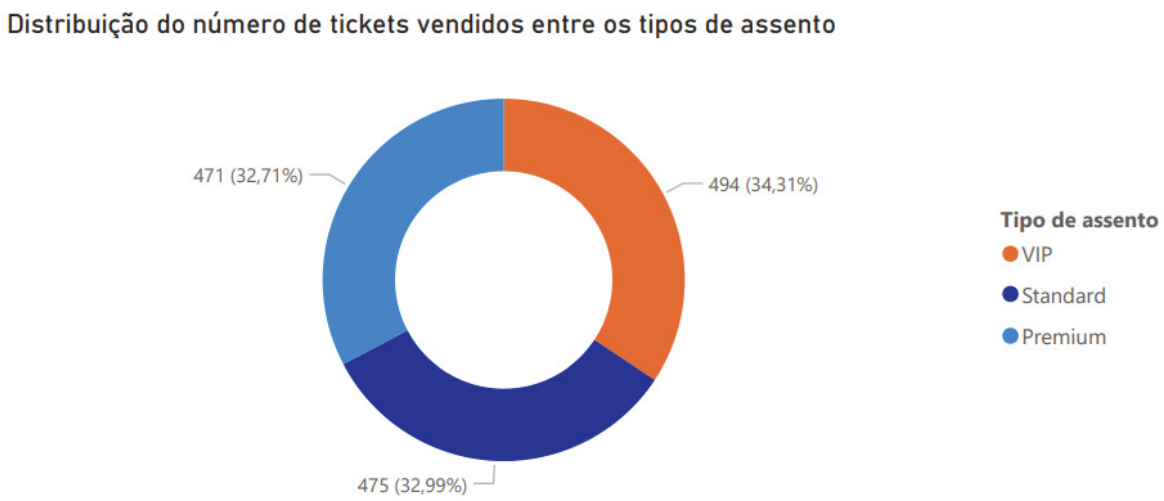
FONTE: O autor (2025)

Tipo de assento preferido

Ao analisarmos a escolha dos tipos de assento pelo público, emergem tendências interessantes que refletem tanto o poder de compra quanto as preferências por uma experiência mais exclusiva. Observamos que os assentos VIP estão em alta, especialmente entre os espectadores que buscam mais conforto e um atendimento diferenciado. Esses assentos atraem principalmente um público disposto a pagar mais por uma experiência premium, o que reflete a crescente busca por valor agregado e exclusividade.

Os dados também revelam que os assentos Standard são a segunda opção mais popular entre os frequentadores do cinema, ficando atrás apenas dos assentos VIP. Esse comportamento sugere um equilíbrio entre conforto e custo-benefício, influenciado por diferentes perfis de espectadores. Observe os gráficos a seguir.

GRÁFICO 33 - DISTRIBUIÇÃO DO NÚMERO DE TICKETS VENDIDOS ENTRE OS TIPOS DE ASSENTOS



FONTE: O autor (2025)

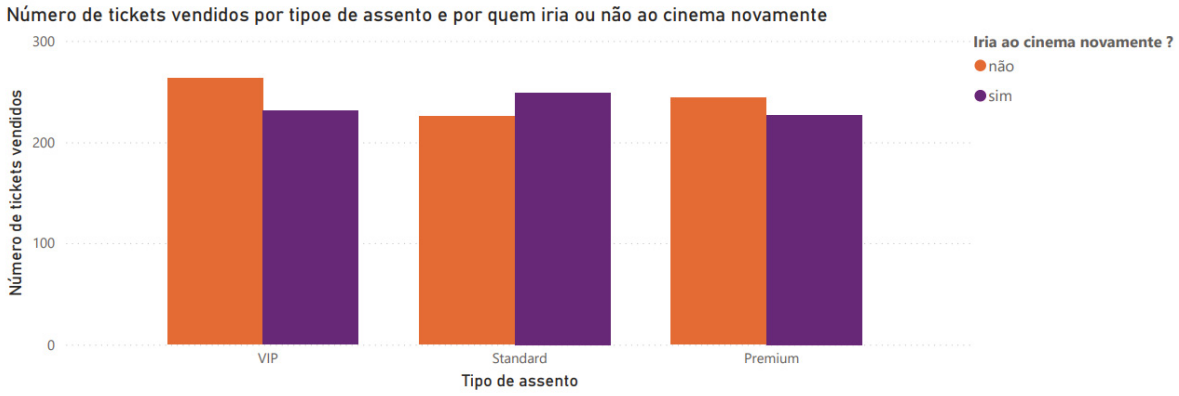
Influência do tipo de assento e a fidelização do cliente

Os dados mostram que o tipo de assento escolhido influencia diretamente a fidelização do cliente. Quanto mais confortável e premium for a experiência, maior a chance de recompra do ingresso. Mas será que apenas o luxo determina a lealdade do público?

Clientes que escolhem assentos VIP ou Premium demonstram maior propensão a comprar ingressos novamente. Isso indica que o conforto e a exclusividade são fatores-chave na experiência do espectador. Para muitos, o cinema é um evento especial, e a experiência imersiva pode incentivar o retorno.

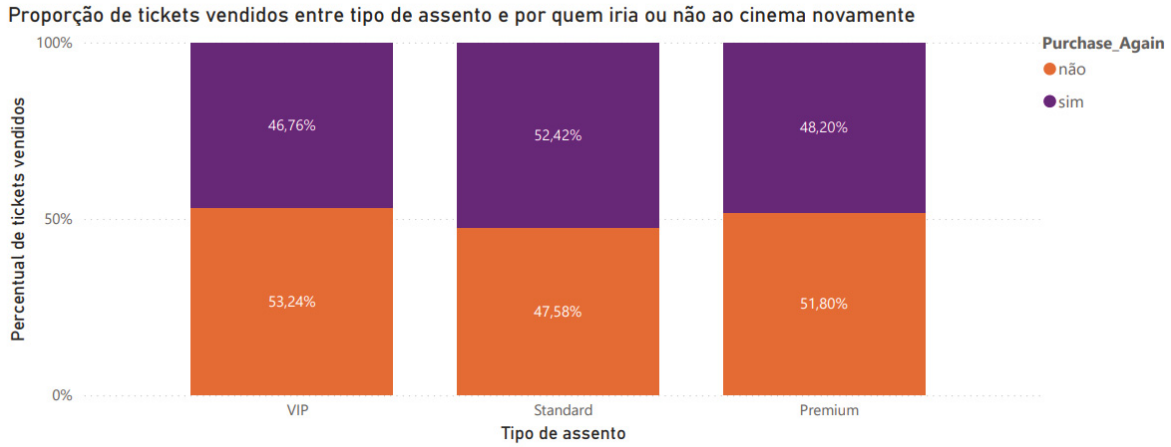
Apesar de serem a segunda opção mais popular, os assentos Standard também apresentam um bom nível de fidelização. O público que opta por esse tipo de assento pode ser mais sensível ao preço, mas ainda assim vê valor na experiência do cinema. Observe os gráficos a seguir.

GRÁFICO 34 - NÚMERO DE TICKETS VENDIDOS POR TIPOS DE ASSENTOS E POR QUEM IRIA AO CINEMA NOVAMENTE



FONTE: O autor (2025)

GRÁFICO 35 - PROPORÇÃO DE TICKETS VENDIDOS ENTRE TIPO DE ASSENTO E POR QUEM IRIA AO CINEMA NOVAMENTE



FONTE: O autor (2025)

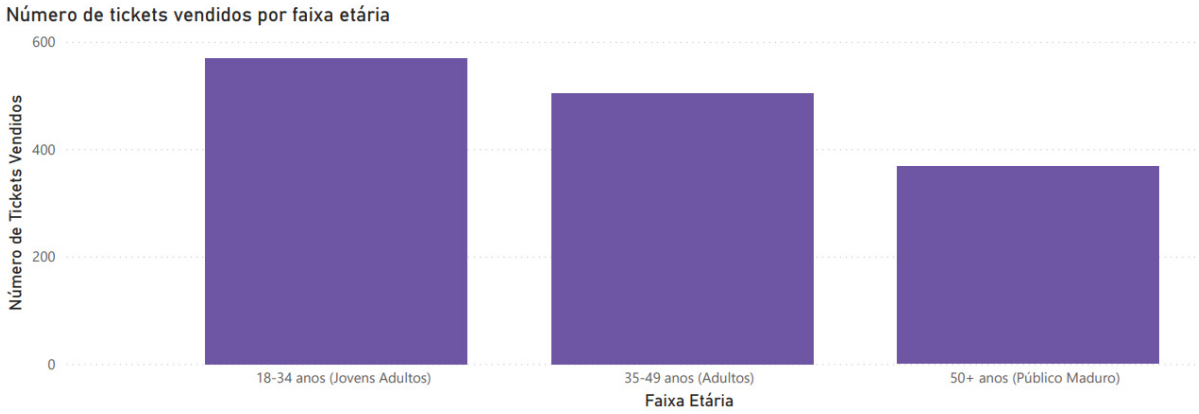
Faixa etária mais frequente ao cinema

Com base nos dados, jovens adultos (18-34 anos) costumam ser os principais frequentadores do cinema. Essa faixa etária tem maior presença nos cinemas, indicando que o cinema ainda é uma atividade popular entre esse grupo. Fatores como maior poder aquisitivo e interesse por lançamentos podem justificar essa tendência.

Há uma queda na presença de público a partir dos 35 anos. Há uma diminuição significativa na quantidade de espectadores conforme a idade avança.

Isso pode estar relacionado a mudanças no estilo de vida, preferências de entretenimento ou falta de tempo. Observe o gráfico a seguir.

GRÁFICO 36 - NÚMERO DE TICKETS VENDIDOS POR FAIXA ETÁRIA



FONTE: O autor (2025)

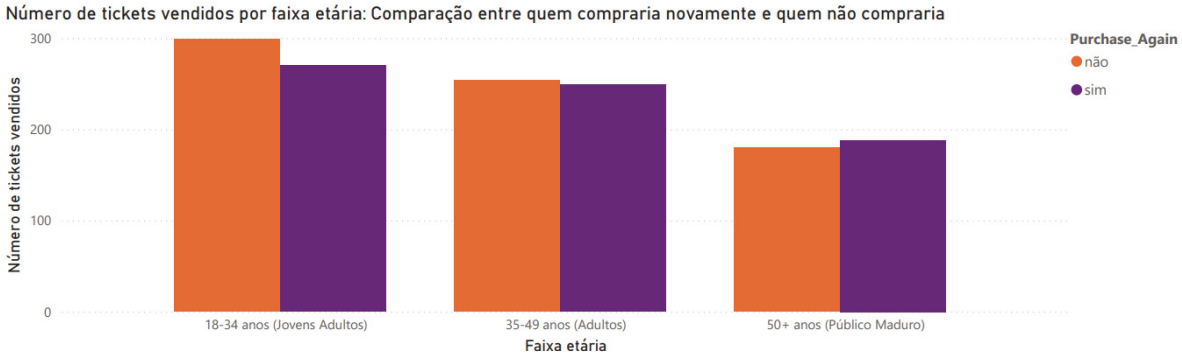
Qual a faixa etária mais fiel ao cinema ?

Analisando os dados, os jovens adultos (18 - 34 anos) são os mais fieis ao cinema. Essa faixa etária representa a maior parte dos espectadores recorrentes. Provavelmente são atraídos por lançamentos, eventos especiais e inovações como IMAX e 4D.

O público acima de 35 anos tem a menor taxa de recompra. Embora vá ao cinema, esse grupo tende a selecionar filmes com mais critério. Preferem experiências mais confortáveis, como salas VIP.

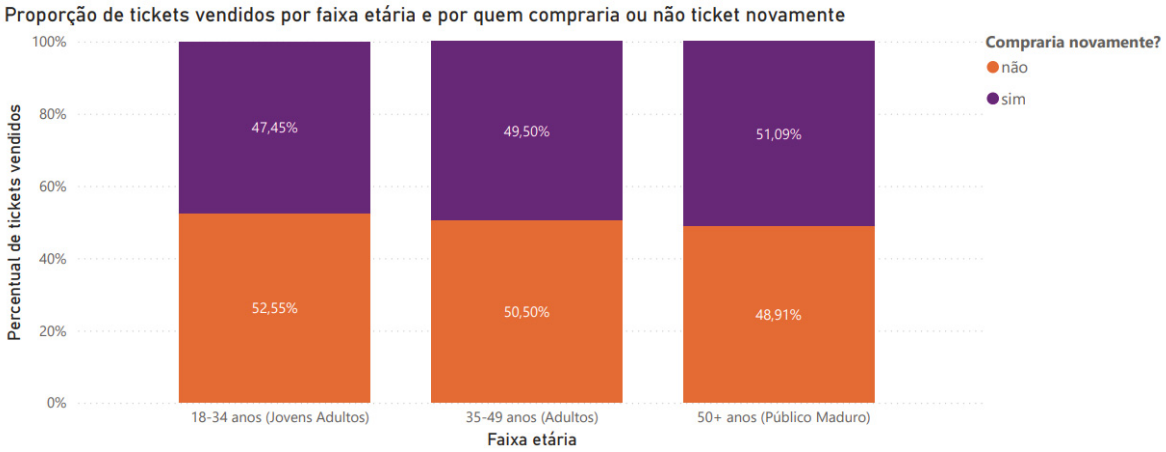
É importante observar que, embora o público jovens adultos apresentam maior fidelidade, a taxa dos que não voltam ao cinema dessa faixa etária é a maior. Isso pode ser explicado pelo fato de esse ser o maior público presente no cinema também. Observe os gráficos a seguir.

GRÁFICO 37 - PROPORÇÃO DE TICKETS VENDIDOS POR FAIXA ETÁRIA E POR QUEM COMPRARIA NOVAMENTE



FONTE: O autor (2025)

GRÁFICO 38 - COMPARAÇÃO DE TICKETS VENDIDOS POR FAIXA ETÁRIA E POR QUEM COMPRARIA NOVAMENTE



FONTE: O autor (2025)

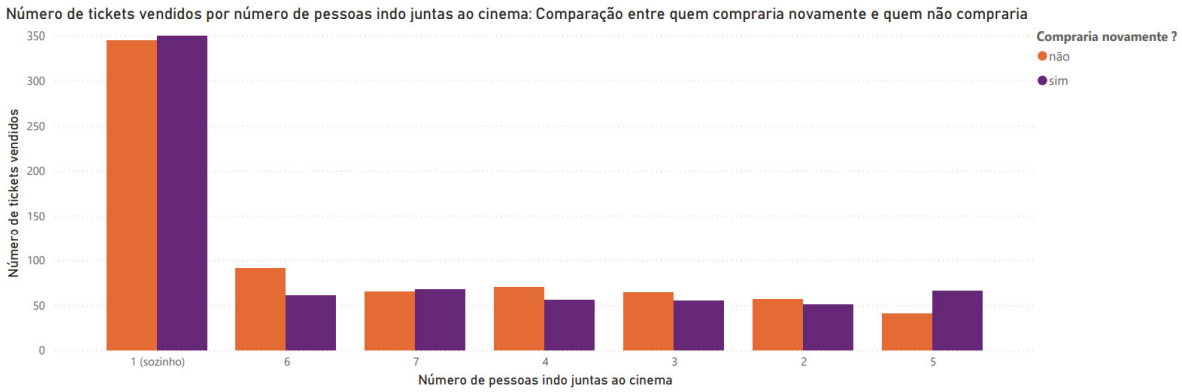
Quem é mais fiel ao cinema ? o que vai sozinho ou em grupo ?

Analisando os dados de recompra de ingressos, pessoas que vão sozinhas tendem a ser mais fieis ao cinema, ou seja, o público que assiste a filmes sozinho tem maior taxa de recompra.

Grupos são menos frequentes, isso pode estar relacionado ao fato de que ir ao cinema em grupo exige coordenação de horários, reduzindo a frequência. Famílias com crianças priorizam filmes infantis e vão apenas em ocasiões específicas.

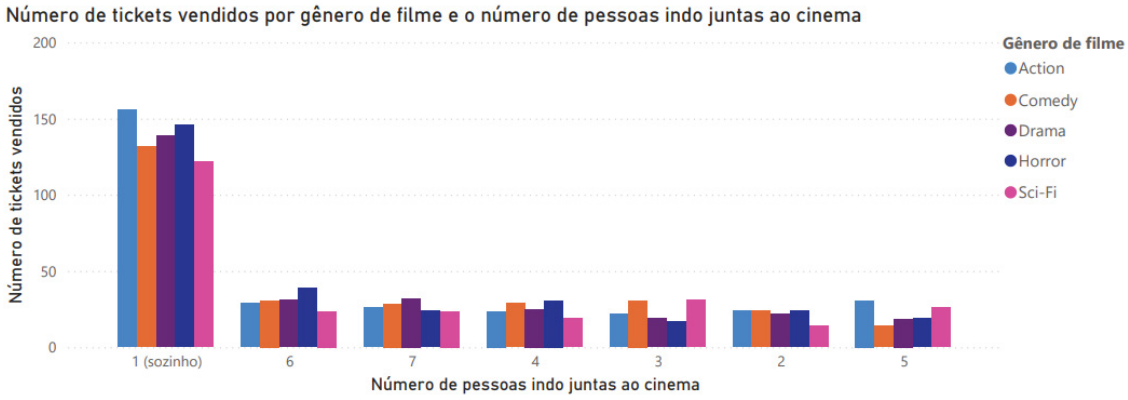
Já casais, por sua vez, possuem fidelidade intermediária. Costumam escolher gêneros específicos (terror, romance, comédia) e retornam mais do que grupos, mas menos do que quem vai sozinho. Observe os gráficos a seguir.

GRÁFICO 39 - COMPARAÇÃO ENTRE NÚMERO DE TICKETS VENDIDOS POR NÚMERO DE PESSOAS QUE VÃO JUNTAS AO CINEMA COM QUEM COMPRARIA NOVAMENTE



FONTE: O autor (2025)

GRÁFICO 40 - NÚMERO DE TICKETS VENDIDOS POR GÊNERO DE FILME E O NÚMERO DE PESSOAS INDO JUNTAS AO CINEMA



FONTE: O autor (2025)

O Cenário Atual e o Futuro do Cinema

Com base nos dados apresentados, podemos traçar um panorama do futuro do cinema:

a) **a experiência no cinema precisa evoluir:** para competir com o streaming, os cinemas podem investir em experiências mais imersivas, como salas premium, eventos exclusivos e interação com os espectadores;

b) **a segmentação do público é essencial**: jovens adultos ainda são o principal público, mas estratégias como ingressos promocionais e benefícios para grupos podem trazer mais diversidade de espectadores;

c) **os gêneros populares podem impulsionar o setor:** filmes de ação e ficção científica continuam atraindo público, indicando que investir nesses gêneros pode ser um caminho para manter os cinemas relevantes.

O Cinema está em declínio?

Embora os dados mostram mudanças significativas nos hábitos dos consumidores, o cinema ainda tem um espaço importante na cultura do entretenimento. A chave para a sobrevivência desse setor será a adaptação, tornando a ida ao cinema uma experiência única que não pode ser substituída pelo streaming. E você? Ainda vai ao cinema ou prefere assistir em casa?

APÊNDICE 15 – TÓPICOS EM INTELIGÊNCIA ARTIFICIAL

A – ENUNCIADO

1) Algoritmo Genético

Problema do Caixeiro Viajante

A Solução poderá ser apresentada em: Python (preferencialmente), ou em R, ou em Matlab, ou em C ou em Java.

Considere o seguinte problema de otimização (a escolha do número de 100 cidades foi feita simplesmente para tornar o problema intratável. A solução ótima para este problema não é conhecida).

Suponha que um caixeiro deva partir de sua cidade, visitar clientes em outras 99 cidades diferentes, e então retornar à sua cidade. Dadas as coordenadas das 100 cidades, descubra o percurso de menor distância que passe uma única vez por todas as cidades e retorne à cidade de origem.

Para tornar a coisa mais interessante, as coordenadas das cidades deverão ser sorteadas (aleatórias), considere que cada cidade possui um par de coordenadas (x e y) em um espaço limitado de 100 por 100 pixels.

O relatório deverá conter no mínimo a primeira melhor solução (obtida aleatoriamente na geração da população inicial) e a melhor solução obtida após um número mínimo de 1000 gerações. Gere as imagens em 2d dos pontos (cidades) e do caminho.

Sugestão:

(1) considere o cromossomo formado pelas cidades, onde a cidade de início (escolhida aleatoriamente) deverá estar na posição 0 e 100 e a ordem das cidades visitadas nas posições de 1 a 99 deverão ser definidas pelo algoritmo genético.
(2) A função de avaliação deverá minimizar a distância euclidiana entre as cidades (os pontos).
(3) Utilize no mínimo uma população com 100 indivíduos;
(4) Utilize no mínimo 1% de novos indivíduos obtidos pelo operador de mutação;
(5) Utilize no mínimo de 90% de novos indivíduos obtidos pelo método de cruzamento (crossover-ox);
(6) Preserve sempre a melhor solução de uma geração para outra.

Importante: A solução deverá implementar os operadores de “cruzamento” e “mutação”.

2) Compare a representação de dois modelos vetoriais

Pegue um texto relativamente pequeno, o objetivo será visualizar a representação vetorial, que poderá ser um vetor por palavra ou por sentença. Seja qual for a situação, considere a quantidade de palavras ou sentenças onde tenha no mínimo duas similares e no mínimo 6 textos, que deverão produzir no mínimo 6 vetores. Também limite o número máximo, para que a visualização fique clara e objetiva.

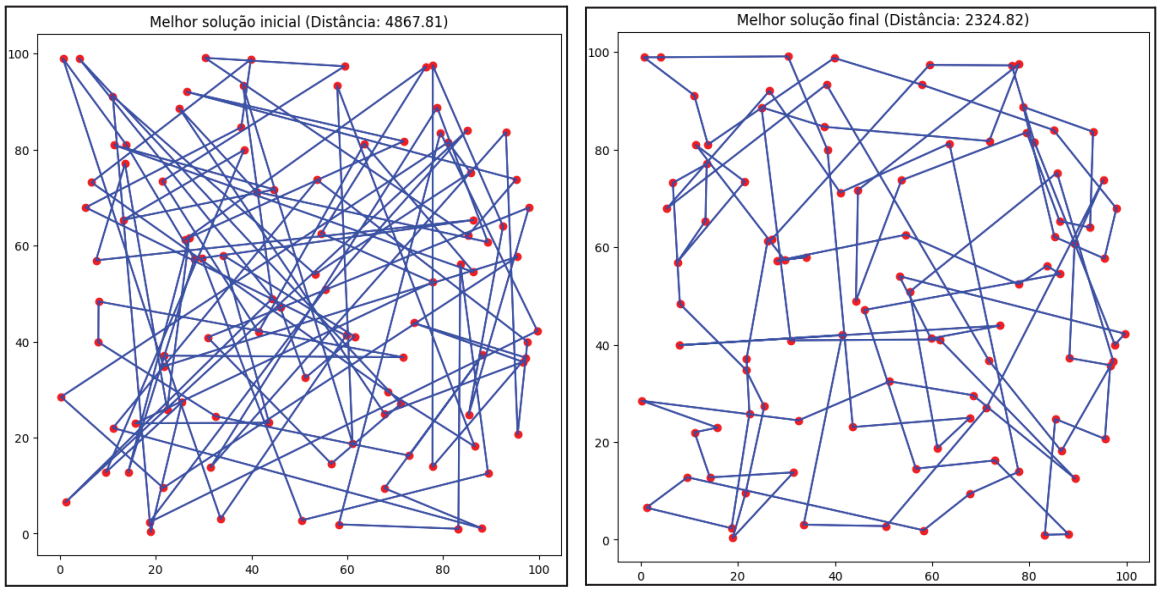
O trabalho consiste em pegar os fragmentos de texto e codificá-las na forma vetorial. Após obter os vetores, imprima-os em figuras (plot) que demonstrem a projeção desses vetores usando a PCA.

O PDF deverá conter o código-fonte e as imagens obtidas.

B – RESOLUÇÃO

Questão 1 - Algoritmo genético

FIGURA 50 - CAMINHO DA MELHOR SOLUÇÃO INICIAL E FINAL



FONTE: O autor (2025)

Questão 2 - Representação de dois modelos vetoriais

O Word2Vec é um modelo de aprendizado profundo utilizado para representar palavras em vetores numéricos (embeddings), permitindo que palavras com significados semelhantes tenham representações vetoriais próximas no espaço multidimensional.

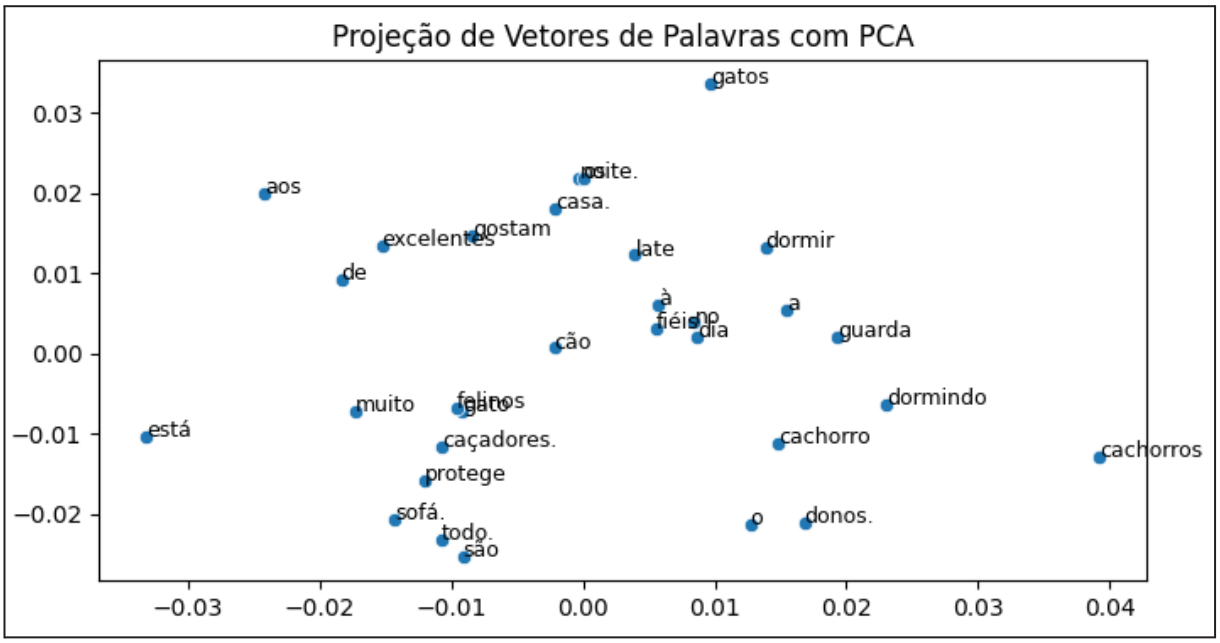
O conjunto de sentenças utilizadas para realizar a sua representação vetorial, onde o vetor foi determinado como sendo uma palavra das sentenças, foi o seguinte:

FIGURA 51 - DEFINIÇÃO DE TEXTOS UTILIZADOS

```
texts = [
  "O gato está dormindo no sofá.",
   "O cachorro late muito à noite.",
   "Os felinos são excelentes caçadores.",
   "O cão de guarda protege a casa.",
   "Gatos gostam de dormir o dia todo.",
   "Cachorros são fiéis aos donos."
]
```

FONTE: O autor (2025)

FIGURA 52 - PROJEÇÃO DE VETORES DE PALAVRAS COM PCA



FONTE: O autor (2025)