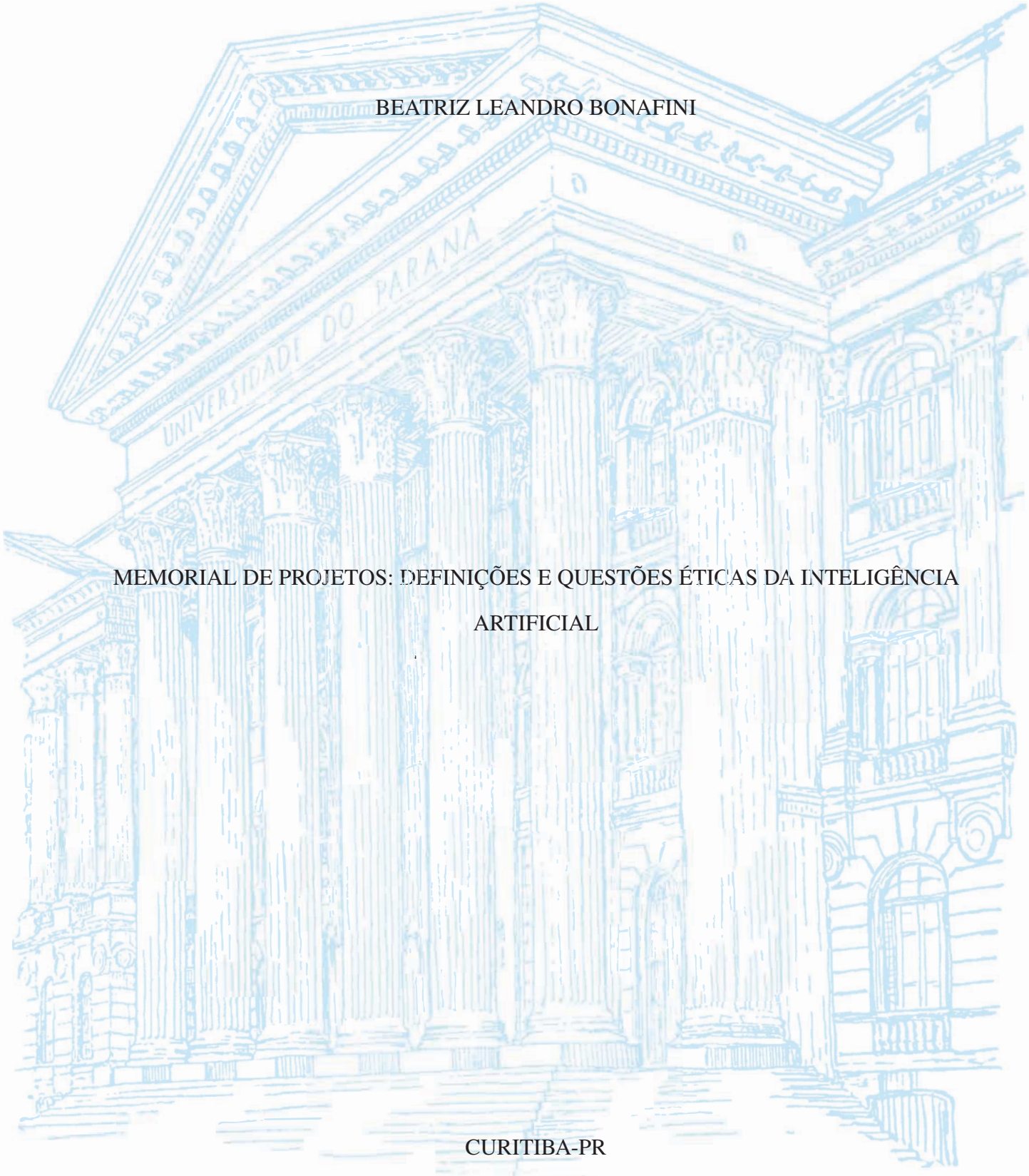
UNIVERSIDADE FEDERAL DO PARANÁ

BEATRIZ LEANDRO BONAFINI



MEMORIAL DE PROJETOS: DEFINIÇÕES E QUESTÕES ÉTICAS DA INTELIGÊNCIA ARTIFICIAL

CURITIBA-PR

2025

BEATRIZ LEANDRO BONAFINI

MEMORIAL DE PROJETOS: DEFINIÇÕES E QUESTÕES ÉTICAS DA INTELIGÊNCIA ARTIFICIAL

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.

Área de concentração: *Inteligência Artificial Aplicada*.

Orientador: Prof. Dr. Razer Anthom Nizer Rojas Montaño.

CURITIBA-PR

2025

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 **BEATRIZ LEANDRO BONAFINI**, intitulada: **MEMORIAL DE PROJETOS:DEFINIÇÕES E QUESTÕES ÉTICAS DA INTELIGÊNCIA ARTIFICIAL**, que após terem inquirido a aluna 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, 22 de Agosto de 2025.

RAZER ANTHOM NIZER ROJAS MONTAÑO
Presidente da Banca Examinadora

RAFAELLA MANTOVANI FONTANA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

AGRADECIMENTOS

A Deus, por conceder-me forças para perseverar, por não me desamparar nos momentos de dificuldade e, igualmente, por permitir que eu reconhecesse, nos instantes de alegria, a perfeição de Seus planos para a minha vida. Durante este percurso, encontrei conforto na Palavra, razão pela qual dedico este trabalho, primeiramente, a Ele.

Aos meus pais, Mari e Lúcio, exemplos de retidão, honestidade e determinação, que me orientaram com amor e dedicação ao longo desta jornada. Sem o apoio incondicional deles, não teria sido possível alcançar esta etapa.

Ao meu irmão Guilherme, cuja influência foi determinante para a minha aproximação da área da computação, e que, ainda na infância, me auxiliou no aprendizado dos primeiros algoritmos.

Ao meu noivo, Eric, pela presença constante, por me motivar, de forma contínua, a perseguir meus objetivos.

Aos colegas de curso, pela disponibilidade em colaborar com as disciplinas, trabalhos e demais desafios que compuseram esta especialização.

À instituição Centro de Estudos e Sistemas Avançados do Recife (CESAR), por oportunizar meu contínuo aprimoramento profissional.

Aos docentes da Especialização em Inteligência Artificial Aplicada, pelo conhecimento transmitido ao longo do curso. De modo especial, ao professor Razer Anthom, pela dedicação, pela disponibilidade em auxiliar os discentes e pela orientação prestada nesta etapa conclusiva.

"Senhor, fazei-me instrumento de vossa paz. Onde houver ódio, que eu leve o amor; Onde houver ofensa, que eu leve o perdão; Onde houver discórdia, que eu leve a união; Onde houver dúvida, que eu leve a fé; Onde houver erro, que eu leve a verdade; Onde houver desespero, que eu leve a esperança; Onde houver tristeza, que eu leve a alegria; Onde houver trevas, que eu leve a luz.

Ó Mestre, fazei que eu procure mais: Consolar, que ser consolado; Compreender, que ser compreendido; Amar, que ser amado.

Pois é dando que se recebe; É perdoando que se é perdoado; E é morrendo que se vive para a vida eterna. Amém"

Oração de São Francisco de Assis

RESUMO

O conceito de Inteligência Artificial (IA) é amplamente difundido, mas permanece subjetivo devido às múltiplas definições e abordagens adotadas por diferentes autores. A partir da disciplina de Introdução à IA, foram analisadas quatro grandes perspectivas: pensar como humanos, agir como humanos, pensar racionalmente e agir racionalmente. Essas abordagens refletem tanto a tentativa de replicar o raciocínio humano quanto a busca por desempenho racional e eficiente em máquinas. Além das definições técnicas, foram discutidas questões éticas e filosóficas relacionadas ao avanço da IA, como a substituição de mão de obra, o uso indevido da tecnologia, a responsabilização por erros e os riscos à segurança e aos direitos civis. Também se destacou a importância da governança e da definição de salvaguardas para garantir o uso responsável da IA. Por fim, enfatiza-se a necessidade de uma abordagem interdisciplinar que una tecnologia, ética, direito e ciências sociais para mitigar os riscos e promover uma IA transparente, segura e alinhada aos valores humanos.

Palavras-chave: Inteligência Artificial; Ética em IA; Agentes Inteligentes; Responsabilidade Algorítmica; Impactos Sociais da IA; Definições de IA.

ABSTRACT

The concept of Artificial Intelligence (AI) is widely disseminated but remains subjective due to the multiple definitions and approaches adopted by different authors. Based on discussions from the Introduction to AI course, four main perspectives were analyzed: thinking like humans, acting like humans, thinking rationally, and acting rationally. These approaches reflect both the attempt to replicate human reasoning and the pursuit of rational and efficient machine performance. In addition to technical definitions, ethical and philosophical issues related to the advancement of AI were addressed, such as the replacement of human labor, the misuse of technology, accountability for errors, and risks to security and civil rights. The importance of governance and the implementation of safeguards was emphasized to ensure the responsible use of AI. Finally, the need for an interdisciplinary approach—integrating technology, ethics, law, and social sciences—is highlighted as essential to mitigate risks and promote transparent, safe, and human-aligned AI systems.

Keywords: Artificial Intelligence; Ethics in AI; Intelligent Agents; Algorithmic Responsibility; Social Impacts of AI; Definitions of AI.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTELIGÊNCIA ARTIFICIAL

1.1 DEFINIÇÃO

Hoje em dia, o termo *Inteligência Artificial* (IA) é muito comum. No entanto, o conceito é subjetivo, pois é usado e aplicado em diferentes contextos, levando a várias interpretações e definições. Durante a disciplina de Introdução à IA, ministrada por Montaño (2024), foi possível observar que diferentes autores oferecem definições conflitantes em relação aos processos de pensamento e raciocínio, bem como ao comportamento exibido por máquinas. Ao analisar o processo de replicação do pensamento e raciocínio humanos por máquinas, alguns autores, definem IA como uma máquina que pensa como um ser humano ou pensa racionalmente. Por outro lado, ao observar o comportamento das máquinas, outros autores definem IA como aquela capaz de agir como um ser humano ou agir racionalmente. Antes de aprofundar cada definição, é importante reconhecer que os pensamentos e ações humanas nem sempre são racionais. Segundo Norvig e Russell (2022), as definições de IA podem ser agrupadas em quatro grandes abordagens: pensar como humanos, agir como humanos, pensar racionalmente e agir racionalmente. Nos parágrafos seguintes, detalha-se cada uma dessas abordagens.

Pensando como humanos: Haugeland (1985) argumenta que a IA está associada ao pensamento humano, visto que se trata de um esforço novo e intrigante para fazer os computadores pensarem como máquinas com mentes, no sentido total e literal. Por outro lado, Bellman (1978) define IA como atividades que associamos ao pensamento humano, como tomada de decisão, resolução de problemas e aprendizado. De acordo com essa perspectiva, o objetivo principal é implementar fielmente o processo de pensamento em um computador por meio de programas. No entanto, é importante lembrar que, embora algoritmos possam produzir resultados corretos, isso não significa que eles simulem o comportamento humano. Para determinar se um programa pensa, devemos primeiro entender como os seres humanos pensam.

Agindo como humanos: A IA, segundo Kurzweil (1990), pode ser definida como “a arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas”. De forma similar, Rich e Knight (1991) a definem como “o estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas”. Nesse contexto, o objetivo da IA seria imitar o comportamento humano, sem necessariamente implicar

a implementação de raciocínio. A relevância reside no fato de o resultado fornecido ser similar ao produzido por um ser humano. Turing (1950) propõe uma forma de medir o nível satisfatório de inteligência por meio do Teste de Turing. O teste consiste em uma conversa entre um humano, outro humano e uma máquina, utilizando linguagem natural. Todos os participantes estão separados. Se não for possível distinguir com segurança a máquina do humano, a máquina é considerada aprovada no teste. É importante ressaltar que, embora a máquina forneça respostas, o teste não verifica a correção dessas respostas, mas sim a proximidade com as respostas humanas.

Pensando racionalmente: De acordo com Charniak e McDermott (1985), a IA consiste no "estudo das faculdades mentais por meio de modelos computacionais". Winston (1992), por sua vez, define a IA como "o estudo das computações que tornam possível perceber, raciocinar e agir". O objetivo, nesse contexto, é a modelagem do pensamento correto, ou seja, a modelagem do processo de raciocínio, a fim de fornecer um resultado logicamente válido. Entretanto, é fundamental compreender como se define um pensamento correto. Assim Aristóteles apresenta um conjunto de frases que formam silogismos. Os silogismos possibilitam a construção do pensamento correto, sendo compostos por três sentenças: duas afirmativas e uma conclusiva. Um exemplo seria: Sócrates é um homem (afirmativa 1); todos os homens são mortais (afirmativa 2); portanto, Sócrates é mortal.

Agindo Racionalmente: para Poole, Mackworth e Goebel (1998), a "Inteligência Computacional é o estudo do projeto de agentes inteligentes". Nilsson (1998) diz que a IA "está relacionada a um desempenho inteligente de artefatos". O objetivo, nesse contexto, é implementar agentes capazes de responder a situações e buscar o melhor resultado possível. Um agente, em sua essência, age sob controle autônomo, percebe seu ambiente, adapta-se às mudanças e é capaz de criar e perseguir metas. Durante a disciplina, foi demonstrado que este é o conceito ideal de IA apresentado por Norvig e Russell (2022), em que "a IA se concentra no estudo e na construção de agentes que fazem a coisa certa".

Diante das diversas definições apresentadas sobre IA, é compreensível que exista certa confusão em relação ao seu real significado e aplicação. A multiplicidade de perspectivas, que vai desde o desempenho inteligente de artefatos até a construção de agentes que fazem a coisa certa, contribui para essa complexidade. Atualmente, observa-se uma tendência crescente de associar a IA a praticamente a maior parte das soluções tecnológicas, criando a impressão de que qualquer inovação deva obrigatoriamente envolver essa tecnologia. No entanto, é importante destacar que nem todas as soluções necessitam da aplicação da IA. Muitas vezes, abordagens

mais simples e tradicionais podem ser mais eficientes, econômicas e adequadas ao contexto. O uso da IA deve ser considerado quando realmente agrega valor, resolve problemas complexos ou otimiza processos de forma significativa. Compreender essa distinção é fundamental para evitar o uso incorreto do termo e para garantir que a tecnologia seja utilizada de maneira ética, responsável e eficaz.

1.2 QUESTÕES ÉTICAS E FILOSÓFICAS DA IA

Ao longo da disciplina de Montaño (2024), foram explorados diversos pontos relacionados às preocupações sobre o avanço da IA na sociedade contemporânea. Os debates concentraram-se em temas como a substituição da mão de obra humana por IA, o uso indevido da tecnologia, questões de responsabilização, riscos de destruição em massa e desafios relacionados ao plágio.

No que diz respeito à substituição da mão de obra, Montaño (2024) destaca que discussões semelhantes já ocorreram em momentos históricos marcados pelo surgimento de novas tecnologias, como a calculadora, a internet e a uberização dos serviços. Ressalta-se que a IA é capaz de executar tarefas que, até recentemente, não existiam ou eram extremamente onerosas para os seres humanos. Entre essas tarefas, destacam-se a análise financeira complexa, a detecção de fraudes, a realização de atividades repetitivas, a manipulação de materiais perigosos e até a exploração de outros planetas. Considerando o avanço exponencial da internet e a imensa geração de dados, torna-se evidente a necessidade de agentes capazes de processar e extrair informações em uma escala que ultrapassa as capacidades humanas. A IA, nesse contexto, surge como uma ferramenta indispensável para lidar com desafios que envolvem grandes volumes de dados e complexidade analítica. Outro ponto neste contexto é que há o surgimento de outras posições de trabalho, melhor remuneradas e qualificadas. Um exemplo disto é o surgimentos de funções como Arquiteto de IA, Cientista de Dados e Engenheiro de *Machine Learning*.

No que diz respeito ao uso indevido da IA, é fundamental destacar que essa preocupação é recorrente em diversas áreas da ciência. Um exemplo discutido durante a disciplina refere-se à tecnologia nuclear, que pode ser aplicada tanto para fins benéficos, como a geração de energia, quanto para propósitos indesejáveis, como a fabricação de armas nucleares. Além disso, o uso da IA em contextos militares também é um tema que gera intensos debates, evidenciando a necessidade de uma reflexão ética sobre suas aplicações.

Durante o período da Especialização em IA Aplicada, a aluna Beatriz Leandro Bonafini foi selecionada para representar o Brasil em uma importante discussão internacional sobre o uso responsável da IA, realizada em junho de 2024, em Lisboa, Portugal. O evento foi promovido pela UNODA (*United Nations Office for Disarmament Affairs*) em parceria com o SIPRI (*Stockholm International Peace Research Institute*) e contou com a participação de novos pesquisadores de 15 países, todos atuantes na área de IA. O principal objetivo do encontro foi fomentar o debate sobre o futuro da IA no contexto da segurança e soberania das nações. Durante as discussões, temas que foram abordados na disciplina da especialização foram vistos, como o controle de Veículos Aéreos Não Tripulados (VANTs) e as estratégias que órgãos governamentais podem adotar para enfrentar desafios relacionados à responsabilização pelo uso da IA. Destacou-se, ainda, a importância da implementação de salvaguardas em soluções de IA, que, conforme também discutido no curso, deixaram de ser opcionais para se tornarem medidas obrigatórias. Em relação à responsabilização por danos causados por sistemas de IA, enfatizou-se que o poder de decisão deve estar nas mãos daqueles que determinam a finalidade do uso da tecnologia, garantindo, assim, maior segurança e ética nas suas aplicações.

Além do uso da IA para fins militares, existem ainda outras formas em que a IA é utilizada de maneira indesejada, tais como:

- Disseminação de *fake news*: A IA pode automatizar a disseminação de notícias falsas, ampliando rapidamente o alcance da desinformação e dificultando a verificação dos fatos;
- *Deepfake* pornográfico: O uso de IA para criar vídeos falsificados de conteúdo adulto representa uma séria violação da privacidade, prejudicando a reputação das vítimas e levantando questões éticas e legais;
- Reforço de preconceitos: Algoritmos de IA podem intensificar preconceitos existentes, caso sejam treinados com dados enviesados, contribuindo para discriminação em áreas como recrutamento, crédito e segurança pública; Plágio acadêmico: Ferramentas de IA podem ser usadas para gerar textos acadêmicos sem a devida originalidade, facilitando casos de plágio e comprometendo a integridade da pesquisa científica;
- Criação de discussões fictícias: *Bots* alimentados por IA podem simular interações humanas em redes sociais, criando debates artificiais que manipulam a percepção da opinião pública;

- Levantamento de *hashtags* em redes sociais: Algoritmos podem ser empregados para impulsionar *hashtags* de forma artificial, influenciando tendências e direcionando a atenção do público para determinados assuntos;

- Manipulação da opinião pública: A IA pode segmentar e direcionar conteúdos com o objetivo de influenciar comportamentos e decisões, especialmente em contextos eleitorais e políticos;

- Direcionamento de embates políticos: O uso de IA para polarizar o debate público, promovendo narrativas extremistas, contribuindo para o aumento da tensão em sociedades democráticas;

- Perda dos direitos civis: com a facilitação da vigilância em massa é possível que pessoas tenham perda da privacidade dado histórico de uso (análise de e-mails, ligações e transações).

De quem é a responsabilidade quando a IA comete um erro? Durante a disciplina, foi abordado um exemplo interessante: um médico, auxiliado por uma IA, fornece um diagnóstico incorreto, gerando o questionamento sobre a quem cabe essa responsabilidade. Nesse caso, foi destacado que o médico mantém o controle da decisão. Enfatizou-se que os sistemas especialistas são ferramentas de apoio, não devendo ser responsáveis pelas decisões finais. O erro é inerente à IA, e isso deve ser considerado sempre que um resultado impreciso for fornecido por ela. O conteúdo apresentado também destacou que, embora sistemas possam superar o desempenho humano em algumas áreas, se um profissional optar por não seguir as recomendações do sistema, ele será o responsável pelas consequências. Deve-se lembrar que robôs e softwares operados por IA são produtos.

Será que uma IA tem a capacidade de aniquilar a raça humana? Esse foi outro tópico abordado ao longo da disciplina, evidenciando que qualquer tecnologia poderia, em teoria, destruir a humanidade. De forma hipotética, imagina-se que a IA poderia se tornar consciente e tomar a decisão equivocada de que os seres humanos geram sofrimento. Assim, para eliminar o sofrimento humano, ela poderia optar por destruir a própria humanidade. Nesse contexto, é necessária a definição de uma função de utilidade, estimativas de estados incorretos e uma função de aprendizagem. A função de utilidade tem o objetivo de verificar se as ações executadas pela IA correspondem à sua real utilidade em determinado contexto. Já a estimativa do estado da

IA busca identificar erros cometidos e implementar ponderações adequadas quando esses erros ocorrem. A função de aprendizagem consiste na validação do que ela está aprendendo.

Por fim, é importante ressaltar que a forma como as tecnologias vêm evoluindo dentro da IA torna-se um desafio para os profissionais da área. Os desafios não se limitam apenas à aquisição de dados, ao desenvolvimento e ao treinamento de uma IA, mas também abrangem os impactos que essas soluções podem gerar em diversas esferas da sociedade. Além disso, destaca-se a necessidade de uma abordagem ética e responsável no desenvolvimento dessas tecnologias, considerando questões como privacidade, segurança e o potencial de viés nos algoritmos. A governança da IA torna-se fundamental para garantir que seu uso seja benéfico e alinhado aos valores humanos, mitigando riscos e promovendo a transparência nas decisões automatizadas. Por conseguinte, a interdisciplinaridade entre tecnologia, filosofia, direito e ciências sociais é essencial para compreender e gerenciar os efeitos da IA no cotidiano, assegurando que o progresso tecnológico contribua para o bem-estar coletivo.

REFERÊNCIAS

BELLMAN, R. E. *An Introduction to Artificial Intelligence: Can Computers Think?* [S.l.]: Boyd & Fraser Publishing Company, 1978.

CHARNIAK, E.; MCDERMOTT, D. *Introduction to Artificial Intelligence*. [S.l.]: Addison-Wesley, 1985.

HAUGELAND, J. *Artificial intelligence: The very idea*. [S.l.]: MIT Press, 1985.

KURZWEIL, R. *The Age of Intelligent Machines*. [S.l.]: MIT Press, 1990.

MONTAÑO, R. A. N. R. Introdução ao aprendizado de máquina. Aula ministrada no Curso de Especialização em Inteligência Artificial Aplicada, Universidade Federal do Paraná. Não há registro público. 2024.

NILSSON, N. J. *Artificial Intelligence: A New Synthesis*. [S.l.]: Morgan Kaufmann, 1998.

NORVIG, P.; RUSSELL, S. *Artificial Intelligence: A Modern Approach, 4th US ed*. [S.l.]: Pearson. http://aima. cs. berkeley. edu, 2022.

POOLE, D.; MACKWORTH, A. K.; GOEBEL, R. *Computational Intelligence: A Logical Approach*. New York: Oxford University Press, 1998.

RICH, E.; KNIGHT, K. *Artificial Intelligence*. 2. ed. [S.l.]: McGraw-Hill, 1991.

TURING, A. M. Computing machinery and intelligence. *Mind*, v. 59, n. 236, p. 433–460, 1950.

WINSTON, P. H. *Artificial Intelligence*. 3. ed. [S.l.]: Addison-Wesley, 1992.

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.

Figura 1.1 – EXERCÍCIO 2: BUSCA HEURÍSTICA - MELHOR CAMINHO PARA BUCHAREST

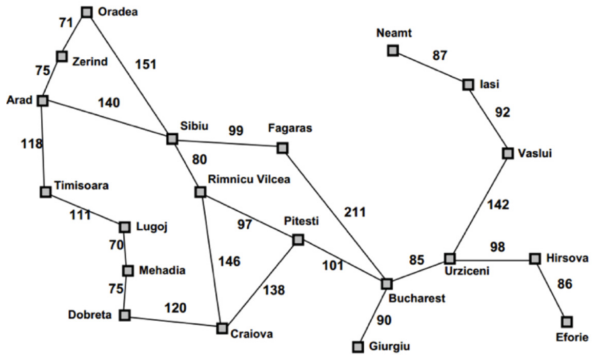


Tabela 1.1: Valores de h_{DLR} — distâncias em linha reta para Bucareste.

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

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: $\alpha, \beta \vdash \gamma \wedge \beta$

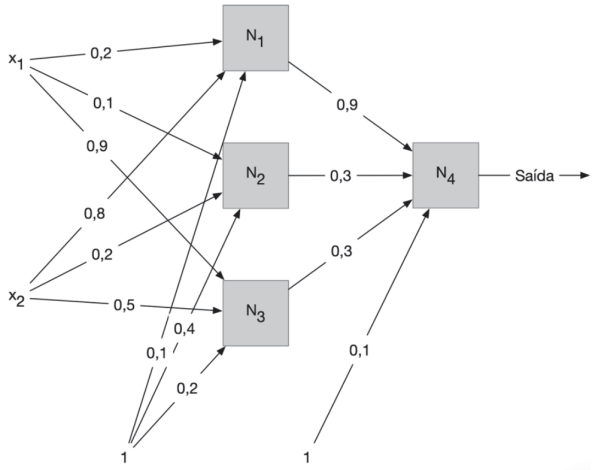
   Equivalência Bicondicional: $(\alpha \leftrightarrow \beta)$ equivale a $(\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha)$

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:

Figura 1.2 – EXERCÍCIO 4: REDE NEURAL ARTIFICIAL



Os neurônios N_1, N_2 e N_3 possuem função de ativação linear. Já N4 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 N_1

b) **(6,25 pontos)** Valor de saída do neurônio N_2

c) **(6,25 pontos)** Valor de saída do neurônio N_3

d) **(6,25 pontos)** Valor de saída da rede como um todo

B - RESOLUÇÃO

1 ChatGPT

a) Inteligência Artificial (IA) refere-se à capacidade de uma máquina ou programa de computador 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).

1. 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.

2. 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 IA 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.

b) É categorizável na abordagem "agir como seres humanos", pois no texto o chat GPT ressaltou a imitação de habilidades inerentemente humanas ao tentar definir o conceito de Inteligência Artificial.

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, *Transfomer 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 e sumariamente através de dados da internet aberta sem nenhuma orientação. Então 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

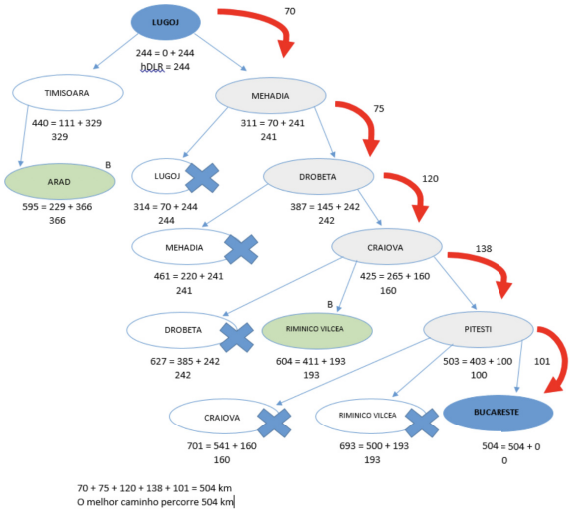
https://zapier.com/blog/how-does-chatgpt-work/

https://investnews.com.br/guias/chatgpt/

d) 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. E ele não entende o que é dito e o que ele responde.

2 Busca Heurística

Figura 1.3 – RESOLUÇÃO EXERCÍCIO 2: CAMINHO DE LUGOJ ATÉ BUCHAREST



3 Lógica

$R4 : r \rightarrow p$ - Equivalência da implicação em R3
$R5 : q \rightarrow p$ - Silogismo hipotético em R2 e R4
$R6 : p \rightarrow q \wedge q \rightarrow p$ - Conjunção em R1 e R5
$R7 : q \leftrightarrow p$ - Equivalência bicondicional em R6

4 Redes Neurais Artificiais

a) $N_1 = 0,3$ função linear onde $fa(u) = u - 3 * 0,2 + 1 * 0,8 + 1 * 0,1 = 0,3$

b) $N_2 = 0,3| - 3 * 0,1 + 1 * 0,2 + 1 * 0,4 = 0,3$

c) $N_3 = -2| - 3 * 0,9 + 1 * 0,5 + 1 * 0,2 = -2$

d) $Sada = -0,139$

$0,3 * 0,9 + 0,3 * 0,3 - 2 * 0,3 + 1 * 0,1 = -0,14$

$TangH = (EXP(-0,14) - EXP(-(-0,14)))/(EXP(-0,14) + EXP(-(-0,14))) = -0,139$

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:

Tabela 2.1: Exercício 2 de LPA: Metados da Base de Dados Carros Brasil

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 uma 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 foram transformadas e **como** foram transformadas;

b) Crie partições contendo 75% dos dados para treino e 25% para teste;

c) Treine modelos `RandomForestRegressor` e `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 Análise Exploratória dos Dados

a. Carregue a base de dados media_precos_carros_brasil.csv

```
# Importando bibliotecas necessárias
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error,
    r2_score

warnings.filterwarnings('ignore')

# Carregando dados do arquivo precos_carros_brasil.csv
dados = pd.read_csv('precos_carros_brasil.csv')

# Listando o nome das colunas
dados.columns

# Imprimindo somente as cinco primeiras linhas
dados.head()

# Imprime o tipo de dado de cada coluna
dados.dtypes

# Número de linhas e colunas
dados.shape
```

b. Verifique se há valores faltantes nos dados. Caso haja, escolha uma tratativa para resolver o problema de valores faltantes

```
# Verificando se há valores faltantes nos dados
dados.isna().any()
dados.isna().sum()

# Verificando se há linhas inteiramente vazias e quantas existem
```

```
dados.isnull().all(axis=1).sum()

# Removendo todas as linhas que têm todos os campos vazios
dados.dropna(axis=0, how='all', inplace=True)

# Verificando novamente se há valores faltantes
dados.isna().sum()
```

c. Verifique se há dados duplicados nos dados

```
# Verificando se há dados duplicados
dados.duplicated().sum()

# Removendo dados duplicados
dados.drop_duplicates(inplace=True)
```

d. Crie duas categorias, para separar colunas numéricas e categóricas. Imprima o resumo de informações das variáveis numéricas e categóricas (estatística descritiva dos dados)

```
# Convertendo colunas numéricas para tipos adequados
dados['engine_size'] = dados['engine_size'].str.replace(',', '.').
    astype(float)
dados['year_of_reference'] = pd.to_numeric(dados['year_of_reference'
    ], errors='coerce').astype('Int64')
dados['year_model'] = pd.to_numeric(dados['year_model'], errors='
    coerce').astype('Int64')
dados['avg_price_brl'] = pd.to_numeric(dados['avg_price_brl'], errors
    ='coerce')

# Criando duas categorias: numéricas e categóricas
numericas_cols = [col for col in dados.columns if dados[col].dtype !=
     'object']

categoricas_cols = [col for col in dados.columns if dados[col].dtype
    == 'object']

# Estatística descritiva das variáveis numéricas
dados[numericas_cols].describe()

# Estatística descritiva das variáveis categóricas
dados[categoricas_cols].describe()
```

e. Imprima a contagem de valores por modelo (model) e marca do carro (brand)

```
# Contagem de valores por modelo e marca
dados.groupby(['model', 'brand']).size().reset_index(name='contagem')
```

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

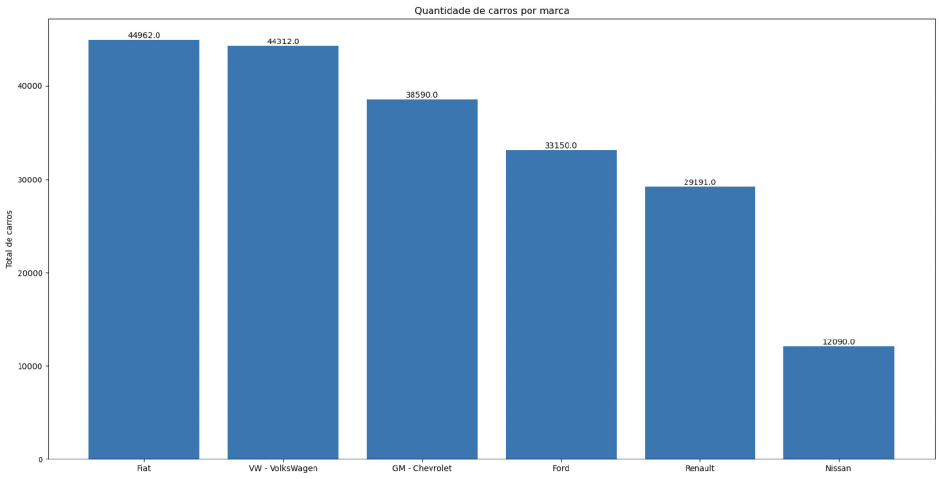
No geral, o preço médio de um carro no Brasil custa aproximadamente R$ 52.756,91. Um dos modelos mais vendidos é Palio Week. Adv/Adv TRYON 1.8 mpi Flex, com câmbio manual e motor 1.6. O menor preço médio de carros foi de R$6.647,00, e o maior preço médio é R$ 979.358,00.

2 Visualização dos Dados

a. Gere um gráfico da distribuição da quantidade de carros por marca

```
# Gráfico: distribuição da quantidade de carros por marca
plt.figure(figsize=(20,10))
grafic_brands_qties = plt.bar(dados['brand'].value_counts().index,
                              dados['brand'].value_counts().values)
plt.title('Quantidade de carros por marca')
plt.ylabel('Total de carros')
plt.bar_label(grafic_brands_qties, size=10, fmt="%.01f", label_type='
    edge')
```

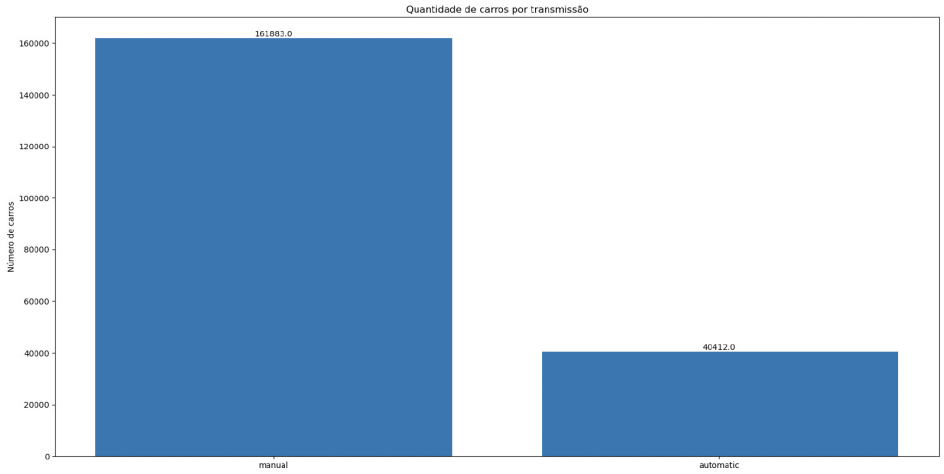
Figura 2.1 – RESOLUÇÃO EXERCÍCIO 2: DISTRIBUIÇÃO DA QUANTIDADE DE CARROS POR MARCA



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

```
# Gráfico: distribuição da quantidade de carros por tipo de
    engrenagem
plt.figure(figsize=(20,10))
grafic_gear_qties = plt.bar(dados['gear'].value_counts().index,
                            dados['gear'].value_counts().values)
plt.title('Quantidade de carros por transmissão')
plt.ylabel('Número de carros')
plt.bar_label(grafic_gear_qties, size=10, fmt="%.01f", label_type='
    edge')
```

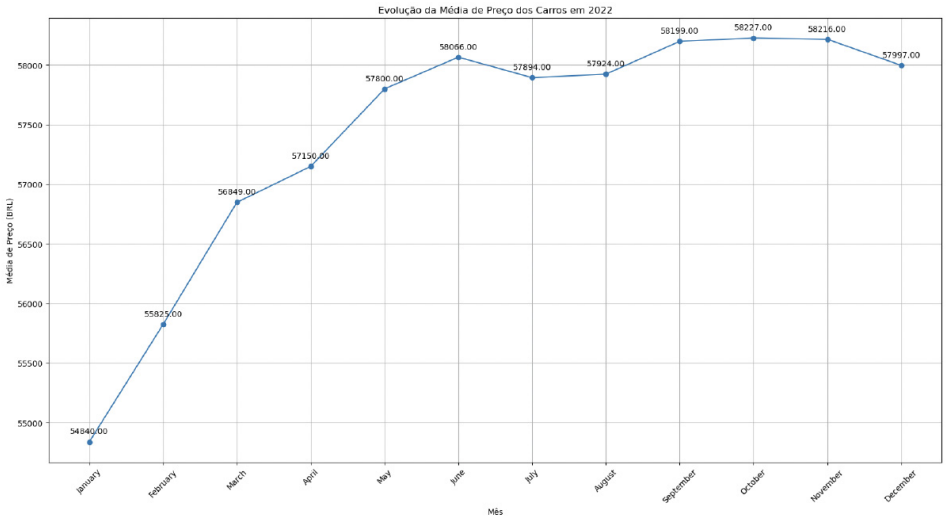
Figura 2.2 – RESOLUÇÃO EXERCÍCIO 2: DISTRIBUIÇÃO DA QUANTIDADE DE CARROS POR TIPO DE ENGRENAGEM



c. Gere um gráfico da evolução da média de preço dos carros ao longo dos meses de 2022 (variável de tempo no eixo X)

```
1  # Evolução da média de preço dos carros ao longo de 2022
2  dados['year_of_reference'] = dados['year_of_reference'].round().
       astype(int)
3  cars_in_2022 = dados[dados['year_of_reference'] == 2022]
4  month_order = ['January','February','March','April','May','June','
       July',
5              'August','September','October','November','December']
6  cars_in_2022['month_of_reference'] = pd.Categorical(cars_in_2022['
       month_of_reference'],
7                                                      categories=month_order, ordered
                                                           =True)
8  cars_avg_price_in_2022 = cars_in_2022.groupby(['month_of_reference'])
       ['avg_price_brl'].mean().round(0).reset_index(name='average_price'
       )
9
10 plt.figure(figsize=(20,10))
11 plt.plot(cars_avg_price_in_2022['month_of_reference'],
12         cars_avg_price_in_2022['average_price'], marker='o', linestyle
               ='-')
13 for x, y in zip(cars_avg_price_in_2022['month_of_reference'],
       cars_avg_price_in_2022['average_price']):
14     plt.annotate(f"{y:.2f}", (x,y), textcoords="offset points", xytext
           =(0,10), ha='center')
15 plt.title('Evolução da Média de Preço dos Carros em 2022')
16 plt.xlabel('Mês')
17 plt.ylabel('Média de Preço (BRL)')
18 plt.xticks(cars_avg_price_in_2022.index, month_order, rotation=45)
19 plt.grid(True)
```

Figura 2.3 – RESOLUÇÃO EXERCÍCIO 2: EVOLUÇÃO DA MÉDIA DE PREÇO DOS CARROS AO LONGO DE 2022

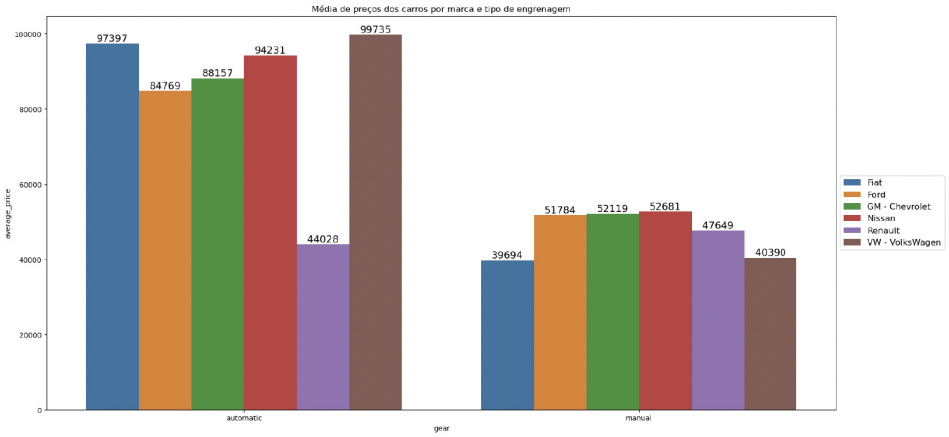


d. Gere um gráfico da distribuição da média de preço dos carros por marca e tipo de engrenagem

```
# Agrupar média de preços médios dos carros por marcar e tipo de engrenagem
car_gear_brand_avg_price = dados.groupby(['brand', 'gear'])['avg_price_brl'].mean().round(0)
car_gear_brand_avg_price.head(12)

# Utilizando a função reset_index para criar uma ordem e facilitar a criação do gráfico
car_gear_brand_avg_price = car_gear_brand_avg_price.reset_index(name='average_price')
car_gear_brand_avg_price.head(12)

# Gerar o gráfico de distruibuição média dos preços dos carros por marca e tipo de engrenagem
plt.figure(figsize=(20,10))
barplot = sns.barplot(
x='gear',
y='average_price',
hue='brand',
data=car_gear_brand_avg_price,
hue_order=car_gear_brand_avg_price['brand'].unique(),
)

barplot.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1, fontsize=12)
barplot.bar_label(barplot.containers[0], fontsize=14);
barplot.bar_label(barplot.containers[1], fontsize=14);
barplot.bar_label(barplot.containers[2], fontsize=14);
barplot.bar_label(barplot.containers[3], fontsize=14);
barplot.bar_label(barplot.containers[4], fontsize=14);
barplot.bar_label(barplot.containers[5], fontsize=14);
```

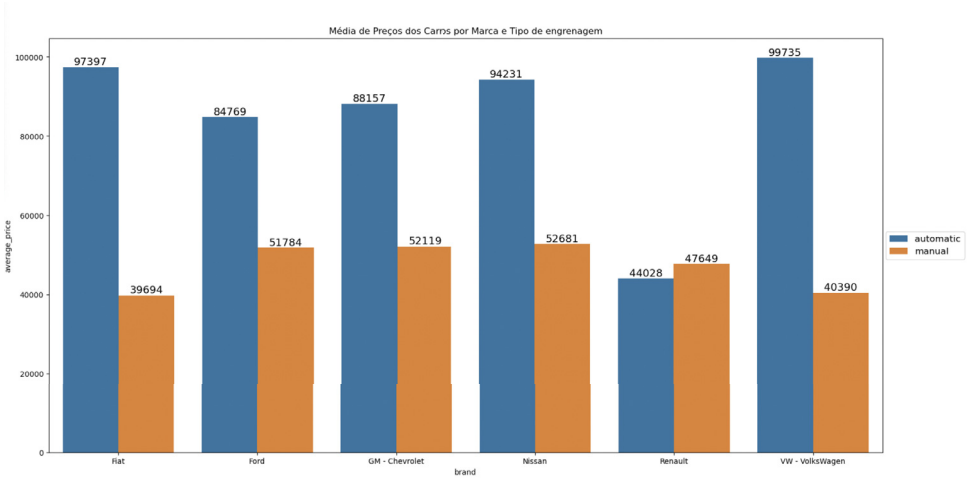
```
26 plt.title('Média de preços dos carros por marca e tipo de engrenagem'
       )
```

Figura 2.4 – RESOLUÇÃO EXERCÍCIO 2: MÉDIA DE PREÇOS DOS CARROS POR MARCA E TIPO DE ENGRENAGEM AGRUPADOS POR TIPO




```
1  # Gerar o segundo tipo de gráfico de distruibuição média dos preços
       dos carros por marca e tipo de engrenagem
2  plt.figure(figsize=(20,10))
3  barplot = sns.barplot(
4  x='brand',
5  y='average_price',
6  hue='gear',
7  data=car_gear_brand_avg_price,
8  hue_order=car_gear_brand_avg_price['gear'].unique(),)
9  barplot.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1,
       fontsize=12)
10 barplot.bar_label(barplot.containers[0], fontsize=14);
11 barplot.bar_label(barplot.containers[1], fontsize=14);
12 plt.title('Média de Preços dos Carros por Marca e Tipo de engrenagem'
       )
```

Figura 2.5 – RESOLUÇÃO EXERCÍCIO 2: MÉDIA DE PREÇOS DOS CARROS POR MARCA E TIPO DE ENGRENAGEM AGRUPADOS POR MARCA



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

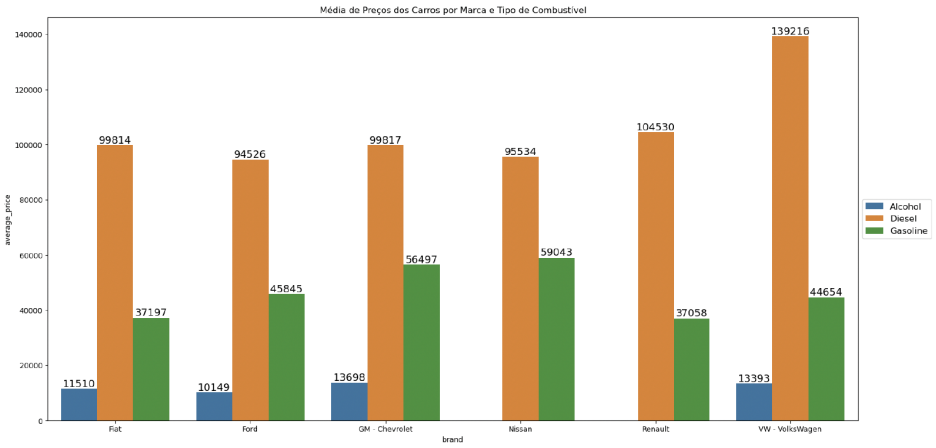
O preço médio do carro com câmbio automático é relativamente mais alto quando comparado aos de câmbio manual para todas as marcas, com exceção da Renault. Volkswagen e Fiat lideram as marcas que possuem maior preço médio para carros de câmbio automático, ao passo que lideram também as marcas com menor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

f. Gere um gráfico da distribuição da média de preço dos carros por marca e tipo de combustível

```
# Agrupar média de preços médios dos carros por marcar e tipo de combustível
cars_avg_price_brand_fuel = dados.groupby(['brand', 'fuel'])['avg_price_brl'].mean().round(0)
cars_avg_price_brand_fuel.head(16)

# Utilizando a função reset_index para criar uma ordem e facilitar a criação do gráfico
cars_avg_price_brand_fuel = cars_avg_price_brand_fuel.reset_index(name='average_price')
cars_avg_price_brand_fuel.head(16)

# Gerar gráfico de média de preços médios de carros por marcar e tipo de combustível
plt.figure(figsize=(20, 10))
barplot = sns.barplot(
x='brand',
y='average_price',
hue='fuel',
data=cars_avg_price_brand_fuel,
hue_order=cars_avg_price_brand_fuel['fuel'].unique()
)
barplot.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1, fontsize=12)
barplot.bar_label(barplot.containers[0], fontsize=14)
barplot.bar_label(barplot.containers[1], fontsize=14)
barplot.bar_label(barplot.containers[2], fontsize=14)
```

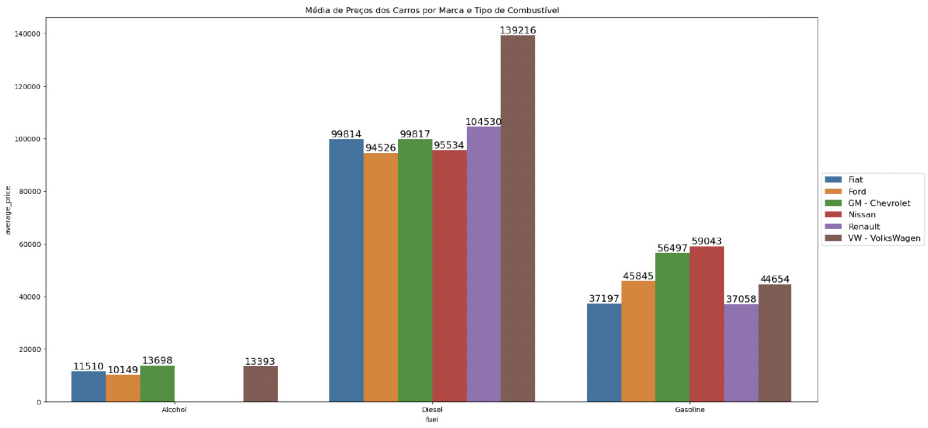
```
22 plt.title('Média de Preços dos Carros por Marca e Tipo de Combustível
       ')
```

Figura 2.6 – RESOLUÇÃO EXERCÍCIO 2: MÉDIA DE PREÇOS DOS CARROS POR MARCA E TIPO DE COMBUSTÍVEL AGRUPADOS POR MARCA




```
1  # Gerar gráfico de média de preços médios de carros por marcar e tipo
       de combustível - Segundo tipo
2  plt.figure(figsize=(20, 10))
3  barplot = sns.barplot(
4  x='fuel',
5  y='average_price',
6  hue='brand',
7  data=cars_avg_price_brand_fuel,
8  hue_order=cars_avg_price_brand_fuel['brand'].unique())
9
10 barplot.legend(loc='center left', bbox_to_anchor=(1, 0.5), ncol=1,
       fontsize=12)
11 barplot.bar_label(barplot.containers[0], fontsize=14)
12 barplot.bar_label(barplot.containers[1], fontsize=14)
13 barplot.bar_label(barplot.containers[2], fontsize=14)
14 barplot.bar_label(barplot.containers[3], fontsize=14)
15 barplot.bar_label(barplot.containers[4], fontsize=14)
16 barplot.bar_label(barplot.containers[5], fontsize=14)
17 plt.title('Média de Preços dos Carros por Marca e Tipo de Combustível
       ')
```

Figura 2.7 – RESOLUÇÃO EXERCÍCIO 2: MÉDIA DE PREÇOS DOS CARROS POR MARCA E TIPO DE COMBUSTÍVEL AGRUPADOS POR TIPO



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 Aplicação de Modelos de Machine Learning para Prever o Preço Médio dos Carros

a) Escolha as variáveis numéricas (modelos de Regressão) para serem as variáveis independentes do modelo. A variável target é avg_price.

```
# Gráfico de distribuição boxplot dos preços médios dos carros
sns.boxplot(dados['avg_price_brl']).set_title('Distribuição dos preços médios')

# Transformar mês de referência em uma variável numérica
dados['month_of_reference'] = LabelEncoder().fit_transform(dados['month_of_reference'])

# Transformar marca de carro em uma variável numérica
dados['brand'] = LabelEncoder().fit_transform(dados['brand'])

# Transformar modelo de carro em uma variável numérica
dados['model'] = LabelEncoder().fit_transform(dados['model'])

# Transformar tipo de combustível em uma variável numérica
dados['fuel'] = LabelEncoder().fit_transform(dados['fuel'])

# Transformar tipo de transmissão em uma variável numérica
dados['gear'] = LabelEncoder().fit_transform(dados['gear'])

# Remover as variáveis de entrada que não são consideradas importantes para a predição
dados_interesting = dados.drop([
```

```
21 'fipe_code',
22 'authentication',
23 ], axis=1)
24
25 # Mapa de correlação das variáveis numéricas com variável Target
26 sns.heatmap(dados_interesting.corr("spearman"), annot = True)
27 plt.title("Mapa de Correlação das Variáveis Numéricas\n", fontsize =
       18)
28 plt.show()
29
30 # Variável X contém apenas variáveis numéricas de interesse para a an
       álise excluindo a variável target
31 X = dados_interesting.drop(['avg_price_brl'], axis=1)
32 X.head()
33
34 # Variável Y contém apenas a variável target - avg_price_brl
35 Y = dados_interesting['avg_price_brl']
36 Y.head()
```

Figura 2.8 – RESOLUÇÃO EXERCÍCIO 3: BOXPLOT DA DISTRIBUIÇÃO DOS PREÇOS MÉDIOS

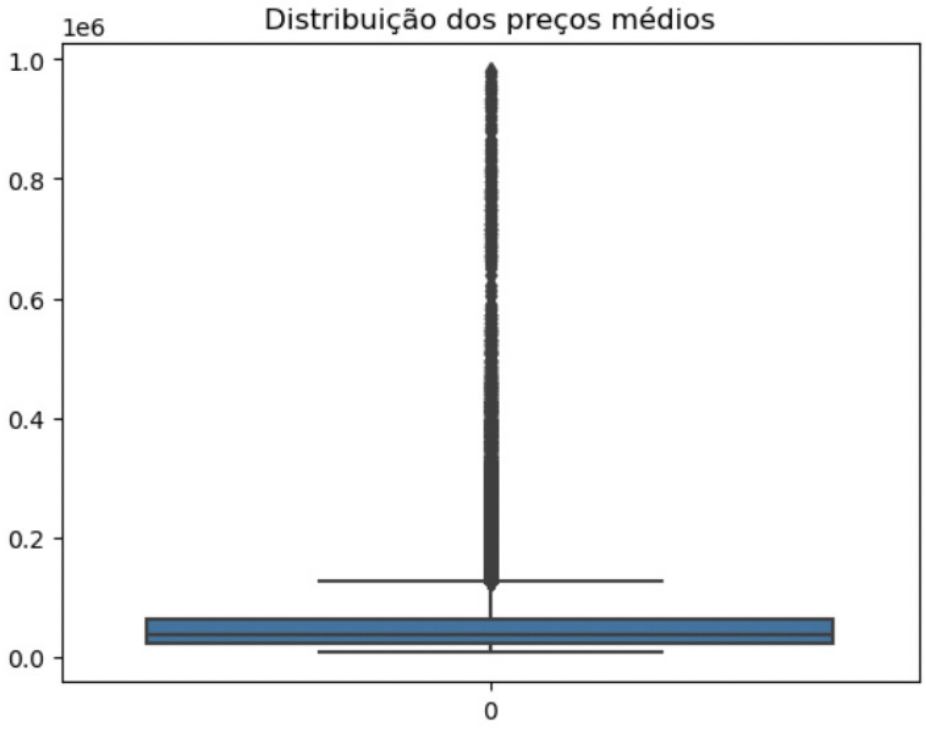
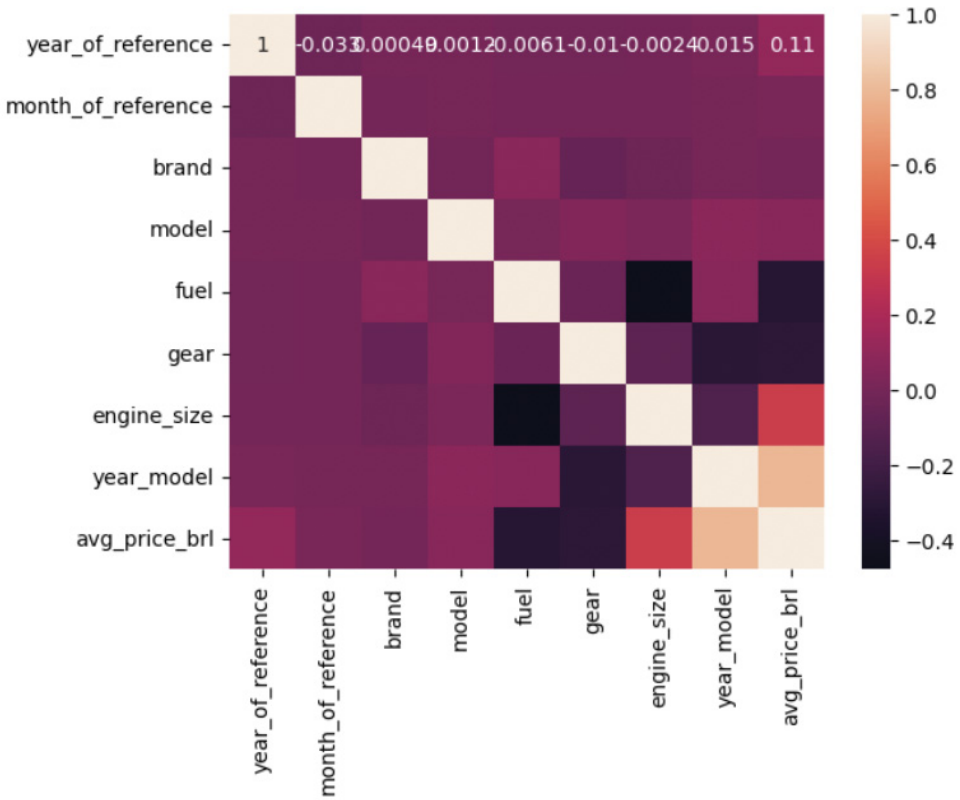


Figura 2.9 – RESOLUÇÃO EXERCÍCIO 3: MAPA DE CORRELAÇÃO ENTRE AS VARIÁVEIS NUMÉRICAS



b. Crie partições contendo 75% dos dados para treino e 25% para teste

```
# Divisão: 25% dos dados são de teste e 75% de treinamento
X_train, x_test, Y_train, y_test = train_test_split(X, Y, test_size =
    0.25,random_state = 42)
```

c. Treine modelos RandomForest (biblioteca RandomForestRegressor) e XGBoost (biblioteca XGBRegressor) para predição dos preços dos carros

```
#Algoritmo Random Forest, sem especificar nenhum parâmetro (número de
    árvores, número de ramificações, etc)
model_rf = RandomForestRegressor()

# Random Forest com os hiperparametros:
model_rf_paramenters = RandomForestRegressor(
max_depth=29,
min_samples_leaf=32,
min_samples_split=28,
n_estimators=208,
random_state=43
)
# Ajuste do modelo, de acordo com as variáveis de treinamento
model_rf_paramenters.fit(X_train, Y_train)

# Ajuste do modelo, de acordo com as variáveis de treinamento
model_rf.fit(X_train, Y_train)

```

```
# Algoritmo XGBoost sem parametros
model_xg = XGBRegressor()

# Ajuste do modelo, de acordo com as variáveis de treinamento
model_xg.fit(X_train, Y_train)

#Algoritmo XGBoost com parametros
model_parameters_xg = XGBRegressor(
    learning_rate=1,
    n_estimators=100,
    max_depth=25,
    min_child_weight=1,
    gamma=0,
    subsample=0.8,
    colsample_bytree=0.8,
    reg_alpha=0,
    reg_lambda=1,
    random_state=42,
    n_jobs=-1
)

# Ajuste do modelo, de acordo com as variáveis de treinamento
model_parameters_xg.fit(X_train, Y_train)
```

d. Grave os valores preditos em variáveis criadas

```
# Predição dos valores de preço médio dos carros com base nos dados de teste Random Forest sem parametros
predicted_values_rf = model_rf.predict(x_test)

# Predição dos valores de preço médio dos carros com base nos dados de teste Random Forest com parametros
predicted_values_parameters_rf = model_rf_paramenters.predict(x_test)

# Predição dos valores de preços médios dos carros com base nos dados de teste XGBoost sem parametros
predicted_values_xg = model_xg.predict(x_test)

# Predição dos valores de preços médios dos carros com base nos dados de teste XGBoost com parametros
predicted_values_parameters_xg = model_parameters_xg.predict(x_test)
```

e. Realize a análise de importância das variáveis para estimar a variável target, para cada modelo treinado

```
# Análise da importancia das variáveis para estimar a variável target - Random Forest sem parametros
model_rf.feature_importances_
feature_importances_rf = pd.DataFrame(
model_rf.feature_importances_,
index = X_train.columns,
columns = ['importances']
).sort_values('importances', ascending=True)
```

month_of_reference 0.005168
year_of_reference 0.012553
brand 0.016695
fuel 0.032713
gear 0.035363
model 0.058640
year_model 0.388397
engine_size 0.450472

```
# Análise da importancia das variáveis para estimar a variável target - Random Forest com parametros
model_rf_paramenters.feature_importances_
feature_importances_parameters_rf = pd.DataFrame(
model_rf_paramenters.feature_importances_,
index = X_train.columns,
columns = ['importances']
).sort_values('importances', ascending=True)
feature_importances_parameters_rf
```

month_of_reference 0.000690
year_of_reference 0.010383
brand 0.016098
gear 0.021959
fuel 0.034132
model 0.037028
year_model 0.410473
engine_size 0.469236

```
# Análise da importancia das variáveis para estimar a variável target - XGBoost com parametros
model_parameters_xg.feature_importances_
feature_importances_parameters_xg = pd.DataFrame(
model_rf_paramenters.feature_importances_,
index = X_train.columns,
columns = ['importances']
).sort_values('importances', ascending=True)
feature_importances_parameters_xg
```

month_of_reference 0.000690
year_of_reference 0.010383
brand 0.016098
gear 0.021959
fuel 0.034132
model 0.037028
year_model 0.410473
engine_size 0.469236

```
# Analisando a importância das variáveis para estimar a variável target - XGBoost sem parametros
model_xg.feature_importances_
feature_importances_xg = pd.DataFrame(
model_xg.feature_importances_,
```

```
index=X_train.columns,
columns=['importances']
).sort_values('importances', ascending=True)
feature_importances_xg
```

month_of_reference 0.004271
year_of_reference 0.017475
model 0.023331
brand 0.056276
gear 0.122657
fuel 0.154437
year_model 0.190992
engine_size 0.430561

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 o modelos treinados.

g. Escolha o melhor modelo com base nas métricas de avaliação MSE, MAE e *R*2:

```
# MSE - calcula o erro quadrático médio das predições do modelo
# Random Forest sem parametro
mse_rf = mean_squared_error(y_test, predicted_values_rf) #12181571.08

# O MAE calcula a média da diferença absoluta entre o valor predito e
    o valor real
# Random Forest sem parametro
mea_rf = mean_absolute_error(y_test, predicted_values_rf) # 1765.88

# O R2 éuma métrica que varia entre 0 e 1 e éuma razão que indica o
   quão bom o nosso modelo
# Random Forest sem parametros
r2_score(y_test, predicted_values_rf) # 0.995
```

```
# O MSE - Random Forest com parametros
mse_parameters_rf = mean_squared_error(y_test,
   predicted_values_parameters_rf) # 151743671.60

# O MAE - Random Forest com parametros
mae_parameters_rf = mean_absolute_error(y_test,
   predicted_values_parameters_rf) # 4577.49

# R2 score - Random Forest com parametros # 0.943
r2_score(y_test, predicted_values_parameters_rf)
```

```
# MSE - XGBoost sem parametros
mse_xg = mean_squared_error(y_test, predicted_values_xg) #
   30952404.229

# MAE - XGBoost sem parametros
mae_xg = mean_absolute_error(y_test, predicted_values_xg) # 3245.31
```

```
6
7 # R2 score - XGBoost sem parametros
8 r2_score(y_test, predicted_values_xg) # 0.988
```

```
1 # MSE - XGBoost com parametros
2 mse_parameters_xg = mean_squared_error(y_test,
     predicted_values_parameters_xg) # 58122007.14
3
4 # MAE - XGBoost com parametros
5 mae_parameters_xg = mean_absolute_error(y_test,
     predicted_values_parameters_xg) # 3988.81
6
7 # R2 score - XGBoost com parametros
8 r2_score(y_test, predicted_values_parameters_xg) # 0.978
```

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

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 parametros para medir a sua acurácia foi utilizado as médidas de acurácia MSE (*mean squared error*) com o valor 12165425.33231638, MAE (*mean absolute error*) com o valor 12165425.33 e R2_score com 0.995 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:

$$Volume = b0 + b1 * dap2 * Ht \tag{3.1}$$

Onde dap é o diâmetro na altura do peito (1,3 metros), 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:

$$Volume = b0 + b1 * dap2 * Ht \tag{3.2}$$

```
1   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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.3}$$

   onde *yi* é o valor observado, *yi* é o valor predito e *y* é a média dos valores *yi* observados. Quanto mais perto de 1 melhor é o modelo;

   - Erro padrão da estimativa: S_{yx}

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{3.4}$$

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

   - Syx%:

$$S_{yx}\% = \frac{S_{yx}}{\bar{y}} \times 100 \tag{3.5}$$

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

7. Escolha o melhor modelo.

B - RESOLUÇÃO

1 Pesquisa com Dados de Satélite (Satellite)

```
#### 1.1 Carregue a base de dados Satellite

#instalar o pacotes necessários
#install.packages("mlbench", repos = "http://cran.us.r-project.org")
#install.packages("e1017", repos = "http://cran.us.r-project.org")
#install.packages("randomForest", repos = "http://cran.us.r-project.org")
#install.packages("kernlab", repos = "http://cran.us.r-project.org")
#install.packages("caret", repos = "http://cran.us.r-project.org")

# usar os pacotes necessários
library(mlbench)
library(caret)

# carregar os dados Satellite
data(Satellite)

# exibir estrutura dos dados Satellite
str(Satellite)

# apresentar alguma medidas estatísticas do dados Satellite
summary(Satellite)

# exibir alguns dados do Satellite
head(Satellite, n = 6)

#### 1.2 Crie partições contendo 80% para treino e 20% para teste

# Para reproductibilidade
set.seed(7)

# particionar em 80% para treino e 20% para teste
indices <- createDataPartition(Satellite[["classes"]], p=0.8, list=F)
treino <- Satellite[indices, ]
teste <- Satellite[-indices, ]

#### 1.3 Treine modelos RandomForest, SVM e RNA para predição destes dados.

# treinar modelos RandomForest, SVM e RNA
```

```
39 rf <- train(classes ~ ., data=treino, method="rf")
40 svm <- train(classes ~ ., data=treino, method="svmRadial")
41 rna <- train(classes ~ ., data=treino, method="nnet", trace=F)
42
43 #### 1.4 Escolha o melhor modelo com base em suas matrizes de confusã
       o.
44
45 # predições
46 predict.rf <- predict(rf, teste)
47 predict.svm <- predict(svm, teste)
48 predict.rna <- predict(rna, teste)
49
50 # matrizes de confusões de cada uma das predições
51
52 # matriz de confusão para o modelo RF
53 conf_matrix.rf <- confusionMatrix(predict.rf, teste[["classes"]])
54
55 print(conf_matrix.rf)
56 cat('\n')
57
58 # matriz de confusão para o modelo SVM
59 conf_matrix.svm <- confusionMatrix(predict.svm, teste[["classes"]])
60
61 print(conf_matrix.svm)
62 cat('\n')
63
64 # matriz de confusão para o modelo RNA
65 conf_matrix.rna <- confusionMatrix(predict.rna, teste[["classes"]])
66
67 print(conf_matrix.rna)
68 cat('\n')
69
70 #### 1.5 Indique qual modelo dá o melhor o resultado e a métrica
       utilizada
71
72 #O melhor modelo foi random forest com acurácia de 0.9182 e kappa de
       0.8987. A métrica utilizada foram a acurácia e kappa.
```

2 Estimativa de Volume de Árvores

```
1  #### 2.1 Carregar o arquivo Volumes.csv (<http://www.razer.net.br/
       datasets/Volumes.csv>)
2
3  dados <- read.csv("http://www.razer.net.br/datasets/Volumes.csv", sep
       =";", dec=",")
4  head(dados)
5
6  #### 2.2 Eliminar a coluna NR, que só apresenta um número sequencial
7
8  dados[["NR"]] <- NULL
9
10 #### 2.3 Criar partição de dados: treinamento 80%, teste 20%
```

```
11
12 regression.indices <- caret::createDataPartition(dados[["VOL"]], p
      =0.8, list=F)
13 regression.treino <- dados[regression.indices, ]
14 regression.teste <- dados[-regression.indices, ]
15
16 #### 2.4 Usando o pacote "caret", treinar os modelos: Random Forest (
      rf), SVM (svmRadial), Redes Neurais (neuralnet) e o modelo alomé
      trico de SPURR
17
18 # Para reproductibilidade
19 set.seed(7)
20
21 regression.rf <- caret::train(VOL ~ ., data=regression.treino, method
      ="rf")
22 regression.svm <- caret::train(VOL ~ ., data=regression.treino,
      method="svmRadial")
23 regression.rna <- caret::train(VOL ~ ., data=regression.treino,
      method="nnet", trControl=trainControl(method = "LOOCV"), trace=F)
24
25 ##### treino do modelo Spurr
26
27 regression.spurr <- nls( VOL ~ b0 + b1*DAP*DAP*HT, data=regression.
      treino, start=list(b0=0.5, b1=0.5))
28
29 ##### visualizar os resultados de Spurr
30
31 summary(regression.spurr)
32
33 #### 2.5 Efetue as predições nos dados de teste
34
35 # predições
36
37 # Para reproductibilidade
38 set.seed(7)
39
40 predict.regression.rf <- predict(regression.rf, regression.teste)
41 predict.regression.svm <- predict(regression.svm, regression.teste)
42 predict.regression.rna <- predict(regression.rna, regression.teste)
43 predict.regression.suprr <- predict(regression.spurr, regression.
      teste)
44
45 #### 2.6 Crie suas próprias funções (UDF) e calcule as seguintes mé
      tricas entre a predição e os dados observados
46
47 # - Erro padrão de estimativa: Syx
48
49 Syx <- function(reals, predicteds, n) {
50   return (sqrt(sum((reals - predicteds)^2)/(n - 2)))
51 }
52
```

```
# - Erro padrão de estimativa em porcentagem: Syx%

SyxPercent <- function(reals, predicteds, n) {
  return ((Syx(reals, predicteds, n)/mean(reals))*100)
}

# - Coeficientededeterminação:R2

R2 <- function (reals, predicteds) {
  return (1 - sum((reals - predicteds)^2)/sum((reals - mean(reals))
      ^2))
}

#### 2.7 Escolha o melhor modelo.

##### métrica de estimativas para o modelo RandomForest - Regressão

# - coeficiente de determinação

R2(regression.teste[["VOL"]], predict.regression.rf)

# - Erro padrão estimativa

n <- nrow(regression.teste)
Syx(regression.teste[["VOL"]], predict.regression.rf, n)

# - Erro padrão estimativa em porcentagem

n <- nrow(regression.teste)
SyxPercent(regression.teste[["VOL"]], predict.regression.rf, n)

##### métrica de estimativas para o modelo SVM - Regressão

# - coeficiente de determinação

R2(regression.teste[["VOL"]], predict.regression.svm)

# - Erro padrão estimativa

n <- nrow(regression.teste)
Syx(regression.teste[["VOL"]], predict.regression.svm, n)

# - Erro padrão estimativa em porcentagem

n <- nrow(regression.teste)
SyxPercent(regression.teste[["VOL"]], predict.regression.svm, n)

##### métricas de estimativas para o modelo nnet - Regressão

# - coeficiente de determinação

```

```
R2(regression.teste[["VOL"]], predict.regression.rna)

# - Erro padrão estimativa

n <- nrow(regression.teste)
Syx(regression.teste[["VOL"]], predict.regression.rna, n)

# - Erro padrão estimativa em porcentagem

n <- nrow(regression.teste)
SyxPercent(regression.teste[["VOL"]], predict.regression.rna, n)

##### métricas de estimativas para o modelo Spurr

# - coeficiente de determinação

R2(regression.teste[["VOL"]], predict.regression.suprr)

# - Erro padrão estimativa

n <- nrow(regression.teste)
Syx(regression.teste[["VOL"]], predict.regression.suprr, n)

# - Erro padrão estimativa em porcentagem

n <- nrow(regression.teste)
SyxPercent(regression.teste[["VOL"]], predict.regression.suprr, n)

##### 2.7 escolha o melhor modelo

##### Resumo dos resultados

###### RF:

# 1. coeficiente de determinação: 0.8223603.
# 2. Erro padrão estimativa: 0.1376052.
# 3. Erro padrão estimativa em porcentagem: 10.42195

###### SVM:

# 1. coeficiente de determinação: 0.6254546
# 2. Erro padrão estimativa: 0.19981
# 3. Erro padrão estimativa em porcentagem: 15.13322

###### nnet:

# 1. coeficiente de determinação: -1.069672
# 2. Erro padrão estimativa: 0.4696948
# 3. Error padrão estimativa em porcentagem: 35.57377

###### spurr:
```

```
154
155 # 1. coeficiente de determinação: 0.7734018
156 # 2. Erro padrão estimativa: 0.1554151
157 # 3. Error padrão estimativa em porcentagem: 11.77084
158
159 # Com base nas métricas, o modelo que se saiu melhor foi o
      RandomForest, com R2 igual 0.8223603, Erro padrão estimativa de
      0.1376052 e Erro padrão de estima- tiva 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

Código

```
install.packages("BSDA")
install.packages("onewaytests")
install.packages("sjPlot")
install.packages("tigerstats")
install.packages("misty")
install.packages("dplyr")
install.packages("ggpubr")
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("stats")

library("BSDA")
library("onewaytests")
library("sjPlot")
library("tigerstats")
library("misty")
library("dplyr")
```

```
library("ggpubr")
library("ggplot2")
library("tidyverse")
library ("car")
library("fdth")
library("stats")
load("salarios.RData")

# 1 Gráficos e Tabelas

# a) Elaborar gráficos box-plot e histograma das variáveis "age"
(idade da esposa) e "husage" (idade do marido) e comparar os
   resultados
# Esposas
Boxplot( ~ age, data=salarios, id=list(method="y"), ylab="Distribuiçã
   o das Idade das Esposas")
# Conclusões: O gráfico demonstra a distribuição das idades das
   esposas
presentes no dataset salarios.
# Desta forma épossível perceber que há maior densidade da distribuiç
   ão
das idades entre os valores aproximados de 30 a 50 anos.
histograma.esposa <- hist(
salarios[["age"]],
xlab="Idade da Esposa",
ylab = "Frequência",
main = "Distribuição das idades das esposas"
)
# Conclusões: Quando observa-se a distribuição das frequências por
   meio
do gráfico de histograma para a idade das esposas,
# épossível notar a densidade observada no gráfico anterior entre os
intervalos de 30 a 50, no entanto, esse tipo de
# visualização fornece mais detalhamento quanto àpresença de esposas
mais jovens e mais velhas.
# Maridos
Boxplot( ~ husage, data=salarios, id=list(method="y"), ylab="
   Distribuição
das Idade dos Maridos")
# Conclusões: A distribuição de idades demonstra a prevalência de
   homens
casados entre intervalos aproximados de 30 a 50 anos.
histograma.marido <- hist(
salarios[["husage"]],
xlab="Idade do Marido",
ylab = "Frequência",
main = "Distribuição das Idade dos Maridos"
)
# Conclusões: A distribuição das frequências das idades demonstram a
prevalência de maridos entre 30 a 50 anos.
# b) Elaborar a tabela de frequências das variáveis "age" (idade da
```

```
esposa) e "husage" (idade do marido) e comparar os resultados
freq.esposa <- table(salarios[["age"]])
# Conclusões: Observando a distribuição das frequências por idades das
esposas épossível observar que nesta base de dados as esposas
# mais novas possuem 18 anos e as mais velhas 59 anos. Outro fato curioso, éque conforme que as ocorrências de esposas mais novas são
# mais raras comparadas as mais velhas.
freq.marido <- table(salarios[["husage"]])
# Conclusões: Já o cenário dos homens demonstra que o mais novo possui 19
anos e o mais velho 86 anos. Outro fato éa pouca incidência
# de homens mais velhos na base de dados. E que comparativamente a
frequência de mulheres mais novas em relação a idade dos homens.

# 2 Medidas de Posição e Dispersão

# a) Calcular a média, mediana e moda das variáveis "age" (idade da
esposa) e "husage" (idade do marido) e comparar os resultados
# Esposa
media.esposa <- mean(salarios[["age"]])
media.esposa # 39.42 anos
mediana.esposa <- median(salarios[["age"]])
mediana.esposa # 39 anos
table(salarios[["age"]])
subset(table(salarios[["age"]]),
table(salarios[["age"]]) == max(table(salarios[["age"]]))) # 217 esposas com
37 anos
# Marido
media.marido <- mean(salarios[["husage"]])
media.marido # 42.45 anos
mediana.marido <- median(salarios[["husage"]])
mediana.marido # 41 anos
table(salarios[["husage"]])
subset(table(salarios[["husage"]]),
table(salarios[["husage"]]) == max(table(salarios[["husage"]]))) # 201
maridos com 44 anos

# Conclusões
media.marido/media.esposa # a média das idades maridos é7,67% superior a
média das idades das esposas.
mediana.marido/mediana.esposa # a mediana das idades dos maridos éde
5,12% superior a mediana das idades das esposas.
44/37 # a moda da idades dos maridos émaior que a moda da idade das
esposas em 18,91%.
# b) Calcular a variância, desvio padrão e coeficiente de variação das
```

```
105 variáveis "age" (idade da esposa) e "husage" (idade do marido) e
    comparar
106 os resultados
107 # Esposa
108 variancia.esposa <- var(salarios[["age"]])
109 variancia.esposa # 99.75
110 desvio.esposa <- sd(salarios[["age"]])
111 desvio.esposa # +- 9.98 anos
112 coeV.esposa <- (desvio.esposa/media.esposa)*100
113 coeV.esposa # 25.33 %
114 # Marido
115 variancia.marido <- var(salarios[["husage"]])
116 variancia.marido # 126.07
117 desvio.marido <- sd(salarios[["husage"]])
118 desvio.marido # +- 11.22 anos
119 coeV.marido <- (desvio.marido/media.marido)*100
120 coeV.marido # 26.44 %
121 # Conclusões
122 variancia.marido/variancia.esposa # a variância das idades dos
    maridos é
123 26,38% superior ao das esposas.
124 desvio.marido/desvio.esposa # o desvio padrão das idades dos maridos
    éde
125 12,42% superior ao das esposas.
126 coeV.marido/coeV.esposa # o coeficiente de variação das idades dos
    marios
127 é de 4,40% superior ao das esposas.
128
129 # 3 Testes paramétricos ou não paramétricos
130
131 # a) Testar se as médias (se você escolher o teste paramétrico) ou as
132 medianas (se você escolher o teste não paramétrico) das variáveis
133 # "age" (idade da esposa) e "husage" (idade do marido) são iguais,
134 construir os intervalos de confiança e comparar os resultados
135 marido.idades <- salarios[["husage"]]
136 esposa.idades <- salarios[["age"]]
137 group.marido <- rep(c('husband'), times=length(marido.idades))
138 group.esposa <- rep(c('wife'), times=length(esposa.idades))
139 data.marido <- data.frame(group=group.marido, ages=marido.idades)
140 data.esposa <- data.frame(group=group.esposa, ages=esposa.idades)
141 data.ages <- rbind(data.marido, data.esposa)
142 # Checando as informações sobre o dataframe gerado
143 group_by(data.ages, group) %>%
144 summarise(
145 count = n(),
146 mean = mean(ages, na.rm = TRUE),
147 median = median(ages, na.rm = TRUE),
148 IQR = IQR(ages, na.rm = TRUE)
149 )
150 ggboxplot(
151 data.ages,
```

```
x = 'group',
y = 'ages',
color = 'group',
palette=c("#00AFBB", "#E7B800"),
ylab = "Ages",
xlab = "Groups"
)
# Premissa 1: Os grupos são independentes? Sim, pois os grupos de idades
de marios e esposas não estão relacionados.Assim não se encontram na
classificação de emparelhados.
# Premissa 2: Os dados possuem distribuição normal?
# H0: dados normalmente distribuídos
# H1: dados não normalmente distribuídos
options(scipen = 999)
amostra <- data.ages[data.ages[["group"]] == 'husband', ]
with(amostra, ks.test(unique(ages), "pnorm", mean = mean(ages), sd =
sd(ages)))
# p-value < 0.05 com rejeição de H0, dados não estão normalmente
distribuídos para idades dos maridos
amostra <- data.ages[data.ages[["group"]] == 'wife', ]
with(amostra, ks.test(unique(ages), "pnorm", mean = mean(ages), sd =
sd(ages)))
# p-value > 0.05 com aceite de H0, dados estão normalmente distribuídos
para idades das esposas
# Premissa 3: As duas amostras possuem homegeneidade das variâncias?
# H0: as variâncias são estatisticamente iguais
# H1: as variâncias não são estatisticamente iguais
res.ftest <- var.test(ages ~ group, data = data.ages)
res.ftest
qf(0.95, 5633, 5633)
dist_f(f = 1.04481, deg.f1 = 5633, deg.f2 = 5633)
dist_f(f = 0.9571118, deg.f1 = 5633, deg.f2 = 5633)
#O teste de F tem valor critico entre 0.9571118 e 1.04481 (regiao de não
rejeição de H0),
#os valores acima de 1.04481 e abaixo de 0.9571118 estão na região de
rejeição de H0 (área azul do gráfico).
#O valor da estatística F calculada é1.2638. Como esse valor se encontra
na região de rejeição de H0,
#então rejeitamos a hipótese de que as variâncias são estatisticamente
iguais.
# Utilizaremos então o teste não paramétrico de Mann-Whitney para
comparar as medianas entre os grupos.
# H0: a idade mediana dos maridos éestatisticamente igual a das esposas.

# H1: a idade mediana dos maridos não éestatisticamente igual a das
```

```
esposas.
res.w <- wilcox.test(ages ~ group, data = data.ages, exact = FALSE,
conf.int = TRUE)
res.w
# p-value < 0.05 com rejeição de H0, concluindo que a idade mediana
   dos
maridos não éestatisticamente igual àdas esposas.
res.w <- wilcox.test(ages ~ group, data = data.ages, exact = FALSE,
conf.int = TRUE)
res.w
# Teste para verificar se a mediana da idade dos maridos émenor que a
mediana da idade das esposas.
# H0: a idade mediana dos maridos não éestatisticamente menor que a
idade mediana das esposas.
# H1: a idade mediana dos maridos éestatisticamente menor que a idade
mediana das esposas.
res.w <- wilcox.test(ages ~ group, data = data.ages, exact = FALSE,
conf.int = TRUE, alternative='less')
res.w
# p-value > 0.05, concluindo que a idade mediana dos maridos não é
estatisticamente menor que a idade mediana das esposas.
# Teste para verificar se a mediana da idade dos maridos émaior que a
mediana da idade das esposas.
# H0: a idade mediana dos maridos não éestatisticamente maior que a
idade mediana das esposas.
# H1: a idade mediana dos maridos éestatisticamente maior que a idade
mediana das esposas.
res.w <- wilcox.test(ages ~ group, data = data.ages, exact = FALSE,
conf.int = TRUE, alternative='greater')
res.w
# p-value < 0.05, concluindo que a idade mediana dos maridos não é
estatisticamente maior que a idade mediana das esposas.
```

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 1 Regressões Ridge, Lasso e ElasticNet

A tabela a seguir apresenta os resultados obtidos no R ao processar os modelos Ridge, Lasso e ElasticNet.

Tabela 5.1: Desempenho dos modelos de regressão nas etapas de treino e teste.

Modelo	Etapa	RMSE	R^2	Lwage (US$)	Lwage inferior (US$)	Lwage superior (US$)
Ridge	Treino	0.83	0.29	7.86	7.69	8.04
	Teste	0.92	0.23	7.86	7.69	8.04
Lasso	Treino	0.83	0.29	8.04	7.86	8.22
	Teste	0.93	0.23	8.04	7.86	8.22
ElasticNet	Treino	0.83	0.29	8.11	7.93	8.29
	Teste	0.92	0.22	8.11	7.93	8.29

Todos os modelos apresentaram RMSE = 0,83 e R^2 = 0,29, ou seja, o mesmo desempenho na etapa de treinamento. Já na etapa de teste, os modelos Ridge e ElasticNet (RMSE = 0,92) apresentaram desempenho preditivo ligeiramente superior ao do Lasso (RMSE = 0,93). O R^2 do ElasticNet foi de 0,22, valor ainda muito próximo ao dos demais modelos. O valor estimado do salário-hora da esposa foi de US$ 7,86 para o Ridge, US$ 8,04 para o Lasso e US$ 8,11 para o ElasticNet. O modelo que apresentou o maior valor estimado foi o ElasticNet. As amplitudes dos intervalos de confiança foram semelhantes, variando aproximadamente entre US$ 0,35 e US$ 0,36, o que indica consistência entre os modelos.

Código

```
1  ############################################################
2  # MODELO RIDGE
3  # Instalando os pacotes necessarios
4  install.packages("plyr")
5  install.packages("readr")
6  install.packages("dplyr")
7  install.packages("caret")
8  install.packages("ggplot2")
9  install.packages("repr")
10 install.packages("glmnet")
11 # Carregando os pacotes necessarios
12 library(plyr)
13 library(readr)
14 library(dplyr)
15 library(caret)
16 library(ggplot2)
17 library(repr)
18 library(glmnet)
19 # Vamos utilizar a base de dados "trabalhosalarios" que eh um dataset
20 # que contem uma amostra com dados sobre a renda de familias
21 # Carregando o dataset
22 load("~/Documents/ESPECIALIZACAO UFPR/ESTATISTICA 2/Bases de Dados
      Usadas nas Aulas Práticas/trabalhosalarios.RData" )
23 # Vamos guardar este dataset em um objeto chamaddo "dat"
24 dat <- trabalhosalarios
25 # Vamos ver o dataset inteiro
26 View(dat)
27 # Vamos visualizar parte do dataset
28 glimpse(dat)
```

```
# Vamos ver como esta nossa memoria
gc()
# Vamos jogar a semente para gerar numeros aleatorios
# Aqui no exemplo a semente eh "1106", mas poderia
# ser qualquer outro numero, se todos usarem a mesma
# semente os resultados serao iguais.
# Essa semente de numeros aleatorios serve para
# particionar o dataset aleatoriamente
set.seed(1106)
# Vamos criar um indice para particionar o dataset em
# 80% para treinamento
index = sample(1:nrow(dat),0.8*nrow(dat))
# Vamos criar a base de dados de treinamento
train = dat[index,]
# Vamos criar a base de dados de teste
test = dat[-index,]
# Vamos checar as dimensoes das bases de treinamento e
# teste
dim(train)
dim(test)
# A base de treinamento possui 2059 linhas (familias) e
# 17 colunas (variaveis)
# A base de teste possui 515 linhas (familias) e 17
# colunas (variaveis)
# Vamos padronizar as variaveis
# Vamos criar um objeto com as variaveis para padronizar
# As variaveis binarias nao sao padronizadas
cols = c('husage', 'husearns', 'huseduc', 'hushrs', 'earns',
'age', 'educ', 'exper', 'lwage')
colnames(train)
# Padronizando a base de treinamento e teste
pre_proc_val <- preProcess(train[,cols], method = c("center", "scale"
    ))
train[,cols] = predict(pre_proc_val, train[,cols])
test[,cols] = predict(pre_proc_val, test[,cols])
# Vamos ver o sumario estatistico das variaveis
# padronizadas de cada dataset
summary(train)
summary(test)

#################################################################
# REGRESSAO RIDGE #
#################################################################
# A regressao Ridge "encolhe os valores dos coeficientes"
# Vamos criar um objeto com as variaveis que usaremos no
# modelo
cols_reg = c('husage',
    'husearns',
    'huseduc',
    'hushrs',
    'age',
```

```
    'educ',
    'husblck',
    'hushisp',
    'kidge6',
    'black',
    'hispanic',
    'union',
    'kidlt6',
    'exper',
    'husunion',
    'lwage')
# Vamos gerar variaveis dummies para organizar os datasets
# em objetos tipo matriz
# Estamos interessados em estimar o salario-hora (lwage)
dummies <- dummyVars(lwage~husage+
    husunion+
    husearns+
    huseduc+
    hushrs+
    husblck+
    hushisp+
    kidge6+
    age+
    educ+
    black+
    hispanic+
    union+
    kidlt6+
    exper,
    data = dat[,cols_reg])
train_dummies = predict(dummies, newdata = train[,cols_reg])
test_dummies = predict(dummies, newdata = test[,cols_reg])
print(dim(train_dummies)); print(dim(test_dummies))
# A regressao Ridge eh uma extens?o da regressao linear em
# que a funcao perda eh alterada para minimizar a
# complexidade do modelo.
# Esta mudanca eh feita ao adicionarmos um parametro de
# penalty igual ao quadrado dos valores dos coeficientes.
# A principal diferenca entre a regressao linear e a
# penalizada eh que a regressao penalizada usa um
# hiperparametro "lambda".
# A funcao glmnet() executa o modelo muitas vezes para
# valores diferentes de lambda.
# Podemos automatizar esta tarefa para encontrar o melhor
# valor de lambda usando a funcao "cv.glmnet()".
# A funcao de perda generica eh:
# Funcao perda = MQO+lambda*soma(quadrados dos valores
# dos coeficientes).
# O lambda eh o parametro de penalty encontrado.
# Vamos guardar a matriz de dados de treinamento das
# variaveis explicativas para o modelo em um objeto
```

```
# chamado "x"
x = as.matrix(train_dummies)
# Vamos guardar o vetor de dados de treinamento da
# variavel dependente para o modelo em um objeto
# chamado "y_train"
y_train = train[["lwage"]]
# Vamos guardar a matriz de dados de teste das variaveis
# explicativas para o modelo em um objeto chamado
# "x_test"
x_test = as.matrix(test_dummies)
# Vamos guardar o vetor de dados de teste da variavel
# dependente para o modelo em um objeto chamado "y_test"
y_test = test[["lwage"]]
# Vamos calcular o valor otimo de lambda;
# alpha = "0", eh para regressao Ridge
# Vamos testar os lambdas de 10^-3 ate 10^2, a cada 0.1
lambdas <- 10^seq(2, -3, by = -.1)
# Calculando o lambda:
ridge_lamb <- cv.glmnet(x, y_train, alpha = 0,
lambda = lambdas)
# Vamos ver qual o lambda otimo
best_lambda_ridge <- ridge_lamb[["lambda"]].min
best_lambda_ridge
# Para contar o tempo de processamento do modelo usamos:
start <- Sys.time()
# Estimando o modelo Ridge
ridge_reg = glmnet(x, y_train, nlambda = 25, alpha = 0,
family = 'gaussian',
lambda = best_lambda_ridge)
end <- Sys.time()
difftime(end, start, units="secs")
# Vamos ver o resultado (valores) da estimativa
# (coeficientes)
ridge_reg[["beta"]]
# Vamos calcular o R^2 dos valores verdadeiros e
# preditos conforme a seguinte funcao:
eval_results <- function(true, predicted, df) {
   SSE <- sum((predicted - true)^2)
   SST <- sum((true - mean(true))^2)
   R_square <- 1 - SSE / SST
   RMSE = sqrt(SSE/nrow(df))
   # As metricas de performace do modelo:
   data.frame(
   RMSE = RMSE,
   Rsquare = R_square
)
}
# Predicao e avaliacao nos dados de treinamento:
predictions_train <- predict(ridge_reg,
s = best_lambda_ridge,
newx = x)
```

```
# As metricas da base de treinamento sao:
eval_results(y_train, predictions_train, train)
# Predicao e avaliacao nos dados de teste:
predictions_test <- predict(ridge_reg,
s = best_lambda_ridge,
newx = x_test)
# As metricas da base de teste sao:
eval_results(y_test, predictions_test, test)
# Vamos fazer uma predicao para o nosso exemplo
# Como os valores do dataset sao padronizados, nos temos
# de padronizar tambem os dados que vamos fazer a predicao
# Note que as variaveis dummies nao devem ser padronizadas
# Vamos fazer uma predicao para:
# husage = 40 anos (idade do marido)
husage = (40-pre_proc_val[["mean"]][["husage"]])/pre_proc_val[["std"
    ]][["husage"]]
husunion = 0
# husearns = 551 (rendimento do marido em dollar)
husearns = (600-pre_proc_val[["mean"]][["husearns"]])/pre_proc_val[["
    std"]]
[["husearns"]]
# huseduc = 13 (anos de estudo do marido)
huseduc = (13-pre_proc_val[["mean"]][["huseduc"]])/pre_proc_val[["std
    "]][["huseduc"]]
# husblck = 1 (o marido eh preto)
husblck = 1
# hushisp = 0 (o marido nao eh hispanico)
hushisp = 0
# hushrs = 40 (o marido trabalha 40 horas semanais)
hushrs = (40-pre_proc_val[["mean"]][["hushrs"]])/pre_proc_val[["std"
    ]][["hushrs"]]
# kidge6 = 1 (tem filhos maiores de 6 anos)
kidge6 = 1
# age = 38 anos (idade da esposa)
age = (38-pre_proc_val[["mean"]][["age"]])/pre_proc_val[["std"]][["
    age"]]
# black = 0 (esposa nao eh preta)
black = 0
# educ = 13 (esposa possui 13 anos de estudo)
educ = (13-pre_proc_val[["mean"]][["educ"]])/pre_proc_val[["std"]][["
    educ"]]
# hispanic = 0 (esposa nao eh hispanica)
hispanic = 0
# union = 0 (o casal nao possui uniao registrada)
union = 0
# exper = 18 anos de experiencia de trabalho da esposa
exper = (18-pre_proc_val[["mean"]][["exper"]])/pre_proc_val[["std"
    ]][["exper"]]
# kidlt6 = 1 (possui filhos com menos de 6 anos)
kidlt6 = 1
dataframe <- data.frame(husage=husage,
```

```
    husunion=husunion,
    husearns=husearns,
    huseduc=huseduc,
    husblck=husblck,
    hushisp=hushisp,
    hushrs=hushrs,
    kidge6=kidge6,
    exper=exper,
    age=age,
    black=black,
    educ=educ,
    hispanic=hispanic,
    union=union,
    kidlt6=kidlt6)
View(dataframe)
colnames(dataframe)
# Vamos construir uma matriz de dados para a predicao
our_pred = as.matrix(dataframe)
# Fazendo a predicao:
predict_our_ridge <- predict(ridge_reg,
s = best_lambda_ridge,
newx = our_pred)
# O resultado da predicao eh:
predict_our_ridge
# O resultado eh um valor padronizado, vamos converte-lo
# para o valor nominal, consistente com o dataset original
wage_pred_ridge=(predict_our_ridge*
pre_proc_val[["std"]][["lwage"]])+
pre_proc_val[["mean"]][["lwage"]]
# O resultado eh:
exp(wage_pred_ridge)
# Este eh o valor predito do salario por hora (USS),
# segundo as caracteristicas que atribuimos
# O intervalo de confianca para o nosso exemplo eh:
n <- nrow(train) # tamanho da amostra
m <- wage_pred_ridge # valor medio predito
s <- pre_proc_val[["std"]][["lwage"]] # desvio padrao
dam <- s/sqrt(n) # distribuicao da amostragem da media
CIlwr_ridge <- m + (qnorm(0.025))*dam # intervalo inferior
CIupr_ridge <- m - (qnorm(0.025))*dam # intervalo superior
# Os valores sao:
exp(CIlwr_ridge)
exp(CIupr_ridge)
# Entao, segundo as caracteristicas que atribuimos o
# salario-hora da esposa eh em media USS7.86 e pode
# variar entre USS7.69 e USS8.04
```

```
############################################################
# MODELO LASSO
# Carregando os pacotes necessarios
library(plyr)
```

```
library(readr)
library(dplyr)
library(caret)
library(ggplot2)
library(repr)
library(glmnet)
# Carregando o dataset
load("~/Documents/ESPECIALIZACAO UFPR/ESTATISTICA 2/Bases de Dados
   Usadas nas Aulas Práticas/trabalhosalarios.RData" )
# Vamos guardar este dataset em um objeto chamaddo "dat"
dat <- trabalhosalarios
# Vamos ver o dataset inteiro
View(dat)
# Vamos visualizar parte do dataset
glimpse(dat)
# Vamos ver como esta nossa memoria
gc()
# Vamos jogar a semente para gerar numeros aleatorios
# Aqui no exemplo a semente eh "1106", mas poderia
# ser qualquer outro numero, se todos usarem a mesma
# semente os resultados serao iguais.
# Essa semente de numeros aleatorios serve para
# particionar o dataset aleatoriamente
set.seed(1106)
# Vamos criar um indice para particionar o dataset em
# 80% para treinamento
index = sample(1:nrow(dat),0.8*nrow(dat))
# Vamos criar a base de dados de treinamento
train = dat[index,]
# Vamos criar a base de dados de teste
test = dat[-index,]
# Vamos checar as dimensoes das bases de treinamento e
# teste
dim(train)
dim(test)
# Vamos padronizar as variaveis
# Vamos criar um objeto com as variaveis para padronizar
# As variaveis binarias nao sao padronizadas
cols = c('husage', 'husearns', 'huseduc', 'hushrs', 'earns',
'age', 'educ', 'exper', 'lwage')
# Padronizando a base de treinamento e teste
pre_proc_val <- preProcess(train[,cols],
method = c("center", "scale"))
train[,cols] = predict(pre_proc_val, train[,cols])
test[,cols] = predict(pre_proc_val, test[,cols])
# Vamos ver o sumario estatistico das variaveis
# padronizadas de cada dataset
summary(train)
summary(test)

#############################################################
```

```
# REGRESSAO LASSO #
#################################################################
# Vamos criar um objeto com as variaveis que usaremos no
# modelo
cols_reg = c('husage',
   'husearns',
   'huseduc',
   'hushrs',
   'age',
   'educ',
   'husblck',
   'hushisp',
   'kidge6',
   'black',
   'hispanic',
   'union',
   'kidlt6',
   'exper',
   'husunion',
   'lwage')
# Vamos gerar variaveis dummies para organizar os datasets
# em objetos tipo matriz
# Estamos interessados em estimar o salario-hora (hrwage)
dummies <- dummyVars(lwage~husage+
   husunion+
   husearns+
   huseduc+
   hushrs+
   husblck+
   hushisp+
   kidge6+
   age+
   educ+
   black+
   hispanic+
   union+
   kidlt6+
   exper,
   data = dat[,cols_reg])
train_dummies = predict(dummies, newdata = train[,cols_reg])
test_dummies = predict(dummies, newdata = test[,cols_reg])
print(dim(train_dummies)); print(dim(test_dummies))
# Vamos guardar a matriz de dados de treinamento das
# variaveis explicativas para o modelo em um objeto
# chamado "x"
x = as.matrix(train_dummies)
# Vamos guardar o vetor de dados de treinamento da
# variavel dependente para o modelo em um objeto
# chamado "y_train"
y_train = train[["lwage"]]
# Vamos guardar a matriz de dados de teste das variaveis
```

```
# explicativas para o modelo em um objeto chamado
# "x_test"
x_test = as.matrix(test_dummies)
# Vamos guardar o vetor de dados de teste da variavel
# dependente para o modelo em um objeto chamado "y_test"
y_test = test[["lwage"]]
# A regressao Lasso reduz os coeficientes nao significativos
# a "zero"
# A regressao Lasso, ou Operador de Encolhimento Absoluto
# Minimo e de Selecao, tambem eh uma variacao da regressao
# linear.
# Para a regressao Lasso, a funcao perda eh alterada para
# minimizar a complexidade do modelo, restringindo a soma
# dos valores absolutos dos coeficientes do modelo
# (tambem chamada de restricao "l1-norm").
# A restricao "l1-norm" forca alguns alguns valores dos
# pesos a "zero" para permitir que outros coeficientes
# tenham valores mais distantes de "zero".
# A funcao perda eh:
# Funcao perda = MQO+lambda*soma(valores absolutos dos
# coeficientes)
# Vamos criar o intervalo para o tunning do melhor lambda
lambdas <- 10^seq(2, -3, by = -.1)
# Vamos atribuir alpha = 1 para implementar a regressao
# lasso
lasso_lamb <- cv.glmnet(x, y_train, alpha = 1,
lambda = lambdas,
standardize = TRUE, nfolds = 5)
# Vamos guardar o lambda "otimo" em um objeto chamado
# best_lambda_lasso
best_lambda_lasso <- lasso_lamb[["lambda"]].min
best_lambda_lasso
# Vamos estimar o modelo Lasso
lasso_model <- glmnet(x, y_train, alpha = 1,
lambda = best_lambda_lasso,
standardize = TRUE)
# Vamos visualizar os coeficientes estimados
lasso_model[["beta"]]
# Perceba que alguns coeficientes estao zerados pois
# nao sao significativos
# Vamos fazer as predicoes na base de treinamento e
# avaliar a regressao Lasso
predictions_train <- predict(lasso_model,
s = best_lambda_lasso,
newx = x)
# Vamos calcular o R^2 dos valores verdadeiros e
# preditos conforme a seguinte funcao:
eval_results <- function(true, predicted, df) {
   SSE <- sum((predicted - true)^2)
   SST <- sum((true - mean(true))^2)
   R_square <- 1 - SSE / SST
```

```
157     RMSE = sqrt(SSE/nrow(df))
158 # As metricas de performace do modelo:
159     data.frame(RMSE = RMSE, Rsquare = R_square)
160 }
161 # As metricas da base de treinamento sao:
162 eval_results(y_train, predictions_train, train)
163 # Vamos fazer as predicoes na base de teste
164 predictions_test <- predict(lasso_model,
165 s = best_lambda_lasso,
166 newx = x_test)
167 # As metricas da base de teste sao:
168 eval_results(y_test, predictions_test, test)
169 # Perceba que os erros nao sao muito diferentes, assim
170 # como os valores de R2.
171 # Vamos fazer uma predicao para o nosso exemplo
172 # Vamos fazer a predicao com base nos mesmos parametros
173 # como fizemos anteriormente na regressao Ridge
174 # Vamos fazer uma predicao para:
175 # husage = 40 anos (idade do marido)
176 husage = (40-pre_proc_val[["mean"]][["husage"]])/pre_proc_val[["std"
    ]][["husage"]]
177 husunion = 0
178 # husearns = 551 (rendimento do marido em USS)
179 husearns = (600-pre_proc_val[["mean"]][["husearns"]])/pre_proc_val[["
    std"]]
180 [["husearns"]]
181 # huseduc = 13 (anos de estudo do marido)
182 huseduc = (13-pre_proc_val[["mean"]][["huseduc"]])/pre_proc_val[["std
    "]][["huseduc"]]
183 # husblck = 1 (o marido eh preto)
184 husblck = 1
185 # hushisp = 0 (o marido nao eh hispanico)
186 hushisp = 0
187 # hushrs = 40 (o marido trabalha 40 horas semanais)
188 hushrs = (40-pre_proc_val[["mean"]][["hushrs"]])/pre_proc_val[["std"
    ]][["hushrs"]]
189 # kidge6 = 1 (tem filhos maiores de 6 anos)
190 kidge6 = 1
191 # age = 38 anos (idade da esposa)
192 age = (38-pre_proc_val[["mean"]][["age"]])/pre_proc_val[["std"]][["
    age"]]
193 # black = 0 (esposa nao eh preta)
194 black = 0
195 # educ = 13 (esposa possui 13 anos de estudo)
196 educ = (13-pre_proc_val[["mean"]][["educ"]])/pre_proc_val[["std"]][["
    educ"]]
197 # hispanic = 0 (esposa nao eh hispanica)
198 hispanic = 0
199 # union = 0 (o casal nao possui uniao registrada)
200 union = 0
201 # exper = 18 anos de experiencia de trabalho da esposa
```

```
exper = (18-pre_proc_val[["mean"]][["exper"]])/pre_proc_val[["std"
    ]][["exper"]]
# kidlt6 = 1 (possui filhos com menos de 6 anos)
kidlt6 = 1
dataframe <- data.frame(husage=husage,
    husunion=husunion,
    husearns=husearns,
    huseduc=huseduc,
    husblck=husblck,
    hushisp=hushisp,
    hushrs=hushrs,
    kidge6=kidge6,
    exper=exper,
    age=age,
    black=black,
    educ=educ,
    hispanic=hispanic,
    union=union,
    kidlt6=kidlt6)
View(dataframe)
colnames(dataframe)
# Vamos construir uma matriz de dados para a predicao
our_pred = as.matrix(dataframe)
# Vamos para a predicao
predict_our_lasso <- predict(lasso_model,
s = best_lambda_lasso,
newx = our_pred)
predict_our_lasso
# Novamente, o resultado esta padronizado, nos temos de
# converte-lo para valor compativel com o dataset original
wage_pred_lasso=(predict_our_lasso*
pre_proc_val[["std"]][["lwage"]])+
pre_proc_val[["mean"]][["lwage"]]
exp(wage_pred_lasso)
# Portanto, o salario-hora da esposa eh USS 2.09
# Vamos criar o intervalo de confianca para o nosso
# exemplo
n <- nrow(train)
m <- wage_pred_lasso
s <- pre_proc_val[["std"]][["lwage"]]
dam <- s/sqrt(n)
CIlwr_lasso <- m + (qnorm(0.025))*dam
CIupr_lasso <- m - (qnorm(0.025))*dam
# O intervalo de confianca eh:
exp(CIlwr_lasso)
exp(CIupr_lasso)
# Entao, o salario hora da esposa eh de USS 8.04 e pode variar
# entre USS 7.93 e USS 8.29
```

```
#############################################################
# MODELO ELASTIC NET
```

```
# Carregando os pacotes necessarios
library(plyr)
library(readr)
library(dplyr)
library(ggplot2)
library(repr)
library(glmnet)
library(caret)
# Carregando o dataset
load("~/Documents/ESPECIALIZACAO UFPR/ESTATISTICA 2/Bases de Dados
   Usadas nas Aulas Práticas/trabalhosalarios.RData" )
# Vamos guardar este dataset em um objeto chamaddo "dat"
dat <- trabalhosalarios
# Vamos ver o dataset inteiro
View(dat)
# Vamos visualizar parte do dataset
glimpse(dat)
# Vamos ver como esta nossa memoria
gc()
# Vamos jogar a semente para gerar numeros aleatorios
# Aqui no exemplo a semente eh "1106", mas poderia
# ser qualquer outro numero, se todos usarem a mesma
# semente os resultados serao iguais.
# Essa semente de numeros aleatorios serve para
# particionar o dataset aleatoriamente
set.seed(1106)
# Vamos criar um indice para particionar o dataset em
# 80% para treinamento
index = sample(1:nrow(dat),0.8*nrow(dat))
# Vamos criar a base de dados de treinamento
train = dat[index,]
# Vamos criar a base de dados de teste
test = dat[-index,]
# Vamos checar as dimensoes das bases de treinamento e
# teste
dim(train)
dim(test)
# Vamos padronizar as variaveis
# Vamos criar um objeto com as variaveis para padronizar
# As variaveis binarias nao sao padronizadas
cols = c('husage', 'husearns', 'huseduc', 'hushrs', 'earns',
'age', 'educ', 'exper', 'lwage')
colnames(train)
# Padronizando a base de treinamento e teste
pre_proc_val <- preProcess(train[,cols],
method = c("center", "scale"))
train[,cols] = predict(pre_proc_val, train[,cols])
test[,cols] = predict(pre_proc_val, test[,cols])
# Vamos ver o sumario estatistico das variaveis
# padronizadas de cada dataset
summary(train)
```

```
summary(test)

#################################################################
# REGRESSAO ELASTICNET #
#################################################################
# Vamos criar um objeto com as variaveis que usaremos no
# modelo
cols_reg = c('husage',
   'husearns',
   'huseduc',
   'hushrs',
   'age',
   'educ',
   'husblck',
   'hushisp',
   'kidge6',
   'black',
   'hispanic',
   'union',
   'kidlt6',
   'exper',
   'husunion',
   'lwage')
# Vamos gerar variaveis dummies para organizar os datasets
# em objetos tipo matriz
# Estamos interessados em estimar o salario-hora (lwage)
dummies <- dummyVars(lwage~husage+
   husunion+
   husearns+
   huseduc+
   hushrs+
   husblck+
   hushisp+
   kidge6+
   age+
   educ+
   black+
   hispanic+
   union+
   kidlt6+
   exper,
   data = dat[,cols_reg])
train_dummies = predict(dummies, newdata = train[,cols_reg])
test_dummies = predict(dummies, newdata = test[,cols_reg])
print(dim(train_dummies)); print(dim(test_dummies))
# Vamos guardar a matriz de dados de treinamento das
# variaveis explicativas para o modelo em um objeto
# chamado "x"
x = as.matrix(train_dummies)
# Vamos guardar o vetor de dados de treinamento da
# variavel dependente para o modelo em um objeto
```

```
# chamado "y_train"
y_train = train[["lwage"]]
# Vamos guardar a matriz de dados de teste das variaveis
# explicativas para o modelo em um objeto chamado
# "x_test"
x_test = as.matrix(test_dummies)
# Vamos guardar o vetor de dados de teste da variavel
# dependente para o modelo em um objeto chamado "y_test"
y_test = test[["lwage"]]
# A regressao ElasticNet combina aspectos das regressoes
# Ridge e Lasso. Ela penaliza o modelo usando ambas
# funcoes de penalidade, l2-norm and l1-norm.
# O modelo pode ser customizado pelo pacote "caret"
# que encontra os valores otimos dos parametros
# automaticamente.
# Vamos configurar o treinamento do modelo por
# cross validation, com 10 folders, 5 repeticoes
# e busca aleatoria dos componentes das amostras
# de treinamento, o "verboseIter" eh soh para
# mostrar o processamento.
train_cont <- trainControl(method = "repeatedcv",
number = 10,
repeats = 5,
search = "random",
verboseIter = TRUE)
# Nos nao temos o parametro "alpha", porque a regressao
# ElasticNet vai encontra-lo automaticamente, cujo valor
# estara entre 0 e 1 (para Ridge ==> alpha = 0; e
# para Lasso ==> alpha = 1).
# O parametro "lambda" tambem eh escolhido por
# cross-validation
# Vamos treinar o modelo
elastic_reg <- train(lwage~husage+
    husunion+
    husearns+
    huseduc+
    hushrs+
    husblck+
    hushisp+
    kidge6+
    age+
    educ+
    black+
    hispanic+
    union+
    kidlt6+
    exper,
    data = train,
    method = "glmnet",
    tuneLength = 10,
    trControl = train_cont)
```

```
# O melhor parametro alpha escolhido eh:
elastic_reg[["bestTune"]]
# E os parametros sao:
elastic_reg[["finalModel"]][["beta"]]
# Vamos fazer as predicoes e avaliar a performance do
# modelo
# Vamos fazer as predicoes no modelo de treinamento:
predictions_train <- predict(elastic_reg, x)
# Vamos calcular o R^2 dos valores verdadeiros e
# preditos conforme a seguinte funcao:
eval_results <- function(true, predicted, df) {
   SSE <- sum((predicted - true)^2)
   SST <- sum((true - mean(true))^2)
   R_square <- 1 - SSE / SST
   RMSE = sqrt(SSE/nrow(df))
   # As metricas de performace do modelo:
   data.frame(RMSE = RMSE, Rsquare = R_square)
}
# As metricas de performance na base de treinamento
# sao:
eval_results(y_train, predictions_train, train)
# Vamos fazer as predicoes na base de teste
# Vamos fazer uma predicao para:
# husage = 40 anos (idade do marido)
husage = (40-pre_proc_val[["mean"]][["husage"]])/pre_proc_val[["std"
   ]][["husage"]]
husunion = 0
# husearns = 551 (rendimento do marido em USS)
husearns = (600-pre_proc_val[["mean"]][["husearns"]])/pre_proc_val[["
   std"]]
[["husearns"]]
# huseduc = 13 (anos de estudo do marido)
huseduc = (13-pre_proc_val[["mean"]][["huseduc"]])/pre_proc_val[["std
   "]][["huseduc"]]
# husblck = 1 (o marido eh preto)
husblck = 1
# hushisp = 0 (o marido nao eh hispanico)
hushisp = 0
# hushrs = 40 (o marido trabalha 40 horas semanais)
hushrs = (40-pre_proc_val[["mean"]][["hushrs"]])/pre_proc_val[["std"
   ]][["hushrs"]]
# kidge6 = 1 (tem filhos maiores de 6 anos)
kidge6 = 1
# age = 38 anos (idade da esposa)
age = (38-pre_proc_val[["mean"]][["age"]])/pre_proc_val[["std"]][["
   age"]]
# black = 0 (esposa nao eh preta)
black = 0
# educ = 13 (esposa possui 13 anos de estudo)
educ = (13-pre_proc_val[["mean"]][["educ"]])/pre_proc_val[["std"]][["
   educ"]]
```

```
# hispanic = 0 (esposa nao eh hispanica)
hispanic = 0
# union = 0 (o casal nao possui uniao registrada)
union = 0
# exper = 18 anos de experiencia de trabalho da esposa
exper = (18-pre_proc_val[["mean"]][["exper"]])/pre_proc_val[["std"
    ]][["exper"]]
# kidlt6 = 1 (possui filhos com menos de 6 anos)
kidlt6 = 1
dataframe <- data.frame(husage=husage,
    husunion=husunion,
    husearns=husearns,
    huseduc=huseduc,
    husblck=husblck,
    hushisp=hushisp,
    hushrs=hushrs,
    kidge6=kidge6,
    exper=exper,
    age=age,
    black=black,
    educ=educ,
    hispanic=hispanic,
    union=union,
    kidlt6=kidlt6)
View(dataframe)
colnames(dataframe)
# Vamos construir uma matriz de dados para a predicao
our_pred = as.matrix(dataframe)
# Vamos fazer a predicao com base nos parametros que
# selecionamos
predict_our_elastic <- predict(elastic_reg,our_pred)
predict_our_elastic
# Novamente, o resultado eh padronizado, nos temos que
# reverte-lo para o nivel dos valores originais do
# dataset, vamos fazer isso:
wage_pred_elastic=(predict_our_elastic*
pre_proc_val[["std"]][["lwage"]])+
pre_proc_val[["mean"]][["lwage"]]
exp(wage_pred_elastic)
# Vamos criar o intervalo de confianca para o nosso
# exemplo
n <- nrow(train)
m <- wage_pred_elastic
s <- pre_proc_val[["std"]][["lwage"]]
dam <- s/sqrt(n)
CIlwr_elastic <- m + (qnorm(0.025))*dam
CIupr_elastic <- m - (qnorm(0.025))*dam
# Os valores minimo e maximo sao:
exp(CIlwr_elastic)
exp(CIupr_elastic)
# Entao, o salario-hora medio da esposa eh de US$8.11
```

```
250 # e pode variar entre USS 7.93 e USS 8.29
251 ##########################################################
```

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

1 Construção de Características: Identificador automático de idioma

```
1  ingles = [
2     "Hello, how are you?",
3     "I love to read books.",
4     "The weather is nice today.",
5     "Where is the nearest restaurant?",
6     "What time is it?",
7     "I enjoy playing soccer.",
8     "Can you help me with this?",
9     "I'm going to the movies tonight.",
10    "This is a beautiful place.",
11    "I like listening to music.",
12    "Do you speak English?",
13    "What is your favorite color?",
14    "I'm learning to play the guitar.",
15    "Have a great day!",
```

```
    "I need to buy some groceries.",
    "Let's go for a walk.",
    "How was your weekend?",
    "I'm excited for the concert.",
    "Could you pass me the salt, please?",
    "I have a meeting at 2 PM.",
    "I'm planning a vacation.",
    "She sings beautifully.",
    "The cat is sleeping.",
    "I want to learn French.",
    "I enjoy going to the beach.",
    "Where can I find a taxi?",
    "I'm sorry for the inconvenience.",
    "I'm studying for my exams.",
    "I like to cook dinner at home.",
    "Do you have any recommendations for restaurants?",
]
espanhol = [
    "Hola, cómo estás?",
    "Me encanta leer libros.",
    "El clima está agradable hoy.",
    "Dónde está el restaurante más cercano?",
    "Qué hora es?",
    "Voy al parque todos los días.",
    "Puedes ayudarme con esto?",
    "Me gustaría ir de vacaciones.",
    "Este es mi libro favorito.",
    "Me gusta bailar salsa.",
    "Hablas espaol?",
    "Cuál es tu comida favorita?",
    "Estoy aprendiendo a tocar el piano.",
    "Que tengas un buen día!",
    "Necesito comprar algunas frutas.",
    "Vamos a dar un paseo.",
    "Cómo estuvo tu fin de semana?",
    "Estoy emocionado por el concierto.",
    "Me pasas la sal, por favor?",
    "Tengo una reunión a las 2 PM.",
    "Estoy planeando unas vacaciones.",
    "Ella canta hermosamente.",
    "El perro está jugando.",
    "Quiero aprender italiano.",
    "Disfruto ir a la playa.",
    "Dónde puedo encontrar un taxi?",
    "Lamento las molestias.",
    "Estoy estudiando para mis exámenes.",
    "Me gusta cocinar la cena en casa.",
    "Tienes alguna recomendación de restaurantes?",
]
portugues = [
    "Estou indo para o trabalho agora.",
```

```
    "Adoro passar tempo com minha família.",
    "Preciso comprar leite e pão.",
    "Vamos ao cinema no sábado.",
    "Gosto de praticar esportes ao ar livre.",
    "O trânsito está terrível hoje.",
    "A comida estava deliciosa!",
    "Você já visitou o Rio de Janeiro?",
    "Tenho uma reunião importante amanhã.",
    "A festa começa às 20h.",
    "Estou cansado depois de um longo dia de trabalho.",
    "Vamos fazer um churrasco no final de semana.",
    "O livro que estou lendo émuito interessante.",
    "Estou aprendendo a cozinhar pratos novos.",
    "Preciso fazer exercícios físicos regularmente.",
    "Vou viajar para o exterior nas férias.",
    "Você gosta de dançar?",
    "Hoje émeu aniversário!",
    "Gosto de ouvir música clássica.",
    "Estou estudando para o vestibular.",
    "Meu time de futebol favorito ganhou o jogo.",
    "Quero aprender a tocar violão.",
    "Vamos fazer uma viagem de carro.",
    "O parque fica cheio aos finais de semana.",
    "O filme que assisti ontem foi ótimo.",
    "Preciso resolver esse problema o mais rápido possível.",
    "Adoro explorar novos lugares.",
    "Vou visitar meus avós no domingo.",
    "Estou ansioso para as férias de verão.",
    "Gosto de fazer caminhadas na natureza.",
    "O restaurante tem uma vista incrível.",
    "Vamos sair para jantar no sábado."
]

import random
import re
import pandas as pd

pre_padroes = []
for frase in ingles:
pre_padroes.append( [frase, 'inglês'])
for frase in espanhol:
pre_padroes.append( [frase, 'espanhol'])

for frase in portugues:
    pre_padroes.append( [frase, 'português'])
random.shuffle(pre_padroes)

dados = pd.DataFrame(pre_padroes)

def tamanhoMedioFrases(texto):
    palavras = re.split("\s", texto)
```

```
118     tamanhos = [len(s) for s in palavras if len(s)>0]
119     #print(palavras)
120     #print(tamanhos)
121     soma = 0
122     for t in tamanhos:
123         soma=soma+t
124     return soma / len(tamanhos)
125 
126 def verificaCaracteresEspeciais(frase):
127 # Criação de dicionário com regras de cada idioma
128     dictRegex = {
129         'nTio': r'[\~n ]',
130         'exclamacaoInvertido': r'[\textexclamdown]',
131         'interrogacaoInvertido': r'[\textquestiondown]',
132         'cedilhas': r'[ç]',
133         'vogalAcentoAgudo': r'[áéíóú]',
134         'vogalAcentoCrase': r'[à]',
135         'vogalAcentoCircunflexo': r'[âêô]',
136         'digrafos': r'ch|lh|nh|gu|qu|rr|ss|sc|sç|xc|ll|nn|cc',
137         'digrafosVogal':'ae|ai|ao|au|ea|ei|eo|eu|ia|ie|io|iu|oa|oe|oi|
              ou|ua|ue|ui|uo',
138         'repeticaoVogal': r'aa|ee|ii|oo',
139         'ditongos': r'ãe|ãe|õe',
140         'apostrofe': r"[']",
141         'formatoHoras': r'PM|AM',
142         'consoantesReservadas': r'[kwxy]'
143     }
144 
145 def qtdOcorrencias(chave): return len(re.findall(dictRegex[chave],
      frase, re.IGNORECASE))
146     dictQtdCaracteresEspeciais = {
147     'qtdNtio': qtdOcorrencias('nTio'),
148     'qtdExclamacaoInvertido': qtdOcorrencias('exclamacaoInvertido'),
149     'qtdInterrogacaoInvertido': qtdOcorrencias('interrogacaoInvertido'
          ),
150     'qtdChedilhas': qtdOcorrencias('cedilhas'),
151     'qtdVogalAcentoCrase': qtdOcorrencias('vogalAcentoCrase'),
152     'qtdVogalAcentoAgudo': qtdOcorrencias('vogalAcentoAgudo'),
153     'qtdVogalAcentoCircunflexo': qtdOcorrencias('
          vogalAcentoCircunflexo'),
154     'qtdDigrafos': qtdOcorrencias('digrafos'),
155     'qtdDigrafosVogal': qtdOcorrencias('digrafosVogal'),
156     'qtdDitongos': qtdOcorrencias('ditongos'),
157     'qtdApostrofe': qtdOcorrencias('apostrofe'),
158     'qtdRepeticaoVogal': qtdOcorrencias('repeticaoVogal'),
159     'qtdHoras': qtdOcorrencias('formatoHoras'),
160     'qtdConsoantesReservadas': qtdOcorrencias('consoantesReservadas')
161     }
162     print(frase, '-', dictQtdCaracteresEspeciais.values())
163     return dictQtdCaracteresEspeciais.values()
164 
```

```
165 def extraiCaracteristicas(frase):
166     texto = frase[0]
167     pattern_regex = re.compile('[^\w+]', re.UNICODE)
168     texto = re.sub(pattern_regex,' ',texto)
169     caracteristica1=tamanhoMedioFrases(texto)
170     caracteristica2=verificaCaracteresEspeciais(frase[0])
171     padrao = []
172     padrao.append(caracteristica1)
173     padrao.extend(caracteristica2)
174     padrao.append(frase[1])
175     return padrao
176 
177 def geraPadroes(frases):
178     padroes = []
179     for frase in frases:
180         padrao = extraiCaracteristicas(frase)
181         padroes.append(padrao)
182     return padroes
183 
184 padroes = geraPadroes(pre_padroes)
185 dados = pd.DataFrame(padroes)
186 
187 from sklearn.model_selection import train_test_split
188 import numpy as np
189 
190 vet = np.array(padroes)
191 classes = vet[:,-1] # classes = [p[-1] for p in padroes]
192 padroes_sem_classe = vet[:,0:-1]
193 
194 X_train, X_test, y_train, y_test = train_test_split(
        padroes_sem_classe, classes, test_size=0.25, stratify=classes)
195 
196 from sklearn import svm
197 from sklearn.metrics import confusion_matrix
198 from sklearn.metrics import classification_report
199 treinador = svm.SVC() #algoritmo escolhido
200 modelo = treinador.fit(X_train, y_train)
201 
202 acuracia = modelo.score(X_train, y_train)
203 print("Acurácia nos dados de treinamento: {:.2f}%".format(acuracia *
        100))
204 
205 # melhor avaliar com a matriz de confusão
206 y_pred = modelo.predict(X_train)
207 cm = confusion_matrix(y_train, y_pred)
208 
209 print(classification_report(y_train, y_pred))
210 
211 # score com os dados de teste
212 acuracia = modelo.score(X_test, y_test)
213 print("Acurácia nos dados de teste: {:.2f}%".format(acuracia * 100))
```

```
214
215 print('métricas mais confiáveis')
216 y_pred2 = modelo.predict(X_test)
217 cm = confusion_matrix(y_test, y_pred2)
218 print(cm)
219 print(classification_report(y_test, y_pred2))
```

Acurácia nos dados de treinamento: 75.36%
[[12 4 6]
[0 21 2]
[5 0 19]]
precision recall f1-score support
espanhol 0.71 0.55 0.62 22
inglês 0.84 0.91 0.87 23
português 0.70 0.79 0.75 24
accuracy 0.75 69
macro avg 0.75 0.75 0.75 69
weighted avg 0.75 0.75 0.75 69
Acurácia nos dados de teste: 82.61%
métricas mais confiáveis
[[5 1 2]
[1 6 0]
[0 0 8]]
precision recall f1-score support
espanhol 0.83 0.62 0.71 8
inglês 0.86 0.86 0.86 7
português 0.80 1.00 0.89 8
accuracy 0.83 23
macro avg 0.83 0.83 0.82 23
weighted avg 0.83 0.83 0.82 23

2 Melhore uma base de dados ruim

```
1 import requests, zipfile, io
2 from io import BytesIO
3 import numpy as np
4 import pandas as pd
5 from sklearn import svm
6 from sklearn.metrics import confusion_matrix
7 from sklearn.metrics import classification_report
8 from sklearn.preprocessing import LabelEncoder
9 # link dos dados: https://www.archive.ics.uci.edu/dataset/14/breast+
     cancer
10
11 data = pd.read_csv('/content/breast-cancer.data', header=None,
     lineterminator='\n')
12 data.columns = ['class', 'age', 'menopause', 'tumor-size', 'inv-nodes
     ', 'node-caps', 'deg-malig', 'breast', 'breast-quad', 'irradiat']
```

Por meio da análise dos dados presentes em cada coluna onde é possível tirar as seguintes conclusões:

- Os dados possuem natureza categórica
- Temos um desbalanceamento grande entre as classes
- Existe pouca prevalência de pessoas mais novas e de pessoas muito mais velhas
- Maior parte dos casos estão na pré-menopausa
- Prevalência de inv-nodes discrepantes aos demais casos
- Ocorrências de câncer nas mamas esquerda e direita são aproximadas
- É possível perceber a presença do carácter ? em alguns campos

```
# transformando os valores presentes nas colunas em inteiros fatorizados
labelEncoder = LabelEncoder()
for col in data.columns: data[col] = labelEncoder.fit_transform(data[col])

new_data = pd.read_csv('/content/breast-cancer.data', header=None, lineterminator='\n')
new_data.columns = data.columns

# remoção das linhas que possuem o caracter ?
new_data_clean = new_data[~new_data.apply(lambda row: row.astype(str).str.contains('\?').any(), axis=1)]

new_data = new_data_clean.copy()
labelEncoder = LabelEncoder()
for col in new_data.columns: new_data[col] = labelEncoder.fit_transform(new_data[col])

# ao remover as colunas 'age', 'tumor-size e 'breast-quad' os valores do classificador melhoraram
# a conclusão acerca disso éque o atribuito idade possui uma grande variedade de intervalos, o que pode dificultar a separação dos grupos
# o mesmo acontece para o tamanho do tumor. Em relação ao local ( quadrante) onde o nódulo não foi significativo para a classificaçã o
# pois a informação sobre qual mama o nódulo já está presente na coluna 'breast', essa nova coluna só inclui o local mais preciso.

new_data = new_data.drop(columns=['age','tumor-size', 'breast-quad'])

X_orig = data.iloc[:, 1:].values
Y_orig = data.iloc[:, 0].values

X = new_data.iloc[:, 1:].values
Y = new_data.iloc[:, 0].values

```

```
28 from sklearn.model_selection import train_test_split
29 import numpy as np
30
31 # com os dados originais
32 X_orig_train, X_orig_test, y_orig_train, y_orig_test =
      train_test_split(X_orig,
33 Y_orig, test_size=0.25, stratify=Y_orig,random_state=10)
34
35 # com os dados tratados
36 X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size
      =0.25,
37 stratify=Y,random_state=10)
38
39 # VERIFICAÇÃO DO MODELO **SEM** TRATAMENTO DOS DADOS
40
41 treinador = svm.SVC() #algoritmo escolhido
42 modelo_orig = treinador.fit(X_orig_train, y_orig_train)
43
44 # predição com os mesmos dados usados para treinar
45 print('Matriz de confusão - com os dados ORIGINAIS usados no
      TREINAMENTO')
46 print('Acurácia:', "{:.2f}%".format(modelo_orig.score(X_orig_train,
      y_orig_train)*100))
47
48 y_orig_pred = modelo_orig.predict(X_orig_train)
49 cm_orig_train = confusion_matrix(y_orig_train, y_orig_pred)
50 print(cm_orig_train)
51 print(classification_report(y_orig_train, y_orig_pred))
52
53 # predição com os mesmos dados usados para testar
54 print('Matriz de confusão - com os dados ORIGINAIS usados para TESTES
      ')
55 print('Acurácia:', "{:.2f}%".format(modelo_orig.score(X_orig_test,
      y_orig_test)*100))
56 y2_orig_pred = modelo_orig.predict(X_orig_test)
57 cm_orig_test = confusion_matrix(y_orig_test, y2_orig_pred)
58 print(cm_orig_test)
59 print(classification_report(y_orig_test, y2_orig_pred))
```

Matriz de confusão - com os dados ORIGINAIS usados no TREINAMENTO
Acurácia: 76.17%
[[146 4]
[47 17]]
precision recall f1-score support
0 0.76 0.97 0.85 150
1 0.81 0.27 0.40 64
accuracy 0.76 214
macro avg 0.78 0.62 0.63 214
weighted avg 0.77 0.76 0.72 214
Matriz de confusão - com os dados ORIGINAIS usados para TESTES
Acurácia: 76.39%

[[47 4]
[13 8]]
precision recall f1-score support
0 0.78 0.92 0.85 51
1 0.67 0.38 0.48 21
accuracy 0.76 72
macro avg 0.72 0.65 0.67 72
weighted avg 0.75 0.76 0.74 72

```
# VERIFICAÇÃO DO MODELO **COM** TRATAMENTO DOS DADOS

treinador = svm.SVC() #algoritmo escolhido
modelo = treinador.fit(X_train, y_train)
# predição com os mesmos dados usados para treinar
print('Matriz de confusão - com os dados TRATADOS usados no TREINAMENTO')
print('Acurácia:', "{:.2f}\%".format(modelo.score(X_train, y_train)*100))
y_pred = modelo.predict(X_train)
cm_train = confusion_matrix(y_train, y_pred)
print(cm_train)
print(classification_report(y_train, y_pred))

# predição com os mesmos dados usados para testar
print('Matriz de confusão - com os dados ORIGINAIS usados para TESTES')
print('Acurácia:', "{:.2f}%".format(modelo.score(X_test, y_test)*100))
y2_pred = modelo.predict(X_test)
cm_test = confusion_matrix(y_test, y2_pred)
print(cm_test)
print(classification_report(y_test, y2_pred))
```

Matriz de confusão - com os dados TRATADOS usados no TREINAMENTO
Acurácia: 78.74%
[[142 4]
[40 21]]
precision recall f1-score support
0 0.78 0.97 0.87 146
1 0.84 0.34 0.49 61
accuracy 0.79 207
macro avg 0.81 0.66 0.68 207
weighted avg 0.80 0.79 0.75 207
Matriz de confusão - com os dados ORIGINAIS usados para TESTES
Acurácia: 81.43%
[[48 2]
[11 9]]
precision recall f1-score support
0 0.81 0.96 0.88 50
1 0.82 0.45 0.58 20
accuracy 0.81 70

macro avg 0.82 0.70 0.73 70
weighted avg 0.81 0.81 0.79 70

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

Classificação - Veículos

Tabela 7.1: Veículos - Resultados por técnica com parâmetros, acurácia e matrizes de confusão.

Técnica	Parâmetro	Acurácia	Matriz de Confusão
SVM – CV	C=100, Sigma=0.01	0.8443	Reference Prediction bus opel saab van bus 40 0 1 0 opel 0 31 10 0 saab 0 10 31 0 van 3 1 1 39
RNA – CV	size=31, decay=0.7	0.8204	Reference Prediction bus opel saab van bus 41 0 0 0 opel 0 27 13 0 saab 0 13 30 0 van 2 0 0 39
SVM – Hold-out	C=1, Sigma=0.07	0.7485	Reference Prediction bus opel saab van bus 40 0 0 1 opel 0 20 13 0 saab 0 20 26 0 van 3 0 0 39
RF – Hold-out	mtry=10	0.7425	Reference Prediction bus opel saab van bus 40 0 1 0 opel 0 19 13 0 saab 0 20 26 0 van 3 3 3 39

Continua na próxima página

Continuação da Tabela 7.1

Técnica	Parâmetro	Acurácia	Matriz de Confusão
RF – CV	mtry=5	0.7305	Reference Prediction bus opel saab van bus 40 0 1 0 opel 0 18 14 0 saab 0 21 25 0 van 3 3 0 39
RNA – Hold-out	size=5, decay=0.1	0.6707	Reference Prediction bus opel saab van bus 35 1 0 0 opel 0 0 0 0 saab 2 9 43 5 van 6 2 0 34
KNN	k=1	0.6353	Reference Prediction bus opel saab van bus 39 1 3 2 opel 1 18 19 2 saab 7 23 15 0 van 0 4 0 36

Classificação - Diabetes

Tabela 7.2: Diabetes - Resultados por técnica com parâmetros, acurácia e matriz de confusão.

Técnica	Parâmetro	Acurácia	Matriz de Confusão
SVM – Hold-out	C=0.5, Sigma=0.14	0.7647	Reference Prediction neg pos neg 92 28 pos 8 25
RF – Hold-out	mtry=2	0.7647	Reference Prediction neg pos neg 86 22 pos 14 31
RF – CV	mtry=9	0.7647	Reference Prediction neg pos neg 83 19 pos 17 34
SVM – CV	C=2, Sigma=0.015	0.7582	Reference Prediction neg pos neg 91 28 pos 9 25

Continua na próxima página

Continuação da Tabela 7.2

Técnica	Parâmetro	Acurácia	Matriz de Confusão
KNN	k=7	0.7273	Reference Prediction neg pos neg 85 31 pos 11 27
RNA – CV	size=21, decay=0.4	0.7059	Reference Prediction neg pos neg 81 26 pos 19 27
RNA – Hold-out	size=5, decay=0.1	0.6732	Reference Prediction neg pos neg 78 28 pos 22 25

Regressão - Admissão

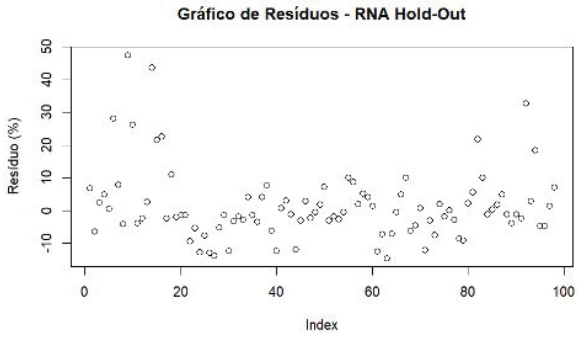
Tabela 7.3: Admissão - Resultados por técnica com parâmetros e métricas (R^2, S_{yx}, coeficiente de Pearson, RMSE e MAE).

Técnica	Parâmetro	R^2	S_{yx}	Pearson	RMSE	MAE
RNA – Hold-out	size=3, decay=1e-4	0.8234	0.0606	0.9083	0.0583	0.0434
RF – Hold-out	mtry=2	0.8231	0.0606	0.9089	0.0584	0.0431
RF – CV	mtry=2	0.8196	0.0612	0.9072	0.0590	0.0436
SVM – CV	C=50, Sigma=0.01	0.8160	0.0618	0.9063	0.0595	0.0430
SVM – Hold-out	C=1, Sigma=0.18	0.8014	0.0642	0.9007	0.0619	0.0438
KNN	k=9	0.7384	0.0737	0.8600	0.0710	0.0547
RNA – CV	size=41, decay=0.1	0.7344	0.0743	0.8703	0.0716	0.0534

Tabela 7.4: Amostra com GRE, TOEFL, rating universitário e métricas preditivas.

GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	predict.rna
316.99	110.72	4.66	1.19	2.20	9.70	0.71	0.89
303.68	99.41	2.67	4.11	2.01	8.08	0.41	0.59
295.93	97.60	4.72	3.48	2.84	9.80	0.80	0.82

Figura 7.1 – ADMISSÃO - GRÁFICO DE RESÍDUOS (MELHOR MODELO)



Regressão - Biomassa

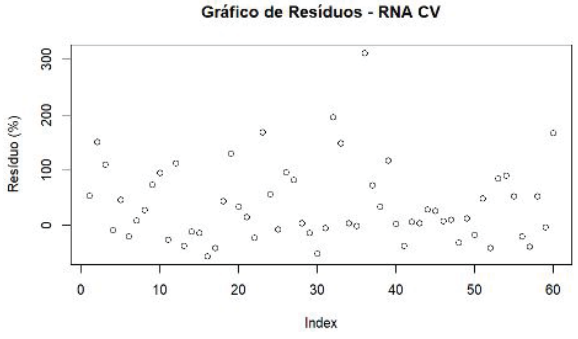
Tabela 7.5: Biomassa - Resultados por técnica com parâmetros e métricas (R^2, S_{yx}, coeficiente de Pearson, RMSE e MAE).

Técnica	**Parâmetro**	R^2	S_{yx}	**Pearson**	**RMSE**	**MAE**
RNA – CV	size=11, decay=0.4	0.9877	141.04	0.9939	137.47	64.407
RF – CV	mtry=5	0.9846	163.73	0.9923	153.88	61.807
RF – Hold-out	mtry=2	0.9843	165.45	0.9921	155.50	63.321
SVM – CV	C=100, Sigma=0.01	0.9448	310.55	0.9881	291.87	94.531
KNN	k=3	0.8576	498.83	0.9578	468.83	124.52
RNA – Hold-out	size=5, decay=0.1	0.5433	861.62	0.8205	839.80	186.86
SVM – Hold-out	C=1, Sigma=0.60	0.3435	1071.3	0.6175	1006.89	210.16

Tabela 7.6: Amostra com variáveis *dap*, *h*, *Me* e predição da RNA.

dap	**h**	**Me**	**predict.rna**
132.31429	23.76053	0.4378101	16744.171
55.74442	48.14359	0.6876391	2724.243
45.74574	14.74192	0.3821269	14206.252

Figura 7.2 – BIOMASSA - GRÁFICO DE RESÍDUOS (MELHOR MODELO)



Agrupamento - Veículo

K-means clustering with 10 clusters of size 75, 67, 54, 131, 24, 137, 78, 109, 70, 101

Tabela 7.7: K-means (médias por cluster) - Parte 1

Cl	Comp	Circ	DCirc	RadRa	PrAxisRa	MaxLRa	ScatRa	Elong	PrAxisRect	MaxLRect
1	89.96	39.17	75.71	168.43	65.49	9.40	149.69	44.17	19.00	135.45
2	95.99	45.96	90.64	196.25	64.34	8.31	182.42	35.90	21.49	148.67
3	85.67	36.74	57.31	122.44	56.19	5.44	120.81	55.72	17.22	128.41
4	85.61	43.20	70.41	138.50	58.93	7.20	148.49	45.30	18.94	143.68
5	107.25	55.54	103.42	190.25	55.79	5.79	248.71	26.88	27.08	168.08
6	104.12	53.66	102.67	201.88	62.42	10.47	217.18	30.70	24.42	169.28
7	102.14	49.71	100.51	204.21	63.67	9.37	200.67	32.81	22.95	156.59
8	88.88	38.89	66.83	138.25	57.81	7.27	134.41	49.92	18.03	135.41
9	93.47	41.33	80.15	188.79	66.96	9.14	165.79	39.61	20.14	139.43
10	91.05	45.54	80.15	157.95	63.08	10.02	156.26	43.02	19.56	152.19

Tabela 7.8: K-means (médias por cluster) - Parte 2

Cl	ScVarMaxis	ScVarmaxis	RaGyr	SkewMaxis	Skewmaxis	Kurtmaxis	KurtMaxis	HolIRa
1	173.88	335.92	142.24	69.19	4.72	14.37	193.47	200.41
2	203.27	502.81	179.22	68.01	6.24	12.97	193.78	200.34
3	140.31	215.50	134.94	74.96	7.00	12.19	185.94	190.83
4	170.03	324.98	173.22	77.76	6.00	9.14	183.64	188.79
5	271.29	910.54	247.17	83.63	6.50	14.21	182.67	183.63
6	227.10	697.25	212.86	71.92	7.58	15.88	187.91	197.39
7	218.59	603.91	199.73	69.38	6.46	14.76	191.15	199.00
8	156.91	267.55	142.82	71.35	6.09	11.05	189.72	194.93
9	192.69	414.83	155.57	69.83	5.40	15.00	194.31	200.69
10	176.25	360.31	178.36	73.12	7.11	9.27	187.40	195.52

Associação - Musculação

Tabela 7.9: Regras de associação — top 10 linhas.

Lhs	Rhs	Support	Confidence	Coverage	Lift	Count
{V2=Gêmeos}	{V3=Bicicleta}	0.026	1.0	0.026	3.2	1
{V2=Gêmeos}	{V4=Esteira}	0.026	1.0	0.026	3.2	1
{V2=Gêmeos}	{V1=Extensor}	0.026	1.0	0.026	1.0	1
{V3=Gêmeos}	{V2=AgachamentoSmith}	0.051	1.0	0.051	3.9	2
{V3=Gêmeos}	{V4=Esteira}	0.051	1.0	0.051	3.2	2
{V4=Crucifixo}	{V3=Afundo}	0.051	1.0	0.051	6.5	2
{V4=Crucifixo}	{V2=Gêmeos}	0.051	1.0	0.051	5.5	2
{V4=Crucifixo}	{V1=LegPress}	0.051	1.0	0.051	2.6	2
{V2=Flexor}	{V3=Bicicleta}	0.051	1.0	0.051	3.2	2
{V2=Flexor}	{V4=Esteira}	0.051	1.0	0.051	3.2	2

Código

```
install.packages("e1071")
install.packages("caret")
install.packages("Metrics")

library("caret")
library("Metrics")

% CLASSIFICACAO - VEICULOS

setwd("/Users/blb/Documents/ESPECIALIZAÇÃO EM IA/APRENDIZADO DE MAQUINA/CLASSIFICACAO/")
dados <- read.csv("6 - Veiculos - Dados.csv")
dados[[a]] <- NULL

% KNN

set.seed(202410)
ran <- sample(1:nrow(dados), 0.8 * nrow(dados))
treino <- dados[ran,]
teste <- dados[-ran,]
set.seed(202410)
tuneGrid <- expand.grid(k = c(1,3,5,7,9))
knn <- train(tipo ~ ., data = treino, method = "knn",tuneGrid= tuneGrid)
predict.knn <- predict(knn, teste)
confusionMatrix(predict.knn, as.factor(teste[["tipo"]]))

% RNA HOLD OUT

set.seed(202410)
indices <- createDataPartition(dados[["tipo"]], p=0.80, list=FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
rna <- train(tipo ~ ., data=treino, method="nnet",trace=FALSE)
predicoes.rna <- predict(rna, teste)
confusionMatrix(predicoes.rna, as.factor(teste[["tipo"]]))

% RNA - Cross-Validation

ctrl <- trainControl(method = "cv", number = 10)
grid <- expand.grid(size = seq(from = 1, to = 45, by = 10),decay = seq(from = 0.1, to = 0.9,
by = 0.3))
set.seed(202410)
rna <- train(
   form = tipo~.,
   data = treino,
   method = "nnet",
   tuneGrid = grid,
   trControl = ctrl,
```

```
49     maxit = 2000,trace=FALSE)
50 predicoes.rna <- predict(rna, teste)
51 confusionMatrix(predicoes.rna, as.factor(teste[["tipo"]]))
52
53 % SVM - Hold-out
54
55 set.seed(202410)
56 indices <- createDataPartition(dados[["tipo"]], p=0.80, list=FALSE)
57 treino <- dados[indices,]
58 teste <- dados[-indices,]
59 set.seed(202410)
60 svm <- train(tipo~., data=treino, method="svmRadial")
61 predicoes.svm <- predict(svm, teste)
62 confusionMatrix(predicoes.svm, as.factor(teste[["tipo"]]))
63
64 % SVM - Cross-Validation
65
66 set.seed(202410)
67 indices <- createDataPartition(dados[["tipo"]], p=0.80, list=FALSE)
68 treino <- dados[indices,]
69 teste <- dados[-indices,]
70 ctrl <- trainControl(method = "cv", number = 10)
71 tuneGrid = expand.grid(C=c(1, 2, 10, 50, 100), sigma=c(.01, .015,
     0.2))
72 set.seed(202410)
73 svm <- train(tipo~., data=treino, method="svmRadial", trControl=ctrl,
      tuneGrid=tuneGrid)
74 predicoes.svm <- predict(svm, teste)
75 confusionMatrix(predicoes.svm, as.factor(teste[["tipo"]]))
76
77 % RANDOM FOREST - Hold-out
78
79 dados_aux <- dados
80 dados_aux[["tipo"]] <- NULL
81 newdata <-lapply(dados_aux, function(coluna) {
82 return(runif(3, min = min(coluna), max = max(coluna)))
83 })
84 predicoes.svm <- predict(svm, data.frame(newdata))
85 resultado <- cbind(data.frame(newdata), predicoes.svm)
86 View(resultado)
87
88 set.seed(202410)
89 indices <- createDataPartition(dados[["tipo"]], p=0.80, list=FALSE)
90 treino <- dados[indices,]
91 teste <- dados[-indices,]
92 set.seed(202410)
93 rf <- train(tipo~., data=treino, method="rf")
94
95 predicoes.rf <- predict(rf, teste)
96 confusionMatrix(predicoes.rf, as.factor(teste[["tipo"]]))
97
```

```
% RANDOM FOREST - Cross-Validation

set.seed(202410)
indices <- createDataPartition(dados[["tipo"]], p=0.80, list=FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
ctrl <- trainControl(method = "cv", number = 10)
tuneGrid = expand.grid(mtry=c(2, 5, 7, 9))
set.seed(202410)
rf <- train(tipo~., data=treino, method="rf", trControl=ctrl,
   tuneGrid=tuneGrid)
predicoes.rf <- predict(rf, teste)
confusionMatrix(predicoes.rf, as.factor(teste[["tipo"]]))
```

```
% CLASSIFICACAO - DIABETES

setwd("/Users/blb/Documents/ESPECIALIZAÇÃO EM IA/APRENDIZADO DE
   MAQUINA/CLASSIFICACAO/")
dados <- read.csv("10 - Diabetes - Dados.csv")
dados[["num"]] <- NULL

% KNN

set.seed(202410)
ran <- sample(1:nrow(dados), 0.8 * nrow(dados))
treino <- dados[ran,]
teste <- dados[-ran,]
set.seed(202410)
tuneGrid <- expand.grid(k = c(1,3,5,7,9))
knn <- train(diabetes ~ ., data = treino, method = "knn",tuneGrid=
   tuneGrid)
predict.knn <- predict(knn, teste)
confusionMatrix(predict.knn, as.factor(teste[["diabetes"]]))

% RNA - Hold-out

set.seed(202410)
indices <- createDataPartition(dados[["diabetes"]], p=0.80, list=
   FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
rna <- train(diabetes ~ ., data=treino, method="nnet",trace=FALSE)
predicoes.rna <- predict(rna, teste)
confusionMatrix(predicoes.rna, as.factor(teste[["diabetes"]]))

% RNA - Cross-Validation

ctrl <- trainControl(method = "cv", number = 10)
   grid <- expand.grid(size = seq(from = 1, to = 45, by = 10),decay =
       seq(from = 0.1, to = 0.9,
```

```
    by = 0.3))
    set.seed(202410)
    rna <- train(
    form = diabetes~.,
    data = treino,
    method = "nnet",
    tuneGrid = grid,
    trControl = ctrl,
    maxit = 2000,trace=FALSE)
predicoes.rna <- predict(rna, teste)
confusionMatrix(predicoes.rna, as.factor(teste[["diabetes"]]))

% SVM - Hold-out

set.seed(202410)
indices <- createDataPartition(dados[["diabetes"]], p=0.80, list=
    FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
svm <- train(diabetes~., data=treino, method="svmRadial")
predicoes.svm <- predict(svm, teste)
confusionMatrix(predicoes.svm, as.factor(teste[["diabetes"]]))

dados_aux <- dados
dados_aux[["diabetes"]] <- NULL
newdata <-lapply(dados_aux, function(coluna) {
return(runif(3, min = min(coluna), max = max(coluna)))
})
predicoes.svm <- predict(svm, data.frame(newdata))
resultado <- cbind(data.frame(newdata), predicoes.svm)

% SVM - Cross-Validation

set.seed(202410)
indices <- createDataPartition(dados[["diabetes"]], p=0.80, list=
    FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
ctrl <- trainControl(method = "cv", number = 10)
tuneGrid = expand.grid(C=c(1, 2, 10, 50, 100), sigma=c(.01, .015,
    0.2))
set.seed(202410)
svm <- train(diabetes~., data=treino, method="svmRadial", trControl=
    ctrl, tuneGrid=tuneGrid)
predicoes.svm <- predict(svm, teste)
confusionMatrix(predicoes.svm, as.factor(teste[["diabetes"]]))

% RANDOM FOREST - Hold-out

set.seed(202410)
```

```
indices <- createDataPartition(dados[["diabetes"]], p=0.80, list=
    FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
rf <- train(diabetes~., data=treino, method="rf")
predicoes.rf <- predict(rf, teste)
confusionMatrix(predicoes.rf, as.factor(teste[["diabetes"]]))

% RANDOM FOREST - Cross-Validation

set.seed(202410)
indices <- createDataPartition(dados[["diabetes"]], p=0.80, list=
    FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
ctrl <- trainControl(method = "cv", number = 10)
tuneGrid = expand.grid(mtry=c(2, 5, 7, 9))
set.seed(202410)
rf <- train(diabetes~., data=treino, method="rf", trControl=ctrl,
    tuneGrid=tuneGrid)
predicoes.rf <- predict(rf, teste)
confusionMatrix(predicoes.rf, as.factor(teste[["diabetes"]]))
```

```
% REGRESSAO - ADMISSAO

setwd("/Users/blb/Documents/ESPECIALIZAÇÃO EM IA/APRENDIZADO DE MÁ
    QUINA/REGRESSÃO/")
dados <- read.csv("9 - Admissao - Dados.csv", header=T)
dados[["num"]] <- NULL

r2 <- function(predito, observado) {
return(1 - (sum((predito-observado)^2) / sum((observado-mean(
    observado))^2)))
}

syx <- function(predito, observado, num_variaveis) {
return(sqrt((sum((predito-observado)^2) / (length(predito) -
    num_variaveis))))
}

regression_metrics <- function(predito, observado){
    r2_aux <- r2(predito, observado)
    syx_aux <- syx(predito, observado, 7)
    pearson_aux <- cor(predito, observado, method='pearson')
    rsme_aux <- rmse(predito, observado)
    mae_aux <- mae(predito, observado)
    return(c(r2_aux, syx_aux, pearson_aux, rsme_aux, mae_aux))
}

set.seed(202410)
```

```
ind <- createDataPartition(dados[["ChanceOfAdmit"]], p=0.80, list = FALSE)
treino <- dados[ind,]
teste <- dados[-ind,]
tuneGrid <- expand.grid(k = c(1,3,5,7,9))
set.seed(202410)
knn <- train(ChanceOfAdmit ~ ., data = treino, method = "knn", tuneGrid=tuneGrid)
predict.knn <- predict(knn, teste)
regression_metrics(predict.knn,teste[["ChanceOfAdmit"]])

% RNA - Hold-out

set.seed(202410)
indices <- createDataPartition(dados[["ChanceOfAdmit"]], p=0.80, list =FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
rna <- train(ChanceOfAdmit ~ ., data=treino, method="nnet",trace= FALSE)
predict.rna <- predict(rna, teste)
regression_metrics(predict.rna, teste[["ChanceOfAdmit"]])
residuos <- function(predito, observado){
return(((predito - observado)/observado) * 100)
}
plot(residuos(predict.rna, teste[["ChanceOfAdmit"]]), ylab='Resíduo (%)', main='Gráfico de Resíduo
s - RNA Hold-Out')

dados_aux <- dados
dados_aux[["ChanceOfAdmit"]] <- NULL
newdata <-lapply(dados_aux, function(coluna) {
   return(runif(3, min = min(coluna), max = max(coluna)))
})

predict.rna <- predict(rna, data.frame(newdata))
resultado <- cbind(data.frame(newdata), predict.rna)

% RNA - Cross-Validation

ctrl <- trainControl(method = "cv", number = 10)
grid <- expand.grid(size = seq(from = 1, to = 45, by = 10),decay = seq(from = 0.1, to = 0.9,
by = 0.3))
set.seed(202410)
rna <- train(
   form = ChanceOfAdmit~.,
   data = treino,
   method = "nnet",
   tuneGrid = grid,
```

```
70    trControl = ctrl,
71    maxit = 2000,trace=FALSE)
72
73  predict.rna <- predict(rna, teste)
74  regression_metrics(predict.rna,teste[["ChanceOfAdmit"]])
75
76  % SVM - Hold-out
77
78  set.seed(202410)
79  indices <- createDataPartition(dados[["ChanceOfAdmit"]], p=0.80, list
      =FALSE)
80  treino <- dados[indices,]
81  teste <- dados[-indices,]
82  set.seed(202410)
83  svm <- train(ChanceOfAdmit~., data=treino, method="svmRadial")
84  predict.svm <- predict(svm, teste)
85  regression_metrics(predict.svm,teste[["ChanceOfAdmit"]])
86
87  % SVM - Cross-Validation
88
89  set.seed(202410)
90  indices <- createDataPartition(dados[["ChanceOfAdmit"]], p=0.80, list
      =FALSE)
91  treino <- dados[indices,]
92  teste <- dados[-indices,]
93  ctrl <- trainControl(method = "cv", number = 10)
94  tuneGrid = expand.grid(C=c(1, 2, 10, 50, 100), sigma=c(.01, .015,
      0.2))
95  set.seed(202410)
96  svm <- train(ChanceOfAdmit~., data=treino, method="svmRadial",
      trControl=ctrl, tuneGrid=tuneG
97  rid)
98  predict.svm <- predict(svm, teste)
99  regression_metrics(predict.svm,teste[["ChanceOfAdmit"]])
100
101 % RANDOM FOREST - Hold-out
102
103 set.seed(202410)
104 indices <- createDataPartition(dados[["ChanceOfAdmit"]], p=0.80, list
      =FALSE)
105 treino <- dados[indices,]
106 teste <- dados[-indices,]
107 set.seed(202410)
108 rf <- train(ChanceOfAdmit~., data=treino, method="rf")
109 predict.rf <- predict(rf, teste)
110 regression_metrics(predict.rf,teste[["ChanceOfAdmit"]])
111
112 % RANDOM FOREST - Cross-Validation
113
114 set.seed(202410)
```

```
indices <- createDataPartition(dados[["ChanceOfAdmit"]], p=0.80, list
    =FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
ctrl <- trainControl(method = "cv", number = 10)
tuneGrid = expand.grid(mtry=c(2, 5, 7, 9))
set.seed(202410)
rf <- train(ChanceOfAdmit~., data=treino, method="rf", trControl=ctrl
    , tuneGrid=tuneGrid)
predict.rf <- predict(rf, teste)
regression_metrics(predict.rf,teste[["ChanceOfAdmit"]])
```

```
% REGRESSAO - BIOMASSA

setwd("/Users/blb/Documents/ESPECIALIZAÇÃO EM IA/APRENDIZADO DE MÁ
    QUINA/REGRESSÃO/")
dados <- read.csv("5 - Biomassa - Dados.csv", header=T)
dados[["num"]] <- NULL

% KNN

r2 <- function(predito, observado) {
return(1 - (sum((predito-observado)^2) / sum((observado-mean(
    observado))^2)))
}
syx <- function(predito, observado, num_variaveis) {
    return(sqrt((sum((predito-observado)^2) / (length(predito) -
        num_variaveis))))
}

regression_metrics <- function(predito, observado){
    r2_aux <- r2(predito, observado)
    syx_aux <- syx(predito, observado, 3)
    pearson_aux <- cor(predito, observado, method='pearson')
    rsme_aux <- rmse(predito, observado)
    mae_aux <- mae(predito, observado)
    return(c(r2_aux, syx_aux, pearson_aux, rsme_aux, mae_aux))
}

residuos <- function(predito, observado){
    return(((predito - observado)/observado) * 100)
}

set.seed(202410)
ind <- createDataPartition(dados[["biomassa"]], p=0.80, list = FALSE)
treino <- dados[ind,]
teste <- dados[-ind,]
tuneGrid <- expand.grid(k = c(1,3,5,7,9))
set.seed(202410)
knn <- train(biomassa ~ ., data = treino, method = "knn",tuneGrid=
    tuneGrid)
```

```
predict.knn <- predict(knn, teste)
regression_metrics(predict.knn,teste[["biomassa"]])


% RNA - Hold-out

set.seed(202410)
indices <- createDataPartition(dados[["biomassa"]], p=0.80, list=
   FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
rna <- train(biomassa ~ ., data=treino, method="nnet", trace=FALSE,
   linout=T)
predict.rna <- predict(rna, teste)
regression_metrics(predict.rna, teste[["biomassa"]])

% RNA - Cross-Validation

ctrl <- trainControl(method = "cv", number = 10)
grid <- expand.grid(size = seq(from = 1, to = 45, by = 10),decay =
   seq(from = 0.1, to = 0.9,
by = 0.3))
set.seed(202410)
rna <- train(
form = biomassa~.,
data = treino,
method = "nnet",
tuneGrid = grid,
trControl = ctrl,
maxit = 2000,trace=FALSE, linout=T)
predict.rna <- predict(rna, teste)
regression_metrics(predict.rna,teste[["biomassa"]])

dap = runif(3, min = min(dados[["dap"]]), max = max(dados[["dap"]]))
h = runif(3, min = min(dados[["h"]]), max = max(dados[["h"]]))
Me = runif(3, min = min(dados[["Me"]]), max = max(dados[["Me"]]))
newdata <- data.frame(dap=dap,h=h, Me=Me)
predict.rna <- predict(rna, newdata)
resultado <- cbind(newdata, predict.rna)

% SVM - Hold-out

set.seed(202410)
indices <- createDataPartition(dados[["biomassa"]], p=0.80, list=
   FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
svm <- train(biomassa~., data=treino, method="svmRadial")
predict.svm <- predict(svm, teste)
```

```
regression_metrics(predict.svm,teste[["biomassa"]])

% SVM - Cross-Validation

set.seed(202410)
indices <- createDataPartition(dados[["biomassa"]], p=0.80, list=
   FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
ctrl <- trainControl(method = "cv", number = 10)
tuneGrid = expand.grid(C=c(1, 2, 10, 50, 100), sigma=c(.01, .015,
   0.2))
set.seed(202410)
svm <- train(biomassa~., data=treino, method="svmRadial", trControl=
   ctrl, tuneGrid=tuneGrid)
predict.svm <- predict(svm, teste)
regression_metrics(predict.svm,teste[["biomassa"]])

% RANDOM FOREST - Hold-out

set.seed(202410)
indices <- createDataPartition(dados[["biomassa"]], p=0.80, list=
   FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
set.seed(202410)
rf <- train(biomassa~., data=treino, method="rf")
predict.rf <- predict(rf, teste)
regression_metrics(predict.rf,teste[["biomassa"]])

% RANDOM FOREST - Cross-Validation

set.seed(202410)
indices <- createDataPartition(dados[["biomassa"]], p=0.80, list=
   FALSE)
treino <- dados[indices,]
teste <- dados[-indices,]
ctrl <- trainControl(method = "cv", number = 10)
tuneGrid = expand.grid(mtry=c(2, 5, 7, 9))
set.seed(202410)
rf <- train(biomassa~., data=treino, method="rf", trControl=ctrl,
   tuneGrid=tuneGrid, linout=
T)
predict.rf <- predict(rf, teste)
regression_metrics(predict.rf,teste[["biomassa"]])
```

```
% AGRUPAMENTO - VEICULOS

setwd("/Users/blb/Documents/ESPECIALIZAÇÃO EM IA/APRENDIZADO DE MÁ
   QUINA/AGRUPAMENTO/")
dados <- read.csv("6 - Veiculos - Dados.csv")
```

```
dados[["a"]] <- NULL
dados[["tipo"]] <- NULL

set.seed(202410)
dados_normalizados <- scale(dados)
km.res = kmeans(dados, 10)
```

```
% REGRAS DE ASSOCIACAO - MUSCULACAO

install.packages('arules', dep=T)
library(arules)
library(datasets)

setwd("/Users/blb/Documents/ESPECIALIZAÇÃO EM IA/APRENDIZADO DE MÁQUINA/ASSOCIAÇÃO/")
dados <- read.csv("2 - Musculacao - Dados(in).csv", header=0, sep=';')

set.seed(202410)
rules <- apriori(dados, parameter = list(supp = 0.001, conf = 0.7, minlen=2))
summary(rules)
options(digits=2)
inspect(sort(rules[1:20], by="confidence"))
```

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

1 Classificação de Imagens

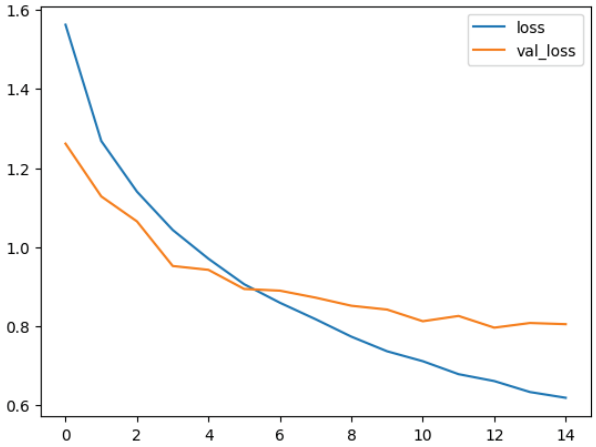
```
1  import tensorflow as tf
2  import numpy as np
3  import matplotlib.pyplot as plt
4  from tensorflow.keras.layers import Input, Conv2D, Dense, Flatten,
       Dropout
5  from tensorflow.keras.models import Model
6  from mlxtend.plotting import plot_confusion_matrix
7  from sklearn.metrics import confusion_matrix
8
9  cifar10 = tf.keras.datasets.cifar10
10 (x_train, y_train), (x_test, y_test) = cifar10.load_data()
11
12 x_train, x_test = x_train/255.0, x_test/255.0
13 y_train, y_test = y_train.flatten(), y_test.flatten()
14
15 print("x_train.shape", x_train.shape)
16 print("y_train.shape", y_train.shape)
17 print("x_test.shape", x_test.shape)
18 print("y_test.shape", y_test.shape)
```

x_train.shape (50000, 32, 32, 3)
y_train.shape (50000,)
x_test.shape (10000, 32, 32, 3)
y_test.shape (10000,)

```
1  K = len(set(y\_train))
2
3  i = Input(shape=x\_train[0].shape)
4
5  x = Conv2D(32, (3,3), strides=2, activation="relu")(i)
```

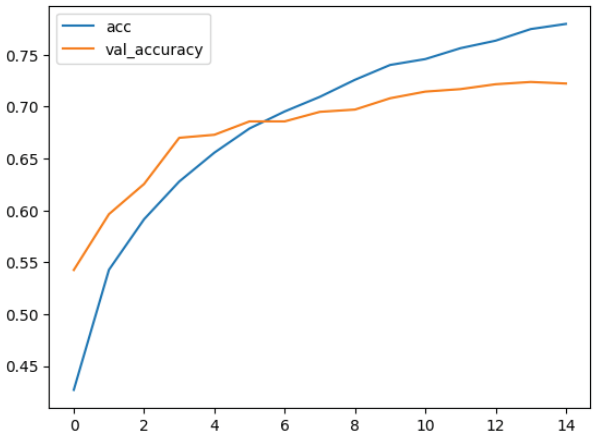
```
x = Conv2D(64, (3,3), strides=2, activation="relu")(x)
x = Conv2D(128, (3,3), strides=2, activation="relu")(x)

x = Flatten()(x)
x = Dropout(0.5)(x)
x = Dense(1024, activation="relu")(x)
x = Dropout(0.2)(x)
x = Dense(K, activation="softmax")(x)
model = Model(i, x)

model.summary()

model.compile(optimizer="adam", loss="sparse_categorical_crossentropy
    ", metrics=["accuracy"])

r = model.fit(x_train, y_train, validation_data=(x_test, y_test),
    epochs=15)

plt.plot(r.history["loss"], label="loss")
plt.plot(r.history["val_loss"], label="val_loss")
plt.legend()
plt.show()
```

Figura 8.1 – CLASSIFICAÇÃO DE IMAGENS COM CNN: GRÁFICO DE PERDA



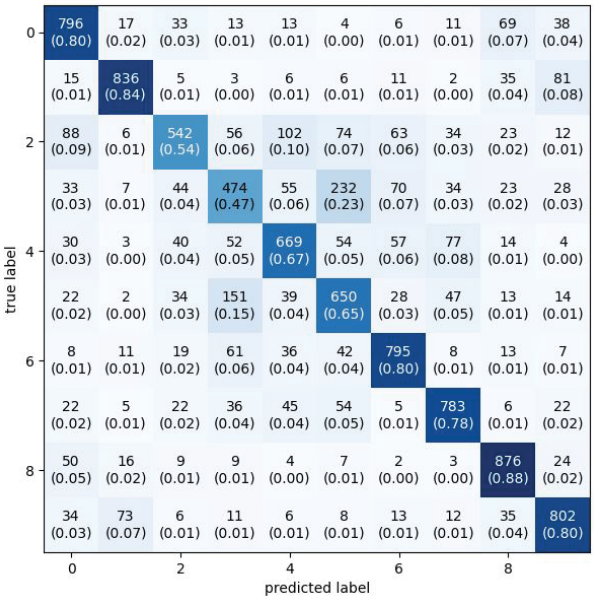

```
plt.plot(r.history["accuracy"], label="acc")
plt.plot(r.history["val_accuracy"], label="val_accuracy")
plt.legend()
plt.show()
```

Figura 8.2 – CLASSIFICAÇÃO DE IMAGENS COM CNN: GRÁFICO DE ACURÁCIA



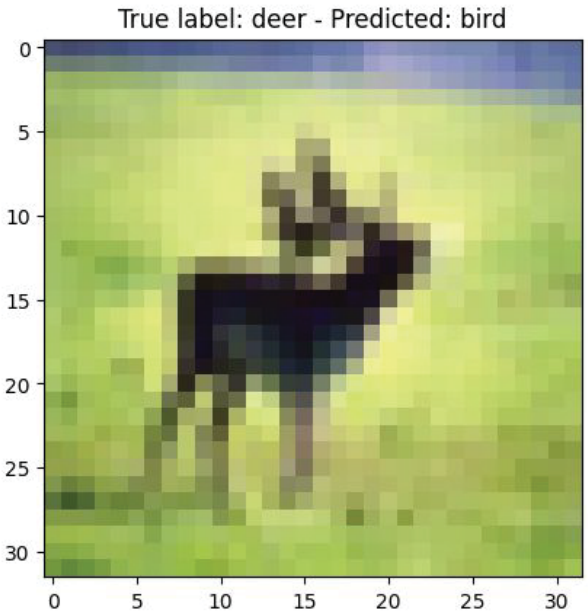

```
y_pred = model.predict(x_test).argmax(axis=1)
cm = confusion_matrix(y_test, y_pred)
plot_confusion_matrix(conf_mat=cm, figsize=(7,7), show_normed=True)
```

Figura 8.3 – CLASSIFICAÇÃO DE IMAGENS COM CNN: MATRIZ DE CONFUSÃO




```

abels = ["airplane", "automobile", "bird", "cat", "deer", "dog", "
    frog", "horse", "ship", "truck"]

misclassified = np.where(y\_pred != y\_test)[0]

i = np.random.choice(misclassified)

plt.imshow(x\_test[i], cmap="gray")
plt.title("True label: \%s - Predicted: \%s" \% (labels[y\_test[i]],
    labels[y\_pred[i]]))
```

Figura 8.4 – CLASSIFICAÇÃO DE IMAGENS COM CNN: RESULTADO DA PREDIÇÃO



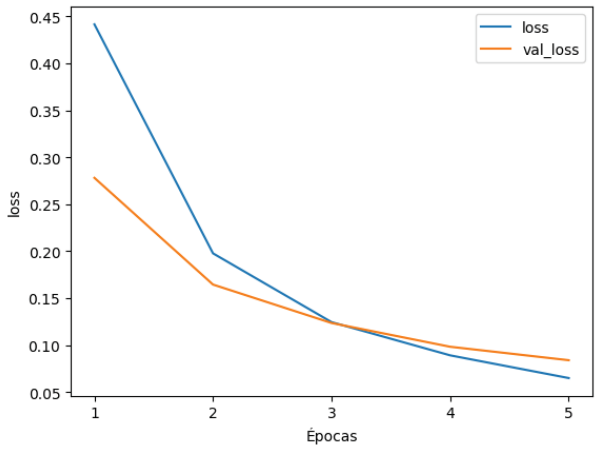
2 Detector de SPAM

```
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
from tensorflow.keras.layers import Input, Embedding, LSTM, Dense
from tensorflow.keras.layers import GlobalMaxPooling1D
from tensorflow.keras.models import Model
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.preprocessing.text import Tokenizer
from google.colab import files
import io

uploaded = files.upload()

df = pd.read_csv(io.BytesIO(uploaded['spam.csv']), encoding="ISO
    -8859-1")
df.head()

df = df.drop(["Unnamed: 2", "Unnamed: 3", "Unnamed: 4"], axis=1)
df.columns = ["labels", "data"]
df["b_labels"] = df["labels"].map({"ham": 0, "spam": 1})
y = df["b_labels"].values

x_train, x_test, y_train, y_test = train_test_split(df["data"], y,
    test_size=0.33)

num_words = 2000
tokenizer = Tokenizer(num_words=num_words)
```

```
tokenizer.fit_on_texts(x_train)
sequences_train = tokenizer.texts_to_sequences(x_train)
sequences_test = tokenizer.texts_to_sequences(x_test)
word2index = tokenizer.word_index
V = len(word2index)
print("%s tokens" % V)

data_train = pad_sequences(sequences_train)
T = data_train.shape[1]
data_test = pad_sequences(sequences_test, maxlen=T)
print("data_train.shape = ", data_train.shape)
print("data_test.shape = ", data_test.shape)

D = 20
M = 5

i = Input(shape=(T,))
x = Embedding(V + 1, D)(i)
x = LSTM(M)(x)
x = Dense(1, activation="sigmoid")(x)

model = Model(i,x)

model.compile(
   loss="binary_crossentropy",
   optimizer="adam",
   metrics=["accuracy"]
)

epochs = 5

r = model.fit(
   data_train, y_train,
   validation_data=(data_test, y_test),
   epochs=epochs
)

plt.plot(r.history["loss"], label="loss")
plt.plot(r.history["val_loss"], label="val_loss")
plt.xlabel("Épocas")
plt.ylabel("loss")
plt.xticks(np.arange(0, epochs, 1), labels=range(1, epochs + 1))
plt.legend()
plt.show()
```

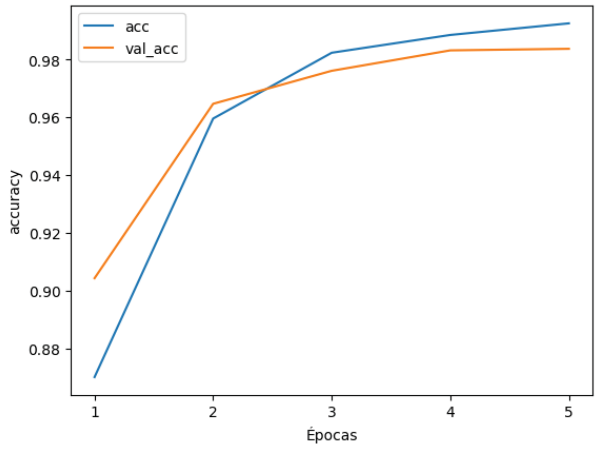
```
7242 tokens
data_train.shape = (3733, 175)
data_test.shape = (1839, 175)
```

Figura 8.5 – DETECTOR DE SPAM COM RNN: GRÁFICO DE PERDA




```
plt.plot(r.history["accuracy"], label="acc")
plt.plot(r.history["val_accuracy"], label="val_acc")
plt.xlabel("Épocas")
plt.ylabel("accuracy")
plt.xticks(np.arange(0, epochs, 1), labels=range(1, epochs + 1))
plt.legend()
plt.show()
```

Figura 8.6 – DETECTOR DE SPAM COM RNN: GRÁFICO DE ACURÁCIA




```
texto = "Is your car dirty? Discover our new product. Free for all.
    Click the link."
seq_texto = tokenizer.texts_to_sequences([texto])
data_texto = pad_sequences(seq_texto, maxlen=T)

pred = model.predict(data_texto)
print(pred)
print("SPAM" if pred >= 0.5 else "OK")
```

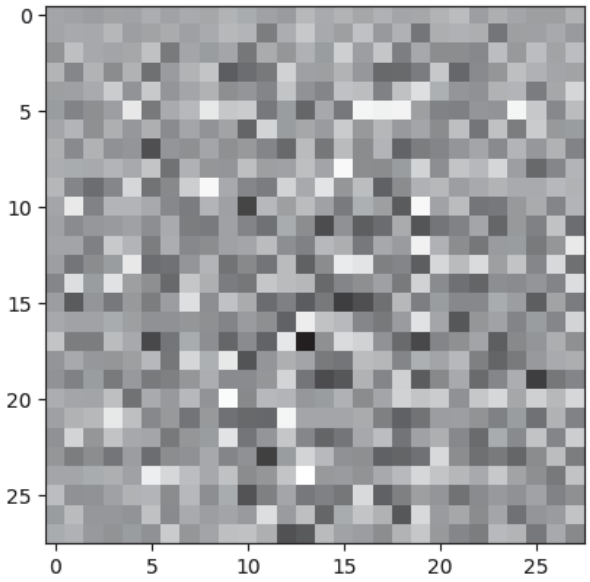
[[0.6281646]] SPAM

3 Gerador de Dígitos Fake

```
%pip install imageio
%pip install git+https://github.com/tensorflow/docs

import glob
import imageio
import numpy as np
import os
import PIL
import time
import tensorflow as tf
import matplotlib.pyplot as plt

from tensorflow.keras import layers
from IPython import display

(train_images, train_labels) , (_,_) = tf.keras.datasets.mnist.load_data()
train_images = train_images.reshape(train_images.shape[0], 28, 28, 1).astype('float32')
train_images = (train_images - 127.5) / 127.5

BUFFER_SIZE = 60000
BATCH_SIZE = 256

train_dataset = tf.data.Dataset.from_tensor_slices(train_images).shuffle(BUFFER_SIZE).batch(BATCH_SIZE)

def make_generator_model():
  model = tf.keras.Sequential()
  model.add(layers.Dense(7*7*256, use_bias=False, input_shape=(100,)))

  model.add(layers.BatchNormalization())
  model.add(layers.LeakyReLU())

  model.add(layers.Reshape((7, 7, 256)))
  assert model.output_shape == (None, 7, 7, 256)

  model.add(layers.Conv2DTranspose(128, (5, 5), strides=1, padding='same', use_bias=False))
  assert model.output_shape == (None, 7, 7, 128)

  model.add(layers.BatchNormalization())
  model.add(layers.LeakyReLU())

  model.add(layers.Conv2DTranspose(64, (5, 5), strides=(2, 2), padding='same', use_bias=False))
  assert model.output_shape == (None, 14, 14, 64)

  model.add(layers.BatchNormalization())
  model.add(layers.LeakyReLU())
```

```
46
47   model.add(layers.Conv2DTranspose(1, (5, 5), strides=(2, 2), padding
         ='same', use_bias=False, activation='tanh'))
48   assert model.output_shape == (None, 28, 28, 1)
49
50   return model
51
52 generator = make_generator_model()
53 noise = tf.random.normal([1, 100])
54 generated_image = generator(noise, training=False)
55 plt.imshow(generated_image[0, :, :, 0], cmap='gray')
```

Figura 8.7 – GERANDO DADOS COM GAN: GERAÇÃO DE IMAGEM RUIDOSA




```
1  def make_discriminator_model():
2    model = tf.keras.Sequential()
3    model.add(layers.Conv2D(64, (5, 5), strides=(2, 2), padding='same',
         input_shape=[28, 28, 1]))
4    model.add(layers.LeakyReLU())
5    model.add(layers.Dropout(0.3))
6
7    model.add(layers.Conv2D(128, (5, 5), strides=(2, 2), padding='same'
         ))
8    model.add(layers.LeakyReLU())
9    model.add(layers.Dropout(0.3))
10
11   model.add(layers.Flatten())
12   model.add(layers.Dense(1))
13
14   return model
15
16 discriminator = make_discriminator_model()
17 decision = discriminator(generated_image)
```

```
18
19 cross_entropy = tf.keras.losses.BinaryCrossentropy(from_logits=True)
20
21 def discriminator_loss(real_output, fake_output):
22   real_loss = cross_entropy(tf.ones_like(real_output), real_output)
23   fake_loss = cross_entropy(tf.zeros_like(fake_output), fake_output)
24   total_loss = real_loss + fake_loss
25   return total_loss
26
27 def generator_loss(fake_output):
28   return cross_entropy(tf.ones_like(fake_output), fake_output)
29
30 generator_optimizer = tf.keras.optimizers.Adam(1e-4)
31 discriminator_optimizer = tf.keras.optimizers.Adam(1e-4)
32
33 checkpoint_dir = './training_checkpoints'
34 checkpoint_prefix = os.path.join(checkpoint_dir, "ckpt")
35 checkpoint = tf.train.Checkpoint(generator_optimizer=
      generator_optimizer,
36                                  discriminator_optimizer=
                                       discriminator_optimizer,
37                                  generator=generator,
38                                  discriminator=discriminator)
39
40 EPOCHS = 100
41 noise_dim = 100
42 num_examples_to_generate = 16
43
44 seed = tf.random.normal([num_examples_to_generate, noise_dim])
45
46 @tf.function
47 def train_step(images):
48   noise = tf.random.normal([BATCH_SIZE, noise_dim])
49
50   with tf.GradientTape() as gen_tape, tf.GradientTape() as disc_tape:
51     generated_images = generator(noise, training=True)
52
53     real_output = discriminator(images, training=True)
54     fake_output = discriminator(generated_images, training=True)
55
56     gen_loss = generator_loss(fake_output)
57     disc_loss = discriminator_loss(real_output, fake_output)
58
59     gradients_of_generator = gen_tape.gradient(gen_loss, generator.
         trainable_variables)
60     gradients_of_discriminator = disc_tape.gradient(disc_loss,
         discriminator.trainable_variables)
61
62     generator_optimizer.apply_gradients(zip(gradients_of_generator,
         generator.trainable_variables))
```

```
    discriminator_optimizer.apply_gradients(zip(
        gradients_of_discriminator, discriminator.trainable_variables))

def train(dataset, epochs):
  for epoch in range(epochs):
    start = time.time()

    for image_batch in dataset:
      train_step(image_batch)

    display.clear_output(wait=True)
    generate_and_save_images(generator, epoch + 1, seed)

    if(epoch + 1) % 15 == 0:
      checkpoint.save(file_prefix = checkpoint_prefix)

    print('Time for epoch {} is {} sec' .format(epoch + 1, time.time()
        - start))

  display.clear_output(wait=True)
  generate_and_save_images(generator, epochs, seed)

def generate_and_save_images(model, epoch, test_input):
  predictions = model(test_input, training=False)
  fig = plt.figure(figsize=(4, 4))

  for i in range(predictions.shape[0]):
    plt.subplot(4, 4, i + 1)
    plt.imshow(predictions[i, :, :, 0] * 127.5 + 127.5, cmap='gray')
    plt.axis('off')

  plt.savefig('image_at_epoch_{:04d}.png'.format(epoch))
  plt.show()

train(train_dataset, EPOCHS)
checkpoint.restore(tf.train.latest_checkpoint(checkpoint_dir))
```

Figura 8.8 – GERANDO DADOS COM GAN: CHECKPOINT DA GERAÇÃO DE DÍGITOS

```
def display_image(epoch_no):
  return PIL.Image.open('image_at_epoch_{:04d}.png'.format(epoch_no))

display_image(EPOCHS)

anim_file = 'dcgan.gif'

with imageio.get_writer(anim_file, mode='I') as writer:
  filenames = glob.glob('image*.png')
  filenames = sorted(filenames)
  for filename in filenames:
    image = imageio.imread(filename)
    writer.append_data(image)
  image = imageio.imread(filename)
  writer.append_data(image)

import tensorflow_docs.vis.embed as embed
embed.embed_file(anim_file)
```

4 Transformer

```
!pip uninstall tensorflow
!pip install tensorflow==2.15.0
!pip install tensorflow_datasets
!pip install -U tensorflow-text==2.15.0
import collections
import logging
import os
import pathlib
import re
import string
import sys
```

```
import time
import numpy as np
import matplotlib.pyplot as plt
import tensorflow_datasets as tfds
import tensorflow_text as text
import tensorflow as tf
logging.getLogger('tensorflow').setLevel(logging.ERROR)

examples, metadata = tfds.load('ted_hrlr_translate/pt_to_en',
with_info=True, as_supervised=True)

train_examples, val_examples = examples['train'], examples['
    validation']

for pt_examples, en_examples in train_examples.batch(3).take(1):
  for pt in pt_examples.numpy():
    print(pt.decode('utf-8'))

  print()

  for en in en_examples.numpy():
    print(en.decode('utf-8'))
```

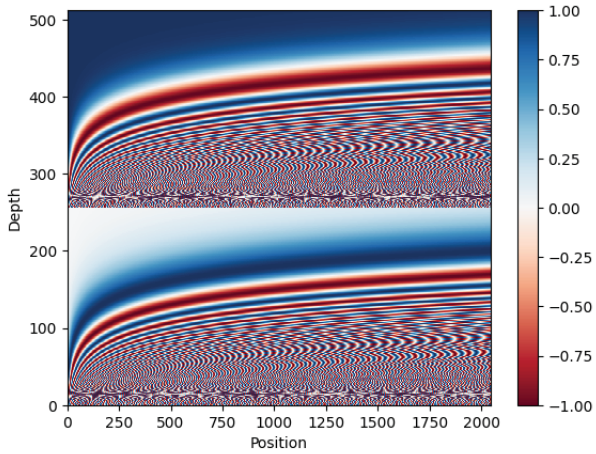
e quando melhoramos a procura , tiramos a única vantagem da impressão ,
que é a serendipidade .
mas e se estes fatores fossem ativos ?
mas eles não tinham a curiosidade de me testar .
and when you improve searchability , you actually take away the one
advantage of print , which is serendipity .
but what if it were active ?
but they did n't test for curiosity .

```
# Tokenização e Destokenização do texto
model_name = "ted_hrlr_translate_pt_en_converter"

tf.keras.utils.get_file(f"{model_name}.zip",
f"https://storage.googleapis.com/download.tensorflow.org/models/{
    model_name}.zip", cache_dir='.', cache_subdir='', extract=True)
# Tem 2 tokenizers: um pt outro em en
# tokenizers.en tokeniza e detokeniza
tokenizers = tf.saved_model.load(model_name)

# PIPELINE DE ENTRADA
# Codificar/tokenizar lotes de texto puro
def tokenize_pairs(pt, en):
  pt = tokenizers.pt.tokenize(pt)
  # Converte ragged (irregular, tam variável) para dense
  # Faz padding com zeros.
  pt = pt.to_tensor()

  en = tokenizers.en.tokenize(en)
  # ragged -> dense
```

```
  en = en.to_tensor()

  return pt, en

# Pipeline simpes: processa, embaralha, agrupa os dados, prefetch
# Datasets de entrada terminam com prefetch
BUFFER_SIZE = 20000
BATCH_SIZE = 64

def make_batches(ds):
  return (
    ds
    .cache()
    .shuffle(BUFFER_SIZE)
    .batch(BATCH_SIZE)
    .map(tokenize_pairs, num_parallel_calls=tf.data.AUTOTUNE)
    .prefetch(tf.data.AUTOTUNE))

train_batches = make_batches(train_examples)
val_batches = make_batches(val_examples)

# CODIFICAÇÃO POSICIONAL
def get_angles(pos, i, d_model):
  angle_rates = 1 / np.power(10000, (2 * (i//2)) / np.float32(d_model
      ))
  return pos * angle_rates

def positional_encoding(position, d_model):
  angle_rads = get_angles(np.arange(position)[:, np.newaxis], np.
      arange(d_model)[np.newaxis, :], d_model)

  # sin em índices pares no array; 2i
  angle_rads[:, 0::2] = np.sin(angle_rads[:, 0::2])

  # cos em índices ímpares no array; 2i+1
  angle_rads[:, 1::2] = np.cos(angle_rads[:, 1::2])

  # newaxis, aumenta a dimensão [] -> [ [] ]
  pos_encoding = angle_rads[np.newaxis, ...]

  return tf.cast(pos_encoding, dtype=tf.float32)

# CODIFICAÇÃO POSICIONAL
n, d = 2048, 512
pos_encoding = positional_encoding(n, d)
print(pos_encoding.shape)
pos_encoding = pos_encoding[0]

# Arrumar as dimensões
pos_encoding = tf.reshape(pos_encoding, (n, d//2, 2))
pos_encoding = tf.transpose(pos_encoding, (2, 1, 0))
```

```
pos_encoding = tf.reshape(pos_encoding, (d, n))

plt.pcolormesh(pos_encoding, cmap='RdBu')
plt.ylabel('Depth')
plt.xlabel('Position')
plt.colorbar()
plt.show()
```

Figura 8.9 – TRANFORMER: CODIFICAÇÃO POSICIONAL




```
# Cria uma máscara de 0 e 1, 0 para quando há valor e 1 quando não há
def create_padding_mask(seq):
  seq = tf.cast(tf.math.equal(seq, 0), tf.float32)

  # add extra dimensions to add the padding
  # to the attention logits.
  return seq[:, tf.newaxis, tf.newaxis, :] # (batch_size, 1, 1, seq_l
     )

# Máscara futura, usada no decoder
def create_look_ahead_mask(size):
  # zera o triângulo inferior
  mask = 1 - tf.linalg.band_part(tf.ones((size, size)), -1, 0)
  return mask # (seq_len, seq_len)

# Função de Atenção
def scaled_dot_product_attention(q, k, v, mask):
  # Q K^T
  matmul_qk = tf.matmul(q, k, transpose_b=True) # (..., seq_len_q,
     seq_len_k)

  # converte matmul_qk para float32
  dk = tf.cast(tf.shape(k)[-1], tf.float32)

  # divide por sqrt(d_k)
  scaled_attention_logits = matmul_qk / tf.math.sqrt(dk)

```

```
  # Soma a máscara, e os valores faltantes serão um número próximo a -inf
  if mask is not None:
    scaled_attention_logits += (mask * -1e9)

  # softmax normaliza os dados, soman 1. // (..., seq_len_q, seq_len_k)
  attention_weights = tf.nn.softmax(scaled_attention_logits, axis=-1)

  output = tf.matmul(attention_weights, v) # (..., seq_len_q, depth_v)

  return output, attention_weights
```

```
# Atenção Multi-cabeças
class MultiHeadAttention(tf.keras.layers.Layer):
  def __init__(self, d_model, num_heads):
    super(MultiHeadAttention, self).__init__()
    self.num_heads = num_heads
    self.d_model = d_model

    assert d_model % self.num_heads == 0

    self.depth = d_model // self.num_heads

    self.wq = tf.keras.layers.Dense(d_model)
    self.wk = tf.keras.layers.Dense(d_model)
    self.wv = tf.keras.layers.Dense(d_model)

    self.dense = tf.keras.layers.Dense(d_model)

  def split_heads(self, x, batch_size):
    """Separa a última dimensão em (num_heads, depth).
    Transpõe o resultado para o shape (batch_size, num_heads, seq_len, depth)
    """
    x = tf.reshape(x, (batch_size, -1, self.num_heads, self.depth))

    return tf.transpose(x, perm=[0, 2, 1, 3])

  def call(self, v, k, q, mask):
    batch_size = tf.shape(q)[0]

    q = self.wq(q) # (batch_size, seq_len, d_model)
    k = self.wk(k) # (batch_size, seq_len, d_model)
    v = self.wv(v) # (batch_size, seq_len, d_model)

    q = self.split_heads(q, batch_size) # (batch_size, num_heads, seq_len_q, depth)
    k = self.split_heads(k, batch_size) # (batch_size, num_heads, seq_len_k, depth)
```

```
    v = self.split_heads(v, batch_size) # (batch_size, num_heads,
        seq_len_v, depth)

    # Calcula a atenção para cada cabeça (de forma matricial)
    # scaled_attention.shape == (batch_size, num_heads, seq_len_q,
        depth)
    # attention_weights.shape == (batch_size, num_heads, seq_len_q,
        seq_len_k)
    scaled_attention, attention_weights = scaled_dot_product_attention
        (q, k, v, mask)

    # Troca a dimensão 2 com 1, para acertar o num_heads
    # (batch_size, seq_len_q, num_heads, depth)
    scaled_attention = tf.transpose(scaled_attention, perm=[0, 2, 1,
        3])

    # Concatena os valores em: (batch_size, seq_len_q, d_model)
    concat_attention = tf.reshape(scaled_attention, (batch_size, -1,
        self.d_model))

    output = self.dense(concat_attention) # (batch_size, seq_len_q,
        d_model)

    return output, attention_weights

def point_wise_feed_forward_network(d_model, dff):
  return tf.keras.Sequential([
    tf.keras.layers.Dense(dff, activation='relu'), # (batch_size,
        seq_len, dff)
    tf.keras.layers.Dense(d_model) # (batch_size, seq_len, d_model)
  ])
```

```
class EncoderLayer(tf.keras.layers.Layer):
  def __init__(self, d_model, num_heads, dff, rate=0.1):
    super(EncoderLayer, self).__init__()

    self.mha = MultiHeadAttention(d_model, num_heads)
    self.ffn = point_wise_feed_forward_network(d_model, dff)

    self.layernorm1 = tf.keras.layers.LayerNormalization(epsilon=1e-6)
    self.layernorm2 = tf.keras.layers.LayerNormalization(epsilon=1e-6)

    self.dropout1 = tf.keras.layers.Dropout(rate)
    self.dropout2 = tf.keras.layers.Dropout(rate)

  def call(self, x, training, mask):
    attn_output, _ = self.mha(x, x, x, mask) # (batch_size,
        input_seq_len, d_model)
    attn_output = self.dropout1(attn_output, training=training)
    out1 = self.layernorm1(x + attn_output) # (batch_size,
        input_seq_len, d_model)
```

```
18
19      ffn_output = self.ffn(out1) # (batch_size, input_seq_len, d_model)
20      ffn_output = self.dropout2(ffn_output, training=training)
21      out2 = self.layernorm2(out1 + ffn_output) # (batch_size,
            input_seq_len, d_model)
22
23      return out2
```

```
1  class DecoderLayer(tf.keras.layers.Layer):
2    def __init__(self, d_model, num_heads, dff, rate=0.1):
3      super(DecoderLayer, self).__init__()
4
5      self.mha1 = MultiHeadAttention(d_model, num_heads)
6      self.mha2 = MultiHeadAttention(d_model, num_heads)
7
8      self.ffn = point_wise_feed_forward_network(d_model, dff)
9
10     self.layernorm1 = tf.keras.layers.LayerNormalization(epsilon=1e-6)
11     self.layernorm2 = tf.keras.layers.LayerNormalization(epsilon=1e-6)
12     self.layernorm3 = tf.keras.layers.LayerNormalization(epsilon=1e-6)
13
14     self.dropout1 = tf.keras.layers.Dropout(rate)
15     self.dropout2 = tf.keras.layers.Dropout(rate)
16     self.dropout3 = tf.keras.layers.Dropout(rate)
17
18   def call(self, x, enc_output, training, look_ahead_mask,
         padding_mask):
19     # enc_output.shape == (batch_size, input_seq_len, d_model)
20     # (batch_size, target_seq_len, d_model)
21     attn1, attn_weights_block1 = self.mha1(x, x, x, look_ahead_mask)
22     attn1 = self.dropout1(attn1, training=training)
23     out1 = self.layernorm1(attn1 + x)
24
25     # (batch_size, target_seq_len, d_model)
26     attn2, attn_weights_block2 = self.mha2(enc_output, enc_output,
           out1, padding_mask)
27     attn2 = self.dropout2(attn2, training=training)
28     out2 = self.layernorm2(attn2 + out1) # (batch_size, target_seq_len
           , d_model)
29
30     ffn_output = self.ffn(out2) # (batch_size, target_seq_len, d_model
           )
31     ffn_output = self.dropout3(ffn_output, training=training)
32     out3 = self.layernorm3(ffn_output + out2) # (batch_size,
           target_seq_len, d_model)
33
34     return out3, attn_weights_block1, attn_weights_block2
```

```
1  class Encoder(tf.keras.layers.Layer):
2    def __init__(self, num_layers, d_model, num_heads, dff,
         input_vocab_size, maximum_position_encoding, rate=0.1):
```

```
    super(Encoder, self).__init__()

    self.d_model = d_model
    self.num_layers = num_layers
    self.embedding = tf.keras.layers.Embedding(input_vocab_size,
       d_model)
    self.pos_encoding = positional_encoding(maximum_position_encoding,
        self.d_model)
    self.enc_layers = [EncoderLayer(d_model, num_heads, dff, rate) for
        _ in range(num_layers)]
    self.dropout = tf.keras.layers.Dropout(rate)

  def call(self, x, training, mask):
    seq_len = tf.shape(x)[1]

    # adding embedding and position encoding.
    x = self.embedding(x) # (batch_size, input_seq_len, d_model)
    x *= tf.math.sqrt(tf.cast(self.d_model, tf.float32))
    x += self.pos_encoding[:, :seq_len, :]
    x = self.dropout(x, training=training)

    for i in range(self.num_layers):
      x = self.enc_layers[i](x, training, mask)

    return x # (batch_size, input_seq_len, d_model)
```

```
class Decoder(tf.keras.layers.Layer):
  def __init__(self, num_layers, d_model, num_heads, dff,
      target_vocab_size, maximum_position_encoding, rate=0.1):
    super(Decoder, self).__init__()

    self.d_model = d_model
    self.num_layers = num_layers

    self.embedding = tf.keras.layers.Embedding(target_vocab_size,
       d_model)
    self.pos_encoding = positional_encoding(maximum_position_encoding,
        d_model)

    self.dec_layers = [DecoderLayer(d_model, num_heads, dff, rate) for
        _ in range(num_layers)]
    self.dropout = tf.keras.layers.Dropout(rate)

  def call(self, x, enc_output, training, look_ahead_mask,
      padding_mask):
    seq_len = tf.shape(x)[1]
    attention_weights = {}

    x = self.embedding(x) # (batch_size, target_seq_len, d_model)
    x *= tf.math.sqrt(tf.cast(self.d_model, tf.float32))
    x += self.pos_encoding[:, :seq_len, :]
```

```
21
22    x = self.dropout(x, training=training)
23
24    for i in range(self.num_layers):
25      x, block1, block2 = self.dec_layers[i](x, enc_output, training,
            look_ahead_mask, padding_mask)
26
27      attention_weights[f'decoder_layer{i+1}_block1'] = block1
28      attention_weights[f'decoder_layer{i+1}_block2'] = block2
29
30    # x.shape == (batch_size, target_seq_len, d_model)
31    return x, attention_weights
```

```
1  class Transformer(tf.keras.Model):
2    def __init__(self, num_layers, d_model, num_heads, dff,
         input_vocab_size, target_vocab_size, pe_input, pe_target, rate
         =0.1):
3      super().__init__()
4      self.encoder = Encoder(num_layers, d_model, num_heads, dff,
           input_vocab_size,
5      pe_input, rate)
6      self.decoder = Decoder(num_layers, d_model, num_heads, dff,
           target_vocab_size,
7      pe_target, rate)
8      self.final_layer = tf.keras.layers.Dense(target_vocab_size)
9
10   def call(self, inputs, training):
11     # Keras models prefer if you pass all your inputs in the first
           argument
12     inp, tar = inputs
13
14     enc_padding_mask, look_ahead_mask, dec_padding_mask = self.
           create_masks(inp, tar)
15
16     # (batch_size, inp_seq_len, d_model)
17     enc_output = self.encoder(inp, training, enc_padding_mask)
18
19     # dec_output.shape == (batch_size, tar_seq_len, d_model)
20     dec_output, attention_weights = self.decoder(tar, enc_output,
           training, look_ahead_mask, dec_padding_mask)
21
22     # (batch_size, tar_seq_len, target_vocab_size)
23     final_output = self.final_layer(dec_output)
24
25     return final_output, attention_weights
26
27   def create_masks(self, inp, tar):
28     # Encoder padding mask
29     enc_padding_mask = create_padding_mask(inp)
30     # Used in the 2nd attention block in the decoder.
31     # This padding mask is used to mask the encoder outputs.
```

```
    dec_padding_mask = create_padding_mask(inp)
    # Used in the 1st attention block in the decoder.
    # It is used to pad and mask future tokens in the input received by
    # the decoder.
    look_ahead_mask = create_look_ahead_mask(tf.shape(tar)[1])
    dec_target_padding_mask = create_padding_mask(tar)
    look_ahead_mask = tf.maximum(dec_target_padding_mask, look_ahead_mask)

    return enc_padding_mask, look_ahead_mask, dec_padding_mask

# Hiperparâmetros
num_layers = 4
d_model = 128
dff = 512
num_heads = 8
dropout_rate = 0.1
```

```

class CustomSchedule(tf.keras.optimizers.schedules.LearningRateSchedule):
  def __init__(self, d_model, warmup_steps=4000):
    super(CustomSchedule, self).__init__()
    self.d_model = d_model
    self.d_model = tf.cast(self.d_model, tf.float32)
    self.warmup_steps = warmup_steps

  def __call__(self, step):
    step = tf.cast(step, tf.float32) # Adicionado para evitar ERRO
    arg1 = tf.math.rsqrt(step)
    arg2 = step * (self.warmup_steps ** -1.5)
    return tf.math.rsqrt(self.d_model) * tf.math.minimum(arg1, arg2)
```

```

learning_rate = CustomSchedule(d_model)
optimizer = tf.keras.optimizers.Adam(learning_rate, beta_1=0.9, beta_2=0.98, epsilon=1e-9)

loss_object = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True, reduction='none')

def loss_function(real, pred):
  mask = tf.math.logical_not(tf.math.equal(real, 0))
  loss_ = loss_object(real, pred)
  mask = tf.cast(mask, dtype=loss_.dtype)
  loss_ *= mask
  return tf.reduce_sum(loss_)/tf.reduce_sum(mask)

def accuracy_function(real, pred):
  accuracies = tf.equal(real, tf.argmax(pred, axis=2))
```

```
  mask = tf.math.logical_not(tf.math.equal(real, 0))
  accuracies = tf.math.logical_and(mask, accuracies)
  accuracies = tf.cast(accuracies, dtype=tf.float32)
  mask = tf.cast(mask, dtype=tf.float32)
  return tf.reduce_sum(accuracies)/tf.reduce_sum(mask)

train_loss = tf.keras.metrics.Mean(name='train_loss')
train_accuracy = tf.keras.metrics.Mean(name='train_accuracy')

transformer = Transformer(
  num_layers=num_layers,
  d_model=d_model,
  num_heads=num_heads,
  dff=dff,
  input_vocab_size=tokenizers.pt.get_vocab_size().numpy(),
  target_vocab_size=tokenizers.en.get_vocab_size().numpy(),
  pe_input=1000,
  pe_target=1000,
  rate=dropout_rate)

# Checkpoint
checkpoint_path = "./checkpoints/train"
ckpt = tf.train.Checkpoint(transformer=transformer, optimizer=
    optimizer)
ckpt_manager = tf.train.CheckpointManager(ckpt, checkpoint_path,
    max_to_keep=5)

# if a checkpoint exists, restore the latest checkpoint.
if ckpt_manager.latest_checkpoint:
  ckpt.restore(ckpt_manager.latest_checkpoint)
  print('Latest checkpoint restored!!')

EPOCHS = 20
train_step_signature = [
  tf.TensorSpec(shape=(None, None), dtype=tf.int64),
  tf.TensorSpec(shape=(None, None), dtype=tf.int64),
]
@tf.function(input_signature=train_step_signature)
def train_step(inp, tar):
  tar_inp = tar[:, :-1]
  tar_real = tar[:, 1:]
  with tf.GradientTape() as tape:
    predictions, _ = transformer([inp, tar_inp], training = True)
    loss = loss_function(tar_real, predictions)

  gradients = tape.gradient(loss, transformer.trainable_variables)
  optimizer.apply_gradients(zip(gradients, transformer.
      trainable_variables))

  train_loss(loss)
  train_accuracy(accuracy_function(tar_real, predictions))
```

```

for epoch in range(EPOCHS):
  start = time.time()
  train_loss.reset_state()
  train_accuracy.reset_state()
  # inp -> portuguese, tar -> english
  for (batch, (inp, tar)) in enumerate(train_batches):
    train_step(inp, tar)
    if batch % 50 == 0:
      print(f'Epoch {epoch + 1} Batch {batch} Loss {train_loss.result
        ():.4f} Accuracy {train_accuracy.result():.4f}')
  if (epoch + 1) % 5 == 0:
    ckpt_save_path = ckpt_manager.save()
    print(f'Saving checkpoint for epoch {epoch+1} at {ckpt_save_path}'
      )

  print(f'Epoch {epoch + 1} Loss {train_loss.result():.4f} Accuracy {
    train_accuracy.result():.4f}')
  print(f'Time taken for 1 epoch: {time.time() - start:.2f} secs\n')
```

```
class Translator(tf.Module):
  def __init__(self, tokenizers, transformer):
    self.tokenizers = tokenizers
    self.transformer = transformer

  def __call__(self, sentence, max_length=20):
    # input sentence is portuguese, hence adding the start and end
      token
    assert isinstance(sentence, tf.Tensor)
    if len(sentence.shape) == 0:
      sentence = sentence[tf.newaxis]

    sentence = self.tokenizers.pt.tokenize(sentence).to_tensor()
    encoder_input = sentence

    # as the target is english, the first token to the transformer
      should be the
    # english start token.
    start_end = self.tokenizers.en.tokenize([''])[0]
    start = start_end[0][tf.newaxis]
    end = start_end[1][tf.newaxis]

    output_array = tf.TensorArray(dtype=tf.int64, size=0, dynamic_size
      =True)
    output_array = output_array.write(0, start)

    for i in tf.range(max_length):
      output = tf.transpose(output_array.stack())
      predictions, _ = self.transformer([encoder_input, output],
        training=False)
```

```
    predictions = predictions[:, -1:, :] # (batch_size, 1,
       vocab_size)
    predicted_id = tf.argmax(predictions, axis=-1)
    output_array = output_array.write(i+1, predicted_id[0])
    if predicted_id == end:
      break

  output = tf.transpose(output_array.stack())
  # output.shape (1, tokens)
  text = tokenizers.en.detokenize(output)[0]
  tokens = tokenizers.en.lookup(output)[0]
  _, attention_weights = self.transformer([encoder_input, output
     [:,:-1]], training=False)

  return text, tokens, attention_weights

translator = Translator(tokenizers, transformer)
sentence = "Eu li sobre triceratops na enciclopédia."
translated_text, translated_tokens, attention_weights = translator(tf
   .constant(sentence))

print(f'{"Prediction":15s}: {translated_text}')
```

Prediction : b'i read about trieps in the encycloody .'

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

Estudo de Caso: A Aplicação de Big Data no Airbnb

Introdução

O Airbnb é uma plataforma online que conecta pessoas que desejam alugar suas casas com viajantes que buscam acomodação. Desde a sua fundação em 2008, a empresa cresceu exponencialmente, operando em mais de 190 países. Esse crescimento é sustentado pela capacidade do Airbnb de coletar, processar e analisar grandes volumes de dados, permitindo uma experiência personalizada para usuários e anfitriões. Neste estudo de caso, analisaremos como os dados do Airbnb podem ser classificados como Big Data, explorando os "Vs"do Big Data presentes em suas operações, e detalharemos as ferramentas utilizadas pela empresa no gerenciamento desses dados.

Classificação dos Dados do Airbnb como Big Data

Os dados gerados e utilizados pelo Airbnb apresentam características que os classificam como Big Data. Os cinco "Vs"do Big Data, Volume, Velocidade, Variedade, Veracidade e Valor, estão presentes de forma evidente nos dados da empresa.

1. Volume: O Airbnb lida com uma quantidade massiva de dados. São milhões de usuários, reservas, avaliações, mensagens e interações diariamente. Esse grande volume requer sistemas capazes de armazenar e processar dados em escala petabyte.

2. Velocidade: Os dados são gerados e precisam ser processados em tempo real ou quase real. Reservas, cancelamentos e atualizações de disponibilidade ocorrem constantemente, exigindo processamento rápido para manter a plataforma atualizada.

3. Variedade: Os dados do Airbnb são diversos, incluindo dados estruturados (informações de reservas, perfis de usuários), semiestruturados (logs de atividades) e não estruturados (comentários, imagens). Essa variedade demanda sistemas flexíveis para gerenciamento.

4. Veracidade: A qualidade e a confiabilidade dos dados são cruciais. O Airbnb precisa garantir que as informações sejam precisas para manter a confiança entre hóspedes e anfitriões.

5. Valor: A análise desses dados gera insights valiosos que permitem melhorar a experiência do usuário, otimizar preços e expandir os negócios. O valor extraído dos dados é fundamental para a estratégia da empresa.

Ferramentas Utilizadas

O Airbnb utiliza uma arquitetura sofisticada de Big Data, incorporando diversas ferramentas e tecnologias para atender às suas necessidades. Abaixo, detalhamos cada ferramenta e seu papel no processo de dados.

1. Amazon Web Services (AWS): S3 (Simple Storage Service): Utilizado para armazenamento de objetos, o S3 permite que o Airbnb armazene grandes volumes de dados de forma escalável e segura. E EC2 (Elastic Compute Cloud): Fornece capacidade computacional elástica, permitindo que a empresa escale seus recursos de processamento conforme a demanda.

2. Hadoop: O Hadoop é um framework de software que facilita o processamento distribuído de grandes conjuntos de dados. No Airbnb, é usado para armazenar e processar dados em clusters de servidores, permitindo a análise de grandes volumes de informação.

3. HBase: Um banco de dados NoSQL que funciona em cima do Hadoop, o HBase é usado para armazenamento em tempo real de dados grandes e dispersos. Ele permite acessos rápidos a dados específicos dentro de conjuntos massivos.

4. Hive: O Hive permite consultas tipo SQL em dados armazenados no Hadoop. Isso facilita para os analistas de dados executarem consultas complexas sem a necessidade de programação em baixo nível.

5. Presto: Presto é um motor de consultas distribuídas que permite ao Airbnb executar consultas SQL interativas e de baixa latência em grandes volumes de dados, espalhados por várias fontes de dados.

6. Apache Spark: O Spark é utilizado para processamento em lote e em tempo real. Com sua capacidade de processamento na memória, o Spark acelera o processamento de tarefas como análise de dados, aprendizado de máquina e processamento de gráficos.

7. Apache Airflow: Desenvolvido pelo próprio Airbnb, o Airflow é uma plataforma de orquestração de fluxos de trabalho que permite a automação de pipelines de dados. Ele gerencia a execução de tarefas, dependências e agendamento de processos.

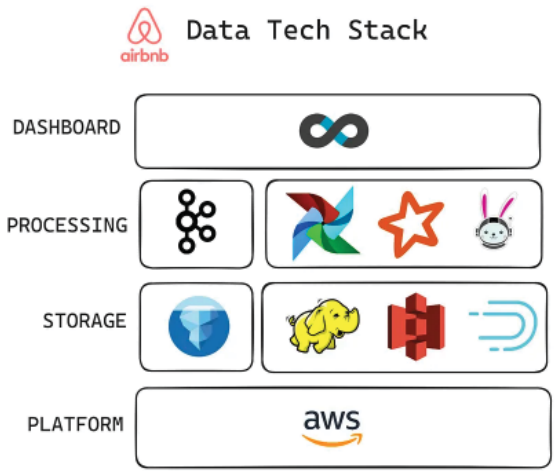
8. Apache Superset: Outra criação do Airbnb, o Superset é uma plataforma de visualização de dados. Permite a criação de dashboards interativos e gráficos, auxiliando na interpretação e análise dos dados.

9. Druid: Druid é um sistema de banco de dados analítico projetado para consultas rápidas em grandes conjuntos de dados. No Airbnb, é usado para análise em tempo real e interativa, especialmente em dados de séries temporais.

Aplicação das Ferramentas

O processo de dados no Airbnb pode ser dividido em várias etapas: coleta, armazenamento, processamento, análise e visualização. A Figura abaixo mostra uma visão geral das ferramentas utilizadas em cada etapa, sendo estas especificadas nas seções seguintes.

Figura 9.1 – VISÃO GERAL DAS FERRAMENTAS UTILIZADAS PELO AIRBNB



1. Coleta: Dados são coletados de diversas fontes, incluindo interações de usuários, informações de reservas, avaliações, logs de atividades e dados de terceiros (como eventos locais e clima).

2. Armazenamento: S3 e Hadoop são utilizados para armazenar os dados coletados. O S3 armazena objetos de dados, enquanto o Hadoop permite armazenamento distribuído e escalável.

3. Processamento: Apache Spark processa dados em lote e em tempo real, permitindo análises complexas e treinamento de modelos de machine learning. Hive e Presto são usados para consultas e manipulação de dados, facilitando o acesso a informações específicas. HBase oferece acesso rápido a dados em tempo real, essencial para aplicações que requerem baixa latência.

4. Análise: Druid permite análises rápidas e interativas, especialmente útil para monitoramento em tempo real e análise de séries temporais. Modelos de machine learning são implementados para diversas finalidades, como previsão de demanda, definição de preços dinâmicos e recomendações personalizadas.

5. Orquestração e Automação: Apache Airflow gerencia e automatiza os fluxos de trabalho de dados, garantindo que os processos sejam executados corretamente e no tempo certo.

6. Visualização: Apache Superset é utilizado para criar dashboards e relatórios que auxiliam na tomada de decisões baseadas em dados.

Conclusão

O Airbnb é um exemplo notável de como Big Data pode ser utilizado para impulsionar negócios digitais. Através da coleta e análise de grandes volumes de dados, a empresa consegue oferecer experiências personalizadas, otimizar operações e inovar continuamente. Os cinco "Vs"do Big Data estão profundamente integrados nas operações do Airbnb, e o uso eficiente de ferramentas como AWS, Hadoop, Spark e outras permite que a empresa transforme dados brutos em insights valiosos. Como estudante de Big Data, observar o caso do Airbnb destaca a importância de uma arquitetura de dados robusta e da integração de diversas tecnologias para lidar com os desafios e oportunidades que o Big Data apresenta.

Referências

https://www.linkedin.com/pulse/what-technology-stack-airbnb-john-doe?utm_source=share utm_medium=member_android&utm_campaign=share_via

https://www.junaideffendi.com/p/airbnb-data-tech-stack

https://vutr.substack.com/p/groupby-40-data-infrastructure-at

https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c

APÊNDICE 10 – VISÃO COMPUTACIONAL

A - ENUNCIADO

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

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

```
1
2  import os
3  import cv2
4  import shutil
5  import numpy as np
6  import pandas as pd
7  import matplotlib.pyplot as plt
8
9  from tqdm import tqdm
10 from pathlib import Path
11
12 from sklearn.svm import SVC
13 from sklearn.metrics import recall_score
14 from sklearn.preprocessing import LabelEncoder
15 from sklearn.ensemble import RandomForestClassifier
16 from sklearn.model_selection import train_test_split
17 from sklearn.metrics import classification_report, confusion_matrix
18
19 from tensorflow.keras.layers import Dense
20 from tensorflow.keras.models import Model
21 from tensorflow.keras.models import Sequential
22 from tensorflow.keras.applications import VGG16
23 from tensorflow.keras.preprocessing import image
24 from tensorflow.keras.applications.vgg16 import preprocess_input
25
26 !unzip Train_Warwick.zip
27
28 # Diretório das amostras de treino
29 images_dir = Path('/content/Train_4cls_amostra')
30
31 # Diretórios de destino para treino, validação e teste
32 train_dir = images_dir.parent / 'train'
33 val_dir = images_dir.parent / 'val'
34
35 # Classes disponíveis
36 classes = ['0', '1', '2', '3']
37
38 patient_dict = {}
39
40 # Percorrer cada classe e organizar pacientes
41 for cls in classes:
42     class_dir = images_dir / cls
43     for img_name in os.listdir(class_dir):
44         if img_name.endswith('.png'):
45             # Extrair o ID do paciente do nome do arquivo (XX_HER_YYYY.
                png)
46             patient_id = img_name.split('_')[0]
```

```
47          img_path = class_dir / img_name
48          if cls not in patient_dict:
49              patient_dict[cls] = {}
50          if patient_id not in patient_dict[cls]:
51              patient_dict[cls][patient_id] = []
52          patient_dict[cls][patient_id].append((img_path, cls))
53
54  # Função para dividir pacientes em 80% treino e 20% validação
55  def split_patients(patients):
56      patient_ids = list(patients.keys())
57      train_size = int(0.8 * len(patient_ids))
58      train_patient_ids = patient_ids[:train_size]
59      val_patient_ids = patient_ids[train_size:]
60      return train_patient_ids, val_patient_ids
61
62  # Função para mover as imagens para os diretórios de treino e validaç
        ão
63  def move_images(patient_ids, patient_dict, target_dir):
64      for pid in patient_ids:
65          for img_path, cls in patient_dict[pid]:
66              # Diretório da classe no conjunto alvo
67              class_dir = target_dir / cls
68              class_dir.mkdir(parents=True, exist_ok=True)
69              # Caminho de destino da imagem
70              dest_path = class_dir / img_path.name
71              # Mover a imagem
72              shutil.move(str(img_path), str(dest_path))
73
74  # Para cada classe, dividir pacientes e mover as imagens
75  for cls in classes:
76      print(f"Processando classe {cls}...")
77
78      # Obter os pacientes dessa classe
79      patients_in_class = patient_dict[cls]
80
81      # Dividir em treino e validação
82      train_patient_ids, val_patient_ids = split_patients(
            patients_in_class)
83
84      print('Número total de pacientes:', len(train_patient_ids) + len(
            val_patient_ids))
85      print('Número de pacientes para treino:', len(train_patient_ids))
86      print('Número de pacientes para validação:', len(val_patient_ids))
87
88      # Mover as imagens de treino
89      move_images(train_patient_ids, patients_in_class, train_dir)
90
91      # Mover as imagens de validação
92      move_images(val_patient_ids, patients_in_class, val_dir)
93
94  print("Divisão de treino e validação concluída.")
```

```
95
96  # Processando classe 0...
97  # Número total de pacientes: 5
98  # Número de pacientes para treino: 4
99  # Número de pacientes para validação: 1
100 # Processando classe 1...
101 # Número total de pacientes: 5
102 # Número de pacientes para treino: 4
103 # Número de pacientes para validação: 1
104 # Processando classe 2...
105 # Número total de pacientes: 5
106 # Número de pacientes para treino: 4
107 # Número de pacientes para validação: 1
108 # Processando classe 3...
109 # Número total de pacientes: 5
110 # Número de pacientes para treino: 4
111 # Número de pacientes para validação: 1
112 # Divisão de treino e validação concluída.
113
114 def get_pixel(img, center, x, y):
115
116       new_value = 0
117
118       try:
119             # If local neighbourhood pixel
120             # value is greater than or equal
121             # to center pixel values then
122             # set it to 1
123             if img[x][y] >= center:
124                   new_value = 1
125
126       except:
127             # Exception is required when
128             # neighbourhood value of a center
129             # pixel value is null i.e. values
130             # present at boundaries.
131             pass
132
133       return new_value
134
135
136 # Function for calculating LBP
137 def lbp_calculated_pixel(img, x, y, radius):
138
139       center = img[x][y]
140
141       val_ar = []
142
143       # top_left
144       val_ar.append(get_pixel(img, center, x-radius, y-radius))
145
```

```
146         # top
147         val_ar.append(get_pixel(img, center, x-radius, y))
148 
149         # top_right
150         val_ar.append(get_pixel(img, center, x-radius, y + radius))
151 
152         # right
153         val_ar.append(get_pixel(img, center, x, y + radius))
154 
155         # bottom_right
156         val_ar.append(get_pixel(img, center, x + radius, y + radius))
157 
158         # bottom
159         val_ar.append(get_pixel(img, center, x + radius, y))
160 
161         # bottom_left
162         val_ar.append(get_pixel(img, center, x + radius, y-radius))
163 
164         # left
165         val_ar.append(get_pixel(img, center, x, y-radius))
166 
167         # Now, we need to convert binary
168         # values to decimal
169         power_val = [1, 2, 4, 8, 16, 32, 64, 128]
170 
171         val = 0
172 
173         for i in range(len(val_ar)):
174               val += val_ar[i] * power_val[i]
175 
176         return val
177 
178 # Função para calcular o LBP
179 def extract_lbp_image(image):
180    img_gray = cv2.cvtColor(image,
181               cv2.COLOR_BGR2GRAY)
182    img_gray = cv2.bilateralFilter(img_gray, 7, 75, 75)
183 
184    height, width, _ = image.shape
185 
186    img_lbp = np.zeros((height, width),
187          np.uint8)
188 
189    for i in range(0, height):
190      for j in range(0, width):
191        img_lbp[i, j] = lbp_calculated_pixel(img_gray, i, j, 1)
192 
193    # Calcular o histograma
194    h = cv2.calcHist([img_lbp], [0], None, [256], [0, 256])
195    hist,bins = np.histogram(h.flatten(),256,[0,256])
```

```
196     h_normalized = 100*((hist - hist.min()) /(hist.max() - hist.min())
            )
197 
198     return img_lbp, h_normalized
199 
200 def extract_lbp_from_directory(directory, classes):
201     features = []
202     labels = []
203 
204     print(f"Extraindo características LBP das imagens no diretório: {
            directory}")
205 
206     # Iterar sobre cada classe
207     for cls in classes:
208         class_dir = os.path.join(directory, cls)
209         image_names = os.listdir(class_dir)
210         for img_name in tqdm(image_names, desc=f"Classe {cls}"):
211             img_path = os.path.join(class_dir, img_name)
212             # Ler a imagem
213             image = cv2.imread(img_path, 1)
214             # Verificar se a imagem tem 4 canais
215             if len(image.shape) == 3 and image.shape[2] == 4:
216                 # Remover o canal alfa
217                 image = image[:, :, :3]
218             # Extrair características LBP
219             img_lbp, h_normalized = extract_lbp_image(image)
220             features.append(h_normalized)
221             # Armazenar o label
222             labels.append(cls)
223 
224     return features, labels
225 
226 # Extrair características do conjunto de treino
227 train_features, train_labels = extract_lbp_from_directory(train_dir,
        classes)
228 
229 # Extrair características do conjunto de validação
230 val_features, val_labels = extract_lbp_from_directory(val_dir,
        classes)
231 
232 # Extraindo características LBP das imagens no diretório: /content/
        val
233 
234 # Converter as listas em arrays numpy
235 train_features = np.array(train_features)
236 train_labels = np.array(train_labels)
237 val_features = np.array(val_features)
238 val_labels = np.array(val_labels)
239 
240 # Criar DataFrames
241 train_df = pd.DataFrame(train_features)
```

```
train_df['label'] = train_labels

val_df = pd.DataFrame(val_features)
val_df['label'] = val_labels

# Salvar em CSV
train_df.to_csv('train_lbp_features.csv', index=False)
val_df.to_csv('val_lbp_features.csv', index=False)

print("Características LBP salvas nos arquivos 'train_lbp_features.csv' e 'val_lbp_features.csv'")

# Características LBP salvas nos arquivos 'train_lbp_features.csv' e 'val_lbp_features.csv'

# Carregar o modelo VGG16 pré-treinado, sem a camada de classificação final
# pooling='avg' para obter um vetor de características de saída
base_model = VGG16(weights='imagenet', include_top=False, pooling='avg')

# Função para carregar uma imagem, redimensioná-la e pré-processá-la
def process_image(img_path):
    img = image.load_img(img_path, target_size=(224, 224))
    img_array = image.img_to_array(img)
    img_array = np.expand_dims(img_array, axis=0)
    img_array = preprocess_input(img_array) # Pré-processamento VGG16
    return img_array

# Função para extrair as características usando o modelo VGG16
def extract_vgg16_features(img_path, model):
    img_array = process_image(img_path)
    features = model.predict(img_array)
    return features.flatten() # Converter para vetor de caracterí
        sticas

  # Extrair características de um diretório de imagens
def extract_features_from_directory(directory, model):
    features = []
    labels = []

    for class_label in os.listdir(directory):
        class_dir = os.path.join(directory, class_label)

        if os.path.isdir(class_dir):
            for img_name in tqdm(os.listdir(class_dir), desc=f"Classe {
                class_label}"):
                img_path = os.path.join(class_dir, img_name)

                # Extrair características
                img_features = extract_vgg16_features(img_path, model)
```

```
287
288             # Armazenar as características e o rótulo (classe)
289             features.append(img_features)
290             labels.append(class_label)
291
292     return np.array(features), np.array(labels)
293
294 # Extração das características das imagens de treino usando o modelo
    VGG16
295 print("Extraindo características do conjunto de treino usando o
    modelo VGG16")
296 train_features, train_labels = extract_features_from_directory(
    train_dir, base_model)
297
298 # Extração das características das imagens de validação usando o
    modelo VGG16
299 print("Extraindo características do conjunto de validação usando o
    modelo VGG16")
300 val_features, val_labels = extract_features_from_directory(val_dir,
    base_model)
301
302
303 # Converter para DataFrame e salvar como CSV
304 train_df = pd.DataFrame(train_features)
305 train_df['label'] = train_labels
306 train_df.to_csv('train_vgg16_features.csv', index=False)
307
308 val_df = pd.DataFrame(val_features)
309 val_df['label'] = val_labels
310 val_df.to_csv('val_vgg16_features.csv', index=False)
311
312 print("Características VGG16 salvas nos arquivos '
    train_vgg16_features.csv' e 'val_vgg16_features.csv'")
313
314 # Carregar características LBP e CNN VGG16
315 train_lbp = pd.read_csv('train_lbp_features.csv')
316 train_vgg16 = pd.read_csv('train_vgg16_features.csv')
317
318 # Separar características e rótulos (labels)
319 X_train_lbp = train_lbp.drop(columns=['label'])
320 y_train_lbp = train_lbp['label']
321
322 X_train_vgg16 = train_vgg16.drop(columns=['label'])
323 y_train_vgg16 = train_vgg16['label']
324
325 # Codificar as labels para valores numéricos
326 label_encoder = LabelEncoder()
327 y_train_lbp = label_encoder.fit_transform(y_train_lbp)
328 y_train_vgg16 = label_encoder.fit_transform(y_train_vgg16)
329
330 # Treinando o modelo SVM com características LBP
```

```
svm_lbp = SVC(kernel='linear', random_state=42)
svm_lbp.fit(X_train_lbp, y_train_lbp)

# Treinando o modelo SVM com características VGG16
svm_vgg16 = SVC(kernel='linear', random_state=42)
svm_vgg16.fit(X_train_vgg16, y_train_vgg16)

# Função para criar o modelo RNA
def create_rna_model(input_dim):
    model = Sequential()
    model.add(Dense(128, activation='relu', input_dim=input_dim))
    model.add(Dense(64, activation='relu'))
    model.add(Dense(32, activation='relu'))
    model.add(Dense(len(label_encoder.classes_), activation='softmax')
        ) # Saída com número de classes
    model.compile(optimizer='adam', loss='
        sparse_categorical_crossentropy', metrics=['accuracy'])
    return model

# Treinando a RNA com características LBP
rna_lbp = create_rna_model(X_train_lbp.shape[1])
rna_lbp.fit(X_train_lbp, y_train_lbp, epochs=10, batch_size=32,
    validation_split=0.2)

# Treinando a RNA com características VGG16
rna_vgg16 = create_rna_model(X_train_vgg16.shape[1])
rna_vgg16.fit(X_train_vgg16, y_train_vgg16, epochs=10, batch_size=32,
     validation_split=0.2)

# Avaliando Random Forest (LBP)
y_pred_rf_lbp = rf_lbp.predict(X_train_lbp)
print("Relatório de Classificação (Random Forest - LBP):")
print(classification_report(y_train_lbp, y_pred_rf_lbp))

# Avaliando SVM (LBP)
y_pred_svm_lbp = svm_lbp.predict(X_train_lbp)
print("Relatório de Classificação (SVM - LBP):")
print(classification_report(y_train_lbp, y_pred_svm_lbp))

# Avaliando RNA (LBP)
y_pred_rna_lbp = rna_lbp.predict(X_train_lbp)
y_pred_rna_lbp = np.argmax(y_pred_rna_lbp, axis=1)
print("Relatório de Classificação (RNA - LBP):")
print(classification_report(y_train_lbp, y_pred_rna_lbp))

# Avaliando Random Forest (VGG16)
y_pred_rf_vgg16 = rf_vgg16.predict(X_train_vgg16)
print("Relatório de Classificação (Random Forest - VGG16):")
print(classification_report(y_train_vgg16, y_pred_rf_vgg16))

# Avaliando SVM (VGG16)
```

```
y_pred_svm_vgg16 = svm_vgg16.predict(X_train_vgg16)
print("Relatório de Classificação (SVM - VGG16):")
print(classification_report(y_train_vgg16, y_pred_svm_vgg16))

# Avaliando RNA (VGG16)
y_pred_rna_vgg16 = rna_vgg16.predict(X_train_vgg16)
y_pred_rna_vgg16 = np.argmax(y_pred_rna_vgg16, axis=1)
print("Relatório de Classificação (RNA - VGG16):")
print(classification_report(y_train_vgg16, y_pred_rna_vgg16))

#Relatório de Classificação (Random Forest - LBP):
#              precision recall f1-score support
#
#           0       1.00    1.00    1.00     116
#           1       1.00    1.00    1.00     117
#           2       1.00    1.00    1.00     120
#           3       1.00    1.00    1.00     120

#    accuracy                         1.00     473
#   macro avg    1.00    1.00    1.00     473
#weighted avg    1.00    1.00    1.00     473

#Relatório de Classificação (SVM - LBP):
#              precision recall f1-score support
#
#           0       1.00    1.00    1.00     116
#           1       1.00    1.00    1.00     117
#           2       1.00    1.00    1.00     120
#           3       1.00    1.00    1.00     120

#    accuracy                         1.00     473
#   macro avg    1.00    1.00    1.00     473
#weighted avg    1.00    1.00    1.00     473

#Relatório de Classificação (RNA - LBP):
#              precision recall f1-score support

#           0       0.76    0.73    0.75     116
#           1       0.71    0.89    0.79     117
#           2       0.43    0.74    0.54     120
#           3       1.00    0.07    0.12     120
#
#    accuracy                         0.60     473
#   macro avg    0.73    0.61    0.55     473
# weighted avg   0.73     0.60     0.55      473

#Relatório de Classificação (Random Forest - VGG16):
#              precision recall f1-score support
#
#           0       1.00    1.00    1.00     116
#           1       1.00    1.00    1.00     117
```

```
#         2     1.00    1.00    1.00     120
#         3     1.00    1.00    1.00     120

#   accuracy                    1.00     473
#  macro avg    1.00    1.00    1.00     473
#weighted avg  1.00    1.00    1.00     473

# Relatório de Classificação (SVM - VGG16):
#           precision recall f1-score support

#         0     1.00    1.00    1.00     116
#         1     1.00    1.00    1.00     117
#         2     1.00    1.00    1.00     120
#         3     1.00    1.00    1.00     120

#   accuracy                    1.00     473
#  macro avg    1.00    1.00    1.00     473
#weighted avg  1.00    1.00    1.00     473

#Relatório de Classificação (RNA - VGG16):
#           precision recall f1-score support

#         0     0.98    1.00    0.99     116
#         1     0.92    1.00    0.96     117
#         2     1.00    0.85    0.92     120
#         3     0.95    1.00    0.98     120

#   accuracy                    0.96     473
#  macro avg    0.96    0.96    0.96     473
#weighted avg  0.96    0.96    0.96     473

# !unzip Test_Warwick.zip

# Diretório das amostras de teste
test_dir = Path('/content/Test_4cl_amostra')

# Extrair características do conjunto de teste
test_features, test_labels = extract_lbp_from_directory(test_dir,
    classes)

# Converter as listas em arrays numpy
test_features = np.array(test_features)
test_labels = np.array(test_labels)

# Criar DataFrame
test_df = pd.DataFrame(test_features)
test_df['label'] = test_labels

# Salvar em CSV
test_df.to_csv('test_lbp_features.csv', index=False)
```

```
479 print("Características LBP salvas no arquivo 'test_lbp_features.csv'")
480
481 # Extração das características das imagens de teste
482 print("Extraindo características do conjunto de teste...")
483 test_features, test_labels = extract_features_from_directory(test_dir, base_model)
484
485 # Converter para DataFrame e salvar como CSV
486 test_df = pd.DataFrame(test_features)
487 test_df['label'] = test_labels
488 test_df.to_csv('test_vgg16_features.csv', index=False)
489
490 print("Características VGG16 salvas nos arquivos 'test_vgg16_features.csv'")
491
492 # Carregar os dados de teste
493 test_lbp = pd.read_csv('test_lbp_features.csv')
494 test_vgg16 = pd.read_csv('test_vgg16_features.csv')
495
496 # Separar as características e os rótulos (labels)
497 X_test_lbp = test_lbp.drop(columns=['label'])
498 y_test_lbp = test_lbp['label']
499
500 X_test_vgg16 = test_vgg16.drop(columns=['label'])
501 y_test_vgg16 = test_vgg16['label']
502
503 # Codificar os rótulos de teste (mesma codificação usada para os dados de treino)
504 label_encoder = LabelEncoder()
505 y_test_lbp = label_encoder.fit_transform(y_test_lbp)
506
507 label_encoder = LabelEncoder()
508 y_test_vgg16 = label_encoder.fit_transform(y_test_vgg16)
509
510 # Fazer previsões com o modelo Random Forest
511 y_pred_rf_lbp = rf_lbp.predict(X_test_lbp)
512
513 y_pred_rf_vgg16 = rf_vgg16.predict(X_test_vgg16)
514
515 # Fazer previsões com o modelo SVM
516 y_pred_svm_lbp = svm_lbp.predict(X_test_lbp)
517
518 y_pred_svm_vgg16 = svm_vgg16.predict(X_test_vgg16)
519
520 # Fazer previsões com a RNA
521 y_pred_rna_lbp = rna_lbp.predict(X_test_lbp)
522 y_pred_rna_lbp = np.argmax(y_pred_rna_lbp, axis=1)
523
524 y_pred_rna_vgg16 = rna_vgg16.predict(X_test_vgg16)
525 y_pred_rna_vgg16 = np.argmax(y_pred_rna_vgg16, axis=1)
```

```
526
527 # Calcular Especificidade a partir da Matriz de Confusão
528 def calculate_specificity(conf_matrix):
529     TN = conf_matrix[0, 0]
530     FP = conf_matrix[0, 1]
531     specificity = TN / (TN + FP)
532     return specificity
533
534 print("Random Forest\n")
535
536 # Avaliar o desempenho do Random Forest (LBP)
537 print("LBP")
538 print(classification_report(y_test_lbp, y_pred_rf_lbp))
539 conf_matrix_rf_lbp = confusion_matrix(y_test_lbp, y_pred_rf_lbp)
540 specificity_rf_lbp = calculate_specificity(conf_matrix_rf_lbp)
541 print(f"Especificidade: {specificity_rf_lbp}")
542
543 # Avaliar o desempenho do Random Forest (VGG16)
544 print("\n\nVGG16")
545 print(classification_report(y_test_vgg16, y_pred_rf_vgg16))
546 conf_matrix_rf_vgg16 = confusion_matrix(y_test_vgg16, y_pred_rf_vgg16
    )
547 specificity_rf_vgg16 = calculate_specificity(conf_matrix_rf_vgg16)
548 print(f"Especificidade: {specificity_rf_vgg16}")
549
550
551 #Random Forest
552
553 #LBP
554 #              precision recall f1-score support
555 #
556 #           0      0.38    0.50    0.43     101
557 #           1      0.29    0.27    0.28      90
558 #           2      0.43    0.29    0.35      90
559 #           3      0.77    0.82    0.80      90
560
561 #    accuracy                    0.47     371
562 #   macro avg    0.47    0.47    0.46     371
563 #weighted avg   0.47    0.47    0.46     371
564
565 #Especificidade: 0.6493506493506493
566
567
568 #VGG16
569 #              precision recall f1-score support
570
571 #           0      0.80    0.99    0.88     101
572 #           1      0.58    0.17    0.26      90
573 #           2      0.59    0.83    0.69      90
574 #           3      0.95    0.98    0.96      90
575
```

```
#   accuracy                           0.75     371
#  macro avg    0.73      0.74      0.70     371
#weighted avg  0.73      0.75      0.70     371

#Especificidade: 0.9900990099009901

print("SVM\n")

# Avaliar o desempenho do SVM (LBP)
print("LBP")
print(classification_report(y_test_lbp, y_pred_svm_lbp))
conf_matrix_svm_lbp = confusion_matrix(y_test_lbp, y_pred_svm_lbp)
specificity_svm_lbp = calculate_specificity(conf_matrix_svm_lbp)
print(f"Especificidade: {specificity_svm_lbp}")

# Avaliar o desempenho do SVM (LBP)
print("\n\nVGG16")
print(classification_report(y_test_vgg16, y_pred_svm_vgg16))
conf_matrix_svm_vgg16 = confusion_matrix(y_test_vgg16,
    y_pred_svm_vgg16)
specificity_svm_vgg16 = calculate_specificity(conf_matrix_svm_vgg16)
print(f"Especificidade: {specificity_svm_vgg16}")

#SVM

#LBP
#              precision recall f1-score support

#           0       0.38      0.50      0.43       101
#           1       0.35      0.32      0.34        90
#           2       0.50      0.43      0.46        90
#           3       0.74      0.63      0.68        90

#   accuracy                           0.47     371
#  macro avg    0.49      0.47      0.48     371
#weighted avg  0.49      0.47      0.48     371

#Especificidade: 0.6538461538461539

#VGG16
#              precision recall f1-score support

#           0       0.85      0.92      0.88       101
#           1       0.77      0.30      0.43        90
#           2       0.64      0.96      0.77        90
#           3       0.98      1.00      0.99        90

#   accuracy                           0.80     371
#  macro avg    0.81      0.79      0.77     371
#weighted avg  0.81      0.80      0.77     371

```

```
#Especificidade: 0.9393939393939394

print("RNA\n")

# Avaliar o desempenho da RNA (LBP)
print("LBP")
print(classification_report(y_test_lbp, y_pred_rna_lbp, zero_division=1))
conf_matrix_rna_lbp = confusion_matrix(y_test_lbp, y_pred_rna_lbp)
specificity_rna_lbp = calculate_specificity(conf_matrix_rna_lbp)
print(f"Especificidade: {specificity_rna_lbp}")

# Avaliar o desempenho da RNA (LBP)
print("\n\nVGG16")
print(classification_report(y_test_vgg16, y_pred_rna_vgg16, zero_division=1))
conf_matrix_rna_vgg16 = confusion_matrix(y_test_vgg16, y_pred_rna_vgg16)
specificity_rna_vgg16 = calculate_specificity(conf_matrix_rna_vgg16)
print(f"Especificidade: {specificity_rna_vgg16}")

#RNA

#LBP
#              precision recall f1-score support

#           0       0.41    0.45    0.43     101
#           1       0.33    0.42    0.37      90
#           2       0.28    0.46    0.35      90
#           3       1.00    0.01    0.02      90

#    accuracy                         0.34     371
#   macro avg    0.51    0.33    0.29     371
#weighted avg    0.50    0.34    0.30     371

#Especificidade: 0.5357142857142857

#VGG16
#              precision recall f1-score support

#           0       0.92    0.95    0.94     101
#           1       0.81    0.58    0.68      90
#           2       0.69    0.73    0.71      90
#           3       0.84    1.00    0.91      90

#    accuracy                         0.82     371
#   macro avg    0.82    0.82    0.81     371
#weighted avg    0.82    0.82    0.81     371

#Especificidade: 0.9504950495049505

```

```
# Resultados: O modelo que deu melhores resultados foi o **RNA**, considerando a métrica de **sensibilidade** (recall).

### RNA - Sensibilidade
#*  CNN VGG16: **0.82**
#*  LBP: **0.34**
```

```
import tensorflow as tf
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.applications import VGG16, ResNet50
from tensorflow.keras.layers import Dense, Flatten, GlobalAveragePooling2D
from tensorflow.keras.models import Model
from tensorflow.keras.optimizers import Adam
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix, classification_report
import seaborn as sns
import os

img_height, img_width = 224, 224
batch_size = 32
num_classes = 4 # Classes: 0, 1, 2, 3

train_datagen = ImageDataGenerator(
    preprocessing_function=tf.keras.applications.vgg16.preprocess_input # Pré-processamento específico
)

val_datagen = ImageDataGenerator(
    preprocessing_function=tf.keras.applications.vgg16.preprocess_input
)

train_generator = train_datagen.flow_from_directory(
    train_dir,
    target_size=(img_height, img_width),
    batch_size=batch_size,
    class_mode='categorical'
)

validation_generator = val_datagen.flow_from_directory(
    val_dir,
    target_size=(img_height, img_width),
    batch_size=batch_size,
    class_mode='categorical'
)

def create_vgg16_model():
    # Carregar o modelo base VGG16 sem as camadas totalmente conectadas (top)
```

```
    base_model_vgg = VGG16(weights='imagenet', include_top=False,
        input_shape=(img_height, img_width, 3))

    # Congelar as camadas convolucionais
    for layer in base_model_vgg.layers:
        layer.trainable = False

    # Adicionar novas camadas
    x = base_model_vgg.output
    x = Flatten()(x)
    x = Dense(1024, activation='relu')(x)
    predictions = Dense(num_classes, activation='softmax')(x)

    # Criar o modelo final
    model_vgg = Model(inputs=base_model_vgg.input, outputs=predictions
        )

    return model_vgg

model_vgg = create_vgg16_model()

model_vgg.compile(optimizer=Adam(learning_rate=1e-4),
                  loss='categorical_crossentropy',
                  metrics=['accuracy'])

epochs = 5

history_vgg = model_vgg.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // batch_size,
    epochs=epochs,
    validation_data=validation_generator,
    validation_steps=validation_generator.samples // batch_size
)

def create_resnet50_model():
    base_model_resnet = ResNet50(weights='imagenet', include_top=False
        , input_shape=(img_height, img_width, 3))
    for layer in base_model_resnet.layers:
        layer.trainable = False
    x = base_model_resnet.output
    x = GlobalAveragePooling2D()(x)
    x = Dense(1024, activation='relu')(x)
    predictions = Dense(num_classes, activation='softmax')(x)
    model_resnet = Model(inputs=base_model_resnet.input, outputs=
        predictions)
    return model_resnet

model_resnet = create_resnet50_model()

model_resnet.compile(optimizer=Adam(learning_rate=1e-4),
```

```
                loss='categorical_crossentropy',
                metrics=['accuracy'])

history_resnet = model_resnet.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // batch_size,
    epochs=epochs,
    validation_data=validation_generator,
    validation_steps=validation_generator.samples // batch_size
)

train_datagen_aug = ImageDataGenerator(
    preprocessing_function=tf.keras.applications.vgg16.
        preprocess_input,
    rotation_range=20,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True
)

train_generator_aug = train_datagen_aug.flow_from_directory(
    train_dir,
    target_size=(img_height, img_width),
    batch_size=batch_size,
    class_mode='categorical'
)

# Treinamento com Data Augmentation
model_vgg_aug = create_vgg16_model()
model_vgg_aug.compile(optimizer=Adam(learning_rate=1e-4),
              loss='categorical_crossentropy',
              metrics=['accuracy'])

history_vgg_aug = model_vgg_aug.fit(
    train_generator_aug,
    steps_per_epoch=train_generator_aug.samples // batch_size,
    epochs=epochs,
    validation_data=validation_generator,
    validation_steps=validation_generator.samples // batch_size
)

# Treinamento com Data Augmentation
model_resnet_aug = create_vgg16_model()
model_resnet_aug.compile(optimizer=Adam(learning_rate=1e-4),
                loss='categorical_crossentropy',
                metrics=['accuracy'])

history_resnet_aug = model_resnet_aug.fit(
    train_generator_aug,
    steps_per_epoch=train_generator_aug.samples // batch_size,
    epochs=epochs,
```

```
    validation_data=validation_generator,
    validation_steps=validation_generator.samples // batch_size
)

test_datagen = ImageDataGenerator(
    preprocessing_function=tf.keras.applications.vgg16.
        preprocess_input
)

test_generator = test_datagen.flow_from_directory(
    test_dir,
    target_size=(img_height, img_width),
    batch_size=1, # Para avaliar imagem por imagem
    class_mode='categorical',
    shuffle=False # Importante para manter a ordem das imagens
)

# VGG sem Data Augmentation
test_generator.reset()
preds_vgg = model_vgg.predict(test_generator, steps=test_generator.
    samples)

# VGG com Data Augmentation
test_generator.reset()
preds_vgg_aug = model_vgg_aug.predict(test_generator, steps=
    test_generator.samples)

# Converter as predições para classes
y_pred_vgg = np.argmax(preds_vgg, axis=1)
y_pred_vgg_aug = np.argmax(preds_vgg_aug, axis=1)
y_pred_resnet = np.argmax(preds_resnet, axis=1)
y_pred_resnet_aug = np.argmax(preds_resnet_aug, axis=1)

# Obter as classes verdadeiras
y_true = test_generator.classes
class_labels = list(test_generator.class_indices.keys())

from sklearn.metrics import confusion_matrix

# VGG16 sem Data Augmentation
cm_vgg = confusion_matrix(y_true, y_pred_vgg)

# VGG16 com Data Augmentation
cm_vgg_aug = confusion_matrix(y_true, y_pred_vgg_aug)

# ResNet50 sem Data Augmentation
cm_resnet = confusion_matrix(y_true, y_pred_resnet)

# ResNet50 com Data Augmentation
cm_resnet_aug = confusion_matrix(y_true, y_pred_resnet_aug)

```

```
186 from sklearn.metrics import classification_report, confusion_matrix
187
188 def calculate_specificity(cm):
189     # cm éa matriz de confusão
190     FP = cm.sum(axis=0) - np.diag(cm)
191     TN = cm.sum() - (FP + cm.sum(axis=1) - np.diag(cm) + np.diag(cm))
192     specificity = TN / (TN + FP)
193     return specificity
194
195 # Avaliar o desempenho do modelo VGG16 sem Data Augmentation
196 print("VGG16 sem Data Augmentation\n")
197 print(classification_report(y_true, y_pred_vgg, target_names=
        class_labels))
198 conf_matrix_vgg = confusion_matrix(y_true, y_pred_vgg)
199 specificity_vgg = calculate_specificity(conf_matrix_vgg)
200 print(f"Especificidade: {specificity_vgg}")
201
202 # Avaliar o desempenho do modelo VGG16 com Data Augmentation
203 print("\nVGG16 com Data Augmentation\n")
204 print(classification_report(y_true, y_pred_vgg_aug, target_names=
        class_labels))
205 conf_matrix_vgg_aug = confusion_matrix(y_true, y_pred_vgg_aug)
206 specificity_vgg_aug = calculate_specificity(conf_matrix_vgg_aug)
207 print(f"Especificidade: {specificity_vgg_aug}")
208
209 # Avaliar o desempenho do modelo ResNet50 sem Data Augmentation
210 print("\nResNet50 sem Data Augmentation\n")
211 print(classification_report(y_true, y_pred_resnet, target_names=
        class_labels))
212 conf_matrix_resnet = confusion_matrix(y_true, y_pred_resnet)
213 specificity_resnet = calculate_specificity(conf_matrix_resnet)
214 print(f"Especificidade: {specificity_resnet}")
215
216 # Avaliar o desempenho do modelo ResNet50 com Data Augmentation
217 print("\nResNet50 com Data Augmentation\n")
218 print(classification_report(y_true, y_pred_resnet_aug, target_names=
        class_labels))
219 conf_matrix_resnet_aug = confusion_matrix(y_true, y_pred_resnet_aug)
220 specificity_resnet_aug = calculate_specificity(conf_matrix_resnet_aug
        )
221 print(f"Especificidade: {specificity_resnet_aug}")
222
223 #VGG16 sem Data Augmentation
224
225 #              precision recall f1-score support
226
227 #           0       0.77    0.90    0.83     101
228 #           1       0.52    0.18    0.26      90
229 #           2       0.63    0.91    0.75      90
230 #           3       0.97    0.99    0.98      90
231
```

```
#   accuracy                           0.75     371
#  macro avg   0.72    0.74    0.70     371
#weighted avg  0.72    0.75    0.71     371

#Especificidade: [0.9 0.94661922 0.82918149 0.98932384]

#VGG16 com Data Augmentation

#              precision recall f1-score support

#         0       0.81    0.99    0.89     101
#         1       0.81    0.29    0.43      90
#         2       0.65    0.84    0.73      90
#         3       0.92    1.00    0.96      90

#   accuracy                           0.79     371
#  macro avg   0.80    0.78    0.75     371
#weighted avg  0.80    0.79    0.76     371

#Especificidade: [0.91111111 0.97864769 0.85409253 0.97153025]

#ResNet50 sem Data Augmentation

#              precision recall f1-score support

#         0       0.99    0.97    0.98     101
#         1       0.93    0.56    0.69      90
#         2       0.69    0.97    0.81      90
#         3       0.98    1.00    0.99      90

#   accuracy                           0.88     371
#  macro avg   0.90    0.87    0.87     371
#weighted avg  0.90    0.88    0.87     371

#Especificidade: [0.9962963 0.98576512 0.86120996 0.99288256]

#ResNet50 com Data Augmentation

#              precision recall f1-score support

#         0       0.75    0.99    0.85     101
#         1       0.58    0.12    0.20      90
#         2       0.61    0.81    0.70      90
#         3       0.91    1.00    0.95      90

#   accuracy                           0.74     371
#  macro avg   0.71    0.73    0.68     371
#weighted avg  0.71    0.74    0.68     371

#Especificidade: [0.87407407 0.97153025 0.83629893 0.96797153]
```

```
283 # Resultados: O modelo que deu melhores resultados foi o **ResNet50
    sem Data Augmentation**, considerando a métrica de **sensibilidade
    ** (recall).
284
285
286 ### ResNet50 sem Data Augmentation - Sensibilidade
287 #*  **0.88**
```

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 do ChatGPT
RESUMO

Este estudo de caso investiga as implicações éticas do uso do ChatGPT em diferentes contextos e propõe soluções responsáveis para lidar com esses dilemas. A análise aborda questões éticas significativas, incluindo a disseminação de informações, privacidade dos usuários, vieses algorítmicos, discriminação e

responsabilidade. Soluções responsáveis são propostas para equilibrar os interesses das partes interessadas e promover valores fundamentais como transparência, justiça e privacidade.
Palavras-chave: Inteligência Artificial, Ética, ChatGPT, Privacidade, Vieses.

ABSTRACT This case study investigates the ethical implications of using ChatGPT in different contexts and proposes responsible solutions to address these dilemmas. The analysis addresses significant ethical issues, including information dissemination, user privacy, algorithmic biases, discrimination, and responsibility. Responsible solutions are proposed to balance stakeholder interests and promote fundamental values such as transparency, fairness, and privacy.
Keywords: Artificial Intelligence, Ethics, ChatGPT, Privacy, Biases.

1. INTRODUÇÃO

De acordo com Aruda (2024), em novembro de 2022, houve uma certa euforia e também assombro quando a empresa OpenAI tornou público e acessível a plataforma ChatGPT. Segundo a empresa OpenAI (2022), o ChatGPT é um modelo de inteligência artificial (IA) que interage de maneira conversacional, ou seja, ele foi programado para que haja um diálogo entre o usuário e a máquina, permitindo que a plataforma responda a perguntas, admita erros, conteste premissas incorretas e rejeite solicitações inadequadas.

Com a definição dada pelos próprios criadores da ferramenta, é possível identificar inúmeras lacunas relacionadas a questões éticas provenientes do processo de geração de respostas por meio de IA. Nesse sentido, alguns dilemas abordam problemáticas éticas relacionadas a: (1) privacidade e segurança dos dados; (2) desinformação e uso indevido; (3) viés discriminatório; (4) impactos nas formas de trabalho; (5) responsabilidade e transparência; (6) interações humanas; e, por último, (7) robustez e precisão das informações geradas automaticamente. A discussão acerca de cada item acima gera um cenário multifacetado, com perspectivas provenientes de diversas áreas. No âmbito da jurisprudência, o contexto torna-se ainda mais complexo, exigindo uma compreensão mais aprofundada para que as questões éticas sejam cuidadosamente examinadas.

Neste trabalho abordaremos três questões éticas, sendo elas: (1) privacidade e segurança dos dados; (2) desinformação e uso indevido e (7) robustez e precisão das informações geradas de forma automática. Quando se fala em privacidade dos dados no uso do ChatGPT, é importante ressaltar que a funcionalidade da ferramenta requer um grande volume de dados para o treinamento do modelo da IA por trás dela. Além disso, é necessário que as requisições sejam feitas pelo usuário. Desta forma, fica evidente que os dados são elementos protagonistas e participam de todas as camadas desta aplicação.

No entanto, muitos debates têm sido levantados acerca da responsabilização pela posse dos dados, sejam eles sensíveis ou não, que a plataforma possui. Uma vez que os dados são fornecidos como entrada, eles se tornam parte do sistema. Neste sentido, há abertura de brechas relacionadas ao uso e compartilhamento de dados sensíveis de forma indevida, causando danos quase que irreversíveis aos afetados. Segundo Wu et al. (2024) embora existam mecanismos de proteção de privacidade no ChatGPT, como o bloqueio de acesso a dados pessoais de indivíduos, não é garantido que não ocorra nenhum vazamento.

Aliado a isso, o modelo de IA (LLM - *Large Language Model*, tradução para o português: Modelo de Linguagem de Grande Escala) usado para a geração de informações pelo ChatGPT não é simples de entender, o que torna o processo pouco transparente e suscetível a gerar informações pouco precisas sobre determinado assunto e até informações falsas. Essa falta de transparência pode manipular a opinião pública, causando desinformação e influenciando decisões de forma negativa. Para elucidar estas questões o trabalho discutirá sobre cada um dos itens de forma mais aprofundada trazendo alguns exemplos reais dos impactos gerados pelos dilemas éticos.

2. DESENVOLVIMENTO

A revisão de literatura foca nos estudos prévios relacionados ao uso ético da IA e os desafios associados à implementação destes sistemas. Neste sentido, esta seção tem como objetivo trazer as definições necessárias presentes na literatura e os debates acerca dos impactos gerados por cada um dos dilemas éticos latentes no assunto.

2.1 Privacidade e segurança dos dados

Para Westin (1967): A privacidade na era digital é um direito fundamental que deve ser protegido com rigor. A coleta massiva de dados pessoais sem o consentimento adequado dos usuários representa uma ameaça significativa à autonomia e à dignidade individual.

A privacidade dos usuários é uma questão em evidência neste momento em que a IA está sendo inserida na sociedade e as pessoas começam a conhecer e utilizar. A partir do momento em que interagimos com o ChatGPT, ele armazena todas as informações em sua base de dados para treinamento e isso implica que esses dados se tornarão públicos, a partir do momento em que se faz uma pergunta e a ferramenta entender que aqueles dados inseridos por outra pessoa são os melhores para uma resposta a qualquer usuário. Então, surge a grande preocupação quanto a privacidade dos dados, que dados serão utilizados para a análise, que dados serão apresentados nas respostas, existe algum contato humano com esta base de dados? Não basta somente informar ao usuário de que ele deve evitar inserir dados sensíveis, e sim, o ChatGPT deve utilizar critérios éticos que aperfeiçoem a ferramenta para que ela analise esses dados e gere uma nova conclusão e resposta, considerando os critérios da ética e segurança e dando a eles o máximo de importância.

No estudo apresentado por Khowaja et. al (2024) são discutidas as preocupações acerca da privacidade dos dados usados para o treinamento de LLMs, em especial do ChatGPT. Os autores mencionam que os dados utilizados para o treinamento do ChatGPT são sistematicamente extraídos de mensagens, websites, artigos, livros e informações pessoais sem a devida autorização. Além disso, de acordo com a política de privacidade do ChatGPT (OpenAI, 2024b), são coletadas informações como dados de interação do usuário com o site, configurações e tipo de navegador, e endereço IP, juntamente com o tipo de conteúdo com o qual os usuários interagem no ChatGPT. Eles também coletam informações sobre atividades de navegação em diferentes sites e ao longo de um determinado período de tempo. A política de privacidade também afirma que "Além disso, de tempos em tempos, podemos analisar o comportamento geral e as características dos usuários de nossos serviços e compartilhar informações agregadas, como estatísticas gerais de usuários, com terceiros, publicar essas informações agregadas ou torná-las geralmente disponíveis. Podemos coletar informações agregadas através dos Serviços, através de cookies e por outros meios descritos nesta política de privacidade"e ainda que "em certas circunstâncias, podemos fornecer suas informações pessoais a terceiros sem aviso prévio, a menos que exigido por lei".

2.2 Desinformação e uso indevido

Segundo Wardle e Derakhshan (2017): A desinformação, especialmente em plataformas digitais, pode ter consequências graves para a sociedade. Ela pode manipular opiniões, influenciar decisões eleitorais e até mesmo causar pânico em situações de crise.

Atualmente, a internet é a principal fonte de informação. Desta forma, o desafio não está apenas em obter um conteúdo relevante, mas também em filtrar as informações incorretas da grande quantidade disponível. Antes do lançamento do ChatGPT, as pessoas utilizavam outros métodos para confirmar a veracidade das informações, avaliando o nível de conhecimento do criador do conteúdo pelo rigor da linguagem, correção do formato e a fonte do texto, sendo esses indicadores de confiabilidade. No entanto, ao utilizar o ChatGPT, é possível observar que esses indicadores são excelentes nesses aspectos, dando a falsa sensação de confiabilidade. Os usuários podem cair na armadilha de confiar cegamente no conteúdo gerado pelo ChatGPT depois de executar testes simples. Essa confiança cega pode levar a julgamentos errados devido a hábitos arraigados. Em áreas críticas, como documentos de pesquisa médica, informações básicas de experimentos em larga escala e conteúdo de políticas, fazer referência a informações errôneas do ChatGPT e tirar conclusões incorretas pode ter consequências imprevisíveis. Esses riscos não decorrem de fraude intencional, mas sim dos erros do ChatGPT, apesar de sua aparência inicialmente confiável (Wu et al., 2024).

Outro aspecto muito importante relacionado ao uso do ChatGPT é a geração de notícias falsas, que têm implicações éticas de bastante relevância social e política e já são um dos principais problemas que vêm sendo enfrentados no Brasil. Estes Modelos de Linguagem Grande (LLMs) são muito populares e podem ser facilmente acessados para a geração de informações falsas ou enganosas que podem facilmente serem disseminadas nas redes sociais e plataforma digitais, podendo ter com consequências graves, como

por exemplo a manipulação de eleições e a incitação à violências (Khowaja et. al, 2024). Até que ponto as informações geradas pelo ChatGPT são verdadeiras? Esse é um grande desafio, pelo motivo de que em muitos momentos as informações necessitam ser validadas, muitas vezes depende de decisões que ainda precisam ser formuladas e votadas pela sociedade, categorias e governos. O fato de que as informações geradas pelo ChatGPT são originadas de uma base de dados e esta base foi construída pelo próprio ChatGPT, leva a conclusão de que esta base de dados deve ser constantemente validada, se possível por um órgão neutro externo que controle a legitimidade das informações e as distribua de forma a guardar a ética e os direitos de todas as partes envolvidas. Atualmente existe a versão gratuita e a versão paga do ChatGPT, e a principal diferença entre elas é a atualização da base de dados e funcionalidades. Já foi constatado que em determinados questionamentos ao ChatGPT, a ferramenta informou erroneamente e isso é muito sério, pois, a maior fatia de utilização é pela opção gratuita, que é a com a base de dados desatualizada e que fornece grandes vulnerabilidades. Uma regulamentação que defina a forma que esses dados são validados é de fundamental importância para garantir a segurança e confiabilidade das informações.

2.3 Robustez e precisão das informações geradas

Para O'Neil (2016): A automação na tomada de decisões promete eficiência e objetividade, mas também apresenta riscos consideráveis. Algoritmos podem perpetuar vieses existentes nos dados de treinamento, levando a decisões injustas ou discriminatórias.

Outro desafio é a tomada de decisões automatizadas, que dá poder para a ferramenta decidir o que fazer com base nas regras que foram criadas por seus programadores. Um grande exemplo são as IAs de recrutamento e seleção, que fazem a triagem com base em critérios definidos e podem muitas vezes serem injustas. Com relação ao ChatGPT, eu posso desde solicitar a criação de um código em determinada linguagem que diante de um perfil pré-selecionado busque a opção ideal em sua base, posso pedir para que ele gere uma imagem como por exemplo pedir uma imagem de uma pessoa bonita, e ele irá criar uma imagem de uma pessoa branca, loira, de olhos claros, ou seja, seguindo critérios já definidos pelos programadores ou seguindo uma tendência das imagens buscadas na internet. Deixar que a IA faça a escolha sozinha oferece riscos, pois ela foi configurada para buscar a melhor solução em sua base de dados e ela pode estar repleta de vieses.

3. CONCLUSÃO

Através de discussões em grupo e colaboração entre os membros da equipe, chegamos a algumas propostas para lidar com as implicações éticas do uso do ChatGPT. Nossa análise destacou a importância de criar regulamentações claras para garantir que o uso da IA respeite a privacidade dos usuários e evite a disseminação de informações falsas. Essas regulamentações devem definir padrões rigorosos para a coleta, armazenamento e uso dos dados, assegurando a proteção das informações pessoais.

Além disso, identificamos a necessidade de promover a transparência nos processos de decisão do ChatGPT. Isso pode ser alcançado através da implementação de mecanismos de auditoria e monitoramento contínuo, permitindo verificar e corrigir vieses algorítmicos, reduzindo o risco de discriminação.

Durante nossas discussões, também enfatizamos a importância de práticas de design ético na criação e atualização dos modelos de IA. Envolver equipes multidisciplinares e incluir feedback de usuários e especialistas externos pode contribuir para a criação de um sistema mais robusto e justo.

Finalmente, destacamos a necessidade de educar os usuários sobre os potenciais riscos e limitações do ChatGPT. Programas de conscientização e treinamentos específicos podem capacitar os usuários a utilizarem a IA de forma crítica e responsável, reconhecendo a importância da verificação das informações fornecidas pela ferramenta.

Em resumo, através da colaboração entre os membros da equipe e discussões aprofundadas, conseguimos desenvolver soluções práticas para lidar com as implicações éticas do uso do ChatGPT. As soluções propostas incluem regulamentações claras, transparência nos processos de decisão, práticas de design ético e educação do usuário, promovendo um uso mais seguro e justo dessa tecnologia.

4. REFERÊNCIAS

Aruda, E. P. Inteligência Artificial Generativa No Contexto Da Transformação do Trabalho Docente. 2024. Disponível em: <https://doi.org/10.1590/0102-469848078>.

Deng, J.; Lin, Y. The benefits and challenges of ChatGPT: An overview. Frontiers in Computing and Intelligent Systems, v. 2, n. 2, p. 81-83, 2022.
Khowaja, S. A.; Khuwaja, P.; Dev, K.; Wang, W.; Nkenyereye, L. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. Cognitive Computation, p. 1-23, 2024.
O'Neil, C. Armas de Destruição Matemática: Como o Big Data Aumenta a Desigualdade e Ameaça a Democracia. Nova York: Crown Publishing Group, 2016.
OpenAI [internet]. Introducing ChatGPT. Disponível em: <https://openai.com/index/chatgpt/>. Acessado em 01 de Julho de 2024a.
OpenAi [internet]. Privacy policy. Disponível em: <https://openai.com/policies/privacy-policy/>. Acesso em 8 de Julho de 2024b.
Wardle, C.; Derakhshan, H. Desordem da Informação: Rumo a uma Estrutura Interdisciplinar para Pesquisa e Formulação de Políticas. Conselho da Europa, 2017.
Westin, A. F. Privacidade e Liberdade. Nova York: Atheneum, 1967. Wu, X.; Ran, D.; Jianbing, N. Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence, v. 2, n. 2, p. 102-115, 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**

Artigo: *Requirements practices and gaps when engineering human-centered Artificial Intelligence systems*

De acordo com os autores, o principal objetivo do estudo apresentado é identificar lacunas nas práticas atuais de engenharia de requisitos (RE) para IA (RE4IA). Os autores também apresentam duas perguntas de pesquisa a serem respondidas ao longo do estudo. A primeira está relacionada ao mapeamento das diretrizes existentes na pesquisa e na indústria para o desenvolvimento de IA centrada no usuário (CNU), bem como aos aspectos que ainda estão ausentes. A segunda pergunta trata da identificação de lacunas entre a pesquisa sobre RE4IA e as práticas aplicadas na indústria.

Qual o objetivo do estudo descrito pelo artigo?

A engenharia de software que utiliza IA é uma área emergente que apresenta inúmeros desafios. Grandes empresas utilizam diretrizes para auxiliar as equipes no desenvolvimento de sistemas de IA centrados no usuário. Entretanto, as práticas adotadas por profissionais para o desenvolvimento de sistemas em RE ainda são pouco relatadas. Os autores destacam que o foco principal da IA nos últimos anos tem sido a construção de sistemas tecnicamente sólidos. No entanto, ao priorizar os aspectos técnicos e negligenciar os aspectos humanos no desenvolvimento de sistemas de IA, podem surgir danos físicos e mentais. Há uma lacuna associada às abordagens centradas no usuário (CNU), o que pode resultar em sistemas de IA enviesados. Os autores também mencionam exemplos de vieses cometidos por sistemas de IA, relacionados a fatores como raça, gênero, cor, religião, estado de saúde e pessoas com deficiência. Recentemente, observa-se uma tendência em explorar aspectos centrados no ser humano no desenvolvimento de softwares com IA. Essa tendência é evidenciada por uma SLR, que revela que a maioria dos softwares ainda carece de uma abordagem centrada no ser humano em relação à modelagem de requisitos. Os aspectos abordados na SLR indicam que as pesquisas existentes em RE para IA têm se concentrado principalmente em temas como ética, confiança e explicabilidade, embora as avaliações empíricas sejam limitadas ou inexistentes. Além disso, nenhum dos estudos investigou os aspectos centrados no ser humano adotados pela indústria ou como eles devem ser abordados. na RE.

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

Ao longo da leitura, percebe-se que os autores realizaram uma pesquisa sistemática da literatura (SLR) sobre as diretrizes do Google, Apple, Microsoft e ML Canvas, com o objetivo de compreender como os sistemas dessas empresas abordam a interação entre o usuário e a IA. Os autores descrevem brevemente o funcionamento de cada uma dessas diretrizes e, em seguida, conduzem uma SLR em trabalhos que utilizam o RE4IA no período de 2010 a 2021. Esses trabalhos são categorizados em dois grupos: CNU e técnicos. Após esse mapeamento, os autores dividem os resultados em categorias e os analisam com base nos seis capítulos das diretrizes do Google, que abrangem as seguintes áreas: necessidade do usuário, necessidade do modelo, necessidade dos dados, feedback e controle, explicabilidade e confiança, erros e falhas. Após o mapeamento, os autores realizaram uma pesquisa com especialistas no desenvolvimento de sistemas de IA para compreender quais aspectos das diretrizes identificadas deveriam ser considerados na RE. Nesse contexto, elaboraram um questionário para investigar os desafios e limitações enfrentados pelos profissionais ao desenvolver sistemas de IA. O questionário também incluía perguntas sobre quais diretrizes em CNU, identificadas na SLR, esses especialistas consideravam ao construir um sistema de IA.

Quais os principais resultados obtidos pelo estudo?

Durante o mapeamento das pesquisas sobre diretrizes realizadas tanto na indústria quanto na academia para a RE, os autores identificaram algumas diferenças importantes. Enquanto a indústria se preocupa principalmente com os tipos e a diversidade dos dados, a academia foca na quantidade e na qualidade dos dados. Na pesquisa com especialistas da área de IA, os autores obtiveram resultados que destacam as áreas de atuação dos participantes, sendo a maioria composta por cientistas de dados e pesquisadores em IA. Outro aspecto interessante identificado foi a presença de especialistas atuando amplamente nos setores de educação e governo. Os autores também relataram os resultados obtidos em relação à forma como os requisitos são construídos pelos especialistas. A análise mostrou que a maior preocupação desses profissionais está em atender aos requisitos funcionais dos sistemas de IA. Sobre as ferramentas utilizadas para levantamento de requisitos, destacaram-se PowerPoint, Word, Email, Excel, Confluence, Jira, entre outras. Além disso, os autores abordaram as linguagens utilizadas pelos especialistas durante o processo de modelagem de requisitos, com o Caso de Uso sendo o formato mais frequente. Um ponto relevante levantado pelos especialistas foi a identificação de desafios recorrentes durante o processo de RE. O principal desafio apontado por eles durante a RE refere-se à “lacuna de viabilidade”, ou seja, as limitações dos modelos de IA em relação ao que podem ou não realizar. Os autores indicam que as diretrizes centradas no ser humano devem ser amplamente consideradas durante RE e que os especialistas também indicaram essa questão em suas respostas para o desenvolvimento de RE4IA.

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.

```
1    data = tf.keras.datasets.fashion_mnist
2    (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.

```
1
2  url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine
       -quality/winequality-red.csv"
3  data = pd.read_csv(url, delimiter=';')
4
5  Dica 1. Para facilitar o trabalho, renomeie o nome das colunas para
       português, dessa forma:
6
7  data.columns = [
8     'acidez_fixa',       # fixed acidity
9     'acidez_volatil',    # volatile acidity
10    'acido_citrico',     # citric acid
11    'acucar_residual',   # residual sugar
12    'cloretos',          # chlorides
13    'dioxido_de_enxofre_livre', # free sulfur dioxide
14    'dioxido_de_enxofre_total', # total sulfur dioxide
15    'densidade',         # density
16    'pH',                # pH
17    'sulfatos',          # sulphates
18    'alcool',            # alcohol
19    'score_qualidade_vinho'      # quality
20 ]
21
22 Dica 2. Separe os dados (x e y) de tal forma que a última coluna (í
       ndice -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.

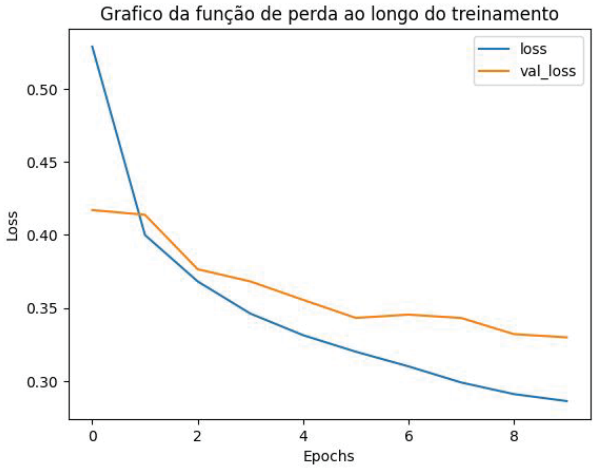
```

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

1 Classificação (RNA)

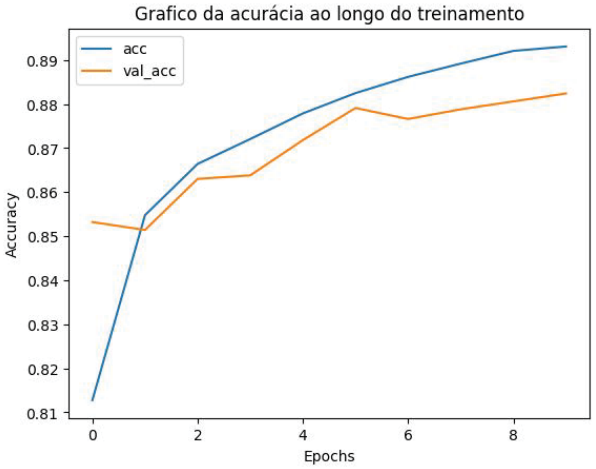
```
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
from mlxtend.plotting import plot_confusion_matrix
from sklearn.metrics import confusion_matrix

(X_train, Y_train), (X_test, Y_test) = tf.keras.datasets.
    fashion_mnist.load_data()

x_train, x_test = X_train/255.0, X_test/255.0

i = tf.keras.layers.Input(shape=(28,28))
x = tf.keras.layers.Flatten()(i)
x = tf.keras.layers.Dense(128, activation='relu')(x)
x = tf.keras.layers.Dropout(0.2)(x)
x = tf.keras.layers.Dense(10, activation='softmax')(x)
model = tf.keras.Model(i, x)

model.compile(optimizer='adam', loss='sparse_categorical_crossentropy
    ', metrics=['accuracy'])

r = model.fit(x_train, Y_train, validation_data=(x_test, Y_test),
    epochs=10)

plt.plot(r.history['loss'], label='loss')
plt.plot(r.history['val_loss'], label='val_loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.title('Grafico da função de perda ao longo do treinamento')
plt.legend()
```

Figura 13.1 – EXERCÍCIO 1: GRÁFICO DE PERDA



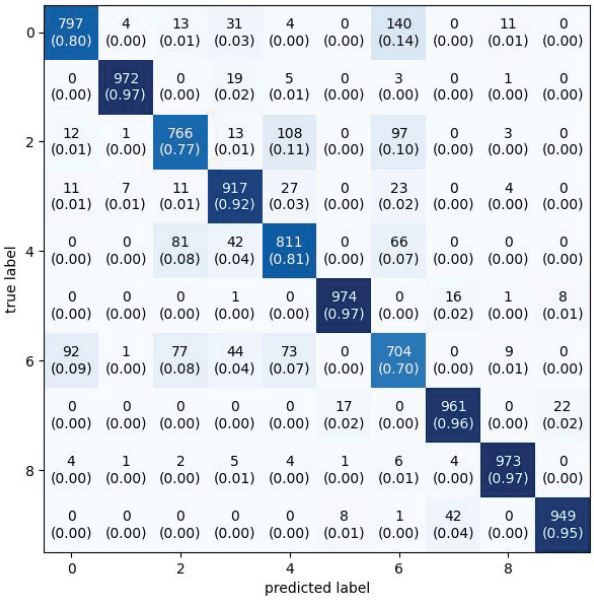

```
plt.plot(r.history['accuracy'], label='acc')
plt.plot(r.history['val_accuracy'], label='val_acc')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.title('Grafico da acurácia ao longo do treinamento')
plt.legend()
```

Figura 13.2 – EXERCÍCIO 1: GRÁFICO DE ACURÁCIA



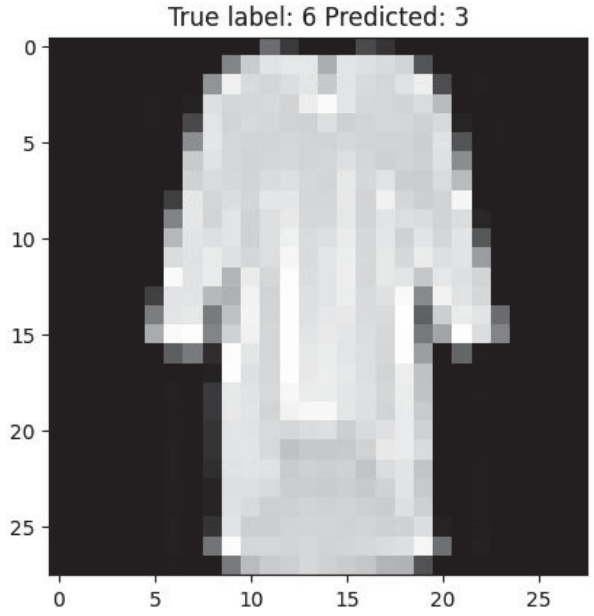

```
y_pred = model.predict(x_test).argmax(axis=1)
cm = confusion_matrix(Y_test, y_pred)
plot_confusion_matrix(conf_mat=cm, figsize=(7,7), show_normed=True)
```

Figura 13.3 – EXERCÍCIO 1: MATRIZ DE CONFUSÃO




```
misclassified = np.where(y_pred != Y_test)[0]
i = np.random.choice(misclassified)
plt.imshow(x_test[i], cmap='gray')
plt.title("True label: %s Predicted: %s" % (Y_test[i], y_pred[i]));
```

Figura 13.4 – EXERCÍCIO 1: APRESENTANDO CLASSIFICAÇÃO INCORRETA



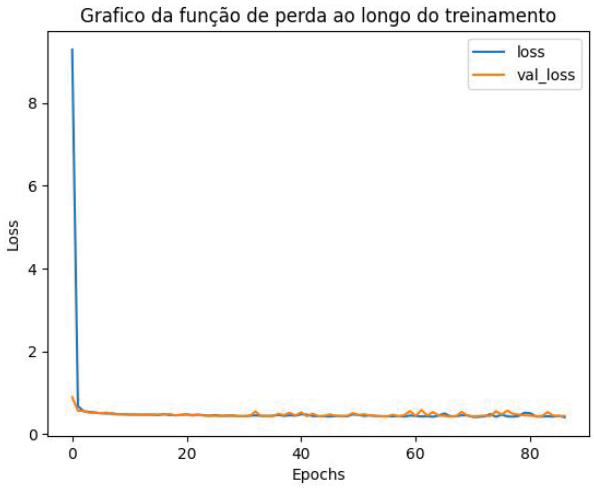
2 Regressão (RNA)

```
import tensorflow as tf
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
5  from tensorflow.python.keras import backend
6  from sklearn.model_selection import train_test_split
7  from sklearn.metrics import r2_score, mean_squared_error
8  from math import sqrt
9
10 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine
       -quality/winequality-red.csv"
11
12 data = pd.read_csv(url, delimiter=';')
13
14 data.columns = [
15 'acidez_fixa', # fixed acidity
16 'acidez_volatil', # volatile acidity
17 'acido_citrico', # citric acid
18 'acucar_residual', # residual sugar
19 'cloretos', # chlorides
20 'dioxido_de_enxofre_livre', # free sulfur dioxide
21 'dioxido_de_enxofre_total', # total sulfur dioxide
22 'densidade', # density
23 'pH', # pH
24 'sulfatos', # sulphates
25 'alcool', # alcohol
26 'score_qualidade_vinho' # quality
27 ]
28
29 data = data.values
30
31 # Dica 2. Separe os dados (x e y) de tal forma que a última coluna (í
       ndice -1), chamada score_qualidade_vinho
32
33 X = data[:, 0:-1].astype(float)
34 Y = data[:, -1].astype(float)
35
36 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size
       =0.25)
37
38 i = tf.keras.layers.Input(shape=(11,))
39 x = tf.keras.layers.Dense(100, activation='relu')(i)
40 x = tf.keras.layers.Dense(1)(x)
41 model = tf.keras.Model(i, x)
42
43 # Criação de funções para as métricas r2 e RMSE serem inseridas no
       modelo
44 def rmse(y_true, y_pred):
45     return backend.sqrt(backend.mean(backend.square(y_true- y_pred),
           axis=-1))
46
47 def r2(y_true, y_pred):
48     media = backend.mean(y_true)
49     num = backend.sum(backend.square(y_true - y_pred))
50     den = backend.sum(backend.square(y_true - media))
```

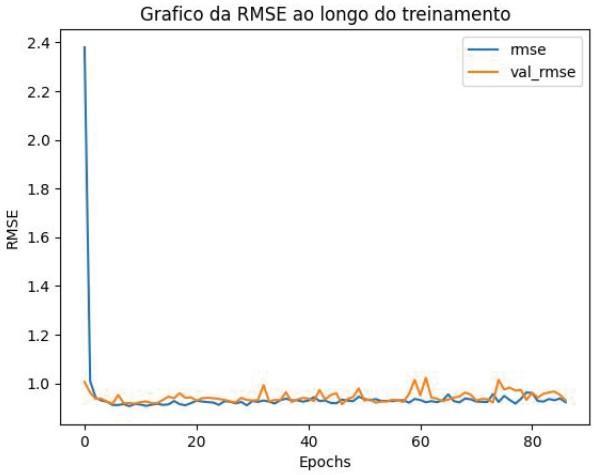
```
    return (1.0 - num/den)

# Compilação
optimizer = tf.keras.optimizers.Adam(learning_rate=0.001)
model.compile(optimizer=optimizer, loss='mse', metrics=[r2, rmse])

# Early Stop para Epochs
early_stop = tf.keras.callbacks.EarlyStopping(monitor='val_loss',
    patience=20, restore_best_weights=True)

r = model.fit(X_train, Y_train, validation_data=(X_test, Y_test),
    epochs=1500, callbacks=[early_stop])

plt.plot(r.history['loss'], label='loss')
plt.plot(r.history['val_loss'], label='val_loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.title('Grafico da função de perda ao longo do treinamento')
plt.legend()
```

Figura 13.5 – EXERCÍCIO 2: GRÁFICO DE PERDA



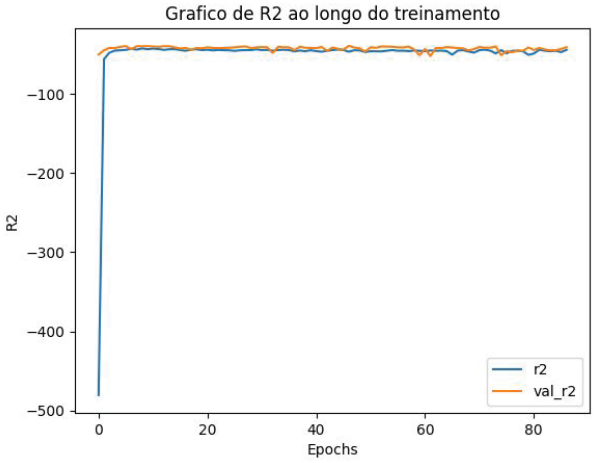

```
plt.plot(r.history['rmse'], label='rmse')
plt.plot(r.history['val_rmse'], label='val_rmse')
plt.xlabel('Epochs')
plt.ylabel('RMSE')
plt.title('Grafico da RMSE ao longo do treinamento')
plt.legend()
```

Figura 13.6 – EXERCÍCIO 2: GRÁFICO DE RMSE




```
plt.plot(r.history['r2'], label='r2')
plt.plot(r.history['val_r2'], label='val_r2')
plt.xlabel('Epochs')
plt.ylabel('R2')
plt.title('Grafico de R2 ao longo do treinamento')
plt.legend()
```

Figura 13.7 – EXERCÍCIO 2: GRÁFICO DE R2




```
y_pred = model.predict(X_test).flatten()
mse = mean_squared_error(Y_test, y_pred)
rmse = sqrt(mse)
r2 = r2_score(Y_test, y_pred)

print("mse:", mse)
print("rmse:", rmse)
print("r2:", r2)
```

mse: 0.4207298387142231
rmse: 0.6486369082269549

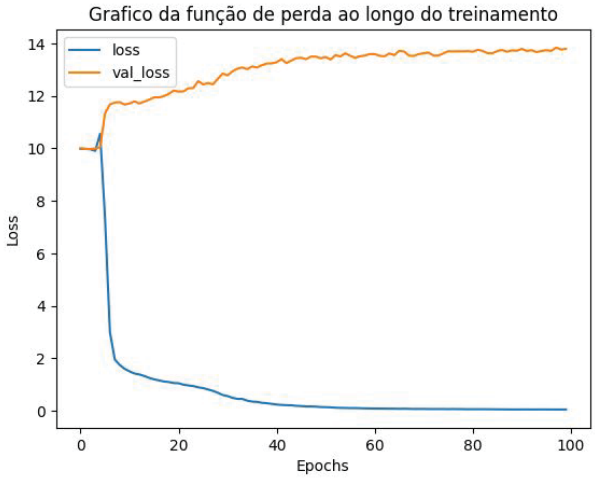
r2: 0.38145588854025325

3 Sistemas de Recomendação

```
import numpy as np
import pandas as pd
import tensorflow as tf
import matplotlib.pyplot as plt
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, Embedding, Dense,
    Embedding, Flatten, Concatenate
from tensorflow.keras.optimizers import SGD, Adam
from sklearn.utils import shuffle

df = pd.read_csv('Base_livros(in).csv', sep=',', on_bad_lines='skip')
df['Notas;;;;;;'] = df['Notas;;;;;;'].str.replace(r'[^0-9]', '',
    regex=True).astype(float)
df.rename(columns={'Notas;;;;;;': 'Notas'}, inplace=True)
df.dropna(inplace=True)

categorical_columns = ['ID_usuario', 'Titulo', 'Autor', 'Editora']
for cat_column in categorical_columns:
    df[cat_column] = pd.Categorical(df[cat_column])
    df['new_' + cat_column] = df[cat_column].cat.codes

N = len(set(df['ID_usuario']))
M = len(set(df['new_Titulo']))

# dimensão do embedding
K = 10

# usuário
u = Input(shape=(1,))
u_emb = Embedding(N, K)(u) # saída : num_samples, 1, K
u_emb = Flatten()(u_emb) # saída : num_samples, K
# filme
m = Input(shape=(1,))
m_emb = Embedding(M, K)(m) # saída : num_samples, 1, K
m_emb = Flatten()(m_emb) # saída : num_samples, K
x = Concatenate()([u_emb, m_emb])
x = Dense(1024, activation="relu")(x)
x = Dense(1)(x)
model = Model(inputs=[u, m], outputs=x)

model.compile(loss="mse", optimizer=SGD(learning_rate=0.08, momentum
    =0.9))

user_ids, movie_ids, ratings = shuffle(df['new_ID_usuario'], df['
    new_Titulo'], df['Notas'])
Ntrain = int(0.8 * len(ratings)) # separar os dados 80% x 20%
train_user = user_ids[:Ntrain]
train_movie = movie_ids[:Ntrain]
```

```
train_ratings = ratings[:Ntrain]
test_user = user_ids[Ntrain:]
test_movie = movie_ids[Ntrain:]
test_ratings = ratings[Ntrain:]

# centralizar as notas
avg_rating = train_ratings.mean()
train_ratings = train_ratings - avg_rating
test_ratings = test_ratings - avg_rating

epochs = 100
r = model.fit(
x=[train_user, train_movie],
y=train_ratings,
epochs=epochs,
batch_size=1024,
verbose=2, # não imprime o progresso
validation_data=([test_user, test_movie], test_ratings)
)

plt.plot(r.history['loss'], label='loss')
plt.plot(r.history['val_loss'], label='val_loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.title('Grafico da função de perda ao longo do treinamento')
plt.legend()
```

Figura 13.8 – EXERCÍCIO 3: GRÁFICO DE PERDA




```
input_usuario = np.repeat(a=8263, repeats=M)
film = np.array(list(set(movie_ids)))
preds = model.predict([input_usuario, film])

# descentraliza as predições
rat = preds.flatten() + avg_rating
# índice da maior nota
```

```
idx = np.argmax(rat)
print("Recomendação: Filme - ", film[idx], " / ", rat[idx] , "*")
```

Recomendação: Filme - 66366 / 19.101202 *

4 Deepdream

```
import tensorflow as tf
import numpy as np
import matplotlib as mpl
import IPython.display as display
import PIL.Image

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

# Download da imagem e gravação em array Numpy
def download(url, max_dim=None):
   name = url.split('/')[-1]
   image_path = tf.keras.utils.get_file(name, origin=url)
   img = PIL.Image.open(image_path)
   if max_dim:
   img.thumbnail((max_dim, max_dim))
   return np.array(img)

# Normalização da imagem
def deprocess(img):
   img = 255*(img + 1.0)/2.0
   return tf.cast(img, tf.uint8)

# Mostra a imagem
def show(img):
   display.display(PIL.Image.fromarray(np.array(img)))

# Redução do tamanho da imagem para facilitar o trabalho da RNN
original_img = download(url, max_dim=500)
show(original_img)
```

Figura 13.9 – EXERCÍCIO 3: IMAGEM BAIXADA DA URL FORNECIDA

```
base_model = tf.keras.applications.InceptionV3(include_top=False,
    weights='imagenet')

# Maximizando as ativações das camadas
names = ['mixed3', 'mixed5']
layers = [base_model.get_layer(name).output for name in names]

# Criação do modelo
dream_model = tf.keras.Model(inputs=base_model.input, outputs=layers)

def calc_loss(img, model):
    # Passe a imagem pelo modelo para recuperar as ativações.
    # Converte a imagem em um batch de tamanho 1.
    img_batch = tf.expand_dims(img, axis=0)
    layer_activations = model(img_batch)

    if len(layer_activations) == 1:
        layer_activations = [layer_activations]

    losses = []
    for act in layer_activations:
        loss = tf.math.reduce_mean(act)
        losses.append(loss)

    return tf.reduce_sum(losses)

class DeepDream(tf.Module):
    def __init__(self, model):
        self.model = model
    @tf.function(
        input_signature=(
            tf.TensorSpec(shape=[None,None,3], dtype=tf.float32),
            tf.TensorSpec(shape=[], dtype=tf.int32),
            tf.TensorSpec(shape=[], dtype=tf.float32),)
    )
    def __call__(self, img, steps, step_size):
        print("Tracing")
        loss = tf.constant(0.0)

        for n in tf.range(steps):
            with tf.GradientTape() as tape:
                # Gradientes relativos a img
                tape.watch(img)
                loss = calc_loss(img, self.model)

            # Calculo do gradiente da perda em relação aos pixels da
              imagem de entrada.
            gradients = tape.gradient(loss, img)

            # Normalizacao dos gradintes
            gradients /= tf.math.reduce_std(gradients) + 1e-8
```

```
50
51        # Na subida gradiente, a "perda" émaximizada.
52        # Você pode atualizar a imagem adicionando diretamente os
              gradientes (porque eles têm o mesmo f
53        img = img + gradients*step_size
54        img = tf.clip_by_value(img, -1, 1)
55
56        return loss, img
57
58  deepdream = DeepDream(dream_model)
59
60  def run_deep_dream_simple(img, steps=100, step_size=0.01):
61     img = tf.keras.applications.inception_v3.preprocess_input(img)
62     img = tf.convert_to_tensor(img)
63     step_size = tf.convert_to_tensor(step_size)
64     steps_remaining = steps
65     step = 0
66     while steps_remaining:
67        if steps_remaining>100:
68           run_steps = tf.constant(100)
69        else:
70           run_steps = tf.constant(steps_remaining)
71
72        steps_remaining -= run_steps
73        step += run_steps
74
75        loss, img = deepdream(img, run_steps, tf.constant(step_size))
76
77        display.clear_output(wait=True)
78        show(deprocess(img))
79        print ("Step {}, loss {}".format(step, loss))
80
81     result = deprocess(img)
82     display.clear_output(wait=True)
83     show(result)
84
85     return result
86
87  dream_img = run_deep_dream_simple(img=original_img, steps=100,
       step_size=0.01)
```

Figura 13.10 – EXERCÍCIO 3: RESULTADO OBTIDO NO PROCESSO DE DEEP DREAM

```
import time
start = time.time()

OCTAVE_SCALE = 1.30
img = tf.constant(np.array(original_img))
base_shape = tf.shape(img)[:-1]
float_base_shape = tf.cast(base_shape, tf.float32)
for n in range(-2, 3):
    new_shape = tf.cast(float_base_shape*(OCTAVE_SCALE**n), tf.int32)
    img = tf.image.resize(img, new_shape).numpy()
    img = run_deep_dream_simple(img=img, steps=50, step_size=0.01)

display.clear_output(wait=True)
img = tf.image.resize(img, base_shape)
img = tf.image.convert_image_dtype(img/255.0, dtype=tf.uint8)
show(img)
end = time.time()
end-start
```

Figura 13.11 – EXERCÍCIO 3: RESULTADO OBTIDO NO PROCESSO DE DEEP DREAM À OITAVA

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

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

Devido ao grande volume de linhas que compõem o arquivo com os dados brutos do conjunto de dados utilizado nesta atividade, esses dados foram posicionados ao final deste documento.

2) Explicação do contexto e o público-alvo da visualização de dados e do storytelling que será desenvolvido;

O conjunto de dados escolhido para esta atividade consiste em um relatório contendo a listagem de todos os filmes, séries e programas disponíveis na Netflix. Os dados foram obtidos por meio da plataforma Kaggle, disponível neste link. Segundo a descrição do conjunto de dados, a Netflix é uma das plataformas de streaming de mídia mais populares, com mais de 8 mil filmes e programas de TV disponíveis. Em 2024, a plataforma contava com mais de 282 milhões de assinantes em todo o mundo.

O público-alvo para a visualização destes dados pode abranger diferentes grupos, dentre eles podemos destacar:

- Pesquisadores que possuem interesse em compreender as tendências de consumo, análise de representatividade, impacto na indústria do entretenimento;
- Profissionais do marketing que possuem interesse em compreender as preferências do público (popularidade) para criar campanhas segmentadas;
- Criadores de conteúdo que possuem interesse em compartilhar assuntos baseados na tendência;
- Críticos do cinema que possuem interesse em analisar padrões de produção e prever o sucesso ou fracasso do gênero;
- Empresas que queiram concorrer com a Netflix no sentindo de compreender como é realizado o engajamento da plataforma com seu público;
- Investidores que possuem interesse em analisar as tendências do setor de streaming e visam enxergar oportunidades de investimentos;

- Jornalistas e especialistas em entretenimento que possuem interesse em relatar sobre a plataforma e realizar comparações com outras soluções do mercado. Sendo estes, também, fortemente relacionados com os criadores de conteúdos citados acima.

As colunas do conjunto de dados são:

- show_id: trata-se de um identificador do item dentro do conjunto;
- type: trata-se de uma especificação entre as categorias a qual a mídia pertence, podendo ser movie (filme) ou TV show (programa de TV);
- title: o nome do título dentro da Netflix;
- director: diretor do título;
- country: país qual o título foi produzido;
- date_added: corresponde a data a qual o título foi adicionado a plataforma;
- realease_year: corresponde ao ano que o título foi lançado;
- rating: classificação do conteúdo do título;
- duration: duração do título em minutos;
- listed_in: categorias ou gêneros em que o título se enquadra;
- description: uma breve descrição sobre o título.

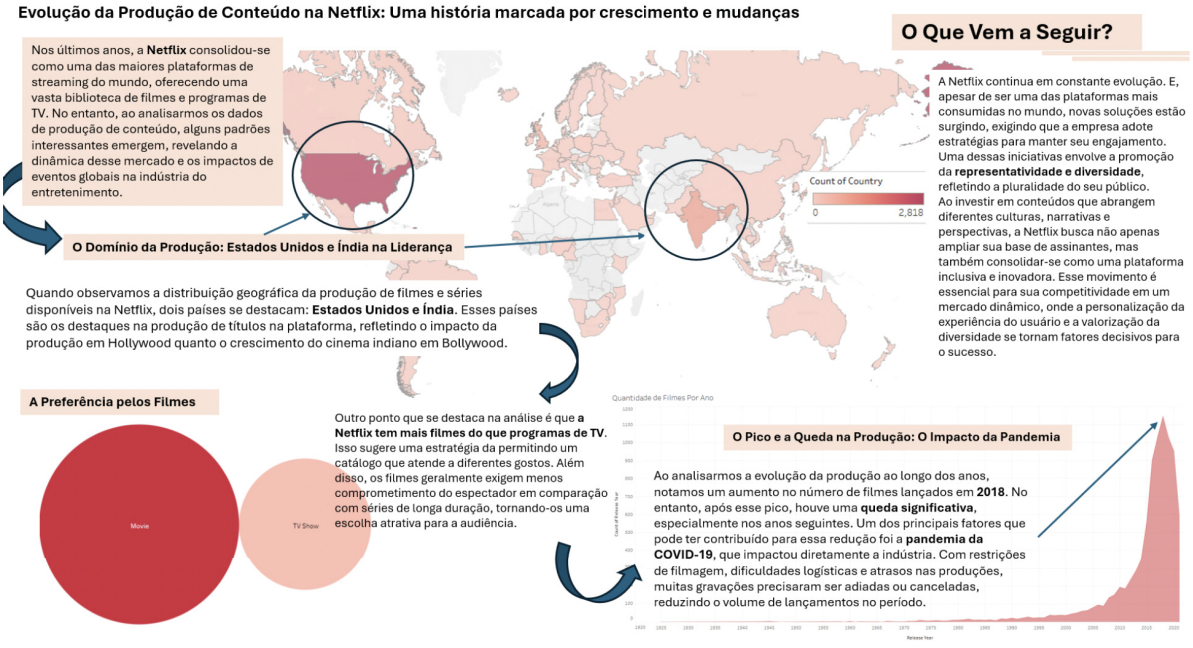
3) A visualização desses dados (de acordo com os dados escolhidos e com a ferramenta de sua escolha) explicando a escolha do tipo de visualização e da ferramenta usada; (50 pontos);

A ferramenta Public Tableau foi selecionada para gerar as visualizações deste trabalho. Ela foi escolhida devido à sua facilidade de uso e à variedade de opções e sugestões automáticas que fornece. Neste trabalho foram escolhidas algumas visualizações para descrever o perfil dos títulos presentes no Netflix. As visualizações escolhidas foram:

- Quantidade de títulos por país, indicando a geolocalização. A prevalência é representada por uma escala de cor vermelha (escolhida por remeter à identidade visual da plataforma), onde tons mais intensos indicam um maior número de filmes produzidos. Nesta visualização, é possível perceber que os Estados Unidos e a Índia são os dois países com o maior número de títulos na Netflix.
- Distribuição por tipos de mídia dentro da plataforma, distinguindo por tamanho o volume de itens que pertencem as categorias filme e programas de TV;
- Quantidade de filmes lançados por ano indicando o crescimento e a queda na produção podendo ser relacionado com acontecimentos no mundo.

4) A descrição da narrativa/storytelling dessa visualização de dados. (50 pontos).

Figura 14.1 – EVOLUÇÃO DA PRODUÇÃO DO CONTEÚDO NA NETFLIX



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:

a) 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.

b) A função de avaliação deverá minimizar a distância euclidiana entre as cidades (os pontos).

c) Utilize no mínimo uma população com 100 indivíduos;

d) Utilize no mínimo 1% de novos indivíduos obtidos pelo operador de mutação;

e) Utilize no mínimo de 90% de novos indivíduos obtidos pelo método de cruzamento (crossover-ox);

f) 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

1) Algoritmo Genético

```
1 import math
2 import random
3 import matplotlib.pyplot as plt
4
5 # Definindo parâmetros do problema
```

```
6  NUM_CIDADES = 100
7  TAMANHO_POPULACAO = 100
8  NUM_GERACOES = 5000
9
10 # Função para calcular a distancia total de uma rota (caminho fechado
       que passa
11 def calcular_distancia_total(rota, coordenadas):
12     """
13     Calcula a distância total percorrida seguindo a sequência de
         cidades em 'rota'.
14     A distância éa soma das distâncias euclidianas entre cada par
         consecutivo.
15     *Importante: A rota começa e termina na cidade de origem (rota[0])
         .
16     """
17     distancia = 0.0
18     # Soma as distâncias entre cada cidade e a próxima na rota
19     for i in range(len(rota) - 1):
20         cidade_atual = rota[i]
21         proxima_cidade = rota[i+1]
22         # Obtém as coordenadas das duas cidades
23         x1, y1 = coordenadas[cidade_atual]
24         x2, y2 = coordenadas[proxima_cidade]
25         # Calcula a distancia Euclidiana entre cidade atual e proxima
             cidade
26         distancia += math.hypot(x2-x1, y2-y1)
27     return distancia
28
29 # Função de seleção por torneio para escolher um individuo (rota)
       como pai
30 def selecionar_pai_por_torneio(populacao, distancias, k=3):
31     """
32     Seleciona e retorna um individuo da população usando seleção por
         torneio.
33     Aleatoriamente escolhe 'k' indivíduos e retorna o que tiver a
         menor distância.
34     """
35     melhor_indice = -1
36     melhor_distancia = float('inf')
37     # Realiza o torneio entre individuos aleatórios
38     for i in range(k):
39         indice = random.randrange(len(populacao)) # escolhe um indice
             aleatorio
40         # Verifica se este individuo éo melhor do torneio até agora
41         if distancias[indice] < melhor_distancia:
42             melhor_distancia = distancias[indice]
43             melhor_indice = indice
44     # Retorna uma cópia da rota vencedora (melhor distância)
45     return list(populacao[melhor_indice])
46
47 def crossover_OX(pai1, pai2):
```

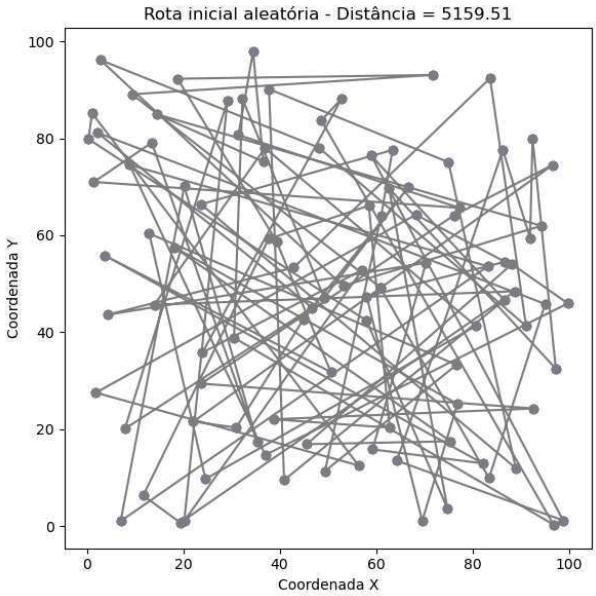
```
48      """
49      Realiza o crossover OX (Order Crossover) entre dois cromossomos (
            pais).
50      Retorna um cromossomo filho válido, com cidade inicial fixa.
51      """
52      tamanho = len(pai1)
53      # Inicializa o filho garantindo cidade inicial e final fixas
54      filho = [None] * tamanho
55      filho[0] = pai1[0]
56      filho[-1] = pai1[-1]
57
58      # Escolhe dois pontos de corte válidos e diferentes
59      corte1, corte2 = sorted(random.sample(range(1, tamanho - 1), 2))
60
61      # Copia segmento do pai1 diretamente para o filho
62      segmento_pai1 = pai1[corte1:corte2 + 1]
63      filho[corte1:corte2+1] = segmento_pai1
64
65      # Guarda as cidades ja inseridas no filho
66      cidades_no_filho = set(segmento_pai1)
67
68      # Preenche o restante das posições com cidades de pai2 (na ordem
            correta)
69      pos_filho = (corte2 + 1) % (tamanho - 1)
70      pos_pai2 = (corte2 + 1) % (tamanho - 1)
71
72      # Repete até o filho ser totalmente preenchido (sem None)
73      while None in filho[1:-1]:
74          cidade = pai2[pos_pai2]
75          if cidade not in cidades_no_filho:
76              filho[pos_filho] = cidade
77              cidades_no_filho.add(cidade)
78              pos_filho = pos_filho + 1 if pos_filho + 1 < tamanho - 1
                    else 1
79          pos_pai2 = pos_pai2 + 1 if pos_pai2 + 1 < tamanho - 1 else 1
80
81      return filho
82
83  # Função de mutação que troca duas cidades de lugar na rota (mutação
        por troca)
84  def mutacao(rota):
85      """
86      Realiza uma mutação simples em uma rota, trocando de posição duas
            cidades aleatórias.
87      A cidade de origem (posição 0 e última posição) não éalterada.
88      Retorna uma nova rota mutada.
89      """
90      # Cria uma cópia da rota para mutação (para não alterar a rota
            original diretamente)
91      rota_mutada = list(rota)
92
```

```
93      # Seleciona aleatoriamente duas posições diferentes na parte
            intermediaria
94      i = random.randint(1, len(rota_mutada) - 2)
95      j = random.randint(1, len(rota_mutada) - 2)
96
97      while i == j:
98          j = random.randint(1, len(rota_mutada) - 2)
99
100     # Troca as cidades das posições i e j
101     rota_mutada[i], rota_mutada[j] = rota_mutada[j], rota_mutada[i]
102     return rota_mutada
103
104 # Gera coordenadas aleatórias para cada cidade dentro de um espaço
        100x100
105 coordenadas = [] # Lista para armazenar tuples (x, y) de coordenadas
        de cada cidade
106 for i in range(NUM_CIDADES):
107     x = random.random() * 100 # coordenada x aleatoria entre 0 e 100
108     y = random.random() * 100 # coordenada y aleatoria entre 0 e 100
109     coordenadas.append((x, y))
110 print("Coordenadas:", coordenadas)
111
112 # Gera a população inicial de rotas aleatórias
113 populacao = []
114 distancias = []
115 cidade_origem = 0 # definindo a cidade de indice 0 como cidade de
        origem
116
117 # Inicializa as variaveis para acompanhar a melhor solução encontrada
118 melhor_rota = None
119 melhor_distancia = float('inf')
120
121 for i in range(TAMANHO_POPULACAO):
122     # Cria uma ordem aleatória das cidades (exceto a cidade de origem)
123     outras_cidades = list(range(1, NUM_CIDADES))
124     random.shuffle(outras_cidades)
125     # Monta a rota completa: cidade de origem no início, seguida das
            outras cidades
126     rota = [cidade_origem] + outras_cidades + [cidade_origem]
127     # Adiciona a rota gerada àpopulação
128     populacao.append(rota)
129     # Calcula a distância dessa rota e guarda
130     distancias.append(calcular_distancia_total(rota, coordenadas))
131
132 # Seleciona uma rota inicial aleatória (por exemplo, a primeira da
        população)
133 rota_inicial = populacao[0]
134 distancia_inicial = distancias[0]
135 print("Distancia inicial:", distancia_inicial)
136
```

```
137 # Identifica o melhor indivíduo da população inicial (com menor
        distancia)
138 for i, rota in enumerate(populacao):
139     d = distancias[i]
140     if d < melhor_distancia:
141         melhor_distancia = d
142         melhor_rota = rota
143 
144 # Loop principal do Algoritmo Genético
145 for geracao in range(NUM_GERACOES):
146     nova_populacao = []
147     nova_distancias = []
148 
149     # Elitismo: preserva a melhor rota da geração atual na próxima
150     nova_populacao.append(melhor_rota)
151     nova_distancias.append(melhor_distancia)
152 
153     # Determina quantos novos indivíduos serão gerados por crossover e
            por mutação
154     num_por_crossover = int(TAMANHO_POPULACAO * 0.90)
155     num_por_mutacao = int(TAMANHO_POPULACAO * 0.09)
156     if num_por_mutacao < 1:
157         num_por_mutacao = 1
158 
159     total_novos = 1 + num_por_crossover + num_por_mutacao # 1 elitismo
            + novos
160     if total_novos > TAMANHO_POPULACAO:
161         num_por_crossover -= (total_novos - TAMANHO_POPULACAO)
162     elif total_novos < TAMANHO_POPULACAO:
163         num_por_crossover += (TAMANHO_POPULACAO - total_novos)
164 
165     # Gera novos individuos por crossover
166     for _ in range(num_por_crossover):
167         pai1 = selecionar_pai_por_torneio(populacao, distancias, k=3)
168         pai2 = selecionar_pai_por_torneio(populacao, distancias, k=3)
169         filho = crossover_OX(pai1, pai2)
170         nova_populacao.append(filho)
171         nova_distancias.append(calcular_distancia_total(filho,
              coordenadas))
172 
173     # Gera novos individuos por mutação
174     for _ in range(num_por_mutacao):
175         base = selecionar_pai_por_torneio(populacao, distancias, k=3)
176         individuo_mutado = mutacao(base)
177         nova_populacao.append(individuo_mutado)
178         nova_distancias.append(calcular_distancia_total(
              individuo_mutado, coordenadas))
179 
180     # Se por algum motivo excedemos o tamanho da população, cortamos o
            excedente
181     if len(nova_populacao) > TAMANHO_POPULACAO:
```

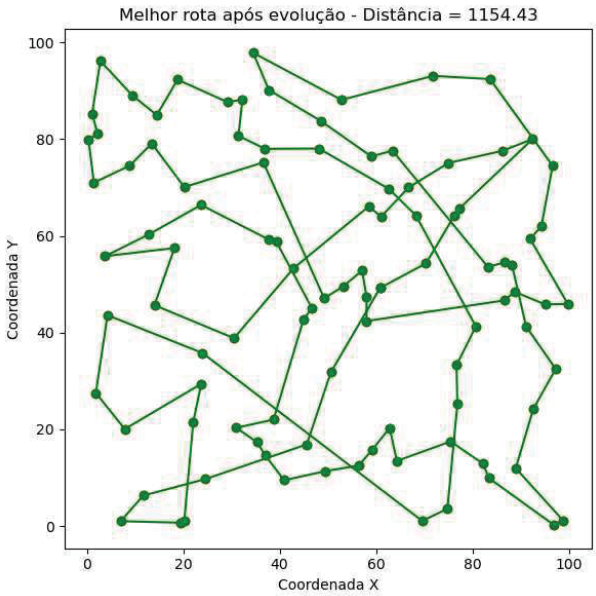
```
182         nova_populacao = nova_populacao[:TAMANHO_POPULACAO]
183         nova_distancias = nova_distancias[:TAMANHO_POPULACAO]
184
185     # Atualiza a população para a proxima geração
186     populacao = nova_populacao
187     distancias = nova_distancias
188
189     # Atualiza o melhor individuo
190     for i, d in enumerate(distancias):
191         if d < melhor_distancia:
192             melhor_distancia = d
193             melhor_rota = populacao[i]
194
195     # A cada 100 gerações, imprime a evolução
196     if (geracao + 1) % 100 == 0 or geracao == 0:
197         print(f"Geração {geracao + 1}: melhor distância encontrada {
            melhor_distancia}")
198
199 # Imprime as distâncias da rota inicial e da melhor rota final
    encontrada
200 print(f"\nDistancia total da rota inicial: {distancia_inicial:.2f}")
201 print(f"Distância total da melhor rota final: {melhor_distancia:.2f}\
    n")
202
203 # Gera o gráfico da rota inicial aleatória
204 plt.figure(figsize=(6,6))
205 xs = [coordenadas[cidade][0] for cidade in rota_inicial]
206 ys = [coordenadas[cidade][1] for cidade in rota_inicial]
207 plt.plot(xs, ys, marker='o', color='gray')
208 plt.scatter(xs, ys, color="blue")
209 plt.title(f"Rota inicial aleatória - Distancia = {distancia_inicial
    :.2f}")
210 plt.xlabel("Coordenada X")
211 plt.ylabel("Coordenada Y")
212 plt.tight_layout()
```

Figura 15.1 – EXERCÍCIO 1: ROTA INICIAL ALEATÓRIA, DISTÂNCIA = 5159,51




```
# Gera o gráfico da melhor rota final apos a evolução
plt.figure(figsize=(6,6))
xs_best = [coordenadas[cidade][0] for cidade in melhor_rota]
ys_best = [coordenadas[cidade][1] for cidade in melhor_rota]
plt.plot(xs_best, ys_best, marker='o', color='green')
plt.scatter(xs_best, ys_best, color='red')
plt.title(f"Melhor rota após evolução - Distância = {melhor_distancia
    :.2f}")
plt.xlabel("Coordenada X")
plt.ylabel("Coordenada Y")
plt.tight_layout()

# Mostra os gráficos gerados
plt.show()
```

Figura 15.2 – EXERCÍCIO 1: ROTA APÓS EVOLUÇÃO, DISTÂNCIA = 1154,43

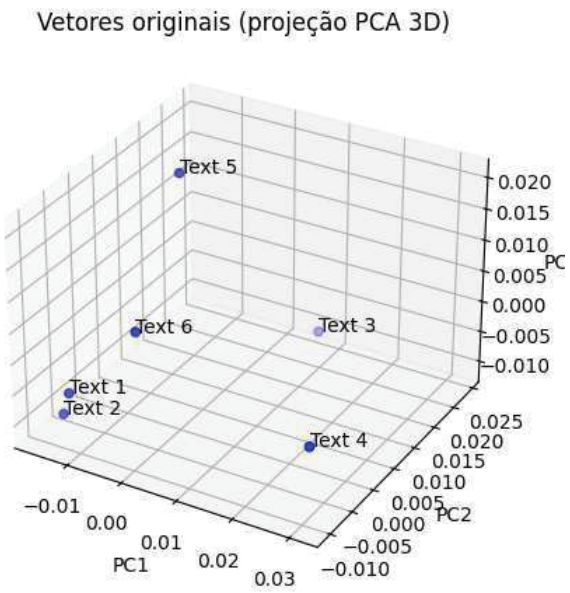


2) Compare a representação de dois modelos vetoriais

```
# Comandos para manipulação de pacotes
!pip uninstall numpy -y
!pip install numpy==1.25.2
!pip uninstall gensim -y
!pip install --no-binary :all: gensim

import numpy as np
import re
from gensim.models import Word2Vec
from sklearn.decomposition import PCA
from scipy.spatial.distance import pdist, squareform
from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt

# Textos curtos de exemplo
textos = [
    "O gato está dormindo no sofá.", # Texto 1
    "O cachorro está dormindo no tapete.", # Texto 2 (similar to Texto
        1)
    "A tecnologia esta mudando nossas vidas.", # Texto 3
    "Estamos vivendo na era da tecnologia.", # Texto 4 (similar to
       Texto 3)
    "O dia está ensolarado e claro.", # Texto 5
    "O sol está brilhando no céu azul." # Texto 6 (similar to Texto 5)
]

# Pre-processamento: impor pontuação e criar lista de tokens por
    texto
sentencas_tokens = []
for txt in textos:
    # Converter para minúsculas e remover pontuação
```

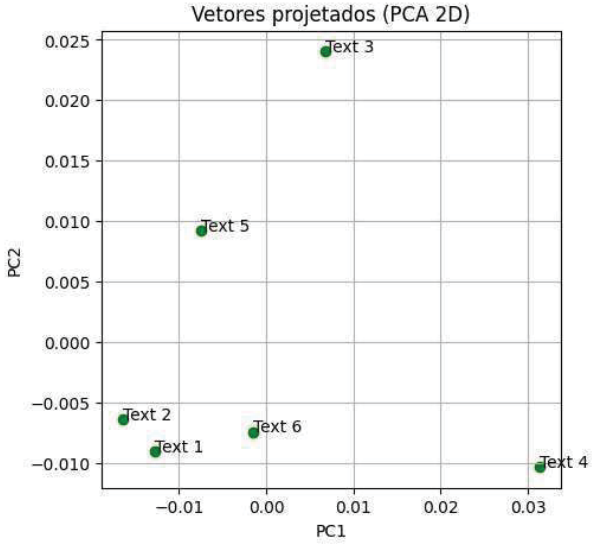
```
    texto_limpo = re.sub(r'[^\w\s]', '', txt.lower())
    tokens = texto_limpo.split()
    if tokens:
        sentencas_tokens.append(tokens)

# Treinar modelo Word2Vec nos tokens
model = Word2Vec(sentencas_tokens, vector_size=50, window=5,
    min_count=1, workers=4)

# Obter vetor de cada sentenca fazendo média dos vetores de palavras
vetores_texto = []
for tokens in sentencas_tokens:
    vetores_palavras = [model.wv[word] for word in tokens if word in
        model.wv.key_to_index]
    if len(vetores_palavras) > 0:
        # Média dos vetores de palavra para representar a sentença
        vetor_sentenca = np.mean(vetores_palavras, axis=0)
    else:
        # Caso sentença sem palavras conhecidas
        vetor_sentenca = np.zeros(model.vector_size)
    vetores_texto.append(vetor_sentenca)
vetores_texto = np.array(vetores_texto)

# Calcular vetor medio e centralizar os vetores de texto
vetor_medio = np.mean(vetores_texto, axis=0)
vetores_centralizados = vetores_texto - vetor_medio

# Aplicar PCA para reduzir dimensão dos vetores
pca_3d = PCA(n_components=3)
pca_2d = PCA(n_components=2)
vetores_3d = pca_3d.fit_transform(vetores_centralizados)
vetores_2d = pca_2d.fit_transform(vetores_centralizados)

# Visualização dos vetores em 3D e 2D
labels = [f"Text {i+1}" for i in range(len(textos))]

# Plot 3D dos vetores originais (após PCA 3D)
fig = plt.figure(figsize=(6,5))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(vetores_3d[:,0], vetores_3d[:,1], vetores_3d[:,2], c='blue
    ', marker='o')
for i, label in enumerate(labels):
    ax.text(vetores_3d[i,0], vetores_3d[i,1], vetores_3d[i,2], label)
ax.set_title('Vetores originais (projeção PCA 3D)')
ax.set_xlabel('PC1')
ax.set_ylabel('PC2')
ax.set_zlabel('PC3')
plt.show()
```

Figura 15.3 – EXERCÍCIO 2: PROJEÇÃO PCA 3D




```
# Plot 2D dos vetores projetados (PCA 2D)
plt.figure(figsize=(5,5))
plt.scatter(vetores_2d[:,0], vetores_2d[:,1], c='green', marker='o')
for i, label in enumerate(labels):
    plt.annotate(label, (vetores_2d[i,0], vetores_2d[i,1]))
plt.title('Vetores projetados (PCA 2D)')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.grid(True)
plt.show()
```

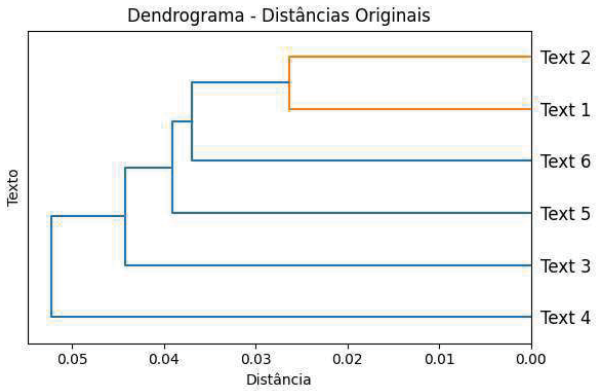
Figura 15.4 – EXERCÍCIO 2: PROJEÇÃO PCA 2D




```
# Calcular matrizes de distancia entre vetores
```

```
dist_original = squareform(pdist(vetores_centralizados, metric='
    euclidean'))
dist_projetado = squareform(pdist(vetores_2d, metric='euclidean'))

# Gerar dendrogramas de agrupamento hierárquico
# Dendrograma para distâncias originais
Z_orig = linkage(vetores_centralizados, method='ward', metric='
    euclidean')
plt.figure(figsize=(6,4))
dendrogram(Z_orig, labels=labels, orientation='left')
plt.title('Dendrograma - Distâncias Originais')
plt.xlabel('Distância')
plt.ylabel('Texto')
plt.tight_layout()
plt.show()
```

Figura 15.5 – EXERCÍCIO 2: DENDOGRAMA - DISTÂNCIAS ORIGINAIS




```
# Dendrograma para distâncias após projeção PCA 2D
Z_proj = linkage(vetores_2d, method='ward', metric='euclidean')
plt.figure(figsize=(6,4))
dendrogram(Z_proj, labels=labels, orientation='left')
plt.title('Dendrograma - Distâncias PCA 2D')
plt.xlabel('Distância')
plt.ylabel('Texto')
plt.tight_layout()
plt.show()
```

Figura 15.6 – EXERCÍCIO 2: DENDOGRAMA - DISTÂNCIAS PCA 2D