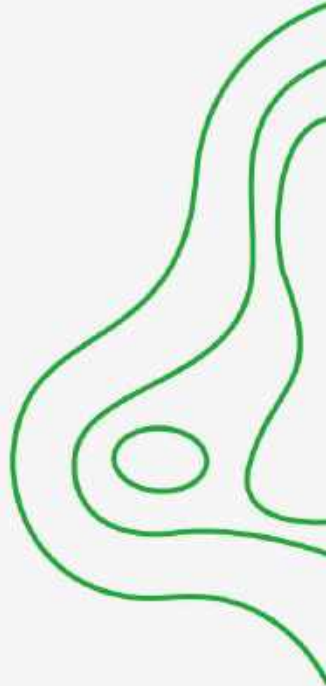


Inteligência Artificial



Autor

Álvaro Farias Pinheiro



GOVERNO DE PERNAMBUCO
SECRETARIA DE ADMINISTRAÇÃO



- P654i Pinheiro, Álvaro Farias
 Inteligência artificial / Álvaro Farias Pinheiro. – Recife : Escola de
 Governo de Administração Pública de Pernambuco, 2024.
 95p. : il.
- Inclui referências.
 Inclui currículo do autor.
 Material produzido pela Escola de Governo de Administração Pública
 de Pernambuco – EGAPE.
1. INTELIGÊNCIA ARTIFICIAL – MÉTODOS – ANÁLISE –
 ESTUDO E ENSINO. I. Título.

CDU 007.52
CDD 006.3

EXPEDIENTE

Governadora de Pernambuco

Raquel Teixeira Lyra Lucena

Vice-governadora de Pernambuco

Priscila Krause Branco

SECRETARIA DE EDUCAÇÃO E ESPORTES

Secretário

Alexandre Alves Schneider

EGAPE**SECRETARIA DE ADMINISTRAÇÃO**

Secretária

Ana Maraíza Sousa Silva

Diretor

Henrique César Freire de Oliveira

Secretária Executiva de Gestão de Pessoas

Luciana Oliveira Pires

Gerência de Educação Profissional da Escola de Governo

Marilene Cordeiro Barbosa Borges

Núcleo de Educação a Distância - NUEAD

Deisiane Gomes Bazante

Revisão de Língua Portuguesa

Paulo Euzébio Bispo

Núcleo de Educação Presencial - NUEDP

Maria Elisete Oliveira

Diagramação

Deisiane Gomes Bazante

Autor

Álvaro Farias Pinheiro

Material produzido pela Escola de Governo de Administração Pública de Pernambuco – EGAPE
outubro, 2024 (1. ed.)

SUMÁRIO

INTRODUÇÃO.....	8
COMPUTAÇÃO CLÁSSICA OU TRADICIONAL.....	8
MÉTODOS DEDUTIVO, INDUTIVO E ABDUTIVO: GENERALIZAÇÃO	8
MÉTODO DEDUTIVO	9
MÉTODO INDUTIVO	9
MÉTODO ABDUTIVO	10
FORMALISMOS OU TIPOS DE APRENDIZADO DE MÁQUINA.....	11
FORMALISMO: UNSUPERVISED.....	12
FORMALISMO: SUPERVISED.....	12
FORMALISMO: SEMI-SUPERVISED	12
FORMALISMO: REINFORCEMENT	13
ALGUMAS APLICABILIDADES DOS FORMALISMOS	13
GENERALIZAÇÃO COGNITIVA	14
ABORDAGENS	15
ABORDAGEM: ESTATÍSTICA	16
ABORDAGEM: ESTATÍSTICA (EXEMPLOS DE ALGORITMOS).....	16
ABORDAGEM: EVOLUCIONÁRIA.....	16
ABORDAGEM: EVOLUCIONÁRIA (EXEMPLOS DE ALGORITMOS).....	17
ABORDAGEM: ENXAMES	17
ABORDAGEM: ENXAMES (EXEMPLOS DE ALGORITMOS).....	18
ABORDAGEM: NEURAL	18
ABORDAGEM: NEURAL (EXEMPLOS DE ALGORITMOS)	18
MÉTRICAS	19
INDICADORES UTILIZADOS EM ALGUMAS MÉTRICAS.....	19
PARA AVALIAR O APRENDIZADO: MÉTRICAS	19
MÉTRICAS COMUNS EM APRENDIZADO DE MÁQUINA	20
MÉTRICAS PARA DADOS CATEGÓRICOS	20
MÉTRICAS PARA DADOS REGRESSIVOS.....	21
MÉTODOS DE AMOSTRAGEM: TEST & SCORE	22
CLASSES DE PROBLEMAS.....	24
CLASSES DE PROBLEMAS EM DIFERENTES PARADIGMAS DE APRENDIZADO	24

APRENDIZADO NÃO SUPERVISIONADO	25
APRENDIZADO SUPERVISIONADO	25
APRENDIZADO POR REFORÇO	26
RECOMENDAÇÃO E RECONHECIMENTO	26
REDES NEURAIS GENERATIVAS	26
CLASSE DE PROBLEMA: MONOMODAL E MULTIMODAL	27
CLASSE DE PROBLEMA: MONO-OBJETIVO E MULTIOBJETIVO	27
EXTRAÇÃO, TRANSFORMAÇÃO E CARGA OU EXTRAÇÃO, TRANSFORMAÇÃO E CARGA (ETL)	28
CARACTERÍSTICAS DOS DADOS	32
DATA VISUALIZATION EXPLORATORY DATA ANALYSIS (EDA)	42
EDA: SCATTER PLOT	43
EDA: VARIÁVEIS DEPENDENTES E INDEPENDENTES	44
EDA: TRATAR VARIÁVEIS COMO INDEPENDENTES	44
EDA: LINHA DE REGRESSÃO	44
EDA: CORRELAÇÃO	45
EDA: DISTRIBUTION PLOT	45
EDA: BOX PLOT	46
EDA: HEAT MAP	46
EDA: SILHOUETT PLOT	47
EDA: LINEAR PROJECTION	48
EDA: LINE PLOT	49
EDA: DENDROGRAM	50
EDA: VENN DIAGRAM	51
EDA: LINE CHART	52
AS VÁRIAS ONDAS DA INTELIGÊNCIA ARTIFICIAL	53
A 4ª ONDA DA INTELIGÊNCIA ARTIFICIAL	53
O QUE FEZ AS PESSOAS ACORDAREM PARA O CHATGPT E A IA?	54
INTELIGÊNCIA ARTIFICIAL NÃO É MÁGICA	54
IA É UM PROGRAMA DE COMPUTADOR COMO QUALQUER OUTRO	55
IA É UM PROGRAMA DE COMPUTADOR COMO QUALQUER OUTRO	55
IA É UMA TECNOLOGIA TRANSFORMADORA	55
OS PRIMEIROS PASSOS PARA O SURGIMENTO DAS REDES NEURAIS	56
A ORIGEM DO TERMO INTELIGÊNCIA ARTIFICIAL (IA)	56
O PRECURSOR DAS REDES NEURAIS	57

A ORIGEM DO TERMO MACHINE LEARNING (ML)	57
A ORIGEM DA NEURAL NETWORK (NN)	58
ALGORITMO TRANSFORM O LAMPEJO DE CRIATIVIDADE	59
ALGORITMO TRANSFORM E O CONCEITO “SELF-ATTENTION”	59
IA NÃO É UM SER SENCIENTE, É MATEMÁTICA!	60
IA GENERATIVA, É UMA FUNÇÃO PROBABILÍSTICA!	60
GEN-AIS REPLICAM FUNÇÕES COGNITIVAS ALGORÍTMICAS	60
REDES NEURAIS ARTIFICIAIS (ROBERT HECHT-NIELSEN, 1987)	61
INSPIRAÇÃO BIOLÓGICA	61
CÉREBRO BIOLÓGICO	62
REDE NEURAL ARTIFICIAL INSPIRADA NA REDE NEURAL BIOLÓGICA	62
REDE NEURAL ARTIFICIAL INSPIRADA NA REDE NEURAL BIOLÓGICA	63
DEFINIÇÕES POSSÍVEIS DE UMA REDE NEURAL ARTIFICIAL	63
PERCEPTRON MODELO MAIS SIMPLES DE REDE NEURAL ARTIFICIAL	64
DO PERCEPTRON PARA AS NEURAL NETWORKS	64
O OBJETIVO DAS REDES NEURAIS É MINIMIZAR O ERRO	66
GRADIENTE DESCENDENTE FORNECE A DIREÇÃO DO ERRO MÍNIMO	67
AS GENAIS SÃO A EVOLUÇÃO DAS DEEP LEARNING	67
O QUE SÃO IAS GENERATIVAS (GEN-AIS)?	68
O QUE SÃO MULTIAGENTES INTELIGENTES NAS GEN-AIS?	68
FERRAMENTAS PARA USAR MULTIAGENTES	68
A TRANSFORMAÇÃO DA ECONOMIA GLOBAL	69
O IMPACTO DA IA NA EMPREGABILIDADE (FMI, DEZ-2023)	69
PARCELA DE EMPREGOS POR DECIS DE RENDA (FMI, DEZ-2023)	70
IA E SEUS SUBCAMPOS	71
DS & AI COM SEUS SUBCAMPOS	71
FAMÍLIAS DOS MODELOS GENERATIVOS MAIS CONHECIDAS	73
INTELIGÊNCIA ARTIFICIAL (IA)	73
CATEGORIAS DE IA	74
FUNDAMENTOS DA IA (RUSSELL & NORVIG, 2013)	75
NATURAL LANGUAGE PROCESSING	77
DISCIPLINAS DE IA	77
INTELIGÊNCIA ARTIFICIAL / INTELIGÊNCIA COMPUTACIONAL	78
ALGORITMOS SIMBÓLICOS	78

ALGORITMOS SUBSIMBÓLICOS	78
CONCEITOS BASILARES	79
ARTIFICIAL NEURAL NETWORK (ANN)	84
COMO UMA REDE NEURAL ARTIFICIAL APRENDE	85
LÓGICA DE UMA REDE NEURAL ARTIFICIAL	85
PERCEPTRON	90
MULTILAYER PERCEPTRON (MLP)	91
CONVOLUTIONAL NEURAL NETWORK (CNN)	92
REFERÊNCIAS	94
SOBRE O AUTOR	96

INTRODUÇÃO

Olá, cursista!

Este documento tem como propósito explorar e esclarecer os métodos dedutivo, indutivo e abdutivo no contexto da inteligência artificial e do aprendizado de máquina. Através da análise desses métodos, busca-se compreender como cada um deles contribui para a generalização dos modelos e a minimização da perda em amostras não vistas. Além disso, o documento pretende ilustrar a íntima relação entre o aprendizado de máquina e a otimização, destacando como os problemas de aprendizado são formulados para minimizar funções de perda e melhorar a acurácia das previsões em dados não previamente conhecidos. A introdução desses conceitos é essencial para profissionais e estudantes da área, visando uma compreensão aprofundada das técnicas e aplicações práticas no desenvolvimento de sistemas inteligentes com uso de Inteligência Artificial.

COMPUTAÇÃO CLÁSSICA OU TRADICIONAL

Na computação clássica ou tradicional, o método dedutivo é amplamente utilizado. Este método baseia-se na aplicação de regras e lógicas predefinidas para resolver problemas específicos. Ao partir de premissas gerais para chegar a conclusões específicas, a computação clássica segue um caminho lógico e estruturado, característico do método dedutivo.

Assim, a computação tradicional emprega o método dedutivo ao aplicar regras e lógicas estabelecidas a dados concretos, resolvendo problemas de maneira previsível e consistente.

MÉTODOS DEDUTIVO, INDUTIVO E ABDUTIVO: GENERALIZAÇÃO

O aprendizado de máquina é estreitamente ligado à otimização, pois muitos problemas são formulados como minimização de funções de perda em exemplos de treinamento. Funções de perda medem a diferença entre as previsões do modelo e as instâncias reais. A otimização foca na redução da perda no conjunto de treinamento, enquanto o aprendizado de máquina busca minimizar a perda em novas amostras.

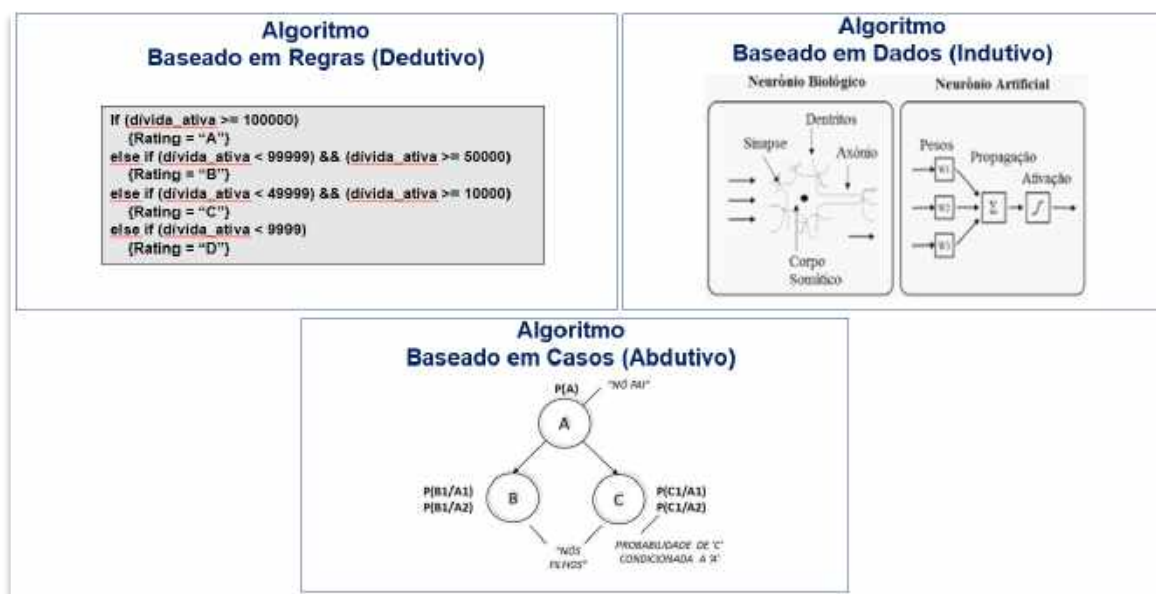


Figura 1: Métodos de Aprendizado
Fonte: Álvaro Pinheiro.

MÉTODO DEDUTIVO

Na inteligência artificial, quando temos a equação $y = f(x)$, onde x são os dados, f é a regra e y são as saídas, e a regra já é conhecida, utilizamos o método dedutivo. Este método parte de uma regra geral (f) e aplica-a a dados específicos (x) para obter uma conclusão particular (y).

Por exemplo, se sabemos que todos os mamíferos têm um coração (regra geral), e que todos os cães são mamíferos (dados específicos), podemos concluir que todos os cães têm um coração (conclusão específica). Assim, na inteligência artificial, ao conhecermos a regra e aplicá-la aos dados para obter as saídas desejadas, estamos empregando o método dedutivo.

MÉTODO INDUTIVO

Quando na inteligência artificial faz uso da equação $y = f(x)$, onde x representa os dados, f a regra e y as saídas, e a regra ainda não é conhecida, mas os dados e as saídas desejadas são, emprega-se o método indutivo. Este método inicia a partir de observações específicas (dados x) para derivar uma conclusão geral (regra f). Em outras palavras, quando a

regra não é conhecida e utilizam-se os dados para inferir essa regra, aplicando-a para obter as saídas desejadas, está-se utilizando o método indutivo.

Observações Específicas (x): Vários exemplos de entrada e saída.

Conclusão Geral (f): A regra ou função que mapeia as entradas para as saídas.

Portanto, na inteligência artificial, quando não se conhece a regra e se faz uso de dados para inferir essa regra e obter as saídas desejadas, está-se utilizando o método indutivo.

MÉTODO ABDUTIVO

Na inteligência artificial, quando trabalhamos com a equação $y = f(x)$, onde x são os dados, f é a regra e y são as saídas, e empregamos o método abdutivo, tentamos encontrar a melhor explicação para um conjunto de observações ou dados. Esse método inicia-se com uma observação específica ou um conjunto de dados (x) e busca determinar a regra mais provável (f) que possa explicar essas observações e prever as saídas (y).

O método abdutivo é amplamente utilizado para formular hipóteses que expliquem fenômenos observados, mesmo que as evidências sejam incompletas ou ambíguas. Por exemplo, se a porta de sua casa está arrombada, pode-se inferir que houve uma tentativa de invasão. Na inteligência artificial, o método abdutivo é empregado para inferir a regra ou modelo que melhor explica os dados observados e, assim, prever as saídas. É um processo de inferência que busca a melhor explicação possível com base nas evidências disponíveis.

$$y=f(x)$$

$$\text{resultado}=\text{regra}(\text{dado})$$

Método Aprendizado	Dados (x)	Regra (f)	Resultado (y)
Dedutivo	Possui	Conhece	?
Indutivo	Possui	?	Conhece
Abdutivo	?	Conhece	Possui

Figura 2: Métodos de Aprendizado

Fonte: Pattern Recognition and Machine Learning de Christopher M. Bishop.

FORMALISMOS OU TIPOS DE APRENDIZADO DE MÁQUINA

O campo do aprendizado de máquina é vasto e diverso, englobando várias abordagens que permitem às máquinas aprenderem e evoluir a partir de dados. Estas abordagens podem ser classificadas em diferentes formalismos, cada um com suas características, métodos e aplicações específicas. A seguir, vamos explorar os principais tipos de aprendizado de máquina: supervisionado, não supervisionado e por reforço, detalhando como cada um deles funciona, suas vantagens e desafios, bem como suas aplicabilidades em diversos contextos.



Figura 3: Formalismos

Fonte: Pattern Recognition and Machine Learning de Christopher M. Bishop.

FORMALISMO: UNSUPERVISED

O aprendizado não supervisionado ocorre quando o algoritmo tenta entender padrões por conta própria, já que os dados não possuem rótulos. Ele é usado para analisar e agrupar dados não rotulados, identificando padrões ocultos ou agrupamentos sem intervenção humana. Técnicas incluem armazenamento em cluster, que agrupa dados similares; associação, que identifica relações entre itens; e redução de dimensionalidade, que reduz variáveis mantendo informações essenciais.

FORMALISMO: SUPERVISED

O aprendizado de máquina supervisionado é uma área que emprega algoritmos de machine learning para examinar conjuntos de dados rotulados. Esses algoritmos são treinados usando um conjunto de dados rotulados, onde os resultados esperados já são conhecidos. A meta é desenvolver um modelo que consiga generalizar esses resultados para novos dados não vistos anteriormente. A escolha do modelo mais adequado é baseada no problema e nos dados disponíveis. Classificação, técnica que categoriza novos dados com base no aprendizado realizado a partir dos dados rotulados do conjunto de treinamento e teste. Predição, que prevê o comportamento dos dados com base no aprendizado obtido do conjunto de dados histórico (legado).

FORMALISMO: SEMI-SUPERVISED

O aprendizado de máquina semisupervisionada situa-se em um meio-termo entre o aprendizado não supervisionado e o supervisionado. Esta abordagem é utilizada quando se possui um conjunto de dados parcialmente rotulado, ou seja, uma parte dos dados possui rótulos, enquanto a outra parte não possui. A ideia é aproveitar os dados rotulados para guiar e melhorar o desempenho do modelo ao analisar os dados não rotulados.

Esta técnica é particularmente útil em situações em que obter rótulos é custoso, demorado ou impraticável. Combinando informações de ambos os conjuntos de dados, o

modelo semisupervisionado pode alcançar uma performance superior àquela obtida usando apenas dados rotulados ou não rotulados isoladamente.

Algoritmos semi-supervisionados podem incluir técnicas como auto-encoder, que aprendem representações compactas dos dados; métodos de propagação de rótulos, onde os rótulos são propagados dos dados rotulados para aqueles não rotulados com base nas similaridades entre eles; e modelos de *co-training*, que treinam dois ou mais classificadores sobre diferentes partes dos dados e utilizam as previsões para rotular novos dados.

As vantagens do aprendizado semisupervisionado incluem a capacidade de melhorar a precisão do modelo e reduzir a necessidade de grandes quantidades de dados rotulados. No entanto, os desafios incluem a complexidade na implementação dos algoritmos e a necessidade de garantir que os rótulos propagados não introduzam ruído ou vieses no modelo.

FORMALISMO: REINFORCEMENT

O Aprendizado por Reforço é uma técnica de IA que permite às máquinas aprenderem a tomar decisões sozinhas, sem programação explícita. O agente aprende a atingir metas em ambientes incertos, tomando uma sequência de decisões usando tentativa e erro para chegar a soluções. A IA recebe recompensas ou penalidades com o objetivo de maximizar a recompensa total. O cientista de dados define a política de recompensa, mas não indica ao modelo como resolver a tarefa. Com feedback não instantâneo, o modelo começa com tentativas aleatórias e evolui até táticas sofisticadas. Esse modelo é usado em robótica, controle de processos, jogos e sistemas autônomos.

ALGUMAS APLICABILIDADES DOS FORMALISMOS

A inteligência artificial (IA) tem transformado significativamente diversos setores, proporcionando soluções inovadoras para problemas complexos. Entre as várias técnicas utilizadas, destacam-se os formalismos que permitem à IA aprender e se adaptar conforme interage com o ambiente. Esses métodos têm mostrado grande eficácia em áreas como

robótica, controle de processos e sistemas autônomos, demonstrando um potencial imenso para o futuro.



Figura 4: Algumas Aplicabilidades dos Formalismos

Fonte: Reinforcement Learning: An Introduction de Richard S. Sutton e Andrew G. Barto.

GENERALIZAÇÃO COGNITIVA

O formalismo do Aprendizado por Reforço é apenas uma das muitas abordagens que demonstram o poder das técnicas de IA em ensinar máquinas a tomar decisões complexas. No entanto, a IA não se restringe a este único método; há uma vasta gama de abordagens que exploram diferentes princípios para a análise de dados e a tomada de decisões.

É a aplicação do princípio ou conceito a um conjunto de casos, isto é, simplificação, onde são abstraídos detalhes particulares ou exceções, atribuindo-se a um grupo de coisas que pertencem ao mesmo gênero algo que já se sabe sobre alguns de seus indivíduos.

$$f(x) = y$$

Mecanismo de inferência	x	f	y
Dedução	x	x	?
Indução	x	?	x
Abdução	?	x	x

Figura 5: Mecanismos de Inferência

Fonte: Reinforcement Learning: An Introduction de Richard S. Sutton e Andrew G. Barto

ABORDAGENS

A crescente sofisticação e aplicabilidade das técnicas de inteligência artificial (IA) transformaram a maneira como abordamos problemas complexos. A figura abaixo ilustra diversos mecanismos de inferência, demonstrando como diferentes abordagens podem ser utilizadas para ensinar máquinas a tomar decisões. Cada uma dessas técnicas se baseia em princípios específicos para analisar dados e gerar soluções, evidenciando a amplitude e a versatilidade da IA na resolução de desafios multifacetados.

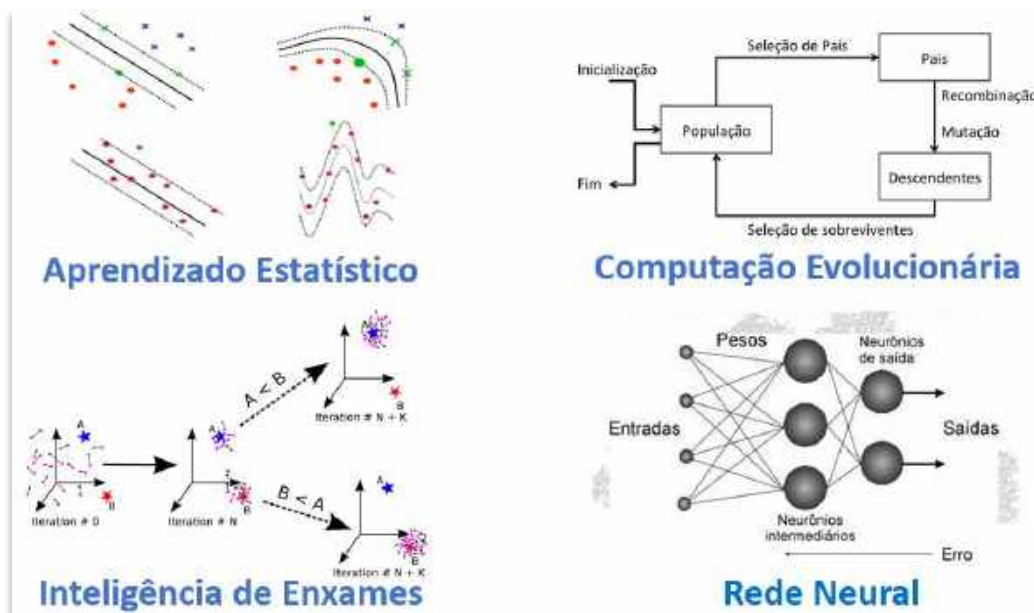


Figura 6: Abordagens da Inteligência Artificial

Fonte: Álvaro Pinheiro

ABORDAGEM: ESTATÍSTICA

Algoritmos que recorrem a técnicas estatísticas para examinar dados, discernir padrões e tomar decisões baseadas em probabilidades se fundamentam na extração de informações valiosas dos dados.

ABORDAGEM: ESTATÍSTICA (EXEMPLOS DE ALGORITMOS)

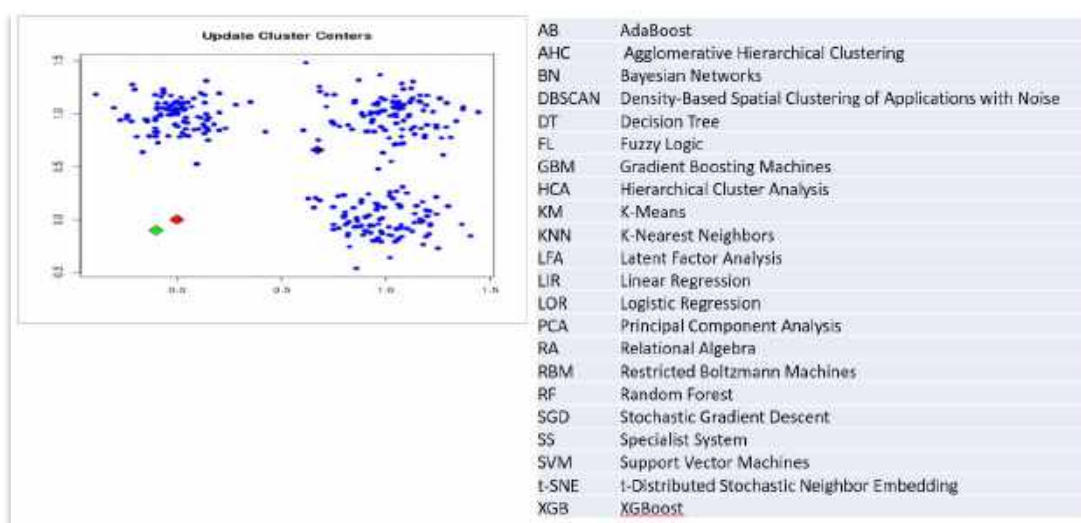


Figura 7: Abordagem Estatística

Fonte: Álvaro Pinheiro

ABORDAGEM: EVOLUCIONÁRIA

Algoritmos que aplicam uma abordagem computacional baseada nos princípios da evolução biológica buscam resolver problemas de otimização, imitando a seleção natural, a recombinação genética e mutações para identificar soluções eficazes dentro de um espaço de busca.

ABORDAGEM: EVOLUCIONÁRIA (EXEMPLOS DE ALGORITMOS)

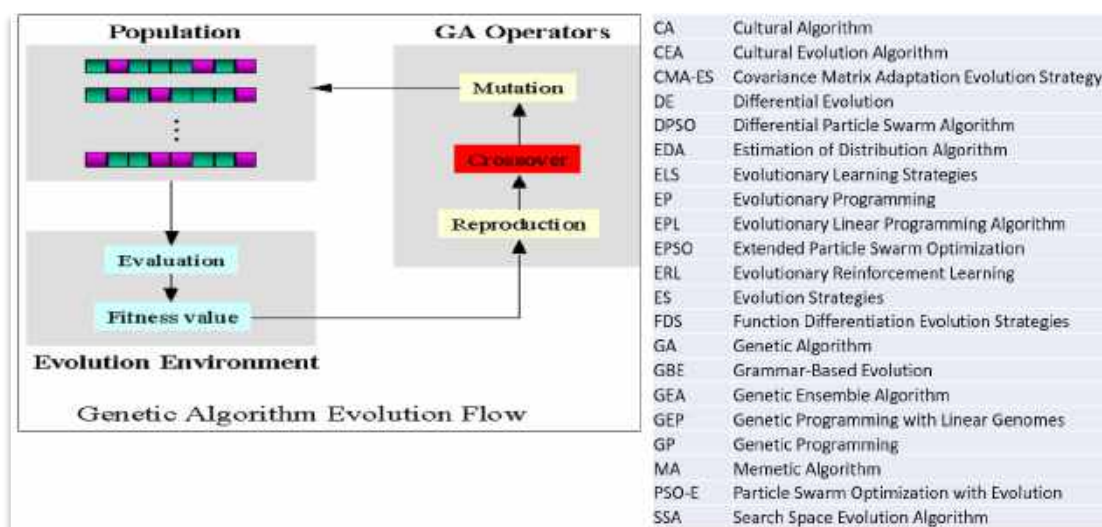


Figura 8: Abordagem Evolucionária
Fonte: Álvaro Pinheiro

ABORDAGEM: ENXAMES

Algoritmos que imitam o comportamento de grupos sociais, como bandos de pássaros ou colônias de insetos, são usados na otimização baseada em enxames. Nessas técnicas, agentes simples chamados "partículas" trabalham juntos para encontrar soluções eficazes.

ABORDAGEM: ENXAMES (EXEMPLOS DE ALGORITMOS)

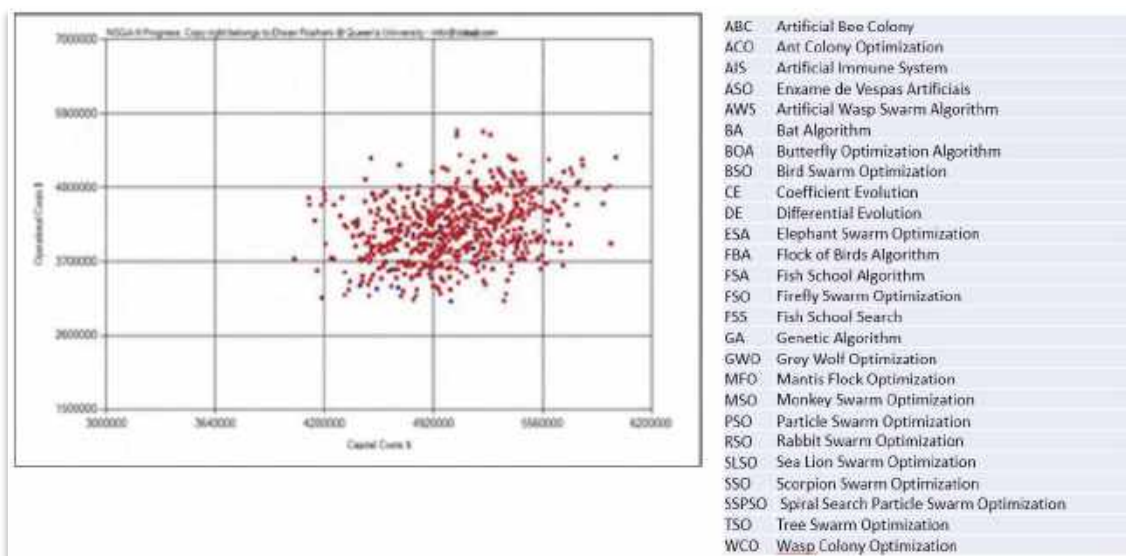


Figura 9: Abordagem de Enxames

Fonte: Álvaro Pinheiro

ABORDAGEM: NEURAL

Algoritmos que realizam tarefas de aprendizado e tomada de decisão, sendo modelos matemáticos inspirados no funcionamento do cérebro humano, compostos por unidades de processamento chamadas de neurônios artificiais.

ABORDAGEM: NEURAL (EXEMPLOS DE ALGORITMOS)

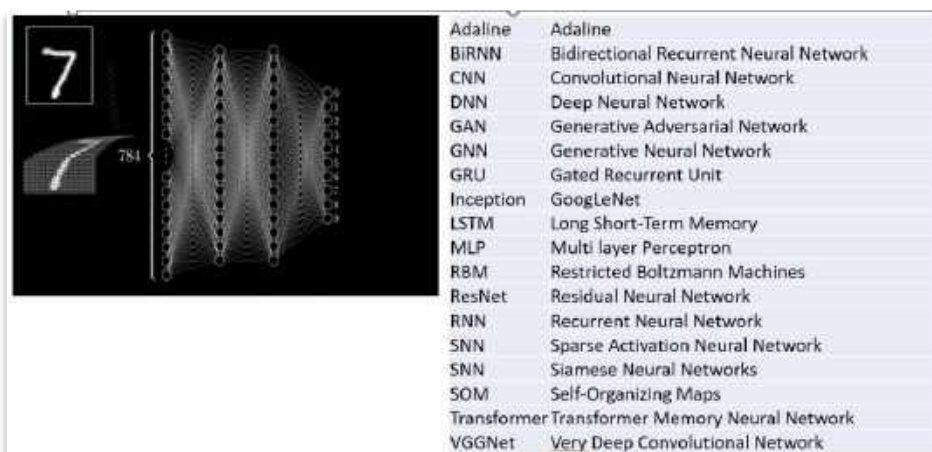


Figura 10: Abordagem Neural

Fonte: Álvaro Pinheiro

MÉTRICAS

Na ciência de dados, métricas são medidas quantitativas que avaliam o desempenho de modelos, processos ou estratégias. Elas podem ser valores numéricos ou não e são essenciais para medir o desempenho com precisão. Como cientista de dados, você encontrará diversos tipos de métricas de distância: na PNL, a métrica de distância do cosseno é usada para encontrar palavras semelhantes; na Visão por Computador, a métrica de distância L2 identifica imagens semelhantes. Para ser considerada uma métrica, deve satisfazer: (1) não negatividade ($d(x, y) \geq 0$) e (2) identidade dos indiscerníveis (se $d(x, y) = 0$, então $x = y$). Escolher a métrica correta é crucial para resultados precisos. No aprendizado de máquina, métricas específicas medem o desempenho de modelos conforme o problema, como em classificações binárias desequilibradas, onde a acurácia pode ser substituída pela área sob a curva ROC para melhores insights.

INDICADORES UTILIZADOS EM ALGUMAS MÉTRICAS

VALOR REAL	VALOR PREDITO	
	SIM	NÃO
SIM	VERDADEIRO POSITIVO	FALSO NEGATIVO
NÃO	FALSO POSITIVO	VERDADEIRO NEGATIVO

Verdadeiro Positivo (VP) é a análise de presença correta
Falso Positivo (FP) é a análise de presença errada
Verdadeiro Negativo (VN) é a análise de ausência correta
Falso Negativo (FN) é a análise de ausência errada

Figura 11: Métrica
Fonte: Álvaro Pinheiro

PARA AVALIAR O APRENDIZADO: MÉTRICAS

Métricas são importantes em diferentes áreas da ciência de dados e aprendizado de máquina para avaliar e otimizar o desempenho dos modelos. Dentre os diversos tipos de métricas, podemos destacar:

MÉTRICAS COMUNS EM APRENDIZADO DE MÁQUINA

No aprendizado de máquina, várias métricas são utilizadas para medir a eficácia de modelos preditivos. Elas variam dependendo do tipo de problema e dos objetivos específicos. As métricas mais comuns incluem acurácia, precisão, revocação, F1-score, e muitas outras que fornecem insights detalhados sobre o desempenho do modelo. A escolha da métrica adequada é vital para garantir uma avaliação justa e precisa dos modelos.

MÉTRICAS PARA DADOS CATEGÓRICOS

As métricas de classificação são particularmente importantes na avaliação de modelos que atribuem categorias a dados. Em problemas de classificação binária e multiclasse, essas métricas ajudam a entender a capacidade do modelo de distinguir entre diferentes classes e a eficácia geral das previsões.

Acurácia (A), razão entre VP e VN para a soma de VP, VN, FP e FN, expressa por, $(VP+VN) / (VP+VN+FP+FN)$.

Precisão (P), razão entre VP para a soma do número de VP e FP, expressa por, $(VP) / (VP + FP)$.

Revocação (R), razão entre VP para a soma do número de VP e FN, expressa por, $(VP) / (VP + FN)$.

Mensuração (F1), dobro da razão entre a multiplicação da precisão pelo Revocação e da soma da precisão com o Revocação, expressa por, $F = 2 * ((Precisão * Revocação) / (Precisão + Revocação))$.

Coeficiente de Correlação de Matthews (MCC), métrica de valor único que resume a matriz de confusão, expressa por, $(VN*VP - FP*FN) / \text{SQRT}((VN+FN) * (FP+VP) * (VN+FP) * (FN+VP))$.

Especificidade (Spec), prever um verdadeiro negativo de cada categoria disponível, expressa por, $VN / (VN + FP)$.

MÉTRICAS PARA DADOS REGRESSIVOS

Perda de Log (LogLoss), mede as previsões de probabilidades de adesão a uma determinada classe, expressa por,

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$$

Área Sob a Curva (AUC) ou Curva Receiver Operating Characteristic (Curva ROC), razão entre a sensibilidade ou taxa de verdadeiros positivos (RPV) pela especificidade ou taxa de verdadeiros negativos (RPF), expressa por $(VP / \text{Positivos Totais}) / (FP / \text{Negativos Totais})$.

Erro Médio Quadrático (MSE), igual à soma da variância e da tendência do quadrado do estimador, expressa por,

$$\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2$$

Raiz do Erro Médio Quadrático (RMSE), é a raiz quadrada da média das previsões subtraindo os valores observados ao quadrado, expressa por,

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - p_i)^2}$$

Erro Médio Absoluto (MAE), mede a média da diferença absoluta entre os valores previstos pelo modelo e os valores observados, expressa por,

$$\frac{1}{n} \sum_{i=1}^n |y_i - p_i|$$

Coefficiente da Determinação (R2), medida para verificar quão próximos os dados estão da linha de regressão ajustada com base na porcentagem da variação como resposta do modelo linear, expressa por,

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

Coeficiente da Variação do RMSE (CVRMSE), medida que calcula o quão próximo os pontos de dados reais estão dos valores previstos pelo modelo e é usado para medir o desvio padrão dos resíduos, expressa por,

$$\frac{1}{\bar{Y}} \sqrt{\frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N}}$$

MÉTODOS DE AMOSTRAGEM: TEST & SCORE

Todas essas métricas são vitais para a avaliação precisa de modelos preditivos, ajudando a garantir que o modelo escolhido não apenas se ajusta bem aos dados de treino, mas também generaliza bem para novos dados. Esses métodos de amostragem e métricas formam a espinha dorsal na validação e verificação dos algoritmos, permitindo a identificação e correção de quaisquer problemas de ajuste ou sobreajuste nos dados.

As técnicas mencionadas, como o Erro Médio Absoluto (MAE), Coeficiente da Determinação (R²), Coeficiente da Variação do RMSE (CVRMSE), Erro Médio Quadrático (MSE), Raiz do Erro Médio Quadrático (RMSE), Área Sob a Curva (AUC), Perda de Log (LogLoss), Especificidade (Spec), Coeficiente de Correlação de Matthews (MCC) e Mensuração (F1), são fundamentais na análise e na validação das performances dos modelos preditivos. Estas métricas são frequentemente utilizadas em diversos métodos de amostragem, como o Test & Score, para garantir que os modelos estão devidamente ajustados e prontos para generalização.

Para aplicar essas métricas de forma eficaz, utilizamos diferentes métodos de testagem. Um dos métodos mais comuns é a *Cross Validation*, que divide os dados em várias partes, testando o algoritmo com uma parte de cada vez. Esse método é crucial para avaliar a robustez do modelo, utilizando métricas como RMSE e MAE para verificar a precisão das previsões em cada subdivisão dos dados. A *Cross Validation Stratified* vai além, assegurando que cada subdivisão seja representativa da distribuição original das categorias, mantendo a diversidade necessária para uma avaliação justa e precisa.

Outra abordagem é a *Random Sampling*, que separa os dados em conjuntos de treinamento e teste conforme uma proporção definida, como 70:30, e repete o processo várias vezes. Durante essas iterações, métricas como AUC e *LogLoss* são calculadas para cada conjunto de amostras, ajudando a identificar qualquer potencial sobreajuste ou subajuste do modelo. O método *Leave-one-out*, por sua vez, é utilizado para deixar uma instância de fora para treinar e testar o modelo com as restantes, sendo especialmente útil em conjuntos de dados menores.

No método *Test on Train Data*, embora seja menos recomendado devido à alta probabilidade de gerar resultados imprecisos e enviesados, métricas como R2 e CVRMSE podem ser usadas para fornecer uma ideia inicial da performance do modelo. Já no *Test on Test Data*, onde são utilizados apenas os dados do conjunto de teste, métricas como MCC e F1 Score são aplicadas para avaliar a capacidade do modelo em prever corretamente as classes ou valores contínuos nos novos dados.

Dessa forma, ao combinar essas métricas com os métodos de testagem apropriados, garantimos uma avaliação robusta e detalhada da performance dos modelos preditivos, permitindo-nos identificar e corrigir quaisquer problemas de ajuste ou sobreajuste, e assegurando que os modelos estejam preparados para generalizar bem em novos dados.

Resumindo, temos o *Cross Validation* divide os dados em várias dobras, testando o algoritmo com uma dobra de cada vez. *Cross Validation Stratified* faz validação cruzada, mas as dobras são definidas por um recurso categórico escolhido a partir dos meta-recursos. *Random Sampling* separa os dados em treinamento e teste conforme uma pro- porção definida (ex: 70:30) e repete o processo várias vezes. *Leave-one-out* mantém uma instância de fora para induzir o modelo com as outras, classificando as instâncias mantidas, sendo estável, confiável, mas lento. *Test on Train Data* usa todos os dados para treinamento e teste, comumente gerando resultados imprecisos. *Test on Test Data* utiliza apenas os dados do sinal de teste.

CLASSES DE PROBLEMAS

É importante destacar as diversas classes de problemas que a Inteligência Artificial (IA) se propõe a resolver. A classificação desses problemas permite uma análise mais detalhada e a aplicação de técnicas específicas para cada tipo de desafio, otimizando os resultados e garantindo soluções mais precisas e eficientes.

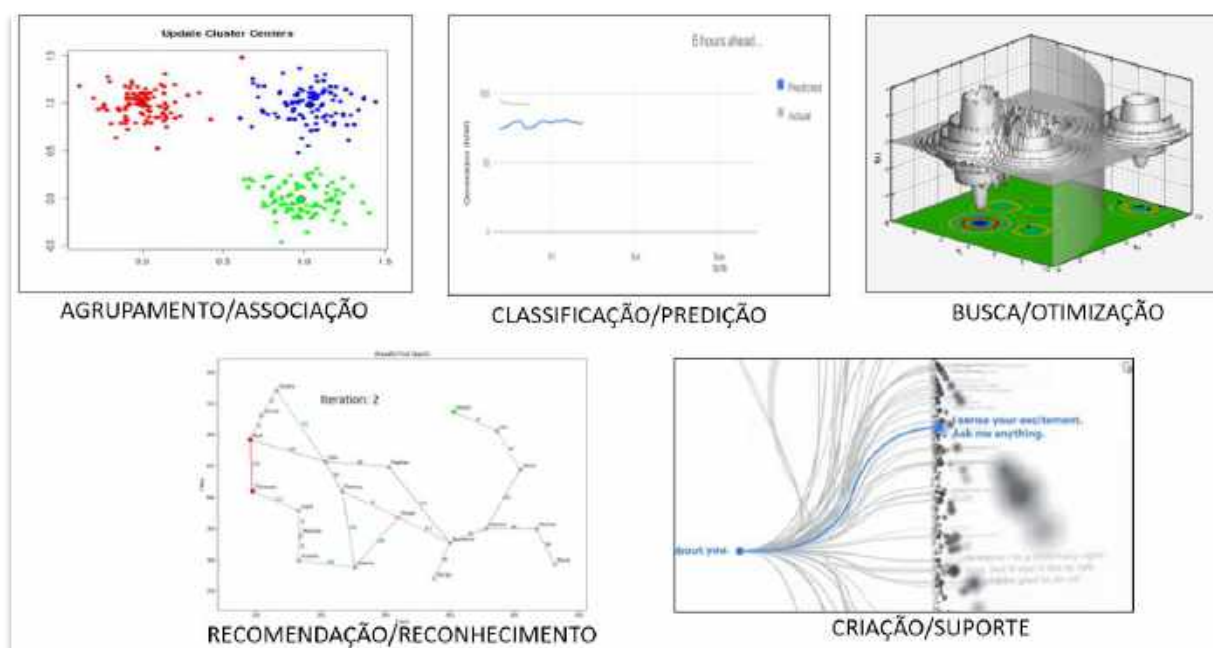


Figura 12: Classes de Problemas

Fonte: Álvaro Pinheiro

CLASSES DE PROBLEMAS EM DIFERENTES PARADIGMAS DE APRENDIZADO

A aplicação de algoritmos de aprendizado de máquina tem revolucionado diversas indústrias, proporcionando análises mais precisas e soluções inovadoras para problemas complexos. No entanto, para garantir o sucesso dessas aplicações, é fundamental compreender as diferentes classes de problemas e os métodos de avaliação correspondentes. Neste contexto, abordaremos algumas das principais classes de problemas em paradigmas de aprendizado supervisionado e não supervisionado, bem como os métodos de validação e métricas de avaliação mais eficazes para cada situação.

É importante destacar as diversas classes de problemas que a Inteligência Artificial (IA) se propõe a resolver. A classificação desses problemas permite uma análise mais detalhada e a aplicação de técnicas específicas para cada tipo de desafio, otimizando os resultados e garantindo soluções mais precisas e eficientes.

APRENDIZADO NÃO SUPERVISIONADO

Agrupamento

O agrupamento é uma técnica de aprendizado não supervisionado que visa dividir um conjunto de dados em grupos ou clusters, com base na similaridade entre os dados. Cada grupo contém itens que são mais semelhantes entre si do que com itens de outros grupos. Algoritmos de agrupamento, como K-means e DBSCAN, são amplamente utilizados em segmentos como análise de comportamento do cliente, segmentação de mercado e análise de redes sociais.

Associação

A associação refere-se ao método de encontrar relações interessantes e úteis entre variáveis em grandes conjuntos de dados. É frequentemente usada em mineração de dados para descobrir padrões frequentes, relações entre produtos e regras de associação, como na análise de cestas de compras. Algoritmos como Apriori e FP-Growth são comuns neste contexto, ajudando a identificar itens que são frequentemente comprados juntos.

APRENDIZADO SUPERVISIONADO

Classificação

A classificação é uma técnica de aprendizado supervisionado onde o objetivo é prever a classe ou categoria de um dado com base em suas características. Algoritmos de classificação, como Árvores de Decisão, Máquinas de Vetores de Suporte (SVM) e Redes Neurais, são aplicados em áreas como diagnósticos médicos, reconhecimento de voz e detecção de spam.

Predição

A predição, ou regressão, é usada para prever valores contínuos com base em dados históricos. Modelos de regressão linear, regressão logística e redes neurais são exemplos de métodos preditivos. Eles são aplicados em previsões meteorológicas, estimativas financeiras e na modelagem de séries temporais.

APRENDIZADO POR REFORÇO

Busca e Otimização

A busca e otimização em aprendizado por reforço envolve encontrar a melhor estratégia ou caminho para maximizar uma recompensa acumulada ao longo do tempo. Algoritmos como *Q-Learning* e *Deep Reinforcement Learning* são usados para treinar agentes autônomos em jogos, robótica e otimização de processos.

RECOMENDAÇÃO E RECONHECIMENTO

Recomendação

Sistemas de recomendação utilizam técnicas de filtragem colaborativa, filtragem baseada em conteúdo e modelos híbridos para sugerir produtos, filmes ou músicas aos usuários com base em suas preferências e comportamento anterior. Modelos de fatoração de matriz e redes neurais colaborativas são frequentemente usados neste contexto.

Reconhecimento

O reconhecimento envolve identificar e classificar objetos, padrões ou sinais em dados. É amplamente utilizado em reconhecimento facial, reconhecimento de fala e sistemas de detecção de anomalias. Algoritmos de redes neurais convolucionais (CNNs) são particularmente eficazes em tarefas de visão computacional.

REDES NEURAIS GENERATIVAS

Criação

Redes neurais generativas, como as Redes Generativas Adversárias (GANs) e os *Autoencoders Variacionais* (VAEs), são usadas para criar dados que são semelhantes ao conjunto de dados de treinamento. Essas técnicas são aplicadas na geração de imagens, música, textos e na simulação de dados sintéticos para treinamento de modelos.

Apoio

O apoio em redes neurais generativas refere-se ao uso dessas redes para melhorar ou enriquecer outros modelos de aprendizado de máquina. Por exemplo, GANs podem ser usados para gerar dados adicionais para treinar modelos de aprendizado supervisionado, melhorando assim a robustez e a performance desses modelos.

A compreensão das diversas classes de problemas em diferentes paradigmas de aprendizado é crucial para a aplicação eficaz de técnicas de inteligência artificial. Cada técnica tem suas vantagens e limitações, e a escolha da abordagem correta depende do tipo de problema e dos objetivos específicos. Ao combinar as técnicas certas com dados de qualidade e métodos robustos de validação, é possível desenvolver soluções inovadoras e eficazes para uma ampla gama de desafios no campo da IA.

CLASSE DE PROBLEMA: MONOMODAL E MULTIMODAL

Os problemas de otimização se dividem em monomodal e multimodal.

Otimização monomodal: há um único objetivo a ser otimizado, com apenas um ótimo global, o que significa que seguir a direção de melhoria sempre levará ao melhor resultado.

Otimização Multimodal: envolve funções com vários ótimos locais, significando várias soluções boas na vizinhança imediata. O desafio é explorar o espaço de soluções para encontrar o ótimo global.

CLASSE DE PROBLEMA: MONO-OBJETIVO E MULTI-OBJETIVO

Os problemas de otimização dividem-se em duas categorias principais: monoobjetivo e multiobjetivo.

Otimização monoobjetivo: Foca-se em um único critério, buscando minimizar ou maximizar uma função específica ao encontrar a melhor combinação de valores das variáveis, respeitando restrições de projeto e operação.

Otimização Multiobjetivo: Envolve otimizar múltiplos objetivos simultaneamente, que geralmente são conflitantes. Não existe uma única solução ótima, mas sim um conjunto de soluções que representam compromissos entre os objetivos.

EXTRAÇÃO, TRANSFORMAÇÃO E CARGA OU EXTRAÇÃO, TRANSFORMAÇÃO E CARGA (ETL)

O processo de ETL é fundamental para a preparação de dados em qualquer solução de inteligência artificial ou análise de dados. Envolve três etapas principais:

Extração: A primeira etapa consiste em extrair dados de várias fontes, que podem incluir bancos de dados relacionais, arquivos de texto, APIs, entre outras. A extração deve ser feita de forma eficiente para garantir que todos os dados relevantes sejam capturados sem perda de informação.

Transformação: Após a extração, os dados brutos são transformados para atender aos requisitos específicos do projeto. Isso pode incluir limpeza, normalização, agregação e outras operações que permitem a conversão dos dados em um formato adequado para análise. A transformação é crucial para garantir a qualidade e a consistência dos dados.

Carga: A etapa final do processo ETL é a carga dos dados transformados em um sistema de destino, como um *data warehouse* ou um banco de dados analítico. A carga pode ser feita em batch ou em tempo real, dependendo das necessidades do projeto e da infraestrutura disponível.

O processo de ETL é essencial para a criação de um pipeline de dados robusto, permitindo que as organizações transformem dados brutos em insights acionáveis.

Figura 13: ETL

Fonte: Álvaro Pinheiro

Dados sujos: com valores faltando e discrepantes em relação ao campo.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	-15	M	F	3000.31	A
2	nulo	F	F	nulo	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	nulo	I
6	21	M	J	9025.76	I
7	32	nulo	F	4002.36	A
8	350	M	J	5001.12	A
9	nulo	nulo	J	7121.33	I

Dados limpos: sem ausência de dados (nulos) e dados íntegros em relação ao campo.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Figura 13: ETL
Fonte: Álvaro Pinheiro

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 1: Dados Discretos
Fonte: Álvaro Pinheiro

Se a idade for medida em anos completos, ela é considerada uma variável discreta, pois pode ser contada e possui um número limitado de valores.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 2: Dados Contínuos

Fonte: Álvaro Pinheiro

A renda é considerada uma variável contínua, isso porque, embora seja expresso em valores monetários, pode assumir um número infinito de valores diferentes dentro de um intervalo.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 3: Dados Categóricos

Fonte: Álvaro Pinheiro

Um campo categórico é um tipo de variável que contém um número finito de categorias ou grupos distintos. Os dados categóricos podem não ter uma ordem lógica e são usados para rotular e classificar elementos em grupos distintos.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 4: Dados Meta
Fonte: Álvaro Pinheiro

Um campo meta é um atributo que não se repete e serve como um identificador, podendo ser uma chave primária ou chave candidata, assim sendo, não é relevante para o processo de generalização.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 5: Dados Alvo
Fonte: Álvaro Pinheiro

Um campo alvo ou target é o objetivo a ser comparado com o resultado obtido no processamento do aprendizado de máquina, isto é, se o obtido com a aplicação do método é igual ao desejado.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 6: Dados para Generalização
Fonte: Álvaro Pinheiro

Um campo característica ou feature é um atributo que se repete e normalmente se trata das características do objeto analisado, podendo ser uma chave estrangeira, assim sendo, é relevante para o processo de generalização.

CARACTERÍSTICAS DOS DADOS

A análise exploratória de dados (EDA) é fundamental para compreender a estrutura e as características de um conjunto de dados antes de aplicar qualquer técnica de modelagem. A EDA envolve a utilização de várias técnicas gráficas e estatísticas para identificar padrões, detectar anomalias e testar hipóteses. Um dos primeiros passos nesse processo é examinar a distribuição dos dados, o que permite entender como os dados estão dispersos e identificar possíveis outliers.

Distribuição

A distribuição de frequência é uma técnica estatística que organiza os dados em uma ordem lógica para facilitar a análise de tendências e a tomada de decisões.

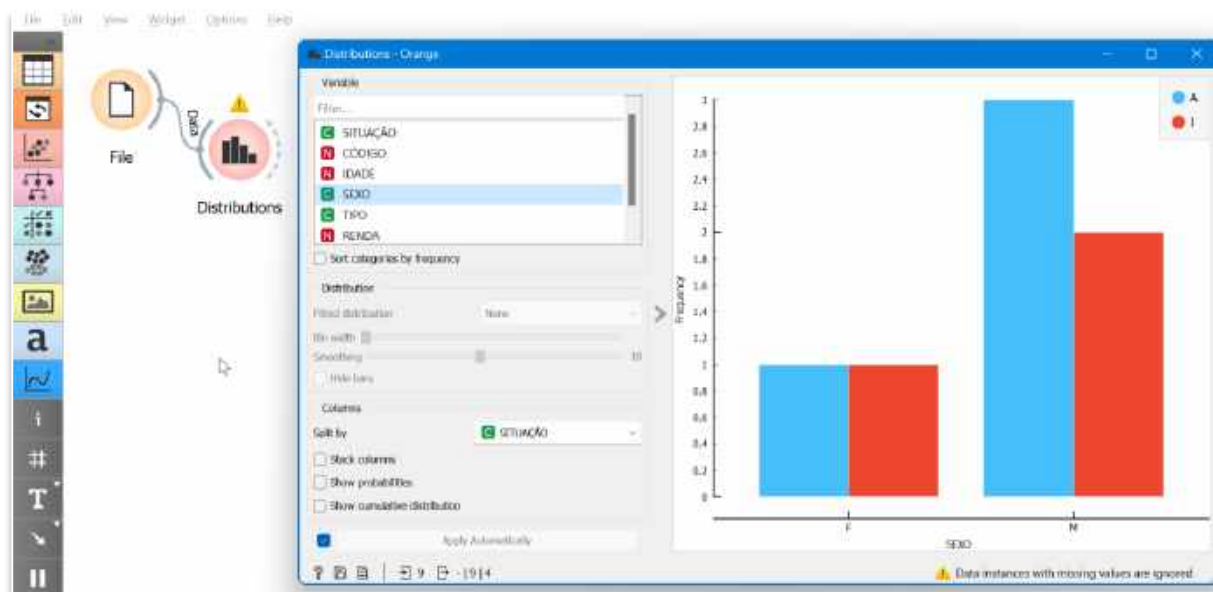


Figura 14: Distribuição
Fonte: Álvaro Pinheiro

Média

A média aritmética é uma medida de tendência central de um conjunto de dados. É calculada somando todos os valores do conjunto e dividindo o resultado pelo número de valores.

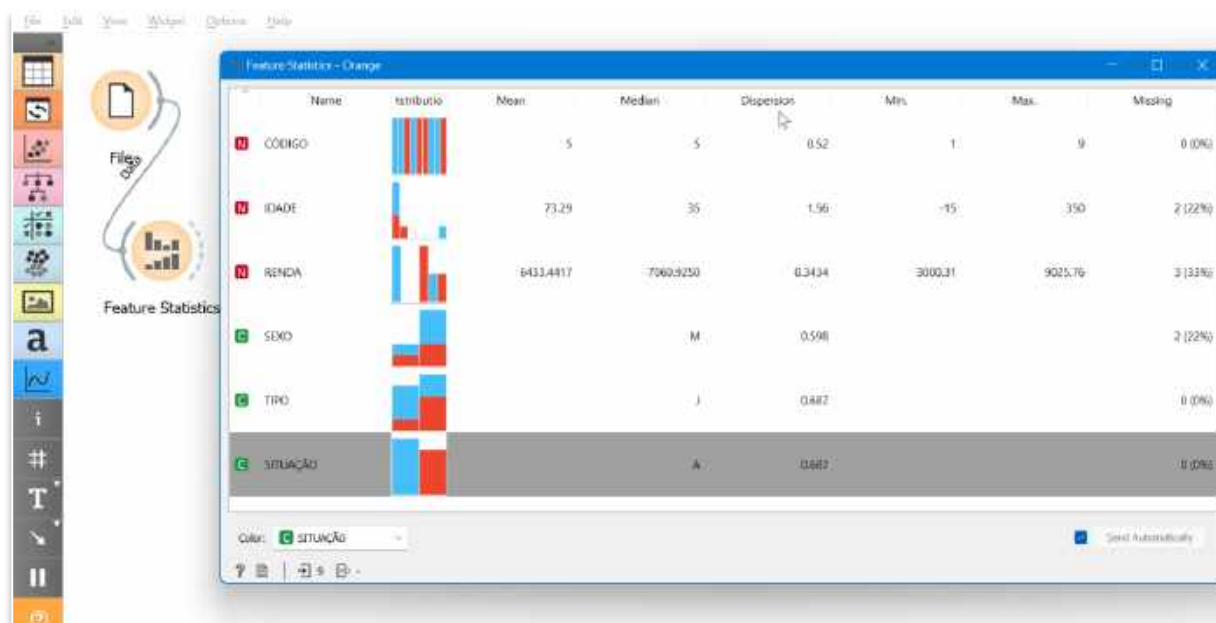


Figura 15: Média
Fonte: Álvaro Pinheiro

Mediana

A mediana é uma medida de tendência central que representa o valor do meio de um conjunto de dados. Para calcular a mediana, primeiro é necessário organizar os dados em ordem crescente ou decrescente. Se o número de observações for ímpar, a mediana é o valor do meio. Se o número de observações for par, a mediana é a média dos dois valores do meio.

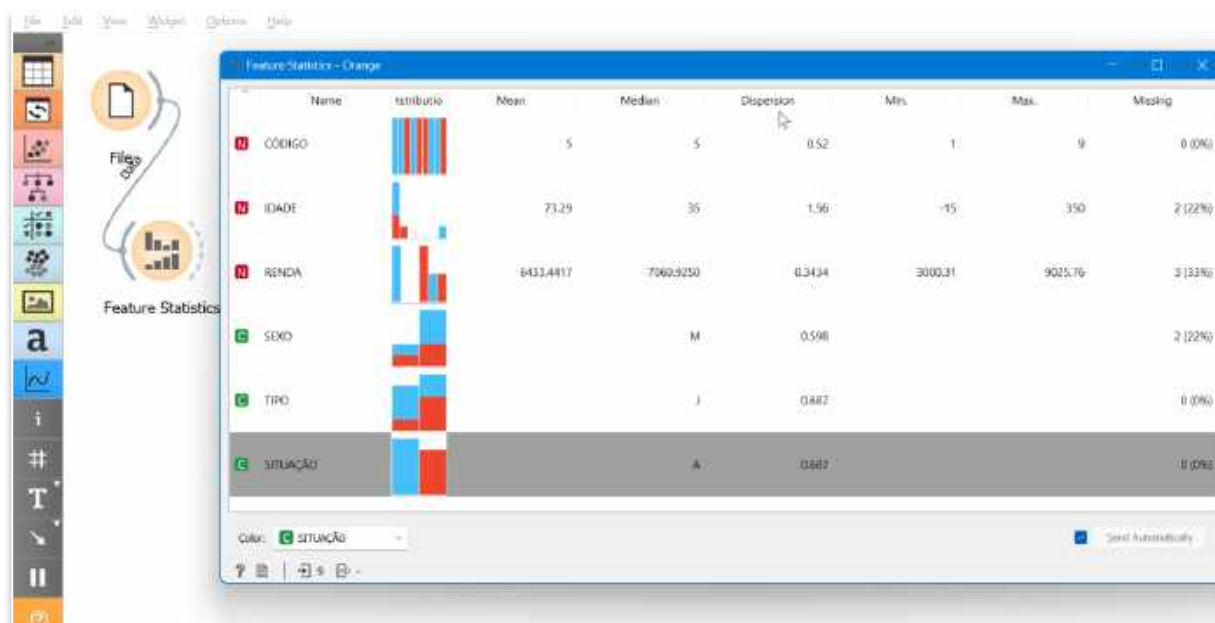


Figura 16: Mediana
Fonte: Álvaro Pinheiro

Dispersão

É uma medida que indica o grau de variação ou flutuação de uma variável aleatória, usada para descrever o quanto os valores de um conjunto de dados estão espalhados em relação à sua média, podendo ser usado desvio padrão, a variância e o intervalo interquartil, ajudando a entender a variabilidade dos dados e a comparar diferentes conjuntos de dados.

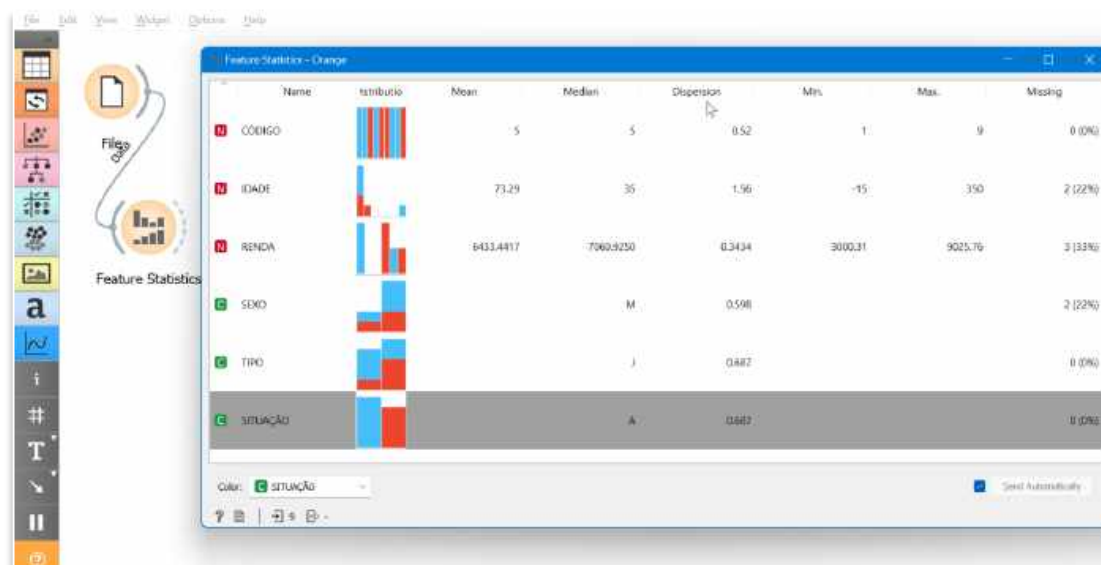


Figura 17: Dispersão
Fonte: Álvaro Pinheiro

Moda

A moda é uma medida de tendência central na estatística que indica o valor mais frequente em um conjunto de dados. Ao contrário da média e da mediana, que consideram todos os valores de um conjunto, a moda se foca no(s) valor(es) que aparecem com maior frequência. Em conjuntos de dados com múltiplos valores que se repetem com a mesma frequência, pode-se ter uma distribuição multimodal. A moda é especialmente útil em situações em que se deseja identificar a categoria mais comum ou popular em um estudo ou pesquisa.

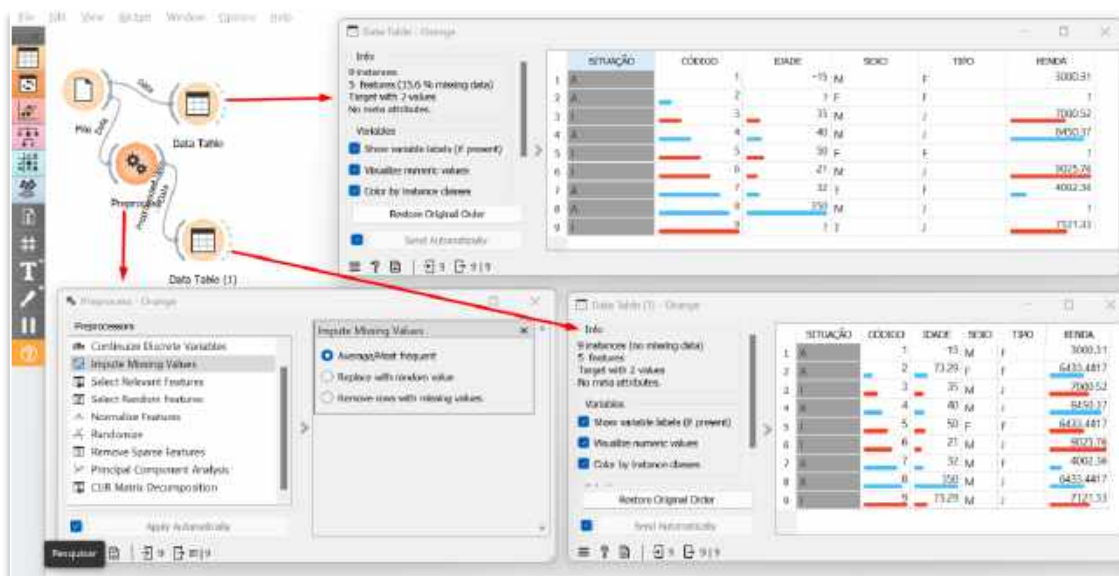


Figura 18: Moda
Fonte: Álvaro Pinheiro

Limpeza

A limpeza de dados é uma das etapas do *Extract, Transform and Load (ETL)*, que consiste em extrair, transformar (limpar) e carregar os *datarow* (i.e., dados brutos) de um ou mais *datasource* (i.e., fonte de dados). Essa limpeza pode consistir em transformar dados nulos em dados de valor, por exemplo, os nulos do campo idade podem ser preenchidos pela média das idades, os nulos do campo sexo pela moda, os nulos do campo renda ou pelo mínimo ou máximo da renda. Essa limpeza pode também ter o objetivo de retirar os outliers.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F		A
3		M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F		I
6	21	M	J	9025.76	I
7	32		F	4002.36	A
8	45	M	J	5001.12	A
9			J	7121.33	I

Tabela 7: Limpeza de Dados
Fonte: Álvaro Pinheiro

Outliers

Observe o campo Idade, existem dois dados discrepantes, o registro de código 2 tem a idade de -15 e o registro de código 8 a idade de 350 anos. Considerando que estamos nos referindo a idade de um humano, esses são dados fora do padrão (outliers).

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	-15	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	350	M	J	5001.12	A
9	51	F	J	7121.33	I

Tabela 8: Outliers
Fonte: Álvaro Pinheiro

Dados Normalizados

Dados não normalizados são aqueles que não foram ajustados para um modelo comum ou padrão, o que pode dificultar a análise e a interpretação. Esses dados podem ter diferentes escalas, unidades de medida ou até mesmo formatos de registro, o que pode levar a distorções ou vieses na análise. Normalizar os dados geralmente envolve transformá-los para uma escala comum, o que pode incluir a padronização (ajustar os dados para uma média de zero e desvio padrão de um) ou a min-max normalização (ajustar os valores dos dados para um intervalo de 0 a 1).

Um exemplo seria se você tivesse um conjunto de dados que medisse alturas em metros e outro em centímetros. Sem normalização, comparar esses dois conjuntos seria difícil e potencialmente enganoso. Normalizar os dados resolveria esse problema ao colocar todas as medidas em uma unidade comum.

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	30	M	F	3000.31	A
2	25	F	F	2500.23	A
3	35	M	J	7000.52	I
4	40	M	J	8450.37	A
5	50	F	F	1057.81	I
6	21	M	J	9025.76	I
7	32	F	F	4002.36	A
8	45	M	J	5001.12	A
9	51	F	J	7121.33	I
IDADE_MIN		IDADE_MAX		RENDA_MIN	RENDA_MAX
21		51		1057.81	9025.76

CÓDIGO	IDADE	SEXO	TIPO	RENDA	SITUAÇÃO
1	0.3000	M	F	0.2438	A
2	0.1333	F	F	0.1810	A
3	0.4667	M	J	0.7458	I
4	0.6333	M	J	0.9278	A
5	0.9667	F	F	0.0000	I
6	0.0000	M	J	1.0000	I
7	0.3667	F	F	0.3695	A
8	0.8000	M	J	0.4949	A
9	1.0000	F	J	0.7610	I
IDADE_MIN		IDADE_MAX		RENDA_MIN	RENDA_MAX
21		51		1057.81	9025.76

$Y = (X - MIN) / (MAX - MIN)$

Figura 19: Dados Normalizados
Fonte: Álvaro Pinheiro

Dados Balanceados

Para garantir uma análise precisa e significativa, é essencial que os dados sejam balanceados. Dados desbalanceados podem introduzir vieses significativos nas análises e levar a conclusões incorretas. Por exemplo, em um cenário de classificação, um conjunto de dados desbalanceado pode fazer com que o modelo aprenda a favorecer a classe majoritária, ignorando a classe minoritária. Isso pode ser especialmente problemático em áreas críticas,

como a detecção de fraudes ou diagnósticos médicos, onde a precisão em identificar a classe minoritária é vital.

Existem várias técnicas para balancear os dados, como a subamostragem da classe majoritária ou a superamostragem da classe minoritária. Essas técnicas ajudam a criar um conjunto de dados mais equilibrado, onde todas as classes têm uma representação igual ou similar, permitindo que os modelos aprendam de forma mais justa e precisa.

Existem 5 amostras de situação "Ativo"

CÓDIGO	IDADE	SEXO_M	SEXO_F	TIPO_F	TIPO_J	RENDA	SITUAÇÃO
1	30	1	0	1	0	3000.31	A
2	25	0	1	1	0	2500.23	A
3	35	1	0	0	1	7000.52	I
4	40	1	0	0	1	8450.37	A
6	21	1	0	0	1	9025.76	I
7	32	0	1	1	0	4002.36	A
8	45	1	0	0	1	5001.12	A

Existem 2 amostras de situação "Inativo"

CÓDIGO	IDADE	SEXO_M	SEXO_F	TIPO_F	TIPO_J	RENDA	SITUAÇÃO
1	30	1	0	1	0	3000.31	A
2	25	0	1	1	0	2500.23	A
3	35	1	0	0	1	7000.52	I
4	40	1	0	0	1	8450.37	A
5	50	0	1	1	0	1057.81	I
6	21	1	0	0	1	9025.76	I
7	32	0	1	1	0	4002.36	A
8	45	1	0	0	1	5001.12	A
9	51	0	1	0	1	7121.33	I

Figura 20: Dados Balanceados
Fonte: Álvaro Pinheiro

Para balancear os dados ou se deve eliminar o excedente de registros do conjunto de dados com maior amostra ou se adicionar registros para o conjunto de dados com menor amostragem.

Correlação

Correlação de Spearman é um teste que mede o grau de correlação, teste desenvolvido pelo psicólogo e estatístico Charles Spearman, descreve a relação entre as variáveis através de uma função que analisa se o valor de uma variável aumenta ou diminui quando o valor da outra variável que se analisa a correlação aumenta ou diminui.

Correlação de Pearson é outra forma de medir o grau de correlação linear entre duas variáveis quantitativas, teste desenvolvido pelo matemático inglês Karl Pearson, que trata do coeficiente de correlação que pode variar de -1 a 1, indicando a direção e a intensidade da associação entre as variáveis.

A correlação de Pearson avalia apenas as relações lineares, já a de Spearman avalia relações lineares e não lineares, quando não houver valores de dados repetidos, uma correlação de Spearman perfeita de +1 ou -1 ocorre quando cada uma das variáveis é uma função monótona perfeita da outra.

Na matemática, uma função monótona ocorre entre dois conjuntos ordenados quando ela preserva (ou inverte) a relação de ordem. Quando a função preserva a relação, ela é chamada de função crescente. Quando ela inverte a relação, ela é chamada de função decrescente.

DATA VISUALIZATION EXPLORATORY DATA ANALYSIS (EDA)

EDA significa Análise Exploratória de Dados, uma abordagem usada para analisar e investigar dados, para: descobrir padrões e anomalias; testar hipóteses, validando suposições usando medidas estatísticas; extrair insights e informações que podem alimentar modelos de aprendizado de máquina; e, direcionar as tomadas de decisões de negócios baseadas em dados. Isso, através de várias técnicas, principalmente gráficas, para examinar e estudar os dados.

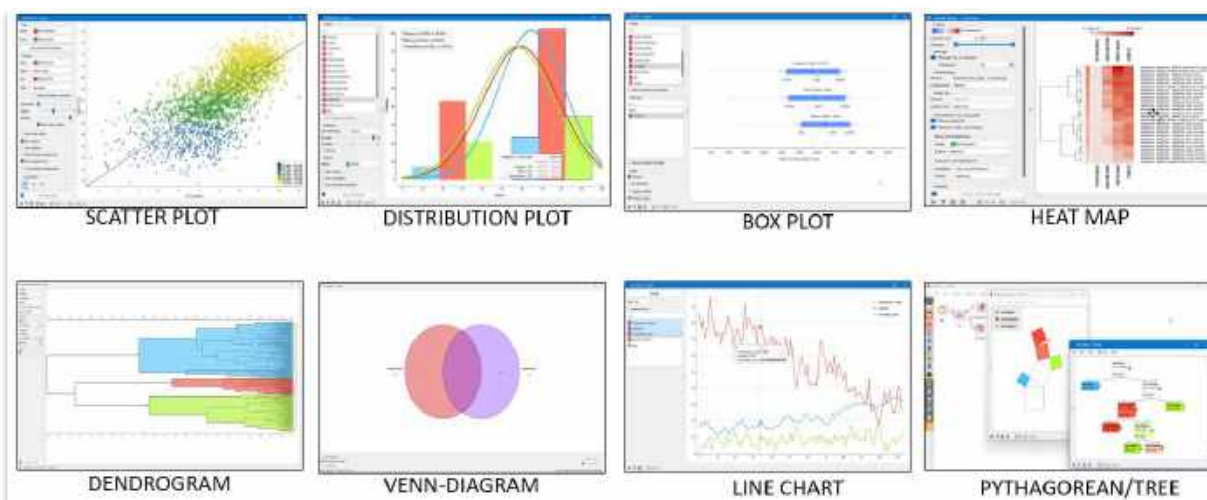


Figura 21: Tipos de Gráficos
Fonte: Álvaro Pinheiro

EDA: SCATTER PLOT

Um gráfico de dispersão exibe a relação entre duas variáveis numéricas através de pontos nos eixos horizontal e vertical, ajudando a identificar correlações, padrões, outliers ou grupos nos dados, desde que haja pelo menos duas variáveis numéricas no conjunto.

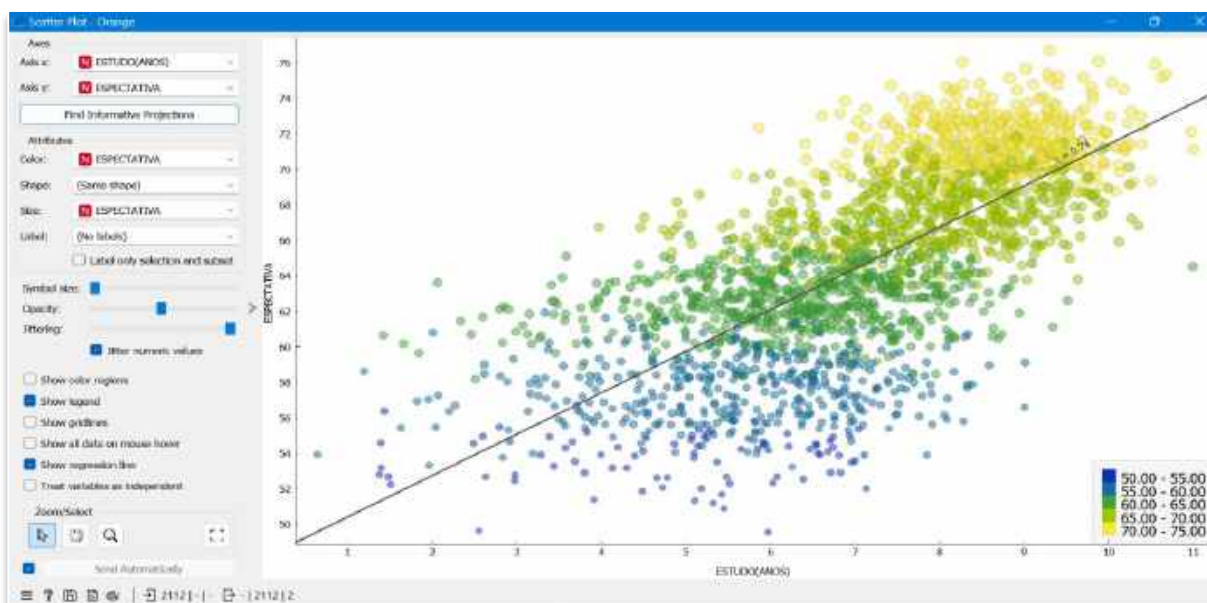


Figura 22: Gráfico de Dispersão
Fonte: Álvaro Pinheiro

EDA: VARIÁVEIS DEPENDENTES E INDEPENDENTES

Ajuda a analisar relações entre diferentes elementos, permitindo testar hipóteses e verificar correlações entre variáveis. Variável Dependente é influenciada pela variável Independente, que representa o efeito ou resultado a ser medido. Variável Independente determina mudanças na variável dependente, sendo o fator manipulado. A relação de causa e efeito é analisada ao alterar a variável independente para observar o impacto na variável dependente.

EDA: TRATAR VARIÁVEIS COMO INDEPENDENTES

Tratar variáveis como independentes significa que a mudança em uma não implica uma mudança na outra. Cada variável tem sua própria distribuição, útil para analisar relações sem supor causalidade. Por exemplo, ao examinar altura e peso das pessoas, você pode tratá-las como independentes. Embora haja correlação, um aumento na altura não necessariamente causa aumento no peso.

EDA: LINHA DE REGRESSÃO

A linha de regressão em um gráfico de dispersão é uma reta que demonstra como a variável dependente (variável de resposta) se altera com mudanças na variável independente (variável explicativa). Essa linha é empregada para prever o valor da variável dependente baseado em um dado valor da variável independente, sendo essencial para a análise do modelo de regressão linear, um modelo estatístico utilizado para prever o valor de uma variável tomando como base outra. Na interpretação, a inclinação da linha de regressão indica a relação entre as variáveis: se a linha sobe, isso revela uma relação positiva; se desce, a relação é negativa. A previsão é outro uso, onde a linha de regressão facilita antecipar valores futuros, como prever vendas futuras a partir dos gastos com publicidade. Em termos de correlação, a linha de regressão ajuda a compreender a força da correlação entre duas variáveis, já que quanto mais próximos os pontos de dados estiverem da linha de regressão, mais forte será a correlação.

EDA: CORRELAÇÃO

Refere-se à relação mútua entre dois termos, sendo uma medida padronizada da conexão entre duas variáveis. Se a correlação estiver próxima de zero, indica que não há relação significativa entre as variáveis. Uma correlação positiva sugere que ambas as variáveis se movem na mesma direção, e a correlação é mais forte quanto mais se aproxima de 1. Por outro lado, uma correlação negativa indica que as variáveis se deslocam em direções opostas, com a força da correlação aumentando à medida que se aproxima de -1. Assim, essa medida descreve o grau de "relação" entre duas variáveis.

EDA: DISTRIBUTION PLOT

Os dados mostram como estão distribuídos, indicando valores mais ou menos frequentes, mínimos e máximos, padrões ou tendências, bem como valores anormais. Isso facilita a compreensão dos dados, a identificação de problemas ou oportunidades, a comparação entre grupos e tomadas de decisões baseadas em evidências. Diferentes tipos de gráficos de distribuição são usados conforme o tipo de variável e objetivo da análise.

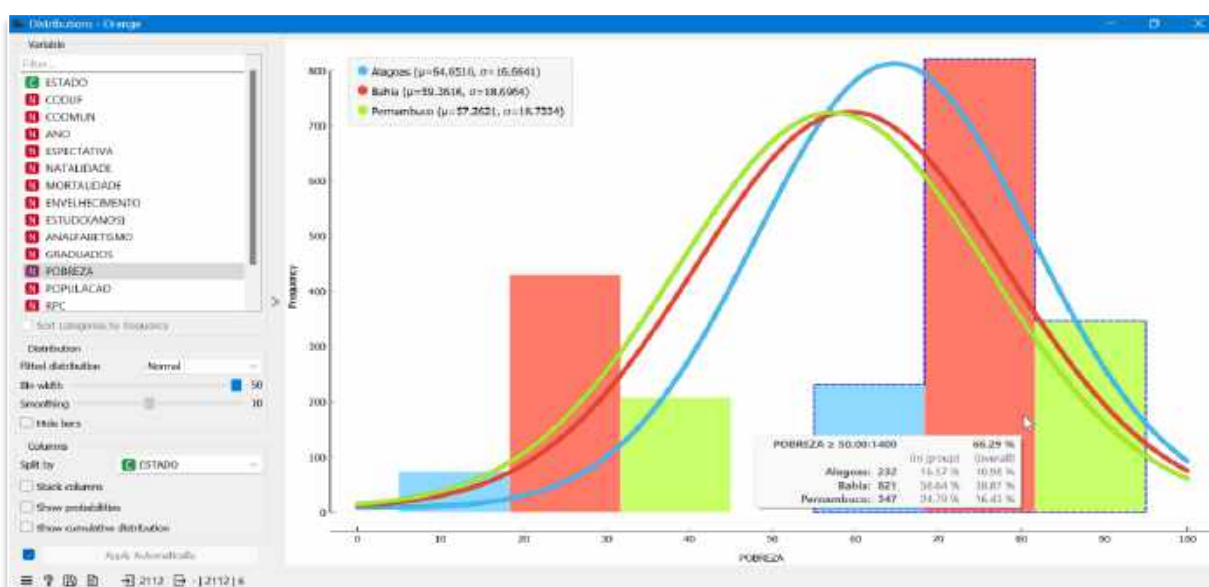


Figura 24: Gráfico de Distribuição
Fonte: Álvaro Pinheiro

EDA: BOX PLOT

O Box Plot, também conhecido como Diagrama de Caixa, é uma ferramenta estatística que representa a distribuição dos dados usando cinco componentes principais: mínimo (sem considerar outliers), primeiro quartil (representando 25% dos dados abaixo deste ponto), mediana (valor central dos dados), terceiro quartil (indicando que 75% dos dados estão abaixo deste ponto) e máximo (sem incluir outliers). É útil para avaliar a tendência central, dispersão, assimetria dos dados e identificar valores atípicos.

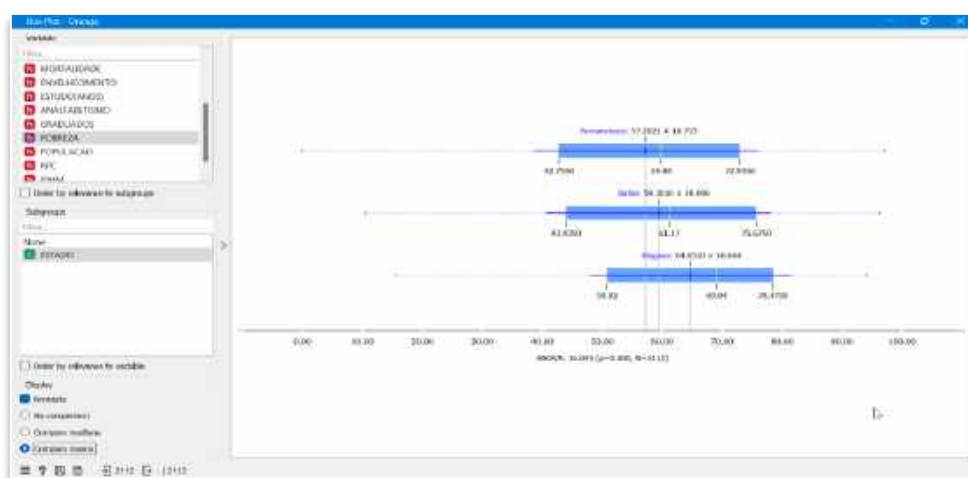


Figura 25: Gráfico de Caixa
Fonte: Álvaro Pinheiro

EDA: HEAT MAP

O gráfico de mapa de calor, ou *heatmap*, é uma ferramenta de visualização que usa a intensidade das cores para exibir grandes quantidades de dados comparativos, como: valores dos dados, mostrando informações em blocos coloridos em formato tabular; padrões e tendências, permitindo identificar facilmente grandes quantidades de dados; concentração de medidas, indicando áreas específicas de interesse; e análise do comportamento, destacando áreas de maior intensidade.

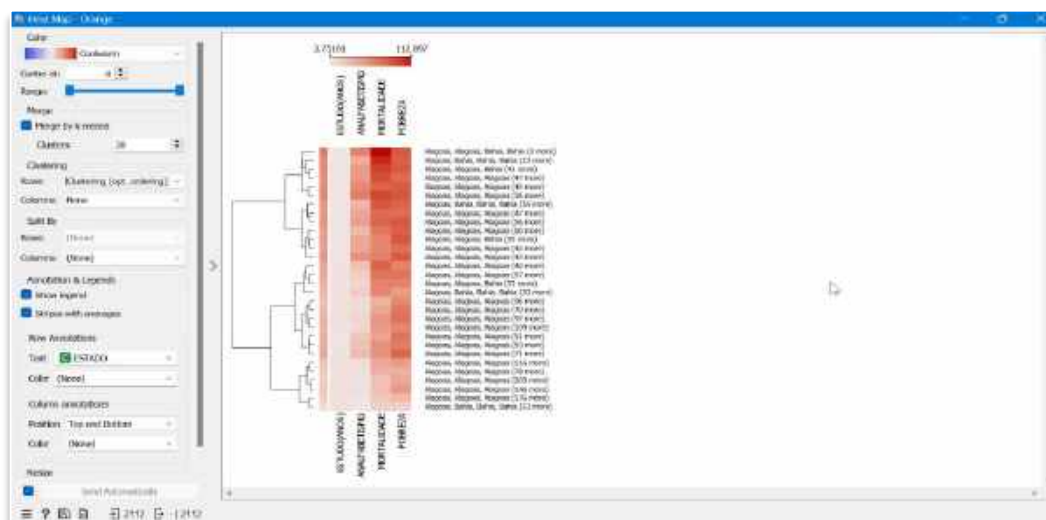


Figura 26: Gráfico de Calor
Fonte: Álvaro Pinheiro

EDA: SILHOUETT PLOT

O gráfico *Silhouette Plot* é uma ferramenta de visualização para interpretar e validar a consistência de clusters de dados. Ele mostra graficamente como cada objeto foi classificado no seu cluster, calculando um coeficiente de silhueta que mede a semelhança do ponto com outros pontos do mesmo cluster e de clusters diferentes. Este coeficiente varia de -1 a 1, onde um valor alto indica boa classificação. O gráfico ajuda a visualizar a consistência dos clusters e a determinar o número ideal de clusters, maximizando o coeficiente médio da silhueta.



A projeção linear, ligada à regressão linear na ciência de dados, modela e analisa relações entre variáveis. Seus objetivos principais são: prever valores com base em outras variáveis; entender a força da relação entre variável dependente e independentes; identificar intercepto e coeficiente que mostram onde a linha cruza o eixo y e como a variável dependente muda para cada unidade independente; testar hipóteses sobre relações existentes. A regressão linear assume uma relação linear; se não for o caso, técnicas como a regressão polinomial podem ser mais adequadas.

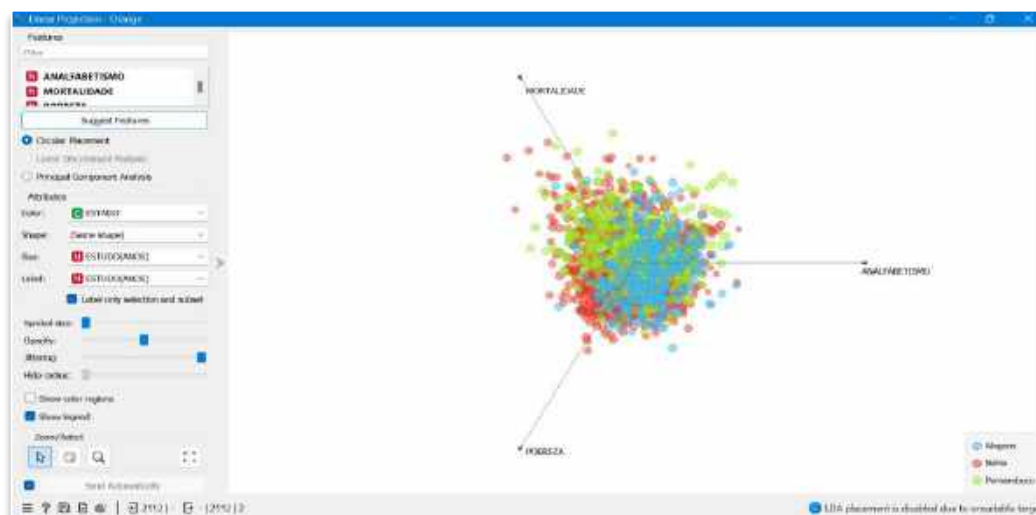


Figura 28: Gráfico de Projeção
Fonte: Álvaro Pinheiro

EDA: LINE PLOT

Os "*line plots*" ou gráficos de linha são ferramentas fundamentais na visualização de dados. Eles ajudam a mostrar tendências ao longo do tempo, como o crescimento de vendas, permitem observar flutuações e variações, comparar diferentes grupos em um único gráfico, identificar a relação entre duas variáveis e fazer previsões com base em dados históricos. Em suma, são poderosos para interpretar séries temporais, revelando padrões, tendências e anomalias de maneira clara.

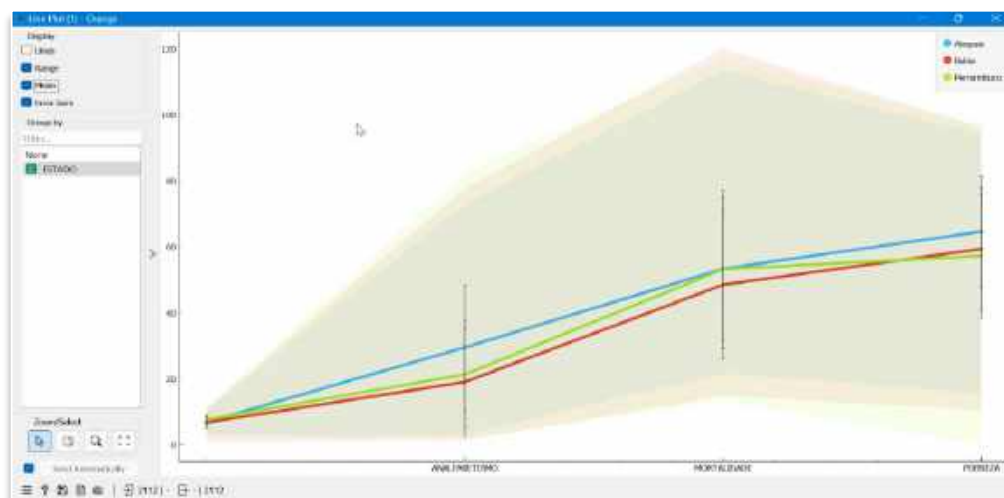


Figura 29: Gráfico de Nanograma
Fonte: Álvaro Pinheiro

EDA: DENDROGRAM

Um dendrograma é um diagrama de árvore que mostra grupos formados pelo agrupamento de observações e seus níveis de similaridade. Ele é frequentemente usado em agrupamento hierárquico. Na interpretação do dendrograma: (1) os nós representam observações ou grupos, conectados por similaridade; (2) a altura indica a dissimilaridade entre nós, conosco unindo-se mais abaixo sendo mais semelhantes; e (3) o corte do dendrograma define o agrupamento final.

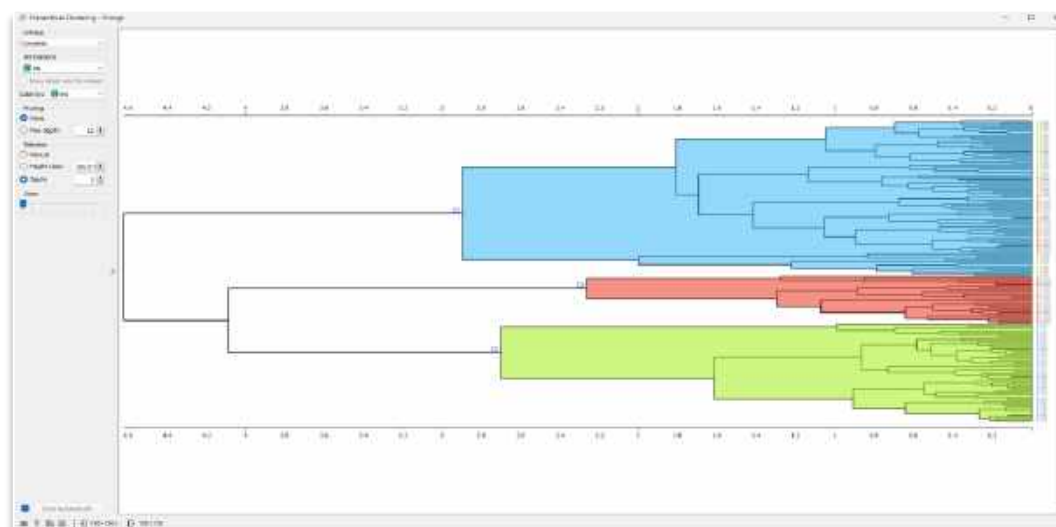


Figura 30: Gráfico de Dendrograma
Fonte: Álvaro Pinheiro

EDA: VENN DIAGRAM

O Diagrama de Venn é uma ferramenta visual usada para mostrar as relações entre conjuntos. Na ciência de dados, ele é útil para visualizar intersecções, como em análises de mercado mostrando quem comprou Produto A e B; comparar dados destacando semelhanças e diferenças; analisar probabilidades, visualizando eventos e suas chances condicionais; e operações booleanas, ilustrando união, intersecção e diferença. Em resumo, Diagramas de Venn são eficazes para clarear as relações entre conjuntos, facilitando a compreensão dos dados.

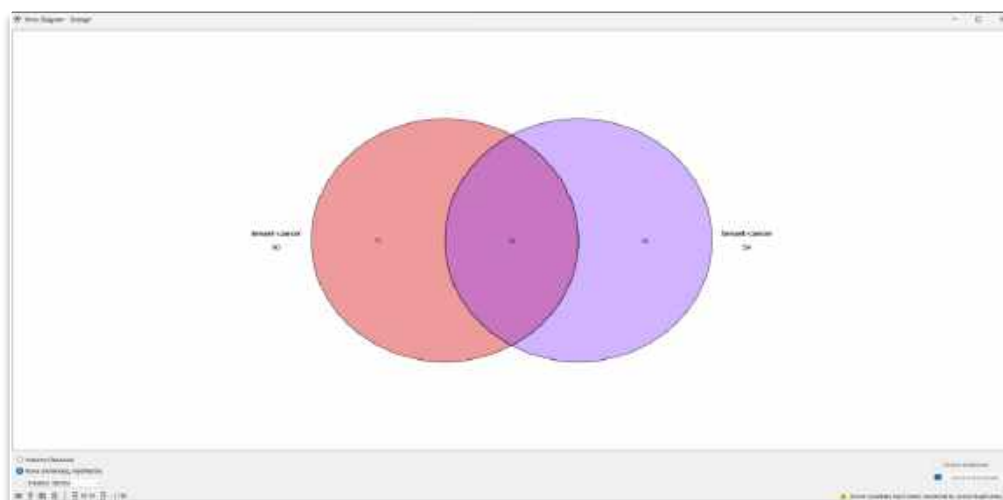


Figura 31: Gráfico de Venn
Fonte: Álvaro Pinheiro

EDA: LINE CHART

O gráfico de linha é ideal para mostrar dados, destacando mudanças e tendências ao longo do tempo ou eventos. Suas principais características incluem: 1. Eixo X (tempo ou categorias) e eixo Y (valores numéricos); 2. Seleção de variáveis a serem plotadas; 3. Customização dos rótulos dos eixos; 4. Personalização visual (cores, estilo e espessura das linhas); 5. Interatividade com zoom, rolagem e destaque de pontos ao passar o mouse. Essas funcionalidades tornam o *Line Chart* útil para análises exploratórias e identificação de padrões ou tendências.

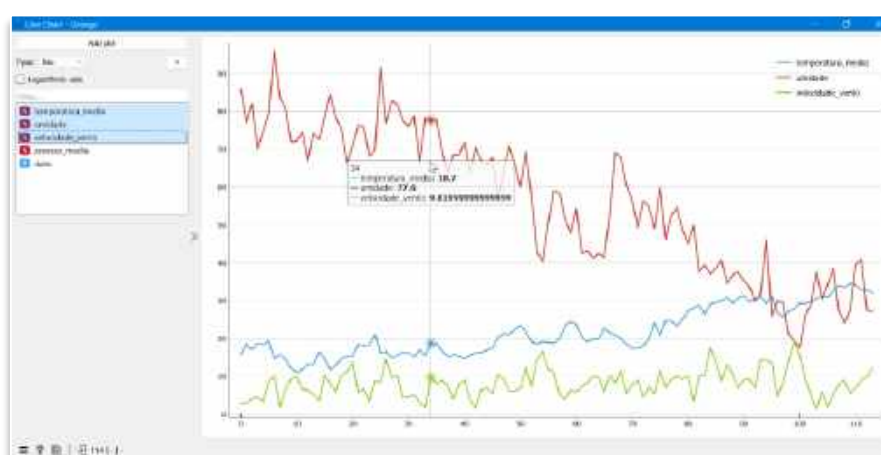


Figura 32: Gráfico de Linha
Fonte: Álvaro Pinheiro

AS VÁRIAS ONDAS DA INTELIGÊNCIA ARTIFICIAL

Ao longo das décadas, a inteligência artificial (IA) passou por diversas fases de desenvolvimento, cada uma marcando avanços significativos na forma como interagimos com a tecnologia. Desde os seus primórdios na década de 1950, quando a internet ainda não existia, até os dias atuais, a IA evoluiu de maneiras que outrora eram inimagináveis. Na primeira onda (1950) a internet não existia. Na segunda onda (1980), as conexões eram limitadas. Na terceira onda (2010), a rede ainda era restrita a centros de pesquisa. Na quarta onda (2018), a OpenAI permitiu amplamente o acesso a algoritmos de IA por meio de diálogos comuns.

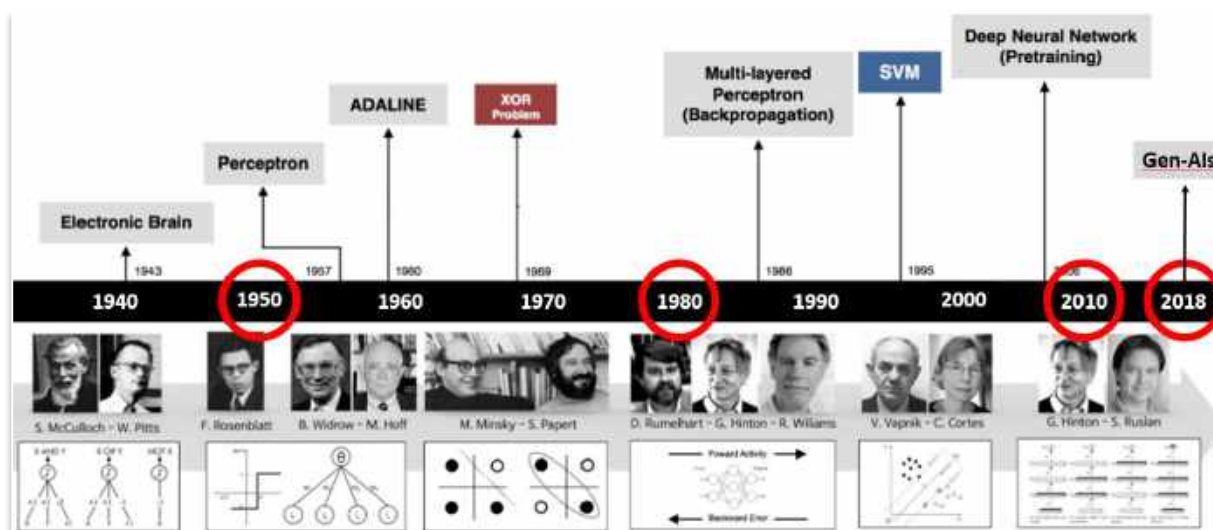


Figura 33: Hypes da IA
Fonte: Sefik Ilkin Serengil

A 4ª ONDA DA INTELIGÊNCIA ARTIFICIAL

O surgimento do ChatGPT desencadeou a 4ª onda da Inteligência Artificial, permitindo que muitos se familiarizassem com a IA e levando outros a acreditar que estavam presenciando algo extraordinário, que apareceu de repente e transformaria o futuro da humanidade rapidamente.



Figura 34: GPT
Fonte: Open-AI

O QUE FEZ AS PESSOAS ACORDAREM PARA O CHATGPT E A IA?

A possibilidade de interagir diretamente com a IA via navegador mudou o cenário. Antes, embora a IA estivesse presente em algoritmos da Amazon, Spotify, Netflix e em sistemas de reconhecimento facial, sua utilização era limitada a pesquisadores, cientistas e grandes organizações. O despertar veio quando a interface do chatGPT permitiu que qualquer pessoa, não apenas especialistas, pudesse interagir diretamente com a IA.

INTELIGÊNCIA ARTIFICIAL NÃO É MÁGICA

A IA utiliza programação e matemática para criar conhecimento usando métodos não simbólicos, de maneira similar à humana.

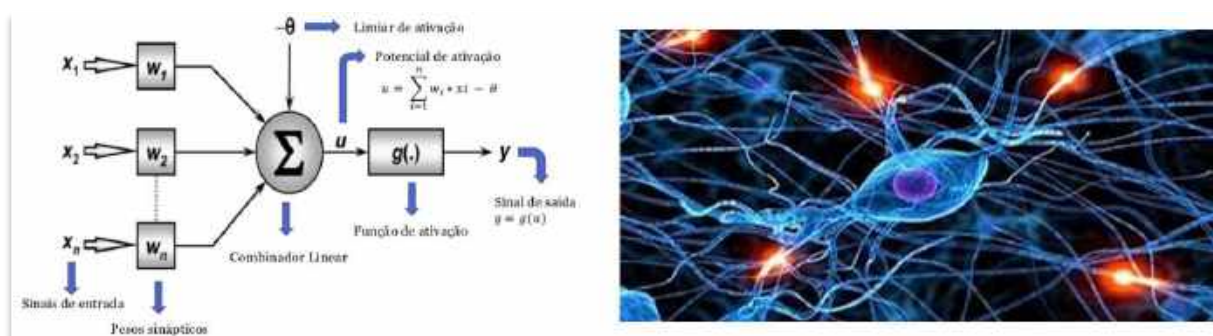


Figura 35: Analogia de uma rede neural artificial e biológica
Fonte: Álvaro Pinheiro

IA É UM PROGRAMA DE COMPUTADOR COMO QUALQUER OUTRO

IA opera em máquinas, recebe dados de entrada, processa e proporciona resultados; são algoritmos criados e controlados por humanos, como qualquer tecnologia.

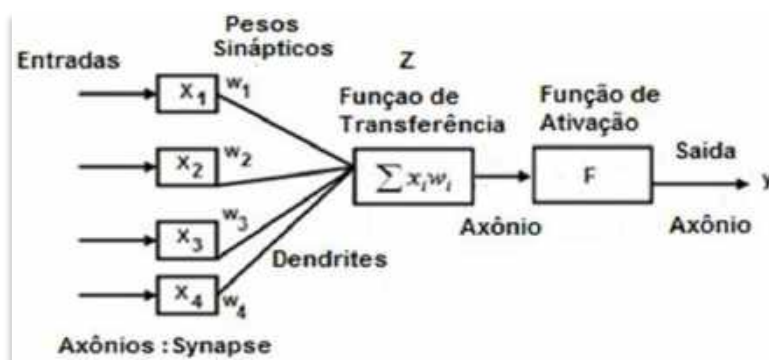


Figura 36: Arquitetura de uma Rede Neural Artificial
Fonte: Álvaro Pinheiro

IA É UM PROGRAMA DE COMPUTADOR COMO QUALQUER OUTRO

As IAs não possuem inteligência real, porque suas saídas são apenas resultados de cálculos probabilísticos. Mesmo que possam escrever textos emocionantes, os algoritmos não compreendem emocionalmente o que produzem.

IA É UMA TECNOLOGIA TRANSFORMADORA

Tecnologias transformadoras geram impactos sociais e econômicos significativos. Elas abrem novas oportunidades e solucionaram desafios de negócios antes insolúveis, beneficiando tanto as empresas quanto a sociedade. As redes generativas são um exemplo dessas tecnologias.



Figura 37: Tecnologias Transformadoras
Fonte: Álvaro Pinheiro

OS PRIMEIROS PASSOS PARA O SURGIMENTO DAS REDES NEURAIS

Warren McCulloch (neurocientista) e Walter Pitts (matemático) colaboraram na criação de modelos computacionais baseados em algoritmos matemáticos conhecidos como lógica de limiar. Em 1943, eles publicaram um artigo seminal sobre redes neurais artificiais, introduzindo o conceito do neurônio artificial, denominado neurônio de McCulloch-Pitts. Este modelo simplificado dos neurônios biológicos se tornou essencial para o desenvolvimento da Inteligência Artificial. Ele foi fundamental para estabelecer uma conexão entre o funcionamento de um neurônio e a lógica proposicional, facilitando a relação com a computação digital.

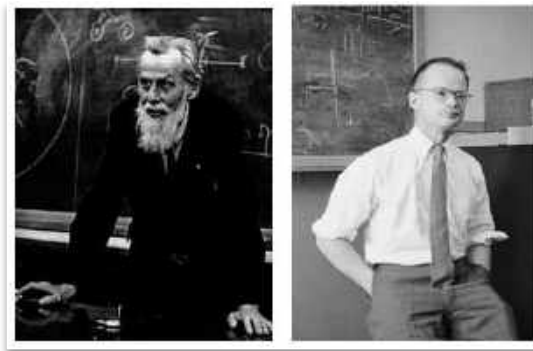


Figura 38: McCulloch e Pitts
Fonte: Internet

A ORIGEM DO TERMO INTELIGÊNCIA ARTIFICIAL (IA)

A expressão pode ser atribuída ao cientista da computação John McCarthy, que a utilizou em uma conferência na Universidade de Dartmouth em 1956. O evento discutiria as “Máquinas Pensantes”, cujos estudos e pesquisas eram conhecidos por termos variados como "Teoria dos Autômatos" e "Cibernética". Para evitar conflitos, McCarthy adotou o termo neutro "Inteligência Artificial".

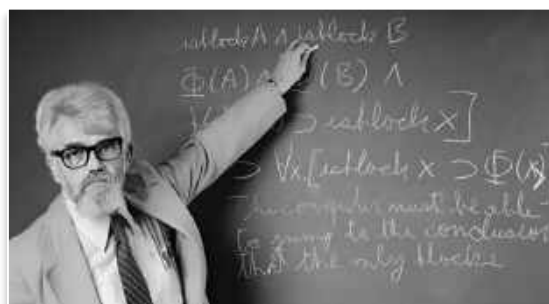


Figura 39: McCarthy

Fonte: Internet

O PRECURSOR DAS REDES NEURAIS

Outro estudo crucial para as redes neurais artificiais foi o Perceptron, desenvolvido pelo psicólogo Frank Rosenblatt em 1957. Este estudo introduziu o aprendizado supervisionado, fundamento do Deep Learning. O Perceptron baseia-se em neurônios, que somam inputs de outros neurônios e disparam sinais ao atingir um limite, valorizando mais conexões fortes e menos as fracas.



Figura 40: Rosenblatt

Fonte: Internet

A ORIGEM DO TERMO MACHINE LEARNING (ML)

Em 1959, o pesquisador Artur Samuel publicou um artigo no IBM Journal of Research and Development sobre “alguns estudos em aprendizado de máquina utilizando o jogo de damas”, marcando a primeira vez que um jogo foi usado para testar o que desde 1956 era

chamado de Inteligência Artificial. O método utilizado baseava-se em um modelo de árvore, uma abordagem comum para a maioria dos modelos de jogos, como o AlphaGo da DeepMind. A cada jogada, o algoritmo calculava as possíveis N jogadas subsequentes, empregando uma função matemática para determinar a sequência mais adequada. Dessa maneira, o programa conseguia “aprender” de forma autônoma.



Figura 41: Samuel
Fonte: Internet

A ORIGEM DA NEURAL NETWORK (NN)

Em 1960, os pesquisadores Bernard Widrow e Marcian Hoff da Universidade de Stanford desenvolveram o algoritmo ADALINE (ADaptive LINear Element), uma rede neural que aprendia com observações repetidas, inspirada nos conceitos do modelo de neurônios biológicos McCulloch-Pitts.

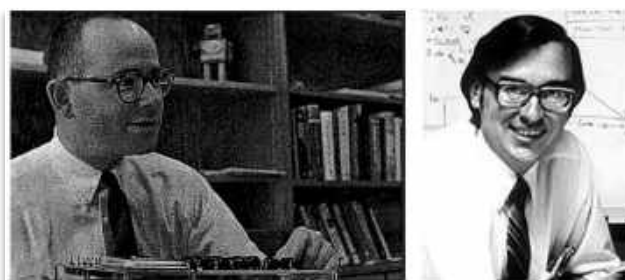


Figura 42: Adaline
Fonte: Internet

ALGORITMO TRANSFORM O LAMPEJO DE CRIATIVIDADE

O algoritmo transformer é fundamental para a IA generativa e facilitou o desenvolvimento do ChatGPT. Em 2017, Ashish Vaswani e Jakob Uszkoreit, pesquisadores do Google, discutiam formas de melhorar o Google Translate enquanto Illia Polusukhin desenvolvia o conceito de "self-attention", revolucionando a velocidade e o método dos algoritmos de processamento de linguagem natural.



Figura 43: Transformer
Fonte: Internet

ALGORITMO TRANSFORM E O CONCEITO “SELF-ATTENTION”

Na época, o *Google Translate*, assim como outros sistemas de PLN, traduzia cada palavra de uma sentença de forma sequencial. Com a introdução do conceito de “self-attention”, que não segue uma sequência linear das palavras, hou-ve uma evolução significativa. O objetivo desse algoritmo é ler a sentença inteira de uma vez só, analisando suas par-tes em conjunto em vez de palavras individualmente, captando todo o contexto para gerar a tradução em paralelo.

Com a contribuição de mais pesquisadores, foi publicado o artigo "Attention is All You Need", uma leitura essencial sobre sistemas generativos. Isso destaca claramente a diferença entre máquinas e humanos: enquanto a criatividade da máquina reside nos algoritmos, a inteligência real está na imaginação dos programadores que os desenvolvem.

IA NÃO É UM SER SENCIENTE, É MATEMÁTICA!

Trata-se de uma matemática onde as equações e os resultados a serem processados superam nossa capacidade humana. Com a atual potência computacional, podemos resolver dados imensos para treinar IAs. “Aprendizado de Máquina” em IA significa encontrar um valor para cada x que dê um y correspondente. A IA é uma vasta equação com trilhões de x e y , utilizando algoritmos sofisticados em constante evolução.

IA GENERATIVA, É UMA FUNÇÃO PROBABILÍSTICA!

A IA generativa é um mecanismo probabilístico de preenchimento automático que foi desenvolvido com uma enorme quantidade de dados. Com essa complexidade matemática, atingimos um ponto onde os sistemas podem imitar funções cognitivas humanas algorítmicas, como aprendizagem e resolução de problemas.



Figura 44: Desempenhos das IAs versus humanos
Fonte: OpenAI

GEN-AIs REPLICAM FUNÇÕES COGNITIVAS ALGORÍTMICAS

Eles não conseguem replicar nossas funções cognitivas não algorítmicas ou não computáveis, como a dor, o amor ou a empatia.

Portanto, apesar de já existirem algoritmos que superam o desempenho humano em certas tarefas específicas, mesmo combinando todos eles, ainda não teríamos um sistema inteligente equiparável ao humano.

“A IA de nível humano ainda está a quinze ou vinte anos de distância”, afirma o cientista cognitivo Steven Pinker.



Figura 45: Pinker
Fonte: Internet

REDES NEURAIIS ARTIFICIAS (ROBERT HECHT-NIELSEN, 1987)

"Um sistema de computação consiste em diversos elementos de processamento simples e bem interconectados, que tratam informações reagindo dinamicamente a entradas externas".



Figura 46: Nielsen
Fonte: Internet

INSPIRAÇÃO BIOLÓGICA

Ela é inspirada no neurônio biológico, a unidade fundamental do cérebro, que contém cerca de 1011 (cento e um bilhões) de neurônios, com cada um formando aproximadamente 1015 (um quatrilhão) de sinapses, resultando em milhares de sinapses por neurônio.

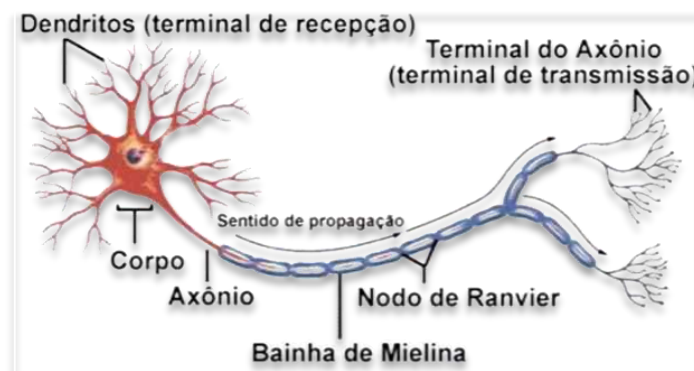


Figura 47: Neurônio Biológico

Fonte: Internet

CÉREBRO BIOLÓGICO

O cérebro biológico é um sistema intrincado formado por unidades simples e altamente interligadas, permitindo realizar funções complexas através da combinação desses elementos básicos.

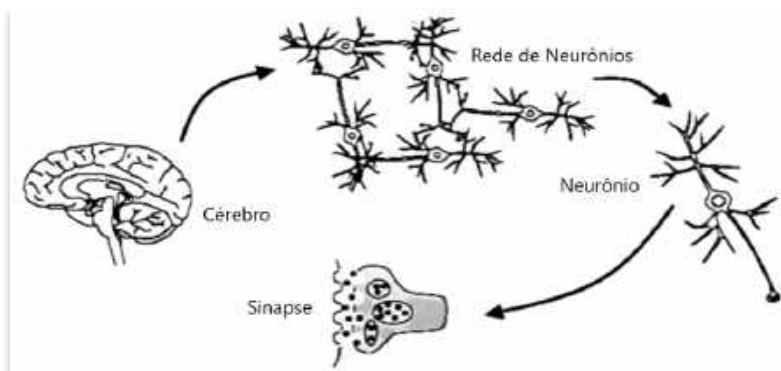


Figura 48: Rede Neural Biológica

Fonte: Internet

REDE NEURAL ARTIFICIAL INSPIRADA NA REDE NEURAL BIOLÓGICA

Utilizar componentes matemáticos básicos para formar uma rede de conexões capaz de executar operações complexas.

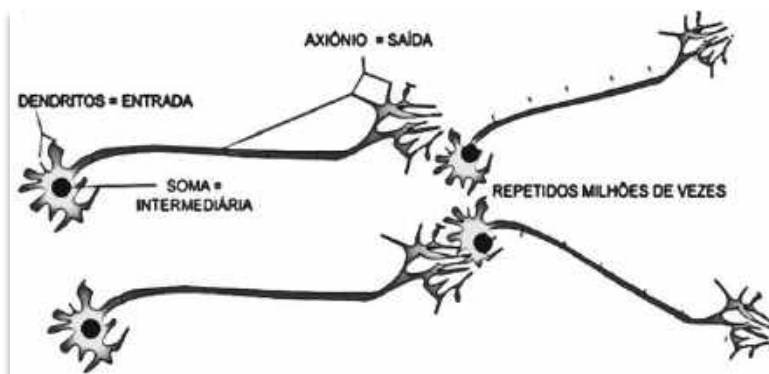


Figura 49: Neurônio Biológico

Fonte: Internet

REDE NEURAL ARTIFICIAL INSPIRADA NA REDE NEURAL BIOLÓGICA

Essas técnicas se baseiam no funcionamento cerebral, onde neurônios artificiais em rede generalizam aproximando funções não lineares.

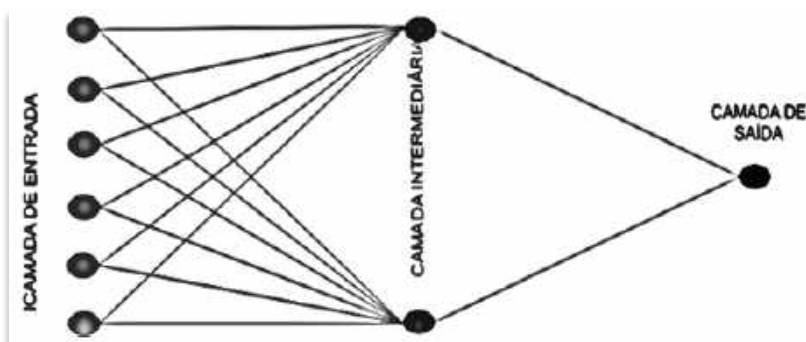


Figura 50: Rede Neural Artificial

Fonte: Internet

DEFINIÇÕES POSSÍVEIS DE UMA REDE NEURAL ARTIFICIAL

Método computacional inspirado nas conexões cerebrais biológicas. Ou, Método matemático para realizar regressão não linear.

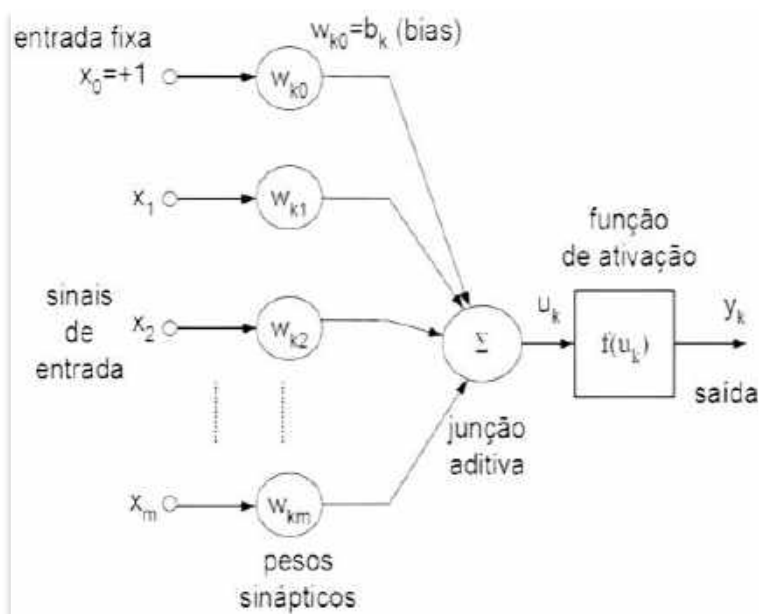


Figura 51: Arquitetura de uma Rede Neural Artificial
Fonte: Álvaro Pinheiro

PERCEPTRON MODELO MAIS SIMPLES DE REDE NEURAL ARTIFICIAL

Os pesos ajudam os separadores a resolverem problemas, ponderando sinais de entrada para que a saída seja positiva ou negativa. O objetivo é ajustar os pesos com base no erro até atingir zero ou o número máximo de épocas. Combinando neurônios, é possível resolver problemas mais complexos.

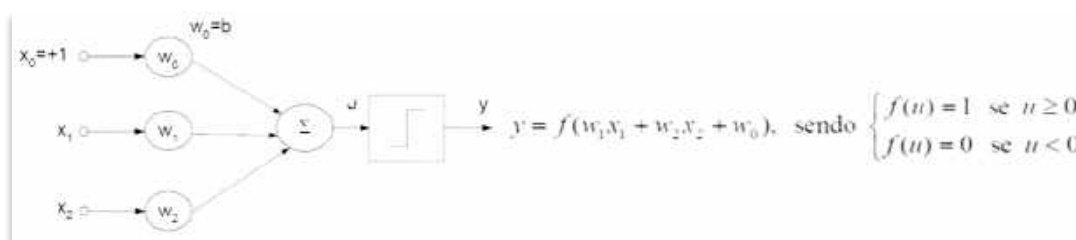


Figura 52: Função Degrau
Fonte: Álvaro Pinheiro

DO PERCEPTRON PARA AS NEURAL NETWORKS

As redes neurais artificiais (RNA) evoluíram do perceptron para arquiteturas mais complexas, permitindo resolver problemas que desafiam a linha reta de separação. A

capacidade de aprendizado e adaptação das redes neurais é o que as torna tão poderosas. Elas ajustam os pesos sinápticos com base em algoritmos de aprendizado, como o gradiente descendente, para minimizar o erro entre a saída prevista e a saída desejada.

A combinação de múltiplas camadas de neurônios, conhecidas como deep learning ou aprendizado profundo, permite que as RNA realizem tarefas complexas, como reconhecimento de voz, processamento de imagem e previsão de séries temporais. Além disso, a introdução de funções de ativação não lineares, como ReLU (Rectified Linear Unit) e sigmoide, traz a necessária complexidade ao modelo, permitindo a captura de padrões não lineares nos dados.

Com o avanço dos métodos de otimização e a disponibilidade de grandes volumes de dados, as redes neurais têm se tornado uma ferramenta essencial em diversas áreas, desde a medicina até a inteligência artificial aplicada à indústria.

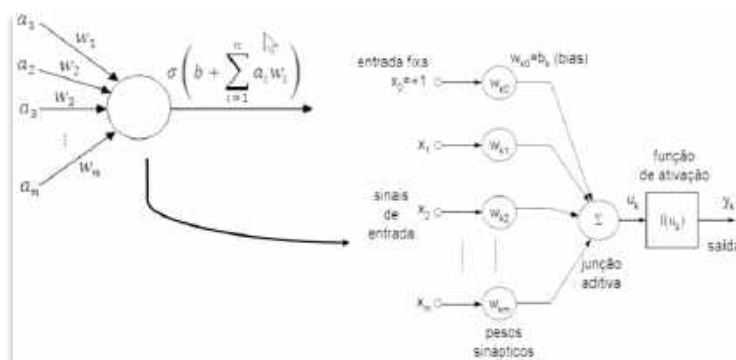


Figura 53: Do Perceptron as MultiLayer Perceptron
Fonte: Álvaro Pinheiro

Os sinais de entrada (x_1 a x_n) são introduzidos na rede; cada sinal é multiplicado por um peso (w_k) para indicar sua influência; A soma ponderada dos sinais produz o nível de atividade (u_k); A função de ativação $f(u_k)$ limita a saída, adicionando não-linearidade ao modelo; O bias (b_k) modifica a influência dos sinais de entrada.

$$y_k = f(u_k) = f\left(\sum_{j=0}^m w_{kj} x_j\right) \quad \text{ou} \quad y_k = f(u_k) = f\left(\sum_{j=1}^m w_{kj} x_j + b_k\right)$$

Figura 54: Função
Fonte: Álvaro Pinheiro

As funções de ativação desempenham um papel crucial nas redes neurais, pois são elas que introduzem a não-linearidade necessária para que o modelo possa aprender e representar funções complexas. Sem as funções de ativação, as redes neurais seriam simplesmente uma combinação linear dos sinais de entrada, limitando sua capacidade de resolver problemas complexos. As funções de ativação mais comuns incluem a sigmoide, a tangente hiperbólica (tanh) e a ReLU (Rectified Linear Unit). A sigmoide e a tangente hiperbólica são funções suaves que limitam a saída em um intervalo finito, enquanto a ReLU permite ativação no intervalo $[0, \infty)$, introduzindo descontinuidade, o que acelera o treinamento das redes neurais profundas.

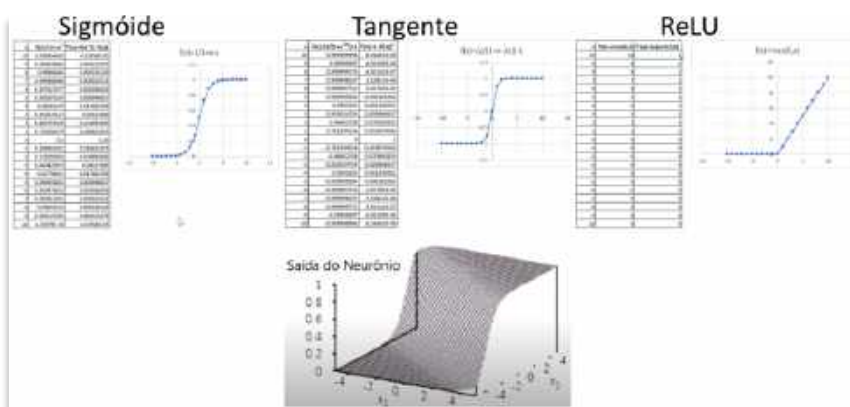


Figura 55: Funções de Ativação
Fonte: Álvaro Pinheiro

O OBJETIVO DAS REDES NEURAIS É MINIMIZAR O ERRO

O gradiente de E , também conhecido como índice de desempenho ou função custo, indica a direção do crescimento mais rápido de E . Redes neurais artificiais ajustam os pesos com base no cálculo do gradiente descendente da função de erro. Aplicando a derivada dessa função, podemos encontrar o ponto que minimiza o erro; invertendo o gradiente, obtemos a direção para ajustar os pesos e minimizar o erro.

$$E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (d_i - y_i)^2$$

Figura 56: Erro
Fonte: Álvaro Pinheiro

GRADIENTE DESCENDENTE FORNECE A DIREÇÃO DO ERRO MÍNIMO

Redes neurais utilizam a derivada da função para encontrar o valor mínimo.

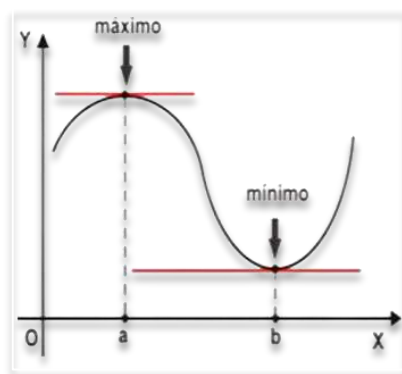


Figura 57: Mínimo da Função
Fonte: Álvaro Pinheiro

AS GENAIS SÃO A EVOLUÇÃO DAS DEEP LEARNING

A evolução das redes neurais tem sido um marco significativo na área da Inteligência Artificial. Desde os primeiros modelos que dependiam de algoritmos simples, as redes neurais progrediram para estruturas complexas, capazes de aprender e adaptar-se com uma precisão impressionante. Um dos avanços mais notáveis foi a introdução do aprendizado profundo, ou "deep learning," que utiliza múltiplas camadas de redes neurais para capturar padrões intrincados nos dados.

O conceito de gradiente descendente tornou-se fundamental nesse processo. Com o uso da derivada da função de erro, as redes neurais podem ajustar seus pesos para minimizar o erro de previsão. Esse método, conhecido como gradiente descendente, dirige a rede neural na direção do erro mínimo, permitindo que aprenda de forma mais eficaz.

Como resultado dessas inovações, emergiram os modelos de IA Generativa, uma forma avançada de IA que pode criar conteúdo original ao responder a prompts.



Figura 58: Evolução das Redes Neurais

Fonte: Álvaro Pinheiro

O QUE SÃO IAs GENERATIVAS (GEN-AIs)?

A IA Generativa refere-se à Inteligência Artificial capaz de produzir conteúdo variados, como textos, imagens, vídeos, áudios ou códigos, em resposta a um comando do usuário utilizando Modelos de Linguagem Grande (LLM).

O QUE SÃO MULTIAGENTES INTELIGENTES NAS GEN-AIs?

Sistemas de Multiagentes Inteligentes consistem em diversos agentes autônomos que colaboram dentro de um mesmo ambiente. Cada agente é capaz de criar conteúdo ou tomar decisões seguindo suas próprias regras e objetivos, influenciando o comportamento geral do sistema.

FERRAMENTAS PARA USAR MULTIAGENTES

CrewAI: desenvolve agentes de IA que colaboram em equipe para solucionar tarefas complexas. Microsoft AutoGen: cria aplicações com LLMs utilizando vários agentes que dialogam para resolver problemas. Flowise: possibilita a criação de modelos personalizados

de linguagem, como chatbots e assistentes virtuais. Spell.so: permite delegar tarefas a agentes de IA autônomos. MindOS: desenvolve assistentes pessoais de IA para sites ou aplicativos. AiAgent.app: cria agentes de IA autônomos para cumprir objetivos definidos pelos usuários. Fine-Tuner.AI: constrói agentes de IA sem necessidade de programação.

A TRANSFORMAÇÃO DA ECONOMIA GLOBAL

A pesquisa indica que a IA afetará cerca de 40% das profissões mundiais, impactando e complementando várias funções. Portanto, é crucial implementar políticas que aproveitem ao máximo o potencial da IA, enquanto consideram as necessidades do mercado de trabalho.



Figura 59: Relatório do FMI
Fonte: FMI de janeiro de 2024

O IMPACTO DA IA NA EMPREGABILIDADE (FMI, DEZ-2023)

Pesquisas apontam que trabalhadores que conseguem integrar a IA em suas atividades diárias desfrutam de maior produtividade e salários mais altos, enquanto aqueles que não se adaptam tendem a experimentar estagnação. Isso evidencia o potencial da IA para impulsionar a eficiência no trabalho. Profissões em vários países têm sido afetadas pela exposição e complementaridade da IA. Nesse contexto, é crucial que os países implementem

programas eficazes de requalificação profissional, visando uma transição mais equitativa para a era da IA, com o objetivo de proteger meios de subsistência e reduzir desigualdades sociais.

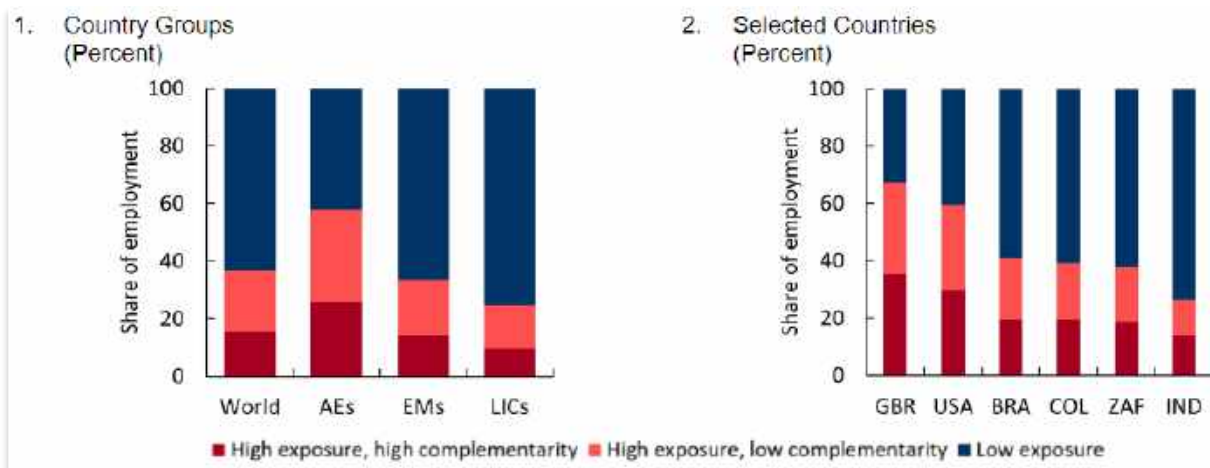


Figura 60: Quotas de emprego por exposição e complemento da IA por grupos de países e selecionados
Fonte: edição de dezembro da revista trimestral Finance and Development – FMI

PARCELA DE EMPREGOS POR DECIS DE RENDA (FMI, DEZ-2023)

Pesquisas indicam que a parcela de empregos em risco devido à IA é semelhante para diferentes faixas de renda, mas tende a aumentar nos níveis de renda mais elevados.

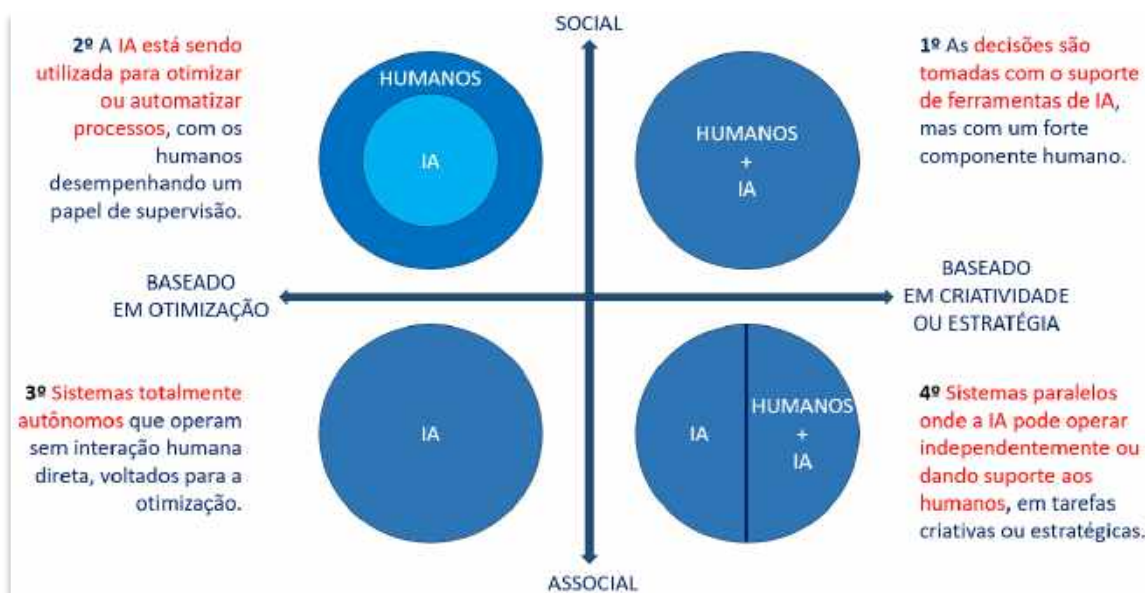


Figura 61: Implementar IA alinhada aos objetivos
Fonte: Kai-Fu Lee, 2019

IA E SEUS SUBCAMPOS

A Inteligência Artificial (IA) é uma das áreas mais fascinantes e revolucionárias da ciência e tecnologia contemporâneas. Ela abrange uma ampla gama de subcampos que se dedicam a criar sistemas capazes de realizar tarefas que, tradicionalmente, requerem inteligência humana. Desde a compreensão e geração de linguagem natural até a simulação de processos de raciocínio complexo, a IA está moldando o futuro de várias indústrias e setores.

Entre os subcampos mais proeminentes da IA, destacam-se os modelos de linguagem generativa, que são projetados para compreender, gerar e interagir com a linguagem humana de maneiras sofisticadas. Esses modelos são categorizados em diferentes "famílias" de acordo com suas arquiteturas e desenvolvedores. Vejamos algumas das famílias mais conhecidas de modelos generativos.

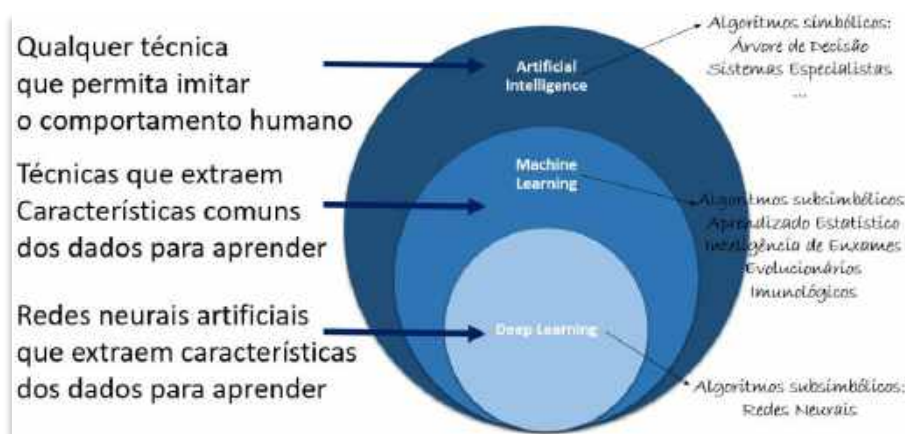


Figura 62: IA e seus Subcampos
Fonte: Álvaro Pinheiro

DS & AI COM SEUS SUBCAMPOS

Dentro do vasto campo da IA, existem subcampos que se destacam não apenas por seu impacto tecnológico, mas também pela forma como estão transformando a sociedade. Um desses subcampos é o aprendizado de máquina, que se concentra em desenvolver algoritmos que permitem que as máquinas aprendam e melhorem com a experiência sem

serem explicitamente programadas para cada tarefa específica. Métodos como redes neurais artificiais, aprendizado supervisionado e não supervisionado, e aprendizado por reforço são apenas algumas das técnicas que impulsionam avanços significativos nesta área.

Outro subcampo crucial é a visão computacional, que envolve capacitar as máquinas a interpretar e entenderem o mundo visual, semelhante à visão humana. Isso inclui tarefas como reconhecimento de objetos, detecção de movimento e análise de imagens e vídeos. A visão computacional tem aplicações diversas, desde a condução autônoma até sistemas de vigilância e diagnóstico médico por imagem.

A robótica é outro subcampo essencial da IA, focado na criação de máquinas inteligentes que possam realizar uma variedade de tarefas físicas no mundo real. Robôs equipados com IA são usados em indústrias de manufatura, agricultura, saúde e até mesmo em missões espaciais, destacando sua versatilidade e importância.

A IA explicável é um subcampo emergente que aborda a necessidade de tornar os processos de tomada de decisão das máquinas mais transparentes e compreensíveis para os seres humanos. Isso é particularmente importante em áreas como finanças, saúde e justiça, onde as decisões automatizadas precisam ser justificáveis e éticas.

No campo da interação humano-computador, a IA está avançando em criar interfaces mais intuitivas e naturais, permitindo uma comunicação mais eficiente entre humanos e máquinas. Assistentes virtuais, sistemas de reconhecimento de fala e tecnologias de realidade aumentada são exemplos de como a IA está facilitando essa interação.

Finalmente, a ética e governança da IA representam um subcampo vital que examina as implicações morais, sociais e legais do desenvolvimento e implementação de tecnologias de IA. Isso inclui questões como privacidade, viés algorítmico, segurança e impacto no emprego, assegurando que o progresso tecnológico seja alcançado de maneira responsável e inclusiva.

Esses subcampos, junto com muitos outros, constituem o vasto e dinâmico campo da Inteligência Artificial, cada um contribuindo para o avanço da ciência e tecnologia e moldando o futuro de inúmeras maneiras.



Figura 63: IA e Subcampos
Fonte: HM Government, 2024

FAMÍLIAS DOS MODELOS GENERATIVOS MAIS CONHECIDAS

1) A Família GPT: Modelos generativos de linguagem desenvolvidos pela OpenAI, como GPT-1, GPT-2 (código aberto) e GPT-3, GPT-4 (código proprietário), acessíveis via APIs.

2) A Família LLaMA: Modelos de linguagem da Meta, de código aberto e disponíveis para a comunidade de pesquisa sob uma licença não comercial, usados para criar LLMs abertos ou específicos para aplicações críticas.

3) A Família PaLM: Desenvolvido pelo Google, o modelo de linguagem PaLM foi lançado em 2022, pré-treinado com 780 bilhões de tokens em chips TPU v4, alcançando resultados de ponta em benchmarks de compreensão e geração de idiomas.

INTELIGÊNCIA ARTIFICIAL (IA)

A Inteligência Artificial (IA) se refere à capacidade do software de realizar tarefas que normalmente requerem inteligência humana. Ela é classificada em quatro categorias:

pensamento vs. comportamento e humano vs. racional. Definições baseadas no pensamento e comportamento humano avaliam o quanto a IA imita a inteligência humana, enquanto definições focadas na racionalidade verificam a eficiência e eficácia do sistema. As quatro abordagens para estudar IA - centrada no humano, empírica, racionalista e engenharia - têm sido usadas por diferentes grupos que impulsionam o campo.

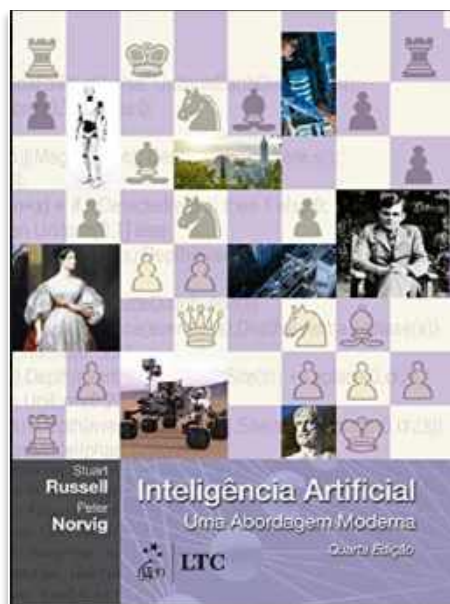


Figura 64: Inteligência Artificial. ISBN: 9788595158870
Fonte: RUSSEL, S.; NORVIG, P.

CATEGORIAS DE IA

As categorias de IA, conforme descritas por Russell e Norvig (2013), fornecem um framework fundamental para entender como a inteligência artificial pode ser classificada e estudada. A figura a seguir ilustra essas categorias, que são divididas em dois eixos principais: pensamento vs. comportamento e humano vs. racional.

No eixo do pensamento vs. comportamento, a análise se concentra em como a IA imita ou simula a inteligência humana através de processos de pensamento ou através de ações observáveis. No eixo humano vs. racional, a ênfase é colocada na comparação de quão humanamente a IA opera em contraste com a eficiência e a lógica racional do sistema.

Essas distinções são cruciais para o desenvolvimento e avaliação de tecnologias de IA, pois ajudam a orientar os pesquisadores e engenheiros a focarem em diferentes aspectos da inteligência de máquina. Por um lado, algumas abordagens podem priorizar a emulação de capacidades humanas específicas, como a cognição ou a percepção, enquanto outras podem buscar maximizar a eficácia e a precisão racional através de algoritmos matemáticos e modelos preditivos.

Estas categorias também servem como uma base para as quatro abordagens principais para o estudo da IA: centrada no humano, empírica, racionalista e engenharia. Cada abordagem oferece uma perspectiva única e contribui para o avanço do campo com uma variedade de metodologias e objetivos de pesquisa.

Pensando como um humano	Pensando racionalmente
“O novo e interessante esforço para fazer os computadores pensarem (...) <i>máquinas com mentes</i>, no sentido total e literal.” (Haugeland, 1985) “[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado...” (Bellman, 1978)	“O estudo das faculdades mentais pelo uso de modelos computacionais.” (Charniak e McDermott, 1985) “O estudo das computações que tornam possível perceber, raciocinar e agir.” (Winston, 1992)
Agindo como seres humanos	Agindo racionalmente
“A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas.” (Kurzweil, 1990) “O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas.” (Rich and Knight, 1991)	“Inteligência Computacional é o estudo do projeto de agentes inteligentes.” (Poole <i>et al.</i>, 1998) “AI... está relacionada a um desempenho inteligente de artefatos.” (Nilsson, 1998)

Figura 65: Categorias de IA
Fonte: RUSSEL, S.; NORVIG, P.

FUNDAMENTOS DA IA (RUSSELL & NORVIG, 2013)

Além das categorias de inteligência artificial descritas por Russell e Norvig, é essencial considerar os fundamentos de IA por meio de várias disciplinas que integram este vasto campo. Cada disciplina contribui de maneira única para o entendimento e desenvolvimento de tecnologias de IA.

(1) Filosofia: A IA levanta questões filosóficas fundamentais como a epistemologia (o estudo do conhecimento), ética (os impactos morais e sociais) e consciência (a possibilidade de máquinas conscientes).

(2) Matemática: Ferramentas matemáticas como álgebra linear, cálculo e estatística são cruciais para o desenvolvimento de algoritmos eficientes e a análise de grandes volumes de dados.

(3) Economia: Teoria dos jogos, otimização de recursos e modelos preditivos ajudam a criar sistemas que podem tomar decisões eficazes em ambientes complexos e dinâmicos.

(4) Neurociência: Redes neurais, processamento sensorial e aprendizagem são inspirados no funcionamento do cérebro humano, buscando replicar esses processos em sistemas artificiais.

(5) Psicologia: A compreensão da cognição, comportamento e memória humanos é fundamental para o desenvolvimento de IA que possa interagir e aprender de maneiras que imitam a inteligência humana.

(6) Engenharia da Computação: Aspectos de hardware, software e programação são a base prática sobre a qual a IA é construída, capacitando as máquinas a executarem tarefas complexas.

(7) Teoria de Controle: Sistemas dinâmicos, automação e loops de feedback são essenciais para a criação de sistemas autônomos que podem se ajustar e regular suas ações em tempo real.

(8) Linguística: O estudo da linguagem natural, semântica, sintaxe e tradução é vital para o desenvolvimento de sistemas de processamento de linguagem natural que possam compreender e gerar texto humano de maneira significativa.

(9) Cibernética: Feedback, interações homem-máquina e regulação são temas centrais na criação de sistemas que podem se adaptar e responder aos estímulos do ambiente de forma inteligente.

NATURAL LANGUAGE PROCESSING

NLP subcampo da IA que ajuda computadores a entender, interpretar e manipular a linguagem humana, permitindo que os computadores compreendam, interpretem e gerem texto ou fala em linguagem humana de forma significativa.

DISCIPLINAS DE IA

NLP subcampo da IA que ajuda computadores a entender, interpretar e manipular a linguagem humana, permitindo que os computadores compreendam, interpretem e gerem texto ou fala em linguagem humana de forma significativa.

Visão Computacional subcampo da IA que permite que computadores e sistemas obtenham informações significativas a partir de imagens digitais, vídeos e outras entradas visuais.

Robótica subcampo da IA que desenvolve dispositivo eletromecânico ou biônico capaz de realizar trabalhos de maneira autônoma ou pré-programada.

Biônica aplicação dos comportamentos inerente a um sistema biológico, estudando seus funcionamentos para desenvolver tecnologia baseado no comportamento.

Semiótica subcampo da IA que estuda todas as formas que o homem utiliza para se comunicar, abrangendo linguagens verbais e não verbais, como a comunicação oral, escrita, desenhada, gestual e corporal.

Aprendizado de Máquina subcampo da IA que se concentra na criação de algoritmos e sistemas que podem aprender e melhorar a partir de dados com base no treinamento.

Aprendizado Profundo subcampo do ML que se concentra em algoritmos que tentam modelar abstrações de alto nível em dados usando um grafo profundo com várias camadas de processamento.

INTELIGÊNCIA ARTIFICIAL / INTELIGÊNCIA COMPUTACIONAL

ALGORITMOS SIMBÓLICOS

Os algoritmos simbólicos utilizam símbolos e regras lógicas explícitas, sendo mais transparentes e adequados para tarefas de interpretação e raciocínio lógico. A escolha depende da aplicação e dos requisitos específicos.

ALGORITMOS SUBSIMBÓLICOS

Algoritmos subsimbólicos representam informações por meio de conexões ponderadas e aprendem padrões automaticamente a partir de dados, ideais para tarefas complexas como reconhecimento de imagens e áudio, mas são menos interpretáveis. A escolha depende da aplicação e dos requisitos específicos.

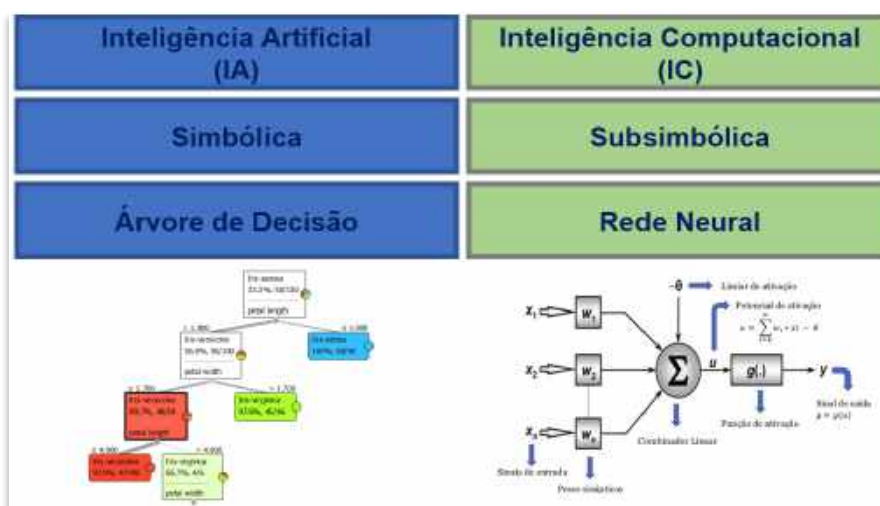


Figura 66: Algoritmos Simbólicos e Não Simbólicos
Fonte: Álvaro Pinheiro

CONCEITOS BASILARES

REDE NEURAL ARTIFICIAL (RNA) Formada por camadas interligadas de software chamadas “neurônios artificiais”, que processam grandes volumes de dados, extraindo e aprendendo características através das várias camadas intermediárias.

REDE NEURAL GENERATIVA (GENAI) Desenvolvida sobre modelos neurais, esta rede inclui recursos para gerar conteúdo, como classificar sentimentos de usuários (negativos ou positivos) a partir de transcrições de áudio ou texto, melhorando consideravelmente os modelos de linguagem.

LARGE LANGUAGE MODEL (LLM) Modelos de Linguagem Grandes são modelos avançados capazes de processar enormes quantidades de dados não estruturados e identificar relações entre palavras ou partes delas. Isso permite que gerem linguagem natural e realizem tarefas como resumo e extração de conhecimento, exemplificados pelo GPT-4, o modelo por trás do BING e GEMINI.

HIPERAUTOMAÇÃO O potencial da hiperautomação (ou seja, automação de processos utilizando algoritmos de IA) diz respeito à fração do tempo de trabalho que pode ser automatizada. Ao analisar o potencial técnico para automação em toda a economia global, considerando as atividades específicas de cada profissão, é possível reduzir custos e aumentar a produtividade em cerca de 850 profissões em aproximadamente 2.100 atividades, com base nas competências exigidas para cada atividade conforme são realizadas atualmente.

AJUSTE FINO O ajuste fino é o ato de adaptar um modelo pré-treinado para aumentar seu desempenho em uma tarefa específica. Esse processo envolve um curto período de treinamento adicional utilizando um conjunto de dados rotulados (i.e., Aprendizado Supervisionado), que é significativamente menor que o conjunto original usado para treinar o modelo (i.e., Aprendizado Não Supervisionado). Esse treinamento extra permite que o modelo aprenda e se ajuste às nuances, terminologias e padrões próprios do novo conjunto de dados.

ÉPOCA Quantas vezes o conjunto de dados de treinamento é totalmente percorrido durante o treinamento.

DATA POINT Na ciência de dados, uma instância refere-se a um exemplo ou observação única dentro de um conjunto de dados. Cada instância inclui um conjunto de características ou atributos que a descrevem. Por exemplo, em um banco de dados sobre casas, cada casa representaria uma instância, e os atributos poderiam incluir o número de quartos, a área do lote, o ano de construção, entre outros. Transformar uma instância em um ponto de dados envolve a representação dessa instância em um espaço dimensional correspondente ao número de atributos. Cada atributo de uma instância torna-se uma dimensão em um espaço multidimensional. Por exemplo, com três atributos (como altura, largura e profundidade), cada instância pode ser mapeada como um ponto em um espaço tridimensional. Em muitos casos, é necessário realizar etapas adicionais de pré-processamento para converter instâncias em pontos de dados utilizáveis. Isso pode envolver a codificação de variáveis categóricas em números, a normalização de variáveis numéricas para colocá-las na mesma escala, o tratamento de valores ausentes, entre outros. Portanto, a transformação de uma instância em um ponto de dados é um processo essencial na preparação de dados para análise na ciência de dados.

No aprendizado de máquina, a dimensão dos pontos de dados indica o número de características que descrevem cada exemplo. Por exemplo, em um conjunto sobre imóveis, isso pode incluir o número de quartos, tamanho da área, localização e preço. Mais características fornecem mais detalhes, mas podem causar problemas como maior complexidade computacional, dificuldades na visualização e overfitting. Técnicas de redução de dimensionalidade, como PCA, MDS ou t-SNE, ajudam a representar os dados em um espaço menor, mantendo informações relevantes.

DATA POINTS (SIMILARITY AND DISSIMILARITY) Dados similares ou simétricos são pontos próximos em termos de características ou atributos, como clientes de um banco que têm idade, renda e histórico de transações semelhantes. Esses pontos podem ser agrupados

em clusters indicando similaridade. Em contraste, dados não similares ou assimétricos são significativamente diferentes, como clientes com idades, rendas e históricos bem distintos, podendo representar perfis únicos ou outliers. Identificar a similaridade ou dissimilaridade é crucial para tarefas de aprendizado de máquina, como clustering, classificação e detecção de anomalias. Esta análise permite agrupar dados semelhantes e identificar padrões ou características distintivas.

INSTANCE TO DATA POINTS Para converter uma instância de um conjunto de dados em um ponto dentro de uma representação dimensional, é crucial utilizar uma técnica conhecida como redução de dimensionalidade. Há vários métodos disponíveis para essa finalidade; um deles é o Multidimensional Scaling (MDS), que permite projetar dados de alta dimensionalidade em um espaço bidimensional, mantendo as distâncias entre as instâncias intactas. Este é o procedimento a ser seguido: (1) calcula-se a matriz de distâncias entre todas as instâncias do conjunto de dados utilizando métricas apropriadas, como distância euclidiana ou correlação; (2) aplica-se o algoritmo MDS à matriz de distâncias para reduzir a dimensionalidade dos dados para 2D. O objetivo do MDS é encontrar uma configuração de pontos que preserve as distâncias relativas entre as instâncias, de modo que as mais semelhantes ou próximas permaneçam próximas na representação em 2D; (3) o resultado será uma representação bidimensional onde cada instância de dados é mapeada para um ponto no espaço, e as coordenadas desses pontos representam a instância na coordenada cartesiana. É relevante destacar que diferentes técnicas de redução de dimensionalidade podem ser mais adequadas dependendo do seu conjunto de dados e do propósito da análise. Além do MDS, métodos populares incluem PCA e t-SNE. Cada técnica possui suposições e considerações específicas, por isso é recomendável explorar várias opções e testar abordagens diversas para encontrar a mais adequada ao seu caso de uso.

DATA NOISE Ruído nos dados indica que as informações presentes podem estar erradas, imprecisas ou inconsistentes. Esse ruído pode originar-se de diversas fontes, como erros de medição, falhas humanas durante a coleta, problemas na transmissão ou corrupção dos dados. O ruído pode impactar negativamente o desempenho dos algoritmos de

aprendizado de máquina, que são suscetíveis a essas imprecisões. Isso pode resultar em modelos imprecisos e previsões incorretas. Portanto, é crucial tratar o ruído nos dados para desenvolver modelos mais robustos e confiáveis, aprimorando a qualidade das previsões e as decisões feitas pelo sistema de aprendizado de máquina.

ENSEMBLE LEARNING O aprendizado em conjunto é uma técnica no campo do aprendizado de máquina que combina múltiplos modelos individuais para criar um modelo final mais robusto e preciso. Ao invés de confiar em um único modelo, o ensemble learning aproveita a diversidade dos modelos para obter melhores resultados de previsão. Cada modelo individual, conhecido como "membro do ensemble", pode ser treinado independentemente, utilizando diferentes algoritmos, configurações, conjuntos de dados ou características.

Existem várias técnicas de ensemble learning, entre as quais podemos destacar:

Bagging (Bootstrap Aggregating): Nesta técnica, diversos modelos são treinados em subconjuntos diferentes do conjunto de dados de treinamento, selecionados por amostragem com reposição. As previsões dos modelos individuais são combinadas usando média (para regressão) ou votação (para classificação).

Boosting: O boosting treina uma sequência de modelos de forma iterativa, onde cada modelo subsequente corrige os erros do anterior. Instâncias de treinamento recebem pesos variáveis ao longo das iterações, dando maior importância às instâncias mal classificadas. As previsões dos modelos são ponderadas conforme sua performance.

Random Forest: Esta técnica envolve a combinação de várias árvores de decisão em um ensemble. Cada árvore é treinada com diferentes subconjuntos do conjunto de dados e características selecionadas aleatoriamente. As previsões das árvores são combinadas por votação.

Stacking: Como explicado anteriormente, essa técnica combina as previsões de vários modelos base com os rótulos reais correspondentes para criar um conjunto de dados empilhado. Um modelo de meta-aprendizado, então, é treinado para realizar a previsão final.

O objetivo do ensemble learning é criar um modelo final mais eficaz, capaz de oferecer previsões mais precisas e mostrar maior resistência ao overfitting. Combinando as capacidades de vários modelos, o ensemble learning compensa as fraquezas de um modelo com as forças de outro, elevando o desempenho geral do sistema de aprendizado de máquina.

HEURISTIC FUNCTION Uma função heurística é uma técnica projetada para resolver problemas de forma mais eficiente quando métodos tradicionais são muito lentos ou incapazes de encontrar uma solução exata. Simplificando um problema de busca, ela pode estimar o custo ou a distância até o objetivo, ajudando o algoritmo a escolher qual caminho seguir no espaço de busca. Por exemplo, em busca de caminhos, a função heurística pode ser a distância em linha reta até o objetivo. Embora não forneça o custo exato, oferece uma boa estimativa para guiar a busca. Escolher uma função heurística adequada é crucial, pois uma escolha ruim pode prejudicar o desempenho do algoritmo.

ÓTIMO LOCAL Uma solução que é a melhor em uma área imediata, mas não necessariamente a melhor no contexto geral. Por exemplo, estar no topo de uma colina faz dela o ponto mais alto localmente, mas pode haver uma montanha mais alta em outro lugar. Em aprendizado de máquina, um algoritmo pode encontrar uma solução ótima para os dados de treinamento, mas essa solução pode não ser a melhor para novos dados. Uma solução que generaliza melhor poderia ser considerada o ótimo global.

ÓTIMO GLOBAL É a melhor solução possível para um problema. Imagine procurar o ponto mais alto em paisagens montanhosas: o ótimo global seria o topo da montanha mais alta. Em aprendizado de máquina, um algoritmo ajusta os dados de treinamento para minimizar o erro, resultando em um ótimo local. Contudo, essa solução pode não se adequar bem a novos dados. Uma solução que talvez não se ajuste tão bem aos dados de treinamento, mas generalize melhor para novos dados, seria o ótimo global. O desafio é encontrar o ótimo

global entre muitos ótimos locais, especialmente em problemas complexos com grandes espaços de soluções.

ARTIFICIAL NEURAL NETWORK (ANN)

Uma Rede Neural Artificial (RNA) é um modelo computacional que imita a estrutura do cérebro, capaz de aprendizado de máquina e reconhecimento de padrões. Consiste em camadas de nós: entrada, ocultas e saída. Cada nó tem um peso e um limite; se o limite for superado, o nó ativa e envia dados para a próxima camada. As RNAs aprendem com dados de treinamento, aprimorando sua precisão ao longo do tempo. Elas são cruciais no aprendizado profundo, resolvendo tarefas complexas como visão computacional e reconhecimento de voz. Cada nó atua como um modelo de regressão linear, processando entradas ponderadas, somando-as e passando por uma função de ativação que determina a saída. Se a saída exceder um nível específico, o nó ativará, enviando dados para a próxima camada.

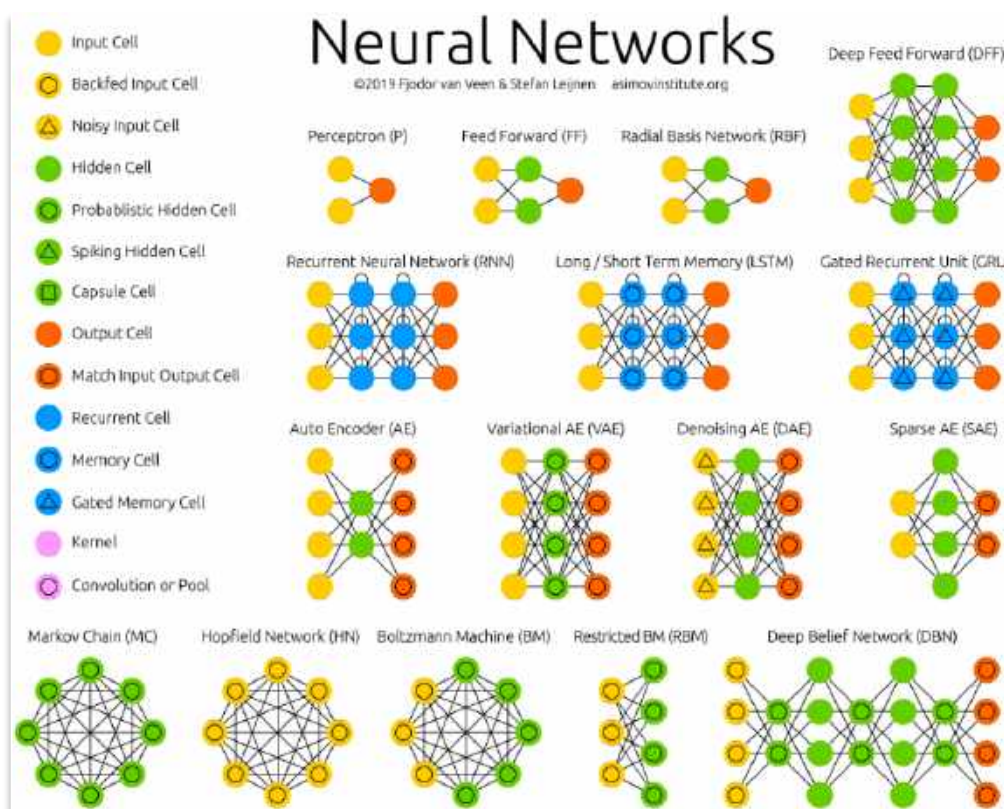


Figura 67: Neural Networks
Fonte: Instituto Asimov

COMO UMA REDE NEURAL ARTIFICIAL APRENDE

Uma Rede Neural Artificial (RNA) é um sistema de processamento de informações que compartilha características com as redes neurais biológicas. Consiste em neurônios artificiais interconectados por sinapses artificiais ponderadas. Cada neurônio multiplica sua entrada pelo peso da sinapse correspondente, e a soma ponderada passa por uma função de ativação para gerar a saída, que pode ser enviada a outros neurônios. O aprendizado em uma RNA ocorre ajustando iterativamente os pesos das sinapses até que a rede encontre uma solução generalizada para um problema.

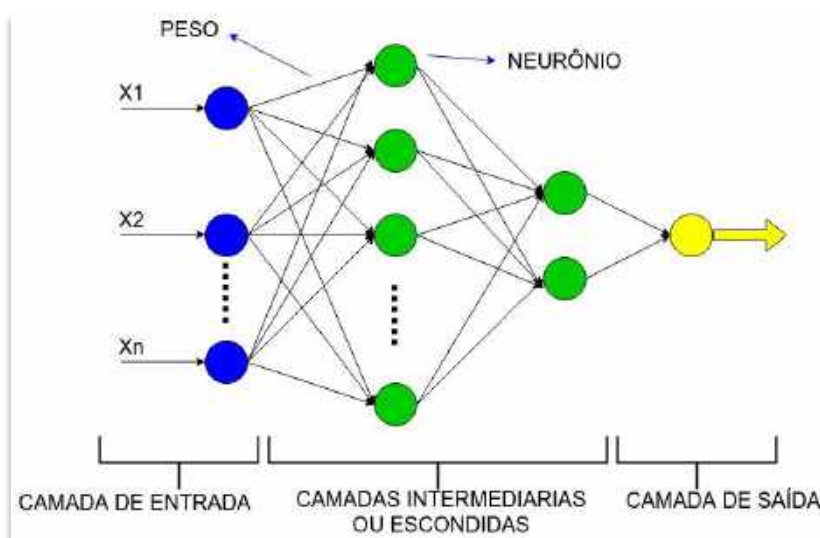


Figura 68: Como a Rede Neural Aprende
Fonte: Álvaro Pinheiro

LÓGICA DE UMA REDE NEURAL ARTIFICIAL

Retropropagação é um algoritmo essencial no treinamento de redes neurais, usado para calcular o gradiente da função de custo em relação aos pesos e vieses da rede. Ele ajusta esses parâmetros para minimizar o erro seguindo estes passos: (1) feedforward, onde a entrada é processada pela rede até a saída final e as ativações são armazenadas; (2) erro, calculado comparando a saída da rede com a desejada; (3) backward pass, onde o erro é propagado de volta pela rede, calculando o gradiente do erro utilizando a regra da cadeia; (4) atualização de pesos e vieses, ajustados proporcionalmente ao negativo do gradiente,

geralmente usando gradiente descendente. Repetir esse processo por várias épocas permite que a rede aprenda padrões complexos e minimize erros nas previsões.

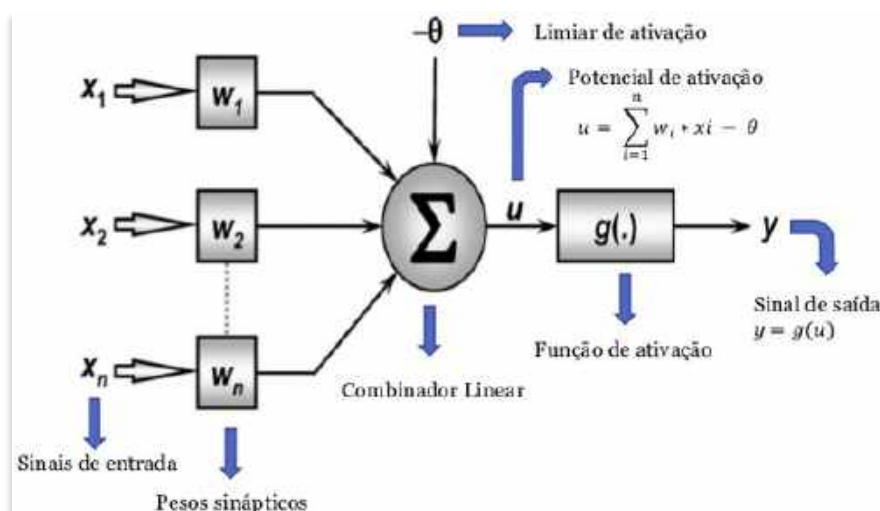


Figura 69: Esquema Lógico de uma Rede Neural Artificial
Fonte: Álvaro Pinheiro

Bias é um termo empregado em redes neurais artificiais para designar um valor constante adicionado à soma ponderada das entradas de um neurônio. Sua função é ajustar o ponto onde a função de ativação é acionada, isto é, o limiar de decisão do neurônio. Além disso, bias permite que a rede aprenda funções que não passam pela origem, aumentando assim a flexibilidade do modelo. Pode ser entendido como uma entrada adicional com valor 1 que possui um peso associado.

Os pesos numa rede neural são valores numéricos ligados às conexões entre neurônios de diferentes camadas. Eles são essenciais para o funcionamento e aprendizado da rede, pois determinam a influência de um neurônio sobre outro. Os pesos passam por: (1) multiplicação pelos valores de entrada, (2) soma ponderada nas entradas para a função de ativação, (3) ajuste durante o treinamento para minimizar erros usando algoritmos como backpropagation, e (4) transmissão de informação, facilitando o aprendizado de padrões complexos.

Os pesos são definidos aleatoriamente antes do início do treinamento e ajustados de forma iterativa para otimizar o desempenho da rede. Embora não sejam exatamente

parâmetros do otimizador, a escolha da inicialização dos pesos é crucial, pois ajuda a rede a aprender os parâmetros necessários para previsões ou classificações precisas. Métodos de inicialização como Xavier ou Glorot são frequentemente utilizados para assegurar que os pesos iniciais da rede sejam adequados. A técnica Xavier é particularmente útil para diminuir problemas durante o treinamento de redes neurais profundas, garantindo que a variância das entradas corresponda à variância das saídas.

A taxa de aprendizagem é um parâmetro crucial, pois determina os passos do algoritmo de otimização ao longo do gradiente. Taxas muito baixas resultam em convergência lenta, enquanto taxas altas podem causar oscilações ou divergência no treinamento. Ela controla a rapidez com que os pesos das conexões entre neurônios são ajustados. A escolha ideal da taxa de aprendizagem depende do problema, dos dados e do modelo. Existem métodos, como o decaimento exponencial, para ajustar essa taxa durante o treinamento.

A "perda" em uma rede neural é uma métrica que indica o desempenho de um modelo após cada iteração do processo de otimização. Um valor mais baixo da perda geralmente indica um modelo melhor (a menos que haja superaste aos dados de treinamento). A perda é calculada tanto no conjunto de treinamento quanto no de validação, e sua interpretação reflete o desempenho do modelo nesses dois conjuntos.

A função de perda em uma rede neural serve para medir o quão bem o modelo se adapta aos dados de treinamento e sua capacidade de generalização para novos dados. A função de perda faz isso ao comparar as saídas previstas pelo modelo com as saídas reais esperadas, gerando um valor numérico que indica o erro ou custo do modelo. O objetivo durante o treinamento de uma rede neural é minimizar a função de perda, ou seja, encontrar os pesos e vieses que resultem na menor diferença possível entre as saídas previstas e as reais. Existem diferentes tipos de funções de perda, dependendo do problema e do tipo de saída da rede neural. Por exemplo, para problemas de regressão, uma função de perda comum é o erro quadrático médio (MSE), que calcula a média dos quadrados das diferenças entre as saídas previstas e as reais. Já para problemas de classificação, uma função de perda frequentemente

utilizada é a entropia cruzada (CE), que mede a discrepância entre as probabilidades atribuídas pelo modelo às classes corretas e incorretas.

O erro em uma rede neural artificial é a diferença entre a saída prevista pelo modelo e a saída real ou desejada. Durante o treinamento de uma rede neural, o objetivo é minimizar esse erro. Isso é feito ajustando os pesos e vieses da rede para produzir uma saída que esteja o mais próximo possível da saída desejada. O erro é calculado usando uma função de perda, que mede a diferença entre a saída prevista e a real. Por exemplo, para problemas de regressão, uma função de perda comum é o erro quadrático médio (MSE), que calcula a média dos quadrados das diferenças entre as saídas previstas e as reais. Se o erro for maior do que um valor aceitável, a rede calcula o erro e propaga a correção para as demais camadas internas até a entrada, ajustando os seus pesos sinápticos. Esse processo é conhecido como retropropagação do erro. Portanto, o erro em uma rede neural artificial é uma medida crucial para entender quão bem o modelo está aprendendo e se ajustando aos dados.

A regularização é uma técnica utilizada para prevenir o overfitting e Underfitting em modelos de aprendizado de máquina, através da adição de um termo de penalidade à função de perda. O overfitting acontece quando um modelo se ajusta excessivamente aos dados de treinamento e não consegue generalizar bem para novos dados. A regularização diminui a complexidade do modelo ao reduzir os coeficientes dos recursos a zero ou próximos de zero. Existem três tipos comuns de regularização no aprendizado de máquina:

Regularização L1 ou Lasso Regression: adiciona o valor absoluto dos coeficientes como um termo de penalidade à função de perda.

Regularização L2 ou Ridge Regression: incorpora o quadrado dos coeficientes como termo de penalidade à função de perda.

Regularização elástica líquida: combina as regularizações L1 e L2, adicionando ambos os valores absolutos e quadrados dos coeficientes como termos de penalidade à função de perda.

Overfitting é um termo usado em estatística e aprendizado de máquina para descrever uma situação em que um modelo corresponde muito próximo ou exatamente a um determinado conjunto de dados. Isso normalmente acontece quando o modelo aprende tão bem os dados de treinamento que não tem um bom desempenho nos dados de teste. Em outras palavras, o modelo aprendeu o ruído dos dados de treinamento junto com o padrão subjacente, fazendo com que ele se ajustasse mal a dados novos.

Muitas vezes, isso é resultado de um modelo excessivamente complexo com muitos parâmetros, o que permite ao modelo “memorizar” os dados de treinamento em vez de “aprender” o verdadeiro padrão. Como resultado, tal modelo tem um desempenho excepcionalmente bom em dados de treinamento, mas fraco em dados novos.

Para evitar sobreajuste: use modelos mais simples com menos parâmetros; use técnicas como validação cruzada; colete mais dados de treinamento, se possível; e, lembre-se de que o objetivo principal do aprendizado de máquina é construir modelos que generalizem bem para dados novos. Portanto, é importante encontrar um equilíbrio entre o Underfitting (modelo muito simples para capturar padrões subjacentes) e o overfitting (modelo muito complexo que captura ruído).

Underfitting é um termo utilizado em estatística e aprendizado de máquina para descrever uma situação em que um modelo é demasiado simples para capturar a estrutura subjacente dos dados. Isso normalmente acontece quando o modelo não é complexo o suficiente para aprender com os dados de treinamento e, portanto, tem um desempenho insatisfatório tanto nos dados de treinamento quanto nos dados de teste.

Em termos práticos, o Underfitting ocorre quando um modelo não consegue encontrar uma relação satisfatória entre características e variável alvo, mesmo na fase de treinamento. Como resultado, tal modelo tem um desempenho insatisfatório tanto em dados de treinamento quanto em dados novos.

Para evitar Underfitting: use modelos mais complexos; adicione mais features; reduza a quantidade de regularização; e, lembre-se de que o objetivo principal do aprendizado de

máquina é construir modelos que generalizem bem para dados novos. Portanto, é importante encontrar um equilíbrio entre o Overfitting e o overfitting.

O momentum é um valor entre 0 e 1 que controla a influência dos gradientes anteriores no ajuste dos pesos. Ele ajuda a acelerar o treinamento e suavizar as oscilações.

Uma função de ativação é uma função usada em redes neurais artificiais que transforma a soma ponderada da entrada em uma saída de um nó ou nós em uma camada da rede. A função de ativação define como o nó é ativado ou não, dependendo do valor da entrada. A função de ativação também normaliza a saída do nó para um intervalo definido, geralmente entre 0 e 1 ou -1 e 1. Existem muitos tipos de funções de ativação, como sigmóide, tangente hiperbólica, ReLU etc. A escolha da função de ativação tem um grande impacto no desempenho e na capacidade da rede neural.

Um otimizador de uma rede neural artificial é uma função que implementa os algoritmos de descida do gradiente e de retropropagação para ajustar os pesos dos neurônios e minimizar a função de perda. O otimizador define como o modelo aprende e converge para uma solução ótima. Existem vários tipos de otimizadores, como SGD, Adam, BFGS, etc. Cada um tem suas vantagens e desvantagens, dependendo do problema e dos dados.

PERCEPTRON

O Perceptron, um modelo matemático inspirado em um neurônio biológico simplificado, foi desenvolvido por Frank Rosenblatt nos anos 1950 e 1960, influenciado pelos trabalhos de Warren McCulloch e Walter Pitts. Considerado a forma mais simples de rede neural, recebe várias entradas e produz uma saída binária. Rosenblatt introduziu pesos para expressar a importância das entradas e determinou a saída pela soma ponderada comparada a um valor limiar. O Perceptron é usado para decisões ao avaliar evidências e exemplifica classificadores binários, sendo parte essencial das redes neurais.

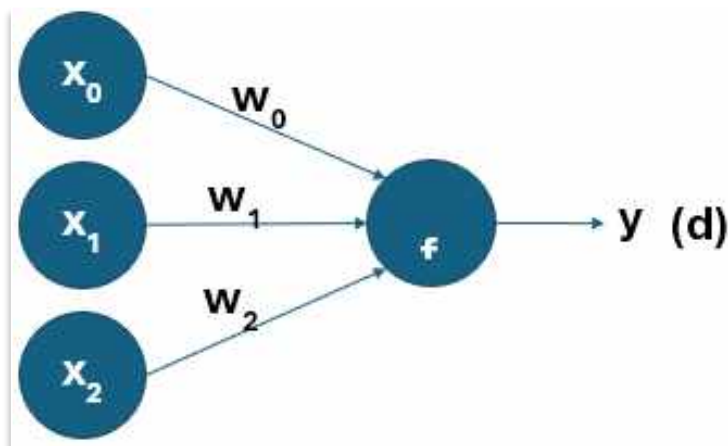


Figura 70: Perceptron
Fonte: Álvaro Pinheiro

MULTILAYER PERCEPTRON (MLP)

O Perceptron Multicamadas (MLP) é uma rede neural com uma ou mais camadas ocultas de neurônios. Essas camadas são chamadas de ocultas porque suas saídas não podem ser previstas diretamente. O algoritmo de aprendizado da MLP, retropropagação, segue estes passos: 1. Inicialização, definindo pesos e limites aleatórios; 2. Ativação, calculando valores dos neurônios das camadas oculta e de saída; 3. Treinamento, ajustando os pesos com base nos erros dos neurônios; 4. Iteração, repetindo até que o erro seja aceitável. O MLP, uma generalização do Perceptron, é progressivo, pois as saídas de uma camada só se conectam às entradas da próxima. O número de neurônios na camada de entrada depende da dimensionalidade do espaço de observação e, na camada de saída, da resposta desejada. Projetar um MLP envolve determinar o número de camadas ocultas, de neurônios em cada camada e especificar os pesos sinápticos entre os neurônios. O MLP é útil para tarefas complexas como predição de voz, análise de dados sociais e visão computacional.

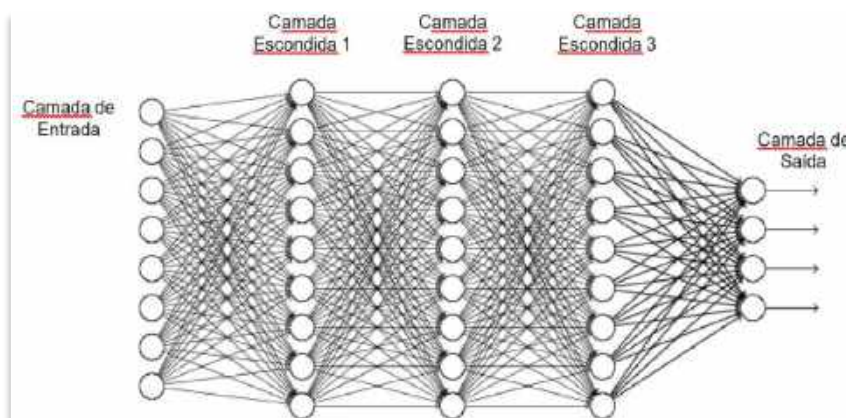


Figura 71: MultiLayer Perceptron
Fonte: Álvaro Pinheiro

CONVOLUTIONAL NEURAL NETWORK (CNN)

As Redes Neurais Convolucionais (CNNs) são modelos de deep learning amplamente utilizados para processamento de imagens. Um modelo CNN básico possui três camadas principais:

Camada Convolucional: Extrai recursos de alto nível dos dados de entrada e cria mapas de recursos.

Camada de Pooling: Reduz as dimensões dos dados aplicada aos mapas de recursos para criar mapas condensados.

Camada Totalmente Conectada: Realiza a classificação, usando a função softmax para calcular as probabilidades de cada classe.

As CNNs se tornaram populares em tarefas de classificação de imagens e aplicações na saúde, sendo uma técnica confiável para previsão automatizada, pois extraem automaticamente recursos úteis das entradas fornecidas.

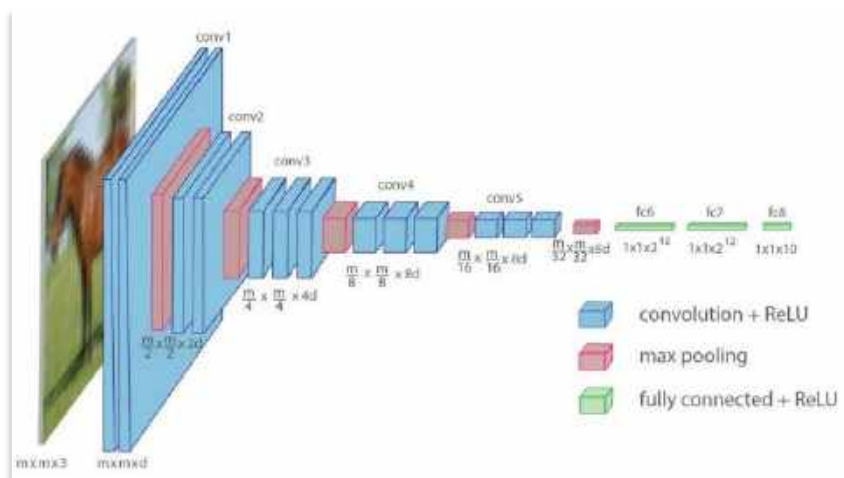


Figura 72: Convolutional Neural Network
Fonte: Álvaro Pinheiro

REFERÊNCIAS

AMARAL, Fernando. **Introdução à ciência de dados: mineração de dados e big data**. Rio de Janeiro: Alta Books, 2016.

BOSTROM, Nick. **Superinteligência: caminhos, perigos, estratégias**. Tradução de Clóvis Marques. Rio de Janeiro: Intrínseca, 2015.

CASTRO, Leandro Nunes De. **Computação Natural - uma Jornada Ilustrada**. São Paulo: Livraria da Física, 2010.

DAVIDOWITZ, Stephens. **Todo Mundo Mente**. 1. ed. Rio de Janeiro: Alta Book, 2018.

FISHER, Max. **Homo Futurus**. São Paulo: Livraria da Física, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data Mining**. Rio de Janeiro: Elsevier Brasil, 2015.

HARARI, Yuval Noah. **21 lições para o século 21**. Tradução de Paulo Geiger. São Paulo: Companhia das Letras, 2018.

KAISER, Brittany. **Manipulados**. 1. ed. São Paulo: HarperCollins Brasil, 2020.

LEE, Kai-Fu. **Inteligência artificial: como os robôs estão mudando o mundo, a forma como amamos, nos relacionamos, trabalhamos e vivemos**. 1. ed. Rio de Janeiro: Globo Livros, 2019.

LINDEN, Ricardo. **Algoritmos genéticos**. Rio de Janeiro: Ciência Moderna, 2008.

O'NEIL, Cathy. **Algoritmos de Destruição em Massa**. 1. ed. São Paulo: Editora Rua do Sabão, 2021.

PINHEIRO, Á. F. **Ambiente de Inteligência Artificial e Computacional para Camada de Serviços no Setor Público**. 2023. Tese (Doutorado) - Universidade de Pernambuco. Disponível em: <https://doi.org/10.5281/zenodo.8000503>.

PINHEIRO, Á. F.; GUIMARAES, R. **Hyperautomation in Government Digital Transformation**. *Journal of Mathematical Techniques and Computational Mathematics*, v. 2, n. 7, p. 330-336, 2023. Disponível em: <https://doi.org/10.33140/JMTCM>.

PINHEIRO, Á. F.; LIMA NETO, F. B. de. **Use of Domain Engineering in Hyperautomation Applied to Decision Making in Government**. *Journal of Advances in Artificial Intelligence*, v. 1, p. 103-116, 2023. Disponível em: <https://doi.org/10.18178/JAAI>.

PINHEIRO, Á. F.; LIMA NETO, F. B. de. **Use of Machine Learning for Active Public Debt Collection with Recommendation for the Method of Collection Via Protest.** In: 8th International Conference on Data Mining and Applications (DMA 2022), Vancouver, Canadá, 2022. p. 99-108. Disponível em: <https://doi.org/10.5121/csit.2022.120909>.

PINHEIRO, Á. F.; SANTOS, W. B. **Prediction of active debt in the State of Pernambuco, Brazil.** *Revista de Engenharia da Universidade de Pernambuco*, 2020. Disponível em: <https://doi.org/10.25286/rep.v5i1>.

PINHEIRO, Á. F.; SANTOS, W. B. **Prioritization and transparency in software development: an action research in public administration.** In: Conference Brazilian Workshop on Social Network Analysis and Mining, 2021. Disponível em: <https://doi.org/10.5753/brasnam.2021.16147>.

PINHEIRO, Á. F.; SANTOS, W. B.; LIMA NETO, F. B. de. **Intelligent Framework to Support Technology and Business Specialists in the Public Sector.** *IEEE Access*, v. 11, p. 15655-15679, 2023. Disponível em: <https://doi.org/10.1109/ACCESS.2023.3243195>.

PINHEIRO, Santiago. **Estatística Básica: a arte de trabalhar com dados.** 2. ed. São Paulo: GEN LTC, 2015.

PROVOST, Foster; FAWCETT, Tom. **Data Science para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados.** 1. ed. Rio de Janeiro: Alta Books, 2016.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial.** 3. ed. Rio de Janeiro: Elsevier, 2013.

SCHWAB, Klaus. **A quarta revolução industrial.** São Paulo: Edipro, 2016.

SIEGEL, Eric. **Análise Preditiva: O poder de prever quem vai clicar, comprar, mentir ou morrer.** Rio de Janeiro: Alta Books, 2018.

SILVA, Ivan Nunes da; SPATTI, Danilo Hernane; FLAUZINO, Rogério Andrade. **Redes neurais artificiais para engenharia e ciências aplicadas.** São Paulo: Artliber Editora, 2010.

SUMPTER, David. **Dominados pelos números.** 1. ed. São Paulo: Bertrand, 2019.

SOBRE O AUTOR

Álvaro Farias Pinheiro

É Analista de Gestão de Tecnologia da Informação e Comunicação (AGTIC) da Agência Estadual de Tecnologia da Informação e Comunicação do Estado de Pernambuco (ATI/PE). Coordenador de Sistemas, Automação Digital e Inovação da Procuradoria Geral do Estado de Pernambuco (PGE/PE). Doutor em Engenharia da Computação com ênfase em Inteligência Computacional pela Escola Politécnica de Pernambuco da Universidade de Pernambuco (POLI/UPE). Mestre em Engenharia de Software pelo Centro de Estudos e Sistemas Avançados do Recife (CESAR). MBA em Inteligência Artificial com ênfase em Marketing Digital pela Unyleya College (UNYLEYA). Especialização Lato Sensu em Governo Digital pela Faculdade Verbo Jurídico (VERBO) e Metodologias de Desenvolvimento em Engenharia de Software pela União Brasileira de Tecnologia (UNIBRATEC). Bacharel em Sistemas de Informação com ênfase em Engenharia de Software pela Faculdade Integrada do Recife (FIR). Para mais detalhes, acesse <http://www.alvarofpinheiro.eti.br>.