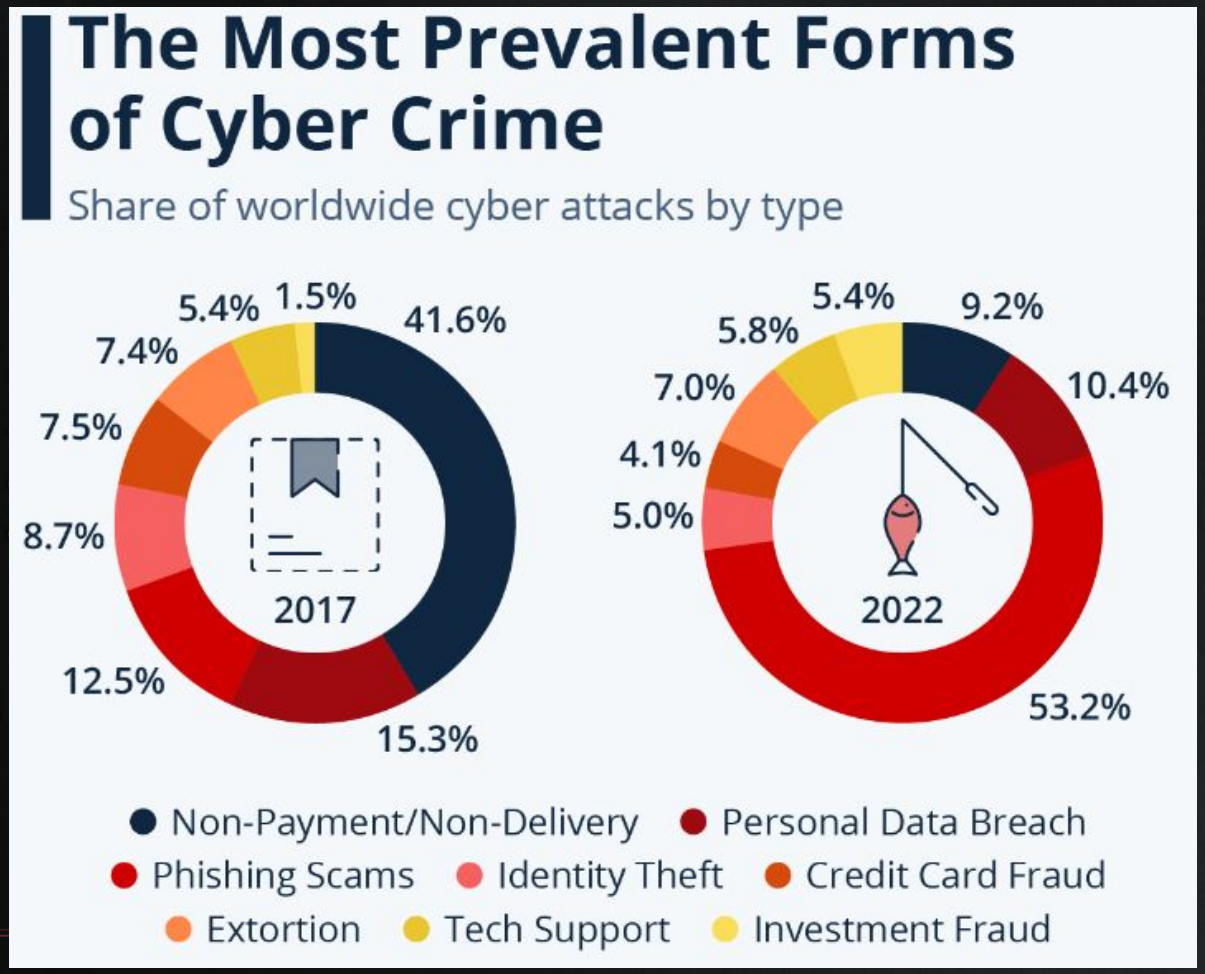
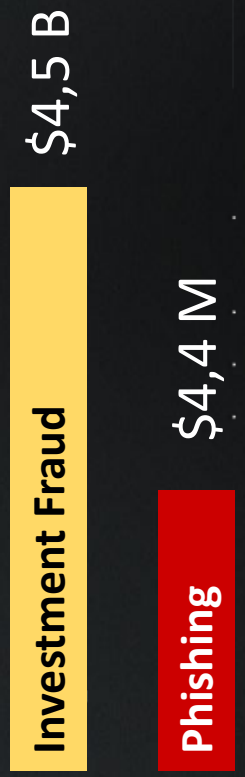

FIAP
LET'S
ROCK
THE
FUTURE

IA & MACHINE LEARNING

[André Marques]
pf2165@fiap.com.br

MACHINE LEARNING FOR CYBERSECURITY





statista.com fieldeffect.com ic3.gov

USA, 2023

MACHINE LEARNING FOR CYBERSECURITY

Front Running:

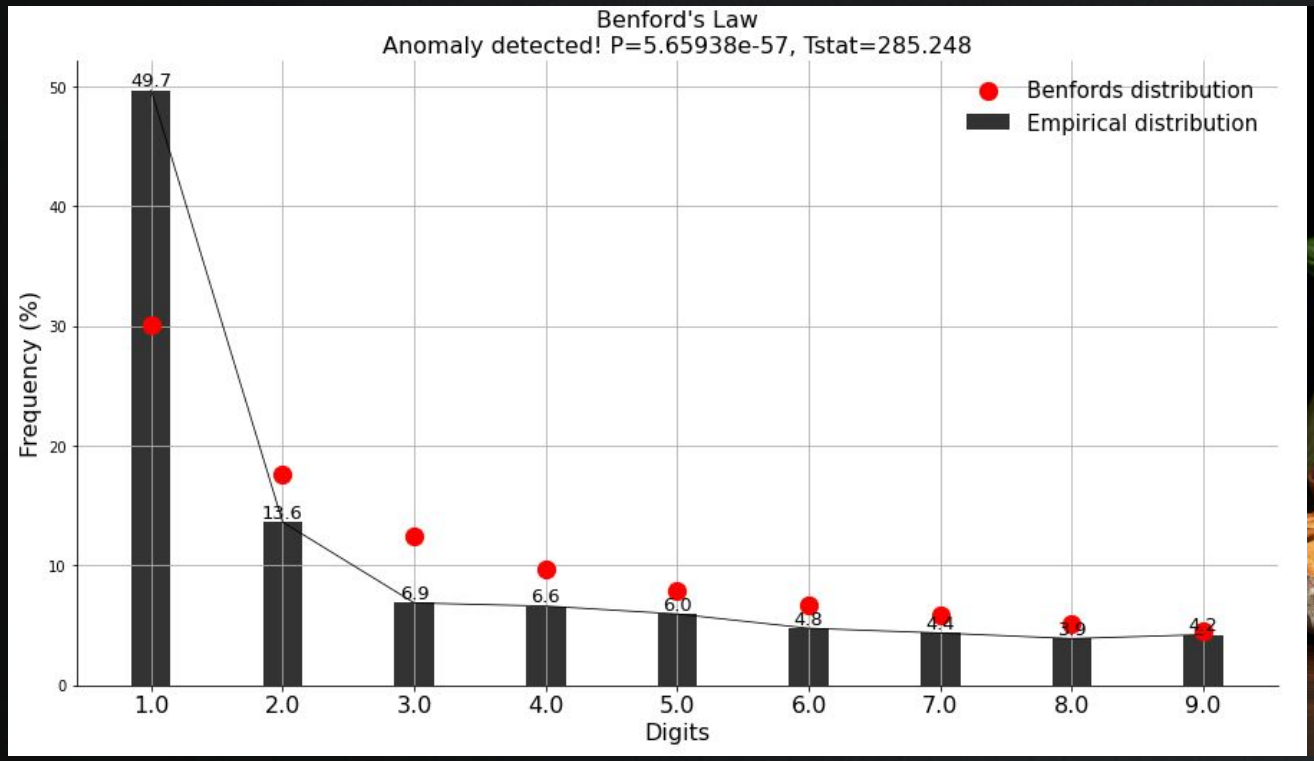
A prática configura o crime de Front Running, quando um operador financeiro antecipa a um investidor que irá realizar uma grande operação, capaz de influenciar no preço de mercado de um ativo. Desse modo, essa obtenção de lucro configura um conflito de interesses por meio do uso de informação privilegiada.

Agência Gov

YouTube/CNN - 07/Ago./2024

BENFORD LAW

Aplicação da lei de Benford para identificar fraudes (acesso por robôs) em visitas a sites.



Github.com

ESTATÍSTICA

Estatística é a ciência que se ocupa da coleta, organização, análise, interpretação e apresentação de dados. Ela fornece as ferramentas e métodos para lidar com a incerteza ao fazer inferências sobre populações a partir de amostras e ao tomar decisões informadas com base em dados observacionais ou experimentais.

População

Conjunto de todas as unidades que são de interesse em um certo estudo.

Amostra

Qualquer subconjunto da população selecionado de acordo com certas regras.

Censo

Estudo que inclui todos os elementos da população.

ESTATÍSTICA

É uma ciência baseada na Teoria da Probabilidade, cujo objetivo principal é nos auxiliar a tomar decisões ou tirar conclusões em situações de incerteza, a partir de dados.

ESTATÍSTICA EM SEGURANÇA CIBERNÉTICA

Na área de Segurança Cibernética, a estatística desempenha um papel crucial ao fornecer as bases para a análise de dados relacionados a ameaças, ataques e a defesa de sistemas.

Análise de Dados e Detecção de Ameaças

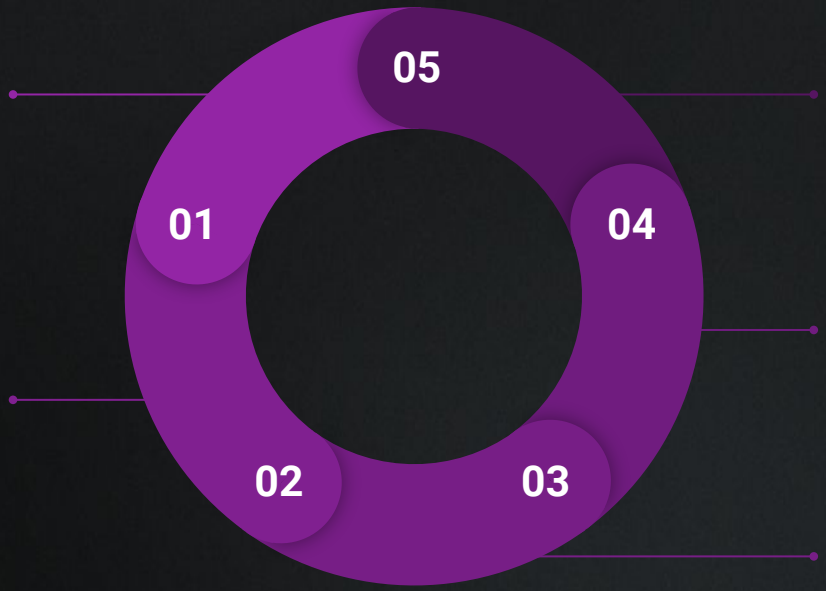
Modelagem Probabilística para identificar padrões de tráfego de rede.

Testes de Hipóteses para identificar se uma atividade observada em uma rede é anômala ou parte de um comportamento normal.

Análise de Riscos e Vulnerabilidades

Análise de Frequência e Gravidade de ataques cibernéticos podem ser usadas para priorizar a alocação de recursos de segurança.

Modelos de Risco, como análise de risco quantitativa, podem ser utilizadas para estimar a probabilidade e impacto de diferentes tipos de ataques



Machine Learning e Estatística

Muitos algoritmos de aprendizado de máquina usados em segurança cibernética, como classificadores, dependem de técnicas estatísticas para modelar a relação entre as características dos dados e o resultado esperados, como a classificação de um e-mail como spam.

Criação de Perfis e Previsão de Comportamento

Estatística Descritiva como médias, medianas, desvio padrão, e percentis são usadas para criar perfis de comportamento normal de usuários ou sistemas. Esses perfis ajudam a identificar comportamentos anômalos que podem ser indicativos de um ataque.

Análise Forense

Análise Estatística de Logs, a estatística é utilizada para analisar grandes volumes de logs, identificando padrões e sequências que possam indicar como um ataque ocorreu.

Inferência Estatística é utilizada para tirar conclusões sobre a origem ou natureza de uma brecha de segurança, baseada em amostras de dados coletados durante a análise forense..

MEDIDAS DE TENDÊNCIA CENTRAL

Em estatística, as medidas de tendência central servem para identificar um valor que representa o centro de um conjunto de dados.

Média

A média de um conjunto de dados é a soma das entradas de dados dividida pelo número de entradas

MÉDIA POPULACIONAL	MÉDIA AMOSTRAL
$\mu = \frac{\sum x}{N}$	$\overline{x} = \frac{\sum x}{n}$

- μ (pronuncia-se "mi") → representa a média populacional;
- $\overline{x}$ (pronuncia-se "x barra") → representa a média amostral;
- N → representa o número de entradas de uma população;
- n → representa o número de entradas de uma amostra.

- **Número de ataques**: Quantas vezes um sistema foi atacado em um determinado período?
- **Tempo de detecção de ameaças**: Quanto tempo leva para identificar uma nova ameaça?
- **Taxa de sucesso de ataques**: Qual a porcentagem de ataques que tiveram sucesso?
- **Custo de incidentes**: Qual o valor médio gasto para remediar um incidente de segurança?

MEDIDAS DE TENDÊNCIA CENTRAL

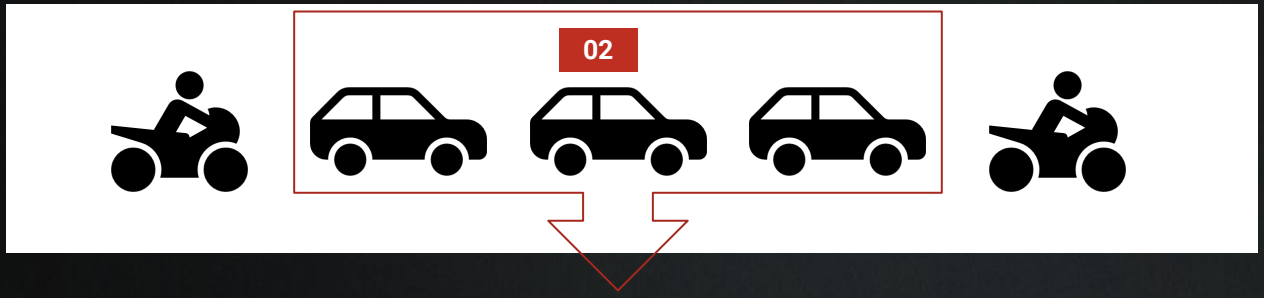
É um valor que representa uma entrada central do conjunto de dados.

01	**Média**	Representa a soma de todos os valores dividida pelo número total de valores. É útil para entender o valor médio de uma métrica, como o custo médio por incidente.
02	**Mediana**	É o valor que divide um conjunto de dados ordenados em duas partes iguais. É útil quando há valores extremos que podem distorcer a média.
03	**Moda**	É o valor que ocorre com maior frequência em um conjunto de dados. É útil para identificar o tipo de ataque mais comum.

MEDIDAS DE TENDÊNCIA CENTRAL

É um valor que representa uma entrada central do conjunto de dados.

01	**Média**	Representa a soma de todos os valores dividida pelo número total de valores. É útil para entender o valor médio de uma métrica, como o custo médio por incidente.
02	**Mediana**	É o valor que divide um conjunto de dados ordenados em duas partes iguais. É útil quando há valores extremos que podem distorcer a média.
03	**Moda**	É o valor que ocorre com maior frequência em um conjunto de dados. É útil para identificar o tipo de ataque mais comum.

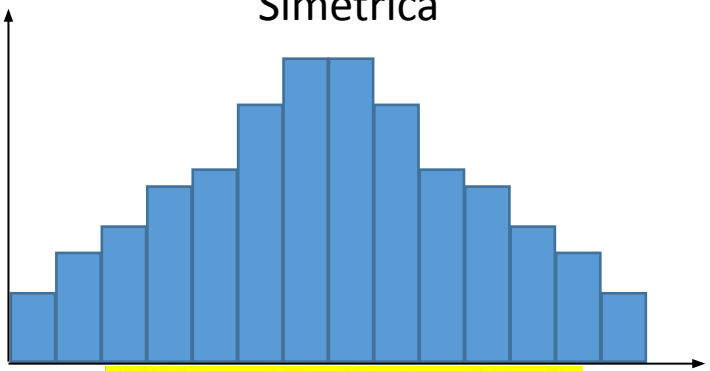


Moda: Carro

MEDIDAS DE TENDÊNCIA CENTRAL

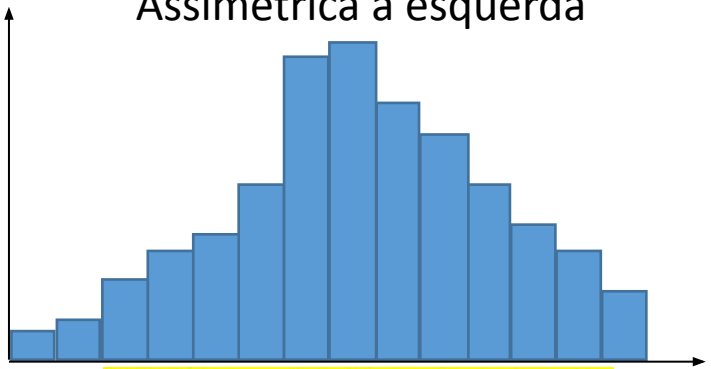
Forma das distribuições

Simétrica



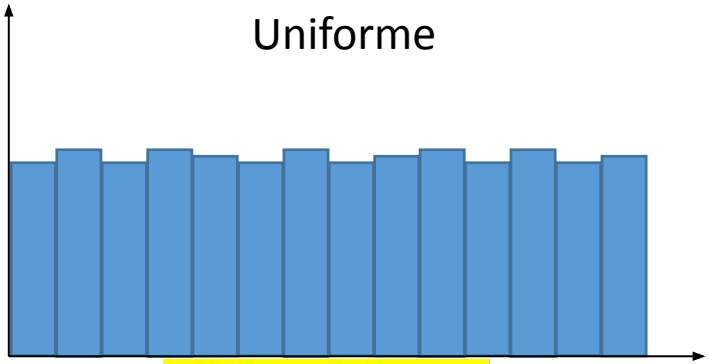
$Média = Mediana = Moda$

Assimétrica à esquerda



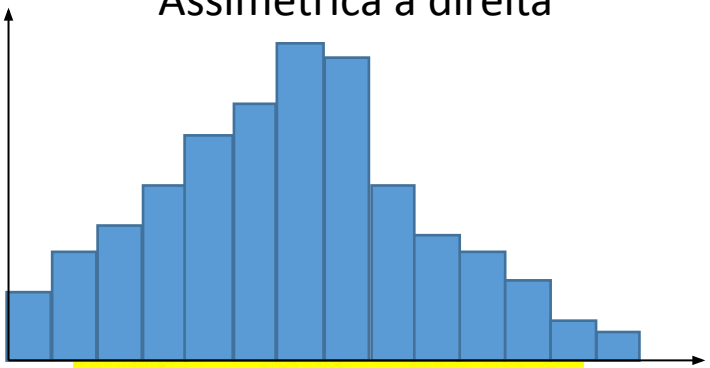
$(Média = Mediana) < Moda$

Uniforme



$Média = Mediana$

Assimétrica à direita



$(Média = Mediana) > Moda$

MEDIDAS DE VARIAÇÃO OU DISPERSÃO

Desvio (DEV), Variância (VAR) e Desvio Padrão (D.P.)

Desvio

O desvio de uma entrada x em uma população é a diferença entre a entrada e a média μ do conjunto de dados.

$$DEV = \mu - Xi$$

Funcionários (X)	Salários
X1	R$ 4.100,00
X2	R$ 3.800,00
X3	R$ 3.900,00
X4	R$ 4.500,00
X5	R$ 4.700,00
X6	R$ 4.100,00
X7	R$ 4.400,00
X8	R$ 4.100,00
X9	R$ 3.700,00
X10	R$ 4.200,00
Σ	R$ 41.500,00

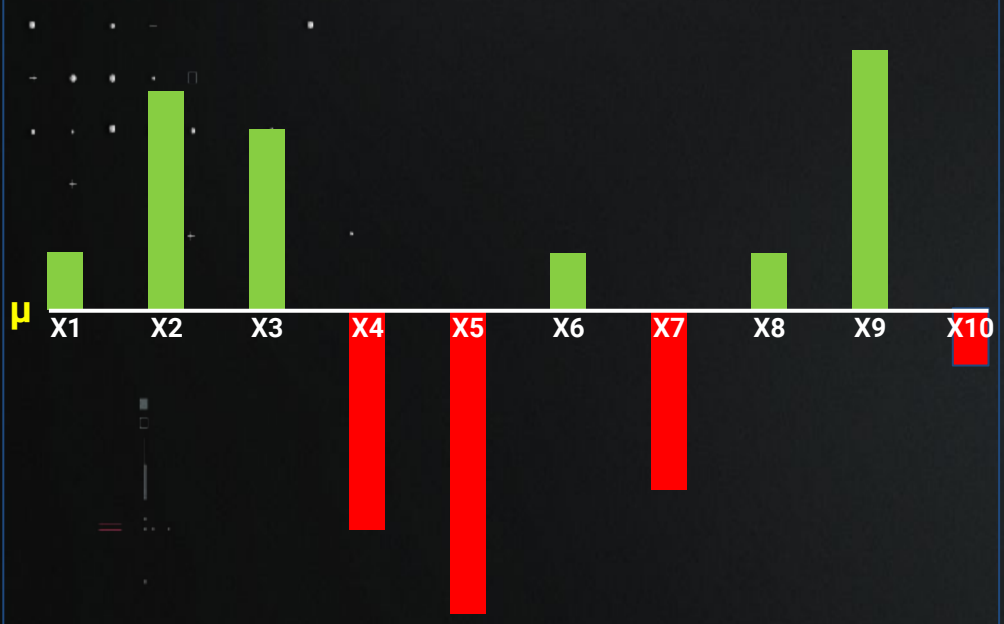
MEDIDAS DE VARIAÇÃO OU DISPERSÃO

Desvio, Variância e Desvio Padrão

$$\mu = \frac{4100 + 3800 + \ldots + 4200}{10} = 4150$$

Desvio

O desvio de uma entrada x em uma população é a diferença entre a entrada e a média μ do conjunto de dados.



Funcionários (X)	Salários	Desvio (μ-x)
X1	R$ 4.100,00	4150 - 4100 = 50
X2	R$ 3.800,00	4150 - 3800 = 350
X3	R$ 3.900,00	4150 - 3900 = 250
X4	R$ 4.500,00	4150 - 4500 = -350
X5	R$ 4.700,00	4150 - 4700 = -550
X6	R$ 4.100,00	4150 - 4100 = 50
X7	R$ 4.400,00	4150 - 4400 = -250
X8	R$ 4.100,00	4150 - 4100 = 50
X9	R$ 3.700,00	4150 - 3700 = 450
X10	R$ 4.200,00	4150 - 4200 = -50
Σ	R$ 41.500,00	Σ(μ-x) = 0

MEDIDAS DE VARIAÇÃO OU DISPERSÃO

Desvio, Variância e Desvio Padrão

Variância

A variância populacional do conjunto de dados populacional de N entradas é a soma dos quadrados de cada entrada dividido pelo total de entradas.

$$\sigma^2 = \frac{\Sigma(\mu - x)^2}{N} \Rightarrow \sigma^2 = \frac{885.000}{10}$$

Variância

$\sigma^2 = 88.500$

*Formalmente descrita por **Ronald A. Fisher** (1890–1962) em seu trabalho sobre estatística inferencial e teoria da amostragem.*

Funcionários (X)	Salários	Desvio (μ-x)	Desvio²
X1	R$ 4.100,00	4150 - 4100 = 50	(4150 - 4100)² = 2.500
X2	R$ 3.800,00	4150 - 3800 = 350	(4150 - 3800)² = 122.500
X3	R$ 3.900,00	4150 - 3900 = 250	(4150 - 3900)² = 62.500
X4	R$ 4.500,00	4150 - 4500 = -350	(4150 - 4500)² = 122.500
X5	R$ 4.700,00	4150 - 4700 = -550	(4150 - 4700)² = 302.500
X6	R$ 4.100,00	4150 - 4100 = 50	(4150 - 4100)² = 2.500
X7	R$ 4.400,00	4150 - 4400 = -250	(4150 - 4400)² = 62.500
X8	R$ 4.100,00	4150 - 4100 = 50	(4150 - 4100)² = 122.500
X9	R$ 3.700,00	4150 - 3700 = 450	(4150 - 3700)² = 202.500
X10	R$ 4.200,00	4150 - 4200 = -50	(4150 - 4200)² = 2.500
Σ	R$ 41.500,00	Σ(μ-x) = 0	Σ(μ-x)² = SSx = 885.000

MEDIDAS DE VARIAÇÃO OU DISPERSÃO

Desvio, Variância e Desvio Padrão

Desvio padrão

O desvio padrão populacional de um conjunto de dados populacional de N entradas é a raiz quadrada da variância populacional.

$$\sigma = \sqrt{\frac{\Sigma(\mu - x)^2}{N}} \Rightarrow \sigma = \sqrt{\frac{885.000}{10}}$$

Desvio Padrão

$\sigma \cong 297{,}50$

O desvio padrão, foi formalmente introduzido por Karl Pearson (1857–1936). Pearson foi um dos fundadores da estatística moderna e desenvolveu o conceito de desvio padrão como parte de seu trabalho na criação da teoria da correlação e regressão.

Funcionários (X)	Salários	Desvio (μ-x)	Desvio²
X1	R$ 4.100,00	4150 - 4100 = 50	(4150 - 4100)² = 2.500
X2	R$ 3.800,00	4150 - 3800 = 350	(4150 - 3800)² = 122.500
X3	R$ 3.900,00	4150 - 3900 = 250	(4150 - 3900)² = 62.500
X4	R$ 4.500,00	4150 - 4500 = -350	(4150 - 4500)² = 122.500
X5	R$ 4.700,00	4150 - 4700 = -550	(4150 - 4700)² = 302.500
X6	R$ 4.100,00	4150 - 4100 = 50	(4150 - 4100)² = 2.500
X7	R$ 4.400,00	4150 - 4400 = -250	(4150 - 4400)² = 62.500
X8	R$ 4.100,00	4150 - 4100 = 50	(4150 - 4100)² = 122.500
X9	R$ 3.700,00	4150 - 3700 = 450	(4150 - 3700)² = 202.500
X10	R$ 4.200,00	4150 - 4200 = -50	(4150 - 4200)² = 2.500
Σ	R$ 41.500,00	Σ(μ-x) = 0	Σ(μ-x)² = SSx = 885.000

MEDIDAS DE VARIAÇÃO OU DISPERSÃO

Desvio, Variância e Desvio Padrão

POPULACIONAL

Variância

$$\sigma^2 = \frac{\sum(\mu - x)^2}{N}$$

$$\sigma^2 = \frac{11.025.000}{10}$$

$$\sigma^2 = 1.102.500$$

Desvio Padrão

$$\sigma = \sqrt{\frac{\sum(\mu - x)^2}{N}}$$

$$\sigma = \sqrt{\frac{11.025.000}{10}}$$

$$\sigma = 1050$$

AMOSTRAL

Variância

$$S^2 = \frac{\sum(\bar{x} - x)^2}{n - 1}$$

$$S^2 = \frac{11.025.000}{9}$$

$$S^2 = 1.225.000$$

Desvio Padrão

$$S = \sqrt{\frac{\sum(\bar{x} - x)^2}{n - 1}}$$

$$S = \sqrt{\frac{11.025.000}{9}}$$

$$S \cong 1.106$$

MEDIDAS DE VARIAÇÃO OU DISPERSÃO

Coeficiente de Variação (CV)

Fórmula do Coeficiente de Variação (CV)

$$CV = \left(\frac{\text{Desvio Padrão}}{\text{Média}} \right) \text{x } 100$$

- O CV é uma medida de dispersão relativa que indica a variabilidade dos dados em relação à média.
- Ele é calculado como a razão entre o desvio padrão e a média, multiplicada por 100 para expressá-lo em porcentagem.
- Ele é útil para comparar a variabilidade de diferentes conjuntos de dados com médias e variâncias diferentes.
- Quanto mais alto o CV, maior é a variabilidade relativa em relação à média.

ATIVIDADE

EDA - Exploratory Data Analysis

1. Carregar as bases de dados pro Github
2. Gerar uma amostra aleatória da base para análise:
   - df.sample(n = 1000, random_state = 43)
3. Fazer um EDA da base principal
4. Identificar quais das outras duas bases têm maior potencial de ter dados de phishing.

MACHINE LEARNING ALGORITHMS

Algoritmos de Aprendizado de Máquina são métodos computacionais que permitem que um sistema aprenda a partir de dados e melhore seu desempenho em tarefas específicas ao longo do tempo. Esses algoritmos são usados para identificar padrões, fazer previsões e tomar decisões. No contexto de Cybersecurity, esses algoritmos são usados para detectar e responder a ameaças cibernéticas, automatizando processos de segurança e melhorando a eficiência na proteção contra ataques.

1

Supervisionados

São treinados com um conjunto de dados rotulados. Após o treinamento, pode prever a saída de novos dados não rotulados.

- Máquina de Vetores de Suporte (SVM) para Detecção de Malware.
- Árvore de Decisão para Filtragem de Spam.

2

Não Supervisionados

Trabalham com dados não rotulados e buscam identificar estruturas ocultas ou padrões dentro dos dados.

- K-means ou DBSCAN para Detecção de Anomalias.
- Clustering Hierárquico para Identificar Padrões de Ataque.

3

Aprendizado por Reforço

Aprende a tomar decisões através de um processo de tentativa e erro, recebendo recompensas ou penalidades com base nas ações que tomam.

- Q-Learning para respostas automatizadas a ameaças e ajustar as regras do firewall.
- Deep Q-Network (DQN) aprende estratégias de defesa contra ataques.

4

Redes Neurais

São modelos computacionais inspirados na estrutura do cérebro humano capazes de aprenderem padrões complexos.

- Redes Neurais Convolucionais (CNNs) ou Redes Neurais Recorrentes (RNNs) para grandes volumes de tráfego de rede e identificar padrões que indicam um ataque.

ARTIFICIAL INTELLIGENCE

São sistemas computacionais capazes de simular a inteligência humana, em atividades específicas:

- Aprendizado: adquire informações a partir de dados e experiências.
- Raciocínio: utiliza as informações para tomar decisões e resolver problemas.
- Percepção: interpreta informações, como imagens e sons.
- Linguagem natural: entende e gera comunicação.

01	IA Fraca (ANI)	• Realiza tarefas específicas, como reconhecimento facial ou tradução de idiomas.
02	IA Forte (AGI)	• Entende, aprende e aplicar conhecimento em qualquer tarefa intelectual. Ainda é um conceito teórico e não existe na prática.
03	Superinteligência	• Uma IA hipotética que excede a capacidade intelectual humana em todos os aspectos.

FERRAMENTAS PARA DATA SCIENCE

- Linguagem de Programação Científica

- Ambiente de Desenvolvimento (IDE)

- Repositórios: códigos / arquivos

- Plataforma de comunicação

- Modelos de linguagem de IA

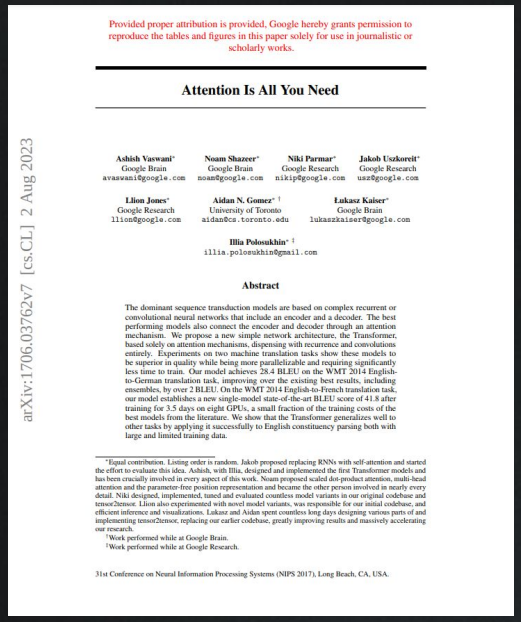
MODELOS DE LINGUAGEM NATURAL

ChatGPT (2022) [OpenAI]	Gemini (2023) [Google DeepMind]	Copilot (2023) [Github & OpenAI]	Maritaca (2022) [ICMC-USP]
Modelo de linguagem treinado para gerar texto de maneira natural e coerente, baseado na arquitetura GPT (Generative Pre-trained Transformer). - Geração de Texto - Tradução - Resumos - Conversação: manter diálogos ... - Matemática e Ciência: resolver equações, explicar conceitos matemáticos e científicos, e criar fórmulas matemáticas.	Modelo de inteligência artificial que integra capacidades de linguagem natural com técnicas avançadas de aprendizado, como o aprendizado por reforço. Oferece funcionalidades em busca, automação de tarefas, e outros serviços baseados em IA.	É um assistente de inteligência artificial generativa que tem suporte para o pacote Microsoft 365, o navegador Edge e o sistema operacional Windows 111. O Copilot é capaz de combinar dados e executar funções nos principais softwares da empresa, como o Word, PowerPoint e Excel.	É uma ferramenta de inteligência artificial desenvolvida no Brasil que é voltada para a interpretação de linguagem natural, com foco em aplicações que envolvem o português. Ele foi criado para facilitar o desenvolvimento de interfaces de voz e texto em projetos que envolvem a interação humano-computador, como assistentes virtuais, chatbots, e outras soluções que requerem processamento de linguagem natural.

MODELOS DE LINGUAGEM NATURAL

Transformer: A Novel Neural Network Architecture for Language Understanding

Google Research/Transformer



Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Vaswani et al. (2017)

É um modelo de IA que foi treinado com um grande conjunto de dados e é capaz de criar novos conteúdos que se assemelham aos dados que foi treinado.

Vantagens:

- Aprendizado eficiente: o pré-treinamento permite acelerar o aprendizado.
- Adaptabilidade: se ajusta a diversas tarefas com quantidades menores de dados rotulados.
- Capacidade de generalização: a grande quantidade de dados permite generalizar melhor para novos exemplos e situações.
- Criação de conteúdo original: abre possibilidades para diversas aplicações criativas e inovadoras.

Modelos pré-treinados

GPT

- Generative Pre-trained Transformer
  - Gerar textos
  - Responder

BERT

- Bidirectional Encoder Representations from Transformers
  - Classificação de textos
  - Responder
  - Análise de sentimentos

DALL-E

- “DALL”: artista Salvador Dalí; “E”: filme WALL-E.
  - Capacidade de gerar imagens criativas e surrealistas a partir de descrições textuais.

MuseNet

- “Gera composições musicais de até 4 minutos, com 10 instrumentos diferentes e estilos que variam desde o country até Mozart e os Beatles.

DICIONÁRIO DE VARIÁVEIS

1. **urls**: urls completas que foram analisadas.
2. **phishing**: indica se uma URL foi classificada como phishing (1) ou não (0).
3. **domain**: extrai o nome de domínio da URL (por exemplo, "techcrunch.com" de "http://techcrunch.com/ ...").
4. **ip**: sinaliza se um endereço IP está presente na URL (1) ou não (0). O uso de endereços IP em URLs pode ser suspeito.
5. **at**: sinaliza se o símbolo "@" está presente na URL (1) ou não (0). O símbolo "@" às vezes é usado em tentativas de phishing para ocultar o destino real.
6. **length_url**: mede o número total de caracteres na URL. URLs longas podem ser um sinal de phishing.
7. **depth_url**: representa o número de níveis ou diretórios no caminho da URL. URLs mais profundas podem ser usadas para ocultar conteúdo malicioso.
8. **double_slash**: sinaliza se barras duplas ("//") estão presentes no caminho da URL (1) ou não (0), o que pode indicar tentativas de redirecionamento.
9. **http_https**: sinaliza se o protocolo "http://" e "https://" estão incluídos no próprio nome de domínio (1) ou não (0), o que é incomum e potencialmente suspeito.
10. **shortening_services**: sinaliza se a URL usa um serviço de encurtamento de URL (1) ou não (0). URLs encurtadas podem ocultar o destino real.
11. **hyphen**: sinaliza se o nome de domínio tem um hífen ("-") no início ou no final (1) ou não (0). Hífens podem ser usados para criar domínios semelhantes.
12. **dns**: indica se o domínio tem um registro DNS correspondente (1) ou não (0). A ausência de um registro DNS pode ser suspeita.
13. **age_domain**: representa há quanto tempo o domínio está registrado, medido em anos. Domínios recém-registrados podem ter maior probabilidade de estar associados a phishing.
14. **end_period_domain**: representa a data de expiração do registro do domínio, no formato “YEAR-MONTH” (YYYY-MM).
15. **country**: indica o país de origem ou associação do domínio da URL. Essa informação pode ser útil para entender a distribuição geográfica de URLs de phishing e identificar regiões com maior atividade maliciosa.

SUMÁRIO EXECUTIVO

Dados (total de observações)

- Treino:
- Teste (“not_label”):

URLs (treino)	Mean	Median	STD
Phishing			
Legítima			

Considerações para tomada de decisão:

Lista de domínios suspeitos (precisa ter 80% de “acurácia”):

- xxxx.com
- yyyy.net
- …
- zzz.edu


FIAP