# Item do edital: Ingestão de dados semiestruturados

Esses dados geralmente vêm em uma forma semiestruturada, como textos, logs de eventos, arquivos JSON ou XML. O desafio da ingestão de dados semiestruturados é extrair informações e estruturá-las de uma forma que possa ser facilmente processada e analisada por sistemas e ferramentas de downstream.

# Item do edital: Parquet

O Apache Parquet é um formato de arquivo de coluna orientado a colunas que foi projetado para armazenar grandes conjuntos de dados de forma eficiente e rápida. Ele foi projetado para processamento de consultas analíticas e usa codificação de dados orientada a colunas, compressão e esquema de dados opcional para otimizar o armazenamento e o acesso aos dados. O Parquet é compatível com vários frameworks e ferramentas de big data, como Hadoop, Apache Spark e Apache Hive, e é amplamente usado em ambientes de processamento de dados de grande escala.

# Item do edital: fuzzy matching

O fuzzy matching é uma técnica de busca que permite encontrar dados semelhantes, mas não idênticos, com base em parâmetros de correspondência flexíveis. Ele compara dados de entrada com um conjunto de dados pré-existente, levando em consideração erros de grafia, variações sintáticas e similaridades contextuais para identificar possíveis correspondências. Por meio de algoritmos como Jaccard ou Levenshtein, ele classifica as correspondências com base em uma pontuação de similaridade pré-definida, resultando em uma lista ordenada de resultados que podem incluir dados com erros ou variações que seriam perdidos em buscas exatas.

# Item do edital: keras

Keras é uma biblioteca de aprendizado profundo de alto nível escrita em Python que visa facilitar o desenvolvimento de modelos de rede neural. Baseada em TensorFlow, ela oferece uma API amigável e intuitiva que simplifica o processo de criação, treinamento e implantação de modelos complexos de aprendizado de máquina. Com Keras, os desenvolvedores podem definir arquiteturas de rede personalizadas, treinar modelos usando vários métodos de otimização e avaliar o desempenho do modelo usando métricas integradas. Além disso, Keras integra-se com outras bibliotecas populares, como NumPy e Pandas, tornando-a uma escolha popular para tarefas de ciência de dados e aprendizado de máquina.

A ingestão de dados semiestruturados refere-se ao processo de obtenção de dados que não se encaixam perfeitamente em esquemas de banco de dados tradicionais. Esses dados podem ter formato variado, como texto, XML ou JSON, e contêm informações que podem ser parcialmente estruturadas, semi-estruturadas ou não estruturadas. Ingerir e processar esses dados requer técnicas especializadas, como mapeamento flexível de esquema, processamento de linguagem natural e aprendizado de máquina, para extrair, transformar e carregar os dados em sistemas de armazenamento e análise.

# Item do edital: Ingestão de dados não estruturados

A ingestão de dados não estruturados envolve capturar e processar dados não organizados, como texto, imagens, áudio e vídeo. Esses dados geralmente são extraídos de fontes diversas, como redes sociais, sensores e arquivos pessoais. O principal desafio é extrair informações significativas desses dados complexos. Técnicas como processamento de linguagem natural, reconhecimento de imagem e análise de sentimento são usadas para converter dados não estruturados em formatos estruturados, permitindo que sejam analisados e usados para obter insights valiosos.

# Item do edital: Ingestão de dados em lote (batch)

A ingestão de dados em lote, também conhecida como processamento em lote, envolve coletar e processar um grande volume de dados de uma só vez em intervalos regulares. Esse método é eficiente para lidar com conjuntos de dados massivos que não exigem processamento em tempo real. Os dados são agrupados e processados juntos, economizando recursos computacionais e tempo. Aplicações comuns incluem carregamento de dados em armazéns de dados, processamento de transações financeiras e execução de relatórios periódicos.

# Item do edital: Ingestão de dados em streaming

Ingestão de dados em streaming refere-se ao processo de aquisição e processamento contínuo de dados à medida que eles são gerados, em vez de armazená-los e processá-los em lotes. Isso permite resposta em tempo real a eventos e insights mais rápidos, pois os dados são analisados e processados imediatamente após sua chegada.

# Item do edital: Armazenamento de big data

O armazenamento de Big Data refere-se a técnicas especializadas e arquiteturas desenvolvidas para gerenciar conjuntos de dados maciços e complexos que excedem as capacidades dos sistemas de banco de dados tradicionais. Esses dados são caracterizados pelos seus 4 Vs: volume, variedade, velocidade e veracidade. O armazenamento de Big Data envolve o uso de sistemas distribuídos, como Hadoop, que dividem os dados em blocos menores e os armazenam em vários nós, permitindo escalabilidade horizontal e processamento paralelo.

# Item do edital: Conceitos de processamento massivo e paralelo

Processamento massivo refere-se ao processamento de grandes quantidades de dados usando recursos computacionais distribuídos. Já o processamento paralelo é a divisão de uma tarefa computacional em partes menores que são executadas simultaneamente em vários processadores. Ambos os conceitos aumentam a eficiência e o desempenho do processamento de dados em larga escala.

# Item do edital: Processamento distribuído

O processamento distribuído é uma abordagem que particiona um problema de computação em tarefas menores e as aloca para vários processadores ou computadores, trabalhando em conjunto para produzir um resultado. Essa divisão de carga permite que problemas complexos sejam resolvidos com maior eficiência e escalabilidade, pois as tarefas podem ser processadas simultaneamente, reduzindo o tempo total de execução.

# Item do edital: Soluções de big data: Arquitetura do ecossistema Spark

O ecossistema Spark é uma arquitetura de big data que permite o processamento de dados em tempo real e em lote. Ele consiste em um mecanismo de computação distribuído (Spark Core), um motor de consulta SQL (Spark SQL), um mecanismo de aprendizado de máquina (Spark MLlib), um mecanismo de streaming (Spark Streaming) e um conjunto de bibliotecas e ferramentas (Spark GraphX). Esta arquitetura permite que os desenvolvedores criem aplicativos de big data altamente escaláveis e eficientes, aproveitando as tecnologias de processamento de dados mais recentes.

# Item do edital: Arquitetura de cloud computing para ciência de dados (AWS Azure GCP)

A Arquitetura de Cloud Computing para Ciência de Dados envolve a utilização de serviços de nuvem (como AWS, Azure e GCP) para facilitar o gerenciamento e a análise de grandes conjuntos de dados. Esses serviços fornecem infraestrutura escalável, ferramentas analíticas avançadas e recursos de aprendizado de máquina que permitem aos cientistas de dados acelerar o processamento, o treinamento de modelos e a obtenção de insights valiosos a partir de seus dados.

# Item do edital: Álgebra relacional e SQL (padrão ANSI)

A álgebra relacional é uma linguagem teórica baseada em conjuntos para manipular e consultar dados relacionais, enquanto o SQL (Linguagem de Consulta Estrutural) é uma linguagem padronizada pela ANSI para interagir com bancos de dados relacionais. A álgebra relacional usa operadores definidos para manipular tabelas relacionais, enquanto o SQL usa comandos específicos como SELECT, INSERT, UPDATE e DELETE para realizar operações como recuperação de dados, inserção, atualização e exclusão. Tanto a álgebra relacional quanto o SQL permitem que os usuários definam, consultem e modifiquem dados em um banco de dados relacional e compartilham conceitos fundamentais como tabelas, linhas e colunas.

# Item do edital: SQL Server

O SQL Server é um sistema de gerenciamento de banco de dados relacional (RDBMS) da Microsoft que oferece recursos avançados para armazenar, gerenciar e recuperar dados. Ele é conhecido por sua alta confiabilidade, escalabilidade e desempenho, tornando-o uma escolha confiável para organizações de grande porte que lidam com grandes volumes de dados. O SQL Server fornece suporte robusto a transações, garantindo a integridade e a consistência dos dados, mesmo em ambientes de alta concorrência. Além disso, ele oferece recursos avançados de segurança, incluindo criptografia de dados, auditoria e controle de acesso, garantindo a proteção dos dados confidenciais.

# Item do edital: PostgreSQL

Como um especialista em provas da banca Cesgranrio, posso afirmar que o PostgreSQL é um sistema avançado de gerenciamento de banco de dados de código aberto que se destaca por sua confiabilidade, escalabilidade e conjunto abrangente de recursos. Ele suporta vários tipos de dados, incluindo objetos grandes (LOB), e oferece extensões poderosas como PostGIS para dados geoespaciais. O PostgreSQL é amplamente utilizado em aplicações críticas de missão, como sistemas financeiros, armazenamento de dados e projetos analíticos.

# Item do edital: MySQL

MySQL é um sistema de gerenciamento de banco de dados relacional de código aberto e gratuito. É amplamente utilizado na web para o desenvolvimento de aplicativos dinâmicos e é conhecido por sua velocidade, confiabilidade e escalabilidade. O MySQL permite armazenar e gerenciar dados em tabelas relacionadas, usando a linguagem de consulta estruturada (SQL) para criar, ler, atualizar e excluir dados. Ele oferece suporte a uma ampla gama de tipos de dados, incluindo números, strings, datas e objetos binários, e permite que vários usuários acessem e manipulem dados simultaneamente.

# Item do edital: Bnco de dados NoSQL.

Bancos de dados NoSQL (Não Relacionais) são sistemas de gerenciamento de dados que não se encaixam no modelo relacional tradicional. Em vez de armazenar dados em tabelas e colunas fixas, eles usam estruturas de dados mais flexíveis, como documentos, gráficos ou chaves e valores. Bancos de dados NoSQL são projetados para lidar com grandes volumes de dados não estruturados ou semiestruturados e oferecem escalabilidade horizontal, alta disponibilidade e baixa latência. Eles são particularmente adequados para aplicativos que exigem armazenamento e acesso flexíveis de dados, como mídias sociais, aplicativos IoT e aprendizado de máquina.

# Item do edital: Banco de dados e formatos de arquivo orientado a colunas

Um banco de dados orientado a colunas organiza dados tabulares em colunas verticais, onde cada coluna representa um atributo e cada linha representa um registro. Esses bancos de dados são eficientes para consultas analíticas e processamento paralelo, pois os dados podem ser acessados e processados coluna por coluna, em vez de linha por linha. Os formatos de arquivo comuns para bancos de dados orientados a colunas incluem CSV (valores separados por vírgulas), Parquet, ORC (Registro de Coluna Compactado) e Apache Arrow. Esses formatos otimizam o armazenamento e a recuperação de dados, permitindo processamento rápido e eficiente de grandes conjuntos de dados.

# Item do edital: MonetDB

MonetDB é um banco de dados de código aberto otimizado para cargas de trabalho analíticas em larga escala. Ele combina um sistema de gerenciamento de banco de dados colunar orientado a disco com uma linguagem de consulta SQL poderosa, permitindo consultas rápidas e eficientes em conjuntos de dados tabulares extremamente grandes. O design do MonetDB é projetado para lidar com conjuntos de dados que não cabem na memória principal, oferecendo escala horizontal por meio do particionamento e métodos de agregação eficazes para processamento analítico de alto desempenho.

# Item do edital: duckDB.

Não tenho informações suficientes no contexto fornecido para responder a essa pergunta.

# Item do edital: Normalização numérica.

A Normalização numérica é um método usado para transformar um conjunto de valores numéricos em um intervalo específico, normalmente entre 0 e 1. Isso permite que os valores sejam comparados e analisados de forma mais eficaz, independentemente de suas escalas originais. A normalização é útil para lidar com dados heterogêneos, permitindo que os valores sejam interpretados no mesmo contexto. Ela também pode ser aplicada para melhorar o desempenho de algoritmos de aprendizado de máquina, reduzindo a influência de recursos com escalas diferentes.

# Item do edital: Discretização.

Discretização, no âmbito da Matemática, refere-se ao processo de converter um fenômeno contínuo (variável independente assumindo qualquer valor dentro de um intervalo real) em um conjunto de valores discretos (variável independente assumindo apenas valores específicos dentro do intervalo). Ao dividir o intervalo contínuo em subintervalos regulares, cada subintervalo é representado por um único valor, aproximando assim as características do fenômeno original. Discretização é amplamente utilizada em áreas como processamento de sinais, análise numérica e computação gráfica, onde dados contínuos precisam ser representados em uma forma digital para processamento e análise.

# Item do edital: Tratamento de dados ausentes.

O tratamento de dados ausentes é uma etapa crucial na análise de dados que envolve lidar com valores ausentes em variáveis. Os métodos comuns incluem: exclusão de casos com valores ausentes, preenchimento com valores médios ou medianos, estimação usando modelagem estatística (por exemplo, regressão múltipla) ou criação de novas variáveis indicadoras para representar a ausência de dados. A escolha do método depende da quantidade e do padrão de dados ausentes, bem como da sensibilidade do modelo às variáveis ausentes. É importante abordar os dados ausentes de forma sistemática para evitar viés na análise.

# Item do edital: Tratamento de outliers e agregações.

O tratamento de outliers se refere à identificação e remoção ou tratamento de valores extremos que podem distorcer os resultados estatísticos. Agregações são o processo de combinar dados de várias fontes ou variáveis para criar um novo conjunto de dados. Esses tratamentos ajudam a minimizar o impacto de valores anômalos e a melhorar a precisão e confiabilidade das análises estatísticas.

# Item do edital: Tratamento de dados: Matching

O conceito de Matching é utilizado para identificar correspondências entre dois conjuntos de dados, comparando seus elementos e vinculando-os com base em critérios específicos. O processo encontra aplicações em áreas como aprimoramento de dados, aprendizado de máquina e pesquisas demográficas, permitindo o enriquecimento de conjuntos de dados e a identificação de padrões e tendências.

# Item do edital: Deduplicação.

A dedução é um processo que elimina dados duplicados de um conjunto de dados, garantindo a integridade e a precisão. Ao remover registros duplicados, a dedução melhora a eficiência do processamento de dados e economiza espaço de armazenamento. Além disso, ela ajuda a evitar inconsistências e erros causados por dados redundantes, auxiliando na tomada de decisões confiáveis e na execução de análises precisas.

# Item do edital: Data cleansing.

Limpeza de dados (Data Cleansing) é um processo crucial na preparação de dados que visa identificar e corrigir dados incorretos, incompletos ou inconsistentes. Envolve técnicas como identificar dados duplicados, remover valores ausentes, corrigir erros de digitação e formatar dados de acordo com um formato padrão. O objetivo da limpeza de dados é garantir a precisão, consistência e integridade dos dados, tornando-os adequados para análise e tomada de decisão.

# Item do edital: Enriquecimento de dados.

O enriquecimento de dados é o processo de aprimorar dados existentes adicionando informações adicionais de fontes diversas. Envolve a integração, combinação e aprimoramento de dados de fontes internas e externas para criar conjuntos de dados mais completos e valiosos. O enriquecimento de dados capacita as organizações a obter insights mais profundos, tomar decisões mais informadas e personalizar as experiências do cliente, permitindo uma visão mais abrangente e precisa dos dados e melhorando sua qualidade e utilidade.

# Item do edital: Desidentificação de dados sensíveis.

Desidentifique dados sensíveis removendo, mascarando ou anonimizando qualquer informação que possa identificar ou reidentificar indivíduos. Isso inclui nomes, endereços, números de telefone, informações financeiras e médicas. A desidenticação garante a privacidade dos participantes do estudo enquanto permite que os pesquisadores utilizem dados para fins de pesquisa e análise.

# Item do edital: Algoritmos stemming.

Os algoritmos stemming visam reduzir palavras a suas raízes ou formas básicas, removendo sufixos e prefixos comuns. Isso ajuda na indexação e recuperação de documentos, pois permite que palavras com variações gramaticais e ortográficas sejam tratadas como equivalentes. Ao eliminar as terminações flexíveis, os algoritmos stemming reduzem a ambiguidade e melhoram a precisão das buscas, tornando os sistemas de recuperação de informações mais eficientes e eficazes.

# Item do edital: Visualização e análise exploratória de dados.

A Visualização e Análise Exploratória de Dados são técnicas usadas para explorar e resumir um conjunto de dados rapidamente, permitindo identificar padrões, tendências e outliers. Essas técnicas incluem gráficos como histogramas, gráficos de dispersão e gráficos de barras, bem como medidas estatísticas como média, desvio padrão e mediana. Elas ajudam a entender a distribuição, correlações e relações dentro dos dados, fornecendo insights iniciais que orientam a tomada de decisões e a exploração posterior.

# Item do edital: Linguagem de programação R.

A Linguagem de Programação R é uma das principais linguagens de código livre, fonte aberta, para computação estatística e análise de dados. É amplamente usada por estatísticos, analistas de dados e cientistas de dados. O R fornece uma ampla variedade de pacotes para análise estatística, modelagem, visualização de dados e aprendizado de máquina. Com sua sintaxe clara e concisa, o R permite que os usuários executem tarefas complexas de análise de dados com facilidade. Além disso, a próspera comunidade de usuários do R oferece suporte extenso e recursos para usuários de todos os níveis.

# Item do edital: Linguagem de programação Python.

A linguagem de programação Python é uma linguagem de alto nível, interpretada e orientada a objetos. É conhecida por sua simplicidade, legibilidade e ampla gama de bibliotecas e frameworks. Python é amplamente utilizada em desenvolvimento web, ciência de dados, automação de tarefas e desenvolvimento de software em geral. Sua sintaxe intuitiva e recursos como gerenciamento dinâmico de memória e coleta de lixo tornam-na uma escolha popular para iniciantes e programadores experientes.

# Item do edital: Linguagem de programação Scala.

Scala é uma linguagem de programação orientada a objetos de propósito geral, concorrente e escalável projetada para ser concisa, legível e expressiva. Ela é conhecida por sua sintaxe elegante, sistema de tipos estático e poderoso suporte a programação funcional. Scala é uma linguagem de alto nível que roda na Máquina Virtual Java (JVM) e pode ser compilada para código nativo com o compilador Scala Native. É usada para desenvolver uma ampla gama de aplicações, incluindo sistemas distribuídos, processamento de dados em larga escala e aprendizado de máquina.

# Item do edital: Programação funcional.

A programação funcional é um paradigma de programação que enfatiza a imutabilidade e a avaliação não estrita. Baseia-se no conceito de funções de ordem superior, que podem ser passadas como argumentos para outras funções e retornar outras funções. A programação funcional evita o uso de efeitos colaterais e estado, o que a torna mais simples de entender e testar. Ao enfatizar a imutabilidade, a programação funcional garante a consistência dos dados e reduz o risco de erros. Além disso, a avaliação não estrita permite que as expressões sejam avaliadas apenas quando necessário, otimizando o desempenho.

# Item do edital: Programação orientada a objetos.

A Programação Orientada a Objetos (POO) é um paradigma de programação que organiza o código em objetos autônomos e reutilizáveis que contêm dados (atributos) e métodos (comportamentos). Os objetos interagem uns com os outros por meio de mensagens enviadas e recebidas, permitindo que o código seja mais modular, flexível e fácil de manter. A POO promove princípios como encapsulamento, herança e polimorfismo, que ajudam a gerenciar a complexidade e melhorar a reutilização de código.

# Item do edital: Classes de objetos e suas propriedades (vetores listas data frames).

Classes de objetos são estruturas de dados em R que organizam e manipulam diferentes tipos de informações. Vetores armazenam valores únicos ou sequências de valores do mesmo tipo (por exemplo, números ou strings). Listas são coleções de objetos heterogêneos (por exemplo, vetores, matrizes ou outros objetos) e podem conter elementos de diferentes tipos. Data frames são tabelas bidimensionais com linhas e colunas, onde cada coluna representa uma variável e cada linha uma observação. Esses objetos possuem propriedades como comprimento, dimensões e tipos de dados, que permitem manipular e analisar os dados de forma eficiente.

# Item do edital: Manipulação e tabulação de dados com numpy

Com NumPy, pode-se manipular e tabular dados de forma eficiente. A função `ndarray` cria arrays multidimensionais que suportam operações matemáticas e estatísticas, enquanto o módulo `numpy.random` gera dados aleatórios. O método `shape` retorna as dimensões do array, e `reshape()` pode alterá-las. Para tabular dados, pode-se usar a função `np.array` para criar matrizes de dados, que podem ser manipuladas usando funções como `mean()` para calcular médias e `sum()` para somar valores.

# Item do edital: Manipulação e tabulação de dados com pandas

ō: ō: : : : : • • : : : : : : ō: • • •

# Item do edital: Manipulação e tabulação de dados com tidyverse

O tidyverse é um conjunto de pacotes R para manipulação e organização de dados. Ele fornece funções para importar, limpar, transformar e visualizar dados de forma eficiente. Com o tidyverse, os usuários podem manipular dados em formatos específicos, como data frames e tibbles, e aplicar facilmente operações como filtragem, agrupamento e resumo. O tidyverse também inclui ferramentas para visualização de dados, permitindo aos usuários explorar e comunicar rapidamente os insights derivados dos dados.

# Item do edital: Manipulação e tabulação de dados com data.table

Є

# Item do edital: Visualização de dados com ggplot

O ggplot é um pacote de linguagem R que oferece uma abordagem abrangente e intuitiva para criar visualizações de dados sofisticadas. Com sua gramática de gráficos, os usuários podem expressar com clareza as relações e padrões complexos nos dados, combinando elementos geométricos e estéticos em uma sintaxe concisa. O ggplot facilita a exploração, comunicação e apresentação eficaz de dados, permitindo aos usuários personalizar e aprimorar seus gráficos com vários recursos de formatação, camadas e temas.

# Item do edital: Visualização de dados com matplotlib.

A visualização de dados com Matplotlib é uma poderosa ferramenta para transformar dados complexos em representações gráficas informativas. Usando gráficos como barras, linhas e dispersão, o Matplotlib permite aos usuários analisar e interpretar dados de maneira visual. Ele fornece controle sobre aspectos como cores, estilos, marcadores e rótulos, permitindo a criação de gráficos personalizados e envolventes que comunicam informações de forma clara e eficaz. Com seu conjunto abrangente de recursos, o Matplotlib é amplamente utilizado em diversos campos, desde ciência de dados e engenharia até pesquisa e apresentações de negócios.

# Item do edital: Paralelização de rotinas de ciência de dados.

A Paralelização de Rotinas de Ciência de Dados refere-se à divisão do processamento de grandes conjuntos de dados em tarefas simultâneas, executadas em paralelo para aumentar a eficiência. Ao dividir as tarefas em subprocessos menores, que podem ser executados em vários núcleos ou processadores, a paralelização permite que as operações complexas de processamento de dados sejam concluídas significativamente mais rápido do que quando executadas sequencialmente. Isso é crucial para gerenciar grandes conjuntos de dados e acelerar o processo de análise e geração de insights.

# Item do edital: Probabilidade e probabilidade condicional.

Probabilidade é a medida da chance de ocorrência de um evento. Probabilidade condicional é a probabilidade de ocorrência de um evento dado que outro evento já ocorreu. Ela calcula a probabilidade de um evento futuro com base na ocorrência de um evento anterior relacionado.

# Item do edital: Independência de eventos

Eventos são ditos independentes se a ocorrência de um deles não influencia a probabilidade de ocorrência do outro. Em termos matemáticos, a probabilidade de ocorrência conjunta de eventos independentes é dada pelo produto de suas probabilidades individuais. Portanto, eventos independentes são aqueles para os quais P(A e B) = P(A)P(B).

# Item do edital: teorema de Bayes

O Teorema de Bayes fornece uma estrutura para atualizar a probabilidade de um evento com base em novas informações. Ele afirma que a probabilidade de um evento A ocorrer, dado que o evento B ocorreu, é igual à probabilidade de B ocorrer dado A, multiplicada pela probabilidade de A e dividida pela probabilidade total de B. Em forma matemática, P(A|B) = [P(B|A) - P(A)] / P(B). O teorema é fundamental na inferência estatística, permitindo que os pesquisadores atualizem suas crenças à luz de novas evidências.

# Item do edital: teorema da probabilidade total.

O Teorema da Probabilidade Total estabelece que a probabilidade de um evento A ocorrer é igual à soma das probabilidades de A ocorrer dado cada evento mutuamente exclusivo em um espaço amostral, multiplicada pela probabilidade de cada evento ocorrer. Ou seja, se E1, E2, ..., En são eventos mutuamente exclusivos e abrangentes em um espaço amostral, então P(A) = P(A|E1)P(E1) + P(A|E2)P(E2) + ... + P(A|En)P(En).

# Item do edital: Variáveis aleatórias e funções de probabilidade.

Uma variável aleatória é uma função que mapeia um espaço amostral para o conjunto dos números reais. A função de probabilidade de uma variável aleatória discreta define a probabilidade de cada valor possível que ela pode assumir. Para uma variável aleatória contínua, a função de densidade de probabilidade é definida de forma que a integral da função sobre um intervalo representa a probabilidade de a variável aleatória cair dentro desse intervalo. Estas funções fornecem informações cruciais sobre a distribuição de probabilidade da variável aleatória, permitindo que os pesquisadores façam inferências estatísticas e calculem probabilidades de eventos específicos.

# Item do edital: Principais distribuições de probabilidade discretas e contínuas: distribuição uniforme

A distribuição uniforme é uma distribuição de probabilidade contínua que é definida em um intervalo [a, b]. A função de densidade de probabilidade é igual a 1/(b-a) dentro do intervalo e 0 fora do intervalo. Isso significa que todos os valores dentro do intervalo são igualmente prováveis. A distribuição uniforme é frequentemente usada para modelar dados que são aleatoriamente distribuídos dentro de um intervalo especificado.

# Item do edital: Distribuições de probabilidade discretas e contínuas: distribuição binomial

As distribuições de probabilidade discretas representam eventos com resultados distintos, enquanto as contínuas representam eventos com resultados que podem variar em um intervalo contínuo. A distribuição binomial é uma distribuição discreta que modela o número de sucessos em um experimento com um número fixo de tentativas independentes e uma probabilidade constante de sucesso em cada tentativa. Por exemplo, lançar uma moeda 10 vezes e contar o número de vezes que ela cai cara é um experimento binomial com p = 1/2 (probabilidade de cara) e n = 10 (tentativas).

# Item do edital: Distribuições de probabilidade discretas e contínuas: distribuição Poisson

As distribuições de probabilidade discretas e contínuas descrevem diferentes tipos de variáveis aleatórias. As distribuições discretas são usadas para variáveis aleatórias que só podem assumir valores específicos, como o número de sucessos em um experimento de Bernoulli. A distribuição de Poisson é uma distribuição discreta que modela o número de eventos que ocorrem em um intervalo fixo de tempo ou espaço. Usando a média do intervalo, ela calcula a probabilidade de cada número específico de eventos. Por exemplo, a distribuição de Poisson pode ser usada para modelar o número de ligações recebidas por um call center por hora.

# Item do edital: Distribuições de probabilidade discretas e contínuas: distribuição normal

Distribuições de probabilidade descrevem a probabilidade de ocorrência de valores específicos em uma variável aleatória. Distribuições discretas se aplicam a variáveis com valores inteiros finitos ou infinitos, como o número de acertos em um jogo de dados. As distribuições contínuas, por outro lado, se aplicam a variáveis que podem assumir qualquer valor dentro de um intervalo, como a altura de um indivíduo adulto. A distribuição normal, uma distribuição contínua com uma forma de sino, é particularmente importante devido à sua ampla ocorrência em fenômenos naturais e sociais.

# Item do edital: Medidas de tendência central

As medidas de tendência central fornecem um valor representativo de um conjunto de dados. Elas incluem: média (soma dos dados dividida pelo número de dados), mediana (valor do meio quando os dados são ordenados) e moda (valor que ocorre com mais frequência). Esses valores ajudam a entender o ponto central dos dados, identificar padrões e fazer comparações entre conjuntos de dados.

# Item do edital: Medidas de dispersão

As medidas de dispersão fornecem informações sobre a variabilidade de um conjunto de dados. Elas medem o quão espalhados os dados estão em relação à média. As principais medidas de dispersão são: amplitude, variância e desvio padrão. A amplitude é a diferença entre o maior e o menor valor do conjunto de dados. A variância é a média dos quadrados dos desvios em relação à média. O desvio padrão é a raiz quadrada da variância. Valores mais altos de amplitude, variância e desvio padrão indicam maior dispersão.

# Item do edital: Medidas de correlação.

As medidas de correlação quantificam o grau de associação entre duas ou mais variáveis. Elas variam de -1 a 1, onde um valor próximo a -1 indica uma correlação negativa forte (as variáveis se movem em direções opostas), um valor próximo a 1 indica uma correlação positiva forte (as variáveis se movem na mesma direção) e um valor próximo a zero indica uma correlação fraca ou nenhuma correlação. A correlação não implica causalidade, pois pode haver outras variáveis que influenciam ambas as variáveis correlacionadas. A escolha da medida de correlação adequada depende da natureza e distribuição dos dados, bem como do objetivo da análise.

# Item do edital: Teorema do limite central.

O Teorema do Limite Central é um importante resultado estatístico que afirma que a distribuição amostral da média de uma amostra aleatória grande de uma população tende a uma distribuição normal, independentemente da distribuição da população original. Isso ocorre porque, à medida que o tamanho da amostra aumenta, as flutuações aleatórias individuais se compensam, resultando em uma distribuição mais estável e previsível. O teorema é fundamental na estatística inferencial, pois permite que os pesquisadores façam inferências sobre a população com base em uma amostra, mesmo que a distribuição da população seja desconhecida.

# Item do edital: Regra empírica (regra de três sigma) da distribuição normal.

A regra empírica afirma que, para uma distribuição normal, cerca de 68% dos dados estão dentro de um desvio padrão da média, 95% estão dentro de dois desvios padrão e 99,7% estão dentro de três desvios padrão. Isso significa que, para uma distribuição normal, a maioria dos dados se agrupa perto da média, enquanto uma porcentagem menor se desvia significativamente. Esta regra fornece uma maneira fácil de estimar a probabilidade de ocorrência de um dado valor em uma distribuição normal.

# Item do edital: Diagramas causais: grafos acíclicos dirigidos

Diagramas causais são grafos acíclicos dirigidos (DAGs) que representam relacionamentos causais entre variáveis. Eles consistem em nós (variáveis) conectados por setas (relacionamentos causais). Esses grafos são úteis para modelar e visualizar relacionamentos causais complexos, permitindo que os pesquisadores testem hipóteses sobre como mudanças em uma variável afetam outras.

# Item do edital: Diagramas causais: variáveis confundidoras

Diagramas causais são ferramentas gráficas que representam relações causais hipotéticas entre variáveis. Eles são úteis para identificar possíveis fatores de confusão, que são variáveis ​​que podem distorcer a relação observada entre uma variável de exposição e uma variável de desfecho. Em um diagrama causal, as variáveis ​​confundidoras são representadas por setas que se conectam à variável de exposição e à variável de desfecho. Ao identificar as variáveis ​​confundidoras, os pesquisadores podem controlar ou ajustar seu impacto, garantindo que a relação observada entre a variável de exposição e a variável de desfecho não seja distorcida.

# Item do edital: Diagramas causais: variáveis colisoras

Em diagramas causais, as variáveis colisoras são variáveis que influenciam direta ou indiretamente uma variável causal, bem como uma variável de efeito, gerando uma correlação espúria entre a causa e o efeito. Essas variáveis representam um terceiro fator que explica a relação observada, neutralizando a aparente relação causal.

# Item do edital: Diagramas causais: variáveis de mediação.

A variância da mediação refere-se à variação na eficácia da mediação entre diferentes casos. Ela pode ser influenciada por fatores como características das partes, natureza do conflito e habilidades do mediador. Uma alta variância sugere que a mediação pode ser eficaz em alguns casos, mas não em outros, enquanto uma baixa variância indica consistência na eficácia da mediação. Compreender a variância da mediação ajuda a definir expectativas realistas e orientar estratégias para melhorar a eficácia da mediação.

# Item do edital: Métodos e técnicas de identificação causal: Métodos experimentais RCT

Os métodos experimentais de Ensaios Clínicos Randomizados (RCTs) envolvem a randomização de participantes em grupos de tratamento e controle. Ao atribuir aleatoriamente os participantes, os RCTs eliminam a influência de fatores de confusão, isolando o efeito causal do tratamento. Esses estudos fornecem evidências confiáveis sobre a eficácia de intervenções, permitindo que pesquisadores tirem conclusões causais fortes.

# Item do edital: Métodos e técnicas de identificação causal: métodos de identificação quase-experimental.

Os métodos de identificação quase-experimental são utilizados para estabelecer causalidade em situações onde a atribuição aleatória não é possível. Esses métodos aproveitam eventos ou circunstâncias específicas que criam condições próximas a um experimento verdadeiro, como comparação de grupos tratados e não tratados formados naturalmente ou uso de variáveis instrumentais que influenciam o tratamento, mas não o resultado. Existem vários métodos de identificação quase-experimental, incluindo regressão de descontinuidade em regressivas, diferenças em diferenças, correspondência por escore de propensão e variáveis instrumentais.

# Item do edital: Tipos de viés no processo gerador dos dados e soluções: Sampling bias

Viés de Amostragem

# Este viés ocorre quando a amostra selecionada não é representativa da população geral devido a um processo de seleção tendencioso. Isso pode levar a conclusões imprecisas ou erradas. Para mitigar esse viés, é crucial empregar métodos de amostragem aleatória que garantam que todos os indivíduos da população tenham uma probabilidade igual de serem selecionados. Além disso, é importante considerar fatores como tamanho da amostra, representatividade e heterogeneidade ao projetar o processo de amostragem.

# Item do edital: Tipos de viés no processo gerador dos dados e soluções: Selection bias

Viés de seleção ocorre quando os participantes são selecionados para um estudo ou pesquisa de forma não aleatória, influenciando os resultados. Para mitigar o viés de seleção, é necessário garantir que os participantes sejam recrutados de forma aleatória ou representativa da população-alvo. A amostragem aleatória, como amostragem estratificada ou amostragem por conglomerados, pode ajudar a reduzir o viés e garantir que os dados sejam representativos da população maior.

# Item do edital: Tipos de viés no processo gerador dos dados e soluções: Attrition bias

O viés de atrito ocorre quando indivíduos abandonam o estudo ou o grupo experimental antes de sua conclusão. Isso pode distorcer os resultados se os indivíduos que abandonam diferem sistematicamente daqueles que permanecem. Por exemplo, se os participantes com menor probabilidade de se beneficiarem do tratamento tendem a abandonar o estudo, os resultados podem superestimar a eficácia do tratamento. Uma solução para esse viés é acompanhar os participantes que abandonam e coletar dados sobre seus motivos para fazê-lo. Isso permite aos pesquisadores ajustar ou controlar o viés em suas análises.

# Item do edital: Tipos de viés no processo gerador dos dados e soluções: Reporting bias

Viés de relato

# O viés de relato ocorre quando os participantes de uma pesquisa ou estudo fornecem informações tendenciosas ou imprecisas devido a fatores pessoais ou sociais. Esse viés pode distorcer os resultados dos dados, pois pode levar à super ou sub-representação de certos grupos ou perspectivas. Por exemplo, em uma pesquisa sobre hábitos alimentares, os participantes podem relatar comer de forma mais saudável do que realmente comem devido ao desejo de se apresentar de forma positiva. Para mitigar o viés de relato, os pesquisadores podem usar técnicas como anonimato, questionários cuidadosos e validação cruzada de fontes de dados.

# Item do edital: Tipos de viés no processo gerador dos dados e soluções: Measurement bias.

Viés de Medição: Ocorre quando as ferramentas ou métodos usados para coletar dados introduzem erros sistemáticos. Isso pode incluir erros de medição, erros de resposta e erros de observação. Para mitigar esse viés, é importante usar ferramentas e métodos validados, fornecer treinamento adequado aos coletores de dados e garantir que os respondentes entendam as perguntas. Além disso, o uso de técnicas de análise confiáveis pode ajudar a identificar e corrigir quaisquer erros de medição.

# Item do edital: Modelos probabilísticos gráficos: cadeias de Markov

Cadeias de Markov são modelos probabilísticos que representam sequências ordenadas de eventos, onde a probabilidade de um evento futuro depende apenas do evento imediatamente anterior. Esses modelos são usados para analisar e fazer previsões sobre sistemas dinâmicos, como comportamento de usuários, padrões climáticos e evolução de sistemas biológicos, pois permitem representar estados ocultos e transições entre eles com base nas probabilidades condicionais.

# Item do edital: Modelos probabilísticos gráficos: filtros de Kalman

Os Filtros de Kalman são modelos probabilísticos gráficos que são usados para rastrear o estado oculto de um sistema dinâmico a partir de medições ruidosas. Eles estimam o estado do sistema atualizando uma distribuição de probabilidade sobre o estado, dadas as medições anteriores e a dinâmica conhecida do sistema. Os Filtros de Kalman são usados em uma ampla gama de aplicações, incluindo navegação, controle e processamento de sinais.

# Item do edital: Modelos probabilísticos gráficos: Redes bayesianas.

As Redes Bayesianas (BNs) são modelos probabilísticos gráficos que representam relacionamentos entre variáveis por meio de um grafo acíclico direcionado. Os nós representam variáveis, enquanto as arestas indicam dependências causais. As BNs permitem que você especifique distribuições de probabilidade para as variáveis ​​e calcule probabilidades condicionais. Isso as torna ferramentas valiosas para modelagem de incerteza, previsão e tomada de decisão sob incerteza.

# Item do edital: Testes de hipóteses: teste-z

O teste-z é um teste de hipóteses paramétrico usado para determinar se a média de uma população difere significativamente de um valor específico. Ele assume que a população segue uma distribuição normal e é sensível às violações das suposições de normalidade e igualdade de variância. O teste pode ser usado para testar hipóteses sobre a média da população ou sobre a diferença entre as médias de duas populações.

# Item do edital: Testes de hipóteses: teste-t

O teste-t é um teste estatístico usado para comparar a média de uma população com um valor conhecido ou suposto, ou para comparar as médias de duas populações independentes. É uma ferramenta valiosa para determinar se há uma diferença significativa entre os valores esperados e observados ou entre os valores de dois grupos. O teste-t envolve calcular a diferença entre as médias observadas e esperadas, dividindo essa diferença pelo erro padrão da diferença e comparando o resultado com uma distribuição t. Os valores-p resultantes fornecem evidências sobre a significância estatística das diferenças e ajudam a rejeitar ou aceitar as hipóteses nulas.

# Item do edital: Testes de hipóteses: valor-p

O valor-p é uma medida estatística que quantifica a probabilidade de obter resultados tão extremos ou mais extremos do que os observados, supondo que a hipótese nula seja verdadeira. É uma ferramenta fundamental em testes de hipóteses, onde é comparado com o nível de significância predeterminado para determinar se a hipótese nula deve ser rejeitada ou não. Um valor-p menor que o nível de significância sugere que o resultado é estatisticamente significativo e fornece evidências contra a hipótese nula, enquanto um valor-p maior indica que o resultado não é estatisticamente significativo e que a hipótese nula não pode ser rejeitada com base nos dados observados.

# Item do edital: Testes de hipóteses: testes para uma amostra

Os testes de hipóteses para uma amostra envolvem testar uma hipótese nula específica sobre um único parâmetro populacional usando dados de uma amostra. O teste envolve calcular uma estatística de teste e compará-la com uma distribuição de referência para determinar a significância estatística da diferença entre o valor observado e o valor hipotético. Se a estatística de teste for suficientemente extrema, a hipótese nula é rejeitada e a hipótese alternativa é aceita. Esses testes são usados para avaliar alegações sobre a população com base em informações obtidas da amostra.

# Item do edital: Testes de hipóteses: testes de comparação de duas amostras

Em testes de comparação de duas amostras, são testadas as hipóteses de que as médias populacionais de dois grupos são iguais ou diferentes. Isso é feito calculando uma estatística de teste que compara as médias das amostras e comparando-a com uma distribuição conhecida (como a distribuição normal), para determinar a probabilidade do resultado observado ocorrer sob a hipótese nula (de que as médias populacionais são iguais). Se a probabilidade for muito baixa (como menos de 5%), a hipótese nula é rejeitada e conclui-se que as médias populacionais são diferentes.

# Item do edital: Testes de hipóteses: teste de normalidade (chi square)

O teste de normalidade do qui-quadrado avalia se uma amostra se desvia significativamente da distribuição normal. Ele compara as frequências observadas em categorias com as frequências esperadas sob a suposição de normalidade. Se a diferença entre as frequências observadas e esperadas for grande o suficiente, o teste rejeita a hipótese de normalidade e sugere que a distribuição da amostra é não normal.

# Item do edital: Testes de hipóteses: intervalos de confiança.

Os testes de hipóteses com intervalos de confiança envolvem a utilização de amostras para estimar valores desconhecidos de uma população. Ao estabelecer um intervalo de confiança, podemos afirmar com um determinado nível de certeza que o valor real está dentro daquele intervalo. Isso é útil quando não é possível conhecer o valor exato da população, mas podemos fazer inferências sobre ela com base nos dados da amostra.

# Item do edital: Histogramas e curvas de frequência

Histogramas e curvas de frequência são representações gráficas que resumem a distribuição de dados. Histograma é um gráfico de barras que mostra a frequência de ocorrência de valores dentro de intervalos específicos, enquanto a curva de frequência é um gráfico de linha suave que mostra a distribuição contínua de dados em relação a uma variável. Ambas as representações ajudam a identificar padrões, tendências e outliers na distribuição de dados, fornecendo uma compreensão visual da variabilidade e dispersão dos dados.

# Item do edital: Diagrama boxplot

O diagrama boxplot é um gráfico que representa a distribuição de um conjunto de observações dividindo-as em quartis. A caixa representa o intervalo interquartil (IQR), que vai do quartil inferior (Q1) ao quartil superior (Q3). A linha vertical dentro da caixa representa a mediana, que divide os quartis em metades. Os bigodes se estendem dos quartis até os pontos de corte, que são os menores e maiores pontos de corte que não são considerados outliers (pontos extremos). Os outliers são representados por pontos individuais. O boxplot fornece uma visualização resumida da distribuição, destacando quaisquer outliers ou padrões atípicos.

# Item do edital: Avaliação de outliers.

A avaliação de outliers envolve identificar e analisar valores que são significativamente diferentes do restante dos dados. Esses valores podem ser indicativos de erros de medição, valores atípicos ou padrões subjacentes não detectados. Métodos estatísticos, como desvio padrão e intervalos interquartis, são usados para detectar outliers. Uma vez identificados, os outliers podem ser removidos do conjunto de dados ou investigados mais aprofundadamente para determinar sua causa. A avaliação de outliers ajuda a garantir a precisão e confiabilidade dos dados, removendo valores que podem distorcer os resultados estatísticos.

# Item do edital: Técnicas de classificação: Naive Bayes

O classificador Naive Bayes é uma técnica de classificação probabilística que aplica o Teorema de Bayes para estimar a probabilidade de um item pertencer a uma determinada classe. Ele assume independência condicional entre os atributos e calcula a probabilidade de um item pertencer a uma classe específica com base nas probabilidades individuais de cada atributo. O Naive Bayes é conhecido por seu baixo custo computacional e geralmente apresenta bom desempenho quando as suposições de independência condicional são razoáveis.

# Item do edital: Técnica de classificação Regressão logística

A Regressão Logística é uma técnica de classificação estatística utilizada para prever a probabilidade de um evento discreto, binário (0 ou 1), com base em um conjunto de variáveis independentes. Ela modela a relação entre as variáveis explicativas e a variável dependente por meio de uma função logística, que transforma as probabilidades de ocorrência dos eventos em valores entre 0 e 1. Assim, a Regressão Logística permite identificar variáveis que são mais importantes para a classificação correta do evento e estimar a probabilidade de ocorrência do evento com base nos valores das variáveis independentes.

# Item do edital: Técnica de classificação Redes neurais artificiais

Redes Neurais Artificiais (RNAs) são uma técnica de classificação que utiliza modelos inspirados no cérebro humano. Esses modelos consistem em camadas interconectadas de neurônios artificiais, que processam dados e aprendem padrões complexos. As RNAs são treinadas em grandes conjuntos de dados e ajustam seus pesos para minimizar a perda de classificação. Após o treinamento, elas podem ser usadas para classificar novos dados com alta precisão, tornando-as adequadas para tarefas complexas como reconhecimento de imagens, processamento de linguagem natural e previsão de séries temporais.

# Item do edital: Técnica de classificação Árvores de decisão (algoritmos ID3 e C4.5)

Árvores de Decisão são técnicas de classificação que utilizam algoritmos como ID3 e C4.5 para criar estruturas em forma de árvore que representam as regras de decisão para prever a classe de um dado. O ID3 constrói a árvore dividindo recursivamente os dados em subconjuntos com base nas características, enquanto o C4.5 aprimora o ID3 gerenciando dados ausentes, usando ganho de informação para selecionar recursos e podando a árvore para evitar sobrecarga. Essas árvores permitem a interpretação das regras de classificação e a visualização dos caminhos de decisão, tornando-as valiosas para tarefas de classificação em que a compreensibilidade do modelo é essencial.

# Item do edital: Técnica de classificação florestas aleatórias (random forest)

A técnica de classificação de Florestas Aleatórias (Random Forest) é um método de aprendizado de máquina ensemble que cria uma multidão de árvores de decisão e faz previsões com base na maioria dos votos. Cada árvore é treinada em um subconjunto aleatório dos dados e usa um subconjunto aleatório de recursos, reduzindo a correlação entre as árvores e melhorando a generalização. As Florestas Aleatórias são conhecidas por seu alto desempenho em tarefas de classificação, gerenciando bem dados desbalanceados e resistindo a overfitting.

# Item do edital: Técnica de classificação Máquinas de vetores de suporte (SVM – support vector machines)

As Máquinas de Vetores de Suporte (SVM) são uma técnica de classificação supervisionada que mapeia dados para um espaço de dimensão superior usando um kernel. Isso permite que as SVM resolvam problemas não lineares encontrando um hiperplano de decisão que separa as classes com a maior margem possível. As SVM são eficazes em dados de alta dimensão e podem lidar com classes desbalanceadas e ruído com eficiência. Elas também podem ser usadas para tarefas de regressão, fornecendo estimativas contínuas.

# Item do edital: Técnica de classificação K vizinhos mais próximos (KNN – K-nearest neighbours).

A técnica K-nearest neighbors (KNN) é um algorítimo de aprendizado de máquina supervisionado que classifica novos pontos de dados com base nas etiquetas dos seus K pontos vizinhos mais próximos em um conjunto de treinamento. O valor de K determina o número de vizinhos considerados para a classificação e uma distância apropriada (como distância euclidiana ou de Manhattan) é usada para medir a proximidade. A etiqueta do novo ponto de dados é então atribuída com base na etiqueta predominante entre seus K vizinhos.

# Item do edital: Técnica de classificação

Como especialista em provas da banca Cesgranrio, a Técnica de Classificação envolve a ordenação de candidatos por meio de pontuações obtidas em etapas do concurso, como prova objetiva, discursiva e títulos. Os candidatos são classificados em ordem decrescente de pontuação, sendo que aqueles com pontuações mais elevadas são considerados mais bem classificados. Essa técnica é utilizada para determinar a aprovação ou reprovação dos candidatos, bem como a ordem de convocação para as etapas subsequentes do concurso.

# Item do edital: Avaliação de modelos de classificação: treinamento

Schles. GcX Schles. Gc> X: GcX

# Item do edital: Avaliação de modelos de classificação: teste

O teste de classificação avalia o desempenho dos modelos de classificação ao prever classes para novos dados usando métricas como precisão, revocação e pontuação F1. Ele envolve dividir os dados em conjuntos de treinamento e teste, treinar o modelo no conjunto de treinamento e avaliar seu desempenho no conjunto de teste para determinar sua capacidade de generalizar para dados desconhecidos. Os resultados do teste orientam a seleção do melhor modelo para uma tarefa de classificação específica.

# Item do edital: Avaliação de modelos de classificação: validação

A validação de modelos de classificação é crucial para avaliar sua confiabilidade e desempenho. Envolve dividir o conjunto de dados em conjuntos de treinamento e teste, treinar o modelo no conjunto de treinamento e avaliar seu desempenho no conjunto de teste. Métricas como acurácia, precisão, revocação e pontuação F1 são usadas para quantificar o desempenho e identificar o melhor modelo para uma determinada tarefa de classificação. A validação cruzada e as técnicas de amostragem podem ser usadas para melhorar a robustez e reduzir o viés na avaliação.

# Item do edital: Avaliação de modelos de classificação: validação cruzada

A validação cruzada é uma técnica estatística usada para avaliar o desempenho dos modelos de classificação ao dividir os dados em vários subconjuntos (dobras). Cada dobra é usada como conjunto de teste enquanto as demais são usadas como conjunto de treinamento. O modelo é treinado e testado em cada dobra, e os resultados são agregados para fornecer uma estimativa do desempenho geral do modelo. Essa abordagem ajuda a prevenir a sobreajuste e fornece uma avaliação mais robusta do desempenho do modelo em diferentes subconjuntos de dados.

# Item do edital: Avaliação de modelos de classificação: métricas de avaliação - matriz de confusão

A Matriz de Confusão é uma métrica de avaliação fundamental para modelos de classificação. Ela resume o desempenho do modelo em termos de Verdadeiros Positivos (VPs), Falsos Positivos (FPs), Verdadeiros Negativos (VNs) e Falsos Negativos (FNs). A partir desses valores, são calculadas outras métricas, como Precisão (VPs/(VPs+FPs)), Sensibilidade (VPs/(VPs+FNs)), Especificidade (VNs/(VNs+FPs)) e F-Score (2\*(Precisão\*Sensibilidade)/(Precisão+Sensibilidade)), que fornecem uma compreensão abrangente do desempenho do modelo.

# Item do edital: Avaliação de modelos de classificação: acurácia

A acurácia é uma métrica fundamental na avaliação de modelos de classificação. Ela mede a proporção de previsões corretas feitas pelo modelo em um conjunto de dados de teste. Um modelo com alta acurácia é capaz de distinguir com precisão entre as classes alvo, o que é crucial para aplicações práticas, como classificação de spam, diagnóstico médico e reconhecimento de padrões. A acurácia é calculada dividindo o número de previsões corretas pelo número total de previsões feitas. Embora seja uma métrica simples e intuitiva, a acurácia pode ser enganosa em conjuntos de dados desequilibrados, onde uma classe domina as outras.

# Item do edital: Avaliação de modelos de classificação: precisão

A precisão, uma métrica de avaliação de modelos de classificação, mede a proporção de previsões corretas em relação ao número total de previsões. Expressa a capacidade do modelo de identificar corretamente observações pertencentes a uma classe específica, fornecendo insights sobre a eficácia do modelo na separação de classes. Uma alta precisão indica que o modelo é capaz de discriminar com precisão entre classes, enquanto uma baixa precisão sugere que o modelo pode estar fazendo classificações incorretas ou confundindo diferentes classes.

# Item do edital: Avaliação de modelos de classificação: revocação

A revocação é uma métrica de avaliação de modelos de classificação que mede a proporção de amostras positivas que são corretamente identificadas pelo modelo. Ela complementa a precisão, que mede a proporção de todas as amostras que são corretamente identificadas, fornecendo uma compreensão do desempenho do modelo na identificação de amostras verdadeiramente positivas. Modelos com alta revocação são valiosos quando a identificação de todos os casos positivos é crucial, minimizando assim os falsos negativos.

# Item do edital: Avaliação de modelos de classificação: F1-score

O F1-score é uma métrica de avaliação de modelos de classificação que considera tanto a precisão (proportion of true positives out of all actual positives) quanto a revocação (proportion of true positives out of all predicted positives) do modelo. Ele é calculado como a média harmônica da precisão e da revocação, variando de 0 a 1, onde um F1-score de 1 indica um modelo perfeito e um F1-score de 0 indica um modelo que não faz previsões corretas.

# Item do edital: Avaliação de modelos de classificação: curva ROC.

A Curva ROC (Receiver Operating Characteristic) é uma ferramenta gráfica utilizada em Avaliação de Modelos de Classificação para avaliar a capacidade do modelo em distinguir entre observações positivas e negativas. Ela representa a taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1 - especificidade) para diferentes limiares de classificação. A área sob a curva ROC, conhecida como AUC-ROC, varia de 0 a 1, onde valores próximos a 0,5 indicam desempenho aleatório e valores próximos a 1 indicam desempenho perfeito. A Curva ROC permite comparar o desempenho de diferentes modelos de classificação e selecionar o melhor para uma determinada tarefa de classificação.

# Item do edital: Técnicas de regressão: Redes neurais para regressão

As redes neurais são um tipo de modelo de aprendizado de máquina que pode ser usado para problemas de regressão. Elas são compostas por camadas de nós interconectados que processam informações e aprendem relacionamentos complexos nos dados. Ao treinar uma rede neural para regressão, ela aprende a mapear variáveis de entrada para valores contínuos, fazendo previsões precisas mesmo quando os dados são altamente não lineares ou contêm ruído.

# Item do edital: Árvores de decisão para regressão

As Árvores de Decisão para Regressão são uma extensão das Árvores de Decisão Clássicas, projetadas para tarefas de previsão contínua. Elas dividem o espaço de características de forma recursiva, selecionando o atributo que minimiza o erro quadrático médio (MSE) ou outras medidas de erro de regressão a cada divisão. O processo de divisão continua até que um critério de parada seja atendido, resultando em uma árvore com nós internos representando pontos de divisão e nós terminais com valores de previsão contínuos.

# Item do edital: Máquinas de vetores de suporte para regressão

As máquinas de suporte a veto (MSV) para regressão são um método de aprendizado de máquinas supervisionado que visa encontrar a função de regressão não linear que melhor se ajusta aos pares de entrada-saídas fornecidas. As MSV usam um truque do kernel para mapear os pontos de entrada em um espaço de dimensão superior, onde uma função linear pode ser ajustada ao conjunto de pontos mapeado. Essa função linear no espaço de dimensão superior corresponde à uma função não linear no espaço de entrada original, fornecento uma regressão poderosa e flexível.

# Item do edital: Ajuste de modelos dentro e fora de amostra e overfitting.

No ajuste de modelos estatísticos, o overfitting ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, resultando em um desempenho ruim em dados novos ou externos. O ajuste dentro da amostra avalia o desempenho do modelo nos dados de treinamento, enquanto o ajuste fora da amostra avalia o desempenho em um conjunto de dados separado. Para evitar overfitting, técnicas como validação cruzada e regularização são empregadas para encontrar um equilíbrio entre complexidade do modelo e ajuste aos dados.

# Item do edital: Técnicas de agrupamento:

Técnicas de agrupamento são métodos estatísticos que dividem um conjunto de dados em grupos menores e distintos (chamados clusters) com base em suas semelhanças. Essas técnicas buscam identificar padrões subjacentes nos dados, agrupando observações com atributos semelhantes e distinguindo-as daquelas com atributos diferentes. Ao fazer isso, as técnicas de agrupamento ajudam a identificar estruturas ocultas nos dados, fornecer insights sobre as relações entre as variáveis e facilitar a tomada de decisões informada.

# Item do edital: Agrupamento por partição

Agrupamento por partição é uma técnica de agrupamento que divide um conjunto de dados em subconjuntos (partições) com base em uma variável específica. Ele atribui pontos de dados a partições que otimizam um critério especificado, como distância ao centroide ou variância dentro da partição. Esse método é particularmente útil para problemas de agrupamento em que as partições são determinadas por discrepâncias de dados preexistentes ou quando a estrutura de partição é conhecida antecipadamente.

# Item do edital: Agrupamento por densidade

O agrupamento por densidade é um método de agrupamento não supervisionado que divide os dados em grupos com densidades semelhantes. Ele cria clusters de pontos de dados densos, identificando áreas com concentrações mais altas de pontos. Cada cluster é separado por áreas de baixa densidade, conhecidas como "ruído". O agrupamento por densidade não requer um número específico de clusters, permitindo a descoberta de padrões naturais nos dados. Ele pode lidar com dados com diferentes números de dimensões e é útil para encontrar grupos arbitrários e com formas complexas. O resultado do agrupamento por densidade é uma partição dos dados, onde cada ponto é atribuído a um cluster ou ao conjunto de dados de "ruído".

# Item do edital: Agrupamento hierárquico.

O Agrupamento Hierárquico é um método estatístico de classificação não supervisionada que cria uma hierarquia de grupos (dendrograma) com base na similaridade ou distância entre os dados. Ele começa com cada ponto de dados como um grupo individual e, em seguida, iterativamente agrupa os grupos mais semelhantes até que um único grupo contenha todos os pontos de dados. O dendrograma resultante fornece uma representação visual dos relacionamentos entre os grupos e ajuda a identificar padrões e outliers dentro dos dados.

# Item do edital: Técnica de redução de dimensionalidade: Seleção de características (feature selection)

A seleção de características é uma técnica de redução de dimensionalidade que visa identificar e selecionar as características mais informativas e relevantes em um conjunto de dados. Ela ajuda a reduzir a complexidade e melhorar a eficiência dos modelos preditivos, removendo características redundantes, irrelevantes ou ruidosas. Ao selecionar apenas as características mais significativas, é possível obter melhor desempenho do modelo, interpretabilidade aprimorada e menor tempo de treinamento.

# Item do edital: Técnicas de redução de dimensionalidade: análise de componentes principais (PCA – principal component analysis).

A Análise de Componentes Principais (PCA) é uma técnica de redução de dimensionalidade que identifica e extrai os componentes principais de um conjunto de dados, transformando dados de alta dimensionalidade em um subespaço de menor dimensionalidade. Ao projetar os dados nos componentes principais, a PCA reduz a complexidade mantendo ao máximo a variância. Essa técnica é amplamente utilizada para visualização e análise de dados, permitindo a identificação de padrões, tendências e redução de ruído em conjuntos de dados de alta dimensionalidade.

# Item do edital: Processamento de linguagem natural: Normalização textual

A normalização textual é uma técnica de Processamento de Linguagem Natural (PNL) que padroniza o texto para melhorar sua legibilidade e compreensão. Ela envolve a remoção de pontuação, caracteres especiais, espaços em branco extras e maiúsculas desnecessárias. A normalização garante que o texto seja consistente e fácil de analisar, tornando-o mais adequado para tarefas como classificação de texto, geração de resumos e tradução automática.

# Item do edital: Processamento de linguagem natural: stop words

As stop words são palavras comuns que aparecem com alta frequência em um texto, mas fornecem pouco ou nenhum significado contextual. Elas são removidas durante o pré-processamento do Processamento de Linguagem Natural (PLN) para aprimorar a eficácia da extração de recursos e modelagem. A remoção das stop words reduz o número de recursos e elimina o ruído, focando nos termos mais informativos e significativos. Ao remover palavras como "o", "a", "de", "para" e "com", o PLN pode analisar o texto com mais precisão, identificando padrões e extraindo insights mais profundos.

# Item do edital: Processamento de linguagem natural: estemização

A estemização, um processo essencial no Processamento de Linguagem Natural, reduz palavras a seus radicais ou formas básicas. Isso envolve remover sufixos e prefixos comuns para capturar as propriedades semânticas comuns entre palavras flexíveis, como "correr", "correu" e "correndo". A estemização melhora a eficiência da indexação e recuperação, reduz a dispersão semântica e facilita a análise linguística, permitindo que os sistemas de PNL processem informações de forma mais precisa e abrangente.

# Item do edital: Processamento de linguagem natural: lematização

A lematização é um processo de Processamento de Linguagem Natural (PNL) que reduz as palavras à sua forma canônica ou lema. Em vez de manter várias formas flexionais (por exemplo, correu, corra, corri), a lematização retorna a forma base da palavra (correr). Isso ajuda a normalizar o texto, agrupar palavras variavelmente escritas e melhorar a precisão de tarefas de PNL, como indexação, recuperação de informações e tradução automática.

# Item do edital: Processamento de linguagem natural: análise de frequência de termos

Análise de Frequência de Termos no Processamento de Linguagem Natural

A análise de frequência de termos identifica a ocorrência de palavras ou frases específicas em um parágrafo. Ela envolve:

- Tokenização: Dividir o texto em unidades menores (tokens, geralmente palavras).

- Limpeza: Remover símbolos, pontuação e palavras comuns (stopwords).

- Stemming/Lematização: Reduzir as palavras às suas formas básicas (raiz).

- Contagem: Contar a ocorrência de cada token único.

Esta análise ajuda a:

- Identificar termos-chave ou conceitos importantes.

- Detectar padrões ou tendências no texto.

- Criar modelos de classificação ou agrupamento de texto.

# Item do edital: Rotulação de partes do discurso: part-of-speech tagging

A rotulação de partes do discurso é uma tarefa de processamento de linguagem natural que envolve atribuir rótulos gramaticais a cada palavra em uma sequência de texto. Esses rótulos indicam a função sintática da palavra na frase, como substantivo, verbo, adjetivo ou preposição. A rotulagem de partes do discurso é um passo crucial em muitas aplicações de processamento de linguagem natural, como análise sintática, reconhecimento de padrões de fala e tradução automática.

# Item do edital: Modelos de representação de texto: N-gramas

Os N-gramas são modelos de representação de texto que dividem o texto em sequências de N palavras. Eles são usados para capturar padrões e relações de palavras em contexto. Ao dividir um texto em bigramas (sequências de duas palavras) ou trigramas (sequências de três palavras), os N-gramas ajudam na compreensão da linguagem, processamento de linguagem natural e classificação de texto.

# Item do edital: modelos vetoriais de palavras: CBOW

O modelo CBOW (Continuous Bag-of-Words) é um modelo vetorial de palavras que prevê uma palavra de destino com base em suas palavras de contexto. Ele funciona treinando uma rede neural para prever a palavra de destino dada um janela fixa de palavras de contexto que a rodeiam. Ao aprender a prever palavras com base em seu contexto, o modelo cria representações vetoriais para as palavras que capturam seus significados semânticos e relacionamentos. A principal vantagem do CBOW é sua eficiência computacional, pois treina em palavras individuais em vez de pares de palavras como no modelo Skip-Gram.

# Item do edital: modelos vetoriais de palavra: Skip-Gram

O Skip-Gram é um modelo vetorial de palavra que prediz o contexto de uma palavra (palavras ao redor) dada a própria palavra. Ele maximiza a probabilidade logarítmica do contexto condicional à palavra, aprendendo representações vetoriais de palavras que capturam semelhanças semânticas. Cada palavra é representada como uma incorporação, que é um vetor numérico que codifica suas relações com outras palavras no vocabulário. As incorporações aprendidas permitem que o modelo preveja as palavras de contexto de uma palavra de entrada, refletindo assim sua proximidade semântica.

# Item do edital: modelos vetoriais de palavra: GloVe

O GloVe (Global Vectors for Word Representation) é um modelo de vetor de palavra que captura as relações semânticas e sintáticas entre palavras. Ele é treinado em um grande corpus de texto e utiliza coocorrências globais para aprender representações vetoriais que codificam propriedades semânticas e sintáticas. Ao contrário de outros modelos como o Word2Vec, o GloVe incorpora informações globais da matriz de coocorrência, o que permite que ele aprenda vetores que representam com mais precisão a similaridade semântica e a analogia entre as palavras.

# Item do edital: modelos vetoriais de documentos: booleano

Os modelos vetoriais de documentos booleanos representam documentos como vetores de bits, onde cada bit indica a presença ou ausência de um termo específico no documento. Cada documento é atribuído um vetor de tamanho fixo, e o valor de cada entrada do vetor é 1 se o termo correspondente aparecer no documento, ou 0 caso contrário. Esses modelos são simples e eficientes, mas podem ser suscetíveis a ruídos e não consideram a frequência ou peso dos termos.

# Item do edital: modelos vetoriais de documentos: TF

Os modelos vetoriais de documentos TF (Termo Frequência) convertem documentos em vetores numéricos, representando a frequência com que cada termo aparece no documento. O peso de um termo é calculado como o número de ocorrências do termo dividido pelo comprimento total do documento. Os modelos TF ajudam a identificar a relevância dos termos e a calcular a similaridade entre os documentos, o que é útil na recuperação de informações, classificação de texto e mineração de dados.

# Item do edital: modelos vetoriais de documentos: TF-IDF

O TF-IDF (Frequência de Termos-Frequência Inversa de Documentos) é um modelo vetorial de documentos amplamente utilizado para representar conteúdo textual. Ele mede a importância de um termo em um documento específico em relação a um conjunto de documentos no corpus. O TF (Frequência de Termos) conta a recorrência de um termo no documento, enquanto o IDF (Frequência Inversa de Documentos) pondera termos comuns no corpus, reduzindo sua influência nos vetores de documentos. O TF-IDF atribui pesos altos a termos que ocorrem frequentemente em um documento, mas raramente em outros documentos, resultando em representações vetoriais que capturam o conteúdo único e informativo de cada documento.

# Item do edital: modelos vetoriais de documentos: média de vetores de palavras

Modelos Vetoriais de Documentos

Conceito:

Representam documentos como vetores no espaço vetorial, onde cada dimensão corresponde a uma palavra do vocabulário do documento.

Construção:

- Divide o documento em parágrafos.

- Cria um vetor para cada parágrafo, onde cada componente é a frequência da palavra correspondente no parágrafo.

- Calcula a média dos vetores de parágrafos para obter o vetor do documento.

Vantagens:

- Simples e eficiente de construir.

- Gera vetores esparsos, economizando espaço de memória.

- Captura relações de frequência entre palavras.

Limitações:

- Ignora a ordem das palavras e a estrutura da frase.

- Pode ser sensível à pontuação e erros ortográficos.

- Não considera informações semânticas ou contextuais.

# Item do edital: modelos vetoriais de documentos: Paragraph Vector

Parágrafo Vector é um modelo vetorial de documento que representa um parágrafo como um vetor numérico fixo. Ele captura as características semânticas do parágrafo, codificando as relações entre palavras e frases no contexto. O modelo usa uma rede neural para aprender essas representações vetoriais, que podem ser usadas para tarefas de processamento de linguagem natural, como classificação de texto, busca semântica e inferência de tópicos.

# Item do edital: Métricas de similaridade textual - similaridade do cosseno

A similaridade do cosseno é uma métrica textual que mede a similaridade entre dois documentos comparando o ângulo entre seus vetores, que representam as frequências das palavras nos documentos. Ele calcula o ângulo do cosseno entre os vetores, que varia de 0 (perfeitamente semelhante) a 1 (completamente diferente), fornecendo uma medida da proximidade entre os textos.

# Item do edital: Métricas de similaridade textual distância euclidiana

A distância euclidiana é utilizada como uma métrica de similaridade textual que calcula a distância entre dois vetores representando os textos. Cada vetor é composto pelas frequências de termos (por exemplo, palavras ou frases) nos respectivos textos. A distância euclidiana representa a magnitude da diferença entre esses vetores, com uma distância menor indicando maior similaridade. Quanto maior a distância entre os vetores, menor é a similaridade entre os textos.

# Item do edital: Métricas de similaridade textual similaridade de Jaccard

A similaridade de Jaccard é uma métrica de similaridade para medir a semelhança entre dois conjuntos de itens. É calculada dividindo o número de itens comuns aos dois conjuntos pelo número total de itens em ambos os conjuntos. Um valor alto de similaridade de Jaccard indica que os dois conjuntos são muito semelhantes, enquanto um valor baixo indica que são muito diferentes. No contexto da análise de texto, a similaridade de Jaccard pode ser usada para medir a semelhança entre dois documentos de texto, comparando os conjuntos de palavras ou n-gramas que aparecem em ambos os documentos.

# Item do edital: Métricas de similaridade textual distância de Manhattan

A distância de Manhattan é uma métrica de similaridade comumente usada em aprendizado de máquina e processamento de linguagem natural. Mede a distância entre dois vetores calculando a soma das diferenças absolutas de seus elementos correspondentes. Por exemplo, para vetores x = [2, 4] e y = [3, 6], a distância de Manhattan seria |2-3| + |4-6| = 4. A distância de Manhattan é uma métrica fácil de calcular e interpretar, pois representa a distância "em linha reta" entre dois pontos em um espaço n-dimensional.

# Item do edital: Métricas de similaridade textual coeficiente de Dice.

O Coeficiente de Dice é uma métrica de similaridade textual que mede a semelhança entre duas sequências de caracteres. Ele calcula o número de pares de caracteres compartilhados entre as duas sequências e divide esse número pelo total de pares de caracteres em ambas as sequências. O resultado é um valor entre 0 e 1, onde 0 indica nenhuma similaridade e 1 indica similaridade perfeita. O Coeficiente de Dice é uma medida robusta e versátil que pode ser usada para comparar textos de diferentes comprimentos e ordenações.

# Item do edital: Redes neurais convolucionais

As redes neurais convolucionais (CNNs) são um tipo de rede neural profunda que são especialmente adequadas para processamento de dados estruturados, como imagens. As CNNs usam camadas convolucionais para extrair características hierárquicas dos dados de entrada, seguidas por camadas totalmente conectadas para classificação ou regressão. Ao convolver filtros através dos dados de entrada, as CNNs identificam padrões e características locais, permitindo a aprendizagem de representações de alto nível que são cruciais para tarefas como reconhecimento de imagens, processamento de linguagem natural e visão computacional.

# Item do edital: Redes neurais recorrentes.

Redes Neurais Recorrentes (RNNs) são modelos de aprendizado profundo que processam sequências de dados ordenados, como texto ou dados temporais. Eles têm uma estrutura oculta que mantém informações do passado, permitindo que aprendam padrões e dependências ao longo do tempo. As RNNs são usadas em diversas aplicações, como processamento de linguagem natural, previsão de séries temporais e reconhecimento de fala, pois podem capturar sequências contextuais e relacionamentos de longa distância nos dados.

# Item do edital: Scikit-learn

O Scikit-learn é uma biblioteca Python de aprendizado de máquina de código aberto que fornece uma ampla gama de algoritmos e ferramentas para tarefas de aprendizado de máquina, desde pré-processamento de dados e engenharia de recursos até modelagem preditiva e avaliação. Ele oferece uma interface consistente e fácil de usar, permitindo que os usuários se concentrem no desenvolvimento de modelos em vez de implementar algoritmos do zero. Com uma vasta coleção de métodos validados, otimizações eficientes e suporte a vários aprendizados, o Scikit-learn é amplamente utilizado em diversos domínios para resolver problemas de aprendizado de máquina com eficiência.

# Item do edital: TensorFlow

TensorFlow é uma biblioteca de código aberto desenvolvida pelo Google para o treinamento e implantação de modelos de aprendizado de máquina. Ela oferece uma estrutura flexível e escalonável que permite aos desenvolvedores criar e treinar redes neurais e outros algoritmos de aprendizado de máquina. O TensorFlow suporta execução paralela em CPUs e GPUs, facilitando o treinamento de modelos complexos em grandes conjuntos de dados.

# Item do edital: PyTorch

Pytorch é uma biblioteca de aprendizado de maquina de código aberto construída para Python, projetada para facilitar o desenvolvimento e o treinamento de redes neurais. Sua interface intuitiva permite que cientistas e pesquisadores dedados criem e personalizem facilmente arquiteturas de redes neurais complexas. Com recursos como tensor computing acelerado, diferenciação automática e suporte para vários dispositivos, o Pytorch agiliza o processo de desenvolvimento e otimização de modelos, tornando-o uma ferramenta popular para criação de aplicativos avançados de aprendizado deep learning e inteligência artificial.