

## Relatório: Card 5- Estatística p/ Aprendizado de Máquina (I)

Jefferson korte junior

### Descrição da atividade:

Neste card aprendi sobre estatística, tipos de dados, construir gráficos usando bibliotecas como matplotlib para facilitar a visualização e até mesmo perceber padrões entre os dados. A estatística é muito bom para aprendizado de máquina pois é a partir dela que podemos ver e melhorar o nosso modelo.

### Sobre tipos de dados temos:

**Dados numéricos:** São dados quantitativos, como altura e preço. Eles podem ser discretos, aqueles que são inteiros e geralmente contabilizam ocorrências, ou contínuo, aqueles que podem ter infinitos valores.

**Dados categóricos:** São dados qualitativos, como legendas de gráficos, gênero.

**Dados ordinal:** Uma mistura de numérico com categórico, podem ser avaliações de aplicativos por exemplo. É um dado categórico com significado numérico.

**Média:** É a soma de um conjunto qualquer de valores divididos pela quantidade de valores.

### *Média Aritmética*

$$\frac{x_1 + \dots + x_n}{n}$$



**Mediana:** É o valor central de um conjunto de dados ordenados, é bom ser usado quando a números muito discrepantes do resto, assim fazendo mais sentido usar a mediana em vez de usar a média.

Nº de elementos par - a mediana é a média dos 2 valores centrais = 6,5

3	4	6	7	8	9
---	---	---	---	---	---

Nº de elementos ímpar - a mediana é o valor central = 7

3	4	6	7	8	9	13
---	---	---	---	---	---	----

**Moda:** É o valor visto com mais frequência no conjunto.

**Variação:** É a distância de cada valor do conjunto em relação ao valor central.

**Desvio padrão:** O desvio padrão é uma medida que indica o quanto os valores de um conjunto de dados variam em relação à média.

**Exemplo de desvio padrão:**

**Exemplo:** Se a média das notas em um exame for 70 e o desvio padrão for baixo, isso significa que a maioria dos alunos tem notas próximas de 70. Um desvio padrão alto indicaria uma maior variação nas notas dos alunos.

**Função de densidade de probabilidade:** a função de densidade de probabilidade é uma ferramenta essencial para modelar e entender como as probabilidades estão distribuídas em variáveis aleatórias contínuas, um exemplo é uma empresa querendo ver como seus salários estão distribuídos na empresa, isso ajuda a identificar padrões de desigualdade salarial.

**Função massa de probabilidade:** Uma função que fala a probabilidade de uma variável aleatória assumir cada um dos seus possíveis valores.

**exemplo** é entender quais momentos um grupo de clientes da netflix estão propensos a cancelar o plano.

**Distribuição normal:** É uma distribuição de dados simétrico em relação a média, muito utilizado em machine learning para tentar visualizar a distribuição normal dos erros.

**Distribuição binomial:** A distribuição binomial é uma ferramenta fundamental para analisar experimentos com dois resultados possíveis, sucesso ou fracasso.  
**exemplo** é para analisar qual a probabilidade de um produto de uma fábrica tem defeito.

**Distribuição de poisson:** A distribuição de Poisson é um modelo estatístico utilizado para descrever a probabilidade de um determinado número de eventos ocorrer em um intervalo fixo de tempo ou espaço.

**Exemplo:** Se uma linha de produção tem uma média de 2 defeitos por hora, podemos usar a distribuição de Poisson para calcular a probabilidade de encontrar exatamente 0, 1, 2, ou mais defeitos em uma hora.

**Percentil:** São uma medida estatística utilizada para entender a posição relativa de um valor dentro de um conjunto de dados.

**Exemplo:** Imagine que temos um grupo de 10 crianças e medimos suas alturas em centímetros: 110, 115, 120, 122, 125, 128, 130, 133, 135 e 138. Queremos saber qual é o 40º percentil dessas alturas. O 40º percentil significa que 40% das crianças têm alturas menores ou

iguais a aproximadamente 122 cm. Em outras palavras, 60% das crianças são mais altas que 122 cm. Esse exemplo simples mostra como podemos usar percentis para entender a distribuição de dados de forma clara e intuitiva.

**Assimetria:** É uma medida da distribuição dos valores de um conjunto de dados em relação à média. Ela indica se a distribuição dos dados é simétrica, tem uma cauda mais longa à direita (assimetria positiva) ou tem uma cauda mais longa à esquerda (assimetria negativa).

**Exemplo:** Imagine uma turma escolar onde a maioria dos alunos tem entre 11 e 12 anos, mas há alguns alunos que repetiram de ano e têm entre 13 e 14 anos. A distribuição das idades será assimétrica à direita porque a maioria das idades está concentrada entre 11 e 12 anos, mas há uma cauda se estendendo para as idades maiores (13 e 14 anos).

**Curtose:** A curtose é uma medida que complementa a assimetria na descrição da forma de uma distribuição. Ao analisar a curtose, podemos obter insights sobre a concentração dos dados

**Covariância:** A covariância é uma medida estatística que indica o grau em que duas variáveis variam juntas. É uma extensão do conceito de variância para duas dimensões. Se a covariância entre duas variáveis é positiva, isso significa que, à medida que uma variável aumenta, a outra tende a aumentar também. Se a covariância é negativa, significa que, à medida que uma variável aumenta, a outra tende a diminuir.







**Probabilidade Condicional:** A probabilidade condicional é a probabilidade de um evento ocorrer, dado que outro evento já aconteceu.








**Teorema de Bayes:** ajuda a calcular a probabilidade de uma hipótese ser verdadeira, considerando as informações que já temos.

### Conclusão:

Neste card aprendi bastante sobre análise de dados utilizando bibliotecas apresentadas no curso. Aprendi bastante sobre visualizar dados de várias maneiras utilizando a biblioteca matplotlib, o que facilitou a compreensão e padrões nos dados. Todos os meus códigos construindo gráficos no curso usando a biblioteca matplotlib estão no meu Git-hub.

### Referências:

1.  [Types of Data \(Numerical, Categorical, Ordinal\)](#)
2.  [Mean, Median, Mode](#)
3.  [Activity Using mean, median, and mode in Python](#)
4.  [Activity Variation and Standard Deviation](#)
5.  [Probability Density Function: Probability Mass Function](#)
6.  [Common Data Distributions \(Normal, Binomial, Poisson, etc\)](#)

7.  **Activity** [Percentiles and Moments](#)
8.  **Activity** [A Crash Course in matplotlib](#)
9.  **Activity** [Advanced Visualization with Seaborn](#)
10.  **Activity** [Covariance and Correlation](#)
11.  **Exercise** [Conditional Probability](#)
12.  **Exercise Solution** [Conditional Probability of Purchase by Age](#)
13.  **Bayes' Theorem**

Essas foram minhas referências sobre este relatório.