**Semana 4: Estatística Básica, Machine Learning (Árvore de Decisão) e Falácias em Dados**

**Estatística, estatística descritiva, inferência estatística**

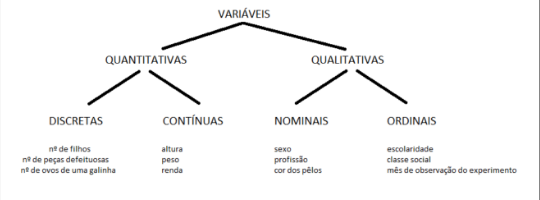
*Estatística* é o estudo de como coletar, organizar, analisar e interpretar dados. Técnicas estatísticas são usadas em todas as instâncias da sociedade, por exemplo, numa pesquisa de eleição pega-se apenas uma amostra de intenções de votos de uma parcela da população, e a partir de estudos estatísticos podemos ter a resposta para a população geral.

*Estatística descritiva* é uma forma de resumir os dados que você possui, calculando a média, mínimo, máximo e desvio padrão, por exemplo.Ela apenas extrai informações da sua amostra, não mais do que isso.

*Inferência estatística* é o conjunto de técnicas para tirar conclusões dos dados.É aqui que a estatística consegue tirar conclusões e deduzir propriedades das informações extraídas pela estatística descritiva.

**Variáveis: Definição e classificação**

É um atributo, uma característica de uma pessoa ou coisa, por exemplo a altura, peso, etc.



*Variáveis quantitativas*: São as numéricas, mensuráveis em uma escala quantitativa, como peso, altura e quantidade de filhos, essas variáveis, por sua vez, podem ser divididas em

1. Discretas: Quando a quantidade de valores possíveis do domínio formam um conjunto finito ou enumerável de números(quantidade de filhos, número de ovos que uma galinha bota, etc
2. Contínuas: São aquelas que em um determinado intervalo, podem adotar quaisquer valores, números reais(altura, peso, renda salarial, etc)

*Variáveis qualitativas*: São variáveis cujo valor não se expressa numericamente, como a cor dos olhos, sexo, etc.

1. Nominais: São aquelas que não possuem ordenação, uma ordem de melhor para pior, maior para menor, etc. Sexo é um exemplo disso, não há como ordenar um sexo.
2. Ordinal: São aquelas que podemos aplicar alguma ordem, como grau de escolaridade, classe social, que podemos claramente achar uma ordem de melhor para pior.

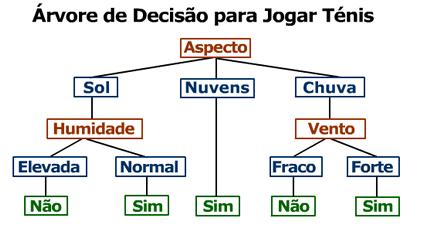
**Falácias de dados**

1. Cherry Picking: É quando pegamos apenas alguns dados que apoiam nossas hipóteses/crenças e ignoramos os demais, evitando achar dados que venham a quebrar nossa hipótese
2. P-Hacking:O p-valor é uma medida estatística para medir a significância dos resultados em um teste de hipótese, para obter um valor alto, muitos analistas forçam tratamentos para obter um p-valor significante, entretanto é perfeitamente possível ter um p-valor significante para relações que não fazem o menor sentido, ou seja, não podemos levar apenas ele como base.
3. Viés de sobrevivência: Acontece quando uma análise é feita a partir uma parte da base de dados ao invés da mesma completa, podendo levar a conclusões totalmente equivocadas, um caso famoso foi do militar que era responsável por diminuir o abatimento de caças americanos, para isso, ele pegou os caças que voltaram das missões, os unicos disponiveis para análise e revestiu os lugares que eles mais tomam tiros, isso foi equivocado pois se eles levaram tiros naquele lugar X e voltaram, possivelmente é por conta daquele lugar não limitar funções vitais do caça, ele deveria olhar para os locais onde os caças abatidos levaram tiros pois são esses os locais graves de se levar tiros, uma vez que foram abatidos.
4. Efeito Cobra: Ocorre é dado um incentivo para ação X, mas acaba aumentando a quantidade de ações Y, diferente do esperado. Por exemplo, o governo na Colômbia teve uma política ao exército de matar mais terroristas, tendo uma remuneração com base no número de mortes, o resultado foi esse exército matando mais inocentes para ganho pessoal.
5. Falsa Causalidade: Correlação não implica em causa, pode acontecer dessa correlação ser apenas uma consciência ou pode acontecer de ter uma terceira variável envolvida para sustentar a causalidade. Por exemplo, imagine que em sua empresa seja descoberto que sucesso tenha uma alta correlação com álcool, em primeira análise não faz muito sentido, mas quando adentramos mais no estudo de caso, pode ser que as pessoas que consomem mais álcool são mais sociáveis, e sociabilidade tem uma forte correlação com sucesso, tendo assim uma terceira variável para se analisar, e por que não uma quarta? quinta?...

**Machine Learning: Árvore de Decisão**

* **Introdução**

O modelo de árvore de decisão é basicamente um fluxograma de IF e ELSES sobre as features da base de dados, chegando ao final em alguma predição. Ele é bastante intuitivo, fácil de codificar, com poucas premissas e com resultado interpretável. É muito usado em concessões de crédito, diagnósticos médicos, análises de investimento, etc.



É importante deixar claro que o modelo de árvore de decisão sempre prioriza colocar as features que melhor separam a base de dados nas primeiras condições, tendo assim um modelo mais assertivo.Por exemplo, vamos supor que uma base de dados para prever se um cliente vai pagar o empréstimo tenha 3 variáveis preditoras, vamos ver qual delas separa melhor os dados:



Com base nessas perguntas o modelo vai escolhendo as perguntas com a menor incerteza possível, buscando grupos com maior separabilidade.

* **Vantagens e desvantagens da árvore de decisão**

1. Vantagens:

* Fáceis de interpretar e visualizar
* Capturam padrões não lineares(aqueles que não podem ser previstos por uma linha reta em um gráfico)
* Requer pouco pré-processamento(ex: Scaling)
* Pode ser usado como feature engineering(Selecionar, modificar e criar novas features)
* Não possui premissas sobre distribuições

1. Desvantagens

* Sensíveis a ruídos(Outliers podem gerar overfitting)
* Pequenas variações podem resultar em árvores totalmente diferentes(modelos de bagging e boosting, como random forest e XG Boosting, respectivamente, amenizam esse problema)
* Cuidado ao lidar com dados desbalanceados