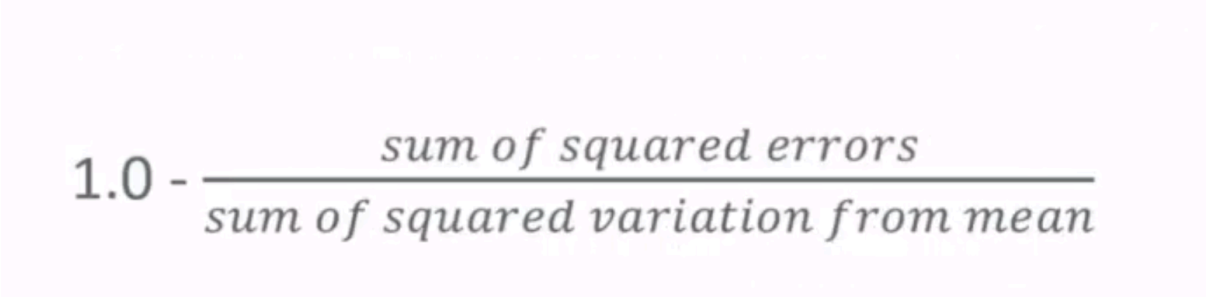
Relatório 12 - Prática: Predição e a Base de Aprendizado de Máquina (II)

Higor Miller Grassi

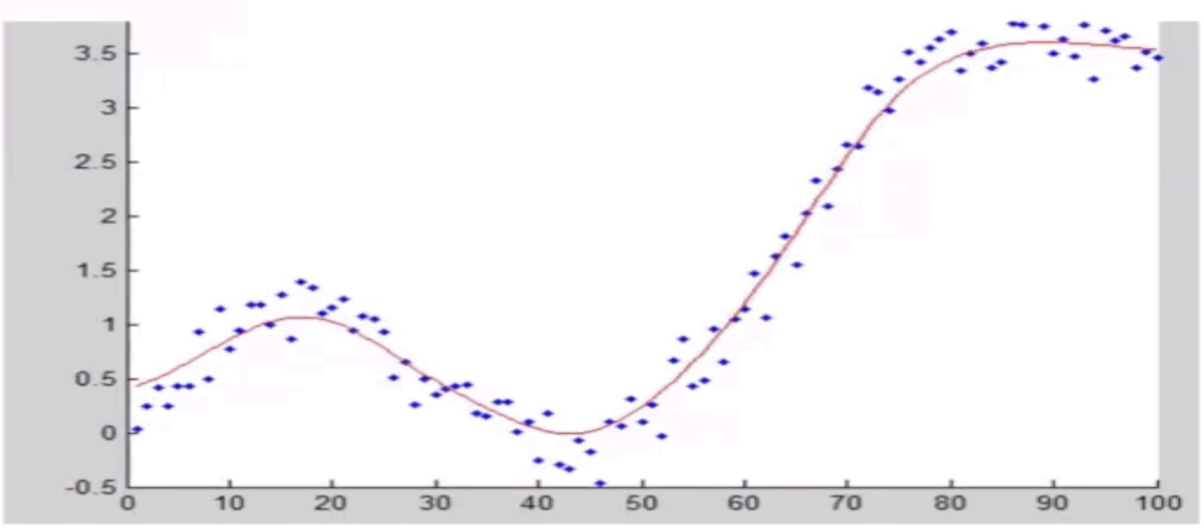
Descrição da Atividade: A atividade abordou conceitos e práticas fundamentais do aprendizado de máquina, explorando técnicas como regressão linear, regressão polinomial e regressão múltipla, com foco na construção de modelos preditivos eficientes.

A regressão linear é uma técnica utilizada para prever resultados futuros, com base na ideia de traçar uma linha que melhor represente os dados disponíveis, onde nesse processo é feito utilizando o método dos mínimos quadrados, que visa minimizar o erro entre os pontos e a linha, ou seja, ajustando a linha de modo que a soma dos quadrados das distâncias entre os pontos e a linha seja a menor possível. No exemplo do vídeo, para estimar o peso de uma pessoa com 1,50m de altura, a regressão linear tenta usar os dados de peso de pessoas com alturas já conhecidas, ajustando a linha de modo que o erro seja mínimo.



O **r-quadrado (Rˆ2)** é uma métrica que indica o quanto da variação dos dados é explicada pelo modelo, variando de 0 a 1, sendo que 0 significa que o modelo não explica nada da variação dos dados (o ajuste é péssimo), enquanto 1 indica um excelente ajuste, onde o modelo consegue capturar todas as variações dos dados. Em um gráfico, se os pontos ficarem próximos da linha de regressão, o Rˆ2 será alto; se os pontos estiverem dispersos, o Rˆ2 será baixo.

A **regressão polinomial** é uma extensão da regressão linear e permite ajustar curvas aos dados, ao invés de apenas uma linha reta, sendo útil quando há uma relação não linear entre as variáveis, mas é importante ter cuidado, pois ao tentar ajustar o modelo de forma muito precisa (colocando muitos graus no polinômio), o modelo pode acabar se ajustando de forma excessiva aos dados, o que pode não refletir uma relação real e levar a um **overfitting** (ajuste excessivo), ou seja, o modelo passa a funcionar bem apenas para os dados de treino, mas falha em generalizar para novos dados.



A **regressão múltipla** é um tipo de regressão que permite prever uma variável dependente com base em várias variáveis independentes, onde no vídeo é citado o exemplo de ao tentar prever o preço de um carro, podemos considerar diversas variáveis, como a quilometragem, a idade do carro e o número de portas, sendo um modelo multivariado, ou seja, leva em consideração múltiplas variáveis, atribuindo um peso a cada uma delas de acordo com sua importância na previsão do valor final.

A **seleção de recursos** é um passo importante na modelagem de dados, onde se escolhe quais variáveis são realmente relevantes para o modelo, evitando o uso de variáveis irrelevantes, que podem confundir o modelo e reduzir sua precisão.

Os **modelos de múltiplos níveis** são usados quando os dados têm uma estrutura hierárquica, ou seja, quando existem diferentes níveis de influências que afetam uma variável, como falado no video, no contexto da saúde, pode ser influenciada pela saúde das células e dos órgãos em seu corpo, que por sua vez são influenciados por fatores externos como a cidade onde a pessoa vive e o nível de estresse a que está exposta. Esses modelos são importantes para entender como os diferentes fatores influenciam em diferentes níveis, e como isso afeta o comportamento geral da variável que estamos tentando prever.

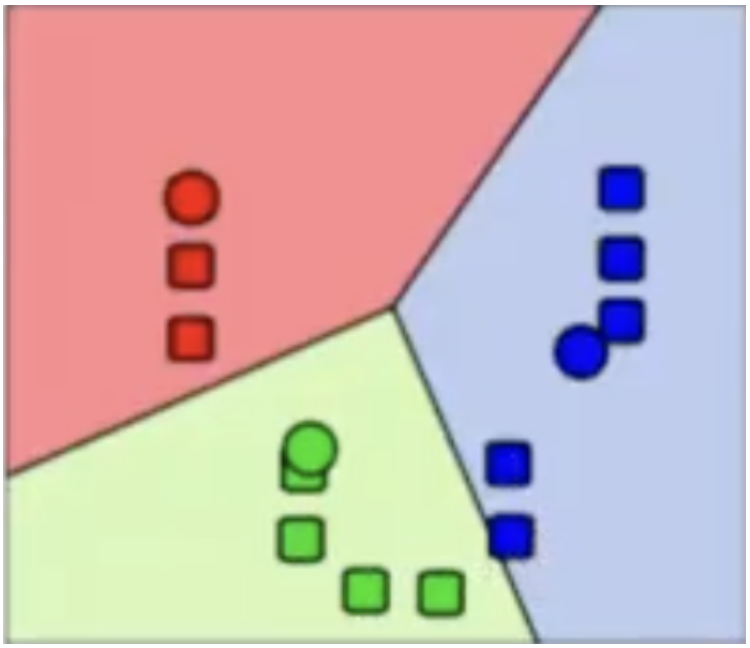
O aprendizado não supervisionado é uma abordagem em que o modelo recebe apenas um conjunto de dados sem rótulos ou respostas associadas com o maior objetivo de identificar padrões ou agrupamentos dentro desses dados. Um exemplo utilizado em vídeo é a análise de agrupamento de formas geométricas, sendo que o modelo deve organizar os dados em grupos com base em semelhanças entre eles, sem saber de como essas semelhanças devem ser definidas. Esse tipo de aprendizado é útil principalmente quando se deseja explorar e descobrir informações ou padrões desconhecidos dentro de um conjunto de dados.

Já o **aprendizado supervisionado** envolve a utilização de um conjunto de dados rotulados anteriormente, assim o modelo aprende a partir desses dados, buscando estabelecer relações entre as características dos dados de entrada e as categorias ou valores de saída, permitindo que o modelo faça previsões para novos dados, com base nas correlações aprendidas durante o treinamento. No aprendizado supervisionado, o processo é dividido em duas fases, treinamento, onde o modelo aprende a partir dos dados rotulados, e teste, onde o modelo é avaliado usando dados novos.

Métodos bayesianos são frequentemente utilizados em classificadores de spam, por exemplo, eles analisam um conjunto de dados, levando em consideração

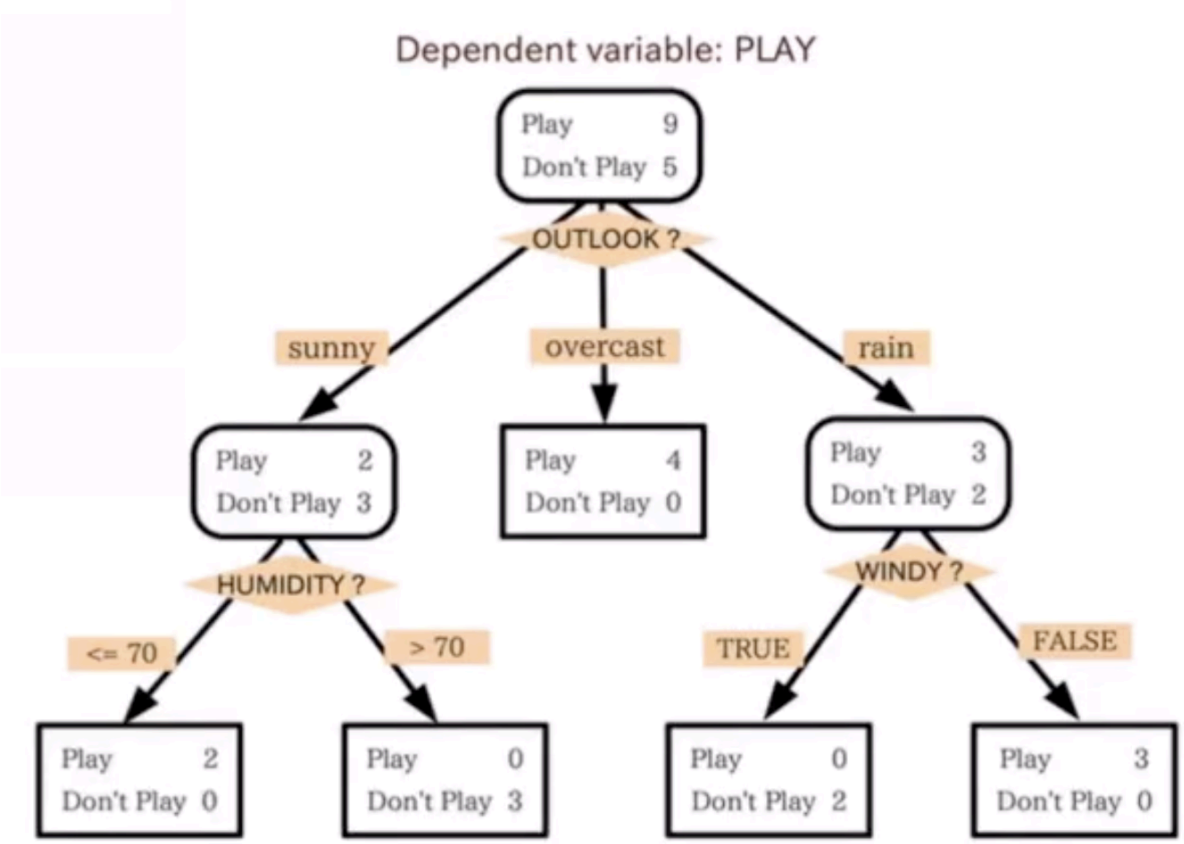
a probabilidade de uma palavra aparecer em mensagens de spam e não spam. O processo envolve o cálculo da probabilidade condicional das palavras nos diferentes tipos de mensagens, tratando as palavras como características independentes entre si, o que simplifica o cálculo e a classificação, transformando assim cada palavra em um número e realizando a contagem para uma maior precisão.

O **K-Means** é uma técnica de aprendizado de máquina não supervisionada (explicado anteriormente)usada para dividir dados em **k** grupos, onde **k** é um número previamente definido. O algoritmo começa selecionando **k** pontos aleatórios, chamados de centroides, que representam o centro de cada grupo, em seguida, ele atribui cada ponto de dados ao centroide mais próximo, formando clusters(grupos), após essa atribuição, os centroides são recalculados como a média dos pontos dentro de cada cluster, e o processo se repete, isto continua até que os centroides não se movam mais, indicando que a solução se estabilizou. A escolha do valor de **k** é importante, pois um valor baixo ou alto demais pode resultar em agrupamentos imprecisos, sendo a melhor abordagem é começar com um valor baixo de **k** e aumentar progressivamente até que a mudança na qualidade dos grupos se estabilize. Como representado na imagem a seguir, sinalizando os centroides juntamente aos grupos:



A entropia é uma medida usada para quantificar a desordem de um conjunto de dados, quanto mais desorganizado ou imprevisível for o conjunto, maior será a entropia, no exemplo citado no vídeo, se você tem um grupo de animais e todos são da mesma espécie, a entropia desse grupo será baixa, pois há pouca diversidade, no entanto, se cada animal for de uma espécie diferente, a entropia será alta, porque a diversidade é maior e o grupo se torna mais "desorganizado".

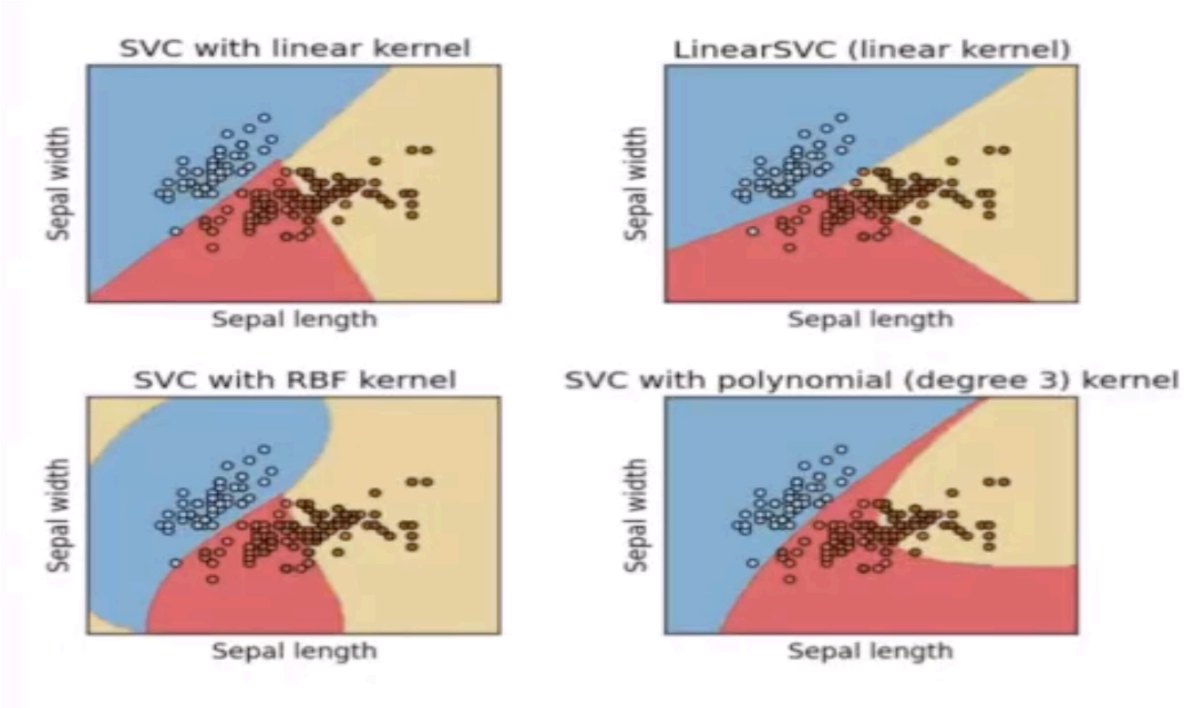
A Árvores de decisão fornece um fluxograma de como fazer algumas decisões, tendo uma variável dependente, como se citado, se deveria ou não brincar lá fora hoje tendo em vista o clima, que pode ter vários aspectos diferente que podem impactar na decisão, e quando vc tem uma decisão dessas que depende de vários outros atributos(variáveis) uma árvore de decisão é muito boa para tal. Mas o seu grande problema é o seu overfitting em casos em que que precisamos de uma classificação correta para novas informações não vistas antes, pois elas são específicas para os dados de treinamentos, e para lidar com este problemas, existe uma técnica chamada florestas aleatórias, tendo diferentes maneiras para várias árvores de decisões diferentes (bootstrap aggregating)



O XGBoost é uma implementação otimizada do método de *gradient boosting*, que constroi árvores sequenciais corrigindo os erros das anteriores. Ele utiliza regularização para evitar *overfitting*, lida automaticamente com valores ausentes e permite validação cruzada integrada para avaliar o benefício real de cada iteração. Oferece suporte a treinamento incremental, possibilitando salvar e retomar o progresso, além de permitir funções de perda personalizadas para maior flexibilidade, implementa *Tree Pruning* para interromper ramificações desnecessárias, resultando em árvores mais eficientes e com esses recursos, o XGBoost é uma solução poderosa para aprendizado supervisionado, combinando eficiência, precisão e flexibilidade.

Máquinas de vetores de suporte, uma maneira avançada de agrupar ou classificar dados de dimensões superiores, com muitos recursos diferentes. Ele utiliza um truque do kernel para realmente encontrar esses vetores de suporte e

existem diferentes kernels que pode ser usado, na imagem seguinte é utilizado diferentes métodos com taxas de complexidades diferentes de SVC:



Conclusão: Abordou de forma abrangente sobre os fundamentos do aprendizado de máquina, destacando tanto métodos básicos como avançados, sendo assim possível aprender a importância de técnicas como regressão linear, polinomial e múltipla, bem como sua aplicação em diferentes contextos. A exploração de algoritmos supervisionados e não supervisionados, como K-Means, Árvores de Decisão, Florestas Aleatórias e XGBoost, demonstrou a versatilidade e o impacto dessas abordagens na análise e predição de dados, e com tudo isso, o estudo de conceitos como entropia, overfitting, seleção de recursos e métodos bayesianos trouxe uma compreensão mais aprofundada sobre os desafios e estratégias para otimizar modelos preditivos.