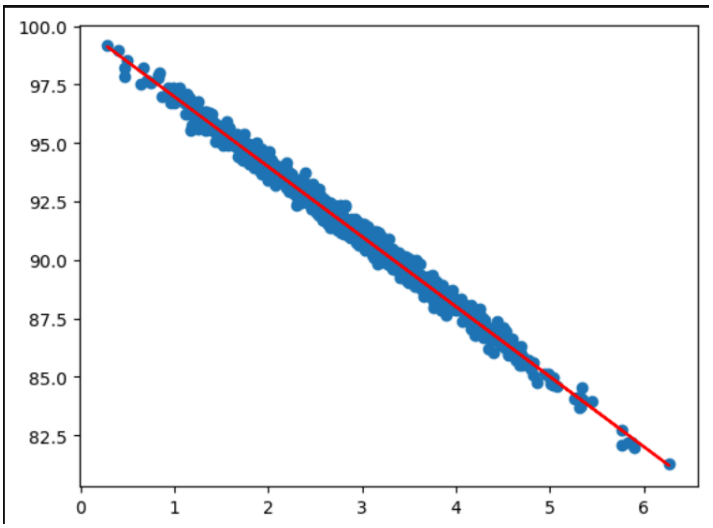
Relatório 11 - Prática: Predição e a Base de Aprendizado de Máquina (II)

Felippe Vernizze Sousa Nadolny

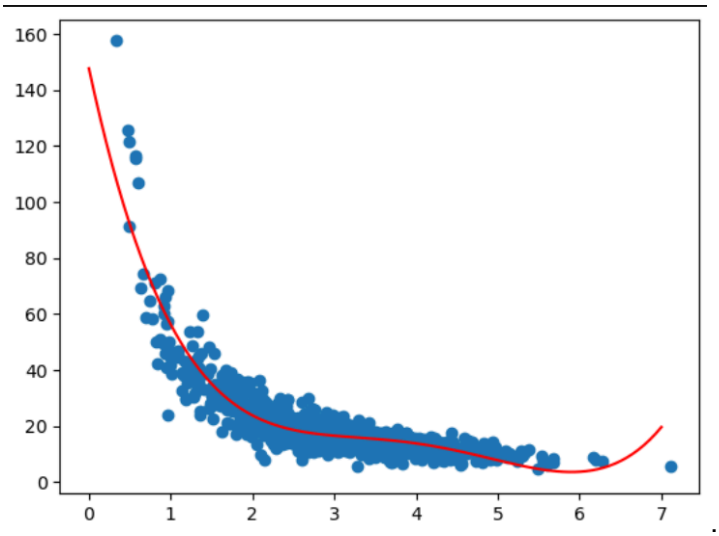
Descrição da atividade

Nesse card entendemos um pouco melhor sobre aprendizado supervisionado e não supervisionado, quando utilizá-los e como dividir dados em teste e treino para a máquina, além disso foram apresentados modelos preditivos utilizados com frequência na estatística como regressão linear, polinomial, múltipla e múltipla com níveis. Dentro desse escopo foram apresentados diversos métodos de classificação e análise de dados baseado em clusters, Bayes, árvores de decisão e máquinas de vetores de suporte.

Regressão Linear: Técnica de modelagem para prever uma variável dependente contínua com base em uma ou outra variável independente. Onde após determinada a linha de regressão, podemos estimar valores futuros e analisar seu grau de eficiência com a utilização do cálculo do R^2.



Regressão Polinomial: Extensão da regressão linear para capturar relações não lineares entre variáveis. A curva da regressão será determinada pela quantidade de variáveis estudadas.



Regressão Múltipla: Usa múltiplas variáveis independentes para prever valores, como o preço de um carro, considerando vários fatores. Como no exemplo, utilizando valores de quilometragem, ano de fabricação e etc.

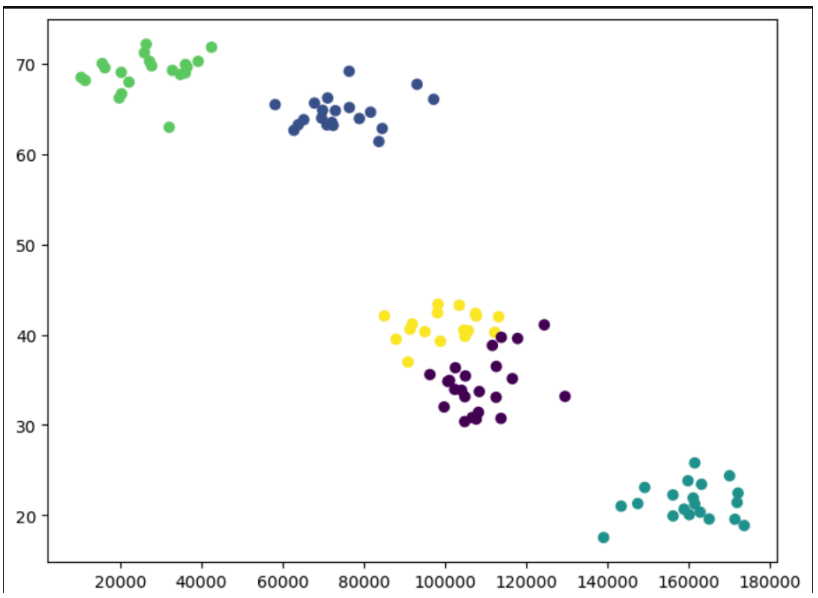
Modelos de Múltiplos Níveis: se baseia em dados hierárquicos observando variáveis em diferentes níveis. Podendo ter uma análise mais detalhada.

Treino / teste: Para prever valores futuros em um modelo é necessário analisar a melhor forma de divisão entre seus dados utilizados para treino e para teste, os dados para teste normalmente são cerca de 20% dos dados coletados. Enquanto os dados de treino são o restante dos dados, assim podemos utilizar tanto os dados para treinar o modelo quanto para validar ele.

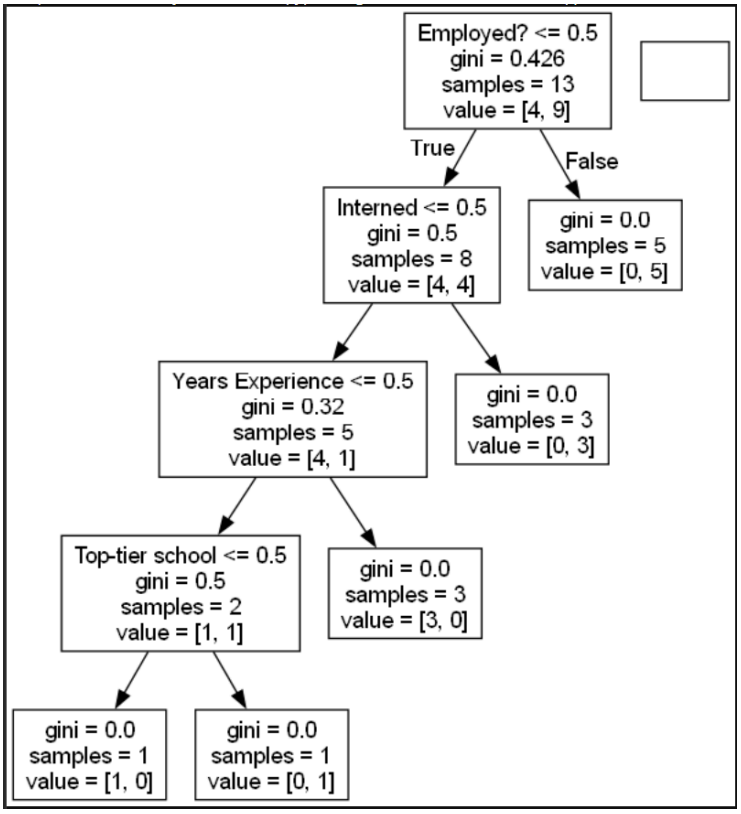
Métodos Bayesianos: Baseados no Teorema de Bayes, onde se utiliza de evidências anteriores para calcular a probabilidade de um evento ocorrer novamente.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

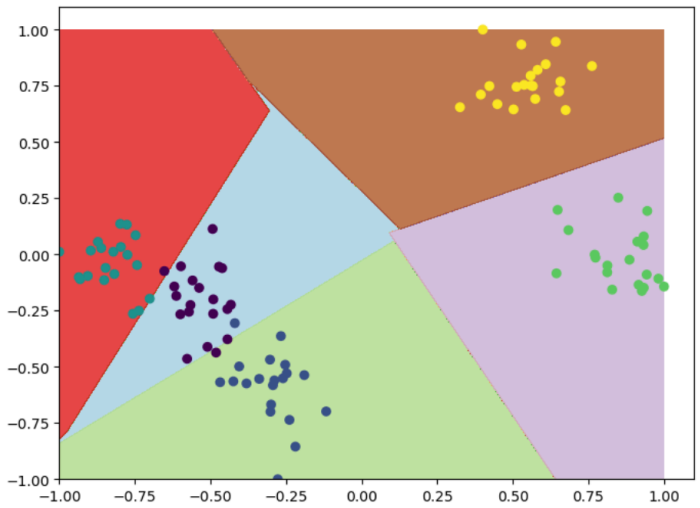
K-Means Clustering: um algoritmo não supervisionado que divide os dados em clusters com base nas similaridades entre os dados. Por fim, cada cluster possui um centróide que se move até que não haja mais mudanças significativas.



Árvores de Decisão: divide os dados em ramos e decide qual caminho deve ser tomado baseado em critérios, quanto mais se divide os dados, reduzimos a complexidade do problema até chegar em um resultado final. Esse método pode ser combinado com Ensemble Learning que agrupa diversos modelos de árvores de decisão melhorando seu resultado para o que se faça necessário.



Máquinas de Vetores de Suporte (SVM): Método de classificação que identifica o “hiperplano” ótimo que separa diferentes classes nos dados. Utiliza vetores de suporte para maximizar a margem entre classes, garantindo separação precisa.



Conclusões

Esses tópicos fornecem diversos métodos de análise de dados e machine learning desde o básico ao mais complexo, além de mostrar métodos para otimizar seus resultados com determinados modelos.