FIAP MBA+

MODELOS DE INTELIGÊNCIA ARTIFICIAL & MACHINE LEARNING

TURMA - 24IA

ANDREY FERREIRA DE ALMEIDA – RM: 344545
DIEGO COHEN MENDES ROSA – RM: 344154

1) PROBLEMA DE APRENDIZADO SUPERVISIONADO

Classificação de SPAM baseado em Aprendizado Supervisionado usando técnicas de Aprendizado de Máquina

INTRODUÇÃO

Segundo os pesquisadores, o uso da internet tem aumentado de maneira extensiva nos últimos anos e, nesse contexto, o e-mail tem sido uma ferramenta poderosa para comunicação e troca de informação.
Embora existam muitas vantagens nesse meio de comunicação, alguns problemas, como é o caso do SPAM, prejudicam seu uso eficiente.
O SPAM é um problema generalizado na internet e ele é tão barato de ser enviado que essas mensagens não solicitadas são enviadas a um grande número de usuários indiscriminadamente.
Assim, o problema identificado pelos autores, é que quando um grande número de mensagens de SPAM são recebidas, leva-se muito tempo para determinar o que é ou não é SPAM.
De acordo com os pesquisadores, existem muitos tipos de e-mail de SPAM, como anúncios com o objetivo de ganhar dinheiro ou vender algo, lendas urbanas com o objetivo de espalhar boatos e etc.
E afirma-se que em geral, o remetente de uma mensagem de SPAM objetiva anunciar produtos, serviços ou ideias para enganar os usuários com as suas informações privadas para entregar software malicioso ou causar falha temporária em um servidor de e-mail.

DATASET

O dataset utilizado pela pesquisa foi coletado do UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/spambase), que consiste em 4601 amostras de mensagens de e-mail com cerca de 58 atributos, que correspondem a:

- 48 atributos reais contínuos [0,100] do tipo word_freq_WORD = porcentagem de palavras no e-mail que correspondem a PALAVRA, ou seja, 100 * (número de vezes que a PALAVRA aparece no e-mail) / número total de palavras no e-mail. Uma "palavra" neste caso é qualquer sequência de caracteres alfanuméricos limitada por caracteres não alfanuméricos ou fim de sequência.
- 6 atributos reais contínuos [0,100] do tipo char_freq_CHAR] = porcentagem de caracteres no e-mail que correspondem a CHAR, ou seja, 100 * (número de ocorrências de CHAR) / total de caracteres no e-mail
- 1 atributo real [1,...] contínuo do tipo capital_run_length_average = comprimento médio de sequências ininterruptas de letras maiúsculas
- 1 atributo inteiro contínuo [1,...] do tipo capital_run_length_longest = comprimento da sequência ininterrupta mais longa de letras maiúsculas
- 1 atributo inteiro contínuo [1,...] do tipo capital_run_length_total = soma do comprimento de sequências ininterruptas de letras maiúsculas = número total de letras maiúsculas no e-mail
- 1 atributo de classe nominal {0,1} do tipo spam = denota se o e-mail foi considerado spam (1) ou não (0), ou seja, e-mail comercial não solicitado.

FONTE:

https://ieeexplore.ieee.org/abstract/document/5979035

D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," 2011 International Conference on Process Automation, Control and Computing, 2011, pp. 1-7, doi: 10.1109/PACC.2011.5979035.

1) PROBLEMA DE APRENDIZADO SUPERVISIONADO

Classificação de SPAM baseado em Aprendizado Supervisionado usando técnicas de Aprendizado de Máquina

Para o problema de classificação de mensagens de SPAM, foi utilizado o software WEKA (que possui um conjunto de algoritmos de Machine Learning implementado em JAVA) que possibilitou a realização de uma análise comparativa entre 3 algoritmos de classificação: MLP, J48 e Naive Bayes.

Durante a execução da pesquisa, afirma-se que os dados foram separados em 2 partes, uma usada para treinamento (essa parcela possui as features e as classificações para cada registro) que produziu o modelo de predição e a outra que foi usada para testar a acuracidade do modelo.

Também, foi utilizado o método k-fold cross validation durante a etapa de teste, cujo valor de k estimado equivaleu a 10.

Multilayer Perceptron (Rede Neural)

Um Perceptron Multicamada é um modelo de rede neural artificial de alimentação direta que mapeia conjuntos de dados de entrada em um conjunto de saída apropriada. O Perceptron Multicamada consiste em três ou mais camadas, uma camada de entrada e uma camada de saída com uma ou mais camadas ocultas. O aprendizado por retropropagação (backpropagation) ocorre no Perceptron, alterando os pesos da conexão após o processamento de cada parte dos dados, com base na quantidade de erro na saída em comparação com o resultado esperado.

J48 (Árvore de Decisão)

A J48 constrói árvores de decisão a partir de um conjunto de dados de treinamento usando o conceito de entropia da informação. A J48 examina o ganho de informação normalizado que resulta da escolha de um atributo para dividir os dados. Ela usa o fato de que cada atributo dos dados pode ser usado para tomar uma decisão dividindo os dados em subconjuntos menores. O classificador J48 classifica recursivamente até que cada folha seja pura, o que significa que os dados foram categorizados o mais próximo possível da perfeição.

Naive Bayes (Probabilístico)

O classificador Naive Bayes é um classificador probabilístico simples baseado na aplicação do teorema de Bayes com fortes hipóteses de independência. Um termo mais descritivo para o modelo de probabilidade subjacente seria "modelo de recurso independente". O indutor Naive-Bayes calcula probabilidades condicionais das classes dadas a instância e escolhe a classe com o posterior mais alto. Dependendo da natureza precisa do modelo de probabilidade, Classificadores Naive Bayes podem ser treinados de forma muito eficiente em um ambiente de aprendizado supervisionado.

FONTE:

https://ieeexplore.ieee.org/abstract/document/5979035

D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," 2011 International Conference on Process Automation, Control and Computing, 2011, pp. 1-7, doi: 10.1109/PACC.2011.5979035.

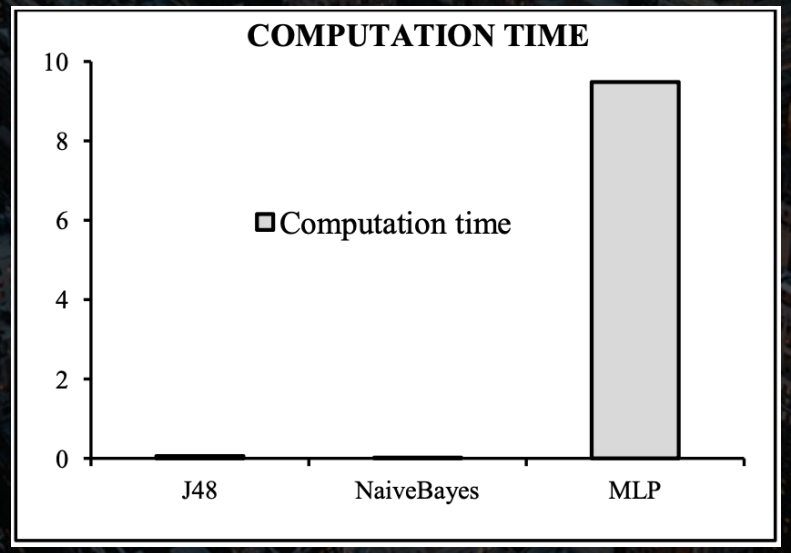
1) PROBLEMA DE APRENDIZADO SUPERVISIONADO

Classificação de SPAM baseado em Aprendizado Supervisionado usando técnicas de Aprendizado de Máquina

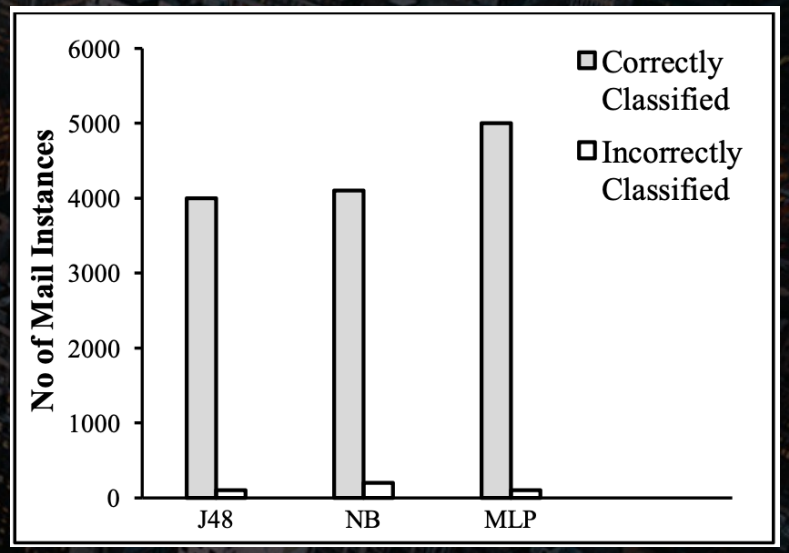
RESULTADOS

Evaluation Criteria	Classifiers		
	J48	Naïve Bayes	MLP
Time taken to build the Model	0.06	0.02	9.48
Correctly Classified Instances	4233	4095	4279
Incorrectly Classified Instances	368	506	322
Prediction Accuracy	92%	89%	93%

Resultados dos critérios de avaliação



Tempo de treinamento dos modelos



Distribuição das instâncias classificadas

CONCLUSÃO

Alguns desses classificadores para diferentes ferramentas de software podem esperar que os classificadores sejam consistentes, pois o teste foi feito no mesmo conjunto de dados. Classificador como Naive Bayes é um bom exemplo. No entanto, alguns classificadores como J48 e Simple Logistic têm um bom desempenho. Mas quando comparado com o MLP parece não ser melhor. Assim, de todas as perspectivas, o MLP teve o melhor desempenho em todos os casos e, portanto, pode ser considerado consistente.

FONTE:

https://ieeexplore.ieee.org/abstract/document/5979035

D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," 2011 International Conference on Process Automation, Control and Computing, 2011, pp. 1-7, doi: 10.1109/PACC.2011.5979035.

2) PROBLEMA DE APRENDIZADO NÃO SUPERVISIONADO

Segmentação de clientes usando Clusterização K-Means

INTRODUÇÃO

Segundo o artigo da plataforma KDNuggets, o problema de segmentação de clientes pode ser um meio poderoso para identificar necessidades insatisfeitas dos clientes e com base nisso, empresas podem desenvolver produtos e serviços exclusivamente atraentes.
Dentre os benefícios da segmentação de clientes, estão:

- Determinar o preço apropriado para o produto
- Desenvolver campanhas de marketing digital personalizadas
- Projetar uma estratégia de distribuição ideal
- Escolher recursos específicos do produto para implantação
- Priorizar o desenvolvimento de novos produtos

Para este problema, o desafio consiste em entender quem são os clientes de um supermercado de um shopping com base em alguns dados básicos obtidos através de um cartão de associação e prover essa informação para o time de marketing planejar a estratégia de acordo com esse público-alvo.

DATASET

O dataset no artigo foi coletado do Kaggle (https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python), que consiste em 200 amostras de informações básicas sobre os clientes, que correspondem a:

- CustomerID = ID Único assinado a cada cliente
- Gênero = Gênero do cliente
- Idade = Idade do cliente
- Annual Income (k$) = Renda anual do cliente
- Spending Score (1-100) = Pontuação atribuída pelo shopping com base no comportamento do cliente e natureza dos gastos

O artigo foi desenvolvido em Python e foram utilizadas as seguintes bibliotecas: scikit-learn, seaborn, numpy, pandas e matplotlib

Durante a execução a coluna CustomerID foi removida do conjunto de dados, pois para o autor ela era irrelevante para o estudo

Foi utilizado o método do cotovelo (Elbow method) para determinar o valor de clusters ideal

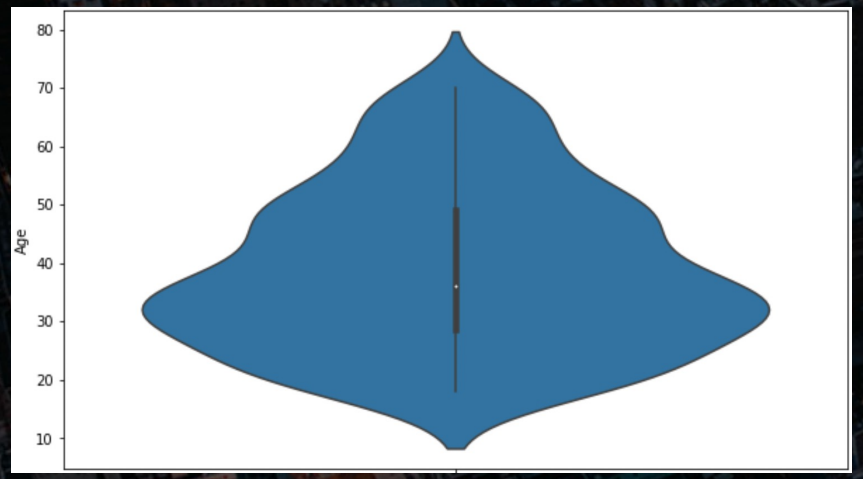
FONTE:

https://www.kdnuggets.com/2019/11/customer-segmentation-using-k-means-clustering.html

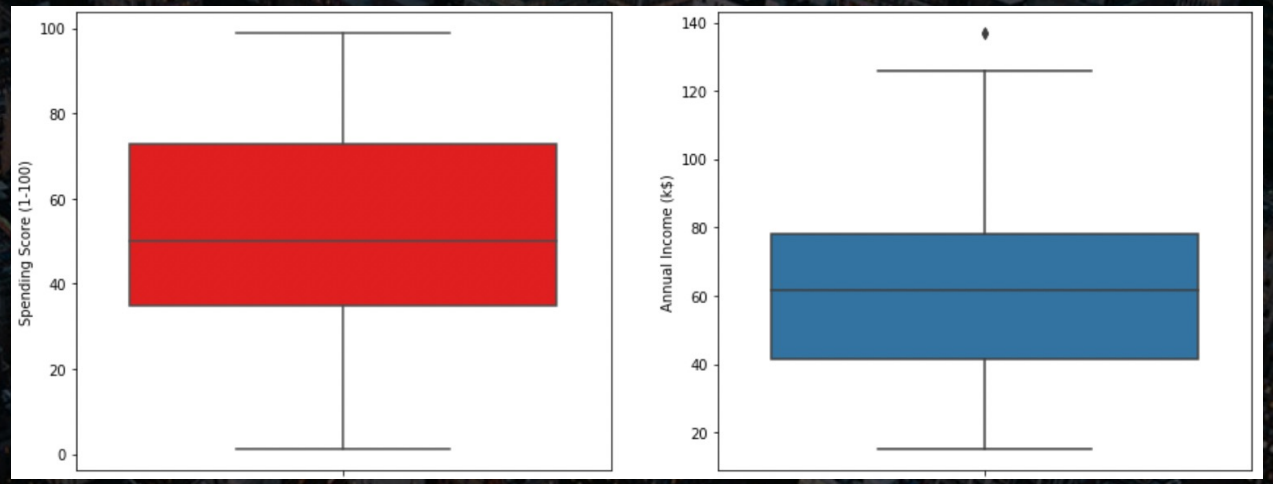
2) PROBLEMA DE APRENDIZADO NÃO SUPERVISIONADO

Segmentação de clientes usando Clusterização K-Means

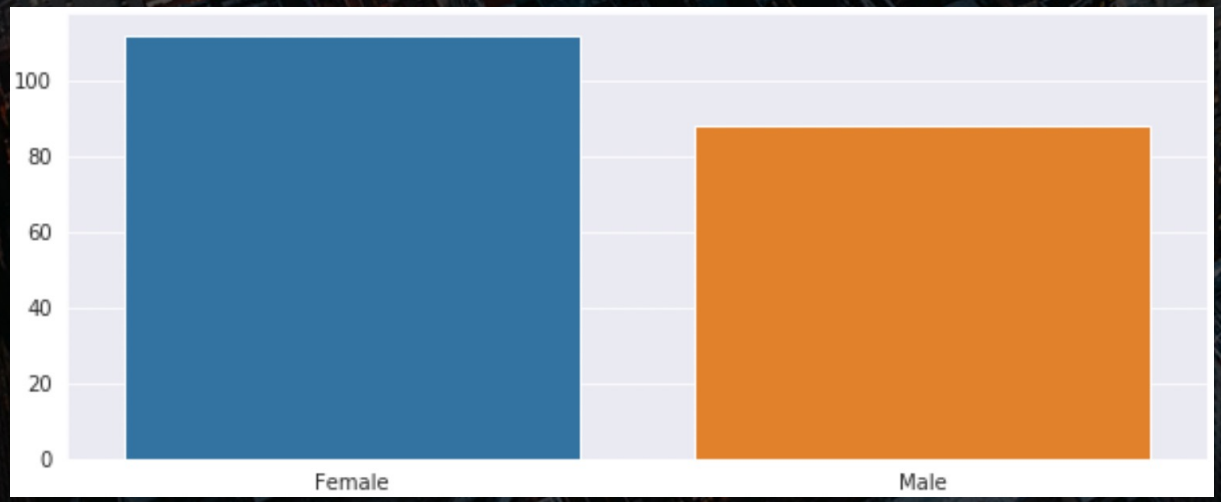
ANÁLISE DE DADOS



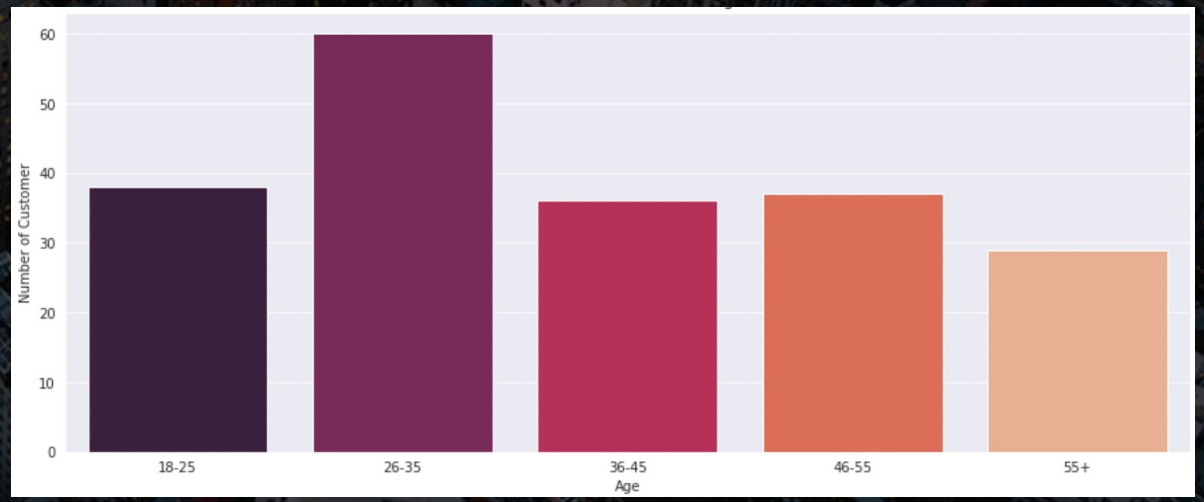
Distribuição de Age dos clientes



Comparativo entre Spending Score x Annual Income



Distribuição por Gender

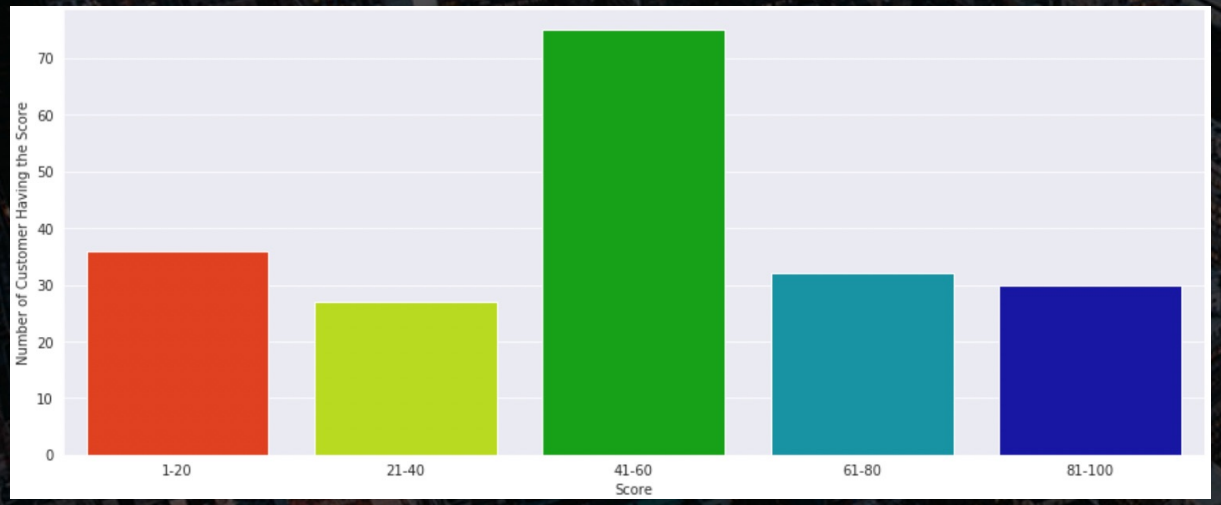


Distribuição de Clientes por range de Age

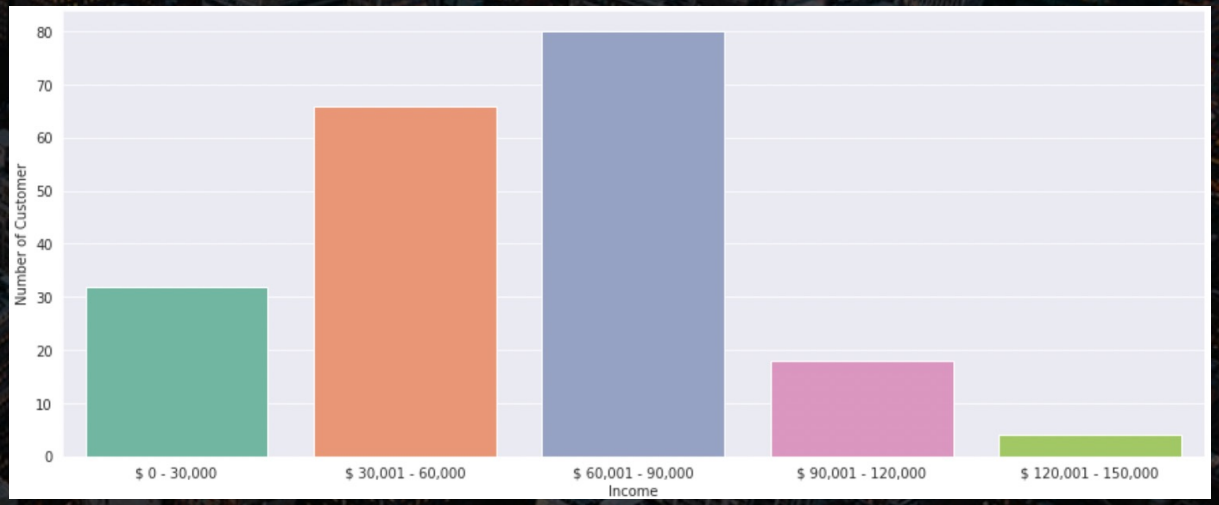
2) PROBLEMA DE APRENDIZADO NÃO SUPERVISIONADO

Segmentação de clientes usando Clusterização K-Means

ANÁLISE DE DADOS

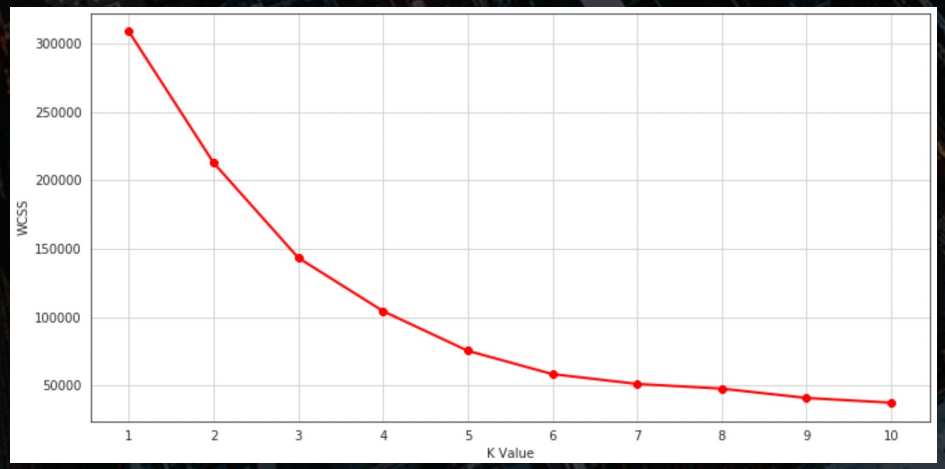


Distribuição de Clientes por range de Spending Score

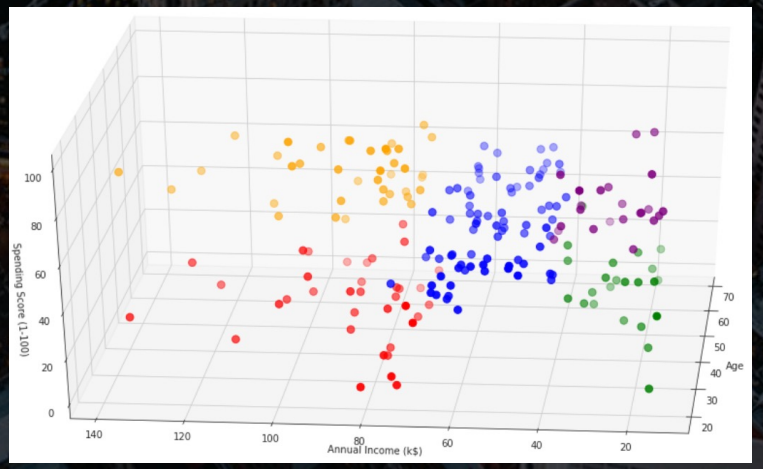


Distribuição de Clientes por range de Annual Income (k$)

RESULTADOS



Quantidade de k = 5



Distribuição espacial dos clusters

CONCLUSÃO

O objetivo dos K-Means é agrupar pontos de dados em subgrupos distintos que não se sobrepujam. Uma das principais aplicações do agrupamento de K-Means é a segmentação de clientes para obter uma melhor compreensão deles, o que, por sua vez, poderia ser usado para aumentar a receita da empresa.

3) PROBLEMA QUE PODE USAR ML, MAS NÃO PRECISA

Após pesquisas, percebemos que algumas referências citam heurísticas que podem ser utilizadas para se definir quando devemos deixar de utilizar o Machine Learning para resolver problemas.

Um exemplo constante nesse artigo do Medium (https://medium.com/geekculture/3-signs-you-dont-need-machine-learning-958199b30c34) são os Chatbots, por exemplo, pois eles nem sempre precisam ser capazes de responder a todas as perguntas. É possível detalhar o problema e prever se uma consulta atende as necessidades do cliente e, caso não aconteça, ele pode ser direcionado para a equipe de atendimento.

OBRIGADO!

FIAP MBA+