INF1771 - INTELIGÊNCIA ARTIFICIAL
TRABALHO 3 – APRENDIZADO DE MÁQUINA

O objetivo do Trabalho 3 na exploração de informações fornecidas através de planilhas para entender como reproduzir os fenômenos apresentados. Trata-se de um trabalho de **DataScience** que deve ser desenvolvido através da exploração dos dados e apresentação do que foi possível aprender ou entender através deles.

São disponibilizados dois datasets:

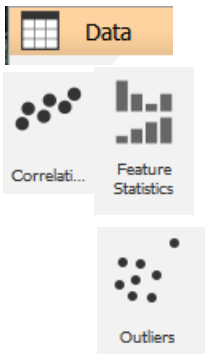
> O primeiro (bank-marketing) tipo apresenta atributos que representam o perfil de possíveis clientes de um banco que realiza uma campanha de marketing. Seu objetivo é captar novos clientes para um produto específico. Duas colunas são marcadas ao final do arquivo CSV – Y representa os clientes que aceitaram a oferta e Response aqueles que ao menos responderam esta campanha. – Trata-se de um problema de classificação
>
> O segundo apresenta atributos de um canal de youtube: views, likes, deslikes, comentários e estimativa de renda do vídeo. Nosso objetivo é estimar quais são os atributos que trazem maior renda e como podemos estimar a renda para o vídeo. – Trata-se de um problema de regressão.

Tarefas:

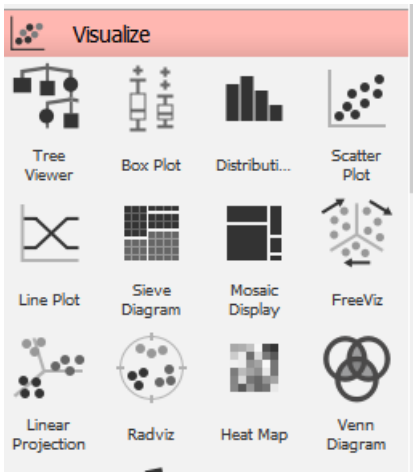
Para este trabalho não será necessário implementar nenhum dos algoritmos de *machine learning* e será totalmente desenvolvido utilizando a ferramenta "Orange". O grupo deverá conduzir as seguintes etapas:

a) Análise Exploratória de dados:

A primeira tarefa que podemos conduzir em dados que não conhecemos é analisar suas principais características. Para isto, utilizando medidas básicas da estatística tais como desvio padrão, média e correlação entre os atributos.



Também podemos verificar se o crescimento do dado segue um padrão, podendo ser linear, polinomial, exponencial ou possivelmente estocástico.



Importante: **O uso de destas informações não é obrigatório!** Utilize estas medidas e gráficos para ter pistas iniciais sobre as características dos dados.

Tree Viewer pode ser interessante para demonstrar o resultado de uma árvore de decisões. Distribution, Scatter Plot, Linear Projection e HeatMap podem ser usados para demonstrar a série de dados e suas relações.

A matriz de correlações pode auxiliar na detecção de possíveis redundâncias entre os dados e a covariância pode ajudar no desempate entre correlações.

b) Seleção de Atributos

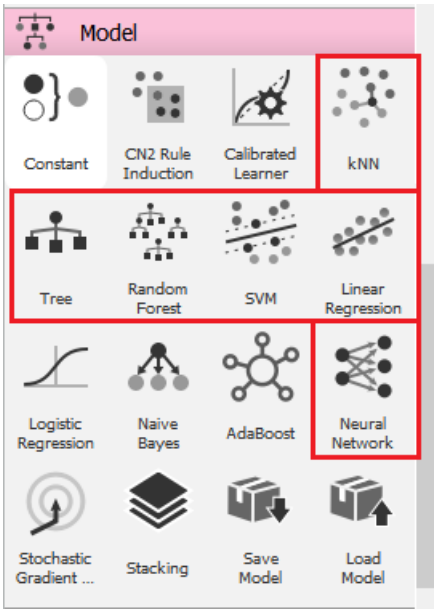
A segunda etapa será importante para remover as redundâncias e definir qual é o melhor conjunto de atributos para a criação do modelo de regressão.

Na aula de dimensionalidade, foram mencionados dois algoritmos: PCA e MDS. Utilize o PCA para encontrar a quantidade mínima de atributos que consegue explicar os dados sem perder informação. Lembre-se que o PCA também calcula pesos que são usados para projetar os dados na dimensão reduzida. Utilize esses pesos para verificar quais são os atributos mais relevantes que podem ser utilizados para a criação do modelo. Experimente apenas selecionar estes atributos ou utilizar o PCA como entrada para a regressão.

O MDS pode auxiliar a encontrar os limiares entre as informações. Utilize o MDS para encontrar (graficamente) como a relação entre os atributos influência nos resultados. O MDS é excelente para encontrar áreas separáveis linearmente de acordo com os resultados. Como trata-se de dados contínuos, perceba que o “Orange” faz a discretização utilizando faixas de valores.

c) Geração do Modelo

Nesta etapa, avalie o uso dos principais algoritmos de machine learning para encontrar aquele (ou aqueles) que encontra(m) o(s) melhor(es) resultado(s). Podem ser usados os seguintes algoritmos:

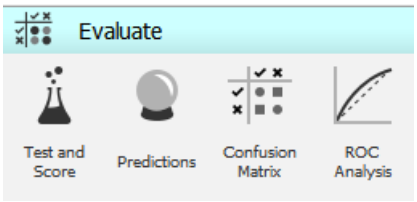


- Árvores de Decisão
- Random Forest
- Support Vector Machine (SVM)
- Rede Neural (usando backpropagation)
- Regressão Linear

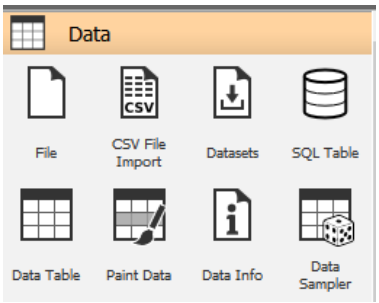
Obs: Apesar de destacado, K-NN não é utilizado para regressões (Por qual motivo?)

Lembre-se que podemos separar os dados em dois conjuntos (teste e treinamento) para avaliarmos o resultado da predição. Utilize o componente "test and score" para comparar os algoritmos de uma única vez.

Nesta etapa, procure verificar se podemos obter melhores resultados através do ajuste dos parâmetros dos algoritmos citados. O grupo deverá responder qual a melhor acurácia obtida, qual o melhor algoritmo (aquele que conseguiu a melhor acurácia) e explicar o resultado (por que ele é o melhor?).

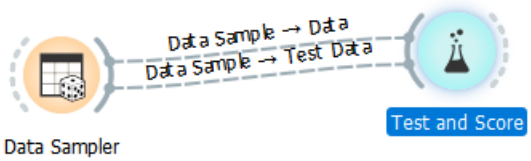


d) Carga de dados e separação de conjuntos

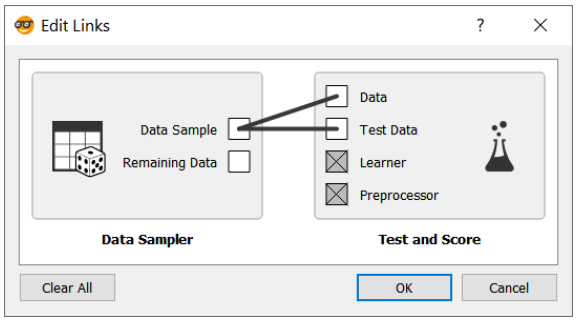


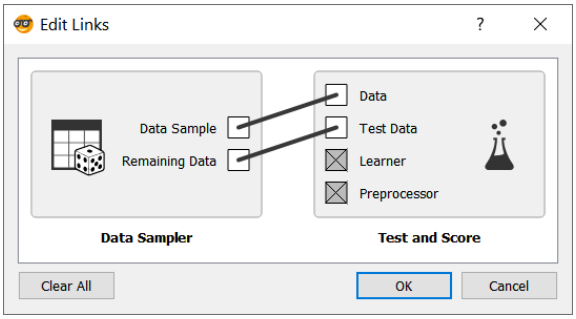
Para carregar os dados, podemos utilizar o componente "File" – que é responsável por carregar arquivos de excel. O componente "Data Table" auxilia na visualização dos dados, sejam eles oriundos do File, Datasets ou qualquer outro componente como PCA, Árvore de Decisão, Redes Neurais, etc...

Para separação dos conjuntos de treinamento e teste, utilize o componente “Data Sampler”. Atenção às saídas deste componente:

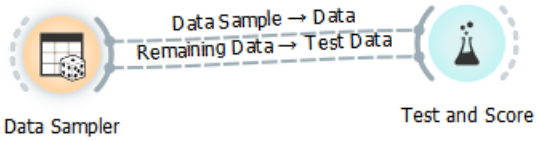


Perceba que quando conectamos duas vezes o data sampler ao Teste and Score, as duas linhas têm origem no mesmo conector: “Data Sample”. Neste exemplo, os dados de teste estão sendo usados tanto para o aprendizado, quanto para o teste. Devemos clicar duas vezes no segundo para reconfigurá-lo:





Puxe uma nova ligação de “Remaining Data” para “Test Data”. Isso sobrescreverá a ligação anterior. O resultado será este:



Utilize a separação 70%-30% para treinamento e teste.

Alguns dados são categóricos nominais (palavras). Para que sejam considerados na regressão, devem ser convertidos em ordinais. Nesta etapa (caso deseje utilizá-los) utilize o componente “Continuize”. Este componente pode associar um número inteiro (0, 1, 2, 3, ...) ao dado, criando um atributo artificial que pode ser utilizado para os cálculos.

Por exemplo:

- Masculino = 0, Feminino = 1 (ou vice versa)
- Baixo = 1, Médio = 2, Alto = 3
- Aprovado = 1, Reprovado = 0

Resumindo os objetivos da Atividade:

1. Procure entender como os dados se expressam e suas características
2. Selecione quais são os atributos relevantes
3. Crie os modelos de dados e analise os resultados
4. Apresente os resultados obtidos e o que pudemos aprender com eles (quais são as informações importantes e como podemos usá-las para prever os resultados futuramente)

Forma de Avaliação:

Será avaliado se todas as etapas do processo foram cumpridas corretamente. A avaliação também será baseada na **apresentação dos resultados** através de vídeo entregue pelo grupo.

Essa apresentação deverá explicar:

- Como o processo foi conduzido;
- Quais foram os experimentos realizados;
- Quais foram os resultados obtidos pelos modelos (basta apresentar no Orange "Test and Score" e explicar);
    - Comparação dos algoritmos analisados;
- Quais foram os atributos selecionados e, consequentemente, os mais relevantes;
    - Justificativa para a escolha dos atributos (basta mostrar o PCA e/ou MDS e/ou gráficos que achar conveniente – diretamente no Orange)
- Descrição dos experimentos realizados:
    - Variação nos parâmetros dos algoritmos;

Observação: O mais importante é a coerência na condução da investigação. Utilize os recursos que achar mais conveniente. Gráficos ajudam a demonstrar visualmente as explicações. Não é necessário criar mais de uma rede principal de componentes no Orange. Utilize a mesma rede em todos os experimentos. Lembrando que componentes como "Continuize", "Impute", "Edit Domain" ou "Select Columns" são customizados por dataset.

Link p/ demonstração simples do uso do Orange:
https://youtu.be/52zug45Gp2c

Canal do Orange com outros tutoriais:
https://www.youtube.com/channel/UClKKWBe2SCAEyv7ZNGhIe4g

Dúvidas? Desabafo? Trocar uma Ideia sobre a abordagem? -> Whasapp!

Forma de Entrega:

Os trabalhos devem ser **apresentados** em vídeo. Entregue a apresentação do grupo, o arquivo Orange e um screenshot da tela do Orange, demonstrando a configuração utilizada (este último me auxiliará durante a avaliação). Envie através do git ou drive.