**APRENDIZADO DE MÁQUINA**

***Curso de especialização em Bancos de Dados***

***Curso de especialização em Business Intelligence***

***Curso de especialização em Big Data***

**Prof. Hugo de Paula**

**ATIVIDADE EM LABORATÓRIO 5**

**REGRESSÃO LINEAR**

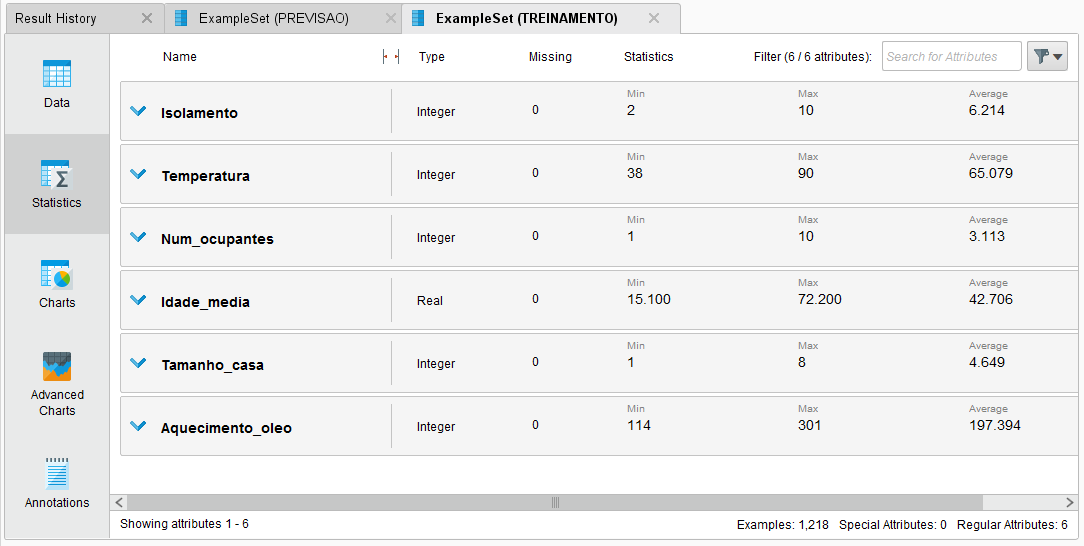
*Adaptado de North, Matthew A. Data Mining for the Masses. 2012.*

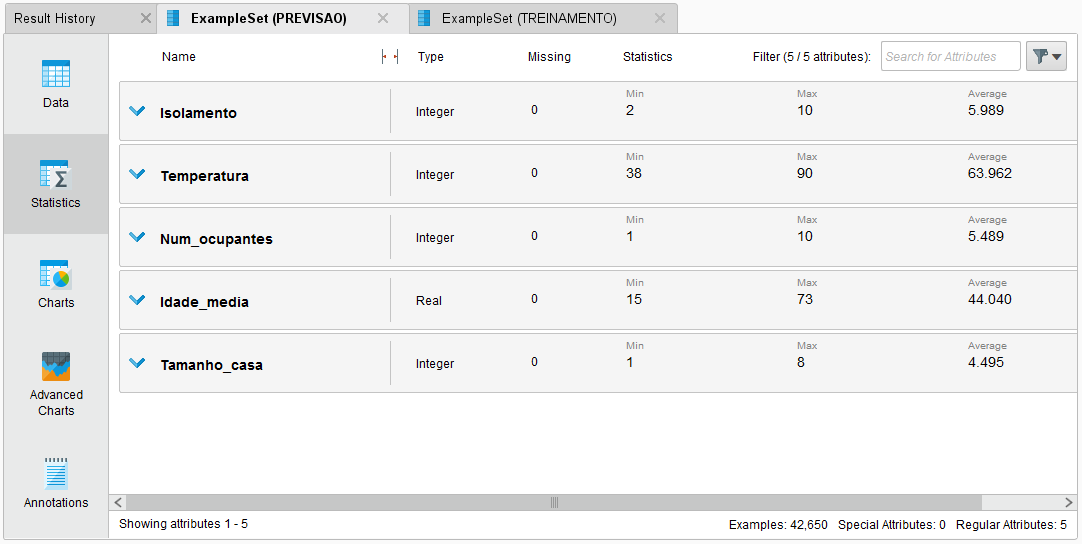
# OBJETIVOS

1. O objetivo dessa atividade é tentar prever necessidades de consumidores baseados em características das casas desses consumidores.
2. Para alcançar este objetivo, iremos realizar a previsão usando regressão linear.
3. Resumo dos dados:
   1. ***Isolamento***: nível de isolamento térmico da casa em uma taxa de 1 a 10, sendo que 10 indica a máxima espessura do isolamento.
   2. ***Temperatura***: temperatura ambiente externa média em cada casa no último ano. As medidas estão em graus Fahrenheit.
   3. ***Num\_ocupantes***: número total de pessoas vivendo na casa.
   4. ***Idade\_media***: idade média dos ocupantes da casa.
   5. ***Tamanho\_casa***: tamanho da casa em uma taxa variando de 1 a 8. Quanto maior o valor, maior a casa.
   6. ***Aquecimento\_oleo***: número total de unidades de aquecedores à óleo compradas pelo proprietário da casa no último ano. Esse é o valor a ser estimado.

# PREPARAÇÃO DOS DADOS

1. Carregue ambas as bases de dados do arquivo ***Atividade 5 - Bases.xlsx***. A planilha chamada ***Bases*** possui 1219 registros e contém a coluna ***Aquecimento\_oleo***, que será utilizada na fase de treinamento. A planilha ***PREVISAO*** contém 42651 registros de perfis de casas, mas não possui a coluna ***Aquecimento\_oleo***, que deverá ser estimada pelo modelo de regressão.
2. Na perspectiva dos resultados, analise a ***Statistics***.

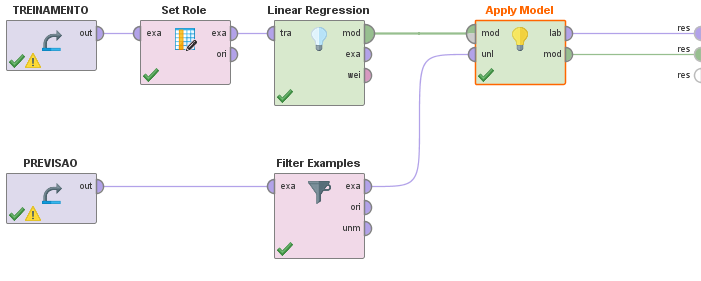




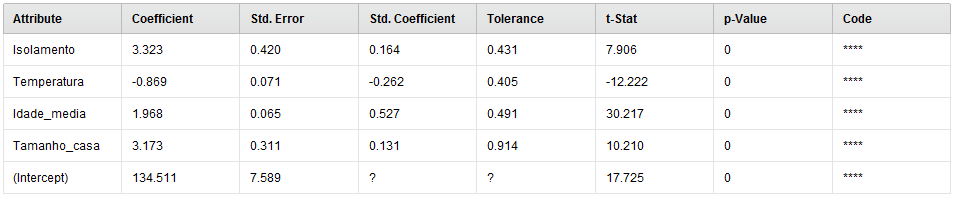
1. Note que os atributos possuem aproximadamente a mesma faixa de valores, com exceção da idade média dos ocupantes, cuja faixa de valores é mais ampla na base de previsão que na base de treinamento. Apesar de podermos acreditar que, provavelmente, essa ligeira diferença na fixa de valores de idade não irá produzir valores imprecisos, a ***REGRESSÃO LINEAR*** não pode ser usada para esses casos. Por isso vamos precisar remover esses registros fora da faixa.
   1. Adicione dois filtros través do operador ***Filter Examples***.
   2. Utilize os filtros ***Idade\_media ≥ 15.1*** e ***Idade\_media ≤ 72.2***.
2. A regressão linear é um modelo semelhante aos modelos de classificação, e precisa de um atributo alvo, que seja designado com a propriedade ***label***. Utilize o operador ***Set Role*** para definir o atributo ***Aquecimento\_oleo*** como ***label***.

# MODELAGEM

1. Vamos construir o nosso modelo de regressão linear. Utilize o operador ***Linear Regression***. Note que o modelo utiliza uma tolerância mínima default ***min tolerance = 0.05***. Essa tolerância também é conhecida como a ***confiança estatística*** ou o ***nível alfa***.
2. Agora vamos aplicar o modelo aos nossos dados de previsão com o operador ***Apply Model***.



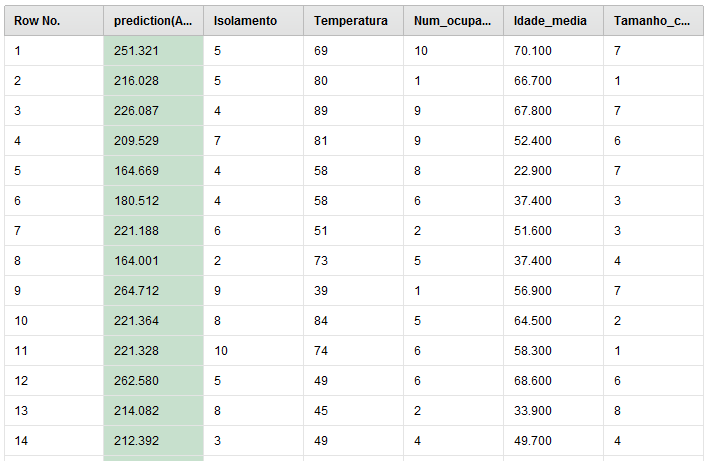
1. Agora vamos avaliar o modelo produzido. A fórmula que representa uma regressão linear é da forma , onde a variável é o alvo (***label***), que no nosso caso ~e a variável ***Aquecimento\_oleo***. Os atributos são as ***variáveis independentes***, que são representadas por . A variável representa os ***coeficientes*** de cada atributo, ou seja, o quanto aquele atributo contribui na predição do valor do atributo objetivo. Vamos observar os coeficientes estimados do nosso modelo:



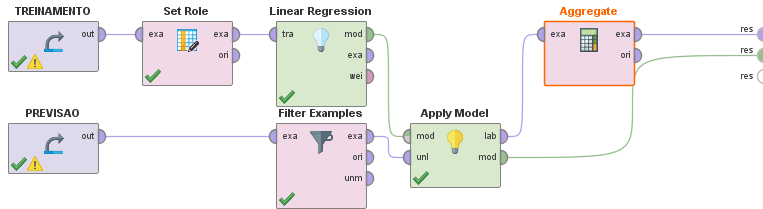
1. Observe que a variável ***Num\_ocupantes*** não foi estatisticamente relevante para a predição de ***Aquecimento\_oleo***. De acordo com o nosso modelo, se tivermos um registro com os seguintes atributos:
   1. ***Isolamento***: 6
   2. ***Temperatura***: 67
   3. ***Idade\_media***: 35.4
   4. ***Tamanho\_casa***: 5

Então, o valor previsto para ***Aquecimento\_oleo*** seria:

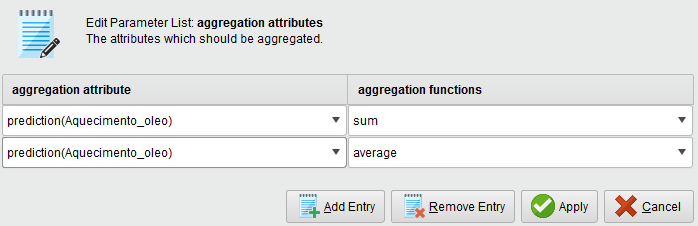
1. O nosso modelo, depois de aplicado na base de previsão, produziu a seguinte estimação:



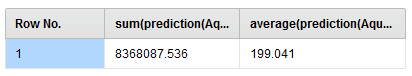
1. Suponha que se deseja prever o estoque de óleo necessário para o próximo ano. Podemos exportar a tabela com a previsão para uma planilha e criarmos um campo total, ou podemos utilizar o operador ***Aggregate***.



1. Clique no botão ***Edit List*** e adicione o campo ***prediction(Aquecimento\_oleo)***. Você pode adicionar quantos agregados desejar. No nosso caso, criaremos a soma e a média do atributo.



1. O resultado da agregação será:



1. Pelo resultado, a empresa pode prever uma probabilidade de venda de 8.368.088 (aproximadamente oito milhões) de unidades de óleo para os novos consumidores, a uma média de 200 unidades cada. Caso possuíssemos alguma informação que desejássemos separar, como região das casas, por exemplo, poderíamos usar o parâmetro ***group by attributes*** do operador ***Aggregate***.

# EXERCÍCIO

1. Baseado nos dados do modelo de ***Linear Regression*** interprete o significado dos pesos encontrado e responda às seguintes questões:
   1. Quais os atributos mais relevantes?
   2. São diretamente relacionados ou inversamente relacionados?
   3. Qual atributo não foi significativo para a construção do modelo?
2. Após a construção do modelo, o processo de Aprendizado de Máquina propõe a etapa de ***Deploy***, onde o modelo será utilizado em produção no apoio à tomada de decisões.
   1. Qual a forma de se fazer o ***deploy*** do modelo de regressão linear? Mostre no RapidMiner alternativas para a realização desta entrega.
3. Escreva um documento mostrando o processo realizado, explicando os operadores e seus parâmetros, e mostrando as regras resultantes.
   1. Tente se limitar a uma ou duas páginas.
   2. Utilize imagens da tela do ***RapidMiner*** para mostrar os operadores, os dados antes e após o processamento.
   3. Este documento não precisa de capa e nem ser formatado em nenhum formato específico de trabalho acadêmico (estilo ABNT).
   4. Basta ter o título **ATIVIDADE 5**, e conter o nome do autor do trabalho.