# INF0613 – Aprendizado de Máquina Não Supervisionado

Trabalho 3 - Técnicas de Agrupamento

O objetivo deste trabalho é exercitar o uso de algoritmos de agrupamento. Neste trabalho, vamos analisar diferentes atributos de carros com o objetivo de verificar se seus atributos são suficientes para indicar um valor de risco de seguro. O conjunto de dados já apresenta o risco calculado no campo `symboling` indicado na Tabela 1. Quanto mais próximo de 3, maior o risco. O conjunto de dados que deve ser usado está disponível na página do Moodle com o nome `imports-85.data`.

## Informações Importantes

- **Prazo de entrega:** 26 de setembro de 2021 (Domingo), até às 23h55.
- **Forma de entrega:** Deverá ser submetido um arquivo no formato **.zip** via Moodle contendo:
  - Arquivo `inf0613-trabalho3.Rmd` com as respostas das atividades, e
  - Arquivo no formato `PDF` gerado pelo `Knit`, gerado a partir do arquivo `inf0613-trabalho3.Rmd` respondido.
- **Pontuação:** Este trabalho será pontuado de 0 a 10, e corresponderá a **40%** da nota final.
- Escreva suas análises com cuidado. Essas análises corresponderão à 85% da sua nota neste trabalho.
- Na página da disciplina no Moodle, fornecemos um arquivo `inf0613-trabalho3.Rmd` que contém um esboço do trabalho, seu uso é obrigatório.
- O arquivo `imports-85.data` não deve ser alterado diretamente, ou seja, todas as alterações devem ser feitas nos objetos lidos no script.
- *Dica para esta tarefa:* revise os slides das aulas, todos os comandos em R necessários estão exemplificados neles.
- *Dica para esta tarefa:* Use as funções `fviz_*` para gerar os gráficos. Veja exemplos de como utilizá-las nos slides da aula.
- Teste o seu código antes de submeter. Códigos com erros sintáticos serão penalizados.
- Funções que não atendem às especificações serão penalizadas.
- Submissões com formatos diferentes dos especificados não serão corrigidas.
- Apenas um membro do grupo deve enviar a solução. Os **nomes completos** dos membros do grupo devem constar no cabeçalho de cada arquivo a ser submetido no local indicado.

## Descrição do Conjunto de Dados

*Tabela 1.* Nome para acesso do atributo no data-frame, nome do atributo, tipo de dado, valores possíveis para cada atributo.

| Nome | Atributo | Tipo | Valores |
|---|---|---|---|
| V1 | symboling | `integer` | -3, -2, -1, 0, 1, 2, 3 |
| V2 | normalized-losses | `numeric` | [65.00,256.00] |
| V3 | make | `string` | 22 marcas |
| V4 | fuel-type | `string` | diesel, gas |
| V5 | aspiration | `string` | std, turbo |
| V6 | num-of-doors | `string` | four, two |
| V7 | body-style | `string` | hardtop, wagon, sedan, hatchback, convertible |
| V8 | drive-wheels | `string` | 4wd, fwd, rwd |
| V9 | engine-location | `string` | front, rear |
| V10 | wheel-base | `numeric` | [86.60,120.90] |
| V11 | length | `numeric` | [141.10,208.10] |
| V12 | width | `numeric` | [60.3,72.3] |
| V13 | height | `numeric` | [47.80,59.80] |
| V14 | curb-weight | `numeric` | [1488.00,4066.00] |
| V15 | engine-type | `string` | dohc, dohcv, l, ohc, ohcf, ohcv, rotor |
| V16 | num-of-cylinders | `string` | eight, five, four, six, three, twelve, two |
| V17 | engine-size | `numeric` | [61.00,326.00] |
| V18 | fuel-system | `string` | 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi |
| V19 | bore | `numeric` | [2.54,3.94] |
| V20 | stroke | `numeric` | [2.07,4.17] |
| V21 | compression-ratio | `numeric` | [7.00,23.00] |
| V22 | horsepower | `numeric` | [48.00,288.00] |
| V23 | peak-rpm | `numeric` | [4150.00,6600.00] |
| V24 | city-mpg | `numeric` | [13.00,49.00] |
| V25 | highway-mpg | `numeric` | [16.00,54.00] |
| V26 | price | `numeric` | [5118.00,45400.00] |

## Atividade 1 – Análise e Preparação dos Dados

O conjunto de dados é composto por 205 amostras com 26 atributos cada descritos na Tabela 1. Os atributos são dos tipos `factor`, `integer` ou `numeric`. O objetivo desta etapa é a análise e preparação desses dados de forma a ser possível agrupá-los nas próximas atividades.

**Implementações:** Nos itens a seguir você implementará a leitura da base e aplicará tratamentos básicos.

a) *Tratamento de dados Incompletos:* Amostras incompletas deverão ser tratadas, e você deve escolher a forma que achar mais adequada. Considere como uma amostra incompleta uma linha na qual faltam dados em alguma das colunas selecionadas anteriormente. Note que, dados faltantes nas amostras podem causar uma conversão do tipo do atributo de todas as amostras e isso pode impactar no item b).

b) *Seleção de Atributos:* Atributos não-numéricos não podem ser usados com as técnicas agrupamento vistas em aula. Portanto, você deve selecionar um conjunto de atributos numéricos que serão usados para o agrupamento. Além disso você deve analisar se os atributos não-numéricos são descritivos para a realização dos agrupamentos. Caso um dos atributos não numéricos seja necessário, use a técnica do *one hot encoding* para transformá-lo em numérico. **Não** aplique essa técnica nos atributos `symboling` e `make` para os agrupamentos subsequentes, eles não devem fazer parte do agrupamento.

### Análises

Após as implementações escreva uma análise da base de dados. Em especial, descreva o conjunto de dados inicial, relate como foi realizado o tratamento, liste quais os atributos escolhidos para manter na base e descreva a base de dados após os tratamentos listados. Explique todos os passos executados, mas sem copiar códigos na análise. Além disso justifique suas escolhas de tratamento nos dados faltantes e seleção de atributos.

## Atividade 2 – Agrupamento com o *K-means*

Nesta atividade, você deverá agrupar os dados com o algoritmo *K-means* e utilizará duas métricas básicas para a escolha do melhor $K$: a soma de distâncias intra-cluster e o coeficiente de silhueta.

**Implementações:** Nos itens a seguir você implementará a geração de gráficos para a análise das distâncias intra-cluster e do coeficiente de silhueta. Em seguida, você implementará o agrupamento dos dados processados na atividade anterior com o algoritmo *K-means* utilizando o valor de $K$ escolhido.

a) *Gráfico Elbow Curve:* Construa um gráfico com a soma das distâncias intra-cluster para $K$ variando de 2 a 30.

b) *Gráfico da Silhueta:* Construa um gráfico com o valor da silhueta para $K$ variando de 2 a 30.

c) *Escolha do K:* Avalie os gráficos gerados nos itens anteriores e escolha o melhor valor de $K$ com base nas informações desses gráficos e na sua análise. Se desejar, use também a função `NbClust` para ajudar nas análises. Com o valor de $K$ definido, utilize o rótulo obtido para cada amostra, indicando o grupo ao qual ela pertence, para gerar um gráfico de dispersão (atribuindo cores diferentes para cada grupo).

### Análises

Descreva cada um dos gráficos gerados nos itens acima e analise-os. Inclua na sua análise as informações mais importantes que podemos retirar desses gráficos. Discuta sobre a escolha do valor $K$ e sobre a apresentação dos dados no gráfico de dispersão.

## Atividade 3 – Agrupamento com o *DBscan*

Nesta atividade, você deverá agrupar os dados com o algoritmo *DBscan*. Para isso será necessário experimentar com diferentes valores de *eps* e *minPts*.

a) *Ajuste de Parâmetros:* Experimente com valores diferentes para os parâmetros *eps* e *minPts*. Verifique o impacto dos diferentes valores nos agrupamentos.

b) *Determinando Ruídos:* Escolha o valor de *minPts* que obteve o melhor resultado no item anterior e use a função `kNNdistplot` do pacote `dbscan` para determinar o melhor valor de *eps* para esse valor de *minPts*. Lembre-se que o objetivo não é remover todos os ruídos.

c) *Visualizando os Grupos:* Após a escolha dos parâmetros *eps* e *minPts*, utilize o rótulo obtido para cada amostra, indicando o grupo ao qual ela pertence, para gerar um gráfico de dispersão (atribuindo cores diferentes para cada grupo).

### Análises

Descreva os experimentos feitos para a escolha dos parâmetros *eps* e *minPts*. Inclua na sua análise as informações mais importantes que podemos retirar dos gráficos gerados. Justifique a escolha dos valores dos parâmetros e analise a apresentação dos dados no gráfico de dispersão.

## Atividade 4 – Comparando os Algoritmos

Com base nas atividades anteriores, faça uma conclusão dos seus experimentos respondendo às seguintes perguntas:

a) Qual dos métodos apresentou melhores resultados? Justifique.

b) Quantos agrupamentos foram obtidos?

c) Analisando o campo `symboling` e o grupo designado para cada amostra, os agrupamentos conseguiram separar os níveis de risco?

d) Analisando o campo `make` que contém as marcas dos carros, os agrupamentos conseguiram separar as marcas?