



Pós-Graduação em Ciência de Dados

Disciplina - CD02.1 - Data Mining e Machine Learning II

(Laboratório 06 – Revisão Geral 1)

<p>Objetivos do Laboratório – Desenvolver projeto de Mineração de Dados utilizando o SAS Enterprise Miner</p>
--

Prof^o Sérgio Côrtes
1^a versão
Setembro de 2018

1. Fontes de Dados

1.1. Estão disponíveis para os seus estudos as seguintes bases de dados:

1.1.1. *ENEM_2015_2017_100p* – Dados (registros) dos candidatos que realizaram as provas dos ENEMs dos anos de 2015 a 2017, contendo 100% dos registros de todos os alunos que realizaram todas as provas.

1.1.2. *ENEM_2015_2017_10p* – Amostra estratificada por (Ano, UF e Município) dos dados (registros) dos candidatos que realizaram as provas dos ENEMs dos anos de 2015 a 2017, contendo 10% dos registros do arquivo citado em 1.1.1.

1.1.3. *ENEM_2015_2017_10p_miss* – Amostra estratificada por (Ano, UF e Município) dos dados (registros) dos candidatos que realizaram as provas dos ENEMs dos anos de 2015 a 2017, contendo 10% dos registros do arquivo citado em 1.1.1. e com algumas variáveis com erro de preenchimento.

2. Análise e Preparação dos dados para Mineração dos Dados

2.1. Utilizando as metodologias estudadas e descritas no Anexo I, vamos desenvolver um projeto para:

2.1.1. *Entender os dados*

2.1.1.1. Utilize o software *Enterprise Miner* ou *Enterprise Guide* para descrever **todas** as variáveis com estatísticas apropriadas e identificar eventuais problemas;

2.1.1.2. Prepare uma apresentação em *power point* com os resultados encontrados e **propostas de correções** para os problemas encontrados.

2.1.2. *Preparar os dados*

2.1.2.1. Utilize o software *Enterprise Miner* e implementar as correções propostas no item 2.1.1.2;

2.1.2.2. Compare os resultados de antes e após as correções realizadas no item 2.1.2.1;

2.1.2.3. Prepare uma apresentação em *power point* com os resultados após a imputação e correção dos dados.

3. Descoberta de conhecimentos

3.1. Utilizando os dados corrigidos no item 2.1.2.1 vamos implementar alguns modelos.

3.1.1. *Modelagem dos dados – Análise de Cluster*

3.1.1.1. *Utilize, inicialmente, as variáveis:*

Variável	Papel
NU_ANO	Input
SG_UF_RESIDENCIA	Input
NU_IDADE	Input
TP_SEXO	Input
TP_ESCOLA	Input
NU_NOTA_MT	Input
NOTA_MEDIA	Target

3.1.1.2. Utilize o software *Enterprise Miner* (Explore → Cluster) e execute três (3) *análises de cluster* utilizando os métodos de Centroide, Média e Ward (analise o dendograma);

3.1.1.3. Compare os três resultados e indicar quais deles é o melhor para este conjunto de dados;

3.1.1.4. Prepare uma apresentação em *power point* com os resultados com as análises dos clusters.

- 3.2. Refaça as suas análises, incorporando ou retirando outras variáveis, para identificar cluters de melhor formação
- 3.3. Desenvolva uma análise de Cluster Hierárquica para Identificar UF/Municípios que possua similaridades em suas notas no território brasileiro.
- 3.4. Processe as suas análises com a base dos três anos (ENEM_2015_2017_100p).
- 3.5. Prepare uma apresentação final sobre os seus resultados, identificando municípios fora das regiões sudeste/sul de excelência na participação dos ENEMs.

Anexo I - Metodologias

O processo da Ciência de Dados



CRISP-DM

Cross Industry Standard Process for Data Mining

Prof. Sérgio C.

- **Entender o Negócio:** foca em entender o objetivo do projeto a partir de uma perspectiva de negócios, definindo um plano preliminar para atingir os objetivos.
- **Entender os Dados:** levantamento de dados e início de atividades para familiarização com os dados, identificando problemas ou conjuntos interessantes.
- **Preparação dos Dados:** construção do conjunto de dados final a partir dos dados iniciais. Normalmente ocorre várias vezes no processo.
- **Modelagem:** várias técnicas de modelagem são aplicadas, e seus parâmetros calibrados para otimização. Assim, é comum retornar à Preparação dos Dados durante essa fase.
- **Avaliação:** é construído um modelo que parece ter grande qualidade de uma perspectiva de análise de dados. No entanto, é necessário verificar se o modelo atinge os objetivos do negócio.
- **Implantação:** o conhecimento adquirido pelo modelo é organizado e apresentado de uma maneira que o cliente possa utilizar.

A001 - Introdução à Ciência de Dados

02. Introdução - 49

O processo da Ciência de Dados



SEMMA (SAS Institute)

S: Sample (Training, Validation, Test)
E: Explore (get an idea of the data at hand)
M: Modify (select, transform)
M: Model (create data mining model)
A: Assess (validate model)

Prof. Sérgio Côrtes

- **Amostragem.** O processo inicia-se com a amostragem de dados, por exemplo, escolhendo os dados estabelecidos para modelagem. O conjunto de dados deve ser grande o suficiente para conter informações suficientes para utilização, mas pequeno o suficiente para ser usado de forma eficiente. Esta fase também trata da partição de dados.
- **Exploração.** Esta fase abrange a compreensão dos dados para descobrir relações antecipadas e inesperadas entre as variáveis, e também anormalidades, com a ajuda da visualização de dados.
- **Modificação/transformação.** A fase Modificar contém métodos para selecionar, criar e transformar variáveis em preparação para modelagem de dados.
- **Modelagem.** O foco é aplicar várias técnicas de modelagem (*mineração de dados*) nas variáveis trabalhadas, a fim de criar modelos que possivelmente forneçam o resultado desejado.
- **Avaliação.** A avaliação dos resultados da modelagem mostra a confiabilidade e utilidade dos modelos criados.

2018.2 - CIA001 - Introdução à Ciência de Dados

02. Introdução - 51

Anexo II - Dicionário de Dados das variáveis categorizadas

```
value $IN_ESTUDA_CLASSE_HOSPITALAR
0='Não'
1='Sim';

value $IN_TREINEIRO
0='Não'
1='Sim';

value $TP_DEPENDENCIA_ADM
1= 'Federal'
2= 'Estadual'
3= 'Municipal'
4= 'Privada';

value $TP_LOCALIZACAO
1= 'Urbana'
2= 'Rural';

value $TP_SIT_FUNC_ESC
1='Em atividade'
2='Paralisada'
3='Extinta';

value $TP_SEXO
M = 'Masculino'
F = 'Feminino';

value $TP_NACIONAL
0= 'Não informado'
1= 'Brasileiro(a)'
2= 'Brasileiro(a) Naturalizado(a)'
3= 'Estrangeiro(a)'
4= 'Brasileiro(a) Nato(a), nascido(a) no exterior';

value $TP_ST_CONCLUSAO
1='Já concluí o Ensino Médio'
2='Estou cursando e concluirei o Ensino Médio em 2017'
3='Estou cursando e concluirei o Ensino Médio após 2017'
4='Não concluí e não estou cursando o Ensino Médio';

value $TP_ANO_CONCLUIU
0= 'Não informado'
1= '2016'
2= '2015'
3= '2014'
4= '2013'
5= '2012'
6= '2011'
7= '2010'
8= '2009'
```

```
9=      '2008'
10=     '2007'
11=     'Anterior a 2007';

value   $TP_ESCOLA
1='Não respondeu'
2='Pública'
3='Privada'
4='Exterior';

value   $TP_ENSINO
1='Ensino Regular'
2='Educação Especial - Modalidade Substitutiva'
3='Educação de Jovens e Adultos';

value   $TP_ESTADO_CIVIL
0='Solteiro(a) '
1='Casado(a) /Mora com um(a) companheiro(a) '
2='Divorciado(a) /Desquitado(a) /Separado(a) '
3='Viúvo(a) ' ;

value   $TP_COR_RACA
0='Não declarado'
1='Branca'
2='Preta'
3='Parda'
4='Amarela'
5='Indígena';

value   $IN_BAIXA_VISAO
1='Sim'
0='Não';

value   $IN_CEGUEIRA
1='Sim'
0='Não';

value   $IN_SURDEZ
1='Sim'
0='Não';

value   $IN_DEFICIENCIA_AUDITIVA
1='Sim'
0='Não';

value   $IN_SURDO_CEGUEIRA
1='Sim'
0='Não';

value   $IN_DEFICIENCIA_FISICA
1='Sim'
0='Não';
```



```
value $IN_DEFICIENCIA_MENTAL
1='Sim'
0='Não';

value $IN_DEFICIT_ATENCAO
1='Sim'
0='Não';

value $IN_DISLEXIA
1='Sim'
0='Não';

value $IN_GESTANTE
1='Sim'
0='Não';

value $IN_LACTANTE
1='Sim'
0='Não';

value $IN_IDOSO
1='Sim'
0='Não';

value $IN_DISCALCULIA
1='Sim'
0='Não';

value $IN_AUTISMO
1='Sim'
0='Não';

value $IN_VISAO_MONOCULAR
1='Sim'
0='Não';

value $IN_OUTRA_DEF
1='Sim'
0='Não';

value $IN_SEM_RECURSO
1='Sim'
0='Não';

value $IN_NOME_SOCIAL
1='Sim'
0='Não';

value $IN_BRILLE
1='Sim'
0='Não';
```

```
value $IN_AMPLIADA
1='Sim'
0='Não';

value $IN_LEDOR
1='Sim'
0='Não';

value $IN_ACESSO
1='Sim'
0='Não';

value $IN_TRANSCRICAO
1='Sim'
0='Não';

value $IN_LIBRAS
1='Sim'
0='Não';

value $IN_LEITURA_LABIAL
1='Sim'
0='Não';

value $IN_MESA_CADEIRA_RODAS
1='Sim'
0='Não';

value $IN_MESA_CADEIRA_SEPARADA
1='Sim'
0='Não';

value $IN_APOIO_PERNA
1='Sim'
0='Não';

value $IN_GUIA_INTERPRETE
1='Sim'
0='Não';

value $IN_COMPUTADOR
1='Sim'
0='Não';

value $IN_CADEIRA_ESPECIAL
1='Sim'
0='Não';

value $IN_CADEIRA_CANHOTO
1='Sim'
0='Não';
```

```
value $IN_CADEIRA_ACOLCHOADA
1='Sim'
0='Não';

value $IN_PROVA_DEITADO
1='Sim'
0='Não';

value $IN_MOBILIARIO_OBESO
1='Sim'
0='Não';

value $IN_LAMINA_OVERLAY
1='Sim'
0='Não';

value $IN_PROTETOR_AURICULAR
1='Sim'
0='Não';

value $IN_MEDIDOR_GLICOSE
1='Sim'
0='Não';

value $IN_MAQUINA_BRAILE
1='Sim'
0='Não';

value $IN_SOROBAN
1='Sim'
0='Não';

value $IN_MARCA_PASSO
1='Sim'
0='Não';

value $IN_SONDA
1='Sim'
0='Não';

value $IN_MEDICAMENTOS
1='Sim'
0='Não';

value $IN_SALA_INDIVIDUAL
1='Sim'
0='Não';

value $IN_SALA_ESPECIAL
1='Sim'
0='Não';
```

```
value $IN_SALA_ACOMPANHANTE
1='Sim'
0='Não';

value $IN_MOBILIARIO_ESPECIFICO
1='Sim'
0='Não';

value $IN_MATERIAL_ESPECIFICO
1='Sim'
0='Não';

value $TP_PRESENCA_CN
0='Faltou à prova'
1='Presente na prova'
2='Eliminado na prova';

value $TP_PRESENCA_CH
0='Faltou à prova'
1='Presente na prova'
2='Eliminado na prova';

value $TP_PRESENCA_LC
0='Faltou à prova'
1='Presente na prova'
2='Eliminado na prova';

value $TP_PRESENCA_MT
0='Faltou à prova'
1='Presente na prova'
2='Eliminado na prova';

value $CO_PROVA_CN
391='Azul'
392='Amarela'
393='Cinza'
394='Rosa'
407='Laranja - Adaptada Ledor'
411='Verde - Videoprova - Libras'
431='Amarela (Reaplicação)'
432='Cinza (Reaplicação)'
433='Azul (Reaplicação)'
434='Rosa (Reaplicação)';

value $CO_PROVA_CH
395='Azul'
396='Amarela'
397='Branca'
398='Rosa'
408='Laranja - Adaptada Ledor'
412='Verde - Videoprova - Libras'
435='Azul (Reaplicação)'
436='Amarelo (Reaplicação)'
```

```
437='Branco (Reaplicação) '
438='Rosa (Reaplicação)';

value $CO_PROVA_LC
399='Azul'
400='Amarela'
401='Rosa'
402='Branca'
409='Laranja - Adaptada Ledor'
413='Verde - Videoprova - Libras'
439='Azul (Reaplicação) '
440='Amarelo (Reaplicação) '
441='Branco (Reaplicação) '
442='Rosa (Reaplicação)';

value $CO_PROVA_MT
403='Azul'
404='Amarela'
405='Rosa'
406='Cinza'
410='Laranja - Adaptada Ledor'
414='Verde - Videoprova - Libras'
443='Amarela (Reaplicação) '
444='Cinza (Reaplicação) '
445='Azul (Reaplicação) '
446='Rosa (Reaplicação)';

value $TP_LINGUA
0='Inglês'
1='Espanhol';

value $TP_STATUS_REDACAO
1='Sem problemas'
2='Anulada'
3='Cópia Texto Motivador'
4='Em Branco'
6='Fuga ao tema'
7='Não atendimento ao tipo'
8='Texto insuficiente'
9='Parte desconectada';

value $Qum
A='Nunca estudou'
B='Não completou a 4ª série/5º ano do ensino fundamental'
C='Completou a 4ª série/5º ano, mas não completou a 8ª
série/9º ano do ensino fundamental'
D='Completou a 8ª série/9º ano do ensino fundamental, mas
não completou o Ensino Médio'
E='Completou o Ensino Médio, mas não completou a
Faculdade'
F='Completou a Faculdade, mas não completou a Pós-
graduação'
G='Completou a Pós-graduação'
```

```
H='Não sei';

value $Qdois
A='Nunca estudou'
B='Não completou a 4ª série/5º ano do ensino fundamental'
C='Completou a 4ª série/5º ano, mas não completou a 8ª
série/9º ano do ensino fundamental'
D='Completou a 8ª série/9º ano do ensino fundamental, mas
não completou o Ensino Médio'
E='Completou o Ensino Médio, mas não completou a
Faculdade'
F='Completou a Faculdade, mas não completou a Pós-
graduação'
G='Completou a Pós-graduação'
H='Não sei';

value $Qtres
A='Grupo 1: Lavrador, agricultor sem empregados, bóia
fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos
etc.), apicultor, pescador, lenhador, seringueiro, extrativista'
B='Grupo 2: Diarista, empregado doméstico, cuidador de
idosos, babá, cozinheiro (em casas particulares), motorista
particular, jardineiro, faxineiro de empresas e prédios, vigilante,
porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja,
auxiliar administrativo, recepcionista, servente de pedreiro,
repositor de mercadoria'
C='Grupo 3: Padeiro, cozinheiro industrial ou em
restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico,
operador de máquinas, soldador, operário de fábrica, trabalhador da
mineração, pedreiro, pintor, eletricista, encanador, motorista,
caminhoneiro, taxista'
D='Grupo 4: Professor (de ensino fundamental ou médio,
idioma, música, artes etc.), técnico (de enfermagem, contabilidade,
eletrônica etc.), policial, militar de baixa patente (soldado, cabo,
sargento), corretor de imóveis, supervisor e gerente, mestre de
obras, pastor, microempresário (proprietário de empresa com menos de
10 empregados), pequeno comerciante, pequeno proprietário de terras,
trabalhador autônomo ou por conta própria'
E='Grupo 5: Médico, engenheiro, dentista, psicólogo,
economista, advogado, juiz, promotor, defensor, delegado, tenente,
capitão, coronel, professor universitário, diretor em empresas
públicas e privadas, político, proprietário de empresas com mais de
10 empregados'
F='Não Sei';

value $Qquatro
A='Grupo 1: Lavradora, agricultora sem empregados, bóia
fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos
etc.), apicultora, pescadora, lenhadora, seringueira, extrativista'
B='Grupo 2: Diarista, empregada doméstica, cuidadora de
idosos, babá, cozinheira (em casas particulares), motorista
particular, jardineira, faxineira de empresas e prédios, vigilante,
porteira, carteira, office-boy, vendedora, caixa, atendente de loja,
```

auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria'

C='Grupo 3: Padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista'

D='Grupo 4: Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora e gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria'

E='Grupo 5: Médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitã, coronel, professora universitária, diretora em empresas públicas e privadas, política, proprietária de empresas com mais de 10 empregados'

F='Não Sei';

```
value $Qcinco
```

```
1='1'  
2='2'  
3='3'  
4='4'  
5='5'  
6='6'  
7='7'  
8='8'  
9='9'  
10='10'  
11='11'  
12='12'  
13='13'  
14='14'  
15='15'  
16='16'  
17='17'  
18='18'  
19='19'  
20='20';
```

```
value $Qseis
```

```
A= "Nenhuma renda"  
B= "Até R$ 937,00"  
C= "De R$ 937,01 até R$ 1.405,50"  
D= "De R$ 1.405,51 até R$ 1.874,00"  
E= "De R$ 1.874,01 até R$ 2.342,50"  
F= "De R$ 2.342,51 até R$ 2.811,00"  
G= "De R$ 2.811,01 até R$ 3.748,00"  
H= "De R$ 3.748,01 até R$ 4.685,00"  
I= "De R$ 4.685,01 até R$ 5.622,00"
```

```
J= "De R$ 5.622,01 até R$ 6.559,00"  
K= "De R$ 6.559,01 até R$ 7.496,00"  
L= "De R$ 7.496,01 até R$ 8.433,00"  
M= "De R$ 8.433,01 até R$ 9.370,00"  
N= "De R$ 9.370,01 até R$ 11.244,00"  
O= "De R$ 11.244,01 até R$ 14.055,00"  
P= "De R$ 14.055,01 até R$ 18.740,00"  
Q= "Mais de R$ 18.740,00";
```

```
value $Qsete  
A='Não'  
B='Sim, um ou dois dias por semana'  
C='Sim, três ou quatro dias por semana'  
D='Sim, pelo menos cinco dias por semana';
```

```
value $Qoito  
A='Não'  
B='Sim, um'  
C='Sim, dois'  
D='Sim, três'  
E='Sim, quatro ou mais';
```

```
value $Qnove  
A='Não'  
B='Sim, um'  
C='Sim, dois'  
D='Sim, três'  
E='Sim, quatro ou mais';
```

```
value $Qdez  
A='Não'  
B='Sim, um'  
C='Sim, dois'  
D='Sim, três'  
E='Sim, quatro ou mais';
```

```
value $Qonze  
A='Não'  
B='Sim, uma'  
C='Sim, duas'  
D='Sim, três'  
E='Sim, quatro ou mais';
```

```
value $Qdoze  
A='Não'  
B='Sim, uma'  
C='Sim, duas'  
D='Sim, três'  
E='Sim, quatro ou mais';
```

```
value $Qtreze  
A='Não'  
B='Sim, um'
```



```
C='Sim, dois'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qcatorze
A='Não'
B='Sim, uma'
C='Sim, duas'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qquinze
A='Não'
B='Sim, uma'
C='Sim, duas'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qdezesesseis
A='Não'
B='Sim, um'
C='Sim, dois'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qdezessete
A='Não'
B='Sim, uma'
C='Sim, duas'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qdezoito
A='Não'
B='Sim';

value $Qdezenove
A='Não'
B='Sim, uma'
C='Sim, duas'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qvinte
A='Não'
B='Sim';

value $Qvinteum
A='Não'
B='Sim';

value $Qvintedois
A='Não'
```

```
B='Sim, um'
C='Sim, dois'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qvintetres
A='Não'
B='Sim';

value $Qvintequatro
A='Não'
B='Sim, um'
C='Sim, dois'
D='Sim, três'
E='Sim, quatro ou mais';

value $Qvintecinco
A='Não'
B='Sim';

value $Qvinteseis
A='Já concluí o Ensino Médio'
B='Estou cursando e concluirei o Ensino Médio em 2017'
C='Estou cursando e concluirei o Ensino Médio após 2017'
D='Não concluí e não estou cursando o Ensino Médio';

value $Qvintesete
A='Somente em escola pública'
B='Parte em escola pública e parte em escola privada sem
bolsa de estudo integral'
C='Parte em escola pública e parte em escola privada com
bolsa de estudo integral'
D='Somente em escola privada sem bolsa de estudo
integral'
E='Somente em escola privada com bolsa de estudo
integral';

run;
```