# Mineração de Dados – Classificação
# Preparação da base de microdados do ENEM 2012

**Jonathan Coutinho Luz de Queiroz**[1]
**Guilherme Lima Bernal**[1]

[1]Instituto de Matemática – Universidade Federal da Bahia

jonathanqueiroz@dcc.ufba.br, ufba@lbguilherme.com

## 1. Introdução

Nesta atividade, preparamos a base de microdados do ENEM 2012 para a futura utilização em um algoritmo de classificação. O objetivo final do trabalho consiste em classificar os candidatos em quatro grupos, de acordo com a nota obtida na prova de redação: "ruim", "regular", "boa" e "ótima". O intervalo de valores para a nota da redação que está associado a cada um desses grupos é baseado na divisão do dataset original em quartis, sendo "ruim" o primeiro quartil e "ótima" o último. Para passar pela qualificação, todas as colunas precisam ser normalizadas para números de 0 a 1.

Optamos, nesta etapa, por remover colunas majoritariamente por razões técnicas (número de identificação do candidato, por exemplo, além de outros campos descritos a seguir). Consequentemente, colunas que simplesmente suspeitamos serem poucos relevantes para a classificação, devido à análise estatística previamente realizada, foram mantidas. Essa abordagem permite que a análise estatística sirva como um guia durante o desenvolvimento do algoritmo de classificação, ao invés de determinar completamente quais dados serão utilizados nele.

## 2. Normalização da base de dados

As seguintes operações foram realizadas com o intuito de normalizar a base de dados:

- Foi criada a coluna ***UF_INSC_IGUAL_PROVA***, que guarda "1" se ***UF_INSC*** for igual ao ***UF_MUNICIPIO_PROVA***, indicando que o candidato não mudou de estado para fazer a prova. "0" caso contrário.
- ***UF_INSC*** foi expandido para um total de 26 colunas, cada uma contendo "1" ou "0" para indicar se o candidato fez a prova nesse estado ou não. Nenhum candidato poder ter mais de um estado marcado como "1". Informações adicionais sobre a localização do candidato (nome do município, código do municício, etc.) foram removidas.
- De forma similar, ***TP_SEXO*** foi expandido para duas colunas: ***GENERO_HOMEM*** e ***GENERO_MULHER***, com apenas uma delas sendo marcada como "1".
- As colunas ***NOTA_CN***, ***NOTA_CH***, ***NOTA_LC*** e ***NOTA_MT*** passaram a guardar a nota do candidato em cada uma das provas. Para que o valor seja consistente de 0 a 1, essa nota foi dividida por 1000.

- Várias colunas que guardavam as opções marcadas pelo candidato, gabarito, identificação da prova e afins foram removidas por serem irrelevantes para obtenção do resultado.
- A coluna ***TP_ESCOLA*** foi expandida para duas colunas: ***ESCOLA_PUBLICA*** e ***ESCOLA_PRIVADA***. Para a maioria dos candidatos, o tipo de escola não foi informado, e portanto ambas as colunas apresentam valor falso. Demais informações sobre a escola e a entidade certificadora do candidato (nome, município, etc.) foram removidas.
- Analogamente, a coluna ***ID_LOCALIZACAO*** foi expandida para duas colunas: ***LOCALIZACAO_URBANA*** e ***LOCALIZACAO_RURAL***. Para a maioria dos candidatos, o tipo de localização não foi informado, e portanto ambas as colunas apresentam valor falso.
- A idade foi normalizada para valores de 0 a 1 através de uma divisão por 100. Dessa forma, a menor idade passou a ser 0.12 e a maior idade passou a ser 0.70.
- ***TP_COR_RACA*** foi dividido em 5 colunas, uma com cada cor/raça. Cada uma possui "1" se ela é verdadeira para o candidato. Nenhum candidato tem duas dessas colunas como verdadeiras.
- Analogamente, ***TP_ESTADO_CIVIL*** foi dividido em 4 colunas booleanas: ***ESTADO_CIVIL_SOLTEIRO, ESTADO_CIVIL_CASADO, ESTADO_CIVIL_DIVORCIADO e ESTADO_CIVIL_VIUVO.***
- ***ST_CONCLUSAO*** foi dividido em 4 colunas seguindo o mesmo princípio: ***ENSINO_MEDIO_CONCLUIDO*** (alunos que já haviam concluído o Ensino Médio quando a prova foi aplicada), ***ENSINO_MEDIO_EM_2012 (***alunos que estavam concluindo o Ensimo Médio no mesmo ano de aplicação da prova***)***, ***ENSINO_MEDIO_DEPOIS_2012*** (alunos que só concluiríam o Ensino Médio em anos posteriores à aplicação da prova) e ***ENSINO_MEDIO_NAO_FAZ*** (alunos que nem haviam concluído previamente nem estavam cursando o Ensino Médio). O ano exato de conclusão do Ensino Médio, quando informado, foi descartado.
- Por fim, a nota da redação foi dividida nos 4 grupos anteriormente cidados: "ruim", "regular", "boa" e "ótima".