Trabalho da Disciplina MATA60 T01 - Banco de dados

Mineração de dados da base aberta do ENEM 2012

Departamento de Ciência da Computação (DCC) / Universidade Federal da Bahia

Integrantes: Mateus Cordeiro, Ramon Dias, Roberto Sales

2ª Etapa: Pré-processamento da Base

Uma base de dados pode apresentar certos problemas, como os seguintes:

- Incompletude: Valores Faltando
- Inconsistência: Valores fora do domínio
- Ruído: Variações inexplicáveis

Na etapa de pré-processamento algumas operações (Ex: limpeza, seleção, redução) são realizadas sobre a base para identificar e corrigir tais problemas e preparar os dados para as etapas de análise.

A seguir estão descritas as operações realizadas no pré-processamento da base do ENEM 2012. Foi utilizada nesse trabalho a linguagem `R`. Cada operação foi implementada dentro de uma função e todas as chamadas dessas funções são realizadas no script `preprocess.R`.

1. Remoção de atributos irrelevantes para o estudo

Função: `eraseColumns(dados, colunas_a_serem_removidas)`

Uma função foi criada para remover atributos (colunas) que não foram considerados importantes para as análises e predições nas próximas etapas. Os atributos removidos foram relacionados aos seguintes dados:

- Local de prova
- Identificação dos tipos de prova
- Discretização das respostas
- Gabaritos das provas
- Alguns outros dados básicos sobre a o candidato e a sua inscrição

2. Normalização de valores para atributos numéricos (Domínio)

Função: `numeric_column(dados, coluna, valor_padrao)`

Essa função recebe uma o identificador de uma coluna numérica e subistitui todos os valores invalidos pelo valor padrão. essa função é chamada para todos os atributos da base do ENEM que foram identificados como numéricos e que continham valores inválidos (total de 19 atributos).

3. Identificação e remoção de outliers

Função: `remove_outliers(dados, coluna)`

Essa função utiliza o `boxplot` para identificar registros muito destoantes do atributo passado. Todo registro dos ouliers são removidos. Essa operação foi realizada sobre alguns atributos de notas e cerca de 700 registros foram removidos.

4. Identificação e correção de incoerências em dados condicionados

Função: `check_conditions(dados, coluna_condicionada, condicoes)`

Essa função verifica se pelo menos uma das condições para atribuição de um valor positivo a uma coluna foi satisfeita. Funciona apenas para colunas de verdadeiro ou falso. Essa operação foi utilizada para corrigir invalidações nos dados de auxílio a necessidades especiais.

5. Predição de idade inválida

Função: `age_correction(dados)`

Foram identificadas manualmente algumas incoerências em relação a idade de alguns candidatos. Com essa função esses registros são alterados seguindo o calculo de que o candidato tivesse 17 anos no quando concluiu o ensino médio.

Ao fim de todas as operações é gerada uma nova base no mesmo formato da original (.csv).