Entendimento dos Dados

21 de dezembro de 2016

Seleção dos Dados

Utilização dos dados de 2000 até 2009 como treino e de 2010 até 2015 para testes.

Atributos que serão levados em consideração foram explicados na fase de entendimento dos dados: sexo, idade, UF, cotista, tipo da escola, raça, curso, forma de ingresso, IRA.

Construção dos Atributos Derivados

Atributos derivados que serão construídos já foram definidos na fase anterior.

Histograma da Distribuição dos Atributos Derivados - Coeficiente de Melhora Acadêmica

Matriz de Correlação

Qual valor de covariância é indicado de modo a eliminar feature

Limpeza dos Dados

Optou-se por não considerar features nos quais mais de 60% das entradas fossem missing values.

Descartou-se assim o feature raça.

Limpeza dos Dados

Para o caso de atributos que podem apresentar missing values, foi feita imputação (exemplo: coeficiente de melhora acadêmica).

Integração dos Dados

Dados originais da SIGRA foram integrados com a informação dos currículos antigos e novos dos cursos, de modo a saber quais matérias são de quais semestres.

Transformação dos dados

Dados textuais serão transformados em dados numéricos.

IRA foi normalizado para 0 a 1, assim como o coeficiente de melhora acadêmica.

Referências