

Aplicação de Data Mining na base de dados do Enade 2012

Ana Flávia F. da Cruz
Thayza S. Guarnier





Limpeza de Dados

- Em seu dicionário de variáveis, possuam 128 variáveis. Nesse primeiro momento era preciso selecionar as variáveis que teriam maior relevância para ser analisadas. Foram escolhidas 13 variáveis para trabalho apresentadas na tabela 1.
- Usou-se um programa em Python para fazer a limpeza, pois possui o módulo CSV.
- Na coluna de Sexo dos candidatos, os atributos estavam de forma nominal, ou seja, em forma de caractere 'M' para masculino e 'F' para feminino. Com essa necessidade de adequar os dados, então catalogamos como 0 os dados de 'M' e 1 para 'F'.



Limpeza de Dados

“Somando a sua renda com a renda dos familiares que moram com você, quanto é, aproximadamente, a renda familiar?”

- A = Nenhuma;
- B = Até 1,5 salários mínimos (até R\$ 697,50);
- C = Acima de 1,5 até 3 salários mínimos (R\$ 697,51 a R\$ 1.395,00);
- D = Acima de 3 até 4,5 salários mínimos (R\$ 1.395,01 a R\$ 2.092,50);
- E = Acima de 4,5 até 6 salários mínimos (R\$ 2.092,51 a R\$ 2.790,00);
- F = Acima de 6 até 10 salários mínimos (R\$ 2.790,01 a R\$ 4.650,00);
- G = Acima de 10 até 30 salários mínimos (R\$ 4.650,01 a R\$ 13.950,00);
- H = Acima de 30 salários mínimos (mais de R\$ 13.950,01)

Normalização

- 0 para A,
- 1 para B,
- 2 para C e D,
- 3 para E e F, e
- 4 para G e H.

Tabela de variáveis

DESCRIÇÃO DA VARIÁVEL	DESCRIÇÃO DAS CATEGORIAS
Código da categoria administrativa da IES	1 = Pública 2 = Privada
Código da região de funcionamento do curso	1 = Norte 2 = Nordeste 3 = Sudeste 4 = Sul 5 = Centro-Oeste
Idade do inscrito em 25/11/2012	-
Sexo do inscrito	M = Masculino F = Feminino
Ano de conclusão do 2º grau	-
Ano de início da graduação	-
Qual o grau de dificuldade desta prova na parte de Formação Geral?	A = Muito fácil. B = Fácil. C = Médio. D = Difícil. E = Muito difícil.
Qual o grau de dificuldade desta prova na parte do Componente Específico?	A = Muito fácil. B = Fácil. C = Médio. D = Difícil. E = Muito difícil.
Somando a sua renda com a renda dos familiares que moram com você, quanto é, aproximadamente, a renda familiar? (Considere a renda de todos os seus familiares que moram na sua casa com você)	A = Nenhuma. B = Até 1,5 salário mínimo (até R\$ 697,50) C = Acima de 1,5 até 3 salários mínimos (R\$ 697,51 a R\$ 1.395,00). D = Acima de 3 até 4,5 salários mínimos (R\$ 1.395,01 a R\$ 2.092,50). E = Acima de 4,5 até 6 salários mínimos (R\$ 2.092,51 a R\$ 2.790,00). F = Acima de 6 até 10 salários mínimos (R\$ 2.790,01 a R\$ 4.650,00). G = Acima de 10 até 30 salários mínimos (R\$ 4.650,01 a R\$ 13.950,00). H = Acima de 30 salários mínimos (mais de R\$ 13.950,01).
Assinale a situação abaixo que melhor descreve seu caso (incluindo bolsa)	A = Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas. B = Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos. C = Tenho renda e me sustento totalmente. D = Tenho renda, me sustento e contribuo com o sustento da família. E = Tenho renda, me sustento e sou o principal responsável pelo sustento da família.
Indique a resposta que melhor descreve sua atual situação de trabalho. (Não contar estágio, bolsas de pesquisa ou monitoria)	A = Não estou trabalhando. B = Trabalho eventualmente. C = Trabalho até 20 horas semanais. D = Trabalho mais de 20 horas semanais e menos de 40 horas semanais. E = Trabalho em tempo integral - 40 horas semanais ou mais.
Durante o curso de graduação:	A = Não fiz nenhum tipo de estágio. B = Fiz ou faço somente estágio obrigatório. C = Fiz ou faço somente estágio não obrigatório. D = Fiz ou faço estágio obrigatório e não obrigatório.



Limpeza de Dados

Nos três últimos, com categorias A,B,C,D e E possuiriam valores 1,2,3,4 e 5, respectivamente.

Classificador Naïve Bayes





Classificador Naïve Bayes

- Método intuitivo que usa as probabilidades de cada atributo pertencente a cada classe para fazer uma previsão
- Abordagem de aprendizado supervisionado para modelar probabilisticamente um problema de modelagem preditiva¹
- O algoritmo “Naive Bayes” é um classificador probabilístico baseado no “Teorema de Bayes”

¹ É a prática de extrair informações de conjuntos de dados, a fim de determinar padrões e resultados futuros.



Cálculo do Teorema de Bayes

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

Diagram illustrating the components of Bayes' Theorem:

- $P(c | x)$ is labeled **Probabilidade posterior** (Posterior Probability).
- $P(x | c)$ is labeled **Probabilidade** (Likelihood).
- $P(c)$ is labeled **Probabilidade original da Classe** (Prior Probability).
- $P(x)$ is labeled **Preditor da probabilidade posterior** (Evidence).

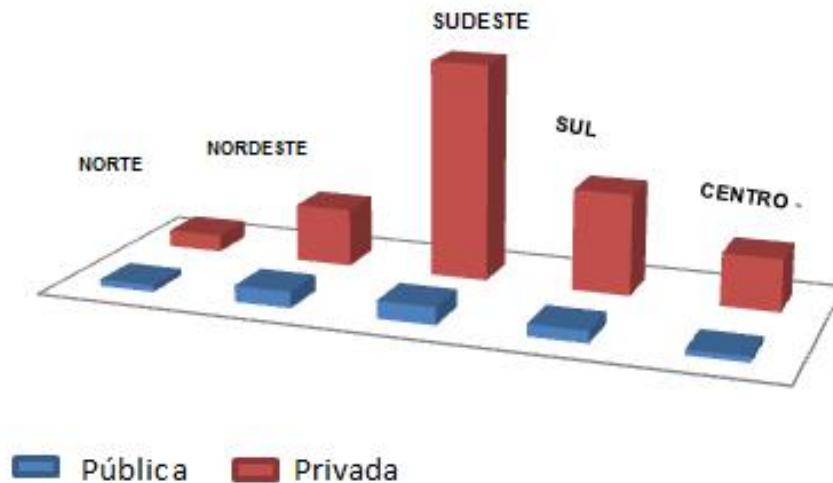
Resultados





Gráfico 1

Quantidade de candidatos de IES's Públicas e Privadas por Região segundo ENADE 2012



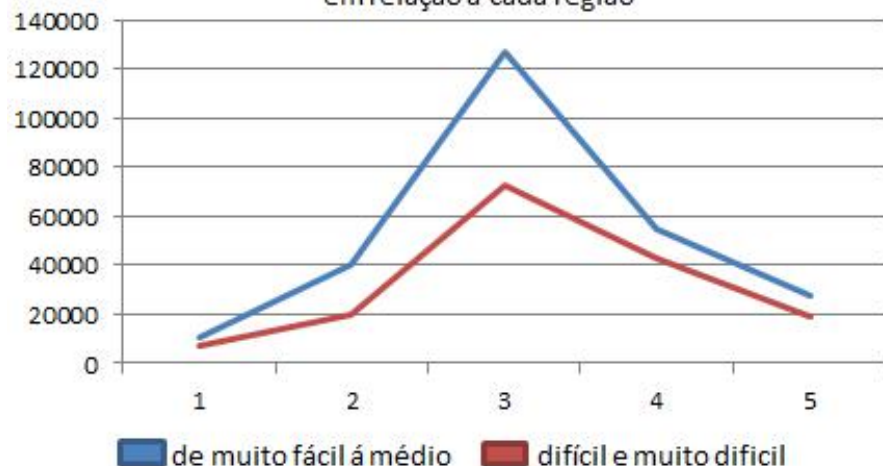


Gráficos 2 e 3

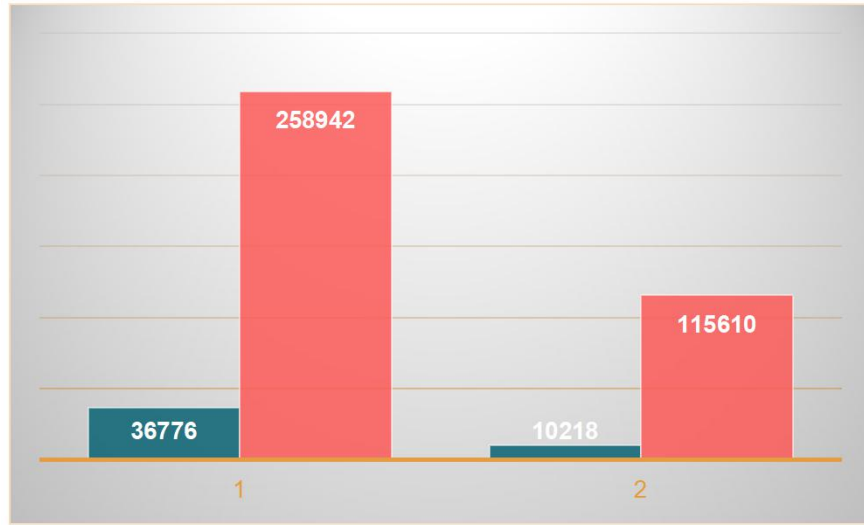
Escala de dificuldade da prova na parte de Formação Geral em relação a cada região





Escala de dificuldade da prova na parte do Componente Específico em relação a cada região



Gráficos 4 e 5



 De muito fácil a médio

 De difícil a muito difícil

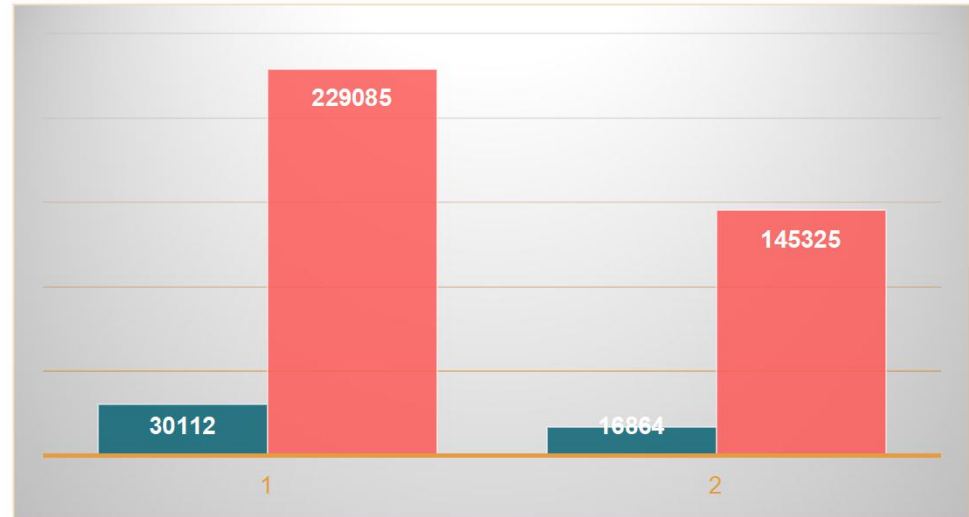
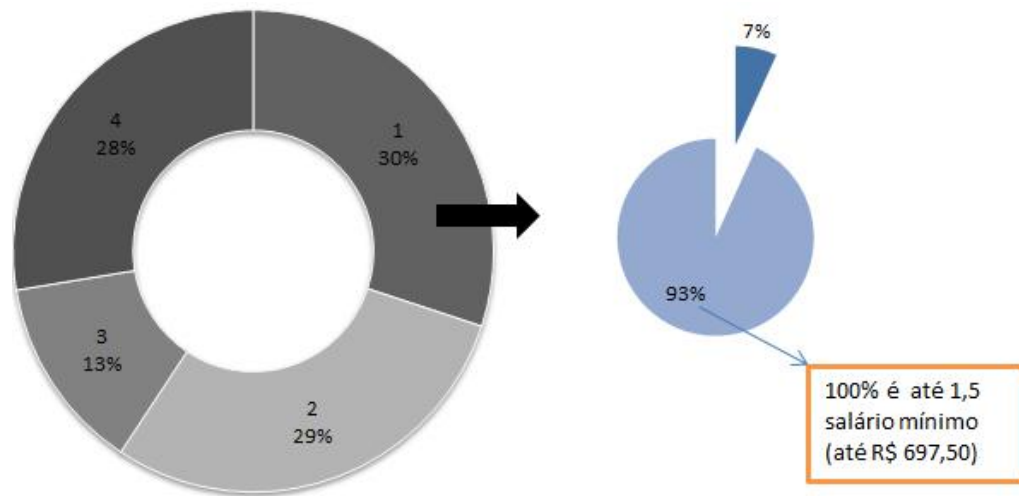


Gráfico 6



O gráfico 4 mostra a porcentagem de candidatos que durante a graduação

1 = Não fez nenhum tipo de estágio.

2 = Fiz ou faço somente estágio obrigatório.

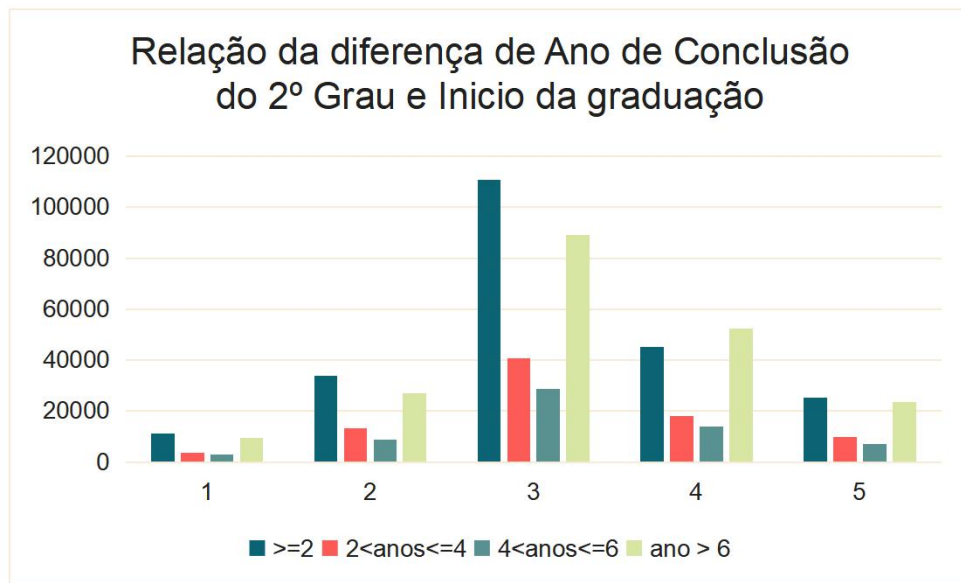
3 = Fiz ou faço somente estágio não obrigatório.

4 = Fiz ou faço estágio obrigatório e não obrigatório

Como mostra no gráfico 30% não fez nenhum tipo de estágio. E desses 93% é da rede privada e 100% dessa rede possui salário mínimo de até 1.5.



Gráfico 7



Obrigada!

