Data Mining na base de dados do ENADE 2012 e a aplicação do classificador Naive Bayes

Ana Flávia Frontino da Cruz
Graduanda em Ciência da Computação na Universidade do Espirito Santo (CEUNES).
anaflaviafrontino@outlook.com

Thayza Sacconi Guarnier
Graduanda em Ciência da Computação na Universidade do Espirito Santo (CEUNES).
thayzasaconi@gmail.com

Resumo

O Enade avalia o rendimento dos alunos dos cursos de graduação, ingressantes e concluintes, em relação aos conteúdos programáticos dos cursos em que estão matriculados. E os resultados do Enade podem produzir resultados gerados pela base de dados disponibilizada pelo mesmo. A metodologia aqui apresentada permite com que, com a utilização de Data Mining (Mineração de dados) investiguem-se dados a procura de padrões que tenham valores explicativos. Vamos gerar esses padrões em cima de cada região. Assim, espera- se descobrir que região tem maior desempenho por instituição de educação superior, categoria administrativa, organização acadêmica e segundo situação socioeconômica.

Palavras – chave

Data Mining, ENADE, Descoberta de conhecimento, Base de dados, Naive Baye.

1. INTRODUÇÃO

1.1 ENADE

O Enade (Exame Nacional de Desempenho de Estudantes) avalia o rendimento dos alunos dos cursos de graduação, ingressantes e concluintes, em relação aos conteúdos programáticos dos cursos em que estão matriculados.

O objetivo principal do exame é acompanhar o processo de aprendizagem e o desempenho acadêmico dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, suas habilidades para ajustamento as exigências decorrentes da evolução do conhecimento e suas competências para compreender temas exteriores ao âmbito especifico de sua profissão, ligados a realidade brasileira e mundial e a outras áreas do conhecimento.

Os resultados do Enade poderão produzir dados por instituição de educação superior, categoria administrativa, organização acadêmica, município, estado e região. Assim, serão constituídos referencias que permitam a definição de ações voltadas para melhoria da qualidade dos cursos de graduação, por parte de professores,

técnicos, dirigentes e autoridades educacionais. (ENADE, 2009)

1.2 DATA MINING

Data Mining, mineração de dados é a uma das três (3) etapas do processo de KDD (Knowledge Discovery in Databases) onde faz– se a descoberta de padrão em grande volume de dados disponível, com o auxilio automatizado onde vários algoritmos e técnicas podem ser usados para encontrar esses padrões.

A mineração de dados está apoiada principalmente em estatística e no aprendizado de máquina (a capacidade do algoritmo aprendizado aprender e melhorar o que ele está fazendo a partir de treino).

FIGURA 1

Esquema do Data Mining

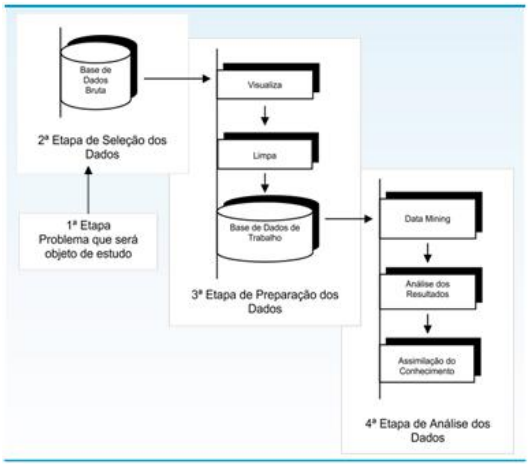
Por que mineração de dados?

É relevante basicamente por três (3) motivos: O avanço nas tecnologias da informação (com o aumento e o barateamento da capacidade de processamento e armazenamento de dados). O aumento do volume dos dados disponíveis, gerados pelas mais diferentes fontes. E a necessidade de transformar os dados em informações e a informação em conhecimento, para gerar a inteligência de negócios.

Uma visão geral das etapas desenvolvidas no DM é mostrada na figura 2. O processo inicia-se com a definição clara do problema: 1ª etapa-, seguida da 2ª etapa, que é a seleção a fim de identificar todas as fontes internas e externas de informação e selecionar o subconjunto de dados necessário para aplicação de DM, que contemple o problema. A 3ª etapa corresponde à preparação dos dados, que inclui o pré- processamento, sendo a que exige maior esforço. Está dividida em ferramentas de visualização e ferramentas de reformatação dos dados. (Quoniam, Tarapanoff, de Araújo Junior, & Alvares, 2001)

FIGURA 2

Etapas do processo de Data Mining



1.3 CLASSIFICADOR NAIVE BAYES

O algoritmo “Naive Bayes” é um classificador probabilístico baseado no “Teorema de Bayes”, o qual foi criado por Thomas Bayes (1701 - 1761) para tentar provar a existência de Deus. Atualmente, o algoritmo se tornou popular na área de Aprendizado de Máquina (*Machine Learning*) para categorizar textos baseado na frequência das palavras usadas, e assim pode ser usado para identificar se determinado e-mail é um SPAM ou sobre qual assunto se refere determinado texto, por exemplo. A principal característica do

algoritmo, e também o motivo de receber “Naive” (ingênuo) no nome, é que ele desconsidera completamente a correlação entre as variáveis (*features*). Ou seja, se determinada fruta é considerada uma “Maçã” se ela for “Vermelha”, “Redonda” e possui “aproximadamente 10cm de diâmetro”, o algoritmo não vai levar em consideração a correlação entre esses fatores, tratando cada um de forma independente. (Candiago, 2017)

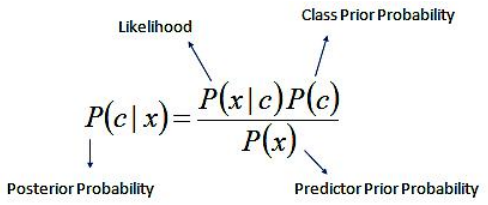
Teorema de Bayes

O teorema de Bayes pode ser resumido pela seguinte fórmula:

Ler se a Figura 3: a probabilidade do valor C ocorrer dado o valor x é igual probabilidade do valor x ocorrer dado o evento C vezes a probabilidade do valor C sobre a probabilidade do valor x.

FIGURA 3

Esquema do cálculo do Teorema de Bayes



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

.

Propõe – se por meio deste trabalho, demonstrar a aplicação da técnica de Data Mining, usando como estudo de caso a base de dados do ENADE 2012. Vamos aplicar em cima de cada região. Assim, espera- se descobrir que região tem maior desempenho por instituição de educação superior, categoria administrativa, organização acadêmica e segundo situação socioeconômica.

2 METOLOGIA

O INEP fornece os resultados do ENADE, com a base de dados que podemos trabalhar em cima. Escolheu- se a base de 2012. 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 12 variáveis para trabalho apresentadas na tabela 1.

Escolheu- se trabalhar com o programa Python, pois possui o módulo CSV. O CSV (Comma Separated Values) módulo implementa classes para ler e gravar dados tabulares no formato CSV. Ele permite que os programadores digam “escreva esses dados no formato preferido pelo Excel” ou “leia os dados desse arquivo que foi gerado pelo Excel”, sem conhecer os detalhes precisos do formato CSV usado pelo Excel. Os programadores também podem descrever os formatos de CSV compreendidos por outros aplicativos ou definir seus próprios formatos de CSV para fins especiais. Apresentado implementação no Código 1.

No entanto, pensou- se em trabalhar apenas com valores inteiros. Assim, alterou- se a base de dados. A variável “sexo do inscrito” com valores “F” ou “M” possuiriam valores 1 e 0, respectivamente.

TABELA 1

DICIONÁRIO DE VARIÁVEIS - ENADE 2012

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.

A variável “Somando a sua renda com a renda dos familiares que moram com você, quanto é, aproximadamente, a renda familiar?” com categorias 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) possuiriam valores 0 para A, 1 para B, 2 para C e D, 3 para E e F, e 4 para G e H. Nos três últimos, com categorias A,B,C,D e E possuiriam valores 1,2,3,4 e 5, respectivamente.

Até aqui se utilizou ferramentas do Excel.

Com a base limpa, temos a base de dados de trabalho, o processo de implementação se deu partindo do ponto do que querer-se descobrir com os dados limpo. Começamos com: Qual a porcentagem por regiões de IES’s Públicas e Privadas?

Essa pergunta foi respondida pelo Gráfico 1 gerado através do Código 2.

E qual o grau de dificuldade que os candidatos responderam desta prova na parte de Formação Geral em cada região? Para responder essa pergunta, dividi em escalar 1, 2 e 3 (de muito fácil á médio) e 4 e 5(difícil e muito difícil). Os resultados mostrados no gráfico 2. Apenas a região 1 (Norte) possuía mais candidatos que acharam a prova 'difícil e muito difícil' do que de 'muito fácil á médio'. Esse resultado foi gerado pelo código 3. O mesmo aplicou - se para o grau de dificuldade desta prova na parte do Componente Específico mostrado no gráfico 3. Observou - se que a escala 'prova difícil e muito difícil' não ultrapassou a escala de 'muito fácil á médio' em nenhuma região, mais a escala de 'prova difícil e muito difícil' obteve um aumento significativo pelo menos para a região 3 (Sudeste).

Os gráficos 4 e 5 mostram a porcentagem de grau de dificuldade por Instituições Publicas e Privadas. Mesma analogia do de cima. Sendo analisado o grau de dificuldade da parte de Formação Geral (Gráfico 4) e o grau de dificuldade da parte de Componente Específicos (Gráfico 5). Podemos ver que não tem diferença significativa das constituições para os candidatos.

O gráfico 6 mostra a porcentagem de candidatos que durante a graduação "1 = Não fiz 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" (implementação Código 4). 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. Conclui – se que candidato de rede privada com renda de até 1.5 salário mínimos não tem oportunidade de estágio (implementação Código 5).

Se perguntarmos para os dados qual a distribuição de probabilidade por região para um candidato ter entrado na faculdade com diferença de até dois (2) anos que conclui o ensino médio. A resposta esta no Gráfico 5, gerado pela implementação Código 6. Nota – se que a região sudeste possui quase metade da proporção. Gerando probabilidade de ser a região maior índice de aluno de ensino médio entrar na faculdade com até dois (2) anos de conclusão do 2º Grau.

Por fim, usaremos o algoritmo de Naive Bayes para classificar a região que será o C pela Figura 3. O meu X será:

X=(2,<=24,<=2,1)

(da categoria administrativa da IES, idade do candidato, diferença de ano de ingresso na IES e ano de termino do ensino médio, renda familiar)

X	Probabilidade
1	0.06522000433028592
2	0.09456170074504
3	0.12202453928386318
4	0.08936735624003095
5	0.1058338752535064

Como P(X|3)P(3) > P(X|5)P(5) > P(X|2)P(2) > P(X|4)P(4) > P(X|5)P(5) => Classe = 3

Assim concluímos que a região 3 possui maior probabilidade de um estudante que ingressar na universidade com diferença de 2 anos que terminou o ensino médio, se formar com com ate 24 anos em uma rede privada e possuir renda de até 1.5 salários mínimos.

RESULTADOS DAS ANÁLISES

GRÁFICO 1

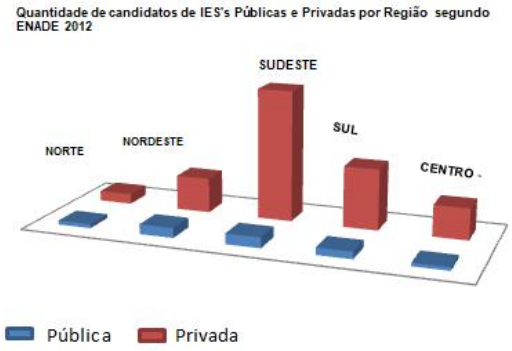


GRÁFICO 2

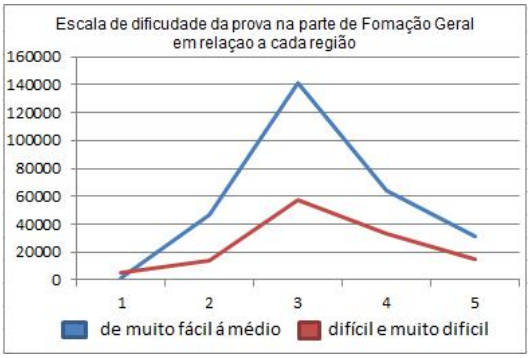


GRÁFICO 3

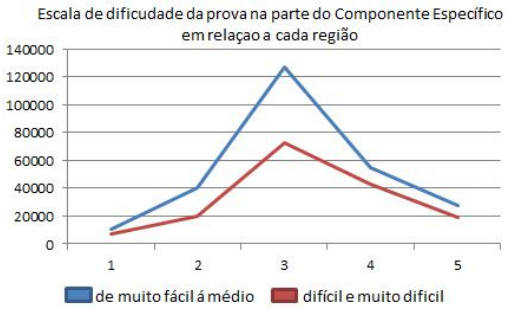


GRÁFICO 4

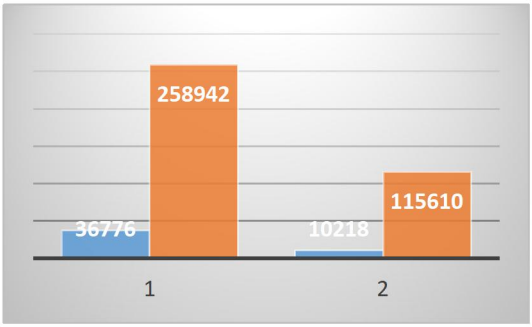


GRÁFICO 5

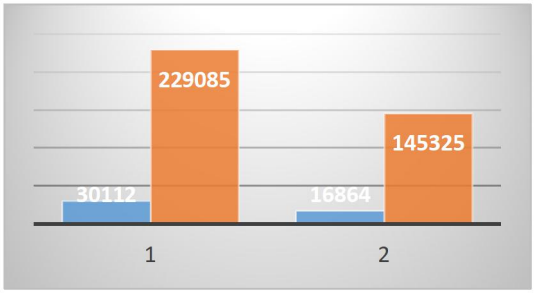


GRÁFICO 6

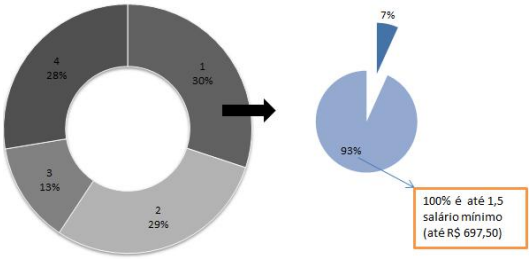
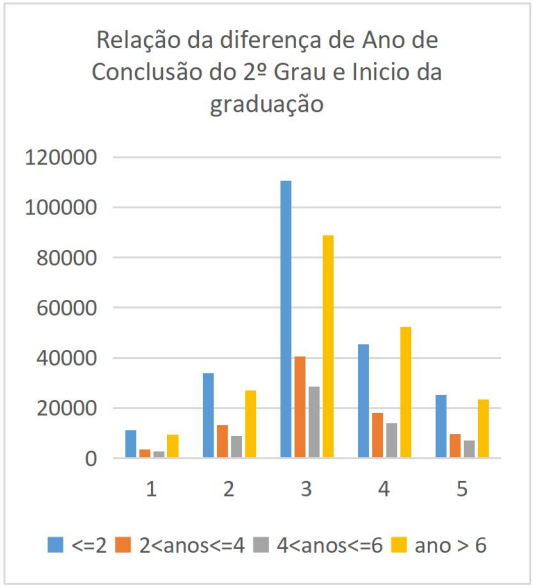


GRÁFICO 7



CÓDIGOS

CÓDIGO 1

```
#le o arquivo...
arq  = open("novo4.csv", "r")

#separa em linhas
linhas =  arq.readlines()
instancias = []
#transforma as linhas em vetores de string
for x in linhas:
   #poe na lista nova de dados a linhas convertida pra um vetor de string
   #strip, retira o ";"
   #split separa pelo ";"
   instancias.append(x.strip().split(';'))
```

CÓDIGO 2

```
def main():
   regiao =1

   while regiao<=5:
      print('Regiao: ', regiao)
      cont1 =0
      cont2 =0
      for x in instancias:
         if(int(x[1])== regiao and (x[0]) == '1'):
            cont1= cont1+1
         if(int(x[1])== regiao and (x[0]) == '2'):
            cont2= cont2+1

      print(" Qnt publica = ",cont1)
      print(" Qnt privada = ",cont2)
      print('\n')
      regiao = regiao +1

main()
```

CÓDIGO 3

```
def prob(regiao):
   cont1 =0
   cont2 =0
   for x in instancias:
      if(int(x[1])== regiao and ((x[6]) == '1' or (x[6]) == '2' or (x[6]) == '3')):
            cont1= cont1+1
      if(int(x[1])== regiao and ((x[6]) == '4' or (x[6]) == '5')):
            cont2= cont2+1

   print("A|B|C = ",cont1)
   print("D|E = ",cont2)

def main():
   regiao =1
   while regiao<=5:
      print('Regiao', regiao)
      prob(regiao)
      regiao = regiao +1

main()
```

CÓDIGO 4

```
def main():
   categoria =1
   while categoria<=4:
      print('categoria: ', categoria)
      cont1 =0
      for x in instancias:
         if(x[12] == str(categoria)):
               cont1= cont1+1
      print("Qnt: ",cont1)
      categoria = categoria +1
      print('\n')"'
```

CÓDIGO 5

```
def main():
      categoria =1
      while categoria<=4:
         cont1=0
         for x in instancias:
            if(x[12]== '1' and  x[0] == '2' and x[8]=='3'):
                  cont1= cont1+1
         print("Qnt de : ", categoria, " = ",cont1)
         categoria = categoria+1
```

CÓDIGO 6

```
def prob(regiao):
   cont1 =0
   for x in instancias:
         if(int(x[1])== regiao and (int(x[5])-int(x[4])) <= 2):
               cont1 =cont1+1
   print(" qnt = ",cont1)

def main():
   regiao =1

   while regiao<=5:

      print('Regiao: ', regiao)
      prob(regiao)
      regiao = regiao +1
```

CÓDIGO 7

```
def cont_valor_atributo(atributo, valor):
   cont =0
   for x in instancias:
      if(int(x[atributo]) == valor):
         cont = cont+ 1
   return cont

def cont_igual( atributo1, valor1, atributo2, valor2):
   cont =0
   for x in instancias:
      if(int(x[atributo1])== valor1 and int(x[atributo2]) == valor2):
            cont= cont+1
   return cont

def cont_menorigual( atributo1, valor1, atributo2, valor2):
   cont =0
   for x in instancias:
      if(int(x[atributo1])== valor1 and int(x[atributo2]) <= valor2):
            cont= cont+1
   return cont

def cont_dif_conclusao_anodeinicio( atributo1, valor1, atributo2, atributo3, valor2):
   cont =0
   for x in instancias:
      if(int(x[atributo1])== valor1 and (int(x[atributo3])-int(x[atributo2])) <= valor2):
            cont= cont+1
   return cont


def prob_cada_reg(regiao):
   qnt_reg = cont_valor_atributo(1,regiao)
   qnt_regeIES2 = cont_igual(1,regiao,0,2)
   qnt_regeidademenor_igual24 = cont_menorigual(1,regiao,2,24)
   qnt_regedif = cont_dif_conclusao_anodeinicio(1,regiao,4,5,2)
   qnt_regerenda = cont_igual(1,regiao,8,1)

   prob = (qnt_regeIES2/qnt_reg)*(qnt_regeidademenor_igual24/qnt_reg)*
         (qnt_regedif/qnt_reg)*(qnt_regerenda/qnt_reg)
   print(prob)

def main():
   regiao =1
   while regiao<=5:
      prob_cada_reg(regiao)
      regiao = regiao +1
```

CONCLUSÕES

A análise da Base de Dados do Enade 2012 foi reveladora em relação às áreas do conhecimento escolhidas, escolher o valor região para trabalhar, mostrou que as regiões estão com diferentes trajetórias em aplicação de suas metodologias. Podem-se encontrar muitas mais informações relevantes.

REFERÊNCIAS BIBLIOGRÁFICAS

Candiago, L. (13 de 04 de 2017). *Seeds*. Acesso em jul de 2018, disponível em Orgânica Digital: https://www.organicadigital.com/seeds/algoritmo-de-classificacao-naive-bayes/

ENADE. (Agosto de 2009). Acesso em Junho de 2018, disponível em INEP: http://download.inep.gov.br/download/enade/2009/Manual_2009_atualizado.pdf

Quoniam, L., Tarapanoff, K., de Araújo Junior, R. H., & Alvares, L. (maio/ago de 2001). Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. *Ci. Inf*, pp. 20-28.