KDD (Knowledge Discovery in Databases)
Documento de Implementação

André Vitor Santana Souza

Paulo Mauricio Dourado Fernandes

Prof. Dr. André Britto Carvalho

Sumário

1. Introdução

Este relatório da disciplina de sistemas de apoio à decisão utiliza a mineração de dados no processo de extração de conhecimento na base de dados. Um sistema de apoio a decisão tem como objetivo auxiliar o administrador na tomada de decisão, fornecendo dados sobre a decisão a ser tomada e suas implicações. O trabalho exposto utilizou o processo de *Knowledge Discovery in Databases(KDD), para seleção e pré-processamento dos dados, seguido da transformação e mineração de dados e por fim avaliação e interpretação da base de dados* Censo de Educação Superior 2019 do INEP. A ferramenta utilizada foi o Weka, que é um software open source para automatizar os processos do KDD. Esse software engloba diversos algoritmos de machine learning voltados para data mining.

1.1 Objetivo

Este relatório tem como objetivo explorar os microdados do INEP do Censo de Educação Superior 2019[3], e utilizá-lo para obtenção de conhecimento, tentar prever a situação(Cursando, Formado, Matrícula Trancada, Desvinculado do curso, Transferido) do estudante perante a universidade.

1.2 Visão geral do documento

1.2.1 O capítulo 2 apresenta o processo de KDD abordando as etapas importantes para concluir o processo.

1.2.2 O capítulo 3 explica como foi aplicada cada uma das etapas do KDD detalhando as decisões tomadas durante o processo.

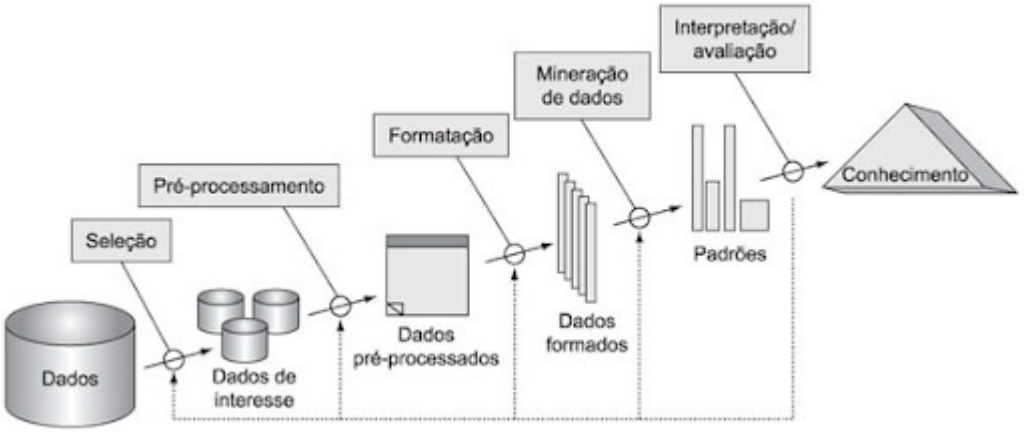
1.2.3 O capítulo 4 apresenta uma discussão dos resultados obtidos(tabelas, gráficos, árvores, etc.) após a aplicação de cada um dos algoritmos de mineração de dados.

1.2.4 O capítulo 5 corresponde a conclusão do projeto.

2. KDD (*Knowledge Discovery in Databases*)

KDD é um processo iterativo e interativo que busca encontrar conhecimento contido em um enorme grupo de dados com o uso de técnicas de Data Mining. O seu uso se faz de suma importância nos dias atuais considerando a enorme quantidade de dados produzida atualmente. Inevitavelmente esses dados podem conter informações relevantes para gerar novos conhecimentos. O processo do KDD supre exatamente essa necessidade, fornecendo um método consistente para entender e analisar esses dados.

Basicamente o KDD é dividido em 5 etapas conforme mostra a figura abaixo. A seleção de dados, pré-processamento dos dados, transformação dos dados, mineração dos dados e a avaliação/análise dos dados. Esse processo pode ser repetido várias vezes para refinar os resultados a fim de se confirmar o objetivo escolhido e consolidar o conhecimento.



2.1 Seleção dos dados

Essa etapa é responsável por selecionar e criar a base de dados onde o processamento ocorre. De acordo com os objetivos definidos deve ser feita uma busca por dados relevantes, talvez até integrar diferentes fontes de dados em um único conjunto. Também é importante selecionar os atributos que vão permanecer no conjunto. É importante selecionar cautelosamente os dados e os atributos, pois impactam diretamente na eficiência da mineração. Caso os dados sejam enviesados ou os atributos importantes sejam descartados, muito provavelmente o resultado da mineração não será preciso.

2.2 Pré-processamento dos dados

Pré-processamento pode ser resumido como uma limpeza dos dados. Nessa etapa o foco é remover inconsistências nos dados, como valores ausentes, outliers ou algum ruído. Existem diversos métodos de realizar essa etapa, incluindo o uso de algum algoritmo de mineração para predizer os valores.

2.3 Transformação dos dados

Nesta etapa é feita a manipulação dos dados para um maior aproveitamento na etapa da mineração. Diversos métodos podem ser utilizados com esta finalidade, incluindo redução de dimensão, conversão de tipos de atributos(discretização ou normalização de atributos numéricos), combinação de atributos, binarização, etc. É crucial que os atributos se adequem para o algoritmo escolhido a seguir, pois, por exemplo, alguns se aplicam somente em atributos nominais ou apenas em atributos numéricos.

2.4 Mineração dos dados

Este pode ser considerado o núcleo do KDD. Primeiro é necessário escolher qual tarefa de mineração vai ser executada, por exemplo, regras de associação, agrupamento ou classificação. A escolha da tarefa está diretamente relacionada com o objetivo escolhido para o KDD. Regras de associação geralmente são utilizadas para encontrar atributos que implicam na presença de outros e dessa forma descobrir como os atributos se relacionam. A classificação, por outro lado, percorre todo o conjunto de dados e atribui cada um a determinadas classes. Essa tarefa foca em encontrar um modelo preditivo para que os dados futuros sejam classificados corretamente. Por último a tarefa de aglomeração que busca encontrar grupos naturais dentro do conjunto de dados por meio da semelhança entre

eles. Geralmente é utilizado quando não se tem informações suficientes sobre o contexto, pois pode encontrar conexões desconhecidas.

Após a escolha da tarefa adequada é o momento de escolher o algoritmo a ser utilizado que é o mais eficiente para o objetivo escolhido. Podem ser escolhidos diferentes algoritmos e testá-los com diferentes parâmetros para obter o melhor resultado possível.

2.5 Avaliação e interpretação

Esta é a última etapa do KDD. Os resultados(Gráficos, Matriz de confusão, Árvore de decisão,etc) obtidos da etapa anterior devem ser analisados e avaliados até que sejam incorporados, a fim de refinar o próprio processo e, caso necessário, o processo deve ser reiniciado utilizando como base o novo conhecimento adquirido.

3. Implementação

3.1 Seleção dos dados

Os dados selecionados foram obtidos através dos dados abertos do INEP, e foram escolhidos os Microdados relativos aos discentes do Censo da Educação Superior 2019. Esses dados possuem 1048575 instâncias e 105 diferentes atributos. Através da seleção de atributo do WEKA com o método de busca ranker e o avaliador de atributos InfoGainAtributeEval, os atributos considerados como mais relevantes (avaliado com mais que 0.01) para o objetivo estão na tabela abaixo. Muitos atributos foram considerados redundantes ou não exibiram uma correlação alta com o nosso objetivo, como por exemplo atributos que especificam deficiência do discente, tipo de financiamento do curso, tipo de cota de ingresso, nacionalidade ou data de nascimento.

```
Ranked attributes:
 0.773212362170368      89 IN_MATRICULA
 0.5175131173342429     90 IN_CONCLUINTE
 0.41419347083795155    34 QT_CARGA_HORARIA_INTEG
 0.2328382366627153      4 CO_CURSO
 0.14451527874000392    35 DT_INGRESSO_CURSO
 0.13685634025394622    93 NU_ANO_INGRESSO
 0.12355313667805268    33 QT_CARGA_HORARIA_TOTAL
 0.11563960053470668     1 CO_IES
 0.08735153773249116    13 NU_IDADE
 0.06947538681974175    10 CO_CINE_ROTULO
 0.0524488006173951     83 TP_SEMESTRE_REFERENCIA
 0.04501492281446362    17 CO_MUNICIPIO_NASCIMENTO
 0.04225716161115178    92 IN_INGRESSO_VAGA_NOVA
 0.04169434298850372    91 IN_INGRESSO_TOTAL
 0.02920407860675089    71 IN_ATIVIDADE_EXTRACURRICULAR
 0.02168479493550213    16 CO_UF_NASCIMENTO
 0.01713618350952917     2 TP_CATEGORIA_ADMINISTRATIVA
 0.01418435235570836    64 IN_APOIO_SOCIAL
 0.01213894993180586    52 IN_FINANCIAMENTO_ESTUDANTIL
 0.01177329158877938    80 TP_ESCOLA_CONCLUSAO_ENS_MEDIO
 0.01013329307211763     6 TP_TURNO
```

Imagem 1. Ranking dos atributos em função do atributo Situação.

NOME DA VARIÁVEL	DESCRIÇÃO	TIPO	DESCRIÇÃO DAS CATEGORIAS
co_ies	Código único de identificação da IES	num	código da Instituição de Ensino.
tp_categoria_administrativa	Código da Categoria Administrativa	num	1. Pública Federal 2. Pública Estadual 3. Pública Municipal 4.Privada com fins lucrativos 5. Privada sem fins lucrativos 7. Especial
co_curso	Código único de identificação do curso	num	código do Curso
tp_turno		num	1.Manhã 2.Tarde 3.Noite 4.Integral
co_cine_rotulo	Código da Área do curso do Discente	num	
nu_idade	Idade que o aluno completa no ano	num	Idade do discente

	de referência do Censo		
co_municipio_nascimento	Código do município de nascimento do aluno	num	código do município de nascimento
tp_situacao	Código do tipo de situação de vínculo do aluno no curso	num	2.Cursando 3. Matrícula trancada 4. Desvinculado do curso 5. Transferido para outro curso da mesma IES 6. Formado 7. Falecido
qt_carga_horaria_total	Carga horária total do curso do discente	num	
qt_carga_horaria_integ	Carga horária integralizada do discente	num	
in_apoio_social	Informa se o aluno recebe algum tipo de apoio social na forma de moradia, transporte, alimentação, material didático e bolsas (trabalho/permanência)	num	0. Não 1. Sim
tp_escola_conclusao_ens_medio	Tipo de escola que o aluno concluiu ensino médio	num	0. Privada 1. Pública 2. não dispõe da informação
tp_semestre_referencia	Informa o semestre de referência do preenchimento do vínculo do curso	num	1. Primeiro semestre 2. Segundo semestre
in_matricula	Informa se o aluno é matriculado no curso	num	1 - situação de matrícula; 0 - situação diferente de matrícula
in_concluinte	Informa se o aluno é concluinte	num	1 - situação de concluinte; 0 - situação diferente de concluinte
in_ingresso_vaga_nova	Informa se o aluno é ingressante no curso por meio de processo seletivo de vaga nova.	num	1. Situação de ingresso por processo seletivo de vaga nova; 0. Situação diferente de ingresso por processo seletivo de vaga nova.
in_ingresso_total	Informa se o aluno é ingressante no curso, não importando a forma de ingresso utilizado.	num	1. Situação de ingresso total; 0. Situação diferente de ingresso total
nu_ano_ingresso	Ano de ingresso do aluno no curso	num	
in_atividade_extracurricular	Informa se o aluno participa de algum tipo de atividade extracurricular (estágio não obrigatório, extensão, monitoria e pesquisa)	num	0. Não 1. Sim
in_financiamento_estudantil	Informa se o aluno utiliza financiamento estudantil	num	0. Não 1. Sim

3.2 Pré-processamento dos dados

Na etapa do Pré-processamento foi feita a limpeza dos dados.Muitos atributos apresentaram taxas altas de *MissingValues, o que significa que algumas instâncias não possuem qualquer valor para estes atributos.* Portanto foi utilizado o filtro do Weka *ReplaceMissingValues* para substituir os valores com o valor médio do atributo.

Para o tratamento de *outliers* foram consideradas apenas as instâncias cujo valor do atributo NU_ANO_INGRESSO fosse superior à 2006, pois representavam menos de 0,001% dos dados e poderiam influenciar negativamente o atributo. Também foram desconsideradas as instâncias que o atributo TP_SITUACAO tivesse valor 7, pois representava discentes que morreram, logo esse valor não tem qualquer correlação com o restante dos atributos e representaria uma dissonância.

3.3 Transformação dos dados

Nessa etapa foi feita a transformação dos dados para um maior aproveitamento dos algoritmos utilizados na mineração de dados. Os atributos numéricos que possuíam apenas 2 valores foram transformados em atributos binários. Já os atributos categóricos que possuíam mais de 2 valores foram transformados em atributos nominais. Por fim, os atributos numéricos foram discretizados para ser feita a conversão para nominal. Portanto, para realizar essas transformações, foram utilizados os filtros fornecidos pelo software Weka. A tabela abaixo descreve qual filtro foi utilizado com cada atributo.

NOME DA VARIÁVEL	FILTRO	TIPO FINAL
co_ies	NumericToNominal	nominal
tp_categoria_administrativa	NumericToNominal	nominal
co_curso	NumericToNominal	nominal
tp_turno	NumericToNominal	nominal
co_cine_rotulo	NumericToNominal	nominal
nu_idade	NumericToNominal Discretize	nominal
co_municipio_nascimento	NumericToNominal	nominal
tp_situacao	NumericToNominal	nominal
qt_carga_horaria_total	NumericToNominal Discretize	nominal
qt_carga_horaria_integ	NumericToNominal Discretize	nominal
in_apoio_social	NumericToBinary	binary
tp_escola_conclusao_ens_medio	NumericToBinary	binary
tp_semestre_referencia	NumericToBinary	binary
in_matricula	NumericToBinary	binary
in_concluinte	NumericToBinary	binary
in_ingresso_vaga_nova	NumericToBinary	binary
in_ingresso_total	NumericToBinary	binary

nu_ano_ingresso	NumericToNominal	nominal
in_atividade_extracurricular	NumericToBinary	binary
in_financiamento_estudantil	NumericToBinary	binary

3.4 Mineração dos dados

Para atingir o objetivo do projeto, que é a predição do atributo TP_SITUACAO, utilizamos a tarefa de classificação. Essa tarefa consiste em encontrar um modelo preditivo que consiga definir a classe de uma instância através da comparação com outra instância já rotulada.Weka dispõe de diversos algoritmos de classificação, dentre eles os algoritmos selecionados foram o Naive Bayes, J48, BayesNet, REPTree.

3.4.1 Naive Bayes

Naive Bayes é um algoritmo classificador probabilístico baseado no teorema de Bayes. Para fazer a predição dos atributos é calculado a probabilidade do valor do atributo usando um conjunto de testes. Após isso, o algoritmo calcula a probabilidade do dado novo possuir cada um dos valores e atribui o valor mais provável. O algoritmo foi utilizado na ferramenta Weka da forma padrão, afinal ele não possui parâmetros.

3.4.2 J48

Esse algoritmo é a implementação em java do algoritmo C4.5. Ele é utilizado para gerar uma árvore de decisão. Essa árvore é gerada com o uso de um conjunto de testes de dados que estão classificados corretamente. Posteriormente o algoritmo faz a predição dos valores com base na árvore de decisão gerada. Os parâmetros utilizados na ferramenta Weka para este algoritmo são confidenceFactor = 0.25 e minNumObj = 2.

3.4.3 BayesNet

Esse é um modelo probabilístico baseado na construção de um grafo estabelecendo a conexão entre os atributos de acordo com a sua dependência. Alguns algoritmos utilizam esse modelo para prever o valor dos atributos de novos dados. A ferramenta Weka permite escolher o algoritmo para construir o modelo e para fazer a predição. Foi utilizado o SimpleEstimator com alpha = 0.5 e o K2 com maxNrOfParents = 1.

3.4.4 REPTree

Esse algoritmo faz o uso de uma técnica chamada *pruning.* Essa técnica consiste em reduzir o tamanho da árvore de decisão removendo nós redundantes ou que não sejam tão importantes. Os parâmetros utilizados foram minNum = 2, minVarianceProp = 0.001, numFolds = 3, maxDepth = -1 e initialCount = 0.0.

4. Resultados

Foram realizados os testes a partir do conjunto de dados já descrito com 104589 instâncias referentes à classificação da situação do aluno na universidade. Para a classificação das instâncias utilizou-se os algoritmos, NaiveBayes, J48, BayesNet e REPTree.e o REPTree. Dentre eles o J48 foi o escolhido como melhor modelo pois apesar de não apresentar o melhor resultado ele é o que possui os erros mais aceitáveis como visto na sua matriz de confusão.

4.1 Naive Bayes

O Naive Bayes teve a maior porcentagem de acerto como exposto anteriormente com 93.4533%, e é o melhor em todas as métricas, sendo sua única desvantagem vista na tabela de confusão onde todos as classes tiveram erro apesar de ser 1 ou 2 em algumas classes ele erra classes que não possuem conexão, para esses pequenos erros.

```
=== Summary ===

Correctly Classified Instances       33232               93.4533 %
Incorrectly Classified Instances      2328                6.5467 %
Kappa statistic                          0.8765
Mean absolute error                      0.0294
Root mean squared error                  0.1404
Relative absolute error                 13.8678 %
Root relative squared error             43.0912 %
Total Number of Instances            35560

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     2
               0,658    0,027    0,701      0,658    0,679      0,649    0,973     0,764     3
               0,775    0,039    0,728      0,775    0,750      0,716    0,976     0,804     4
               0,492    0,006    0,560      0,492    0,524      0,518    0,963     0,525     5
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     6
Weighted Avg.  0,935    0,007    0,934      0,935    0,934      0,927    0,994     0,948

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 23405     1     0     2     0 |     a = 2
     1  2089  1027    60     0 |     b = 3
     1   779  3237   162     0 |     c = 4
     0   109   185   285     0 |     d = 5
     1     0     0     0  4216 |     e = 6
```

4.2 J48

o J48 teve uma taxa de acerto de 93.4196%, e ele possui a vantagem de so errarem na classificação entre as classes Formando, matrícula trancada e desvinculado do curso que no fim podem ser classificados como Não Cursando, pois não há atributos que ajude a diferenciar esses 3.

```
=== Summary ===

Correctly Classified Instances       33220               93.4196 %
```

```
Incorrectly Classified Instances      2340               6.5804 %
Kappa statistic                          0.8758
Mean absolute error                      0.0314
Root mean squared error                  0.1428
Relative absolute error                 14.8244 %
Root relative squared error             43.832  %
Total Number of Instances            35560

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     2
               0,660    0,031    0,673      0,660    0,667      0,634    0,920     0,638     3
               0,767    0,039    0,726      0,767    0,746      0,711    0,940     0,709     4
               0,508    0,003    0,721      0,508    0,596      0,600    0,858     0,468     5
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     6
Weighted Avg.  0,934    0,007    0,934      0,934    0,934      0,927    0,984     0,925

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 23408     0     0     0     0 |     a = 2
     0  2097  1041    39     0 |     b = 3
     0   900  3204    75     0 |     c = 4
     0   117   168   294     0 |     d = 5
     0     0     0     0  4217 |     e = 6
```

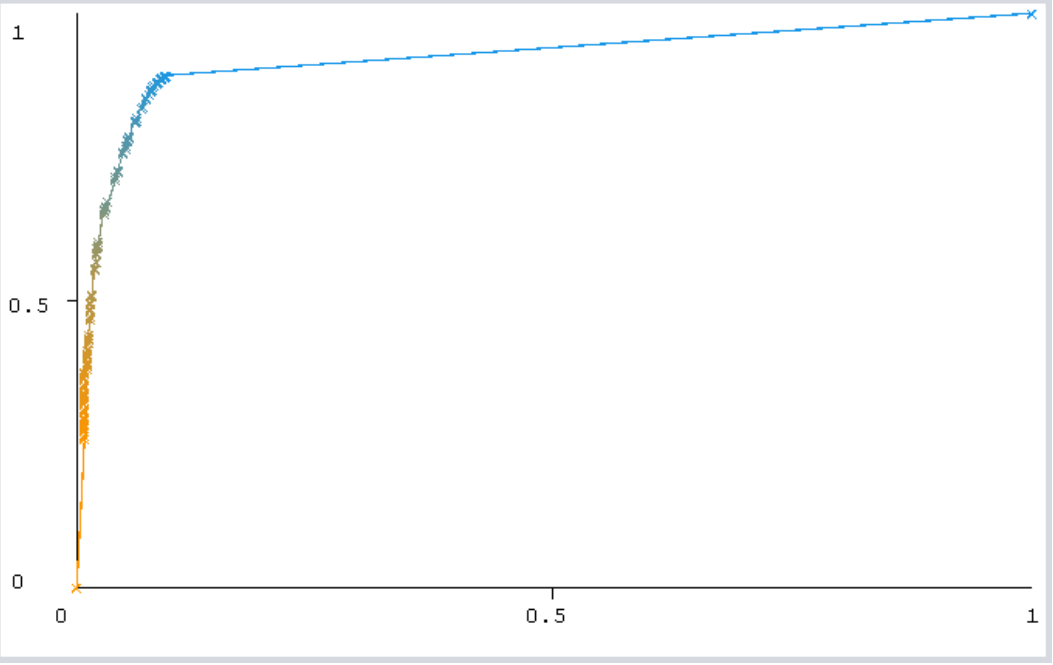


Gráfico 1. Threshold da classe b

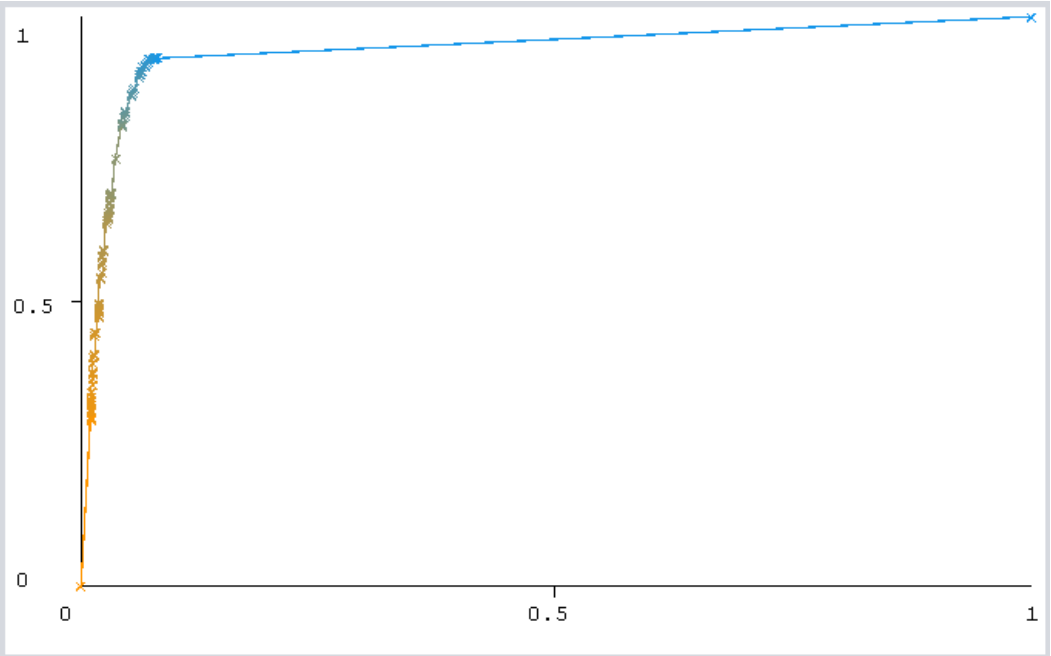


Gráfico 2. Threshold da classe c

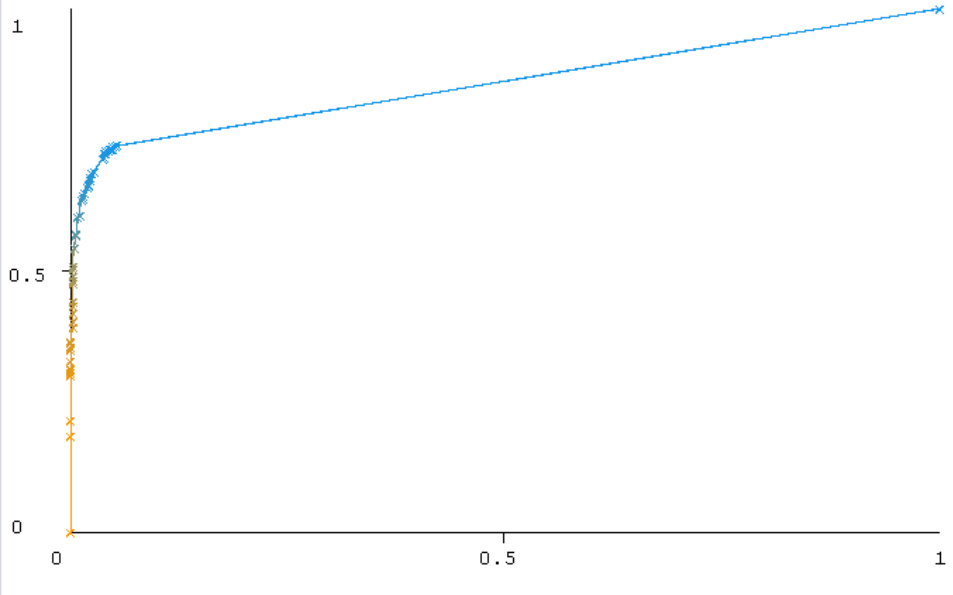


Gráfico 3. Threshold da classe d

4.3 BayesNet

O BayesNet teve 93.4308% de taxa de acerto e o é o segundo melhor nas métricas e ele possui o mesmo problema do naive Bayes visto na tabela de confusão.

```
=== Summary ===

Correctly Classified Instances       33224               93.4308 %
Incorrectly Classified Instances      2336                6.5692 %
Kappa statistic                          0.8761
```

```
Mean absolute error                      0.0292
Root mean squared error                  0.1416
Relative absolute error                 13.751  %
Root relative squared error             43.4818 %
Total Number of Instances            35560

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     2
               0,666    0,029    0,695      0,666    0,680      0,650    0,973     0,765     3
               0,766    0,038    0,731      0,766    0,748      0,714    0,976     0,804     4
               0,494    0,007    0,555      0,494    0,523      0,516    0,967     0,535     5
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     6
Weighted Avg.  0,934    0,007    0,934      0,934    0,934      0,927    0,994     0,948

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 23403     0     0     5     0 |     a = 2
     0  2116   995    66     0 |     b = 3
     1   817  3203   158     0 |     c = 4
     0   111   182   286     0 |     d = 5
     1     0     0     0  4216 |     e = 6
```

4.4 REPTree

O REPTree teve uma porcentagem de 93.1046% de taxa de acerto e possui a mesma vantagem do J48 visto na tabela de confusão.

```
=== Summary ===

Correctly Classified Instances       33108               93.1046 %
Incorrectly Classified Instances      2452                6.8954 %
Kappa statistic                          0.8699
Mean absolute error                      0.0328
Root mean squared error                  0.1426
Relative absolute error                 15.4469 %
Root relative squared error             43.7836 %
Total Number of Instances            35560

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     2
```

```
                 0,661    0,035    0,653      0,661    0,657      0,623    0,925     0,646     3
                 0,747    0,039    0,718      0,747    0,732      0,696    0,947     0,713     4
                 0,453    0,003    0,710      0,453    0,553      0,561    0,863     0,463     5
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     6
Weighted Avg.    0,931    0,008    0,931      0,931    0,931      0,923    0,985     0,926

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 23408     0     0     0     0 |     a = 2
     0  2101  1040    36     0 |     b = 3
     0   988  3120    71     0 |     c = 4
     0   130   187   262     0 |     d = 5
     0     0     0     0  4217 |     e = 6
```

5. Conclusão

Foi apresentado neste relatório o uso do WEKA para a mineração de dados do Censo de Educação Superior 2019 do INEP. Foram feitos os seguintes procedimentos: seleção dos dados, pré-processamento, transformação dos dados, mineração de dados. Foram testados alguns algoritmos de classificação e foram expostos os 4 melhores resultados, dentre eles o J48 se apresentou o melhor modelo apesar dele não possuir o melhor desempenho seu erro como visto anteriormente deve-se às classes serem bastantes semelhantes. Por fim, esse relatório conseguiu apresentar um bom modelo de classificação quanto à situação de um estudante da universidade.

Referências

Censo da Educação Superior. Governo do Brasil. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Acesso em: 24, de julho de 2021

Weka 3: Machine Learning Software in Java. Waikato. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 20, de julho de 2021.