RANKING DE UNIVERSIDADES NO BRASIL: UM ESTUDO DE PREDIÇÃO COM MACHINE LEARNING

Cleverson Tabajara Vianna
Mestre em Administração
Doutorando em Engenharia e Gestão do Conhecimento na Universidade Federal de Santa Catarina
tabajara@ifsc.edu.br

Sergio Nicolau da Silva
Especialista em Engenharia de Projetos de Software
Analista de Tecnologia da Informação no Instituto Federal de Santa Catarina
sergio.silva@ifsc.edu.br

Fernando Alvaro Ostuni Gauthier
Doutor em Engenharia de Produção
Professor na Universidade Federal de Santa Catarina
gauthier@egc.ufsc.br

Antônio Pereira Cãndido
Doutor em Engenharia de Produção
Professor no Instituto Federal de Santa Catarina
apec@ifsc.eu.br

RESUMO

Objetivo: Utilizar técnicas de *Machine Learning* para predição do Ranking Universitário da Folha (RUF), com base no histórico do ano anterior para treinar o algoritmo *Naïve Bayes.*
Design/Metodologia/Abordagem: Pesquisa aplicada, descritiva, de objetivo exploratório e abordagem qualitativa e quantitativa. Foram extraídos dados do RUF, da CAPES, homogeneizados, e aplicaram-se métodos de engenharia valendo-se de diversas ferramentas (*WEKA*, *KDD*, *Data Mining*, *Postgre*, *ETL Pentaho*).
Resultados: A predição realizada com precisão de 61,5%. Podendo reduzir a incerteza ao reduzir o número de classificados no ranking.
Limitações da pesquisa (se aplicável): Utilização de software *opensource* disponíveis que possibilitam a aplicação dos conceitos, sem necessidade de desenvolvimento de software.
Implicações práticas (se aplicável): Evidencia a viabilidade de um modelo com base estatística para determinar a qualidade de uma instituição de ensino
Implicações sociais (se aplicável): apresenta uma alternativa possível de pré-avaliação da qualidade do ensino, sem necessidade de aguardar o fechamento do RUF
Originalidade/valor: Uso de técnicas de *machine learning* para aferição/predição da qualidade de instituições de ensino, índice que está inserido em um contexto complexo e interdisciplinar.

Palavras-chave: Ranking Universitário da Folha. Algoritmo *Naïve Baye*. Data Mining. KDD.

BRAZIL'S UNIVERSITY RANKING: A PREDICTION STUDY WITH MACHINE LEARNING

ABSTRACT

Goal: *The use of Machine Learning techniques to predict the Ranking Universitário da Folha (RUF), using previous year's history to train the Naïve Bayes algorithm*
Design / Methodology / Approach: *Applied research, descriptive, exploratory objective and qualitative and quantitative abortion. Data were extracted from the RUF, CAPES, homogenized, and engineering methods were applied using several tools (WEKA, KDD, Data Mining, Postgre, ETL Pentaho).*
Results: *The accuracy predicted was 61.5%. Uncertainty van be reduced, by reducing the number of classified Universities in the ranking.*
Limitations of the research (if applicable): *Use of available opensource software that enables the application of concepts, without the need for software development*
Practical implications (if applicable): *Proposes a statistical-based model to determine the quality of an educational institution*
Social implications (if applicable): *presents a possible alternative of pre-evaluation of teaching quality, without waiting for RUF final ranking*
Originality / value: *Use of machine learning techniques to gauge / predict the quality of Higher Education, an index that is inserted in a complex and interdisciplinary context.*

Keywords: *Brazil's University Ranking - RUF. Naïve Baye Algorithm. Data Mining. KDD.*

1 INTRODUÇÃO

Qualidade da Educação Superior, avaliação, ranqueamentos são os temas abordados neste artigo, no entanto, mesmo dada à relevância do tema, ponto central explorado, é a utilização de algoritmos de mineração para a predição deste ranqueamento. No entanto, se faz obrigatória a contextualização do surgimento e importância deste ranqueamentos.

O tópico preliminar, que evidencia a relevância do tema, se refere à qualidade do Ensino Superior:

A qualidade se tornou um tema central na agenda da educação superior. Em que pese ser amplamente utilizado, esse termo não consegue reunir consensos no campo educacional. Porém, para todos os efeitos práticos, a falta de entendimentos quanto ao conceito não chega a ser problema. Mais ainda, o conceito de qualidade nem mesmo é posto em foco de discussão. Juntamente com o tema da qualidade surgem as questões da garantia da qualidade e da acreditação. (SOBRINHO, 2008, p.817)

A partir dos anos 1990, a maioria dos países latino-americanos, criaram seus organismos de avaliação do Ensino Superior (SOBRINHO, 2006). No Brasil, a acreditação, que no Brasil significa em última instância, a "autorização de funcionamento", é atribuição governamental regulada pelo Sistema Nacional de Avaliação da Educação Superior - SINAES. (SOBRINHO; VESSURI, 2006) (RISH, 2001).

Uma vez que todas as IES no Brasil têm obrigatoriamente uma acreditação, ou autorização do governo para atuar, como distinguir as melhores ou as piores? Esta é uma pergunta que permeia a mente e corações de pais, alunos e professores, pois educação é um investimento no futuro pessoal e da nação. Há alguns recursos disponíveis, mas sendo governamentais, tem a mesma origem e não evidenciam uma independência de avaliação como tendo origem na própria sociedade.

Precisamente neste "vácuo de informação", surge o Ranking Universitário da Folha - RUF. Conhecido por sua tradicional avaliação, o Ranking da Folha, é tido como um instrumento independente de avaliação e que proporciona um ranqueamento das melhores universidades brasileiras. O RUF é elaborado sob a responsabilidade da Folha de São Paulo (iniciada em 1921),

e envolve diversos mecanismos, tendo o objetivo de ranquear as 195 melhores universidades do país, públicas ou privadas. Sua execução fica a cargo do DATAFOLHA[1].

Segundo a Folha de São Paulo (2016), em seu próprio site temos: O RUF avalia as 195 universidades brasileiras com base em 5 indicadores: Pesquisa científica; Qualidade do Ensino; Internacionalização; Mercado de trabalho; Inovação.

Os dados são obtidos através de várias fontes, incluindo duas pesquisas anuais, envolvendo milhares de respondentes, e os dados são coletados em bases como:

a. Inep-MEC
b. Web of Science
c. SciELO
d. Inpi
e. FAPs
f. CNPq
g. Capes
h. Duas pesquisas anuais Datafolha

A pergunta que nos motivou a esta pesquisa é:

Com que grau de certeza, analisando apenas os dados fornecidos pela CAPES, relativos à pós-graduação das IES, poderemos predizer se uma IES estará ou não no RUF?

Para responder a esta pergunta, estabelecemos o objetivo de fazer predição do RUF com base nos dados de pós-graduação. Estes dados são fornecidos através de dados abertos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). A seguir, determinou-se sua representatividade através de uma matriz de confusão.

Para atingir estes objetivos foram utlizadas ferramentas da Engenharia do Conhecimento visando estabelecer um ranqueamento e comparando-o com o RUF. Utilizamos ferramentas e técnicas de Mineração de Dados, Classificação, *Machine Learning* e Algoritmos de Recomendação e Predição e probabilidades.

2 CONSTRUÇÃO TEÓRICA

As pesquisas nas Universidades, estão geralmente atreladas aos Grupos de pesquisa, liderados por pós-graduados, em sua maioria doutores. Assim é plausível a hipótese de que a influência da estrutura e funcionamento dos programas de Mestrado e Doutorado, é elevada no RUF, ainda mais que "pesquisa e qualidade de ensino" são partes relevantes do RUF, como se

[1] DATAFOLHA: Criado nos anos 70, o Banco de Dados de da Folha de São Paulo, abrange os arquivos de foto, texto e a biblioteca da Folha de São Paulo

apresenta ao analisar a construção do RUF e sua estrutura. Nesta seção, observa-se como o RUF é construído e as ferramentas de Mineração de dados que darão suporte ao experimento.

Descrevemos brevemente o que é e como é composto o RUF

2.1 ESTRUTURA DO RUF E OS DADOS ABERTOS DA CAPES

Ao observar-se a estrutura de formação do RUF, temos que 74% dos dados voltam-se diretamente para a Pesquisa Científica (42%) e a Qualidade de Ensino (32%) sendo que os demais tópicos Mercado de Trabalho, Internacionalização e Inovação, se somados representam os 26% restantes[2]. Diante disso, havendo uma predominância de dados relativos à pesquisa e como a pesquisa é em geral atribuição da pós-graduação, nos surgiu a percepção de que embora o ranking seja voltado para o ensino superior, os dados da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), poderiam ter papel preponderante neste ranking.

Para Gonçalves (2006) várias abordagens de base estatísticas são propostas para a predição e aprendizagem de máquina, as quais se valem de algoritmos de agrupamento visando o estabelecimento de padrões: *k-means* e Bayesianos são exemplos.

Bayes era um filósofo inglês do século XVIII que expôs sua teoria da probabilidade em 1763. A regra que leva seu nome tem sido uma pedra angular da teoria da probabilidade desde então. A dificuldade com a aplicação da regra de Bayes na prática é a atribuição de probabilidades prévias (WITTEN; FRANK, 2011).

Neste trabalho, foi utilizado o algoritmo de *Naïve Bayes*, com a abordagem de *Supervised Learning* (aprendizado de máquina supervisionado) o qual é baseado em métodos probabilísticos (FULMARI; CHANDAK, 2014). Utilizou-se o ano de 2015, com os dados abertos da CAPES e o RUF 2015 como treinamento e efetuou-se a predição do RUF-2016, valendo-se dos dados da CAPES de 2015. A seguir comparou-se a predição (pelo algoritmo) do ano de 2016 do RUF, com os resultados divulgados do RUF. Também através do algoritmo J48, buscou-se estabelecer uma

[2] Há inclusive uma inconsistência de valores que deve ser fruto de digitação, pois apresenta os percentuais diferentes entre o total e a soma das partes; vide item “qualidade de ensino” com 30% e não 32% que é o correto.

árvore de decisão, mas que devido ao elevado número de ramos não é viável e com a "poda", torna-se pouco significativo. Como ferramenta, utilizamos *WEKA*.

2.2 DESCOBERTA DE CONEHCIMENTO

Descoberta de conhecimento é o processo aplicado sobre dados estruturados, semiestruturados e não estruturados, com o objetivo de verificação da hipótese de usuários ou a descoberta de novos padrões. Essa ainda se pode subdividir em outros dois objetivos: a previsão de comportamento futuro baseado na análise dos dados históricos e a apresentação de padrões identificados na análise de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p.7).

Descoberta de Conhecimento em Banco de Dados – do inglês *KDD Knowledge Data Discovery* – é a área que dispõe de mecanismos e técnicas para análise de dados estruturados. Para Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 40), o *KDD* pode ser visto como uma atividade multidisciplinar, pois engloba técnicas e além do escopo da disciplina, como a aprendizagem de máquina. Como parte do *KDD*, *Data Mining* atua na extração de informações úteis de banco de dados.

2.3 DATA MINING

No mundo atual o volume de dados digitais armazenados em repositórios eletrônicos cresce em um ritmo acelerado, fazendo uma grande migração de empresas de software para atuar nas tecnologias de *big data* e dados abertos, isto conforme estudos publicados em 2015 da IDC: *Worldwide Tecnologia Big Data* e *Serviços Forecast*, 2015-2019 (IDC # 259532) e o *Worldwide Big Data Forecast* por *Vertical Market*, 2014-2019 (IDC # US40544915).

O termo *Data Mining*, que significa mineração de dados em português, tem sido utilizado por estatísticos, analistas de dados e comunidades de sistemas de informação na área de gestão e mais popularmente relacionados diretamente com banco de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 39). Tal processo de mineração de dados é suportado por técnicas que atuam na forma de treinamento e teste com base em dados históricos, reconhecendo, assim, padrões. Este método é característico das técnicas de aprendizagem de máquina para

reconhecimento de padrões como a classificação, a clusterização, o agrupamento, entre outros (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 43).

Cada uma das técnicas possuem uma diversidade de algoritmos possíveis disponíveis e suas variações. Para o objetivo deste artigo será abordado a classificação por meio do algoritmo *Naïve Bayes*.

2.4 CLASSIFICAÇÃO

A classificação é um processo que constantemente estamos realizando ao longo da história e em nosso cotidiano. Classificamos os meios de transporte em aéreo, terrestre e marítimo, pessoas maiores de idade e menores de idade e as classes econômicas da população são alguns exemplos.

O processo de classificação consiste em examinar as características de determinado objeto a ser classificado e atribuir a este uma ou mais classes (LINOFF; BERRY, 2011, p. 9). Quando é apresentada a idade de uma pessoa, por exemplo, aplicando a regra da maioridade vigente, é possível classificar o indivíduo em maior ou menor de idade.

Em *data mining*, os objetos a serem classificados geralmente são representados por registros em uma tabela de banco de dados ou um arquivo, nos quais é adicionada uma coluna que representará sua classe. A tarefa de classificação é caracterizada por uma definição de classes distintas que é identificada a partir de um conjunto de treinamento composto de exemplos pré-classificados (LINOFF; BERRY, 2011, p. 9).

A classificação por si só não é suficiente em casos complexos para tomada de decisão automatizada, mas é um excelente norteador para tomada de decisão em **atividades intensivas de conhecimento**. Assim, ao buscar identificar o risco do cliente em cumprir suas obrigações, a técnica busca predizer o futuro. Por exemplo: com base na experiência passada, poderá estabelecer numa instituição financeira quais riscos/confiança para receber empréstimos concedidos.

Para Linoff e Berry (2011, p. 10):

> […] qualquer uma das técnicas utilizadas para a classificação e estimativa pode ser adaptada para uso na previsão usando exemplos de treinamento onde o valor da variável a ser predito já é conhecido, juntamente com dados históricos para esses exemplos. Os dados históricos são usados para construir um modelo que explique o comportamento

> atual. Quando este modelo é aplicado às entradas atuais, o resultado é uma previsão de comportamento futuro.

Desta forma, é aplicado a classificação para o caso de estudo em questão.

Existem inúmeros algoritmos para classificação de informações. ID3 e C4.5 são alguns exemplos de algoritmos de classificação, que utilizam abordagem simbólica[3]. Outros algoritmos como *Naïve Bayes* e *K-Neighboards* possuem uma abordagem estatísticas[4] e com várias implementações. O software *WEKA*[5] é exemplo de software que implementa diversos algoritmos relacionados a data minig e extração do conhecimento.

Para o caso de estudo que prevê predição baseado em casos, utilizou-se algoritmo *Naïve Bayes*. Para Witten, Frank e Hall (2011, p.97), "*Naïve Bayes* é uma técnica popular para esta aplicação, porque é muito rápido e bastante preciso". O algoritmo *Naïve Bayes* é bastante efetivo quando aplicado em conjuntos de dados e combinados com procedimentos de seleção e eliminando redundâncias.

Algoritmos baseados em *Naïve Bayes*, que calculam probabilidades explícitas para hipóteses estão entre as abordagens mais práticas para certos tipos de problemas de aprendizagem. Pesquisas mostram que o classificador *Naïve Bayes* pode superar o desempenho de algoritmos baseados em árvore de decisão e até mesmo redes neurais (MITCHELL, 1997, p. 154).

3 METODOLOGIA

3.1 CLASSIFICAÇÃO METODOLÓGICA

A classificação metodológica desta pesquisa a caracteriza como de natureza Aplicada, uma vez que produz resultados imediatos, no entanto é também básica ao servir de base para outras pesquisas (MARCONI; LAKATOS, 2010). A pesquisa tem objetivo descritivo, na medida que descreve características de um fenômeno e estabelece relações entre variáveis. Ao buscar estabelecer limites, e abordagens para novas pesquisas, delimitando uma área desconhecia, é

[3] classifica com base em árvores de decisão como se tem sol então não chove

[4] verificam qual a probabilidade de um evento ocorrer

[5] Disponível em http://www.cs.waikato.ac.nz/ml/weka/

caracterizada também como tendo objetivo exploratório. Apresenta também um objetivo explicativo, pois "aprofunda o conhecimento da realidade porque explica a razão, o porquê das coisas. (GIL, 2002, p.28). Possui uma abordagem qualitativa, na medida que os pesquisadores atribuem significados aos dados; por outro lado é quantitativa, pois segue os rigores estatísticos, não valendo-se apenas de amostras, mas de todo o universo que envolve as IES. Utilizou-se procedimentos bibliográficos, documentais e experimentais (GIL, 2008).

3.2 METODOLOGIA DA PESQUISA

A pesquisa propriamente dita, seguiu os passos seguintes:

1. Obtenção dos dados abertos da CAPES relativos aos anos de 2014 e 2015, referente a discentes, docentes e cursos. Estes dados foram tratados e preparados, compondo um banco de dados relacional. A seguir foram obtidos os dados RUF de 2015 e 2016, e da mesma forma foram tratados e carregados em tabelas de banco relacional.
2. Tivemos a necessidade então de minerar os dados, preparando uma tabela de conversão DE/PARA, compatibilizando as SIGLAS IES de ambos os sistemas (CAPES e RUF). Esta foi um tarefa exaustiva que inclusive, apresentou 2 incompatibilidades que não puderam ser resolvidas e que fazem parte da análise geral dos dados.
3. A seguir os dados da CAPES foram sumarizados, contemplando Cursos de Mestrado e Doutorado de cada IES, número de professores, alunos concluintes. Os preditores foram cada um destes campos sumarizados, e a decisão obtida era se pertencia ou não ao RUF, compatibilizando assim as Tabelas RUF e CAPES.
4. Seguindo os conceitos de aprendizagem de máquina, utilizamos os dados de 2014 como "teste", ensinando a "máquina". Para tanto, nos valemos do Software WEKA[6], onde aplicamos o algoritmo *Naïve Bayes*.
5. A seguir, submetemos os dados de 2015, visando estabelecer a predição, de quais IES estariam no RUF 2016 e o comparamos com o resultado efetivo.
6. Estes dados foram então comparados, e se estabeleceu uma matriz de confusão, indicando tanto os falsos positivos como os negativos. Os resultados são interessantes, pois com apenas dados abertos se obteve um resultado significativo, não necessitando de *surveys*, entrevistas, e outros dados não abertos (como quantidade de publicações, citações, etc.).

[6] WEKA é software livre, produzido na Universidade de Waikatu (NZ) e que se constitui numa coleção de algoritmos para o aprendizado de máquina a serem utilizados nas tarefas de mineração de dados.

4 O EXPERIMENTO EM DADOS ABERTOS: CAPES E RUF

O primeiro passo é a coleta dos dados brutos, tanto do RUF quanto da CAPEs.

Do RUF foram coletados os dados do ranking de 2015 e 2016 do site e gerado um arquivo no formato CSV, tal arquivo contém todos os dados do RUF, acrescentando o ano de referência.

Figura 1 – RUF como é apresentado no site da Folha

Ranking 2016 ▲	Nome da Universidade	UF	Pública / Privada	Ensino	Pesquisa	Mercado	Inovação	Internacionalização	Nota
1º	Universidade Federal do Rio de Janeiro (UFRJ)	RJ		31,17	3º	3º	6º	3º	97,46
2º	Universidade de São Paulo (USP)	SP		29,96	1º	1º	1º	5º	97,03
3º	Universidade Estadual de Campinas (UNICAMP)	SP		31,21	2º	11º	2º	11º	96,77
4º	Universidade Federal de Minas Gerais (UFMG)	MG		31,30	7º	2º	3º	9º	96,54
5º	Universidade Federal do Rio Grande	RS		31,15	5º	12º	5º	13º	95,72

Fonte: Ranking Universitário Folha - Ruf (2016)

A seguir os dados foram "normalizados", isto é, preparados para que pudessem ser tratados pelas ferramentas de software.

Os dados abertos da CAPEs são fornecidos em formato CSV, que são um formato mais adequado para processamento do dado em relação ao RUF que é uma página web. Por estar em CSV o tratamento do dado é mais simples que o aplicado ao RUF.

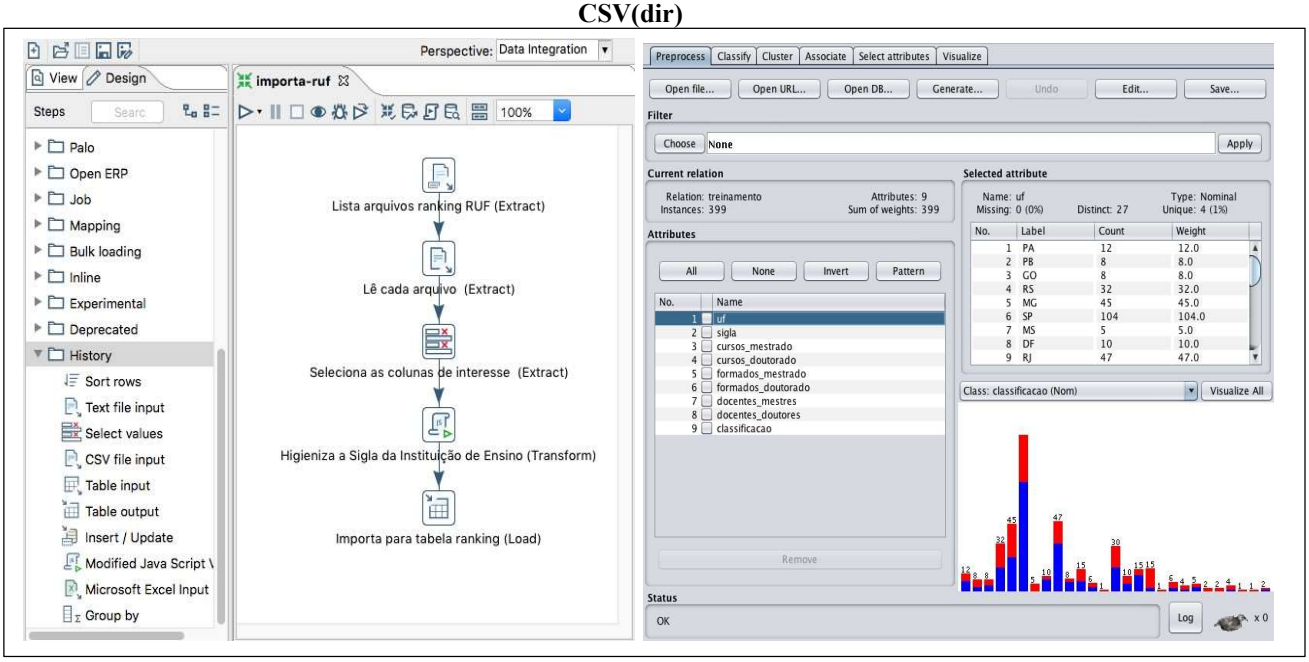
Do site da CAPEs de dados abertos, foram coletados os arquivos referentes a seguintes informações sopre programas de pós-graduação para os anos de 2014 e 2015: cursos, corpo docente e discentes formandos.

De posse dos dados brutos, realizamos a importação dos mesmos para um banco de dados *PostgreSQL* para facilitar o processo de normalização dos dados e extração no formato esperado pelo software *WEKA*. Embora em padrão *CSV*, isto não significa que seus dados estão higienizados[7]. Para este processo de higienização e importação para o banco de dados foi utilizada uma ferramenta de *Extract, Transform and Load (ETL)* (Extrair, Transformar e Carregar) utilizada

[7] processo de padroniza os dados, mantendo sua validade

no processo de *KDD*, mais especificamente na fase de pré-processamento dos da dos. A ferramenta escolhida é a *Data-integration* da Pentaho[8] (figura 2).

Figura 2 – Exemplo ETL aplicado ao RUF (esquerda) e Interface do WEKA com a importação do arquivo CSV(dir)



Fonte: Extraído das ferramentas de software utilizadas

Com a ferramenta *Data-integration*, todos os dados de interesse para a pesquisa foram importados. Mesmo que ambas as fontes de dados tratam do mesmo domínio – instituições de ensino – há divergências entre as siglas das instituições entre base de dados. Mesmo com a higienização, aproximadamente 50 instituições que fazem parte do RUF não eram localizadas nos dados da CAPEs. Para minimizar esta diferença, foi necessária uma análise manual dos dados.

O foi fornecido para o aprendizado da máquina no caso em estudo, é justamente a união entre os dados da CAPES de 2014 com o RUF2015, É que se chama de “massa de treino”, para posteriormente passar para o algoritmo já “treinado”, os novos dados de CAPEs e o mesmo

[8] pode ser obtido na comunidade Pentaho, no link http://community.pentaho.com/ projects/data-integration/

classificar e fazer as predições com base no conhecimento percebido no treinamento. Os dados utilizados foram da CAPES 2014 obtendo-se UF e Sigla da Instituição compatibilizada com RUF.

O arquivo de treinamento foi então importado no *WEKA*, via interface gráfica, para aplicar o algoritmo *Naïve Bayes* e análise da precisão (Figura 2).

Foram realizadas várias análises e testes para identificar uma configuração com o melhor resultado possível. Importante ressaltar que este é uma atividade extremamente importante para o processo de extração do conhecimento e que está ligada a interdisciplinaridade exigida, onde um especialista no assunto contribui para estes ajustes.

Aplicando o algoritmo aos dados de treinamento, o resultado foi um percentual de 78,95% de acerto. A matriz de confusão gerada está a seguir:

```
a   b    <- Classificado como
191 33   a = N
51 124 b = S
```

Com este nível de precisão, os dados de treinamento foram exportados, por meio do *WEKA*, para o formato *ARFF*.

Após o treinamento, temos então de fazer a predição do RUF para 2015. Para tanto foi gerado o arquivo *ARFF* com os dados da CAPEs para classificação pelo algoritmo aprendido pela máquina com os dados de 2015. Novamente o *ARFF* conterá, as mesmas colunas, no entanto a decisão conterá "?", indicando ao algoritmo para calcular:

Utilizando um terminal[9] do sistema operacional OS X, foi executado o comando a seguir para determinar ao *WEKA* que realize a classificação com base nos dados de treinamento:

java -cp weka.jar weka.classifiers.bayes.NaiveBayes -t treinamento.arff -T classificar.arff -p 3-8 –D

Como resultado o *WEKA* apresenta a classificação realizada para cada instância do arquivo a ser classificado. A figura 3 apresenta o resultado parcial do processamento do *WEKA*.

[9] também conhecido como linha de comando ou shell, permite ao usuário solicitar ao sistema operacional qye execute algumas açõescomo listar arquivos, criar diretórios, executar uma aplicação, entre outros

Figura 3 – Dados obtidos do Weka

=== Predictions on test data === inst# actual **predicted** error prediction (cursos_mestrado, cursos_doutorado, formados_mestrado, formados_doutorado, docentes_mestres, docentes_doutores)

Inst#				
01	1:?	1:N	0.996	**(0,2,0,88,22,2)**
02	1:?	1:N	0.994	**(0,2,0,169,28,0)**
03	1:?	1:N	0.988	**(0,3,0,75,95,0)**
…	…	…	…	…
…	…	…	…	…
23	1:?	1:N	0.998	**(1,0,0,0,40,0)**

Fonte: Weka Output

A ultima coluna *predicted* apresenta a predição de cada entrada (dados entre parênteses), informando assim que, uma instituição com aquelas características de cursos, docentes e discentes tende ou não fazer parte do ranking do RUF. O resultado da predição foi normalizado e importado para o banco de dados *PostgreSQL*. Após isto, foi comparado o RUF2016 com o resultado das predições, chegando ao seguinte resultado: das 195 instituições que compõe o RUF 2016, 120 foram preditas pelo processo aplicando *Naïve Bayes* com de acerto de 61,5%.

5 CONSIDERAÇÕES FINAIS

Sobrinho e Ristoff (2005), destacam a importância de critérios objetivos para a avaliação institucional, quando nos afirma que se exigem-se critérios objetivos e procedimentos que privilegiem os aspectos quantitativos e comparáveis.

A publicação de dados abertos ligados, amplia este processo de comparação, permitindo que agentes humanos e não humanos possam tratar e analisar informações. Há assim a percepção da importância dos dados abertos, especialmente aqueles que futuramente possam ser classificados com 5 estrelas, conforme proposta de Berners-Lee (1989):

Figura 1 Representação da classificação 5 Stars Open Data

Classificação	Descrição
	Disponível na web (qualquer formato), mas com uma licença aberta, a ser Open Data
	Disponível como dados estruturados legíveis por máquina (por exemplo, Excel em vez de digitalização de uma imagem de uma tabela)
	Como (2) mais formato não proprietário (por exemplo, CSV ao invés de Excel)
	Tudo o mais acima, use padrões abertos da W3C (RDF e SPARQL) para identificar as coisas, de modo que as pessoas podem apontar em seu material
	Todos os acima, mais: Vincular seus dados para dados de outras pessoas para fornecer contexto

Fonte: Adaptado de Berners-Lee (1989, 2006)

Para Vianna (2014, p.4), o ato de medir embora seja uma parte do processo avaliativo da sociedade sobre as IES, não pode ser isoladamente considerado:

> A avaliação irá expressar as ações, atitudes e valores e tanto de indivíduos, quanto comunidades, ou a própria ciência em si; se possível deverá contemplar as suas múltiplas dimensões e inter-relações. Sempre produzirá efeitos ao longo do tempo, sejam eles políticos ou pedagógicos. Uma parte importante da avaliação se refere aos testes aplicados, questionários a responder e os resultados obtidos – esta é o que chama-se de parte técnica da avaliação; logo, medir faz parte da avaliação, mas a avaliação não se esgota na medição. Isto significa que não basta atribuir-se notas, pesos e conceitos.

Um percentual acima de 60% de acerto do ranking RUF mostra que é possível, com aprofundamento de estudos e análise mais detalhadas das técnicas, a predição com certo grau de confiança. Cabe observar que segundo o RUF a Pesquisa Científica (majoritariamente na pós-graduação) corresponde a um peso de 42% no ranqueamento.

Outra hipótese é fazermos um corte, selecionado as 60 primeiras universidades. Desta forma, um algoritmo para prever as 40, 50 ou 60 melhores universidades brasileiras, baseadas estritamente em dados abertos da CAPEs poderá apresentar um grau de confiança maior.

É observado ainda que há reflexos positivos (acima 60%) dos processos da CAPES sobre a de gestão da qualidade dos Programas de Pós-Graduação das Instituições de Ensino, intrinsecamente ligadas à qualidade do ensino superior.

REFERÊNCIAS

BERNERS-LEE, T. Information management: A proposal. 1989.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37, 1996.

FULMARI, A.; CHANDAK, M. B. An approach for word sense disambiguation using modified naïve bayes classifier. International Journal of Innovative Research in Computer and Communication Engineering, v. 2, Abril 2014.

GIL, A. C. Como elaborar projetos de pesquisa. São Paulo, v. 5, 2002.

GIL, A. C. Métodos e técnicas de pesquisa social. In: Métodos e técnicas de pesquisa social. [S.l.]: Atlas, 2008.

GONÇALVES, A. L. Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento. 2006. 196 f. Tese (Doutorado) — Tese (Doutorado em Engenharia de Produção)-Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2006.

LINOFF, G. S.; BERRY, M. J. Data mining techniques: for marketing, sales, and customer relationship management. [S.l.]: John Wiley & Sons, 2011.

MARCONI, M. d. A.; LAKATOS, E. M. Fundamentos de metodologia científica. In: Fundamentos de metodologia científica. [S.l.]: Atlas, 2010.

MITCHELL, T. M. Machine learning. New York, 1997.

RISH, I. An empirical study of the naive bayes classifier. In: IBM NEW YORK. IJCAI 2001 workshop on empirical methods in artificial intelligence. [S.l.], 2001. v. 3, n. 22, p. 41–46.
SOBRINHO, J. D. Acreditación de la educación superior en américa latina y el caribe. In: TRES, J.; SANYK, B. C. (Ed.). La educación superior en el Mundo 2007. Acreditación para la garantía de la calidad: ¿Qué está en juego? [S.l.]: Global University Network for Innovation, 2006.

SOBRINHO, J. D. Quality, evaluation: from sinaes to indexes. Avaliação: Revista da Avaliação da Educação Superior (Campinas), SciELO Brasil, v. 13, n. 3, p. 815–825, 2008.

SOBRINHO, J. D.; RISTOFF, D. I. Avaliação como instrumento da formação cidadã e do desenvolvimento da sociedade democrática: por uma ético-epistemologia da avaliação. Ristoff,

Dilvo; Almeida JR, Vicente (organizadores). Avaliação Participativa, Perspectivas e Debates, série Educação Superior em Debate, n. 1, p. 15–38, 2005.

RANKING UNIVERSITÁRIO FOLHA (RUF). O que é o RUF. 2016. Disponível em: < http://ruf.folha.uol.com.br/2016/ranking-de-universidades/ >. Acesso em: 9 out. 2017.

SOBRINHO, J. D.; VESSURI, H. Paradigmas e políticas de avaliação da educação superior. autonomia e heteronomia. Universidad e investigación científica: convergências y tensiones. Vessuri H, org. Buenos Aires: CLACSO, Consejo Latinoamericano de Ciencias Sociales, p. 169–191, 2006.

VIANNA, C. T. Avaliação institucional e o desafio da implantação da cultura da autoavaliação - autoavaliação e cpa. In: Anais dos seminários regionais sobre autoavaliação institucional e comissões próprias de avaliação - CPA. [S.l.]: INEP, 2014.

WITTEN, I. H.; FRANK, E. Data Mining: Practical machine learning tools and techniques. 3rd. ed. [S.l.]: Morgan Kaufmann, 2011.