Descoberta de Conhecimento em Base de Dados sobre o perfil de estudantes brasileiros de Tecnologia da Informação

Aline de Campos[1], Sílvio César Cazella[1 2]

[1] Programa de Pós-Graduação em Informática na Educação - Universidade Federal do Rio Grande do Sul (UFRGS)

[2] Programa de Pós-Graduação em Ensino na Saúde - Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)

alinedecampos@gmail.com, silvioc@ufcspa.edu.br

Abstract. *This paper presents the process for Knowledge Discovery in Databases (KDD), focusing on the dataset obtained through a global survey conducted by the Stack Overflow platform in 2018. For this study, we selected answers coming from Brazilian students, totaling 885 instances. The procedures were performed based on the CRISP-DM methodology and the use of four computational tools in order to verify its potentialities of application in KDD. After several experiments focused on the distillation of data for human judgment and clustering, we reached some exploratory interpretations regarding the Brazilian profile of information technology students.*

Resumo. *Este artigo apresenta o processo de Descoberta de Conhecimento em Base de Dados (DCBD) em um conjunto de dados obtido através de um questionário (survey) promovido pela plataforma Stack Overflow no ano de 2018. Selecionou-se o recorte de respostas advindas de estudantes brasileiros, totalizando 885 instâncias. Para tanto, fez-se uso da metodologia CRISP-DM e de quatro ferramentas computacionais a fim de verificar suas potencialidades. Após diversos experimentos focou-se em tarefas de destilação de dados para o julgamento humano (sumarização)* e agrupamentos *(clusterização) conduzindo a algumas interpretações relativas ao perfil brasileiro de estudantes de Tecnologia da Informação*.

1. Introdução

A Descoberta de Conhecimento em Base de Dados, conhecido pela sigla DCBD *(do inglês, Knowledge Discovery in Databases (KDD))* constitui-se em um processo que envolve as etapas de pré-processamento, mineração de dados (MD) e pós-processamento no sentido de extrair conhecimento útil de grandes concentrações de dados. Segundo Fayyad et al. (1996), trata-se do processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis implícitos nos dados.

O uso do DCBD no domínio educacional tem tido aplicação cada vez mais abrangente, tendo oportunizado o surgimento de uma área de pesquisa denominada de Mineração de Dados Educacionais. Conhecida também por sua designação em inglês *Educational Data Mining (EDM)*, trata do desenvolvimento, pesquisa e aplicação de métodos computadorizados para detecção de padrões em grandes coleções de dados educacionais onde seria muito complexa uma análise aprofundada devido ao volume de dados. Sendo assim, pode ser definida como a aplicação de técnicas de mineração em conjunto de dados que são provenientes do âmbito educacional buscando responder importantes questões deste domínio [Romero e Ventura, 2013].

DOI: 10.5753/cbie.wcbie.2018.449

Todos os anos o portal de perguntas e respostas *Stack Overflow*[1] aplica o *"Stack Overflow Developer Survey"* que busca realizar diversas perguntas em um questionário abrangente para verificar o perfil de pessoas que atuam na área de Tecnologia de Informação em um nível mundial. Grande parte dos usuários da plataforma são estudantes de diversos níveis educacionais que buscam esclarecer dúvidas no fórum. Assim, verificou-se o potencial de descoberta de informações relevantes acerca do respondentes uma vez que são coletados dados do perfil social, demográfico, econômico, educacional e profissional, bem como práticas adotadas por pessoas que atuam na área de Tecnologia da Informação.

Com o objetivo de explorar os recursos da área de DCBD no que diz respeito aos aspectos teóricos das tarefas de mineração, bem como os aspectos técnicos de algumas das ferramentas computacionais existentes, apresenta-se este estudo experimental no conjunto de dados correspondente aos estudantes da área de Tecnologia da Informação (TI) tendo em vista um recorte nos dados provenientes do Brasil. Portanto, além da experiência de aplicação, ao usar esse conjunto coletado através do *survey* de 2018 espera-se verificar *insights* interessantes sobre características gerais e conjuntos de perfis, bem como possíveis correlações de elementos.

Além desta introdução, o artigo apresenta na seção 2 a seleção e descrição de ferramentas computacionais utilizadas, na seção 3 os materiais e métodos adotados, bem como a descrição dos procedimentos realizados no experimento e a discussão de seus resultados, respectivamente nas seções 4 e 5. Encerra-se este estudo com a seção 6 destinada às considerações finais e possibilidades futuras.

2. Análise de recursos computacionais

Em se tratando de um projeto experimental, com foco na análise exploratória tanto dos dados, quanto dos recursos computacionais disponíveis para os diversos processos envolvidos na DCBD, optou-se pelo uso de quatro ferramentas distintas: *Scikit-Learn*, *Pandas*, *Orange* e *Weka*. Essas ferramentas foram escolhidas, sobretudo, por sua popularidade na área de Ciência de Dados, a disponibilização gratuita, bem como as possibilidades de parametrização e de integração de seus recursos.

O *Scikit-Learn*[2] é uma biblioteca de distribuição aberta lançada em 2007 e voltada para aprendizado de máquina com uso da linguagem Python. É considerado um dos melhores recursos para a área de mineração de dados, uma vez que além de gratuito, possui alta qualidade em termos de código, uma comunidade ativa na manutenção de funcionalidades e documentação vasta e detalhada. Apresenta uma série de algoritmos implementados, tais como *Support Vector Machine (SVM), Random Forests* e *K-means*, além de auxiliar em tarefas de pré-processamento [Pedregosa *et al.*, 2011].

Outra biblioteca que faz uso da linguagem de programação Python é o *Pandas*[3]. Surgiu em 2008 também com código aberto e tem como especialidade a manipulação de tabelas numéricas e séries temporais. Trata-se de uma ótima ferramenta para análise de grandes volumes de dados e de fácil integração com outras biblioteca, como o *Scikit-Learn*, por exemplo [McKinney, 2018].

[1] Acesso disponível em: https://stackoverflow.com/
[2] Biblioteca disponível para download em: http://scikit-learn.org/stable/index.html
[3] Biblioteca disponível para download em: https://pandas.pydata.org/

Já a ferramenta *Orange*[4] apresenta uma interface mais intuitiva e facilitada para quando não se tem conhecimentos em linguagens de programação, tornando os processos de DCBD mais acessíveis. Apresenta funcionalidades para análise de dados, visualização, modelagem, avaliação e aplicação de algoritmos de aprendizagem supervisionada e não supervisionada. Faz uso de bibliotecas em Python, porém através de um painel configurável proporciona a realização das tarefas sem a implementação de código fonte. Elaborado pela Universidade de Ljubljana na Eslovênia com sua primeira versão em 1997, ganhou maior visibilidade a partir de 2013 [Demsar *et al.,* 2013].

Por fim o *Weka (Waikato Environment for Knowledge Analysis)*[5] é amplamente conhecido, sobretudo na área acadêmica. Teve a primeira versão em 1993 desenvolvida pela Universidade de Waikato da Nova Zelândia, sendo adquirido por uma empresa em 2006. Apresenta diversos algoritmos para as tarefas de classificação, associação e clusterização, além de recursos para pré-processamento e grande possibilidade de customização com a instalação de pacotes com recursos adicionais [Hall *et al.,* 2009].

3. Materiais e métodos

Para este estudo optou-se por adotar o processo CRISP-DM (*Cross-Industry Standard Process for Data Mining*) proposto por Chapman *et al.* (2000) e amplamente difundido. O ciclo de vida consiste em seis fases que se interligam e não possuem necessariamente uma ordenação rígida, ou seja, trata-se de um processo cíclico onde o resultado de cada fase é que determina qual fase deverá ser realizada posteriormente. São elas:

a) entendimento do negócio: avaliação das situações e recursos, determinação dos objetivos de MD em relação ao domínio e criação de um plano de projeto;
b) entendimento dos dados: coleta dos dados iniciais, realização da descrição dos dados, bem como sua exploração e análise de qualidade;
c) preparação dos dados: processos de seleção, etapas de pré-processamento, com limpeza, construção, integração e formatação dos dados;
d) modelagem: seleção de técnicas de modelagem com a geração de testes para construção de um modelo e posterior revisão;
e) avaliação: avaliação dos resultados e revisão dos procedimentos;
f) disponibilização: distribuição dos resultados, monitoramento, manutenção e produção de relatório final e documentação da experiência.

3.1 Entendimento do negócio

O *Stack Overflow* é um ambiente colaborativo de perguntas e respostas. Lançado em 2008, é voltado para área de tecnologia da informação e amplamente utilizado na comunidade de estudantes e profissionais para esclarecimento de dúvidas e discussão acerca de tecnologias. Há um processo de auto-regulação onde a comunidade pode avaliar as respostas realizadas por seus pares num sistema de votação e pontuações por interações realizadas [Joorabchi *et al.*, 2016]. Atualmente possui mais de 9 milhões de usuários registrados e passa dos 16 milhões de questões[6].

Há muitos anos este ambiente é fonte de informação por parte de estudantes da área de tecnologia da informação, chamando atenção de pesquisadores no sentido de

[4] Ferramenta disponível para download em: https://orange.biolab.si/
[5] Ferramenta disponível para download em: https://www.cs.waikato.ac.nz/ml/weka/
[6] Dados de Setembro de 2018. Atualização disponível em: https://stackexchange.com/sites?view=list#users

analisar os diversos fenômenos associados a seu uso, tais como os processos colaborativos, as práticas gamificadas da plataforma e a construção e validação de conteúdo em formato *crowdsourcing.*

Assim, entende-se o *Stack Overflow* como espaço para verificação do perfil dos estudantes da área de Tecnologia da Informação que fazem uso da plataforma e responderam ao questionário anual, através da execução de tarefas de sumarização e clusterização para descoberta de conhecimento associada a este contexto em uma abordagem heurística, ou seja, que leva em consideração o processo de descoberta, o trabalho experimental e o raciocínio lógico.

3.2 Interpretação do conjunto de dados

A obtenção do *dataset* se deu através do portal de Ciência de Dados *Kaggle*[7]. Trata-se de um ambiente colaborativo online que realiza competições relacionadas às áreas de estatística e mineração de dados para produção de modelos a partir dos dados cedidos pela plataforma. Adquirida pela Google em 2017, essa plataforma além de um espaço para disponibilização de conjunto de dados, configura-se em um interessante ambiente de aprendizado e compartilhamento de conhecimento nestas áreas.

Através de uma pesquisa nos *datasets* em âmbito educacional disponíveis na plataforma, chegou-se ao *"Stack Overflow Developer Survey"* que está distribuído sob a licença ODbL. Os dados são relativos a pesquisa realizada no início do ano de 2018 e estão disponibilizados em dois arquivos em formato CSV *(comma separated values):*

a) *survey_results_schema.csv:* descrição detalhada dos 129 atributos do conjunto de dados que correspondem a todas as perguntas realizadas no questionário;
b) *survey_results_public.csv:* dados em diversos tipos relativos às respostas ao questionário com total de 98.855 registros e tamanho de 195 megabytes.

4. Experimento

Após o entendimento do contexto e do *dataset*, passou-se a fase de experimentações iniciando pela preparação dos dados com atividades de pré-processamento e posteriormente com os processos de modelagem dos dados e aplicação de tarefas.

4.1 Pré-processamento dos dados

Em se tratando de um ambiente com ampla popularidade, o questionário global resulta em um *dataset* com grande quantidade de registros. O *dataset* de 2018 conta com 98.855 registros, quantidade grande de registros para manipulação e que poderia resultar em análises muito genéricas. Sendo assim, buscou-se realizar procedimentos de filtros nos dados iniciando pela seleção de registros provenientes apenas de estudantes chegando ao número de 34.502. Ao realizar o filtro por respostas vindas apenas do Brasil, apresentaram-se 2.505 registros. Ao aplicar o filtro conjunto de respostas de estudantes e que moram no Brasil o número final de registros ficou em 885 (Tabela 1).

Tabela 1. Filtros aplicados nas instâncias

	Total	Estudantes	Brasileiros	Estudantes Brasileiros
Instâncias	98.855	34.502	2.505	885

[7] Acesso disponível em: https://www.kaggle.com/datasets

Esse processo foi realizando de maneira exploratória, verificando-se diretamente o *dataset* em formato CSV no software Microsoft Excel, realizando filtros através da ferramenta. Ressalta-se que o atributo *Student* comportou no *dataset* os valores *"Yes, part-time", "Yes, full-time", "No"* e *"NA"*. Os registros que constavam *"No"* e *"NA"* foram retirados. Os dados resultantes foram transpostos para outro arquivo em formato CSV gerando o conjunto de registros a ser analisado neste estudo. No que diz respeito aos atributos, originalmente constavam 129 colunas de diversos tipos de dados. Em se tratando de um questionário, em sua maioria existiam dados nominais, entretanto alguns dados eram apresentados em escala de razão (salário, por exemplo) e valores ordinais (formação, graus de satisfação e etc).

O arquivo relativo ao *dataset schemma* continha todos os atributos e sua descrição. Nessa grande quantidade de atributos, optou-se por classificar manualmente em categorias a fim de selecionar adequadamente quais atributos seriam relevantes ao processo de mineração dos dados tendo em vista os objetivos (Tabela 2).

Tabela 2. Categorização de atributos

SIGLA	CATEGORIA	DESCRIÇÃO	QUESTÕES
PES	PESSOAL	Dados pessoais (idade, gênero, orientação sexual e hábitos pessoais)	20
EDU	EDUCACIONAL	Dados educacionais (formação, área, impressões a respeito da área educacional)	8
PRO	PROFISSIONAL	Dados profissionais (satisfação com carreira, tipo de trabalho, tempo de experiência)	11
EMP	EMPREGO	Percepção quanto a busca por empregos, benefícios, ambiente de trabalho.	21
COM	COMUNICAÇÃO	Dados de preferências de ferramentas comunicacionais.	12
ECO	ECONOMICO	Dados econômicos do respondente, como faixa salarial, por exemplo.	5
TEC	TECNICO	Conhecimentos técnicos, linguagens de programação e banco de dados utilizados.	15
ETI	ETICA	Dados de percepção éticas do respondente com algumas perguntas hipotéticas.	4
DEM	DEMOGRÁFICO	Questões relativas a dados de localização geográfica do respondente.	1
OUT	OUTROS	Outras questões como a percepção sobre publicidade online, os avanços tecnológicos, a plataforma *Stack Overflow* e potencial para novas ferramentas e serviços.	32

As categorias que mais interessam a este estudo são as que possuem relação com aspectos pessoais, educacionais e profissionais. Assim, levando-se em consideração a relevância das questões em cada categoria, chegou-se ao conjunto de atributos selecionados para análise apresentado na Tabela 3.

Tabela 3. Conjunto de atributos selecionados para análise

CAT	ATRIBUTO	DESCRIÇÃO	DADOS
PES	Hobby	*Do you code as a hobby?*	Binário
EDC	Student	*Are you currently enrolled in a formal, degree-granting college or university program?*	Binário
EDU	FormalEducation	*Which of the following best describes the highest level of formal education that you've completed?*	9 opções
EDU	UndergradMajor	*Which of the following best describes your main field of study (aka 'major')*	12 opções
PES	YearsCoding	*Including any education, for how many years have you been coding?*	11 opções
PRO	YearsCodingProf	*For how many years have you coded professionally (as a part of your work)?*	11 opções
PRO	HopeFiveYears	*Which of the following best describes what you hope to be doing in five years?*	7 opções
PRO	JobSearchStatus	*Which of the following best describes your current job-seeking status?*	3 opções
EDU	TimeAfterBootcamp	*How long did it take you to get a full-time job as a developer after graduating?*	7 opções
PES	EducationParents	**What is the highest level of education received by either of your parents?*	9 opções
PES	Age	**What is your age?*	11 opções
PES	Dependents	**Do you have any children or other dependents that you care for?*	Binário

*Questão de resposta não obrigatória.

Não havia duplicação de dados no conjunto analisado, entretanto alguns ajustes precisaram ser realizados quanto a codificação de caracteres. Usando a função *PreProcessor* da ferramenta Orange foram encontradas 885 instâncias e 12 atributos, sendo 27.3% de *missing values* (Figura 1). Considerou-se esse valor elevado e portanto

foram analisadas formas de ajustar essa questão. Ao ajustar a função *Impute missing values* aplicando a opção *Remove* restaram apenas 39 instâncias, portanto escolheu-se a opção *Average/Most Frequent* que substitui os *missing values* pela média ou resultados mais frequentes daquele atributo.

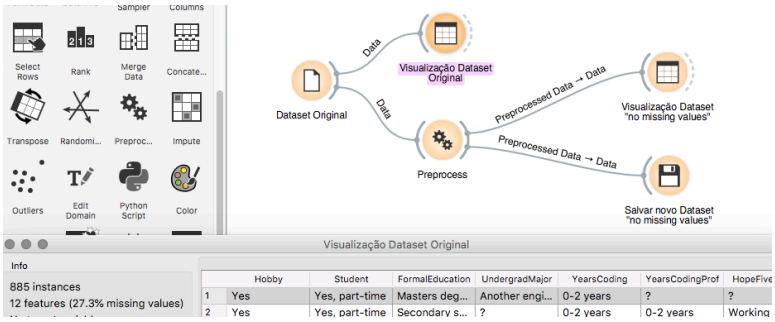

Figura 1. Aplicação de tarefa de pré-processamento usando Orange

Alguns atributos que possibilitavam múltiplas respostas foram isolados para avaliação posterior por demonstrarem relevância para o contexto (Tabela 4). Para estes atributos realizou-se um tratamento diferenciado, separando-os em registros únicos para serem contabilizados nas estatísticas.

Tabela 4. Conjunto de atributos que apresentavam múltiplas respostas

CAT	ATRIBUTO	TIPO DE DADOS
EDU	SelfTaughtTypes	*You indicated that you had taught yourself a programming technology without taking a course. What resources did you use to do that? If you've done it more than once, please think about the most recent time you've done so.*
EDU	HackathonReasons	*You indicated previously that you had participated in an online coding competition or hackathon. Which of the following best describe your reasons for doing so?*
EDU	EducationTypes	*Which of the following types of non-degree education have you used or participated in?*
PES	Gender	**Which of the following do you currently identify as?*
PES	SexualOrientation	**Which of the following do you currently identify as?*
PES	RaceEthnicity	**Which of the following do you identify as?*

*Questão de resposta não obrigatória.

4.2 Modelagem dos dados

Pela natureza dos dados optou-se pela realização de uma análise exploratória, uma vez que não se tinha conhecimento aprofundado sobre o domínio dos dados, bem como sobre o *dataset*. Segundo Amaral (2016) a análise exploratória é indicada para aprofundar o conhecimento dos dados antes da aplicação de técnicas explícitas ou implícitas. Tendo em vista a natureza dos dados que compõem o *dataset* analisado, optou-se por executar atividades de sumarização e visualização *(distillation of data for human judgement)*, bem como tarefas de agrupamentos *(clustering)*.

A destilação de dados para julgamento humano *(distillation of data for human judgement)* segundo Romero e Ventura (2013) apresenta como objetivo a representação inteligível usando sumarização, visualização e interfaces interativas para destacar informações úteis e dar suporte a intepretação dos dados e tomadas de decisão.

Os agrupamentos *(clustering)* são tarefas de MD capazes de identificar grupos de instâncias que apresentam similaridade em algum aspecto. Tipicamente algum tipo de distância é mensurada para decisão de características das instâncias. Dá suporte a possibilidade de execução posterior de tarefas de classificação, uma vez que quando conjuntos de instâncias são determinados, novas instâncias podem ser classificadas de acordo com o grupo mais próximo [Romero e Ventura, 2013].

Essas são tarefas de aprendizagem de máquina não supervisionadas, uma vez que não existe uma classe, ou seja, algo específico para prever ou descrever. Existem diversos algoritmos associados a esta tarefa, tais como os particionais, onde dividem-se as instâncias em grupos, os hierárquicos onde há formação de grupos e subgrupos e os difusos nos quais são atribuídos pesos as instâncias [Amaral, 2016].

5. Discussão dos Resultados

Será apresentada a seguir a avaliação dos dados através de aplicações de sumarização e clusterização, bem como a discussão sobre os resultados e possíveis interpretações.

5.1 Sumarização dos atributos com respostas únicas

Usando as bibliotecas Pandas fez-se uma série de análises exploratórias nos dados a fim de conhece-los. O arquivo em formato CSV com os tratamentos de pré-processamento já realizados foi carregado para análise e inserido em um *dataframe*, ou seja, uma estrutura de dados onde foram aplicados os diversos processos. Para cada um dos atributos foi realizada a análise de frequência de respostas e gerado um gráfico para melhor visualizar as informações. Esta atividade permitiu analisar de modo geral o perfil dos respondentes.

Pela análise através dos recursos de visualização de dados pode-se verificar a predominância de respostas por maior frequência, caracterizando de modo geral o perfil dos conjunto de respondentes selecionados. A partir da consolidação desses dados, criou-se um gráfico para melhor entendimento como apresenta a Figura 2.

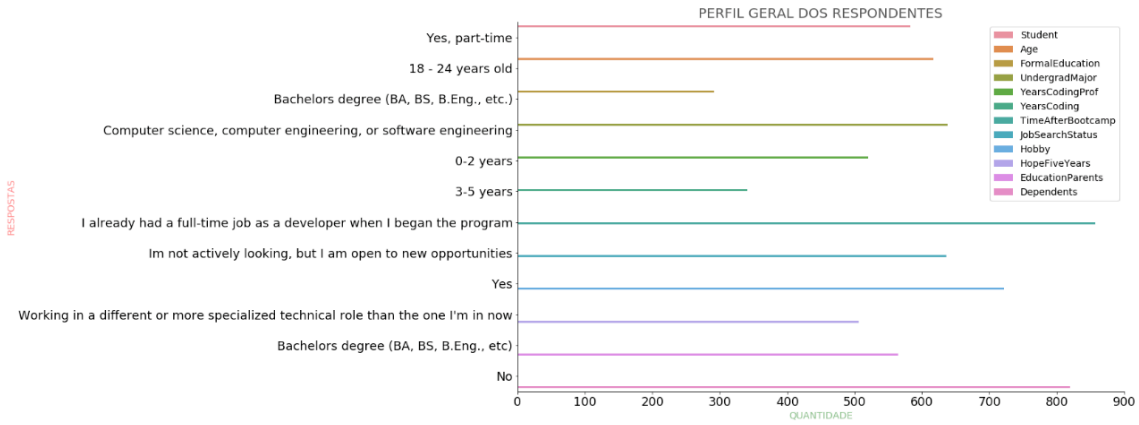

Figura 2. Visão geral do conjunto de dados conforme frequência de respostas

A pergunta do atributo *"Student"* era relativa a vinculação do respondente a um programa formal de estudos de qualquer natureza. Como apresentado no pré-processamento realizou-se o filtro selecionando apenas os registros que marcaram afirmativamente essa questão. Entretanto, os que afirmaram estar vinculados podiam especificar se a dedicação era exclusiva ou parcial. O que percebe-se é que a maior parte dos respondentes estuda apenas parte do tempo, o que se mostra muito comum na área de tecnologia na informação, sobretudo no Brasil.

A maior faixa etária ficou entre 18 e 24 anos demonstrando a predominância de pessoas jovens. Ainda, é possível perceber que a maior parte dos respondentes possui o nível de graduação na área de Ciência da Computação, Engenharia da Computação ou Engenharia de Software. Grande parte apresenta pouco tempo de experiência como

desenvolvedores em âmbito profissional (menos de 2 anos) e entre 3 a 5 anos de experiências em desenvolvimento envolvendo autodidatismo e educação formal.

A questão relativa ao atributo *"TimeAfterBootcamp"* dizia respeito ao tempo que a pessoa demorou para conseguir um emprego como desenvolvedor depois de sua formação. Pode-se verificar que quase a totalidade das respostas relativas a questão dão conta de *"eu já tinha um trabalho como desenvolvedor quando comecei o programa de estudos"*. Em se tratando de uma área não regulamentada, não existe a exigência de formações específicas para ingresso em atividades como desenvolvedor. Entretanto, percebe-se que cada vez mais há a preocupação em buscar qualificação devido a grande concorrência e a ampla gama de atividades na área de tecnologia da informação.

Pode-se verificar, que em se tratando de um público jovem, grande parte ainda programa como um *hobby*. Quanto a projeção em cinco anos, a maior parte indica querer trabalhar em uma área diferente ou cargo, o que corrobora com a questão da perspectiva de evolução dentro da área que possui um espectro de atividades de diferentes graus de complexidade, salários e projeção profissional.

Tendo em vista o contexto brasileiro de grandes desigualdades sociais é curioso perceber que boa parte dos respondentes possui pais com nível de graduação e quase a totalidade dos respondentes não possuem dependentes. Esses dois fatores podem auxiliar tanto no acesso a formação, quanto no processo de realização e finalização dos estudos em nível superior, por exemplo. O questionário não apresentava se os estudantes frequentam instituições públicas ou particulares e nem o porte das instituições, o que seria um indicativo interessante para o contexto brasileiro.

5.2 Sumarização dos atributos com respostas múltiplas

Alguns dos atributos selecionados permitiam mais de uma resposta, bem como a possibilidade de não responder. Nesses casos realizou-se a sumarização de todas as respostas concedidas. Em relação ao gênero, grande parte dos respondentes entende-se como do gênero masculino. A orientação sexual não se mostrou diversa, onde a maioria considera-se heterossexual. Quanto a etnia, a maior parte considera-se branco, havendo uma parte que considera-se hispânico/latino, com baixa ocorrência de pessoas negras.

Para este estudo três questões relativas a visão sobre atividades educacionais e forma de aprendizagem se mostraram relevantes para análise. A primeira delas diz respeito ao tipos não formais de educação que são usados ou dos quais participam. As duas respostas mais apresentadas foram o autodidatismo de uma nova linguagem, *framework* e ferramenta sem realizar um curso formal e a realização de um curso online em programação ou desenvolvimento de software. Ao encontro dessa questão, foi realizada uma pergunta relativa às formas de autoaprendizagem e grande parte sinalizou o uso da plataforma de perguntas e respostas no *Stack Overflow*, a documentação oficial de tecnologias, bem como o acesso e participação em comunidades online. Ainda, no sentido de verificar as razões para se participar de atividades tais como Hackathons, grande parte indicou a intenção de melhoria nas habilidades técnicas e de programação, o fato de ser uma atividade agradável e também possibilitar a melhoria de conhecimento em determinada linguagem de programação, *framework* ou tecnologia.

5.3 Teste com agrupamentos

A fim de verificar dentro do conjunto de dados padrões que pudessem gerar agrupamentos optou-se pela aplicação do algoritmo *K-means* na tarefa de Clusterização. Realizou-se a experiência com duas ferramentas: a execução através da ferramenta *Weka* e a construção usando Python e a biblioteca *Scikit-Learn*.

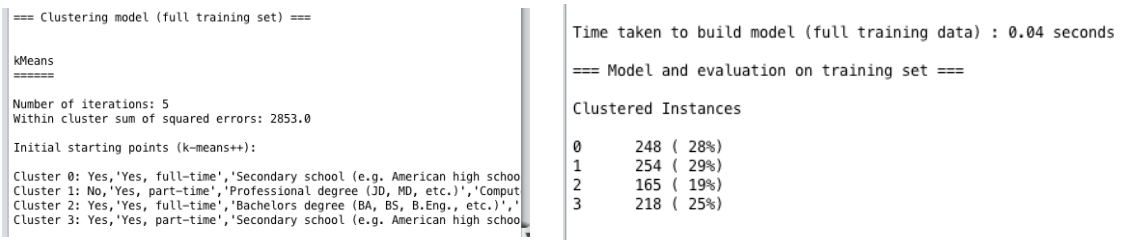

Figura 3. Resultado da aplicação do algoritmo k-means usando Weka

Com o conjunto de dados pré-processado e analisado, fez-se uso da ferramenta *Weka* para a tarefa de agrupamento *(clustering)*. Usou-se o algoritmo *SimpleKMeans* realizando alguns testes com a modificação do número de clusters e a função de distância entre Euclidiana e *Manhattan*. Ainda, com a opção de inicialização em k-means++ ou randômica. Testou-se também com algoritmos *EM* e *FarthestFirst* e diversos resultados foram apresentados, mas acredita-se que o mais consistente é o apresentado na figura 3 com as seguintes configurações: a) algoritmo: *SimpleKMeans*; b) número de clusters: 4; c) função de distância: distância Euclidiana; d) número máximo de iterações: 300; e) método de inicialização: *k-means++*.

Percebe-se a ocorrência de quatro *clusters* equilibrados. O *cluster 0* apresenta pessoas entre 18 e 24 anos que já tem conhecimento em programação em média 10 anos, ou seja, um público que teve acesso muito cedo aos conhecimentos em programação. Estão cursando Ciência da Computação e trabalham profissionalmente na área há menos de 2 anos. Não estão interessados em um novo emprego no momento, mas pretendem trabalhar em uma área mais especializada do que estão atualmente. Já o conjunto apresentado no *cluster 1* apresenta pessoas entre 25 e 34 anos com cerca de 10 anos de experiência em programação e que já possuem uma certificação profissional. Estão abertos a novas vagas, mas não estão procurando ativamente e futuramente pretendem trabalhar como gestores. O *cluster 2,* apresenta pessoas entre 18 e 24 anos que conhecem programação desde muito cedo, mas trabalham na área há menos de 5 anos e estão cursando pós-graduação em sistemas de informação e gestão de tecnologia e pretendem se fundar sua própria empresa. Por fim, o *cluster 3* indica pessoas entre 18 e 24 anos que tem menos de 2 anos de experiência na área de programação e estão cursando graduação na área de Computação. Estão abertos a novas vagas, mas não procurando ativamente e pretendem trabalhar como gestores.

6. Considerações finais

Este estudo permitiu entender DCBD através de um processo prático e analítico. Foram executadas todas as etapas em uma atividade exploratória, onde além de ampliar os conhecimentos relativos a temática, pode-se reconhecer as potencialidades de alguns ambientes computacionais disponíveis para estas práticas.

As ferramentas *Scikit-Learn* e *Pandas* se mostraram muito úteis na análise exploratória na fase de entendimento do negócio e dos dados. Já a ferramenta *Orange* se mostrou interessante nos procedimentos de pré-processamento, bem como na análise visual dos dados. Por fim, para o processo de mineração de dados usando agrupamentos, o *Weka* mostrou avançado nas possibilidades de parametrização e personalização.

Em relação aos dados percebeu-se uma tendência a homogeneidade, uma vez que houve grande frequência de respostas semelhantes dos respondentes, o que denota o perfil de acesso a este questionário. Verificou-se um público jovem, branco, masculino, heterossexual e com formação superior, que teve acesso ao conhecimento de programação desde cedo, entretanto com pouca experiência profissional na área e que considera com muita relevância formas alternativas de aprendizado em sua formação.

Tendo em vista o *dataset* selecionado com grande quantidade de atributos e dados em nível mundial, novos estudos experimentais podem ser realizados no sentido de comparar resultados entre diferentes países, analisar possíveis correlações entre países de um mesmo continente ou então com aspectos sociais semelhantes, bem como aprofundar análises de agrupamentos que possam fornecer processos posteriores de classificação e associação.

Referências

Amaral, F. (2016) "Introdução à Ciência de Dados: mineração de dados e Big Data". Rio de Janeiro: Alta Books. 320 p.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000) “CRISP-DM 1.0 (CRoss-Industry Standard Process for Data Mining): Step-by-step data mining guide”. SPSS.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., Zupan, B. (2013) "Orange: Data Mining Toolbox in Python". Journal of Machine Learning Research 14, 2349-2353.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996) “Advances in Knowledge Discovery and Data Mining”. AAAI Press, Menlo Park, CA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. (2009) "The WEKA Data Mining Software: An Update". ACM SIGKDD Explorations 11 (1).

Joorabchi, A., English, M., & Mahdi, A.E. (2016) “Text mining stack overflow: An insight into challenges and subject-related difficulties faced by computer science learners”. J. Enterprise Inf. Management, 29, 255-275.

McKinney, W. (2018). "Python para Análise de Dados: tratamento de dados com Pandas, Numpy e iPhyton". São Paulo: Novatec. 615 p.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D. (2011) "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research 12, 2825-2830.

Romero, C.; Ventura, S. (2013) “Data Mining in Education”. WIREs Data Mining Knowl Discov, 3: 12-27.