Mineração de dados para identificação de *clusters*: um estudo na área de tecnologia da informação.

Gilberto Alves Filho, Anderson Silva do Nascimento

Escola de Ciência e Tecnologia – Universidade do Grande Rio Professor José de Souza Herdy (Unigranrio)
Duque de Caxias – Rio de Janeiro - Brasil

gilberto.filho@unigranrio.br, anderson.nascimento@unigranrio.edu.br

Abstract. *Nowadays it is a great challenge for young people to get jobs in the area of Information Technology after completing higher education. This is because they do not know which way to go or which attributes to acquire during graduation. The data mining along with its several algorithms aims to help the process of discovery of knowledge. In this work, the use of data mining will be presented along with its aggregation algorithms, with the objective of providing an accurate knowledge of what are the best attributes for students in the area of Information Technology to develop and better prepare for the work market.*

Resumo. *Hoje em dia é um grande desafio para os jovens conseguirem emprego na área de tecnologia da informação após concluir o ensino superior. Isso se dá pelo fato deles não saberem qual caminho seguir ou quais atributos adquirirem durante a graduação. A mineração de dados junto aos seus diversos algoritmos tem como objetivo auxiliar o processo de descoberta de conhecimento. Neste trabalho será apresentado o uso da mineração de dados junto aos seus algoritmos de agregação, com o objetivo de prover um conhecimento preciso sobre quais são os melhores atributos para que os estudantes da área de tecnologia da informação possam desenvolvê-los e melhor se prepararem para o mercado de trabalho.*

1. Introdução

A mineração de dados, ou *data mining*, é um processo de extração de conhecimento em grandes bases de dados, tendo como foco analisar e explorar bases de dados buscando padrões, relacionamento entre os dados, e até realizar o agrupamento desses dados através de suas características, com foco em previsões para a tomada de decisões. De acordo com Fayyad et al. (1996), a mineração de dados é apenas uma das etapas de um processo conhecido como KDD (*Knowledge Discovery in databases*), também conhecido como descoberta de conhecimento em bases de dados. Fayyad et al. (1996) define o KDD como um método tradicional de transformar dados em conhecimento. Ainda segundo Fayyad et al. (1996), o KDD pode ser aplicada em diversas áreas importantes, por exemplo: marketing, finanças, detecção de fraude, fabricação, telecomunicações, etc. Dentro da área de mineração de dados possuímos atualmente diversos tipos de algoritmos que podem ser utilizados para resolver diversos tipos de problemas. São eles: algoritmos de associação, algoritmos de classificação, algoritmos de padrões sequenciais e algoritmos de agrupamento.

Costa (2015) afirma que entre janeiro e março de 2015, o total de desempregados com diploma de curso superior cresceu 21,25% em relação a 2014. E esse panorama tende e aumentar com o tempo.

Sendo assim, para esse trabalho será utilizado a mineração de dados com o algoritmo de agregação junto à ferramenta *Weka* de mineração para podermos identificar quais são os atributos que mais influenciam a empregabilidade na área de TI e assim resolver o seguinte problema: como prover um conhecimento preciso de quais são os melhores atributos para que os estudantes de TI possam desenvolvê-los e melhor se prepararem para o mercado de trabalho?

2. Weka

A ferramenta *Weka* foi desenvolvida na Universidade de Waikato (Nova Zelândia), sendo um software de código aberto (*open source*) onde o usuário pode alterar seu código fonte da forma que achar melhor. O Weka foi desenvolvido em Java e pode ser utilizado tanto pelo modo console quanto por sua GUI onde o usuário poderá interagir com os dados a serem analisados (ABERNETHY, 2010).

De acordo com Ferreira (2016) o Weka é uma excelente ferramenta de mineração para iniciantes, por possuir uma curva de aprendizado menor comparado ao R e com outras ferramentas. Sendo esse o motivo da utilização da ferramenta neste trabalho. Na Figura 1 é mostrado o ambiente *explorer* do *Weka*.

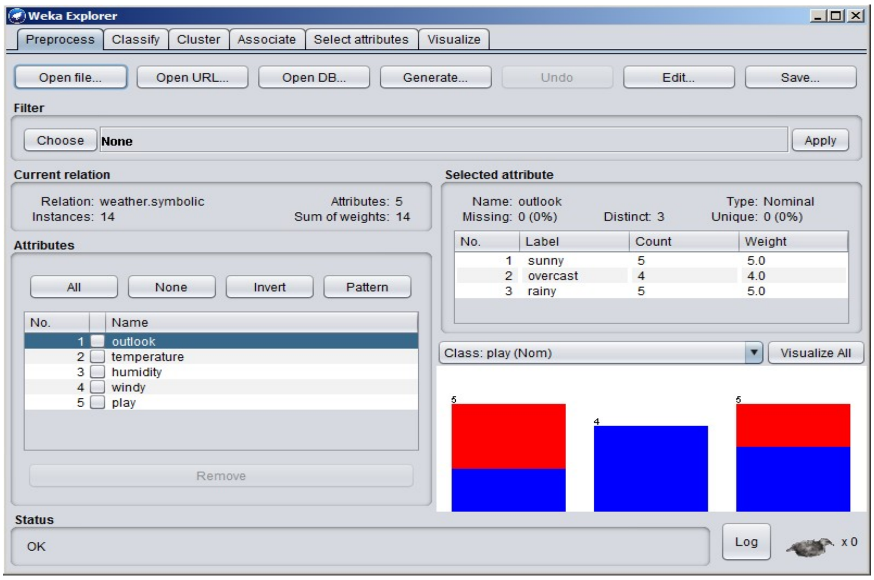

Figura 1. Ambiente *explorer* do *Weka*.

3. Perguntas contidas no questionário e formas de compartilhamento

O questionário criado para este trabalho foi produzido através do Google Formulários. O formulário foi do tipo aberto e fechado, ou seja, serão apresentadas perguntas abertas, na qual o respondente poderá responder de forma livre. E também serão apresentadas questões objetivas, na qual as respostas estão definidas junto às alternativas.

Para alcançarmos o maior número de pessoas possíveis, o questionário foi compartilhado em redes sociais como o facebook, e em redes sociais de negócios como o Linkedln. Após o questionário permanecer online por duas semanas, foram obtidas 408 respostas, um número bom considerando o pouco tempo em que ficou online, porém para a mineração pode ser considerado um número baixo.

As perguntas contidas no questionário irão servir tanto para a parte de mineração como também para podermos analisar o perfil dos respondentes. Abaixo serão mostradas as perguntas contidas no questionário.

- ✓ Sexo
- ✓ Idade
- ✓ Já concluiu a graduação?
- ✓ Ano em que concluiu a graduação?
- ✓ CR Final (Coeficiente de rendimento)
- ✓ Qual foi a modalidade de graduação realizada?
- ✓ Realizou estágio durante a graduação?
- ✓ Trabalhou na área de TI durante a graduação?
- ✓ Realizou alguma certificação na área de TI durante a graduação?
- ✓ Qual certificação foi feita?
- ✓ Em qual universidade foi realizada a graduação?
- ✓ Realizou algum curso de TI durante a graduação?
- ✓ Realizou ou participou de algum programa de iniciação cientifica?
- ✓ Conseguiu ingressar na área de TI após a conclusão do curso?
- ✓ Atualmente está empregado(a) na área de TI?

Para este trabalho também não foram utilizadas todas as perguntas contidas no formulário, algumas foram descartadas e foram utilizadas apenas para analisar o perfil dos respondentes. As outras colunas serão mantidas e farão parte do processo de mineração.

Colunas Descartadas: ID, Sexo, Idade, Ano em que concluiu a graduação, modalidade da graduação, qual certificação foi realizada, universidade em que se graduou, já concluiu a graduação.

Colunas Mantidas: CR Final, realizou estágio durante a graduação, trabalhou na área de TI durante a graduação, realizou alguma certificação na área de TI, realizou algum curso de TI durante a graduação, realizou ou participou de alguma iniciação científica na área de TI, atualmente está empregado na área de TI, conseguiu ingressar na área de TI após concluir a graduação.

4. Perfil dos respondentes

Através da base de dados adquirida foi possível ter conhecimento do perfil dos respondentes que colaboraram com o preenchimento do formulário. Abaixo será apresentado as tabelas com os resultados adquiridos dos respondentes.

Sexo

Masculino	Feminino
347 (85%)	**61 (15%)**

Tabela 1. Tabela com sexo dos respondentes.

Modalidade da Graduação

Presencial	Semipresencial	A Distância
378(92,6%)	**9(2,3%)**	**21(5,1%)**

Tabela 2. Tabela com a modalidade de graduação.

Atributos	Sim	Não
Concluíram a graduação?	307(76%)	100(24%)
CR Final >=7?	333(82,5%)	70(17,5%)
Realizou estágio durante a graduação?	252(61,8%)	156(38,2%)
Trabalhou durante a graduação?	296(72,5%)	112(27,5%)
Realizou alguma certificação na área de TI?	68(16,7%)	340(83,3%)
Realizou algum curso de TI durante a graduação?	275(69,1%)	124(30,9)
Realizou ou participou de alguma iniciação científica na área de TI?	72(16,9%)	333(83,1%)
Conseguiu ingressar na área de TI após a graduação?	283(69,9%)	123(30,1%)

Atualmente está empregado na área de TI?	273(68,1%)	132(31,9%)

Tabela 3. Tabela contendo os outros resultados dos respondentes.

5. Algoritmo de Agrupamento (*clustering*) e Aplicação do Algoritmo *K-Means*

De acordo com Honda (2017), a clusterização é o agrupamento de dados similares, uma classificação não-supervisionada. Esse algoritmo de clusterização classifica os dados em conjuntos que se assemelham de alguma forma, independente de classes predefinidas. Os conjuntos criados com esses dados são definidos como *clusters*.

Um dos algoritmos de agrupamento muito conhecido é o *K-Means*, mais conhecido como o algoritmo das K-Medias. De acordo com Santana (2017), a principal função desse algoritmo é encontrar a similaridade entre os dados analisados e agrupá-los de acordo com o número de *clusters* definidos pelo argumento K. O 'K' de *K-Means* é o número de *centroids* que serão criados. Ainda segundo Santana (2017), o *K-Means* chegará na sua fase final assim que a movimentação dos *centroids* se encerrarem e os clusters se tornarem estáticos, ou seja, sem que os dados alterem de *cluster*. Na Figura 2 podemos observar um exemplo de grupos criados e seus respectivos centroids marcados com X.

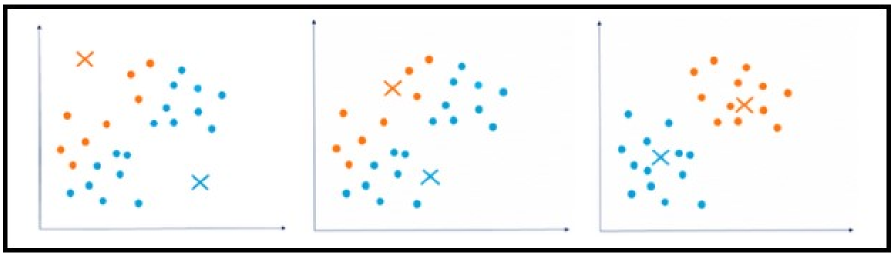

Figura 2. Exemplo de dados sendo agrupados em seus *centroids* onde K=2.

Primeiramente definimos a quantidade de *clusters* que serão criados no algoritmo *K-Means*, sendo K o número de centroides que serão criados. Essa é uma das tarefas mais difíceis, pois, dependendo da quantidade de *clusters* que forem criados pode acabar gerando grupos de dados com poucos valores, ou mais conhecidos como pontos fora da curva, que por serem poucos valores, não podem ser considerados como um padrão e logo são descartados.

Das 408 linhas obtidas, nem todas serão utilizadas, pois, entre as linhas obtidas têm muitos respondentes que ainda não concluíram a graduação, totalizando 100 linhas, nos restando apenas 308 linhas para realizar o estudo.

Após importar a base de dados para o *Weka* e aplicar o algoritmo foram criados os seis *clusters* como mostrado na Tabela 4.

Atributos	***Cluster 0* (80)**	***Cluster 1* (35)**	***Cluster 2* (21)**	***Cluster 3* (48)**	***Cluster 4* (20)**	***Cluster 5* (103)**
CR Final	>=7	<7	>=7	>=7	>=7	>=7
Realizou estágio durante a graduação?	Não	Sim	Não	Não	Sim	Sim
Trabalhou na área de TI durante a graduação?	Sim	Sim	Sim	Não	Não	Sim
Realizou alguma certificação durante a graduação?	Não	Não	Não	Não	Não	Não
Realizou algum curso de TI durante a graduação?	Não	Sim	Não	Sim	Sim	Sim
Realizou ou participou de alguma iniciação científica na área de TI?	Não	Não	Sim	Não	Não	Não
Atualmente está empregado na área de TI?	Sim	Sim	Sim	Não	Sim	Sim
Conseguiu ingressar na área de TI após concluir a graduação?	Sim	Sim	Sim	Não	Sim	Sim

Tabela 4. Resultado do agrupamento utilizando o algoritmo *K-Means*.

***Cluster* zero:** possui 80 instâncias (26% do total de instâncias). Nesse *cluster* podemos observar que os respondentes mesmo não tendo realizado algum estágio isso não influenciou a conseguirem ingressar na área de TI e atualmente estarem empregados. Talvez por terem trabalhado na área de TI durante a graduação, talvez eles possam ter ganhado experiência na área e com isso terem sido efetivados após o término da graduação.

***Cluster* um:** possui 35 instâncias (11% do total de instâncias), diferente do *cluster* zero, nesse os respondentes além de terem trabalhado na área de TI durante a graduação, também realizaram estágio e fizeram algum curso na área de TI. Mesmo os respondentes desse *cluster* possuírem CR <7, isso aparenta não ter influenciado no resultado final. Pode-se dizer que o esforço e dedicação deles em ter realizado curso e estágio possa ter influenciado no momento de ingressar na área de TI e de atualmente estarem empregados também.

***Cluster* dois:** possui 21 instâncias (7% do total de instâncias), esse *cluster* é muito parecido com o *cluster zero*, a única diferença é que esse é o único *cluster* onde os respondentes realizaram ou participaram de uma iniciação cientifica.

***Cluster* três:** possui 48 instâncias (16% do total de instâncias), nesse *cluster* os respondentes não realizaram estágio, não trabalharam na área de TI durante a graduação, não realizaram certificação, não realizaram ou participaram de algum programa de iniciação cientifica. Porém possuem CR Final>=7 e realizaram cursos na área de TI durante a graduação. Por fim acabaram não conseguindo ingressar na área TI após concluir a graduação. Analisando o *cluster* podemos concluir que possuir CR >=7 e realizar cursos na área de TI não garantem que o estudando irá conseguir trabalhar na área de TI após concluir a graduação.

***Cluster* quatro:** possui 20 instâncias (7% do total de instâncias), esse *cluster* é parecido com o *cluster* três, a diferença é que nesse os respondentes realizaram estágio. Pode-se notar que o estágio foi o diferencial nesse *cluster*, pelo fato de todos terem conseguido ingressar na área de TI e atualmente estarem empregados.

***Cluster* cinco:** possui 103 instâncias (34% do total de instâncias), entre os *clusters* anteriores é o com maior número de respondentes. Esse cluster aparente ser igual em relação ao cluster um. O que difere os dois é o CR, que nesse cluster é >=7. Nota-se que o CR não fez muita diferença entre o cluster um e cinco, pois, em ambos os respondentes conseguiram ingressar na área de TI e atualmente estão empregados.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou a mineração de dados com algoritmos de agrupamento (*cluster*), para a identificação de quais atributos mais influenciam a empregabilidade na área de TI. Foram criados seis *clusters* onde todos apresentaram resultados satisfatórios. Um fato interessante é a respeito do CR Final, que parece não ter tido muita influência nos resultados finais. Com esses resultados tivemos certeza de que o algoritmo de agregação tenha sido o melhor para realizar este trabalho. Como trabalhos futuros, foi pensado na ideia de aplicar esse mesmo estudo, porém, em uma base de dados maior (10.000 linhas), por exemplo. Isto tornaria o resultado ainda mais real, com maior nível de confiança. Outra ideia para trabalhos futuros seria a utilização de ferramentas mais modernas atualmente, como “SAS, R, etc.” Focando também na utilização de outros algoritmos de mineração, como o de classificação.

Referências

Abernethy, M. (2010). *Mineração de Dados com WEKA*. Fonte: IBM: https://www.ibm.com/developerworks/br/opensource/library/os-weka1/index.html

Costa, R. (2015). *Cresce o número de desempregados com diploma de curso superior no Brasil*. Fonte: Correio Braziliense: https://www.correiobraziliense.com.br/app/noticia/economia/2015/06/07/internas_economia,485744/cresce-o-numero-de-desempregados-com-diploma-de-curso-superior-no-bras.shtml

Fayyad, U., Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Al Magazine Volume 17 Number 3*, 18.

Ferreira, M. (2016). *RapidMiner x WEKA*. Fonte: Haiku Deck: https://www.haikudeck.com/rapidminer-x-weka-uncategorized-presentation-pFQetySeqZ#slide14

Honda, H. (2017). *Introdução básica à clusterização*. Fonte: LAMFO: https://lamfo-unb.github.io/2017/10/05/Introducao_basica_a_clusterizacao/

Santana, F. (2017). *Entenda o algoritmo k-means e saiba como aplicar essa técnica*. Fonte: Minerando Dados: https://minerandodados.com.br/entenda-o-algoritmo-k-means/