APLICAÇÃO DA MINERAÇÃO DE DADOS: ESTUDO DE CASO DAS CARACTÉRISTICAS DOS ALUNOS INGRESSANTES NA UNIVERSIDADE FEDERAL RURAL DO SEMIÁRIDO*

Lucas Borcard Cancela - UEMG/Unidade Carangola
Adriano Simioni Alvim – UEMG/Unidade Carangola
Flávio Aparecido de Almeida - UEMG/Unidade Carangola
Luciano Dias de Sousa - UEMG/Unidade Carangola
Patrícia Aparecida Romeiro Campos Cancela - UEMG/Unidade Carangola

RESUMO: Este trabalho procura abordar algumas análises realizadas em base de dados disponibilizados pela Universidade Federal Rural do Semiárido (UFERSA), referentes aos alunos ingressos nesta Universidade. A partir da utilização de técnicas de mineração de dados, informações relevantes foram encontradas através de processo de seleção da base de dados, limpeza e transformação dos dados, escolha dos atributos que melhor retornam informações entre outros, considerando indicadores socioeconômico dos candidatos, divididos entre as classes A, C, D e E, a dependência da escola de onde o ingressante é oriundo, se da rede pública ou da rede privada de ensino e raça dos alunos, considerando o estudo entre os anos de 2005 e 2016. Após a análise dos resultados foi possível chegar ao objetivo que era o de descobrir a origem das classes socioeconômicas, rede de ensino e a raça do maior número de ingressantes, de acordo com o ano e o período de ingresso. Utilizando métodos e o algoritmo SimpleKMeans, utilizando como ferramenta o software WEKA para minerar os dados coletados, percorrendo pelas fases de seleção, processamento, transformação, *data mining* e interpretação dos resultados, com intuito de gerar conhecimento, obteve-se informações relevantes relacionadas ao número de aprovações na UFERSA.

PALAVRAS-CHAVE: Mineração de Dados; Estudo de Caso; Ensino Superior; Processamento de Dados; WEKA.

1. Introdução

A UFERSA foi instalada em 29 de julho de 2005, a partir da Lei nº 11.155. Ela é a única Instituição Federal de ensino superior localizada no Semiárido e especializada no desenvolvimento e da ciência e tecnologia voltadas para o agronegócio e para o fortalecimento da agricultura familiar. Oferece atualmente dez cursos de graduação e cinco de pós-graduação (FILHO, 2017).

Por isso, é uma escola devidamente aparelhada para servir a uma região carente não apenas de chuvas. Mas, necessita também, do conhecimento científico e tecnológico que está sendo levado por uma escola e agora Universidade federal do semiárido.

A aplicação de técnicas de mineração de dados auxilia na tomada de decisões, podendo ser utilizada em situações diversas devido a sua variedade de algoritmos. Pensando em uma solução eficiente e gratuita que concentre vários algoritmos na mesma localidade, utilizou-se para este trabalho o software WEKA.

Sendo um software livre e multiplataforma, ou seja, que pode ser utilizado em diferentes sistemas operacionais que disponham em sua máquina o Java, assim como um software acessível também devido ao usuário conseguir manipular sem muitas dificuldades grandes bases de dados sem conhecimento muito profundo de computação, o WEKA mostrou-se eficiente para a execução das tarefas propostas.

Para encontrar informações relevantes diante de grande base de dados, foi escolhido o um algoritmo de agrupamento ou clustering. De acordo com Dias (2001), o agrupamento trata-

*XVI Encontro Virtual de Documentação em Software Livre e XIII Congresso Internacional de Linguagem e Tecnologia Online.

se de uma técnica utilizada com objetivo de particionar registros de determinada base de dados em subconjuntos ou cluster, fornecendo padrões de critérios e semelhança no tratamento dos registros.

Sendo assim, o intuito deste trabalho é apresentar a aplicação do algoritmo de agrupamento implementado pelo WEKA e baseado no K-Means, que foi o SimpleKMeans, em um resultado da análise dos alunos ingressos na Universidade Federal Rural do Semiárido do Rio Grande do Norte, para assim encontrar o cluster que apresenta dados referentes aos ingressantes baseado em informações socioeconômicas.

2. Metodologia

Para o desenvolver o presente trabalho, utilizou-se de pesquisas bibliográficas em artigos similares ao tema e, para a mineração de dados, utilizou-se a ferramenta Weka pensando em agrupamento de informações através de algoritmos de cluster, com o algoritmo SimpleKMeans do software Weka The University Waikato 3.6.12. Desenvolvido em linguagem de programação Java, um sistema de código aberto, o que permite que qualquer pessoa possa implementá-lo e criar melhorias. Esse sistema faz análises em base dados, estabelecendo relações e extraindo algum tipo de conhecimento até então não conhecido. Para a inserção de dados, é preciso obedecer a alguns padrões. A base de dados precisa estar nos formatos que são suportados pelo sistema. O formato padrão é o ARFF, mas pode ser utilizado outros, como o CSV, por exemplo.

A mineração de dados consiste em análises e padrões. Antes de iniciar o trabalho com o Weka, foi necessário criar padrões na base a ser analisada e posteriormente a converter para um formato aceitável. Segundo Steiner et al (2006), pode-se definir padrão como um tipo de acordo ou compromisso para criar boas práticas de definição e classificação de grupos de dados. Estes dados devem estar em conformidade à aceitação de valores e regras. A isso, chamamos de processo KDD, do inglês, Knowledge Discovery in Databases.

3. Processo KDD

Para encontrar coisas interessantes ou nunca vistas antes em uma grande base dados, primeiramente deve-se criar condições de poder extrair esse conhecimento. Essas condições são criadas pelo Processo KDD. O processo é caracterizado por cinco etapas básicas: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados (Data Mining); e interpretação e avaliação dos resultados (FAYYAD ET AL. al.,1996).

Para Fayyad Et Al (1996), o processo KDD é todo conhecimento útil extraído de uma base de dados, e classifica o Data Mining como uma etapa principal do KDD onde temos métodos para análise dos dados.

A base de dados para análise, será do banco de dados de alunos ingressos (indicador socioeconômico) na universidade federal rural do semiárido, onde constam os candidatos e as opções às vagas de cursos de capacitação profissional de todo o Estado.

Como a maioria das bases de dados provenientes aos processos de seleção, esta, possui inúmeras informações como: número de alunos ingressantes, ano de ingresso, período no qual o aluno ingressou, número de ingressantes vindos da rede pública e privada, raça, número de alunos pertencentes as classes A, C, D e E. Esses dados são de extrema importância, no entanto precisam sofrer algumas alterações.

3.1. Pré-processamento e limpeza dos dados

Nesta etapa foram realizadas limpezas no banco e uma organização nos números de

alunos ingressos com o intuito de melhorar e facilitar o processo de análise dos dados contidos no banco de dados estudado.

Durante a limpeza do banco de dados referente aos alunos ingressos na Universidade Federal Rural do Semiárido, foram realizadas algumas adaptações com o intuito de facilitar uma futura análise dos dados, tais adaptações ocorreram nas colunas de Alunos ingressos; alunos ingressos vindos de escolas públicas; alunos ingressos vindos de escolas privadas; alunos ingressos portadores de necessidades especiais; alunos ingressos provindos da classe A; alunos ingressos provindos da classe C; alunos ingressos provindos da classe D; e, alunos ingressos provindos da classe E.

Dentre as limpezas feitas no banco de dados para uma posterior análise dos dados ali contidos, podem-se destacar a retirada da coluna período assim como a retirada dos dados dos anos de 2003 e 2004, a retirada dos dados correspondentes a esses anos se deve ao número pequeno de ingressantes nesse período, logo não traziam relevância a mineração.

A coluna período foi retirada pois a variação que ocorria dentro dessa coluna, não trazia nenhuma relevância para a mineração e posterior análise dos dados.

Os demais números não foram divididos em intervalos ou porcentagens devido à pequena quantidade de dados contida no referido banco. Apesar de não ter havido essa divisão dos dados durante a limpeza, a análise pode ser feita perfeitamente com a ferramenta Cluster, obtendo os resultados desejados de forma satisfatória.

3.2. Mineração dos dados (WEKA)

Uma base de dados é um aglomerado de informações que podem ser lidas e interpretadas. Têm-se como exemplos planilhas eletrônicas, bancos de dados, listas e qualquer forma de se representar diversas informações de forma organizada.

A noção de encontrar padrões úteis em grandes volumes de dados pode ser conhecido por diversos nomes, tais como mineração de dados, extração de conhecimento, descoberta de informações, arqueologia de dados e processamento de padrões de dados. O termo mineração de dados é o mais usado por profissionais da computação, estatísticos e analistas de dados.

A fase de mineração de dados é uma fase do processo de Descoberta de Conhecimento em Banco de Dados (DCBD). Esta etapa é responsável pela aplicação dos algoritmos que são capazes de identificar e extrair padrões relevantes presente nos dados (HAN, 2001; WITTEN, 2000).

De acordo com Gordon & Gordon (2006), a etapa de mineração de dados permite, a partir dos dados, extrair informações imprevisíveis, antes não conhecidas e são potencialmente úteis.

A mineração de dados explora grandes quantidades de dados, com o objetivo de organizar informações, procurando padrões, tendências e associações, mutações, irregularidades e exceções. Este processo torna-se dificultosa a análise a um ser humano, desprovido de ferramentas apropriadas (GORDON & GORDON, 2006).

O processo DCBD é interdisciplinar, tanto em sua aplicação, quanto das suas fundamentações teóricas. O processo pode ser aplicado a qualquer problema de identificação de padrões em dados e contém fundamentação de diversas áreas como a banco de dados, inteligência artificial, estatística, probabilidade e visualização de dados.

WEKA é uma suíte de mineração de dados muito popular no meio acadêmico, desenvolvido utilizando a linguagem Java. Criada nas dependências da Universidade de Waikato, Nova Zelândia. Atualmente é mantida por uma comunidade de entusiastas por ser um software livre disponível sobre a licença GPL.

3.2.1. Uso do algoritmo SimpleKMeans

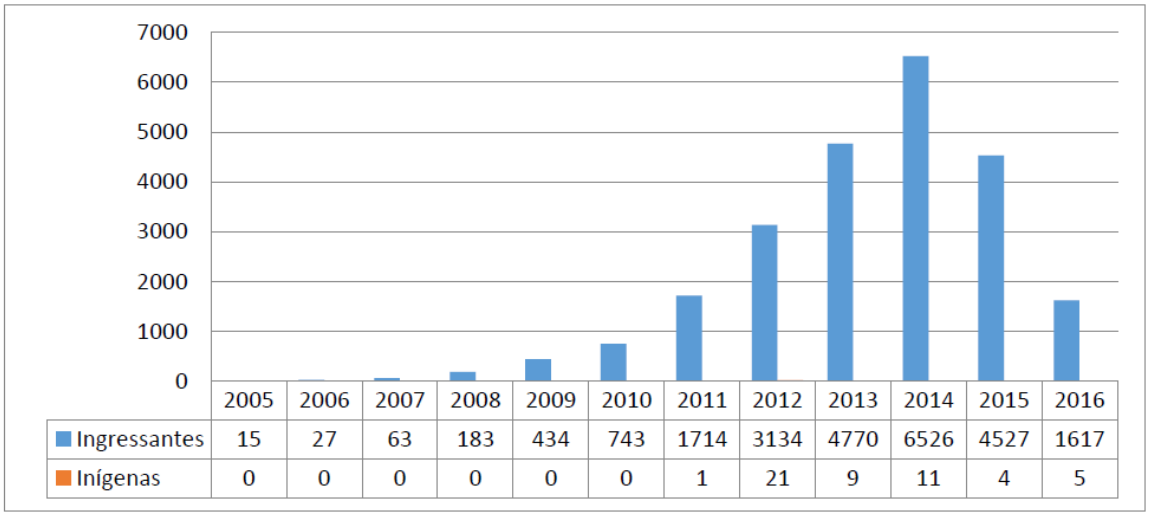
O SimpleKMeans é um algoritmo que cria grupos fazendo uso da média aritmética, cria-se assim a quantidade de grupos solicitada pelo usuário, através de uma fórmula, onde a média e a soma das observações são divididas pelo número delas (FERREIRA, 2005).

Este algoritmo apresenta sua eficiência através de um dado chamado de sum of squared errors (SSE), que é a soma dos quadrados dos erros. O quadrado de erros é uma medida para quantificar a diferença entre os valores da média aritmética, assim pode-se mostrar a precisão desta média. Por fazer o uso de média aritmética, pode-se encontrar valores que não foram testados. Isso pode ter vantagens ou desvantagens, sendo que isso depende dos dados que estão sendo trabalhados.

A ideia do algoritmo SimpleKMeans (também conhecido por K-Médias) é possibilitar uma classificação das informações em conformidade com os próprios dados, fundamentada em comparações e análises. Assim sendo, o algoritmo fornece uma classificação automática, ou seja, sem supervisão humana, sendo considerado por esta característica própria como um algoritmo de mineração de dados não supervisionado.

3.3. Análise de dados e resultados obtidos

Entre os dados obtidos durante a análise do banco de dados alguns resultados foram observados e selecionados para serem apresentados neste artigo. Dentre eles podemos citar: O número de ingressantes indígenas entre os anos de 2005 e 2016. Entre os anos de 2005 e 2010 não houveram ingressantes indígenas na universidade. A partir de 2011, alguns ingressantes indígenas começaram a surgir, tendo neste ano um total de 1714 ingressantes sendo 1 (um) ingressante indígena. No ano de 2012, houve um crescimento, chegando ao número de 21 (vinte e um) ingressantes indígenas, um total de 3134 (três mil cento e trinta e quatro) ingressantes no geral. No ano de 2013, o número de ingressantes indígenas diminuiu para 9 (nove) mesmo com o número de ingressantes geral tendo se elevado para 4770 (quatro mil setecentos e setenta). Já no ano de 2014, o número de ingressantes total foi de 6526 e o número de ingressantes indígenas foi de 11 (onze) ingressantes, em 2015 o número de ingressantes total foi de 4527 sendo que foram 4 (quatro) ingressantes indígenas e em 2016, o número total de ingressantes foi de 1617 sendo 5 ingressantes indígenas. Como pode ser visto no gráfico abaixo:



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Ingressantes	15	27	63	183	434	743	1714	3134	4770	6526	4527	1617
Inígenas	0	0	0	0	0	0	1	21	9	11	4	5

Gráfico 1 – Número total de ingressantes e de indígenas.

Fonte: os autores

Outros resultados observados foram do número de ingressantes portadores de necessidades especiais (PNE) ao longo dos anos de 2005 e 2016. No ano de 2005, o número total de ingressantes foi de 15 e o número de ingressantes PNE foi de 4 (quatro). Já no ano de 2006, o número de ingressantes total foi de 27 sendo que o número de ingressantes PNE diminuiu para 1 (um). Já em 2007 o número de ingressantes total foi de 63, e o número de ingressantes PNE se elevou para 5 (cinco). No ano de 2008, o número de ingressantes total foi de 183 e 29 ingressante PNE. Em 2009, o número de ingressantes total foi de 434 e o de ingressantes PNE foi de 18. Em 2010, o número de ingressantes total foi de 743 e o número de ingressantes PNE foi de 43. Em 2011, o número de ingressantes total foi de 1714 e o número de ingressantes PNE foi de 67. Já em 2012, o número de ingressantes total foi de 3134 e o número de ingressantes PNE teve um crescimento satisfatório, alcançando 114 ingressantes PNE. Em 2013, o número de ingressantes total foi de 4770 e o número de ingressantes PNE se elevou a 119, se elevando mais ainda no ano de 2014 para 202 ingressantes PNE de um número de ingressantes total de 6525. Em 2015, o número de ingressantes total foi de 4527 e o número de ingressantes PNE foi de 110 e em 2016 o número total de ingressantes foi de 1617 e ingressantes PNE foi de 52. Como pode ser observado no gráfico abaixo:

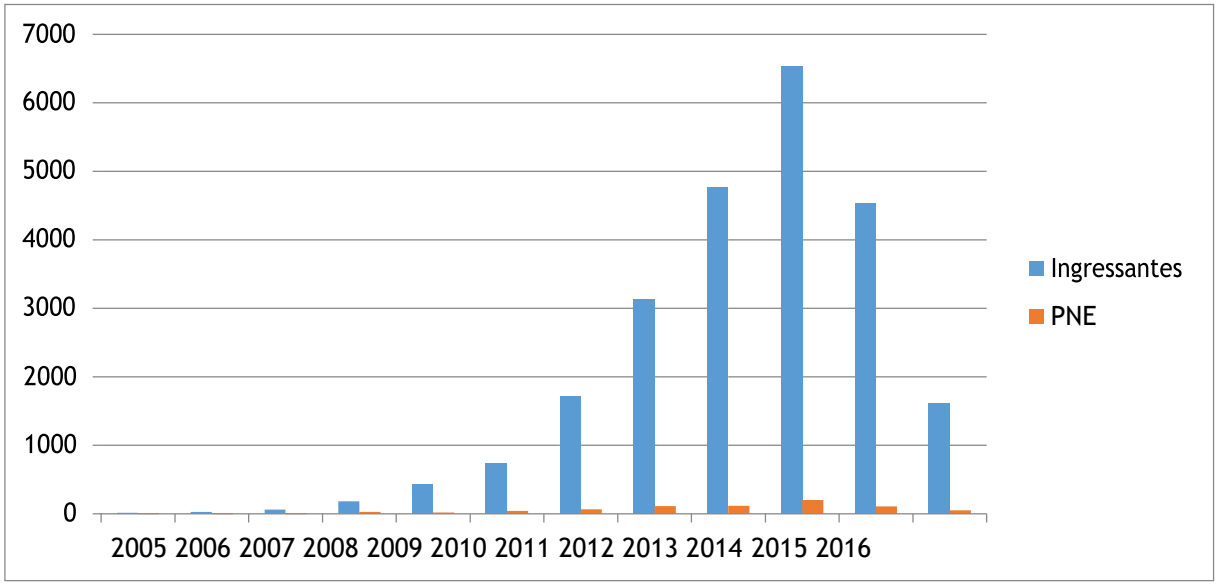


Gráfico 2 – Número total de ingressantes e de PNEs.

Fonte: os autores

Outros resultados observados foram o número de ingressantes provindos da rede privada que pertenciam a classe A e a classe C. Curiosamente, pode-se observar em todos os anos que apesar desses ingressantes virem de escolas privadas pouquíssimos vinham da classe A, tendo em todos os anos um número de ingressantes provindos da classe C de escolas privadas maior que o de ingressantes membros da classe A. Seguem os números: No ano de 2005, o número de ingressantes da rede privada foi de 10 com nenhum da classe A e seis da classe C. Em 2006, o número de ingressantes provindos da rede privada foi 15 sendo nenhum da classe A e 6 da classe C. No ano de 2007 o número de ingressantes vindo da rede privada foi de 34, sendo nenhum da classe A e 26 da classe C. Em 2008, o número de ingressantes da rede privada foi 108 sendo apenas 1 da classe A e 70 da classe C. Em 2009 o número de ingressantes vindos da rede privada foi de 201, sendo 2 da classe A e 111 da classe C. Em 2010 o número de ingressantes vindos da rede privada foi de 340 sendo estes 8 da classe A e 198 da classe C. Em 2011, o número de ingressantes da rede privada foi de 786 sendo estes 6 da classe A e 383 da classe C. Em 2012 o número de ingressantes vindos da rede privada foi de 1517, sendo estes 29 da classe A e 728 da

classe C. Em 2013, o número de ingressantes vindos da rede privada foi de 2101, sendo estes 62 da classe A e 983 da classe C. Em 2014 o número de ingressantes vindo da rede privada foi de 2037 sendo estes 32 da classe A e 1115 da classe C. Em 2015 o número de ingressantes da rede privada foi de 1501 sendo estes 24 da classe A e 943 da classe

C. Em 2016 o número de ingressantes provenientes da rede privada foi de 558 sendo estes 12 da classe a e 328 da classe C. De acordo com o gráfico abaixo:

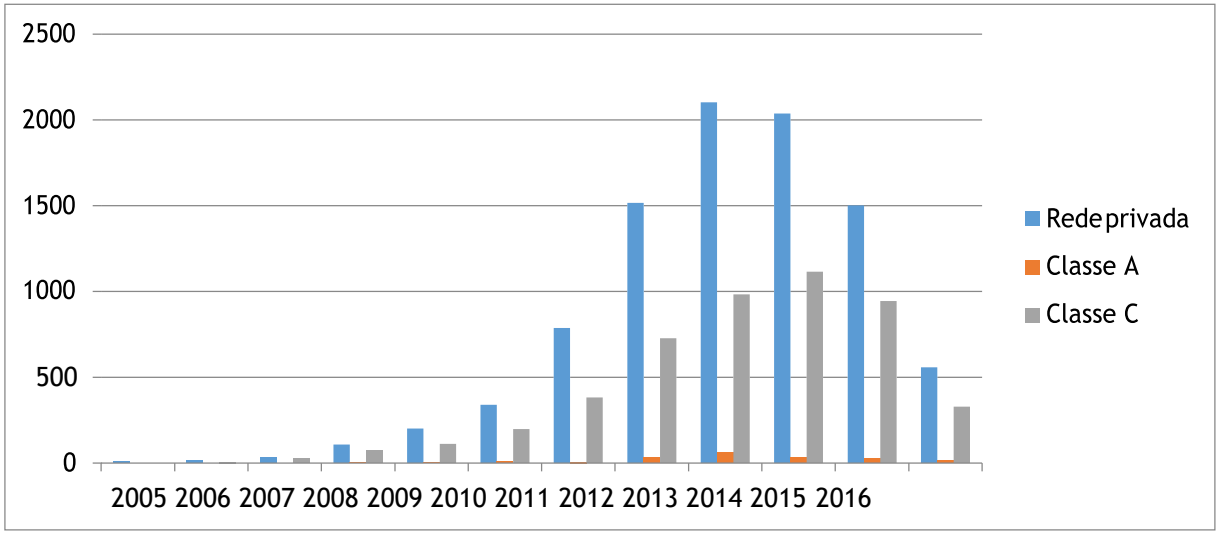


Gráfico 3 – número de ingressantes da rede privada que pertencem às classes A e C.
Fonte: os autores

Outros resultados observados foram o dos ingressantes provenientes da rede pública da raça negra e da raça branca. Um dado interessante observado em todos os anos foi que o número de ingressantes vindos da rede pública da raça branca foi maior que os da raça negra. Seguem os números: Em 2005, o número de ingressantes vindos da rede pública foi de 15 sendo estes pertencentes da raça branca 1 e da raça negra nenhum. Em 2006 o número de ingressantes vindos da rede pública foi de 27 sendo estes 12 da raça branca e 3 da raça negra. Em 2007 o número de ingressantes vindos da rede pública foi de 18 sendo estes 6 da raça branca e 5 da raça negra. Em 2008 o número de ingressantes vindos da rede pública foi de 63 sendo estes 21 da raça branca e 1 da raça negra. Em 2009 o número de ingressantes vindos da rede pública sendo estes 55 da raça branca e 51 da raça negra. Em 2010 o número de ingressantes vindos da rede pública foi de 354 sendo estes 162 da raça branca e 14 da raça negra. Em 2011 o número de ingressantes vindos da rede pública foi de 842 sendo estes 355 da raça branca e 97 da raça negra. Em 2012 o número de ingressantes vindos da rede pública foi de 1404 sendo estes 611 da raça branca e 95 da raça negra. Em 2013 o número de ingressantes vindos da rede pública foi de 2284 sendo estes 944 da raça branca e 181 da raça negra. Em 2014 o número de ingressantes vindos da rede pública foi de 3987 sendo estes 1392 da raça branca e 388 da raça negra. Em 2015 o número de ingressantes da rede pública foi de 2518 sendo estes 935 da raça branca e 261 da raça negra. Em 2016 o número de ingressantes vindos da rede pública foi de 1059 sendo estes 367 da raça branca e 122 da raça negra.

Outros resultados observados foram os do número de ingressantes provenientes da classe D e da classe E. Tendo estes resultados tido uma alternância ao longo dos anos sem manter uma prevalência exata do número superior de alguma classe em relação a outro. Seguem os números: No ano de 2016, de 1617 ingressantes, 407 pertenciam a classe D e 790 pertenciam a classe E. Em 2015, de 4257 ingressantes, 1103 pertenciam a classe D e 2007 pertenciam a classe E. Em 2014, de 6526 ingressantes, 1642 pertenciam a classe D e 3492 pertenciam a classe

E. Em 2013, de 4770 ingressantes, 1277 pertenciam a classe D e 2174 pertenciam a classe

E. Em 2012, de 3134 ingressantes, 921 pertenciam a classe D e 1262 pertenciam a classe E. Em 2011, de 1714 ingressantes, 597 pertenciam a classe D e 641 pertenciam a classe E. Em 2010, de 743 ingressantes, 249 pertenciam a classe D e 225 pertenciam a classe E. Em 2009, de 434 ingressantes, 115 pertenciam a classe D e 161 pertenciam a classe E. Em 2008, de 183 ingressantes, 79 pertenciam a classe D e 25 pertenciam a classe E. Em 2007, de 66 ingressantes, 14 pertenciam a classe D e 20 pertenciam a classe E. Em 2006, de 27 ingressantes, 9 pertenciam a classe D e 12 pertenciam a classe E. Em 2005, de 15 ingressantes, 7 pertenciam a classe D e 2 pertenciam a classe E. Como está exposto na tabela abaixo:

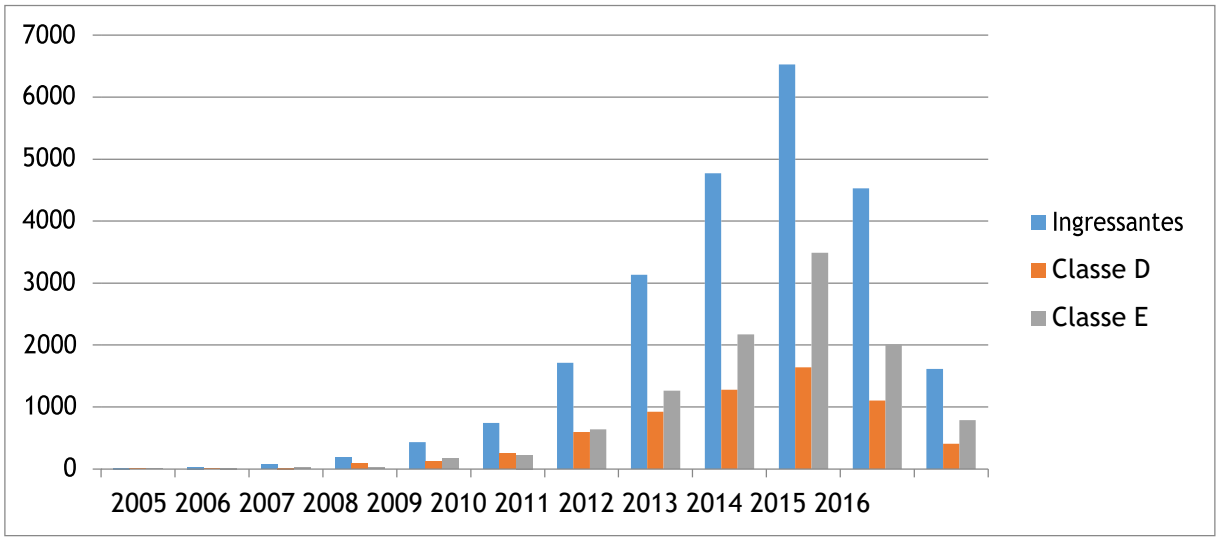


Gráfico 4 – número de ingressantes total e das classes A e C.

Fonte: os autores

4. Considerações finais

A partir desse estudo, pois possível mostrar como a técnica de Mineração de Dados, consegue e permite identificar que algumas raças, classes e redes tiveram maior aprovação e ingresso na Universidade. Pôde-se observar que certas classes, como D e E tiveram um número de ingressantes maior do que as classes A e C, assim como os alunos da rede pública tiveram um número maior em relação aos da rede privada.

Pela observação dos aspectos analisados através da utilização de ferramentas e do algoritmo SimpleKMeans, chegou-se a um resultado, onde foram extraídos dados da base da Universidade Federal Rural do semiárido em Mossoró, e passados por processos de mineração, desde a seleção, ao processo de KDD, que abrangem diversas áreas, transformações dos dados, data mining até o processo de interpretação e a geração dos resultados.

Desde a coleta até o pré-processamento, onde é feita a limpeza dos dados eliminando o que se tornaria irrelevante para chegar ao resultado maior que seriam os números extraídos a partir do estudo. Foram executados algoritmos, como o SimpleKMeans de modo que foram gerados números para comparações e, também, o software WEKA, gerando gráficos para melhor entendimento dos resultados.

Assim, é possível entender que a Universidade tem abrigado e formado grandes números de jovens e adultos, independentemente do nível socioeconômico, raça ou cor. A sua única missão é dar a esses estudantes conhecimentos científicos e tecnológicos.

5. Referências

DIAS.M.M. **Um modelo de formalização do processo de sistema de descoberta de**

Conhecimento em banco de dados. Tese de Doutorado. Florianópolis-SC: Universidade Federal de Santa Catarina, 2001.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery:** An Overview**.** In: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

FERREIRA.D.F. **Estatística Básica.** 1 Ed. Lavras-MG: Universidade Federal de Lavras, 2005.

FILHO, Garibalde Alves. Universidade Federal Rural do Semiárido. UFERSA, 2017. Disponível em: <https://ufersa.edu.br/>. Acesso em: 21 Jun. 2019.

GORDON, Steven R.; GORDON, Judith R. **Sistemas de Informação:** Uma abordagem gerencial. Trad. Oscar Kronmeyer Filho. 3. ed. Rio de Janeiro: Ltc, 2006. 377p.

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. Morgan Kaufmann, 2001.

STEINER Et al. **Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados.** Gestão & Produção. v.13, n.2, p. 325-337, mai-ago. 2006. Disponível em: <http://www.scielo.br/pdf/gp/v13n2/31177.pdf>. Acesso em: 02 jun. 2019.

UNIVERSITY OF WAIKATO. **Weka 3 – Machine Learning Software in Java.** Disponível no site da University of Waikato. <URL: http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 29 mai. 2019.