Utilização de agrupamento para identificar perfis de alunos com dificuldades na utilização de ambientes virtuais de aprendizagem

Lucielton Manoel[1], José Antonio[1]

[1]DC(Departamento de computação) – Universidade Federal Rural de Pernambuco (UFRPE)

lucielton.silva@ufrpe.br, antonio.barboza@ufrpe.br

1. Introdução

Com a popularização da internet nos últimos anos a educação avançou com a utilização de plataformas de ensino a distância(EaD) e os cursos online massivos abertos(MOOC). Porém, existem indivíduos que nunca estiveram em contato ou não são acostumados com o uso dessa tecnologia que podem acabar por ter seu processo de aprendizagem dificultado e essas dificuldades podem levar ao desengajamento, perda de motivação e até evasão do discente no curso.

Como as plataformas geram grandes quantidades de dados sobre o comportamento dos estudantes, estes dados em conjunto com técnicas de agrupamentos podem ajudar a encontrar grupos de alunos com dificuldades para usar a plataforma. Desta forma os tutores e professores podem auxiliar os discentes de forma mais precisa.

Este trabalho tem o objetivo de compreender padrões de comportamentos de usuários baseados em suas interações com o ambiente para encontrar melhores medidas de suporte ao uso da plataforma.

2. Motivação

Devido a grande quantidades de alunos nos curso EaDs, fica praticamente impossível para o professor acompanhar todos alunos atentamente. Para os discentes obterem um aproveitamento nas disciplinas é necessário que se familiarize com a plataforma. Assim, faz-se necessário um suporte que identifique discentes que estão tendo dificuldades, com isto em mente e com o avanço das técnicas de mineração de dados, sabemos que aplicando estas técnicas podemos levantar grupos de perfis de alunos com padrões comportamentais similares.

Conhecendo esses grupos de antemão e analisando padrões procurando por grupos com dificuldades podem ajudar gestores a identificar alunos com problemas na utilização do ambiente e tomar as medidas necessárias.

3. Metodologia

- **Descrição da base de Dados**

A base é composta por alunos de graduação que utilizaram a plataforma moodle. São

estudantes de vários cursos e disciplinas e às variáveis disponibilizadas estão divididas em duas partes:

1. Variáveis comportamentais.

Variável	Descrição
VAR01	Quantidade de diferentes locais (IP's) a partir dos quais a(o) aluna(o) acessou o ambiente.
VAR02	Quantidade de mensagens enviadas por aluna(o) as (os) Professoras(es) pelo ambiente.
VAR03	Quantidade de mensagens enviadas por aluna(o) ás(os) Tutor(es) pelo ambiente.
VAR04	Quantidade geral de mensagens enviadas pela(o) aluna(o) dentro do ambiente.
VAR05	Quantidade geral de mensagens recebidas pela(o) aluna(o) dentro do ambiente.
VAR06	Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo "tira-dúvidas".
VAR07	Quantidade de postagens no "Fórum tira dúvidas";
VAR08	Quant. de postagens de um(a) aluno(a) em fóruns que foram respondidas por outros(as) alunos(as).
VAR09	Quantidade de postagens de um(a) aluno(a) em fóruns que foram respondidas pelo(a) professor(a) ou tutor(a).
VAR10	Quantidade de colegas diferentes para quem o(a) aluno(a) enviou mensagens dentro do ambiente.
VAR12	Quantidade de visualizações da aba "Conteúdo" do curso, onde constam os arquivos com o conteúdo programático do curso
VAR13	Horário que mais realizou atividades;
VAR14	Turno do dia em que realizou mais atividades.
VAR16	Quantidade de atividades entregues por um(a) aluno(a) fora do prazo, por disciplina;
VAR17	Tempo médio entre a abertura da atividade e sua submissão;
VAR18	Quantidade de leituras feitas ao fórum (pageviews);
VAR20	Quantidade de respostas ao tópico principal (refazer opinião em fórum);

VAR21	Quantidade de pageviews ao quadro de notas;
VAR22	Quantidade de vezes que o aluno visualiza o (Checklist Atividades)
VAR23	Quantidade de visualizações de notas por atividade;
VAR24	Média semanal da quantidade de acessos de um(a) aluno(a) ao ambiente.
VAR25	Tempo médio entre a criação de um tópico no fórum temático e a primeira postagem do aluno;
VAR28	Quantidades de Time Out;
VAR31	Quantidade de acessos do(a) aluno(a) ao ambiente.
VAR31b	Quantidade de dias distintos que o aluno entrou na disciplina
VAR31c	Quantidade de dias distintos que o aluno entrou na plataforma
VAR32a	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Manhã).
VAR32b	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Tarde).
VAR32c	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Noite).
VAR32d	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Madrugada).
VAR33	Quantidade de atividades entregues por um(a) aluno(a) no prazo, por disciplina;
VAR34	Quantidade geral de postagens de um(a) aluno(a) em fóruns.
VAR35	Quantidade de respostas de um(a) professor(a) para as dúvidas de alunos(as) em fóruns.

2. Variáveis Relacionadas a Desempenho e Nota

Variável	**Descrição**
PROVA01	Nota da primeira prova presencial
PROVA01_2CHAMADA	Segunda chamada da primeira prova presencial
PRIMEIRA_PROVA	Nota da primeira prova ou da segunda chamada
PROVA02	Nota da segunda prova presencial
PROVA02_2CHAMAD	Segunda chamada da segunda prova presencial

A	
SEGUNDA_PROVA	Nota da segunda prova ou da segunda chamada
MEDIA_PROVAS	Media geral das duas provas presencial
FORUM01	Nota do primeiro fórum
FORUM02	Nota do segundo fórum
FORUM03	Nota do terceiro fórum
FORUM04	Nota do quarto fórum
MEDIA_FORUM	Média geral dos quatro fóruns
WEBQUEST01	Nota da primeira atividade (webquest)
WEBQUEST02	Nota da segunda atividade (webquest)
MEDIA_WEBQUEST	Média geral das duas atividades
DESEMPENHO	Desempenho final

Além dessas variáveis , temos às do tipo categóricas que representam o curso, período e disciplina dos alunos.

- **Abordagem ao Problema**

Aplicamos o algoritmo de clustering particional k-Means para formar os agrupamentos. Para começar, teremos que montar um vetor de características apropriado para tal tarefa. Como o objetivo é agrupar alunos com similaridades de uso da plataforma serão dadas preferência as variáveis mais relacionadas a plataforma e interação direta do estudante com o ambiente. Eliminado variáveis relacionada a notas e desempenho, temos um vetor de característica que dê ênfase a interação do usuário com a plataforma:

1. Quantidade de diferentes locais (IP's) a partir dos quais a(o) aluna(o) acessou o ambiente. (VAR1)
2. Quantidade de mensagens enviadas por aluna(o) ás (os) Tutor(es) pelo ambiente. (VAR2)
3. Quantidade de mensagens enviadas por aluna(o) ás (os) Tutor(es) pelo ambiente. (VAR3).
4. Quantidade geral de mensagens enviadas pela(o) aluna(o) dentro do ambiente. (VAR4)
5. Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo "tira-dúvidas". (VAR6)
6. Quant. de postagens de um(a) aluno(a) em fóruns que foram respondidas por outros(as) alunos(as). (VAR7)
7. Quantidade de postagens de um(a) aluno(a) em fóruns que foram respondidas pelo(a) professor(a) ou tutor(a). (VAR10)

8. Quantidade de visualizações da aba "Conteúdo" do curso, onde constam os arquivos com o conteúdo programático do curso. (VAR 12)
9. Quantidade de atividades entregues por um(a) aluno(a) fora do prazo, por disciplina. (VAR 16)
10. Tempo médio entre a abertura da atividade e sua submissão. (VAR 17)
11. Quantidade de leituras feitas ao fórum (pageviews). (VAR 18)
12. Quantidade de respostas ao tópico principal (refazer opinião em fórum). (VAR 20)
13. Quantidade de pageviews ao quadro de notas. (VAR 21)
14. Quantidade de vezes que o aluno visualiza o (Checklist Atividades). (VAR 22)
15. Quantidade de visualizações de notas por atividade. (VAR 23);
16. Média semanal da quantidade de acessos de um(a) aluno(a) ao ambiente. (VAR 24)
17. Tempo médio entre a criação de um tópico no fórum temático e a primeira postagem do aluno. (VAR 25)
18. Quantidades de Time Out. (VAR 28)
19. Quantidade de acessos do(a) aluno(a) ao ambiente. (VAR 31)
20. Quantidade de dias distintos que o aluno entrou na plataforma (VAR 31c)
21. Quantidade de atividades entregues por um(a) aluno(a) no prazo, por disciplina; (VAR 33)
22. Quantidade geral de postagens de um(a) aluno(a) em fóruns. (VAR 34)

Com esse vetor de características, é esperado que os grupos de alunos que tenham comportamentos similares. A partir da geração dos grupos, poderemos julgar quais grupos têm dificuldades com a plataforma.

- **O algoritmo particional K-Means**

O funcionamento do K-Means é bem simples.

1. Escolha arbitrariamente k centróides iniciais.
2. Repita até Convergir ou o número de iterações máximas seja atingido:

   2.1 Para cada vetor de característica calcule sua distância até os centróides

   2.2 O vetor pertencerá ao grupo em que tiver a menor distância euclidiana entre os centróides. A label do grupo será o respectivo centróide.

   2.3 Os novos centróides serão a soma de todos vetores pertencentes a um determinado grupo divididos pela quantidade de elementos desse grupo .

O K-Means tem algumas limitações. A primeira é a tendência a formar clusters esféricos. Um segunda é que ele depende de uma boa inicialização. Clusters com uma má inicialização pode gerar um agrupamento diferente do esperado. E por último, é preciso fornecer a quantidade de clusters a serem formados.Uma outra limitação é que ele não nos diz sobre a quantidade de de clusters a serem formados. Tem que descobrir. Para esta último problema, utilizaremos o índice de Dunn que analisa se elementos de um grupo estão próximos e grupos estão bem separados.

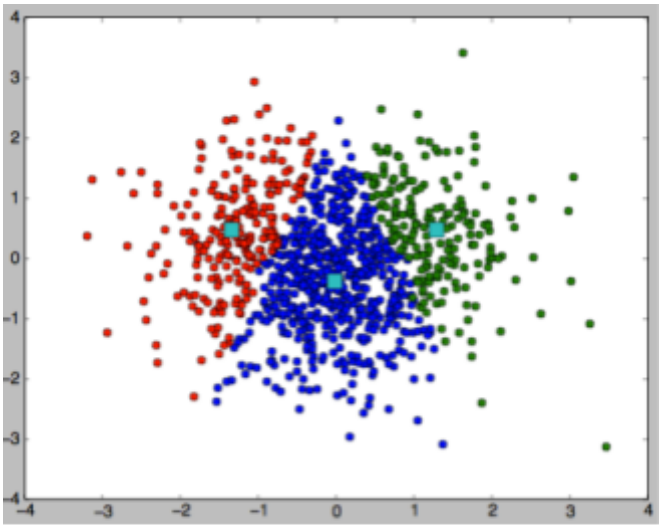


Fig.1. Tendência Esférica do Agrupamento.

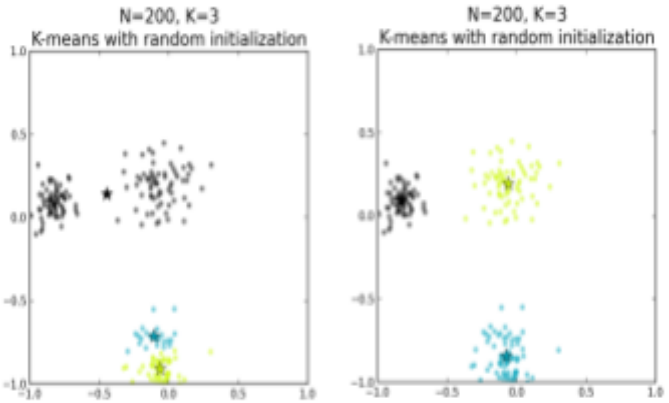


Fig. 2.Má inicialização pode não agrupar eficientemente.

- **Experimentos**

Para haver um bom agrupamento se faz necessário encontrar os parâmetros ideais. Para isso utilizaremos o índice de Dunn como critério de avaliação. Quanto maior o índice, melhor o agrupamento. Os parâmetros serão o número de cluster, a normalização dos dados e a inicialização.

Primeiramente iremos começar com os dados não-normalizados e procurar por uma boa inicialização. Fazendo o k variando de 2 até 10, para cada valor de k faremos dez inicializações diferente e pegamos a que tiver maior índice de Dunn.

Nas mesmas interações de inicialização iremos analisar qual foi a quantidade de grupos ideais para o conjunto entre os valores dois e dez, utilizando o índice de Dunn que continua sendo o critério de avaliação.

Depois, será realizada a mesma bateria de teste, porém com os dados normalizados.Com isso, teremos parâmetros mais eficientes para análise do dados. Assim, poderemos analisar os grupos em si formados pela soluções que encontramos. Para função de normalização iremos utilizar as seguintes fórmula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

O processo relatado acima será executado duas vezes

4. Resultados

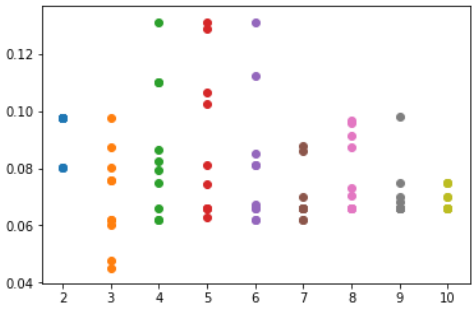
- **Dados da 1º bateria de teste**

Nestas tabelas temos os o número de grupos utilizados neste experimento. Para isso, foram dez iterações para cada número k de grupos.

- **Com a normalização:**

2	3	4	5	6	7	8
0.097602	0.061957	0.086707	0.102346	0.061957	0.066136	0.066013
0.097602	0.061957	0.079262	0.062976	0.130980	0.061957	0.066136
0.097602	0.060176	0.109931	0.129039	0.067388	0.086256	0.070428
0.080066	0.097602	0.061957	0.081344	0.061957	0.065816	0.065816
0.097602	0.075661	0.109931	0.065816	0.081344	0.070087	0.087467
0.080066	0.080066	0.130980	0.066136	0.066136	0.066136	0.095754
0.097602	0.044889	0.066136	0.065816	0.081344	0.087677	0.065816
0.097602	0.075661	0.075099	0.074614	0.085362	0.061957	0.096968
0.097602	0.087569	0.061957	0.130980	0.065816	0.065816	0.073261
0.080066	0.047535	0.082338	0.106719	0.112404	0.066136	0.091336

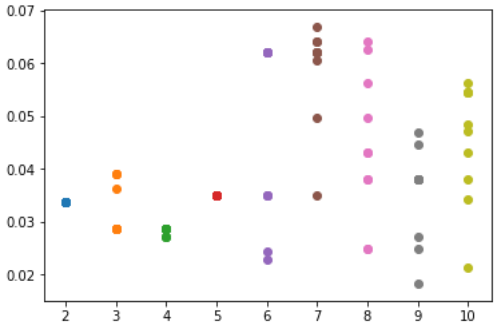
9	10
0.065816	0.065816
0.065816	0.070087
0.066136	0.074771
0.066136	0.065816
0.066136	0.070665
0.098065	0.069790
0.070214	0.066136
0.074948	0.066136
0.068057	0.065816
0.065816	0.051990



Nestes gráficos temos as dez iterações de cada número k distribuída de plano cartesiano.

- **Sem a normalização**

2	3	4	5	6	7	8	9	10
0.033603	0.038906	0.027216	0.035008	0.024301	0.064131	0.062541	0.038043	0.021258
0.033603	0.038906	0.027216	0.035008	0.022680	0.060567	0.024820	0.046863	0.048390
0.033603	0.036343	0.028728	0.035008	0.062107	0.062107	0.064131	0.027083	0.056121
0.033603	0.028578	0.028728	0.035008	0.062107	0.066892	0.024820	0.018235	0.034194
0.033603	0.028578	0.028728	0.035008	0.062107	0.062107	0.043113	0.038043	0.047044
0.033603	0.038906	0.028728	0.035008	0.035008	0.062107	0.049557	0.044518	0.038043
0.033603	0.028578	0.028728	0.035008	0.035008	0.035008	0.038043	0.024820	0.043113
0.033603	0.028578	0.028728	0.035008	0.062107	0.049557	0.043113	0.038043	0.054474
0.033603	0.028578	0.028728	0.035008	0.035008	0.064131	0.056121	0.038043	0.054474
0.033603	0.028578	0.027216	0.035008	0.062107	0.062107	0.038043	0.038043	0.054474



Nestes gráficos temos as dez iteração de cada número k distribuída de plano cartesiano.

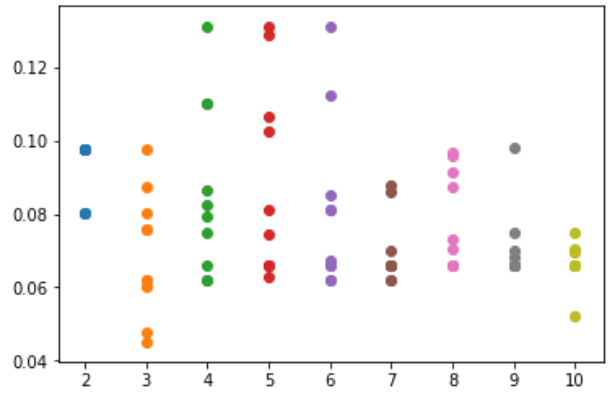
Através destes dados podemos inferir qual foi a melhor inicialização para este caso específico. Os dados normalizados têm melhores índices de Dunn em aproximadamente 80% em relação aos dados não normalizados. Vejamos um gráfico que exalta essa diferença:

Os melhores índices foram os 4,5,6.

- **Dados da 2º bateria de teste**

    - **Com a normalização**

2	3	4	5	6	7	8	9	10
0.097602	0.061957	0.086707	0.102346	0.061957	0.066136	0.066013	0.065816	0.065816
0.097602	0.061957	0.079262	0.062976	0.130980	0.061957	0.066136	0.065816	0.070087
0.097602	0.060176	0.109931	0.129039	0.067388	0.086256	0.070428	0.066136	0.074771
0.080066	0.097602	0.061957	0.081344	0.061957	0.065816	0.065816	0.066136	0.065816
0.097602	0.075661	0.109931	0.065816	0.081344	0.070087	0.087467	0.066136	0.070665
0.080066	0.080066	0.130980	0.066136	0.066136	0.066136	0.095754	0.098065	0.069790
0.097602	0.044889	0.066136	0.065816	0.081344	0.087677	0.065816	0.070214	0.066136
0.097602	0.075661	0.075099	0.074614	0.085362	0.061957	0.096968	0.074948	0.066136
0.097602	0.087569	0.061957	0.130980	0.065816	0.065816	0.073261	0.068057	0.065816
0.080066	0.047535	0.082338	0.106719	0.112404	0.066136	0.091336	0.065816	0.051990

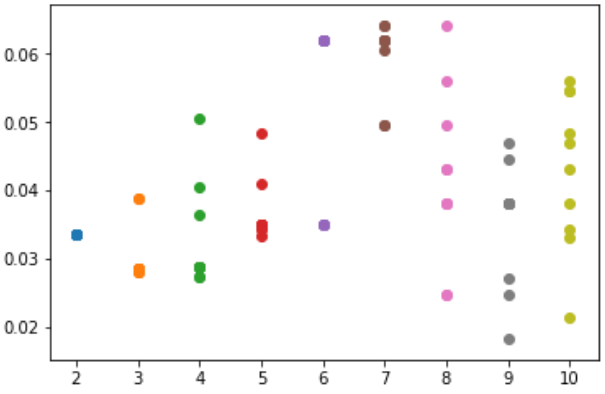


Nesse gráficos temos as dez iteração de cada número k distribuída de plano cartesiano.

- **Sem normalização**

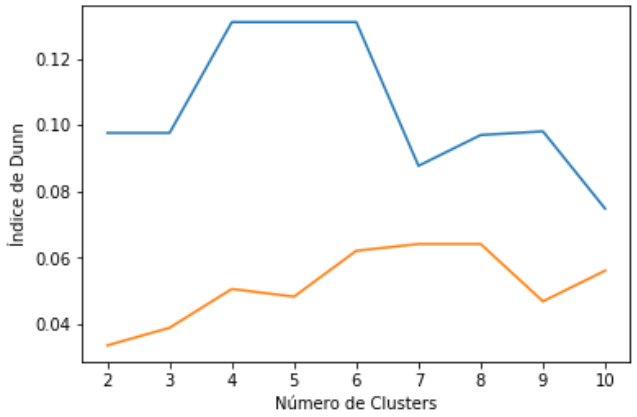
2	3	4	5	6	7	8
0.033603	0.028578	0.050592	0.035008	0.062107	0.049557	0.024820
0.033603	0.028578	0.028728	0.034296	0.035008	0.062107	0.064131
0.033603	0.027934	0.028728	0.035008	0.062107	0.062107	0.024820
0.033603	0.038906	0.027216	0.033259	0.035008	0.060567	0.043113
0.033603	0.028578	0.027216	0.035008	0.062107	0.049557	0.049557
0.033603	0.038906	0.036518	0.035008	0.062107	0.062107	0.038043
0.033603	0.028578	0.028728	0.035008	0.035008	0.062107	0.043113
0.033603	0.027934	0.040418	0.048284	0.062107	0.064131	0.056121
0.033603	0.027934	0.028728	0.035008	0.062107	0.062107	0.038043
0.033603	0.028578	0.028728	0.040868	0.035008	0.064131	0.038043

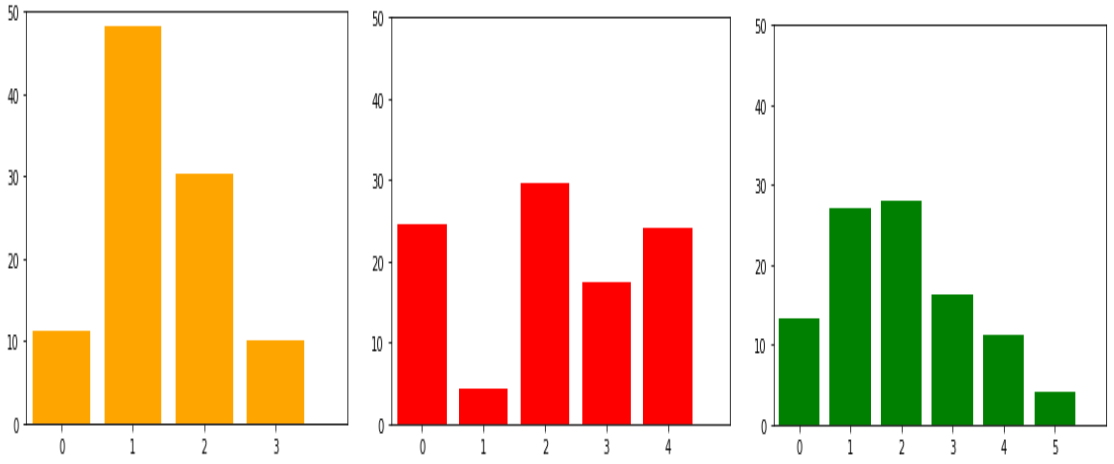
9	10
0.038043	0.021258
0.046863	0.048390
0.027083	0.056121
0.018235	0.034194
0.038043	0.047044
0.044518	0.038043
0.024820	0.043113
0.038043	0.054474
0.038043	0.054474
0.038043	0.033113



Nesse gráficos temos as dez iteração de cada número k distribuída de plano cartesiano.

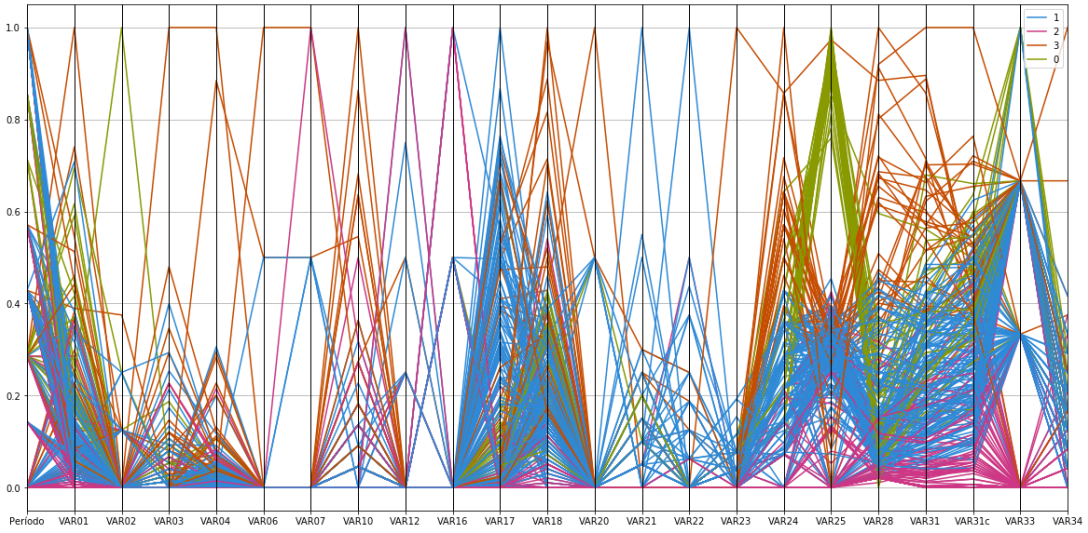
Para Segunda bateria temos o mesmo resultado: Os dados normalizados têm vantagem sobre os dados não-normalizados. Temos que os números de grupos(k) com maiores índices continuam com o 4,5,6 e normalização feita. E também os centros iniciais do grupo.





Porcentagem dos grupos gerados para k=4,5,6.

E por fim temos um gráfico de coordenadas paralelas para analisar os grupo gerados por k=4 com o centro que gerou o melhor índice de Dunn.



5. Conclusões

Através deste trabalho é possível entender como os parâmetros inseridos num algoritmo de agrupamento têm o potencial de mudar drasticamente o processo completo de agrupamento. Os grupos que estão normalizados tem uma índice de qualidade superior aos grupos que não possuem normalização e que a inicialização é um fator importante e decisivo na formação dos grupos gerados.