



UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

TÉRCIO JORGE DA SILVA

**EXTRAÇÃO DE CONHECIMENTO NOS DADOS DA UNIVERSIDADE
FEDERAL DO CEARÁ VIA MINERAÇÃO DE DADOS: DESCOBERTA
E ANÁLISE DOS PERFIS DOS ALUNOS.**

**QUIXADÁ
2014**

TÉRCIO JORGE DA SILVA

**EXTRAÇÃO DE CONHECIMENTO NOS DADOS DA UNIVERSIDADE
FEDERAL DO CEARÁ VIA MINERAÇÃO DE DADOS: DESCOBERTA
E ANÁLISE DOS PERFIS DOS ALUNOS.**

Trabalho de Conclusão de Curso submetido à Coordenação do
Curso Bacharelado em Sistemas de Informação da Universidade
Federal do Ceará como requisito parcial para obtenção do grau
de Bacharel.

Área de concentração: Computação

Orientadora Prof^a. Ticiane Linhares Coelho da Silva

**QUIXADÁ
2014**

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

S581e Silva, Tércio Jorge da
Extração de conhecimento nos dados da Universidade Federal do Ceará via mineração de dados: descoberta e análise dos perfis dos alunos / Tércio Jorge da Silva. – 2014.
65 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2014.
Orientação: Profa. Me. Ticiane Linhares Coelho da Silva
Área de concentração: Computação

1. Mineração de dados (Computação) 2. Clusters 3. Educação superior I. Título.

CDD 001.642

TÉRCIO JORGE DA SILVA

Extração de Conhecimento nos Dados da Universidade Federal do Ceará via Mineração de Dados: Descoberta e Análise dos Perfis dos Alunos.

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Sistemas de Informação da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel.

Área de concentração: computação

Aprovado em: _____ / Dezembro / 2014.

BANCA EXAMINADORA

Prof^a. MSc. Ticiane Linhares Coelho da Silva
(Orientadora)
Universidade Federal do Ceará-UFC

Prof. Dr. Críston Pereira de Souza
Universidade Federal do Ceará-UFC

Prof. Dr. Davi Romero de Vasconcelos
Universidade Federal do Ceará-UFC

Dedico este trabalho a minha Mãe, a minha Irmã, aos meus amigos, aos meus professores, e todos aqueles que contribuíram de forma direta ou indireta para este trabalho.

AGRADECIMENTOS

Primeiramente agradeço a Deus, por toda força e discernimento, pois acredito que sem a companhia Dele, eu não conseguiria superar todas as dificuldades encontradas ao longo dessa graduação.

A minha mãe Tamar Jorge, que tem sido até aqui meu pai e minha mãe, agradecer todo o esforço, toda paciência e toda dedicação que me possibilitaram ter a oportunidade de concluir essa graduação. Um agradecimento especial a minha irmã Thais Jorge, por todo o companheirismo, risadas e brincadeiras. A todos os meus familiares que torceram para que eu concluísse essa graduação.

Aos professores da Universidade Federal do Ceará campus Quixadá, em especial, a professora Msc. Ticiane Linhares, minha orientadora, por todo companheirismo, amizade e paciência ao longo desses 12 meses que dediquei à produção dessa monografia, ao professor Dr. Davi Romero, meu tutor no PET, que contribuiu grandemente para que esse trabalho pudesse ser feito, aos professores Drs. Críston Pereira e Tânia Pinheiro, que também contribuíram para este trabalho.

E por fim, agradeço aos colegas do grupo PET, aos colegas de graduação que sempre estiveram presentes, em especial, Ana Klyssia, Araújo Filho, Adeilson, Alex, Anderson, Claudio, Cinthia Maria, Gerlan, Guilherme, Junior Leonel, Junior Holanda, Ricardo Lopes, Salomão, Savio de Castro, Warnly, William, pelas muitas madrugadas de estudos, brincadeiras e companheirismo, ao longo desse curso.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.” (Arthur Schopenhauer)

RESUMO

Algumas mudanças têm ocorrido no sistema de avaliação para ingresso no ensino superior no Brasil. Talvez uma das mais impactantes mudanças tenha sido a implantação do Exame Nacional do Ensino Médio (Enem) como critério único de seleção em diversas universidades públicas, entre elas a Universidade Federal do Ceará (UFC). Nesse contexto, é importante conhecer o perfil dos alunos que ingressam na universidade todo ano. Esse conhecimento pode ser obtido utilizando os dados da própria universidade. Assim como as grandes empresas e organizações, a universidade também gera dados que sem o processo de análise têm seu significado limitado. Eles podem ser usados como matéria-prima na aplicação da descoberta do conhecimento. A descoberta do conhecimento consiste em extrair conhecimento útil de uma grande quantidade de dados. Dentre suas várias etapas, que serão citadas neste trabalho, a mineração de dados é uma das mais importantes e consiste em um conjunto de técnicas para analisar grandes quantidades de dados tentando encontrar um padrão consistente, sumarizar os dados, extrair conhecimento ou realizar previsões. É possível utilizar algumas dessas técnicas nos dados da universidade. Neste estudo a técnica escolhida é a clusterização, que se fundamenta no agrupamento de elementos similares em grupos, e elementos de grupos distintos são menos semelhantes, cada um desses grupos formando um perfil de aluno da Universidade Federal do Ceará, assim procurando responder questionamentos a respeito de medidas adotadas pela universidade ao longo dos anos, como a adoção ao SISU e a nova regra de frequência adota no segundo semestre de 2008.

Palavras chave: Mineração de dados, Clusterização, Descoberta do Conhecimento.

ABSTRACT

Some changes have occurred in the evaluation system for admission to higher education in Brazil. Perhaps one of the most striking changes has been the implementation of the National Examination of Secondary Education (ENEM), as the only way of admission in several public universities, including the Federal University of Ceará(UFC). In this context, it is important to know the profile of the students who enter the university every year. This knowledge can be obtained using the university's own data. As well as large companies and organizations, the university also generates data that without the analysis process have limited meaning. The data can be used as feedstock in the application of knowledge-discovery. Knowledge-discovery in extracting useful knowledge from a large amount of data. Among its various stages, which will be mentioned in this term paper, Data Mining is one of the most important and consists of a set of techniques to analyze large amounts of data trying to find a consistent pattern, summarize the data, extract knowledge or make predictions. It is possible use some of these techniques in the university data. In this study, the technique chosen is clustering, which is based on grouping similar items into groups, and elements of different groups are less similar, each group forming a profile of student of the Federal University of Ceará, seeking to respond to questions about measures taken by the university over the years , such as adoption to SISU and the new frequency rule adopted in the second half of 2008.

Keywords: Data Mining, Clustering, Knowledge-discovery

LISTA DE ILUSTRAÇÕES

Figura 1 – Etapas do processo de descoberta do conhecimento.....	21
Figura 2 – Definição da Distância Euclidiana entre os vetores.	24
Figura 3 – Parte de um DAG da Disciplina de FUP entre 2006.1 e 2007.1.....	24
Figura 4 – Definição da Similaridade de Jaccard.....	25
Figura 5 – Parte de um DAG entre perfis da disciplina de ED 2009.2 e AG 2010.1	25
Figura 6 – Dados recebidos e adicionados em nova tabela	26
Figura 7 – CSV gerado para a disciplina de FUP 2006.1	28
Figura 8 – Centróides encontrados na ferramenta RapidMiner para FUP 2006.1	30
Figura 9 – Perfil mais frequente em FUP de 2006 até 2012.....	32
Figura 10 – Perfis mais frequentes em ED de 2009 até 2012.....	33
Figura 11 – Perfis mais frequentes em AG de 2007 até 2012	35
Figura 12 – Perfis mais frequentes em CANA de 2007 até 2012.....	36
Figura 13 – DAG entre FUP 2006.1 e CANA 2007.2.....	46
Figura 14 – DAG entre FUP 2007.1 e CANA 2008.2.....	47
Figura 15 – DAG entre FUP 2008.1 e CANA 2009.2.....	48
Figura 16 – DAG entre FUP 2009.1 e CANA 2010.2.....	49
Figura 17 – DAG entre FUP 2010.1 e CANA 2011.2.....	50
Figura 18 – DAG entre FUP 2011.1 e CANA 2012.2.....	51
Figura 19 – DAG entre FUP 2012.1 e ED 2012.2.....	52
Figura 20 – DAG entre FUP 2009.1 e FUP 2012.1.....	53
Figura 21 – DAG entre ED 2009.2 e ED 2012.2.....	54
Figura 22 – DAG entre AG 2010.1 e AG 2012.1.....	55
Figura 23 – DAG entre CANA 2010.2 e CANA 2012.2.....	56
Figura 24 – DAG entre FUP 2007.1 e FUP 2010.1.....	57
Figura 25 – DAG entre ED 2006.2 e ED 2010.2.....	58
Figura 26 – DAG entre AG 2007.1 e AG 2010.1.....	59
Figura 27 – DAG entre CANA 2007.2 e CANA 2010.2.....	60
Figura 28 – DAG entre FUP 2006.1 e FUP 2009.1.....	61
Figura 29 – DAG entre FUP 2009.1 e FUP 2012.1.....	62
Figura 30 – DAG entre ED 2006.2 e ED 2009.2.....	63
Figura 31 – DAG entre ED 2009.2 e ED 2012.2.....	64
Figura 32 – DAG entre AG 2007.1 e AG 2009.1.....	65
Figura 33 – DAG entre AG 2009.1 e AG 2012.1.....	66
Figura 34 – DAG entre CANA 2007.2 e CANA 2009.2.....	67
Figura 35 – DAG entre CANA 2009.2 e CANA 2012.2.....	68

SUMÁRIO

1 INTRODUÇÃO.....	15
2 TRABALHOS RELACIONADOS	18
3 FUNDAMENTAÇÃO TEÓRICA	20
3.1 Descoberta do Conhecimento	20
3.2 Mineração de Dados	21
3.3 Clusterização.....	23
3.4 Distância Euclidiana	24
3.5 Similaridade de Jaccard	25
4 PROCEDIMENTOS.....	26
4.1 Recebimento, limpeza e transformação dos dados	26
4.2 Definição de quais análises/questões deveriam ser feitas sobre os dados	28
4.3 Aplicação da técnica de clusterização com a ferramenta RapidMiner	29
4.4 Aplicação da Similaridade de Jaccard e da Distância Euclidiana.....	30
4.5 Criação de DAGs para análise dos resultados	30
5 RESULTADOS	32
5.1 Perfis de alunos encontrados por disciplina.....	32
5.1.1 Fundamentos de Programação.....	32
5.1.2 Estrutura de Dados.....	33
5.1.3 Algoritmos em Grafos	34
5.1.4 Construção e Análise de Algoritmos	35
5.2 Impacto da adesão ao SISU nos perfis dos alunos.....	37
5.2.1 Impacto da adesão ao SISU na disciplina de Fundamentos de Programação	37
5.2.2 Impacto da adesão ao SISU na disciplina de Estrutura de Dados	38
5.2.3 Impacto da adesão ao SISU na disciplina de Algoritmos em Grafos	39
5.2.4 Impacto da adesão ao SISU na disciplina de Construção e Análise de Algoritmos.....	40
5.3 Impacto da nova regra de frequência de 2008.2 nos perfis dos alunos.....	40
5.3.1 Impacto da nova regra de frequência nos perfis dos alunos de Fundamentos de Programação	41
5.3.2 Impacto da nova regra de frequência nos perfis dos alunos de Estrutura de Dados	41
5.3.3 Impacto da nova regra de frequência nos perfis dos alunos de Algoritmos em Grafos	41
5.3.4 Impacto da nova regra de frequência nos perfis dos alunos de Construção e Análise de Algoritmos	42
6 CONSIDERAÇÕES FINAIS	43
REFERÊNCIAS	45
APÊNDICES	46
APÊNDICE A – DAGs relacionados às trilhas de disciplinas de 2006 até 2012	46

APÊNDICE B – DAGs relacionados às análises do impacto do SISU nos perfis de alunos entre 2009 até 2012.....	53
APÊNDICE C – DAGs relacionados às análises do impacto da nova regra de frequência de 2008.2 nos perfis de alunos entre 2006.2 até 2010.2.....	57
APÊNDICE D – DAGs relacionados às descobertas de perfis de alunos por disciplina entre 2006 até 2012.....	61

1 INTRODUÇÃO

Algumas mudanças têm ocorrido no sistema de avaliação para ingresso no ensino superior no Brasil. Talvez uma das mais impactantes mudanças tenha sido a implantação do Exame Nacional do Ensino Médio (Enem) como critério único de seleção em diversas universidades públicas, entre elas a Universidade Federal do Ceará (UFC). Nesse contexto, é importante conhecer o perfil dos alunos que ingressam na universidade todo ano. Esse conhecimento pode ser obtido utilizando os dados da própria universidade. Assim como as grandes empresas e organizações, a universidade também gera dados que sem o processo de análise têm seu significado limitado. Eles podem ser usados como matéria-prima na aplicação da descoberta do conhecimento.

A descoberta do conhecimento consiste em extrair conhecimento útil de uma grande quantidade de dados. Dentre suas várias etapas, que serão citadas na Seção 4.1 deste trabalho, a mineração de dados é uma das mais importantes e consiste em um conjunto de técnicas para analisar grandes quantidades de dados tentando encontrar um padrão consistente, sumarizar os dados, extrair conhecimento ou realizar previsões. É possível utilizar algumas dessas técnicas nos dados da universidade. Neste estudo a técnica escolhida é a clusterização, que se fundamenta no agrupamento de elementos similares em grupos, e elementos de grupos distintos são menos semelhantes.

No contexto deste trabalho, aplicando clusterização aos dados da universidade, encontrou-se vários grupos de alunos, e cada um desses grupos com suas características próprias formou um perfil de aluno. Exemplificando, existe um grupo com as seguintes características: boas notas, boa frequência e sem reprovações. Esse é o grupo formado pelos melhores alunos de uma determinada disciplina. Os dados da universidade são compostos por: identificação do aluno, nota do aluno por disciplina, quantidade de faltas do aluno por disciplina, média final do aluno na disciplina, semestre em que cursou a disciplina, ano, situação do aluno (aprovado, reprovado, trancado, abandono, aproveitamento externo, aproveitamento interno, etc), entre outros atributos. Desses dados foi retirada uma amostra com os alunos das disciplinas de Fundamentos de Programação, Estrutura de Dados, Algoritmos em Grafos e Construção e Análise de Algoritmo, do curso de Computação da Universidade Federal do Ceará em Fortaleza, entre os anos de 2006 até 2012, que são os dados mais familiares ao autor deste trabalho, por ser aluno da área e conhecedor das dificuldades de cada uma das disciplinas escolhidas, o que auxilia na hora de entender os

padrões encontrados. Além disso, a escolha das disciplinas não foi ao acaso, são disciplinas em que existe uma relação de dependência entre elas, para que o aluno possa cursar uma disciplina é necessário que tenha obtido êxito em todas as disciplinas que são pré-requisitos, por exemplo, para que o aluno esteja apto a cursar a disciplina de Estrutura de Dados é necessário ele tenha cursado a disciplina de Fundamentos de Programação e tenha sido aprovado.

Neste trabalho foram analisados os perfis de alunos encontrados por disciplina e por semestre, a fim de saber como os perfis foram formados seguindo uma trilha de dependência de disciplina, a exemplificar, quantos alunos que cursaram a disciplina de Fundamentos de Programação de um determinado Perfil, formam outro Perfil na disciplina de Estrutura de Dados no semestre seguinte ao que cursaram a disciplina de Fundamentos de Programação. Isso foi possível aplicando a técnica utilizada por Coelho da Silva et al (2014), no trabalho intitulado *Discovering Frequent Mobility Patterns on Moving Object Data*, onde é usada uma medida de similaridade, Jaccard, é aplicada para determinar as semelhanças entre clusters que variam com o tempo, quanto mais próximo de 1 (um) esse valor se encontra mais próximo da totalidade de alunos que formavam um perfil anterior. No exemplo, Fundamentos de Programação, estejam presentes no perfil da disciplina de um semestre seguinte, seguindo o exemplo, Estrutura de Dados. Além disso, uma análise auxiliar foi usada para determinar os padrões dos perfis de uma mesma disciplina, variando por semestre, para isso foi utilizada a distância euclidiana, que é a distância entre dois pontos que pode ser provada pela aplicação repetida do teorema de Pitágoras, e foi utilizada para medir a dissimilaridade entre os perfis encontrados por uma disciplina. Exemplificando, comparamos a dissimilaridade entre os cada um dos perfis da disciplina de Fundamentos de Programação entre os anos de 2006 até 2012, verificando a repetição desses perfis ao longo dos anos.

Por fim, este trabalho visou aplicar a técnica de clusterização aos dados da Universidade Federal do Ceará, a procura dos perfis dos alunos que cursam as disciplinas analisadas e responder a questionamentos sobre medidas adotadas pela UFC (Universidade Federal do Ceará) ao longo dos anos, como por exemplo, a nova regra de frequência de 2008.2, que mudou a quantidade de reprovações por faltas o aluno teria direito até ter sua matrícula cancelada, entre outras medidas, e a adesão ao SISU (Sistema de Seleção Unificado) em 2011.

Como resultados foram encontrados impactos das medidas citadas acima nos perfis das disciplinas, como por exemplo, a nova regra de frequência de 2008.2, influenciou positivamente no número de reprovações por falta, na análise feita usando dados de 2 anos

antes da medida ser colocada em pratica e 2 anos após, em todas as disciplinas. Enquanto a adesão ao SISU como única forma de ingresso a universidade, influenciou positivamente nas notas dos perfis de disciplina como Estrutura de Dados e Algoritmos em Grafos. Esses resultados e questionamentos estão disponíveis nas Seções: 5.1 – perfis de alunos encontrados por disciplina, 5.2 – impacto da adesão ao SISU pela Universidade Federal do Ceará nos perfis dos alunos, 5.3 – impacto da nova regra de frequência de 2008.2 nos perfis de alunos. Também foram encontrados casos de quebra de requisitos, que são quando o aluno cursa uma disciplina mesmo tendo reprovado a disciplina que servia de pré-requisito para ela, ao todo foram encontrados 5 casos, indicados nas Figuras 17 e 18 no apêndice A deste trabalho. Mostrando em aresta vermelha onde ocorrem os casos de quebras de requisito, com apenas um sucesso, do aluno que cursou Fundamentos de Programação no primeiro semestre de 2011 e reprovou, porém foi aprovado em Estrutura de Dados no segundo semestre do mesmo ano, os casos foram encontrados comparando as situações encontradas com os dados originais sem modificações.

Em Leonel Junior et al (2014) é apresentado um estudo de caso do qual participaram a orientadora e o autor deste trabalho. Em Leonel Junior et al (2014) foram coletadas cerca de 3 (três) milhões de tuplas referentes a informações socioeconômicas de aluno que prestaram o Exame Nacional do Ensino Médio (Enem) de 2011. Nesses dados, foi aplicada uma técnica de Mineração semelhante à que será usada neste trabalho, no intuito de descobrir perfis dos candidatos que prestam o exame, no total cerca de 7 (sete) perfis foram descobertos variando por renda da família, escolaridade dos pais e a disponibilidade, ou não, de internet.

Na seção 2, são apresentados os trabalhos relacionados. Na seção 3, são descritos os conceitos importantes para o entendimento desse estudo, na seção 4 são apresentados os procedimentos, na seção 5 é são detalhados os resultados, e por fim na seção 6 as considerações finais.

2 TRABALHOS RELACIONADOS

A seguir serão apresentados os trabalhos que foram usados como base para a criação desse estudo e uma breve descrição de cada, acompanhada das semelhanças e diferenças com relação a este trabalho.

No trabalho Kabakchieva et al (2012), foram utilizados dados dos alunos, referentes ao período da faculdade e ensino médio. Mais de 20 (vinte) atributos foram avaliados, incluindo: data de nascimento e lugar, local de nascimento e nacionalidade, tipo, perfil, colocação e pontuação total no ensino médio, ano de admissão na universidade, exame e pontuação obtida, semestre atual, pontuação total na universidade. Por meio da aplicação de técnicas de mineração de dados, foi possível classificar os alunos em grupos de acordo com o desempenho (ruim, mediano, bom, muito bom e excelente).

O trabalho de Kabakchieva et al (2012) assemelha-se a este trabalho na análise de dados dos alunos de ensino superior, com base nos atributos, encontrando padrões que podem classificar os alunos em grupos de acordo com o desempenho dos mesmos. Porém se diferencia do trabalho proposto por possuir dados dos alunos não só durante o período da universidade, como também dados durante o período de ensino secundário. Outra diferença é a técnica de mineração de dados escolhida para ser aplicada nos dados. No trabalho de Kabakchieva et al (2012) foi utilizada a classificação, que é um método preditivo, que tenta encontrar um padrão para um atributo que foi definido como classe.

Em Silva et al (2012) foi usada a base de dados do Enem de 2010, do qual foi retirada uma amostra somente dos estudantes da região Sul do Brasil. Os quesitos levados em consideração foram: o desempenho do aluno, e alguns fatores socioeconômicos, tais como, grau de instrução dos pais, acesso à internet e o tipo de escola no ensino médio – pública ou particular. O resultado do estudo foi capaz de estabelecer correlações sobre o desempenho dos estudantes. Por exemplo, se o estudante não tem internet e estuda em escola pública, provavelmente ele terá um resultado insatisfatório na prova do Enem. A semelhança deste trabalho se deve ao uso da técnica de clusterização, que foi aplicada aos dados dos alunos que prestaram o Exame Nacional do Ensino Médio (Enem), e que também deve ser aplicada neste estudo. As diferenças são principalmente na escolha dos atributos levados em consideração, em Silva et al (2012) fatores socioeconômicos foram priorizados durante a análise dos dados, e outras técnicas de mineração de dados foram utilizadas com o intuito de conseguir melhores resultados, foram elas: classificação e regras de associação. Regras de Associação tem como

princípio descobrir elementos que implicam no aparecimento de outros nos dados, ou seja, encontrar relacionamento/padrões no conjunto de dados. Outra diferença é a base de dados escolhida, neste trabalho é usada a base de dados dos alunos da Universidade Federal do Ceará.

Em De Moraes (2010) foram colhidos dados dos currículos dos professores da Escola Politécnica de Pernambuco (Poli) armazenados na plataforma Lattes – um sistema desenvolvido pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para auxiliar na pesquisa no Brasil. Foi desenvolvida uma ferramenta para coletar informações importantes desses currículos e aplicar técnicas de mineração de dados nelas, com o intuito de traçar um perfil dos profissionais, auxiliando assim os gestores nas tomadas de decisão com relação ao investimento nos cursos e nos docentes da instituição. A equivalência com esse trabalho está na técnica de clusterização que foi aplicada na base de dados. Entretanto, em De Moraes (2010) ocorre a aplicação de outra técnica de mineração, regras de associação, que foi utilizada em parte dos dados, tentando obter outros tipos de conhecimento do perfil dos profissionais que mais publicavam na instituição. Outro diferencial para este trabalho é a base de dados utilizada.

3 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados conceitos e definições importantes para o entendimento desse trabalho. Na Seção 3.1 é apresentada a definição de KDD (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados) e as etapas do processo. Logo depois, na Seção 3.2, é apresentada de forma mais detalhada a etapa de mineração de dados, que é um dos passos do processo de Descoberta do Conhecimento. Na Seção 3.3 é exposta a definição de Clusterização e o algoritmo *K-means* usado nessa técnica de mineração de dados e ferramentas que auxiliam na aplicação da técnica. Na Seção 3.4 é conceituada a Distância Euclidiana E por fim, na Seção 3.5 é definida a Similaridade de Jaccard, medida usada no trabalho de Coelho da Silva(2014), para validar a técnica de análise da evolução dos clusters.

3.1 Descoberta do Conhecimento

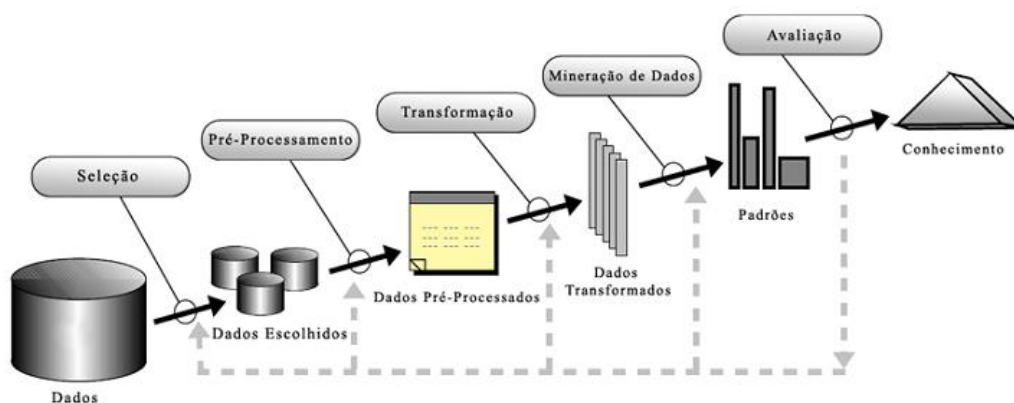
Segundo Fayyad (1996), a descoberta do conhecimento é o processo não-trivial de descoberta de padrões válidos, novos, potencialmente úteis e compreensíveis, implícitos em um conjunto de dados.

De acordo com Silva (2004), as etapas do KDD são definidas desta maneira: o processo de descoberta de conhecimento em banco de dados é feito de forma iterativo, iterativo, cognitivo e exploratório, nos quais estão envolvidos vários passos e as decisões são feitas pelo analista do domínio ou analista de dados. As etapas descritas a seguir também podem ser visualizadas na Figura 1:

1. Definição de qual o conhecimento se quer descobrir se entendendo o domínio, e o tipo de decisão que o conhecimento proporcionará em termos de melhora;
2. Seleção: É elegido o conjunto de dados com o qual se quer trabalhar, desses dados são tirados aqueles mais relevantes para a extração do conhecimento.
3. Limpeza de dados e pré-processamento: nesta fase são retirados dados fora do padrão (*outliers*), dados inconsistentes (dados incorretos ou redundantes), e é escolhida uma abordagem para tratar dados ausentes;
4. Transformação ou Formatação: consiste em redução e projeção dos dados para tornar mais fácil a análise dos mesmos.
5. Mineração de dados: fase na qual se escolhe a técnica e o algoritmo a ser utilizado nos dados para a obtenção do conhecimento.

6. Interpretação/Avaliação: os resultados obtidos com a mineração dos dados são avaliados por especialistas e serão julgados úteis ou não. O conhecimento útil deve ser interpretado de maneira que os usuários possam entender.
7. Conhecimento: Aqui o conhecimento obtido deve ser utilizado por quem é de interesse.

Figura 1 – Etapas do processo de descoberta do conhecimento



Fonte: Fayyad et al. (1996).

Todas as etapas da Descoberta de Conhecimento serão usadas neste trabalho, para realizar a análise dos dados da Universidade Federal do Ceará (UFC).

3.2 Mineração de Dados

A Mineração de dados é multidisciplinar e atinge as mais diversas áreas, de ciência da computação à estatística.

Em Hand et al (2001, p.6), a definição é dada de uma perspectiva estatística: “Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados”.

Em Cabena et al (1998, p.12), a definição é dada de uma perspectiva de banco de dados: “Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados”.

Em Fayyad et al (1996, p.40), a definição é dada da perspectiva do aprendizado de máquina: “Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados.”

A definição que usaremos neste trabalho se assemelha mais a de Fayyad (1996), e foi retirada do livro *Introduction to Data Mining*: “Mineração de dados é o processo automático de descoberta de conhecimentos úteis em grandes bases de dados” (TAN et al, 2006, p.2, tradução livre).

Existem diferentes técnicas utilizadas no processo de minerar os dados, que são de duas categorias. As preditivas, que tentam prever valores desconhecidos ou valores futuros dos dados, e as descritivas, que tentam encontrar padrões que descrevem os dados.

Dentre as técnicas preditivas, temos: a regressão/estimativa e a classificação. Regressão/Estimativa é semelhante à classificação, porém é usada quando a classe, atributo escolhido do conjunto dados para ser previsto, é um valor numérico que pode ser estimado matematicamente através dos demais atributos do conjunto de dados. Por exemplo, um proprietário deseja saber o valor de venda do imóvel, para isso ele pesquisa algumas variáveis de imóveis próximos, tais como: tamanho da propriedade, área construída, quantidade de quartos, quantidade de banheiros, valor da propriedade. Aplicando regressão nesses dados obtidos, é possível encontrar uma fórmula matemática que define o valor de venda do imóvel para o proprietário.

Classificar é prever a que classe um determinado registro do conjunto de dados pertence, para isso é fornecido um conjunto de registro já com as classes aos quais pertencem para que o algoritmo possa aprender cada padrão presente nos dados. Dessa forma, quando um novo registro é adicionado aos dados, o algoritmo é capaz de prever a que classe ele pertence. Por exemplo, temos um conjunto de dados de pacientes que são ou não predispostos a ter diabetes, quando um novo registro é adicionado o algoritmo deve ser capaz de dizer se o novo paciente é ou não propenso a ter diabetes.

Dentre as técnicas descritivas, temos: regras de associação e clusterização. Regras de associação é uma das técnicas mais conhecidas da mineração de dados devido ao seu uso, e consiste em analisar em um conjunto de dados a influência de um determinado dado sobre outro. Exemplificando, analisando lista de compras de um supermercado, podemos dizer que a compra de um determinado produto, influencia na compra de outro, assim como no conhecido caso do Walmart, em que se observou que, nos finais de semana, os clientes que compravam fralda, também compravam cerveja. Clusterização será vista na Seção 3.3.

A Mineração de dados será usada para obter o conhecimento desejado, através da aplicação de técnicas descritivas nos dados da Universidade Federal do Ceará (UFC) para obter os perfis acadêmicos dos alunos. E a técnica descritiva escolhida para ser aplicada neste trabalho é a clusterização.

3.3 Clusterização

Clusterização é uma técnica descritiva de mineração de dados, e consiste em encontrar grupos de objetos, visto que, objetos pertencentes a um determinado grupo devem ser similares entre si e diferentes dos objetos de outros grupos.

O algoritmo *k-means* é um algoritmo de clusterização simples e eficiente. Ele começa escolhendo aleatoriamente k registros para representar os centróides (médias) dos clusters. Todos os registros são colocados em determinado cluster com base na distância entre o registro e o centróide. Quando todos os registros tiverem sido colocados inicialmente em um cluster, o centróide do cluster é recalculado. O processo se repete, examinando cada registro novamente e atribuindo-o o cluster cujo centróide é mais próximo. Várias iterações podem ser necessárias, mas o algoritmo convergirá para um valor próximo ao ótimo.

Existem ferramentas de código aberto para auxiliar na aplicação da técnica de mineração de dados e os algoritmos, e dentre elas estão: Weka (*Waikato Environment for Knowledge Analysis*), Projeto R – ou somente “R”, e o Rapid Miner que será a ferramenta usada neste trabalho.

RapidMiner é uma suíte de mineração de dados Open Source desenvolvido pela empresa de mesmo nome, que fornece um ambiente integrado para a aprendizagem de máquina, mineração de dados, mineração de texto, análise preditiva e análise de negócios. O RapidMiner é escrito na linguagem JAVA. Ele é usado para aplicações comerciais e industriais, bem como para pesquisas.

Neste trabalho a clusterização será usada nos dados da universidade esperando-se que vários grupos de alunos sejam descobertos, e cada um desses grupos com suas características próprias formará um perfil de aluno.

3.4 Distância Euclidiana

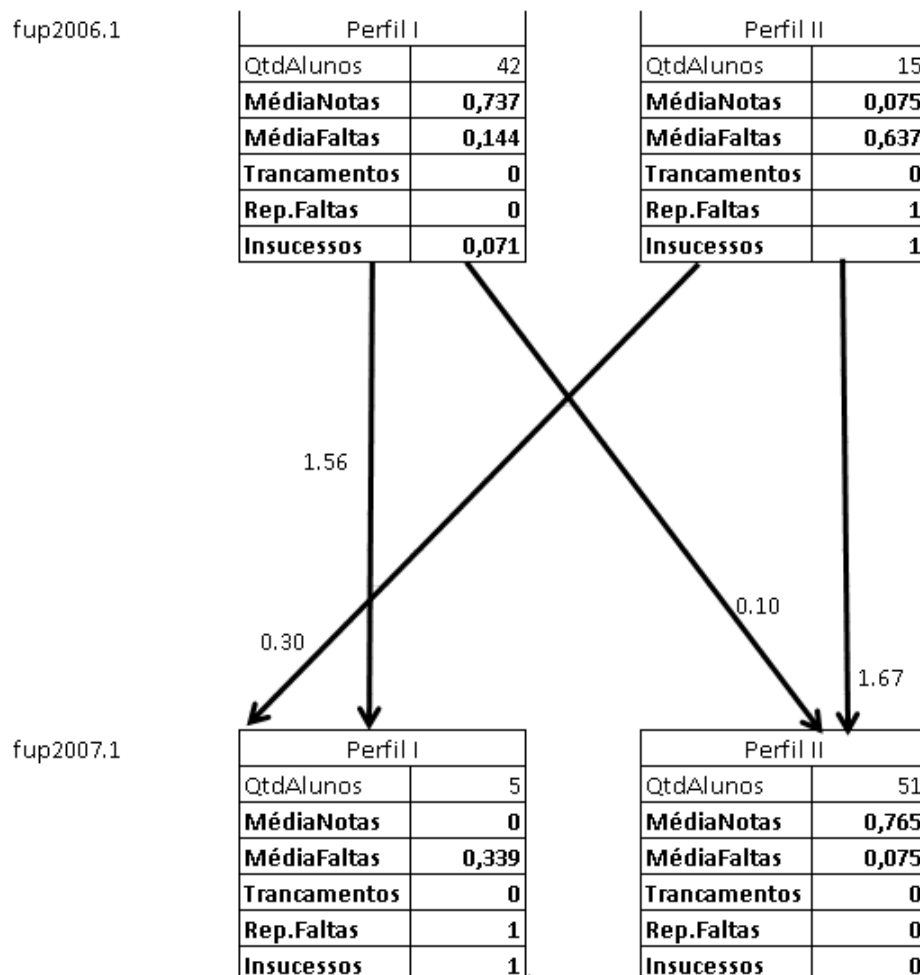
A Distância Euclidiana é uma medida de dissimilaridade entre dois vetores(O e O'), calculando a diferença ao quadrado das diferenças de suas dimensões, e tirando a raiz do somatório dessas diferenças. Quanto mais longe de zero está essa diferença, mais diferentes são esses vetores.

Figura 2 – Definição da Distância Euclidiana entre os vetores.

$$\sqrt{(O[1] - O'[1])^2 + (O[2] - O'[2])^2 + \dots + (O[n] - O'[n])^2}$$

Neste trabalho a Distância Euclidiana é utilizada para notar-se repetições de um mesmo perfil ao longo dos anos de uma disciplina. Na figura 5 temos parte de um DAG onde os vértices são os perfis e as arestas são as distâncias euclidianas entre os perfis, utilizado para analisar a disciplina de Fundamentos de Programação (FUP) de 2006 até 2012, onde cada perfil é equivalente a vetor, notamos que Perfil I da disciplina de FUP 2006.1 é similar ao Perfil II de FUP 2007.2, a distância euclidiana é próxima à zero (0.10), é possível afirmar que esse perfil se repetiu de um ano pra outro.

Figura 3 – Parte de um DAG da Disciplina de FUP entre 2006.1 e 2007.1.



3.5 Similaridade de Jaccard

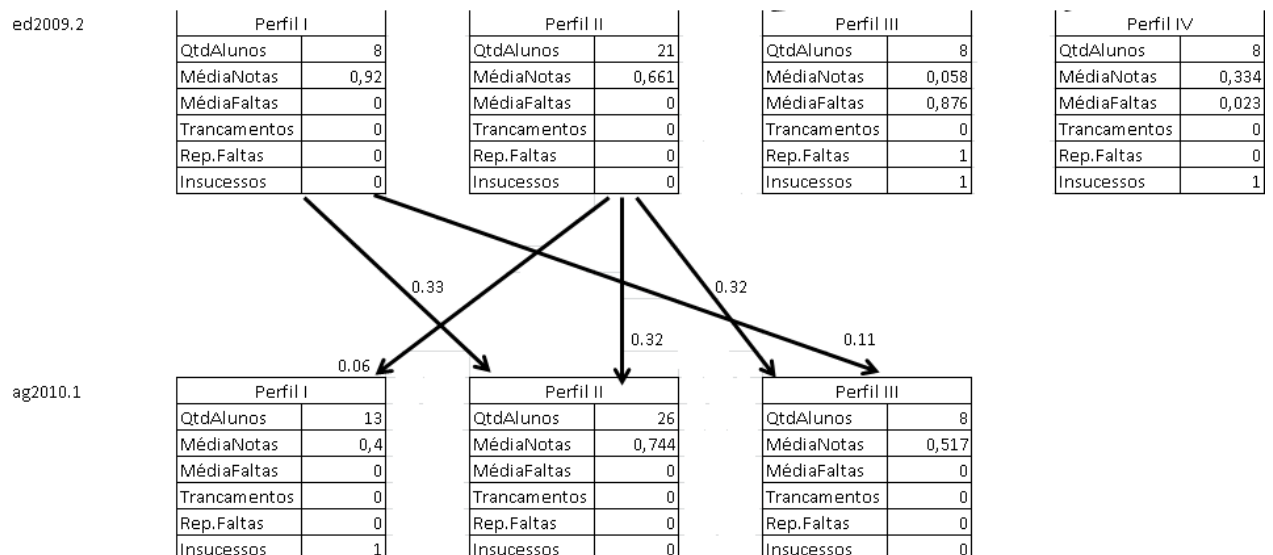
A Similaridade de Jaccard é uma medida usada para comparar a semelhança entre conjuntos de dados finitos, onde é feita a divisão do tamanho do conjunto interseção entre os conjuntos de dados pelo tamanho do conjunto união entre esses mesmos conjuntos de dados. Quanto mais próximo de 1(um) esse valor se encontra, mais semelhantes são esses conjuntos de dados. Na Figura 4, tem o cálculo da similaridade entre clusters C1 e C2, que no trabalho correspondem a perfis de alunos. Onde Oc1 são os identificadores de alunos que pertencem a C1, e Oc2 são os identificadores de alunos que pertencem a C2.

Figura 4 – Definição da Similaridade de Jaccard

$$J(Oc1, Oc2) = \frac{|Oc1 \cap Oc2|}{|Oc1 \cup Oc2|}$$

Neste trabalho a Similaridade de Jaccard foi usada para comparar a semelhança entre perfis de disciplina que possuem relação de dependência entre si. Na Figura 3 temos parte de um DAG (*Directed Acyclic Graph* – Grafo Direcionado Acíclico), onde os valores da aresta são as similaridades de Jaccard e os vértices são os perfis com seus atributos. No caso é possível observar que o perfil II da disciplina de Algoritmos em Grafos (AG) de 2010.1 é formado por 0.33 (33 por cento) dos alunos que estavam no Perfil I da disciplina de Estrutura de Dados (ED) de 2009.2, com mais 0.32(32 por cento) dos alunos que estavam no Perfil II da disciplina de ED, e os demais alunos que não vieram dos perfis de Estrutura de Dados em 2009.2.

Figura 5 – Parte de um DAG entre perfis da disciplina de ED 2009.2 e AG 2010.1



4 PROCEDIMENTOS

Este trabalho foi realizado em cinco etapas: (i) recebimento dos dados da Universidade Federal do Ceará, limpeza e transformação dos dados; (ii) definição de quais questões/análises deveriam ser feitas sobre os dados; (iii) aplicação da técnica de clusterização com o auxílio da ferramenta RapidMiner nos dados; (iv) Aplicação da similaridade de Jaccard e da distância euclidiana; e (v) criação de DAGs para análise dos resultados. A primeira etapa deste trabalho corresponde a 3 etapas da Descoberta do Conhecimento, totalizando assim as 7 etapas do processo.

4.1 Recebimento, limpeza e transformação dos dados

Em reunião com o diretor do campus da Universidade Federal do Ceará em Quixadá, na pessoa do Professor Doutor Davi Romero, foi decidido quais dados seriam passados para o autor deste trabalho. Para preservar o sigilo dos alunos presentes na base de dados, foi solicitada a criação de um procedimento no banco de dados para que algumas informações com respeito à identificação dos alunos fossem mascaradas. O id, identificador do aluno, foi trocado por um número sequencial qualquer, para que o aluno não pudesse ser rastreado através do cruzamento dessa informação com outras bases de dados da universidade.

Figura 6 – Dados recebidos e adicionados em nova tabela

Edit Data - PostgreSQL 9.3 (localhost:5433) - mineracao - tabelao2								
	id_situacao integer	id_discente integer	media_final numeric(4,2)	numero_falta integer	ano integer	periodo integer	id_componen integer	nome character varying(150)
1	4	6666	8.71	4	2010	1	658927	NAVEGACAO II
2	4	6666	8.60	3	2010	1	658940	AQUICULTURA II
3	4	6666	7.90	4	2007	2	697290	QUIMICA ANALITICA APLICADA
4	4	6666	7.85	18	2007	2	658919	BIOLOGIA PESQUEIRA I
5	4	6666	5.40	7	2006	1	659793	QUIMICA GERAL
6	4	6666	6.00	0	2006	2	659492	INTRODUCAO A ESTATISTICA
7	4	6666	7.70	1	2008	1	658918	BIOLOGIA PESQUEIRA II
8	4	6666	5.10	0	2006	2	659784	QUIMICA ORGANICA I
9	4	6666	8.40	6	2007	1	659842	INTRODUCAO A CIENCIA DA COMPUTA
10	4	6666	10.00	0	2010	2	658935	TRABALHO SUPERVISIONADO
11	4	6666	10.00	0	2009	2	658948	ADMINISTRACAO E LEGISLACAO PESQ
12	4	6666	9.30	8	2006	2	660047	ECOLOGIA BASICA
13	4	6666	7.00	2	2007	2	660078	INTRODUCAO A BIOQUIMICA
14	4	6666	7.80	7	2006	1	658914	PRINCIPIO DA CIENCIA PESQUEIRA
15	4	6666	6.30	11	2010	2	658923	PROCESSAMENTO DO PESCADOR II
16	4	6666	5.40	0	2008	1	697287	FUNDAMENTOS DE FISICO-QUIMICA
17	4	6666	7.00	3	2006	1	696135	BIOLOGIA GERAL I
18	4	6666	9.00	8	2006	1	659657	FISICA BASICA I
19	4	6666	7.80	5	2008	2	658908	OCEANOGRAFIA PESQUEIRA
20	6	6666	4.40	6	2007	1	697290	QUIMICA ANALITICA APLICADA
21	4	6666	5.90	0	2008	2	658909	PLANCTOLOGIA

De todos os dados recebidos da universidade foram selecionados somente os atributos que o autor deste trabalho, juntamente com a professora orientadora, consideraram mais importantes, e foram armazenados no SGBD(Sistema de Gerenciamento de Banco de

Dados) Postgres 9.3 em uma única tabela, como mostra a Figura 6. Os atributos escolhidos foram: “id_situacao_matricula” – Atributo que informa a atual situação do aluno na disciplina – Matriculado, Trancado, etc., “id_discente” – Identificador do Aluno, “media_final” – A média final do aluno na disciplina, “numero_faltas” – quantidade de faltas do aluno naquela disciplina durante o semestre, “ano” – ano em que o aluno cursou a disciplina, “período” – semestre em que o aluno cursou a disciplina, “id_componente_detalhes” – identificador da disciplina e “nome” – o nome da disciplina cursada. Foram retirados dos dados recebidos originalmente, aquelas linhas em que o valor da “media_final” do aluno era inexistente, vale ressaltar que valores iguais a zero foram mantidos, das mais de 5(cinco) milhões de tuplas presente no banco de dados, cerca de 3 milhões foram adicionadas nesse novo banco. Essa medida foi tomada porque sem as notas não faria sentido analisar os perfis dos alunos, levando em conta o desempenho dos mesmos na disciplina.

Foram então realizadas consultas no banco de dados para gerar arquivos CSV(Comma Separated Values), que são arquivos separados por vírgula contendo informação ordenada, como mostra a Figura 7. Esses arquivos CSVs foram gerados por disciplina – Fundamentos de Programação(FUP), Estrutura de Dados(ED), Algoritmos em Grafos(AG) e Construção e Análise de Algoritmos(CANA), por semestre, entre os anos de 2006 até 2012, do curso de computação da Universidade Federal do Ceará do campus de Fortaleza, contendo aproximadamente 1400 tuplas, com os seguintes atributos: a identificação do aluno, a média de notas, a média de faltas, a quantidade de vezes em que o aluno se matriculou na disciplina, a quantidade de trancamentos daquele aluno na disciplina, se o aluno reprovou por falta e os insucessos, métrica que reflete se ocorreu reprovação por nota, reprovação por falta ou trancamento no perfil. Todos esses atributos foram normalizados para que os valores ficassem no intervalo entre 0(zero) e 1(um), um caso especial dessa normalização foi o caso das médias de faltas, onde a quantidade de faltas do aluno foi dividida pela quantidade de horas-aula totais da disciplina. Vale ressaltar que todos os passos citados acima correspondem às fases do processo de Descoberta do Conhecimento:

- Seleção: A escolha da amostra de dados referentes às disciplinas de Fundamento de Programação, Estrutura de Dados, Algoritmos em Grafo e Construção e Análise de Algoritmos, entre os anos de 2006 até 2012, do curso de computação da Universidade Federal do Ceará em Fortaleza.
- Limpeza e pré-processamento dos dados: retirada das tuplas do banco de dados com valores de média final inexistente.

- **Transformação ou Formatação:** A normalização dos valores dos atributos entre 0(zero) e 1(um).

Figura 7 – CSV gerado para a disciplina de FUP 2006.1

```
C:\Users\Tercio\Downloads\Oitavo Semestre\TCC\gup2006 - cana2007\gup_2006.1.csv - Notepad++
Arquivo  Editar  Localizar  Visualizar  Formatar  Linguagem  Configurações  Macro  Executar  Plugins  Janela  ?
[Icons]
gup_2006.1.csv x
1 id_discente,media_notas,media_faltas,qtd_vezes_matriculas,trancamentos,rep_faltas,insucessos
2 152530,0.88,0.0625,1,0,0,0
3 154528,0.76,0,1,0,0,0
4 102543,0.77,0.234375,1,0,0,0
5 121747,0.04,0.5,1,0,1,1
6 141373,0.7,0.0625,1,0,0,0
7 125985,0.88,0.125,1,0,0,0
8 144213,0.77,0.1875,1,0,0,0
9 43274,0.7,0.234375,1,0,0,0
10 117689,0.7,0.25,1,0,0,0
11 157629,0.47,0.328125,1,0,1,1
12 128670,0.79,0.25,1,0,0,0
13 142706,0.54,0.125,1,0,0,0
14 96465,0,1,1,0,1,1
15 153781,0.67,0,1,0,0,0
16 69057,0.05,0.5625,1,0,1,1
17 154951,0.98,0,1,0,0,0
18 158407,0.7,0.125,1,0,0,0
19 72370,0.61,0.25,1,0,0,0
20 62845,0.7,0.0625,1,0,0,0
21 129098,0.06,0.625,1,0,1,1
22 40651,0.76,0.1875,1,0,0,0
23 165309,0.44,0.0625,1,0,0,1
24 57515,0.79,0.234375,1,0,0,0
25 121132,0,0.28125,1,0,1,1
```

4.2 Definição de quais análises/questões deveriam ser feitas sobre os dados

Esta fase do trabalho foi marcada por uma série de reuniões com o diretor do campus da Universidade Federal do Ceará em Quixadá, Professor Doutor Davi Romero, com a coordenadora do curso de Sistemas de Informação do campus da UFC em Quixadá, Professora Doutora Tânia Pinheiro, com o coordenador do curso de Ciência da Computação do campus da UFC em Quixadá, Professor Doutor Críston Pereira, a orientadora deste trabalho, Professora Mestra Ticiane Linhares e o autor deste trabalho. Para decidir que questões deveriam ser respondidas sobre os dados. Foram levantados os seguintes questionamentos:

- Quais os perfis de alunos baseados nas disciplinas de Fundamentos de Programação, Estrutura de Dados, Algoritmos em Grafos e Construção e Análise de Algoritmos?
- A adesão da Universidade Federal do Ceará ao SISU em 2011 teve impacto negativo nos perfis de alunos?

- A nova regra de frequência de 2008.2 teve impacto na permanência dos alunos nos anos subsequentes?
- Os aproveitamentos e quebras de requisitos impactam na continuidade dos alunos no curso?

4.3 Aplicação da técnica de clusterização com a ferramenta RapidMiner

Nesse ponto, foi aplicada a técnica de clusterização aos dados da universidade, com o auxílio da ferramenta *open source* RapidMiner. Os conjuntos de dados em formato CSV, foram passados para o programa, em que foram os seguintes parâmetros:

- Tipo de entrada de dados: CSV
- O atributo “id_discente” foi definido como label, para seu valor não interferir nos centroides dos clusters.
- Algoritmo *X-Means* (versão estendida do *K-means*).
- *k min*: 2 (quantidade de clusters mínimos)
- *k max*: 60 (quantidade de clusters máximos)
- *measure types*: *NumericalMeasures*
- *numerical measures*: *EuclideanDistance*
- *clustering algorithm*: *KMeans*
- *max runs*: 10 (número máximo de rodadas do algoritmo)
- *max optimization steps*: 100 (número máximo de passos por rodada)

X-means é uma variação do *K-means* onde a quantidade de clusters é definida pela pelo ponto onde os erros quadráticos dos clusters é máximo, isso define que a partir daquela quantidade, não vale a pena continuar dividindo em mais clusters.

Encontrou-se vários clusters de alunos com características distintas, cada um correspondendo a um perfil de aluno de uma determinada disciplina e de um determinado semestre analisado, como mostrado na Figura 8.

Figura 8 – Centróides encontrados na ferramenta RapidMiner para FUP 2006.1

Attribute	cluster_0	cluster_1
media_nota:	0.737	0.075
media_faltas:	0.144	0.634
qtd_vezes_n	1	1
trancamento	0	0
rep_faltas	0	1
insucessos	0.071	1

Centróides correspondem às médias dos atributos de cada cluster. Cada cluster corresponde a um perfil de aluno, como é mostrado na Figura 8, onde temos 2(dois) perfis de alunos para a disciplina de Fundamentos de Programação no ano de 2006 no primeiro semestre, um em que a média de notas é 0.734 – Valor discretizado entre 0(zero) e 1(um) para nota, que corresponde a 7,3 de nota em uma escala de 0(zero) a 10(dez), e o outro que possui notas por volta de 0.075 que em escala entre 0(zero) e 10(dez) corresponde a 0,7 de nota.

O mesmo procedimento foi feito para as disciplinas de Estrutura de Dados, Algoritmos em Grafos e Construção e Análise de Algoritmos entre os anos de 2006 até 2012.

4.4 Aplicação da Similaridade de Jaccard e da Distância Euclidiana

Ainda com o auxílio da ferramenta RapidMiner é possível descobrir quais alunos estão em cada perfil, através do identificador do aluno, “id_discente”, que são armazenados em um arquivo texto. Esse arquivo de texto é passado como parâmetro para um programa, feito pelo autor deste trabalho em linguagem de programação Java, onde são retirados os “id_discente” e seus respectivos clusters para que seja aplicada a Similaridade de Jaccard, a exemplo do que é feito no trabalho de Coelho da Silva et al(2014) para validar a técnica que mostra a evolução dos clusters (como eles surgem a partir de outros clusters). O programa retorna os valores das similaridades entre cada par de clusters.

Para complementar as análises também foi feito o cálculo da distância euclidiana entre perfis de uma mesma disciplina ao longo dos anos. Para isso também foi criado um programa em linguagem Java que recebe duas listas ordenadas, contendo os atributos de dois perfis e retorna a distância euclidiana entre os dois perfis.

4.5 Criação de DAGs para análise dos resultados

Nesta fase do trabalho foram criados DAGs(em tradução livre – Grafos Acíclicos Direcionados) para tornar mais simples a visualização dos resultados obtidos, tanto para

facilitar a análise como para tornar mais fácil a exposição dos resultados para os gestores da universidade. Na seção de Apêndices estão disponíveis todos os DAGs gerados neste trabalho. Os DAGs foram gerados com o auxílio da ferramenta *Excel* presente no pacote *Microsoft Office*.

5 RESULTADOS

Nesta seção estão todos os resultados encontrados das análises feitas sobre os dados da Universidade Federal do Ceará com a aplicação da técnica de clusterização. Na Seção 5.1 são mostrados os resultados referentes aos perfis dos alunos encontrados por disciplina. Na Seção 5.2 é mostrado o impacto da adesão ao SISU pela Universidade Federal do Ceará nos perfis dos alunos. Na Seção 5.3 é mostrado o impacto da nova regra de frequência de 2008.2 nos perfis de alunos.

5.1 Perfis de alunos encontrados por disciplina

Na subseção 5.1.1 são apresentados os perfis de alunos mais frequentes na disciplina de Fundamentos de Programação. Na subseção 5.1.2 são apresentados os perfis de alunos mais frequentes na disciplina de Estrutura de Dados. Na subseção 5.1.3 são apresentados os perfis de alunos mais frequentes na disciplina de Algoritmos em Grafos. E para finalizar, na subseção 5.1.4 são mostrados os perfis de alunos mais frequentes da disciplina de Construção e Análise de Algoritmos.

5.1.1 Fundamentos de Programação

Analizando os DAGs referentes à disciplina de Fundamentos de Programação(FUP) entre os anos de 2006 até 2012, que estão disponíveis no Apêndice D deste trabalho, Figuras 28 e 29, é notável a presença de um perfil que se repete em todos os anos sem muita alteração, como mostrado na Figura 9.

Figura 9 – Perfil mais frequente em FUP de 2006 até 2012

Perfil I	
MédiaNotas	0,8
MédiaFaltas	0,05
Trancamentos	0
Rep.Faltas	0
Insucessos	0

Se o grafo for percorrido desde a origem até o ano de 2012, esse perfil se mantém com pouca variação durante todos os anos (ele surge no perfil I de 2006). Com auxílio da distância euclidiana, podemos ver que as características desse perfil se repetem no perfil II de 2007, até chegar ao perfil II de 2012, sempre escolhendo as aresta onde o valor da distância euclidiana é mais próximo de 0(zero), que indica uma maior semelhança entres os perfis. As médias de notas dos alunos variam de valores entre 0.7 até 0.9 ao longo dos anos, que resulta

em uma média de notas em torno de 0.8, aproximadamente 8(oito) em escala de 0 a 10, como mostrado na Figura 9. A médias de faltas variam entre valores próximos à zero a no máximo 0.1 que resulta em uma média de 0.05, que equivale a uma porcentagem de 0(zero) a 100(cem) da quantidade de faltas do aluno na disciplina, assim esse valor de 0.05 equivale a 5%(5 por cento) em média de faltas para os alunos do perfil.

Os demais perfis possuem variações maiores em quase todos os atributos, e acabam se descaracterizando bastante com o passar dos anos para que seja possível afirma que são os mesmos perfis ao longo dos anos.

A disciplina de Fundamentos de Programação entre os anos de 2006 até 2012 apresenta um perfil bem definido de aluno, que tem notas altas e poucas faltas, sem reprovações ou trancamentos.

5.1.2 Estrutura de Dados

Analizando os DAGs referentes à disciplina de Estrutura de Dados entre os anos de 2006 até 2012, que estão disponíveis em Apêndice D deste trabalho, Figuras 30 e 31, é possível notar que a partir do ano de 2009 são mostrados 3(três) perfis bem formados e com pouca variação, como mostrados na Figura 10.

Figura 10 – Perfis mais frequentes em ED de 2009 até 2012

Perfil I			Perfil II			Perfil III	
MédiaNotas	0,8		MédiaNotas	0,5		MédiaNotas	0,2
MédiaFaltas	0		MédiaFaltas	0		MédiaFaltas	0
Trancamentos	0		Trancamentos	0		Trancamentos	0
Rep.Faltas	0		Rep.Faltas	0		Rep.Faltas	0
Insucessos	0		Insucessos	0		Insucessos	1

O perfil I da Figura 10 aparece a primeira vez no perfil I de ED do ano de 2009, percorrendo o grafo até 2012. Sempre escolhendo a aresta com a distância euclidiana mais próxima de 0(zero), é possível notar que ele se repete no perfil II de ED do ano de 2010, no perfil III de ED do ano de 2011, e por fim no perfil I de ED do ano de 2012. As notas dos alunos dos perfis semelhantes ao Perfil I da Figura 10, variam entre 0.7 e 0.9, durante os anos, o que resulta em uma média de aproximadamente 0.8 que em escala de 0(zero) a 10(dez) equivale a aproximadamente 8(oito), não possuem faltas, trancamentos ou reprovação.

O perfil II da Figura 10 aparece a primeira vez no perfil II de ED do ano de 2009, percorrendo o grafo até 2012, sempre escolhendo a aresta com a distância euclidiana mais próxima de 0(zero), pode-se observar que ele se repete no perfil I de ED do ano de 2010, fato

que volta a acontecer no perfil IV de ED do ano de 2011, e por fim acontece pela última vez no perfil II de ED do ano de 2012. As notas dos alunos dos perfis semelhantes ao perfil II da Figura 15, variam entre 0.4 até 0.6, que em média é aproximadamente 0.5, que em escala de 0(zero) a 10(dez) equivale aproximadamente a nota 5(cinco), sem faltas ou trancamentos.

O perfil III na Figura 10, aparece a primeira vez no perfil IV de ED do ano de 2009, percorrendo o grafo até 2012, sempre escolhendo a aresta com distância euclidiana mais próxima de 0(zero), nota-se que ele volta a aparecer no perfil III de ED do ano de 2010, no perfil I de ED do ano de 2011 e por fim no perfil III de ED do ano de 2012. Os alunos dos perfis semelhantes ao perfil III da Figura 15 tem notas entre 0.2 e 0.3, que em média ficou próximo ao valor 0.2, que equivale em escala de 0(zero) a 10(dez) a aproximadamente a nota 2(dois), sem faltas e sempre com valor 1(um) de insucessos – essa métrica analisa reprovações por falta, reprovações por nota e trancamentos. Como não existem faltas e nem tão pouco trancamentos nesse perfil, esse valor de insucesso igual a 1(um), significa que 100% (cem por cento) dos alunos desse perfil reprovaram a disciplina por nota.

Na disciplina de Estrutura de Dados entre os anos de 2009 até 2012, foram encontrados 3(três) perfis bem definidos de alunos, o perfil I com notas elevadas, sem faltas ou reprovações; perfil II com notas próximas a média e poucas faltas, sem reprovações; e o perfil III com notas baixas, poucas faltas e apresentando reprovações por nota.

5.1.3 Algoritmos em Grafos

Analizando os DAGs referentes à disciplina de Algoritmos em Grafos, entre os anos de 2007 até 2012, não foi analisado o ano de 2006, por essa disciplina ser de terceiro semestre do curso de computação, os dados referentes à disciplina de Fundamentos de Programação e Estrutura de Dados do ano de 2005 não estão presentes na amostra retirada dos dados, e para analisar a turma de 2006.1 de Algoritmo em Grafos seria importante acompanhá-la durante o ano anterior. Os DAGs estão no Apêndice D deste trabalho, Figuras 32 e 33. Pode-se observar no DAG referente à disciplina de Algoritmos em Grafos que existem 2(dois) perfis que se repetem ao longo dos anos, como mostrados na Figura 11.

Figura 11 – Perfis mais frequentes em AG de 2007 até 2012

Perfil I		Perfil II	
MédiaNotas	0,6	MédiaNotas	0,3
MédiaFaltas	0	MédiaFaltas	0
Trancamentos	0	Trancamentos	0
Rep.Faltas	0	Rep.Faltas	0
Insucessos	0	Insucessos	1

O Perfil I da Figura 11, surge a primeira vez no DAG no perfil II de AG do ano de 2007, percorrendo o grafo até 2012, escolhendo sempre a aresta com distância euclidiana próxima a 0(zero), nota-se as repetições desse perfil ao longo dos anos, em AG 2008 ele é semelhante ao perfil II, em AG 2009 é equivalente ao perfil I, em AG 2010 é semelhante ao perfil II, em AG 2011 é bem próximo ao perfil I, e por fim em AG 2012 é semelhante ao perfil II. Em todos os perfis que são semelhantes ao perfil I da Figura 11, a média de notas dos alunos varia entre 0.5 e 0.7, tirando média temos notas próximas de 0.6, que em escala de 0(zero) a 10(dez) equivale a nota 6(seis), não possuem reprovações ou trancamentos.

O Perfil II da Figura 11, surge a primeira vez no DAG no perfil I de AG do ano de 2007, percorrendo o grafo até 2012, escolhendo sempre a aresta com distância euclidiana próxima a 0(zero), nota-se as repetições desse perfil ao longo dos anos, em AG 2008 ele é semelhante ao perfil III, em AG 2009 ele não aparece, voltando a surgir em AG 2010 sendo semelhante ao perfil I, em AG 2011 é bem próximo ao perfil II, e por fim em AG 2012 é semelhante ao perfil III. Em todos os perfis que são semelhantes ao perfil I da Figura 11, a média de notas dos alunos varia entre 0.3 e 0.4, tirando média temos notas próximas de 0.3, que em escala de 0(zero) a 10(dez) equivale a nota 3(seis), possuem o valor de insucesso igual 1(um), como não temos reprovações por falta ou trancamentos, o valor de insucesso igual 1(um) significa que 100%(cem por cento) dos alunos desse perfil reprovam a disciplina por nota..

Em Algoritmos em Grafo entre os anos de 2007 até 2012, temos um perfil que aparece em todos os anos, que é o perfil I da Figura 11, composto por alunos que tiram notas próxima de 6(seis), com poucas faltas e sem reprovações. E o perfil II, que só não aparece no ano de 2009, mas é presente nos demais anos, com alunos que tiram notas próximas a 3(três), reprovam a disciplina por notas e possuem poucas faltas.

5.1.4 Construção e Análise de Algoritmos

A exemplo da disciplina Algoritmos em Grafos, não se tem os dados das disciplinas de Fundamentos de Programação e Estrutura de dados do ano de 2005, por isso a

análise dessa disciplina começa em 2007, para acompanhar a turma desde sua entrada em 2006 até chegar a disciplina de Construção e Análise de algoritmos que é do quarto semestre do curso de computação. Os DAGs referentes à disciplina de Construção e Análise de Algoritmos entre 2007 até 2012 estão no Apêndice D deste trabalho, Figuras 34 e 35. É possível observar 2 (perfis) que aparecem frequentemente ao longo dos anos, como mostrados na Figura 12.

Figura 12 – Perfis mais frequentes em CANA de 2007 até 2012

Perfil I			Perfil I	
MédiaNotas	0,6		MédiaNotas	0,2
MédiaFaltas	0		MédiaFaltas	0
Trancamentos	0		Trancamentos	0
Rep.Faltas	0		Rep.Faltas	0
Insucessos	0		Insucessos	1

O perfil I da Figura 12 aparece a primeira vez no DAG no perfil II de CANA em 2007, percorrendo o grafo até 2012, escolhendo sempre a aresta com o valor da distância euclidiana, nota-se as repetições do perfil I da Figura 12 ao longo dos anos, em CANA 2008 é próximo ao perfil I, em CANA 2009 é equivalente ao perfil III, em CANA 2010 é semelhante ao perfil I, em CANA 2011 é equivalente ao perfil III, e por fim em CANA 2012 é próximo ao perfil IV. Em todos os perfis semelhantes ao perfil I da Figura 12 ao longo dos anos, a média de notas dos alunos varia entre 0.5 a 0.6, que em média é próxima a 0.6, que em uma escala de 0(zero) a 10(dez) equivale a nota 6(seis), não apresentam faltas, reprovações ou trancamentos.

O perfil II da Figura 12 aparece a primeira vez no DAG no perfil I de CANA em 2007, percorrendo o grafo até 2012, escolhendo sempre a aresta com o valor da distância euclidiana, nota-se as repetições do perfil II da Figura 12 ao longo dos anos, em CANA 2008 é próximo ao perfil II, em CANA 2009 é equivalente ao perfil I, em CANA 2010 é semelhante ao perfil II, em CANA 2011 é equivalente ao perfil I, e por fim em CANA 2012 é próximo ao perfil II. Em todos os perfis semelhantes ao perfil I da Figura 12 ao longo dos anos, a média de notas dos alunos varia entre 0.1 a 0.4, que em média é próxima a 0.2, que em uma escala de 0(zero) a 10(dez) equivale a nota 2(seis), possuem o valor de insucesso igual 1(um), como não temos reprovações por falta ou trancamentos, o valor de insucesso igual 1(um) significa que 100%(cem por cento) dos alunos desse perfil reprovam a disciplina por nota.

Em Construção e Análise de Algoritmos, entre os anos de 2007 até 2012, temos 2(dois) perfis bem definidos de alunos, o perfil I de alunos com notas próximas a 6(seis), sem

faltas ou reprovações. E o perfil II de alunos com notas baixas, que sempre reprovam a disciplina.

5.2 Impacto da adesão ao SISU nos perfis dos alunos

Para essa análise foram considerados dados das disciplinas de Fundamentos de Programação e Estrutura de Dados entre 2009 e 2010 que correspondem a 2(dois) anos antes da adesão da Universidade Federal do Ceará ao SISU, e entre 2011 e 2012 que correspondem aos 2(dois) anos com o SISU como única forma de ingresso a universidade que continham nos dados. As disciplinas de Algoritmos em Grafos e Construção e Análise de Algoritmos foram analisadas dois anos antes da adesão ao SISU, os anos de 2010 e 2011, vale ressaltar que os alunos que cursam essas disciplinas em 2011 – ano de adesão ao SISU – ainda são alunos que entraram na universidade por vestibular tradicional, considerando que as disciplinas de Algoritmos em Grafo e Construção e Análise de Algoritmos, são de terceiro e quarto semestres respectivamente do curso de computação. E foram analisados apenas os perfis de um ano dessas disciplinas após a adesão ao SISU, que foram os alunos que cursaram as disciplinas no ano de 2012. Na Seção 5.2.1 é mostrado o impacto da adesão ao SISU nos perfis da disciplina de Fundamentos de Programação. Na Seção 5.2.2 é mostrado o impacto da adesão ao SISU nos perfis da disciplina de Estrutura de Dados. Na Seção 5.2.3 é mostrado o impacto da adesão ao SISU nos perfis da disciplina de Algoritmos em Grafos. Por fim, na Seção 5.2.4 é mostrado o impacto da adesão ao SISU nos perfis da disciplina de Construção e Análise de Algoritmos.

5.2.1 Impacto da adesão ao SISU na disciplina de Fundamentos de Programação

Na Figura 20, presente no Apêndice B deste trabalho, é mostrado o DAG gerado para a disciplina de Fundamentos de Programação entre os anos de 2009 até 2012, onde os vértices são os perfis e as arestas são as distâncias euclidianas entre os perfis, levando em consideração para o cálculo das distâncias apenas os dois atributos em negrito – “MédiaNotas” e “MédiaFaltas”.

Com relação às médias de notas, nos anos anteriores a adesão ao SISU, temos em 2009 o perfil I composto por 54 alunos, com média de notas próxima a 8(oito), sem faltas. Em 2010 temos o perfil II, composto por 29 alunos, que é bastante semelhante ao perfil I do ano de 2009 – com distância euclidiana por volta de 0.04 – que indica uma semelhança elevada entre os dois perfis. Ainda em 2010, temos o perfil II, composto de 28 alunos, com média de

notas por volta de 6,5. Se fizermos uma combinação do perfil I e II de 2010, temos um perfil novo, composto de 57 alunos, com média de notas por volta de 7(sete), que é bem semelhante ao perfil I do ano de 2009, é possível dizer que esse perfil de alunos com boas notas que é maioria em 2009, se repete em 2010. Nos anos de 2011 e 2012, que são os anos de adesão ao SISU, temos a repetição desse perfil de alunos com boas notas, em 2011 temos o perfil III, composto de 31 alunos, com média de notas por volta de 9(nove); e o perfil IV, composto de 19 alunos com média de notas por volta de 6.3. Se esses perfis de 2011 forem combinados, a exemplo do que foi feito em 2010, temos um novo perfil em 2011, composto de 50 alunos, com média de notas por volta de 7.6, bastante semelhante ao perfil I do ano de 2009, inclusive semelhante na quantidade de alunos e não só nas características. Em 2012, temos o perfil II, composto de 46 alunos, com média de notas em torno de 7.8, semelhante também ao perfil I do ano de 2009.

Analizando o impacto da adesão ao SISU com relação às notas dos alunos, os perfis de alunos com notas boas se mantêm constante durante os anos antes da adesão e depois, em quantidade de alunos e em média de notas, para a disciplina de Fundamentos de Programação.

Com relação à média de faltas, nos anos anteriores a adesão ao SISU, existe em 2009 o perfil II composto de 10 alunos, que possui a média de faltas mais elevada entre os perfis desse ano, cerca de 55%(cinquenta e cinco por cento). Em 2010, o perfil III composto de 11 alunos, é o que possui a média de faltas mais elevada, cerca de 21% (vinte e um por cento) de faltas. Nos anos de adesão ao SISU, em 2011 existe o perfil II composto por 6 alunos, com média de faltas em torno de 76%(setenta e seis por cento). Em 2012, tem o perfil I composto de 7 alunos, com média de faltas em torno de 65% (sessenta e cinco por cento).

Analizando o impacto da adesão ao SISU com relação às médias de faltas, nos anos antes do SISU o maior percentual de falta era 55%, depois do SISU aparecem perfis com médias de faltas entre 65% e 76%.

5.2.2 Impacto da adesão ao SISU na disciplina de Estrutura de Dados

Na Figura 21, presente no apêndice B deste trabalho, é mostrado o DAG gerado para a disciplina de Estrutura de Dados entre os anos de 2009 até 2012, onde os vértices são os perfis e as arestas são as distâncias euclidianas entre os perfis, levando em consideração para o cálculo das distâncias apenas os dois atributos em negrito – “MédiaNotas” e “MédiaFaltas”.

Com relação às médias de notas, nos anos anteriores a adesão ao SISU, em 2009 existe o Perfil I de boas notas, com média por volta de 9 (nove) e formado por 8(oito) alunos. Em 2010 tem perfil II com média de notas próxima de 7 (sete) composto por 17(dezessete) alunos. Nos anos de adesão ao SISU, em 2011 o perfil III de médias de notas por volta de 8(oito) é formado por 21(vinte e um) alunos. Em 2012 tem o perfil I com médias de notas por volta de 8(oito), composto por 28 alunos.

Analisando o impacto da adesão ao SISU com relação às notas dos alunos, nos anos antes do SISU a quantidade de alunos nos perfis com boas notas correspondem a 27% do total de alunos, e após a adesão do SISU essa porcentagem corresponde a 43%.

Com relação às médias de faltas, nos anos anteriores a adesão ao SISU, em 2009 aparece o perfil III composto por 8 alunos com média de faltas por volta de 87%(oitenta e sete por cento), esse perfil possui um semelhante em 2010, o perfil IV do ano de 2010 é composto por 3 alunos com média de faltas em torno de 100% (cem por cento). Esses perfis desaparecem nos anos da adesão ao SISU, em 2011 e 2012 não é possível encontrar semelhantes.

Analisando o impacto da adesão ao SISU com relação às médias de faltas, nos anos anteriores ao SISU aparecem perfis com médias de faltas elevadas, esses perfis desaparecem nos anos após a adesão ao SISU, o que demonstra uma diminuição significativa da média de faltas.

5.2.3 Impacto da adesão ao SISU na disciplina de Algoritmos em Grafos

Na Figura 22, presente no apêndice B deste trabalho, é mostrado o DAG gerado para a disciplina de Algoritmos em Grafos entre os anos de 2010 até 2012, onde os vértices são os perfis e as arestas são as distâncias euclidianas entre os perfis, levando em consideração para o cálculo das distâncias apenas os dois atributos em negrito – “MédiaNotas” e “MédiaFaltas”. Os alunos que cursam a disciplina no ano de 2011 ainda são aqueles que entraram por meio de vestibular tradicional.

Com relação às médias de notas, nos anos anteriores a adesão ao SISU em 2010 tem o perfil I composto de 26 alunos com notas em média 7,4. Em 2011 tem o perfil I composto por 35 alunos com notas em média 6,4. No ano de 2012, onde o predomínio de alunos que ingressaram pelo SISU é maior, têm 2(dois) perfis que se unidos refletem a situação de boas notas do anos anteriores, são eles: o perfil I composto por 18 alunos e com média de notas por volta de 8,7, e o perfil II composto por 16 alunos com notas médias por volta de 6,1.

Analisando o impacto da adesão ao SISU com relação às notas dos alunos, nos anos antes da adesão ao SISU o percentual de alunos nos perfis com boas notas correspondem a 64%, após o SISU esse percentual é 72%.

Com relação às médias de faltas, apenas no ano de 2011 são encontrados dois perfis em que a média dos alunos é superior 90%, são eles os perfis III e IV.

Devido à falta de um padrão mais consistente, não é possível afirmar se houve uma melhora ou piora na média de faltas após a adesão do SISU.

5.2.4 Impacto da adesão ao SISU na disciplina de Construção e Análise de Algoritmos

Na Figura 23, presente no apêndice B deste trabalho, é mostrado o DAG gerado para a disciplina de Construção e Análise de Algoritmos entre os anos de 2010 até 2012, onde os vértices são os perfis e as arestas são as distâncias euclidianas entre os perfis, levando em consideração para o cálculo das distâncias apenas os dois atributos em negrito – “MédiaNotas” e “MédiaFaltas”. Os alunos que cursam a disciplina no ano de 2011 ainda são aqueles que entraram por meio de vestibular tradicional.

Com relação às medias de notas, elas se mantem as mesmas com uma pequena melhora nas notas boas de 2010 para 2011, que sem mantém na turma com predominância de alunos que ingressaram pelo SISU em 2012.

Analisando o impacto da adesão ao SISU nas notas dos alunos, não houve alterações nas notas dos alunos para a disciplina de Construção e Análise de Algoritmo.

Com relação às médias de faltas, todos os perfis possuem média de faltas igual 0, que corresponde a 100%(cem por cento) de presença – se consideramos o valor 1 nesse atributo como sendo 100%(cem por cento) de faltas. Esse é um comportamento observado apenas nessa disciplina.

5.3 Impacto da nova regra de frequência de 2008.2 nos perfis dos alunos

Nesta parte do trabalho, foram analisadas as disciplinas 2(dois) anos antes da implantação da nova regra, e 2(dois) anos depois, o intervalo entre o segundo semestre de 2006 até o segundo semestre de 2010. A Seção 5.3.1 mostra a análise feita para a disciplina de Fundamentos de Programação. A Seção 5.3.2 mostra a análise feita para a disciplina de Estrutura de Dados. A Seção 5.3.3 mostra a análise feita para a disciplina de Algoritmos em Grafos. E por fim, a Seção 5.3.4 mostra análise feita para a disciplina de Construção e Análise de Algoritmos.

5.3.1 Impacto da nova regra de frequência nos perfis dos alunos de Fundamentos de Programação

No Apêndice C deste trabalho, a Figura 24 mostra o DAG criado para analisar o impacto da nova regra de frequência nos perfis dos alunos da disciplina de Fundamentos de Programação, vale ressaltar que a análise dessa disciplina começa no primeiro semestre de 2007, devido a ser uma disciplina do primeiro semestre e não ter dados disponíveis no segundo semestre de 2006 sobre ela. No DAG os vértices são os perfis e as arestas são as distâncias euclidianas levando em consideração apenas o atributo “Rep.Faltas”.

Nesta análise, somente foi feita uma contagem dos alunos reprovados antes e depois da nova regra de frequência. Nos anos anteriores a implantação da nova regra de frequência, temos 16 alunos reprovados por falta. Nos anos após a implantação da nova regra de frequência, temos 8 alunos reprovados por falta.

Analisando o impacto da nova regra de frequência nas reprovações por faltas para a disciplina de Fundamentos de Programação durante os anos de 2007 até 2010, nos anos anteriores a implantação da regra, os reprovados por falta correspondem a 13% do total de alunos, após a implantação da nova regra essa porcentagem corresponde a 6% do total de alunos.

5.3.2 Impacto da nova regra de frequência nos perfis dos alunos de Estrutura de Dados

No Apêndice C deste trabalho, a Figura 25 mostra o DAG criado para analisar o impacto da nova regra de frequência nos perfis dos alunos da disciplina de Estrutura de Dados, a análise é feita entre o segundo semestre de 2006 até o segundo semestre de 2010, diferente de Fundamentos de Programação, nessa análise são considerados 2(dois) anos antes de implantação da nova regra de frequência e 3(três) anos após. No DAG os vértices são os perfis e as arestas são as distâncias euclidianas levando em consideração apenas o atributo “Rep.Faltas”.

Nesta análise as reprovações por falta se mantêm constantes, praticamente sem alterações que possam determinar que a nova regra de frequência de 2008.2 teve algum impacto positivo ou negativo sobre os perfis desta disciplina.

5.3.3 Impacto da nova regra de frequência nos perfis dos alunos de Algoritmos em Grafos

No Apêndice C deste trabalho, a Figura 26 mostra o DAG criado para analisar o impacto da nova regra de frequência nos perfis dos alunos da disciplina de Algoritmos em Grafos, a análise é feita entre o primeiro semestre de 2007 até o primeiro semestre de 2010. No DAG os vértices são os perfis e as arestas são as distâncias euclidianas levando em consideração apenas o atributo “Rep.Faltas”.

Nos anos anteriores a implantação da nova regra de frequência existem 14 alunos reprovados por falta. Nos anos após a implantação existem 4 alunos reprovados por falta.

Analizando o impacto da implantação da nova regra de frequência, nos anos anteriores a implantação da regra os alunos reprovados correspondem a 13% do total de alunos. Nos anos após a implantação das regras, esse percentual é 4%.

5.3.4 Impacto da nova regra de frequência nos perfis dos alunos de Construção e Análise de Algoritmos

No Apêndice C deste trabalho, a Figura 27 mostra o DAG criado para analisar o impacto da nova regra de frequência nos perfis dos alunos da disciplina de Construção e Análise de Algoritmos, a análise é feita entre o segundo semestre de 2007 até o segundo semestre de 2010. No DAG os vértices são os perfis e as arestas são as distâncias euclidianas levando em consideração apenas o atributo “Rep.Faltas”.

Em 2007, único ano antes da implantação da nova regra de frequência que faz parte dessa análise, aparecem 8(oito) alunos reprovados por falta. Nos anos após a implantação da nova regra de frequência temos 4(quatro) alunos reprovados por falta.

Analizando o impacto da implantação da nova regra de frequência, antes da nova regra de frequência os alunos reprovados por falta correspondem a 17% do total de alunos. Nos anos após a implantação da regra, esse percentual é de 3% do total de alunos.

6 CONSIDERAÇÕES FINAIS

Este trabalho procurou demonstrar que é possível encontrar os perfis dos alunos da Universidade Federal do Ceará aplicando técnicas de mineração aos dados da universidade. Analisando apenas uma amostra dos dados referentes às disciplinas de Fundamentos de Programação, Estrutura de Dados, Algoritmos em Grafo e Construção e Análise de Algoritmo dos anos de 2006 até 2012, do curso de Computação da UFC em Fortaleza. Usando medidas para determinar a consistência desses perfis encontrados como, por exemplo, a distância euclidiana que foi usada para encontrar os perfis mais frequentes por disciplinas, ou a similaridade de Jaccard que foi usada para definir como os perfis eram formados ao longo dos anos, respeitando a dependência entre as disciplinas e analisando uma mesma turma que cursa todas as disciplinas.

Além disso, este trabalho analisou o impacto de algumas medidas adotadas pela Universidade Federal do Ceará ao longo dos anos, como a implantação da nova regra de frequência de 2008.2, analisando os pontos negativos e positivos que apareceram no perfil dos alunos antes e depois da medida. A implantação da nova regra frequência teve um impacto positivo sobre o índice de reprovação por falta dos alunos nas disciplinas de Fundamentos de Programação, Algoritmos em Grafo e Construção e Análise de Algoritmos. Além disso, a adesão ao SISU no ano de 2011 como única forma de ingresso dos alunos a Universidade Federal do Ceará, apresentou impacto nas notas dos alunos. Em Fundamentos de Programação e Construção e Análise de Algoritmos, não se alteraram, e até melhoraram, como ocorreu em Estrutura de Dados e Algoritmos em Grafos. Com relação à frequência dos alunos, tivemos dois casos: na disciplina de Fundamentos de Programação aconteceu um aumento na média de falta dos alunos após a adesão ao SISU, enquanto na disciplina de Estrutura de Dados ocorreu uma melhora significativa dessas médias, o que não ocorreu nas disciplinas de Algoritmos em Grafos e Construção e Análise de Algoritmos. Sem mais dados para analisar, não é possível afirmar se houve uma melhoria ou uma piora da média de faltas após a adesão ao SISU. O que foi encontrado pode ser reflexo da implantação da nova regra de frequência adota em 2008.2.

Durante o trabalho, o aprofundamento na área trouxe um amadurecimento que sem essa pesquisa não seria alcançado. As dificuldades encontradas ao longo do caminho também auxiliaram nesse amadurecimento, como por exemplo o pouco tempo para realizar uma pesquisa que deveria ser mais longa, entre outros desafios que surgiram ao longo do caminho.

Os dados geraram uma infinidade de questionamentos que poderiam ser feitos sobre eles, fazendo assim com que o escopo do trabalho fosse reduzido. Apesar de serem analisadas apenas 1400 tuplas, após tirar apenas uma parcela da base original, o que foi feito nesse trabalho pode ser replicado para qualquer quantidade de dados.

Como trabalhos futuros, ficaram repetir as análises feitas nesse trabalho considerando apenas 2 atributos, a média de notas e a média de faltas, para analisar se a quantidade de perfis muda, também é possível usar o professor que ministrou a disciplina como uma das variáveis de análise. Além disso, ficam questionamentos sobre a evasão na universidade e a criação de um sistema integrado aos sistemas já existentes na universidade para mostrar as análises.

REFERÊNCIAS

- BERKHIN, Pavel. 2006. *Survey of Clustering Data Mining Techniques.*, Springer Berlin Heidelberg, p.56
- CARREIRA, Suely da Silva, et al. **Aplicação de Data Mining na Base de Dados do Processo Seletivo do Exame Nacional do Ensino Médio - ENEM 2010** In: CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO, II, 2012, Ponta Grossa. p.12
- CABENA, P; HADJINIAN, P; STADLER, R; JAAPVERHEES; ZANASI, A. **Discovering Data Mining: From Concept to Implementation.** Prentice Hall, 1998.
- COELHO DA SILVA, Ticiane Linhares, et al. Discovering Frequent Mobility Patterns on Moving Object Data In: **MobiGIS, ACM Sigspatial 2014**, Texas, EUA, p.8;
- DE MORAIS, Bruno Carlos Sales. **Extração de conhecimento da plataforma lattes utilizando técnicas de mineração de dados: Estudo de caso Poli/UPE.** 2010. 54 f. Trabalho de Conclusão de Curso (Graduação em Engenharia da Computação) – Universidade de Pernambuco, Recife, 2010.
- FAYYAD, Usama; PIATETSKY-SHAPIO, Gregory.; SMYTH, Padhraic. From *Data Mining to Knowledge Discovery in Databases*. **AI MAGAZINE**, Providence, p.37-54, 1997. Disponível em: <<http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>>. Acesso em: 20 mar. 2014.
- HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining.** MIT Press, 2001.
- KABACHIEVA, D.; STEFANOVA, K.; KISIMOV, V. **Analyzing University Data for Determining Student Profiles and Predicting Performance.** Sofia, Bulgaria: Sofia University, 2011. Disponível em: <http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_poster15_Kabakchieva.pdf>. Acesso em: 10 mar. 2014.
- LEONEL JUNIOR, Raimundo de Acácio, et al. Mineração em Dados Abertos In: **Jornada científica de sistemas de informação, IV**, 2014, Parnaíba p.20
- SILVA, Marcelino P. Santos. (2004) “**Mineração de Dados- Conceitos, Aplicações e Experimentos com Weka**”. In: Artigo. Instituto Nacional de Pesquisas Espaciais (INEP), 2004. São José dos Campos-SP.
- TAN, P.N.; STEINBACH, M; KUMAR, V. **Introduction to Data Mining.** New York: Pearson Addison Wesley, 2006.

APÊNDICES

APÊNDICE A – DAGs relacionados às trilhas de disciplinas de 2006 até 2012

Esses DAGs os vértices são os perfis de alunos por disciplina e as arestas são os valores das similaridades de Jaccard calculadas entres os perfis.

Figura 13 – DAG entre FUP 2006.1 e CANA 2007.2.

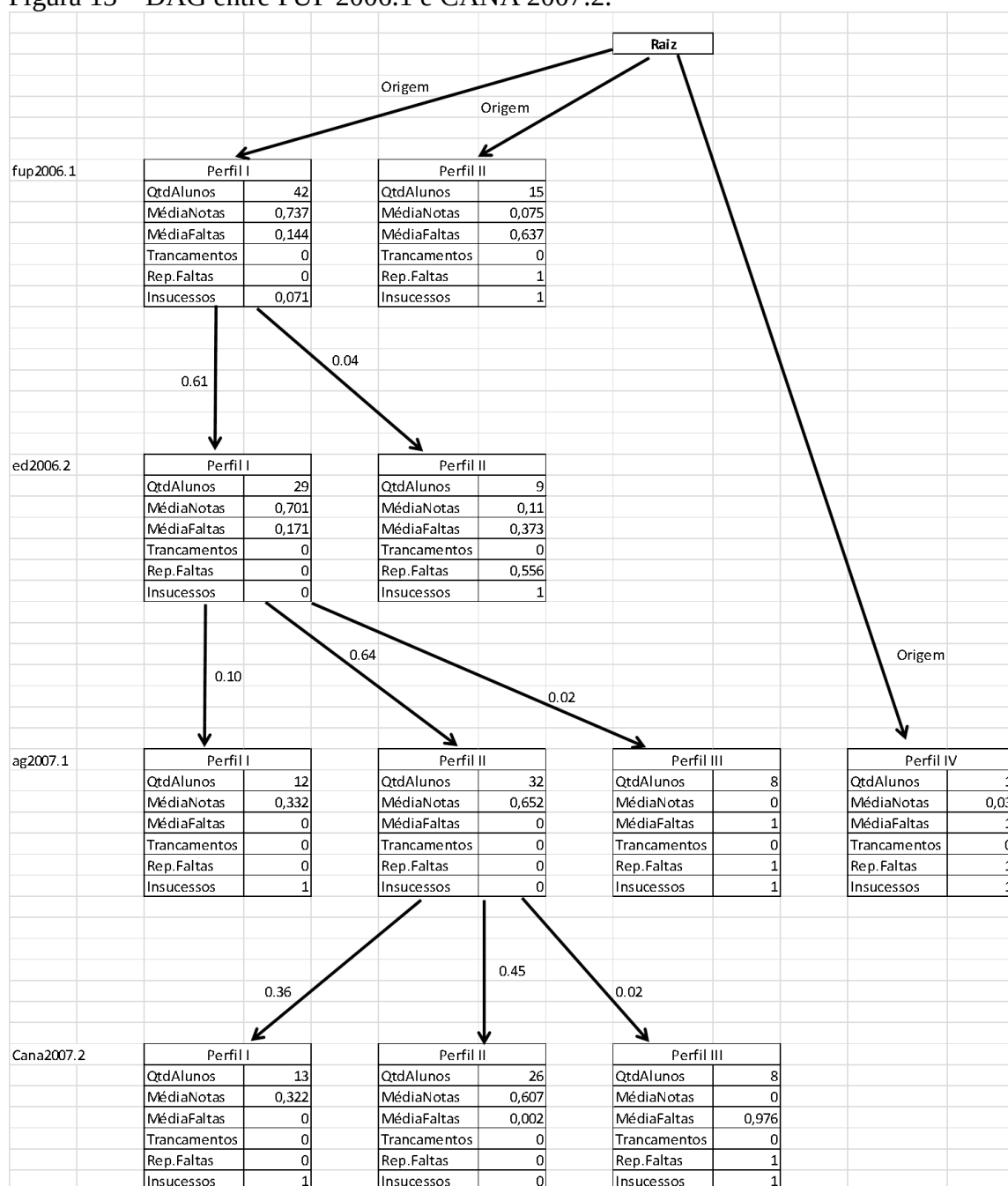


Figura 14 – DAG entre FUP 2007.1 e CANA 2008.2.

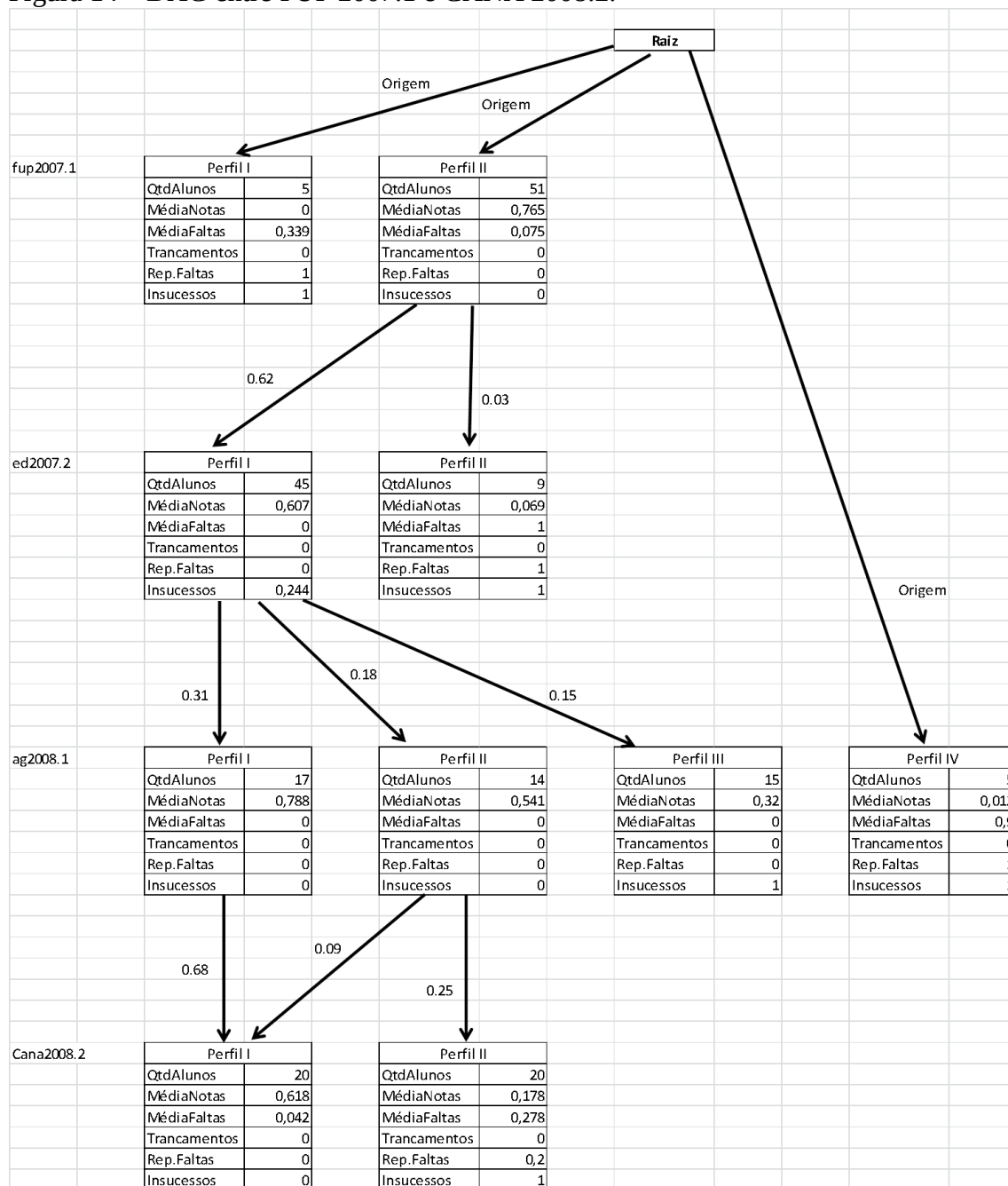


Figura 15 – DAG entre FUP 2008.1 e CANA 2009.2.

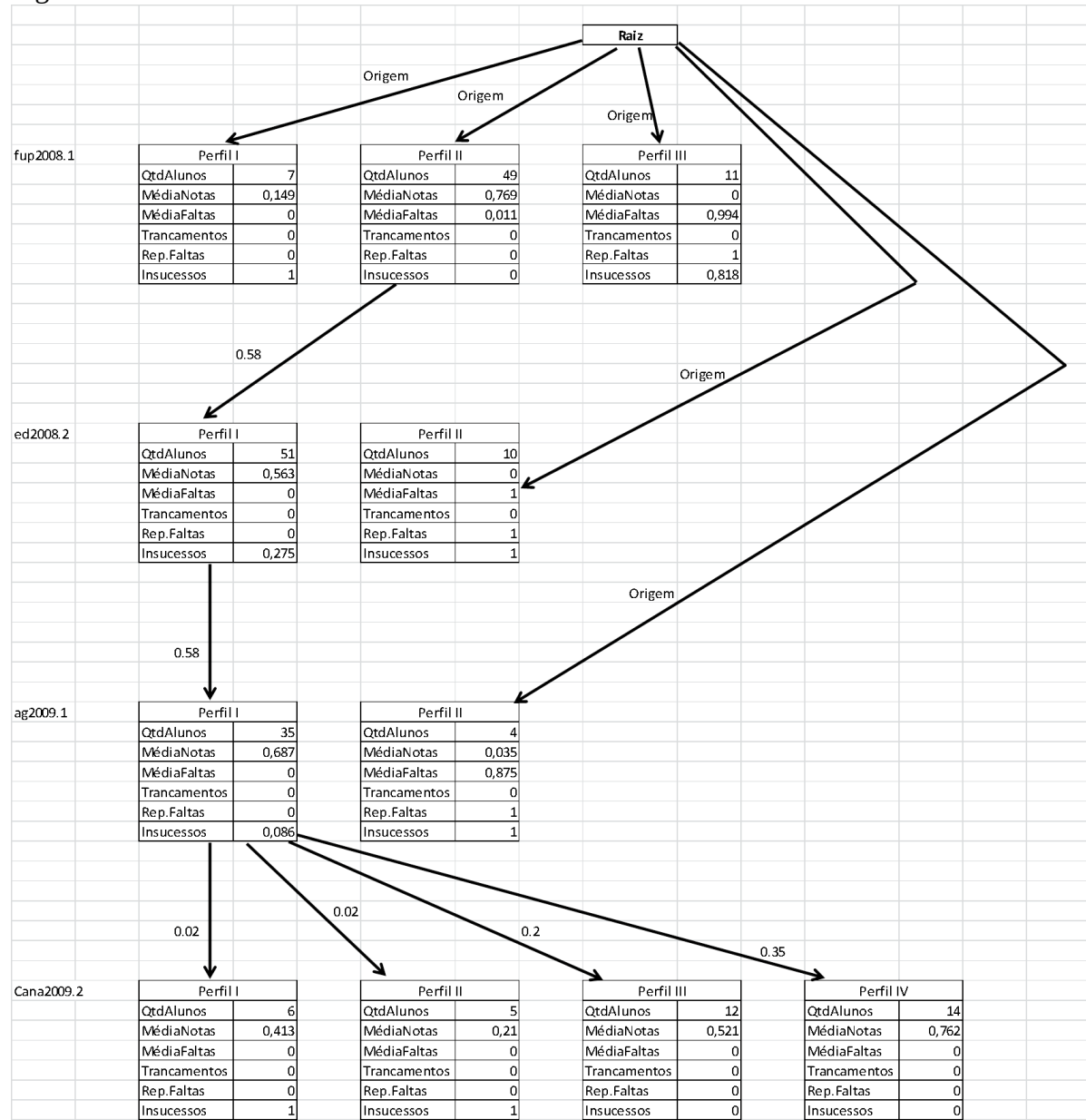


Figura 16 – DAG entre FUP 2009.1 e CANA 2010.2.

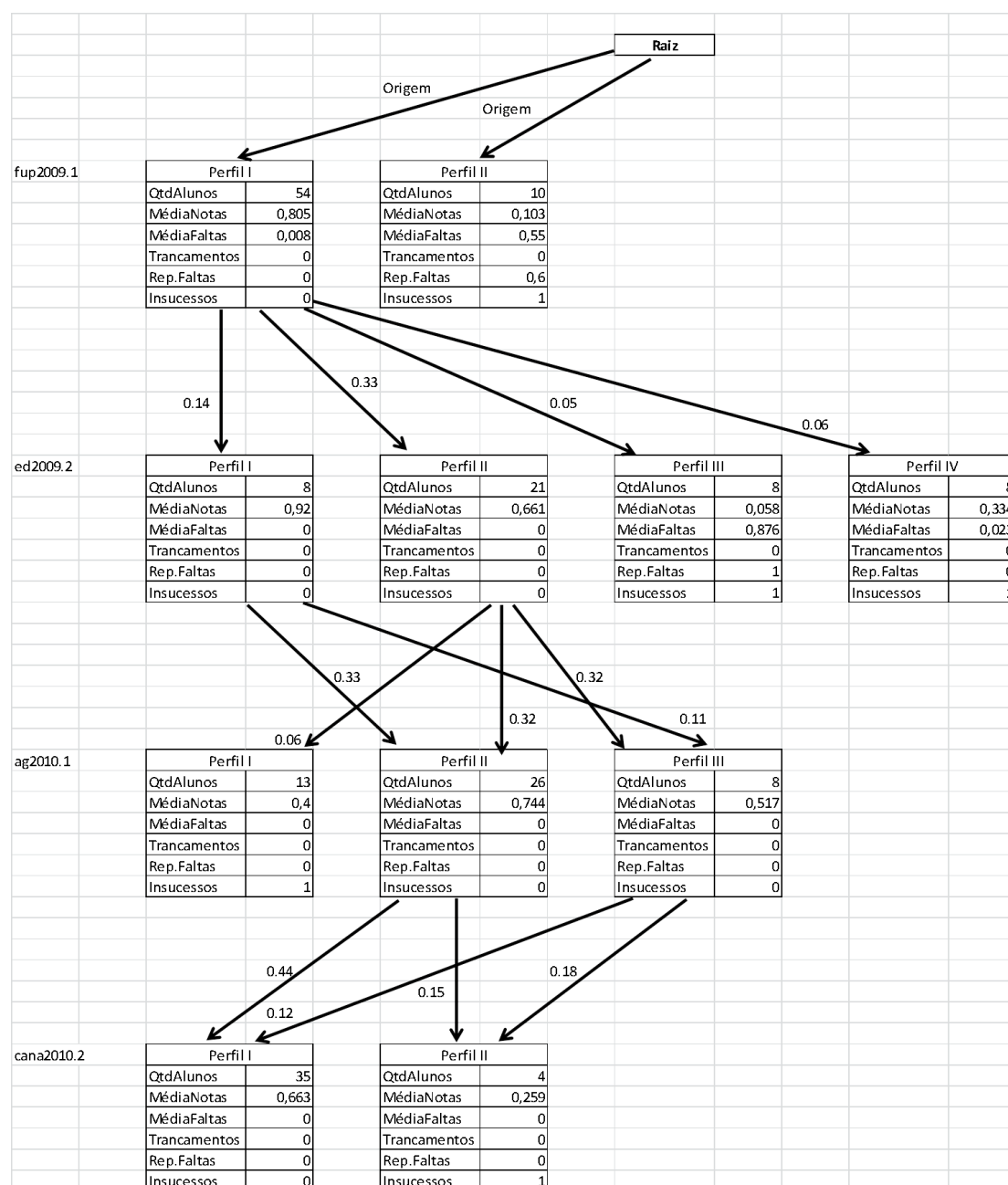


Figura 17 – DAG entre FUP 2010.1 e CANA 2011.2.

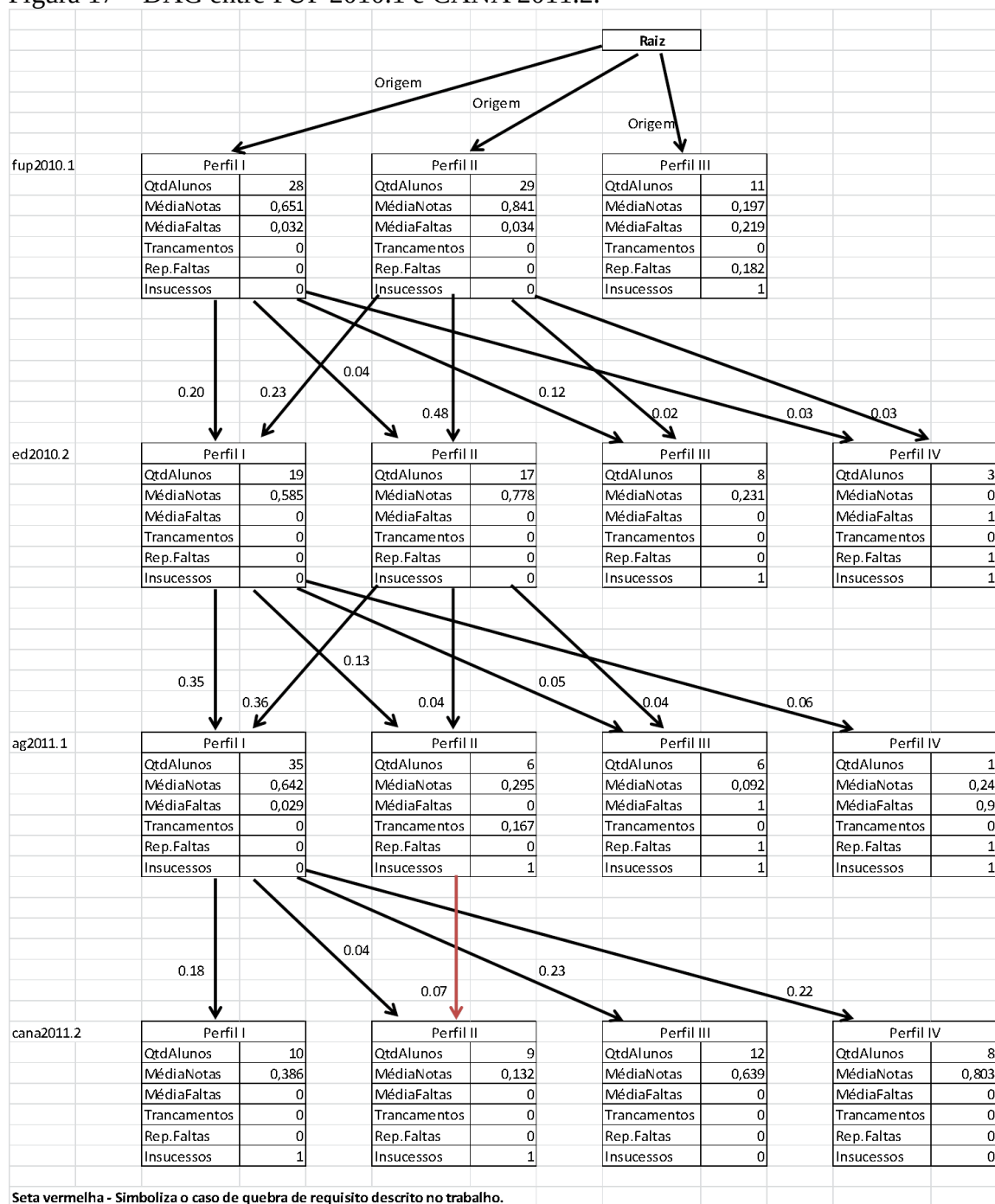


Figura 18 – DAG entre FUP 2011.1 e CANA 2012.2

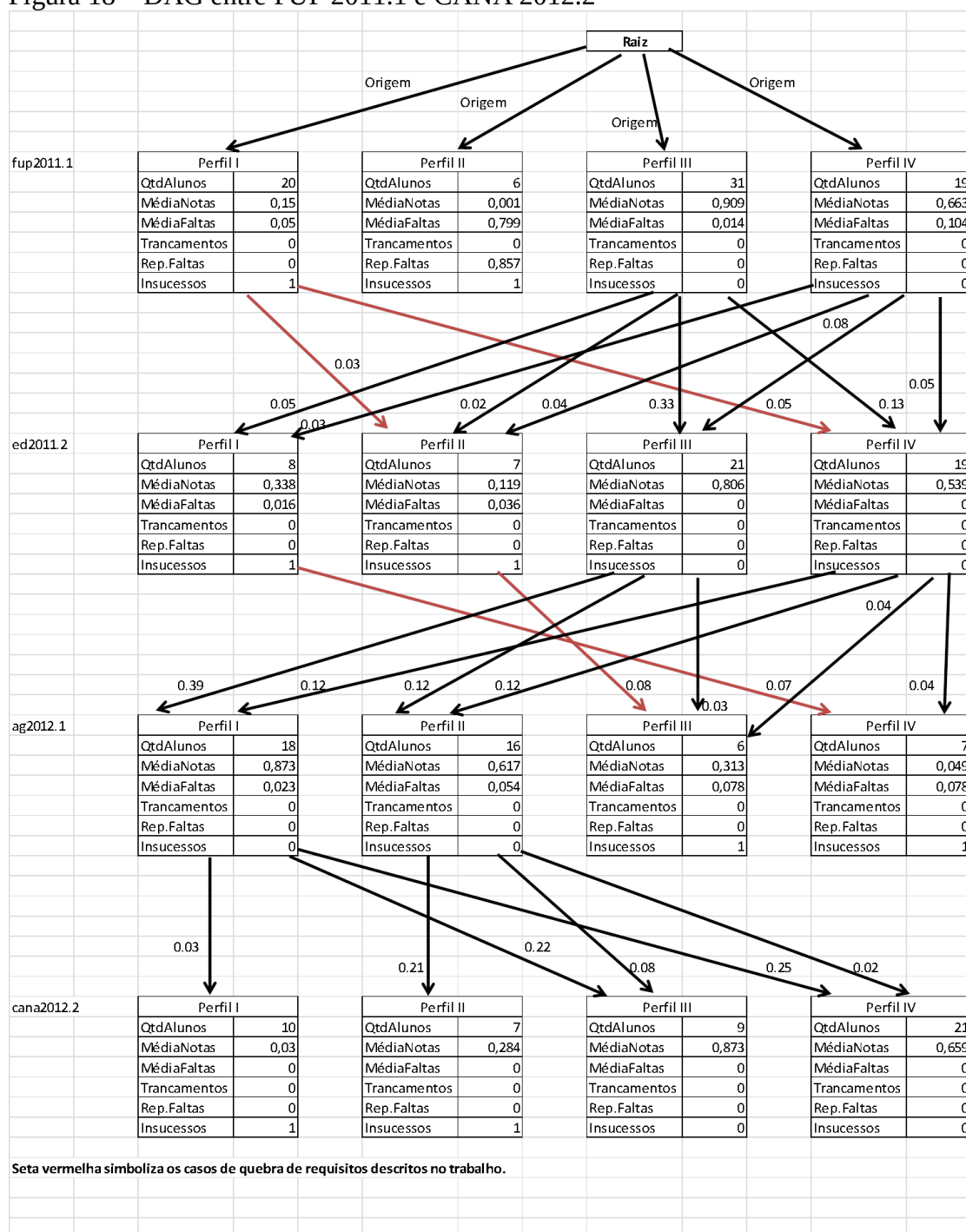
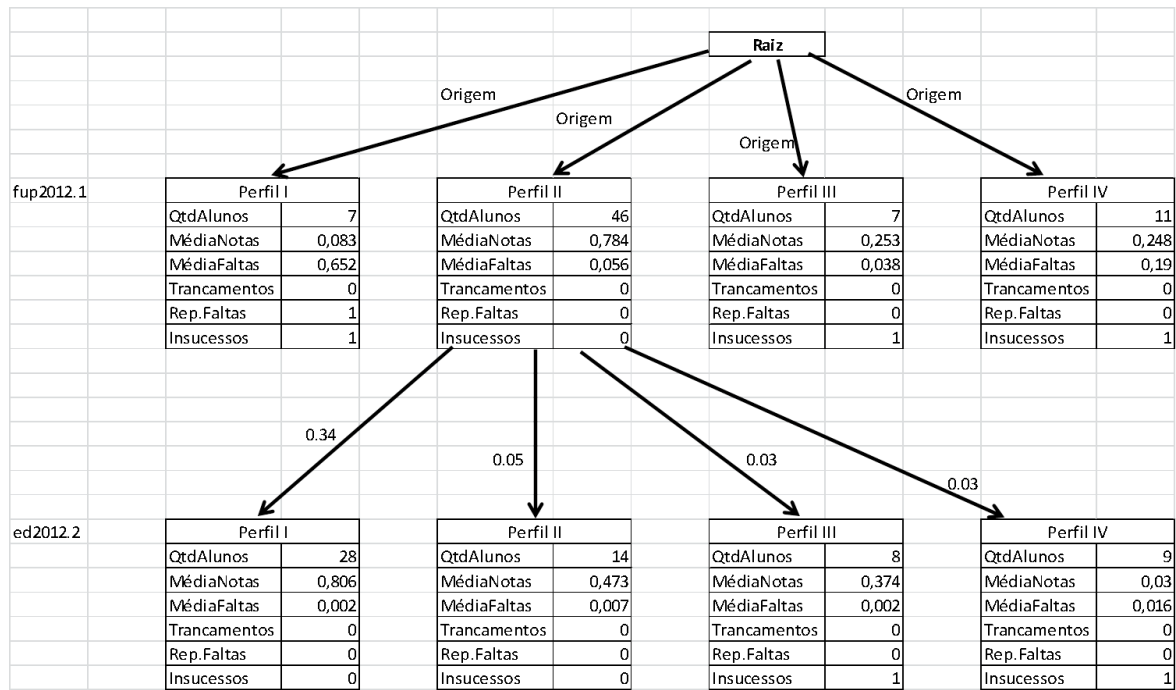


Figura 19 – DAG entre FUP 2012.1 e ED 2012.2



APÊNDICE B – DAGs relacionados às análises do impacto do SISU nos perfis de alunos entre 2009 até 2012

Esses DAGs os vértices são os perfis de alunos por disciplina e as arestas são os valores das distâncias euclidianas calculadas entres os perfis.

Figura 20 – DAG entre FUP 2009.1 e FUP 2012.1

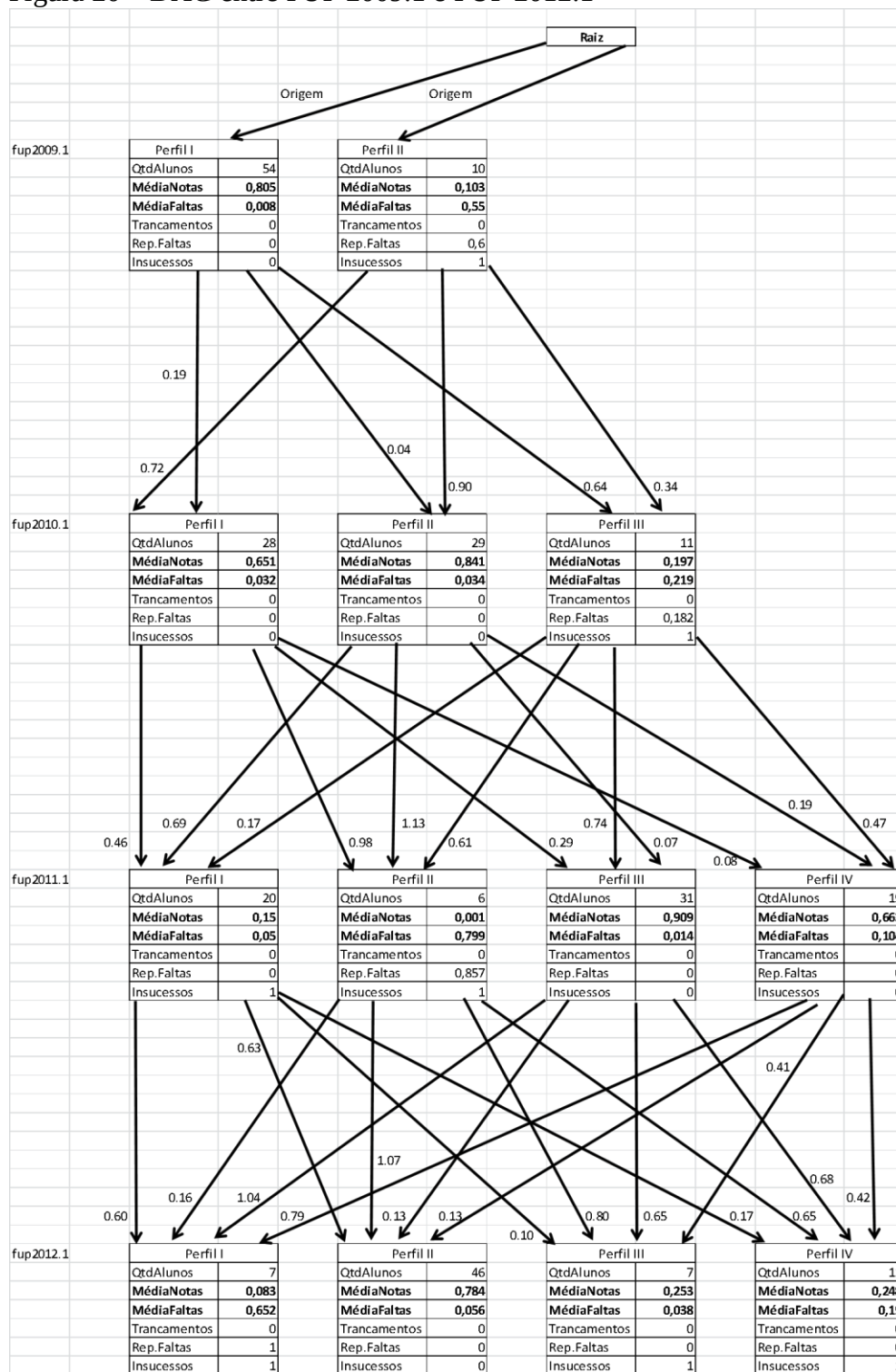


Figura 21 – DAG entre ED 2009.2 e ED 2012.2

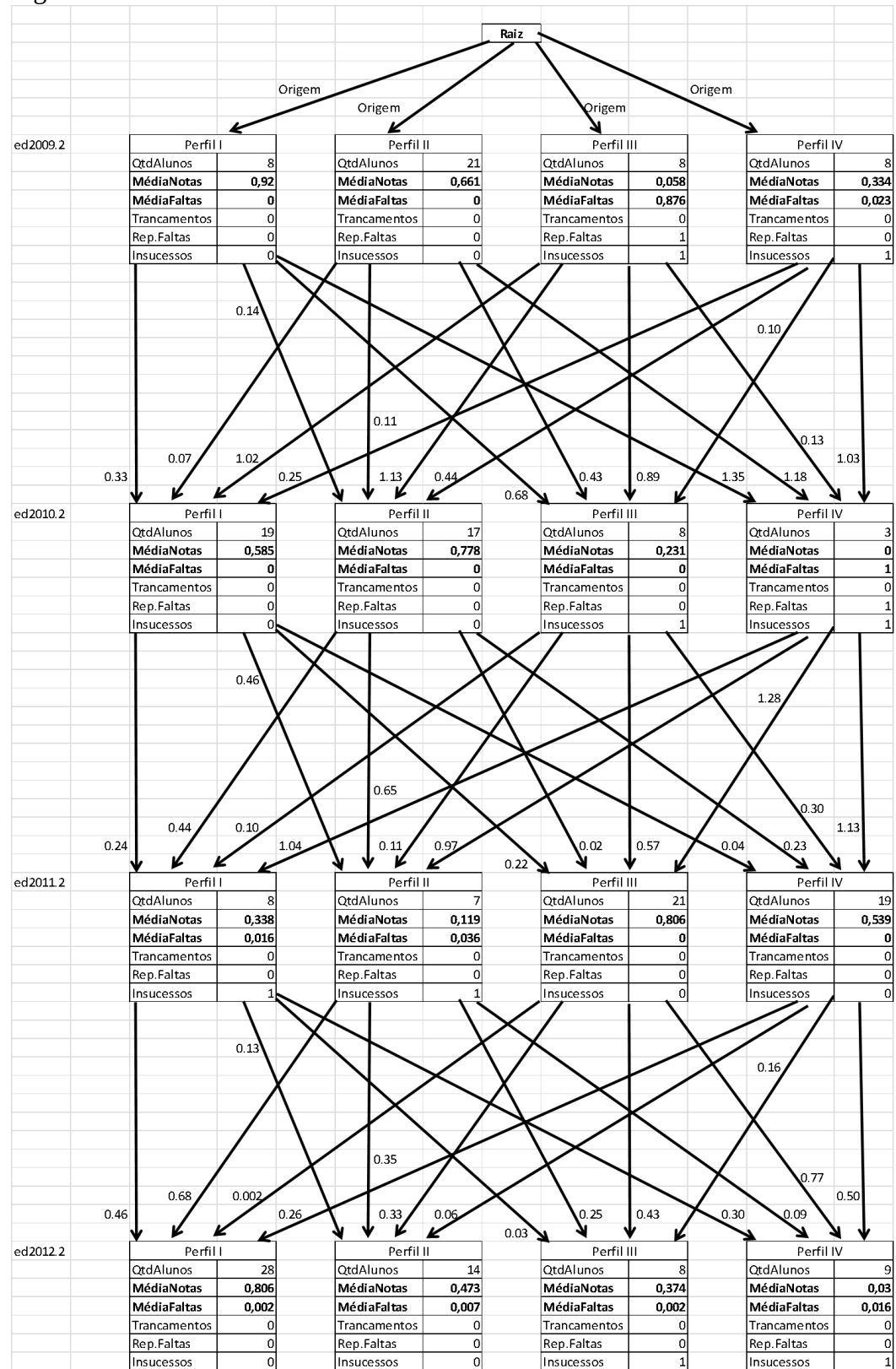


Figura 22 – DAG entre AG 2010.1 e AG 2012.1

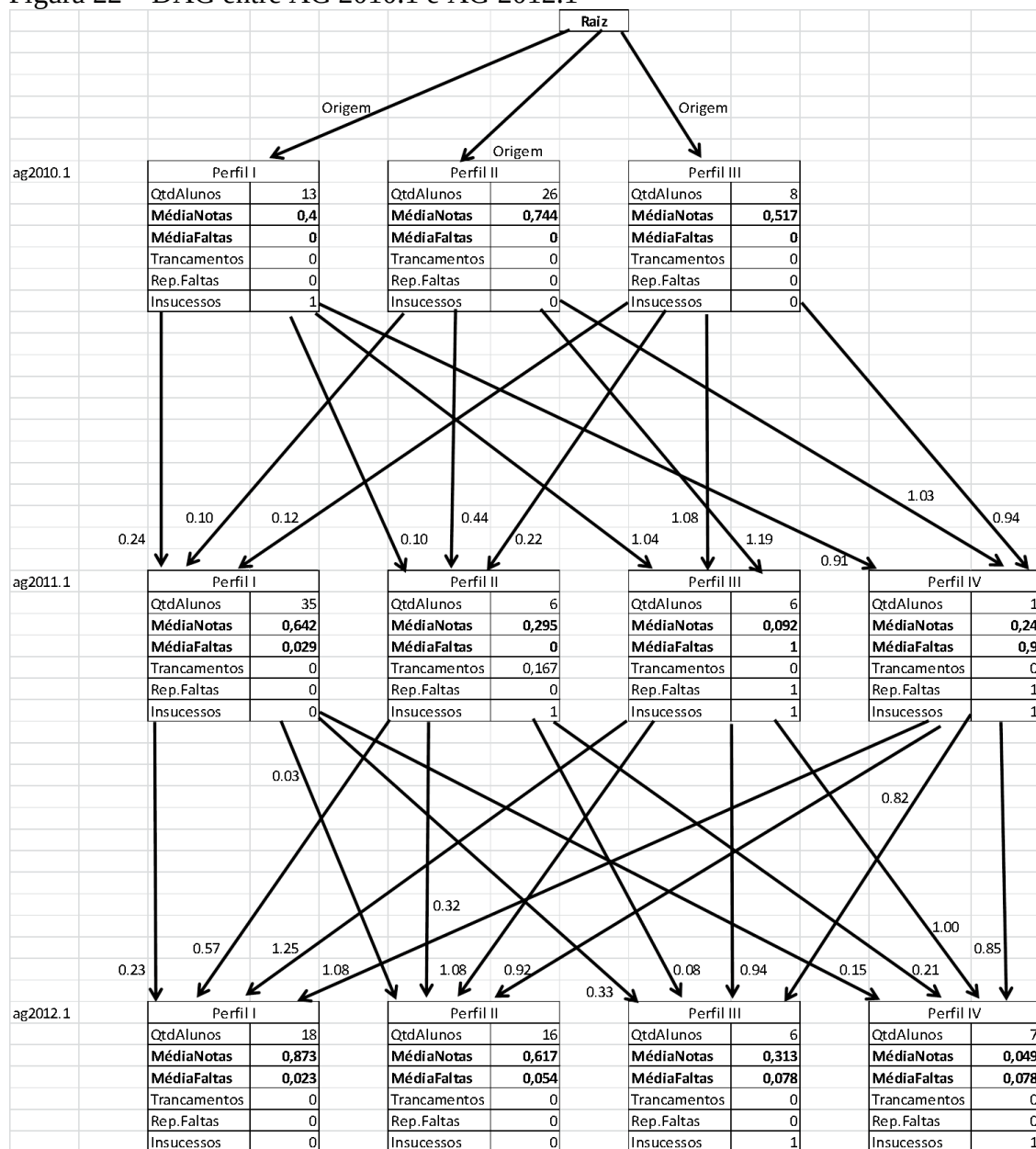
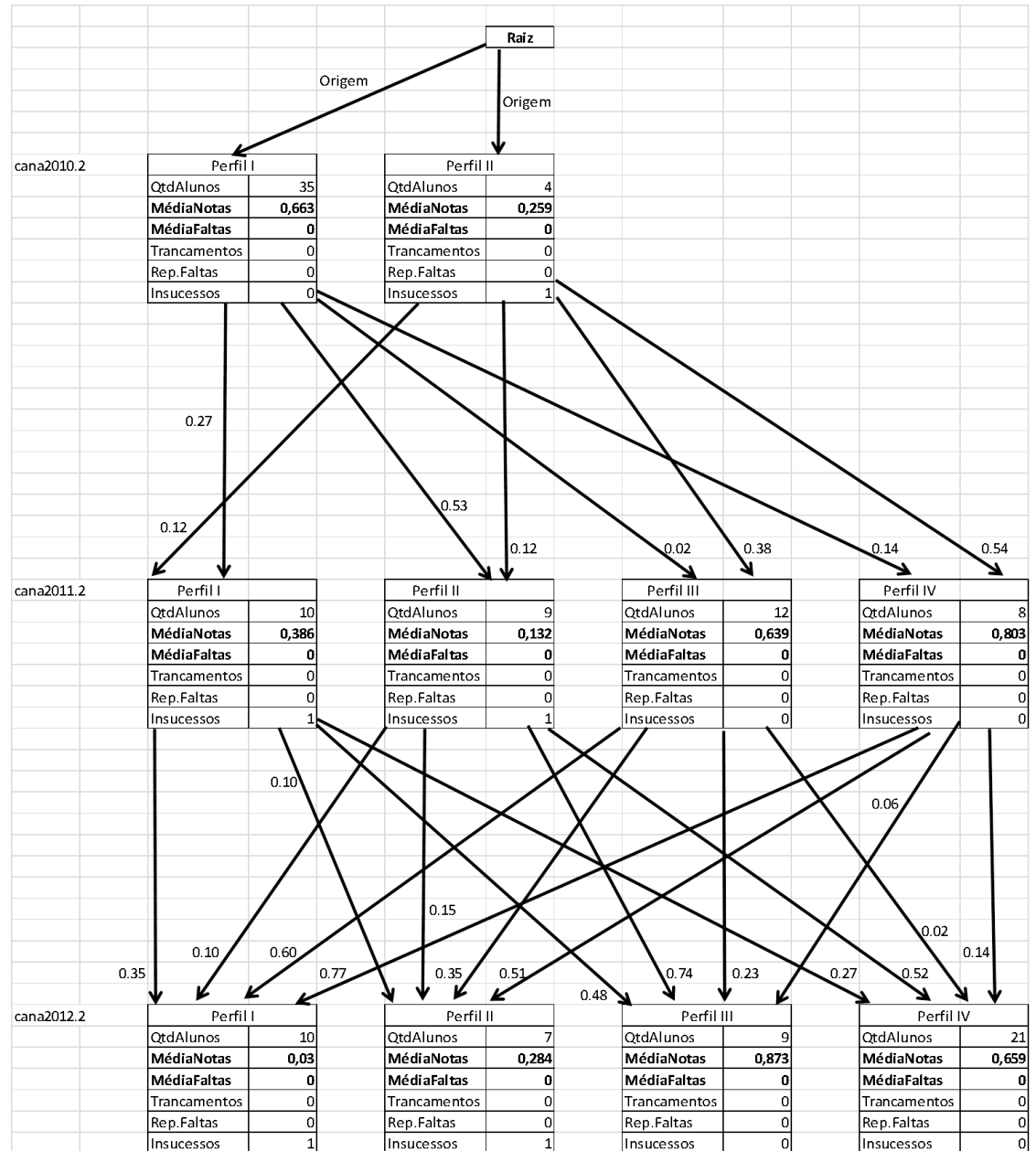


Figura 23 – DAG entre CANA 2010.2 e CANA 2012.2



APÊNDICE C – DAGs relacionados às análises do impacto da nova regra de frequência de 2008.2 nos perfis de alunos entre 2006.2 até 2010.2

Esses DAGs os vértices são os perfis de alunos por disciplina e as arestas são os valores das distâncias euclidianas calculadas entres os perfis.

Figura 24 – DAG entre FUP 2007.1 e FUP 2010.1

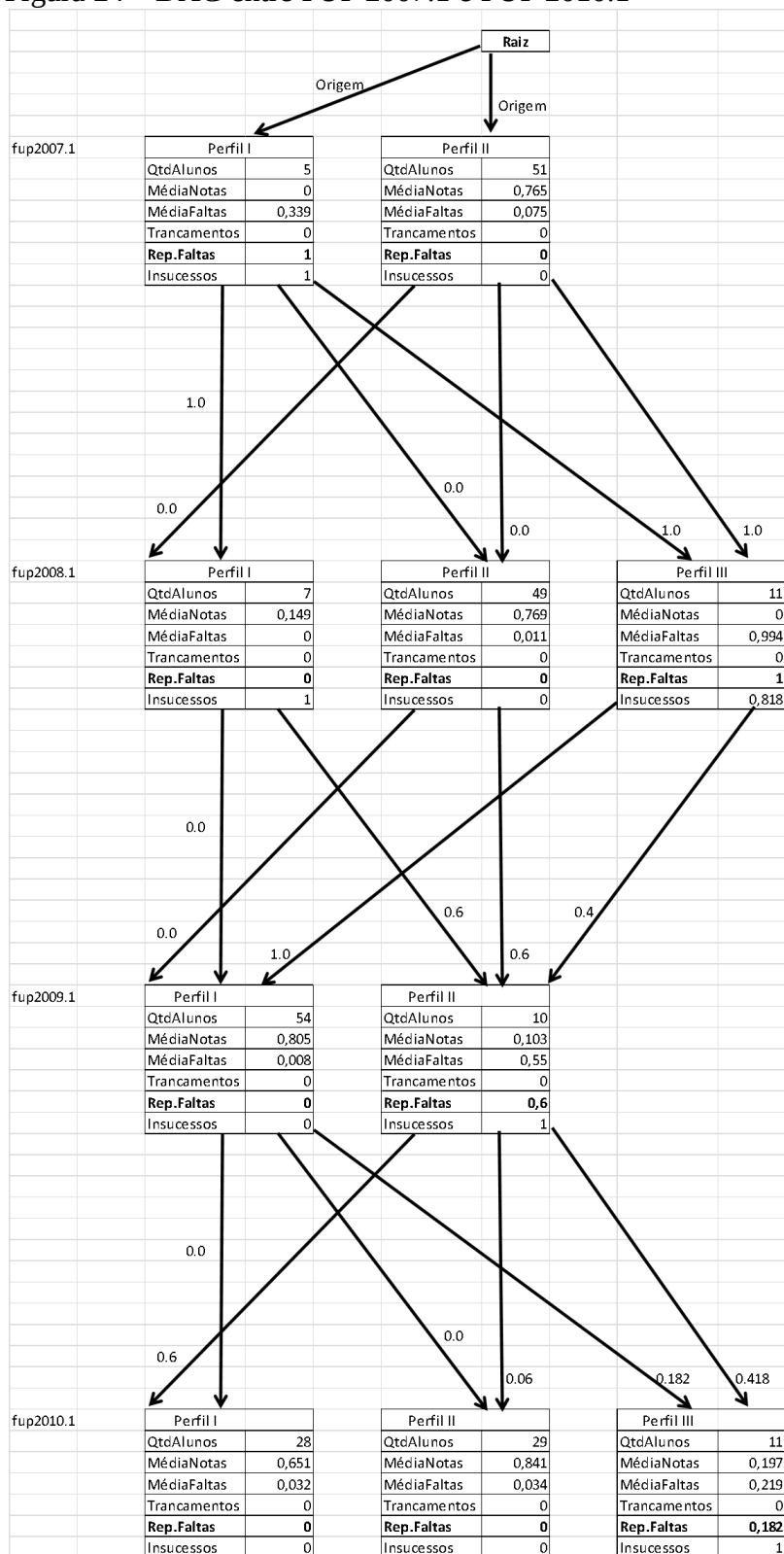


Figura 25 – DAG entre ED 2006.2 e ED 2010.2

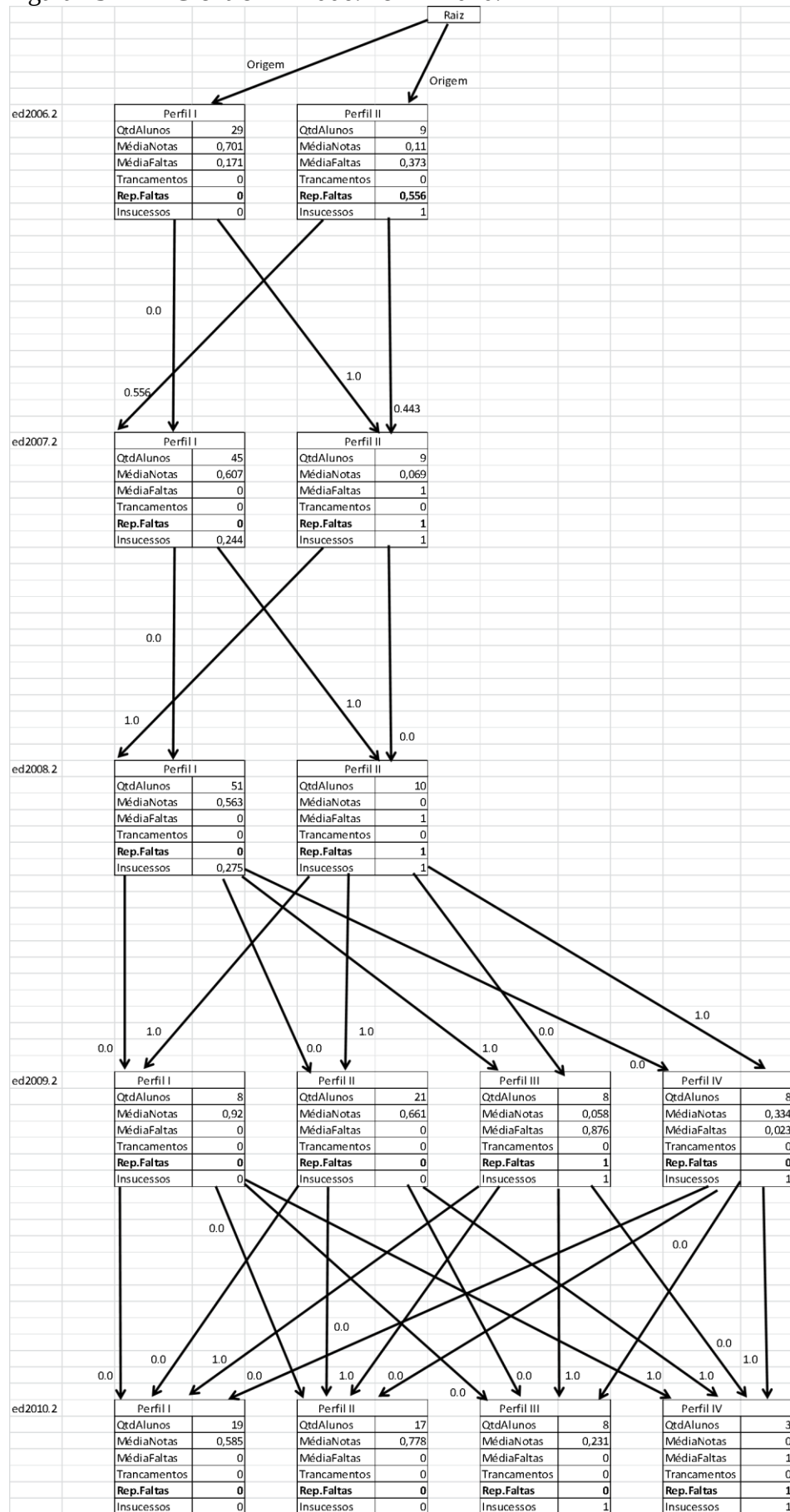


Figura 26 – DAG entre AG 2007.1 e AG 2010.1

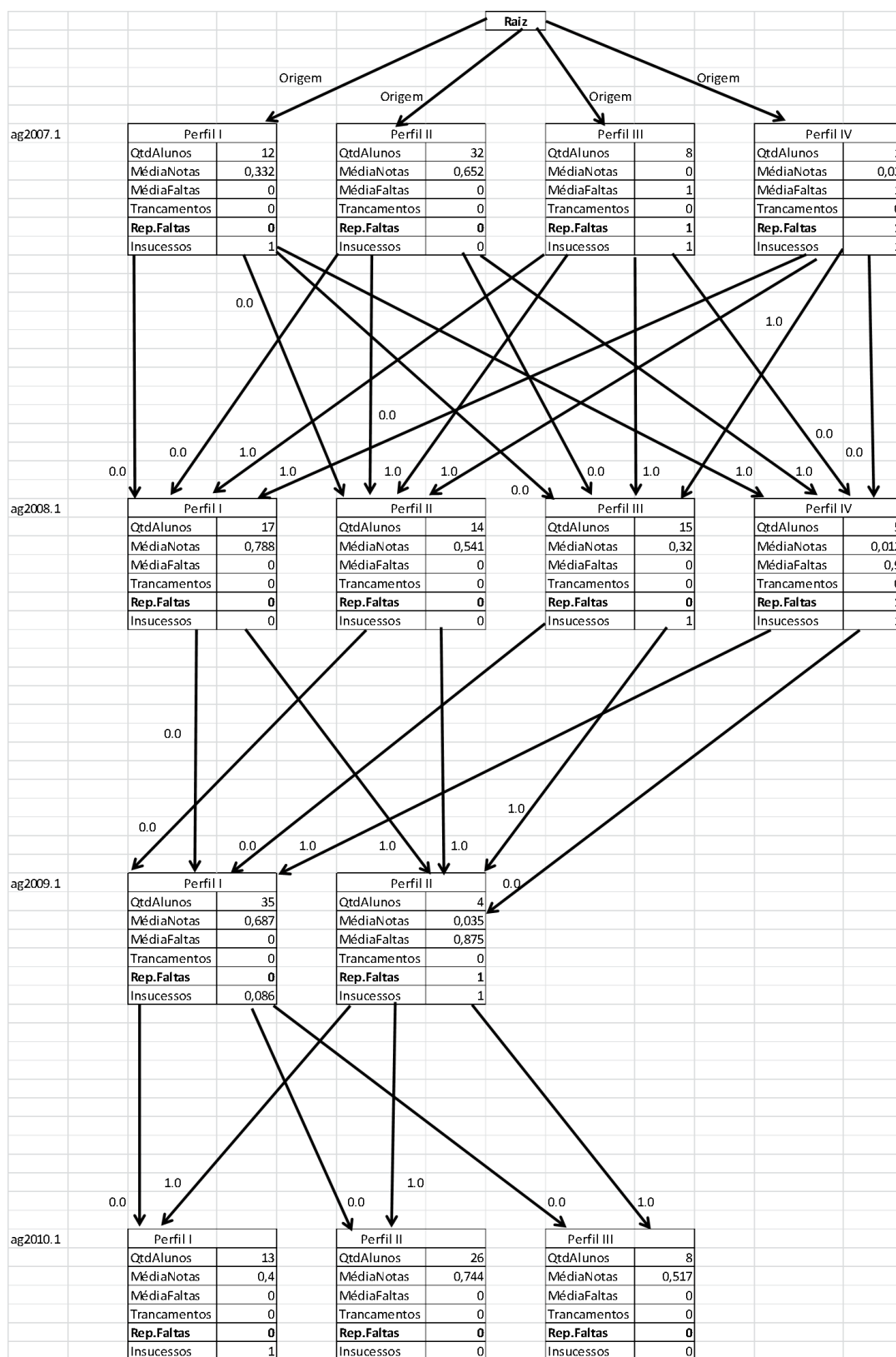
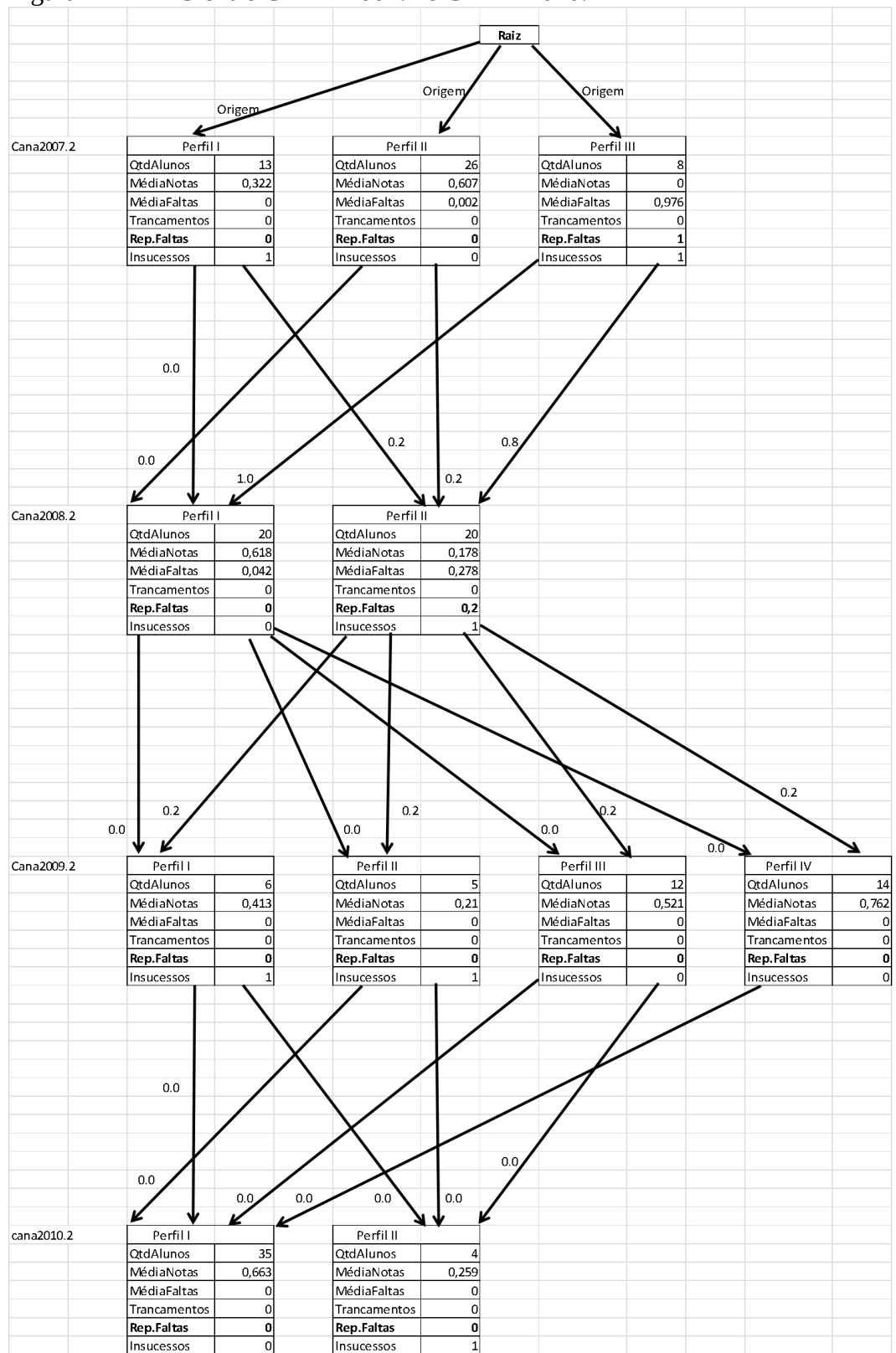


Figura 27 – DAG entre CANA 2007.2 e CANA 2010.2



APÊNDICE D – DAGs relacionados às descobertas de perfis de alunos por disciplina entre 2006 até 2012

Esses DAGs os vértices são os perfis de alunos por disciplina e as arestas são os valores das distâncias euclidianas calculadas entres os perfis.

Figura 28 – DAG entre FUP 2006.1 e FUP 2009.1

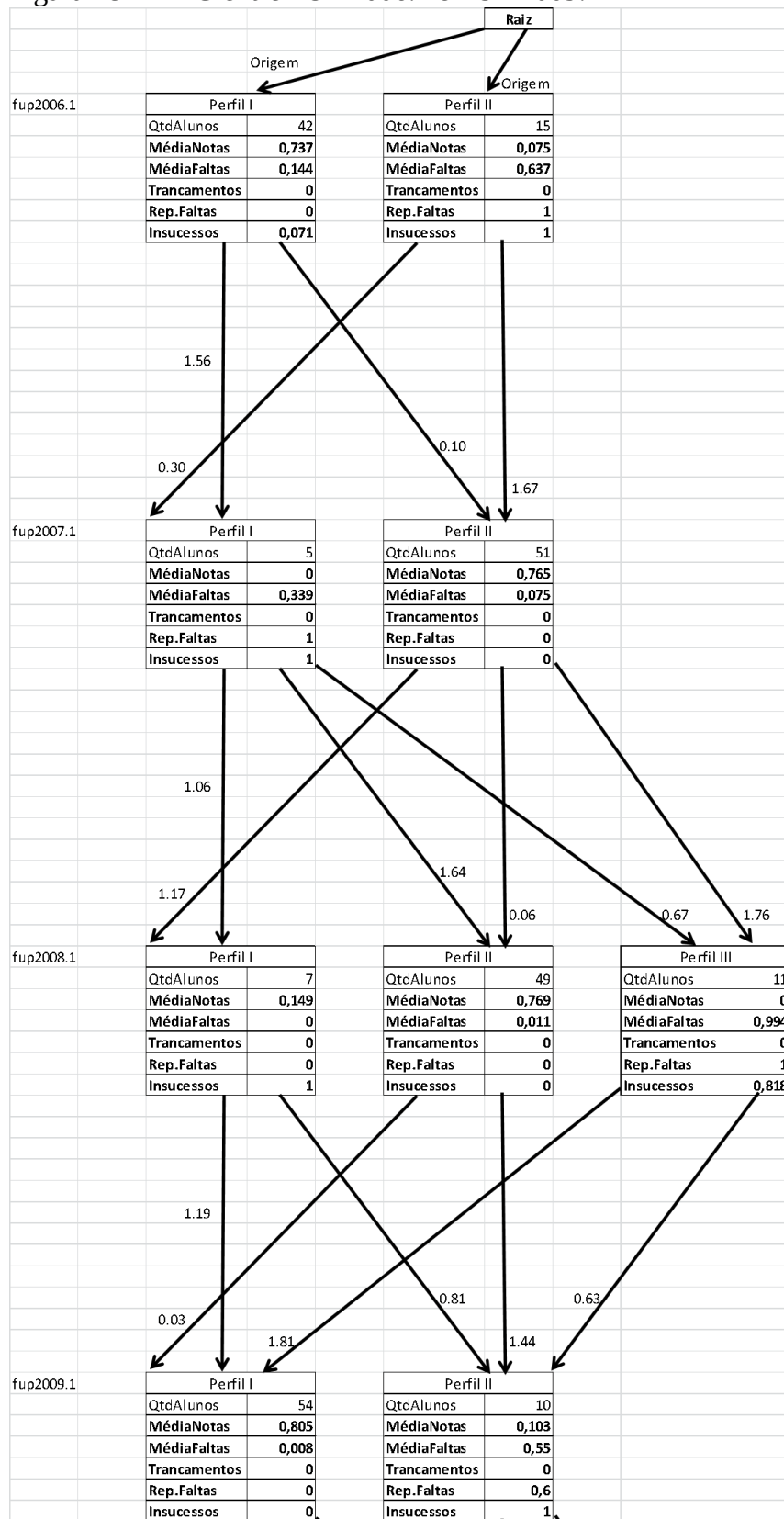


Figura 29 – DAG entre FUP 2009.1 e FUP 2012.1

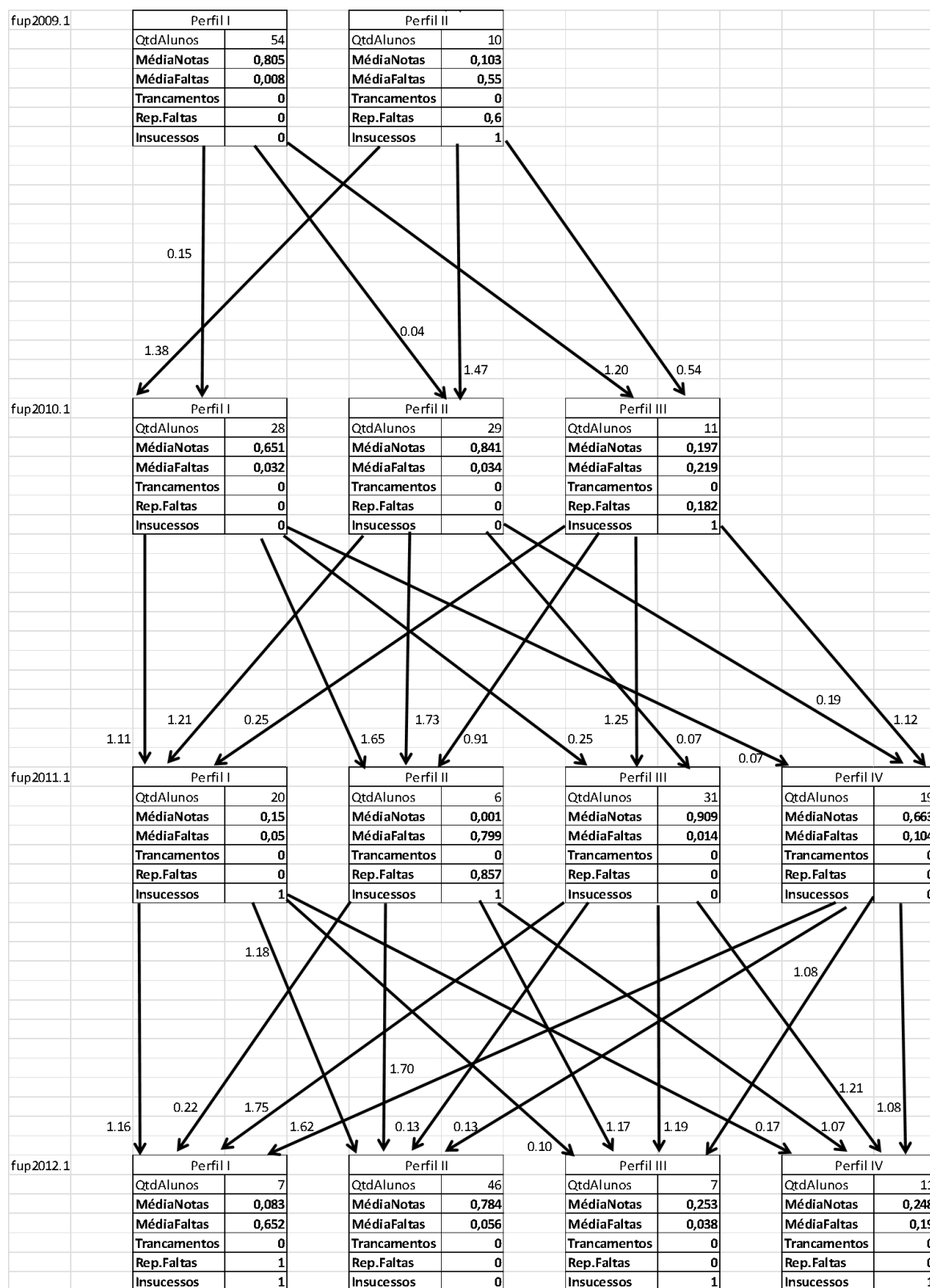


Figura 30 – DAG entre ED 2006.2 e ED 2009.2

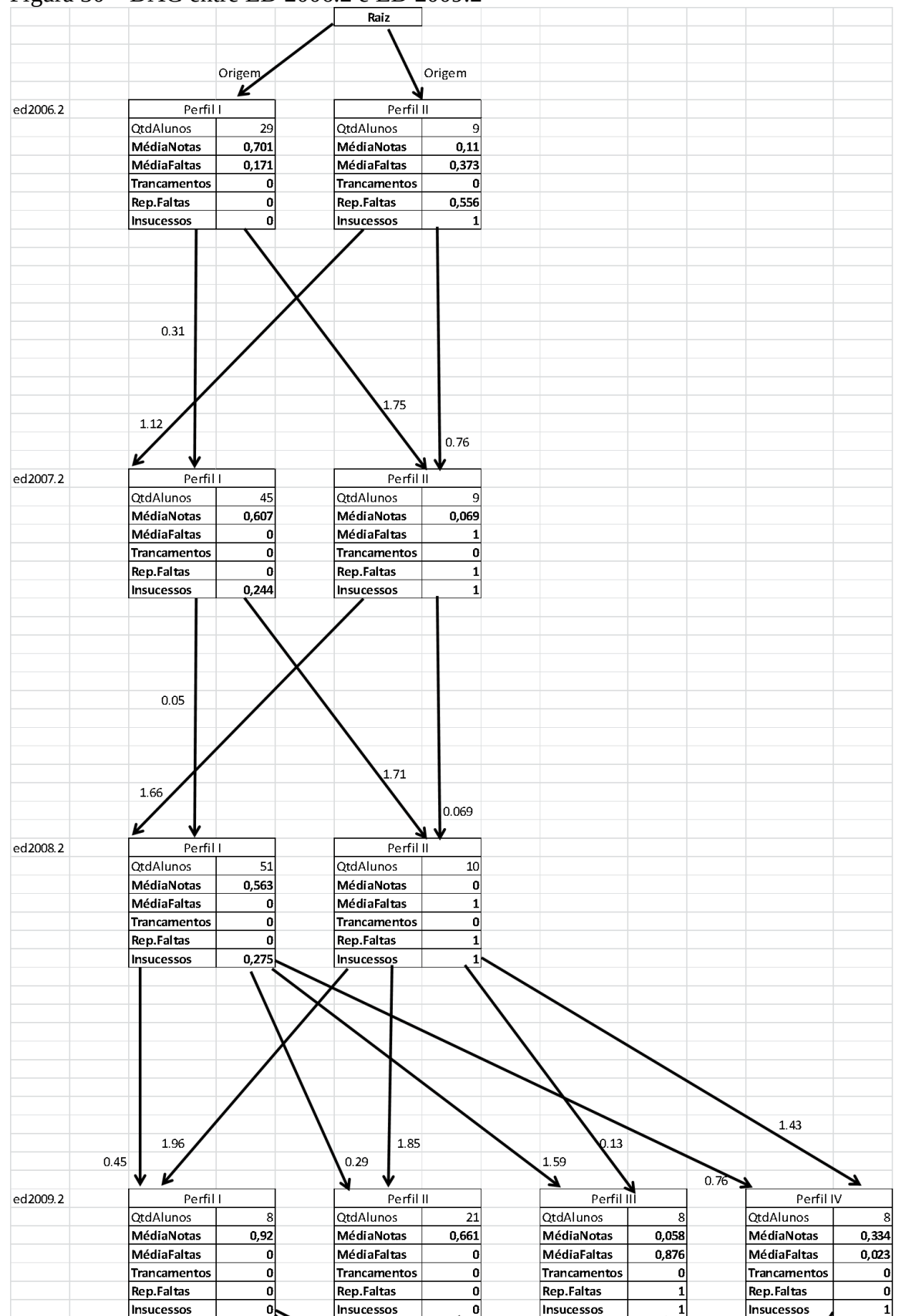


Figura 31 – DAG entre ED 2009.2 e ED 2012.2

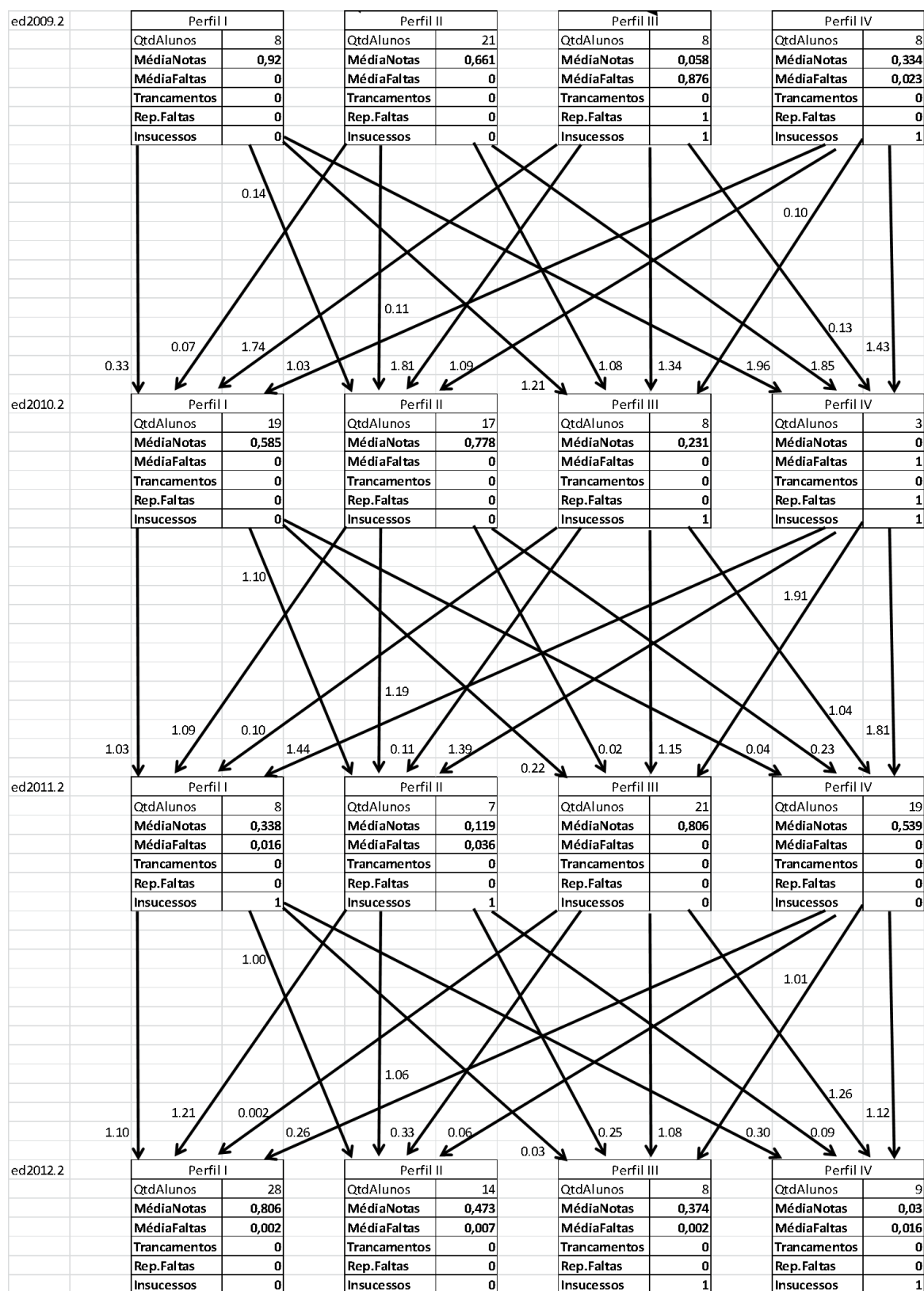


Figura 32 – DAG entre AG 2007.1 e AG 2009.1

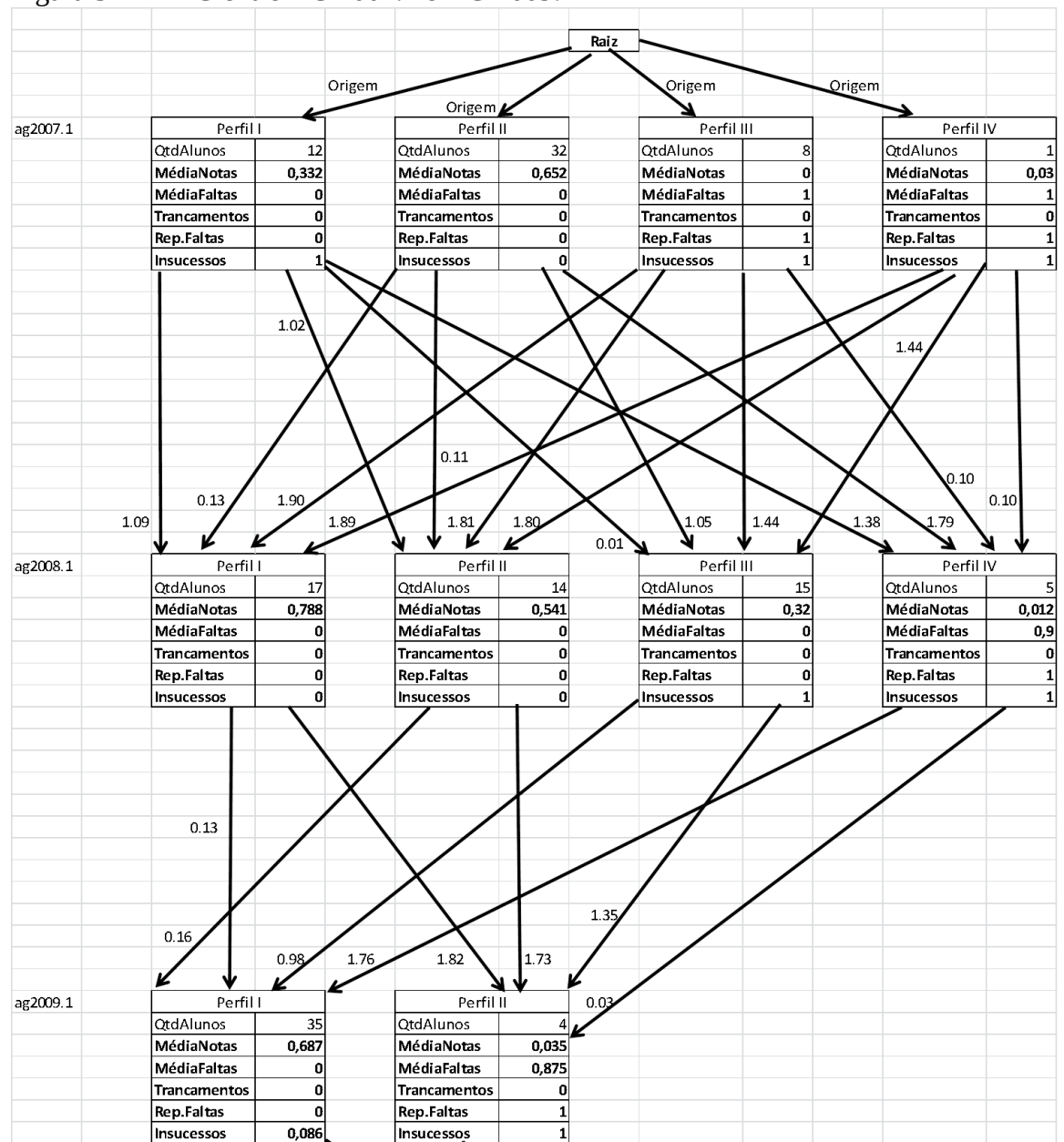


Figura 33 – DAG entre AG 2009.1 e AG 2012.1

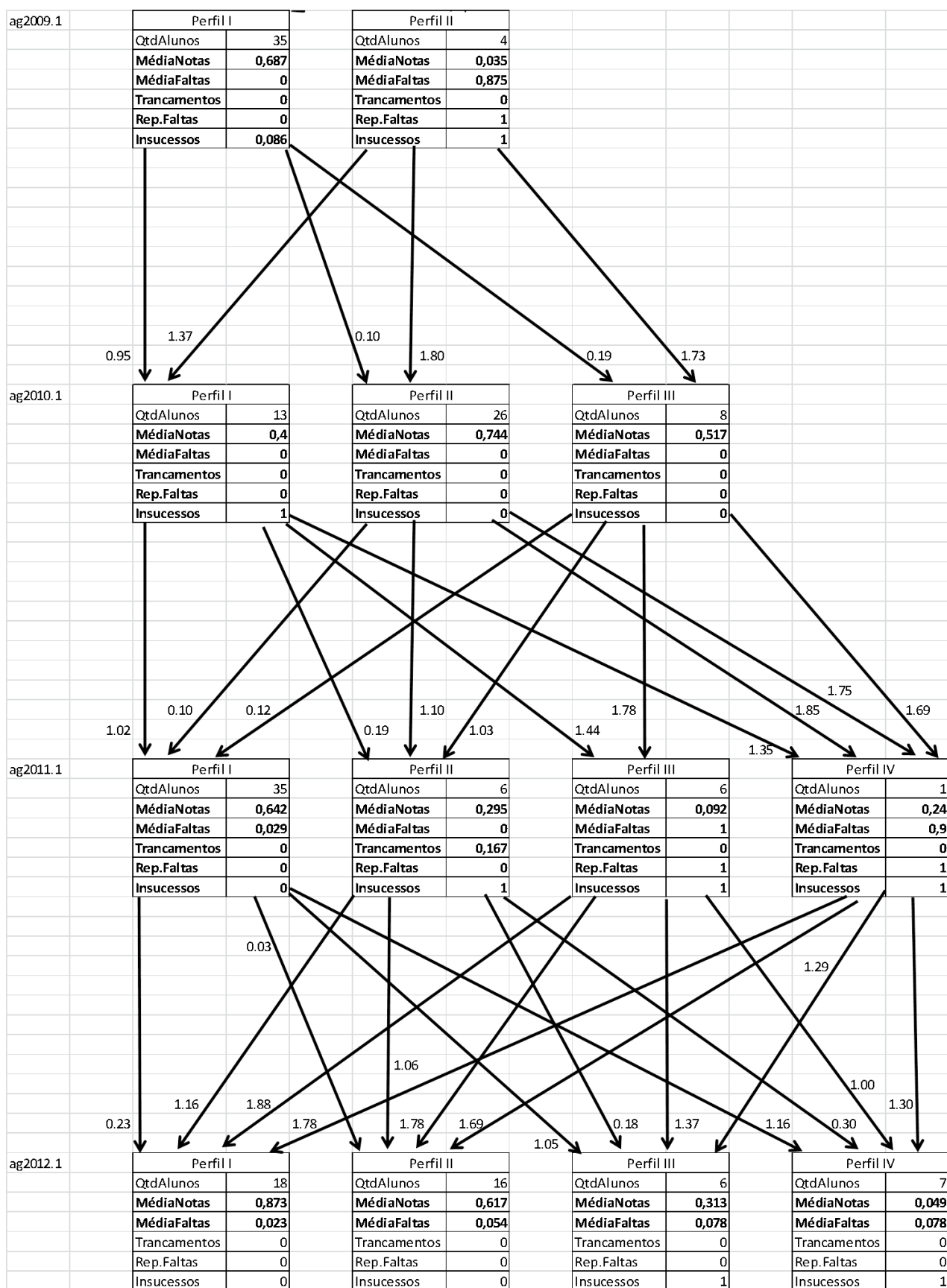


Figura 34 – DAG entre CANA 2007.2 e CANA 2009.2

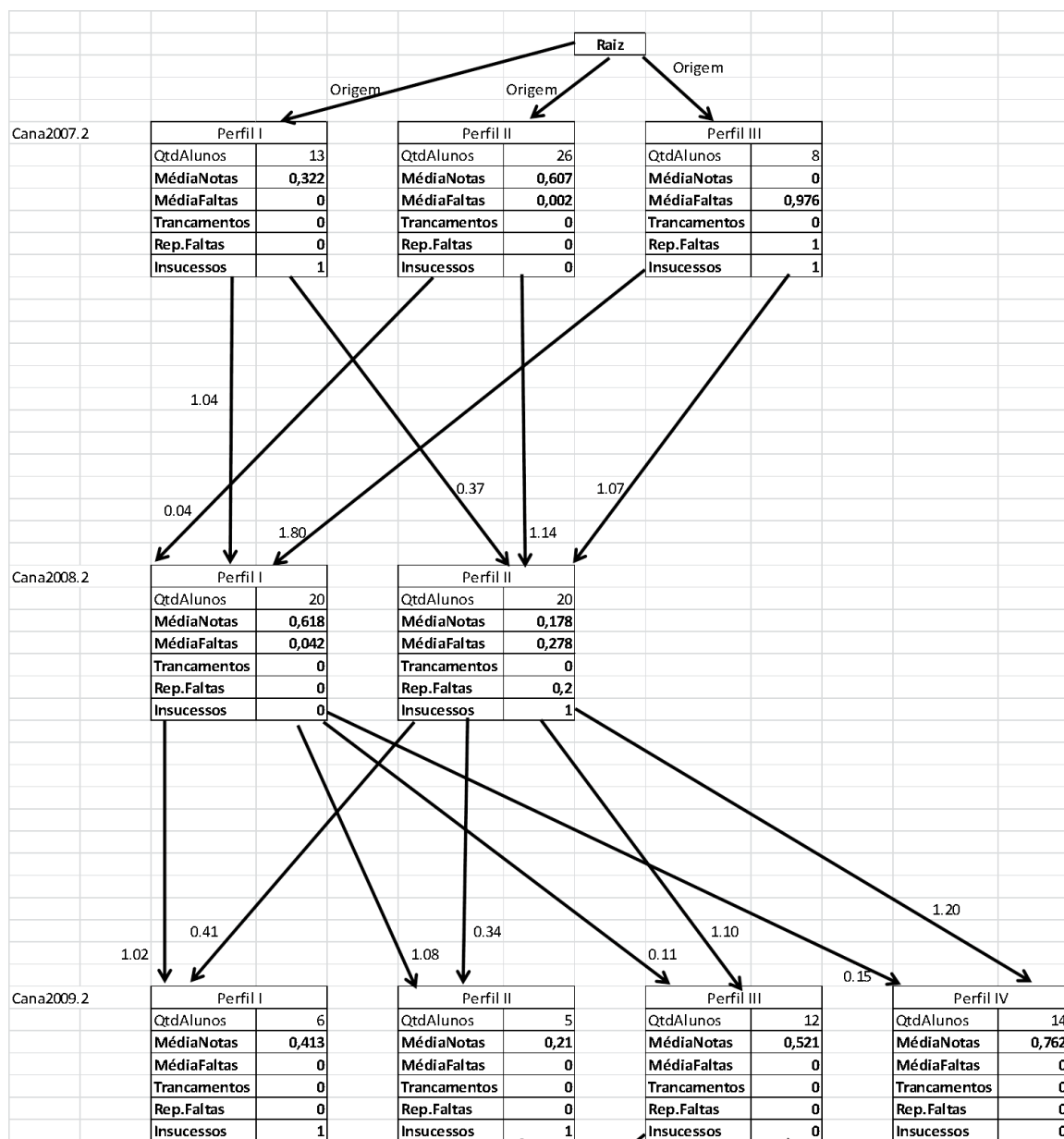


Figura 35 – DAG entre CANA 2009.2 e CANA 2012.2

