Ana Luiza de Abreu Esteves

Empregado, estudante ou desempregado? Um estudo sobre agrupamento de egressos de cursos técnicos

Monografia de Pós-Graduação Lato Sensu

Monografia apresentada como requisito parcial para conclusão do curso de Sistemas Inteligentes de Apoio à Decisão em Negócios / Business Intelligence Master pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: André Vargas Abs da Cruz

Rio de janeiro, setembro de 2011

Dedicatória

Ao meu pai amado João Luiz da Cruz Esteves que lá do céu está muito orgulhoso por mais esta nossa conquista.

Agradecimentos

Ao meu futuro marido Luiz Gustavo Fernandes Ramos pela colaboração no desenvolvimento desta monografia.

À minha mãe Maria Luiza de Abreu Esteves pelo incentivo para realização do curso.

Ao amigo Leandro Magalhães Gonçalves pela ajuda ao longo do curso.

Ao professor André Vargas Abs da Cruz pela orientação desta monografia.

À minha estagiária Adriana Carolina Carbonel Velarde pelo esforço em prol desta monografia.

Ao Sistema FIRJAN pela oportunidade concedida.

Resumo

Numa sociedade marcada pelas crescentes exigências do mundo corporativo que, por conseqüência, levam a alterações no mercado de trabalho, fica evidente a necessidade de transformações aceleradas, que pedem práticas pedagógicas flexíveis capazes de promover sintonia entre egressos e o mundo do trabalho.

É nesse contexto que atuam os cursos técnicos do SENAI-RJ[1], que objetivam a capacitação de recursos humanos na área industrial, favorecendo a elevação da competitividade empresarial do estado do Rio de Janeiro.

Para mensurar a qualidade dos cursos técnicos é realizada a Pesquisa de Impacto dos Cursos Técnicos do SENAI-RJ. No entanto, a análise dos resultados, atualmente, é simplória e pode conter informações importantes que devido à complexidade, ainda não foram extraídas. Com isso advém a necessidade de analisá-las automaticamente, extraindo informação útil que pode agregar conhecimento.

Esta monografia investigará a situação profissional dos egressos do curso técnico do SENAI-RJ mediante análise de agrupamentos (*cluster* de dados ou *clusterização*), que procura encontrar grupos de dados semelhantes entre si, revelando como os dados estão estruturados e resultando em um melhor entendimento sobre a situação dos egressos: empregados, estudantes ou desempregados.

[1] Serviço Nacional de Aprendizagem Industrial.

Abstract

In a society determined by the increased demands of the corporate world, which, consequently cause changes in the labor market, it´s palpable the need of accelerated changes that requires flexible pedagogical practices capable of promoting the correct tune of graduate ones and the labor market.

Is, in that context, that SENAI-RJ technical courses operate by qualifying human resources in the industrial area and favoring the elevation of corporate competitiveness at Rio de Janeiro.

To assure the quality of its technical courses, SENAI-RJ organizes the SENAI-RJ Technical Courses Impact Survey. However, the results analysis is quite simple and lack of some important information, that because of their complexity, weren´t yet extracted.

This work will explore the professional status of SENAI-RJ technical courses graduates by grouping analysis (data cluster) which intends to find similar data groups, revealing how data is structured and resulting in a better understanding of the subject.

Sumário

Lista de Figuras

Lista de Tabelas

1
Introdução

1.1
Motivação

O Sistema FIRJAN[2] (Federação das Indústrias do estado do Rio de Janeiro) é composto por cinco organizações, sendo uma delas o SENAI-RJ que promove a capacitação tecnológica por meio de formação profissional, qualificação e especialização de trabalhadores e da sociedade.

Há mais de dez anos, o SENAI-RJ realiza anualmente a Pesquisa de Impacto dos Cursos Técnicos com o intuito de prover dados que norteiem o constante aperfeiçoamento dos mesmos. Tais informações estão relacionadas ao perfil dos egressos e suas trajetórias profissionais antes e depois do curso, avaliando as conquistas após a participação no curso, as expectativas que foram ou não atendidas e os motivos para tal.

Por razões metodológicas, a Pesquisa de Impacto dos Cursos Técnicos realiza-se cerca de um ano após os alunos terem concluído seus cursos. O planejado lapso de tempo é proposital e consiste no período de maturação considerado necessário para a efetiva inserção do egresso no mercado de trabalho, tornando possível avaliar a contribuição do curso realizado em suas vidas profissionais. A metodologia da pesquisa envolve entrevista telefônica mediante questionário estruturado. Posteriormente o relatório é enviado à equipe de educação profissional do Sistema FIRJAN, que engloba todo o estado do Rio de Janeiro.

A partir do universo de 2.272 concluintes de 2009 foi definida uma amostra de 461 alunos, ou seja, foram entrevistados 20,3% do total de egressos. Com a amostragem no lugar do censo, segundo Aaker, pode-se dedicar mais atenção a cada entrevista, aumentando a qualidade da resposta. Além disso, para o plano

[2] Visão “ser reconhecido pela sociedade em 2014 como uma organização privada prestadora de serviços, indispensável ao desenvolvimento sustentável do estado do Rio de Janeiro” e missão “promover a competitividade empresarial, a educação e a qualidade de vida do trabalhador e da sociedade, contribuindo para o desenvolvimento sustentável do estado do Rio de Janeiro”.

amostral houve a preocupação de criar uma amostra proporcional representativa da população dos concluintes dos cursos técnicos do SENAI-RJ.

Devido aos tamanhos do universo e da amostra, eliminou-se a hipótese de encontrar resultados com significativas distorções da realidade, uma vez que, tendo em vista um intervalo de 95,0% de confiança, foi obtida margem de erro[3] de 4,1%. Assim, pôde-se afirmar com 95,0% de segurança que os resultados mostrados na pesquisa refletiam a opinião e percepção dos alunos, variando num intervalo de 4,1% para menos a 4,1% para mais.

Atualmente, a análise da Pesquisa de Impacto dos Cursos Técnicos envolve:

1) Frequência de todos os resultados – informações individuais das frequências relativas de todas as variáveis da pesquisa divulgadas na forma de percentagem;
2) Cruzamentos de algumas variáveis juntamente com teste estatístico Qui-Quadrado para medir independência das mesmas – exposição do comportamento simultâneo de duas variáveis na forma de percentagem e descoberta de relação de independência ou dependência entre tais;
3) Elaboração de duas taxas (taxa de inserção e taxa de empregabilidade) – taxa de inserção representa a priori os egressos que efetivamente conseguiram emprego após o curso, ou seja, inserção apenas dos alunos que se declararam desempregados ou estudantes antes do

[3] De acordo com Cochran, para uma população finita (abaixo de 100.000 objetos), utiliza-se a seguinte fórmula para calcular a margem de erro em uma amostra:

$$\varepsilon = z\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$$

Onde:

ε - margem de erro

z - grau de confiança - valor tabelado da distribuição N(0,1)

p - proporção amostral - valor estimado a partir da amostra

n - tamanho da amostra

N - tamanho da população

Para a determinação da margem de erro, ou seja, do erro máximo desejado/aceitado pelo pesquisador é preciso fixar o grau de confiança e possuir algum conhecimento a priori da variabilidade da população. O primeiro é escolhido pelo pesquisador, já o segundo, requer o conhecimento de pesquisas passadas. No entanto, com a ausência de uma amostra piloto, estima-se que p =0,5, pois qualquer que seja o tamanho da amostra, o erro máximo ocorre quando taxamos esta proporção amostral.

curso com sua situação no mercado de trabalho após o curso e taxa de empregabilidade é a diferença entre a proporção de egressos ocupados antes e depois da conclusão do curso;

4) Teste estatístico de correlação de Kendall da cauda superior – detecta associação positiva entre duas variáveis ordenáveis segundo algum critério, neste caso, o ano;
5) Informações extraídas da Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE) – comparação de dados entre a população brasileira e os egressos do SENAI-RJ.

Desta forma, percebe-se, em linhas gerais, que a atual análise da pesquisa não permite maior conhecimento sobre a situação profissional dos egressos do SENAI-RJ após a realização do curso técnico. O estudo apresenta investigação meramente descritiva do banco de dados sem tentar entender a relação comportamental das variáveis.

A idéia dessa compreensão abrangente possibilitaria a evolução por parte dos programas de educação do SENAI-RJ para que possam atingir e manter a excelência, atendendo e talvez até superando constantemente às expectativas do mercado contratante.

Para as lacunas existentes na atual análise da pesquisa, considerou uma ferramenta capaz de formar grupos, que permita levantar hipóteses sobre o relacionamento de todas as variáveis. Desta forma, a opção pela análise de agrupamentos permitirá o enriquecimento dos resultados e facilitará as tomadas de decisão quanto às diretrizes pedagógicas dos cursos técnicos do SENAI-RJ.

As técnicas de análise de agrupamentos estão ganhando cada vez mais mercado e se mostram bastante interessantes, pois revelam como os dados estão estruturados e possibilitam melhor entendimento sobre o negócio. Assim pode ser possível encontrar grupos que identifiquem diferentes tipos de egressos/clientes do SENAI-RJ, permitindo, dessa forma, que a área de educação do Sistema FIRJAN trabalhe de maneira diferenciada para cada grupo de egresso/cliente, de acordo com as características intrínsecas do grupo em questão.

Como definição de Jain, análise de agrupamentos é a classificação não-supervisionada de dados, formando agrupamentos ou *clusters*: representa uma das principais etapas de processos de análise de dados. A idéia é agrupar um conjunto de padrões em grupos que possuam algum significado, ou seja, de tal modo que os padrões apresentem alguma propriedade comum. Neste sentido, esta análise pode fornecer novas hipóteses a respeito dos interrelacionamentos dos dados e de sua estrutura.

O processo de agrupamento pode ser dividido basicamente em três etapas: seleção e tratamento de dados, agrupamento de dados e análise dos resultados. Os principais itens que devem ser considerados nesse processo são (Berkhin):

1) Tipo de atributos que o algoritmo opera;
2) Escalabilidade para grandes conjuntos de dados;
3) Habilidade de operar com uma dimensão grande de variáveis;
4) Habilidade de encontrar agrupamentos de forma irregular;
5) Tratar valores discrepantes (*outliers*);
6) Tempo de execução;
7) Dependência de ordem dos dados;
8) Classificação;
9) Segurança no conhecimento a priori e parâmetros definidos pelos usuários (coleta de definições formais dos termos envolvidos);
10) Interpretabilidade dos resultados.

1.2
Objetivos

Deste modo, esta monografia tem como objetivos principais:

- Estudo do processo de agrupamento de dados – as técnicas de agrupamentos de dados estão sendo cada vez mais pesquisadas e utilizadas, principalmente nas grandes empresas onde é importante a aquisição de informações estratégicas (Vale);
- Análise de agrupamentos nos dados da Pesquisa de Impacto dos Cursos Técnicos do SENAI-RJ para servir de instrumento de informação exploratória para a equipe de educação profissional do

Sistema FIRJAN no sentido de colaborar com a descrição de pontos relevantes acerca da situação profissional do egresso do SENAI-RJ, podendo inclusive ser utilizado para ajudar na avaliação dos cursos técnicos garantindo constante melhoria;

- Estudar e descrever o perfil e o comportamento de cada grupo criado pela análise de agrupamentos.

1.3
Descrição do Trabalho

1.3.1
Análise de Agrupamentos

O algoritmo deve criar grupos (classes) através da produção de um particionamento do banco de dados em um conjunto de objetos. Um bom agrupamento caracteriza-se pela produção de grupos de alta qualidade, onde a similaridade intraclasse é alta e a interclasse é baixa. A qualidade do resultado do particionamento também depende da medida utilizada para aferir a similaridade entre os objetos e entre os agrupamentos.

A análise de agrupamentos permite ainda identificar registros com valores aberrantes, formular hipóteses de relacionamento dos dados e estudar sua dimensionalidade.

Os métodos mais utilizados para agrupar objetos estão divididos em duas categorias principais:

- Hierárquicos – criam uma (de)composição hierárquica do conjunto de dados segundo algum método;
- Não-hierárquicos – partem de um agrupamento inicial dos dados e procuram melhorá-los através da otimização de algum critério. Estão nesta categoria os algoritmos *k-means*, *Fuzzy C-means* e *Kohonen*.

1.3.1.1
Seleção da Amostra para Trabalho

Em muitos casos, segundo Azevedo, a base de dados a ser analisada é muito extensa, tornando inviável a utilização de certos algoritmos de agrupamento. Nessas situações, por uma questão de economia, é comum extrair da base de dados uma ou mais amostras menores para se trabalhar. Após esta decisão, é necessário garantir que todos os elementos da base tenham a mesma probabilidade de serem escolhidos e que as amostras extraídas sejam representativas do total da base de dados.

1.3.1.2
Pré-processamento dos Dados

Dada a sua complexidade do problema, esta etapa do trabalho é normalmente dividida em três fases menores, descritas a seguir.

1.3.1.2.1
Identificação de Inconsistências e de Valores Aberrantes (limpeza dos dados)

A limpeza dos dados envolve a verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de registros contendo valores nulos, redundantes ou aberrantes (*outliers*). Tal etapa tem o intuito de corrigir a base de dados de maneira a eliminar registros que possam comprometer a qualidade dos resultados e evitar consultas desnecessárias pelo algoritmo de agrupamento, comprometendo o desempenho.

1.3.1.2.2

Seleção dos Atributos e Redução de Dimensionalidade (seleção dos dados)

Esta fase tem o propósito de escolher um subconjunto dos atributos da base de dados que seja mais relevante ao problema. Este subconjunto pode conter os próprios atributos ou combinações dos mesmos, e será posteriormente fornecido ao algoritmo de agrupamento.

Uma das principais motivações desta fase é otimizar o tempo de processamento do algoritmo, visto que precisará trabalhar com apenas um subconjunto do total de atributos. Além de contribuir para o aumento da interpretabilidade dos dados, já que é normalmente mais fácil interpretar a base utilizando menos atributos, especialmente se eles não estiverem correlacionados.

Existem duas principais categorias de métodos para a seleção de atributos: os que utilizam o algoritmo de agrupamento no processo de seleção e os que são independentes do mesmo.

Nos métodos que pertencem a primeira categoria, os dados a serem agrupados são normalmente (embora não necessariamente) divididos em dois subconjuntos: treinamento e avaliação. Os métodos mais utilizados são: *Forward Selection* (FS) e *Backward Elimination* (BE).

Já os métodos da segunda categoria são independentes do algoritmo e um mesmo conjunto de atributos (ou combinação de atributos) é fornecido para qualquer algoritmo de agrupamento. Pertencem a essa categoria métodos tradicionais como Análise de Fatores (Anderson), Análise de Componentes Principais e Seleção por Entropia.

Muitas vezes os atributos são selecionados também com base na experiência ou bom senso do analista de dados.

1.3.1.2.3
Codificação e Transformação dos Atributos

Normalmente, visando melhorar o desempenho do processo, são feitas algumas codificações e transformações sobre os atributos da base de dados. Embora conceitualmente simples, esta tarefa demanda habilidade, exigindo grande experiência do analista de dados e conhecimento sobre o problema.

Existem diversas técnicas de codificações. Um algoritmo de codificação tradicional divide os valores contínuos dos atributos (inteiros ou reais) numa lista de intervalos representados por um código: efetivamente converte valores numéricos em valores categóricos, ou seja, cada intervalo resulta num valor discreto do atributo.

Além das codificações, transformações (Weiss) e (Johnson) sobre os atributos podem também vir a ser bastante úteis. É muito comum normalizar atributos, extraindo o logaritmo ou reduzindo o domínio para o intervalo $[0,1]$.

Há inúmeras vantagens em se codificar e transformar atributos: melhorar a compreensão do conhecimento descoberto; reduzir o tempo do processamento para o algoritmo de agrupamento, diminuindo o seu espaço de busca; facilitar o algoritmo a tomar decisões globais, já que os valores dos atributos foram englobados em faixas, etc. No entanto, como desvantagens há perda de detalhes dos dados que poderiam ou não vir a ser relevantes posteriormente.

1.3.1.3
Interpretação e Validação dos Resultados (pós-processamento dos dados)

Essa fase envolve a interpretação e validação dos resultados obtidos. Eventualmente, é necessário ainda algum processamento desses resultados e frequentemente tem que se decidir entre diferentes modelos propostos para os dados.

1.4
Organização da Monografia

Esta monografia está organizada em cinco capítulos, descritos a seguir:

O capítulo 2 descreve o problema a ser tratado, pontuando todas as dificuldades atuais.

O capítulo 3 detalha alguns conceitos fundamentais em análise de agrupamentos, descrevendo todo o processo basicamente dividido em três etapas: seleção e tratamento de dados, agrupamento de dados e análise dos resultados.

O capítulo 4 apresenta os resultados obtidos com a análise de agrupamentos na base de dados da Pesquisa de Impacto dos Cursos Técnicos do SENAI-RJ, justificando as opções tomadas.

Finalmente, o capítulo 5 mostra a conclusão do trabalho.

2
Descrição do Problema

Como supracitado, a análise da Pesquisa de Impacto dos Cursos Técnicos envolve, atualmente, informações individuais meramente descritivas em relação às variáveis do banco de dados da pesquisa. Desta forma, a equipe de educação profissional do Sistema FIRJAN possui relatório analítico da situação dos egressos, porém com poucos elementos para significativas tomadas de decisão quanto aos rumos pedagógicos e de marketing em prol da excelência que a clientela (pessoas física e jurídica) espera da marca SENAI.

De maneira sucinta, as etapas da atual análise podem ser descritas a seguir:

Primeiramente, todas as 63 variáveis pertencentes à pesquisa são relatadas em percentuais. Para melhor entendimento, a pesquisa é composta por 50 variáveis categóricas nominais (34 fechadas única/uma opção de resposta e 16 abertas/texto) e 13 variáveis escalares. Tais dados indicam o perfil dos egressos como gênero, faixa etária, segmento industrial do curso realizado, situação ocupacional antes e depois do curso, renda, benefícios adquiridos com o curso, nível de satisfação com o SENAI-RJ, assuntos de empreendedorismo e continuação dos estudos.

Uma segunda etapa da análise é composta por cruzamentos de algumas variáveis juntamente com teste estatístico no intuito de inferir os resultados. Será que um determinado segmento do curso garante emprego após o curso? Será que há relação entre escola do SENAI-RJ e renda? Será que a área da atividade profissional é igual à do curso e influencia na situação ocupacional atual do egresso?

Dois conceitos (taxas de inserção e de empregabilidade) mensuram e comparam quais variáveis são maiores empregadoras dos egressos, ou seja, o que determina a inserção ou permanência dos concluintes no mercado de trabalho. Além de comparar ao longo dos anos, quais épocas tiveram melhores cenários empregatícios aos ex-alunos dos cursos técnicos do SENAI-RJ.

Para todas as comparações no histórico evolutivo da pesquisa, o teste estatístico de correlação de Kendall da cauda superior é usado para detectar associação positiva entre duas variáveis ordenáveis, isto é, descobre se duas variáveis são proporcionalmente crescentes ao longo dos anos.

Por fim, como dados secundários, informações da PNAD são usadas como parâmetro no intuito de avaliar a situação profissional dos egressos com a população brasileira em termos de renda, empregabilidade no mercado formal e grau de instrução.

Neste sentido, os atuais resultados não colaboram com uma compreensão global acerca da situação profissional dos egressos após os cursos técnicos do SENAI-RJ. Tal investigação meramente descritiva do banco de dados não esclarece quem procura o SENAI-RJ para estudar nos cursos técnicos e nem como vivem após a realização do curso.

A idéia da compreensão abrangente da pesquisa mediante construção de grupos com diferentes perfis da clientela possibilitaria melhorias dos programas de educação com foco certeiro a todos os tipos de clientes do SENAI-RJ. Além de colaborar, para a equipe de marketing, com propagandas direcionadas a cada público-alvo dos cursos técnicos.

3
Metodologia[4]

A organização dos objetos em cada *cluster* deve ser feita de forma que haja:

- Alta similaridade entre os objetos pertencentes a um mesmo cluster (similaridade intracluster), ou seja, minimizar a distância intragrupo;
- Baixa similaridade entre elementos que pertencem a *clusters* diferentes (similaridade *intercluster*), ou seja, maximizar a distância intergrupos.

Os algoritmos de agrupamento são capazes de segmentar o banco de dados e não são guiados por um atributo classe. Todos os atributos são tratados da mesma forma e os dados são segmentados em grupos que não estavam pré-definidos.

3.1
Processo de Agrupamentos

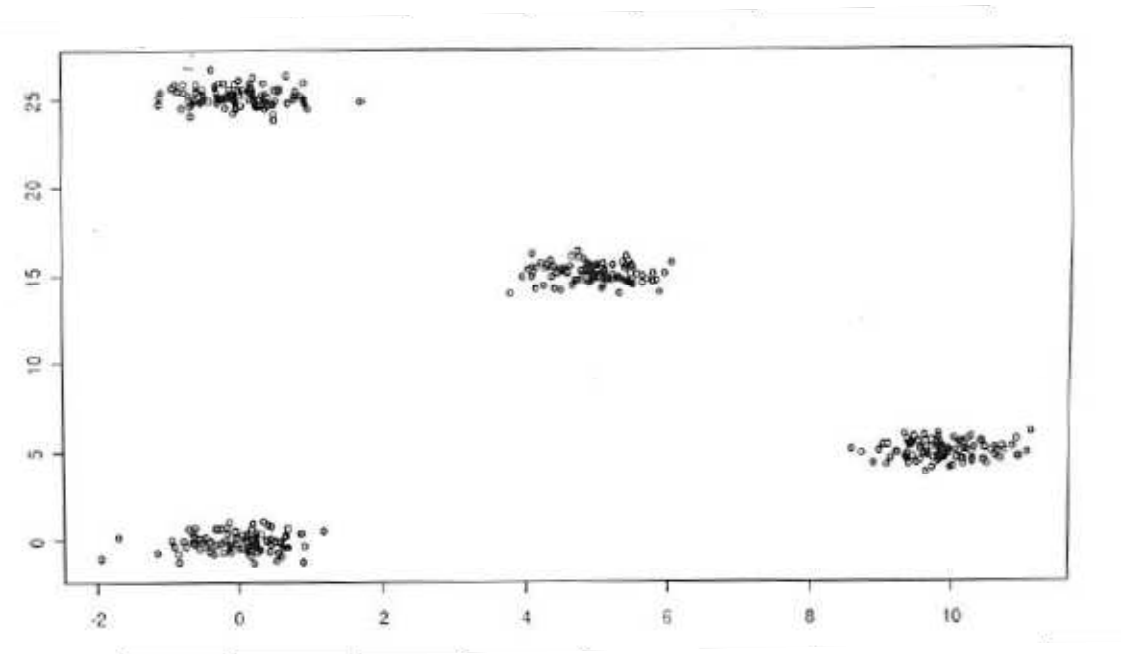

Figura 1: Gráfico ilustrativo de dados agrupados em quatro grupos.

[4] Este capítulo teve forte influência e inspiração da tese do Vale descrita nas referências bibliográficas.

A análise de agrupamentos pode ser definida como o processo de determinação de k grupos em um conjunto de dados.

Observa-se na distribuição acima a presença de quatro grupos distintos. O processo automático de descobrimento desses grupos é o principal objetivo da análise de agrupamentos, onde basicamente se buscam grupos de objetos (dados) similares entre si.

O processo de agrupamento de dados pode ser dividido em três etapas que serão abordadas nas próximas seções:

- Seleção e tratamento de dados;
- Agrupamento de dados;
- Análise de resultados.

3.1.1
Seleção e Tratamento de Dados

Na seleção dos dados, o objetivo é extrair do total de variáveis da base de dados apenas os atributos que possuam maior relevância ao processo de agrupamento de dados, eliminando atributos irrelevantes ou redundantes. Esse passo é importante para diminuir o tempo de processamento como também para evitar que seja prejudicado por atributos irrelevantes.

No tratamento dos dados, o objetivo é preparar esses dados de modo a assegurar sua qualidade e eficiência no processo de agrupamento. Os itens mais importantes para o tratamento dos dados são:

- Eliminação de dados duplicados ou corrompidos;
- Remoção de *outliers* (dados com valores inválidos significativamente fora do esperado para uma variável);
- Retirada de valores faltantes ou inválidos do conjunto selecionado;
- Transformação de dados – essa etapa pode ser subdividida em duas tarefas:
  - Tratamento de Atributos – adequar os diferentes tipos de atributos para o processo de agrupamento;

- Normalização – tratar dados com atributos de diferentes dimensões, quando de pretende que eles tenham a mesma influência no processo.

3.1.1.1
Tratamento de Atributos

O primeiro objetivo dessa etapa consiste em transformar os dados de maneira que seja possível realizar o agrupamento de dados de forma adequada. Uma base de dados pode conter dados numéricos ou categóricos, sendo necessário lidar adequadamente com cada um destes casos.

Os tipos de atributos são divididos em seis classes. Entretanto, para um problema real de agrupamento de dados podem ser considerados apenas duas grandes classes de atributos, a saber:

- Atributos Quantitativos – expressam numericamente a medida de uma dada variável. Estes podem ser de dois tipos:
  - Contínuos – atributos assumem valores reais. Exemplo: salário, renda, altura, etc.;
  - Discretas – atributos assumem valores inteiros. Exemplo: idade, número de empregados, CPF, etc.
- Atributos Categóricos – são variáveis não numéricas de valores finitos. Podem assumir dois tipos:
  - Binários – possuem apenas dois tipos de valor. Exemplo: sexo (masculino ou feminino), fumante (sim ou não), etc.;
  - Nominais – possuem mais do que dois tipos de valor. Exemplo: escolaridade (analfabeto, ensino fundamental, ensino médio ou superior), estado civil (casado, solteiro, separado ou viúvo), etc.

Os atributos categóricos necessitam de uma representação numérica para que o algoritmo de agrupamento de dados consiga operar sobre os seus valores.

3.1.1.2
Normalização dos Atributos

A normalização dos dados é importante para garantir que cada variável tenha o mesmo peso, exercendo a mesma influência na execução do algoritmo. Essa influência acontece predominantemente ao se calcular as medidas de semelhança ou dessemelhança entre os dados, conhecida como medidas de proximidades. Sem a normalização, as variáveis com maior escala se tornam dominantes.

Supondo que a medida de proximidade adotada seja a distância euclidiana e o número de variáveis seja igual a dois, a distância entre dois pontos será dada pela fórmula abaixo:

$$DE = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2} \quad (3.1)$$

onde x_i e μ_i são respectivamente o valor do ponto e média da i-ésima variável.

Caso x_1 tenha uma ordem de grandeza muito maior que x_2, a distância será dominada pela primeira variável.

Para evitar esse problema, é aconselhável que se normalize os dados. As técnicas mais utilizadas são: *z-score* e *min-max cutoff*.

3.1.2
Agrupamento de Dados

A segunda etapa do processo de análise de agrupamentos é o agrupamento de dados, que tem como objetivo dividir um determinado conjunto de dados em grupos com características similares entre si.

Os algoritmos de agrupamento de dados podem ser classificados em duas grandes classes de métodos:

- Métodos Hierárquicos:
  - Algoritmos Aglomerativos;
  - Algoritmos Divisivos.

- Métodos Particionais:
  - Algoritmos Exclusivos;
  - Algoritmos Não Exclusivos.

Em geral, os métodos de agrupamento de dados buscam dados similares entre si através de uma medida de proximidade, conforme visto a seguir.

3.1.2.1
Medidas de Proximidade

A medida de proximidade pode ser definida como a medida de similaridade ou dissimilaridade entre os dados.

A matriz de similaridades é uma matriz de dimensão nxn contendo as medidas de similaridades/dissimilaridades entre os n objetos. Essa matriz é bastante utilizada em diversos algoritmos de agrupamento de dados.

3.1.2.1.1
Dissimilaridade

Dissimilaridade é a medida de diferença entre dois objetos. Existem várias maneiras possíveis de se obter essa medida. Métricas muito conhecidas, como a distância Euclidiana e a distância de Manhattan, podem e são utilizadas, mas medidas de dissimilaridade baseadas no Coeficiente de Correlação de Pearson são muito úteis quando o objetivo é o agrupamento de dados, pois ele mede o nível de relacionamento entre duas variáveis.

O Coeficiente de Correlação de Pearson $R(x, y)$ é dado por:

$$R(x,y) = \frac{\sum_{i=1}^{p} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{p} (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^{p} (y_i - \mu_y)^2}} \quad (3.2)$$

onde x_i e y_i são, respectivamente, os valores dos i-ésimos atributos dos dados x e y, e μ_x e μ_y são, respectivamente, os valores de média dos dados x e y.

Os valores do coeficiente de correlação de Pearson estão no intervalo $-1 \le R \le 1$. Quanto mais próximo de 1, mais correlacionados estão os dados, da mesma forma que, quanto mais próximos de -1, menos correlacionados estão os dados.

As Medidas de Dissimilaridade baseadas no Coeficiente de Correlação de Pearson são:

$$d(x,y) = \frac{(1 - R(x,y))}{2} \quad (3.3)$$

$$d(x,y) = 1 - |R(x,y)| \quad (3.4)$$

onde $d(x,y)$ é a medida de dissimilaridade entre os dados x e y.

Com a expressão 3.3, variáveis com uma correlação positiva alta terão um coeficiente de dissimilaridade perto de zero, enquanto que variáveis com uma correlação negativa forte serão consideradas muito dissimilares. Em outras aplicações pode ser preferível usar a expressão 3.4, onde variáveis com uma correlação negativa ou positiva alta receberão um coeficiente de dissimilaridade perto de zero, e serão considerados muito dissimilares quando o coeficiente de correlação for próximo de zero.

A escolha de uma das expressões depende muito do problema e do entendimento que se tem sobre a base de dados. Pode ser conveniente usar a expressão 3.4 quando se deseja que variáveis muito correlacionadas ou pouco correlacionadas sejam similares. Já na expressão 3.3, a relação entre a correlação e a dissimilaridade é linear.

Comparações entre as duas equações sobre dados reais mostraram que a equação 3.3 apresentou resultados bem melhores, embora a equação 3.4 tenha apresentado resultados relativamente bons.

3.1.2.1.2
Similaridade

Similaridade é a medida de igualdade entre dois objetos. Tipicamente a similaridade s entre dois objetos x e y assume valores entre 0 e 1, onde 0 expressa que os dois objetos não são similares, enquanto que 1 expressa máxima similaridade. Geralmente são consideradas as seguintes condições para se definir similaridade:

- $0 \le s(x, y) \le 1$
- $s(x, x) = 1$
- $s(x, y) = s(y, x)$

Como as medidas de similaridades não podem ser calculadas diretamente através do coeficiente de correlação de Pearson, é necessário efetuar algumas transformações a fim de se respeitar as condições de similaridade. Existem essencialmente duas maneiras para isso, dependendo do significado dos dados e do propósito da aplicação. Supondo que variáveis com uma correlação negativa forte são variáveis muito diferentes, considere-se a expressão de similaridade abaixo.

$$s(x, y) = \frac{1 + R(x, y)}{2} \quad (3.5)$$

onde x e y são objetos em um conjunto de dados, e $s(x, y)$ e $R(x, y)$ são, respectivamente, a similaridade e o coeficiente de correlação entre x e y.

A expressão 3.5 indica que sempre que o coeficiente de correlação for próximo de -1, a similaridade será próxima de 0, enquanto que valores de correlação próximos de 1 representam valores de similaridade próximos de 1.

Há existem situações onde variáveis com uma correlação positiva ou negativa forte representam essencialmente o mesmo significado. Para esses casos pode ser usada a expressão abaixo:

$$s(x, y) = |R(x, y)| \quad (3.6)$$

onde x e y são objetos em um conjunto de dados, $s(x,y)$ e $R(x,y)$ são, respectivamente, a similaridade e o coeficiente de correlação entre x e y.

A expressão 3.6 indica que para coeficiente de correlação próximo de 1 ou -1, a similaridade será próxima de 1, enquanto que valores de correlação próximos de 0 representam objetos pouco similares, com valores de similaridade próximos de 0.

3.1.2.2
Métricas Comuns em Medidas de Proximidade

As métricas mais comuns e utilizadas na prática são a distância Euclidiana, distância de Mahalanobis e distância de Manhattan. A Figura 2 mostra as superfícies observadas por cada uma dessas métricas.

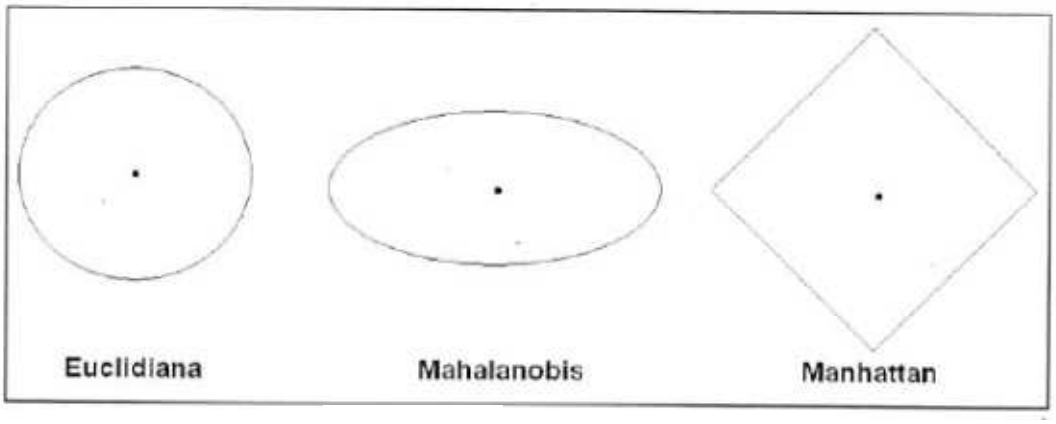

Figura 2: Superfícies observadas pelas distâncias Euclidiana, Mahalanobis e Manhattan.

Segundo Vale, em geral prefere-se usar a distância Euclidiana, pois a distância de Manhattan é uma simplificação da distância Euclidiana e a de Mahalanobis dificulta a determinação precisa das matrizes de covariância e o custo computacional cresce muito com o número de variáveis envolvidas. Assim, segue definição da métrica de distância d usual entre os dados x e y.

- Distância Euclidiana é uma medida invariante a translações, porém assume covariâncias iguais entre as classes e em geral não é invariante a transformações lineares. É a métrica mais utilizada na prática.

$$d(x,y) = \|x-y\| = \sqrt{(x-y)^T (x-y)} \quad (3.7)$$

3.1.3
Análise dos Resultados

A análise dos resultados compreende primeiramente a avaliação da qualidade dos agrupamentos. É importante ressaltar que alguns métodos possuem métricas específicas para cálculo da qualidade do agrupamento. O passo seguinte é a compreensão e interpretação dos agrupamentos gerados, a fim de inferir regras ou características que expliquem cada grupo.

3.2
Métodos de Agrupamento de Dados

Esse é o item de maior destaque no processo de agrupamento de dados, pois é o responsável pelo agrupamento propriamente dito.

Conforme visto anteriormente, os métodos de agrupamentos de dados podem ser divididos em duas grandes categorias, cada uma delas compreendendo diferentes tipos de algoritmos:

- Métodos Hierárquicos:
  - Algoritmos Aglomerativos;
  - Algoritmos Divisivos.
- Métodos Particionais:
  - Algoritmos Exclusivos;
  - Algoritmos Não-exclusivos.

A seguir são descritas as características de cada um desses conjuntos de métodos, bem como os métodos mais conhecido ou que tenham alguma relevância no processo de agrupamento.

3.2.1
Métodos Hierárquicos

Os métodos hierárquicos são técnicas simples onde os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos. Essa representação facilita a visualização sobre a formação dos agrupamentos em cada estágio e com que grau de semelhança. Os métodos hierárquicos não requerem que seja definido um número *a priori de* agrupamentos. Os métodos hierárquicos são subdivididos em Métodos Aglomerativos e Métodos Divisivos.

3.2.1.1
Métodos Aglomerativos

Os métodos aglomerativos são os mais comuns entre os métodos hierárquicos. Nesse tipo de método inicia-se com cada padrão formando seu próprio agrupamento e gradualmente os grupos são unidos até que um único agrupamento contendo todos os dados seja gerado. Logo no início do processo, os agrupamentos são pequenos e os elementos de cada grupo possuem um alto grau de similaridade. Ao final do processo, há poucos agrupamentos, cada podendo conter muitos elementos e menos similares entre si.

As principais desvantagens dos métodos hierárquicos aglomerativos são:

- Os agrupamentos não podem ser corrigidos, ou seja, os padrões de um determinado agrupamento permanecerão nesse agrupamento até o final da execução do algoritmo;
- Requerem muito espaço de memória e tempo de processamento.

3.2.1.2
Métodos Divisivos

Os métodos divisivos são os menos comuns entre os métodos hierárquicos devido à ineficiência e exigem capacidade computacional maior que os métodos hierárquicos aglomerativos.

Esse método começa com um único agrupamento formado por todos os padrões e gradualmente vai dividindo os agrupamentos em agrupamentos menores até que termine com um agrupamento por padrão.

O primeiro passo do algoritmo envolve todas as divisões possíveis dos dados em dois agrupamentos, o que o tornaria impraticável para um número grande de elementos, envolvendo, dessa forma, um grande número de combinações.

Os métodos divisivos possuem a vantagem de considerar muitas divisões no primeiro passo, diminuindo a probabilidade de uma decisão errada, sendo assim, mais seguros que os métodos hierárquicos aglomerativos (Vale).

3.2.1.3
Métodos Hierárquicos Conhecidos

3.2.1.3.1
Agglomerative Nesting (AGNES)

AGNES é um método hierárquico aglomerativo. A união entre agrupamentos é feita entre os agrupamentos com a menor dissimilaridade entre si. A principal vantagem do método reside no tempo de computação.

3.2.1.3.2
Divisive Analysis (DIANA)

DIANA é um método hierárquico divisivo. O algoritmo consiste de $n-1$ divisões sucessivas, onde n é o número de dados do conjunto de dados. Em cada passo é selecionado o agrupamento C com o maior diâmetro.

3.2.1.3.3
Monothetic Analysis (MONA)

O método MONA é destinado exclusivamente a dados do tipo binário. Apesar do algoritmo ser hierárquico divisivo, ele não usa dissimilaridades entre objetos e por isso a matriz de dissimilaridades não é computada. A divisão em agrupamentos utiliza as variáveis diretamente.

3.2.2
Métodos Particionais

Os métodos particionais são métodos baseados na minimização de uma função de custo, onde os padrões são agrupados em um número k de agrupamentos escolhido a priori. Cada padrão é agrupado na classe em que essa função de custo é minimizada.

Uma das principais vantagens dos métodos particionais em relação aos métodos hierárquicos é a possibilidade de um padrão poder mudar de agrupamento com a evolução do algoritmo e a possibilidade de se operar com bases de dados maiores. Os métodos particionais são extremamente mais rápidos que os métodos hierárquicos. No entanto, as principais desvantagens dos métodos particionais estão no fato do número de agrupamentos ter que ser escolhido *a priori*, o que poderá sugerir interpretações erradas sobre a estrutura dos dados caso o número de agrupamentos não seja o ideal e no fato de que o algoritmo é em geral sensível às condições iniciais, podendo gerar resultados diferentes a cada rodada. O problema quando se escolhe erroneamente o

número de agrupamentos é que o método irá impor uma estrutura aos dados, no lugar de buscar a estrutura inerente a estes.

3.2.2.1
Métodos Não-exclusivos

A segmentação de dados numéricos forma uma base de muitos algoritmos de classificação. O seu propósito é identificar agrupamentos naturais de dados para produzir uma representação concisa do comportamento do sistema.

É comum em um processo de agrupamento de dados que cada objeto, representado pelo dado, pertença a um único agrupamento. Os métodos não-exclusivos, conhecidos também como métodos *fuzzy*, permitem alguma ambiguidade entre os dados, o que geralmente acontece.

Esses métodos são técnicas de agrupamento de dados onde cada padrão pertence a um agrupamento com certo grau de pertinência.

A principal vantagem dos agrupamentos *fuzzy* em relação aos demais métodos particionais está no fato de representar com muito mais detalhes a informação sobre a estrutura dos dados. Por outro lado, isso poderia ser considerado uma desvantagem, pois a quantidade de informação gerada cresce muito rapidamente com o número de objetos e o número de agrupamentos, tornando a compreensão mais difícil. Outra desvantagem é a ausência de objetos representativos dos agrupamentos formados e o fato de que geralmente os algoritmos são mais complicados e consomem um maior tempo de computação. No entanto, os princípios *fuzzy* são muito interessantes, pois permitem a descrição de algumas incertezas que geralmente estão presentes em dados reais.

3.2.2.2
Métodos Particionais Conhecidos

3.2.2.2.1
K-MEANS

O método *k-means* é um dos métodos mais populares das técnicas particionais. O método particiona os dados em k agrupamentos mutuamente exclusivos. Diferentemente dos métodos hierárquicos, o *k-means* não cria uma estrutura em árvore para descrever o agrupamento dos dados e é mais adequado para uma grande quantidade de dados.

O algoritmo procura dentro do possível, a partição em que os padrões de cada agrupamento estão mais próximos entre si e mais distantes dos padrões dos outros agrupamentos. O *k-means* é um algoritmo iterativo que minimiza a soma das distâncias de cada padrão ao centróide de cada agrupamento até que a função objetivo não se altere ou se altere muito pouco, ou até que o número de iterações máximo pré-determinado tenha sido alcançado. O resultado é um conjunto de agrupamentos compactos e bem separados tanto quando possível.

Em resumo, cada agrupamento é representado pelo centro do grupo e cada padrão é atribuído ao agrupamento que está mais próximo.

O procedimento geral pode ser descrito em poucos passos:

1) Inicializar as médias das k partições;
2) Para cada padrão determinar a partição mais próxima;
3) Calcular a média de cada partição;
4) Se houver mudança na média das partições, voltar ao passo 2;
5) Resultado: as médias das k partições.

Como muitos outros problemas de minimização, a solução encontrada pelo *k-means* geralmente depende do ponto de partida, mas em geral o algoritmo encontra um mínimo local. Trata-se de um método prático e computacionalmente eficiente, embora seja sensível a ruído e *outliers* e não aplicável para agrupamentos não-convexos.

O resultado deste método pode, em muitos casos, ser drasticamente afetado pela escolha das condições iniciais. Entretanto, em base de dados bem estruturadas, em geral espera-se a convergência para um mínimo global. Comportamentos como convergência lenta e resultado de agrupamentos bastante diferentes para diferentes configurações iniciais podem indicar que o número de agrupamentos escolhido esteja errado, ou que os dados não possuam estrutura de agrupamentos.

O método apresenta bons resultados apenas quando os agrupamentos são hiperesféricos e possuem aproximadamente o mesmo número de padrões em cada agrupamento. O bom desempenho do algoritmo depende muito também da escolha adequada da medida de distância e do ponto inicial de partida do algoritmo.

3.2.2.2.2
Fuzzy C-Means

Fuzzy c-means (FCM) é uma técnica de agrupamento de dados *fuzzy*. Esta técnica foi originalmente introduzida por Jim Bezdek em 1981 como uma evolução das técnicas de agrupamento de dados mais recentes e fornece um método que mostra como agrupar padrões que pertencem a um espaço multidimensional em um número específico de diferentes agrupamentos.

3.2.2.2.3
Fuzzy Analysis (FANNY)

Fuzzy Analysis (FANNY) é uma técnica de agrupamento de dados *fuzzy* proposta por Kauffman. Assim como o FCM, este método atribui a cada padrão um grau de pertinência aos agrupamentos envolvidos, com a vantagem de ser mais robusto. O algoritmo roda iterativamente e pára quando a função objetivo converge, gerando assim estimativas para os k agrupamentos.

Implicitamente, no FCM considera-se que cada objeto, representado como o centro de cada agrupamento, é dado pela média das coordenadas em um

espaço p-dimensional. O método FANNY não possui representações de tais objetos, sendo necessárias apenas as distâncias ou as dissimilaridades entre os dados.

3.2.2.2.4
Gustafson-Kessel

Esse método é uma extensão do método FCM e tem o objetivo de detectar agrupamentos de formas geométricas diferentes, se adaptando, dessa forma, a diferentes estruturas.

Este método pode ser definido como um método adaptativo: a principal motivação é encontrar agrupamentos não-esféricos. A idéia básica do algoritmo está na utilização de medidas não-esféricas de distâncias específicas para cada agrupamento. Essas medidas de distância evoluem com o tempo.

3.2.2.2.5
Gath-Geva

Gath-Geva é uma técnica de agrupamento de dados *fuzzy* que tem como propósito detectar agrupamentos de formas, tamanhos e densidades diferentes. O algoritmo contém apenas o parâmetro de *fuzzificação* m como entrada para o algoritmo e utiliza uma medida de distância baseada na estimativa de máxima probabilidade *fuzzy*.

3.2.2.2.6
Mistura de Densidades

Nesse método os dados são vistos como provenientes de uma função de densidade de probabilidade, cada função representando um agrupamento diferente. Dessa forma, os agrupamentos podem ser considerados como uma soma de gaussianas ponderadas pela probabilidade *a priori* de cada agrupamento.

O resultado do método se assemelha ao Teorema de *Fourier* onde qualquer função pode ser aproximada por uma soma de cossenoides. Neste caso, qualquer função de densidade de probabilidade pode ser aproximada por uma soma de gaussianas com parâmetros desconhecidos a serem estimados.

3.2.2.2.7
Partitioning Around Medoids (PAM)

Esse algoritmo procura por k objetos chamados de medóides, que são representativos de cada agrupamento e contêm os padrões onde a dissimilaridade média dos padrões pertencentes a um dado agrupamento é mínima. Em outras palavras, esse algoritmo minimiza a soma das dissimilaridades.

3.2.2.2.8
Clustering Large Applications (CLARA)

Esse método é uma adaptação do método PAM para um conjunto grande de dados.

O método PAM trabalha com uma matriz de dissimilaridades contendo todos os n dados, consumindo, dessa forma, muito espaço de memória para um conjunto grande de dados. Por esse motivo se torna impraticável o uso desse método. Já o método CLARA não trabalha com toda a matriz de dissimilaridades de uma só vez. Trabalha com subconjuntos de tamanhos previamente definidos por vez.

4
Resultados

Um dos objetivos deste estudo foi, a partir de uma base contendo dados da Pesquisa de Impacto de Cursos Técnicos do SENAI-RJ, identificar grupos de egressos com perfis semelhantes. Para atingir tal objetivo, foi realizada a análise de agrupamentos na tentativa de se encontrar tais grupos.

Antes da análise, houve pré-processamento dos dados fornecidos pela pesquisa. Nesta etapa do trabalho, foram realizadas as seguintes tarefas:

- Seleção de variáveis;
- Codificação de variáveis;
- Transformação de variáveis;
- Identificação de inconsistências;
- Identificação de valores aberrantes;
- Análise de fatores;
- Seleção da amostra para trabalho.

Na etapa seguinte, da análise de agrupamentos propriamente dita, particionou a base em grupos e visualizou as características principais de cada um. Várias tentativas de particionamento foram feitas, utilizando diferentes técnicas e parâmetros em cada uma e escolheu a melhor no final.

Todas as etapas foram guiadas de maneira a fazer com que os resultados encontrados estivessem em sincronia com o que estava sendo procurado pela pesquisa.

4.1
Descrição da Base de Dados

A base de dados analisada foi extraída do banco de dados da Pesquisa de Impacto dos Cursos Técnicos do SENAI-RJ com 461 respostas e constitui um subconjunto representativo do total de concluintes. A base compreendia 63 atributos relacionados a:

- Perfil;

- Continuidade dos estudos;
- Empregabilidade;
- Empreendedorismo;
- Benefícios;
- Satisfação.

A tabela 1 contém a listagem dos atributos presentes na base de dados fornecida pela pesquisa.

Atributo	Descrição
Sexo	Sexo:
Idade	Faixa etária:
Segme	Segmento do curso:
Unop	Unidade operacional do SENAI-RJ (escola):
Reali	P1. Porque decidiu realizar um curso profissionalizante?
Outre	P1. Por outro motivo. Qual?
Estud	P2. Atualmente você está estudando?
Areap	P3. O curso que está realizando é da mesma área ao curso realizado?
Outro	P4. Você já fez ou pretende fazer outro curso no SENAI?
Motiv	P5. Se você ainda não fez outro curso no SENAI, qual o motivo?
Outmo	P5. Por outro motivo. Qual?
Antes	P6. Qual era sua principal ocupação antes de realizar o curso?
Outan	P6. Outra ocupação. Qual?
Depoi	P7. Qual a sua principal ocupação atual?
Outde	P7. Outra ocupação. Qual?
Tempo	P8. Após início do curso, quanto tempo conseguiu emprego/trabalho?
Ativi	P9. Sua principal atividade profissional atual é:
Difer	P10. Se trabalha numa área diferente da do SENAI, qual o motivo?
Outdi	P10. Por outro motivo. Qual?
	P11. Grau de dificuldade geral que você tem para realizar seu trabalho:
Difia	a) Uso de máquina;
Difib	b) Uso de ferramentas;
Dific	c) Uso de instrumento de medição;
Difid	d) Aplicação de conhecimentos técnicos;
Difie	e) Leitura e interpretação de desenho;
Difif	f) Leitura e interpretação de normas técnicas;
Difig	g) Manutenção de máquinas / ferramentas;
Difih	h) Processamento de materiais;
Difii	i) Uso de equipamentos de proteção individual e coletivo;
Difij	j) Controle de qualidade;
Difik	k) Relacionamento com chefia;
Difil	l) Trabalho em grupo.
Renda	P12. Quanto você ganha por mês?
Empre	P13. Nome da empresa em que trabalha?
Emare	P13. Área em que atua na empresa?

Emtem	P13. Tempo na empresa?
Emnom	P13. Nome do chefe?
Emtel	P13. Telefone de contato?
Emema	P13. Email da empresa?
Naotr	P14. Se você não está trabalhando atualmente, qual o principal motivo?
Outna	P14. Outro:
Desmo	P15. Se procurou emprego mas não encontrou, qual o principal motivo?
Outde	P15. Outro:
Propr	P16. Você já pensou em abrir seu próprio negócio?
Monta	P17. Qual motivo que pretenda montar/ tenha montado próprio negócio?
Outmo	P17. Outro:
Inici	P18. Condições necessárias para iniciar um empreendimento, diria que:
Outin	P18. Outro:
Apren	P19. Você utiliza o que aprendeu no curso do SENAI?
Novo	P20.a) O curso possibilitou que arranjasse um novo emprego/trabalho?
Manti	P20.b) O curso colaborou para que você se mantivesse empregado?
Promo	P20.c) O curso possibilitou que você obtivesse uma promoção?
Melho	P20.d) O curso possibilitou uma melhoria salarial ou de renda?
Conti	P20.e) O curso melhorou a sua base para continuar os estudos?
Conhe	P20.f) O curso ofereceu novos conhecimentos da sua área profissional?
Desem	P20.g) O curso melhorou o seu desempenho no trabalho?
Relac	P20.h) O curso melhorou o relacionamento com as pessoas no trabalho?
Visão	P20.i) O curso aumentou sua visão do processo produtivo?
Atuac	P20.j) O curso ampliou as suas possibilidades de atuação profissional?
Busca	P20.k) O curso orientou na busca de emprego?
Abrir	P20.l) O curso colaborou para que pensasse / abrisse o próprio negócio?
Recom	P21. Você recomendaria o curso para algum amigo/conhecido?
Nível	P22. Qual seu nível de satisfação com o SENAI numa escala de 1 a 10?
Comen	P23. Deseja fazer crítica, sugestão, comentário ou elogio ao curso?

Tabela 1: Descrição dos atributos.

4.2
Pré-processamento dos Dados

4.2.1
Seleção de variáveis

Como observado na tabela 1, existe uma enorme quantidade de atributos, o que traz muita complexidade ao processo de análise. Felizmente, alguns atributos são perguntas do tipo ‘Outro’ e não foram levados em consideração, bem como a última variável de caráter subjetivo e também com infinidade de respostas.

Outros atributos também foram previamente descartados como, por exemplo, as questões sobre a empresa em que os egressos trabalham, que formam cadastro para outra pesquisa do Sistema FIRJAN.

Desta forma, os 47 atributos (variáveis) selecionados foram:

Atributo	Descrição
Sexo	Sexo:
Idade	Faixa etária:
Segme	Segmento do curso:
Unop	Unidade operacional do SENAI-RJ (escola):
Reali	P1. Porque decidiu realizar um curso profissionalizante?
Estud	P2. Atualmente você está estudando?
Areap	P3. O curso que está realizando é da mesma área ao curso realizado?
Outro	P4. Você já fez ou pretende fazer outro curso no SENAI?
Motiv	P5. Se você ainda não fez outro curso no SENAI, qual o motivo?
Antes	P6. Qual era sua principal ocupação antes de realizar o curso?
Depoi	P7. Qual a sua principal ocupação atual?
Tempo	P8. Após início do curso, quanto tempo conseguiu emprego/trabalho?
Ativi	P9. Sua principal atividade profissional atual é:
Difer	P10. Se trabalha numa área diferente da do SENAI, qual o motivo?
	P11. Grau de dificuldade geral que você tem para realizar seu trabalho:
Difia	a) Uso de máquina;
Difib	b) Uso de ferramentas;
Dific	c) Uso de instrumento de medição;
Difid	d) Aplicação de conhecimentos técnicos;
Difie	e) Leitura e interpretação de desenho;
Difif	f) Leitura e interpretação de normas técnicas;
Difig	g) Manutenção de máquinas / ferramentas;
Difih	h) Processamento de materiais;
Difii	i) Uso de equipamentos de proteção individual e coletivo;
Difij	j) Controle de qualidade;
Difik	k) Relacionamento com chefia;
Difil	l) Trabalho em grupo.
Renda	P12. Quanto você ganha por mês?
Naotr	P14. Se você não está trabalhando atualmente, qual o principal motivo?
Desmo	P15. Se procurou emprego mas não encontrou, qual o principal motivo?
Propr	P16. Você já pensou em abrir seu próprio negócio?
Monta	P17. Qual motivo que pretenda montar/ tenha montado próprio negócio?
Inici	P18. Condições necessárias para iniciar um empreendimento, diria que:
Apren	P19. Você utiliza o que aprendeu no curso do SENAI?
Novo	P20.a) O curso possibilitou que arranjasse um novo emprego/trabalho?
Manti	P20.b) O curso colaborou para que você se mantivesse empregado?
Promo	P20.c) O curso possibilitou que você obtivesse uma promoção?
Melho	P20.d) O curso possibilitou uma melhoria salarial ou de renda?
Conti	P20.e) O curso melhorou a sua base para continuar os estudos?
Conhe	P20.f) O curso ofereceu novos conhecimentos da sua área profissional?

Desem	P20.g) O curso melhorou o seu desempenho no trabalho?
Relac	P20.h) O curso melhorou o relacionamento com as pessoas no trabalho?
Visão	P20.i) O curso aumentou sua visão do processo produtivo?
Atuac	P20.j) O curso ampliou as suas possibilidades de atuação profissional?
Busca	P20.k) O curso orientou na busca de emprego?
Abrir	P20.l) O curso colaborou para que pensasse / abrisse o próprio negócio?
Recom	P21. Você recomendaria o curso para algum amigo/conhecido?
Nível	P22. Qual seu nível de satisfação com o SENAI numa escala de 1 a 10?

Tabela 2: Descrição dos atributos selecionados.

4.2.2
Codificação e Transformação das Variáveis

Como dito anteriormente, a maioria das variáveis são categóricas nominais. No intuito de determinar melhor o perfil de cada egresso e garantir que cada variável tenha o mesmo peso, todos os atributos foram codificados e transformados num intervalo $[0,1]$. Desta forma, houve a formação de agrupamentos envolvendo valores numéricos, que são fáceis de serem comparados e terem a distância medida.

4.2.3
Identificação de Inconsistências e de Valores Aberrantes

Outro desafio usual para a área de análise de agrupamentos é melhorar a qualidade dos dados disponíveis. Muitas instituições ainda dão pouca atenção a esta questão. Felizmente, essa prática parece estar mudando nas empresas. Em todas as pesquisas do Sistema FIRJAN há a preocupação pela qualidade da base de dados.

Normalmente nesta fase do pré-processamento, há procura na base de dados por registros com campos nulos ou por valores distorcidos (*outliers*). Como a Pesquisa de Impacto dos Cursos Técnicos do SENAI-RJ foi realizada por telefone e em parceria com experiente instituto de pesquisa, não foram detectados nenhum desses problemas.

4.2.4
Análise de Fatores

Esta etapa procurou encontrar fatores latentes no conjunto de variáveis. Foram feitas diversas análises distintas, utilizando diferentes métodos de extração de fatores e de rotação. Os melhores resultados, no entanto, foram obtidos utilizando o método de extração de *Principal Axis Factoring* e o método de rotação *Varimax* (Kaiser). Com estes métodos, obteve melhor separação entre os fatores, além de maior capacidade explicativa.

Algumas variáveis foram eliminadas, visto que havia dependência linear. E posteriormente a esta análise, os 12 atributos relacionados ao grau de dificuldade para realizar o trabalho foram transformados em apenas um atributo oriundo da média dos originais.

Por fim, os métodos acima aplicados neste novo conjunto de variáveis resultaram em quatro fatores relevantes, responsáveis por explicar 77,2% da variação dos dados.

Segundo Azevedo, em muitos casos, os atributos são selecionados também com base na experiência ou bom senso do especialista dos dados.

A tabela 3 descreve os 22 atributos finais:

Atributo	Descrição
Sexo	Sexo:
Idade	Faixa etária:
Segme	Segmento do curso:
Unop	Unidade operacional do SENAI-RJ (escola):
Reali	P1. Porque decidiu realizar um curso profissionalizante?
Estud	P2. Atualmente você está estudando?
Areap	P3. O curso que está realizando é da mesma área ao curso realizado?
Outro	P4. Você já fez ou pretende fazer outro curso no SENAI?
Antes	P6. Qual era sua principal ocupação antes de realizar o curso?
Depoi	P7. Qual a sua principal ocupação atual?
Tempo	P8. Após início do curso, quanto tempo conseguiu emprego/trabalho?
Ativi	P9. Sua principal atividade profissional atual é:

Dific	P11. Grau de dificuldade geral que você tem para realizar seu trabalho:
Renda	P12. Quanto você ganha por mês?
Propr	P16. Você já pensou em abrir seu próprio negócio?
Apren	P19. Você utiliza o que aprendeu no curso do SENAI?
Novo	P20.a) O curso possibilitou que arranjasse um novo emprego/trabalho?
Manti	P20.b) O curso colaborou para que você se mantivesse empregado?
Promo	P20.c) O curso possibilitou que você obtivesse uma promoção?
Melho	P20.d) O curso possibilitou uma melhoria salarial ou de renda?
Recom	P21. Você recomendaria o curso para algum amigo/conhecido?
Nível	P22. Qual seu nível de satisfação com o SENAI numa escala de 1 a 10?

Tabela 3: Descrição dos atributos finais

4.3
Seleção da Amostra

Usualmente nesta etapa de pré-processamento dos dados, uma amostra é selecionada da base original para ser utilizada nas próximas análises. De acordo com Azevedo, grande quantidade de objetos não deve ser aplicada devido à alta complexidade computacional de alguns algoritmos de agrupamento, o que levaria o processo de análise a ser bastante demorado.

No presente estudo, o banco de dados da pesquisa já é uma amostra representativa da população de concluintes dos cursos técnicos do SENAI-RJ.

4.4
Análise de Agrupamentos

Esta etapa teve como intuito particionar a base em grupos de egressos com características semelhantes. Foram feitos diversos particionamentos, utilizando diferentes métodos, mas sempre com a distância Euclidiana como medida de dissimilaridade.

Por fim, o método escolhido foi um dos métodos mais populares das técnicas particionais, o *k-means*. Duas das razões iniciais pela escolha foram que o método particiona os dados em agrupamentos mutuamente exclusivos, além de ser mais adequado para grande quantidade de dados.

Primeiramente, os resultados dos particionamentos de 2, 3 e 4 grupos foram observados. Na primeira tentativa, o algoritmo basicamente dividiu em um grupo formado pelos empregados e outro por desempregados após o curso. Na segunda tentativa, além da segmentação anterior também levou em consideração a idade dos egressos. Já na última partição, a variável sexo também foi determinante, uma vez que a base apresenta apenas 17,6% de mulheres.

Na tentativa de obter resultados melhores, foram feitos particionamentos em 5, 6 e 7 grupos, mas os resultados não compensaram o número elevado de grupos.

Desta forma, a decisão final foi por quatro grupos, classificados em:

- Egressos adultos;
- Mulheres;
- Egressos jovens;
- Homens adolescentes.

Num trabalho conjunto entre a análise de fatores e análise de agrupamentos, 15 variáveis apresentaram relevância significativa no particionamento:

- Sexo;
- Faixa etária;
- Segmento do curso;
- P1. Porque decidiu realizar um curso profissionalizante?
- P2. Atualmente você está estudando?
- P6. Qual era sua principal ocupação antes de realizar o curso?
- P7. Qual a sua principal ocupação atual?
- P8. Após início do curso, quanto tempo conseguiu emprego/trabalho?
- P9. Sua principal atividade profissional atual é:
- P11. Grau de dificuldade geral que você tem para realizar seu trabalho:
- P12. Quanto você ganha por mês?
- P19. Você utiliza o que aprendeu no curso do SENAI?
- P20.a) O curso possibilitou que arranjasse um novo emprego/trabalho?
- P20.b) O curso colaborou para que você se mantivesse empregado?
- P20.d) O curso possibilitou uma melhoria salarial ou de renda?

5
Conclusões

Segundo Weinstein, segmentação é o processo de dividir mercados em grupos de potenciais consumidores com necessidades e/ou características similares, que, provavelmente, exibirão comportamento de compra similar. Surgiu como uma importante ferramenta de planejamento de marketing e como um dos fundamentos para a efetiva formulação de estratégias em muitas empresas norte-americanas e também de outros países.

Como um dos objetivos principais desta monografia, os dados foram aglutinados em *clusters* a fim de permitirem uma segmentação da Pesquisa de Impacto dos Cursos Técnicos do SENAI-RJ segundo os perfis. O segundo objetivo foi o estudo da análise de agrupamentos, que resultou no particionamento de quatro grupos de egressos classificados em ‘Egressos adultos’, ‘Mulheres’, ‘Egressos jovens’ e ‘Homens adolescentes’. Por fim, o último objetivo foi estudar o comportamento de cada grupo criado. A partir das análises de fatores e de agrupamentos encontradas, chegou a seguinte constituição dos *clusters*:

- Grupo 1 - Egressos adultos;

Os componentes possuem idade superior a 30 anos. Decidiram realizar o curso técnico principalmente por dois motivos: se requalificar/aprimorar na área profissional em que já atuava e adquirir uma nova profissão (diferente da que já exercia). Antes do curso, esses egressos já estavam inseridos no mercado de trabalho e eram empregados com ou sem carteira de trabalho assinada. Os benefícios adquiridos pelo curso foram: o curso colaborou para que se mantivessem empregados e possibilitou melhoria salarial, uma vez que com a conclusão do curso, esses egressos estão inseridos apenas no mercado formal, ou seja, são empregados com carteira de trabalho assinada. A renda mensal varia de R$ 1.090 a R$ 2.725, podendo também alcançar o intervalo de R$ 3.815 a R$ 5.450. De acordo com a escala proposta pela pesquisa, este grupo não possui dificuldade em realizar o trabalho. Por fim, utilizam o que aprenderam no curso tanto na vida pessoal como na profissional.

- Grupo 2 – Mulheres;

Os integrantes escolheram basicamente o segmento industrial de segurança no trabalho para estudar no SENAI-RJ. A principal razão para realizar o curso técnico foi conhecer melhor uma área profissional em que possuem interesse. Antes do curso essas mulheres eram estagiárias / *trainees*. O curso não colaborou para que se mantivessem empregadas, uma vez que ao término deste, elas se declararam estudantes ou desempregadas. Este fato corrobora para a questão, que infelizmente, a indústria ainda prefere contratar homens mesmo em áreas brandas, ou seja, sem esforço físico como o segmento de segurança no trabalho. Por não estarem inseridas no mercado de trabalho, utilizam o que aprenderam no curso apenas na vida pessoal.

- Grupo 3 - Egressos jovens;

As pessoas que constituem este grupo têm idade entre 20 e 24 anos. Quando da escolha do curso, apresentaram forte propensão ao segmento industrial de alimentos e bebidas e decidiram realizar o curso técnico para adquirir uma profissão e ingressar no mercado de trabalho além de ocupar o tempo livre. Antes do curso esses egressos eram estudantes ou desempregados. Como benefícios oriundos do curso, há a possibilidade de arrumar um novo emprego/trabalho bem como a possibilidade de melhoria salarial. Neste sentido, após o curso, são empregados com carteira de trabalho assinada ou estagiários / *trainees*. Conseguiram emprego/trabalho durante o curso ou de um a doze meses após a conclusão do mesmo. A atividade profissional atual é numa área igual a que se qualificou no SENAI-RJ e a renda varia até R$ 1.635.

- Grupo 4 – Homens adolescentes.

É o grupo mais jovem composto por rapazes de até 20 anos, que realizaram preferencialmente cursos nos segmentos industriais de gás e metalurgia. Atualmente, estudam em cursos pré-vestibulares e/ou outro curso técnico. Antes do curso técnico do SENAI-RJ, esses egressos eram estudantes. Infelizmente, para eles, o curso não possibilitou que arranjassem um emprego/trabalho, pois depois do curso continuam estudantes ou desempregados. Assim, utilizam o que aprenderam no curso apenas na vida pessoal ou simplesmente não utilizam.

De acordo com Azevedo, hoje, muitas empresas se deparam com a seguinte questão: “Mais dados significa mais conhecimento?”. Não basta apenas adquirir e disponibilizar os dados, mas é preciso, também, analisá-los, interpretá-los e relacioná-los para que se cheguem a conclusões que possibilitem desenvolver corretamente a estratégia de ação da empresa. Essas análises permitem obter novos conhecimentos e que podem vir a ser muitos úteis, já que conhecimento é um recurso econômico com impacto direto sobre a capacidade de competitividade de uma empresa, qualquer que seja seu setor de atuação.

Conclui-se que as análises de fatores e de agrupamentos construídas no presente estudo, apresentaram conjuntamente eficientes resultados na aquisição de informações estratégicas para o Sistema FIRJAN. Neste sentido, este trabalho, ao demonstrar em termos práticos como a *clusterização* pode ser utilizada, serve como decisivo instrumento de informação exploratória para a equipe de educação profissional da instituição no intuito de garantir constante melhoria dos cursos técnicos oferecidos.

Esta segmentação dos perfis da clientela dos cursos técnicos do SENAI-RJ promove benefícios para o Sistema FIRJAN: capacitação personalizada de recursos humanos na área industrial, contribuindo para o alcance da missão da instituição; capacidade de projetar produtos (novos cursos) que atendam eficazmente às necessidades do mercado de trabalho; contribuição para elaborar estratégias promocionais eficazes e de menor custo; avaliar a concorrência, especialmente entender como a empresa é percebida por seus consumidores reais e potenciais, relativamente à concorrência e, por fim, prover *insights* junto às estratégias de marketing atuais.

Esse tipo de análise de agrupamentos auxilia em encontrar grupos de pessoas, empresas, mercados, produtos dentro de uma categoria, de modo que os itens dentro de cada grupo sejam semelhantes entre si e diferentes dos itens em outros grupos. Assim, os resultados desse tipo de análise são bastante úteis, pois uma vez encontrados os grupos, pode, entre outras decisões, criar pacotes, ofertas e produtos específicos às necessidades e características de cada um aumentando assim a expectativa de resposta, lucratividade e nível de satisfação dos clientes.

Finalmente, desse modo, segundo Bussab, após essa divisão em *clusters*, pode restringir o estudo a um representante de cada grupo, obtendo resultados mais variados e menos custosos. Neste sentido, como sugestão de trabalhos futuros, seria interessante uma avaliação mais minuciosa com representantes “típicos” da população, que poderão ser escolhidos para tentar traçar diferentes históricos, mediante aplicação de questionários mais complexos.

6
Referências Bibliográficas

AAKER, D. A., KUMAR, V., DAY, G. S., **Pesquisa de Marketing**, São Paulo: Atlas, 2001.

ANDERSON, T. W., **An Introduction to Multivariate Statistical Methods**, New York: John Willey, 1984.

AZEVEDO, H. L. C., **Mineração de Dados Aplicada na Solução de Problemas de Marketing Direto e Segmentação de Mercado**, Rio de Janeiro: PUC-Rio, 2001.

BERKHIN, P., **Survey of Clustering Data Mining Techniques**, 2002. Disponível em <www.citeulike.org/user/metamerist/article/556827>. Acesso em setembro de 2011.

BUSSAB, W. O., MIAZAKI, E. S., ANDRADE, D. F., **Introdução à Análise de Agrupamentos**, São Paulo: ABE,1990.

COCHRAN, W. G., **Sampling Techniques**, New York: John Wiley, 1977.

JAIN, A.K., MURTY, M.N. & FLYNN, P.J. **Data Clustering: A Review**, ACM Computing Surveys, 1999.

JOHNSON, R., W., D., **Applied Multivariate Statistical Analysis**, Prentice Hall: 1999.

KAISER, H. F., **The Varimax Criterion for Analytical Rotation in Factor Analysis**, Psychometrika: 1958.

VALE, M. N., **Agrupamentos de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos**, Rio de Janeiro: PUC-Rio, 2005.

WEINSTEIN, A., tradução de RIMOLI, C. A., **Segmentação de Mercado**, São Paulo: Atlas, 1995.

WEISS, S. M., I. N., **Predictive Data Mining**, Morgan Kaufmann Publishers, Inc: 1998