UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

MATHEUS DOS REIS LOHMANN

ANÁLISE DE PERFIS HUMANOS EM CENÁRIOS INDUSTRIAIS E ACADÊMICOS BALIZADA POR FERRAMENTAS MULTIVARIADAS

Porto Alegre
2019

Matheus dos Reis Lohmann

Análise de perfis humanos em cenários industriais e acadêmicos balizada por ferramentas multivariadas

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Gestão de Operações em Universidades Públicas Federais.

Orientador: Michel José Anzanello, *Ph.D.*

Porto Alegre

2019

Matheus dos Reis Lohmann

Análise de perfis humanos em cenários industriais e acadêmicos balizada por ferramentas multivariadas

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel José Anzanello*, Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. Alejandro German Frank, Dr.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor Milad Yousefi, Dr. (PPGEP/UFRGS)

Professor Flávio Sanson Fogliatto, *Ph.D.* (PPGEP/UFRGS)

Professor Alessandro Kahmann, Dr. (IMEF/FURG)

LOHMANN, Matheus dos Reis. ***Análise de perfis humanos em cenários industriais e acadêmicos balizada por ferramentas multivariadas***, 2019. Dissertação (Mestrado em Engenharia) – Universidade Ferderal do Rio Grande do Sul, Brasil.

RESUMO

A aplicação de técnicas multivariadas encontra diversas aplicações práticas tanto na análise de agrupamentos quanto na classificação e desperta interesse nos mais diversos setores. Em segmentos industriais, técnicas multivariadas são tipicamente utilizadas na programação de produção e monitoramento de processos produtivos, mas suas aplicações não ficam restritas a este segmento. Setores como o educacional tem apresentado interesse crescente na aplicação de técnicas multivariada para melhor análise de dados e geração de estratégias. Esta dissertação propõe métodos para a análise de perfis humanos através de ferramentas multivariadas com propósitos de agrupamentos e classificação em diferentes segmentos. Para tal, o primeiro artigo propõe uma estrutura multivariada para formar grupos consistentes de trabalhadores com base em seus padrões de aprendizado. Em termos de sua operacionalização, aplica-se a análise de componentes principais (ACP) a parâmetros oriundos da modelagem de curvas de aprendizagem (CAs) sobre os dados de desempenho de tais trabalhadores; a manipulação dos parâmetros gerou um índice de importância de variável (IIV) que orientou um processo iterativo de remoção das variáveis. Ao aplicar a estrutura proposta a um processo de fabricação de calçados, descobriu-se que apenas 8 dos 29 parâmetros oriundos das CAs foram relevantes na inserção dos trabalhadores em dois grupos distintos por suas características de aprendizagem. Em seguida, no artigo 2, é apresentada uma estrutura para selecionar um subconjunto de parâmetros das CAs com o propósito de classificar trabalhadores de acordo com seus padrões de aprendizado; um índice de importância de parâmetro (IIP) é gerado como base nas saídas das regressões *Partial Least Square* (PLS) e *Least Absolute Shrinkage and Selection Operator* (LASSO). Quando aplicado a dados reais, identificou-se que 3 dos 29 parâmetros originais foram relevantes na classificação dos trabalhadores em duas linhas de produção existentes; a estrutura proposta atingiu 100% de classificações corretas com as três ferramentas de classificação utilizadas. Por fim, o artigo 3 traz uma abordagem multivariada para selecionar as variáveis com maior influência sobre três possíveis desfechos de alunos de graduação:

diplomação, evasão interna (troca de curso dentro da mesma IES) ou evasão externa. Variáveis do perfil acadêmico e dados de desempenho foram analisadas através da técnica “omita uma variável por vez” (OUVV) em conjunto com ferramentas de classificação. Ao ser aplicada a dados de ingressantes em cursos de engenharias, a abordagem obteve acurácia de 91,22%, retendo 22,22% das variáveis originais; destaca-se o fato da maioria dos procedimentos realizados apontar as variáveis de desempenho acadêmico (aprovações e reprovações) como as mais influentes no processo.

Palavras-chave: Análise de cluster, Classificação, Perfis humanos, Índice de importância

LOHMANN, Matheus dos Reis. ***Analysis of human profiles in industrial and academic scenarios using multivariate techniques***, 2019. Dissertation (Master in Engineering) - Federal University of do Rio Grande do Sul, Brazil.

ABSTRACT

The application of multivariate techniques finds several practical applications such as cluster analysis and classification, and generates interest in the most diverse sectors. In industrial segments, multivariate techniques are typically used in production scheduling and production process monitoring, but their applications are not restricted to this segment. Other sectors, such as education, has shown increasing interest in the application of multivariate techniques to do better data analysis and generation strategies. This thesis proposes methods for the analysis of human profiles through multivariate techniques for the purpose of clustering and classification in different segments. For that matter, the first paper proposes a multivariate structure to form consistent groups of workers based on their learning patterns. In terms of its operationalization, principal component analysis (PCA) is applied to parameters derived from learning curve (LC) modeling on the workers' performance data; the manipulation of the parameters generated an importance index that guided an iterative process of variable removal. Applying the proposed structure to a shoe manufacturing process, 8 out of the original 29 LC parameters were deemed relevant for inserting workers into two distinct clusters by their learning characteristics. Next, in paper 2, a framework for selecting a subset of LC parameters is presented for purpose of classifying workers according to their learning patterns; a parameter importance index (PII) is generated based on the outputs of the Partial Least Square (PLS) and Least Absolute Shrinkage and Selection Operator (LASSO) regressions. When applied to real data, it was identified that 3 out of the original 29 parameters were relevant in the classification of workers in two existing production lines; the proposed structure reached 100% correct classifications with three classification techniques used. Finally, paper 3 presents a multivariate approach to select the variables with the greatest influence on three possible outcomes of undergraduate students: graduation, internal dropout or external dropout. Academic profile variables and performance data were analyzed using the "omit one variable at a time" method combined with classification techniques. When applied to data from freshmen in engineering courses, the approach obtained

accuracy of 91.22%, retaining 22.22% of the original variables; most of the procedures performed indicated the academic performance variables as the most influential in the process.

Keywords: Clustering analysis, Classification, Human profiles, Importance index

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. Introdução

1.1 Considerações Iniciais

Análise Multivariada é constituída por um conjunto de métodos estatísticos que podem ser utilizados para analisar simultaneamente múltiplas medidas de um indivíduo ou objeto de uma ou mais amostras (RENCHER, 2002). O crescente número de pesquisas nos mais variados segmentos apoiados em ferramentas multivariadas demostra a sua importância e elevada aplicabilidade nos mais diversos setores e segmentos.

No segmento industrial, técnicas multivariadas têm sido aplicadas com sucesso em setores que vão desde a indústria alimentícia (CALLAO; RUISÁNCHEZ, 2018) até a de combustíveis (SOARES et al., 2017). A análise de perfis humanos através de técnicas multivariadas também não é uma novidade na literatura. Proposições neste sentido foram realizadas utilizando o padrão de aprendizagem de trabalhadores para se obter estimativas de tempos de produção e de fornecimento, levando a planos de produção mais precisos e realistas (ANZANELLO; FOGLIATTO, 2011).

Tornando-se cada vez mais frequente a busca por maior eficiência em todos os segmentos, abre-se espaço para que as técnicas multivariadas também sejam aplicadas a outros segmentos não muito explorados na literatura. Um destes segmentos é o educacional, que apesar de não apresentar números expressivos em termos de aplicação estruturada de ferramentas multivariadas, tem demonstrado um crescimento constante de pesquisas. Percebe-se que o setor educacional brasileiro vem buscando, de diversas maneiras, uma maior eficiência frente aos recursos disponíveis (ROSANO-PEÑA; ALBUQUERQUE; MARCIO, 2012). Reduzir as evasões e aumentar a retenção no ensino superior são alguns dos objetivos do governo federal (BRASIL, 2007) e das instituições de ensinos superior brasileiras, buscando melhor aproveitamento dos recursos aplicados.

De modo similar a análises multivariadas em dados de perfis de aprendizado de trabalhadores, banco de dados de cunho educacionais, como perfis de alunos e seus respectivos desempenhos acadêmicos, abrem um leque de opções a serem analisados por ferramentas com diversos propósitos, conforme apontado por diversos autores (CARVAJAL; CERVANTES, 2017; CHITTUM; JONES; CARTER, 2019; SALES et al., 2016). A aplicação de técnicas

multivariadas aos dados acadêmicos, em nível de educação superior, pode ser útil na identificação de alunos com risco de evasões, bem como entregar um entendimento de quais os principais fatores que influenciam na permanência do aluno até a diplomação ou no seu desligamento precoce.

Independente do segmento em que a análise multivariada é realizada, uma grande quantidade de variáveis pode prejudicar os resultados, de modo que estes sejam imprecisos ou não confiáveis. A seleção de variáveis, a qual visa reduzir a dimensionalidade dos dados a serem analisados, pode proporcionar modelos mais robustos, precisos e interpretáveis, eliminando as variáveis irrelevantes ou ruidosas, e reduzindo o esforço computacional na geração dos modelos (ANDERSEN; BRO, 2010).

A presente dissertação é composta por três artigos que abordam a análise de perfis humanos em diferentes cenários, com propósitos de agrupar e classificar, fazendo uso de ferramentas multivariadas. O primeiro artigo aborda o agrupamento de trabalhadores em linhas de produção homogêneas de acordo com suas características de aprendizado. Os dados de desempenho coletados são modelados através de diferentes curvas de aprendizados (CAs); as saídas da análise de componentes principais (ACP), quando aplicada aos parâmetros oriundos das CAs, dão origem a um índice de importância de variável (IIV), que é utilizado para identificar quais as variáveis são mais relevantes com o propósito de agrupamento. As métricas *Silhouette Index* (SI), *Calinski-Harabasz* (CH) e *Davies-Bouldin* (DB) são utilizadas para verificar as qualidades dos grupos formados através da estrutura proposta. No segundo artigo é proposta uma estrutura multivariada para a seleção de parâmetros, a qual visa classificar trabalhadores de forma consistente em grupos pré-existentes (linhas produtivas ou células de produção) de acordo com suas características de aprendizagem. As regressões PLS e LASSO são aplicadas aos parâmetros oriundos das CAs, e os coeficientes das regressões, juntamente com os coeficientes de determinação (R^2) das regressões, dão origem a um índice de importância de parâmetros (IIP). Na sequência, os trabalhadores são alocados a grupos pré-existentes através de uma ferramenta de classificação, fazendo uso de todos parâmetros. Iterativamente os dados são classificados e o parâmetro de menor importância de acordo com o IIP é removido, repetindo-se até restar apenas um parâmetro. O terceiro artigo propõe uma abordagem multivariada para selecionar as variáveis com maior influência sobre três possíveis

desfechos de alunos de graduação: diplomação, evasão interna (troca de curso dentro da mesma IES) ou evasão externa. As variáveis analisadas compreendem dados do perfil acadêmico dos alunos no momento de ingresso na graduação e do desempenho acadêmico do primeiro ao quarto semestre. A sistemática “omita uma variável por vez” em conjunto com uma ferramenta de classificação são utilizadas com a finalidade de identificar o subconjunto de variáveis que melhor classifica o destino dos graduandos. Cinco ferramentas de classificação distintas são utilizadas para efeito de comparação dos resultados: k-vizinhos mais próximos (KNN), Rede Neural Probabilística (PNN), Análise Discriminante Linear (LDA), Máquina de Suporte Vetorial (SVM) e *Naïve Bayes* (NB).

1.2 Objetivos

O objetivo principal da dissertação é propor sistemáticas de seleção de variáveis com vistas à clusterização e classificação de perfis humanos em contextos distintos (dados oriundos dos meios industrial e acadêmico).

Os seguintes objetivos específicos são apresentados:

- Identificar os parâmetros oriundos da modelagem por CA que dão origem a grupos (*clusters*) consistentes de trabalhadores de acordo com seus perfis de aprendizagem;
- Desenvolver um novo índice de importância para classificar trabalhadores de acordo com seus padrões de aprendizado com base nas regressões LASSO e PLS;
- Identificar as variáveis mais informativas em cenário acadêmico com vistas à classificação de alunos de graduação de acordo com seu desfecho (diplomação, transferência ou evasão).

1.3 Justificativa do Tema e dos Objetivos

Nos âmbitos práticos e teóricos, observa-se o constante interesse no desenvolvimento de novas abordagem que se apoiem em métodos multivariados. Apesar do número relevante de pesquisas disponíveis na literatura, não estão esgotadas todas as aplicações de ferramentas multivariadas sobre dados de perfis humanos. Desta forma, a realização desta pesquisa justifica-se em função de reunir a aplicação de análises multivariadas em dados de perfis humanos de

diferentes segmentos, da área industrial e acadêmica, onde são explorados métodos supervisionados e não supervisionados.

A busca por maior eficiência impacta desde a indústria às universidades. Com propósitos semelhantes aos realizados nesta pesquisa, outros estudos utilizaram dados de perfis humanos no desenvolvimento e aplicação de seus métodos. Pesquisas relacionadas a agrupamentos de trabalhadores foram realizados, entre outros, por Uzumeri e Nembhard (1998) e Stroieke, Fogliatto e Anzanello (2012), com o objetivo de uma formação mais eficiente de equipes ao considerar os perfis de aprendizagem dos trabalhadores. Pesquisas realizadas com o objetivo de identificar alunos com maiores riscos de desligarem-se precocemente de seus cursos de graduação foram desenvolvidas por Kantorski et al. (2015) e Sales et al. (2016), sendo este um dos fatores que o governo federal busca solucionar, de forma à melhorar a eficiência das universidades (BRASIL, 2007). Portanto, os métodos propostos neste trabalho encontram respaldo prático, tanto na formação e manutenção de grupos consistentes de trabalhadores, quanto na identificação dos fatores que influenciam na evasão ou retenção de alunos de graduação.

1.4 Procedimentos Metodológicos

Quanto a natureza, a presente dissertação pode ser classificada como aplicada, uma vez que o método de pesquisa utilizado objetiva gerar conhecimentos para aplicação prática em problemas específicos. Em relação aos objetivos, é tido como exploratória, pois busca a resolução de problemas práticos a partir da análise das hipóteses construídas. Apresenta ainda uma abordagem quantitativa, visto que utiliza ferramentas matemáticas e estatísticas para análise e solução dos problemas apresentados (GIL, 2017; SILVA; MENEZES, 2005).

No primeiro artigo a estrutura multivariada proposta para identificar as variáveis mais relevantes na formação de grupos homogêneos de trabalhadores com base em seus perfis de aprendizagem é composta por quatro etapas operacionais. Inicialmente é feita a seleção dos trabalhadores e a coleta de dados de desempenho. Em seguida, os dados de desempenho são modelados utilizando as diferentes CAs, obtendo o perfil de aprendizagem dos trabalhadores. No terceiro passo é aplicada a análise de componentes principais (ACP) aos parâmetros oriundos das CAs e um índice de importância de variável (IIV) é gerado para orientar a remoção

de parâmetros menos informativos. O quarto passo, iterativamente agrupa os trabalhadores ao aplicar o *Fuzzy C-Means* (FCM), avalia-se a qualidade dos grupos através de três métricas, e remove-se a variável menos informativa, reiniciando a etapa até que apenas uma variável resulte do processo. O subconjunto de variáveis que produzir melhor desempenho nas métricas é escolhido.

O segundo artigo propõe uma estrutura para selecionar um subconjunto reduzido de parâmetros de CAs com o propósito de classificar trabalhadores. As primeiras etapas da estrutura visam selecionar os trabalhadores que terão seus dados de desempenho coletados e modelados de acordo com diferentes CAs. Na sequência, as regressões PLS e LASSO são aplicadas e os parâmetros resultantes dão origem ao Índice de Importância de Parâmetros (IIP). O último passo, iterativamente classifica os trabalhadores utilizando uma ferramenta de classificação e remove o parâmetro de menor importância até restar apenas um único parâmetro. Um gráfico associando os parâmetros removidos e a acurácia de classificação é utilizado no monitoramento do processo de eliminação de variáveis e contribui na definição do melhor subconjunto de parâmetros.

Com o objetivo identificar os fatores mais influentes sobre três possíveis desfechos de alunos de graduação, o terceiro artigo propõe uma abordagem multivariada para selecionar variáveis. A abordagem proposta inicialmente coleta dados de dois subgrupos de variáveis: (i) variáveis do perfil acadêmico do aluno no momento de ingresso na graduação; e (ii) variáveis de desempenho acadêmico dos quatro primeiros semestres. Na sequência, aplica-se a sistemática de seleção de variáveis "omita uma variável por vez" (OUVV) em conjunto com uma ferramenta de classificação na porção de treinamento (T_r). Na sistemática OUVV, a cada repetição omite-se uma das variáveis e as acurácias de classificação são computadas; a variável responsável pela maior acurácia de classificação enquanto omitida é permanentemente removida. Iterações são conduzidas sobre as variáveis remanescentes até restar apenas uma variável. Após a finalização da sistemática OUVV, os melhores subconjuntos de variáveis são selecionados e utilizados na classificação das observações pertencentes à porção de teste (T_s). Cinco ferramentas de classificação são utilizadas individualmente, de modo a comparar o seu desempenho: k-vizinhos mais próximos (KNN), Rede Neural Probabilística (PNN), Análise Discriminante Linear (LDA), Máquina de Suporte Vetorial (SVM) e *Naïve Bayes* (NB).

1.5 Estrutura da Dissertação

A dissertação encontra-se dividida em 5 capítulos. O primeiro capítulo introduz o trabalho, apresentando os objetivos e as justificativas, bem como o método de pesquisa adotado. O capítulo ainda apresenta a estrutura do trabalho e a delimitação do estudo.

No segundo capítulo é apresentado o primeiro artigo, que propõe uma estrutura multivariada para gerar grupos consistentes de trabalhadores baseados em seus perfis de aprendizado, através da integração de modelagem de curva de aprendizado (CA) e análise de cluster (AC). É proposto um Índice de Importância de Variáveis (IIV) baseado nos parâmetros oriundos da Análise de Componentes Principais (ACP), quando aplicada ao conjunto de dados dos parâmetros de CAs. Tal índice orienta um processo iterativo de remoção de parâmetros, ao passo que a cada iteração novos agrupamentos são formados através da técnica *Fuzzy C-Means* (FCM) e um parâmetro é removido. O método proposto é aplicado a um processo de fabricação de calçados e três métricas independentes (SI, CH e DB) são utilizadas para avaliar os agrupamentos formados.

O terceiro capítulo apresenta o segundo artigo, o qual propõe uma estrutura para seleção de parâmetros de CAs com vistas à alocação de trabalhadores em grupos já constituídos (como linhas de produção ou células produtivas já concebidas). Com a finalidade de identificar a relevância dos parâmetros de CAs, um Índice de Importância de Parâmetros (IIP) é proposto com base nos parâmetros oriundos das regressões PLS e LASSO. Inicia-se um processo iterativo de classificações e remoções de parâmetros, guiadas pelo IIP proposto, para seleção dos parâmetros de CAs mais relevantes ao processo de alocação dos trabalhadores. Três ferramentas de classificação, k-vizinhos mais próximos (KNN), Máquina de Vetor de Suporte (SVM) e *Naïve Bayes* (NB), foram utilizadas para fins de comparação dos desempenhos das classificações realizadas no setor de costura de uma indústria calçadista.

O quarto capítulo da dissertação traz o terceiro artigo, que busca identificar as variáveis com maior influência sobre três possíveis desfechos de alunos de graduação: diplomação, evasão interna (troca de curso dentro da mesma IES) ou evasão externa, através de uma abordagem multivariada. A análise é balizada por variáveis que compreendem os dados do perfil acadêmico dos alunos no momento de ingresso no curso de graduação e dados de desempenho acadêmico do primeiro ao quarto semestre dos estudantes. A abordagem proposta

faz uso de ferramentas de classificação integradas à técnica “omita uma variável por vez” para identificar o subconjunto de variáveis que melhor descreve o destino dos graduandos. Cinco ferramentas foram utilizadas de modo distinto, com propósito de comparação de resultados: k-vizinhos mais próximos (KNN), Rede Neural Probabilística (PNN), Análise Discriminante Linear (LDA), Máquina de Suporte Vetorial (SVM) e *Naïve Bayes* (NB). Tal abordagem prioriza um subconjunto reduzido de variáveis juntamente com a melhor acurácia de classificação. Quatro subconjuntos de dados são selecionados, de modo a representar individualmente do primeiro ao quarto semestre acadêmico dos estudantes, e análises são realizadas sobre tais subconjuntos.

O quinto e último capítulo apresenta conclusão do trabalho, onde são avaliados os principais resultados frente aos objetivos traçados e as delimitações do estudo, e sugestões para desdobramentos futuros desta pesquisa.

1.6 Delimitações do Estudo

Constituem-se em restrições do presente estudo:

- O trabalho não irá apresentar novas ferramentas para classificação ou clusterização, restringindo-se a combinar a utilização de tais ferramentas com outros métodos para atingir os objetivos;
- Não são realizadas análises financeiras decorrentes da aplicação das abordagens no contexto industrial; e
- As estruturas propostas são aplicadas apenas aos segmentos propostos, não garantindo resultados conclusivos quando aplicadas a outros segmentos.

1.7 Referências

ANDERSEN, C. M.; BRO, R. Variable selection in regression-a tutorial. **Journal of Chemometrics**, v. 24, n. 11–12, p. 728–737, nov. 2010.

ANZANELLO, M. J.; FOGLIATTO, F. S. Selecting the best clustering variables for grouping mass-customized products involving workers learning. **International Journal of Production Economics**, v. 130, n. 2, p. 268–276, 1 abr. 2011.

BRASIL. **Decreto Nº 6.096, de 24 de abril 2007. Institui o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI**, Brasil, 2007.

CALLAO, M. P.; RUISÁNCHEZ, I. An overview of multivariate qualitative methods for food fraud detection. **Food Control**, v. 86, p. 283–293, 1 abr. 2018.

CARVAJAL, R. A.; CERVANTES, C. T. Aproximaciones a la deserción universitaria en Chile. **Educação e Pesquisa**, v. 44, n. 0, 4 set. 2017.

CHITTUM, J. R.; JONES, B. D.; CARTER, D. M. A person-centered investigation of patterns in college students' perceptions of motivation in a course. **Learning and Individual Differences**, v. 69, p. 94–107, 1 jan. 2019.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 6. ed. São Paulo: Atlas, 2017.

KANTORSKI, G. Z.; HOFFMANN, I. L.; LIMBERGER, S. J.; MULLER, F. Uma Visão Do Futuro: Previsão De Evasão Em Cursos De Graduação Presenciais De Universidades Públicas: O Caso Do Curso De Zootecnia. **XV Colóquio Internacional de Gestão Universitária**, 4 dez. 2015.

RENCHER, A. C. **Methods of multivariate analysis**. [s.l.] J. Wiley, 2002.

ROSANO-PEÑA, C.; ALBUQUERQUE, P. H. M.; MARCIO, C. J. A eficiência dos gastos públicos em educação: evidências georreferenciadas nos municípios goianos. **Economia Aplicada**, v. 16, n. 3, p. 421–443, set. 2012.

SALES, J. S.; BRASIL, G. H..; CARNEIRO, T. C. J.; CORASSA, M. A. C. Fatores Associados à Evasão e Conclusão de Cursos de Graduação Presenciais na UFES. **Meta: Avaliacao**, v. 8, n. 24, p. 488–514, 8 dez. 2016.

SILVA, E. L.; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: Laboratório de Ensino da Universidade Federal de Santa Catarina, 2005.

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, p. 171–178, 15 ago. 2017.

STROIEKE, R. E.; FOGLIATTO, F. S.; ANZANELLO, M. J. Análise de conglomerados em curvas de aprendizado para formação de agrupamentos homogêneos de trabalhadores. **Production**, v. 23, n. 3, p. 537–547, 30 out. 2012.

UZUMERI, M.; NEMBHARD, D. A population of learners: A new way to measure organizational learning. **Journal of Operations Management**, v. 16, n. 5, p. 515–528, 1 out. 1998.

2. Primeiro artigo: Sistemática para agrupamento de trabalhadores com perfis de aprendizado semelhantes em linhas de produção customizadas

Resumo

A Customização em Massa (CM) implica em uma grande variedade de produtos e lotes de produção de tamanhos reduzidos. As tarefas que dependem das habilidades humanas são especialmente afetadas neste contexto, uma vez que os trabalhadores precisam se adaptar rapidamente aos requisitos dos novos modelos. Tal processo ocorre de maneira diferente de acordo com o trabalhador, justificando o desenvolvimento de estratégias para agrupar trabalhadores com comportamentos de aprendizagem semelhantes. Este artigo propõe uma estrutura para formar grupos homogêneos de trabalhadores com base em seus perfis de aprendizagem, integrando modelagem de curva de aprendizado (CA) e análise de *cluster* (AC). Para isso, os dados de desempenho são coletados e modelados através de CAs, de tal forma que os parâmetros do modelo quantifiquem a adaptação dos trabalhadores às tarefas. A Análise de Componentes Principais (ACP) é aplicada ao conjunto de dados que consiste dos parâmetros de CA; as saídas da ACP dão origem a um índice de importância integrado a um processo de seleção de variáveis. Depois que cada parâmetro de CA é removido do conjunto de dados, um novo agrupamento usando a técnica de agrupamento *Fuzzy C-Means* (FCM) é realizado usando os parâmetros das CAs restantes, e a qualidade dos grupos formados é avaliado por meio de três métricas. Quando aplicado a um processo de fabricação de calçados, 8 dos 29 parâmetros originais das CAs foram considerados relevantes para a inserção de trabalhadores em dois grupos. O subconjunto escolhido elevou a qualidade do procedimento de agrupamento em 29,4% (de 0,476 para 0,616) quando avaliado através do *Silhouette Index* (uma das três métricas testadas); melhorias semelhantes foram sugeridas pelas outras métricas de qualidade. Os parâmetros retidos estão relacionados ao desempenho dos trabalhadores na primeira repetição e experiência prévia na tarefa, destacando a importância de programas de treinamento bem projetados.

Palavras-Chaves: Customização em Massa, Curvas de Aprendizado, Índice de Importância de Variáveis

2.1 Introdução

Customização em massa (CM) é definida como a capacidade de fornecer produtos e serviços projetados individualmente para cada cliente, por meio de alta flexibilidade e integração de processos. Ela tem sido adotada por muitas empresas de produção em massa com o objetivo de repensar estratégias de produção em cenários onde os ciclos de vida dos produtos são reduzidos. A definição foi posteriormente refinada por vários autores, que associaram a CM a processos de produção flexíveis voltados à geração de elevados volumes de itens personalizados a custos razoavelmente baixos (DA SILVEIRA; BORENSTEIN; FOGLIATTO, 2001; FOGLIATTO; DA SILVEIRA; BORENSTEIN, 2012; MACCARTHY; BRABAZON; BRAMHAM, 2003).

Diversos produtos customizados são fabricados por meio de operações manuais, geralmente em pequenos lotes (ANZANELLO; FOGLIATTO, 2011a; ANZANELLO; FOGLIATTO; SANTOS, 2014; FOGLIATTO; DA SILVEIRA; BORENSTEIN, 2012). Programações de produção são mais numerosas à medida que as opções de catálogo aumentam, expondo os funcionários a novos modelos, tecnologias e processos de produtos. Em pequenos lotes de produção, os procedimentos exigidos pelos novos produtos tentem a reduzir o rendimento dos trabalhadores e a qualidade do produto devido ao processo de aprendizagem a que os trabalhadores são submetidos (NEMBHARD; UZUMERI, 2000b). Neste contexto, a modelagem de curva de aprendizado (CA) é uma das técnicas disponíveis para superar essas desvantagens da CM.

CAs são representações matemáticas do desempenho de trabalhadores em tarefas repetitivas (ANZANELLO; FOGLIATTO, 2011a, 2011b; FIORETTI, 2007; GROSSE; GLOCK; MÜLLER, 2015; JABER; KHAN, 2010; REID; MIRKA, 2007; TILINDIS; KLEIZA, 2017). As CAs têm sido amplamente utilizadas para estudar mudanças no desempenho dos trabalhadores em função dos efeitos da aprendizagem (PLAZA; ROHLF, 2008). Na análise de produção, modelagem de CAs podem ser usadas para se obter estimativas de tempos de produção e de fornecimento, levando a planos de produção mais precisos (ANZANELLO; FOGLIATTO, 2011a). As CAs também podem auxiliar na caracterização de trabalhadores de acordo com sua adequação a tarefas de diferentes complexidades (NEMBHARD; UZUMERI, 2000b; STROIEKE; FOGLIATTO; ANZANELLO, 2011;

UZUMERI; NEMBHARD, 1998), e planejar lotes de produção quando a aprendizagem é considerada (ANZANELLO; FOGLIATTO; SANTOS, 2014; LIU et al., 2018).

Em cenários onde a produção em larga escala ocorre, os trabalhadores são tipicamente agrupados com base nos tempos médios de execução de tarefas destinadas a minimizar a variação de tempo. Esse curso de ação não é adequado para a maioria dos cenários de CM, que normalmente são caracterizados por pequenos lotes de produção. Assim, uma formação mais eficiente de equipes pode ser obtida ao considerar os perfis de aprendizagem dos trabalhadores, capturando seu desempenho desde o início dos ciclos de produção (ANZANELLO; FOGLIATTO, 2007). Trabalhadores com padrões de aprendizagem semelhantes são então designados à mesma equipe, evitando gargalos nas linhas de produção e reduzindo os níveis de ociosidade. Grupos de trabalhadores com comportamento semelhante de aprendizado também permitem que os gerentes desenvolvam e implantem programas de treinamentos mais específicos, tendo como base os níveis de destreza esperados em cada grupo. Além disso, grupos homogêneos de trabalhadores também incentivam uma melhor atitude entre eles, devido à uniformidade no processo de aprendizagem e desempenho de tarefas.

Apesar de sua relevância e aplicabilidade, um número limitado de autores na literatura de CA aborda o tópico de agrupar trabalhadores de acordo com seu perfil de aprendizagem. Uzumeri e Nembhard (1998) identificaram distintos perfis de trabalhadores em uma empresa de manufatura analisando os parâmetros de CA, destacando como os trabalhadores se adaptam de maneiras diferentes às tarefas. Wong, Cheung e Wu (2010) identificaram diferentes perfis entre os empregados que receberam *feedback* regular e que foram agrupados de acordo com seus parâmetros de CA. Técnicas de agrupamento foram aplicadas a dados de desempenho por Stroieke, Fogliatto e Anzanello (2011) objetivando caracterizar diferentes grupos de trabalhadores, enquanto Grosse e Glock (2015) avaliaram o efeito da aprendizagem dos trabalhadores nos processos de colheita manual. Entende-se que este número reduzido de referências não esgotou as possibilidades de agrupar trabalhadores com base em perfis de aprendizado utilizando técnicas de análise multivariada.

Este artigo propõem uma estrutura multivariada para formar grupos consistentes de trabalhadores com base em seus padrões de aprendizado em linhas de produção personalizadas. A estrutura proposta começa reunindo dados de desempenho dos trabalhadores, que são

ajustados a um conjunto de 10 modelos de CA disponíveis na literatura. Esse processo de modelagem produz um vetor composto por 29 parâmetros que descrevem o perfil de aprendizado do trabalhador. Em seguida, aplica-se a Análise de Componentes Principais (ACP) a esses parâmetros; os resultados emergentes dão origem a um índice que mede a importância de cada parâmetro de CA com o propósito de realizar o agrupamento. Os trabalhadores são inicialmente inseridos em grupos usando todos os parâmetros através da técnica de agrupamento *Fuzzy C-Means* (FCM), e a qualidade do procedimento de agrupamento é avaliada independentemente por meio de três métricas: *Silhouette Index* (SI), *Calinski-Harabasz* (CH) e *Davies-Bouldin* (DB). Em seguida, o parâmetro que apresenta o menor índice de importância é removido do conjunto de dados e os trabalhadores são agrupados com base nos parâmetros restantes. Esse processo iterativo é repetido até que reste um único parâmetro. O subconjunto de parâmetros que produz o melhor desempenho de agrupamento para cada uma das métricas avaliadas é definido.

O método proposto foi aplicado a 70 trabalhadores em um processo de fabricação de calçados. Verificou-se que 8 dos 29 parâmetros originais das CA mostraram-se relevantes para a inserção dos trabalhadores em dois grupos que dependem de diferentes características de aprendizagem. O uso dos parâmetros selecionados elevou a qualidade do agrupamento em 29,4% (de 0,476 para 0,616), de acordo com o *Silhouette Index*; melhorias semelhantes foram obtidas ao usar as outras métricas de avaliação de qualidade. Os parâmetros retidos estão relacionados principalmente ao desempenho dos trabalhadores na primeira repetição e na experiência prévia dos trabalhadores.

2.2 Referencial Teórico

Nesta seção, são fornecidas informações básicas sobre as ferramentas utilizadas no método proposto neste artigo: Curvas de aprendizado (CA), técnicas multivariadas voltadas à redução da dimensão de dados e agrupamento de observações.

CA são modelos de regressão não-linear que associam o desempenho dos trabalhadores (normalmente dado em número de unidades produzidas) e períodos de operação (ANZANELLO; FOGLIATTO, 2007). Os principais modelos de CA da literatura podem ser classificados em três grupos, de acordo com sua estrutura matemática: (*i*) potencial, (*ii*)

exponencial e (*iii*) hiperbólico. A Tabela 2.1 apresenta um resumo dos modelos, definições de parâmetros e aplicativos recomendados. Tais parâmetros serão utilizados como variáveis de agrupamento nas proposições descritas na Seção 2.3.

Tabela 2.1 - Modelos de Curvas de Aprendizado (adaptado de WONG; ON CHEUNG; HARDCASTLE, 2007)

(continua)

Modelo	Equação	Parâmetros	Aplicações	Referências
Modelos Potenciais				
Wright	$Y = Y_1 x^b$ (2.1)	Y = tempo médio (ou custo) por unidade demandada para produzir x unidades; Y1 = tempo (ou custo) para produzir a primeira unidade; b = taxa de aprendizado, onde -1 < b < 0, e b próximo -1 denota alta taxa de aprendizado	Representa o efeito da aprendizagem, como a diminuição do custo médio (ou tempo) à medida que ocorrem repetições de tarefas.	Wright (1936)
Stanford-b	$Y = Y_1 (x + B)^b$ (2.2)	B = número de unidade de experiência prévia *	Incorporar a experiência prévia do trabalhador no início de um ciclo de produção	Anzanello e Fogliatto (2011b); Yeh e Rubin (2012)
DeJong	$Y = Y_1[M + (1 - M)x^b]$ (2.3)	M = fator de incompressibilidade (0 < M <1) que informa a fração da tarefa executada pelas máquinas; quando M =0, obtemos o modelo de Wright *	Incorporar a influência das máquinas no processo de aprendizagem	Anzanello e Fogliatto (2011b)
Curva S	$Y = Y_1[M + (1 - M)(x + B)^b]$ (2.4)	Os parâmetros são definidos nos modelos DeJong e Stanford-B	Descreve a aprendizagem quando ocorre a intervenção de maquinário e os primeiros ciclos de operação exigem uma análise aprofundada	Nembhard e Uzumeri (2000a); Anzanello e Fogliatto (2011b)
Plateau	$Y = C + Y_1 x^b$ (2.5)	C = constante aditiva que descreve o desempenho do trabalhador em estado estacionário *	Descrever o estado estacionário do processo de aprendizagem	Anzanello e Fogliatto (2011b)
Modelos Exponenciais				
Knecht	$Y = Y_1 x^b e^{c\prime x}$ (2.6)	C' = segunda constante *	Integra funções exponenciais e log-lineares para melhorar as previsões em execuções de produção de longa duração	Anzanello e Fogliatto (2011b)
Exponencial de 3 parâmetros	$Y = k(1 - e^{-(x+p)/r})$ (2.7)	Y = desempenho do trabalhador em termos de número de itens produzidos após x unidades de tempo de operação; k = desempenho máximo do trabalhador quando o processo de aprendizagem é concluído, dado em número de itens produzidos por tempo de operação (k ≥0); p= experiência prévia do trabalhador, avaliada em unidades de tempo (p ≥0); r= taxa de aprendizado dada em unidades de tempo	Adequado para situações em que os trabalhadores têm experiência prévia na tarefa	Mazur and Hastie (1978); Anzanello e Fogliatto (2011b)
Tempo constante	$Y = y_c + y_f(1 - e^{-t/\tau})$ (2.8)	yc = desempenho inicial do trabalhador em termos de número de itens produzidos por tempo; yf = máximo desempenho quando a aprendizagem é concluída, dada nas mesmas unidades; t = tempo de operação cumulativa, análogo a x nos modelos anteriores	Recomendado para processos em que a coleta de dados de desempenho é iniciada após uma breve adaptação dos trabalhadores à tarefa	Towill (1990); Dardan; Busch e Sward (2006); Anzanello e Fogliatto (2011b)

Tabela 2.1: Modelos de Curvas de Aprendizado (adaptado de WONG; ON CHEUNG; HARDCASTLE, 2007)

(conclusão)

Modelo	Equação	Parâmetros	Aplicações	Referências
Modelos Hiperbólicos				
Hiperbólico de 2 parâmetros	$Y = k(\frac{x}{x+r})$ (2.9)	Y = número de itens produzidos em x unidades de tempo de operação; k = nível máximo de desempenho; r = taxa de aprendizado	Relaciona o número de unidades em conformidade com o número total de unidades produzidas	Mazur and Hastie (1978)
Hiperbólico de 3 parâmetros	$Y = k(\frac{x+p}{x+r+p})$ (2.10)	p = representa a experiência prévia do trabalhador, expressa em unidades de tempo *	Mesma aplicação que o modelo anterior, mas considerando a experiência anterior do trabalhador	Mazur and Hastie (1978)

* Outros parâmetros no modelo são definidos anteriormente.

O primeiro modelo potencial deve-se a Wright (1936), que observou a redução nos custos de montagem de aviões à medida que as repetições aconteciam. De acordo com a equação (2.1), os custos cumulativos de montagem são reduzidos em 20% quando o número de unidades é duplicado. Dada a sua simplicidade matemática e capacidade de se ajustar aos dados empíricos (ANZANELLO; FOGLIATTO, 2011b), o modelo de Wright tem sido amplamente aplicado na modelagem de aprendizado. Jaber, Bonney e Guiffrida (2010) utilizaram o modelo para analisar uma cadeia de suprimentos onde a produção foi submetida a melhoria contínua, concluindo que considerar os efeitos de aprendizagem no processo de produção é benéfico para toda a cadeia, reduzindo os custos de fornecimento associados. Zorgios, Vlismas e Venieris (2009) usaram o modelo de Wright para descobrir que as variações na taxa de aprendizado são devidas à variabilidade no capital humano, e que os custos envolvidos em um ambiente produtivo podem ser relacionados e monitorados através da análise de CA. Outras aplicações do modelo de Wright podem ser encontradas em Jaber e Glock (2013) e Yeh e Rubin (2012).

O modelo de Wright foi adaptado por vários autores. O modelo de Stanford-b [equação (2.2)] adiciona um parâmetro ao modelo da equação (2.1) para incorporar a experiência prévia do trabalhador na tarefa (YEH; RUBIN, 2012). O modelo de DeJong na equação (2.3) foi desenvolvido para incorporar a influência das máquinas no processo de aprendizagem (NEMBHARD; UZUMERI, 2000a). Mesclando os modelos de Stanford-b e DeJong, o modelo Curva S na equação (2.4) foi proposto para descrever a aprendizagem quando a intervenção de máquinas está presente (NEMBHARD; UZUMERI, 2000a). Para superar o comportamento assintótico do modelo de Wright após um grande número de repetições, o modelo Plateau exibe

uma constante C que representa o desempenho estacionário do trabalhador ao qual a produtividade converge após o término do aprendizado ou limitações de maquinário que bloqueiam o desempenho dos trabalhadores (ANZANELLO; FOGLIATTO, 2011b); ver equação (2.5).

Modelos exponenciais de CA foram criados para melhorar as previsões de produtividade de processos caracterizados por longos períodos de produção. O primeiro modelo exponencial [equação (2.6)] foi proposto por Knecht (1974), combinando funções exponenciais e potenciais. Mazur e Hastie (1978) apresentaram o modelo exponencial de 3 parâmetros [equação (2.7)] para considerar três aspectos no processo de aprendizagem: o desempenho máximo dos trabalhadores quando a aquisição do conhecimento é concluída, a experiência prévia dos trabalhadores na execução da tarefa observada e a taxa de aprendizado determinada pela velocidade da aquisição de destreza. Estudos realizados pelos autores mostram que o modelo é deficiente frente a situações em que os trabalhadores são submetidos a tarefas complexas que exigem uma grande quantidade de novos conhecimentos, mas adequados para representar situações em que os trabalhadores têm experiência prévia na tarefa (ANZANELLO; FOGLIATTO, 2007). Towill (1990) propôs um modelo de tempo constante [equação (2.8)] para representar situações em que a coleta de dados de desempenho é iniciada após uma breve adaptação dos trabalhadores à tarefa. O modelo foi utilizado por Dardan, Busch e Sward (2006) para levar em conta os impactos da aprendizagem nas avaliações de investimento em tecnologia.

Finalmente, Mazur e Hastie (1978) propuseram uma CA que relaciona o número de unidades em conformidade com o número total de unidades produzidas. Sua representação matemática dada na equação (2.9) consiste em uma curva hiperbólica de 2 parâmetros. Com o objetivo de considerar a experiência prévia dos trabalhadores na tarefa, Mazur e Hastie (1978) expandiram o modelo da equação (2.9), propondo o modelo hiperbólico de 3 parâmetros na equação (2.10), que tem sido aplicado em uma variedade de setores industriais. Por exemplo, Wong, On Cheung e Hardcastle (2007) utilizaram o modelo para prever o desempenho de empreiteiros em projetos de construção civil, enquanto Guimarães, Anzanello e Renner (2012) obtiveram uma redução significativa em acidentes e absenteísmo em uma indústria de calçados usando o modelo para projetar a rotação dos trabalhadores em uma linha de montagem.

Analisa-se agora duas técnicas de análise multivariada usadas neste artigo: Análise de Componentes Principais (ACP) e Análise de *Cluster* (AC). ACP é uma técnica estatística utilizada para reescrever um conjunto de observações de variáveis correlacionadas em um conjunto de valores de variáveis lineares não correlacionadas, chamadas componentes principais. Para isso, uma decomposição ortogonal da matriz de covariância (ou correlação) do conjunto de dados deve ser realizada (geralmente uma decomposição de autovalor). Idealmente, o número de componentes principais retidos é menor que o número de variáveis originais, e a redução de dimensionalidade do conjunto de dados original é alcançada. A transformação ortogonal é definida de tal forma que o primeiro componente principal tem a maior variância possível, e cada componente restante tem a maior variância possível, sendo ortogonal aos componentes anteriores (JOLLIFFE, 2002; RENCHER, 2002).

Os Componentes Principais (CP), denotados por z, são compostos de combinações lineares das variáveis originais y; isto é:. $z = a_1\, y_1 + a_2\, y_2 + \cdots + a_n\, y_n$. Os resultados da ACP são geralmente analisados em termos de escores de CP z, que são os valores de variáveis transformadas correspondentes a um determinado ponto de dados, e cargas a, que são os pesos pelos quais as variáveis originais são multiplicadas para obter os scores da ACP. Os scores dão a composição do CP em relação aos pontos de dados, enquanto as cargas dão a mesma composição em relação às variáveis (JOLLIFFE, 2002). O número de CPs a serem retidos em uma ACP pode ser definido ordenando os componentes de acordo com a porcentagem da variância total no conjunto de dados explicado por cada um deles, do maior para o menor, e mantendo os que adicionam a uma variação esperada. Tal operação pode ser realizada usando o *Scree Graph* (RENCHER, 2002).

A Análise de *Cluster* (AC), também conhecida como Análise de Agrupamento, é um método estatístico para agrupar observações em grupos homogêneos com base em seu grau de similaridade (FÁVERO et al., 2009; HAIR et al., 1998). Os agrupamentos resultantes devem ter alta homogeneidade interna e heterogeneidade externa. Quando representadas graficamente, as observações dentro dos grupos estarão próximas, enquanto as observações de diferentes grupos estarão distantes (HAIR et al., 1998). Na AC, variáveis de agrupamento (ou seja, conjuntos de variáveis que caracterizam observações) são especificadas pelo pesquisador, não sendo estimadas empiricamente (FÁVERO et al., 2009; HAIR et al., 1998).

Uma técnica de agrupamento amplamente conhecida é a *Fuzzy C-Mean*s (FCM), em que cada observação tem uma probabilidade de pertencer a um cluster de destino, em vez de pertencer inteiramente a um único *cluster* (como sugerido por outras técnicas de agrupamento não hierárquicas). De tal forma, o FCM calcula um "grau de pertinência" de cada observação a cada grupo. Esse grau é inversamente relacionado à distância de uma observação específica aos centroides dos grupos em torno dessa observação (AHMED et al., 2002). Cada observação é inserida no grupo que apresenta a maior probabilidade de que essa observação pertença, ou seja, o grupo com o maior grau de pertinência. Detalhes adicionais sobre o FCM estão disponíveis em Ahmed et al. (2002) e Nock e Nielsen (2006).

Por fim, a qualidade do procedimento de agrupamento pode ser avaliada por várias métricas. Rousseeuw (1987) propôs o *Silhouette Index* (SI), que mede o grau de similaridade de uma observação com outras no mesmo grupo em comparação com observações no grupo mais próximo a ela. O valor de SI está compreendido no intervalo [-1,+1]; um valor SI próximo a -1 indica uma observação que foi erroneamente atribuída ao grupo, enquanto um valor SI próximo de +1 indica uma observação corretamente atribuída ao grupo. O índice de *Calinski-Harabasz* (CH) (CALINSKI; HARABASZ, 1974), por sua vez, avalia a qualidade do agrupamento com base na soma medida da distância ao quadrado entre e dentro de grupos. Altos valores de CH denotam agrupamentos adequados, pois as observações dentro de um agrupamento estão próximas umas das outras e distantes das observações inseridas em outros grupos. Finalmente, o índice de *Davies-Bouldin* (DB) (DAVIES; BOULDIN, 1979) primeiro calcula uma medida que avalia a dispersão dentro de cada um dos grupos; valores menores são preferíveis, denotando grande proximidade entre observações inseridas no mesmo grupo. Em seguida, o índice DB determina uma segunda medida de separação entre cada par de grupos, que deve ser o maior possível. A melhor partição é obtida minimizando a proporção entre essas duas medidas.

2.3 Método

O método proposto objetiva definir um subconjunto reduzido de parâmetros de Curvas de Aprendizado que conduzam a grupos consistentes de trabalhadores de acordo com seus perfis de aprendizado. A estrutura sugerida segue quatro etapas operacionais: (*i*) seleção de

trabalhadores e coleta de dados de desempenho; (*ii*) modelagem dos dados de desempenho utilizando as diferentes CAs; (*iii*) aplicação da ACP aos parâmetros oriundos das CA e geração do índice de importância, e (*iv*) agrupamento iterativo dos trabalhadores e remoção dos parâmetros menos relevantes. Estes passos são detalhados a seguir.

A primeira etapa do método proposto identifica ($j = 1, \ldots, J$) trabalhadores que terão sus perfis de aprendizagem monitorado. É desejado que tais trabalhadores estejam familiarizados com as operações analisadas. O desempenho pode ser medido em termos de tempo requerido por repetição da tarefa analisada, ou número de unidades produzidas em um período de tempo. Recomenda-se coletar dados de desempenho até que não sejam observadas alterações significativas no desempenho.

Na segundo etapa, os dados de desempenho dos trabalhadores são modelados utilizando as CAs presentes na Tabela 2.1, obtendo os perfis de aprendizagem dos trabalhadores. Os parâmetros de referência podem ser estimados utilizando rotinas de regressão não linear disponíveis em pacotes estatísticos. No procedimento de modelagem, o número de unidades produzidas (ou tempo necessário em cada repetição) é a variável de resposta *y*, enquanto o tempo de trabalho (ou o número de repetições) é a variável independente *x*. Após o procedimento de modelagem ser concluído, cada trabalhador passa a ser caracterizado por um vetor de estimativas de parâmetros de CA derivados dos 10 modelos de CA na Tabela 2.1. Recomenda-se a normalização dos parâmetros para evitar efeitos de escala no processo de agrupamento.

Na terceira etapa aplica-se ACP aos parâmetros de CA organizados em uma matriz que consiste de J trabalhadores (linhas) e P parâmetros (colunas), onde $P = 29$ neste artigo. As saídas consistem nos pesos dos componentes (w_{pa}) e a porcentagem de variação explicada por cada componente retido $(a = 1, \ldots, A)$, λ_a. Em seguida, gera-se um Índice de Importância de Variável (IIV) para orientar a remoção de parâmetros menos informativos com base em w_{pa} e λ_a. O IIV é denotado por v_p, $p = 1, \ldots, P$, gerado pela equação (2.11). Parâmetros com grandes w_{pa} derivados de componentes principais com grande λ_a são preferidos nestas proposições devido sua alta variabilidade (DUDA; HART; STORK, 2001), sugerindo que parâmetros com

maior v_p permitem um melhor agrupamento de trabalhadores de acordo com seus perfis de aprendizagem.

$$v_p = \sum_{a=1}^{A} \lambda_a |w_{pa}| \qquad p = 1, \ldots, P \tag{2.11}$$

A quarta etapa agrupa iterativamente os trabalhadores e remove as variáveis menos informativas. Inicia agrupando os trabalhadores em *t clusters* e aplicando o FCM aos *P* parâmetros. O número de grupos a ser gerado (*t*) pode ser definido por meio de dendograma [ver Rencher (2002)], ou por meio de testes de diferentes *t's* dentro de um intervalo razoável (por exemplo *t*=2 a *t*=5); outros valores de *t* podem ser testados no caso de um grande número de trabalhadores ser agrupado. Em seguida, avalia-se a qualidade dos grupos computando cada uma das três métricas (SI, CH, e DB).

Em seguida, remove-se a variável com o menor v_p, e executa-se um novo procedimento de agrupamento utilizando os parâmetros restantes; recalculam-se então as métricas de qualidade SI, CH, e DB. Esse processo é repetido iterativamente (ou seja, remove-se a próxima variável com menor v_p e agrupam-se os trabalhadores utilizando as variáveis restantes) até que reste apenas uma variável. Este processo de remoção pode ser ilustrado por um perfil que apresenta a qualidade do agrupamento no eixo vertical e o número de parâmetros retidos no eixo horizontal; observe que três perfis são gerados, pois a qualidade de agrupamentos foi avaliada por meio de três métricas diferentes (SI, CH, e DB). Os subconjuntos que produzem o maior SI e CH (e o menor DB) são escolhidos, pois denotam o melhor desempenho de agrupamento. No caso de não haver convergência no número de parâmetros apontados pelas 3 métricas, recomenda-se escolher o subconjunto apontado pela métrica que retém o menor número de parâmetros.

2.4 Estudo de caso

O método proposto foi aplicado a uma fábrica de calçados no sul do Brasil. A maior parte dos 250 modelos de calçados produzidos pela empresa é direcionada para o mercado externo; a taxa de produção é de cerca de 200.000 pares de calçados por mês. A customização em massa está ganhando força em industrias como a deste estudo; consequentemente, os tamanhos de lotes estão diminuindo à medida que a variedade de modelos aumenta. No entanto,

independentemente do modelo que está sendo produzido, a costura – uma operação que é altamente dependente da destreza manual dos trabalhadores – é o gargalo no processo de produção (ANZANELLO; FOGLIATTO; SANTOS, 2014).

Dados de desempenho de 70 trabalhadores foram coletados e analisados usando os 10 modelos de CA da Tabela 2.1. Um total de 29 parâmetros foi gerado ajustando os dados de desempenho de cada trabalhador para estes modelos de CA; parte dos parâmetros estimados estão ilustrados no Apêndice. ACP foi então aplicada aos parâmetros padronizados das CA; 3 componentes principais foram retidos com base no Scree Graph (RENCHER, 2002), representando 78,1% da variabilidade dos dados. A Tabela 2.2 apresenta o índice de importância de variável de CA estimado pela equação (2.11); valores mais altos denotam parâmetros com maior capacidade de estratificação de observações. Quanto ao número de *clusters* a serem formados, uma inspeção visual de um dendrograma (RENCHER, 2002) apontou dois grupos como a melhor alternativa para estratificar os trabalhadores. Todos os experimentos computacionais foram realizados no Matlab 7.8.

Tabela 2.2 - Índice de Importância de Variável (v_p)

Modelo de CA	Parâmetro	v_p	Modelo de CA	Parâmetro	v_p
Stanford-b	$\boldsymbol{B}$	**2.0800**	Exponencial 3-parâmetros	k	0.3348
Curva S	$\boldsymbol{Y_1}$	**1.8689**	Hiperbólico 2-parâmetros	k	0.3324
Stanford-b	$\boldsymbol{Y_1}$	**1.6247**	Hiperbólico 2-parâmetros	r	0.2340
Plateau	$\boldsymbol{Y_1}$	**1.3336**	Tempo Constante	y_c	0.2070
Curva S	$\boldsymbol{B}$	**1.3051**	Tempo Constante	y_f	0.2010
Knecht	$\boldsymbol{Y_1}$	**1.2443**	Plateau	b	0.1277
Wright	$\boldsymbol{Y_1}$	**1.1949**	DeJong	b	0.1273
DeJong	$\boldsymbol{Y_1}$	**1.1547**	Knecht	b	0.1258
Exponencial 3-parâmetros	r	1.1480	Wright	b	0.1245
Tempo Constante	t	1.0574	Knecht	c'	0.1241
Hiperbólico 3-parâmetros	r	0.9467	Curva S	b	0.1229
Exponencial 3-parâmetros	p	0.7902	Stanford-b	b	0.1228
Hiperbólico 3-parâmetros	p	0.4072	DeJong	M	0.1200
Plateau	C	0.3891	Curva S	M	0.1191
Hiperbólico 3-parâmetros	k	0.3816			

O procedimento iterativo descrito na quarta etapa da seção 2.3 foi então realizado. As Figuras 2.1.a a 2.1.c retratam a variação das três métricas para avaliação da qualidade do *cluster* a medida que os parâmetros foram removidos de acordo com a ordem definida pelo índice v_p.

A semelhança entre os perfis é notável: existe um platô (da extremidade direita para a esquerda do gráfico) onde nenhuma variação substancial na qualidade dos grupos formados é observada quando parâmetros são eliminados do conjunto de dados. De acordo com a Tabela 2.2, os parâmetros inicialmente excluídos (menor v_p) estão associados ao efeito de máquinas sobre o processo de aprendizado (parâmetros M da curva S e DeJong CA, que foram prontamente removidos), seguido pelo parâmetro b de taxa de aprendizado dos modelos Stanford-B, Curva S, Knetch, DeJong e Plateau. Em seguida, os parâmetros relacionados ao desempenho final (k) dos trabalhadores das CAs hiperbólicas foram descartados do procedimento de agrupamento.

Maiores SI (Figura 2.1.a) e CH (Figura 2.1.b) foram encontrados quando um conjunto de 8 parâmetros (assinalados em negrito na Tabela 2.2 e discutidos adiante na Tabela 2.3) foram retidos para a formação dos grupos. Os parâmetros retidos estão associados ao desempenho dos trabalhadores na primeira repetição (Y_1) e experiência prévia (B). A remoção adicional de parâmetros levou a uma redução substancial na qualidade de agrupamentos, sugerindo que as informações mais relevantes dependem dos parâmetros Y_1 e B para agrupar os trabalhadores de acordo com a sua adaptação às tarefas.

Embora um índice menor de DB seja alcançado quando um único parâmetro é retido, a Figura 2.1.c mostra que o segundo menor valor é atingido quando os mesmos 8 parâmetros escolhidos por SI e CH são mantidos. Assim, entende-se que as três métricas são condizentes ao sugerir 8 como o número ideal de parâmetros a serem retidos para fins de agrupamentos.

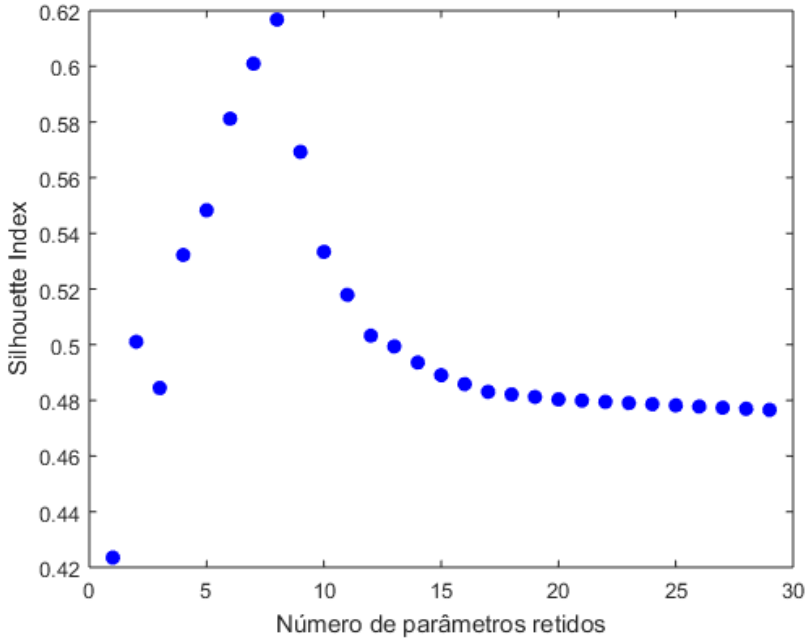

Figura 2.1.a - Perfil do índice Silhouette (SI) com a remoção dos parâmetros de acordo com v_p

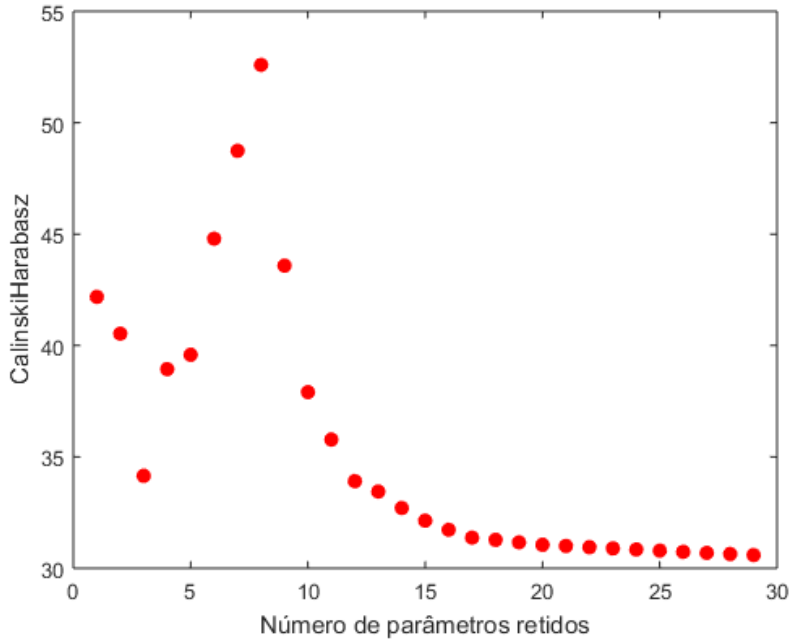

Figura 2.1.b - Perfil do índice Calinski-Harabasz (CH) com a remoção dos parâmetros de acordo com v_p

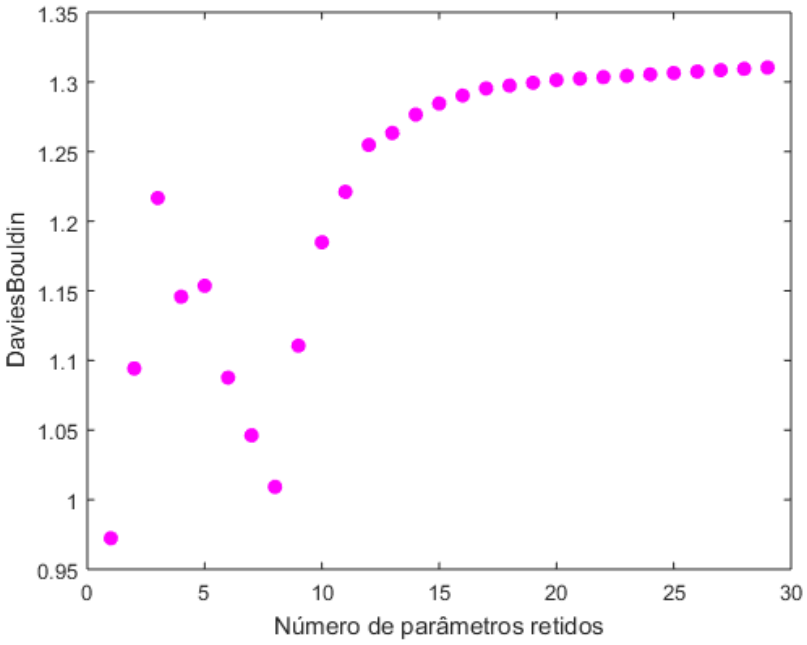

Figura 2.1.c - Perfil do índice Davies-Bouldin (DB) com a remoção dos parâmetros de acordo com v_p

A Tabela 2.3 mostra a média dos 8 parâmetros retidos para cada um dos dois grupos; 46 trabalhadores foram inseridos no grupo 1 e 24 no grupo 2. Há uma diferença substancial em tais parâmetros ao avaliar cada *cluster* formado, corroborando os distintos perfis de aprendizagem dos trabalhadores pertencentes a cada grupo. O *cluster* 1 é formado principalmente por trabalhadores que tendem a exigir mais tempo para concluir o primeiro ciclo de produção, quando comparados aos trabalhadores inseridos no *cluster* 2 (ou seja, todos os parâmetros Y_1 na Tabela 2.3 assumem valores mais altos para o *cluster* 1, com exceção da CA Plateau). O parâmetro B sugere que os membros pertencentes ao *cluster* 2 apresentam uma experiência prévia maior do que aqueles inseridos no *cluster* 1. Tais resultados sugerem que o grupo 2 é principalmente formado por trabalhadores com aprendizagem mais rápida; nenhuma conclusão a respeito do desempenho final é possível, uma vez que tais parâmetros não foram considerados relevantes pela estrutura proposta.

Tabela 2.3 - Média dos parâmetros retidos para cada cluster

CA	Parâmetro	*Cluster* 1	*Cluster* 2
Stanford-b	B	51,6	139,3
Curva S	Y_1	168,2	121,7
Stanford-b	Y_1	196,6	129,2
Plateau	Y_1	158,6	174,4
Curva S	B	23,8	80,3
Knecht	Y_1	170,8	85,0
Wright	Y_1	179,0	95,6
DeJong	Y_1	182,9	98,8

A partir de uma perspectiva gerencial, a prevalência de tais parâmetros para os agrupamentos de trabalhadores enfatiza a relevância da oferta de programas de formação adequada, de forma que os primeiros ciclos produtivos sejam bem sucedidos (que tendem a refletir sobre valores pequenos de Y_1). Além disso, a experiência prévia também é considerada como um parâmetro relevante para o agrupamento de trabalhadores. Surpreendentemente, os parâmetros relacionados à taxa de aprendizado e desempenho final não foram retidos neste estudo, sugerindo que os trabalhadores avaliados tendem a apresentar padrões semelhantes em tais dimensões de aprendizagem.

A fim de visualizar os agrupamentos gerados pelos 8 parâmetros selecionados, representou-se graficamente os três primeiros componentes principais através da ACP (ver Figura 2.2). Os três componentes retidos explicam 100% da variabilidade dos dados, capturando todas as informações possíveis para esta análise. A Figura 2.2 sugere uma separação clara entre os dois grupos de trabalhadores gerados, atestando as diferenças entre os trabalhadores avaliados em termos de seus perfis de aprendizagem.

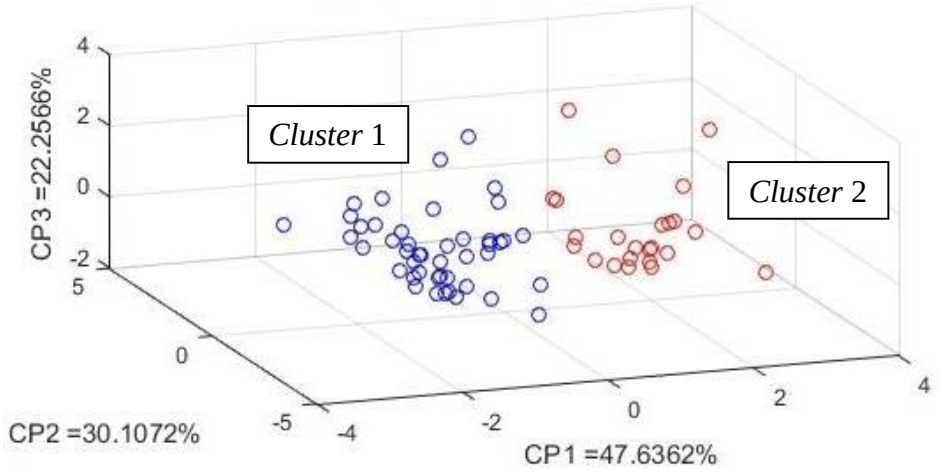

Figura 2.2 - Separação visual dos *clusters* gerados

Finalmente, a Figura 2.3.a retrata o gráfico *Silhouette* quando todos os parâmetros originais são utilizados na formação dos grupos, enquanto que a Figura 2.3.b denota o agrupamento utilizando os 8 parâmetros selecionados (neste gráfico, cada barra horizontal representa um trabalhador agrupado; quanto mais próximo de 1, maior a precisão do agrupamento). Há um incremento substancial na qualidade do agrupamento, já que o subconjunto selecionado conduz a um aumento de 29,4% do SI médio (de 0.476 para 0.616). Comportamento semelhante foi observado para as métricas CH e DB. Tais resultados corroboram a importância de selecionar os parâmetros mais informativos para garantir uma formação robusta de agrupamentos de trabalhadores.

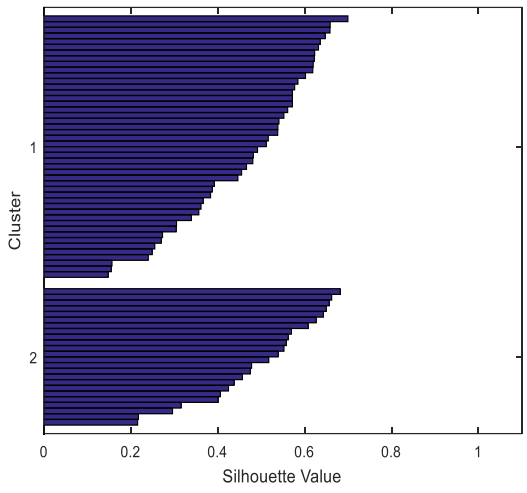

Figura 2.3.a - Gráfico *Silhouette* utilizando os 29 parâmetros originais

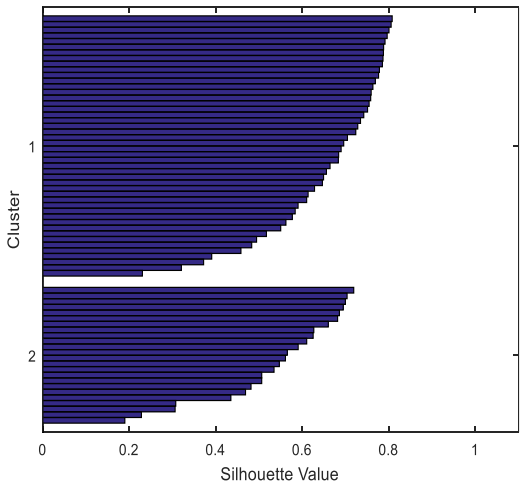

Figura 2.3.b - Gráfico *Silhouette* utilizando os 8 parâmetros selecionados

2.5 Conclusão

Este artigo apresentou um novo método para obter grupos homogêneos de trabalhadores de acordo com seus perfis de aprendizagem. O método integra modelagem de curva de aprendizado e análise de agrupamentos em suas etapas operacionais. A partir de modelos de CA, obtiveram-se parâmetros que descrevem diferentes aspectos da aprendizagem dos trabalhadores. Os parâmetros mais relevantes para o agrupamento foram selecionados com base em um índice de importância de variáveis derivado dos parâmetros da ACP. A qualidade do agrupamento foi avaliada usando três métricas: *Silhouette Index* (SI), *Calinski-Harabasz* (CH) e *Davies-Bouldin* (DB).

Quando aplicado a dados reais de um fabricante de calçados, a estrutura proposta sugeriu que os trabalhadores fossem inseridos em dois grupos, apresentando distintos perfis de aprendizagem. Além disso, verificou-se que 8 dos 29 parâmetros originais das CA foram relevantes para o agrupamento dos trabalhadores; tal subconjunto de parâmetros elevou a qualidade do agrupamento de 0,476 a 0,616, quando medido pelo SI. Os parâmetros retidos estão relacionados com o desempenho dos trabalhadores na produção da primeira unidade de um novo ciclo e sua experiência prévia, destacando a importância da realização de programas de treinamento bem projetados. Um gráfico com 3 CPs corroborou a capacidade do método de separar os trabalhadores de acordo com o seu perfil de aprendizagem.

Como desdobramentos futuros, objetiva-se utilizar transformações Kernel previamente ao procedimento de agrupamento para melhorar a estratificação dos trabalhadores. O uso de escores da ACP em vez dos parâmetros originais da CA como variável de agrupamento também é promissor a fim de evitar o efeito de variáveis correlacionadas no processo de agrupamento.

2.6 Referências

AHMED, M. N.; YAMANY, S. N.; MOHAMED, N.; FARAG, A.A.; Moriarty, T. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. **IEEE Transactions on Medical Imaging**, v. 21, n. 3, p. 193–199, mar. 2002.

ANZANELLO, M. J.; FOGLIATTO, F. S. Learning curve modelling of work assignment in mass customized assembly lines. **International Journal of Production Research**, v. 45, n. 13, p. 2919–2938, jul. 2007.

ANZANELLO, M. J.; FOGLIATTO, F. S. Selecting the best clustering variables for grouping mass-customized products involving workers learning**. International Journal of Production Economics**, v. 130, n. 2, p. 268–276, 1 abr. 2011a.

ANZANELLO, M. J.; FOGLIATTO, F. S. Learning curve models and applications: Literature review and research directions. **International Journal of Industrial Ergonomics**, v. 41, n. 5, p. 573–583, set. 2011b.

ANZANELLO, M. J.; FOGLIATTO, F. S.; SANTOS, L. Learning dependent job scheduling in mass customized scenarios considering ergonomic factors. **International Journal of Production Economics**, v. 154, p. 136–145, ago. 2014.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, v. 3, n. 1, p. 1–27, 1974.

DA SILVEIRA, G.; BORENSTEIN, D.; FOGLIATTO, F. S. Mass customization: Literature review and research directions. **International Journal of Production Economics**, v. 72, n. 1, p. 1–13, 30 jun. 2001.

DARDAN, S.; BUSCH, D.; SWARD, D. An application of the learning curve and the nonconstant-growth dividend model: IT investment valuations at Intel® Corporation. **Decision Support Systems**, v. 41, n. 4, p. 688–697, maio 2006.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-1, n. 2, p. 224–227, abr. 1979.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [s.l.] Wiley, 2001.

EPPLE, D.; ARGOTE, L.; DEVADAS, R. Organizational Learning Curves: A Method for Investigating Intra-Plant Transfer of Knowledge Acquired Through Learning by Doing. **Organization Science**, v. 2, n. 1, p. 58–70, 1991.

FÁVERO, L. P. L.; BELFIORE, P. P.; SILVA, F. L.; CHAN, B. L. **Análise de Dados - Modelagem Multivariada para Tomada de Decisões**. 3a ed. Rio de Janeiro: Elsevier, 2009.

FIORETTI, G. The organizational learning curve. **European Journal of Operational Research**, v. 177, n. 3, p. 1375–1384, 16 mar. 2007.

FOGLIATTO, F. S.; DA SILVEIRA, G. J. C.; BORENSTEIN, D. The mass customization decade: An updated review of the literature**. International Journal of Production Economics**, v. 138, n. 1, p. 14–25, 1 jul. 2012.

GROSSE, E. H.; GLOCK, C. H. The effect of worker learning on manual order picking processes. **International Journal of Production Economics**, v. 170, p. 882–890, 1 dez. 2015.

GROSSE, E. H.; GLOCK, C. H.; MÜLLER, S. Production economics and the learning curve: A meta-analysis. **International Journal of Production Economics**, v. 170, p. 401–412, 1 dez. 2015.

GUIMARÃES, L. B. D. M.; ANZANELLO, M. J.; RENNER, J. S. A learning curve-based method to implement multifunctional work teams in the Brazilian footwear sector. **Applied Ergonomics**, v. 43, n. 3, p. 541–547, 1 maio 2012.

HAIR, J.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E.; THATAM, R.L. **Multivariate Data Analysis**. [s.l.] Prentice Hall, 1998.

JABER, M. Y.; BONNEY, M.; GUIFFRIDA, A. L. Coordinating a three-level supply chain with learning-based continuous improvement. **International Journal of Production Economics**, v. 127, n. 1, p. 27–38, 1 set. 2010.

JABER, M. Y.; GLOCK, C. H. A learning curve for tasks with cognitive and motor elements. **Computers and Industrial Engineering**, v. 64, n. 3, p. 866–871, 1 mar. 2013.

JABER, M. Y.; KHAN, M. Managing yield by lot splitting in a serial production line with learning, rework and scrap. **International Journal of Production Economics**, v. 124, n. 1, p. 32–39, 1 mar. 2010.

JOLLIFFE, I. T. T. **Principal Component Analysis, Second Edition**. [s.l: s.n.].

KNECHT, G. R. COSTING, TECHNOLOGICAL GROWTH AND GENERALIZED LEARNING CURVES. **Operational Research Quarterly**, v. 25, n. 3, p. 487–491, 19 set. 1974.

LIU, C.; WANG, C.; ZHENG,Z.; ZHENG, L. Scheduling with job-splitting considering learning and the vital-few law. **Computers and Operations Research**, v. 90, p. 264–274, 1 fev. 2018.

MACCARTHY, B.; BRABAZON, P. G.; BRAMHAM, J. Fundamental modes of operation for mass customization. **International Journal of Production Economics**, v. 85, n. 3, p. 289–304, 11 set. 2003.

MAZUR, J. E.; HASTIE, R. Learning as accumulation: A reexamination of the learning curve. **Psychological Bulletin**, v. 85, n. 6, p. 1256–1274, 1978.

NEMBHARD, D. A.; UZUMERI, M. V. An individual-based description of learning within an organization. **IEEE Transactions on Engineering Management**, v. 47, n. 3, p. 370–378, 2000a.

NEMBHARD, D. A.; UZUMERI, M. V. Experiential learning and forgetting for manual and cognitive tasks. **International Journal of Industrial Ergonomics**, v. 25, n. 4, p. 315–326, 1 maio 2000b.

NOCK, R.; NIELSEN, F. On weighting clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 28, n. 8, p. 1223–1235, ago. 2006.

PLAZA, M.; ROHLF, K. Learning and performance in ERP implementation projects: A learning-curve model for analyzing and managing consulting costs. **International Journal of Production Economics**, v. 115, n. 1, p. 72–85, 1 set. 2008.

REID, S. A.; MIRKA, G. A. Learning curve analysis of a patient lift-assist device**. Applied Ergonomics**, v. 38, n. 6, p. 765–771, nov. 2007.

RENCHER, A. C. **Methods of Multivariate Analysis** . New York: [s.n.].

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1 nov. 1987.

STROIEKE, R. E.; FOGLIATTO, F. S.; ANZANELLO, M. J. FORMAÇÃO DE AGRUPAMENTOS HOMOGÊNEOS DE TRABALHADORES ATRAVÉS DE CURVAS DE APRENDIZADO. **Pré-anais XLIII Simpósio Brasileiro de Pesquisa Operacional**, 2011.

STROIEKE, R. E.; FOGLIATTO, F. S.; ANZANELLO, M. J. Estado da arte das aplicações de curvas de aprendizado. **Gestão & Produção**, v. 20, n. 3, p. 681–694, 2013.

TILINDIS, J.; KLEIZA, V. Learning curve parameter estimation beyond traditional statistics. **Applied Mathematical Modelling**, v. 45, p. 768–783, 1 maio 2017.

TOWILL, D. R. Forecasting learning curves. **International Journal of Forecasting**, v. 6, n. 1, p. 25–38, 1 jan. 1990.

UZUMERI, M.; NEMBHARD, D. A population of learners: A new way to measure organizational learning. **Journal of Operations Management**, v. 16, n. 5, p. 515–528, 1 out. 1998.

WONG, P. S. P.; CHEUNG, S. O.; WU, R. T. H. Learning from project monitoring feedback: A case of optimizing behavior of contractors. **International Journal of Project Management**, v. 28, n. 5, p. 469–481, jul. 2010.

WONG, P. S. P.; ON CHEUNG, S.; HARDCASTLE, C. Embodying Learning Effect in Performance Prediction. **Journal of Construction Engineering and Management**, v. 133, n. 6, p. 474–482, jun. 2007.

WRIGHT, T. P. Factors Affecting the Cost of Airplanes. **Journal of the Aeronautical Sciences**, v. 3, n. 4, p. 122–128, fev. 1936.

YEH, S.; RUBIN, E. S. A review of uncertainties in technology experience curves. **Energy Economics**, v. 34, n. 3, p. 762–771, 2012.

ZORGIOS, Y.; VLISMAS, O.; VENIERIS, G. A learning curve explanatory theory for team learning valuation. **VINE**, v. 39, n. 1, p. 20–39, 10 abr. 2009.

Apêndice – Parâmetros de CAs (Apresentação parcial)

Trabalhador	Wright	Stanford-b	DeJong	Curva S	Plateau	Knecht	Exponencial de 3 parâmetros	Tempo Constante	Hiperbólico de 2 parâmetros	Hiperbólico de 3 parâmetros
1	Y_1= 90,49 b = -0,15	Y_1= 97,9 B = 1,2 b = -0,17	Y_1= 90,49 M = 0 b = -0,15	Y_1= 97,9 M = 0 B = 1,2 b = -0,17	C = 0 Y_1= 90,39 b = -0,15	Y_1= 90,49 b = -0,15 c'= 0	k = 14,78 p = 36,1 r = 35,09	y_c= 7,11 y_f= 7,68 t = 55,09	k = 13,96 r = 8,4	k = 17,6 p = 22,86 r = 36,76
2	Y_1 = 246,67 b = -0,27	Y_1= 250,2 B = 0,08 b = -0,28	Y_1= 257,57 M = 0,12 b = -0,36	Y_1= 250,19 M = 0 B = 0,08 b = -0,28	C = 30,45 Y_1= 227,11 b = 0,36	Y_1= 221 b = -0,24 c'= 0	k = 8,46 p = 17,23 r = 61,18	y_c= 2,08 y_f= 6,38 t = 61,18	k = 9,28 r = 28,98	k = 10,19 p = 4,5 r = 45,09
3	Y_1 = 60,7 b = -0,09	Y_1= 70,78 B = 5,58 b = -0,12	Y_1= 60,7 M = 0 b = -0,09	Y_1= 70,78 M = 0 B = 5,58 b = -0,12	C = 0 Y_1= 60,67 b = -0,09	Y_1= 56 b = -0,06 c'= 0	k = 16,37 p = 49,31 r = 50,96	y_c= 10,15 y_f= 6,22 t = 50,96	k = 15,75 r = 4,97	k = 18,66 p = 29,55 r = 26,28
4	Y_1 = 334,95 b = -0,26	Y_1= 105,23 B = 11,68 b = -0,53	Y_1= 334,95 M = 0 b = -0,26	Y_1= 105,23 M = 0 B = 11,68 b = -0,53	C = 0 Y_1= 328,05 b = -0,25	Y_1= 298 b = -0,22 c'= 0	k = 7,35 p = 38,54 r = 157,89	y_c= 1,59 y_f= 5,58 t = 157,89	k = 6,99 r = 56,61	k = 11,15 p = 37,12 r = 220,62
5	Y_1 = 202,08 b = -0,12	Y_1= 324,31 B = 15,03 b = -0,22	Y_1= 202,08 M = 0 b = -0,12	Y_1= 324,28 M = 1 B = 15,03 b = -0,22	C = 0 Y_1= 201,58 b = -0,12	Y_1= 185 b = -0,09 c'= 0	k = 7,15 p = 168,55 r = 280,79	y_c= 3,23 y_f= 3,93 t = 280,79	k = 5,25 r = 12,06	k = 8,06 p = 105,2 r = 167,02
6	Y_1 = 155,83 b = -0,11	Y_1= 155,83 B = 0 b = -0,11	Y_1= 170,87 M = 0,55 b = -0,5	Y_1= 155,83 M = 0 B = 0 b = -0,11	C = 94,3 Y_1= 76,57 b = -0,5	Y_1= 183 b = -0,19 c'= 0	k = 11,43 p = 117,75 r = 281,09	y_c= 3,91 y_f= 7,52 t = 281,16	k = 6,48 r = 8,82	k = 11,38 p = 72,52 r = 146,82
7	Y_1 = 205,84 b = -0,08	Y_1= 278,82 B = 15,09 b = -0,14	Y_1= 205,84 M = 0 b = -0,08	Y_1= 279,97 M = 0 B = 8,85 b = -0,16	C = 0 Y_1= 205,75 b = -0,08	Y_1= 196 b = -0,06 c'= 0	k = 5,08 p = 115,37 r = 142,29	y_c= 2,82 y_f= 2,26 t = 142,29	k = 4,19 r = 6,31	k = 6,45 p = 94,85 r = 122,73
8	Y_1 = 101,43 b = -0,18	Y_1= 101,43 B = 0 b = -0,18	Y_1= 108,81 M = 0,27 b = -0,35	Y_1= 108,81 M = 0,27 B = 1 b = -0,35	C = 29,69 Y_1= 79,12 b = -0,35	Y_1= 109,57 b = -0,24 c'= 0	k = 12,6 p = 0 r = 12,3	y_c= -2,19 y_f= 14,84 t = 12,32	k = 14,87 r = 11,54	k = 13,63 p = -6 r = 5,1
9	Y_1 = 190,75 b = -0,25	Y_1= 314,89 B = 4,28 b = -0,38	Y_1= 190,75 M = 0 b = -0,25	Y_1= 314,89 M = 0 B = 4,28 b = -0,38	C = 0 Y_1= 187,53 b = -0,24	Y_1= 85,49 b = -0,02 c'= 0	k = 10,49 p = 14,91 r = 60,57	y_c= 2,29 y_f= 8,2 t = 60,57	k = 11,3 r = 29,57	k = 14,7 p = 12,98 r = 71,85
10	Y_1 = 184,77 b = -0,28	Y_1= 220,81 B = 1,13 b = -0,33	Y_1= 184,77 M = 0 b = -0,28	Y_1= 220,81 M = 0 B = 1,13 b = -0,33	C = 0 Y_1= 183,13 b = -0,28	Y_1= 163 b = -0,25 c'= 0	k = 15,34 p = 13,85 r = 82,91	y_c= 2,36 y_f= 12,98 t = 82,91	k = 15,03 r = 38,08	k = 19,7 p = 8,35 r = 79,91

3. Segundo artigo: Seleção de variáveis para alocação de trabalhadores a linhas de produção existentes com base em seus perfis de aprendizagem

Resumo

A manufatura de produtos personalizados implica no aumento da variedade de modelos e na redução do tamanho dos lotes de produção, impactando na adaptação dos trabalhadores aos requisitos dos novos modelos. De tal forma, estratégias capazes de alocar trabalhadores a grupos pré-existentes (linhas produtivas ou células de produção) de acordo com seu perfil de aprendizagem passam a ser entendidas como relevantes. Este artigo propõe uma estrutura para selecionar um subconjunto de parâmetros oriundos da modelagem por curvas de aprendizado (CAs) com vistas a classificar trabalhadores em agrupamentos homogêneos em termos de seus perfis de aprendizado. Para tanto, os dados de desempenho de trabalhadores atuando sobre novos modelos de produtos são coletados e avaliados através de CAs; os parâmetros oriundos da modelagem permitem avaliar a adaptação dos trabalhadores às tarefas. Na sequência, as regressões *Partial Least Square* (PLS) e *Least Absolute Shrinkage and Selection Operator* (LASSO) são aplicadas aos parâmetros oriundos das CAs, e os coeficientes das regressões, juntamente com seus coeficientes de determinação, dão origem a um índice de importância de parâmetros (IIP), permitindo medir a relevância de cada um dos parâmetros para a alocação dos trabalhadores. Três ferramentas de classificação são testadas no método proposto: k-vizinhos mais próximos (KNN), *Naïve Bayes* (NB) e Máquina de Suporte Vetorial (SVM). Inicia-se então um processo iterativo de seleção dos parâmetros mais relevantes, o qual integra a ordem sugerida pelo IIP a uma sistemática de seleção do tipo backward (ou seja, um parâmetro é removido a cada iteração). O subconjunto de parâmetros que conduz à alocação mais acurada é escolhido. Quando aplicado ao setor de costura de uma indústria calçadista, 3 dos 29 parâmetros das CAs foram considerados relevantes para a inserção de trabalhadores em dois grupos já existentes, conduzindo a classificações 100% acuradas na porção de teste.

Palavras-Chaves: Classificação, Curvas de Aprendizado, Índice de Importância de Parâmetros

3.1 Introdução

A customização em massa (CM) refere-se à capacidade de fornecer produtos e serviços projetados individualmente para cada cliente (DA SILVEIRA; BORENSTEIN; FOGLIATTO, 2001; FOGLIATTO; DA SILVEIRA; BORENSTEIN, 2012; MACCARTHY; BRABAZON; BRAMHAM, 2003), o que oferece desafios à gestão de operações. Dentre eles destacam-se a exposição constante de trabalhadores a novos modelos de produtos, os quais modificam constantemente as opções do catálogo e acarretam redução no rendimento e qualidade dos novos produtos (NEMBHARD; UZUMERI, 2000b). Novas estratégias e técnicas surgem para superar esses desafios da CM. A modelagem de curva de aprendizado (CA) é uma destas técnicas, conforme cita Azevedo e Anzanello (2015).

CAs são representações matemáticas do desempenho de trabalhadores em tarefas repetitivas (ANZANELLO; FOGLIATTO, 2011a; FIORETTI, 2007; GROSSE; GLOCK; MÜLLER, 2015; JABER; KHAN, 2010a; REID; MIRKA, 2007; TILINDIS; KLEIZA, 2017). Estudos relacionados a mudanças no desempenho de trabalhadores levando em consideração os efeitos da aprendizagem também têm se apoiado em CAs (PLAZA; ROHLF, 2008). São numerosos os benefícios de aplicar o conceito de curva de aprendizado no gerenciamento de produção e operações, como melhorias de desempenho de trabalhadores no que diz respeito a repetições e experiência nas tarefas (Glock et al. 2018).

Na literatura encontram-se diversos estudos nos quais técnicas de agrupamentos de trabalhadores apoiam-se em parâmetros provenientes de curvas de aprendizado. Uzumeri e Nembhard (1998), através da análise dos parâmetros gerados após a modelagem de CAs, identificaram perfis de trabalhadores e evidenciaram suas diferenças na adequação de tarefas. Diferentes perfis de trabalhadores também foram identificados por Wong, Cheung e Wu (2010) através da análise e agrupamento dos parâmetros de CAs. Com o objetivo de caracterizar diferentes grupos de trabalhadores, Stroieke, Fogliatto e Anzanello (2011) e Azevedo e Anzanello (2015) também aplicaram técnicas de agrupamentos a dados oriundos de curvas de aprendizado. Técnicas de classificação com intuito de, no entanto, não foram relatadas na literatura.

Embora possa-se utilizar os parâmetros derivados de um único modelo de curva de aprendizado com vistas à classificação dos trabalhadores, tais parâmetros podem não ser capazes de definir com precisão as diferenças entre os grupos. Neste sentido, a utilização de parâmetros resultantes de vários modelos de CA tende a revelar mais detalhes sobre o processo de aprendizagem dos trabalhadores, melhorando a acurácia de classificação. Por outro lado, nem todos os parâmetros obtidos após o processo de modelagem são relevantes (ou apresentam a mesma relevância) para caracterizar o processo de aprendizagem dos trabalhadores (LOHMANN et al., 2019). Assim, identificar o subconjunto de parâmetros mais informativos com vistas à alocação de trabalhadores a grupos pré-existentes torna-se uma tarefa extremamente importante.

A sistemática proposta neste artigo objetiva selecionar um subconjunto informativo de parâmetros de CAs com vistas a alocar novos trabalhadores em grupos de trabalho já existentes, garantindo a homogeneidade de tais grupos no que diz respeito aos seus perfis de aprendizado. A estrutura proposta inicialmente reúne dados de desempenho dos trabalhadores; tais dados são então ajustados a um conjunto de 10 modelos de CA da literatura. Esse processo de modelagem produz um vetor composto por 29 parâmetros de CAs, os quais descrevem o perfil de aprendizado do trabalhador. Em seguida, as regressões PLS e LASSO são aplicadas sobre esses parâmetros; os coeficientes das regressões, juntamente com os coeficientes de determinação (R^2) das regressões, dão origem a um índice que mede a importância de cada parâmetro da CA. Na sequência, os trabalhadores são alocados a grupos pré-existentes utilizando todos os parâmetros através da técnica de classificação k-vizinhos mais próximos (KNN). Em seguida, o parâmetro de CA que apresenta o menor índice de importância é removido do conjunto de dados e os trabalhadores são classificados com base nos parâmetros restantes. Esse processo iterativo é repetido até que reste um único parâmetro. O subconjunto de parâmetros que conduz à melhor acurácia é definido, e estes parâmetros são discutidos. Por fim, os procedimentos acima são repetidos substituindo-se KNN por *Naïve Bayes* (NB) e Máquina de Vetor de Suporte (SVM), com vistas a identificar a melhor técnica de classificação.

A estrutura proposta foi aplicada a 300 trabalhadores do setor de costura em uma indústria de calçados. 3 dos 29 parâmetros originais das CA mostraram-se relevantes para a classificação dos trabalhadores em dois grupos de trabalhadores (duas linhas de produção

existentes). O uso dos parâmetros selecionados elevou a predição das classificações, alcançando 100% de classificações corretas com as três técnicas de classificação utilizadas. Em termos qualitativos, os parâmetros retidos dizem respeito ao desempenho dos trabalhadores na primeira repetição do ciclo produtivo e na experiência prévia dos trabalhadores.

O restante do artigo está organizado da seguinte maneira. A seção 2 apresenta uma introdução sobre Curvas de Aprendizado e seus principais modelos, assim como um breve referencial sobre regressões e ferramentas de classificações. Na seção 3 é apresentado o método desenvolvido com o propósito de classificar trabalhadores de acordo com seu perfil de aprendizagem. A seção 4 apresenta um estudo de caso onde o método é aplicado em uma indústria calçadista e os resultados obtidos. A seção 5 traz a conclusão do artigo e sugestões de pesquisas futuras.

3.2 Referencial Teórico

3.2.1 Curvas de Aprendizado (CA)

Curvas de Aprendizado (CA) são representações matemáticas do desempenho de um trabalhador submetido a repetições de operações ou tarefas manuais (ANZANELLO; FOGLIATTO, 2007; JABER; EL SAADANY, 2011). O trabalhador realiza as tarefas em menor tempo, de acordo com as repetições que são efetuadas, dando-se o fato pela adaptação às ferramentas utilizadas nas tarefas ou pela descoberta de atalhos para a realização das mesmas (ANZANELLO; FOGLIATTO, 2007; WRIGHT, 1936). Há décadas as CAs têm sido valiosas ferramentas de gerenciamento, sendo aplicadas em diversos setores para prever e monitorar o desempenho de indivíduos, grupos de indivíduos e organizações (JABER; GLOCK, 2013).

Cada modelo de CA é constituído por funções matemáticas diversas, o que possibilita a aplicação e descrição do processo de aprendizagem em diversos setores (ANZANELLO; FOGLIATTO, 2007). Os principais modelos de curvas de aprendizado são classificados em: (*i*) potenciais, (*ii*) exponenciais e (*iii*) hiperbólicos, sendo detalhados na sequência.

3.2.1.1. Modelos Potenciais

Os modelos potenciais foram os primeiros a serem criados. A Wright (1936) deve-se o desenvolvimento do primeiro modelo potencial e a inclusão das curvas de aprendizado na Engenharia. O autor desenvolveu o primeiro modelo potencial ao observar redução de 20% dos custos na montagem de aviões a cada duplicação da quantidade produzida, chegando à expressão conhecida como a "regra dos 80%". O modelo de Wright é dado pela equação (3.1).

$$Y = Y_1 x^b \tag{3.1}$$

Onde Y representa o tempo (custo) necessário para a execução da repetição x, e Y_1 é o tempo (custo) para a produção da primeira unidade. O declive da curva de aprendizado (velocidade de aprendizado) é representado pelo parâmetro b; seu valores variam entre -1 e 0, sendo que quanto mais próximo de -1, maior é o aprendizado (ANZANELLO; FOGLIATTO, 2007).

O modelo de Wright tem sido amplamente aplicado em cenários práticos devido a sua simplicidade matemática e capacidade de ajustar-se a dados empíricos (ANZANELLO; FOGLIATTO, 2011b). Tal modelo foi utilizado por Jaber e Khan (2010b) com o propósito de analisar uma cadeia de suprimentos em que a produção foi submetida a preceitos de melhoria contínua, mostrando-se benéfico para toda a cadeia ao considerar os efeitos da aprendizagem no processo de produção. Zorgios, Vlismas e Venieris (2009) utilizaram este modelo de CA e verificaram que provêm da variabilidade no capital humano as variações na taxa de aprendizado, e que os custos envolvidos em um ambiente produtivo podem ser racionalizados e monitorados por meio de análise de CA. Outras aplicações do modelo de Wright podem ser encontradas em Jaber e Glock (2013) e Yeh e Rubin (2012).

Transformações no modelo de Wright surgiram com o intuito de adaptar o modelo a situações específicas e, em seguida, foram reconhecidos como modelos específicos (ANZANELLO; FOGLIATTO, 2011b). O modelo de Stanford-B, na equação (3.2), é um desses modelos.

$$Y = Y_1 (x + B)^b \tag{3.2}$$

O modelo de Stanford-B acrescenta um parâmetro ao modelo de Wright, na equação (3.1). O Parâmetro B representa unidades equivalentes a experiência prévia do trabalhador. Os demais parâmetros mantêm as mesmas definições do modelo de Wright (YEH; RUBIN, 2012).

Outra transformação da equação (3.1) surgiu para incorporar a participação de maquinários no processo de aprendizado (NEMBHARD; UZUMERI, 2000a). A equação (3.3) representa o modelo de DeJong. O parâmetro M da equação (3.3) é o fator de incompressibilidade, variando entre 0 e 1. Este parâmetro representa a proporção o tempo total de operações constituído por procedimentos automatizados. Os demais parâmetros apresentam as mesmas definições da equação (3.1). $M = 1$ indica que o processo é inteiramente automatizado por maquinário; em não havendo maquinário no processo, a expressão se reduz ao modelo da equação (3.1) (ANZANELLO; FOGLIATTO, 2011b).

$$Y = Y_1[M + (1 - M)\, x^b] \tag{3.3}$$

Conhecido como Curva S, o modelo representado pela equação (3.4) contempla operações onde tanto a experiência prévia do trabalhador e interferência do maquinário estão presentes. Este modelo foi desenvolvido através da junção das equações (3.2) e (3.3) (ANZANELLO; FOGLIATTO, 2011b).

$$Y = Y_1[M + (1 - M)\, (x + B)^b] \tag{3.4}$$

Por fim, o modelo de Plateau contorna uma dificuldade apresentada pelo modelo potencial de Wright, onde o tempo de execução de uma tarefa tende a zero quando há um grande número de repetições. Mantendo os demais parâmetros, o modelo de Plateau acrescenta uma constante aditiva C, que representa o desempenho do trabalhador ao atingir o estado estacionário (ANZANELLO; FOGLIATTO, 2011b). A equação (3.5) representa o modelo de Plateau.

$$Y = C + Y_1 x^b \tag{3.5}$$

3.2.1.2. Modelos Exponenciais

Os modelos exponenciais foram criados para melhorar as previsões de produtividade de processos caracterizados por longos períodos de produção, permitindo extrair mais informações a respeito do aprendizado individual (ANZANELLO; FOGLIATTO, 2011b; NEMBHARD; UZUMERI, 2000a). Os primeiros estudos de modelo de curvas de aprendizado exponenciais foram introduzidos por Knecht (1974), o qual buscava aprimorar a modelagem de processos com elevado número de repetições. Assim, foi proposta a utilização combinada de funções exponenciais e potenciais. O modelo de Knecht é representado pela equação (3.6).

$$Y = Y_1 x^b e^{cx} \tag{3.6}$$

Onde c é uma segunda constante, e os demais parâmetros apresentam as mesmas definições das equações anteriores.

Por sua vez, o modelo exponencial de três parâmetros [equação (3.7)] foi apresentado por Mazur e Hastie (1978). Nesta equação, Y representa o desempenho do trabalhador na execução da tarefa, expresso em unidades produzidas após x unidades de tempo de operação na tarefa, k' representa o máximo de desempenho a ser atingido após a integral aquisição de conhecimento, p' designa a experiência prévia na realização da tarefa e r representa a taxa de aprendizado do trabalhador.

$$Y = k'(1 - e^{-(x+p')/r}) \tag{3.7}$$

Estudos realizados pelos autores do modelo revelam que o mesmo é adequado para representar situações em que os trabalhadores têm experiência prévia na tarefa realizada. No entanto, o modelo está mal adaptado a situações em que os trabalhadores são submetidos a tarefas complexas e que exigem aquisição de novos conhecimentos (ANZANELLO; FOGLIATTO, 2011b).

O modelo exponencial de tempo constante, representado pela equação (3.8), foi desenvolvido por Towill (1990). Segundo o autor, este é o modelo mais apropriado para modelagem de processos em que a coleta dos dados de desempenho se dá após um breve período de adaptação do trabalhador à tarefa.

$$Y = Y_c + Y_f\,(1 - e^{-t/r}) \tag{3.8}$$

Onde Y_c representa o desempenho inicial do trabalhador (unidades/tempo) e Y_f o desempenho máximo do trabalhador após atingir o estado estacionário de aprendizagem. No modelo, a variável t representa o tempo acumulado de operação, mesmo significado de x nos modelos anteriormente apresentados (ANZANELLO; FOGLIATTO, 2007). O modelo exponencial de tempo constante foi utilizado por Dardan, Busch e Sward (2006) para avaliar o investimento em tecnologia levando em consideração os impactos de aprendizagem.

3.2.1.3. Modelos Hiperbólicos

Mazur e Hastie (1978) introduziram os modelos de curvas de aprendizado hiperbólicos. Os autores propuseram uma curva de aprendizado baseada na razão entre o número de unidades consideradas conformes e o número total de unidades produzidas (ANZANELLO; FOGLIATTO, 2007). O modelo hiperbólico de 2 parâmetros é representado pela equação (3.9).

$$Y = k'\left(\frac{x}{x+r}\right) \tag{3.9}$$

A equação (3.9) tem Y como número de unidades produzidas em um intervalo de tempo de operação x, o parâmetro que indica o aprendizado é representado por r, e k' corresponde ao máximo desempenho a ser atingido.

Posteriormente, Mazur e Hastie (1978) introduziram um parâmetro p' na equação (3.9), representando a experiência prévia no trabalhador na execução da tarefa em questão. Da adição desde parâmetro resultou a equação (3.10), que representa o modelo hiperbólico de 3 parâmetros, o qual tem sido aplicado em diversos setores industriais. Wong, On Cheung e Hardcastle (2007) o aplicaram em um projeto de construção civil para prever o desempenho de empreiteiros, enquanto Guimarães, Anzanello e Renner (2012) utilizaram o modelo em uma indústria de calçados para projetar a rotação dos trabalhadores em uma linha de montagem, resultando em redução significativa de acidentes e absenteísmo.

$$Y = k'\left(\frac{x+p'}{x+r+p'}\right) \tag{3.10}$$

3.2.2 Regressões PLS e LASSO

A regressão PLS (*Partial Least Square* ou Quadrados Parciais Mínimos), proposta originalmente por Wold (1982), é uma técnica de análise multivariada que relaciona variáveis variáveis independentes (V) e dependentes (Q). Considere uma matriz V com N observações para cada uma das variáveis independentes J, e uma matriz Q consistindo de N observações para cada uma das variáveis dependentes H (neste estudo $H = 1$). A observação independente i é representada pelo vetor v_o $(v_{i1}, v_{i2}, \dots, v_{iJ})$, enquanto a observação dependente i é denotada por q_o $(q_{i1}, q_{i2}, \dots, q_{iH})$, para $i = 1, \dots, N$ (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012).

A regressão PLS constrói A combinações lineares indepentendes $l_{ia} = w_{1a}v_{i1} + w_{2a}v_{i2} + \cdots + w_{Ja}v_{iJ} = w'_a v_i$, das variáveis independentes, onde $A \leq J$. O vetor $w_{a=}(w_{1a}, w_{2a}, \dots, w_{Ja})'$ representa os pesos, e o elemento w_{ja} é o peso da variável independentes j no componente a. Da mesma forma, os componentes são construidos para as variáveis dependentes em Z; $u_{ia=} z_{1a}q_{i1} + z_{2a}q_{i2} + z_{Ha}q_{iH} = z'_a q_i$ onde $z_a = (z_{1a}, z_{2a}, \dots, z_{Ha})'$ são os pesos das variáveis dependentes (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

Os vetores de peso w_a e z_a, são selecionados de modo a maximizar a covariância entre os componentes PLS l_a e u_a. Além disso, os pesos são definidos de modo a produzir componentes ortogonais (os compondentes l_a são independentes entre si, e os componentes u_a são independentes entre si), de acordo com Wold, Sjöström e Eriksson (2001) e Xu e Albin (2002). De modo geral o número de componentes, A, retidos é pequeno comparado ao número de variáveis originais, que podem muito numerosas, e pode ser definido pelo método de validação cruzada em Hoskuldsson (1988).

O algoritmo Nonlinear Iterative Partial Least Squares (NIPALS) pode ser utilizado para calcular os parâmetros da regressão PLS; ver Goutis (1997), Geladi; Kowalski (1986) e Abdi (2003). Os principais parâmetros resultantes do algoritmo NIPALS são pesos e cargas, os quais podem ser manipulados para gerar a regressão. Os coeficientes da regressão são representado por d_{hj}, e dependem dos pesos dos componentes previamente estimados, w_a e z_a, conforme a equação (3.11). Esta regressão foi utilizada por Anzanello, Albin e Chaovalitwongse (2009)

com o propósito de criar um indice de importancia de variaveis baseando-se nos parâmetros da regressão PLS, ordenando as variáveis de processo de acordo com sua relevância para caracterização das variáveis de produto. A regressão PLS pode ser realizada utilizandi o *toolbox* PLS em pacotes estatísticos como Matlab® e R®.

$$d_{hj} = \sum_{a=1}^{A} z_{ha} w_{ja}^{*} \quad h = 1, \dots, H \text{ e j= } 1, \dots, J \tag{3.11}$$

A regressão LASSO (*Least Absolute Shrinkage and Selection Operator*), proposta originalmente por Tibshirani (1996), é uma regressão de mínimos quadrados que realiza tanto a seleção de variáveis quanto a regularização através de um fator de retenção. A regularização ajuda a resolver problemas de ajuste excessivo, onde o modelo tem um bom desempenho no treinamento, mas apresenta desempenho ruim na validação. Para isso é adicionado o termo de penalidade a função objetivo da regressão.

Desta forma, a LASSO é capaz de aumentar a precisão e capacidade de interpretação dos métodos clássicos de regressão. Esta regressão utiliza um parâmetro de ajuste, λ (lambda), o qual pode ser determinado por validação cruzada e penaliza coeficientes que tornam-se grandes. O resultado é um modelo que zera estimativas de alguns coeficientes, cujas variáveis independentes associadas são entendidas como irrelevantes. Entende-se, de tal forma, que a regressão LASSO promova a seleção de variáveis (MELKUMOVA; SHATSKIKH, 2017). A equação (3.12) representa o vetor de $\hat{\beta}$, entendido como vetor dos coeficientes da regressão LASSO.

$$\hat{\beta} = \sum_{i=1}^{N} (q_i - \sum_j v_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{f} |\beta_j| \tag{3.12}$$

3.2.3 Ferramentas de classificação

A ferramentas k-vizinhos mais próximos (KNN - k-Nearest Neighbor) é um método de classificação supervisionado, proposto por Fix e Hodges (1951). Esta técnica destaca-se por sua simplicidade, sendo necessário definir apenas um parâmetro k (número de vizinhos a serem analisados), que pode ser definido por validação cruzada (BARBON et al., 2016). O algoritmo considera cada observação como um ponto em seu espaço amostral n-dimensional. A distância entre uma nova observação e suas vizinhas é medida, e à nova observação é atribuída a classe

predominante entre suas observações vizinhas mais próximas. Tipicamente a distância Euclidiana é utilizada para medir a distância entre as observações.

Naïve Bayes (NB), segunda ferramenta de classificação utilizada, é um classificador probabilístico supervisionado. O NB é considerado uma ferramenta de alto desempenho e de fácil aplicação, sendo amplamente utilizada em problemas de classificações e aprendizado de máquina (FARID et al., 2014). Baseado no Teorema de *Bayes*, este classificador desconsidera qualquer correlação entre as variáveis, calculando a probabilidade de uma amostra pertencer a cada uma das possíveis classes, sendo a observação classificada na classe de maior probabilidade.

A terceira ferramenta de classificação utilizada foi a Máquina de Suporte Vetorial (SVM - Support Vector Machine), um algoritmo de aprendizagem de máquina supervisionado introduzido por Cortes e Vapnik (1995). Em um espaço amostral n-dimensional onde estão dispostas as observações, a ferramenta SVM constrói um hiperplano de modo a separar as observações em duas classes. Esse hiperplano objetiva maximizar a distância entre as duas observações mais próximas, cada uma pertencente a uma classe distinta. Mais detalhes podem ser vistos em Cristianini e Shawe-Taylor (2000).

3.3 Método

O método proposto visa identificar um subconjunto reduzido de parâmetros oriundos da modelagem por Curvas de Aprendizagem com vistas à inserção de trabalhadores em grupos pré-existentes (linhas ou células de produção) de acordo com semelhanças entre seus perfis de aprendizado. A estrutura sugerida apoia-se em quatro passos operacionais: (i) selecionar trabalhadores e coletar dados de desempenho; (ii) modelar os dados de desempenho utilizando diferentes modelos de CAs e separar os parâmetros oriundos da modelagem em porções de treino e teste; (iii) aplicar as regressões e gerar o Índice de Importância de Parâmetros (IIP) com base nos parâmetros oriundos das regressões; (iv) iterativamente classificar os trabalhadores e remover os parâmetros menos relevantes. Estes passos são agora detalhados.

O primeiro passo do método proposto visa identificar os trabalhadores que terão seus perfis de aprendizagem monitorado. O desempenho do trabalhador pode ser medido em número

de unidades produzidas em determinado período de tempo, ou em termos de tempo requerido por repetição. Deseja-se que o trabalhador esteja familiarizado com as operações que serão analisadas. As coletas de dados de desempenhos devem ser realizadas até que alterações significativas em seu desempenho não sejam mais observadas.

No segundo passo, os dados de desempenho coletados dos trabalhadores são modelados utilizando as 10 CA apresentadas na seção 3.2, obtendo os perfis de aprendizagem dos trabalhadores. Esses dados são organizados em uma tabela que traz 300 linhas (referentes ao número de trabalhadores monitorados); 29 colunas (associadas ao número de parâmetros gerados pelos 10 modelos de CAs testados) de variáveis independentes; e uma coluna (associando o trabalhador à estação de trabalho, sendo 0 para trabalhadores da primeira estação, e 1 para trabalhadores da segunda estação) como variável dependente. Na sequência os dados são normalizados e divididos em duas porções: treinamento (T_r), contendo n_{Tr} observações, e teste (T_s), contendo n_{Ts} observações, de modo que $n_{Tr} + n_{Ts} = n$.

As regressões PLS e LASSO são então aplicadas aos dados da porção de treino, no terceiro passo. Para geração do IIP utiliza-se os coeficientes (d_p) da regressão PLS e o parâmetro $\hat{\beta}_p$ da regressão LASSO, juntamente com os coeficientes de determinação (R^2) das regressões. O Índice de Importância de Parâmetro é obtido através da equação (3.13). A utilização do IIP proposto para seleção de parâmetros relevantes é justificada i) pela capacidade em identificar os parâmetros detentores da maior variância em si; ii) por selecionar os parâmetros mais aptos a explicarem a variância das variáveis dependentes v; e iii) pelo mecanismo de penalização da regressão LASSO aos coeficientes com alto grau de correlação entre si, o qual naturalmente elimina parâmetros menos relevantes (TIBSHIRANI, 1996).

$$IIP_p = \left|d_p\right| * R^{2\,(PLS)} + \left|\hat{\beta}_p\right| * R^{2\,(Lasso)} \quad p = 1, \dots, P \text{ parâmetros} \tag{3.13}$$

O quarto passo consiste na eliminação iterativa de parâmetros com o objetivo de definir os parâmetros mais relevantes para a alocação dos trabalhadores em grupos pré-existentes (linhas ou células). Uma classificação inicial utilizando P parâmetros é realizada e a acurácia do processo é computada. Neste estudo a acurácia é definida como a razão entre o número de classificações corretas e o número total de casos classificados. Quando a classe prevista pela

ferramenta utilizada para classificação é idêntica à classe anteriormente estabelecida ao trabalhador, esta classificação é dita correta.

Na sequência, o parâmetro com menor IIP é eliminado das porções de treino e teste, e uma nova classificação utilizando os $P - 1$ parâmetros restantes é realizada; a acurácia da classificação é computada. O procedimento iterativo de remoção de parâmetros e classificação é executado até que haja apenas um parâmetro remanescente. Para fins de comparação do desempenho de classificação, o procedimento iterativo acima descrito é repetido utilizando os 3 classificadores testados (KNN, NB e SVM).

Um gráfico associando o número de parâmetros removidos e a acurácia para cada técnica de classificação é gerado para monitorar o processo de eliminação dos parâmetros. O ponto de máxima acurácia indica o subconjunto de parâmetros recomendados para classificação de trabalhadores em grupos homogêneos. No caso de distintos subconjuntos de parâmetros conduzirem a picos idênticos de acurácia, opta-se pelo subconjunto com menor número de parâmetros (um menor número de parâmetros é desejado).

3.4 Estudo de caso

O método proposto foi aplicado em uma indústria calçadista no sul do Brasil. Indústrias deste segmento são fortemente impactadas pela tendência de customização em massa ao invés de produções tradicionais, fazendo com que tamanhos de lotes diminuam à medida que a variedade de modelos aumenta (ANZANELLO; FOGLIATTO; SANTOS, 2014).

Dados de desempenho (tempo de processamento por repetição) de 300 trabalhadores foram modelados utilizando as 10 equações de CAs apresentadas na seção 3.2.1. Um total de 29 parâmetros foram gerados ajustando os dados de desempenho de cada trabalhador aos modelos de CA; parte dos parâmetros estimados estão ilustrados no Apêndice. As observações do banco de dados (parâmetros de CA dos trabalhadores) foram rotuladas em duas categorias, de acordo com a estação de trabalho em que o trabalhador operava (0 para trabalhadores pertencentes à primeira estação, e 1 para trabalhadores da segunda estação). Tal divisão já se encontrava consolidada no momento da geração do modelo de classificação, tendo sido indicada por especialistas da empresa analisada.

O conjunto de parâmetros foi normalizado e dividido nas porções de treino (75%) e teste (25%) através do algoritmo de Kennard-Stone (KS); mais detalhes sobre o algoritmo podem ser encontrados em Kennard e Stone (1969). Na sequência, as regressões PLS e LASSO foram aplicadas à porção de treino. Os parâmetros gerados pelas regressões foram manipulados matematicamente de acordo com a equação (3.13) para obtenção dos IIP_p apresentados na Tabela 3.1. Valores mais altos denotam parâmetros com maior capacidade de alocação de trabalhadores a uma das 2 classes especificadas anteriormente.

Tabela 3.1 - Índice de Importância de Parâmetros (IIP_p)

Modelo CA	Parâmetro	IIP_p	Modelo CA	Parâmetro	IIP_p
Knecht	$\boldsymbol{Y_1}$	**0,3826**	Hiperb. de 3 parâmetros	p'	0,0613
Stanford-b	$\boldsymbol{B}$	**0,3617**	Knecht	c'	0,0601
Curva S	$\boldsymbol{Y_1}$	**0,2433**	Hiperb. de 2 parâmetros	k'	0,0521
Wright	b	0,1611	Tempo Constante	Y_c	0,0418
Tempo Constante	Y_f	0,1549	Tempo Constante	t	0,0360
Hiperb. de 2 parâmetros	r	0,1321	Hiperb. de 3 parâmetros	r	0,0334
DeJong	M	0,1190	Exp. de 3 parâmetros	p'	0,0297
Stanford-b	b	0,1036	DeJong	b	0,0245
Stanford-b	Y_1	0,1019	Curva S	b	0,0232
Plateau	Y_1	0,1001	Hiperb. de 3 parâmetros	k'	0,0084
DeJong	Y_1	0,0994	Exp. de 3 parâmetros	k'	0,0062
Curva S	B	0,0946	Knecht	B	0,0062
Curva S	M	0,0946	Exp. de 3 parâmetros	r	0,0041
Wright	Y_1	0,0906	Plateau	C	0,0010
Plateau	b	0,0704			

O procedimento iterativo descrito no quarto passo foi então realizado. As Figuras 3.1 a 3.3 trazem a variação do perfil de acurácia com a remoção dos parâmetros de acordo com a ordem definida pelo IIP para as três ferramentas de classificação testadas.

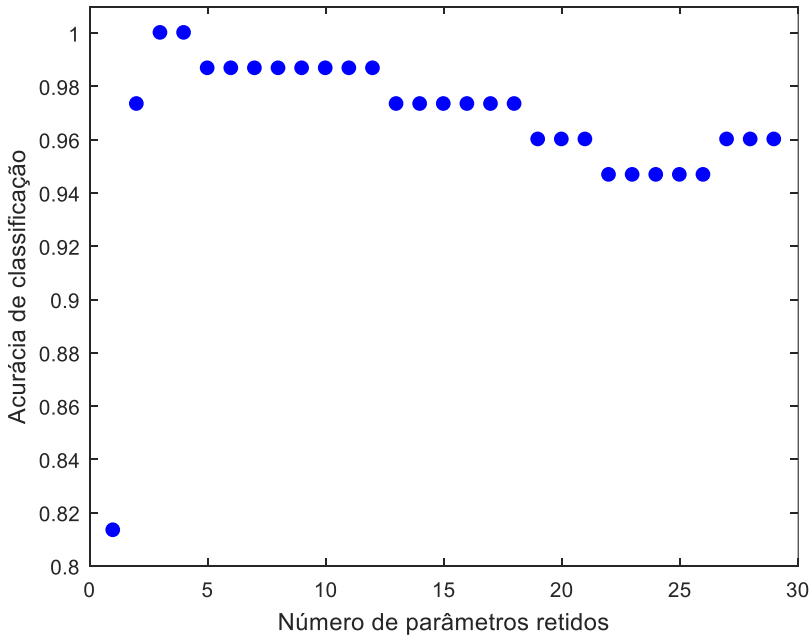

Figura 3.1 - KNN - remoção de parâmetros de acordo com IIP

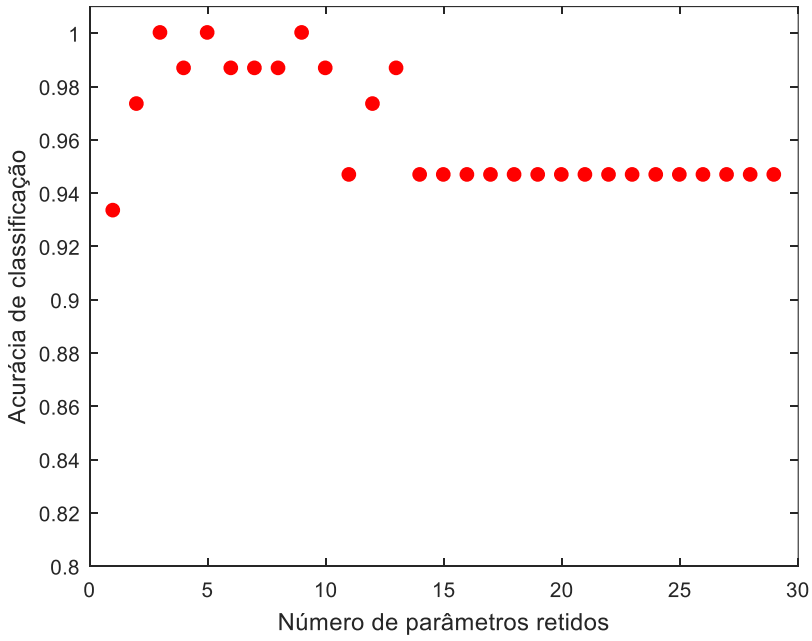

Figura 3.2 - NB - remoção de parâmetros de acordo com IIP

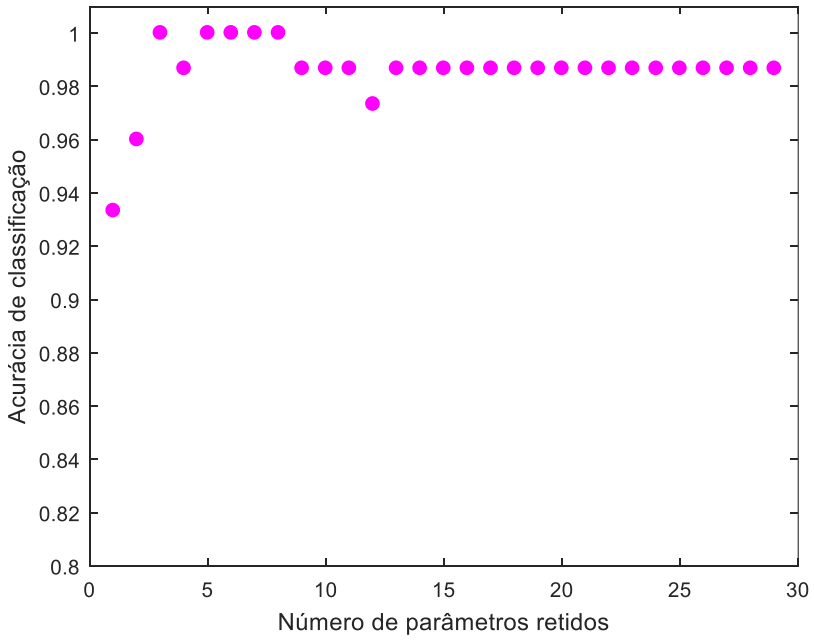

Figura 3.3 - SVM - remoção de parâmetros de acordo com IIP

De acordo com as três ferramentas de classificação utilizadas (KNN, NB e SVM), as melhores acurácias (100%) são obtidas ao reter-se apenas 3 parâmetros (assinalados em negrito na Tabela 3.1). Os parâmetros retidos estão associados ao desempenho dos trabalhadores na primeira repetição (Y_1) e à experiência prévia dos trabalhadores na tarefa desempenhada (B). A remoção adicional de parâmetros levou a uma redução da capacidade de predição, sugerindo que estes três parâmetros são os mais relevantes para alocar trabalhadores a grupos pré-existentes que podem designar estações de trabalho, células de manufatura ou linhas de produção.

A Tabela 3.2 apresenta a média para cada um dos parâmetros retidos para cada um dos grupos pré-existentes. Ao assumir valores mais altos na primeira estação de trabalho, os parâmetros Y_1, das CAs Knecht e Curva S, indicam que esta estação é formada principalmente por trabalhadores que tendem a exigir mais tempo para concluir o primeiro ciclo de produção. Ainda, observa-se o fato dos trabalhadores da segunda estação apresentarem maior experiência prévia nas tarefas desempenhadas, o que é indicado pelo parâmetro B. Destaca-se os perfis de aprendizagem distintos dos trabalhadores de cada estação, sugerindo que a segunda estação foi formada tendo como base trabalhadores que apresentam um aprendizado mais rápido.

Tabela 3.2 - Média dos parâmetros retidos para cada estação pré-existente

CA	Parâmetro	Estação 1	Estação 2
Knecht	Y_1	176,9	89,8
Stanford-b	B	69,2	157,1
Curva S	Y_1	166,7	122,8

Com base nos resultados do estudo, entende-se que é possível manter a homogeneidade de grupos pré-existentes, de modo a preservar suas características. O método proposto apresenta vantagens com vistas a diminuir os possíveis impactos negativos que o desbalanceamento dos perfis dos trabalhadores poderia causar ao ritmo de produção. Assim, mantendo o mesmo perfil de aprendizagem quando da alocação de um novo trabalhador à determinada estação ou substituição de um de seus membros, evitando a formação de gargalos e mantem o ritmo da produção estável.

3.5 Conclusão

Este artigo apresentou um novo método de seleção de variáveis, o qual foi aplicado na classificação de trabalhadores de acordo com seu perfil de aprendizagem. O método integra modelagem de curva de aprendizado e técnicas multivariadas para selecionar as melhores variáveis e posteriormente classificar os trabalhadores em suas etapas operacionais. Parâmetros que descrevem diferentes aspectos do perfil de aprendizado de cada trabalhador foram obtidos através de modelos de CAs; os parâmetros mais relevantes para o propósito de classificação foram selecionados guiando-se por um índice de importância de parâmetros proposto, o qual é gerado através da integração dos coeficientes das regressões PLS e LASSO. Três ferramentas de classificação foram testadas no método proposto: k-vizinhos mais próximos (KNN), *Naïve Bayes* (NB) e Máquina de Vetor de Suporte (SVM).

Ao ser aplicado a dados coletados no setor de costura de uma indústria calçadista, o método satisfatoriamente alocou os trabalhadores em duas estações de trabalho pré-existentes, apresentando acurácia de 100% com qualquer dos três classificadores ao reter apenas 3 parâmetros dos 29 parâmetros originais. Os parâmetros retidos estão relacionados com o desempenho dos trabalhadores na primeira unidade de um novo ciclo e sua experiência prévia.

Do ponto de vista prático, a aplicação do método proposto visa garantir a preservação homogeneidade dos perfis de aprendizagem dos trabalhadores alocados à cada estação de trabalho. Ao modo que a substituição ou inserção de novos trabalhadores em tais estações não inicie cenários produtivos de gargalos, gerando ociosidade de trabalhadores, ou altere de modo negativo os ritmos de produção.

Trabalhos futuros visam voltar ferramentas multivariadas para o meio acadêmico ao criar uma sistemática que identifique as variáveis mais informativas na predição do destino de alunos de graduação (diplomação ou abandono de curso) após os primeiros semestres letivos.

3.6 Referências

ABDI, H. Partial Least Squares (PLS) Regression. **Encyclopedia for research methods for the social sciences**, p. 792–795, 2003.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratory Systems**, v. 97, n. 2, p. 111–117, jul. 2009.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, n. 1, p. 97–105, 1 abr. 2012.

ANZANELLO, M. J.; FOGLIATTO, F. S. Learning curve modelling of work assignment in mass customized assembly lines. **International Journal of Production Research**, v. 45, n. 13, p. 2919–2938, jul. 2007.

ANZANELLO, M. J.; FOGLIATTO, F. S. Selecting the best clustering variables for grouping mass-customized products involving workers learning. **International Journal of Production Economics**, v. 130, n. 2, p. 268–276, 1 abr. 2011a.

ANZANELLO, M. J.; FOGLIATTO, F. S. Learning curve models and applications: Literature review and research directions. **International Journal of Industrial Ergonomics**, v. 41, n. 5, p. 573–583, 1 set. 2011b.

ANZANELLO, M. J.; FOGLIATTO, F. S.; SANTOS, L. Learning dependent job scheduling in mass customized scenarios considering ergonomic factors. **International Journal of Production Economics**, v. 154, p. 136–145, ago. 2014.

AZEVEDO, B. B.; ANZANELLO, M. J. Agrupamento de trabalhadores com perfis semelhantes de aprendizado apoiado em Análise de Componentes Principais. **Gestão & Produção**, v. 22, n. 1, p. 35–52, mar. 2015.

BARBON, A. P. A. C.; BARBON, S.; MANTOVANI, R. G.; FUZYI, E. M.; PERES, L. M.; BRIDI, A. M. Storage time prediction of pork by Computational Intelligence. **Computers and Electronics in Agriculture**, v. 127, p. 368–375, set. 2016.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines : and other kernel-based learning methods**. [s.l.] Cambridge University Press, 2000.

DA SILVEIRA, G.; BORENSTEIN, D.; FOGLIATTO, F. S. Mass customization: Literature review and research directions. **International Journal of Production Economics**, v. 72, n. 1, p. 1–13, 30 jun. 2001.

DARDAN, S.; BUSCH, D.; SWARD, D. An application of the learning curve and the nonconstant-growth dividend model: IT investment valuations at Intel® Corporation. **Decision Support Systems**, v. 41, n. 4, p. 688–697, 1 maio 2006.

FARID, D. M.; ZHANG, L.; RAHMAN, C. M.; HOSSAIN, M.A.; STRACHAN, R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. **Expert Systems with Applications**, v. 41, n. 4, p. 1937–1946, mar. 2014.

FIORETTI, G. The organizational learning curve. **European Journal of Operational Research**, v. 177, n. 3, p. 1375–1384, 16 mar. 2007.

FIX, EVELYN; HODGES JR, J. L. **Discriminatory analysis-nonparametric discrimination: consistency properties**. 1951.

FOGLIATTO, F. S.; DA SILVEIRA, G. J. C.; BORENSTEIN, D. The mass customization decade: An updated review of the literature. **International Journal of Production Economics**, v. 138, n. 1, p. 14–25, 1 jul. 2012.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, p. 1–17, 1 jan. 1986.

GOUTIS, C. A fast method to compute orthogonal loadings partial least squares. **Journal of Chemometrics**, v. 11, n. 1, p. 33–38, 1 jan. 1997.

GROSSE, E. H.; GLOCK, C. H.; MÜLLER, S. Production economics and the learning curve: A meta-analysis. **International Journal of Production Economics**, v. 170, p. 401–412, 1 dez. 2015.

GUIMARÃES, L. B. D. M.; ANZANELLO, M. J.; RENNER, J. S. A learning curve-based method to implement multifunctional work teams in the Brazilian footwear sector. **Applied Ergonomics**, v. 43, n. 3, p. 541–547, 1 maio 2012.

HÖSKULDSSON, A. PLS regression methods. **Journal of Chemometrics**, v. 2, n. 3, p. 211–228, 1 jun. 1988.

JABER, M. Y.; EL SAADANY, A. M. A. An economic production and remanufacturing model with learning effects. **International Journal of Production Economics**, v. 131, n. 1, p. 115–127, 1 maio 2011.

JABER, M. Y.; GLOCK, C. H. A learning curve for tasks with cognitive and motor elements. **Computers & Industrial Engineering**, v. 64, n. 3, p. 866–871, 1 mar. 2013.

JABER, M. Y.; KHAN, M. Managing yield by lot splitting in a serial production line with learning, rework and scrap. **International Journal of Production Economics**, v. 124, n. 1, p. 32–39, 1 mar. 2010a.

JABER, M. Y.; KHAN, M. Managing yield by lot splitting in a serial production line with learning, rework and scrap. **International Journal of Production Economics**, v. 124, n. 1, p. 32–39, 1 mar. 2010b.

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, v. 11, n. 1, p. 137–148, fev. 1969.

KNECHT, G. R. Costing, Technological Growth and Generalized Learning Curves. **Journal of the Operational Research Societ**y, v. 25, n. 3, p. 487–491, 19 set. 1974.

LOHMANN, M.; ANZANELLO, M. J.; FOGLIATTO, F. S.; SILVEIRA, G. C. Grouping workers with similar learning profiles in mass customization production lines. **Computers & Industrial Engineering**, v. 131, p. 542–551, 1 maio 2019.

MACCARTHY, B.; BRABAZON, P. G.; BRAMHAM, J. Fundamental modes of operation for mass customization. **International Journal of Production Economics**, v. 85, n. 3, p. 289–304, 11 set. 2003.

MAZUR, J. E.; HASTIE, R. Learning as accumulation: A reexamination of the learning curve. **Psychological Bulletin**, v. 85, n. 6, p. 1256–1274, 1978.

MELKUMOVA, L. E.; SHATSKIKH, S. Y. Comparing Ridge and LASSO estimators for data analysis. **Procedia Engineering**, v. 201, p. 746–755, 1 jan. 2017.

NEMBHARD, D. A.; UZUMERI, M. V. An individual-based description of learning within an organization. **IEEE Transactions on Engineering Management**, v. 47, n. 3, p. 370–378, 2000a.

NEMBHARD, D. A.; UZUMERI, M. V. Experiential learning and forgetting for manual and cognitive tasks. **International Journal of Industrial Ergonomics**, v. 25, n. 4, p. 315–326, 1 maio 2000b.

PLAZA, M.; ROHLF, K. Learning and performance in ERP implementation projects: A learning-curve model for analyzing and managing consulting costs. **International Journal of Production Economics**, v. 115, n. 1, p. 72–85, 1 set. 2008.

REID, S. A.; MIRKA, G. A. Learning curve analysis of a patient lift-assist device. **Applied Ergonomics**, v. 38, n. 6, p. 765–771, nov. 2007.

STROIEKE, R. E.; FOGLIATTO, F. S.; ANZANELLO, M. J. Formação de agrupamentos homogêneos de trabalhadores através de Curvas de Aprendizado. **Pré-anais XLIII Simpósio Brasileiro de Pesquisa Operacional**, 2011.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, p. 267–288, 1996.

TILINDIS, J.; KLEIZA, V. Learning curve parameter estimation beyond traditional statistics. **Applied Mathematical Modelling**, v. 45, p. 768–783, 1 maio 2017.

TOWILL, D. R. Forecasting learning curves. **International Journal of Forecasting**, v. 6, n. 1, p. 25–38, 1 jan. 1990.

UZUMERI, M.; NEMBHARD, D. A population of learners: A new way to measure organizational learning. **Journal of Operations Management**, v. 16, n. 5, p. 515–528, 1 out. 1998.

WOLD, H. Soft modeling : the basic design and some extensions**. Systems under indirect observation : causality, structure, prediction**, v. 2, 1982.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109–130, 28 out. 2001.

WONG, P. S. P.; CHEUNG, S. O.; WU, R. T. H. Learning from project monitoring feedback: A case of optimizing behavior of contractors. **International Journal of Project Management**, v. 28, n. 5, p. 469–481, jul. 2010.

WONG, P. S. P.; ON CHEUNG, S.; HARDCASTLE, C. Embodying Learning Effect in Performance Prediction. **Journal of Construction Engineering and Management**, v. 133, n. 6, p. 474–482, jun. 2007.

WRIGHT, T. P. Factors Affecting the Cost of Airplanes. **Journal of the Aeronautical** Sciences, v. 3, n. 4, p. 122–128, fev. 1936.

XU, D.; ALBIN, S. L. Manufacturing start-up problem solved by mixed-integer quadratic programming and multivariate statistical modelling. **International Journal of Production Research**, v. 40, n. 3, p. 625–640, jan. 2002.

YEH, S.; RUBIN, E. S. A review of uncertainties in technology experience curves. **Energy Economics**, v. 34, n. 3, p. 762–771, 1 maio 2012.

ZORGIOS, Y.; VLISMAS, O.; VENIERIS, G. A learning curve explanatory theory for team learning valuation. **VINE**, v. 39, n. 1, p. 20–39, 10 abr. 2009.

Apêndice – Parâmetros de CAs (Apresentação parcial)

Trabalhador	Wright	Stanford-b	DeJong	Curva S	Plateau	Knecht	Exponencial de 3 parâmetros	Tempo Constante	Hiperbólico de 2 parâmetros	Hiperbólico de 3 parâmetros
1	Y_1= 90,49 b = -0,15	Y_1= 97,9 B = 1,2 b = -0,17	Y_1= 90,49 M = 0 b = -0,15	Y_1= 97,9 M = 0 B = 1,2 b = -0,17	C = 0 Y_1= 90,39 b = -0,15	Y_1= 90,49 b = -0,15 c'= 0	k = 14,78 p = 36,1 r = 35,09	y_c= 7,11 y_f= 7,68 t = 55,09	k = 13,96 r = 8,4	k = 17,6 p = 22,86 r = 36,76
2	Y_1 = 246,67 b = -0,27	Y_1= 250,2 B = 0,08 b = -0,28	Y_1= 257,57 M = 0,12 b = -0,36	Y_1= 250,19 M = 0 B = 0,08 b = -0,28	C = 30,45 Y_1= 227,11 b = 0,36	Y_1= 221 b = -0,24 c'= 0	k = 8,46 p = 17,23 r = 61,18	y_c= 2,08 y_f= 6,38 t = 61,18	k = 9,28 r = 28,98	k = 10,19 p = 4,5 r = 45,09
3	Y_1 = 60,7 b = -0,09	Y_1= 70,78 B = 5,58 b = -0,12	Y_1= 60,7 M = 0 b = -0,09	Y_1= 70,78 M = 0 B = 5,58 b = -0,12	C = 0 Y_1= 60,67 b = -0,09	Y_1= 56 b = -0,06 c'= 0	k = 16,37 p = 49,31 r = 50,96	y_c= 10,15 y_f= 6,22 t = 50,96	k = 15,75 r = 4,97	k = 18,66 p = 29,55 r = 26,28
4	Y_1 = 334,95 b = -0,26	Y_1= 105,23 B = 11,68 b = -0,53	Y_1= 334,95 M = 0 b = -0,26	Y_1= 105,23 M = 0 B = 11,68 b = -0,53	C = 0 Y_1= 328,05 b = -0,25	Y_1= 298 b = -0,22 c'= 0	k = 7,35 p = 38,54 r = 157,89	y_c= 1,59 y_f= 5,58 t = 157,89	k = 6,99 r = 56,61	k = 11,15 p = 37,12 r = 220,62
5	Y_1 = 202,08 b = -0,12	Y_1= 324,31 B = 15,03 b = -0,22	Y_1= 202,08 M = 0 b = -0,12	Y_1= 324,28 M = 1 B = 15,03 b = -0,22	C = 0 Y_1= 201,58 b = -0,12	Y_1= 185 b = -0,09 c'= 0	k = 7,15 p = 168,55 r = 280,79	y_c= 3,23 y_f= 3,93 t = 280,79	k = 5,25 r = 12,06	k = 8,06 p = 105,2 r = 167,02
6	Y_1 = 155,83 b = -0,11	Y_1= 155,83 B = 0 b = -0,11	Y_1= 170,87 M = 0,55 b = -0,5	Y_1= 155,83 M = 0 B = 0 b = -0,11	C = 94,3 Y_1= 76,57 b = -0,5	Y_1= 183 b = -0,19 c'= 0	k = 11,43 p = 117,75 r = 281,09	y_c= 3,91 y_f= 7,52 t = 281,16	k = 6,48 r = 8,82	k = 11,38 p = 72,52 r = 146,82
7	Y_1 = 205,84 b = -0,08	Y_1= 278,82 B = 15,09 b = -0,14	Y_1= 205,84 M = 0 b = -0,08	Y_1= 279,97 M = 0 B = 8,85 b = -0,16	C = 0 Y_1= 205,75 b = -0,08	Y_1= 196 b = -0,06 c'= 0	k = 5,08 p = 115,37 r = 142,29	y_c= 2,82 y_f= 2,26 t = 142,29	k = 4,19 r = 6,31	k = 6,45 p = 94,85 r = 122,73
8	Y_1 = 101,43 b = -0,18	Y_1= 101,43 B = 0 b = -0,18	Y_1= 108,81 M = 0,27 b = -0,35	Y_1= 108,81 M = 0,27 B = 1 b = -0,35	C = 29,69 Y_1= 79,12 b = -0,35	Y_1= 109,57 b = -0,24 c'= 0	k = 12,6 p = 0 r = 12,3	y_c= -2,19 y_f= 14,84 t = 12,32	k = 14,87 r = 11,54	k = 13,63 p = -6 r = 5,1
9	Y_1 = 190,75 b = -0,25	Y_1= 314,89 B = 4,28 b = -0,38	Y_1= 190,75 M = 0 b = -0,25	Y_1= 314,89 M = 0 B = 4,28 b = -0,38	C = 0 Y_1= 187,53 b = -0,24	Y_1= 85,49 b = -0,02 c'= 0	k = 10,49 p = 14,91 r = 60,57	y_c= 2,29 y_f= 8,2 t = 60,57	k = 11,3 r = 29,57	k = 14,7 p = 12,98 r = 71,85
10	Y_1 = 184,77 b = -0,28	Y_1= 220,81 B = 1,13 b = -0,33	Y_1= 184,77 M = 0 b = -0,28	Y_1= 220,81 M = 0 B = 1,13 b = -0,33	C = 0 Y_1= 183,13 b = -0,28	Y_1= 163 b = -0,25 c'= 0	k = 15,34 p = 13,85 r = 82,91	y_c= 2,36 y_f= 12,98 t = 82,91	k = 15,03 r = 38,08	k = 19,7 p = 8,35 r = 79,91

4 Terceiro artigo: Identificação das variáveis mais relevantes para classificação de estudantes de acordo com seu desfecho acadêmico

Resumo

A evasão no ensino superior é um problema que ocasiona perdas em diversas dimensões e contextos. No Brasil, altos índices de evasão são detectados nas instituições de ensino superior (IES), justificando o desenvolvimento de métodos que sejam capazes de detectar previamente qual será o desfecho dos alunos perante seus cursos de graduação. Tal recurso permite aos gestores identificar os fatores que preponderantemente levam alunos a deixar seus cursos de graduação, auxiliando na elaboração de estratégias voltas para a redução dos índices de evasão. Este artigo propõe uma abordagem multivariada para selecionar as variáveis com maior influência sobre três possíveis desfechos de alunos de graduação: diplomação, evasão interna (troca de curso dentro da mesma IES) ou evasão externa. As variáveis analisadas compreendem dados do perfil acadêmico dos alunos no momento de ingresso na graduação e do desempenho acadêmico do primeiro ao quarto semestre. Cinco ferramentas de classificação são integradas à técnica "omita uma variável por vez" com a finalidade de identificar o subconjunto de variáveis que melhor descreve o destino dos graduandos: k-vizinhos mais próximos (KNN), Rede Neural Probabilística (PNN), Análise Discriminante Linear (LDA), Máquina de Suporte Vetorial (SVM) e *Naïve Bayes* (NB). Quando aplicada a dados de 1421 alunos de graduação de cursos de engenharia ingressantes nos anos de 2008 e 2009, a abordagem proposta obteve acurácia de 91,22% ao reter 22,22% das variáveis originais e utilizar a ferramenta de classificação NB. Informações relativas aos resultados obtidos no terceiro semestre acadêmico e informações de desempenho acadêmico mostraram-se as mais relevantes na determinação do desfecho. Alinhada com os resultados obtidos por outros estudos, a abordagem proposta identificou que dados de desempenho acadêmico, como número de créditos cursados e aprovados, são fatores determinantes para definir se um aluno permanecerá no curso até a sua diplomação ou se evadirá antes do seu término.

Palavras-Chaves: Classificação, Alunos de graduação, Omita uma variável por vez, Evasão

4.1 Introdução

A evasão no ensino superior é um problema que acarreta desperdícios sociais, acadêmicos e econômicos em nível mundial (SILVA FILHO et al., 2007). No Brasil, o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (Reuni) foi instituído em 2007, sendo a redução dos índices de evasão um de seus objetivos (BRASIL, 2007). Apesar de tais esforços, elevados índices de evasão continuam a serem detectados no ensino superior do país, justificando estudos voltados à identificação dos fatores que levam alunos a concluir suas graduações ou evadir (tanto trocar de curso como abandonar a formação universitária).

Detectar os alunos que possuem maior risco de evadir, bem como identificar variáveis confiáveis que sinalizem uma potencial evasão, são desafios para as instituições, gestores e pesquisadores. Mecanismos capazes de antecipar tais situações podem permitir aos gestores um melhor entendimento dos fatores que levam a tais desfechos, bem como levar à elaboração de ações que reduzam as evasões e retenham alunos até sua diplomação. No entanto, a construção de um modelo robusto, capaz de prever o desfecho dos alunos com base em dados do perfil acadêmico do aluno no momento de ingresso na graduação e de seu desempenho nas etapas iniciais do curso, mostra-se uma tarefa desafiadora (XING et al., 2015).

Embora tenham muita relevância, pesquisas relacionadas aos temas de evasão, retenção e diplomação no ensino superior ainda não são frequentes no Brasil, apesar de serem encontradas em números consideráveis em países da Europa e Estados Unidos. Os trabalhos existentes exploram tal situação sob diversas perspectivas, incluindo causas, consequências, ações a serem tomadas e previsões. Vandamme, Meskens e Superby (2007) conduziram pesquisa com o objetivo classificar alunos de graduação de acordo com o risco de evasão (baixo, médio e alto riscos), alcançando 40,63% de acurácia de classificação. Em outros estudos, Allen et al. (2008) e Carvajal e Cervantes (2017) identificaram que o desempenho acadêmico tem influência na decisão de um aluno desligar-se do curso de graduação ou continuar até sua diplomação. Analisando os fatores que poderiam vir a influenciar no tempo de conclusão do curso ou evasão de alunos de graduação de uma IFES brasileira, Costa, Bispo e Pereira (2018) identificaram que o número de semestres do curso, gênero e desempenho acadêmico influenciam nas decisões dos alunos. Com propósitos semelhantes, Kantorski et al. (2015)

buscaram prever a probabilidade de evasão de alunos do curso de Zootecnia de outra IFES brasileira; os experimentos alcançaram uma taxa de sucesso de 74% na previsão dos alunos que abandonam a graduação. Neste contexto, entende-se que não estão esgotadas as possibilidades de pesquisas que possibilitem identificar antecipadamente quais alunos possivelmente terão êxito na conclusão da graduação e quais irão desvincular-se dos cursos de graduação antes do término, bem como a identificação dos fatores que levam a tais desfechos.

De maneira análoga a processos que reúnem grandes quantidades de informações, classificar alunos quanto à probabilidade de desligamento do curso valendo-se de elevado número de variáveis disponíveis em bancos de dados institucionais pode levar a resultados instáveis e pouco precisos. De tal forma, justifica-se a aplicação de abordagens voltadas à remoção das variáveis menos relevantes e ruidosas nos modelos de classificação (ANZANELLO et al., 2015). Além de aumentar a precisão de alocação de alunos à sua potencial classe de desfecho (diplomação, evasão para outro curso ou evasão da universidade), modelos mais enxutos facilitam a interpretação dos fatores que efetivamente influenciam e determinam o desfecho do aluno, servindo como norteadores para tomada de decisão dos gestores acadêmicos.

Este artigo propõe um método para selecionar as variáveis mais relevantes com vistas à classificação de alunos de graduação de acordo com 3 potenciais desfecho na universidade (diplomação, evasão interna ou evasão externa). O método proposto inicialmente coleta dados do perfil acadêmico do aluno ao ingressar no curso de graduação (por exemplo: semestre de ingresso, idade ao ingressar na graduação e média harmônica no vestibular), bem como dados relativos ao desempenho acadêmico nos semestres iniciais do aluno na universidade (por exemplo: número de créditos cursados e aprovados).

Na sequência, o método é aplicado a cada um dos semestres que tiveram dados coletados. Para isto, são formados conjuntos de dados compostos por dois subgrupos de variáveis: o subgrupo (i) de variáveis, relativas ao perfil acadêmico dos alunos no momento de ingresso na graduação, com informações em comum a todos os conjuntos; e o subgrupo (ii) de variáveis, relativas ao desempenho acadêmico, onde os dados apresentam o desempenho acadêmico dos alunos do início da graduação até a finalização do semestre que o conjunto representa. Na sequência, os quatro conjuntos de dados (1º semestre, 2º semestre, 3º semestre e

4º semestre), formados a partir da união dos dados do perfil acadêmico dos alunos e dados de desempenho acadêmico referentes a cada semestre, são divididos em porções de treinamento (T_r) e teste (T_s). Aplica-se então a sistemática de seleção de variáveis "omita uma variável por vez" (OUVV), a qual é integrada a cada uma das cinco ferramentas de classificação testadas (KNN, PNN, LDA, SVM e NB). Na sistemática OUVV, a cada repetição omite-se uma das variáveis e as acurácias de classificação são computadas; a variável responsável pela maior acurácia de classificação enquanto omitida é permanentemente removida. Iterações semelhantes são conduzidas sobre as variáveis remanescentes até restar apenas uma variável. Concluída a OUVV, os melhores subconjuntos de variáveis apontados por cada classificador são utilizados na classificação das observações inseridas na porção de teste (T_s) e a acurácia resultante é calculada.

O método foi aplicado a dados de 1421 alunos de graduação de cursos de engenharia, ingressantes por vestibulares nos anos de 2008 e 2009, em uma instituição federal de ensino superior (IFES). Dados do perfil dos alunos e de desempenho acadêmico dos quatro primeiros semestres de cada um dos alunos, totalizando nove variáveis, serviram como base para validação do método. O método proposto alcançou acurácia de classificação de 91,22%, retendo apenas duas variáveis (total de créditos cursados e total de créditos aprovados) das 9 variáveis originais, utilizando a ferramenta *Naïve Bayes* (NB) como técnica de classificação. Além disso, informações relativas aos resultados obtidos no terceiro semestre acadêmico e informações de desempenho acadêmico mostraram-se mais relevantes na determinação do desfecho dos alunos.

4.2 Referencial Teórico

Esta seção traz a fundamentação acerca da identificação de fatores que influenciam estudantes a concluir ou não suas graduações, bem como propostas de abordagens para predição do desfecho de alunos com base em dados que descrevem seus perfis. Por fim, são apresentadas as ferramentas de classificação utilizadas no método proposto.

4.2.1 Estudos relacionados ao desempenho de estudantes

Pesquisas relacionadas ao desempenho de estudantes de graduação são desenvolvidas mundialmente, especialmente com foco no desempenho acadêmico discente e nos fatores que levam alunos a abandonarem os estudos. Baker e Yacef (2009) enfatizam o surgimento de uma nova área de pesquisa nos Estados Unidos e Europa: a Mineração de Dados Educacionais (EDM – Educational Data Mining). A EDM objetiva desenvolver novos métodos para análise de dados coletados em ambientes educacionais. Tendências e mudanças na EDM são discutidas por diversos autores, especialmente com vistas a estimar a probabilidade de alunos de graduação desvincularem-se dos cursos antes do término, bem como identificar os fatores que influenciam os alunos a finalizarem ou não suas graduações. Baker, Isotani e Carvalho (2011) relacionam a crescente utilização da EDM no Brasil, apresentando resultados de pesquisas realizadas na área e listando oportunidades decorrentes da aplicação da EDM no país. Entre as pesquisas realizadas no tema, percebem-se dois grupos com propósitos distintos: aquelas voltadas à identificação dos fatores que influenciam na decisão dos alunos por concluir com êxito a graduação ou desligar-se do curso antes do término, e aquelas que buscam estimar a probabilidade de os alunos virem a evadir dos cursos de graduação.

Allen et al. (2008) identificaram que o desempenho acadêmico durante os primeiros períodos da graduação é um dos principais fatores que contribuem para conclusão com êxito ou não das graduações. Além do desempenho acadêmico na graduação, o estudo aponta que desempenhos acadêmicos e educacionais pré-universitários contribuem nas decisões dos alunos. Carvajal e Cervantes (2017) tiveram por objetivo identificar os principais motivadores nas decisões de alunos de abandonarem a graduação do turno noturno de universidades no Chile. Alinhado aos resultados da pesquisa de Allen et al. (2008), o estudo apontou o desempenho acadêmico como fator relevante, além de sinalizar questões pessoais como fatores motivadores na evasão dos alunos. Em estudo realizado por Ma e Cragg (2013), fatores como idade, sexo e etnia foram listados como influenciadores nas evasões ou diplomações, além do desempenho acadêmico. A mesma pesquisa ainda aponta que alunos que evadiram de seus cursos após longo período (evasões tardias) não possuíam registros de novos ingressos em faculdades. Diferente das pesquisas anteriores que estavam centradas nas características dos alunos, a pesquisa realizada por Chen (2012) priorizou a identificação das características

institucionais capazes de contribuir na redução do risco de evasão de estudantes de graduação. Constatou-se que gastos institucionais em serviços voltados aos alunos estão negativamente relacionados com o comportamento de desligamento dos estudantes.

Com objetivo de classificar alunos de graduação em três grupos de acordo com o risco de evasão (baixo, médio e alto riscos) antes do término do primeiro ano, Vandamme, Meskens e Superby (2007) desenvolveram uma abordagem que alcançou 40,63% de acurácia média, sendo 48,65% para estudantes de alto risco, 18,46% para estudantes de médio risco e 60,34% para estudantes de baixo risco. O levantamento de dados foi feito através de questionário aplicado a 533 alunos, e as variáveis identificadas como mais significativas para classificar os alunos relacionavam-se à frequência semanal e à sensação por parte dos alunos de terem escolhido corretamente a universidade. Com propósitos semelhantes (avaliar os riscos de evasão e analisar os fatores capazes de manter os alunos nos cursos de graduação), Delen (2010) analisou 5 anos de dados institucionais através de técnicas de mineração de dados, obtendo 81,18% de acurácia utilizando a ferramenta de classificação SVM. O autor concluiu ainda que fatores educacionais e financeiros são os mais importantes para manutenção dos alunos em seus cursos.

Apesar da grande relevância do tema, no Brasil poucas pesquisas foram desenvolvidas com vistas a identificar os fatores que mais influenciam no desfecho acadêmico de alunos de graduação. Costa, Bispo e Pereira (2018) analisaram os fatores tidos como relevantes para o tempo de conclusão do curso de graduação de uma IFES brasileira; concluíram que o número de semestres do curso, gênero e desempenho acadêmico impactam tanto no tempo de permanência do discente na universidade quanto no risco de evasão. Além disso, o estudo apontou que variáveis como idade no momento de ingresso, estado civil, raça e ter estudado em escolas públicas ou privadas durante a educação básica não demonstraram relevância para permanência ou evasão de alunos. Estudo realizado por Bonaldo e Pereira (2016) em instituição privada de ensino superior apontou que idade, mudança de estado civil durante o curso, ter bolsa de estudo e financiamento educacional são fatores que influenciam na permanência ou não dos alunos na instituição. Com propósitos semelhantes, Sales et al. (2016) analisaram 35 variáveis descrevendo estudantes da Universidade Federal do Espírito Santos (UFES) com o propósito de identificar as mais relevantes para diplomação ou evasão dos estudantes. Sete

variáveis foram apontadas como mais informativas: duas são características individuais do estudante, quatro estão associadas a experiências institucionais e apenas uma variável refere-se ao desempenho acadêmico do estudante.

Com o propósito de prever o risco de evasão de estudantes de graduação de Engenharia Civil da Universidade Federal do Rio de Janeiro (UFRJ), Manhães et al. (2011) alcançaram acurácias variando entre 75% e 80% utilizando distintas técnicas de mineração de dados aplicadas aos coeficientes de rendimentos dos alunos em disciplinas do primeiro semestre acadêmico. Com proposta semelhante, Kantorski et al. (2015, 2016) buscaram prever a evasão de alunos de graduação utilizando dados pessoais, acadêmicos, sócio econômicos e institucionais dos estudantes da Universidade Federal de Santa Maria (UFSM). No primeiro estudo, Kantorski et al. (2015) utilizaram dados de estudantes do curso de graduação em Zootecnia e alcançaram acurácia média de 75% das previsões de alunos que abandonaram o curso. Em Kantorski et al. (2016), dados de estudantes do curso de graduação em Administração os estudos permitiram obter acurácia de 73% das previsões de alunos que abandonariam os cursos. Além destes estudos, Júnior, Noronha e Kaestner (2018) alcançaram acurácias entre 83% e 97% na classificação de alunos com risco de evasão. Ainda, Lanes e Alcântara (2018), com o objetivo de identificar alunos de graduação que apresentam risco de evasão na Universidade Federal do Rio Grande (FURG), alcançaram 90,7% de acurácia de classificação.

Apesar da importância do tema, entende-se que este número reduzido de referências não esgota a possibilidade de desenvolvimento de novos métodos capazes de classificar estudantes de acordo com seu desfecho acadêmico focada na identificação das variáveis mais relevantes na definição do desfecho do aluno perante a universidade. O alinhamento de sistemáticas capazes de identificar subconjuntos de variáveis mais relevantes no processo de classificação, como a sistemática “omita uma variável por vez” (OUVV), com diversas ferramentas de classificações disponíveis na literatura pode contribuir tanto no avanço técnico dos métodos quanto na gestão acadêmica.

4.2.2 Ferramentas de classificação

O método proposto foi construído utilizando cinco ferramentas de classificação: k-vizinhos mais próximos (KNN), Rede Neural Probabilística (PNN), Análise Discriminante

Linear (LDA), Máquina de Suporte Vetorial (SVM) e *Naïve Bayes* (NB). Os fundamentos de tais classificadores são agora apresentados.

KNN (*k-Nearest Neighbor*) é um método de classificação supervisionado, proposto por Fix e Hodges (1989). O algoritmo considera cada observação como um ponto em seu espaço amostral n-dimensional. Para cada nova observação uma classe é atribuída, tendo como base a distância (tipicamente Euclidiana) entre ela e as *k* observações vizinhas mais próximas dentro do espaço amostral. As classes das *k* observações vizinhas são conhecidas, e a nova observação pertencerá a mesma classe em que a maioria das *k* observações vizinhas estão classificadas. Esta técnica destaca-se por sua simplicidade e por requerer apenas um parâmetro, *k* (número de vizinhos a ser analisados), que pode ser definido por validação cruzada (BARBON et al., 2016).

A segunda ferramenta de classificação utilizada foi a Rede Neural Probabilística (PNN - *Probabilistic Neural Network*). A PNN utiliza todas as observações existentes, calculando a distância euclidiana entre uma observação e todas as outras para definir a qual classe tal observação pertence. A distância medida de uma observação em relação a todas as outras é transformada por uma função exponencial padrão, que através de um parâmetro sigma escala a similaridade entre a observação e todas as outras. Os valores transformados são separadamente somados de acordo com cada classe, assim a PNN atribui à observação a classe que apresentar maior somatório. Mais detalhes sobre redes neurais probabilísticas podem ser encontrados em Duda, Hart e Stork (2001).

A Análise Discriminante Linear (LDA - *Linear Discriminant Analysis*), terceira ferramenta utilizada, foi proposta por Fisher (1936), e deriva combinações lineares (funções discriminantes). A LDA realiza uma projeção dos dados originais em um espaço dimensional menor através de funções discriminantes; o número de funções a serem construídas é proporcional ao número de classes. Os coeficientes das funções discriminantes são estimados para que seja maximizada a variância entre as classes e minimizada a variância dentro das classes (HAIR et al., 2009). Para classificar as observações, as variáveis são inseridas nas funções discriminantes, e o valor gerado pelas funções é comparado com um limite que define a qual classe a observação pertence. Mais detalhes e fundamentos matemáticos podem ser encontrados em Rencher (2002).

A ferramenta de classificação Máquina de Suporte Vetorial (SVM - Support Vector Machine), quarta ferramenta utilizada, é um algoritmo de aprendizagem de máquina supervisionado introduzido por Cortes e Vapnik (1995). Em um espaço amostral *n* dimensional onde estão dispostas as observações, a ferramenta SVM constrói um hiperplano de modo a separar as observações em classes autênticas e não autênticas. Seu algoritmo objetiva encontrar um hiperplano que maximize a distância entre as duas observações mais próximas, uma pertencente à classe autêntica e outra pertencente à classe não autêntica. Mais detalhes podem ser vistos em Cristianini e Shawe-Taylor (2000).

A quinta e última ferramenta utilizada foi *Naïve Bayes* (NB), um classificador probabilístico supervisionado. Este classificador, baseado no Teorema de Bayes, calcula a probabilidade de determinada observação pertencer a cada uma das classes existentes com base em termos de probabilidade Por ter fácil aplicação e elevado desempenho de classificação, a ferramenta NB é amplamente utilizada em problemas de classificação e aprendizado de máquina (FARID et al., 2014)..

4.3 Descrição das Variáveis Analisadas

Esta seção apresenta as 9 variáveis que são avaliadas no método proposto. As variáveis foram divididas em dois subgrupos: (i) variáveis do perfil acadêmico do aluno no momento de ingresso na graduação; e (ii) variáveis de desempenho acadêmico. O primeiro subgrupo é composto por 6 variáveis que correspondem ao perfil acadêmico do aluno no momento de ingresso no curso de graduação. O segundo subgrupo, com 3 variáveis, retrata o desempenho acadêmico dos estudantes desde o início do curso até a finalização de cada semestre acadêmico avaliado. No estudo de caso apresentado foram analisados do primeiro ao quarto semestre acadêmico, desse modo quatro coletas foram realizadas para as variáveis do subgrupo (ii); a Figura 4.1 ilustra o procedimento de coleta desse subgrupo de variáveis. A Tabela 4.1 apresenta as 9 variáveis, bem como suas descrições e os valores que as variáveis podem assumir. Ressalta-se que os valores apresentados na Tabela 4.1 (como média harmônica e idade no início do curso) foram obtidos com base nos dados coletados nos anos de 2008 e 2009, e retratam apenas tal conjunto, de tal forma que alunos ingressantes em outros anos podem gerar alterações nestes valores.

Tabela 4.1 - Variáveis selecionadas para o estudo de caso (alunos ingressantes em 2008 e 2009)

		Variável	Descrição da variável	Possíveis valores
Subgrupo (i)	(a)	Semestre de ingresso	Representa semestre letivo em que o aluno inicia suas as atividades acadêmicas junto à IES.	0 - Primeiro; 1 – Segundo
	(b)	Reserva de vaga de ingresso	Representa a opção de reserva de vaga que o aluno concorreu no concurso vestibular no qual foi aprovado seu ingresso no curso de graduação.	0 - Acesso Universal; 1 - Ensino Público; 2 - Ensino Público autodeclarado Negro
	(c)	Idade no início do curso	Representa a idade do aluno no momento do início do semestre de ingresso.	Valores entre 16 e 54
	(d)	Média harmônica do vestibular	Representa o cálculo realizado sobre o escore padronizado e o peso de cada prova do vestibular, definindo a pontuação do aluno.	Valores entre 377,12 e 759,04
	(e)	Tempo de conclusão do ensino médio	Representa a diferença entre os anos de início da graduação e o de término do ensino médio.	Valores entre 1 e 36
	(f)	Benefícios	Representa se o aluno recebeu algum benefício durante os semestres letivos.	0 - Não recebeu; 1 - Recebeu
Subgrupo (ii)	($g^{(*)}$)	Total de créditos cursados	Representa o total de créditos cursados em atividades acadêmicas pelo aluno do início do curso até o fim do semestre acadêmico analisado.	Valores entre 0 e 151
	($h^{(*)}$)	Total de créditos aprovados	Representa o total de créditos aprovados em atividades acadêmicas pelo aluno do início do curso até o fim do semestre acadêmico analisado.	Valores entre 0 e 132
	($i^{(*)}$)	Taxa de aprovação	Representa a taxa entre Total de créditos cursados e Total de créditos aprovados.	Valores entre 0 e 1

* Variável coletada de forma independente quatro vezes. Do início do curso até a finalização do semestre que representou.

Fonte: Elaborada pelos autores.

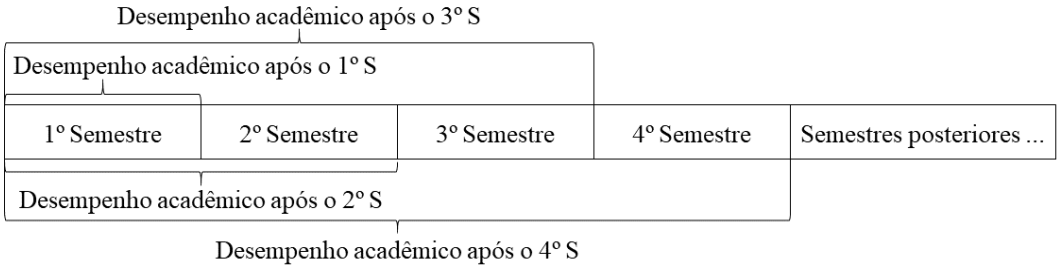

Figura 4.1 - Ilustração da coleta de dados de desempenho para o subgrupo (ii)

Conforme ilustrado na Figura 4.1, as variáveis relativas aos dados de desempenho acadêmicos foram coletadas do início do curso de graduação até a finalização de cada um dos quatro semestres analisados. As variáveis relativas ao primeiro semestre ($g^{(1)}$, $h^{(1)}$, $i^{(1)}$)

compreendem os créditos dos estudantes do início do curso até a finalização do primeiro semestre. Da mesma maneira, as variáveis relativas aos dados de desempenho acadêmicos do segundo semestre ($g^{(2)}$, $h^{(2)}$, $i^{(2)}$) foram coletadas de modo a representar a trajetória acadêmica dos estudantes desde o início do curso até a finalização do segundo semestre (acumulando, de tal forma, acumulando os desempenhos dos semestres anteriores). A coleta das variáveis para o terceiro ($g^{(3)}$, $h^{(3)}$, $i^{(3)}$) e quarto ($g^{(4)}$, $h^{(4)}$, $i^{(4)}$) semestre deu-se de forma semelhante, do início do curso de graduação até a finalização do referido semestre, de modo a compreender não somente o desempenho do estudante no semestre em questão, mas acumular os desempenhos de semestres anteriores.

4.4 Método

O método proposto visa identificar um subconjunto reduzido de variáveis mais informativas capazes de aprimorar a classificação de alunos de graduação de acordo com 3 potenciais desfechos no curso (entendidas como classes). Para tal, utilizam-se variáveis de dados do perfil do aluno no momento do ingresso no curso de graduação e dados de desempenho acadêmico. É requerido que, ao final de seu ciclo acadêmico, os estudantes que terão seus dados coletados estejam inseridos em classes relativas a uma das seguintes situações: diplomação, evasão interna (troca de curso dentro da mesma IES) ou evasão externa. O método sugerido se apoia em quatro passos operacionais: (i) seleção dos alunos de graduação e coleta dos dados que os caracterizam; (ii) formação de conjuntos de dados; (iii) aplicação da sistemática OUVV em conjunto com ferramentas de classificação, e (iv) determinação dos melhores subconjuntos de variáveis. Estes passos são detalhados na sequência.

4.4.1 Primeiro passo - Seleção dos alunos de graduação e coleta dos dados que os caracterizam

O primeiro passo do método proposto visa identificar os alunos de graduação que terão seus dados coletados. As variáveis coletadas e que representam os alunos são apresentadas na seção 4.3, estando divididas em 2 subgrupos: (i) perfil acadêmico do aluno no momento de ingresso na graduação, e (ii) dados de desempenho acadêmico, sendo este subgrupo diferente para cada um dos conjuntos de dados a serem formados e posteriormente analisados. Deseja-se

que os alunos cujos dados serão coletados tenham concluído seu ciclo acadêmico no referido curso de graduação, e que seu desfecho tenha se dado em uma das seguintes situações: diplomação, evasão interna ou evasão externa. Tal condição permitirá ao modelo avaliar como flutuações das variáveis de entrada impactam no enquadramento dos alunos em termos de seus desfechos.

4.4.2 Segundo passo – Formação dos conjuntos de dados

Quatro conjuntos de dados são formados a partir da totalidade de dados coletados. A formação dos conjuntos de dados se dá após a coleta das variáveis dos subgrupos (i) e (ii). A Figura 4.2 ilustra a formação destes conjuntos de variáveis independentes, que recebem as nomenclaturas a seguir, para facilitar a compreensão: 1º semestre, 2º semestre, 3º semestre e 4º semestre. Em comum, os conjuntos possuem as variáveis do subgrupo (i), e individualmente, as variáveis do subgrupo (ii) são alocadas de acordo com semestre que o conjunto representa. As observações coletadas devem estar classificadas em uma das 3 classes de desfecho acadêmico possíveis (0 - diplomação, 1 - evasão interna ou 2 - evasão externa). Na sequência, os conjuntos de dados são subdivididos em porções de treinamento (T_r), contendo n_{Tr} observações, e teste (T_s), contendo n_{Ts} observações, de modo que $n_{Tr} + n_{Ts} = n$.

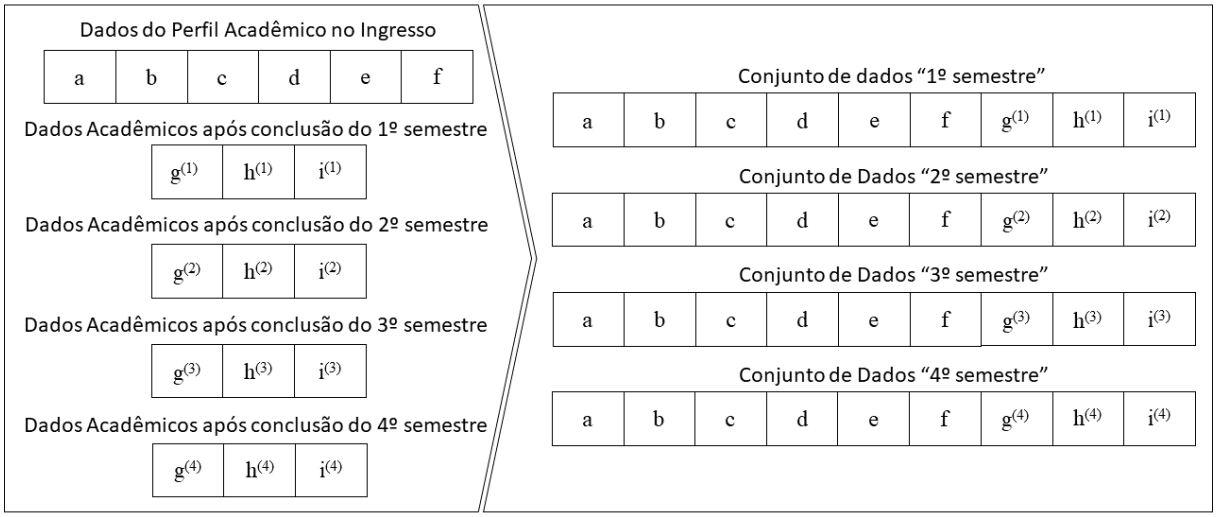

Figura 4.2 - Formação dos quatro conjuntos de dados

4.4.3 Terceiro passo – Aplicação da sistemática OUVV em conjunto com ferramentas de classificação

O terceiro passo consiste na aplicação da sistemática “omita uma variável por vez” (OUVV) à porção de treinamento (T_r) dos conjuntos de dados, utilizando cada uma das cinco ferramentas de classificação (KNN, PNN, LDA, SVM e NB). Tal procedimento é detalhado na sequência.

A sistemática “omita uma variável por vez” (OUVV), descrita em Caruana e Freitag (1994), é utilizada com a finalidade de selecionar as variáveis mais informativas para classificação dos estudantes. A cada iteração da sistemática OUVV, as acurácias de classificação são computadas com a finalidade de auxiliar na escolha do subconjunto de variáveis com maior capacidade de classificar corretamente os estudantes de acordo com seu desfecho acadêmico. Neste estudo a acurácia é definida como a razão entre o número de classificações corretas e o número total de classificações.

Um processo iterativo apoiado na sistemática OUVV é iniciado para remover as variáveis menos informativas. Para determinar qual variável deve ser removida, a cada iteração uma variável é omitida e a acurácia de classificação é verificada. Remove-se permanentemente a variável que, quando omitida, resultou na maior acurácia de classificação, visto que sua omissão elevou a acurácia de classificação, comprovando sua contribuição limitada no processo. Caso mais de uma variável apresentar a maior acurácia quando omitida, seleciona-se aleatoriamente uma delas para ser removida. O processo é executado até restar apenas uma variável. Esse procedimento é executado para cada uma das 5 ferramentas de classificação apresentadas.

4.4.4 Quarto passo - Determinar o melhor subconjunto de variáveis para cada semestre em análise

No quarto passo do método, analisam-se as acurácias obtidas no passo anterior a fim de selecionar os subconjuntos de variáveis a serem utilizadas em classificações futuras. É recomendado reter o subconjunto de variáveis que obteve maior acurácia de classificação após a finalização do passo anterior. Caso mais de um subconjunto de variáveis conduzir à mesma

acurácia de classificação, recomenda-se que o subconjunto escolhido seja o que retém o menor número de variáveis.

Após selecionar o melhor subconjunto de variáveis, aplica-se novamente a ferramenta de classificação sobre as observações da porção de teste (T_s). As acurácias resultantes destas classificações compõem os resultados finais deste método.

4.5 Estudo de Caso

O método proposto foi aplicado em uma instituição federal de ensino superior (IFES) brasileira. Altos índices de evasão vêm sendo verificados nas instituições de ensino brasileiras, ocasionando perdas em diversas dimensões e contextos. Detectar alunos que possam vir a desligar-se prematuramente de seus cursos de graduação é uma tarefa importante para auxiliar os gestores na elaboração de estratégias voltadas à redução dos índices de evasão. É importante ressaltar que o método proposto não é capaz de identificar em qual período do curso o aluno irá desligar-se, apenas prever o potencial desfecho do aluno com base nas suas variáveis de entrada e desempenho ao longo dos 4 primeiros semestres. A universidade em análise possui grande número de cursos de graduação divididos em diversas áreas (engenharias, artes, saúde, entre outras). Trabalhar com diversos perfis diferentes de alunos oriundos de áreas diversas poderia causar distorções nas acurácias de classificação; desta forma, apenas cursos da área de engenharia foram avaliados pelo método proposto.

Foram coletados dados de 1421 alunos que ingressaram em seus cursos por meio do vestibular nos anos de 2008 e 2009. Este período foi escolhido por considerar que grande parte dos ingressantes destes anos já estariam desligados de seus cursos, visto que o período de integralização de créditos dos cursos de engenharias estão entre 5 e 10 anos. As coletas de dados foram realizadas de modo a obter as 9 variáveis (ver Tabela 4.1), divididas em dois subgrupos, necessárias à formação de quatro conjuntos de dados (1º semestre, 2º semestre, 3º semestre e 4º semestre) a serem analisados. Os quatro conjuntos de dados foram formados seguindo o especificado no segundo passo do método proposto, onde foram unidas as variáveis do subgrupo (i) às variáveis do subgrupo (ii), de modo a representar o perfil acadêmico do aluno ingressante, e o desempenho acadêmico de tal aluno no semestre que o conjunto representou. Na sequência, os conjuntos de dados foram normalizados e divididos em porções de

treinamento e teste através do algoritmo Kennard-Stone (KS), sendo utilizadas as proporções de 75% (treinamento) / 25% (teste). Mais detalhes sobre o KS podem ser encontrados em Kennard e Stone (1969).

A sistemática OUVV foi aplicada as porções de treinamento (T_r) dos conjuntos de dados. Para cada um dos quatro conjuntos de dados (1º semestre, 2º semestre, 3º semestre e 4º semestre), a OUVV foi aplicada utilizando as 5 ferramentas de classificação descritas na seção 4.2.2. A cada iteração da sistemática, as acurácias de classificação foram armazenadas, de modo a selecionar o subconjunto de variáveis que conduziu às maiores acurácias nesta etapa do método proposto.

Fazendo uso das porções de treinamento (T_r) para ajuste do modelo e das porções de testes (T_s) para validá-lo, os subconjuntos de variáveis retidas foram novamente classificados de modo a obter os resultados finais do método. Na sequência são apresentados os resultados obtidos nos terceiro e quarto passos do método para os quatro conjuntos de dados. Quando ferramentas de classificação distintas conduziram a acurácias idênticas, optou-se pela ferramenta que apresentou o menor número de variáveis (menos variáveis são desejadas). Para melhor apresentação dos resultados, as variáveis foram representadas por letras conforme a Tabela 4.1.

4.5.1 1º semestre

Quando aplicado o método ao conjunto de dados que representa o primeiro semestre acadêmico, a ferramenta de classificação *Naïve Bayes* (NB) apresentou a melhor acurácia de classificação dentre as ferramentas utilizadas. Apesar da ferramenta Máquina de Vetor de Suporte (SVM) obter a mesma acurácia de classificação que a ferramenta NB, esta última reteve menor número de variáveis, portanto considerada a melhor ferramenta neste conjunto de dados. Perceba que 3 níveis distintos de vizinhos mais próximos são testados para a ferramenta de classificação KNN (1, 3 e 5).

Finalizando o processo iterativo promovido pela sistemática OUVV, verificou-se quais subconjuntos de variáveis apresentaram as melhores acurácias para cada uma das ferramentas de classificação; essas foram então utilizadas para classificar a porção de testes (T_s) do conjunto

de dados 1º semestre. Ao reter apenas 3 das 9 variáveis originais, a ferramenta NB obteve 81,46% de classificações corretas conforme apresentado na Tabela 4.2. As 3 variáveis retidas pertencem ao subgrupo (ii) de variáveis (total de créditos cursados, total de créditos aprovados, e taxa de aprovação). Verificaram-se ainda as acurácias de classificação obtidas em cada uma das 3 classes com a ferramenta NB, que foram de aproximadamente 93%, 28% e 28% para as classes diplomação, evasão interna e evasão externa, respectivamente.

Tabela 4.2 - Acurácia na porção de testes e variáveis retidas para cada ferramenta de classificação no conjunto de dados 1º semestre

Ferramenta	Acurácia (%)	Variáveis
KNN (k=1)	69,66	d;i
KNN (k=3)	80,61	a;b;c;d;e;f;g;i
KNN (k=5)	80,61	a;b;c;e;h;i
PNN	80,89	a;b;c;d;e;f;g;h;i
LDA	81,18	b;c;d;h;i
SVM	81,46	a;b;c;h;i
NB	**81,46**	**g;h;i**

4.5.2 2º semestre

Ao aplicar o método proposto ao conjunto de dados 2º semestre, novamente a ferramenta de classificação Naïve Bayes (NB) obteve a melhor acurácia de classificação, conforme detalhado na Tabela 4.3, onde são exibidas acurácias de classificação da porção de teste (T_s) obtidas com as cinco ferramentas de classificação utilizadas. Retendo apenas duas das 9 variáveis originais, a ferramenta obteve acurácia de 88,38% nas classificações. Apesar de obter a mesma acurácia, a ferramenta de classificação LDA reteve maior número de variáveis.

Finalizada a aplicação da sistemática OUVV, pode-se verificar que apenas duas variáveis foram retidas para sequência do método com a ferramenta NB . Pode-se verificar que as duas variáveis a serem retidas (total de créditos cursados e total de créditos aprovados). Assim como na aplicação do método com a ferramenta NB ao conjunto de dados 1º semestre, as variáveis retidas pertencem exclusivamente ao subgrupo (ii) de variáveis, que representam o desempenho acadêmico dos alunos. Retendo apenas o subconjunto das duas variáveis destacadas a ferramenta de classificação NB obteve a acurácia de classificação de 88,38% na

porção de testes (T_s). Acurácias de classificação de aproximadamente 96%, 22% e 64% foram obtidas para as classes diplomação, evasão interna e evasão externa respectivamente com tal ferramenta.

Tabela 4.3 - Acurácia na porção de testes e variáveis retidas para cada ferramenta de classificação no conjunto de dados 2º semestre

Ferramenta	Acurácia (%)	Variáveis
KNN (k=1)	74,22	d;i
KNN (k=3)	86,40	a;b;c;d;e;f;g;h;i
KNN (k=5)	86,68	b;c;d;e;f;g;h;i
PNN	86,40	a;b;c;d;e;f;g;h;i
LDA	88,38	d;g;h
SVM	87,81	a;b;c;d;e;f;h;i
NB	**88,38**	**g;h**

4.5.3 3º semestre

Da mesma forma que na aplicação do método aos dois conjuntos de dados anteriores, a ferramenta *Naïve Bayes* (NB) obteve o melhor resultado na aplicação do método, tendo como base o reduzido número de variáveis requeridas para classificação das observações. Apesar de obter acurácia de classificação levemente inferior a ferramenta PNN, essa foi obtida com menor número de variáveis retidas.

O processo promovido pela sistemática OUVV apontou a retenção de apenas duas das 9 variáveis originais ao utilizar a ferramenta NB. Retendo tais variáveis (total de créditos cursados e total de créditos aprovados), a ferramenta de classificação NB obteve acurácia de 91,22% quando aplicada à porção de testes (T_s) do conjunto de dados 3º semestre. Os resultados da aplicação das cinco ferramentas de classificação às porções de testes (T_s) são apresentados na Tabela 4.4. Ressalta-se novamente o fato do subconjunto de variáveis retidas pertencer exclusivamente ao subgrupo (ii), que representam o desempenho acadêmico dos alunos. Ainda, com a ferramenta NB aplicada aos dados do 3º semestre, foram obtidas as acurácias de aproximadamente 95%, 31% e 77% para as classes diplomação, evasão interna e evasão externa, respectivamente.

Tabela 4.4 - Acurácia na porção de testes e variáveis retidas para cada ferramenta de classificação no conjunto de dados 3º semestre

Ferramenta	Acurácia (%)	Variáveis
KNN (k=1)	85,38	d;i
KNN (k=3)	90,64	a;b;c;d;e;h;i
KNN (k=5)	91,22	b;c;d;g;i
PNN	91,81	a;b;c;d;e;f;g;h;i
LDA	90,93	c;d;e;g;h
SVM	91,52	a;b;c;d;e;f;h;i
NB	**91,22**	**g;h**

4.5.4 4º semestre

Aplicando o método ao conjunto de dados que representava o quarto semestre acadêmico a ferramenta Máquina de Vetor de Suporte (SVM) obteve a melhor acurácia de classificação quando aplicada a porção de testes (T_s), conforme a Tabela 4.5. A ferramenta LDA conduziu a mesma acurácia de classificação, porém reteve maior número de variáveis.

A sistemática OUVV apontou o subconjunto de 5 variáveis a serem retidas para sequência da aplicação do método com a ferramenta SVM. Verifica-se que o subconjunto de variáveis retidas foi representado por 5 variáveis (Semestre de ingresso, Reserva de vaga de ingresso, Média harmônica do vestibular, Total de créditos cursados, e Total de créditos aprovados). O subconjunto de 5 variáveis conduziu a 91,69% de classificações acuradas na porção de testes (T_s) com a ferramenta SVM, sendo obtidas as acurácias de aproximadamente 95%, 56% e 62% respectivamente para as classes diplomação, evasão interna e evasão externa.

Tabela 4.5 - Acurácia na porção de testes e variáveis retidas para cada ferramenta de classificação no conjunto de dados 4º semestre

Ferramenta	Acurácia (%)	Variáveis
KNN (k=1)	78,63	d;i
KNN (k=3)	86,94	a;c;d;g;i
KNN (k=5)	90,20	c;d;e;g;h;i
PNN	90,20	a;b;c;d;e;f;g;h;i
LDA	91,69	a;b;d;e;g;i
SVM	**91,69**	**a;b;d;g;i**
NB	91,39	b;d;e;f;h;i

4.5.5 Discussão dos resultados

O método proposto alcançou acurácia de classificação de 91,22% ao utilizar-se o conjunto de dados do terceiro semestre acadêmico (3º semestre), retendo apenas 22% das variáveis, ou seja, duas das 9 variáveis originais. Além disso, ressalta-se que todos os resultados obtidos com este conjunto de dados foram superiores aos resultados obtidos com os conjuntos de dados que representavam os outros semestres acadêmicos (1º semestre, 2 º semestre e 4 º semestre), sugerindo informações relativas ao terceiro semestre como recomendadas para identificação dos estudantes que terão êxito ou não na graduação.

Paralelamente o estudo de caso mostra que, dentre as variáveis avaliadas, as variáveis de desempenho acadêmico são as mais relevantes na determinação do desfecho de estudantes de graduação. Considerando os melhores resultados obtidos nas análises, os três primeiros conjuntos de dados obtiveram as melhores acurácias de classificação utilizando apenas variáveis do subgrupo (ii) de variáveis de desempenho acadêmico dos estudantes (número total de créditos cursados, número total de créditos aprovados, taxa de aprovação). Tal fato vai ao encontro de outros estudos que apontaram o desempenho acadêmico como fator determinante na conclusão ou evasão dos estudantes (ALLEN et al., 2008; CARVAJAL; CERVANTES, 2017).

Considerando as ferramentas selecionadas nas análises realizadas, pode-se realizar uma análise mais detalhada relativas ao desempenho de cada uma das 3 classes. Variáveis de desempenho acadêmico dos estudantes foram retidas de forma exclusiva pelas ferramentas escolhidas em 3 das 4 análises. Dentre as 3 classes, a que se mostrou menos acurada foi a de evasão interna. Tais acurácias podem estar diretamente ligadas ao fato do desempenho acadêmico dos estudantes que realizam transferência de curso dentro da universidade ser muito semelhante aos demais estudantes, os quais diplomaram em seus cursos de origem. Possivelmente esses estudantes apenas repensaram sua escolha na graduação, vindo a diplomar-se em seus novos cursos.

Por outro lado, o desempenho das ferramentas de classificação se mostrou mais acurado na classe de alunos que abandonaram a universidade, evasão externa, especialmente a partir do segundo semestre acadêmico. A maior acurácia se deu com a ferramenta de classificação NB

quando aplicada ao conjunto de dados 3º semestre, obtendo 75% de classificações corretas. Ao realizar uma análise das variáveis retidas, observa-se que estudantes que abandonam a universidade após o 3º semestre possuem média aproximada de 47 créditos cursados ao final o 3º semestre, e aproveitamento inferior a metade. Enquanto isso, os outros estudantes possuem media aproximada de 74 créditos cursados com aproveitamento superior a 80% após a finalização do mesmo semestre.

Por fim, entende-se que os resultados obtidos podem fornecer aos gestores uma nova perspectiva dos fatores que preponderantemente levam estudantes a desligarem-se dos cursos, permitindo o desenvolvimento de medidas com o propósito de aprimorar o monitoramento do desempenho acadêmico dos estudantes, levando esses a obter êxito na conclusão de suas graduações.

4.6 Conclusão

Este artigo apresentou um método para selecionar as variáveis mais informativas para predição do desfecho de alunos em cursos de graduação de engenharia: diplomação, evasão interna (troca de curso dentro da mesma IES) ou evasão externa. As variáveis mais informativas foram selecionadas após a aplicação da sistemática “omita uma variável por vez” (OUVV). Cinco ferramentas de classificação foram testadas no método proposto: k-vizinhos mais próximos (KNN), Rede Neural Probabilística (PNN), Análise Discriminante Linear (LDA), Máquina de Vetor de Suporte (SVM) e N*aïve Bayes* (NB).

Quando aplicado a dados reais de uma instituição federal de ensino superior (IFES) brasileira, o método proposto alcançou acurácia de classificação de 91,22% na porção de testes, retendo apenas 2 das 9 variáveis originais (número total de créditos cursados e número total de créditos aprovados), com a ferramenta de classificação NB. Verificou-se ainda que informações do terceiro semestre acadêmico se mostraram as mais relevantes na determinação do desfecho dos estudantes. Além disso, alinhada com os resultados obtidos nos estudos realizados por Allen et al. (2008) e Carvajal e Cervantes (2017), a abordagem proposta identificou que o desempenho acadêmico de alunos nos primeiros semestres é fator determinante para definir se o aluno permanecerá matriculado até sua diplomação ou deixará o curso antes do término. Desta maneira, o estudo sugere que o monitoramento constante do desempenho acadêmico pode ser

uma medida eficaz para conclusão com êxito da graduação de tal estudante.

Trabalhos futuros visam obter mais informações sobre a vida acadêmica e socioeconômica dos alunos com o propósito de obter acurácias de classificação mais precisas, bem como identificar os fatores que mais influenciam nas decisões de estudantes desvincularem-se de seus cursos de graduação. Sugere-se ainda, com a finalidade de avaliação dos resultados, a aplicação de um método que realize a classificação hierárquica dos conjuntos de dados, de modo a ser capaz de identificar melhores ferramentas e subconjuntos de variáveis que caracterizem cada uma das classes quando comparada com as restantes. Maiores conjuntos de dados, como inclusão de outros cursos de graduação, novos métodos e outras ferramentas também podem ser estudos promissores com o propósito de aprimoramento dos resultados, auxiliando no entendimento das causas e possíveis soluções para assegurar a diplomação dos estudantes.

4.7 Referências

ALLEN, J.; ROBBINS, S. B.; CASILLAS, A.; OH, I. S. Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. **Research in Higher Education**, v. 49, n. 7, p. 647–664, 12 nov. 2008.

ANZANELLO, M. J.; KAHMANN, A.; MARCELO, M. C.A.; MARIOTTI, K. C.; FERRÃO, M. F.; ORTIZ, R. S. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, v. 115, p. 562–569, nov. 2015.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 31 ago. 2011.

BAKER, R. S. J. D. R.; YACEF, K. The State of Educational Data Mining in 2009 : A Review and Future Visions. **Journal of Educational Data Mining**, v. 1, n. 1, p. 3–16, 1 out. 2009.

BARBON, A. P. A. C.; BARBON, S.; MANTOVANI, R. G.; FUZYI, E. M.; PERES, L. M.; BRIDI, A. M. Storage time prediction of pork by Computational Intelligence. **Computers and Electronics in Agriculture**, v. 127, p. 368–375, set. 2016.

BONALDO, L.; PEREIRA, L. N. Dropout: Demographic Profile of Brazilian University Students. **Procedia - Social and Behavioral Sciences**, v. 228, p. 138–143, 20 jul. 2016.

BRASIL. **Decreto Nº 6.096, de 24 de abril 2007. Institui o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI**, Brasil, 2007.

CARUANA, R.; FREITAG, D. Greedy Attribute Selection. In: **Machine Learning Proceedings 1994**. [s.l.] Elsevier, 1994. p. 28–36.

CARVAJAL, R. A.; CERVANTES, C. T. Aproximaciones a la deserción universitaria en Chile**. Educação e Pesquisa**, v. 44, n. 0, 4 set. 2017.

CHEN, R. Institutional Characteristics and College Student Dropout Risks: A Multilevel Event History Analysis. **Research in Higher Education**, v. 53, n. 5, p. 487–505, 15 ago. 2012.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

COSTA, F. J. DA; BISPO, M. DE S.; PEREIRA, R. DE C. DE F. Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. **RAUSP Management Journal**, v. 53, n. 1, p. 74–85, 1 jan. 2018.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines : and other kernel-based learning methods**. [s.l.] Cambridge University Press, 2000.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. **Decision Support Systems**, v. 49, n. 4, p. 498–506, 1 nov. 2010.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. New York: Wiley, 2001.

FARID, D. M.; ZHANG, L.; RAHMAN, C. M.; HOSSAIN, M.A.; STRACHAN, R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. **Expert Systems with Applications**, v. 41, n. 4, p. 1937–1946, mar. 2014.

FISHER, R. A. the use of multiple measurements in taxonomic problems. ANNALS OF EUGENICS, v. 7, n. 2, p. 179–188, 1 set. 1936.

FIX, E.; HODGES, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. **International Statistical Review / Revue Internationale de Statistique**, v. 57, n. 3, p. 238, dez. 1989.

HAIR, J.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E.; THATAM, R.L. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

JÚNIOR, J. G. DE O.; NORONHA, R. V.; KAESTNER, C. A. A. Método de Seleção de Atributos Aplicados na Previsão da Evasão de Cursos de Graduação**. Revista de Informática Aplicada**, v. 13, n. 2, 26 fev. 2018.

KANTORSKI, G. Z.; HOFFMANN, I. L.; LIMBERGER, S. J.; MULLER, F. Uma Visão Do Futuro: Previsão De Evasão Em Cursos De Graduação Presenciais De Universidades Públicas: O Caso Do Curso De Zootecnia. **XV Colóquio Internacional de Gestão Universitária**, 4 dez. 2015.

KANTORSKI, G. Z.; FLORES, E. G.; SCHMITT, J.; HOFFMANN, I.; BARBOSA, F. Predição da Evasão em Cursos de Graduação em Instituições Públicas. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 27, n. 1, p. 906, 7 nov. 2016.

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, v. 11, n. 1, p. 137–148, fev. 1969.

LANES, M.; ALCÂNTARA, C. Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE),** v. 29, n. Cbie, p. 1921, 28 out. 2018.

MA, Y.; CRAGG, K. M. So Close, Yet So Far Away: Early Vs. Late Dropouts. **Journal of College Student Retention: Research, Theory & Practice**, v. 14, n. 4, p. 533–548, 9 fev. 2013.

MANHÃES, L. M. B.; CRUZ, S. M. S.; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. **Anais do XXII SBIE - XVII WIE**, v. 1, n. 1, p. 150–159, 2011.

RENCHER, A. C. **Methods of multivariate analysis**. [s.l.] J. Wiley, 2002.

SALES, J. S.; BRASIL, G. H..; CARNEIRO, T. C. J.; CORASSA, M. A. C. Fatores Associados à Evasão e Conclusão de Cursos de Graduação Presenciais na UFES. **Meta: Avaliacao**, v. 8, n. 24, p. 488–514, 8 dez. 2016.

SILVA FILHO, R. L. L. E.; BRASIL, G. H..; CARNEIRO, T. C. J.; CORASSA, M. A. C. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641–659, dez. 2007.

VANDAMME, J. P.; MESKENS, N.; SUPERBY, J. F. Predicting Academic Performance by Data Mining Methods. **Education Economics**, v. 15, n. 4, p. 405–419, dez. 2007.

XING, W.; GUO, R.; PETAKOVIC, E.; GOGGINS, S. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. **Computers in Human Behavior**, v. 47, p. 168–181, jun. 2015.

5. Considerações finais

Este capítulo apresenta as conclusões da dissertação, além de sugestões para trabalhos futuros.

5.1 Conclusões

Esta dissertação teve como principal objetivo o desenvolvimento de estruturas multivariadas com vistas à análise de perfis humanos em diferentes segmentos. Técnicas multivariadas possuem uma ampla aplicabilidade prática, tanto na indústria quando no setor acadêmico. Através da análise da literatura científica, objetivos específicos foram definidos. São eles: (*i*) identificar os parâmetros oriundos da modelagem por CA que dão origem a grupos (clusters) consistentes de trabalhadores de acordo com seus perfis de aprendizagem; (*ii*) desenvolver um novo índice de importância para classificar trabalhadores de acordo com seus padrões de aprendizado com base nas regressões LASSO e PLS; e (*iii*) identificar as variáveis mais informativas em cenário acadêmico com vistas à classificação de alunos de graduação de acordo com seu desfecho (diplomação, transferência ou evasão).

O objetivo (*i*) foi atingido no primeiro artigo, o qual apresentou uma estrutura multivariada para obter grupos homogêneos de trabalhadores de acordo com suas características de aprendizado. Tal estrutura, integrou modelagem de curva de aprendizado e análise de agrupamento em suas etapas operacionais. Diferentes aspectos da aprendizagem dos trabalhadores foram obtidos através da modelagem de CA; um índice de importância de variável (IIV) baseado nas saídas da ACP, quando aplicada a dados oriundos das CAs, serviu como base para a seleção dos parâmetros mais relevantes para o propósito de agrupamento. Três métricas foram utilizados para avaliar a qualidade dos grupos formados: *Silhouette Index* (SI), *Calinski-Harabasz* (CH) e *Davies-Bouldin* (DB). Aplicado a dados reais, a estrutura sugeriu a formação de dois grupos com distintos padrões de aprendizado, sendo que 8 dos 29 parâmetros originais das CA foram relevantes; tal subconjunto de parâmetros elevou a qualidade do agrupamento de 0,476 a 0,616 medido pelo SI.

O objetivo (*ii*) foi atingido no segundo artigo, que em sua estrutura propôs o desenvolvimento de um novo índice de importância de parâmetros (IIP) com o propósito de identificar os parâmetros de CAs mais relevantes com o propósito de classificar trabalhadores

de acordo com seus padrões de aprendizado. Modelagem de CAs e técnicas multivariadas integraram as etapas operacionais do método proposto. Os parâmetros de CAs mais relevantes, para classificação dos trabalhadores, foram obtidos através do IIP proposto, o qual foi gerado através da integração dos coeficientes das regressões PLS e LASSO. KNN, NB e SVM foram as ferramentas de classificação utilizadas. Aplicado a dados de uma indústria, o método apresentou acurácia de 100% com as três ferramentas de classificação, utilizando apenas um subconjunto de 3 dos 29 parâmetros originais.

O objetivo (*iii*) foi atingido no terceiro artigo, onde foi desenvolvida uma abordagem multivariada para identificar os fatores mais influentes em três possíveis desfechos (diplomação, evasão interna ou evasão externa) de alunos de graduação. A abordagem proposta aplicou a sistemática "omita uma variável por vez" (OUVV) em conjunto com uma ferramenta de classificação, de modo a identificar as variáveis mais influentes para o desfecho dos alunos. Cinco ferramentas de classificação foram utilizadas para efeito de comparação dos resultados obtidos: KNN, PNN, LDA, SVM e NB. Quatro conjuntos de dados foram utilizados para representar do primeiro ao quarto semestre acadêmico dos estudantes. Quando aplicado a dados reais de uma universidade, a abordagem obteve acurácia de 91,22% nas classificações, fazendo uso de apenas 22,22% das variáveis originais. Tais resultados foram obtidos com a ferramenta de classificação NB aplicada aos dados que representaram o 3º semestre acadêmico. Destaca-se o fato da maioria dos procedimentos realizados apontar as variáveis de desempenho acadêmico (aprovações e reprovações) como as mais influentes, fato que vai ao encontro de outros estudos realizados na área.

5.2 Sugestões para trabalhos futuros

Como extensões das proposições apresentadas nessa dissertação, sugerem-se as seguintes pesquisas futuras:

- Testar outras técnicas de classificação, como *Random Forest* e Redes Neurais;
- Aplicação das estruturas propostas nessa dissertação a conjuntos de dados de outros segmentos; e
- Desenvolvimento de novos índices de importância voltados a análise de variáveis em diferentes segmentos.