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

# ABORDAGENS DE SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO, PREDIÇÃO E AGRUPAMENTO DE AMOSTRAS INDUSTRIAIS

GILBERTO MÜLLER BEUREN

Porto Alegre
2019

GILBERTO MÜLLER BEUREN

# ABORDAGENS DE SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO, PREDIÇÃO E AGRUPAMENTO DE AMOSTRAS INDUSTRIAIS

Tese 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 Doutor em Engenharia, na área de concentração em Sistemas de Qualidade.

Orientador: Prof. Michel José Anzanello, PhD.

Porto Alegre

2019

GILBERTO MÜLLER BEUREN

**ABORDAGENS DE SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO, PREDIÇÃO E AGRUPAMENTO DE AMOSTRAS INDUSTRIAIS**

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia de Produção 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, PhD.**
Orientador PPGEP/UFRGS

_____

**Prof. Alejandro Germán Frank, Dr.**
Coordenador PPGEP/UFRGS

**Banca Examinadora:**

Professor Alessandro Kahmann, Dr. (IMEF/FURG)

Professora Liane Werner, Dra. (IME/UFRGS)

Professora Marcia Elisa Soares Echeveste, Dra. (PPGEP/UFRGS)

## RESUMO

Com os avanços tecnológicos nos mais diversos processos industriais, é cada vez mais frequente a coleta de grandes volumes de dados e seu armazenamento com vistas ao monitoramento de tais processos. Entretanto, a análise precisa das informações coletadas pode ser comprometida pelo volume excessivo de variáveis, provocando ruído e distorções nos resultados. Neste contexto, a seleção de variáveis consideradas mais importantes para a correta interpretação dos dados surge como uma alternativa para a identificação de padrões com propósitos que incluem classificação, predição e agrupamento de amostras, removendo aquelas que apresentam ruídos ou alta correlação. Dentro do escopo desta tese, a seleção de variáveis tem por objetivo criar modelos inovadores que se adaptem aos mais variados tipos de objetivos de classificação, predição e agrupamento, reduzindo o número de variáveis irrelevantes, ruidosas e redundantes, bem como apresentando maior eficiência computacional para a realização das análises. Tais metodologias são apresentadas em três artigos, visando a resolução de problemas específicos. No primeiro artigo, um índice de importância de variáveis é apresentado para selecionar as variáveis mais relevantes na construção de um modelo de predição, através de Informação Mútua; no segundo artigo, uma nova metodologia apoiada em duas fases para identificar as variáveis mais relevantes ao agrupamento de amostras de medicamentos similares quanto a aspectos químicos é proposta; no terceiro artigo, uma abordagem para seleção das variáveis mais informativas para classificação de bateladas produtivas em sete bancos de dados supervisionados é proposta através de três testes não-paramétricos. A aplicação dos métodos em distintos bancos de dados industriais, sua validação e comparação com abordagens da literatura corroboram a adequabilidade das proposições desta tese.



**Palavras-chave:** Seleção de variáveis. Classificação. Agrupamento. Predição. Dados industriais.

## ABSTRACT

The recent developments in technology area allowed the collection of larger amounts of data and its storage in industrial sector. However, the excessive number of variables, which generate results comprised of noise and distortion, may compromise the correct analysis of such information. In this context, the selection of most informative variables to analyze data precisely emerges as an alternative to pattern identification with purposes that include classification, prediction and clustering of samples, removing noisy and high collinear features. Within the scope of the thesis, variable selection has the objective to create groundbreaking models that can adapt to a large variety of model classification, prediction and clustering, reducing the number of irrelevant, noisy and redundant features, as well as presenting a higher computational efficiency in the data analysis. Such methodologies are presented in three scientific articles, aiming the solution of specific problems: the first one presents a variable importance index to select the most relevant features to build a prediction model, through Mutual Information; the second article proposes a new framework to identify the most informative variables to cluster similar medicine samples based on their chemical aspects, where a new feature selection in two phases is conducted; the third article proposes a new methodology to select the most important variables through three non-parametric tests to classify production batches in seven supervised datasets. The application of the framework in different industrial datasets as well as the validation and comparison with other studies and methodologies corroborates the suitability of the thesis' propositions.



**Keywords:** Variable selection. Classification. Clustering. Prediction. Industrial data.

**SUMÁRIO**

## 1. INTRODUÇÃO

A coleta de grandes volumes de dados vem se tornando uma prática cada vez mais frequente devido aos avanços na área de tecnologia que permitem a análise e o armazenamento destes. Estes dados possibilitam diversos tipos de análise, como classificação de amostras produtivas, predição de propriedades de itens e identificação de padrões nos mais variados processos produtivos. Entretanto, a análise mais minuciosa e precisa desses dados pode ser comprometida pelo volume excessivo de variáveis, provocando ruído e distorções nos resultados. Além disto, grande parte das ferramentas paramétricas de análise multivariada tem a sua eficiência comprometida quando os dados apresentam elevada multicolinearidade, o que é frequentemente verificado em bancos com muitas dimensões [1].

Nos últimos anos, um grande número de estudos e abordagens têm sido propostos com o objetivo de adequadamente analisar bancos com elevada dimensionalidade, a fim de permitir a melhor interpretação das informações disponíveis. Estes estudos englobam as mais variadas áreas do conhecimento, como, por exemplo, medicina [2], química [3,4], farmácia [5] e biologia [6]. A presença de variáveis com características desfavoráveis à aplicação de diversas técnicas de análise e monitoramento reforça a necessidade do uso de mineração de dados para identificação precisa de padrões presentes nestes dados [7].

Com o objetivo de identificar padrões em grandes bancos de dados com propósitos que incluem classificação, predição ou clusterização, uma vasta gama de abordagens é oferecida pela literatura. Dentre estas abordagens destaca-se a seleção de variáveis, que tem como principal objetivo a remoção de variáveis ruidosas ou altamente correlacionadas, retendo apenas aquelas consideradas mais importantes para a correta interpretação dos dados. Tal curso de ação permite um aumento da acurácia de técnicas estatísticas com um menor custo computacional na geração de modelos preditivos e classificatórios. Diversos estudos [8-10] apontam outros benefícios da seleção de variáveis, que incluem: *(i)* reduzir o grau de subjetividade na identificação das variáveis mais importantes para o processo, minimizando a influência do conhecimento empírico de especialistas; *(ii)* aumentar o conhecimento a respeito do processo e das variáveis que mais o impactam; *(iii)* evitar modelos com sobreajuste (*overfitting*); e *(iv)* gerar modelos com melhor relação custo-benefício de processamento.

Operacionalmente, percebe-se que grande volume de estudos focados na seleção de variáveis para fins de classificação ou predição de bancos supervisionados (variável resposta conhecida) usualmente apoiam-se em ferramentas de estatística paramétrica, como Teste Qui-Quadrado [11], Análise de Componentes Principais [12] e Programação Quadrática [13] com

vistas à definição das variáveis mais informativas. Tais técnicas, segundo Rönkkö *et al.* [14], podem ter a sua capacidade preditiva substancialmente reduzida quando aplicadas a dados ruidosos, criando vieses. Desta forma, existe espaço para o desenvolvimento de novas metodologias que envolvam a utilização de ferramentas não-paramétricas que se apresentam como boas alternativas para seleção de variáveis, especialmente para o desenvolvimento de índices de importância [15]. As análises não-paramétricas apresentam diversas vantagens com relação a seus pares paramétricos, como: *(i)* são menos afetadas por valores extremos ou atípicos, *(ii)* são mais estáveis quando o número de observações coletadas é baixo em comparação com o número de variáveis e *(iii)* são mais adequadas quando aplicadas a dados que não seguem uma distribuição de probabilidade conhecida [16].

Além disto, quando se tem por objetivo o agrupamento de observações em dados não-supervisionados (variável resposta não conhecida), a vasta maioria das abordagens de seleção das variáveis mais relevantes conta com apenas uma fase, sem fazer uma eliminação prévia daquelas altamente correlacionadas. Tal curso de ação, segundo Xiaobo *et al.* [17], pode fazer com que informações relevantes sejam suprimidas e até mesmo criar vieses na seleção das variáveis. Assim, percebe-se a possibilidade de desenvolvimento de metodologias alternativas, onde a utilização de duas fases para seleção de variáveis é adotada. A utilização de uma fase preliminar de seleção traz algumas vantagens, como: *(i)* excluir preliminarmente variáveis ruidosas ou redundantes; *(ii)* reduzir o esforço computacional da próxima etapa de seleção e *(iii)* evitar vieses de seleção [18].

A seleção de variáveis, dentro do escopo desta tese, tem por objetivo criar modelos inovadores que se adaptem aos mais variados tipos de objetivos de classificação, predição e agrupamento, reduzindo o número de variáveis irrelevantes, ruidosas e correlacionadas, bem como apresentando maior eficiência computacional para a realização das análises. Isto traz como resultado a criação de modelos mais simples e fáceis de serem interpretados, reduzindo a sua complexidade e resultando em ganhos de precisão e acurácia [19].

Neste contexto, as seguintes questões norteiam esta pesquisa: *(i)* como utilizar ferramentas não-paramétricas para a criação de um índice de importância de variáveis para a classificação em um banco de dados supervisionado?; *(ii)* a seleção de variáveis realizada em duas fases pode ser mais eficiente em bancos de dados não-supervisionados para realizar o agrupamento das observações?; e *(iii)* em relação à predição, a criação de um índice de importância de variáveis, visando a sua seleção através de ferramentas multivariadas gera uma melhoria na precisão?

## 1.1 Tema e objetivos

O tema desta tese é a seleção de variáveis mais relevantes para o desenvolvimento de modelos mais robustos. O objetivo desta tese consiste na proposição de novas abordagens para a seleção de variáveis apoiadas em ferramentas multivariadas com vistas à classificação, predição e agrupamento de amostras industriais. Para tanto, os objetivos específicos, listados a seguir, são importantes para a construção deste trabalho:

a) Propor novos índices de importância de variáveis baseados em testes não-paramétricos para fins de classificação;
b) Propor novas abordagens apoiadas em duas fases (fase inicial de filtragem - *filter* - e fase secundária de avaliação das variáveis - *wrapper*) para identificação das variáveis mais informativas;
c) Propor um novo método de seleção de variáveis para predição de resultados industriais baseado em um índice de importância de variáveis;
d) Validar as abordagens propostas em bancos de dados dos tipos supervisionados e não-supervisionados coletados de processos industriais; e
e) Comparar os resultados dos métodos propostos com as metodologias existentes na literatura.

## 1.2 Justificativa

O fenômeno de coleta excessiva de informações pode ser observado nos mais variados setores da indústria, como, por exemplo, farmacêutico, alimentício e químico, onde variáveis descrevendo concentrações, temperatura e pressão são frequentes, gerando dados ruidosos, redundantes e irrelevantes [20]. Além disto, a formulação de modelos interpretáveis e que reflitam de forma acurada a realidade em que estão inseridos pode ser comprometida, dificultando assim a classificação, predição ou agrupamento de amostras de forma precisa. Desta forma, a identificação das variáveis mais relevantes torna-se crucial, tanto para a geração de modelos mais precisos e parcimoniosos quanto para uma maior eficiência computacional [21]. Tais benefícios justificam as abordagens propostas neste estudo em termos práticos.

Entretanto, percebe-se que as proposições presentes na literatura para identificar as variáveis mais informativas [22,23], tanto para bancos com dados supervisionados quanto para não-supervisionados, nem sempre apresentam modelos satisfatórios, pois os dados podem apresentar distribuições de probabilidade desconhecidas, observações extremas, valores atípicos ou alta colinearidade. Estas características são comuns em dados industriais e, assim,

as técnicas paramétricas comumente utilizadas para realizar a seleção de variáveis em dados supervisionados podem gerar índices de importância instáveis [24]. Apesar da utilização de técnicas não-paramétricas para este fim apresentar resultados mais robustos [16], geralmente aquelas são utilizadas apenas em fases de iniciais de filtragem de variáveis. Desta forma, justifica-se a elaboração de novas abordagens onde se utilize testes não-paramétricos para a criação de índices de importância de variáveis em dados supervisionados, a fim de otimizar o conjunto de dados que serão utilizadas para classificação.

Da mesma forma, bancos de dados industriais não-supervisionados apresentam multicolinearidade entre as variáveis, fazendo com que vieses na seleção possam ser criados e até mesmo dados relevantes sejam suprimidos [18]. Entretanto, as metodologias existentes na literatura para a seleção de variáveis com fins de agrupamento de amostras baseiam-se na utilização de índices de importância aplicados diretamente no banco original, sem uma filtragem inicial de variáveis altamente correlacionadas ou ruidosas. Assim, apresenta-se também a necessidade do desenvolvimento de metodologias alternativas em bancos de dados supervisionados que selecionem variáveis em duas fases, onde uma etapa preliminar de filtragem é adotada.

Nesta mesma direção, apesar de estudos indicando a importância e eficiência da utilização de índices de importância de variáveis para criação de modelos de predição [25], a utilização de Informação Mútua para geração destes índices não é frequente, apesar de apresentar resultados interessantes em termos de redução da dimensionalidade dos dados e melhoramento da acurácia [26]. Assim, também existe a possibilidade de se testar abordagens de seleção de variáveis para a predição de respostas em bancos de dados industriais utilizando-se a Informação Mútua.

Portanto, é possível perceber no âmbito acadêmico a necessidade do desenvolvimento de novos índices de seleção de variáveis que levem em consideração estas características presentes em dados industriais, justificando de forma teórica o tema e os objetivos desta tese.

## 1.3 Delineamento do estudo

Esta seção apresenta o enquadramento da pesquisa do ponto de vista metodológico, alinhado aos objetivos e justificativa desta tese. Além disto, traz o método de trabalho idealizado para alcançar os objetivos propostos e responder as questões de pesquisas levantadas.

### *1.3.1 Método de Pesquisa*

Do ponto de vista da natureza, esta pesquisa é classificada com aplicada, visto que visa resolver problemas práticos propondo novas metodologias [27]. Já do ponto de vista de abordagem, classifica-se como quantitativa, uma vez que relações de causa e efeito são explicadas por métodos lógico-dedutivos, permitindo a sua replicação por meio da generalização [28]. Em relação aos seus objetivos, esta pesquisa é classificada como exploratória, visto que tem por finalidade desenvolver, esclarecer e modificar ideias e conceitos [29].

### *1.3.2 Método de Trabalho*

A pesquisa é realizada a partir de três artigos que visam atender os objetivos específicos supracitados, abordando de diferentes formas de seleção de variáveis, apresentando contribuições inéditas para a literatura acadêmica. Cada um dos artigos representa um dos capítulos subsequentes à introdução na presente tese. Abaixo são descritos brevemente cada um dos três artigos e, na Tabela 1.1 são apresentadas as ferramentas utilizadas e contribuição científica de cada um deles. É importante ressaltar que os artigos são apresentados no formato de submissão realizada aos periódicos internacionais e, desta forma, encontram-se escritos em língua inglesa.

O Artigo 1 – *Ranking-based variable selection using Mutual Information for enhancing the prediction of industrial batch properties* (Seleção de variáveis baseada em ranqueamento utilizando Informação Mútua para aprimorar a previsão de propriedades de bateladas industriais) – objetiva propor uma metodologia para o desenvolvimento de um índice de importância de variáveis onde as mais relevantes são selecionadas para prever propriedades de produtos industriais. A abordagem segue as seguintes etapas: *(i)* divisão do banco de dados em treino e teste utilizando o Algoritmo de Kennard Stone; *(ii)* ordenamento das variáveis de acordo com a sua Informação Mútua com a variável resposta; e *(iii)* predição iterativa das amostras de acordo com a classificação gerada na etapa anterior e verificação da eficiência desta predição através do Erro Percentual Absoluto Médio (EPAM). A metodologia proposta foi aplicada a três bancos de dados industriais oriundos de processos distintos, e o EPAM da predição foi comparada com a obtida em outras abordagens da literatura. Em média, a qualidade da predição foi incrementada em 76,15% ao utilizar-se 79,19% menos variáveis.

O Artigo 2 – *Improving the clustering quality of medicine samples towards a two-phase variable selection approach* (Aprimorando a qualidade de agrupamento de amostras de

medicamentos por meio de uma abordagem de seleção de variáveis em duas etapas) – propõe uma nova abordagem para identificar as variáveis mais relevantes para agrupar amostras de medicamentos similares quanto a aspectos químicos. Esta seleção é particularmente importante para auxiliar as autoridades a identificarem medicamentos com padrões químicos fora dos definidos por órgãos reguladores, resultando assim em falsificações e em risco para a saúde pública. O método baseia-se em uma seleção de variáveis em duas fases. Na primeira fase, variáveis que apresentam alta similaridade (e potencial redundância) são eliminadas através de Informação Mútua; na segunda fase, as variáveis remanescentes são ranqueadas através de índice de importância de variáveis baseado em parâmetros oriundos da Análise Fatorial. Quando aplicado a um conjunto de dados composto de 1016 amostras descritas por 1803 variáveis, a estrutura proposta reteve 241 variáveis (13,37% dos dados originais) responsáveis pelo melhor agrupamento dos medicamentos. Cinco grupos foram formados e avaliados qualitativamente em relação às suas similaridades.

O Artigo 3 – *Variable selection using statistical non-parametric tests for classifying production batches into multiple classes* (Seleção de variáveis utilizando testes estatísticos não-paramétricos para classificar bateladas produtivas em múltiplas classes) – objetiva identificar as variáveis mais informativas para realizar uma classificação de bateladas produtivas conforme classes pré-estabelecidas. A abordagem apoia-se nas seguintes etapas: *(i)* eliminação de variáveis redundantes através de Informação Mútua; *(ii)* separação do banco de dados em treino e teste com o Algoritmo de Kennard Stone; *(iii)* classificação das variáveis por ordem de importância através de três diferentes testes não-paramétricos e; *(iv)* classificação das amostras de acordo com o ranqueamento gerado na etapa anterior e verificação da acurácia de classificação das bateladas. A abordagem proposta foi aplicada a sete bancos de dados industriais e a acurácia da classificação foi comparada com a obtida por outras abordagens da literatura. Em média, a estrutura proposta resultou em classificações 17,04% mais precisas utilizando 78,65% menos variáveis do que os métodos de seleção tradicionais.

Tabela 1.1: Descrição dos artigos da tese

| Artigo | Título | Ferramentas utilizadas | Contribuição científica |
|---|---|---|---|
| **1** | Ranking-based variable selection using Mutual Information for enhancing the prediction of industrial batch properties | Informação Mútua, Algoritmo de Kennard Stone, Análise de Regressão, Erro Percentual Absoluto Médio | Proposição de um novo método de seleção de variáveis para a predição de dados industriais baseado em um índice de importância de variáveis |
| **2** | Improving the clustering quality of medicine samples towards a two-phase variable selection approach | Informação Mútua, Análise Fatorial, Análise de Cluster, *k*-Médias, Índice de Silhueta | Proposição de um novo método de seleção de variáveis para o agrupamento de amostras de medicamentos baseada em duas etapas |
| **3** | Variable selection using statistical non-parametric tests for classifying production batches into multiple classes | Informação Mútua, Algoritmo de Kennard Stone, Teste de Anderson-Darling, Teste de Kruskal-Wallis, Teste de Steel, *k*-Vizinhos Próximos, Máquina de Vetores de Suporte | Proposição de um novo método de seleção de variáveis para a classificação de amostras industriais baseada em testes estatísticos não-paramétricos |

## 1.4 Delimitações

A presente pesquisa se concentra apenas na utilização de ferramentas já consolidadas na área de estatística multivariada e bancos de dados da área industrial e farmacêutica, comparando seus resultados com outras técnicas presentes na literatura. Desta forma, esta pesquisa não considerou:

a) Técnicas de análise multivariada inovadoras;
b) Bancos de dados públicos de outras áreas do conhecimento;
c) Métodos alternativos de inclusão/exclusão de variáveis que não envolvam a utilização ordenada de uma variável por vez;
d) Análises de ganho financeiro por conta da redução do número de variáveis a serem coletadas e analisadas; e
e) Avaliação do desempenho dos modelos por mais do que uma métrica (por exemplo, somente acurácia é considerada em problemas de classificação).

## 1.5 Estrutura da Tese

Esta tese está organizada em cinco capítulos. O primeiro apresenta a contextualização do problema de pesquisa, o seu tema, objetivos e justificativas, além do delineamento do estudo, delimitações e estrutura. Os capítulos 2, 3 e 4 apresentam os artigos 1, 2 e 3, respectivamente. O capítulo 5 apresenta as considerações finais, conclusões da tese e sugestões para trabalhos futuros.

## 1.6 Referências

[1] Liu, H., Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491–502. doi: 10.1109/TKDE.2005.66.

[2] Wang, H., Huang, L., Jing, R., Yang, Y., Liu, K., Li, M., Wen, Z. (2015). Identifying oncogenes as features for clinical cancer prognosis by Bayesian nonparametric variable selection algorithm. *Chemometrics and Intelligent Laboratory Systems, 146*, 464-471. doi: 10.1016/j.chemolab.2015.07.004.

[3] de Figueiredo, M., Cordella, C. B. Y., Bouveresse, D. J., Archer, X., Bégué. J., Rutledge, D. N. (2018). A variable selection method for multiclass classification problems using two-class ROC analysis. *Chemometrics and Intelligent Laboratory Systems*, *177*, 35-46. doi: 10.1016/j.chemolab.2018.04.005.

[4] Gu, Z., Deun, K. V. (2016). A variable selection method for simultaneous component based data integration. *Chemometrics and Intelligent Laboratory Systems*, *158*, 187-199. doi: 10.1016/j.chemolab.2016.07.013.

[5] Kahmann, A., Anzanello, M. J., Fogliatto, F. S., Marcelo, M. C. A., Ferrão, M. F., Ortiz, R. S., Mariotti, K. C. (2018). Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. *Journal of Pharmaceutical and Biomedical Analysis*, *152*, 120-127. doi: 10.1016/j.jpba.2018.01.050.

[6] Chiang, L. H., Pell, R. J. (2004). Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control, 14*(2), 143–155. doi: 10.1016/S0959-1524(03)00029-5.

[7] Maione, C., Lemos, B., Dobal, A., Barbosa, F., Melgaço, R. (2016). Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. *Computers and Electronics in Agriculture, 121*, 101–107. doi: 10.1016/j.compag.2015.11.009

[8] Guyon, I., Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research, 3*(3), 1157–1182. doi: 10.1023/A:1012487302797.

[9] Kettaneh, N., Berglund, A., Wold, S. (2005). PCA and PLS with very large data sets. *Computational Statistics and Data Analysis, 48*(1), 69–85. doi: 10.1016/j.csda.2003.11.027.

[10] Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517. doi: 10.1093/bioinformatics/btm344.

[11] Austin, P. C., Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology, 57*, 1138-1146. doi: 10.1016/j.jclinepi.2004.04.003.

[12] Gu, Z., Deun, K. V. (2016). A variable selection method for simultaneous component based data integration. *Chemometrics and Intelligent Laboratory Systems*, *158*, 187-199. doi: 10.1016/j.chemolab.2016.07.013.

[13] Kahmann, A., Anzanello, M. J., Fogliatto, F. S., Marcelo, M. C. A., Ferrão, M. F., Ortiz, R. S., Mariotti, K. C. (2018). Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. *Journal of Pharmaceutical and Biomedical Analysis*, *152*, 120-127. doi: 10.1016/j.jpba.2018.01.050.

[14] Rönkkö, M., McIntosh, C. N., Antonakis, J. (2015). On the adoption of partial least squares in psychological research: Caveat emptor. *Personality and Individual Differences, 87*, 76-84. doi: 10.1016/j.paid.2015.07.019.

[15] Hettmansperger, T. P., McKean, J. W., Sheather, S. J. (2000). Robust nonparametric methods. *Journal of the American Statistical Association, 95*(452), 1308-1312. doi: 10.2307/2669777.

[16] Walsh, J. E. (1962). *Handbook of nonparametric statistics*. Van Nostrand: New York.

[17] Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M., Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta, 667*(1-2), 14-32. doi: 10.1016/j.aca.2010.03.048.

[18] Yu, L., Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research, 5*, 1205-1224.

[19] Chen, M., Khare, S., Huang, B., Zhang, H., Lau, E., Feng, E. (2013). Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. *Industrial and Engineering Chemistry Research, 52*(23), 7886-7895. doi: 10.1021/ie4008248.

[20] Chong, I. G., Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems, 78*(1), 103–112. doi: 10.1016/j.chemolab.2004.12.011.

[21] He, K., Cheng, H., Du, W., Qian, F. (2014). Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. *Chemometrics and Intelligent Laboratory Systems, 134*, 79–88. doi: 10.1016/j.chemolab.2014.03.007.

[22] Anzanello, M. J., Fogliatto, F. S., Rossini, K. (2011). Data mining-based method for identifying discriminant attributes in sensory profiling. *Food Quality and Preference, 22*(1), 139–148. doi: 10.1016/j.foodqual.2010.08.010.

[23] Anzanello, M., Fogliatto, F., Marcelo, M. C. A., Pozebon, D., Ferrão, M. F. (2016). Wavelength selection framework for classifying food and pharmaceutical samples into multiple classes. *Journal of Chemometrics, 30*(6), 346–353. doi: 10.1002/cem.2799.

[24] Anzanello, M. J., Albin, S. L., Chaovalitwongse, W. A. (2009). Selecting the best variables for classifying production batches into two quality levels. *Chemometrics and Intelligent Laboratory Systems, 97*(2), 111-117. doi: 10.1016/j.chemolab.2009.03.004.

[25] Nathans, L. I., Oswald, F. L., Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation, 17*(9).

[26] Sotoca, J. M., Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition, 43*(6), 2068-2081. doi: 10.1016/j.patcog.2009.12.013.

[27] da Silva, E. L., Menezes, E. M. (2005). *Metodologia da Pesquisa e Elaboração de Dissertação*. 4. ed. Universidade Federal de Santa Catarina – UFSC: Florianópolis.

[28] Berto, R. M. V. S., Nakano, D. N. (1999). A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. *Production, 9*(2), 65-75. doi: 10.1590/S0103-65131999000200005.

[29] Gil, A. C. (2008). *Métodos e técnicas de pesquisa social*. 6. ed. Editora Atlas S.A.: São Paulo.

# 5. CONSIDERAÇÕES FINAIS

Este capítulo apresenta as discussões finais sobre o trabalho apresentado como tese de doutorado. A discussão apresentada neste capítulo divide-se em três partes: *(i)* uma discussão sobre os resultados obtidos nos três artigos; *(ii)* a apresentação das limitações presentes nesta pesquisa; e *(iii)* sugestões para futuros trabalhos feitos nesta área.

## 5.1 Conclusões

A pesquisa apresentada nesta tese foi originada através da constatação de que, com o desenvolvimento de novas tecnologias e capacidade de armazenamento, um número substancialmente maior de informações vem sendo coletada no ambiente industrial, gerando dados ruidosos e altamente correlacionados. Assim, surge a necessidade do desenvolvimento de técnicas que permitam, através de ferramentas estatísticas, selecionar variáveis relevantes para a classificação, predição e agrupamento destas amostras industriais, resultando em modelos mais precisos e parcimoniosos.

Para suprir esta necessidade, esta tese apresenta como objetivo principal a proposição de novas metodologias para a seleção de variáveis baseadas em ferramentas de estatística multivariada em bancos de dados industriais supervisionados e não-supervisionados, com a aplicação em dados coletados de processos industriais. Para tanto, o trabalho foi dividido em três artigos, a fim de realizar esta seleção com fins de classificação, agrupamento e predição, respondendo as seguintes questões de pesquisa: *(i)* como utilizar ferramentas não-paramétricas para a criação de um índice de importância de variáveis para a classificação em um banco de dados supervisionado?; *(ii)* a seleção de variáveis realizada em duas fases pode ser mais eficiente em bancos de dados não-supervisionados para realizar o agrupamento das observações?; e *(iii)* em relação à predição, a criação de um índice de importância de variáveis, visando a sua seleção através de ferramentas multivariadas gera uma melhoria na precisão?

No primeiro artigo foi sugerida uma nova estrutura para seleção de variáveis, através de Informação Mútua, com o intuito de selecionar as variáveis que mais impactam na predição de propriedades químicas verificadas no processamento de bateladas industriais. Os resultados encontrados indicam uma grande melhoria nos modelos de predição, com uma redução no Erro Percentual Absoluto Médio (EPAM) observada em todos os três bancos de dados analisados, quando comparados com a utilização de todas variáveis para compor o modelo: de 38,8% para 2,4%, de 49,2% para 5,4% e de 33,5% para 3,8%, utilizando-se, respectivamente, apenas 8%,

8,55% e 7,29% das variáveis originais. A metodologia proposta também apresentou melhores resultados em comparação com outras técnicas presentes na literatura. Estes resultados corroboram a hipótese de que o uso de índices de importância de variáveis robustos para a seleção de variáveis em cenários onde há uma variável resposta contínua gera uma melhoria na precisão do modelo de predição.

Também foi proposto e validado nesta tese, no artigo 2, um novo índice de seleção de variáveis baseado em duas etapas, utilizando-se ferramentas multivariadas para agrupar amostras de medicamentos em um banco de dados não-supervisionado. A eficiência desta metodologia foi verificada através do aumento de 0,07 para 0,81 no Índice de Silhueta no agrupamento realizado dos medicamentos, utilizando-se apenas 13,37% das variáveis originais. Além disto, a metodologia proposta apresentou desempenho superior às demais estruturas existentes na literatura. Estes resultados corroboram a eficácia do uso de uma fase preliminar de seleção de variáveis correlacionadas em bancos de dados não-supervisionados. Além disto, a metodologia empregada auxiliou na detecção de medicamentos potencialmente falsificados, através do agrupamento gerado após a realização da seleção das variáveis mais informativas.

Por fim, no artigo 3 foi apresentada e validada (em sete bancos de dados industriais supervisionados) uma metodologia inovadora de seleção de variáveis que utiliza testes estatísticos não-paramétricos para ranquear as variáveis de acordo com a sua relevância para a classificação de bateladas produtivas. Esta metodologia proposta apresentou, em média classificações 17,04% mais precisas utilizando 78,65% menos variáveis do que a metodologia tradicional de seleção *stepwise*. Também superou outras abordagens propostas na literatura que utilizaram os mesmos bancos de dados. Estes resultados indicam a relevância do uso de testes não-paramétricos para o ranqueamento das variáveis mais importantes em dados supervisionados que tipicamente apresentam distribuições não parametrizadas e muticolinearidade entre as suas variáveis.

Assim, para o cumprimento do objetivo desta tese, os objetivos específicos definidos foram executados ao longo dos três artigos: foi proposto um novo índice de importância de variáveis baseado em testes não-paramétricos no artigo 3; foi proposta uma nova abordagem apoiada em duas fases para identificação das variáveis mais informativas no artigo 2 e; foi proposto um novo método de seleção de variáveis para predição de resultados industriais baseado em um índice de importância de variáveis no artigo 1. Os outros dois objetivos específicos propostos foram atingidos ao longo dos três artigos, ao validar cada uma das

abordagens propostas em bancos de dados coletados de processos industriais e comparar os resultados encontrados com outras metodologias existentes na literatura.

Com base nestas contribuições e resultados apresentados, conclui-se que todos os objetivos específicos propostos neste trabalho foram alcançados, permitindo afirmar que o objetivo principal da tese foi obtido. Uma vez satisfeitos estes objetivos, pode-se concluir também que a pesquisa conseguiu satisfatoriamente suprir as necessidades constatadas e que todas as três etapas da tese apresentaram contribuições complementares para este fim.

## 5.2 Limitações

As limitações desta pesquisa devem ser ressaltadas, a fim de permitir a elaboração de futuras complementações. Com relação aos modelos utilizados para predição, apenas a Regressão Linear Múltipla foi utilizada, não considerando outros tipos de regressão, como por exemplo a não-linear, a não-paramétrica ou a Bayesiana. Além disto, a metodologia proposta no artigo 2 foi aplicada apenas em um banco de dados e as comparações com outros estudos foram também realizadas apenas neste banco.

Da mesma forma, o tamanho dos bancos de dados utilizados nos artigos 1 e 3 também mostra algumas limitações, visto que nenhum deles apresenta mais do que 300 observações ou 3 classes. A utilização destes dados, porém, foi escolhida visto que estas informações já estão presentes em demais pesquisas na temática de seleção de variáveis, permitindo uma comparação mais ampla com o que já havia sido realizado nesta área.

Por fim, todas as avaliações do desempenho dos modelos apresentados nesta pesquisa foram realizadas por apenas uma métrica (por exemplo, somente Índice de Silhueta é considerado em problemas de agrupamento), impondo assim restrições com relação a sua generalização, uma vez que os resultados podem variar a medida em que se altera a métrica utilizada.

## 5.3 Sugestões para trabalhos futuros

Como sugestões para possíveis extensões do estudo apresentado neste trabalho, seguem as frentes para pesquisas futuras nesta área:

i. Desenvolvimento de abordagens de seleção de variáveis utilizando técnicas de análise multivariada inovadoras e não existentes na literatura;

ii. Proposição de abordagens que não utilizem a inclusão/exclusão de variáveis de forma ordenada (uma por vez);
iii. Desenvolvimento de índices de importância de variáveis que envolvam também aspectos qualitativos, como opinião de especialistas.